Powder-Diffraction-Based Structural Comparison for Crystal Structure Prediction without Prior Indexing

Alberto Otero-de-la-Roza [a]*

[a] *Departamento de Química Física y Analítica and MALTA Consolider Team, Facultad de Química, Universidad de Oviedo, 33006 Oviedo, Spain.*

E-mail: oteroalberto@uniovi.es

Abstract

The objective of crystal structure prediction (CSP) is to predict computationally the thermodynamically stable crystal structure of a compound from its stoichiometry or its molecular diagram. Crystal similarity indices measure the degree of similarity between two crystal structures, and are essential in CSP because they are used to identify duplicates. In addition, powder-based indices, which are based on comparing X-ray diffraction patterns, also allow the use of experimental X-ray powder diffraction data to inform the CSP search. Powder-assisted CSP presents two unique difficulties: i) the experimental and computational structures are not entirely comparable because the former is subject to thermal expansion from lattice vibrations, and ii) experimental patterns present features (noise, background contribution, varying peak shapes, etc.) that are not easily predictable computationally. In this work, we present a powder-based similarity index (GPWDF) based on a modification of de Gelder et al.'s index using cross-correlation functions that can be calculated analytically. Based on

GPWDF, we also propose a variable-cell similarity index (VC-GPWDF) that assigns a high similarity score to structures that differ only by a lattice deformation and takes advantage of the analytical derivatives of GPWDF with respect to the lattice parameters. VC-GPWDF can be used to identify similarity between: two computational structures generated using different methods, a computational and a experimental structure, and two experimental structures measured under different conditions (e.g. different temperature and pressure). In addition, VC-GPWDF can also be used to compare crystal structures with experimental patterns in combination with an automatic pre-processing step. The proposed similarity indices are simple, efficient, and fully automatic. They require no indexing of the experimental pattern or a guess of the space group, account for deformations caused by varying experimental conditions, give meaningful results even when the experimental pattern is of very poor quality, and have a cost does not increase with the flexibility of the molecular motif.

1. Introduction

The purpose of crystal structure prediction (CSP) is the computational prediction of the crystal structure of a compound from its stoichiometry or its molecular diagram (Price *et al.*, 2016; Price, 2018; Bowskill *et al.*, 2021; Lommerse *et al.*, 2000; Motherwell *et al.*, 2002; Day *et al.*, 2005; Day *et al.*, 2009; Bardwell *et al.*, 2011; Reilly *et al.*, 2016). In a CSP protocol, multiple candidate structures are generated randomly, and subsequently ranked by energy, with the purpose of identifying the thermodynamically stable phase and metastable phases likely to be observed (Nyman & Day, 2015). In the more sophisticated (and successful) CSP protocols, the final ranking function is based on dispersion-corrected density-functional theory (DFT), perhaps in combination with a method to estimate the vibrational contribution to the free energy (Bardwell *et al.*, 2011; Reilly *et al.*, 2016; Whittleton *et al.*, 2017*a*; Whittleton *et al.*, 2017*b*). (The latter

is used mostly in molecular CSP.) CSP is a global optimization problem and, due to its complex nature, even the most successful CSP protocols often fail to predict the experimentally observed phases (Reilly *et al.*, 2016). In this context, any additional information about the experimental crystal structure is invaluable to help restrict the search.

One of the most common experimental observables are X-ray powder diffraction (XRPD) patterns (Pecharsky & Zavalij, 2005), due to the relative easiness with which they can be obtained. Powder diffraction data contains structural information and, if the pattern is of very high quality (e.g. obtained at a synchrotron facility and free from defects) a structure can be derived from it. This is the objective in "structure determination from powder data (SDPD)" (David *et al.*, 2006; Padgett *et al.*, 2007; David & Shankland, 2008; Brüning & Schmidt, 2015; Gao *et al.*, 2017; Habermehl *et al.*, 2014; Schlesinger *et al.*, 2021; Habermehl *et al.*, 2022; Altomare, 2022). A successful application of SDPD requires expert knowledge from the user as well as a high quality pattern. In contrast, low-quality patterns (with broadened peaks because of poor crystallinity, preferred orientation effects, etc.) are common in actual practice. The ultimate goal of this work is to provide a simple way of extracting information from XRPD patterns—perhaps very low quality patterns—to help restrict the search for candidate structures in an already-existing CSP method (Schmidt *et al.*, 2005; Price *et al.*, 2016). A reliable powder-assisted CSP protocol would be invaluable for the structural determination of high-pressure phases due to the low quality of XRPD patterns measured in diamond-anvil cells and the difficulty in carrying out single-crystal diffraction experiments under such conditions. Examples abound of high-pressure mineral phases for which the XRPD pattern has been measured but the structure has not been solved (Chuliá-Jordan *et al.*, 2020; Chuliá-Jordán *et al.*, 2021; Santamaría-Pérez *et al.*, 2024).

A similarity index is a quantitative measure of the similarity between two crystal structures. A CSP protocol uses similarity indices to identify duplicate structures (Wei *et al.*, 2024), but they are also useful in other contexts such as classification and database searching (de Gelder *et al.*, 2001; Van De Streek & Motherwell, 2005; Van De Streek, 2006; Sacchi *et al.*, 2020; Özer *et al.*, 2022). Similarity indices based on the comparison of powder diffraction patterns (in the following, "powder-based indices") are attractive because they can be used for both duplicate identification and to compare the candidate structures with experimental XRPD patterns. A further advantage of powder-based indices is that they identify structurally related compounds (conformational phases, isomorphous systems) as similar (de Gelder *et al.*, 2001). De Gelder et al.'s powder-based index using cross-correlation functions (de Gelder *et al.*, 2001), based on similar previously proposed measures (Stephenson & Binsch, 1980; Karfunkel *et al.*, 1993; Lawton & Bartell, 1994), is a very popular example(Guzei *et al.*, 2011; Wood *et al.*, 2012; Nyman *et al.*, 2019; Fredericks *et al.*, 2021; Otero-de-la-Roza *et al.*, 2014; Habermehl *et al.*, 2014).

Comparing an *in silico* crystal structure with an experimental pattern using a powder-based index requires overcoming two difficulties. First, the equilibrium crystal structure from a computational method is rarely directly comparable with the experimental structure. The reason is computational methods commonly employed (like DFT or empirical potentials) do not account for lattice vibrations. Vibrational effects, both zero-point and temperature-dependent, impact the crystal geometry, typically cause a slight anisotropic expansion of the lattice (Mayo & Johnson, 2021). XRPD patterns, and therefore powder-based indices, are particularly sensitive to lattice deformations, which causes a powder-based index to identify as different structures that share motif and merely differ by a slight lattice deformation (Hofmann & Kuleshova, 2005; Van De Streek & Motherwell, 2005; Price *et al.*, 2016). Second,

experimental XRPD patterns present peak shapes determined by a variety of factors such as sample preparation and quality, experimental setup, and others, as well as experimental noise and a background contribution (Pecharsky & Zavalij, 2005). These features need to be taken into account when comparing with a pattern derived from a computed structure, for which only reflection angles and intensities are available.

In a previous work, we proposed a variable-cell powder-based index built on de Gelder et al.'s index (de Gelder *et al.*, 2001) to compare crystal structures allowing for slight lattice deformations (VC-PWDF) (Mayo *et al.*, 2022). VC-PWDF introduced the idea of a similarity index that considers two structures as exactly equal if there exists an affine transformation (translation, rotation, and anisotropic scaling) that brings both lattices and motifs into exact coincidence. VC-PWDF works by choosing the reduced (Niggli) basis of one of the structures as "target" and then exhaustively exploring all possible bases for the other ("candidate") structure. If a basis is found for the candidate structure that is similar to the target (within a certain allowance for cell angle and length deformation), the lattice parameters of the target are adopted by the candidate and de Gelder's similarity index is calculated. The VC-PWDF index is the minimum of all indices calculated in this way. In a subsequent work, Mayo et al. applied the VC-PWDF method to compare CSP-generated crystal structures with experimental patterns (VC-xPWDF)(Mayo *et al.*, 2023). However, VC-xPWDF does not account for experimental peak shapes and noise and no background correction is performed beyond subtracting a constant. More important, VC-xPWDF requires indexing the experimental pattern, which often cannot be done unequivocally (Hageman *et al.*, 2003; Hofmann & Kuleshova, 2005; Schmidt *et al.*, 2005; Habermehl *et al.*, 2014; Harris, 2022). Furthermore, even if the experimental lattice constants are known with certainty, a transformation may not exist that converts the calculated structure into the experimental lattice even if the structures are similar, which would

result in an erroneous mismatch. This problem, which affects both VC-PWDF and VC-xPWDF, can happen, for instance, if the calculated structure has lower translational symmetry than the experimental structure.

In this work, we propose a powder-based similarity index using Gaussian functions (GPWDF) built on a modification of de Gelder's cross-correlation function (de Gelder *et al.*, 2001). Instead of comparing powder diffraction patterns directly, GPWDF operates on lists of reflections and intensity pairs (θ_i, I_i). Compared to de Gelder's index, GPWDF has a closed analytical form and is therefore exactly differentiable with respect to the structural parameters of the crystals being compared, which enables the efficient optimization of the similarity index with respect to the lattice parameters. The optimization of the similarity enables distorting one of the structures to match the other, thus accounting for the mismatch caused by lattice vibrations and other effects. In addition, we define the associated variable-cell similarity index (VC-GPWDF) that uses a quasirandom exploration over lattice distortions, combined with the optimization of GPWDF with respect to the lattice parameters. VC-GPWDF can be applied to compare a calculated structure with another structure (calculated or experimental) as well as with an experimental powder pattern, and therefore can serve as a similarity index for both a normal CSP and a powder-assisted CSP method. A pre-processing step for the experimental pattern is proposed to extract the reflection angle and intensity pairs (θ_i, I_i), circumventing difficulties caused by experimental noise, peak shapes, and background contribution. Compared to similar approaches used in SDPD, VC-GPWDF works as a much simpler, standalone similarity index: it does not attempt a global fit to the experimental pattern, does not restrict the search to a given crystal symmetry and, because it searches only over lattice deformations, its cost does not scale with the complexity of the molecular motif. VC-GPWDF also circumvents the problems with VC-PWDF and VC-xPWDF mentioned above. The calculation of

VC-GPWDF is fully automatic and takes seconds to minutes on a desktop PC, which makes it a viable tool to augment an existing CSP protocol and make use of available experimental XRPD information. It is implemented in the open-source critic2 program (Otero-de-la-Roza *et al.*, 2014).

2. Gaussian Powder-Based Similarity Index

Similarity indices based on powder diffraction pattern comparisons, particularly de Gelder et al.'s index,(de Gelder *et al.*, 2001) are among the most popular in the literature for the purpose of structural comparison (Hofmann & Kuleshova, 2005; Habermehl *et al.*, 2014; Habermehl *et al.*, 2022; Mayo *et al.*, 2022; Mayo *et al.*, 2023). Indices based on real-space distances (Willighagen *et al.*, 2005; Zhu *et al.*, 2016; Rohlíček & Skořepová, 2020; Mosca & Kurlin, 2020; Terban & Billinge, 2021; Schlesinger *et al.*, 2021; Widdowson *et al.*, 2022; Widdowson & Kurlin, 2022; Chisholm & Motherwell, 2005; Nessler *et al.*, 2022). in particular the popular COMPACK method (Chisholm & Motherwell, 2005). are also widely used, despite the limitations of COMPACK regarding its behavior with respect to variations in the algorithm tolerances (Mayo *et al.*, 2022). Powder-based indices have as downside that they do not fulfill the mathematical requirements of a metric (Widdowson *et al.*, 2022; Widdowson & Kurlin, 2022). because two different crystals can share the exact same diffraction pattern (Patterson, 1939; Widdowson & Kurlin, 2022). (Schlesinger et al. recently report a case of four different structures that satisfactorily match an experimental XRPD pattern (Schlesinger *et al.*, 2022; Altomare, 2022).) Nonetheless, these coincidences are rare, and powder-based indices are the natural choice when experimental XRPD patterns for the compounds of interest are available.

In this article, our ultimate goal is the design of a powder-based index that identifies as similar crystal structures that differ by a lattice distortion. To do this, we run

a global search over the lattice parameter space in combination with local optimizations that maximize the agreement between the candidate structure and the target pattern (experimental or calculated) as a function of the distortion of the former. Similar methods have been used in the field of SDPD (Padgett *et al.*, 2007; Kariuki *et al.*, 1999; Habermehl *et al.*, 2022), and the computational cost of the global optimization has been noted as the bottleneck of the approach (Habermehl *et al.*, 2022). Therefore, for this method to be useful in CSP, where many complex structures need to be compared, it is essential that the individual local optimizations are computationally very efficient. Our first objective is to create a powder-based index that has an analytical, differentiable dependence on the structural parameters of the crystals being compared to enable efficient local optimizations.

We achieve this by modifying de Gelder's index, which for two crystals A and B is calculated as:

$$D_{\mathrm{AB}} = \frac{I_{\mathrm{AB}}}{\sqrt{I_{\mathrm{AA}} I_{\mathrm{BB}}}} \tag{1}$$

where:

$$I_{\mathrm{AB}} = I_{\mathrm{AB}}^{\mathrm{dG}} = \int w^{\mathrm{dG}}(r) S_{\mathrm{AB}}(r)\, dr \tag{2}$$

and $S_{\mathrm{AB}}(r)$ is the cross-correlation function of A and B, i.e. the overlap integral between both patterns (p_A and p_B) displaced by r:

$$S_{\mathrm{AB}}(r) = \int p_A(x) p_B(x + r)\, dx \tag{3}$$

The $w^{\mathrm{dG}}(r)$ is a weight function, symmetric around $r = 0$, whose purpose is to allow the two patterns to deviate somewhat from each other. In this way, similar patterns have a high similarity index even if the peaks are slightly shifted. In de Gelder's work a triangle function function was proposed as weight function (de Gelder *et al.*, 2001) (Figure 1):

$$w^{\mathrm{dG}}(r) = \begin{cases} 1 - \frac{|r|}{l}, & |r| < l \\ 0, & |r| \geq l \end{cases} \tag{4}$$

Compared to other previously proposed weight functions, de Gelder et al. showed by example that this choice is optimal for maximum discrimination of experimental powder patterns (de Gelder *et al.*, 2001). D_{AB} equals one if the patterns match exactly and zero if they have no overlap at all, corresponding to maximum crystal dissimilarity.

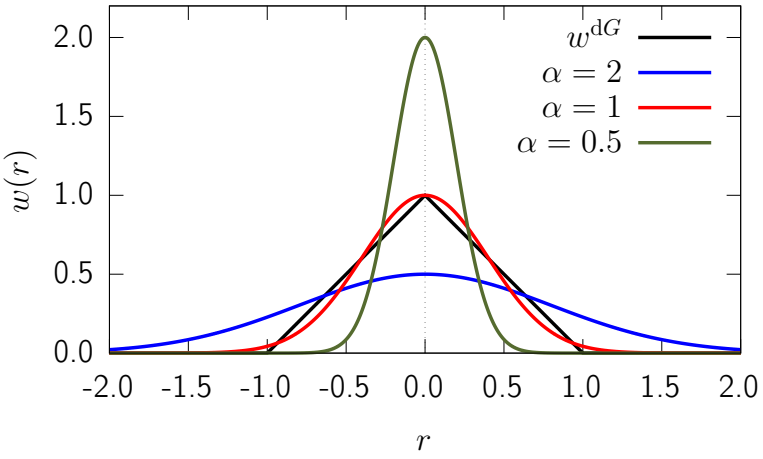

Fig. 1. De Gelder's triangle function (de Gelder *et al.*, 2001) (Eq. 4) compared with the alternative function proposed in this work (Eq. 6) for various width parameters (α).

De Gelder's index has shown to be quite robust at measuring crystal similarity (Mayo *et al.*, 2022). However, an analytical expression for $I_{\mathrm{AB}}^{\mathrm{dG}}$ and its derivatives with respect to the structural parameters of crystals A and B is not easily calculable, precluding the analytical calculation of the derivatives of D_{AB} with respect to the structural parameters, and therefore the efficient optimization of the similarity index.

To overcome these difficulties, we propose working not with the full pattern profile, but with the corresponding set of (θ_i, I_i) pairs, where I_i is the peak intensity (area) corresponding to reflection angle θ_i. If the crystal structure is known, this information is easily obtained. For an experimental pattern, we show how to obtain these quantities from the experimental data in Section 3. Our powder-based index uses de Gelder's expression (Eqs. 1 to 3) but the patterns are generated as a sum of Gaussian functions

with constant standard deviation σ:

$$p_A(x) = \sum_i \frac{I_i}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta_i)^2}{2\sigma^2}\right) \tag{5}$$

where the sum goes over all angle/intensity pairs in the considered range. In addition, de Gelder's triangle function is replaced with a normalized Gaussian function centered around zero:

$$w(r) = \frac{1}{\alpha} \exp\left(-\frac{\pi r^2}{\alpha^2}\right) \tag{6}$$

where α is a parameter that controls the width of the function, and therefore how relatively strict the index is with respect to deviations in angle or intensity (the equivalent of l in the triangle function, Eq. 4). In the limit $\alpha \to 0$, the weight function in Eq. 6 becomes a delta function and $I_{\mathrm{AB}} = S_{\mathrm{AB}}(0)$, maximizing the penalty for deviations between the two patterns. A comparison of de Gelder's triangle function and Eq. 6 for various values of α is shown in Figure 1. Note the denominator in Eq. 1 makes D_{AB} invariant with respect to a constant scaling of the intensities (de Gelder *et al.*, 2001). In the rest of this work, $\alpha = 1$ is used.

The proposed index is calculable exactly in terms of the (θ_i, I_i) pairs for the two structures. It is straightforward to show that:

$$I_{\mathrm{AB}} = \sqrt{z} \sum_i^{\mathrm{A}} \sum_j^{\mathrm{B}} I_i I_j \exp\left(-\pi z(\theta_i - \theta_j)^2\right) \tag{7}$$

where indices i and j run over the angle/intensity pairs of pattern A and B, respectively, and z is:

$$z = \frac{1}{4\pi\sigma^2 + \alpha^2} \tag{8}$$

The Gaussian powder-based similarity index (GPWDF) is defined as:

$$G_{\mathrm{AB}} = 1 - D_{\mathrm{AB}} = 1 - \frac{I_{\mathrm{AB}}}{\sqrt{I_{\mathrm{AA}} I_{\mathrm{BB}}}} \tag{9}$$

G_{AB} is zero for exactly equal structures and one for maximally dissimilar structures. (This makes G_{AB} a "dissimilarity index" instead of a "similarity index", but we will

keep using the latter term for simplicity.) G_{AB} is also readily differentiable with respect to the structural parameters of the crystals under comparison. The formulas for the derivatives of G_{AB} with respect to the lattice parameters are given in Appendix A.

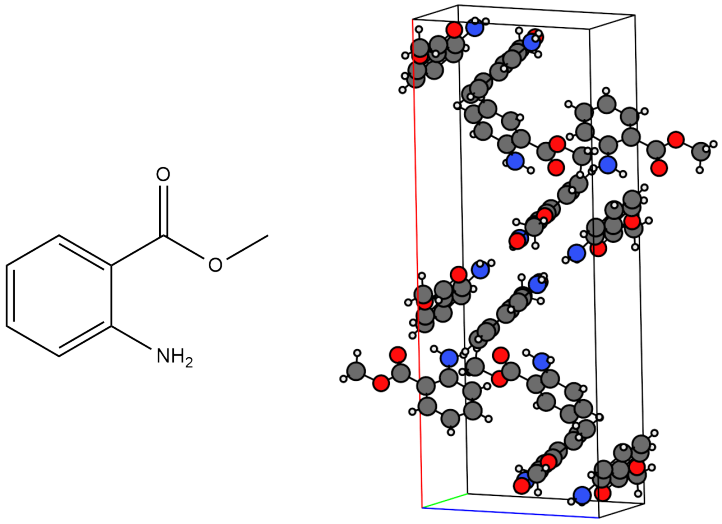

Fig. 2. Left: the molecular diagram of molecule XXIX (methyl-anthranilate) from the powder-assisted challenge of the seventh blind test. Right: a ball-and-stick representation of the experimental crystal structure.

For illustration purposes in this article, we use molecule XXIX (methyl-anthranilate) from the powder-assisted challenge within the seventh blind test run by the Cambridge Crystallographic Database Centre (CCDC) (Hunnisett & et al., 2024*a*; Hunnisett & et al., 2024*b*). The molecular diagram and unit cell for the experimental crystal structure are shown in Figure 2. The challenge consisted in predicting the crystal structure of this molecule, and the contestants were given an experimental XRPD pattern that had been deliberately modified to decrease its quality in order to simulate instances when only poor-quality powder data is available. The data for this challenge was available from the CCDC website before the publication of the blind test results (CCDC, 2024). The low-quality XRPD diffraction pattern is shown in Figure 4. For testing the new method we will use the experimental structure, as well as 310 candidate structures from our group submission. The structures were relaxed using the B86bPBE functional

(Becke, 1986; Perdew *et al.*, 1996) with the exchange-hole dipole moment (XDM) model dispersion correction (Otero-de-la-Roza & Johnson, 2012) and the Quantum ESPRESSO program; (Giannozzi *et al.*, 2017) further details can be found in the 7th blind test articles (Hunnisett & et al., 2024*a*; Hunnisett & et al., 2024*b*). Our candidate list contains the experimental structure (at the equilibrium B86bPBE-XDM geometry). Note that, while molecular crystals are used as a test case in this work, there is nothing in the presented method that prevents it from being used for other solids.

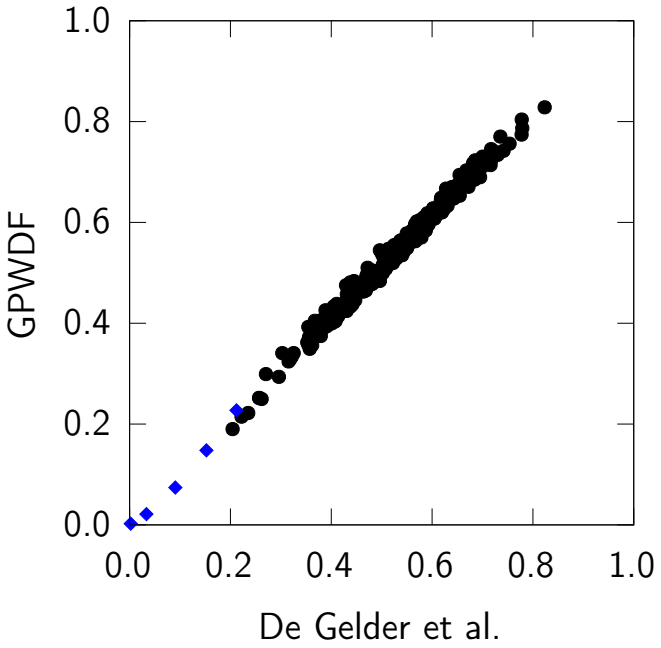

Fig. 3. Comparison between de Gelder's index and GPWDF values for the 310 candidate structures (molecule XXIX) relaxed using DFT, against the experimental structure. The blue diamonds represent an additional set of structures with the same atomic positions as the DFT-relaxed experimental structure and lattice parameters interpolated between the DFT and the experimental unit cell.

Figure 3 presents the outcome of comparing the 310 candidate crystal structures for molecule XXIX with the experimental structure, using de Gelder's index as well as GPWDF with $\alpha = 1$. There is a very strong correlation between the two, showing that

GPWDF is equivalent to de Gelder's index at measuring crystal similarity. GPWDF is calculated analytically, resulting in an average 4.5 factor speedup over de Gelder's index, as well as providing the derivatives of the index with respect to the lattice parameters of the structures under comparison.

One important thing to note about the results in Figure 3 is that the experimental structure is not very similar to the same structure after DFT relaxation, according to either of the two indices (de Gelder's= 0.27, GPWDF=0.30). This exemplifies how powder-based indices are very sensitive to changes in lattice parameters, since the calculated structure does not account for distortions caused by lattice vibrations (Hofmann & Kuleshova, 2005; Van De Streek & Motherwell, 2005). If the fractional atomic coordinates from the DFT-relaxed structure are fixed, and the lattice is continuously deformed to match experiment, the difference between the structures decreases, as shown by the blue diamonds in Figure 3. For a structure built from the DFT-relaxed atomic coordinates and the experimental cell, both indices give a difference of 0.002, a very high similarity. In agreement with our previous work (Mayo *et al.*, 2022), this suggests that a variable-cell variant of these indices could be helpful for the successful comparison between structures from different sources.

3. Pre-processing of Experimental Powder Patterns

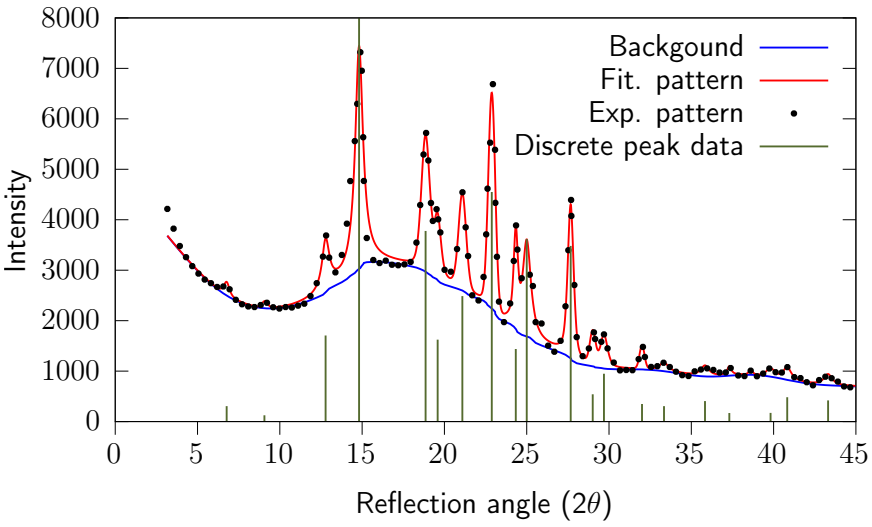

Fig. 4. Deliberately obfuscated experimental pattern for molecule XXIX (points), background determined using David and Sivia's method (David & Sivia, 2001) (blue), pattern fitted using the method proposed in this work (red), and discrete impulse representation of the fitted (θ_i, I_i) data (green).

To compare a structure with an experimental pattern, our method requires first that the experimental data is converted into a set of reflections and their corresponding intensities, (θ_i, I_i). This is achieved with a pre-processing step, in the spirit of similar approaches previously discussed in the literature (Ivanisevic *et al.*, 2005; Gao *et al.*, 2017), which has the advantage that factors such as varying peak shapes, background contribution, and experimental noise, are removed from the actual comparison. Our approach fits automatically a model pattern comprising a sum of pseudo-Voigt functions to the experimental data; these functions are commonly used to fit experimental XRPD peaks (Thompson *et al.*, 1987). Unlike whole-pattern fitting approaches like the LeBail or Pawley methods (Pecharsky & Zavalij, 2005), our algorithm does not carry out an indexing of the pattern; it merely fits a mathematical model to the data (Schreiner & Jenkins, 1982). The advantage of this approach in combination with a GPWDF similarity calculation, relative to methods that index the pattern, is that

peak overlap and small peak shifts do not have a substantial impact on the outcome (Harris, 2022). For instance, the contribution to the GPWDF value from two very near peaks (θ_i, I_i) and (θ_j, I_j) is essentially the same as that of a single peak with intensity $I_i + I_j$ at the same reflection angle.

The obfuscated XRPD pattern for molecule XXIX to which we wish to compare our candidate DFT-relaxed structures is shown in Figure 4 (black dots). The first step in our approach is the estimation of the pattern background using the method by David and Sivia based on Bayesian analysis (David & Sivia, 2001), and implemented in the FoX program (Favre-Nicolin & Černỳ, 2002). This process requires no user intervention. The calculated background for the XXIX pattern is shown as the blue line in Figure 4. Next, the background is subtracted from the experimental data, and the remaining peaks are fitted using a pattern model consisting of a linear combination of pseudo-Voigt functions:

$$P(\theta) = \sum_i I_i V(\theta; \eta_i, \theta_i, \gamma_i) \tag{10}$$

where the sum goes over the set of all discrete peaks in the pattern. Pseudo-Voigt functions are themselves convex linear combinations of normalized Gaussian and Lorentzian functions:

$$V(\theta; \eta_i, \theta_i, \gamma_i) = \eta_i G(\theta; \theta_i, \gamma_i) + (1 - \eta_i) L(\theta; \theta_i, \gamma_i) \tag{11}$$

$$G(\theta; \theta_i, \gamma_i) = \frac{1}{\sqrt{2\pi} s_i} \exp\left(-\frac{(\theta - \theta_i)^2}{2 s_i^2}\right) \tag{12}$$

$$L(\theta; \theta_i, \gamma_i) = \frac{\gamma_i}{2\pi} \frac{1}{(\theta - \theta_i)^2 + (\gamma_i/2)^2} \tag{13}$$

with $s_i = \gamma_i / (2\sqrt{2 \ln 2})$. For each peak i in the model there are four adjustable parameters: the intensity of the peak (I_i), its position (θ_i), its full width at half maximum (FWHM, γ_i), and the coefficient that controls the relative weight of the Gaussian and Lorentzian contributions ($0 \leq \eta_i \leq 1$). The derivatives of Eq. 10 with respect to these

parameters are easily calculated, enabling efficient non-linear least-squares fitting of the pattern model to the experimental data.

Typical experimental patterns comprise dozens of peaks, and the result of a straightforward non-linear least-squares fit of the pattern model (Eq. 10) to the data would depends strongly on the initial set of parameters. Therefore, a more sophisticated approach is required to fit the model. In our method, we employ a pre-fitting step in which reasonable initial values for the model parameters are determined. The approach follows these steps:

1. Find the positions of all candidate peak centers. A point is considered the center of a candidate peak if the profile intensity at that point is higher than the adjacent points. Given that experimental patterns show considerable noise, the user can input an intensity value below which candidate peaks are disregarded as noise. The impact of this choice of threshold on the model is not severe because most candidates peaks are eventually pruned by the fitting procedure, as described below.

2. For each candidate peak, define initial values for the peak parameters as well as their associated lower and upper bounds. The center position (θ_i) is allowed to shift by at most two times the distance between adjacent points in the pattern, the FWHM (γ_i) is constrained to a maximum pre-set value to prevent spurious solutions, η_i is bound between 0 and 1, and I_i is forced to remain positive.

3. A non-linear least-squares fit is carried out individually for each candidate peak, in order of decreasing intensity. After a given peak is fitted, its contribution is subtracted from the experimental pattern, and the resulting data is used for subsequent peaks. The process continues until the list of candidate peaks is exhausted. The least-squares fit of many candidate peaks, particularly those

with low intensity, results in an intensity of zero. These zero-intensity peaks are discarded, simplifying the model.

4. The model pattern comprising the remaining peaks is now fitted in its entirety to the experimental pattern (minus background) using the previously determined values for the initial parameters. The peaks whose intensity decreases to zero are pruned again, resulting in the final model pattern, and the corresponding set of (θ_i, I_i) pairs.

The proposed method requires no user intervention (beyond perhaps setting a noise threshold for peak detection) and may take between a few seconds to several minutes on a desktop computer, depending on the density of the XRPD data and the number of peaks in the pattern. It is important to note that the pre-processing of the experimental pattern needs to be done only once in a powder-assisted CSP. Figure 4 shows the remarkable performance of our method for the obfuscated XXIX pattern, in which there are only a few data points per reflection peak. Similar good performance is obtained for the seven "low-quality" experimental patterns (2-minute scans), corresponding to seven different molecules, in the work of Mayo et al. (Mayo *et al.*, 2023), as shown in Figure 5.

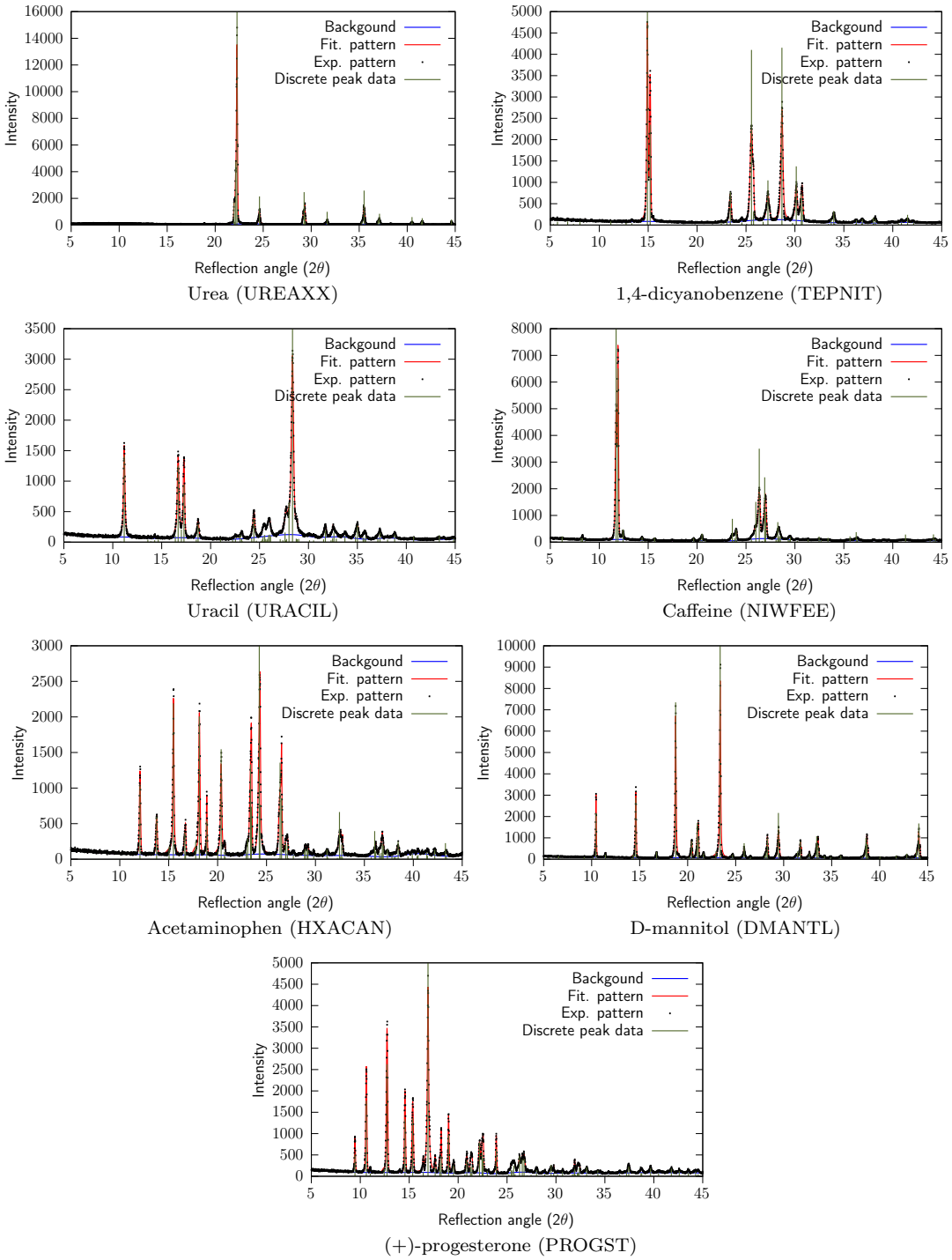

Fig. 5. Pattern models fitted to the X-ray powder diffraction data in Mayo et al.(Mayo *et al.*, 2023). Compound name and CSD refcode are indicated under the plot. The plots show the experimental data (black dots), fitted pattern (red), calculated background contribution (blue), and discrete angle/intensity pairs (green).

Another nice feature of the pre-processing method is that it gives the user some flexibility regarding how to treat the experimental pattern. For instance, the initial parameter set that enters the whole-pattern least-squares fit in step 4 can be modified by the user to remove candidate peaks. This is important if an XRPD pattern contains spurious peaks or consists of a known mixture of two different phases. A common occurrence is the presence of the pressure calibrant in an XRPD pattern from a high-pressure diamond-anvil cell experiment. Likewise, specific corrections (preferred orientation, zero-point error) can be applied during pre-processing, although we have not considered them in this work for simplicity.

4. Variable-Cell Similarity Index

We now consider the case of comparing two crystals (A and B), given as sets of reflections and intensity pairs. The two structures may originate from different sources: A and B could be equilibrium structures corresponding to different computational methods, or they could be experimental structures. In the last case, if XRPD is used as the experimental method, the structure itself may not be available, and only the reflections and intensities resulting from the pre-processing step in Section 3 may be known. As noted in Section 2, structures from different sources are not directly comparable because variations in lattice parameters between otherwise similar structures preclude the use of structural similarity indices such as GPWDF (Hofmann & Kuleshova, 2005; Van De Streek & Motherwell, 2005; Price *et al.*, 2016).

The question of whether two structures correspond to redeterminations of the same polymorph has been debated in the literature (Sacchi *et al.*, 2020). The crux of this question is that a purely mathematical description of structural similarity encapsulated in some similarity indices does not necessarily align with our chemical understanding of similarity: from a chemical perspective, we consider two structures as

similar if their molecular geometries, orientations and positions relative to each other are similar, even if the crystal lattice distorts under the effect of temperature, pressure, or as a result of a geometry relaxation with a particular method. As shown in Section 2, this is a crucial point in CSP, where this lattice distortion plays an important role in the comparison.

The problem of comparing structures from different sources was addressed by Hofmann et al. (Hofmann & Kuleshova, 2005; Hofmann & Kuleshova, 2006; Hofmann *et al.*, 2009), who proposed using integrated diffraction patterns. In this work, we first designate structure B as the "target" and define the variable-cell similarity index (VC-GPWDF) as:

$$G_{\mathrm{AB}}^{\mathrm{vc}} = \min_{\varepsilon} G_{\mathrm{A}(\varepsilon)\mathrm{B}} \tag{14}$$

where G_{AB} is the GPWDF index in Eq. 9 and $\mathrm{A}(\varepsilon)$ represents the structure A deformed by the strain ε, but with the same fractional atomic coordinates. A similar definition that is symmetric with respect to A and B can be made:

$$G_{\mathrm{AB}}^{\mathrm{vc}} = \frac{1}{2}\left(\min_{\varepsilon} G_{\mathrm{A}(\varepsilon)\mathrm{B}} + G_{\mathrm{AB}(\varepsilon^{-1})}\right) \tag{15}$$

However, in this work, one of the structures may correspond to a XRPD pattern, and therefore it would not be possible to calculate its deformation under a strain without indexing. Therefore, we will use the first definition in the rest of the article.

The calculation of the index in Eq. 14 is a global minimization problem. VC-PWDF addressed the calculation of $G_{\mathrm{AB}}^{\mathrm{vc}}$ by trying to build unit cells for A that coincide with the reduced unit cell of B. This method may fail to calculate $G_{\mathrm{AB}}^{\mathrm{vc}}$ properly if no such transformation exists (for instance, if the reduced cell of A spans a lattice corresponding to a sublattice of B). Instead, we carry out a global minimization search by exploring possible deformations of structure A within a certain range, and carrying out local minimizations of G_{AB} as a function of the lattice parameters of

A. The idea of using global minimization to fit a structure to an experimental pattern has been used for the interpretation and solution of powder diffraction data (Kariuki *et al.*, 1999; David & Shankland, 2008; Schlesinger *et al.*, 2021; Habermehl *et al.*, 2022; Altomare, 2022), and notably it is the approach taken by FIDEL-GO (Habermehl *et al.*, 2022). Unlike FIDEL-GO, we simplify the global minimization problem so that the resulting algorithm can be used efficiently and automatically to filter candidate structures in an arbitrary CSP protocol, given experimental XRPD data about the target compound. This is achieved in two ways. First, the minimization problem consists only of six parameters, corresponding to the lattice strain: the atomic coordinates of the candidate structure are fixed, and the need for parameters regarding peak shape, background, etc. has been removed by the pre-processing step. Second, the use of the analytical GPWDF index in Section 2 is faster than the numerical integration in the original de Gelder's index, and its derivatives are calculated analytically. This greatly increases the efficiency of each individual local minimization. For comparison, local minimizations with GPWDF take a fraction of a second, whereas a single FIDEL local optimization takes between a few minutes and half an hour (Habermehl *et al.*, 2014). Importantly, there are no parameters associated with the molecular motif, and therefore the cost of VC-GPWDF does not increase with the complexity of the target molecule.

The details for the calculation of $G^{\mathrm{vc}}_{\mathrm{AB}}$ (Eq. 14) are as follows. The global minimization is carried out using the NLOPT library (Johnson, 2007), specifically its implementation of the multi-level single-linkage (MLSL) method (Kan & Timmer, 1987), which searches for the global minimum by carrying out a sequence of local minimizations starting from a set of points with quasi-random distribution. Each local minimization is carried out using NLOPT's implementation of a modified sequential quadratic programming method (SLSQP) by Kraft (Kraft, 1994). The user can control the search

region and the stopping criterion for the global minimization. By default, the global search is constrained to a region around the unstrained structure of crystal A that is determined by a maximum cell length elongation of up to 10% and maximum cell angle deformation of 5°. The global minimization stops once there has been no improvement in the GPWDF value ($G_{\mathrm{A}(\varepsilon)\mathrm{B}}$) for 5000 GPWDF evaluations, although typically much fewer evaluations are required to arrive at a stable solution thanks to there being only six parameters in the minimization.

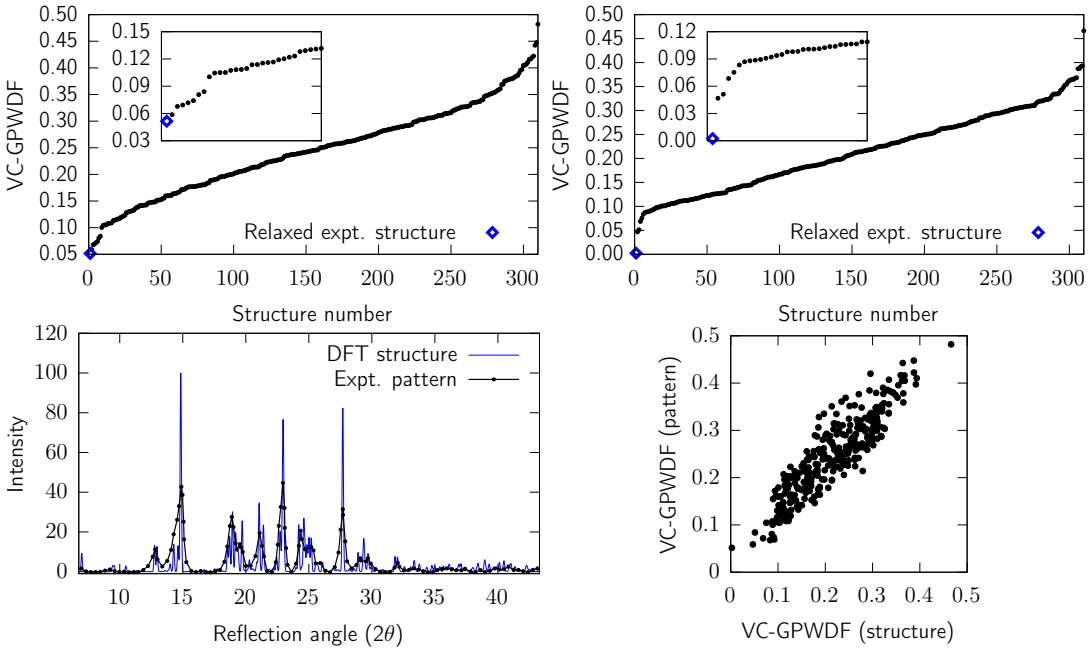

Fig. 6. Results using VC-GPWDF on the list of 310 DFT-relaxed candidates against the experimental structure and against the obfuscated XRPD pattern in molecule XXIX: a) comparison with the experimental pattern; b) Comparison with the experimental structures; c) experimental pattern of molecule XXIX compared with the best VC-GPWDF match, after application of the structural distortion used in the calculation of the similarity index; d) comparison of VC-GPWDF calculated against the experimental structure (y-axis) vs. the experimental pattern (x-axis).

The results for the VC-GPWDF comparison between the list of 310 DFT-relaxed candidate structures and the experimental structure and XRPD pattern of molecule XXIX are shown in Figure 6. The VC-GPWDF values for the comparison between

the candidate structures and the experimental pattern (Figure 4) are given in Figure 6a. The experimental structure is identified from the list of candidates as the first rank with the lowest VC-GPWDF score, 0.052. Figure 6c shows the experimental pattern compared with a synthetic pattern calculated from the DFT-relaxed experimental structure after the deformation applied to match the experimental data applied by VC-GPWDF. Even with an experimental pattern as poor as the one used in this example, our new method still identifies correctly the candidate corresponding to the DFT-relaxed experimental structure as the best match, and in the case of a higher-quality XRPD pattern, our approach could be followed by a Rietveld refinement, same as in other SDPD approaches (Habermehl *et al.*, 2022). It is clear that a powder-assisted CSP protocol may produce several candidates that reasonably match the XRPD pattern, and VC-GPWDF is clearly useful to filter out most of the non-matching candidate structures, even if the structure cannot be assigned conclusively (Altomare, 2022). The average run time for the VC-GPWDF calculations in Figure 6a is 31 seconds on a single processor, with a maximum run time of 102 seconds. This allows comparing the whole candidate list in a few minutes by parallelizing over structures, which is a remarkable performance considering a global minimization is carried out at each point in the figure, and a negligible cost compared to that of a typical CSP protocol.

The VC-GPWDF index can also be used if the experimental structure is available. The results for comparison with the experimental structure of molecule XXIX are shown in Figure 6b. The DFT-relaxed experimental structure matches the actual experimental structure with a VC-GPWDF index of 0.0024, an order of magnitude lower than the next candidate (0.047). Clearly, when the experimental structure is available, VC-GPWDF predicts a clearer match than when comparing only to the XRPD pattern, indicating that the information loss caused by the low quality of

the pattern is affecting the comparison. Figure 6d shows the VC-GPWDF calculated against the pattern (y-axis) and against the experimental structure (x-axis) for the 310 candidates. There is a clear correlation between the two, although the spread in the points indicates that the comparison to the XRPD pattern suffers from the low quality of the experimental data.

Lastly, we applied the new VC-GPWDF index to compare the candidate structures to the XRPD patterns for the seven molecules used in the work of Mayo et al.(Mayo *et al.*, 2023), with the list of candidate structures obtained from the Control and Prediction of the Organic Solid State (CPOSS) database (Price & Price, 2024). The outcome of this comparison is shown in Figure 7. Same as in the case of VC-xPWDF (Mayo *et al.*, 2023), VC-GPWDF identifies the candidate structures derived from experiment as the top ranks for each molecule. However, unlike VC-xPWDF, VC-GPWDF does not require indexing the patterns.

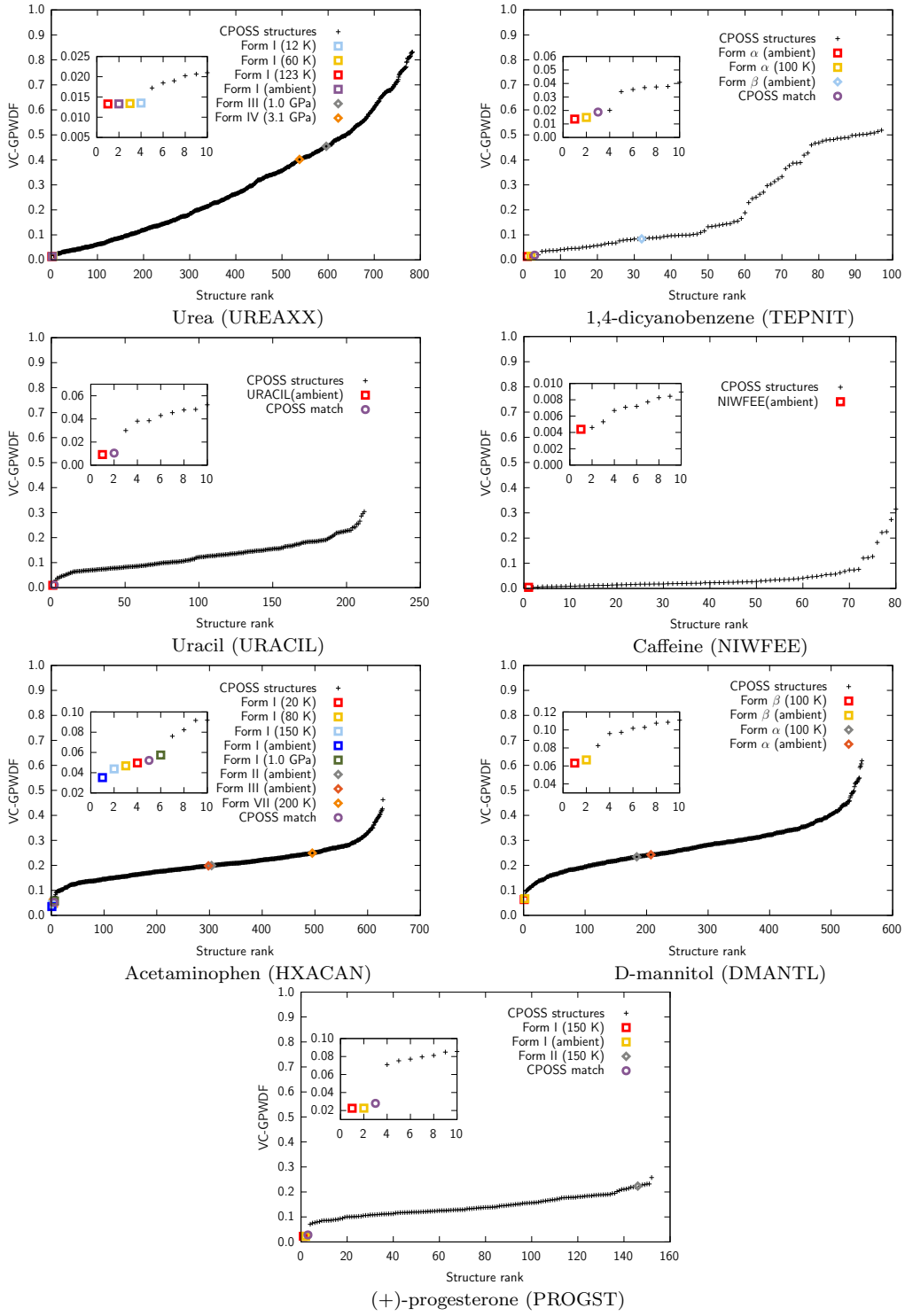

Fig. 7. Variable-cell similarity index (VC-GPWDF) applied to the set of crystal structures in Mayo et al.(Mayo *et al.*, 2023), compared to the corresponding experimental patterns. The colored points correspond to the experimental forms in Table 1 of the reference article.

5. Conclusions

In this article, we presented two crystal structure similarity indices (GPWDF and VC-GPWDF) based on the comparison of X-ray powder diffraction (XRPD) patterns. Our objective is to enable the use of experimental powder diffraction data to filter candidate structures in any crystal structure prediction (CSP) protocol, in an automatic, reliable, and straightforward manner, even if the powder data is of low quality.

GPWDF is a modification of de Gelder's similarity index using cross-correlation functions. In addition to using a modified weight function, GPWDF operates not on the powder pattern profiles but on discrete sets of reflection angles and intensities (θ_i, I_i). These two modifications enable the analytical calculation of GPWDF and its derivatives as a function of the structural parameters of the crystals being compared. The derivatives of the GPWDF index with respect to the lattice parameters are given (Appendix A). It was shown that GPWDF produces similarity scores that correlate strongly with de Gelder's index, but the calculation has a much lower cost, in addition to providing analytical derivatives. This makes local optimizations of GPWDF with respect to the structural parameters of the two crystals particularly efficient.

In order to use GPWDF to compare structures with experimental XRPD patterns, we present an automatic pre-processing method that extracts the set of reflection angles and intensities from an XRPD profile. The pre-processing step is an automatic algorithm for fitting a linear combination of pseudo-Voigt functions to the XRPD pattern. This step, which needs to be done only once in each CSP run, requires minimal user intervention, and does not perform an indexing of the pattern. The pre-processing step was tested on molecule XXIX from the powder-assisted challenge within the Cambridge Crystallographic Database Centre (CCDC) seventh blind test. The provided experimental XRPD pattern is of very poor quality—it was deliberately obfuscated as part of the challenge—but nonetheless our method successfully extracts the relevant

features from it.

Crystal structures from different sources (e.g. computational and experimental) are not comparable because of lattice distortions caused by temperature, pressure, or the idiosyncrasies of a particular computational method. This is a problem that needs to be overcome for our method to be useful in powder-assisted CSP, where *in silico* structures are matched against experimental ones. For this purpose, we define the variable-cell GPWDF (VC-GPWDF) as the minimum GPWDF when considering all possible lattice deformations of one of the structures. VC-GPWDF is calculated using a global minimization of GPWDF as a function of lattice strain, which is carried out by sampling a region around the provided structure and performing local minimizations of the GPWDF value with respect to the lattice parameters. Application of the VC-GPWDF index to a list of 310 candidate structures for molecule XXIX identifies the DFT-relaxed experimental structure as the best match, both when comparing with the experimental pattern as well as with the experimental structure.

Each VC-GPWDF comparison takes seconds to minutes, which is a negligible amount of time compared to the cost of a CSP run itself. Calculating VC-GPWDF requires neither user intervention nor an indexing of the experimental XRPD pattern. Because the minimization does not involve the molecular motif, the cost of VC-GPWDF does not depend directly on the compound's molecular complexity (e.g. the number of rotatable bonds) and, as evidenced by the performance on molecule XXIX, meaningful information can be extracted even from experimental data with very poor quality. We believe this will be an important tool when combined with existing molecular CSP protocols, particularly in the search of high-pressure material phases, for which only low-quality XRPD data is available.

6. Funding information

The following funding is acknowledged: Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación, projects PGC2021-125518NB-I00 and RED2022-134388-T cofinanced by EU FEDER funds; the Principality of Asturias (FICYT), project AYUD/2021/51036 cofinanced by EU FEDER; the Spanish MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR for grants TED2021-129457B-I00 and CNS2023-144958. The author thankfully acknowledges the computer resources at MareNostrum5 and the technical support provided by BSC (RES-AECT-2024-2-0010).

Appendix A
Derivatives of the GPWDF Similarity Index

The VC-GPWDF variable-cell similarity index (Eq. 14) is based on the minimization of the GPWDF similarity index (Eq. 9) with respect to the lattice parameters of one of the structures (in the following, A). The derivative of G_{AB} with respect to a metric tensor component of structure A (G_{ij}) is:

$$G'_{\mathrm{AB}} = -D'_{\mathrm{AB}} = -D_{\mathrm{AB}} \left[\frac{I'_{\mathrm{AB}}}{I_{\mathrm{AB}}} - \frac{I'_{\mathrm{AA}}}{I_{\mathrm{AA}}} \right] \tag{16}$$

where the prime symbol indicates differentiation, and the derivative in I'_{AA} is taken with respect to one of the patterns but not the other. Assuming the six G'_{AB} are known, the calculation of the G_{AB} derivatives with respect to the lattice parameters

is straightforward using the chain rule:

$$\frac{1}{2}\frac{\partial G_{\rm AB}}{\partial a} = a\frac{\partial G_{\rm AB}}{\partial G_{11}} + b\cos\gamma\frac{\partial G_{\rm AB}}{\partial G_{12}} + c\cos\beta\frac{\partial G_{\rm AB}}{\partial G_{13}} \tag{17}$$

$$\frac{1}{2}\frac{\partial G_{\rm AB}}{\partial b} = a\cos\gamma\frac{\partial G_{\rm AB}}{\partial G_{12}} + b\frac{\partial G_{\rm AB}}{\partial G_{22}} + c\cos\alpha\frac{\partial G_{\rm AB}}{\partial G_{23}} \tag{18}$$

$$\frac{1}{2}\frac{\partial G_{\rm AB}}{\partial c} = a\cos\beta\frac{\partial G_{\rm AB}}{\partial G_{13}} + b\cos\alpha\frac{\partial G_{\rm AB}}{\partial G_{23}} + c\frac{\partial G_{\rm AB}}{\partial G_{33}} \tag{19}$$

$$\frac{1}{2}\frac{\partial G_{\rm AB}}{\partial \alpha} = -bc\sin\alpha\frac{\partial G_{\rm AB}}{\partial G_{23}} \tag{20}$$

$$\frac{1}{2}\frac{\partial G_{\rm AB}}{\partial \beta} = -ac\sin\beta\frac{\partial G_{\rm AB}}{\partial G_{13}} \tag{21}$$

$$\frac{1}{2}\frac{\partial G_{\rm AB}}{\partial \gamma} = -ab\sin\gamma\frac{\partial G_{\rm AB}}{\partial G_{12}} \tag{22}$$

In Eq. 16, the derivative of $I_{\rm AB}$ (Eq. 2) with respect to G_{ij} is:

$$I'_{\rm AB} = \sqrt{z}\sum_{i}^{\rm A}\sum_{j}^{\rm B} I_j \exp\left(-\pi z(\theta_i-\theta_j)^2\right) \times \\ \times \left[I'_i - 2\pi z I_i(\theta_i-\theta_j)\theta'_i\right] \tag{23}$$

where I'_i and θ'_i are the derivatives of the reflection intensity (peak area) and diffraction angle for reciprocal lattice vector i (with coordinates hkl). The intensity and its derivative are given by (Pecharsky & Zavalij, 2005):

$$I_i = L|F_i|^2 \tag{24}$$

$$I'_i = L'|F_i|^2 + L(|F_i|^2)' \tag{25}$$

where $L(\theta_i)$ is the Lorentz-Polarization factor and $|F_i|^2$ is the structure factor corresponding to the i reflection. In this equation we have disregarded multiplicative terms because $G_{\rm AB}$ is invariant to them. Likewise, absorption and preferred orientation corrections are also disregarded for simplicity, although a more complex implementation considering these as well as thermal effects is possible. The Lorentz-polarization factor for the integrated intensity (Yinghua, 1987; Pecharsky & Zavalij, 2005) is:

$$L = \frac{1+A(\cos 2\theta_i)^2}{(1+A)\cos\theta_i\sin^2\theta_i} \tag{26}$$

$$L' = -\theta'_i\left[\frac{8A\cos 2\theta_i}{(1+A)\sin\theta_i} + L\frac{\cos\theta_i}{\sin\theta_i} + L\frac{2\cos 2\theta_i}{\sin 2\theta_i}\right] \tag{27}$$

with $A = \cos^2 2\theta_M$ and θ_M is the monochromator angle. The structure factor is:

$$F_i = \sum_k A_{ik} \exp(i\xi_{ik}) \tag{28}$$

where k runs over the atoms in the unit cell, $\xi_{ik} = 2\pi h_i \cdot x_k$, h_i is the i reciprocal lattice vector, and x_k are the coordinates of atom k. The contribution to the intensity is:

$$|F_i|^2 = \left(\sum_k A_{ik} \cos\xi_{ik}\right)^2 + \left(\sum_k A_{ik} \sin\xi_{ij}\right)^2 \tag{29}$$

$$\begin{aligned}(|F_i|^2)' = {} & 2\left(\sum_j A_{ik} \cos\xi_{ij}\right)\left(\sum_j A'_{ik} \cos\xi_{ij}\right) \\ & + 2\left(\sum_j A_{ik} \sin\xi_{ij}\right)\left(\sum_j A'_{ik} \sin\xi_{ij}\right)\end{aligned} \tag{30}$$

The A_{ik} factors combine the atomic scattering factors (f_k) and the temperature factors:

$$A_{ik} = \exp\left(-B_k \frac{\sin^2\theta_i}{\lambda^2}\right) f_k \tag{31}$$

$$A'_{ik} = \theta'_i \exp\left(-B_k \frac{\sin^2\theta_i}{\lambda^2}\right)\left[f'_k - \frac{2B_k \sin\theta_i f_k}{\lambda}\right]\frac{\cos\theta_i}{\lambda} \tag{32}$$

To avoid having too many parameters in the global minimization required for the VC-GPWDF calculation, we assume a constant value for the atomic displacement parameters ($B_k = 1^2$) for all atoms in the structure. A cursory exploration using the molecule XXIX candidate list revealed that the choice of B does not impact significantly the performance of the index in actual comparisons. The scattering factors are calculated with the usual interpolation formulas (Prince, 2004), which are easily differentiated. For instance:

$$f_k = \sum_{m=1}^{4} a_m \exp\left(-b_m \frac{\sin^2\theta_i}{\lambda^2}\right) + c_m \tag{33}$$

$$f'_k = -\theta'_i \frac{\sin 2\theta_i}{\lambda} \sum_{m=1}^{4} a_m b_m \exp\left(-b_m \frac{\sin^2\theta_i}{\lambda^2}\right) \tag{34}$$

Lastly, the diffraction angle is related to the distance between adjacent planes perpendicular to the i reciprocal lattice vector by Bragg's law:

$$\sin\theta_i = \frac{\lambda}{2d_i} \tag{35}$$

$$\theta_i' = -\frac{\sin\theta_i}{d_i\cos\theta_i} d_i' \tag{36}$$

and the relation between the interplanar distance and its derivative with respect to the metric tensor of the structure is:

$$d_i = \sqrt{hG^{-1}h^T} \tag{37}$$

$$\frac{\partial d_i}{\partial G} = -\frac{G^{-1}h^T hG^{-1}}{2d_i} \tag{38}$$

References

Altomare, A. (2022). *IUCrJ*, **9**(4), 403–405.

Bardwell, D. A., Adjiman, C. S., Arnautova, Y. A., Bartashevich, E., Boerrigter, S. X. M. & et al. (2011). *Acta Crystallogr.* B**67**, 535–551.

Becke, A. D. (1986). *J. Chem. Phys.* **85**, 7184.

Bowskill, D. H., Sugden, I. J., Konstantinopoulos, S., Adjiman, C. S. & Pantelides, C. C. (2021). *Annu. Rev. Chem. Biomol. Eng.* **12**, 593–623.

Brüning, J. & Schmidt, M. U. (2015). *J. Pharm. Pharmacol.* **67**(6), 773–781.

CCDC (2024). CSP blind test structure reveal target XXIX—Flavouring compound. Accessed on March 15th, 2024. https://www.ccdc.cam.ac.uk/discover/blog/csp-blind-test-reveal-xxix/

Chisholm, J. A. & Motherwell, S. (2005). *J. Appl. Crystallogr.* **38**(1), 228–231.

Chuliá-Jordan, R., Santamaría-Pérez, D., Pereira, A., García-Domene, B., Vilaplana, R., Sans, J., Martinez-Garcia, D., Morales-Garcia, A., Popescu, C., Muehle, C., Jansen, M. & FJ, M. (2020). *J. Alloys Compd.* **830**, 154646.

Chuliá-Jordán, R., Santamaria-Perez, D., Ruiz-Fuertes, J., Otero-de-la Roza, A. & Popescu, C. (2021). *Minerals*, **11**(6), 607.

David, W. & Sivia, D. (2001). *J. Appl. Crystallogr.* **34**(3), 318–324.

David, W. I. & Shankland, K. (2008). *Acta Crystallogr. A*, **64**(1), 52–64.

David, W. I., Shankland, K., Van De Streek, J., Pidcock, E., Motherwell, W. S. & Cole, J. C. (2006). *J. Appl. Crystallogr.* **39**(6), 910–915.

Day, G. M., Cooper, T. G., Cruz-Cabeza, A. J., Hejczyk, K. E., Ammon, H. L. & et al. (2009). *Acta Crystallogr.* B**65**, 107–125.

Day, G. M., Motherwell, W. D. S., Ammon, H. L., Boerrigter, S. X. M., Della Valle, R. G. & et al. (2005). *Acta Crystallogr.* B**61**, 511–527.

Favre-Nicolin, V. & Černỳ, R. (2002). *J. Appl. Crystallogr.* **35**(6), 734–743.

Fredericks, S., Parrish, K., Sayre, D. & Zhu, Q. (2021). *Comput. Phys. Commun.* **261**, 107810.

Gao, P., Tong, Q., Lv, J., Wang, Y. & Ma, Y. (2017). *Comput. Phys. Commun.* **213**, 40–45.

de Gelder, R., Wehrens, R. & Hageman, J. A. (2001). *J. Comput. Chem.* **22**(3), 273–289.

Giannozzi, P., Andreussi, O., Brumme, T., Bunau, O., Nardelli, M. B., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Cococcioni, M., Colonna, N., Carnimeo, I., Corso, A. D., de Gironcoli, S., Delugas, P., DiStasio, R. A., Ferretti, A., Floris, A., Fratesi, G., Fugallo, G., Gebauer, R., Gerstmann, U., Giustino, F., Gorni, T., Jia, J., Kawamura, M., Ko, H.-Y., Kokalj, A., Küçükbenli, E., Lazzeri, M., Marsili, M., Marzari, N., Mauri, F., Nguyen, N. L., Nguyen, H.-V., de-la Roza, A. O., Paulatto, L., Poncé, S., Rocca, D., Sabatini, R., Santra, B., Schlipf, M., Seitsonen, A. P., Smogunov, A., Timrov, I., Thonhauser, T., Umari, P., Vast, N., Wu, X. & Baroni, S. (2017). *J. Phys.: Condens. Matter*, **29**(46), 465901.

Guzei, I. A., Gunn, E. M., Spencer, L. C., Schomaker, J. M. & Rigoli, J. W. (2011). *CrystEngComm*, **13**(10), 3444–3450.

Habermehl, S., Mörschel, P., Eisenbrandt, P., Hammer, S. M. & Schmidt, M. U. (2014). *Acta Crystallogr. B*, **70**(2), 347–359.

Habermehl, S., Schlesinger, C. & Schmidt, M. U. (2022). *Acta Crystallogr. B*, **78**(2), 195–213.

Hageman, J. A., Wehrens, R., De Gelder, R. & Buydens, L. M. (2003). *J. Comput. Chem.* **24**(9), 1043–1051.

Harris, K. D. (2022). *Acta Crystallogr. B*, **78**(2), 96–99.

Hofmann, D. & Kuleshova, L. (2006). *Crystallogr. Rep.* **51**, 419–427.

Hofmann, D., Kuleshova, L., Hofmann, F. & D'Aguanno, B. (2009). *Chem. Phys. Lett.* **475**(1-3), 149–155.

Hofmann, D. W. M. & Kuleshova, L. (2005). *J. Appl. Crystallogr.* **38**(6), 861–866.

Hunnisett, L. M. & et al. (2024*a*). *Acta Crystallogr. B*. (in preparation).

Hunnisett, L. M. & et al. (2024*b*). *Acta Crystallogr. B*. (in preparation).

Ivanisevic, I., Bugay, D. E. & Bates, S. (2005). *J. Phys. Chem. B*, **109**(16), 7781–7787.

Johnson, S. G. (2007). The NLopt nonlinear-optimization package. `https://github.com/stevengj/nlopt`.

Kan, A. H. G. R. & Timmer, G. T. (1987). *Math. Program.* **39**, 57–78.

Karfunkel, H., Rohde, B., Leusen, F., Gdanitz, R. J. & Rihs, G. (1993). *J. Comput. Chem.* **14**(10), 1125–1135.

Kariuki, B. M., Belmonte, S. A., McMahon, M. I., Johnston, R. L., Harris, K. D. & Nelmes, R. J. (1999). *J. Synchrot. Radiat.* **6**(2), 87–92.

Kraft, D. (1994). *ACM Trans. Math. Softw.* **20**, 262–281.

Lawton, S. L. & Bartell, L. S. (1994). *Powder Diffr.* **9**(2), 124–135.

Lommerse, J. P. M., Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Gavezzotti, A. & et al. (2000). *Acta Crystallogr.* B**58**, 647–661.

Mayo, R. A. & Johnson, E. R. (2021). *CrystEngComm*, **23**(40), 7118–7131.

Mayo, R. A., Marczenko, K. M. & Johnson, E. R. (2023). *Chem. Sci.* **14**(18), 4777–4785.

Mayo, R. A., Otero-de-la Roza, A. & Johnson, E. R. (2022). *CrystEngComm*, **24**(47), 8326–8338.

Mosca, M. M. & Kurlin, V. (2020). *Cryst. Res. Technol.* **55**(5), 1900197.

Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Dzyabchenko, A., Erk, P. & et al. (2002). *Acta Crystallogr.* B**58**, 647–661.

Nessler, A. J., Okada, O., Hermon, M. J., Nagata, H. & Schnieders, M. J. (2022). *J. Appl. Crystallogr.* **55**(6), 1528–1537.

Nyman, J. & Day, G. M. (2015). *CrystEngComm*, **17**, 5154–5165.

Nyman, J., Yu, L. & Reutzel-Edens, S. M. (2019). *CrystEngComm*, **21**(13), 2080–2088.

Otero-de-la-Roza, A. & Johnson, E. R. (2012). *J. Chem. Phys.* **136**, 174109.

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

Özer, B., Karlsen, M. A., Thatcher, Z., Lan, L., McMahon, B., Strickland, P. R., Westrip, S. P., Sang, K. S., Billing, D. G., Ravnsbæk, D. B. & Billinge, S. J. L. (2022). *Acta Crystallogr. A*, **78**(5), 386–394.

Padgett, C. W., Arman, H. D. & Pennington, W. T. (2007). *Cryst. Growth Des.* **7**(2), 367–372.

Patterson, A. (1939). *Nature*, **143**(3631), 939–940.

Pecharsky, V. & Zavalij, P. (2005). *Fundamentals of Powder Diffraction and Structural Characterization of Materials*. Springer US.

Perdew, J. P., Burke, K. & Ernzerhof, M. (1996). *Phys. Rev. Lett.* **77**(18), 3865.

Price, L. S. & Price, S. L. (2024). Control and prediction of the organic solid state database. Accessed on March 23rd, 2024. http://www.chem.ucl.ac.uk/cposs/index.htm

Price, S. L. (2018). *Faraday Discuss.* **211**, 9–30.

Price, S. L., Braun, D. E. & Reutzel-Edens, S. M. (2016). *Chem. Commun.* **52**(44), 7065–7077.

Prince, E. (2004). *International Tables for Crystallography, Volume C: Mathematical, physical and chemical tables*. Springer Science & Business Media.

Reilly, A. M., Cooper, R. I., Adjiman, C. S., Bhattacharya, S., Boese, A. D., Brandenburg, J. G., Bygrave, P. J., Bylsma, R., Campbell, J. E., Car, R., Case, D. H., Chadha, R., Cole, J. C., Cosburn, K., Cuppen, H. M., Curtis, F., Day, G. M., DiStasio Jr, R. A., Dzyabchenko, A., van Eijck, B. P., Elking, D. M., van den Ende, J. A., Facelli, J. C., Ferraro, M. B., Fusti-Molnar, L., Gatsiou, C.-A., Gee, T. S., de Gelder, R., Ghiringhelli, L. M., Goto, H., Grimme, S., Guo, R., Hofmann, D. W. M., Hoja, J., Hylton, R. K., Iuzzolino, L., Jankiewicz, W., de Jong, D. T., Kendrick, J., de Klerk, N. J. J., Ko, H.-Y., Kuleshova, L. N., Li, X., Lohani, S., Leusen, F. J. J., Lund, A. M., Lv, J., Ma, Y., Marom, N., Masunov, A. E., McCabe, P., McMahon, D. P., Meekes, H., Metz, M. P., Misquitta, A. J., Mohamed, S., Monserrat, B., Needs, R. J., Neumann, M. A., Nyman, J., Obata, S., Oberhofer, H., Oganov, A. R., Orendt, A. M., Pagola, G. I., Pantelides, C. C., Pickard, C. J., Podeszwa, R., Price, L. S., Price, S. L., Pulido, A., Read, M. G., Reuter, K., Schneider, E., Schober, C., Shields, G. P., Singh, P., Sugden, I. J., Szalewicz, K., Taylor, C. R., Tkatchenko, A., Tuckerman, M. E., Vacarro, F., Vasileiadis, M., Vazquez-Mayagoitia, A., Vogt, L., Wang, Y., Watson, R. E., de Wijs, G. A., Yang, J., Zhu, Q. & Groom, C. R. (2016). *Acta Crystallogr. B*, **72**(4), 439–459.

Rohlíček, J. & Skořepová, E. (2020). *J. Appl. Crystallogr.* **53**(3), 841–847.

Sacchi, P., Lusi, M., Cruz-Cabeza, A. J., Nauha, E. & Bernstein, J. (2020). *CrystEngComm*, **22**(43), 7170–7185.

Santamaría-Pérez, D., Chulia-Jordan, R., González-Platas, J., Otero-de-la-Roza, A., Ruiz-Fuertes, J., Oliva, R. & Popescu, C. (2024). *Cryst. Growth Des.* **24**, 1159–1169.

Schlesinger, C., Fitterer, A., Buchsbaum, C., Habermehl, S., Chierotti, M. R., Nervi, C. & Schmidt, M. U. (2022). *IUCrJ*, **9**(4), 406–424.

Schlesinger, C., Habermehl, S. & Prill, D. (2021). *J. Appl. Crystallogr.* **54**(3), 776–786.

Schmidt, M. U., Ermrich, M. & Dinnebier, R. E. (2005). *Acta Crystallogr. B*, **61**(1), 37–45.

Schreiner, W. N. & Jenkins, R. (1982). *Adv. X-ray Anal.* **26**, 141–147.

Stephenson, D. S. & Binsch, G. (1980). *J. Magn. Res.* **37**(3), 409–430.

Terban, M. W. & Billinge, S. J. (2021). *Chem. Rev.* **122**(1), 1208–1272.

Thompson, P., Cox, D. & Hastings, J. (1987). *J. Appl. Crystallogr.* **20**(2), 79–83.

Van De Streek, J. (2006). *Acta Crystallogr. B*, **62**(4), 567–579.

Van De Streek, J. & Motherwell, S. (2005). *Acta Crystallogr. B*, **61**(5), 504–510.

Wei, L., Li, Q., Omee, S. S. & Hu, J. (2024). *Comput. Mater. Sci.* **235**, 112802.

Whittleton, S. R., Otero-de-la-Roza, A. & Johnson, E. R. (2017*a*). *J. Chem. Theory Comput.* **13**, 441–450.

Whittleton, S. R., Otero-de-la-Roza, A. & Johnson, E. R. (2017*b*). *J. Chem. Theory Comput.* **13**, 5332–5342.

Widdowson, D. & Kurlin, V. (2022). In *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=4wrB7Mo9$_O$Q

Widdowson, D., Mosca, M. M., Pulido, A., Kurlin, V. & Cooper, A. I. (2022). *MATCH Commun. Math. Comput. Chem.* **87**(3), 529–559.

Willighagen, E., Wehrens, R., Verwer, P., De Gelder, R. & Buydens, L. (2005). *Acta Crystallogr. B*, **61**(1), 29–36.

Wood, P. A., Oliveira, M. A., Zink, A. & Hickey, M. B. (2012). *CrystEngComm*, **14**(7), 2413–2421.

Yinghua, W. (1987). *J. Appl. Crystallogr.* **20**(3), 258–259.

Zhu, L., Amsler, M., Fuhrer, T., Schaefer, B., Faraji, S., Rostami, S., Ghasemi, S. A., Sadeghi, A., Grauzinyte, M., Wolverton, C. & Goedecker, S. (2016). *J. Chem. Phys.* **144**(3), 034203.

Synopsis

A new method is presented for utilizing experimental powder diffraction patterns to rank in silico structures generated using a crystal structure prediction protocol.