

Multi-Dimensional Structural Fingerprinting for Quantitative Polymorph Comparison: A Comprehensive Topological Analysis of Sulfathiazole, Galunisertib and Aripiprazole Polymorphs

Nikolaos Galanakis

ng1807@nyu.edu

New York University

David Hughes

University of Cambridge

Mark Tuckerman

New York University

Article

Keywords:

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10148467/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multi-Dimensional Structural Fingerprinting for Quantitative Polymorph Comparison: A Comprehensive Topological Analysis of Sulfathiazole, Galunisertib and Aripiprazole Polymorphs

Nikolaos Galanakis^{1,*}, David S. Hughes², and Mark E. Tuckerman^{1,3,4,5}

¹Department of Chemistry, New York University, New York, 10003, New York, USA

²Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

³Courant Institute of Mathematical Sciences, New York University, New York, 10012, New York, USA

⁴NYU-ECNU Center for Computational Chemistry, NYU Shanghai, 3663 Zhongshan Road North, Shanghai, 200062, China

⁵Simons Center for Computational Physical Chemistry at New York University, New York, 10003, New York, USA

*Correspondence: ng1807@nyu.edu

October 22, 2026

Abstract

Polymorphism in organic molecular crystals presents significant challenges for pharmaceutical and materials science applications, requiring robust methods for quantitative structural comparisons. We present a multi-dimensional topological fingerprinting approach that quantifies structural relationships through five complementary categories: crystal packing, intermolecular interaction networks, molecular conformation, fragment orientation, and spatial geometry. Distribution-based similarity metrics enable ranking and clustering of structures, and every descriptor is independent of the unit cell, space-group setting and Cartesian frame. We demonstrate the approach on three pharmaceutical polymorph families: five DFT-optimized sulfathiazole polymorphs, ten experimental galunisertib polymorphs and nine experimental forms of the flexible, partly non-aromatic aripiprazole. Which categories are conserved depends on the molecule. For sulfathiazole, spatial geometry is the most consistent category (0.78–0.86), whereas conformation and interaction networks separate Forms II, III and IV from Forms I and V. For galunisertib, spatial geometry and conformation are the most consistent, crystal packing varies most. For aripiprazole, spatial geometry is much less conserved (0.41–0.74), and the decomposition distinguishes forms that share their packing but not their conformation (Forms I and X) from forms that share their conformation but not their lattice (Forms III, IV and V). The overall similarity agrees with COMPACK cluster matching for sulfathiazole (Spearman $\rho = -0.89$ with RMSD) and follows the number of matched molecules for galunisertib ($\rho = 0.44$) and aripiprazole (0.67). Compared with the invariant descriptors AMD, PDD and SOAP, the fingerprint identifies the form of 47 sulfathiazole and 21 aripiprazole structures equally reliably and ranks distinct polymorphs closer to COMPACK cluster matching (aripiprazole: $\rho = 0.67$ versus 0.14–0.37), though not to simulated PXRD similarity. Rankings are robust to coordinate perturbations and cutoff choices, except where an H-bond lies at the criterion or few aromatic contacts occur. The framework supports polymorph classification, structural relationship analysis, validation of crystal structure prediction and solid-form screening.

1 Introduction

Predicting polymorphism, the ability of a compound to exist in multiple crystalline forms with different three-dimensional arrangements, in molecular crystals starting with nothing more than a molecular diagram represents one of the most significant challenges in solid-state chemistry and pharmaceutical development [1]. Different crystal polymorphs forms of the same molecule can exhibit dramatically different physical, chemical, and pharmaceutical properties, including solubility, bioavailability, mechanical stability, and other characteristics [2, 3, 4]. The pharmaceutical industry faces particular challenges, as regulatory agencies require comprehensive characterization of all known polymorphic forms during drug development [5].

Traditional approaches to polymorph comparison rely primarily on powder X-ray diffraction pattern matching, individual unit cell parameter comparisons or on subjective visual inspection of crystal structures [6]. While these methods provide valuable insights, they often fail to capture the full complexity of structural relationships between polymorphic forms. Moreover, such qualitative assessments lack the quantitative rigor necessary for systematic classification and prediction applications.

Recent advances in computational crystallography have enabled more sophisticated approaches to structural comparison. Methods such as crystal structure similarity metrics [7], energy landscape analysis [8], and topological crystal structure prediction approaches [9] have shown promising results. Cluster-overlay comparison with COMPACK has become the reference criterion for deciding whether two structures are the same, for example in the seventh CCDC blind test of crystal structure prediction [10, 11], and the five sulfathiazole polymorphs have been compared within a landscape of 96 sulfathiazole-containing structures using COMPACK, CrystalCMP, XPac, interaction energies [12] and hydrogen-bond topology [13]. However, these approaches typically focus on specific aspects of a crystal structure rather than providing comprehensive, multi-dimensional characterization. The recent development of mathematical crystal structure prediction methods [9] has demonstrated that topological principles can successfully identify polymorphic forms without relying on energy calculations, inspiring new approaches to structural comparison.

Methods for comparing crystal structures fall broadly into three groups. *Overlay methods* superimpose clusters of molecules: COMPACK [7] reports how many molecules of a 15- or 20-molecule cluster match within distance and angle tolerances and the RMSD of the match; progressive alignment of crystals (PAC) [14] reproduces COMPACK results at lower cost; CrystalCMP [15] superimposes clusters through selected molecular fragments; and XPac [16] identifies the supramolecular constructs that two structures share. These methods are precise and widely trusted, but they return one score, or one common construct, per pair of structures and depend on the cluster size and the tolerances. *Pattern-based methods* compare simulated powder patterns, for example by weighted cross-correlation [17], or Hirshfeld-surface fingerprint plots [18]; they need no alignment, but one pattern superposes all the ways in which two structures differ. *Invariant descriptors* represent a crystal independently of the cell and the frame: SOAP with the REMatch kernel [19, 20]; pointwise distance distributions (PDD) and average minimum distances (AMD), which are complete invariants for generic periodic point sets [21, 22]; and GRID, compared with the earth mover’s distance [23]. Topological representations have also been applied to polymorphism: in more than 19,000 inorganic polymorphs, local polyhedral environments are often retained across space groups [24], and space-group diversity has been built into a crystal structure prediction algorithm [25]. Invariant descriptors are mostly atom-based and establish whether two crystals are the same, or how far apart they are; they do not report in which structural respect two molecular crystals differ.

None of these approaches separates the different ways in which two polymorphs of the same

molecule can differ. The challenge in developing robust polymorph comparison methods lies in the multi-scale nature of crystal structure differences. Polymorphic variations can manifest at multiple levels: unit cell parameters and symmetry (crystal packing), intermolecular interaction patterns (hydrogen bonding and van der Waals contacts), molecular conformational changes, relative molecular orientations, and overall spatial organization [26]. A notable subclass of polymorphism is polytypism, in which layers of identical composition are stacked in different sequences along one crystallographic direction, producing distinct crystal structures with potentially different physical properties despite sharing the same chemical makeup [27]. Each level contributes differently to the overall structural relationship between polymorphs, requiring appropriate weighting and integration strategies. Early efforts to understand molecular crystal packing based on mathematical principles date back decades [28, 29], with the vision that crystal structures should be derivable through mathematical reasoning alone. Recent work has shown that such approaches can indeed be successful when combined with appropriate topological descriptors and physical constraints [9].

This work details a comparative study within three polymorphic families using a variety of geometric analytical methods. Topological and computational methods are combined, offering insights into the quantitative assessment of crystal packing in all three sets of polymorphs. The work is designed with the application of machine learning in pharmaceutical solid-form research in mind [30, 31], which, combined with advances in high-throughput crystallization [32, 33], requires structural descriptors that can be computed automatically for large numbers of crystal forms.

Distribution-based similarity metrics have emerged as powerful tools for comparing complex structural data. The Wasserstein distance, in particular, provides a mathematically rigorous framework for quantifying differences between probability distributions while maintaining physical interpretability [34]. This metric has found successful applications in various fields of structural biology and materials science, demonstrating its utility for comparing molecular arrangements. With this in mind, in this paper, we present a comprehensive multi-dimensional fingerprinting method employing a modular architecture that systematically quantifies structural relationships between crystal polymorphs. Existing approaches to polymorph comparison predominantly rely on geometric overlay methods such as COMPACT [7], which compute root-mean-square deviations (RMSD) over clusters of nearest-neighbor molecules. While powerful for pairwise alignment, these methods provide a single aggregate similarity score that conflates fundamentally different sources of structural variation—conformational differences, symmetry changes, and packing rearrangements—making it difficult to identify *why* two structures differ. Energy landscape approaches [8] offer thermodynamic context but require expensive force field or electronic structure calculations and do not directly quantify geometric structural relationships. Topological descriptors [9], while effective for structure prediction, have not been systematically extended to multi-dimensional polymorph comparison. Our approach addresses these limitations by integrating five independent yet complementary structural categories through weighted similarity calculations using distribution-based metrics, primarily the Wasserstein distance. All descriptors are built from the full periodic crystal and are invariant to the choice of unit cell, space-group setting and Cartesian frame, so that equivalent descriptions of the same crystal (for example $P2_1/c$ and $P2_1/n$) give identical fingerprints. The fingerprint introduced here differs from these approaches in four respects. (i) It resolves the similarity of two structures into five physically distinct components, reported as a similarity vector, so that the reason why two polymorphs are similar or different can be read from the result. (ii) Its descriptors are defined for molecules and rigid fragments rather than for atoms: contact directions are measured in molecule-fixed frames, torsions are compared bond by bond after matching the molecular graphs, and coordination shells are defined by contacts between whole molecules. (iii) It is evaluated on the full periodic crystal, without a cluster size or iterative superposition, and is invariant to the cell, the space-group setting and the frame by construction, which we verify numer-

ically. (iv) Its fingerprints and similarity vectors are fixed-format and machine-readable, suitable for machine learning and for the analysis of crystal structure prediction landscapes. Established tools, such as space-group operations and the Wasserstein distance, are used where appropriate; the contribution lies in the descriptors built on them and in the decomposition they enable. The method employs equal category weighting ($w = 0.20$ per category) to provide an unbiased overall similarity assessment while maintaining flexibility for application-specific optimization. Validation using five DFT-optimized sulfathiazole polymorphs, ten experimental galunisertib polymorphs and nine experimental forms of the flexible drug aripiprazole demonstrates the method’s capability to distinguish subtle structural variations while identifying robust structural invariants across polymorphic forms.

2 Methods

2.1 Data Representation and Preprocessing

Crystal structure data is processed from standard crystallographic information files (CIF format) and stored in hierarchical data format (HDF5) for efficient access and manipulation. Each structure undergoes systematic feature extraction to generate comprehensive structural fingerprints across five primary categories. Only the asymmetric unit and the space-group symbol are stored; the full periodic crystal is regenerated from the symmetry operations of the stated setting (International Tables [35]), and all descriptors are evaluated on this periodic crystal. Equivalent space-group settings therefore need no manual transformation: the descriptors are constructed to be setting-invariant (Section 2.4). Molecular fragments are identified by connectivity analysis. Hydrogen atoms are added using standard geometric criteria when absent from the original structure, with X–H bond lengths set to neutron diffraction values. Intermolecular contacts are identified using a van der Waals criterion, $d \leq r_i + r_j + 0.6 \text{ \AA}$, with Bondi radii [36]. Each contact is counted once. Hydrogen bonds are D–H $\cdots$ A contacts with D, A $\in \{\text{N, O}\}$, H $\cdots$ A not exceeding the sum of the van der Waals radii and D–H $\cdots$ A $\geq 120^\circ$.

Three polymorphic systems are analyzed: five DFT-optimized polymorphs of sulfathiazole, ten experimental polymorphs of galunisertib and nine experimental forms of aripiprazole (Tables 1–3). Forms are assigned as in the CSD. Further CSD depositions of the same forms are used in the benchmark (Section 2.5).

Table 1 summarizes the crystallographic parameters for the five sulfathiazole polymorphs analyzed in this study. All structures are DFT-optimized, representing local energy minima. Forms I, III and V contain two symmetry-independent molecules in the asymmetric unit ($Z'=2$), while Forms II and IV contain one ($Z'=1$). All polymorphs crystallize in space group No. 14, described in the $P2_1/c$ (Forms I, II, III) or $P2_1/n$ (Forms IV, V) setting, demonstrating the strong preference for this symmetry in sulfonamide compounds. Cell parameters and starting coordinates for all five polymorphs were taken from Gelbrich *et al.* [37]. The asymmetric units were transformed so that the molecular centroid(s) lie within the unit cell and as close as possible to the origin, and molecule 1 of every asymmetric unit was set to the same (R) conformational enantiomer [12]. Prior to energy minimization, the positions of all H atoms in the X-ray structures were normalized using the default settings in Mercury to give starting positions close to the nuclear positions. Dispersion-corrected DFT (DFT-D) calculations were made with CASTEP [38] through the Materials Studio interface, using the PBE exchange–correlation functional [39] with the Grimme dispersion correction [40], a plane-wave basis-set cut-off of 340 eV, and all other parameters at the “Fine” defaults of Materials Studio 6.0. Unit-cell parameters were constrained to those of the reported crystal structure and the space-group symmetry was imposed. Minimized structures were compared with the starting

structures using the method of van de Streek and Neumann [41], which confirmed that the minimization does not significantly distort the crystal structures. The minimized structures are used to reduce differences due to thermal motion and other experimental variables, so that the comparison reflects intrinsic packing differences; they are provided as Supplementary Data.

Table 2 summarizes the crystallographic parameters for the ten galunisertib polymorphs analyzed in this study. Unlike the sulfathiazole dataset, these are experimental crystal structures available in the Cambridge Structural Database (CSD refcode family DORDUM [42]), and have not been subjected to DFT optimization; thermal effects and experimental uncertainties are therefore present. DORDUM08 is a second deposition of Polymorph X (DORDUM07); COMPACK overlays all 20 molecules of the two clusters with an RMSD of 0.083 Å and a simulated PXRD similarity of 0.998, so DORDUM08 was excluded to avoid counting Polymorph X twice; it is used however in the identification test of Section 3.4. All forms are $Z' = 1$ structures, except polymorphs IV and VII which contain two symmetry-independent molecules in the asymmetric unit ($Z' = 2$). The dataset exhibits greater crystallographic diversity than sulfathiazole: four polymorphs crystallize in triclinic structures (space group $P\bar{1}$; Polymorphs II, III, VII, VIII), while six adopt monoclinic structures ($P2_1/n$, $P2_1/c$, or $C2/c$; Polymorphs I, IV, V, VI, IX, X).

Table 3 summarizes the crystallographic parameters for aripiprazole ($C_{23}H_{27}Cl_2N_3O_2$), a flexible molecule with two aromatic and two non-aromatic rings that shows extensive conformational polymorphism [43, 44, 45, 46, 47, 48, 49]. All 21 CSD depositions (refcode family MELFIT; nine forms, CSD numbering) were processed in the same way as the other structures (Supplementary Section S1). Each form is represented by the deposition with the highest mean overall similarity to the other depositions of its form; the others are used in the benchmark (Section 3.4). For the disordered Form IX (MELFIT19) the major component was used. Because cutting the molecule at its rotatable bonds also produces one- to three-atom linker fragments without defined principal axes, the fragment-level descriptors use the three ring systems only; all atoms still enter the contacts, H-bonds, torsions and coordination shells.

2.2 Multi-Dimensional Fingerprinting Framework

Our method employs a modular architecture that decomposes crystal structure comparison into five independent yet complementary categories, each capturing distinct physical phenomena operating at different structural scales. This modular design is implemented through five specialized comparator classes (Crystal Packing Comparator, Interaction Networks Comparator, Conformation Comparator, Fragment Orientation Comparator, Spatial Geometry Comparator), each responsible for data extraction, comparison metric calculation, and category-specific analysis. The independent category implementations enable systematic validation while maintaining consistent interfaces for integration. This separation of concerns facilitates extension to new descriptor types and adaptation to system-specific requirements without affecting the overall framework architecture.

The five categories are integrated using weighted similarity scores, with default equal weighting to provide unbiased multi-dimensional comparison: $w_{cp}=0.20$ (crystal packing), $w_{in}=0.20$ (interaction networks), $w_{cf}=0.20$ (conformation), $w_{fo}=0.20$ (fragment orientation), $w_{sg}=0.20$ (spatial geometry). This equal weighting strategy avoids *a priori* assumptions about category importance and provides robust baseline comparison suitable for diverse chemical systems. The framework supports custom weight optimization for specific applications. Within each category, the sub-metric similarities are likewise combined with equal weights (Eq. 9). A step-by-step description of every category, the content of the input files and a worked example that lists every intermediate quantity of one comparison are given in Supplementary Section S2 (Table S3).

For each structure the method produces a fingerprint: the set of descriptor distributions and

scalars listed in the following subsections (tabulated for all fifteen structures in Supplementary Tables S6 and S7). Comparing two fingerprints gives a five-component similarity vector $\mathbf{s}(i, j) = (s_{\text{cp}}, s_{\text{in}}, s_{\text{cf}}, s_{\text{fo}}, s_{\text{sg}})$, from which the overall similarity $S(i, j)$ is obtained (Eq. 10). The similarity vectors for all 55 pairs are listed in Supplementary Tables S9 and S10; the fingerprints and similarity vectors of aripiprazole are given in Supplementary Tables S8 and S11.

2.2.1 Crystal Packing Analysis

Crystal packing analysis characterizes unit cell-level organization through four primary descriptor groups. **Packing efficiency** is quantified via the packing coefficient (ratio of molecular volume to unit cell volume per molecule) and crystal density measurements, both providing thermodynamic indicators of structural stability and compactness. **Unit cell geometry** is analyzed on the Niggli-reduced primitive cell [50], which is unique for a given lattice and therefore independent of the cell chosen in the deposited structure. Reduced cell lengths are compared through $\exp(-\|\hat{\mathbf{l}}_1 - \hat{\mathbf{l}}_2\|)$, where $\hat{\mathbf{l}}$ is the length vector divided by its mean, and the obliquity of the reduced cell through the sorted deviations $|90^\circ - \alpha|$, $|90^\circ - \beta|$, $|90^\circ - \gamma|$, whose differences Δ are scored with the same transform as all other angles, $\exp(-|\Delta|/\sigma_{\text{angle}})$ (Eq. 8), and averaged. Because the normalized length vectors are dimensionless, their distance is used directly (scale 1) rather than with one of the scale parameters of Eq. 8. **Symmetry operations** form the core of this category: each space group’s complete set of rotation-translation operations is extracted and represented as 4×4 augmented matrices combining 3×3 rotation components $\mathbf{M}$ with translation vectors $\mathbf{t}$ given by

$$\mathbf{G} = \begin{pmatrix} M_{11} & M_{12} & M_{13} & t_x \\ M_{21} & M_{22} & M_{23} & t_y \\ M_{31} & M_{32} & M_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (1)$$

which represent elements of the space group, a discrete subgroup of the Euclidean group $E(3)$ (proper and improper operations), and which act on a 4-vector of fractional coordinates,

$$\mathbf{R} = (x \ y \ z \ 1)^T.$$

Comparison employs Jaccard similarity coefficients on the set of unique symmetry operations, capturing fundamental symmetry relationships governing molecular arrangements in a manner consistent with the topological principles established for crystal structure prediction [9]. To make this comparison independent of the setting, the operations are always taken for the space-group type in its standard ITA setting; two descriptions of the same type (for example $P2_1/c$ and $P2_1/n$) therefore give $J = 1$, while different types are compared through the operations they share (e.g. $J = 0.5$ between $P\bar{1}$ and $P2_1/c$). **Molecular multiplicity** parameters (Z , Z') complement the analysis by characterizing packing density and asymmetric unit composition, and are compared through the ratio $\min/\max$. The category score is the unweighted mean of the seven sub-metrics (packing coefficient, density, reduced-cell lengths, reduced-cell obliquity, symmetry operations, Z and Z').

2.2.2 Intermolecular Interaction Networks

Intermolecular interaction analysis characterizes the distribution of contact lengths, the geometry of hydrogen bonds and the directions in which each molecule makes contacts. Because the crystal is periodic and every symmetry-equivalent molecule has an identical environment, the contacts of the asymmetric unit describe the complete interaction network. Contact directionality is characterized

in the local principal-axes frame of the fragment that carries the contacting atom: for each contact vector $\hat{\mathbf{u}}$ (from the fragment centroid to the contact atom) we record the polar angle to the fragment normal and the in-plane angle to the fragment long axis, both folded to $[0^\circ, 90^\circ]$. These angles describe where on a molecule its contacts are made, and are invariant to the unit cell and to the Cartesian frame, unlike angles measured with respect to the crystallographic or Cartesian axes. To summarize the anisotropy of the full network, the contact vectors are expanded with the rotational part R of each of the n symmetry operations G (Eq. 1),

$$\{\mathbf{d}_k^{(l)}\} = \left\{ R^{(l)} \mathbf{d}_k \mid l = 1, \dots, n; k = 1, \dots, m \right\}, \quad (2)$$

and the contact direction tensor

$$T_{ij} = \frac{1}{N} \sum_{k=1}^N \hat{u}_k^{(i)} \hat{u}_k^{(j)}, \quad (3)$$

where $\hat{\mathbf{u}}_k$ is the unit vector of the k -th contact and N is the total number of symmetry-expanded contacts. The contact anisotropy index, calculated from the eigenvalue spread of $\mathbf{T}$,

$$\mathcal{A} = \frac{\lambda_1 - \lambda_3}{\lambda_1}, \quad (4)$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the eigenvalues of $\mathbf{T}$, quantifies the degree of directional preference (0=isotropic, 1=completely anisotropic). $\mathcal{A}$ depends only on eigenvalues and is therefore frame-invariant; it is reported as a descriptor but not scored.

Contact statistics are analyzed through distribution-based comparison using the Wasserstein distance, which quantifies the minimal ‘work’ required to transform one distribution into another. For two distributions $P(x)$ and $Q(x)$ with cumulative distribution functions $F_P(x)$ and $F_Q(x)$, given by

$$F_P(x) = \int_{-\infty}^x P(y) dy, \quad F_Q(x) = \int_{-\infty}^x Q(y) dy \quad (5)$$

such that $F_{P/Q}(x) \in [0, 1]$, the Wasserstein distance is given by

$$W(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| dx, \quad (6)$$

providing physically interpretable distance measures in the units of the underlying quantity ($\text{\AA}$ for distances, degrees for angles, *etc.*). This metric naturally handles varying sample sizes and provides physically interpretable distance measures. **Hydrogen bonding geometry** receives special attention due to its structural importance: we analyze both donor–acceptor distances and D–H $\cdots$ A angles through their full distributions, with angle distributions providing conformational information about H-bonding motifs (linear versus bent arrangements) [51]. The category score is the unweighted mean of six sub-metrics: contact lengths, contact polar angles, contact in-plane angles, H-bond D $\cdots$ A distances, H-bond angles, and the number of H-bonds per molecule. Three of the six sub-metrics describe hydrogen bonds, so H-bonding contributes half of the category score; this reflects its importance for the packing of H-bonded molecular crystals such as the two studied here, and should be kept in mind when the method is applied to systems without strong H-bond donors.

2.2.3 Molecular Conformation Analysis

Molecular conformation characterization focuses on internal molecular geometry through four complementary descriptors that separate genuine conformational differences from rigid-body molecular reorientation. **Torsion (dihedral) angles** serve as the primary conformational descriptor, capturing rotational freedom around single bonds and distinguishing conformers with different three-dimensional arrangements while remaining invariant to overall molecular rotation and translation. All heavy-atom torsions about rotatable bonds are compared bond by bond: the heavy-atom graphs of the two molecules are matched by graph isomorphism, so that every torsion has a defined partner, and the mean circular difference $D = \langle |\Delta\tau| \rangle$ over the matched torsions is converted to a similarity $\exp(-D/\sigma_\tau)$. Equivalent atom mappings arising from molecular symmetry (for sulfathiazole, the ring flip of the aniline group and the exchange of the two sulfonyl O atoms) are all tried and the best is kept, as is the better of τ' and $-\tau'$, since mirror-image conformers are equivalent in centrosymmetric crystals; for $Z' > 1$ each molecule is paired with its best partner in the other structure, in both directions. Matching is necessary because a pooled distribution of torsion values does not record which bond each value belongs to: galunisertib Polymorphs II and III, for example, have almost the same pooled distribution of $|\tau|$ (Wasserstein distance 4.7°) but differ by 54° on average bond by bond. Torsions are the soft coordinate of a molecule and differ between polymorphs by tens of degrees, so they are given their own scale, $\sigma_\tau = 30^\circ$; with the 10° used for stiffer angles, every difference above about 25° would give a similarity close to zero.

Bond angle distributions complement torsional analysis by characterizing local geometric strain and deviations from ideal geometries. Only bond angles at vertices that carry a rotatable bond are included, since angles within rigid rings carry no conformational information. Changes in bond angle patterns indicate conformational adaptation to crystal packing forces, with systematic shifts suggesting energetically costly deformations. The conservation of bond angle distributions across polymorphs indicates that conformational changes occur primarily through torsional rotation rather than angle bending, consistent with the larger energetic cost of the latter.

Improper (out-of-plane) angles are computed at every three-coordinate atom as the mean Wilson angle between each bond and the plane of the other two. They capture pyramidalization, which in sulfathiazole and galunisertib is concentrated at the amine and amide N atoms and is sensitive to their hydrogen-bonding environment. **Bond lengths** are computed and reported but are not scored: in the DFT-optimized sulfathiazole structures they are fixed by the functional to within ~ 0.01 Å (Wasserstein distances between forms 0.001–0.007 Å), and in the X-ray galunisertib structures they differ between forms by Wasserstein distances of only 0.002–0.034 Å, of the order expected from differences in data quality and collection temperature, so including them would add noise rather than conformational information.

Intramolecular contact networks, identified using the same van der Waals-based distance criteria applied for intermolecular contacts, reveal through-space interactions within the molecule that may stabilize particular conformations. These include 1,4-interactions, interactions across rings, and longer-range contacts that create constraints on conformational flexibility. The distribution of intramolecular contact lengths provides a fingerprint of the molecular shape in three-dimensional space. Together, these four descriptor classes enable robust conformational comparison that distinguishes true conformational polymorphs from packing polymorphs where molecules adopt the same conformation but different orientations. The category score is their unweighted mean.

2.2.4 Fragment Orientation Analysis

Fragment orientation analysis characterizes the relative spatial arrangement and alignment of molecular fragments through two complementary approaches: inertia tensor analysis and comprehensive π - π stacking interactions.

Inertial tensor analysis provides shape-dependent orientation characterization. For each molecular fragment containing n atoms with coordinates $\mathbf{r}_1, \dots, \mathbf{r}_n$ with masses $m_1, \dots, m_n$ identified by connectivity analysis, we compute the inertial tensor given by

$$I_{\alpha\beta} = \sum_{i=1}^n m_i (|\mathbf{r}_i|^2 \delta_{\alpha\beta} - r_{i\alpha} r_{i\beta}) \quad (7)$$

where $r_{i\alpha}$ is the α th component of coordinate vector $\mathbf{r}_i$ relative to the fragment centre of mass. The fragment shape is summarized by the relative shape anisotropy $\kappa^2 = 1 - 3(g_1 g_2 + g_2 g_3 + g_3 g_1)/(g_1 + g_2 + g_3)^2$ of the mass-weighted gyration tensor, whose eigenvalues g_k follow from the principal moments of inertia; $\kappa^2 = 0$ for a spherical, 0.25 for a planar disc-like and 1 for a linear fragment. For sulfathiazole the mean fragment anisotropy is 0.399–0.406 across the five forms (Scheme 1a), and for galunisertib 0.447–0.459 across the ten forms (Scheme 1b); in both systems the per-fragment values span the planar rings and the smaller sulfonyl or amide groups. For aripiprazole, whose fragment-level descriptors use the three ring systems (Section 2.1), the mean is 0.293–0.296 across the nine forms (Scheme 1c), with per-fragment values of about 0.27 for the planar dichlorophenyl ring, 0.44 for the elongated bicyclic dihydroquinolinone and 0.17 for the compact chair piperazine.

The eigenvectors define fragment-fixed coordinate systems that enable orientation comparisons across structures: for every pair of neighbouring fragments belonging to different molecules (centroid separation ≤ 8 Å) we record the angles between their corresponding principal axes (long axis, in-plane axis and plane normal), folded to $[0^\circ, 90^\circ]$. These relative orientations are invariant to the unit cell and the Cartesian frame, and their distributions are compared via the Wasserstein distance. Pairs of fragments of different types are included: for them the angles describe how the long axes, in-plane axes and planes of the two fragments are aligned, not the alignment of chemically equivalent directions. Restricting the analysis to fragments of the same type, or comparing each type pair separately, leaves too few pairs per structure for stable distributions (7–23 same-type pairs per galunisertib structure).

π - π stacking analysis characterizes aromatic interactions throughout the crystal structure. This analysis proceeds through several steps: (1) *Neighbour search*: all images of the aromatic fragments within 6 Å (centroid–centroid) of each asymmetric-unit fragment are generated from the full periodic crystal, so that interactions extending beyond the unit cell boundaries are included. (2) *Fragment identification*: Aromatic fragments are identified through either user-specified chemical formulas (enabling selective analysis of specific moieties such as the aniline $\text{C}_6\text{H}_6\text{N}$ and thiazole $\text{C}_3\text{H}_3\text{N}_2\text{S}$ rings in sulfathiazole, or the methylpyridine $\text{C}_6\text{H}_6\text{N}$, pyrazole–dihydropyrrole $\text{C}_6\text{H}_6\text{N}_2$, and quinoline $\text{C}_9\text{H}_5\text{N}$ rings in galunisertib) or automatic planarity detection. The user-defined approach provides chemical selectivity and avoids spurious detection of accidentally planar non-aromatic groups. (3) *Geometric parameter calculation*: For each fragment pair, we compute the plane normals by a least-squares fit to the ring atoms (equivalently, the principal axis of the largest moment of inertia), the interplanar angle from $|\hat{\mathbf{n}}_1 \cdot \hat{\mathbf{n}}_2|$ so that parallel and antiparallel normals are treated alike, perpendicular distance (the smaller of the projections of the centroid–centroid vector onto the two normals), parallel offset (in-plane displacement of fragment centroids), and centroid-to-centroid distance (overall separation). (4) *Motif classification*: interactions with an interplanar angle below 30° are classified as face-to-face (offset < 1.5 Å) or parallel-displaced, those above 60° as edge-to-face (T-shaped), and the remainder as tilted. (5) *Distribution-based comparison*:

The distributions of geometric parameters (distances, offsets, angles) across all interactions are compared between structures using the Wasserstein distance, and combined with the ratio of the number of aromatic contacts per molecule into a single aromatic-contact score (unweighted mean).

This π - π stacking analysis provides insights into aromatic interaction motifs that stabilize crystal structures. The relative populations of these motifs constitute the motif distribution, which often serves as a fingerprint of packing preferences characteristic of a given crystal form. The fragment-orientation score is the unweighted mean of the relative-orientation and aromatic-contact similarities. The fragment shape κ^2 is reported but not scored: it is a property of the molecule rather than of its orientation, and is nearly identical in all polymorphs of the same molecule (similarities 0.94–1.00).

2.2.5 Spatial Geometry Analysis

Spatial geometry characterization describes the arrangement of fragment centres around each fragment, from the first coordination shell to intermediate-range ordering. Unlike the previous categories that focus on unit cell-level or pairwise properties, spatial geometry analysis requires sampling beyond the unit cell to capture periodic ordering effects.

Fragment environments: for every asymmetric-unit fragment, the centroids of all fragments of other molecules within 12 Å are generated from the periodic crystal (fragments of the same molecule are excluded, because their distances describe the conformation rather than the packing), so that each reference fragment is completely surrounded. Environments of individual fragments are used because statistics pooled over all pairs of fragments inside a single sphere are dictated, for any dense packing, by the sphere itself (for a 20 Å sphere, mean pair distance $36R/35 = 20.6$ Å and per-axis spread $R/\sqrt{5} = 8.9$ Å) rather than by the crystal structure.

Fragment distance distributions are computed from the distances between each reference fragment and its neighbours, yielding a fingerprint of characteristic length scales governing spatial organization. The distribution typically shows peaks corresponding to nearest-neighbor separations, second-shell neighbors, and longer-range ordered positions. Wasserstein distance comparison between these distributions quantifies spatial organizational similarity.

Radial distribution function (RDF) analysis provides multi-scale characterization of spatial ordering analogous to liquid structure analysis. The RDF, $g(r)$, quantifies the probability of finding a fragment center at distance r from a reference fragment, normalized by the expected probability for a random distribution at the same density (the number of fragments in the unit cell divided by its volume), so that $g(r) \rightarrow 1$ at large r . $g(r)$ is computed with 0.1 Å bins and 0.2 Å Gaussian smoothing, and two RDFs are compared by their overlap $1 - \int |g_1 - g_2| dr / \int (g_1 + g_2) dr$, which, unlike a Wasserstein distance between $g(r)$ values, is sensitive to peak positions. The neighbour distances and $g(r)$ are derived from the same set of distances but respond to different aspects of it: the Wasserstein distance to overall shifts of the distribution, the overlap to its peak structure (Spearman correlation between the two sub-metrics over all pairs 0.93 for sulfathiazole and 0.75 for galunisertib); together they give radial information half of the weight of the category. Peaks in $g(r)$ indicate preferred fragment-fragment separations at increasing radial distances, with peak positions revealing short-range, intermediate-range, and long-range ordering patterns.

Coordination number analysis characterizes local environments by the number of distinct neighbouring molecules in van der Waals contact with each molecule, using the contact criterion of Section 2.1 ($d \leq r_i + r_j + 0.6$ Å). A contact-defined first shell is used because the first minimum of a fragment-level $g(r)$ is ill-defined for molecules made of several fragments. This metric provides a local environment descriptor analogous to coordination numbers in inorganic crystals, enabling classification of packing environments (*e.g.*, close-packed versus loosely packed).

First-shell anisotropy: the vectors from the centroid of a molecule to the centroids of its contacting neighbours define a gyration tensor whose relative shape anisotropy κ_s^2 (0 for an isotropic shell) quantifies whether the neighbours are distributed evenly or concentrated in layers or columns. Unlike a spread along Cartesian axes, κ_s^2 is frame-invariant.

This combination of metrics spanning microscopic (pairwise distances) to mesoscopic (RDF, coordination) scales provides comprehensive spatial characterization that captures both local packing efficiency and longer-range structural ordering. The multi-scale nature enables detection of similarities in overall spatial organization even when unit cell parameters differ significantly, as similar spatial distributions can be achieved through different symmetry operations and cell geometries. The category score is the unweighted mean of the four sub-metrics.

2.3 Similarity Metrics and Integration

Distribution-based similarity is computed primarily using the Wasserstein distance (Eq. 6), which provides an intuitive metric for comparing probability distributions. This metric accounts for continuous distributions rather than requiring binning (as in histogram comparisons) or kernel density estimation, and provides physically interpretable distances in the units of the underlying quantity being compared ($\text{\AA}$ for distances, degrees for angles, *etc.*).

The Wasserstein distance is particularly well-suited to comparing molecular structural distributions where: (1) the continuous nature of geometric parameters is important, (2) tail behavior of distributions carries information, (3) sample sizes may differ between structures, and (4) physical interpretation of similarity magnitude is desired.

Conversion to similarities: Wasserstein distances are converted to bounded similarity scores via exponential decay:

$$s = \exp(-W/\sigma) \quad (8)$$

where σ is a scale parameter for each physical unit: $\sigma = 0.20 \text{ \AA}$ for all distances, $\sigma = 10^\circ$ for all angles except the matched torsions, for which $\sigma_\tau = 30^\circ$ (Section 2.2.3), and $\sigma = 0.20$ for the dimensionless anisotropies. A distance difference of σ therefore reduces the similarity to $e^{-1} = 0.37$; the values were chosen so that differences that are structurally meaningful between polymorphs (a few tenths of an $\text{\AA}$, or several degrees) give intermediate similarities. Scalar quantities are compared by $1 - |a - b|/\max(a, b)$ (packing coefficient, density) or $\min(a, b)/\max(a, b)$ (counts). A single transform with one scale per unit makes the sub-metrics directly comparable, and Supplementary Table S23 shows that halving or doubling all σ changes the absolute similarities but not the ranking of polymorph pairs (Spearman $\rho \geq 0.97$).

Category-level integration: Within each category, individual descriptor similarities are combined as an unweighted average to give the category-level similarity score:

$$s_k(i, j) = \frac{1}{N_k} \sum_{l=1}^{N_k} s_k^{(l)}(i, j), \quad (9)$$

where N_k is the number of sub-metrics in category k and $s_k^{(l)}(i, j)$ is the l -th sub-metric similarity, each lying in $[0, 1]$. The sub-metrics of each category are listed in the corresponding subsection ($N_{\text{cp}} = 7$, $N_{\text{in}} = 6$, $N_{\text{cf}} = 4$, $N_{\text{fo}} = 2$, $N_{\text{sg}} = 4$). Weighting the sub-metrics unequally instead (50/30/20% for conformation and 30/30/40% for fragment orientation, with the fragment shape scored) changes the overall similarities by at most 0.047 and leaves the ranking of the pairs essentially unchanged (Spearman $\rho = 0.98, 0.97$ and 0.95 for sulfathiazole, galunisertib and aripiprazole; Supplementary Table S24). The overall structural similarity between structures i and j is then the

weighted sum:

$$S(i, j) = \sum_{k \in \text{categories}} w_k s_k(i, j), \quad (10)$$

where w_k are non-negative weights for each component ($w_k = 0.20$).

Interpretation scale. To describe similarity levels consistently, we use the following bands for category means: high ($\bar{s} \geq 0.85$), moderate ($0.70 \leq \bar{s} < 0.85$) and low ($\bar{s} < 0.70$); a category is called consistent across polymorph pairs if its standard deviation over all pairs is $\sigma_s \leq 0.05$ and variable otherwise. Because absolute similarity values depend on the scale parameters of Eq. 8, these bands are used only to compare categories and systems analysed with the same parameters.

2.4 Invariance to Unit Cell, Setting and Frame

Descriptors computed from the cell as deposited, from symmetry operations in the deposited setting or from angles measured in the Cartesian frame give different values for equivalent descriptions of the same crystal. For example, if sulfathiazole Forms IV and V are re-expressed in the $P2_1/c$ setting (an exact change of basis, $\mathbf{c}' = \mathbf{a} + \mathbf{c}$), crystal-packing similarities computed in this way change by up to 0.11, and the most similar pair changes from Form I–III to Form I–V (Supplementary Table S4). In the present method, every descriptor is invariant: the cell is Niggli-reduced, symmetry operations are taken for the space-group type in its standard setting, and all directions are measured in molecule-fixed frames or between pairs of molecules. We verified this directly by recomputing all 91 pairwise comparisons (10 sulfathiazole, 45 galunisertib and 36 aripiprazole pairs) after (i) transforming every $P2_1/c$ structure to $P2_1/n$ and every $P2_1/n$ structure to $P2_1/c$, (ii) rotating the Cartesian frame of every structure by an arbitrary rotation, and (iii) re-describing every triclinic cell with $\mathbf{c}' = \mathbf{a} + \mathbf{b} + \mathbf{c}$. In all three tests the category and overall similarities are unchanged to within 4×10^{-9} , i.e. to numerical precision (Supplementary Table S5). The conversion of all structures to $P1$ suggested as an alternative would also be invariant, but it would discard the symmetry descriptors that are part of the crystal-packing category.

2.5 Benchmark Against Other Comparison Methods

To establish what the fingerprint adds to existing methods, we compared it with three invariant descriptors, two simple baselines and COMPACK. **Data.** In addition to the structures above, we processed all 42 CSD depositions of sulfathiazole, which COMPACK assigns to the same forms as the CSD (Supplementary Table S1), the second deposition of galunisertib polymorph X (DOR-DUM08) and all 21 aripiprazole depositions. This gives 47 sulfathiazole structures in five forms (including the five DFT-D structures), 11 galunisertib structures in ten forms and 21 aripiprazole structures in nine forms. **Methods.** Average minimum distances (AMD) and pointwise distance distributions (PDD) [21, 22] with $k = 100$ neighbours; the SOAP power spectrum [19] compared with the REMatch kernel [20] and converted to a distance $d = \sqrt{2 - 2K}$; the Niggli-reduced cell; the density; and COMPACK (20-molecule clusters), which gives the number of matched molecules N_{match} , the RMSD and the simulated PXRD similarity. AMD, PDD and SOAP were computed with the `average-minimum-distance` [21] and `DScribe` [52] packages on the same periodic crystals as the fingerprint, without hydrogen atoms; parameters and variants are given in Supplementary Section S7. **Tests.** (1) *Identification*: each structure is used as a query and the others are ranked by distance ($1 - S$ for the fingerprint); we report the mean average precision (mAP), the area under the ROC curve for separating same-form from different-form pairs (AUC) and, for sulfathiazole,

the adjusted Rand index of clustering into five groups, and for galunisertib the distance between DORDUM07 and DORDUM08 divided by the smallest distance between two distinct polymorphs. (2) *Variation between depositions of the same form*, for each category. (3) *Agreement with COMPACK for distinct forms*: Spearman rank correlations over the 10 sulfathiazole, 45 galunisertib and 36 aripiprazole pairs, with pair-bootstrap confidence intervals. (4) *Category response*: Spearman correlations between the distance of each method and each category similarity. (5) *Invariance*: the tests of Section 2.4, applied to all methods.

2.6 Computational Implementation

The method is implemented in Python (version 3.8+) using NumPy (v1.21+) for numerical operations, SciPy (v1.7+) for statistical calculations including Wasserstein distance computation and hierarchical clustering, spglib [53] for space-group operations and Niggli reduction, H5py (v3.0+) for efficient hierarchical data management, and Matplotlib/Seaborn for visualization. The modular architecture employs five independent comparator classes, each implementing standard interfaces for data extraction, pairwise comparison, visualization, and reporting. A master Polymorph Comparator class orchestrates the analysis pipeline, manages weighting schemes, and integrates results across categories. This object-oriented design enables independent validation and extension of individual categories while maintaining consistent data flow. All figures, tables and numbers in this paper are generated by the same scripts from a single run of the pipeline. A complete analysis of any of the three polymorph families (fingerprint extraction for all structures and all pairwise comparisons) runs in under 3 s on a standard workstation, enabling high-throughput screening applications.

3 Results

3.1 Polymorph Datasets

Sulfathiazole:

We validate our method using five DFT-optimized polymorphs of sulfathiazole ($\text{C}_9\text{H}_9\text{N}_3\text{O}_2\text{S}_2$), representing distinct crystallographic forms extensively characterized in the pharmaceutical and crystallographic literature [37, 12, 13]. All polymorphs crystallize in space group No. 14 ($P2_1/c$ or $P2_1/n$ setting), reflecting the strong preference for this symmetry in sulfonamide compounds. Forms I, III and V contain two symmetry-independent molecules in the asymmetric unit ($Z'=2$), while Forms II and IV contain one ($Z'=1$), providing an opportunity to examine how Z' affects structural similarity patterns. Table 1 summarizes the crystallographic parameters. The crystal packing of the five polymorphs is illustrated in Figure 1. Despite common crystal systems and space group types, the structures exhibit significant variation in unit cell dimensions (cell volumes ranging from 1037–2247 Å³), intermolecular interaction patterns, and molecular conformations, providing an appropriate test case for multi-dimensional comparison. The DFT-optimized structures eliminate thermal disorder and experimental uncertainties, enabling focus on intrinsic structural relationships rather than artifacts of data collection conditions. In the absence of prior information favoring any single structural category, we assign equal weights $w_k = 1/N_{\text{cat}} = 0.20$ across all five categories, providing an unbiased baseline for this validation study.

Galunisertib:

We further validate our method using ten experimental polymorphs of galunisertib ($\text{C}_{22}\text{H}_{19}\text{N}_5\text{O}$), a TGF- β receptor inhibitor whose polymorphic landscape has been characterized in the Cambridge

Structural Database (CSD refcode family DORDUM). Unlike the sulfathiazole dataset, these are experimental structures available as deposits into the CSD and have not been subjected to DFT optimization, providing a test of the method’s applicability to structures carrying thermal effects and experimental uncertainties. The dataset exhibits greater crystallographic diversity than sulfathiazole: six polymorphs crystallize in monoclinic systems ($P2_1/n$, $P2_1/c$, or $C2/c$; Polymorphs I, IV, V, VI, IX, X), while four adopt triclinic symmetry ($P\bar{1}$; Polymorphs II, III, VII, VIII). Polymorphs IV and VII contain two symmetry-independent molecules in the asymmetric unit ($Z' = 2$), while the remaining eight are $Z' = 1$ structures, again providing an opportunity to examine how Z' affects structural similarity patterns. Table 2 summarizes the crystallographic parameters. The structures exhibit substantial variation in unit cell dimensions (cell volumes ranging from 917–3765 Å³), intermolecular interaction patterns, and molecular conformations, providing a more demanding test case for multi-dimensional comparison. As with sulfathiazole, we assign equal weights across all five categories, providing an unbiased and directly comparable baseline.

Aripiprazole:

We finally apply the method to nine experimental forms of aripiprazole (C₂₃H₂₇Cl₂N₃O₂; CSD refcode family MELFIT), an antipsychotic whose flexible butoxy linker joins a 2,3-dichlorophenylpiperazine to a dihydroquinolinone. Like the galunisertib structures, these are experimental structures, but the molecule differs from the other two in the respects that matter most for the five categories: it is flexible, two of its four rings are non-aromatic, and its lactam N–H is the only H-bond donor, providing a test of the method on extensive conformational polymorphism. Four space-group types occur: $P2_1$ (Forms I and X), $Pna2_1$ (II and VIII), $P\bar{1}$ (III, IV, V and VII) and $P2_1/n$ (IX); Forms III, IV and V have $Z' = 2$, the other forms $Z' = 1$. Cell volumes range from 1084 to 2269 Å³ (538–569 Å³ per molecule) and densities from 1.308 to 1.385 g/cm³ (Table 3). As for the other systems, equal category weights are used.

3.2 Category-Specific Similarity Patterns

Analysis of category-specific contributions reveals distinct patterns of conservation and variation across the five structural dimensions (Table 4). The category rankings provide insights into which structural features are most constrained by molecular chemistry and which provide polymorphic diversity. Figures 2, 8 and 14 show the similarity matrices of each category for sulfathiazole, galunisertib and aripiprazole, respectively, and reveal distinct clustering behavior. A detailed analysis of the similarity scores is presented in table 4.

Crystal packing has mean similarities of 0.84 for sulfathiazole, 0.76 for galunisertib and 0.74 for aripiprazole. In all three systems the packing coefficients and densities of different forms are similar (similarities ≥ 0.87 for every pair), so the variation comes from the reduced-cell lengths and obliquity, from differences in Z and Z' and, for galunisertib and aripiprazole, from differences in space-group type; the most similar crystal packing of all three systems is found for aripiprazole Forms I and X (0.983). **Interaction networks** vary strongly for sulfathiazole (0.58–0.87) and galunisertib (0.48–0.75), driven by the H-bond geometry rather than by the contact lengths. For aripiprazole they are more uniform (0.70–0.87, $\sigma = 0.04$), because every form makes the same single N–H $\cdots$ O bond per molecule; here too the H-bond angle is the most variable sub-metric (similarities 0.18–1.00, compared with 0.63–0.88 for the contact lengths). **Conformation** varies most for sulfathiazole (0.70–0.95, $\sigma = 0.10$), where it separates Forms II, III and IV from Forms I and V. For galunisertib and aripiprazole the category score is moderated by the bond and improper angles, which are nearly identical in all forms (similarities 0.82–0.98), whereas the matched torsions vary widely (0.11–0.90 and 0.24–0.87, respectively); for galunisertib they span the widest

range of any sub-metric outside crystal packing, comparable only to the H-bond geometry, and for aripiprazole the torsional differences are concentrated in the butoxy linker (Section 3.2.3). **Fragment orientation** has the lowest mean similarity in all three systems (0.49, 0.35 and 0.26): the relative orientations of neighbouring rings and the geometry of aromatic contacts differ appreciably between forms, while the fragment shapes, which are reported but not scored, are nearly identical (similarities ≥ 0.94). It is lowest for aripiprazole, which makes only two to four aromatic contacts per molecule. **Spatial geometry** is the most consistent category for sulfathiazole (0.78–0.86, $\sigma = 0.03$) and one of the two most consistent for galunisertib (0.65–0.84, $\sigma = 0.05$): in these fifteen structures every molecule is surrounded by 13–16 contacting neighbours, and the first coordination shells have similar shapes even when the cells and symmetry differ. For aripiprazole it is much less conserved (0.41–0.74, mean 0.56, $\sigma = 0.10$), because the layered triclinic $Z' = 2$ forms have 13–14 neighbours in a strongly anisotropic first shell and the other forms 15–18 neighbours in a less anisotropic one.

For sulfathiazole and galunisertib the category means follow the same order, crystal packing > spatial geometry > conformation > interaction networks > fragment orientation, although crystal packing exceeds spatial geometry by only about 0.02 in both. This order is unchanged when all scale parameters are halved; when they are doubled, spatial geometry and conformation move above crystal packing, six of whose seven sub-metrics do not depend on σ (Supplementary Table S23). For aripiprazole the order is different: interaction networks, conformation and crystal packing have similar means (0.74–0.77) and change places when the scale parameters are halved, while spatial geometry and fragment orientation remain the two lowest categories. The order of the category means is therefore a property of the molecule, which the fingerprint measures rather than assumes.

The similarity matrices of aripiprazole (Figure 14) show three groups of forms: Forms I and X have nearly the same crystal packing (0.983) but the least similar conformations of all pairs (0.629); Forms III, IV and V have almost the same conformation (0.89–0.91) but different reduced cells; and Forms II and VIII are the most similar pair overall (0.787). Like sulfathiazole, aripiprazole therefore shows both packing and conformational polymorphism, and the decomposition separates the two.

3.2.1 Crystal Packing Patterns

Sulfathiazole:

All five polymorphs belong to space group type No. 14, so the symmetry similarity is 1 for every pair once the settings are treated as equivalent; crystal packing therefore distinguishes the forms through their reduced cells, packing efficiency and multiplicity. Crystal packing groups the forms by multiplicity: pairs within Forms I, III and V ($Z' = 2$) or within Forms II and IV ($Z' = 1$) score 0.91–0.94, whereas pairs across the two groups score 0.76–0.82. The highest crystal-packing similarity is found for Form I versus Form III (0.937; Figure 3): both are $P2_1/c$ with similarly oblique reduced cells (largest deviation from 90° 17.8° versus 22.8° , similarity 0.868), and they differ mainly in packing efficiency (packing coefficient 0.696 versus 0.732, similarity 0.951; density 1.520 versus 1.600 g/cm³, similarity 0.950) and reduced cell lengths (similarity 0.792). Form I versus Form V follows at 0.936. These two forms are described in the $P2_1/c$ and $P2_1/n$ settings, respectively, but have identical space-group type and nearly identical packing efficiency (packing coefficient 0.696 versus 0.688, similarity 0.988; density 1.520 versus 1.510 g/cm³, similarity 0.993); they differ in their reduced cells, which are oblique for Form I and almost rectangular for Form V (17.8° versus 1.2° , similarity 0.730; lengths 10.53, 12.94, 17.19 Å versus 10.40, 14.28, 15.13 Å, similarity 0.838). Comparing symmetry operations in the deposited settings would give this pair a Jaccard similarity of only 0.333, an artifact of the setting rather than a difference in symmetry

(Section 2.4). The lowest similarity is found for Form III versus Form II (0.765), which differ in molecular multiplicity ($Z' = 2$ versus 1; similarity 0.5 for both Z and Z'), in reduced cell lengths (similarity 0.670) and in cell obliquity (22.8° versus 4.0° ; similarity 0.718), although their packing coefficients (0.732 versus 0.717) and densities (1.600 versus 1.574 g/cm³) agree to within about 2%. This illustrates how polymorphs can achieve equivalent densities through distinct packing solutions.

Galunisertib:

The highest crystal packing similarity is observed between Polymorphs V and IX (0.966; Figure 9). Both crystallize in $P2_1/c$ with $Z=4$ and $Z' = 1$, and their reduced cells (8.26, 11.64, 19.61 Å versus 8.11, 11.30, 20.85 Å, similarity 0.923; obliquity 12.8° versus 8.9° , similarity 0.891), densities (1.334 versus 1.300 g/cm³; 0.974) and packing coefficients (0.708 versus 0.691; 0.976) are close. The lowest similarity, 0.603, is observed between Polymorphs II and IV, which differ in space-group type ($P\bar{1}$ versus $P2_1/n$; Jaccard similarity 0.500), multiplicity ($Z=2$ versus 8, similarity 0.25; $Z' = 1$ versus 2, similarity 0.5), packing efficiency (density 1.209 versus 1.304 g/cm³, packing coefficient 0.619 versus 0.670), reduced cell lengths (similarity 0.771) and cell obliquity (the triclinic cell of Polymorph II deviates from 90° in all three angles, by 15.4° , 12.4° and 12.0° , the monoclinic cell of Polymorph IV in one, by 23.2° ; similarity 0.349). This illustrates the same principle observed in sulfathiazole: similar packing efficiencies can be achieved through fundamentally different crystallographic solutions.

Aripiprazole:

The packing coefficients of the nine forms vary little (0.652–0.697), and crystal packing distinguishes them mainly by space-group type, multiplicity and reduced cell. Forms I and X have the highest crystal-packing similarity of all three systems (0.983; Figure 15): both are $P2_1$ with $Z=2$ and have nearly the same reduced cells (similarity 0.932), obliquity (0.961), density and packing coefficient (both ≥ 0.99). Forms II and VIII, both $Pna2_1$, follow (0.973) and differ mainly in packing efficiency (packing coefficients 0.663 and 0.693), consistent with the report that Form VIII arises from a low-temperature phase transition [47]. The lowest similarity is found for Forms I and V (0.599), which differ in space-group type ($P2_1$ versus $P\bar{1}$), multiplicity ($Z' = 1$ versus 2), reduced cell lengths (0.565) and obliquity (0.323), although their packing efficiencies agree to within 1.5%.

3.2.2 Interaction Network Analysis

Sulfathiazole:

Each sulfathiazole molecule makes 85–105 unique intermolecular contacts and 3–4 hydrogen bonds (all N–H $\cdots$ N or N–H $\cdots$ O, consistent with the three N–H donors of the molecule), with mean D $\cdots$ A distances of 2.92–3.02 Å. The average contact lengths cluster tightly around 3.21–3.25 Å across all forms, indicating that optimal packing distances are maintained regardless of specific packing arrangement. Contact anisotropy indices range from 0.26 to 0.42, i.e. weakly to moderately directional networks. The highest interaction-network similarity is found for Form III versus Form II (0.866; Figure 4). Both forms make three H-bonds per molecule with nearly identical geometry (mean D $\cdots$ A 2.95 versus 2.92 Å, Wasserstein distance 0.047 Å, similarity 0.792; mean D–H $\cdots$ A 162.2° versus 162.7° , Wasserstein distance 1.19° , similarity 0.888), and the directions in which their contacts are made relative to the fragment planes are almost the same (polar-angle Wasserstein distance 1.12° , similarity 0.894). The lowest similarity is found for Form I versus Form II (0.584). Their contact length distributions remain similar (Wasserstein distance 0.049 Å, similarity 0.784), but Form I accommodates more bent H-bonds (mean angle 154.8° versus 162.7° ; Wasserstein distance 14.6° , similarity 0.233) and makes its contacts in different directions (in-plane

angle Wasserstein distance 9.3° , similarity 0.394). H-bond angle distributions are the most discriminating interaction descriptor, providing conformational information about interaction motifs not accessible from length analysis alone.

Galunisertib:

Galunisertib molecules make 125–172 unique intermolecular contacts and 1–3 H-bonds per molecule. The highest interaction-network similarity is observed between Polymorphs II and X (0.755; Figure 10), which both form two H-bonds per molecule of similar length (mean $D\cdots A$ 2.96 versus 2.98 Å, similarity 0.809) and have similar distributions of contact lengths and polar angles (similarities 0.787 and 0.854), although their mean H-bond angles differ (162.8° versus 151.8° , similarity 0.331). The lowest similarity is observed between Polymorphs I and V (0.476): Polymorph I forms a single $N-H\cdots O$ hydrogen bond per molecule ($D\cdots A$ 2.82 Å), whereas Polymorph V forms two (mean 3.11 Å), giving H-bond length and angle similarities of only 0.229 and 0.303. Contact anisotropy indices vary more widely across galunisertib polymorphs (0.03–0.43) than for sulfathiazole (0.26–0.42), indicating greater diversity in directional interaction preferences across this more crystallographically diverse dataset.

Aripiprazole:

Every aripiprazole form makes one H-bond per molecule, an $N-H\cdots O$ bond from the lactam $N-H$ ($D\cdots A$ 2.80–2.93 Å, $D-H\cdots A$ 158 – 175°), and 110–168 unique contacts. The interaction networks therefore differ only in H-bond geometry and contact directions, and the category is more uniform than for the other systems (0.70–0.87, $\sigma = 0.04$). The highest similarity is found for Forms II and VIII (0.868; Figure 16), whose H-bonds are almost identical (similarities 0.935 for length and 0.999 for angle) and whose contacts point in similar directions (0.850). The lowest is found for Forms III and X (0.705), whose contact lengths agree (0.792) but whose H-bond angles differ (175.1° versus 159.2° ; similarity 0.204). As for the other systems, the H-bond angle is the most discriminating interaction descriptor.

3.2.3 Conformational Differences

Sulfathiazole:

Conformation varies more than any other category for sulfathiazole (0.70–0.95, $\sigma = 0.10$). The matched torsions resolve three groups. Forms II, III and IV adopt nearly the same conformation (pairwise similarities 0.91–0.95; mean torsion differences 1.5 – 3.6°); Form V differs from them by 3.8 – 7.0° (0.78–0.80); and Form I is the most distinct, differing from all other forms by 13.6 – 18.5° (0.70–0.71). The highest similarity is found for Form III versus Form II (0.950; Figure 5), whose torsions (mean $|\Delta\tau| = 1.5^\circ$, similarity 0.953), bond angles (Wasserstein distance 0.27° , similarity 0.974), improper angles (0.50° , 0.951) and intramolecular contacts (0.016 Å, 0.922) all agree closely. The lowest is found for Form I versus Form IV (0.695): their torsions differ by 15.8° on average (similarity 0.590), with the largest differences, of 18 – 30° , about the C–S bond, whereas the torsions about the S–N bond agree to within 4.5° ; and their intramolecular contacts (9 versus 8 per molecule; Wasserstein distance 0.12 Å, similarity 0.559) and N pyramidalization (mean out-of-plane angle 16.3° versus 27.5° ; improper-angle similarity 0.727) also differ, whereas their bond angles are well conserved (1.0° , 0.905). Sulfathiazole therefore shows both packing polymorphism and conformational polymorphism within the same family. Across all pairs, bond angles are better conserved (similarities 0.87–0.98) than torsions (0.54–0.95), indicating that conformational changes occur primarily through torsional rotation rather than angle bending, consistent with the larger energetic cost of the latter.

Galunisertib:

Polymorphs II and VI represent the highest conformational similarity pair (similarity=0.850; Figure 11). Their matched torsions differ by only 3.3° on average (similarity 0.896), bond angles (1.16°, 0.891) and improper angles (0.95°, 0.909) are well conserved, and they have the same number of intramolecular contacts (22; Wasserstein distance 0.071 Å, similarity 0.702). The lowest conformational similarity is observed between forms IX and X (similarity=0.611). Their matched torsions differ by 45.2° on average (similarity 0.222); about all three inter-ring and ring–amide bonds the torsions differ by 10–82°, indicating that these two polymorphs represent genuine conformational polymorphs with fundamentally different rotameric states. Bond angle distributions remain well conserved (Wasserstein distance=1.23°, similarity=0.884), as do the improper angles (0.42°, 0.959), confirming that conformational changes occur primarily through torsional rotation rather than bond angle deformation, consistent with the energetic cost argument made for sulfathiazole. Intramolecular contact networks show marked differences (21 versus 30 contacts, Wasserstein distance=0.193 Å, similarity=0.381), with the larger contact count in form X reflecting rotational changes that bring additional atom pairs into through-space proximity.

Aripiprazole:

Twelve torsions are matched per molecule: six about the two piperazine N–C bonds, four along the butoxy linker and two about the aryl ether bond. The highest similarity is found for Forms IV and V (0.912; Figure 17; mean torsion difference 4.3°), and Forms III and IV are nearly as similar (0.910). The lowest is found for Forms I and X (0.629; mean torsion difference 38.4°), the pair with the most similar crystal packing: their piperazine torsions agree to within 11°, but two linker torsions that are gauche in Form I are anti in Form X (N–C–C–C 60° versus –171°; C–C–C–O –63° versus –174°), and the dihydroquinolinone ring is turned about the aryl ether bond (C–O–C–C 105° versus –173°), which also changes the intramolecular contacts (similarity 0.364). Forms II and VIII (0.717) differ in the same way in two linker torsions. Across all pairs, bond and improper angles are conserved (similarities 0.82–0.97) and torsions are not (0.24–0.87): as for the other systems, conformational differences arise from torsional rotation, here concentrated in the linker.

3.2.4 Fragment Orientation and π - π Stacking*Sulfathiazole:*

The fragment shapes are almost identical in all forms (shape-anisotropy similarities 0.96–1.00, reported but not scored), and fragment orientation distinguishes the forms through the relative orientation of neighbouring fragments and the geometry of their aromatic contacts. Each molecule makes 10–13 aromatic ring contacts (centroid–centroid ≤ 6 Å; mean distances 4.86–5.10 Å, mean perpendicular distances 2.03–2.23 Å). Edge-to-face contacts are the most common motif in all forms (42–85% of aromatic contacts), with mean interplanar angles of 52.7°–68.0°; Form V has the largest edge-to-face fraction (85%) and Form I the smallest (42%, with 33% tilted contacts). The highest fragment-orientation similarity is found for Form III versus Form IV (0.634; Figure 6), whose neighbouring fragments are oriented almost identically (relative-orientation similarity 0.790; Wasserstein distances 2.0–2.5° for the three axes), and the lowest for Form I versus Form II (0.384), where both the relative orientations (similarity 0.472) and the interplanar angles of the aromatic contacts (Wasserstein distance 16.2°) differ.

Galunisertib:

Galunisertib shows a very different aromatic packing: parallel-displaced contacts are the largest group in five of the ten forms (II, V, VII, VIII and X; in VI they tie with tilted contacts) and tilted contacts in Polymorphs III, IV and IX (parallel-displaced fractions 22–82%; mean interplanar angles 15.0°–53.8°), with 9–14 aromatic contacts per molecule. Polymorphs V, VIII and X are almost entirely parallel-stacked (80–82% parallel-displaced, mean interplanar angles 15–20°), whereas Polymorph I is the only form in which edge-to-face contacts are the largest group (44%). The highest fragment-orientation similarity is observed between Polymorphs IV and VI (0.497; Figure 12), and the lowest between Polymorphs I and V (0.249): Polymorph I is the form with the most edge-to-face contacts (44%) and Polymorph V one of the most parallel-stacked (80% parallel-displaced), so their interplanar-angle distributions differ by 38.7° (similarity 0.021), and the relative orientations of their neighbouring fragments also differ (similarity 0.279). This greater motif diversity reflects the more complex three-fragment aromatic architecture of galunisertib compared to the two-fragment sulfathiazole system.

Aripiprazole:

Each aripiprazole molecule makes only two to four aromatic contacts (centroid–centroid ≤ 6 Å), far fewer than sulfathiazole (10–13) or galunisertib (9–14); parallel-displaced contacts dominate in Forms II, IV, V and VIII and tilted contacts in Forms IX and X, and Forms I, III and VII combine two or three motifs. The fragment shapes are again nearly identical (similarities 0.98–1.00). The highest similarity is found for Forms II and VIII (0.634; Figure 18), which make two parallel-displaced contacts per molecule with the same interplanar angle (14.1°), and the lowest for Forms V and VII (0.147), whose relative ring orientations (0.181) and interplanar angles (0.023) both differ. With so few contacts, a single contact changes these distributions substantially, so fragment orientation is the least conserved category of aripiprazole (mean 0.26) and the one most sensitive to the aromatic-contact cutoff (Section 3.5).

3.2.5 Spatial Geometry Patterns

Sulfathiazole:

Spatial geometry is the most consistent category for sulfathiazole (0.78–0.86, $\sigma = 0.03$). Every molecule has 13–16 contacting neighbours (13.5 for Forms I and III, 13.0 for Forms II and V, and 16.0 for Form IV), and the first-shell anisotropies are small ($\kappa_s^2 = 0.05$ –0.08), indicating nearly isotropic first coordination shells. The first peak of the intermolecular fragment-centroid $g(r)$ lies at 4.45–5.75 Å. The highest similarity is found for Form III versus Form II (0.861; Figure 7): their $g(r)$ have the same first peak (4.55 Å; overlap 0.932) and their coordination numbers agree closely (13.5 versus 13.0; similarity 0.963). Form I versus Form III follows closely (0.859): both have 13.5 neighbours per molecule, similar first-shell shapes ($\kappa_s^2 = 0.074$ and 0.054; similarity 0.900) and closely matching $g(r)$ (overlap 0.891). The lowest similarity is found for Form V versus Form IV (0.784), which differ mainly in coordination number (13.0 versus 16.0; similarity 0.812), in the position of the first $g(r)$ peak (5.75 versus 4.45 Å) and in their neighbour-distance distributions (Wasserstein distance 0.129 Å, similarity 0.525). The fact that Form I–III is among the most similar pairs here, while it is one of the least similar in conformation (0.711) and interaction networks (0.629), illustrates the central point of the decomposition: two forms can share the spatial arrangement of their molecules while differing in how those molecules are shaped and bonded.

Galunisertib:

Spatial geometry similarities range from 0.65 to 0.84 (mean 0.74, $\sigma = 0.05$). Coordination numbers are 14–16 in all ten forms. The highest similarity is observed between Polymorphs VI and VII

(0.841; Figure 13): their coordination numbers (16.0 versus 15.5; similarity 0.969) and first-shell anisotropies ($\kappa_s^2 = 0.113$ versus 0.101; similarity 0.942) agree closely and their $g(r)$ overlap is 0.895, even though one is monoclinic ($P2_1/n$, $Z' = 1$) and the other triclinic ($P\bar{1}$, $Z' = 2$). The lowest similarity is observed between Polymorphs II and VIII (0.646), whose neighbour-distance distributions (Wasserstein distance 0.259 Å, similarity 0.274), first-shell shapes ($\kappa_s^2 = 0.156$ versus 0.070; similarity 0.652) and $g(r)$ (first peaks at 5.75 and 4.95 Å; overlap 0.784) all differ, although both are $P\bar{1}$ structures with $Z' = 1$. This demonstrates that similar spatial molecular distributions can be achieved through fundamentally different crystallographic realizations, and that the same space group does not imply a similar spatial arrangement.

Aripiprazole:

Spatial geometry is far less conserved than for sulfathiazole and galunisertib (0.41–0.74, mean 0.56), because two types of first coordination shell occur. The monoclinic and orthorhombic forms have 16–18 contacting neighbours in a less anisotropic shell ($\kappa_s^2 = 0.20$ –0.46), whereas the triclinic $Z' = 2$ Forms III, IV and V have 13–14 neighbours in a strongly anisotropic shell (0.68–0.72), as expected for a rod-like molecule packed in layers; Form VII is intermediate (15 neighbours, 0.49). The highest similarity is found for Forms II and VIII (0.744, the same similarity as Forms I and X; Figure 19; both forms have 18 neighbours and the first $g(r)$ peak at 5.45 Å), and the lowest for Forms II and V (0.413; 18 versus 14 neighbours, $\kappa_s^2 = 0.226$ versus 0.707). Unlike in sulfathiazole and galunisertib, where similar shells occur in forms with different cells and symmetry, spatial geometry in aripiprazole follows crystal packing (rank correlation 0.77; Section 3.5).

3.3 Comparison with COMPACK

Sulfathiazole:

Comparison with COMPACK [7] (20-molecule clusters) shows a strong correlation with our overall similarity over the 10 sulfathiazole pairs: Spearman $\rho = -0.89$ ($p < 0.001$) with the cluster RMSD, $\rho = 0.85$ ($p = 0.002$) with the simulated PXRD similarity, and $\rho = 0.73$ ($p = 0.016$) with the number of matched molecules (Pearson $r = -0.92$, 0.90 and 0.90, respectively; per-pair values in Supplementary Table S12). Both methods identify Forms II, III and IV as the most closely related group: COMPACK matches 16 of 20 molecules for Forms II–III and III–IV and 10 of 20 for II–IV, and these are the three pairs with the highest overall similarity in our method (0.814, 0.805 and 0.788). All pairs involving Form I or V match only one or two molecules in COMPACK. The decomposition adds information that COMPACK does not provide: Forms I–V and I–III have the highest crystal-packing similarities of all pairs (0.936 and 0.937) although COMPACK matches only one molecule for each, because these forms share their multiplicity and space-group type and have similar packing efficiency, but not the same molecular arrangement. Among the categories, conformation ($\rho = 0.95$), fragment orientation (0.69) and interaction networks (0.73) correlate with the number of matched molecules, whereas crystal packing ($\rho = -0.38$) and spatial geometry (-0.02) do not (Supplementary Table S14). The methods therefore provide complementary insights: COMPACK excels at precise geometric overlay and RMSD calculations for nearest-neighbor molecular clusters (typically 15–20 molecules), while this method offers (1) category-specific decomposition revealing why structures are similar, (2) distribution-based metrics capturing statistical properties beyond geometry, (3) complete periodic-crystal characterization rather than cluster-based sampling, and (4) computational efficiency enabling larger-scale comparisons. This complementarity suggests that integrated workflows combining both approaches could provide comprehensive polymorph assessment for crystal structure prediction and polymorph assessment for pharmaceutical development.

Galunisertib:

Comparison with COMPACK distance-based similarity analysis for the ten galunisertib polymorphs reveals an instructive contrast with the sulfathiazole results. COMPACK simulated PXRD similarities range from 0.749 to 0.981 (mean=0.905) and RMSD values from 0.211 to 1.800 Å (mean=0.959 Å) across all 45 pairs. The lowest RMSD pair is Polymorph IV versus Polymorph VI (RMSD=0.211 Å, matching 2 out of 20 molecules), while the highest PXRD similarity is observed between Polymorph IV and Polymorph I (PXRD=0.981, RMSD=1.364 Å), illustrating that RMSD and PXRD similarity do not necessarily rank pairs consistently even within the same method. Only one pair (V–VIII) matches more than five of the 20 molecules (10/20).

Unlike the sulfathiazole case, the correlation between COMPACK and our overall similarity is considerably weaker for galunisertib: Spearman $\rho = 0.44$ ($p = 0.002$) with the number of matched molecules, -0.33 ($p = 0.029$) with RMSD, and 0.38 ($p = 0.009$) with PXRD similarity. Polymorphs IV and VI, the pair with the lowest COMPACK RMSD, are also among the four most similar pairs in our method (overall 0.690; highest fragment-orientation similarity of all pairs, 0.497). This weaker correlation is itself a scientifically meaningful result: it reflects the greater structural diversity of the galunisertib dataset, where COMPACK’s cluster-based geometric overlay finds few matching molecules for most pairs (one or two of 20 for 38 of the 45 pairs) and therefore discriminates little between them, while our method captures a broader multi-dimensional picture including conformation, fragment orientation, and spatial organization. For highly diverse polymorph families spanning multiple crystal systems and space groups, the two methods sample complementary aspects of structural similarity rather than converging on the same ranking. This further motivates the use of our multi-dimensional framework as a complement to geometric overlay methods, particularly for systems with high crystallographic diversity.

Aripiprazole:

COMPACK matches only one molecule of 20 for 28 of the 36 pairs of aripiprazole forms; the largest numbers are found for Forms II–X (9), II–VIII (8), I–X and VIII–X (7) and III–IV (6) (Supplementary Table S13). The overall similarity correlates with the number of matched molecules ($\rho = 0.67$, $p < 0.001$), but not with the RMSD (-0.18 , $p = 0.28$) or the PXRD similarity (0.19 , $p = 0.25$). The correlation comes from spatial geometry (0.69) and crystal packing (0.60) (Supplementary Table S14): COMPACK matches large clusters for forms that share their lattice and coordination shell, such as I and X or II and VIII, whatever their conformations, whereas Forms III, IV and V, which share their conformation but not their reduced cells, match two to six molecules. For this flexible molecule, a large COMPACK cluster thus reflects shared packing rather than shared conformation.

3.4 Comparison with Invariant Descriptors

Table 5 and Figure 20 summarize the benchmark (Section 2.5; details in Supplementary Section S7). Sulfathiazole (47 structures: the 42 CSD depositions and the five DFT-D structures) and aripiprazole (21 CSD depositions) contain several depositions of most forms; galunisertib contains one form deposited twice, Polymorph X (DORDUM07 and DORDUM08).

Identifying the form of a deposition.

For sulfathiazole, the nearest neighbour of each of the 47 structures under the overall similarity belongs to its own form, as with AMD, PDD and SOAP, and all members of the form are ranked first

(mAP 1.00), as with PDD and SOAP but not AMD (0.85). Clustering into five groups recovers the forms less well than with PDD and SOAP (adjusted Rand index 0.70 versus 1.00; AUC 0.99 versus 1.00), because the deposited hydrogen positions of two CSD entries, SUTHAZ01 and SUTHAZ06, satisfy the H-bond criterion for fewer H-bonds than in the other depositions of their forms; without them, the overall similarity also gives an adjusted Rand index and AUC of 1.00 (Supplementary Table S15). For galunisertib, DORDUM07 and DORDUM08 are closer to each other than any two distinct polymorphs, by a factor of 3.4 for the fingerprint ($S = 0.917$), 2.6 for PDD, 7.3 for SOAP and only 1.1 for AMD. For aripiprazole, the nearest neighbour of each of the 16 depositions of forms with more than one entry belongs to its own form (mAP 1.00, as with PDD and SOAP; AMD 0.85), and every same-form pair is more similar (at least 0.871) than every pair of different forms (at most 0.792) (Supplementary Table S16). Identifying a known form is therefore not where the methods differ.

Variation between depositions of the same form.

Pairs of independent depositions of the same form show how much each category varies for reasons other than polymorphism (Figure 20c; Supplementary Tables S17 and S18). For sulfathiazole (210 pairs, including DFT-D versus experimental structures), the mean same-form similarities are 0.99 for crystal packing, 0.92 for spatial geometry, 0.91 for conformation, 0.85 for interaction networks and 0.83 for fragment orientation (overall 0.90). For the galunisertib pair they are 0.81–1.00 (overall 0.917), and for aripiprazole (33 pairs, all experimental) 0.91–1.00 (overall 0.95). Differences between distinct forms are best judged against these values: the fragment-orientation similarity of different forms (means 0.48 for sulfathiazole and 0.34 for aripiprazole), for example, lies far below the same-form range.

Ranking distinct polymorphs.

The methods differ when the forms being compared are distinct, the situation the fingerprint is designed for. The overall similarity correlates with the number of molecules matched by COMPACT with $\rho = 0.73$ for sulfathiazole (10 pairs), 0.44 for galunisertib (45 pairs) and 0.67 for aripiprazole (36 pairs), compared with 0.12–0.51, 0.08–0.16 and 0.14–0.37 for AMD, PDD and SOAP (Table 5). With 10 pairs for sulfathiazole, and for galunisertib, where the pair-bootstrap 95% intervals of the differences include zero, this is a consistent trend; for aripiprazole the intervals exclude zero ($[0.19, 0.86]$, $[0.05, 0.58]$ and $[0.01, 0.61]$ for AMD, PDD and SOAP), although COMPACT matches a single molecule for 28 of its 36 pairs (Supplementary Tables S19 and S16). The fingerprint does not agree better with every COMPACT measure: for galunisertib SOAP agrees better with the PXRD similarity (0.55 versus 0.38), and for aripiprazole AMD, PDD and SOAP follow the RMSD more closely (correlations with $-\text{RMSD}$ of 0.32–0.43 versus 0.18).

Differences between the methods

Figure 20d shows which categories the AMD, PDD and SOAP distances follow (Spearman correlation of the negative distance with each category similarity; Supplementary Table S20). In none of the three systems do they follow crystal packing ($\rho = -0.20$ to 0.08). What they do follow depends on the molecule: for sulfathiazole all other categories (up to 0.81, for spatial geometry), for galunisertib mainly spatial geometry (0.28–0.41) and fragment orientation (0.19–0.27), and for aripiprazole conformation (0.22–0.41), fragment orientation (0.22–0.31) and, for SOAP, interaction networks (0.40). The categories through which the fingerprint follows COMPACT also differ between the systems: conformation for sulfathiazole ($\rho = 0.95$ with N_{match}), fragment orientation and conformation for galunisertib (0.50 and 0.41), and spatial geometry and crystal packing for aripiprazole (0.69 and 0.60; Supplementary Table S14). Atom-based invariants respond to local atomic

environments, which polymorphs of the same molecule largely share, and do not indicate which of these aspects separates two forms. Pairs at the same distance can therefore differ in different respects: galunisertib Polymorphs IV–VIII and V–IX have the same SOAP distance (0.092) but crystal-packing similarities of 0.62 and 0.97 (Figure 20e; examples for AMD and PDD in Supplementary Section S7).

Invariance and cost.

AMD, PDD, SOAP, the reduced cell and the density are invariant to setting, frame and cell choice, as is the fingerprint, whereas the cell as deposited is not (Supplementary Table S22); invariance is therefore a property the fingerprint shares with the invariant descriptors, not one that distinguishes it from them. Fingerprints and all 1081 comparisons of the 47 sulfathiazole structures take about 10 s, and AMD, PDD and SOAP about 6 s; for the 11 galunisertib structures both take about 1 s.

3.5 Robustness, Uncertainty and Redundancy

For the distinct forms of all three systems, we tested how the similarities respond to uncertainty in the coordinates and to the cutoffs of the method, and whether the categories carry independent information (Supplementary Section S8, Figure S1).

Coordinate uncertainty.

Every atom was displaced by an isotropic Gaussian vector of 0.01, 0.02, 0.05 or 0.10 Å rms (ten realizations per level; cell and contact lists fixed), and all coordinate-dependent quantities were recomputed. For comparison, positional uncertainties of non-hydrogen atoms in well-refined structures are about 0.01 Å or less, and the DFT-D minimizations moved the sulfathiazole atoms by 0.04–0.11 Å. In all three systems, the similarity of a structure to its perturbed copy is 0.97–0.98 at 0.01 Å and 0.86–0.88 at 0.10 Å; spatial geometry changes least (0.96 or higher) and conformation most (0.73–0.76 at 0.10 Å) (Supplementary Table S26). The mean same-form similarities of the CSD depositions (0.90 for sulfathiazole, 0.95 for aripiprazole) thus correspond to displacements of 0.05–0.10 Å and about 0.02 Å, respectively. For sulfathiazole and galunisertib, the ranking of the pairs of distinct forms is preserved (mean rank correlation with the reference ≥ 0.99 at 0.02 Å and ≥ 0.92 at 0.10 Å; standard deviation of a pair similarity on average at most 0.011, against a spread of 0.03–0.06 between pairs). For aripiprazole it is less stable (0.89 already at 0.01 Å), because Form VII has a second N–H $\cdots$ N contact with H $\cdots$ N = 2.75 Å, at the H-bond criterion, which displacements of 0.01 Å switch on and off; without the eight pairs involving Form VII, the rank correlation is 0.99 at 0.01 Å and 0.92 at 0.10 Å.

Cutoffs.

Each of the five cutoffs of the method (aromatic-contact, orientation and environment cutoffs, minimum H-bond angle, $g(r)$ smoothing) was set to a lower and a higher value (Supplementary Table S27). For sulfathiazole the overall similarity changes by at most 0.014 and its ranking is preserved (rank correlation ≥ 0.98), and for galunisertib by at most 0.041 (≥ 0.94). The affected categories respond more, most strongly the galunisertib fragment orientation to a 6 Å orientation cutoff (0.72) and the galunisertib interaction networks to a 130° minimum H-bond angle (0.78), because most galunisertib forms make only one or two H-bonds per molecule. For aripiprazole all cutoffs except the aromatic-contact cutoff change the overall similarity by at most 0.034 (≥ 0.92); the aromatic-contact cutoff matters more because aripiprazole makes only two to four aromatic contacts per molecule, and it changes the fragment-orientation ranking substantially (0.45 at 5 Å

instead of 6 Å) and the overall ranking moderately (0.75 at 5 Å and 0.90 at 7 Å; largest change 0.081). The van der Waals contact criterion is applied during preprocessing and was not varied.

Redundancy of the categories.

Over the pairs of distinct forms (Supplementary Table S25), the category similarities of sulfathiazole are partly correlated: interaction networks, conformation and fragment orientation ($\rho = 0.55\text{--}0.94$), because all three separate Forms II, III and IV from Forms I and V, but not crystal packing and spatial geometry ($|\rho| \leq 0.32$). For galunisertib all correlations are small ($|\rho| \leq 0.40$), and for aripiprazole too ($|\rho| \leq 0.21$), except crystal packing and spatial geometry (0.77), which reflect the same division into forms with and without layered packing. Omitting any one category changes the ranking of the pairs (rank correlation with the five-category value 0.95–1.00 for sulfathiazole, 0.79–0.96 for galunisertib and 0.82–0.96 for aripiprazole; Supplementary Table S24). No category is therefore redundant across the three systems.

4 Discussion

4.1 Multi-Dimensional Structural Signatures

The category-specific analysis (Table 4) shows which structural features are conserved, which distinguish the forms, and how this depends on the molecule. Two observations hold for all three systems. Fragment orientation is the least conserved category (means 0.49, 0.35 and 0.26 for sulfathiazole, galunisertib and aripiprazole), because the relative orientations of neighbouring rings and the geometry of aromatic contacts change between forms while the fragment shapes do not; and bond angles are conserved (similarities 0.82–0.98), so that conformational differences arise from torsional rotation. Beyond these, the order of the categories is a property of the molecule. For sulfathiazole and galunisertib, crystal packing (0.84 and 0.76) and spatial geometry (0.82 and 0.74) have the highest means, followed by conformation, interaction networks and fragment orientation: packing efficiencies are similar, and every molecule has 13–16 contacting neighbours in a nearly isotropic first shell. Molecular shape and the overall spatial arrangement of molecules are thus the most strongly conserved features of these two families, one DFT-optimized and one experimental, while conformation, hydrogen-bond geometry and the relative orientation of aromatic rings provide the adjustable parameters for polymorphic variation. For aripiprazole, a flexible, rod-like molecule with non-aromatic rings, spatial geometry is far less conserved (mean 0.56), because its forms pack in two types of coordination shell, and interaction networks, conformation and crystal packing have similar means (0.74–0.77) (Section 3.2). The fingerprint therefore measures the relative conservation of the categories rather than assuming it.

Where the first coordination shells are conserved, they reveal structural degeneracy: different crystallographic solutions achieve similar spatial outcomes. This parallels the large-scale analysis of inorganic polymorphs by Dey *et al.*, who found that polymorphs frequently retain the same local polyhedral environments in different space groups [24]. In sulfathiazole and galunisertib, the first shells are similar even between forms of different space-group type; for example, galunisertib Polymorphs VI ($P2_1/n$, $Z' = 1$) and VII ($P\bar{1}$, $Z' = 2$) have the highest spatial-geometry similarity of all galunisertib pairs (0.841). In aripiprazole, by contrast, the shell follows the packing (rank correlation between spatial geometry and crystal packing 0.77; Section 3.5): forms that share their lattice share their shell, and forms with layered packing do not. The fingerprint thus extends to molecular crystals the observation that local environments can be more conserved than symmetry, shows when they are not, and quantifies in the remaining categories the respects in which such forms nevertheless differ.

4.2 Methodological Advantages

The modular architecture with distribution-based metrics provides several advantages. These arise from the way established techniques (symmetry expansion of the asymmetric unit, periodic neighbour search and the Wasserstein comparison of distributions) are combined into molecule-level descriptors, rather than from the techniques themselves. **Invariance:** all descriptors are independent of the unit cell, space-group setting and Cartesian frame, which we verified to numerical precision (Section 2.4). **Complete interaction network characterization:** contacts are evaluated on the full periodic crystal and their directions are expressed in molecule-fixed frames, so they can be compared between structures with different cells and symmetry. **Comprehensive π - π stacking analysis:** neighbour search in the periodic crystal with user-defined fragment filtering captures the complete aromatic interaction network with geometric motif classification. **Multi-scale spatial characterization:** fragment environments within 12 Å capture ordering from the first coordination shell ($\sim$ 4–5 Å) to the third shell. **Distribution-based robustness:** Wasserstein distance metrics handle varying sample sizes and capture full distribution shapes rather than relying on means/variances, providing more discriminatory power. **Independent category validation:** modular design enables separate validation of each component while maintaining systematic integration. **Computational efficiency:** no iterative structural alignment required; a complete analysis of any of the three polymorph families runs in under 3 s. The benchmark of Section 3.4 quantifies what the decomposition adds to invariant descriptors. The fingerprint identifies the form of a deposition as reliably as PDD and SOAP, and more reliably than AMD, but does not outperform them. For distinct polymorphs, AMD, PDD and SOAP follow the local atomic environments, never crystal packing, and agree less with COMPACK cluster matching than the fingerprint does in all three systems, significantly so only for aripiprazole; the fingerprint in addition states in which respect two forms differ. Invariant descriptors are therefore the natural choice for establishing whether two structures are the same and for searching large databases, and the fingerprint for describing how polymorphs differ.

4.3 Extension to Other Polymorph Families

While demonstrated here for sulfathiazole, galunisertib and aripiprazole, our method is readily applicable to other pharmaceutically relevant polymorph families with diverse characteristics. **Ritonavir** (2 polymorphs): the case where Form II appeared unexpectedly during manufacturing, compromising Form I production. Our method could quantify structural relationships and potentially predict Form II-like structures from Form I data. **Sulfapyridine** (5 polymorphs): another sulfonamide with extensive polymorphism, providing a close chemical analog to sulfathiazole for methodology validation. **Furosemide** (3 polymorphs): a loop diuretic with known polymorphic variations affecting bioavailability. **Paracetamol** (3 well-characterized polymorphs): the extensively-studied analgesic provides a benchmark system with significant pharmaceutical importance. **Carbamazepine** (3 polymorphs): an anticonvulsant with documented polymorphism-related formulation challenges. **ROY** (14 polymorphs): the record-holder for documented polymorphs in a single compound, presenting maximum structural diversity for methodology testing. Sulfathiazole solvates, co-crystals and salts [12] are a natural next step. Examination of these systems will be undertaken in future work.

Each system may exhibit different category-specific signatures reflecting chemical and structural characteristics: highly flexible molecules might show greater conformational variation, while rigid aromatics may display conserved conformations with variable packing. The modular framework enables systematic exploration of these patterns, potentially revealing general principles of

polymorphic variation across chemical space.

4.4 Future Directions

Several extensions emerge naturally from this work. **Polymorph prediction integration:** the structural similarity metrics could inform crystal structure prediction by (1) identifying characteristic fingerprints of experimentally observed forms, (2) screening computationally generated structures for ‘experimental-like’ signatures before energy evaluation, (3) guiding structure generation toward regions of fingerprint space known to produce stable polymorphs, and (4) validating predicted structures by comparing similarity profiles to known polymorphs. CrystalMath [9] generates candidate structures by aligning molecular inertia axes and ring normals with crystallographic directions and ranks them by contact distributions and free volume, the fragment-level quantities that enter the fragment-orientation and interaction-network categories; the fingerprint is therefore a natural, energy-free criterion for clustering, de-duplicating and ranking its candidates. For molecular crystals, diversity in fingerprint space, which also distinguishes forms that share a space group, could similarly complement the space-group diversity control used by Omeo *et al.* [25]. For sulfathiazole specifically, this could enable prediction of previously undiscovered polymorphs by generating structures with fingerprints intermediate between known forms or extrapolating beyond the characterized space.

Cocrystal-polymorph relationships: analysis of sulfathiazole multicomponent crystal structures (cocrystals and salts) using our framework could reveal structural relationships with polymorphs, potentially identifying cocrystal-polymorph continuums. Questions include: Do cocrystals exhibit similar spatial geometry as polymorphs? Are π - π stacking patterns preserved? This could inform cocrystal design strategies.

Machine learning integration: the five-category fingerprints provide natural feature vectors of fixed dimensionality for supervised learning applications. For *stability ranking prediction*, fingerprint vectors could serve as inputs to regression models trained on computed lattice energies, learning to predict relative polymorph stability without explicit energy calculations; once trained on a sufficiently diverse set of structures, such models could rapidly rank candidate polymorphs generated by crystal structure prediction. For *automated form classification*, the fingerprint space enables k -nearest-neighbor or support vector machine classifiers to assign newly generated structures to known polymorphic forms, useful for high-throughput screening applications. Unsupervised learning could identify latent structural patterns not captured by pre-defined categories, potentially revealing new organizing principles. The modular design facilitates these extensions without requiring fundamental architecture changes, positioning the framework for integration into next-generation computational crystallography workflows.

4.5 Limitations

Several limitations merit discussion. The method requires high-quality crystal structure data with complete atomic coordinates. Disordered structures or those with missing atoms may show reduced performance. The DFT-optimized sulfathiazole structures eliminate experimental uncertainties, but real-world applications must account for data quality variations. Sulfathiazole is relatively rigid compared to many pharmaceuticals. The inclusion of ten experimental galunisertib polymorphs from the CSD demonstrates that the method remains effective in the presence of thermal effects and experimental uncertainties, although observed structural differences may partially reflect data collection conditions rather than intrinsic polymorphic variation. The greater conformational flexibility of galunisertib relative to sulfathiazole is reflected in its matched tor-

sions, which vary far more between forms than any other conformational descriptor (similarities 0.11–0.90, compared with 0.82–0.98 for bond and improper angles), while spatial geometry remains consistent. Highly flexible molecules may show reduced spatial geometry conservation, as aripiprazole does, where forms with layered and non-layered packing of the rod-like molecule have different first coordination shells. For flexible molecules, the fragment-level descriptors must be restricted to user-specified ring systems (as for aripiprazole), and contacts of the chain enter only through their lengths. H-bond candidates at the H-bond criterion (as in aripiprazole Form VII) make the interaction networks sensitive to small coordinate changes, and molecules with few aromatic contacts give fragment-orientation similarities that depend on the aromatic-contact cutoff. The three test systems are single-component pharmaceutical crystals; solvates, co-crystals, salts and non-pharmaceutical molecular crystals have not yet been tested. The interaction-network category depends on the deposited hydrogen positions: for two of the 42 experimental sulfathiazole depositions (SUTHAZ01 and SUTHAZ06) the H-bond criterion recognizes fewer H-bonds than in the other depositions of the same form, which lowers their similarity to those depositions (Section 3.4); placing or normalizing hydrogen atoms consistently before the analysis would reduce this sensitivity. Where a molecule makes only one or two hydrogen bonds, as in most galunisertib forms, the H-bond length and angle distributions contain only a few values per structure, so their Wasserstein distances compare individual H-bonds rather than distributions and are correspondingly sensitive to single bonds. Absolute similarity values depend on the scale parameters σ of Eq. 8; rankings of polymorph pairs do not (Supplementary Table S23), and similarity values should therefore be compared only when computed with the same parameters. Category weights currently use equal values (0.20 each) providing unbiased comparison. System-specific optimization could improve discriminatory power but requires training data. Automated weight learning from machine learning approaches could address this. While efficient for typical crystals ($Z < 16$, < 100 atoms/molecule), very large systems with $Z' > 2$ or very large molecules may require computational optimizations. The purely geometric approach provides orthogonal information to energy-based methods but cannot predict thermodynamic stability rankings. Integration with lattice energy calculations could provide complementary insights.

5 Conclusions

We have presented a comprehensive multi-dimensional fingerprinting method for quantitative polymorph comparison that employs modular architecture, distribution-based similarity metrics, and systematic multi-scale characterization. The five-category framework (crystal packing, intermolecular interactions, molecular conformation, fragment orientation, spatial geometry) separates structural variation into independent yet complementary dimensions, each weighted equally (0.20) to provide unbiased overall similarity assessment. All descriptors are computed on the full periodic crystal and are invariant to the unit cell, space-group setting and Cartesian frame, which we verified to numerical precision; this invariance is essential, since with setting-dependent descriptors a change of setting alone alters crystal-packing similarities by up to 0.11 (Section 2.4).

For five DFT-optimized sulfathiazole polymorphs, spatial geometry is the most consistent category (0.78–0.86), conformation and interaction networks vary most (0.70–0.95 and 0.58–0.87), and the conformational and interaction-network similarities separate Forms II, III and IV from Forms I and V, in agreement with COMPACT cluster matching (Spearman $\rho = -0.89$ with RMSD). Each molecule has 13–16 contacting neighbours and makes 3–4 hydrogen bonds; H-bond angle distributions are the most discriminating interaction descriptor, and edge-to-face contacts are the most common aromatic motif in every form (42–85%).

For ten experimental galunisertib polymorphs, spatial geometry and conformation are the most consistent categories, crystal packing and interaction networks vary most, and the matched torsions, together with the H-bond geometry, are the most variable descriptors outside crystal packing, consistent with the greater flexibility of the molecule; parallel-displaced and tilted contacts, rather than edge-to-face contacts, dominate the aromatic packing. The order of the category means is the same as for sulfathiazole. The weaker correlation with COMPACK for galunisertib ($\rho = 0.44$ with the number of matched molecules) reflects the crystallographic diversity of this family, for which COMPACK matches only one or two molecules for most pairs, and demonstrates the complementary nature of the two approaches.

For nine experimental aripiprazole forms, interaction networks, conformation and crystal packing have similar means, spatial geometry is far less conserved than for sulfathiazole and galunisertib, and fragment orientation is again the least conserved category. The decomposition separates conformational polymorphs that pack in almost the same lattice (Forms I and X) from packing polymorphs of the same conformer (Forms III, IV and V), and the overall similarity follows the number of molecules matched by COMPACK ($\rho = 0.67$).

In a benchmark on all CSD depositions of the three systems, the fingerprint identifies the form of every sulfathiazole and aripiprazole deposition as reliably as the invariant descriptors AMD, PDD and SOAP, and recognizes the two galunisertib depositions of Polymorph X. For distinct forms it follows COMPACK cluster matching more closely than these descriptors ($\rho = 0.73, 0.44$ and 0.67 with the number of matched molecules, against $0.08\text{--}0.51$), clearly so for aripiprazole, although not the simulated PXRD similarity. The depositions of the same form define an experimental noise floor for each category, lowest for crystal packing and highest for interaction networks and fragment orientation.

The method’s computational efficiency, interpretability through category-specific decomposition, and validation against established benchmarks make it suitable for high-throughput applications in solid-form screening and crystal structure prediction. Structural similarity alone does not determine the choice of a pharmaceutical form, which also depends on stability, solubility, dissolution and mechanical properties [3, 4]; the fingerprints are intended as a fast structural input to that process, for example to flag new forms that are structurally distinct from known ones. Extension to other polymorph families (ritonavir, paracetamol, carbamazepine, ROY) could reveal general principles of polymorphic variation, while integration with machine learning frameworks and crystal structure prediction workflows positions the method for next-generation computational crystallography applications and will be considered in future studies. By providing analytical tools for polymorph classification, structural relationship analysis, and prediction validation, this work advances quantitative methods for solid-state chemistry with potential applications in solid-form screening, crystal structure prediction and materials discovery.

6 Acknowledgements

M.E.T. and N.G. acknowledge funding from the National Science Foundation, DMR-2118890.

Tables

Table 1: Crystallographic data for sulfathiazole polymorphs

Form	Refcode ^a	Space Group	Z	Z'	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	β (°)	<i>V</i> (Å ³)	ρ (g/cm ³)	η
Form I	SUTHAZ16	P2 ₁ /c	8	2	10.534	12.936	17.191	107.77	2230.81	1.520	0.696
Form II	SUTHAZ18	P2 ₁ /c	4	1	8.193	8.538	15.437	94.01	1077.20	1.574	0.717
Form III	SUTHAZ17	P2 ₁ /c	8	2	17.448	8.498	15.511	112.81	2120.00	1.600	0.732
Form IV	SUTHAZ19	P2 ₁ /n	4	1	10.774	8.467	11.367	91.65	1036.51	1.636	0.747
Form V	SUTHAZ05	P2 ₁ /n	8	2	10.399	15.132	14.280	91.21	2246.57	1.510	0.688

η =packing coefficient; ρ =crystal density. All structures DFT-optimized. ^aCSD entry from which the DFT-D structure was minimized.

Table 2: Crystallographic data for galunisertib polymorphs

Form	Refcode	Space Group	Z	Z'	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	α (°)	β (°)	γ (°)	<i>V</i> (Å ³)	ρ (g/cm ³)	η
I	DORDUM09	P2 ₁ /n	4	1	19.431	11.325	9.151	90	95.302	90	2005.00	1.224	0.660
II	DORDUM	$P\bar{1}$	2	1	8.212	11.510	11.875	102.001	105.368	102.349	1014.63	1.209	0.619
III	DORDUM01	$P\bar{1}$	2	1	7.712	8.911	14.617	81.906	78.799	75.258	948.50	1.293	0.665
IV	DORDUM02	P2 ₁ /n	8	2	18.998	9.853	21.883	90	113.219	90	3764.52	1.304	0.670
V	DORDUM03	P2 ₁ /c	4	1	8.260	19.607	12.691	90	116.545	90	1838.73	1.334	0.708
VI	DORDUM04	P2 ₁ /n	4	1	10.727	14.471	12.705	90	107.112	90	1884.91	1.302	0.669
VII	DORDUM10	$P\bar{1}$	4	2	8.134	12.620	18.059	86.707	100.094	83.303	1806.87	1.358	0.708
VIII	DORDUM05	$P\bar{1}$	2	1	8.554	11.372	11.567	65.784	105.495	77.315	916.89	1.338	0.709
IX	DORDUM06	P2 ₁ /c	4	1	11.301	20.847	8.106	90	98.879	90	1886.95	1.300	0.691
X	DORDUM07	$C2/c$	8	1	27.938	8.082	22.004	90	133.550	90	3600.82	1.363	0.700

η =packing coefficient; ρ =crystal density. All structures are experimental CSD depositions (refcodes DORDUM–DORDUM10, excluding DORDUM08, a second deposition of Polymorph X).

Table 3: Crystallographic data for aripiprazole polymorphs (one representative deposition per form)

Form	Refcode	Space Group	Z	Z'	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	α (°)	β (°)	γ (°)	<i>V</i> (Å ³)	ρ (g/cm ³)	η
I	MELFIT01	P2 ₁	2	1	8.679	7.568	17.381	90	94.500	90	1138.14	1.308	0.652
II	MELFIT15	<i>Pna</i> 2 ₁	4	1	23.608	12.471	7.706	90	90	90	2268.80	1.313	0.663
III	MELFIT	<i>P</i> $\bar{1}$	4	2	10.150	12.080	18.603	82.759	81.447	83.246	2226.16	1.338	0.671
IV	MELFIT04	<i>P</i> $\bar{1}$	4	2	8.518	9.035	30.417	88.072	86.550	73.874	2244.36	1.327	0.661
V	MELFIT06	<i>P</i> $\bar{1}$	4	2	12.263	13.787	14.741	101.396	108.921	98.847	2246.31	1.326	0.660
VII	MELFIT07	<i>P</i> $\bar{1}$	2	1	7.027	9.977	15.945	81.773	78.728	85.888	1083.87	1.374	0.692
VIII	MELFIT11	<i>Pna</i> 2 ₁	4	1	21.671	13.125	7.646	90	90	90	2174.86	1.369	0.693
IX	MELFIT19	P2 ₁ /n	4	1	16.613	6.953	19.463	90	106.990	90	2150.21	1.385	0.697
X	MELFIT05	P2 ₁	2	1	8.867	7.762	16.485	90	93.250	90	1132.80	1.315	0.657

η =packing coefficient; ρ =crystal density. All structures are experimental CSD depositions (refcode family MELFIT); no structure is available for Form VI. The other deposition of each form are listed in Supplementary Table S2.

Table 4: Category-specific similarity ranges across all polymorph pairs(sulfathiazole, galunisertib and aripiprazole)

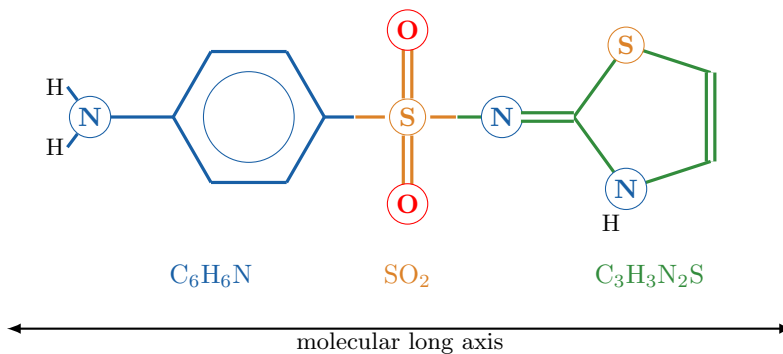
Category	Min	Max	Mean $\pm \sigma$	Range	Interpretation
<i>Sulfathiazole (10 pairs)</i>					
Crystal packing	0.76	0.94	0.84 $\pm$ 0.08	0.17	Moderate, variable
Interaction networks	0.58	0.87	0.70 $\pm$ 0.10	0.28	Moderate, variable
Conformation	0.70	0.95	0.80 $\pm$ 0.10	0.25	Moderate, variable
Fragment orientation	0.38	0.63	0.49 $\pm$ 0.08	0.25	Low, variable
Spatial geometry	0.78	0.86	0.82 $\pm$ 0.03	0.08	Moderate, consistent
Overall	0.66	0.81	0.73 $\pm$ 0.05	0.15	Moderate, variable
<i>Galunisertib (45 pairs)</i>					
Crystal packing	0.60	0.97	0.76 $\pm$ 0.10	0.36	Moderate, variable
Interaction networks	0.48	0.75	0.62 $\pm$ 0.07	0.28	Low, variable
Conformation	0.61	0.85	0.71 $\pm$ 0.05	0.24	Moderate, consistent
Fragment orientation	0.25	0.50	0.35 $\pm$ 0.06	0.25	Low, variable
Spatial geometry	0.65	0.84	0.74 $\pm$ 0.05	0.19	Moderate, consistent
Overall	0.57	0.72	0.64 $\pm$ 0.03	0.15	Low, consistent
<i>Aripiprazole (36 pairs)</i>					
Crystal packing	0.60	0.98	0.74 $\pm$ 0.09	0.38	Moderate, variable
Interaction networks	0.70	0.87	0.77 $\pm$ 0.04	0.16	Moderate, consistent
Conformation	0.63	0.91	0.76 $\pm$ 0.08	0.28	Moderate, variable
Fragment orientation	0.15	0.63	0.26 $\pm$ 0.09	0.49	Low, variable
Spatial geometry	0.41	0.74	0.56 $\pm$ 0.10	0.33	Low, variable
Overall	0.53	0.79	0.62 $\pm$ 0.05	0.25	Low, variable

Values are computed over all pairwise comparisons of the five sulfathiazole (10 pairs) and ten galunisertib (45 pairs) polymorphs and the nine aripiprazole forms (36 pairs); σ is the sample standard deviation; the bands are applied to the unrounded values. Interpretation bands: high $\bar{s} \geq 0.85$, moderate $0.70 \leq \bar{s} < 0.85$, low $\bar{s} < 0.70$; consistent $\sigma \leq 0.05$, variable $\sigma > 0.05$ (see Methods).

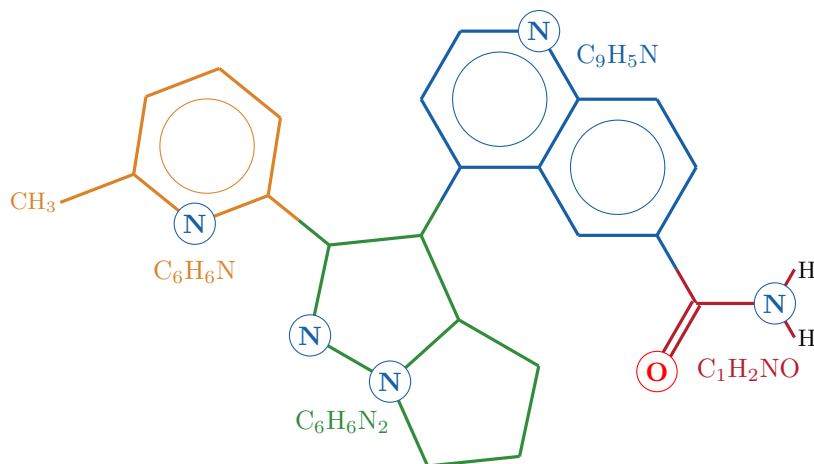
Table 5: **Benchmark against other comparison methods.** Identification of the form: mean average precision (mAP) and ROC area for same-form versus different-form pairs (AUC) for the 47 sulfathiazole and 21 aripiprazole structures; for galunisertib, the distance between DORDUM07 and DORDUM08 (the same polymorph) divided by the smallest distance between distinct polymorphs (below 1: the pair is recognized). Aripiprazole: 16 queries (forms with more than one depositions). COMPACK PXR: 46 sulfathiazole structures (no COMPACK results for SUTHAZ01). Right: Spearman correlation with the number of molecules matched by COMPACK for the pairs of distinct forms. Per-category results and all other measures are given in Supplementary Section S7.

Method	Identification of the form					ρ with COMPACK N_{match}		
	SLFZ (47)		ARIP (21)		GAL (11)	SLFZ	GAL	ARIP
	mAP	AUC	mAP	AUC	ratio	(10)	(45)	(36)
This work, overall S	1.00	0.99	1.00	1.00	0.29	+0.73	+0.44	+0.67
AMD ($k = 100$)	0.85	0.94	0.85	0.99	0.89	+0.12	+0.13	+0.14
PDD ($k = 100$)	1.00	1.00	1.00	1.00	0.38	+0.51	+0.16	+0.36
SOAP/REMatch	1.00	1.00	1.00	1.00	0.14	+0.42	+0.08	+0.37
Reduced cell	0.96	0.99	1.00	1.00	0.12	-0.54	+0.01	+0.13
Density	0.62	0.83	0.55	0.83	7.60	+0.24	+0.04	+0.25
COMPACK PXR	1.00	1.00	1.00	1.00	0.12	-	-	-

Figures



Scheme 1a. Molecular structure of sulfathiazole with the three fragments used in the fingerprinting analysis highlighted: the aniline ring (blue), the sulfonyl bridge (amber), and the thiazole ring (green). The aniline and thiazole rings are the aromatic fragments used in the π -contact analysis.



Scheme 1b. Molecular structure of galunisertib with the four fragments used in the fingerprinting analysis highlighted: the methylpyridine ring (amber), the pyrazole-dihydropyrrrole system (green), the quinoline bicyclic system (blue), and the amide group (red). The first three are the aromatic fragments used in the π -contact analysis.

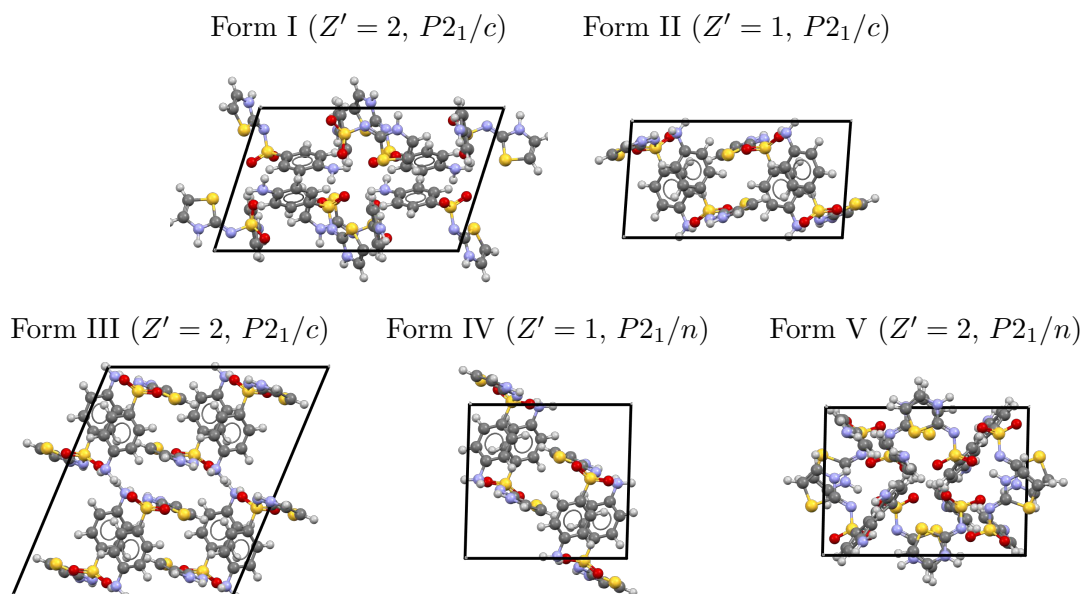
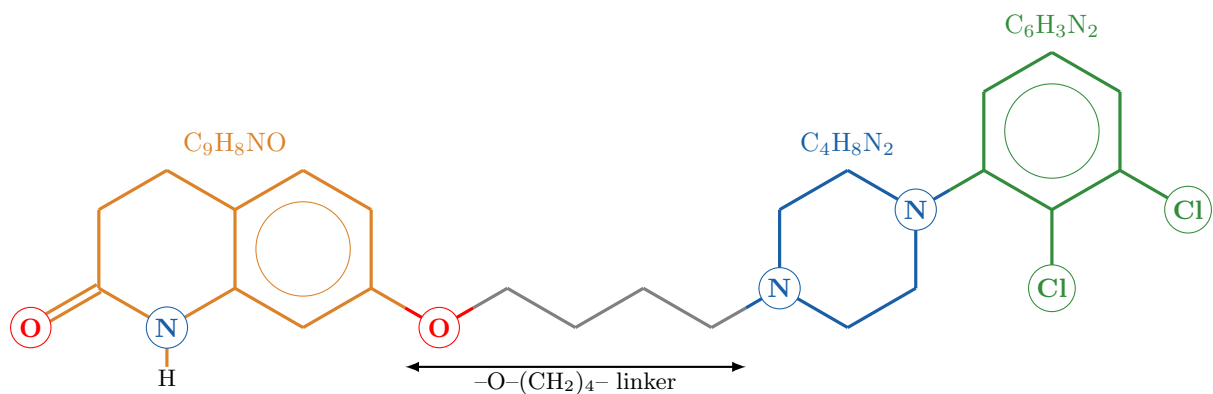


Figure 1: Crystal structures of the five DFT-optimized sulfathiazole polymorphs visualized using Mercury, showing the unit cell packing viewed along the b -axis. Each panel indicates the space group and number of symmetry-independent molecules in the asymmetric unit (Z'). Forms I, III, and V ($Z' = 2$) contain two symmetry-independent molecules, while Forms II and IV ($Z' = 1$) contain one. All polymorphs crystallize in space group No. 14 ($P2_1/c$ or $P2_1/n$ setting). Forms are numbered as in the CSD.

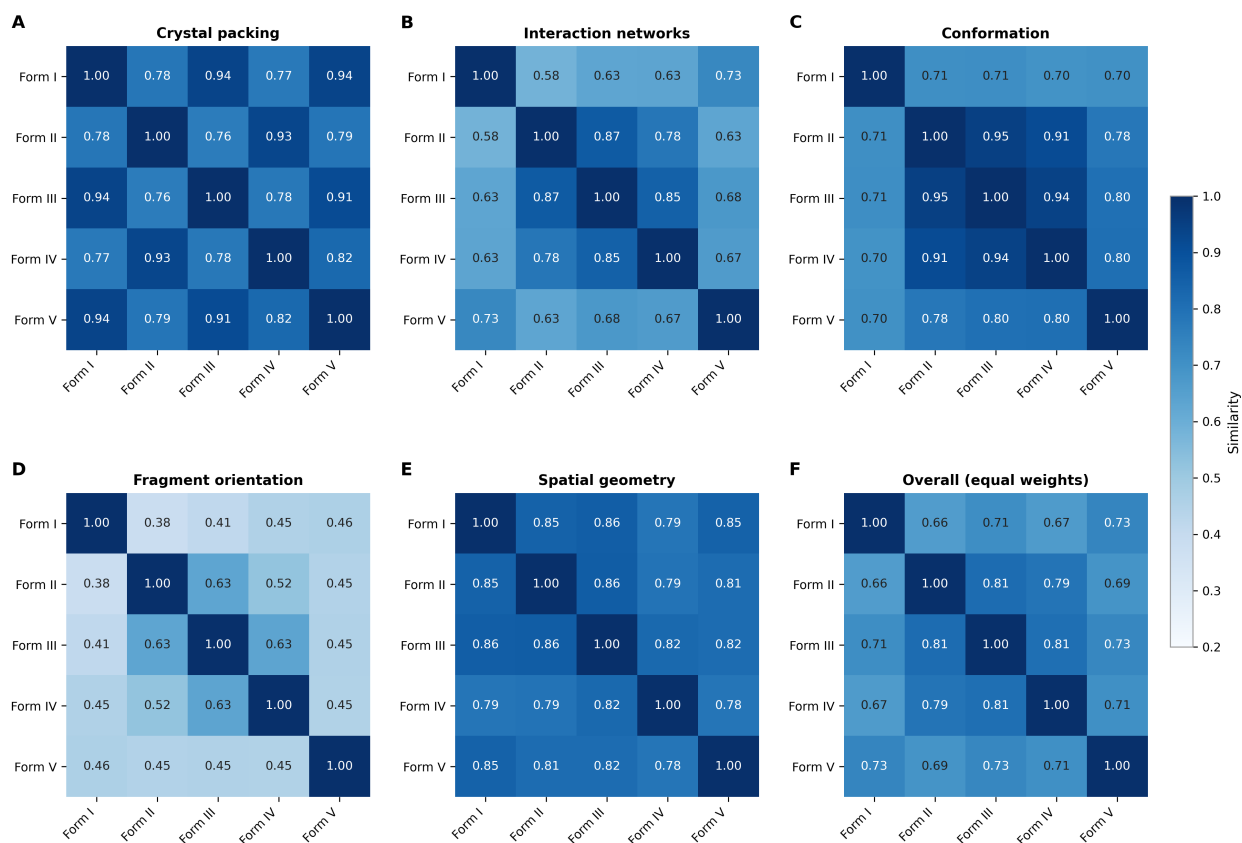
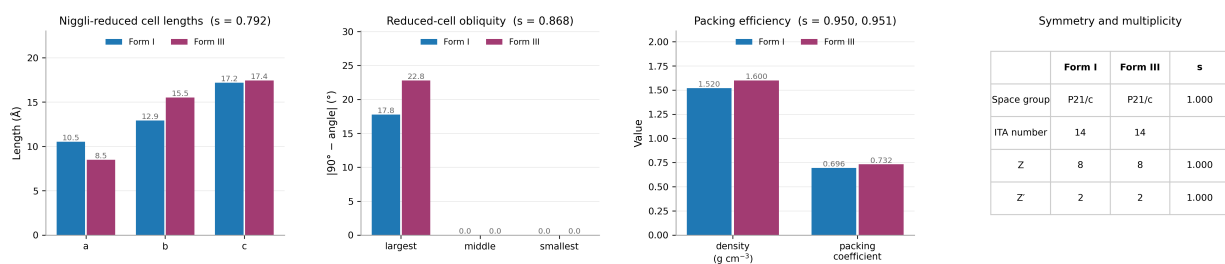


Figure 2: **Sulfathiazole.** Similarity matrices comparing all pairwise combinations of DFT-optimized sulfathiazole polymorphs (Forms I–V) across five independent structural categories: (A) Crystal Packing, (B) Intermolecular Interaction Networks, (C) Molecular Conformation, (D) Fragment Orientation, and (E) Spatial Geometry, with (F) Overall similarity (equal category weights, Eq. 10). Colour intensity represents the similarity score, from light (low) to dark blue (high); the numerical value is printed in each cell. Each category employs distribution-based metrics using Wasserstein distance calculations.

Highest similarity: Form I vs Form III (crystal packing similarity = 0.937)



Lowest similarity: Form II vs Form III (crystal packing similarity = 0.765)

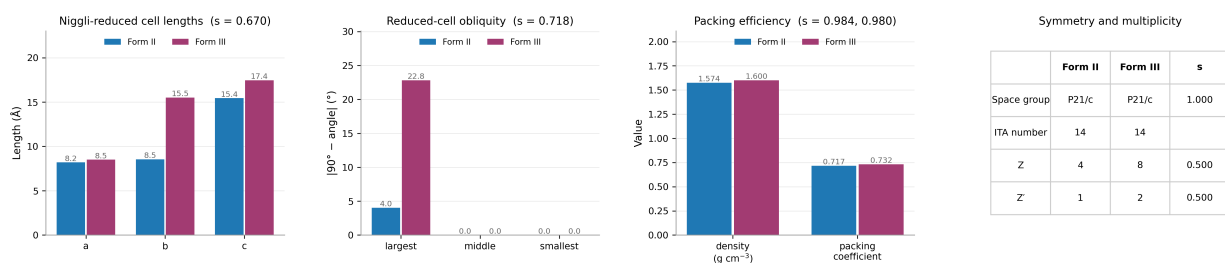
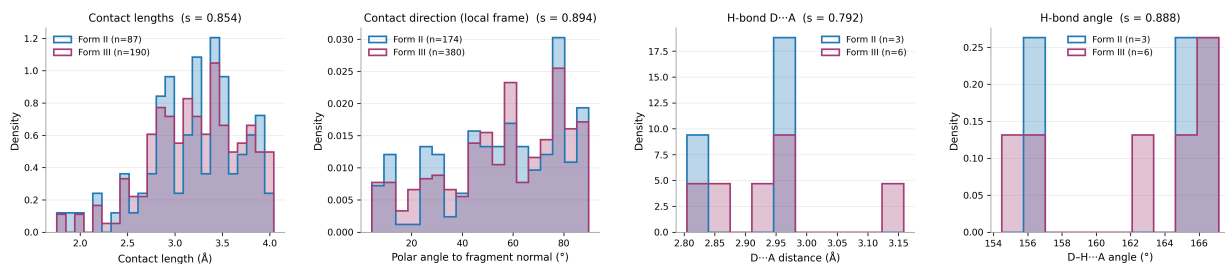


Figure 3: **Sulfathiazole: crystal packing** for the pairs with the highest (top, Form I vs Form III, similarity 0.937) and lowest (bottom, Form III vs Form II, 0.765) crystal-packing similarity. From left: Niggli-reduced cell lengths; obliquity of the reduced cell (sorted deviations of the cell angles from 90°); packing efficiency (density and packing coefficient); space group, ITA number, Z and Z' with the corresponding sub-metric similarities. The sub-metric similarity is given in each panel title.

Highest similarity: Form II vs Form III (interaction networks similarity = 0.866)



Lowest similarity: Form I vs Form II (interaction networks similarity = 0.584)

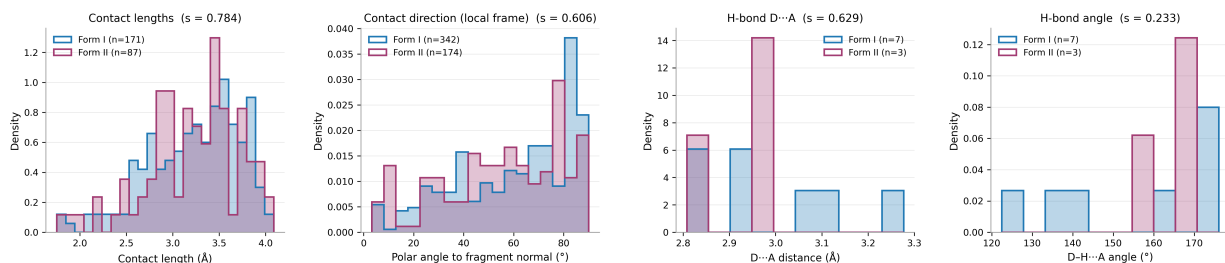
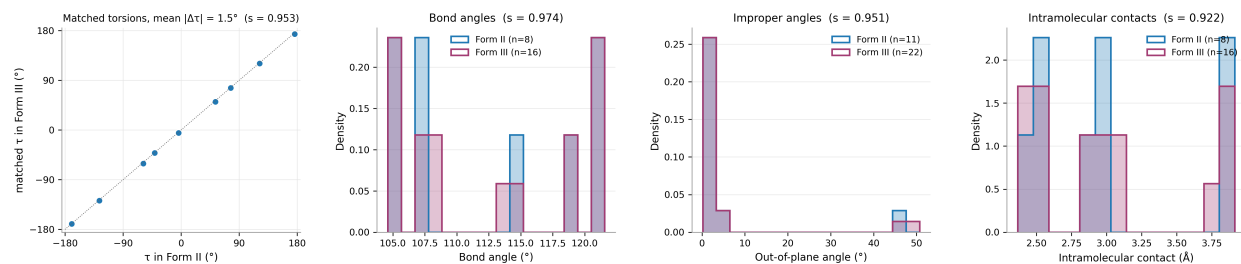


Figure 4: **Sulfathiazole: intermolecular interaction networks** for the pairs with the highest (top, Form III vs Form II, similarity 0.866) and lowest (bottom, Form I vs Form II, 0.584) interaction-network similarity. From left: distributions of contact lengths; polar angle of the contact direction to the plane normal of the fragment making the contact; H-bond D...A distances; D-H...A angles. n is the number of contacts or H-bonds in each distribution; the sub-metric similarity is given in each panel title.

Highest similarity: Form II vs Form III (conformation similarity = 0.950)



Lowest similarity: Form I vs Form IV (conformation similarity = 0.695)

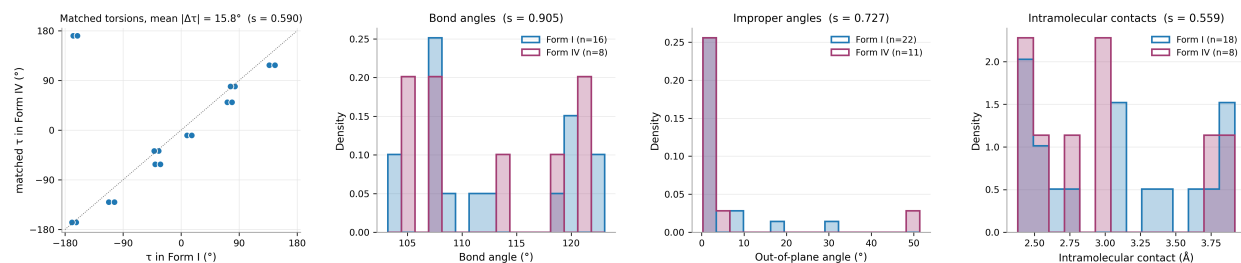
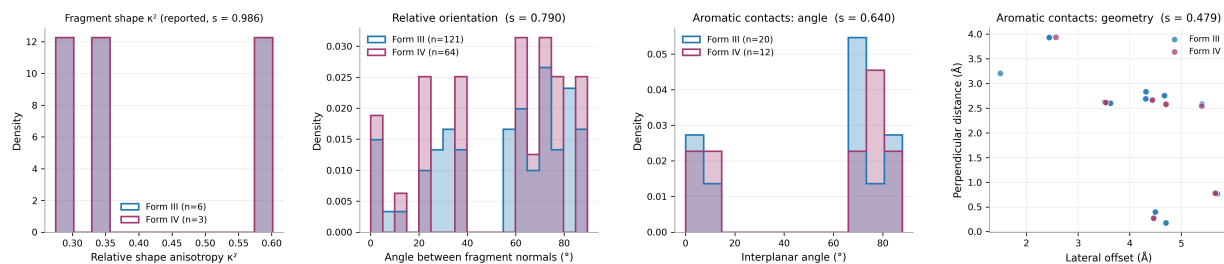


Figure 5: **Sulfathiazole: molecular conformation** for the pairs with the highest (top, Form III vs Form II, similarity 0.950) and lowest (bottom, Form I vs Form IV, 0.695) conformational similarity. From left: matched torsions about rotatable bonds, each point being one torsion in the first form plotted against its partner in the second (points on the dotted diagonal are identical); bond angles at rotatable-bond vertices; improper (out-of-plane) angles at three-coordinate atoms; intramolecular contact lengths. The four sub-metrics contribute equally to the category score.

Highest similarity: Form III vs Form IV (fragment orientation similarity = 0.634)



Lowest similarity: Form I vs Form II (fragment orientation similarity = 0.384)

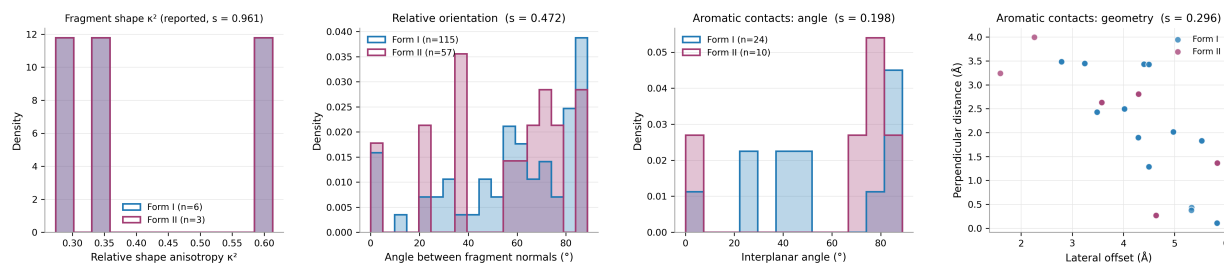
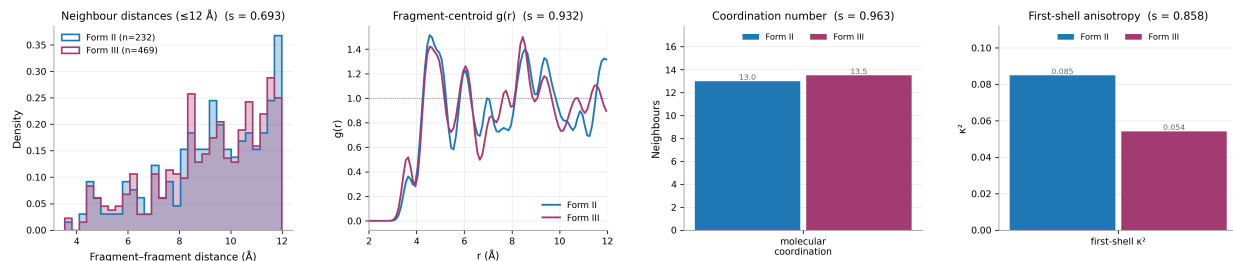


Figure 6: **Sulfathiazole: fragment orientation** for the pairs with the highest (top, Form III vs Form IV, similarity 0.634) and lowest (bottom, Form I vs Form II, 0.384) fragment-orientation similarity. From left: fragment shape anisotropy κ^2 (reported, not scored); angle between the plane normals of neighbouring fragments of different molecules ($\leq 8^\circ$); interplanar angles of aromatic contacts ($\leq 6^\circ$); perpendicular distance versus lateral offset of the aromatic contacts. Relative orientation and aromatic contacts contribute equally to the category score.

Highest similarity: Form II vs Form III (spatial geometry similarity = 0.861)



Lowest similarity: Form IV vs Form V (spatial geometry similarity = 0.784)

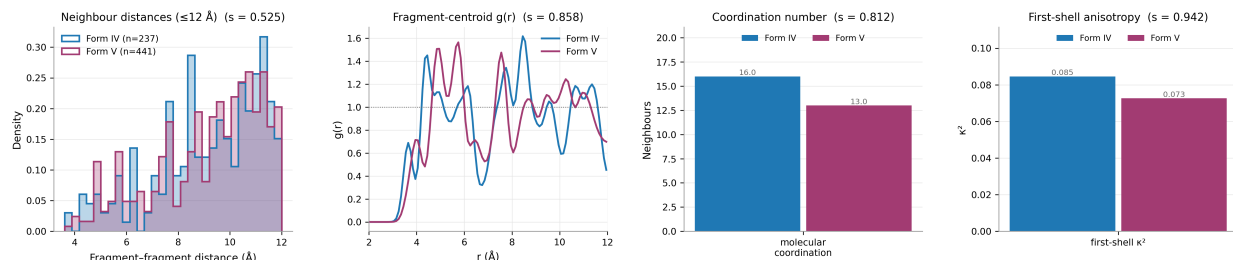


Figure 7: **Sulfathiazole: spatial geometry** for the pairs with the highest (top, Form III vs Form II, similarity 0.861) and lowest (bottom, Form V vs Form IV, 0.784) spatial-geometry similarity. From left: distances from each fragment to the centroids of all fragments of other molecules within 12 Å; intermolecular fragment-centroid radial distribution function $g(r)$ (dotted line, $g = 1$); molecular coordination number (contacting neighbour molecules); first-shell anisotropy κ_s^2 . The four sub-metrics contribute equally to the category score; Wasserstein distances are converted with $s = \exp(-W/\sigma)$ (Eq. 8).

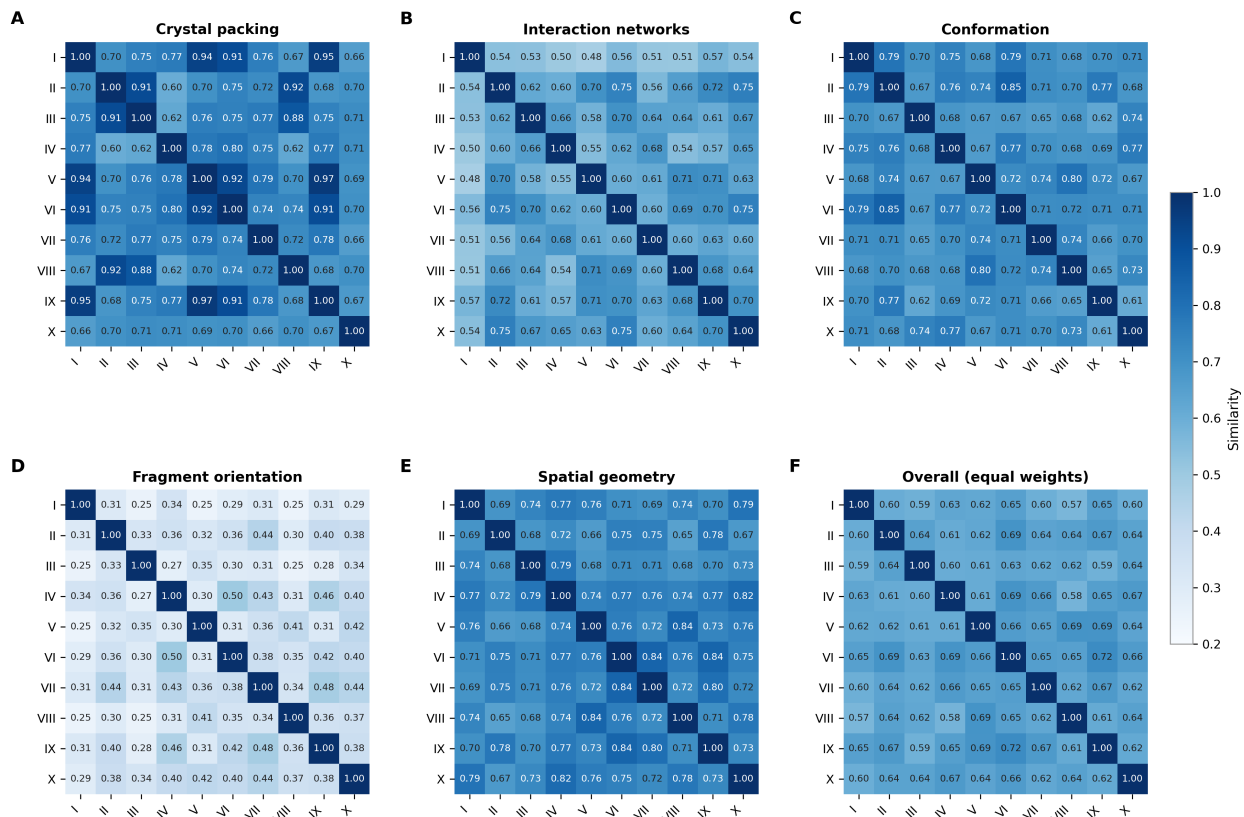
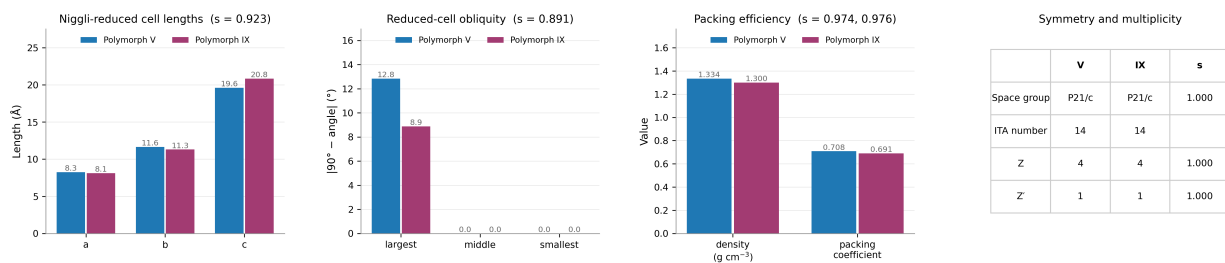


Figure 8: **Galunisertib**. Similarity matrices comparing all pairwise combinations of experimental galunisertib polymorphs (Polymorphs I–X) across five independent structural categories: (A) Crystal Packing, (B) Intermolecular Interaction Networks, (C) Molecular Conformation, (D) Fragment Orientation, and (E) Spatial Geometry, with (F) Overall similarity (equal category weights). Colour intensity represents the similarity score, from light (low) to dark blue (high); the numerical value is printed in each cell.

Highest similarity: Polymorph V vs Polymorph IX (crystal packing similarity = 0.966)



Lowest similarity: Polymorph II vs Polymorph IV (crystal packing similarity = 0.603)

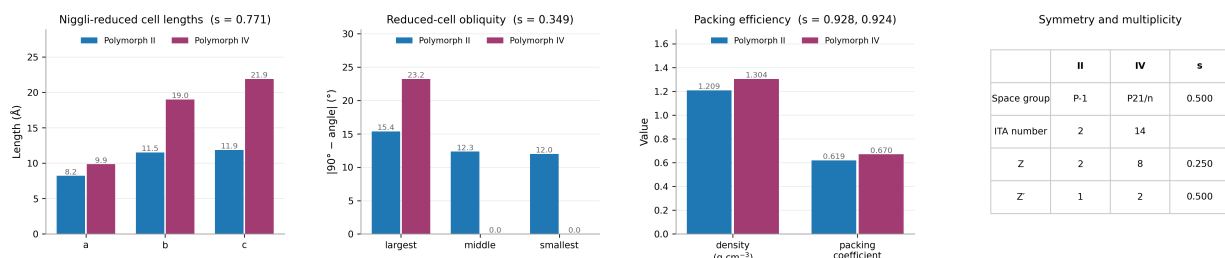
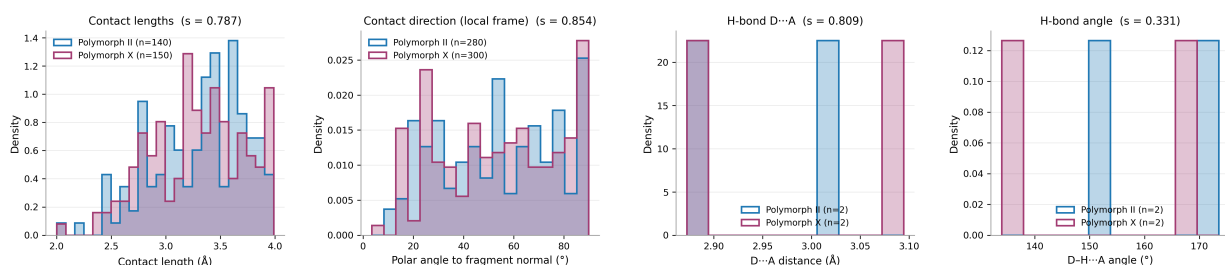


Figure 9: **Galunisertib: crystal packing** for the pairs with the highest (top, Polymorph V vs Polymorph IX, similarity 0.966) and lowest (bottom, Polymorph II vs Polymorph IV, 0.603) crystal-packing similarity. Panels as in Figure 3.

Highest similarity: Polymorph II vs Polymorph X (interaction networks similarity = 0.755)



Lowest similarity: Polymorph I vs Polymorph V (interaction networks similarity = 0.476)

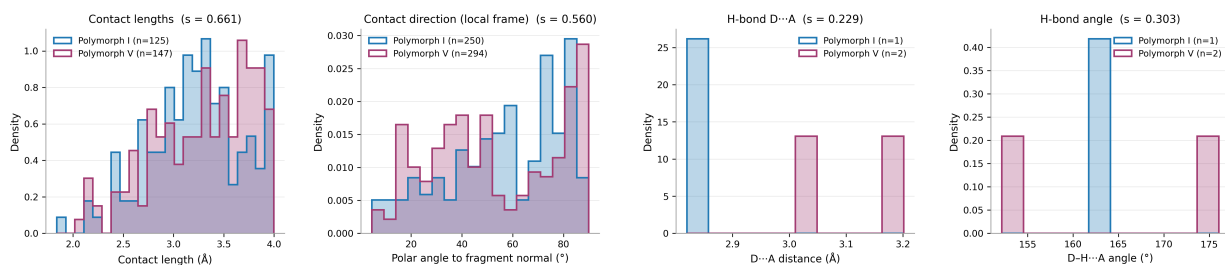
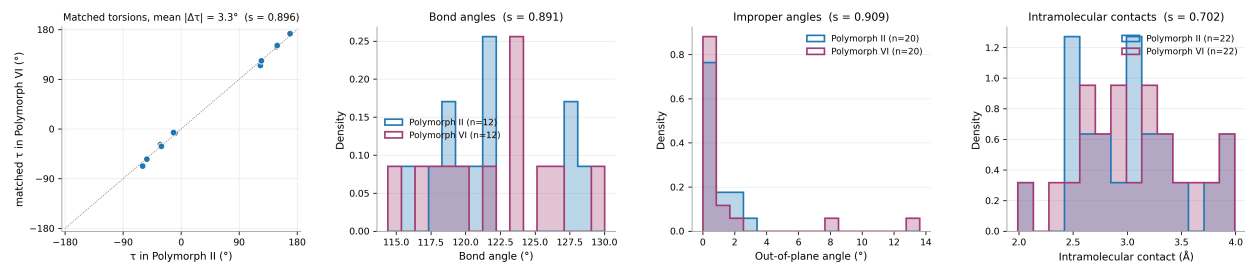


Figure 10: **Galunisertib: intermolecular interaction networks** for the pairs with the highest (top, Polymorph II vs Polymorph X, similarity 0.755) and lowest (bottom, Polymorph I vs Polymorph V, 0.476) interaction-network similarity. Panels as in Figure 4.

Highest similarity: Polymorph II vs Polymorph VI (conformation similarity = 0.850)



Lowest similarity: Polymorph IX vs Polymorph X (conformation similarity = 0.611)

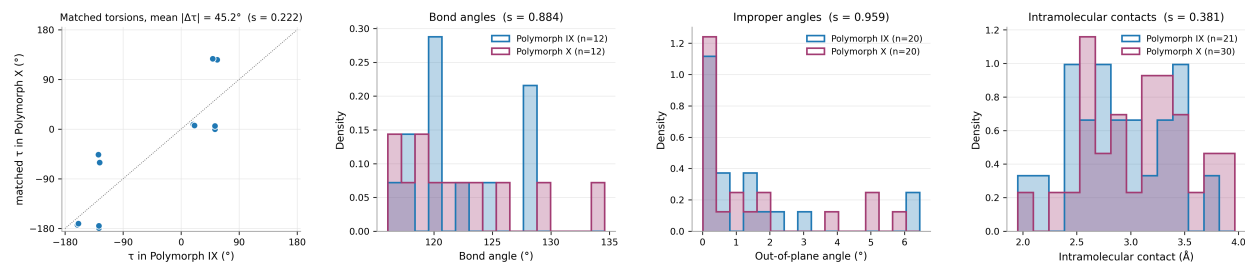
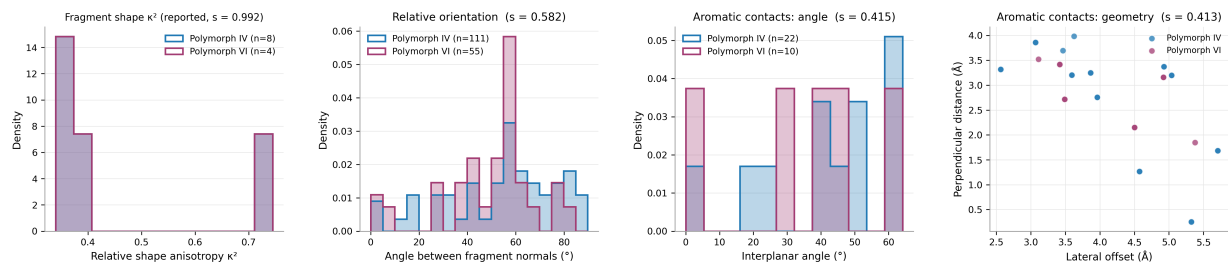


Figure 11: **Galunisertib: molecular conformation** for the pairs with the highest (top, Polymorph II vs Polymorph VI, similarity 0.850) and lowest (bottom, Polymorph IX vs Polymorph X, 0.611) conformational similarity. Panels as in Figure 5.

Highest similarity: Polymorph IV vs Polymorph VI (fragment orientation similarity = 0.497)



Lowest similarity: Polymorph I vs Polymorph V (fragment orientation similarity = 0.249)

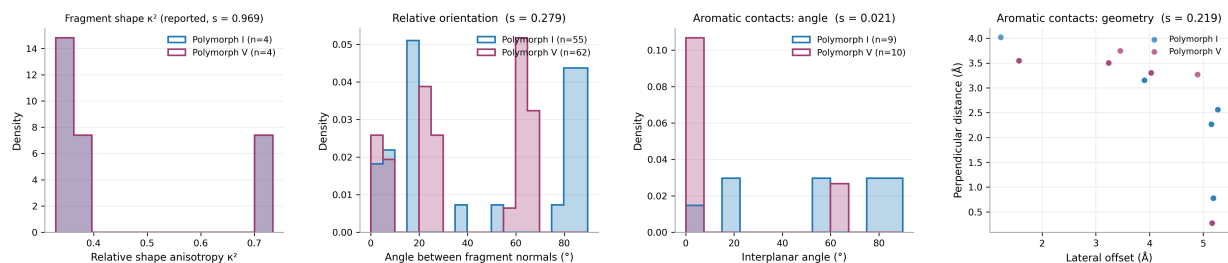
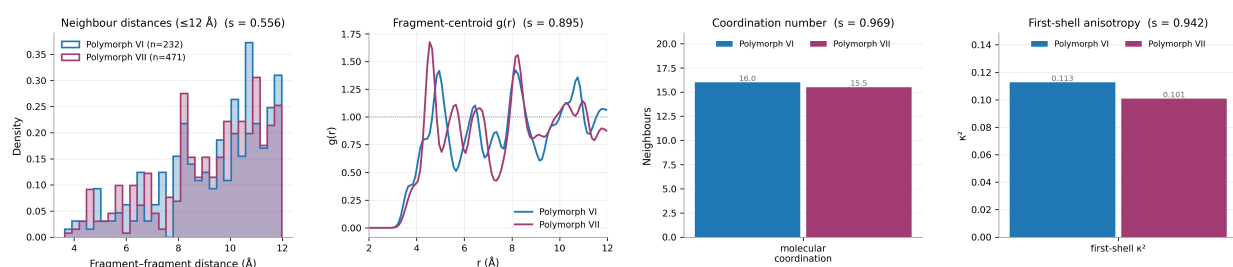


Figure 12: **Galunisertib: fragment orientation** for the pairs with the highest (top, Polymorph IV vs Polymorph VI, similarity 0.497) and lowest (bottom, Polymorph I vs Polymorph V, 0.249) fragment-orientation similarity. Panels as in Figure 6.

Highest similarity: Polymorph VI vs Polymorph VII (spatial geometry similarity = 0.841)



Lowest similarity: Polymorph II vs Polymorph VIII (spatial geometry similarity = 0.646)

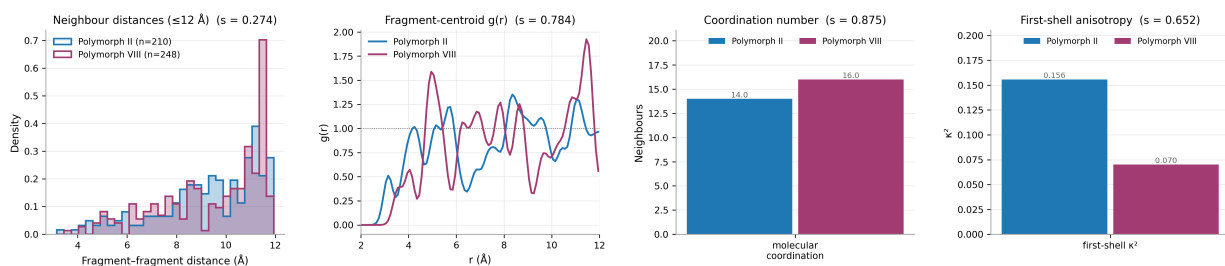


Figure 13: **Galunisertib: spatial geometry** for the pairs with the highest (top, Polymorph VI vs Polymorph VII, similarity 0.841) and lowest (bottom, Polymorph II vs Polymorph VIII, 0.646) spatial-geometry similarity. Panels as in Figure 7.

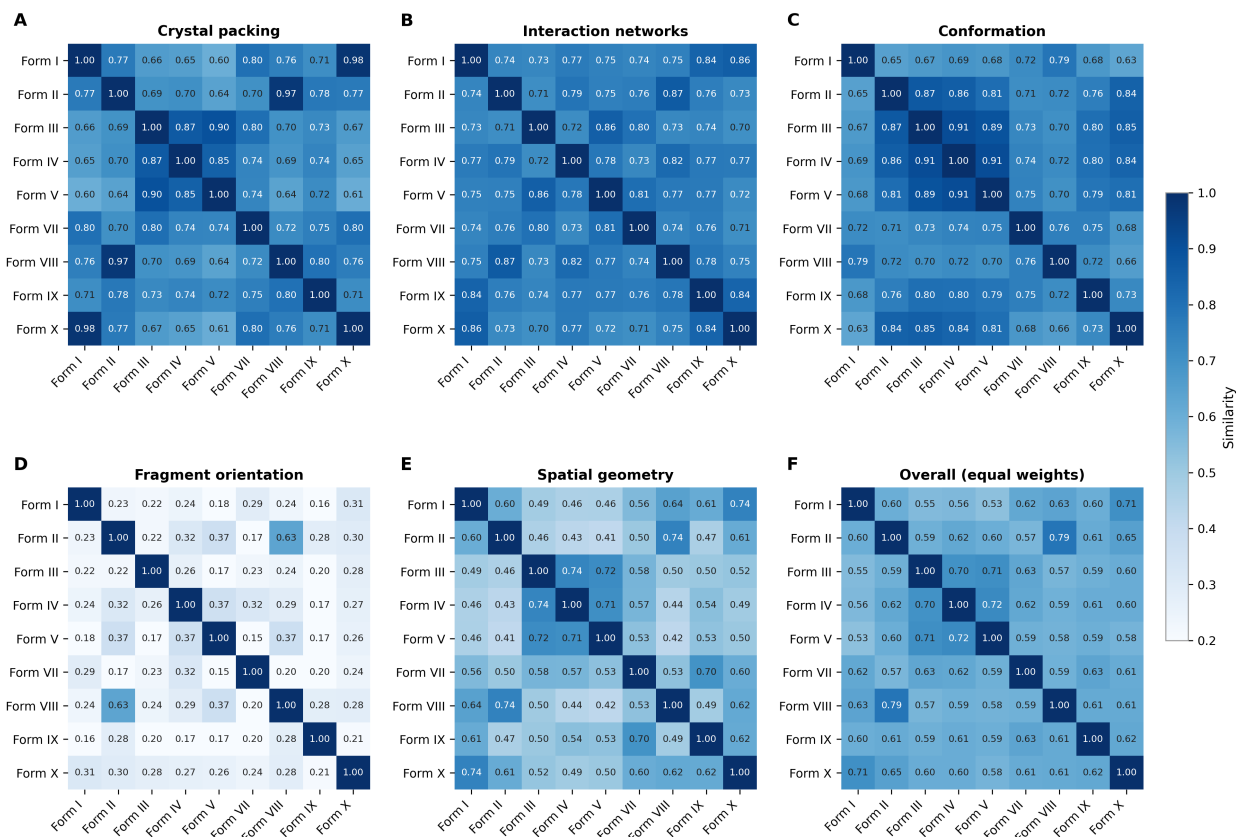
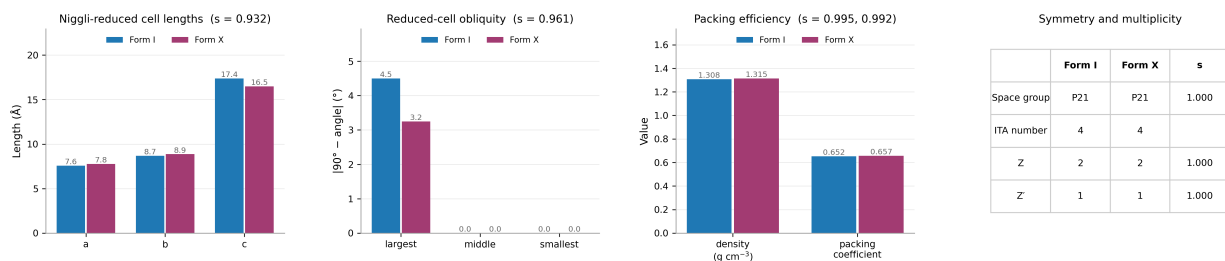


Figure 14: **Aripiprazole**. Similarity matrices of the nine aripiprazole forms (one representative depositions per form, Table 3) for the five categories (A–E) and the overall similarity (F). Forms I and X share their crystal packing but not their conformation; Forms III, IV and V share their conformation but not their reduced cells; Forms II and VIII are the most similar pair overall.

Highest similarity: Form I vs Form X (crystal packing similarity = 0.983)



Lowest similarity: Form I vs Form V (crystal packing similarity = 0.599)

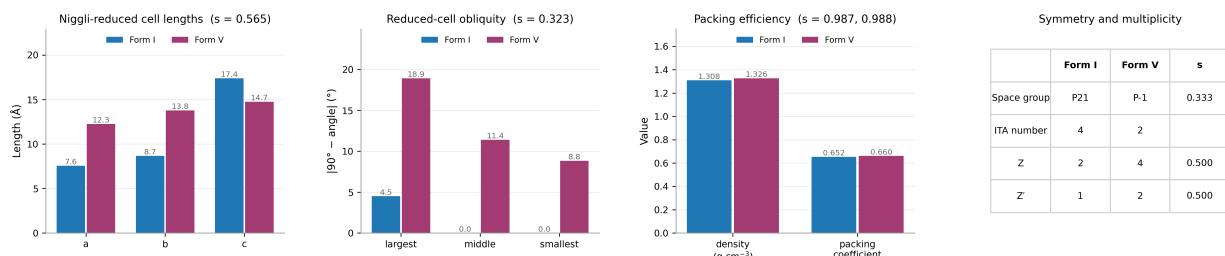
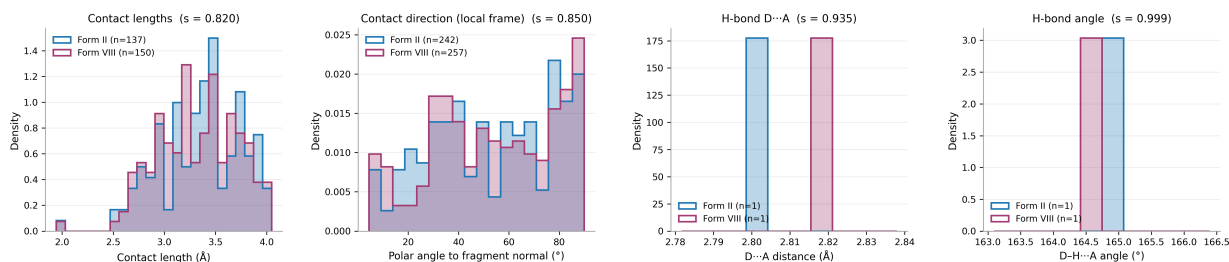


Figure 15: **Aripiprazole: crystal packing** for the pairs with the highest (top, Form I vs Form X, similarity 0.983) and lowest (bottom, Form I vs Form V, 0.599) crystal-packing similarity. Panels as in Figure 3.

Highest similarity: Form II vs Form VIII (interaction networks similarity = 0.868)



Lowest similarity: Form III vs Form X (interaction networks similarity = 0.705)

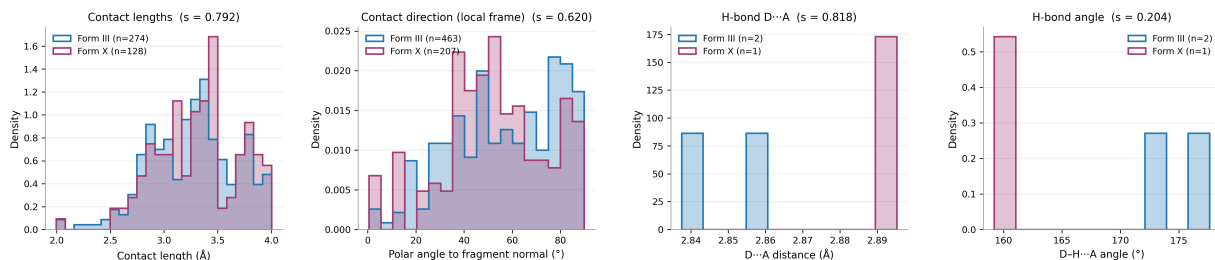
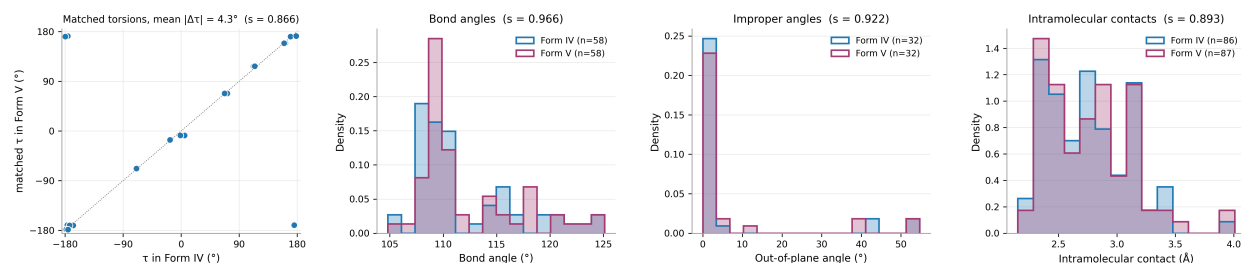


Figure 16: **Aripiprazole: intermolecular interaction networks** for the pairs with the highest (top, Form II vs Form VIII, similarity 0.868) and lowest (bottom, Form III vs Form X, 0.705) interaction-network similarity. Panels as in Figure 4.

Highest similarity: Form IV vs Form V (conformation similarity = 0.912)



Lowest similarity: Form I vs Form X (conformation similarity = 0.629)

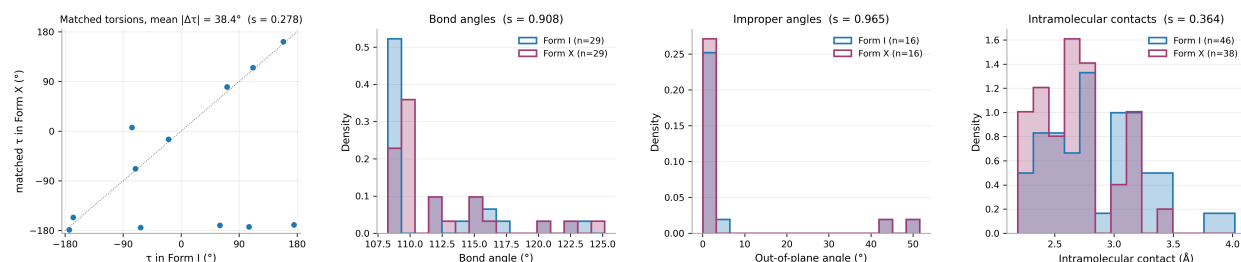
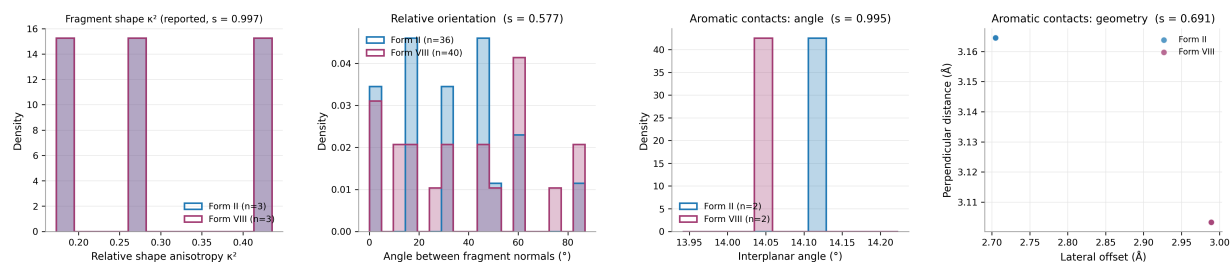


Figure 17: **Aripiprazole: molecular conformation** for the pairs with the highest (top, Form IV vs Form V, similarity 0.912) and lowest (bottom, Form I vs Form X, 0.629) conformational similarity. Panels as in Figure 5; for the $Z' = 2$ Forms IV and V the torsions of both independent molecules are shown.

Highest similarity: Form II vs Form VIII (fragment orientation similarity = 0.634)



Lowest similarity: Form V vs Form VII (fragment orientation similarity = 0.147)

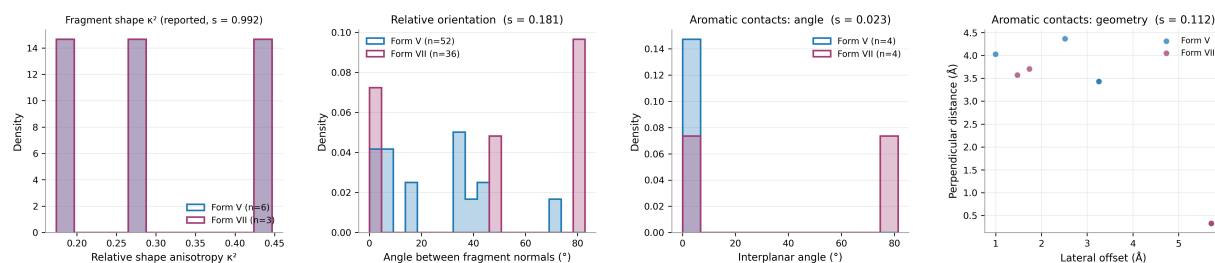
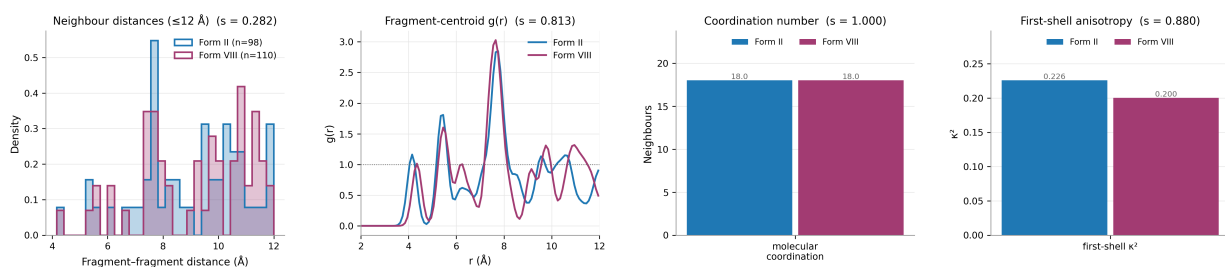


Figure 18: **Aripiprazole: fragment orientation** for the pairs with the highest (top, Form II vs Form VIII, similarity 0.634) and lowest (bottom, Form V vs Form VII, 0.147) fragment-orientation similarity. Panels as in Figure 6; the fragment-level descriptors use the three ring systems (Methods).

Highest similarity: Form II vs Form VIII (spatial geometry similarity = 0.744)



Lowest similarity: Form II vs Form V (spatial geometry similarity = 0.413)

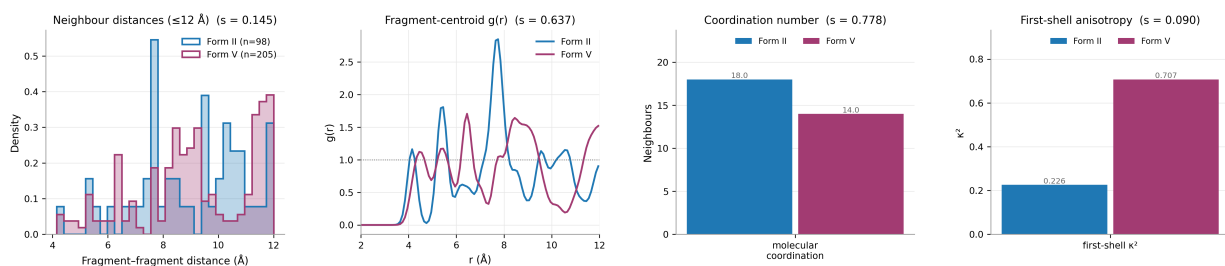


Figure 19: **Aripiprazole: spatial geometry** for the pairs with the highest (top, Form II vs Form VIII, similarity 0.744) and lowest (bottom, Form II vs Form V, 0.413) spatial-geometry similarity. Panels as in Figure 7.

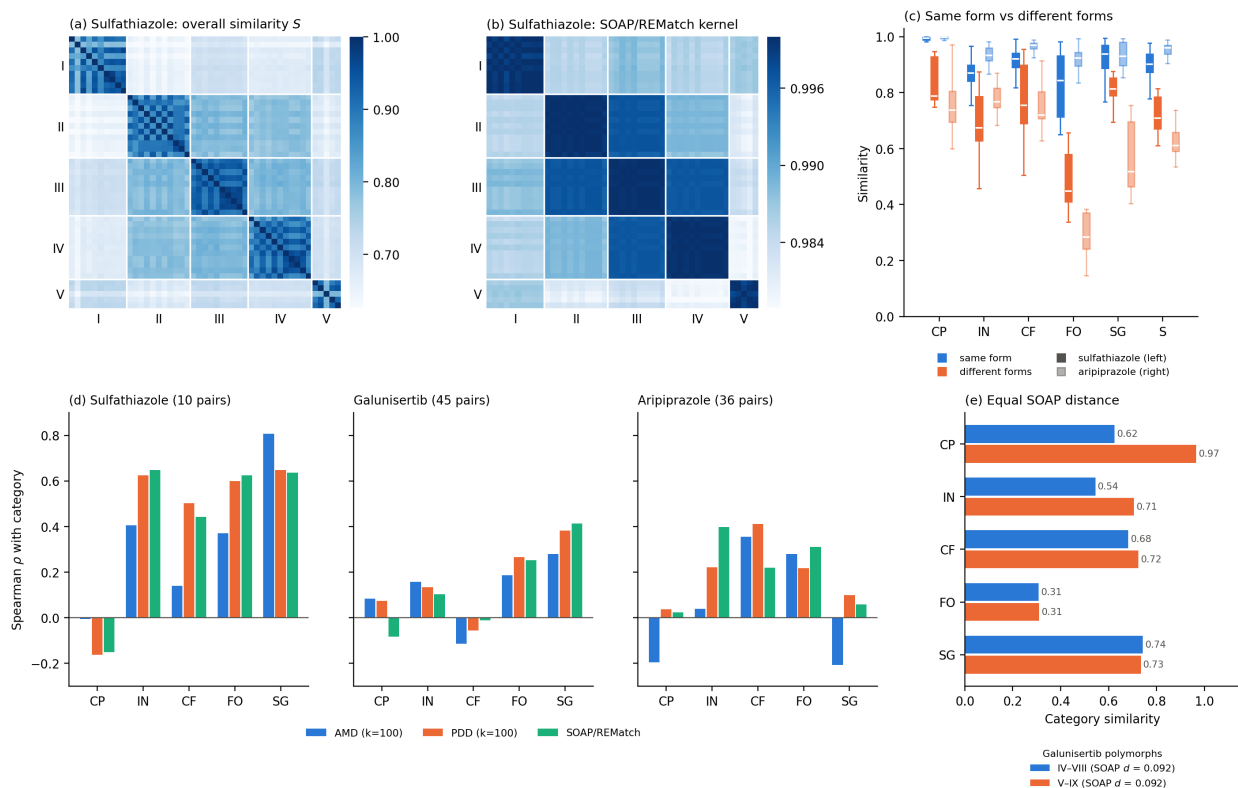


Figure 20: **Benchmark against invariant descriptors.** (a) Overall similarity S and (b) SOAP/REMatch kernel for the 47 sulfathiazole structures, ordered by form (white lines separate the forms). (c) Category (CP, crystal packing; IN, interaction networks; CF, conformation; FO, fragment orientation; SG, spatial geometry) and overall similarities of pairs of depositions of the same form (blue) and of different forms (orange), for sulfathiazole (47 structures, left of each pair of boxes) and aripiprazole (21 structures, right, lighter). (d) Spearman correlation between the similarity implied by the AMD, PDD and SOAP distances (negative distance) and each category similarity, for the pairs of distinct forms of the three systems. (e) Two galunisertib pairs with the same SOAP distance but different similarity vectors.

References

- [1] Bernstein, J. *Polymorphism in Molecular Crystals*. Oxford University Press, 2020.
- [2] Price, S.L. The computational prediction of pharmaceutical crystal structures and polymorphism. *Advanced Drug Delivery Reviews*, 56(3):301–319, 2004.
- [3] Shi, Q., Chen, H., Wang, Y., Xu, J., Liu, Z. and Zhang, C. Recent advances in drug polymorphs: Aspects of pharmaceutical properties and selective crystallization. *International Journal of Pharmaceutics*, 611:121320, 2022.
- [4] Chen, A., Peng, Y., Chen, Z., Lu, Y., Guo, M. and Cai, T. Crystal engineering for poorly water-soluble drugs: From design to applications. *Acta Pharmaceutica Sinica B*, 16(4):1848–1882, 2026.
- [5] Hilfiker, R., editor. *Polymorphism in the Pharmaceutical Industry*. Wiley-VCH, Weinheim, 2006.
- [6] Groom, C.R., Bruno, I.J., Lightfoot, M.P. and Ward, S.C. The Cambridge Structural Database. *Acta Crystallographica Section B*, 72(2):171–179, 2016.
- [7] Chisholm, J.A. and Motherwell, S. COMPACT: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography*, 38(1):228–231, 2005.
- [8] Price, S.L. Computed crystal energy landscapes for understanding and predicting organic crystal structures and polymorphism. *Accounts of Chemical Research*, 51(1):14–21, 2018.
- [9] Galanakis, N. and Tuckerman, M.E. Rapid prediction of molecular crystal structures using simple topological and physical descriptors. *Nature Communications*, 15:9757, 2024.
- [10] Hunnisett, L.M. *et al.* The seventh blind test of crystal structure prediction: structure generation methods. *Acta Crystallographica Section B*, 80:517–547, 2024.
- [11] Hunnisett, L.M. *et al.* The seventh blind test of crystal structure prediction: structure ranking methods. *Acta Crystallographica Section B*, 80:548–574, 2024.
- [12] Hughes, D.S. *et al.* The extensive solid-form landscape of sulfathiazole: geometrical similarity and interaction energies. *CrystEngComm*, 24:609, 2022.
- [13] Hughes, D.S. *et al.* The extensive solid-form landscape of sulfathiazole: hydrogen bond topology and node shape. *CrystEngComm*, 24:6578, 2022.
- [14] Nessler, A.J., Okada, O., Hermon, M.J., Nagata, H. and Schnieders, M.J. Progressive alignment of crystals: reproducible and efficient assessment of crystal structure similarity. *Journal of Applied Crystallography*, 55:1528–1537, 2022.
- [15] Rohlíček, J., Skořepová, E., Babor, M. and Čejka, J. CrystalCMP: an easy-to-use tool for fast comparison of molecular packing. *Journal of Applied Crystallography*, 49:2172–2183, 2016.
- [16] Gelbrich, T. and Hursthouse, M.B. A versatile procedure for the identification, description and quantification of structural similarity in molecular crystals. *CrystEngComm*, 7:324–336, 2005.

- [17] de Gelder, R., Wehrens, R. and Hageman, J.A. A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22:273–289, 2001.
- [18] Spackman, M.A. and McKinnon, J.J. Fingerprinting intermolecular interactions in molecular crystals. *CrystEngComm*, 4:378–392, 2002.
- [19] Bartók, A.P., Kondor, R. and Csányi, G. On representing chemical environments. *Physical Review B*, 87:184115, 2013.
- [20] De, S., Bartók, A.P., Csányi, G. and Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18:13754–13769, 2016.
- [21] Widdowson, D., Mosca, M.M., Pulido, A., Cooper, A.I. and Kurlin, V. Average minimum distances of periodic point sets – foundational invariants for mapping periodic crystals. *MATCH Communications in Mathematical and in Computer Chemistry*, 87:529–559, 2022.
- [22] Widdowson, D. and Kurlin, V. Resolving the data ambiguity for periodic crystals. *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 35:24625–24638, 2022.
- [23] Zhang, R., Seth, S. and Cumby, J. Grouped representation of interatomic distances as a similarity measure for crystal structures. *Digital Discovery*, 2:81–90, 2023.
- [24] Dey, S., Miklaucic, N., Omeo, S.S., Dong, R., Wei, L., Li, Q., Fu, N. and Hu, J. Data-driven topological analysis of polymorphic crystal structures. *The Journal of Physical Chemistry C*, 2026. doi:10.1021/acs.jpcc.5c06442.
- [25] Omeo, S.S. *et al.* Polymorphism crystal structure prediction with adaptive space group diversity control. *Advanced Science*, e10792, 2025. doi:10.1002/advs.202510792.
- [26] Cruz-Cabeza, A.J., Reutzel-Edens, S.M. and Bernstein, J. Facts and fictions about polymorphism. *Chemical Society Reviews*, 44:8619–8635, 2015.
- [27] Verma, A.R. and Krishna, P. *Polymorphism and Polytypism in Crystals*. Wiley, New York, 1966.
- [28] Pepinsky, R. Crystal Engineering – A new concept in crystallography. *Physical Review*, 100:971–972, 1955.
- [29] Burckhardt, J.J. Zur Geschichte der Entdeckung der 230 Raumgruppen. *Archive for History of Exact Sciences*, 4:235–246, 1967.
- [30] Wang, A.Y.-T., Murdock, R.J., Kauwe, S.K., Oliynyk, A.O., Gurlo, A., Brgoch, J., Persson, K.A. and Sparks, T.D. Machine learning for materials scientists: an introductory guide toward best practices. *Chemistry of Materials*, 32:4954–4965, 2020.
- [31] Frade, A.P., McCabe, P. and Cooper, R.I. Increasing the performance, trustworthiness and practical value of machine learning models: a case study predicting hydrogen bond network dimensionalities from molecular diagrams. *CrystEngComm*, 22:7186–7192, 2020.
- [32] Cui, P. *et al.* Mining predicted crystal structure landscapes with high throughput crystallisation: old molecules, new insights. *Chemical Science*, 10:9988, 2019.

- [33] Rosso, V. *et al.* High-throughput crystallization screening technique with transmission PXRD analysis. *Organic Process Research & Development*, 27:1437–1444, 2023.
- [34] Arjovsky, M., Chintala, S. and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.
- [35] Aroyo, M.I. (Ed.) *International Tables for Crystallography, Volume A: Space-Group Symmetry*, 6th ed. Wiley, Chichester, 2016.
- [36] Bondi, A. van der Waals volumes and radii. *Journal of Physical Chemistry*, 68(3):441–451, 1964.
- [37] Gelbrich, T., Hughes, D.S., Hursthouse, M.B. and Threlfall, T.L. Packing similarity in polymorphs of sulfathiazole. *CrystEngComm*, 10:1328–1334, 2008.
- [38] Clark, S.J., Segall, M.D., Pickard, C.J., Hasnip, P.J., Probert, M.J., Refson, K. and Payne, M.C. First principles methods using CASTEP. *Zeitschrift für Kristallographie*, 220:567–570, 2005.
- [39] Perdew, J.P., Burke, K. and Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters*, 77:3865–3868, 1996.
- [40] Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry*, 27:1787–1799, 2006.
- [41] van de Streek, J. and Neumann, M.A. Validation of experimental molecular crystal structures with dispersion-corrected density functional theory calculations. *Acta Crystallographica Section B*, 66:544–558, 2010.
- [42] Bhardwaj, R.M., *et al.* A Prolific Solvate Former, Galunisertib, under the Pressure of Crystal Structure Prediction, Produces Ten Diverse Polymorphs *J. Am. Chem. Soc.* 141(35), 13887–13897, 2019
- [43] Tessler, L. and Goldberg, I. Crystal structures of aripiprazole, a new anti-psychotic drug, and of its inclusion compounds with methanol, ethanol and water. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, 55(3):255–261, 2006.
- [44] Braun, D.E., Gelbrich, T., Kahlenberg, V., Tessadri, R., Wieser, J. and Griesser, U.J. Conformational polymorphism in aripiprazole: preparation, stability and structure of five modifications. *Journal of Pharmaceutical Sciences*, 98(6):2010–2026, 2009.
- [45] Nanubolu, J.B., Sridhar, B., Babu, V.S.P., Jagadeesh, B. and Ravikumar, K. Sixth polymorph of aripiprazole – an antipsychotic drug. *CrystEngComm*, 14(14):4677–4685, 2012.
- [46] Delaney, S.P., Pan, D., Yin, S.X., Smith, T.M. and Korter, T.M. Evaluating the roles of conformational strain and cohesive binding in crystalline polymorphs of aripiprazole. *Crystal Growth & Design*, 13(7):2943–2952, 2013.
- [47] Delaney, S.P., Smith, T.M., Pan, D., Yin, S.X. and Korter, T.M. Low-temperature phase transition in crystalline aripiprazole leads to an eighth polymorph. *Crystal Growth & Design*, 14(10):5004–5010, 2014.

- [48] Zeidan, T.A., Trotta, J.T., Tilak, P.A., Oliveira, M.A., Chiarella, R.A., Foxman, B.M., Almarsson, Ö. and Hickey, M.B. An unprecedented case of dodecamorphism: the twelfth polymorph of aripiprazole formed by seeding with its active metabolite. *CrystEngComm*, 18(9):1486–1488, 2016.
- [49] Tyler, A.R., Ragbirsingh, R., McMonagle, C.J., Waddell, P.G., Heaps, S.E., Steed, J.W., Thaw, P., Hall, M.J. and Probert, M.R. Encapsulated nanodroplet crystallization of organic-soluble small molecules. *Chem*, 6(7):1755–1765, 2020.
- [50] Křivý, I. and Gruber, B. A unified algorithm for determining the reduced (Niggli) cell. *Acta Crystallographica Section A*, 32:297–298, 1976.
- [51] Jeffrey, G.A. *An Introduction to Hydrogen Bonding*. Oxford University Press, New York, 1997.
- [52] Himanen, L., Jäger, M.O.J., Morooka, E.V., Federici Canova, F., Ranawat, Y.S., Gao, D.Z., Rinke, P. and Foster, A.S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.
- [53] Togo, A., Shinohara, K. and Tanaka, I. Spglib: a software library for crystal symmetry search. *Science and Technology of Advanced Materials: Methods*, 4:2384822, 2024.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [si.pdf](#)