Progressive Alignment of Crystals:
Reproducible and Efficient Assessment of Crystal Structure Similarity

Aaron J. Nessler¹, Okimasa Okada², Mitchell J. Hermon¹, Hiroomi Nagata³,*, and Michael J. Schnieders¹,*
¹The University of Iowa, Computational Biomolecular Engineering Lab
²Mitsubishi Tanabe Pharma Corporation, Sohyaku, Innovative Research Division
³Mitsubishi Tanabe Pharma Corporation, Production Technology & Supply Chain Management Division, CMC Modality Technology Laboratories
¹Denotes joint first authorship
*Corresponding authors (nagata.hiroomi@md.mt-pharma.co.jp, michael-schnieders@uiowa.edu)

Synopsis
Evaluating crystal structure packings in a reproducible manner requires not only calculating a coordinate root-mean-square deviation (RSMD) for N molecules (or N asymmetric units) but should also include metrics to describe the shape of the compared molecular clusters to account for alternative approaches used to prioritize selection of molecules. Here we describe a fast algorithm called Progressive Alignment of Crystals (PAC) to evaluate crystal packing similarity using coordinate RMSD and introduce radius of gyration (R_g) as a metric to quantify the shape of the superimposed clusters.

Abstract
During in silico crystal structure prediction of organic molecules, millions of candidate structures are often generated. These candidates must be compared to remove duplicates prior to further analysis (e.g., optimization with electronic structure methods) and ultimately compared with structures determined experimentally. The agreement of predicted and experimental structures forms the basis of evaluating the results from the Cambridge Crystallographic Data Centre (CCDC) blind assessment of crystal structure prediction, which further motivates the importance of rigorous alignments. Evaluating crystal structure packings in a reproducible manner requires not only calculating a coordinate root-mean-square deviation (RSMD) for N molecules (or N asymmetric units), but we argue should also include metrics to describe the shape of the compared molecular clusters to account for alternative approaches used to prioritize selection of molecules. Here we describe a flexible algorithm called Progressive Alignment of Crystals (PAC) to evaluate crystal packing similarity using coordinate RMSD and introduce radius of gyration (R_g) as a metric to quantify the shape of the superimposed clusters. We show that the absence of metrics to describe cluster shape adds ambiguity to the results of the CCDC blind assessments because it is not possible to determine whether the superposition algorithm prioritized tightly packed molecular clusters (i.e., to minimize R_g) or prioritized reduced RMSD (i.e., via possibly elongated clusters with relatively larger R_g). For example, we show that when the PAC algorithm described here uses “single linkage” to prioritize molecules for inclusion in the superimposed clusters, our results are nearly identical to those calculated by the widely used program COMPACK. However, we favor the
lower $R_g$ values obtained by use of “average linkage” for molecule prioritization because the resulting RMSDs more equally reflect the importance of packing along each dimension. We conclude by showing that the PAC algorithm is faster than COMPACK when using a single process, demonstrate its utility for biomolecular crystals, and finally present parallel scaling up to 64 processes in the open-source code Force Field X.

Keywords: structure comparison; crystal packing; crystal structure prediction; radius of gyration

1. Introduction

Organic crystals are significant due to their role in causing diseases such as gout (Terkeltaub, 2010) (monosodium urate monohydrate) and kidney stones (Moe, 2006) (calcium oxalate), their potential use in the low-pressure storage of gases within crystalline metal-organic frameworks (James, 2003; Furukawa et al., 2010), and their use in the oral delivery of pharmaceuticals (Blagden et al., 2007) such as paracetamol (Haisa et al., 1976; Haisa et al., 1974) (acetaminophen) and acetylsalicylic acid (Wheatley, 1964; Vishweshwar et al., 2005) (aspirin). During the pharmaceutical formulation process, crystallization screens often discover more than one crystal packing arrangement (i.e., polymorphs) based on testing an array of experimental conditions (e.g., solvent, pH, salt, temperature, and pressure). Each solid form has unique physical properties (e.g., density, thermodynamic stability, melting temperature, solubility) driven by both intramolecular conformation and intermolecular interactions. For this reason, each polymorph can be covered by a unique patent, and, in the case of a pharmaceutical solid form, must be considered individually for FDA approval (Kapczynski et al., 2012). Crystal structure prediction can be performed in silico to complement experimental polymorph screens and thereby reduce the risk of a previously unknown stable polymorph emerging (Leelananda & Lindert, 2016). A variety of computational methods have been used to predict crystal structures (Day, 2011; Reilly et al., 2016; Burger et al., 2018; Price, 2008; Price & Price, 2011; Price, 2014; Karamertzanis et al., 2009), each of which includes one or more steps to compare predicted crystal packings and remove duplicates (Day, 2011).

Each polymorph is defined by its space group, lattice parameters and the atomic coordinates of its asymmetric unit. The asymmetric unit of a space group is that part of the crystallographic unit cell that can be used to generate a complete unit cell by the symmetry of the space group. Throughout this work, comparisons are described in terms of clusters of $N$ molecules, rather than more cumbersome terminology such as $N$ asymmetric units. Constructing an optimal, reproducible comparison of two crystal polymorphs is a challenge because simply superimposing a single molecule from each conformer does not quantify intermolecular orientations. For this reason, crystal packing coordinate root mean square deviations (RMSD) generally consider a cluster of $N$ molecules (denoted RMSD$_N$), where $N$ is often chosen to be in the range $\sim$10 to 20. Coordinate RMSD$_N$ increases with $N$ because small discrepancies between the lattice parameters of two polymorphs are magnified as cluster size
increases. The requirement to prioritize N molecules (or N times the number of molecules in the asymmetric unit when more than one molecule is present) from each polymorph and match them prior to calculation of the RMSDs can lead to ambiguous results unless the shape of the superimposed clusters is reported via a simple metric such as radius of gyration ($R_g$).

Multiple algorithms have been proposed to quantify crystal structure similarity. In addition to their own algorithm (named CMPZ), Hundt et al. present a thorough history of early crystal comparison approaches (Hundt et al., 2006). There are a plethora of crystal comparison algorithms currently available with a variety of methods ranging from reductions in the dimensionality of input structures into more manageable representations based on intrinsic properties (e.g., periodic point sets, crystallographic information, X-ray powder diffraction, etc.) to transformations of the crystallographic information into a many dimensional configurations (or fingerprint) space (Sadeghi et al., 2013; Valle & Oganov, 2010; Willighagen et al., 2005; Gelder et al., 2001; Karfunkel et al., 1993; Verwer & Leusen, 1998; Mosca & Kurlin, 2020; Thomas et al., 2021; Widdowson et al., 2022; Edelsbrunner et al., 2021; de la Flor et al., 2016; Ferré et al., 2015; Hicks et al., 2021; De et al., 2016; Gelato & Parthe, 1987; Dzyabchenko, 1994; Lonie & Zurek, 2012; Chuanxun et al., 2017; Ong et al., 2013). These methods can mitigate complexities that arise when dealing with a direct comparison of atomic positions (e.g., atom labeling, special positions, space group conversions, etc.). However, comparisons produced via this approach can be difficult to visualize. Another genre of comparisons consists of overlapping packing shells (i.e., sub-clusters) of the desired crystals before calculating a metric that is usually based on distances and/or angles (Gelbrich & Hursthouse, 2005; Rohlicek & Skorepova, 2020; Rohlicek et al., 2016; Chisholm & Motherwell, 2005).

A widely used algorithm that follows this final classification is COMPACK (Chisholm & Motherwell, 2005), which was proposed by the Cambridge Crystallography Data Centre (CCDC, London, UK) (Groom et al., 2016). COMPACK represents the molecular distribution of a specified number of molecules by recording interatomic distances and creates triangular subsets to generate a unique representation of a given crystal that can be compared to other crystals. Two molecules within the sub-crystals match when the difference between their distances is less than a specified distance tolerance (as a percentage) and the angle of their triangles differ by less than a specified angle tolerance (in degrees). This method quantifies crystal similarity regardless of the space group and lattice parameters. However, the implementation of the COMPACK algorithm is relatively slow and has difficulties scaling up to large entities (e.g., proteins and nucleic acids).

In this study, we describe an algorithm for evaluating crystal packing similarity called Progressive Alignment of Crystals (PAC). This algorithm relies on a progressive series of coordinate superpositions to align N molecules. The algorithm performs similarly to COMPACK on small molecule crystals but scales up to biomolecular crystal comparisons. The implementation is faster than available alternatives using a single process and shows favorable parallel scaling to 64 processes. Finally, we introduce the use of metrics to quantify the shape of superimposed clusters (e.g., $R_g$ and/or anisotropy) to avoid ambiguity when reporting results (e.g., for the CCDC blind assessment of CSP).
and help to prioritize molecules during CSP workflows.

2. Materials

2.1. Software

The PAC algorithm is maintained within the Force Field X (FFX) software package that is freely available from GitHub (https://github.com/SchniedersLab/forcefieldx). Further documentation can be found on the Schnieders Laboratory website (https://ffx.biochem.uiowa.edu/). Like most programs in FFX, PAC is written in Java facilitated by a Groovy script and requires version 10 or later of the Java Development Kit. Further assistance for the installation process can be found at the GitHub link above.

The 2021 Cambridge Structural Database Software (version 3.0.4) was utilized for the COMPACK comparisons. A default number of 20 molecules was utilized unless otherwise stated. All COMPACK comparisons were performed with a distance tolerance of 25% and an angle tolerance of 25° unless higher values were necessary for the comparison to succeed (such cases will have the tolerances labeled). All single process timing comparisons were performed on a desktop CPU with an Intel® Core™ i7-9800X CPU (16 cores) at 3.80 GHz running x86_64.

2.2. Data for the Evaluation of the Progressive Alignment of Crystals Algorithm

We have designed the PAC algorithm to be applicable to a wide range of crystal structures. Therefore, the test crystals include molecules/proteins that scale in atom count (4 - 20,409 non-hydrogen atoms) and include both chemical and biological crystals. Each entity, depicted in Figure 1, will be listed as follows: IUPAC name or abbreviation (database abbreviation; molecular formula; space groups).

[ Figure 1 Around Here]

The biological crystals in this study were obtained from the RCSB PDB (http://www.rcsb.org/) (Berman et al., 2000) and are used to demonstrate PAC on larger systems. Two polymorphs were selected for the NNQQ (composed of two asparagine and two glutamine residues) amyloid of the yeast prion sup35 with 35 non-hydrogen atoms (2olx; C18H30N6O9; P212121) and (2onx; C18H30N6O9; P21) (Sawaya et al., 2007). The hen egg white lysozyme (HEWL) hydrolase with 1,001 non-hydrogen atoms (2vb1; P1) (Wang et al., 2007) was chosen to represent small proteins and a cholesterol reductase from Brevibacterium sterolicum with 3,834 non-hydrogen atoms (4rek; P21) (Zarychta et al., 2015) was selected as a midsize protein. The largest protein utilized in this study was ethyl-coenzyme M reductase from Candidatus Ethanoperedens Thermophilum with 20,409 non-hydrogen atoms (7b1s; P21) (Hahn et al., 2021). Both water and co-solutes were removed prior to applying PAC.

All of the chemical structures were accessed from the Cambridge Structural Database(Groom et al., 2016). The smallest molecule included was acetamide with 4 non-hydrogen atoms (ACEMID; C2H5NO; Pccn, H3c). Carbamazepine (5H-dibenzo[b,f]azepine-5-carboxamide) with 18 non-hydrogen atoms (CBMZPN; C15H12N2O; P21/c, P21/n, H3, P1, C2/c, Pbcn) serves as a classic example of crystal
polymorphism (Reboul et al., 1981; Arlin et al., 2011; Lang et al., 2002; Lowes et al., 1987). The largest chemical structure included in this study is ritonavir (1,3-thiazol-5-ylmethyl N-[(2S,3S,5S)-3-hydroxy-5-[(2S)-3-methyl-2-[(methyl-[(2-propan-2-yl]-1,3-thiazol-4-yl) methyl] carbamoyl] amino-] butanoyl] amino]-1,6-diphenylhexan-2-yl] carbamate) with 50 non-hydrogen atoms (YIGPIO; C_{37}H_{48}N_{6}O_{5}S_{2}; P21, P212121). Additionally, the Cambridge Crystallography Data Centre (CCDC, London, UK) has hosted several blind crystal structure prediction (BCSP) competitions which allow participants to apply their algorithms on crystal structures determined via physical experiments (e.g., x-ray crystallography), which have yet to be released to the public. In the BCSP contest held in 2015, participants started from a two-dimensional chemical diagram and predicted one to two list(s) that contained up to 100 predicted crystal structures (Reilly et al., 2016). Compound XXIII or 2-((4-(2-(3,4-dichlorophenyl) ethyl) phenyl amino) benzoic acid with 26 non-hydrogen atoms (XAFPAY; C_{21}H_{17}Cl_{2}N_{2}O_{2}; P1, P21/c, P21/n) (Samas, 2016) was selected to demonstrate how RMSD$_{N}$ rank and R$_{g}$ are effected for participant submissions based on the molecular prioritization criterion for cluster inclusion (i.e., single linkage vs. average linkage).

AMOEBA (Ponder et al., 2010) parameters were generated using the PolType2 (Wu et al., 2012) automatic parameterization program on SDF files obtained from PubChem (Kim et al., 2021). Local optimization of coordinates and lattice parameters of each experimental structure to an energetic convergence criterion of 0.1 kcal/mol/Å was performed according with AMOEBA using Force Field X. The AMOEBA minimization produced crystal polymorphs that were compared to experimental structures using both COMPACK and PAC.

3. The Progressive Alignment of Crystals Algorithm

The six main steps to compare two crystals according to the PAC algorithm follow the flow chart and images in Figure 2 (images and values obtained from single linkage comparison). All alignments in this algorithm are performed via quaternion superposition (Horn, 1987; Kearsley, 1989). Inputs to PAC include the atomic coordinates of asymmetric unit atoms, the space group, and the lattice parameters for two crystals. Although PAC can handle multiple molecules/proteins in the asymmetric unit, the algorithm will be described assuming the asymmetric unit contains a single molecule for simplicity. A subset of atoms can be selected for the comparison (e.g., non-hydrogen atoms, alpha carbons, etc.), which will be more thoroughly discussed in the Discussion section. Mass weighting can be utilized, but comparisons in this work were performed utilizing geometric centers. By default, PAC does not use mass weighting to avoid overprioritizing third period or higher elements (e.g., phosphorous, chlorine, etc.) relative to second period elements. Hydrogen atoms are not included by default as their experimental coordinates are often more uncertain than those for heavier atoms.

[ Figure 2 Around Here]

(I) The molecular coordinates from each structure are expanded through the provided crystallographic information until each crystal occupies a scalar (default of 4) times the expected volume of the final cluster. The expected volume for an RMSD$_{N}$ is calculated by
dividing the volume of the unit cell by the number of molecules it contains and multiplying by \( N \).

**II** The unique molecules are paired between crystals based on a molecular RMSD (i.e., \( \text{RMSD}_1 \)). The number of unique molecules in each crystal is determined based on space group and the number of molecules in the asymmetric unit (\( Z' \)). Crystals in a Sohnece space group are non-enantiogenic (i.e., do not create a non-superimposable copy of the entity) and will have the same number of conformations as \( Z' \). However, enantiogenic space groups create two times \( Z' \) conformations. Therefore, PAC loops through molecules in each crystal (prioritizing molecules closest to the center) and identifies the unique molecular conformations in each crystal.

**III** Molecules are then ranked by the distance of their geometric center to the center of all atoms in the expanded crystal.

**IV** Both crystals are translated so the geometric center of their center-most molecules are at the origin. The central molecule of the second crystal is rotated to achieve optimal superposition on that from the first crystal. For the example in Figure 2, the central molecule has an \( \text{RMSD}_1 \) of 0.068 Å, whereas the \( \text{RMSD}_{20} \) at this stage is 0.684 Å.

**V** The second and third closest molecules from the first crystal (using a specified linkage criterion discussed in the next paragraph) are matched via geometric distance to molecules within the second crystal. The two crystals are aligned based on the three molecules that have been matched between both crystals. The \( \text{RMSD}_3 \) in Figure 2 for this alignment is 0.227 Å, while the \( \text{RMSD}_{20} \) has been reduced to 0.444 Å.

**VI** Finally, \( N \) molecules closest to the central molecule of the first crystal are matched with those from the second crystal and a final coordinate alignment is performed. Coordinates for the selected atoms produced from the final alignment (step VI) are utilized to compute the \( \text{RMSD}_N \). Using this procedure, the example in Figure 2 has an \( \text{RMSD}_{20} \) of 0.302 Å.

The selected molecules for the cluster of the first crystal are known prior to consideration of the second crystal because selection is based only on the linkage method (linkage description to follow). However, the selected molecules for the cluster of the second crystal depend on the distance of molecules between the crystals, which change during steps IV, V and VI above. If the crystals are sufficiently similar (e.g., the example used in Figure 2), then the selected \( N \) molecules for the cluster of the second crystal remain the same and the \( \text{RMSD}_N \) progressively decreases. Steps IV-VI are repeated for each pair of unique molecules between the two crystals. The final \( \text{RMSD}_N \) between the compared crystals is the minimum value produced from the repeated comparisons.

The PAC algorithm supports three linkage criteria, which follow those widely used for hierarchical clustering, to select molecules for cluster inclusion:

- single (shortest atomic distance between two molecules).
- average (shortest distance between average atomic position of two molecules).
- complete (shortest atomic distance for the farthest apart atoms between two molecules).

Depending on the selected linkage criterion, the final cluster shape and \( \text{RMSD}_N \) usually differ as shown
in Figure 3. Structure metrics have previously been used to characterize proteins to assess characteristics of their 3D structures (Šolc, 1971; Blavatska & Janke, 2010). The gyration tensor quantifies the deviation of atoms from their geometric center.

\[
S_{ij} = \frac{1}{N} \sum_{k=1}^{N} (r_i^{(k)} - r_i^{(GC)})(r_j^{(k)} - r_j^{(GC)})
\]

Equation 1

The elements of the gyration tensor (\(S_{ij}\) from Eq. 1) are defined as the summation of the coordinate distance to the geometric center for each of \(N\) atoms where \(i\) and \(j\) denote the \(x\)-, \(y\)-, or \(z\)-coordinate.

\[
R_g^2 = \lambda_{\text{min}} + \lambda_{\text{med}} + \lambda_{\text{max}}
\]

Equation 2

The principal moments of the gyration tensor (with eigenvalues: \(l_{\text{min}}\), \(l_{\text{med}}\), \(l_{\text{max}}\)) equate to the squared characteristic semi-axis lengths that describe the ellipsoid containing the cluster of atoms. The summation of principal moments results in the squared radius of gyration. Reporting \(R_g\) along with \(\text{RMSD}_N\) quantifies if the packing comparison achieved a cluster geometry that equally weights each crystal axis. For the structures compared in this study, single linkage performs most similarly to COMPACK, however, average linkage generally provides a preferable compromise between low \(\text{RMSD}_N\) and low \(R_g\). Other descriptive metrics such as moments of inertia, asphericity, acylindricity, and anisotropy are also reported by the PAC algorithm, but \(R_g\) is generally sufficient to assess the impact of linkage choice. All data generated via complete linkage are in the supplementary information (SI).

4. Results
4.1. Accuracy

Each of the experimentally determined structures listed in Materials were compared to minimized coordinates and lattice parameters (minimization via the AMOEBA force field) utilizing COMPACK, PAC with single linkage, and PAC with average linkage. The comparisons were performed at a comparison shell size of twenty molecules and did not include hydrogen. The RMSDs between the experimental crystals and AMOEBA lattice minimized crystals are plotted in Figure 4.

The average radius of gyration was calculated for each pair of sub crystals generated in the comparisons that produced Figure 4. The radii of gyration for these comparisons are plotted in Figure 5.

We obtained the crystal submissions from the 2015 BCSP and reproduced the COMPACK comparisons (20 molecule shells, distance tolerance of 25% and an angle tolerance of 25°). The crystal structures that successfully produced \(\text{RMSD}_{20}\) values for COMPACK as compared to the experimentally determined polymorphs for XAFPAY were also compared with PAC. The results of the
2015 BCSP competition focused on the ability of contestants to rank their own submissions (i.e., the team that ranked a submission with an RMSD < 0.8 Å higher than another group was considered to have a better prediction regardless of the RMSD from experiment). The ability of the contestants to accurately predict experimental structures (i.e., produce crystals that obtain a low RMSD) is also important. Table 1 contains the RMSD values to the experimental structure XAFPAY01 (polymorph B) for COMPACK and PAC using average linkage (single and complete linkages can be found in SI Table 2). Two such crystal comparisons that were originally included in the SI of the BCSP paper were not reproducible with our version of COMPACK at the reported tolerances. Therefore, we used the values reported previously and replaced the Rg for the clusters with a dash. The structures are ordered based on the computed COMPACK RMSD and their corresponding ranks are presented for PAC using average linkage. Additionally, the average Rg between the compared sub-crystal molecular clusters is reported for each comparison. These PAC comparisons were completed on the Fugaku supercomputer at the Riken Center for Computational Science in Kobe, Japan.

4.2. Performance

COMPACK and PAC were used to perform all vs. all comparisons between 100 crystal structures obtained from a molecular dynamics simulation on the experimental crystal structure using the AMOEBA force field. Relative to COMPACK, all PAC linkage methods display similar comparison times, therefore average linkage will be presented for all figures in the main text. Timing figures utilizing single and complete linkage are included in the SI Figures 3-5. The times presented in Figure 6 are the fastest elapsed CPU time for a single 20 molecule comparison when comparing each of the 100 structures generated from simulation to themselves (total 10,000 comparisons).

The 100 molecular dynamics snapshots for each carbamazepine crystal underwent all vs. all RMSD packing comparisons for increasing values of N = {20, 40, 80} with results shown in Figure 7 (other molecules display similar trends). CBMZPN11 (P1) was left out of the graph as the COMPACK timings extend above 0.2 seconds and would lower its resolution. All PAC comparisons were at least 8 times faster than the corresponding COMPACK timings.

As seen in Figures 6 and 7, an increase in the number of atoms within a cluster increases the computational time necessary to perform a packing comparison. Therefore, it is useful to restrict the number of atoms being compared when possible. In addition to limiting comparisons to non-hydrogen atoms, PAC can operate on protein α-carbon atoms or a custom subset. The use of α-carbon atoms significantly decreases the duration of each comparison as shown in Figure 8.

The RMSD values of the protein crystal comparisons change moderately through exclusion of side chains as shown in Figure 9.
The PAC algorithm can divide comparisons between multiple processes. The comparisons of the 100 molecular dynamics snapshots (RMSDexcluding hydrogen) were scaled up to an all vs all comparison of 1,024 structures (for a total of 1,048,576 comparisons). The parallel comparisons were performed utilizing the Argon HPC cluster maintained at the University of Iowa, with nodes containing two Intel® Xeon® E5-2680 v4 CPUs at 2.40 GHz. Each parallel comparison (regardless of the number of processes) was allocated three 512 GB memory node, which consisted of 56 hyperthreaded cores (28 physical cores). Two hyperthreaded cores were assigned to each process, which limited each Argon node to a maximum of 28 processes. Algorithm logging was reduced and comparison results were written to a text file to promote maximum efficiency. The same PAC comparisons were performed while doubling the number of processes as shown in Figure 10. PAC presents moderately decreasing efficiency gains as more nodes are utilized ranging from 1.99 x speed up at two nodes to a 44.4 x speed up at 64 nodes (~70% efficiency resulting in more than 5,500 comparisons per second at 64 nodes).

5. Discussion

Crystal packing comparison methods compute the coordinate RMSD for a cluster of N molecules, however, the shape of the compared clusters is typically not reported. While the lowest possible RMSD may result from elongated clusters that prioritize accurate packing along a single dimension, uniform prioritization of packing in all 3 dimensions serves to minimize the radius of gyration. Just as the global distance test (GDT) is of central importance the Critical Assessment of Structure Prediction (Moult et al., 1995), RMSD serves as the gold standard for comparing entries to the CCDC CSP blind tests with experiment. By reporting radius of gyration along with RMSD the shape of the compared clusters (i.e., elongated vs. spherical) can be appreciated and ambiguity reduced. Generally, single linkage yields lower RMSD at the cost of higher Rg, and more closely replicates COMPACK (Figures 4a and 5a). Based on the data reported here, average linkage results in clusters that more equally prioritize all 3 dimensions and thereby lowers Rg with only modestly higher RMSD values (Figures 4b and 5b). As seen in Table 1, the order of crystals based on RMSD seems to change minimally between COMPACK and PAC with single linkage, however, average linkage had several structures whose rank increased significantly (highlighted in light blue). Each of these predictions had their rank increased by at least 15 places when using average linkage, which suggests that their crystal packing is more closely related to experiment when more spherical clusters are selected. The incorporation of Rg improves the robustness of PAC by encouraging a selection of molecules that do not favor a specific orientation. When the unit cell volumes differ dramatically between two crystals, it is possible that PAC (and COMPACK) can inappropriately quantify the crystal similarity with a low RMSD if large sections of the two crystals are similar (See SI Table 2). Increasing the number of molecules included in the comparison can improve the fidelity of PAC with modest loss of efficiency. Multiplying the default number of molecules by a factor of volume change worked well for the provided test systems (e.g., if one unit cell is roughly four times greater than the other then a comparison cluster of 80 molecules could be used).
The efficiency increase of the PAC algorithm has implications for crystal structure prediction, where many candidate packings are generated and must be compared. Relative to COMPACK, the computational cost of PAC comparisons scales more favorably as the number of atoms increases, which allows it to scale up to larger crystals (e.g., proteins, nucleic acids, etc.). PAC also maintains efficiency for packing comparisons as the number of molecules N increases (Figures 6 and 7). Finally, PAC leverages the non-enantiomorphic nature of Sohncke groups featured in most biological crystals for additional efficiency. Inclusion of all non-hydrogen atoms in the packing comparison is recommended when efficiency is not a limiting factor, however, the ability to select a subset of atoms provides performance improvements (Figures 8 and 9). For example, the exclusion of side-chain atoms tends to slightly reduce the RMSD_N for large proteins, as the algorithm focuses exclusively on alignment of amino acid backbone conformation. The PAC algorithm is parallelized over processes using MPI to accelerate performing large batches of comparisons. Comparison times can be significantly reduced using parallel processors (Figure 10). Furthermore, average linkage has improved efficiency over the other PAC linkage methods (single and complete) as all the atoms per constituent are condensed into a single point, which vastly reduces the number of distances that need to be evaluated.

6. Conclusions

We propose the PAC algorithm for evaluating the similarity of two crystal structures. The results demonstrate that PAC is an accurate and efficient method to evaluate the similarity of two crystal structures. PAC employs a progressive series of coordinate alignments to optimize RMSD_N. The RMSD_N values obtained by PAC agree with those obtained from the widely used program COMPACK when using single linkage to prioritize molecules for inclusion in the superimposed clusters. PAC performed an average of 17 times faster than COMPACK when computing multiple comparisons for the carbamazepine polymorphs. We suggest that utilization of cluster shape metrics such as radius of gyration helps to avoid the ambiguity inherent to reporting RMSD_N alone. PAC has many potential applications, including identification and removal of duplicate crystal structure candidates during CSP and the comparison of optimized structures to experiment.

Acknowledgements

Computations were performed on 1) the University of Iowa Argon cluster with support and guidance from Joe Hetrick, Glenn Johnson, and John Saxton and 2) on the Fugaku supercomputer provided by Riken Center through the HPCI System Research Project (Project: hp210200). AJN, MJH and MJS were supported by NSF CHE-1751688. Mitsubishi Tanabe Pharma Corp. provided partial support for Aaron Nessler during the preparation of this work.

Contributions

Aaron Nessler and Okimasa Okada contributed equally to the work and are joint first authors. Okimasa Okada conceived the novel algorithm for crystal packing comparison via a series of
coordinate transformations to optimize $RMSD_N$. Aaron Nessler enhanced the original algorithm to develop PAC, ported it into the publicly available Force Field X software package, and incorporated $R_g$ as a metric to remove ambiguity in the shape of the compared clusters. Hiroomi Nagata and Michael Schnieders are both senior authors for this collaboration. They contributed by supervising the work, suggesting refinements to the PAC algorithm, and assisting with drafting the manuscript.
Figure 1  PyMol renderings of the molecules and proteins used to test the PAC algorithm (Schrödinger, 2015). Structures with four alphanumerical characters are from the PDB and those with six letters are from the CSD (see text for citations).
Figure 2 A general overview of the PAC algorithm consists of a progressive series of alignments to optimize RMSD between superimposed clusters with N molecules. The six basic steps for the algorithm are listed in the flow chart with crystal alignments emphasized as superimposed images. This example comparison was performed using single linkage to prioritize addition of molecules into the clusters. The RMSD between similar crystals improves as the alignment progresses.
Figure 3 Different linkage methods effect molecular cluster shape, RMSD$_{20}$, and radius of gyration ($R_g$).
The diagram shows a scatter plot with a linear regression line. The x-axis represents COMPACK RMSD₂₀ (Å), and the y-axis represents Single Linkage PAC RMSD₂₀ (Å). The regression line equation is $y = 0.992x$ with $R^2 = 0.997$. The data points are labeled with various crystallographic space groups.
Figure 4: Output metrics for COMPACK are plotted on the X-axis while PAC results are plotted on the Y-axis. PAC with single linkage produces similar RMSD values to COMPACK as demonstrated by the regression slope of 0.992 in the left panel. The RMSD values for PAC with average linkage tend to be slightly larger than those for COMPACK and for PAC with single linkage which is plotted in the right panel.
Figure 5 Crystal packing comparison algorithms use a range of criteria to prioritize molecules for inclusion in superimposed clusters, which effects both RMSD$_{20}$ and cluster shape as quantified by radius of gyration. Radius of gyration values from COMPACK are similar to those from PAC with single linkage based on clusters selected for RMSD$_{20}$ as demonstrated by the left panel. Radius of gyration values from PAC with average linkage are significantly smaller which is shown in the right panel.
Figure 6 Packing comparison computational cost increases with number of atoms. COMPACK and PAC timings are represented by diamonds and circles, respectively. Entity is color coded according to the legend. The time presented is the fastest out of 100 RMSD$_{20}$ trials.
Figure 7 Packing comparison computational cost increases with the number of molecules $N$ included in the cluster. COMPACK and PAC are represented by diamonds and circles, respectively. The time presented is the fastest out of 100 identical trials.
Comparisons using a specified subset of atoms can significantly decrease calculation time. The durations shown are the fastest RMSD comparison out of 100 trials between two protein crystals. Circles represent the default PAC algorithm whereas triangles depict a comparison limited to alpha carbons. Log scales are utilized to allow all proteins to be displayed on the same graph.

**Figure 8** Comparisons using a specified subset of atoms can significantly decrease calculation time. The durations shown are the fastest RMSD comparison out of 100 trials between two protein crystals. Circles represent the default PAC algorithm whereas triangles depict a comparison limited to alpha carbons. Log scales are utilized to allow all proteins to be displayed on the same graph.
Figure 9 Restricting protein comparisons to only consider alpha carbons results in a modest change in the RMSD\textsubscript{20} values for the PAC algorithm. The abscissa shows RMSD\textsubscript{20} values when using all heavy atoms for the comparison while the ordinate is restricted to alpha carbons. The left panel displays results with single linkage whereas the right panel contains average linkage.
Figure 10 Ritonavir packing comparison performance is shown for the PAC algorithm when utilizing 1 to 32 processes. The ordinate shows the wall clock time necessary for PAC to perform over one million (1,048,576) comparisons with the number of processes given on the abscissa.
Table 1 RMSD$_{20}$ values for packing comparisons between experiment (XAFPAY01) and submissions to the CCDC’s 2015 BCSP assessment depend on the algorithm used. The ranking for many entries using PAC with average linkage is similar to that from COMPACK, but a few deviate significantly (highlighted in light blue).
| Submission: Team (Rank, List) | COMPACK | Average Linkage |
|-----------------------------|---------|----------------|
| Neuman, Kendrick, Leusen (R26, L2) | 1 0.218 13.37 | 1 0.323 11.37 |
| Neuman, Kendrick, Leusen (R04, L2) | 2 0.229 13.37 | 2 0.328 11.36 |
| Neuman, Kendrick, Leusen (R02, L1) | 2 0.229 13.41 | 2 0.328 11.36 |
| Price et al. (R05, L1) | 4 0.286 15.92 | 4 0.359 11.38 |
| Tkatchenko et al. (Price) (R05, L2) | 5 0.294 14.20 | 5 0.435 11.34 |
| Tkatchenko et al. (Price) (R02, L1) | 5 0.294 14.22 | 5 0.435 11.34 |
| Brandenburg & Grimme (Price) (R04, L2) | 7 0.330 14.29 | 10 0.498 11.31 |
| Brandenburg & Grimme (Price) (R08, L2) | 8 0.334 14.23 | 9 0.469 11.32 |
| Price et al. (R02, L2) | 9 0.339 15.70 | 8 0.444 11.38 |
| Price et al. (R01, L1) | 10 0.340 15.74 | 7 0.442 11.38 |
| Brandenburg & Grimme (Price) (R02, L2) | 11 0.349 14.34 | 12 0.529 11.31 |
| Brandenburg & Grimme (Price) (R03, L2) | 12 0.369 14.32 | 16 0.550 11.31 |
| Brandenburg & Grimme (Price) (R01, L2) | 13 0.391 14.36 | 18 0.573 11.35 |
| Brandenburg & Grimme (Price) (R26, L1) | 14 0.392 15.54 | 17 0.554 11.30 |
| Brandenburg & Grimme (Price) (R31, L1) | 15 0.394 15.36 | 27 0.625 11.35 |
| Brandenburg & Grimme (Price) (R06, L2) | 16 0.396 14.25 | 19 0.586 11.32 |
| Brandenburg & Grimme (Price) (R37, L1) | 17 0.403 14.91 | 34 0.648 11.35 |
| Brandenburg & Grimme (Price) (R38, L1) | 18 0.405 14.85 | 20 0.589 11.30 |
| Brandenburg & Grimme (Price) (R45, L1) | 19 0.409 14.64 | 35 0.657 11.35 |
| Brandenburg & Grimme (Price) (R07, L2) | 20 0.412 14.23 | 24 0.613 11.35 |
| Brandenburg & Grimme (Price) (R39, L1) | 21 0.412 14.79 | 23 0.608 11.30 |
| Brandenburg & Grimme (Price) (R05, L2) | 22 0.414 14.27 | 22 0.601 11.35 |
| Brandenburg & Grimme (Price) (R57, L1) | 23 0.416 14.44 | 31 0.632 11.31 |
| Brandenburg & Grimme (Price) (R34, L1) | 24 0.418 15.09 | 25 0.618 11.30 |
| Brandenburg & Grimme (Price) (R36, L1) | 25 0.420 14.99 | 37 0.675 11.35 |
| Brandenburg & Grimme (Price) (R32, L1) | 26 0.421 15.20 | 26 0.624 11.30 |
| Brandenburg & Grimme (Price) (R46, L1) | 27 0.424 14.60 | 38 0.683 11.35 |
| Brandenburg & Grimme (Price) (R61, L1) | 28 0.425 14.39 | 30 0.628 11.30 |
| Brandenburg & Grimme (Price) (R47, L1) | 29 0.426 14.56 | 33 0.644 11.31 |
| Brandenburg & Grimme (Price) (R59, L1) | 30 0.427 14.42 | 28 0.628 11.30 |
| Brandenburg & Grimme (Price) (R56, L1) | 31 0.428 14.47 | 29 0.628 11.30 |
| van Eijck (R20, L1) | 32 0.430 14.23 | 13 0.533 11.47 |
| Brandenburg & Grimme (Price) (R52, L1) | 33 0.434 14.51 | 32 0.639 11.30 |
| Elking & Fusti-Molnar (R78, L1) | 34 0.434 14.23 | 14 0.536 11.43 |
| Brandenburg & Grimme (Price) (R42, L1) | 35 0.437 14.72 | 39 0.701 11.35 |
| Brandenburg & Grimme (Price) (R44, L1) | 36 0.448 14.68 | 36 0.658 11.30 |
| Pantelides, Adjiman et al. (R21, L1) | 37 0.455 14.08 | 15 0.544 11.40 |
| Obata & Goto (R13, L1) | 38 0.495 14.22 | 21 0.595 11.54 |
| Brandenburg & Grimme (Price) (R11, L1) | 39 0.524 14.19 | 11 0.515 11.33 |
| Day et al. (R75, L2) | 40 0.601 14.19 | 40 0.741 11.48 |
| Pantelides, Adjiman et al. (R13, L1) | 41 0.604 13.37 | 41 0.793 11.45 |
| Mohamed (R88, L1) | 42 0.827 14.52 | 42 0.843 11.55 |
| Average Values | - 0.408 14.52 | - 0.573 11.36 |

26
References
Arlin, J.-B., Price, L. S., Price, S. L. & Florence, A. J. (2011). ChemComm 47, 7074-7076.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). Nucleic Acids Res. 28, 235-242.

Blagden, N., de Matas, M., Gavan, P. T. & York, P. (2007). Adv Drug Deliv. Rev. 59, 617-630.

Blavatska, V. & Janke, W. (2010). J. Chem. Phys. 133, 184903.

Burger, V., Claeyssens, F., Davies, D. W., Day, G. M., Dyer, M. S., Hare, A., Li, Y., Mellot-Draznieks, C., Mitchell, J. B. O., Mohamed, S., Oganov, A. R., Price, S. L., Ruggiero, M., Ryder, M. R., Sastre, G., Schön, J. C., Spackman, P., Woodley, S. M. & Zhu, Q. (2018). Faraday Discuss. 211, 613-642.

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

Chuanxun, S., Lv, J., Li, Q., Wang, H., Zhang, L., Wang, Y. & Ma, Y. (2017). Journal of Condensed Matter Physics 29.

Day, G. M. (2011). Crystallogr. Rev. 17, 3-52.

de la Flor, G., Orobengoa, D., Tasci, E., Perez-Mato, J. M. & Aroyo, M. I. (2016). J. Appl. Crystallogr. 49, 653-664.

De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. (2016). Phys. Chem. Chem. Phys. 18, 13754-13769.

Dzyabchenko, A. (1994). Acta Crystallogr. B 50, 414-425.

Edelsbrunner, H., Heiss, T., Vitaliy, K., Smith, P. & Wintraecken, M. (2021). arXiv 32, 1-16.

Ferré, G., Maillet, J.-B. & Stoltz, G. (2015). J. Chem. Phys. 143, 104114.

Furukawa, H., Ko, N., Go, Y. B., Aratani, N., Choi, S. B., Choi, E., Yazaydin, A. O., Snurr, R. Q., O'Keeffe, M., Kim, J. & Yaghi, O. M. (2010). Science 329, 424-428.

Gelato, L. M. & Parthe, E. (1987). J. Appl. Crystallogr. 20, 139-143.

Gelbrich, T. & Hursthouse, M. B. (2005). CrystEngComm 7, 324-336.

Gelder, R., Wehrens, R. & Hageman, J. (2001). J. Comput. Chem. 22, 273-289.

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). Acta Crystallogr. B 72, 171-179.
Hahn, C. J., Lemaire, O. N., Kahnt, J., Engilberge, S., Wegener, G. & Wagner, T. (2021). *Science* **373**, 118-121.

Haisa, M., Kashino, S., Kawai, R. & Maeda, H. (1976). *Acta Crystallogr. B* **32**, 1283-1285.

Haisa, M., Kashino, S. & Maeda, H. (1974). *Acta Crystallogr. B* **30**, 2510-2512.

Hicks, D., Toher, C., Ford, D. C., Rose, F., Santo, C. D., Levy, O., Mehl, M. J. & Curtarolo, S. (2021). *NPJ Comput. Mater.* **7**, 30.

Horn, B. K. P. (1987). *J. Opt. Soc. Am. A* **4**, 629-642.

Hundt, R., Schon, J. C. & Jansen, M. (2006). *J. Appl. Crystallogr.* **39**, 6-16.

James, S. L. (2003). *Chem. Soc. Rev.* **32**, 276-288.

Kapczynski, A., Park, C. & Sampat, B. (2012). *PLOS ONE* **7**, e49470.

Karamertzantis, P. G., Kazantsev, A. V., Issa, N., Welch, G. W. A., Adjiman, C. S., Pantelides, C. C. & Price, S. L. (2009). *J. Chem. Theory Comput.* **5**, 1432-1448.

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

Kearsley, S. (1989). *Acta Crystallogr. A* **45**, 208-210.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J. & Bolton, E. E. (2021). *Nucleic Acids Res.* **49**, D1388-D1395.

Lang, M., Kampf, J. W. & Matzger, A. J. (2002). *J. Pharm. Sci.* **91**, 1186-1190.

Leelananda, S. P. & Lindert, S. (2016). *Beilstein J. Org. Chem.* **12**, 2694-2718.

Lonie, D. C. & Zurek, E. (2012). *Comput. Phys. Commun.* **183**, 690-697.

Lowes, M. M. J., Caira, M. R., Lötter, A. P. & Van Der Watt, J. G. (1987). *J. Pharm. Sci.* **76**, 744-752.

Moe, O. W. (2006). *Lancet* **367**, 333-344.

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

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. (1995). *Proteins* **23**, ii-v.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A. & Ceder, G. (2013). *Comput. Mater. Sci.* **68**, 314-319.
Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio, R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E. & Head-Gordon, T. (2010). *J. Phys. Chem. B* **114**, 2549-2564.

Price, S. (2008). *Acc. Chem. Res.* **42**, 117-126.

Price, S. L. (2014). *Chem. Soc. Rev.* **43**, 2098-2111.

Price, S. L. & Price, L. S. (2011). *Solid State Characterization of Pharmaceuticals*, pp. 427-450.

Reboul, J. P., Cristau, B., Soifer, J. C. & Astier, J. P. (1981). *Acta Crystallogr. B* **37**, 1844-1848.

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., Fusi-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**, 439-459.

Rohlicek, J. & Skorepova, E. (2020). *J. Appl. Crystallogr.* **53**, 841-847.

Rohlicek, J., Skorepova, E., Babor, M. & Cejka, J. (2016). *J. Appl. Crystallogr.* **49**, 2172-2183.

Sadeghi, A., Ghasemi, S. A., Schaefer, B., Mohr, S., Lill, M. A. & Goedecker, S. (2013). *J. Chem. Phys.* **139**, 184118.

Samas, B. (2016).

Sawaya, M. R., Sambashivan, S., Nelson, R., Ivanova, M. I., Sievers, S. A., Apostol, M. I., Thompson, M. J., Balbirnie, M., Wiltzius, J. J., McFarlane, H. T., Madsen, A., Riekel, C. & Eisenberg, D. (2007). *Nature* **447**, 453-457.

Schrödinger, L. (2015). *PyMOL The PyMOL molecular graphics system, version 2.4.0*. Schrödinger,
Šolc, K. (1971). *J. Chem. Phys.* **55**, 335-344.

Terkeltaub, R. (2010). *Nature Reviews Rheumatology* **6**, 30-38.

Thomas, J. C., Natarajan, A. R. & Van der Ven, A. (2021). *NPJ Comput. Mater.* **7**, 164.

Valle, M. & Oganov, A. R. (2010). *Acta Crystallogr. A* **66**, 507-517.

Verwer, P. & Leusen, F. J. J. (1998). *Rev. Comput. Chem.* **12**, 327-365.

Vishweshwar, P., McMahon, J. A., Oliveira, M., Peterson, M. L. & Zaworotko, M. J. (2005). *Journal of the American Chemical Society* **127**, 16802-16803.

Wang, J., Dauter, M., Alkire, R., Joachimiak, A. & Dauter, Z. (2007). *Acta Crystallographica Section D: Structural Biology* **63**, 1254-1268.

Wheatley, P. J. (1964). *J. Chem. Soc.* 6036-6048.

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

Willighagen, E. L., Wehrens, R., Verwer, P., de Gelder, R. & Buydens, L. M. C. (2005). *Acta Crystallogr. B* **61**, 29--36.

Wu, J. C., Chattree, G. & Ren, P. (2012). *Theor. Chem. Acc.* **131**, 1138.

Zarychta, B., Lyubimov, A., Ahmed, M., Munshi, P., Guillot, B., Vrielink, A. & Jelsch, C. (2015). *Acta Crystallographica Section D: Structural Biology* **71**, 954-968.

*Supplementary Information*
Figure S1 Comparing the use of complete linkage to select asymmetric units with COMPACK produces the RMSD plot above.
Figure S2 The use of complete linkage to select asymmetric units and comparing the $R_g$ values produces the plot above.
Figure S3 Comparison times for each of the linkages are relatively similar when compared to COMPACK. Although single and complete linkage have very similar comparison times, average linkage shows reduced comparison times when applied to larger systems.
Figure S4 Increasing the number of molecules included in the comparison ($RMSD_x$) results in slight differences between single and average linkages. Timings for single and complete linkages are essentially equivalent.
Figure S5 Comparisons gain a similar increase to efficiency when limiting protein comparisons to alpha carbons regardless of linkage used. The left panel uses single linkage whereas the right panel uses complete linkage.
Figure S6 This figure demonstrates the accuracy change for utilizing only alpha carbons with complete linkage.
Figure S7 When scaling up to larger systems, the radii of gyration appear to be consistent between single, average, and complete linkages.
Table S1 PAC and COMPACK RMSD and \( R_g \) values for crystal comparisons of the chosen structures for this study.

| CCDC ID | COMPACK | Single | Average | Complete |
|---------|----------|--------|---------|----------|
|         | RMSD (Å) | Rg (Å) | RMSD (Å) | Rg (Å) | RMSD (Å) | Rg (Å) | RMSD (Å) | Rg (Å) |
| ACEDIM  | 0.204 | 5.775 | 0.198 | 5.867 | 0.225 | 5.843 | 0.230 | 5.877 |
| ACEDIM01| 0.274 | 6.127 | 0.265 | 6.125 | 0.341 | 5.824 | 0.362 | 6.120 |
| CBMZPN01| 0.367 | 9.123 | 0.376 | 9.149 | 0.363 | 9.065 | 0.372 | 9.007 |
| CBMZPN02| 0.223 | 9.162 | 0.229 | 9.178 | 0.226 | 9.074 | 0.233 | 9.026 |
| CBMZPN03| 0.285 | 9.482 | 0.305 | 9.398 | 0.302 | 9.278 | 0.293 | 9.149 |
| CBMZPN11| 0.150 | 9.209 | 0.174 | 9.259 | 0.165 | 9.284 | 0.177 | 9.045 |
| CBMZPN12| 0.385 | 9.191 | 0.377 | 8.963 | 0.393 | 9.112 | 0.394 | 8.966 |
| CBMZPN16| 0.396 | 9.093 | 0.396 | 9.157 | 0.398 | 9.005 | 0.402 | 8.957 |
| XAFPAY  | 0.655 | 13.738 | 0.673 | 13.606 | 0.782 | 11.612 | 0.931 | 12.599 |
| XAFPAY01| 0.219 | 13.741 | 0.217 | 11.827 | 0.214 | 11.456 | 0.213 | 12.165 |
| XAFPAY02| 0.333 | 13.885 | 0.322 | 13.838 | 0.334 | 11.485 | 0.334 | 11.997 |
| XAFPAY03| 0.215 | 14.263 | 0.204 | 12.188 | 0.208 | 11.534 | 0.204 | 12.552 |
| XAFPAY04| 0.296 | 13.739 | 0.297 | 13.781 | 0.342 | 11.472 | 0.356 | 12.076 |
| 2olx    | 0.926 | 14.749 | 0.915 | 13.563 | 0.892 | 12.358 | 0.921 | 12.283 |
| 2onx    | 0.979 | 13.518 | 0.944 | 13.395 | 0.963 | 12.301 | 0.952 | 12.520 |
| YIGPIO02| 0.683 | 14.950 | 0.685 | 14.931 | 0.743 | 14.170 | 0.823 | 14.942 |
| YIGPIO03| 0.352 | 16.234 | 0.353 | 16.239 | 0.348 | 14.512 | 0.368 | 15.067 |
| 2vb1    | 1.019 | 41.826 | 1.016 | 40.135 | 1.017 | 41.338 |
| 4rek    | 0.845 | 65.410 | 0.856 | 66.135 | 0.856 | 66.976 |
| 7b1s    | 0.977 | 150.718 | 0.995 | 135.945 | 0.995 | 139.601 |
| 2olx    |         |         | 0.962 | 13.086 | 0.927 | 11.866 | 0.966 | 11.889 |
| 2onx    |         |         | 0.989 | 13.208 | 0.944 | 11.938 | 0.961 | 11.875 |
| 2vb1    | 0.812 | 41.428 | 0.807 | 40.054 | 0.812 | 41.375 |
| 4rek    | 0.693 | 65.312 | 0.685 | 66.164 | 0.693 | 66.787 |
| 7b1s    | 0.823 | 150.559 | 0.844 | 135.779 | 0.844 | 139.433 |
Table S2 BCSP comparison values for all submitted structures that were comparable to experimental polymorph B of XAFPAY at 25% bond and 25° angle tolerances. The comparison shaded in orange corresponds to a submitted crystal with vastly increased unit cell size. Single linkage happened to utilize portions of the crystals that do not align well which better represents the similarity between this structure and experiment more so than the similar portions of the crystals alone. As described in the discussion, comparisons featuring crystals with large differences in unit cell volume will benefit from including more molecules (i.e., larger RMSDx).
| Submission: Team (Rank, List) | Single Linkage | Complete Linkage |
|-----------------------------|---------------|------------------|
|                             | RANK | RMSD<sub>20</sub> (Å) | R<sub>g</sub> (Å) | RANK | RMSD<sub>20</sub> (Å) | R<sub>g</sub> (Å) |
| Neuman, Kendrick, Leusen (R26, L2) | 42   | 2.150             | 13.83        | 1    | 0.391             | 12.07        |
| Neuman, Kendrick, Leusen (R04, L2) | 1    | 0.244             | 14.01        | 2    | 0.408             | 12.08        |
| Neuman, Kendrick, Leusen (R02, L1) | 1    | 0.244             | 14.01        | 2    | 0.408             | 12.08        |
| Price et al. (R05, L1) | 3    | 0.288             | 14.07        | 4    | 0.414             | 12.10        |
| Tkatchenko et al. (Price) (R05, L2) | 5    | 0.352             | 12.16        | 6    | 0.550             | 12.15        |
| Tkatchenko et al. (Price) (R02, L1) | 5    | 0.352             | 12.16        | 6    | 0.550             | 12.15        |
| Brandenburg & Grimme (Price) (R04, L2) | 7    | 0.352             | 13.99        | 12   | 0.588             | 12.03        |
| Brandenburg & Grimme (Price) (R08, L2) | 4    | 0.350             | 14.34        | 10   | 0.585             | 12.01        |
| Price et al. (R02, L2) | 10   | 0.398             | 11.84        | 8    | 0.551             | 12.18        |
| Price et al. (R01, L1) | 9    | 0.396             | 11.84        | 5    | 0.546             | 12.18        |
| Brandenburg & Grimme (Price) (R02, L2) | 8    | 0.393             | 13.28        | 18   | 0.666             | 12.11        |
| Brandenburg & Grimme (Price) (R03, L2) | 17   | 0.441             | 12.12        | 16   | 0.651             | 12.02        |
| Brandenburg & Grimme (Price) (R01, L2) | 11   | 0.411             | 13.99        | 24   | 0.714             | 12.03        |
| Brandenburg & Grimme (Price) (R26, L1) | 12   | 0.415             | 13.99        | 14   | 0.642             | 12.00        |
| Brandenburg & Grimme (Price) (R31, L1) | 13   | 0.424             | 14.03        | 17   | 0.666             | 11.99        |
| Brandenburg & Grimme (Price) (R06, L2) | 16   | 0.435             | 13.29        | 28   | 0.734             | 12.12        |
| Brandenburg & Grimme (Price) (R37, L1) | 28   | 0.494             | 12.12        | 21   | 0.698             | 11.99        |
| Brandenburg & Grimme (Price) (R38, L1) | 29   | 0.495             | 12.12        | 19   | 0.679             | 11.99        |
| Brandenburg & Grimme (Price) (R45, L1) | 32   | 0.502             | 12.12        | 22   | 0.706             | 11.99        |
| Brandenburg & Grimme (Price) (R07, L2) | 19   | 0.450             | 13.28        | 30   | 0.743             | 12.04        |
| Brandenburg & Grimme (Price) (R39, L1) | 18   | 0.442             | 14.03        | 23   | 0.712             | 11.99        |
| Brandenburg & Grimme (Price) (R05, L2) | 14   | 0.432             | 13.99        | 31   | 0.750             | 12.04        |
| Brandenburg & Grimme (Price) (R57, L1) | 33   | 0.509             | 12.12        | 25   | 0.717             | 12.00        |
| Brandenburg & Grimme (Price) (R34, L1) | 34   | 0.513             | 12.12        | 32   | 0.773             | 12.08        |
| Brandenburg & Grimme (Price) (R36, L1) | 35   | 0.516             | 12.12        | 33   | 0.775             | 12.08        |
| Brandenburg & Grimme (Price) (R32, L1) | 36   | 0.518             | 12.12        | 34   | 0.781             | 12.08        |
| Brandenburg & Grimme (Price) (R46, L1) | 37   | 0.520             | 12.12        | 35   | 0.788             | 12.08        |
| Brandenburg & Grimme (Price) (R61, L1) | 20   | 0.455             | 14.02        | 36   | 0.792             | 12.09        |
| Brandenburg & Grimme (Price) (R47, L1) | 24   | 0.486             | 13.30        | 26   | 0.726             | 12.00        |
| Brandenburg & Grimme (Price) (R59, L1) | 26   | 0.492             | 12.73        | 29   | 0.738             | 12.00        |
| Brandenburg & Grimme (Price) (R56, L1) | 25   | 0.488             | 13.30        | 27   | 0.734             | 12.00        |
| van Eijck (R20, L1) | 21   | 0.463             | 13.52        | 13   | 0.592             | 12.21        |
| Brandenburg & Grimme (Price) (R52, L1) | 31   | 0.502             | 12.74        | 38   | 0.799             | 12.08        |
| Elking & Fusti-Molnar (R78, L1) | 22   | 0.473             | 13.52        | 11   | 0.586             | 12.14        |
| Brandenburg & Grimme (Price) (R42, L1) | 30   | 0.499             | 13.30        | 39   | 0.808             | 12.08        |
| Brandenburg & Grimme (Price) (R44, L1) | 38   | 0.547             | 12.12        | 40   | 0.820             | 12.08        |
| Pantelides, Adjiman et al. (R21, L1) | 23   | 0.483             | 12.51        | 15   | 0.646             | 12.22        |
| Obata & Goto (R13, L1) | 27   | 0.494             | 14.11        | 20   | 0.682             | 12.33        |
| Brandenburg & Grimme (Price) (R11, L1) | 15   | 0.434             | 13.94        | 9    | 0.568             | 12.01        |
| Day et al. (R75, L2) | 39   | 0.637             | 13.53        | 37   | 0.797             | 12.26        |
| Pantelides, Adjiman et al. (R13, L1) | 40   | 0.644             | 12.95        | 41   | 0.924             | 12.24        |
| Mohamed (R88, L1) | 41   | 0.721             | 14.09        | 42   | 1.069             | 12.41        |
| **Average Values** | -    | 0.496             | 13.12        | -    | 0.678             | 12.09        |
Table S3 PAC and COMPACK RMSD\textsuperscript{20} timings for the fastest single comparison of many.

| CCDC ID | COMPACK | Single | Average | Complete |
|---------|---------|--------|---------|----------|
| ACEMID  | 0.0190  | 0.0010 | 0.0011  | 0.0009   |
| ACEMID01| 0.0486  | 0.0019 | 0.0020  | 0.0018   |
| CBMZPN01| 0.0430  | 0.0028 | 0.0014  | 0.0026   |
| CBMZPN02| 0.0431  | 0.0024 | 0.0012  | 0.0022   |
| CBMZPN03| 0.0413  | 0.0047 | 0.0023  | 0.0043   |
| CBMZPN11| 0.2284  | 0.0065 | 0.0035  | 0.0062   |
| CBMZPN12| 0.0415  | 0.0028 | 0.0014  | 0.0026   |
| CBMZPN16| 0.0404  | 0.0026 | 0.0013  | 0.0023   |
| XAFPAY  | 0.0547  | 0.0035 | 0.0012  | 0.0030   |
| XAFPAY01| 0.0564  | 0.0040 | 0.0015  | 0.0035   |
| XAFPAY02| 0.1240  | 0.0075 | 0.0026  | 0.0065   |
| XAFPAY03| 0.0547  | 0.0041 | 0.0015  | 0.0036   |
| XAFPAY04| 0.1154  | 0.0060 | 0.0021  | 0.0053   |
| 2olx    | 0.1599  | 0.0035 | 0.0012  | 0.0044   |
| 2onx    | 0.1349  | 0.0029 | 0.0010  | 0.0035   |
| YIGPIO02| 0.2218  | 0.0077 | 0.0019  | 0.0077   |
| YIGPIO03| 0.2833  | 0.0054 | 0.0014  | 0.0058   |
| 2vb1    | 1.5295  | 0.0244 | 1.5305  |           |
| 4rek    | 25.4530 | 0.1173 | 34.9250 |           |
| 7b1s    | 718.4660| 0.7000 | 1076.4070|          |
### Table S4 Comparison timings for increasing shell sizes of carbamazepine polymorphs (RMSDₜ).  

| CCDC ID | CBMZPN01 | CBMZPN02 | CBMZPN03 | CBMZPN11 | CBMZPN12 | CBMZPN16 |
|---------|----------|----------|----------|----------|----------|----------|
| **COMPACK** | | | | | | |
| RMSD₂₀ | 0.043 | 0.043 | 0.041 | 0.228 | 0.042 | 0.040 |
| Time (s) | 0.101 | 0.103 | 0.090 | 0.486 | 0.095 | 0.095 |
| RMSD₄₀ | 0.185 | 0.186 | 0.161 | 0.829 | 0.174 | 0.168 |
| Time (s) | 0.185 | 0.186 | 0.161 | 0.829 | 0.174 | 0.168 |
| **Single** | | | | | | |
| RMSD₂₀ | 0.003 | 0.002 | 0.005 | 0.007 | 0.003 | 0.003 |
| Time (s) | 0.004 | 0.004 | 0.006 | 0.018 | 0.004 | 0.004 |
| RMSD₄₀ | 0.009 | 0.008 | 0.008 | 0.024 | 0.011 | 0.010 |
| Time (s) | 0.009 | 0.008 | 0.008 | 0.024 | 0.011 | 0.010 |
| **Average** | | | | | | |
| RMSD₂₀ | 0.001 | 0.001 | 0.002 | 0.004 | 0.001 | 0.001 |
| Time (s) | 0.002 | 0.002 | 0.003 | 0.009 | 0.002 | 0.002 |
| RMSD₄₀ | 0.005 | 0.004 | 0.004 | 0.013 | 0.006 | 0.005 |
| Time (s) | 0.005 | 0.004 | 0.004 | 0.013 | 0.006 | 0.005 |
| **Complete** | | | | | | |
| RMSD₂₀ | 0.003 | 0.002 | 0.004 | 0.006 | 0.003 | 0.002 |
| Time (s) | 0.004 | 0.004 | 0.006 | 0.016 | 0.004 | 0.004 |
| RMSD₄₀ | 0.008 | 0.007 | 0.007 | 0.022 | 0.010 | 0.009 |
| Time (s) | 0.008 | 0.007 | 0.007 | 0.022 | 0.010 | 0.009 |
Table S5 Explicit values of the parallelized computation timings for the ritonavir comparisons.

| Number of Nodes | CCDC ID YIGPIO02 Time (min) | CCDC ID YIGPIO03 Time (min) |
|-----------------|-----------------------------|-----------------------------|
| 1               | 105.1                       | 140.8                       |
| 2               | 52.8                        | 70.6                        |
| 4               | 26.6                        | 35.5                        |
| 8               | 14.0                        | 18.7                        |
| 16              | 7.9                         | 9.9                         |
| 32              | 4.1                         | 5.6                         |
| 64              | 2.4                         | 3.2                         |