

Supplementary Information

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

Nikolaos Galanakis, David S. Hughes, Mark E. Tuckerman

All values were computed with the setting-invariant implementation described in the main text, with the scale parameters $\sigma = 0.20$ Å (distances), 10° (angles), 30° (matched torsions) and 0.20 (dimensionless anisotropies), unweighted means within categories (Eq. 9) and equal category weights $w_k = 0.20$ (Eq. 10). Sulfathiazole: Forms I–V, numbered as in the CSD (DFT-D minimized; the minimized structures are provided as Supplementary Data); Form I from SUTHAZ16 (file 1p_DFT), II from SUTHAZ18 (4p_DFT), III from SUTHAZ17 (3p_DFT), IV from SUTHAZ19 (5p_DFT) and V from SUTHAZ05 (2p_DFT). Aripiprazole: Section S1. Galunisertib: Polymorphs I–X with CSD refcodes DORDUM09 (I), DORDUM (II), DORDUM01 (III), DORDUM02 (IV), DORDUM03 (V), DORDUM04 (VI), DORDUM10 (VII), DORDUM05 (VIII), DORDUM06 (IX) and DORDUM07 (X). DORDUM08 is a redetermination of the same polymorph as DORDUM07 (COMPACT: 20/20 molecules matched, RMSD 0.083 Å, PXR similarity 0.998) and was therefore not included.

S1. Datasets

Sulfathiazole: the five DFT-D structures of the main analysis and the 42 CSD determinations used in the benchmark (Table S1). Galunisertib: Polymorphs I–X (main text, Table 2) and DORDUM08, a second determination of Polymorph X. Aripiprazole: the 21 CSD determinations of nine forms (Table S2); forms are numbered as in the CSD, and no structure is available for Form VI. For aripiprazole, the fragment-level descriptors (fragment frames, relative orientations, aromatic contacts and fragment environments) use the three ring systems (2,3-dichlorophenyl, dihydroquinolinone, piperazine), because cutting the molecule at its rotatable bonds also produces five one- to three-atom fragments of the butoxy linker whose principal axes are not defined; all atoms still enter the contacts, H-bonds, torsions and coordination shells. Aromatic contacts use the dichlorophenyl and dihydroquinolinone fragments.

Table S1: The 47 sulfathiazole structures used in the benchmark. Forms are numbered as in the CSD. Each DFT-D structure was minimized from the CSD entry given in parentheses; COMPACK matches all 20 molecules of every entry of a form (all except SUTHAZ01, for which COMPACK results were not available) with the DFT-D structure of that form, and at most 16 molecules for any pair of structures of different forms.

Form	DFT-D structure	CSD determinations
I	1p_DFT (from SUTHAZ16)	SUTHAZ01, SUTHAZ07, SUTHAZ08, SUTHAZ16, SUTHAZ28, SUTHAZ29, SUTHAZ43, SUTHAZ46, SUTHAZ47
II	4p_DFT (from SUTHAZ18)	SUTHAZ, SUTHAZ03, SUTHAZ09, SUTHAZ10, SUTHAZ18, SUTHAZ24, SUTHAZ30, SUTHAZ31, SUTHAZ32, SUTHAZ48
III	3p_DFT (from SUTHAZ17)	SUTHAZ11, SUTHAZ12, SUTHAZ17, SUTHAZ25, SUTHAZ33, SUTHAZ34, SUTHAZ35, SUTHAZ44, SUTHAZ49
IV	5p_DFT (from SUTHAZ19)	SUTHAZ04, SUTHAZ13, SUTHAZ14, SUTHAZ19, SUTHAZ26, SUTHAZ36, SUTHAZ37, SUTHAZ38, SUTHAZ45, SUTHAZ50
V	2p_DFT (from SUTHAZ05)	SUTHAZ05, SUTHAZ06, SUTHAZ27, SUTHAZ51

Table S2: The 21 aripiprazole determinations in the CSD. The representative of each form (bold) is the determination with the highest mean overall similarity to the other determinations of the same form (ties broken by the lower refcode). MELFIT19 is disordered; its major component was used.

Form	Space group	Z'	CSD determinations (representative in bold)
I	$P2_1$	1	MELFIT01
II	$Pna2_1$	1	MELFIT02, MELFIT08, MELFIT15 , MELFIT16, MELFIT17
III	$P\bar{1}$	2	MELFIT , MELFIT03
IV	$P\bar{1}$	2	MELFIT04 , MELFIT20
V	$P\bar{1}$	2	MELFIT06
VII	$P\bar{1}$	1	MELFIT07
VIII	$Pna2_1$	1	MELFIT09, MELFIT10, MELFIT11 , MELFIT12, MELFIT13, MELFIT14, MELFIT18
IX	$P2_1/n$	1	MELFIT19
X	$P2_1$	1	MELFIT05

S2. Algorithmic details and worked example

This section describes every step from the input file to the similarity vector, and lists all intermediate quantities for one comparison (sulfathiazole Form II versus Form III; Table S3). The code that produces Table S3 is part of the code release.

S2.1 Input

Each structure is one record of an HDF5 file produced from the CIF. The record contains: the cell matrix, the space-group symbol of the deposited setting, Z , Z' , the density and the packing coefficient; the atoms of the asymmetric unit (element, label, fractional coordinates, mass); the bonds, with their lengths and a flag for rotatable bonds; the rigid fragment to which each atom belongs (atoms connected without crossing a rotatable bond; for sulfathiazole the aminophenyl group C_6H_5N , the sulfonyl group SO_2 and the thiazolyl-amide group $C_3H_3N_2S$; for galunisertib four fragments, C_9H_5N , C_6H_6N , $C_6H_6N_2$ and

the carboxamide CH_2NO ; for aripiprazole see Section S1); the torsions and bond angles; intramolecular contacts; and intermolecular contacts and H-bond candidates. Each intermolecular entry gives an atom of the asymmetric unit, the contacting atom (an image of an asymmetric-unit atom under a symmetry operation and a lattice translation), the coordinates of both and their distance. Contacts are atom pairs with $d \leq r_i + r_j + 0.6 \text{ \AA}$ (Bondi radii); every contact is stored from both ends and is counted once.

S2.2 Periodic crystal

1. The symmetry operations of the deposited setting are taken from the spglib database, matched by the full Hermann–Mauguin symbol, so that any setting (e.g. $P2_1/n$ or $P112_1/b$) is reproduced exactly.
2. Applying all operations to the asymmetric unit gives every molecule of the unit cell; each fragment image carries its molecule identity (operation, asymmetric-unit molecule, lattice translation), so that intramolecular and intermolecular fragment pairs can be distinguished.
3. For every fragment: centroid, centre of mass, inertia tensor and principal axes (long axis = smallest moment, plane normal = largest moment), and the best-plane normal of its atoms (singular value decomposition), which is used for aromatic contacts.
4. The neighbours of every asymmetric-unit fragment within a cutoff are found by adding lattice translations to the fragment images of the unit cell.

S2.3 Descriptors and sub-metrics of the five categories

Crystal packing (7 sub-metrics). (i) Packing coefficient and (ii) density, $s = 1 - |a - b| / \max(a, b)$. (iii) Lengths of the Niggli-reduced primitive cell, normalized by their mean, $d = \|\mathbf{l}_1/\bar{l}_1 - \mathbf{l}_2/\bar{l}_2\|$, $s = e^{-d}$. (iv) Obliquity: the three deviations $|90^\circ - \text{angle}|$ of the reduced cell, sorted, $s = \text{mean of } e^{-|\Delta|/10^\circ}$. (v) Symmetry: Jaccard index of the sets of operations of the two space-group types in their standard settings. (vi) Z and (vii) Z' , $s = \min / \max$.

Interaction networks (6 sub-metrics). (i) Lengths of the unique intermolecular contacts. (ii, iii) Direction of each contact, as the unit vector from the centroid of the fragment that contains the asymmetric-unit atom to the contacting atom, expressed in that fragment’s principal-axis frame: polar angle to the plane normal and in-plane angle to the long axis, both folded to $[0^\circ, 90^\circ]$; every contact contributes once from each end. (iv, v) $\text{D} \cdots \text{A}$ distance and $\text{D}-\text{H} \cdots \text{A}$ angle of the H-bonds ($\text{D}, \text{A} \in \{\text{N}, \text{O}\}$; $\text{H} \cdots \text{A} \leq r_{\text{H}} + r_{\text{A}}$ and angle $\geq 120^\circ$ after normalizing $\text{D}-\text{H}$ to the neutron length). (vi) H-bonds per molecule, $s = \min / \max$. Sub-metrics (i)–(v) are compared with the Wasserstein distance W between the two distributions and $s = e^{-W/\sigma}$. Contacts per molecule and the contact anisotropy are reported but not scored.

Conformation (4 sub-metrics). (i) Matched torsions: the heavy-atom graphs of the molecules are matched by graph isomorphism; for every automorphism, and for the mirror image (all torsions of opposite sign), the mean circular difference $D = \langle |\Delta\tau| \rangle$ over all torsions about rotatable bonds is computed and the smallest value is kept; for $Z' > 1$ every molecule is paired with its best partner in the other structure, in both directions, and the two directions are averaged; $s = e^{-D/30^\circ}$. (ii) Bond angles that contain a rotatable bond, (iii) improper angles (mean Wilson out-of-plane angle at every three-coordinate atom) and (iv) intramolecular contact lengths, each compared as distributions ($s = e^{-W/\sigma}$). Bond lengths are reported but not scored.

Fragment orientation (2 sub-metrics). (i) Relative orientation: for every pair of fragments of different molecules with centroid separation $\leq 8 \text{ \AA}$, the angles between their corresponding principal axes (long, in-plane, normal), folded to $[0^\circ, 90^\circ]$; the three angle distributions are compared separately and averaged. (ii) Aromatic contacts between the aromatic fragments of different molecules with centroid separation $\leq 6 \text{ \AA}$: number per molecule ($s = \min / \max$) and the distributions of centroid distance,

perpendicular distance (the smaller of the projections of the centroid vector onto the two plane normals), lateral offset and interplanar angle (from $|\hat{\mathbf{n}}_1 \cdot \hat{\mathbf{n}}_2|$), averaged. Motifs (face-to-face, parallel-displaced, tilted, edge-to-face) are reported. The fragment shape κ_f^2 is reported but not scored.

Spatial geometry (4 sub-metrics). (i) Distances from each asymmetric-unit fragment to the centroids of the fragments of other molecules within 12 Å. (ii) Radial distribution function $g(r)$ of these distances (0.1 Å bins, normalized by the bulk fragment number density, Gaussian smoothing 0.2 Å), compared by the overlap $1 - \int |g_1 - g_2| dr / \int (g_1 + g_2) dr$. (iii) Coordination number: the number of distinct neighbouring molecules with at least one atom in van der Waals contact with the molecule, $s = \min / \max$. (iv) First-shell anisotropy: relative shape anisotropy κ_s^2 of the vectors from the molecule to its contacting neighbours, compared as a distribution.

S2.4 Similarity and aggregation

Every distribution is compared with the one-dimensional Wasserstein distance W and converted to $s = e^{-W/\sigma}$, with $\sigma = 0.20$ Å for lengths, 10° for angles, 30° for matched torsions and 0.20 for dimensionless quantities (Eq. 8). The category similarity is the unweighted mean of its sub-metrics (Eq. 9) and the overall similarity the mean of the five categories (Eq. 10).

Table S3: Worked example: all intermediate quantities of the comparison of sulfathiazole Form II ($Z' = 1$) and Form III ($Z' = 2$). For distributions the number of values and their mean are given (“ n ; mean”). W , Wasserstein distance; D , mean circular torsion difference. The distributions of several of these quantities are plotted in Figures 4, 5 and 7 of the main text, for which this pair has the highest similarity. ^aUnique intermolecular contacts of the asymmetric unit. ^bEach contact counted from both ends. ^cTorsions about the two rotatable bonds over heavy atoms (9 per molecule), matched bond by bond. ^dNeighbouring fragments of other molecules within 8 Å. ^eAminophenyl and thiazolyl fragments, centroid separation ≤ 6 Å. ^fFragments of other molecules within 12 Å.

	Quantity	Form II	Form III	Distance	Transform	s
CP	Packing coefficient	0.717	0.732	–	$1 - \Delta /\max$	0.980
	Density (g cm^{-3})	1.574	1.600	–	$1 - \Delta /\max$	0.984
	Reduced-cell lengths (Å)	8.19, 8.54, 15.44	8.50, 15.51, 17.45	$d = 0.400$	e^{-d}	0.670
	Obliquity $ 90^\circ - \text{angle} $ ($^\circ$)	4.0, 0.0, 0.0	22.8, 0.0, 0.0	$ \Delta $ per angle	mean $e^{- \Delta /10^\circ}$	0.718
	Symmetry operations (standard setting)	$P2_1/c$ (No. 14)	$P2_1/c$ (No. 14)	4/4 shared	Jaccard	1.000
	Z	4	8	–	min / max	0.500
	Z'	1	2	–	min / max	0.500
	Crystal packing s_{cp} (mean of 7)					0.765
IN	Contact lengths (Å) ^a	87; 3.208	190; 3.220	$W = 0.031$ Å	$e^{-W/0.2}$ Å	0.854
	Contact polar angle ($^\circ$) ^b	174; 55.4	380; 56.0	$W = 1.12^\circ$	$e^{-W/10^\circ}$	0.894
	Contact in-plane angle ($^\circ$) ^b	174; 50.0	380; 47.5	$W = 2.67^\circ$	$e^{-W/10^\circ}$	0.766
	H-bond $\text{D} \cdots \text{A}$ (Å)	3; 2.916	6; 2.951	$W = 0.047$ Å	$e^{-W/0.2}$ Å	0.792
	H-bond angle $\text{D}-\text{H} \cdots \text{A}$ ($^\circ$)	3; 162.7	6; 162.2	$W = 1.19^\circ$	$e^{-W/10^\circ}$	0.888
	H-bonds per molecule	3	3	–	min / max	1.000
	Interaction networks s_{in} (mean of 6)					0.866
	reported: contacts per molecule; anisotropy	87; 0.321	95; 0.310			
CF	Matched torsions ($^\circ$) ^c	9	18	$D = 1.46^\circ$	$e^{-D/30^\circ}$	0.953
	Bond angles at rotatable bonds ($^\circ$)	8; 112.7	16; 112.6	$W = 0.27^\circ$	$e^{-W/10^\circ}$	0.974
	Improper (out-of-plane) angles ($^\circ$)	11; 5.5	22; 5.8	$W = 0.50^\circ$	$e^{-W/10^\circ}$	0.951
	Intramolecular contacts (Å)	8; 2.998	16; 2.993	$W = 0.016$ Å	$e^{-W/0.2}$ Å	0.922
	Conformation s_{cf} (mean of 4)					0.950
FO	Relative orientation, long axis ($^\circ$) ^d	57; 51.0	121; 53.6	$W = 3.31^\circ$	$e^{-W/10^\circ}$	0.718
	in-plane axis ($^\circ$)	57; 48.7	121; 49.9	$W = 2.06^\circ$	$e^{-W/10^\circ}$	0.813
	plane normal ($^\circ$)	57; 54.3	121; 53.9	$W = 2.29^\circ$	$e^{-W/10^\circ}$	0.795
	relative orientation (mean of 3 axes)					0.776
	Aromatic contacts per molecule ^e	10	10	–	min / max	1.000
	centroid distance (Å)	10; 4.86	20; 4.89	$W = 0.159$ Å	$e^{-W/0.2}$ Å	0.452
	perpendicular distance (Å)	10; 2.14	20; 2.17	$W = 0.250$ Å	$e^{-W/0.2}$ Å	0.286
	lateral offset (Å)	10; 4.05	20; 4.12	$W = 0.283$ Å	$e^{-W/0.2}$ Å	0.242
	interplanar angle ($^\circ$)	10; 61.4	20; 54.4	$W = 8.34^\circ$	$e^{-W/10^\circ}$	0.434
	aromatic contacts (mean of 5)					0.483
	Fragment orientation s_{fo} (mean of 2)					0.629
	reported: fragment shape κ_f^2	3; 0.399	6; 0.401	$W = 0.0022$	$e^{-W/0.2}$	
SG	Neighbour-fragment distances (Å) ^f	232; 9.21	469; 9.18	$W = 0.073$ Å	$e^{-W/0.2}$ Å	0.693
	$g(r)$, first peak (Å)	4.55	4.55	curve overlap	$1 - \frac{\int g_1 - g_2 }{\int (g_1 + g_2)}$	0.932
	Coordination number (molecules)	13	13.5	–	min / max	0.963
	First-shell anisotropy κ_s^2	1; 0.085	2; 0.054	$W = 0.031$	$e^{-W/0.2}$	0.858
	Spatial geometry s_{sg} (mean of 4)					0.861
	Overall $S = \sum_k 0.20 s_k$					0.814

S3. Invariance to the cell setting

S3.1 Setting-dependent descriptors

Table S4: Crystal-packing similarity of the sulfathiazole pairs computed with *setting-dependent* descriptors (cell as deposited, symmetry operations in the deposited setting), for the structures as deposited (Forms IV and V in $P2_1/n$) and after re-expressing Forms IV and V in $P2_1/c$ by the exact change of basis $\mathbf{c}' = \mathbf{a} + \mathbf{c}$ (obtuse- β convention: $\beta' = 132.6^\circ$ for Form IV and 125.27° for Form V). The two descriptions are the same crystals, yet the similarities change by up to 0.11 and the most similar pair changes from I–III to I–V. The Spearman rank correlation between the two columns is 0.77.

Pair	As deposited	IV, V in $P2_1/c$	Δ
I–II	0.817	0.817	+0.000
I–III	0.920	0.920	+0.000
I–IV	0.699	0.798	+0.099
I–V	0.867	0.979	+0.112
II–III	0.784	0.784	+0.000
II–IV	0.843	0.952	+0.109
II–V	0.694	0.797	+0.103
III–IV	0.716	0.800	+0.084
III–V	0.812	0.910	+0.098
IV–V	0.786	0.787	+0.001

S3.2 Setting-invariant descriptors

Table S5: Maximum absolute change in category and overall similarities over all 91 pairs (10 sulfathiazole, 45 galunisertib, 36 aripiprazole) computed with the setting-invariant descriptors of the main text after (i) transforming every $P2_1/c$ structure to $P2_1/n$ and every $P2_1/n$ structure to $P2_1/c$ ($\mathbf{c}' = \mathbf{c} \pm \mathbf{a}$), (ii) rotating the Cartesian frame of every structure by an arbitrary rotation, and (iii) re-describing every triclinic cell with $\mathbf{c}' = \mathbf{a} + \mathbf{b} + \mathbf{c}$. All differences are at the level of floating-point precision. The slightly larger residuals for fragment orientation arise from arccos evaluated at angles close to 0° , where rounding errors of order 10^{-16} in the cosine are amplified.

Category	$P2_1/c \leftrightarrow P2_1/n$	Rotated frame	Triclinic cell
Crystal packing	0	3×10^{-16}	0
Interaction networks	4×10^{-16}	3×10^{-15}	2×10^{-16}
Conformation	8×10^{-16}	1×10^{-15}	2×10^{-16}
Fragment orientation	2×10^{-09}	4×10^{-09}	2×10^{-09}
Spatial geometry	6×10^{-16}	7×10^{-16}	3×10^{-16}
Overall	4×10^{-10}	7×10^{-10}	4×10^{-10}

S4. Structure fingerprints

Column definitions (Tables S6–S8): a, b, c , lengths of the Niggli-reduced primitive cell; C/mol, unique intermolecular van der Waals contacts per molecule; $\bar{d}_c$, mean contact length (Å); HB/mol, unique hydrogen bonds per molecule; $\bar{d}_{DA}$, mean donor–acceptor distance (Å); IC/mol, intramolecular contacts per molecule; $\bar{\omega}_N$, mean improper (out-of-plane) angle at three-coordinate N atoms ($^\circ$); $\bar{\kappa}_f^2$, mean relative shape anisotropy of the fragments; π /mol, aromatic contacts (centroid–centroid ≤ 6 Å) per molecule; $\bar{\theta}_\pi$, mean interplanar angle of those contacts ($^\circ$); CN, molecular coordination number (distinct molecules in

van der Waals contact); $\bar{\kappa}_s^2$, mean relative shape anisotropy of the first coordination shell. The complete fingerprints (all distributions) are provided in machine-readable form with the code.

Table S6: Fingerprint summary for the five sulfathiazole polymorphs.

Form	Reduced cell (Å)			Interactions				Conformation		Fragments			Spatial	
	<i>a</i>	<i>b</i>	<i>c</i>	C/mol	$\bar{d}_c$	HB/mol	$\bar{d}_{DA}$	IC/mol	$\bar{\omega}_N$	$\bar{\kappa}_f^2$	π /mol	$\bar{\theta}_\pi$	CN	$\bar{\kappa}_s^2$
I	10.53	12.94	17.19	85.5	3.23	3.5	2.99	9.0	16.3	0.406	12.0	54.4	13.5	0.074
II	8.19	8.54	15.44	87.0	3.21	3.0	2.92	8.0	25.7	0.399	10.0	61.4	13.0	0.085
III	8.50	15.51	17.45	95.0	3.22	3.0	2.95	8.0	26.6	0.401	10.0	54.4	13.5	0.054
IV	8.47	10.77	11.37	105.0	3.21	3.0	2.98	8.0	27.5	0.403	12.0	52.7	16.0	0.085
V	10.40	14.28	15.13	85.0	3.25	4.0	3.02	7.5	11.0	0.399	13.0	68.0	13.0	0.073

Table S7: Fingerprint summary for the ten galunisertib polymorphs.

Form	Reduced cell (Å)			Interactions				Conformation		Fragments			Spatial	
	<i>a</i>	<i>b</i>	<i>c</i>	C/mol	$\bar{d}_c$	HB/mol	$\bar{d}_{DA}$	IC/mol	$\bar{\omega}_N$	$\bar{\kappa}_f^2$	π /mol	$\bar{\theta}_\pi$	CN	$\bar{\kappa}_s^2$
I	9.15	11.32	19.43	125.0	3.21	1.0	2.82	16.0	0.2	0.447	9.0	53.8	14.0	0.051
II	8.21	11.51	11.88	140.0	3.30	2.0	2.96	22.0	1.2	0.458	14.0	32.6	14.0	0.156
III	7.71	8.91	14.62	144.0	3.29	2.0	3.20	23.0	1.9	0.456	10.0	29.5	14.0	0.026
IV	9.85	19.00	21.88	124.5	3.28	3.0	3.08	25.5	1.3	0.459	11.0	42.9	15.0	0.045
V	8.26	11.64	19.61	147.0	3.28	2.0	3.11	21.0	7.4	0.453	10.0	15.0	16.0	0.068
VI	10.73	12.71	14.47	133.0	3.27	2.0	3.01	22.0	7.0	0.457	10.0	34.2	16.0	0.113
VII	8.13	12.62	18.06	146.0	3.22	2.5	3.24	26.5	4.8	0.453	13.0	29.9	15.5	0.101
VIII	8.55	11.37	11.52	172.0	3.26	2.0	3.06	24.0	9.4	0.449	10.0	17.3	16.0	0.070
IX	8.11	11.30	20.85	146.0	3.23	2.0	3.09	21.0	3.3	0.453	12.0	38.2	15.0	0.132
X	8.08	14.54	16.50	150.0	3.28	2.0	2.98	30.0	2.6	0.457	11.0	19.7	15.0	0.058

Table S8: Fingerprint summary for the nine aripiprazole forms.

Form	Reduced cell (Å)			Interactions				Conformation		Fragments			Spatial	
	<i>a</i>	<i>b</i>	<i>c</i>	C/mol	$\bar{d}_c$	HB/mol	$\bar{d}_{DA}$	IC/mol	$\bar{\omega}_N$	$\bar{\kappa}_f^2$	π /mol	$\bar{\theta}_\pi$	CN	$\bar{\kappa}_s^2$
I	7.57	8.68	17.38	110.0	3.34	1.0	2.88	46.0	31.2	0.293	4.0	61.0	18.0	0.340
II	7.71	12.47	23.61	137.0	3.35	1.0	2.80	39.0	31.8	0.293	2.0	14.1	18.0	0.226
III	10.15	12.08	18.60	137.0	3.27	1.0	2.85	40.5	33.5	0.296	3.5	24.5	13.0	0.718
IV	8.52	9.04	30.42	130.5	3.32	1.0	2.93	43.0	32.1	0.295	3.5	7.7	13.5	0.678
V	12.26	13.79	14.74	125.0	3.30	1.0	2.86	43.5	33.1	0.296	2.0	3.0	14.0	0.707
VII	7.03	9.98	15.95	166.0	3.36	1.0	2.86	46.0	32.5	0.296	4.0	40.7	15.0	0.485
VIII	7.65	13.12	21.67	150.0	3.32	1.0	2.82	45.0	31.7	0.293	2.0	14.1	18.0	0.200
IX	6.95	16.61	19.46	168.0	3.30	1.0	2.86	46.0	34.4	0.295	2.0	43.9	16.0	0.463
X	7.76	8.87	16.48	128.0	3.32	1.0	2.89	38.0	31.0	0.294	4.0	39.5	18.0	0.369

S5. Similarity vectors

CP, crystal packing; IN, interaction networks; CF, conformation; FO, fragment orientation; SG, spatial geometry; Overall, Eq. 10 with $w_k = 0.20$.

Table S9: Five-component similarity vectors and overall similarity for all sulfathiazole pairs.

Pair	CP	IN	CF	FO	SG	Overall
Form I–Form II	0.783	0.584	0.709	0.384	0.846	0.661
Form I–Form III	0.937	0.629	0.711	0.413	0.859	0.710
Form I–Form IV	0.774	0.634	0.695	0.455	0.786	0.669
Form I–Form V	0.936	0.731	0.702	0.460	0.846	0.735
Form II–Form III	0.765	0.866	0.950	0.629	0.861	0.814
Form II–Form IV	0.930	0.784	0.915	0.520	0.794	0.788
Form II–Form V	0.786	0.628	0.780	0.449	0.815	0.692
Form III–Form IV	0.775	0.850	0.942	0.634	0.825	0.805
Form III–Form V	0.914	0.677	0.800	0.454	0.815	0.732
Form IV–Form V	0.825	0.665	0.804	0.453	0.784	0.706

Table S10: Five-component similarity vectors and overall similarity for all galunisertib pairs.

Pair	CP	IN	CF	FO	SG	Overall
I-II	0.700	0.535	0.788	0.309	0.689	0.604
I-III	0.750	0.534	0.695	0.251	0.738	0.594
I-IV	0.766	0.502	0.748	0.344	0.774	0.627
I-V	0.943	0.476	0.684	0.249	0.756	0.622
I-VI	0.911	0.555	0.787	0.286	0.705	0.649
I-VII	0.758	0.513	0.707	0.312	0.693	0.597
I-VIII	0.672	0.511	0.681	0.250	0.736	0.570
I-IX	0.954	0.573	0.699	0.314	0.702	0.648
I-X	0.657	0.541	0.711	0.294	0.792	0.599
II-III	0.914	0.617	0.674	0.334	0.684	0.644
II-IV	0.603	0.598	0.762	0.362	0.720	0.609
II-V	0.697	0.699	0.743	0.322	0.663	0.625
II-VI	0.746	0.746	0.850	0.361	0.754	0.691
II-VII	0.716	0.564	0.712	0.442	0.752	0.637
II-VIII	0.920	0.656	0.698	0.300	0.646	0.644
II-IX	0.684	0.720	0.773	0.396	0.781	0.671
II-X	0.698	0.755	0.684	0.382	0.672	0.638
III-IV	0.622	0.656	0.684	0.275	0.786	0.604
III-V	0.758	0.582	0.667	0.350	0.677	0.607
III-VI	0.748	0.700	0.669	0.303	0.714	0.627
III-VII	0.767	0.644	0.653	0.307	0.711	0.616
III-VIII	0.879	0.639	0.676	0.251	0.676	0.624
III-IX	0.745	0.605	0.623	0.276	0.697	0.589
III-X	0.713	0.673	0.742	0.342	0.726	0.639
IV-V	0.777	0.547	0.665	0.299	0.745	0.607
IV-VI	0.795	0.621	0.768	0.497	0.769	0.690
IV-VII	0.748	0.679	0.697	0.431	0.759	0.663
IV-VIII	0.625	0.545	0.681	0.308	0.742	0.580
IV-IX	0.771	0.573	0.686	0.458	0.772	0.652
IV-X	0.709	0.655	0.768	0.400	0.817	0.670
V-VI	0.923	0.602	0.722	0.306	0.757	0.662
V-VII	0.789	0.615	0.736	0.363	0.723	0.645
V-VIII	0.699	0.713	0.797	0.410	0.836	0.691
V-IX	0.966	0.706	0.724	0.311	0.735	0.688
V-X	0.695	0.632	0.672	0.421	0.764	0.637
VI-VII	0.742	0.597	0.713	0.377	0.841	0.654
VI-VIII	0.736	0.690	0.717	0.345	0.758	0.650
VI-IX	0.912	0.697	0.712	0.422	0.836	0.716
VI-X	0.700	0.751	0.712	0.404	0.749	0.663
VII-VIII	0.719	0.600	0.739	0.343	0.721	0.624
VII-IX	0.779	0.628	0.660	0.479	0.804	0.670
VII-X	0.659	0.599	0.697	0.440	0.719	0.623
VIII-IX	0.680	0.682	0.647	0.356	0.708	0.615
VIII-X	0.701	0.640	0.731	0.368	0.781	0.644
IX-X	0.672	0.696	0.611	0.375	0.730	0.617

Table S11: Five-component similarity vectors and overall similarity for all aripiprazole pairs.

Pair	CP	IN	CF	FO	SG	Overall
Form I–Form II	0.771	0.736	0.650	0.233	0.601	0.598
Form I–Form III	0.662	0.731	0.671	0.219	0.486	0.554
Form I–Form IV	0.653	0.768	0.686	0.243	0.461	0.562
Form I–Form V	0.599	0.748	0.678	0.183	0.463	0.534
Form I–Form VII	0.801	0.737	0.716	0.290	0.557	0.620
Form I–Form VIII	0.756	0.750	0.786	0.237	0.644	0.634
Form I–Form IX	0.710	0.842	0.676	0.159	0.608	0.599
Form I–Form X	0.983	0.856	0.629	0.313	0.744	0.705
Form II–Form III	0.694	0.712	0.873	0.219	0.460	0.592
Form II–Form IV	0.704	0.788	0.858	0.317	0.428	0.619
Form II–Form V	0.642	0.753	0.812	0.370	0.413	0.598
Form II–Form VII	0.700	0.760	0.710	0.172	0.503	0.569
Form II–Form VIII	0.973	0.868	0.717	0.634	0.744	0.787
Form II–Form IX	0.777	0.756	0.759	0.284	0.468	0.609
Form II–Form X	0.771	0.726	0.845	0.297	0.609	0.649
Form III–Form IV	0.869	0.715	0.910	0.264	0.737	0.699
Form III–Form V	0.899	0.855	0.892	0.169	0.718	0.707
Form III–Form VII	0.804	0.798	0.732	0.225	0.577	0.627
Form III–Form VIII	0.698	0.729	0.704	0.241	0.497	0.574
Form III–Form IX	0.727	0.735	0.803	0.197	0.504	0.593
Form III–Form X	0.668	0.705	0.854	0.277	0.518	0.604
Form IV–Form V	0.848	0.778	0.912	0.367	0.709	0.723
Form IV–Form VII	0.738	0.729	0.743	0.320	0.572	0.620
Form IV–Form VIII	0.685	0.822	0.717	0.290	0.439	0.591
Form IV–Form IX	0.739	0.774	0.802	0.173	0.538	0.605
Form IV–Form X	0.647	0.771	0.838	0.269	0.494	0.604
Form V–Form VII	0.736	0.807	0.745	0.147	0.533	0.594
Form V–Form VIII	0.639	0.774	0.702	0.372	0.419	0.581
Form V–Form IX	0.717	0.770	0.785	0.174	0.528	0.595
Form V–Form X	0.606	0.721	0.806	0.260	0.504	0.579
Form VII–Form VIII	0.721	0.743	0.756	0.196	0.530	0.589
Form VII–Form IX	0.754	0.764	0.755	0.197	0.696	0.633
Form VII–Form X	0.805	0.708	0.679	0.240	0.604	0.607
Form VIII–Form IX	0.800	0.782	0.715	0.280	0.486	0.613
Form VIII–Form X	0.761	0.746	0.659	0.284	0.616	0.613
Form IX–Form X	0.713	0.839	0.727	0.208	0.615	0.620

S6. Comparison with COMPACK

Table S12: COMPACK results (20-molecule clusters) and overall similarity S for the sulfathiazole pairs.

Pair	Matched molecules	RMSD ($\text{\AA}$)	PXRD similarity	S
I-II	1/20	0.310	0.912	0.661
I-III	1/20	0.306	0.909	0.710
I-IV	1/20	0.329	0.885	0.669
I-V	1/20	0.248	0.943	0.735
II-III	16/20	0.121	0.985	0.814
II-IV	10/20	0.165	0.954	0.788
II-V	2/20	0.313	0.939	0.692
III-IV	16/20	0.144	0.973	0.805
III-V	2/20	0.289	0.934	0.732
IV-V	2/20	0.216	0.933	0.706

Table S13: COMPACK results (20-molecule clusters) and overall similarity S for the aripiprazole pairs.

Pair	Matched molecules	RMSD (Å)	PXRD similarity	S
I-II	1/20	0.831	0.883	0.598
I-III	1/20	1.238	0.946	0.554
I-IV	1/20	1.138	0.956	0.562
I-V	1/20	1.240	0.904	0.534
I-VII	1/20	2.018	0.932	0.621
I-VIII	3/20	0.775	0.918	0.634
I-IX	1/20	1.001	0.950	0.599
I-X	7/20	1.149	0.928	0.705
II-III	1/20	0.915	0.875	0.592
II-IV	1/20	0.902	0.889	0.619
II-V	1/20	0.950	0.904	0.598
II-VII	1/20	2.366	0.901	0.569
II-VIII	8/20	1.064	0.966	0.787
II-IX	1/20	1.391	0.903	0.609
II-X	9/20	0.550	0.834	0.649
III-IV	6/20	1.032	0.964	0.699
III-V	3/20	0.967	0.929	0.707
III-VII	1/20	2.786	0.962	0.627
III-VIII	1/20	1.198	0.905	0.574
III-IX	1/20	1.984	0.949	0.593
III-X	1/20	0.618	0.959	0.604
IV-V	2/20	0.514	0.922	0.723
IV-VII	1/20	2.685	0.962	0.620
IV-VIII	1/20	1.141	0.910	0.591
IV-IX	1/20	1.814	0.948	0.605
IV-X	1/20	0.764	0.946	0.604
V-VII	1/20	2.872	0.962	0.594
V-VIII	1/20	1.213	0.920	0.581
V-IX	1/20	2.004	0.909	0.595
V-X	1/20	0.738	0.944	0.579
VII-VIII	1/20	2.161	0.932	0.589
VII-IX	1/20	1.352	0.962	0.633
VII-X	1/20	2.506	0.963	0.607
VIII-IX	1/20	1.249	0.924	0.613
VIII-X	7/20	0.849	0.878	0.613
IX-X	1/20	1.552	0.938	0.620

Table S14: Spearman rank correlations of the category and overall similarities with the number of molecules matched by COMPACK (N_{match}) and with the COMPACK simulated PXRD similarity.

Category	Sulfathiazole (10 pairs)		Galunisertib (45 pairs)		Aripiprazole (36 pairs)	
	N_{match}	PXRD	N_{match}	PXRD	N_{match}	PXRD
Crystal packing	-0.38	-0.19	+0.02	+0.14	+0.60	+0.20
Interaction networks	+0.73	+0.81	+0.17	-0.06	+0.13	+0.08
Conformation	+0.95	+0.76	+0.41	+0.18	+0.15	+0.01
Fragment orientation	+0.69	+0.76	+0.50	+0.45	+0.32	-0.13
Spatial geometry	-0.02	+0.26	+0.32	+0.47	+0.69	+0.36
Overall	+0.73	+0.85	+0.44	+0.38	+0.67	+0.19

S7. Benchmark against other comparison methods

Methods and tests are described in Section 2.5 of the main text and the results are discussed in Section 3.6. Distances: $1 - S$ for the fingerprint and its categories; L_∞ between AMD vectors; earth mover’s distance between PDDs; $\sqrt{2 - 2K}$ for the SOAP/REMatch kernel K ; Euclidean distance between Niggli-reduced cell parameters ($\text{\AA}$, 10°); absolute density difference; $20 - N_{\text{match}}$ and $1 - \text{PXR}$ similarity for COMPACK. AMD and PDD use $k = 100$ neighbours; SOAP uses $r_{\text{cut}} = 6 \text{ \AA}$, $n_{\text{max}} = 8$, $l_{\text{max}} = 6$ and $\sigma = 0.5 \text{ \AA}$, for every atom of the unit cell, with the REMatch kernel ($\alpha = 0.1$, linear base kernel on normalized spectra); the reduced cell is compared with lengths in $\text{\AA}$ and angles in units of 10° . AMD, PDD and SOAP are computed without hydrogen atoms, whose positions in X-ray structures are systematically short, unless stated otherwise (Table S21). Identification measures: 1-NN, fraction of queries whose nearest neighbour is of the same form; mAP, mean average precision; AUC, area under the ROC curve for separating same-form from different-form pairs; ARI, adjusted Rand index of average-linkage clustering into as many clusters as there are forms. Pairs with nearly the same scalar distance can differ in different respects, as shown for SOAP in Figure 20e of the main text; for AMD, galunisertib Polymorphs V–IX and VII–VIII have distances of 0.097 and 0.093 but crystal-packing similarities of 0.97 and 0.72, and for PDD, Polymorphs I–IX and II–IV have distances of 0.435 and 0.433 and crystal-packing similarities of 0.95 and 0.60. The code that produces every number in this section is part of the code release (steps 8–12 of the analysis pipeline).

S7.1 Identification of the form

Table S15: Polymorph identification on the 47 sulfathiazole structures for every method and every category. Margin: smallest distance between structures of different forms divided by the largest distance between structures of the same form (above 1: the forms are completely separated by a distance threshold); $d_{\text{within}}^{\text{max}}$ and $d_{\text{between}}^{\text{min}}$ are these two distances. COMPACK results were not available for SUTHAZ01 ($n = 46$); COMPACK matches all 20 molecules for every same-form pair and at most 16 for every different-form pair, so its margin is infinite. Without SUTHAZ01 and SUTHAZ06, the overall similarity gives AUC 1.00, ARI 1.00 and margin 1.01.

Method	n	1-NN	mAP	AUC	ARI	Margin	$d_{\text{within}}^{\text{max}}$	$d_{\text{between}}^{\text{min}}$
This work, overall S	47	1.00	1.00	0.99	0.70	0.83	0.222	0.185
Crystal packing	47	1.00	1.00	1.00	1.00	3.02	0.0175	0.0529
Interaction networks	47	0.96	0.87	0.90	0.34	0.26	0.483	0.125
Conformation	47	0.87	0.72	0.84	0.34	0.14	0.313	0.0445
Fragment orientation	47	1.00	1.00	1.00	1.00	0.98	0.35	0.344
Spatial geometry	47	1.00	0.90	0.89	0.52	0.53	0.232	0.124
AMD ($k = 100$)	47	1.00	0.85	0.94	0.64	0.37	0.117	0.0437
PDD ($k = 100$)	47	1.00	1.00	1.00	1.00	0.91	0.226	0.205
SOAP/REMatch	47	1.00	1.00	1.00	1.00	1.15	0.0648	0.0744
Reduced cell	47	1.00	0.96	0.99	0.84	0.74	3.61	2.67
Deposited cell	47	0.98	0.95	0.95	0.61	0.37	9.9	3.62
Density	47	0.83	0.62	0.83	0.49	0.00	0.0692	6.39e-05
COMPACK N_{match}	46	1.00	1.00	1.00	1.00	∞	0	4
COMPACK PXR	46	1.00	1.00	1.00	1.00	0.86	0.0147	0.0126

Table S16: Polymorph identification on the 21 aripiprazole determinations and agreement with COMPACK. Identification: each of the 16 determinations of forms with more than one entry is used as a query against the other 20 determinations; the margin is the smallest distance between determinations of different forms divided by the largest distance between determinations of the same form. Right: Spearman correlations with COMPACK over the 36 pairs of representatives (similarity against N_{match} , $-\text{RMSD}$ and PXR similarity). Pair-bootstrap 95% intervals of the difference between the N_{match} correlations of the overall similarity and of AMD, PDD and SOAP: [0.19, 0.86], [0.05, 0.58] and [0.01, 0.61].

Method	Identification (21 structures, 16 queries)				ρ with COMPACK (36 pairs)		
	1-NN	mAP	AUC	Margin	N_{match}	RMSD	PXR
This work, overall S	1.00	1.00	1.00	1.61	+0.67	+0.18	+0.19
Crystal packing	1.00	1.00	1.00	1.02	+0.60	−0.13	+0.20
Interaction networks	0.81	0.89	0.99	0.71	+0.13	−0.11	+0.08
Conformation	1.00	1.00	1.00	1.14	+0.15	+0.34	+0.01
Fragment orientation	1.00	1.00	1.00	1.35	+0.32	+0.37	−0.13
Spatial geometry	1.00	1.00	1.00	1.68	+0.69	+0.06	+0.36
AMD ($k = 100$)	0.81	0.85	0.99	0.59	+0.14	+0.43	−0.06
PDD ($k = 100$)	1.00	1.00	1.00	2.05	+0.36	+0.32	+0.10
SOAP/REMatch	1.00	1.00	1.00	1.29	+0.37	+0.38	+0.03
Reduced cell	1.00	1.00	1.00	3.02	+0.13	−0.14	+0.00
Deposited cell	1.00	1.00	1.00	3.02	+0.03	−0.29	+0.64
Density	0.31	0.55	0.83	0.00	+0.25	+0.48	−0.12
COMPACK N_{match}	1.00	1.00	1.00	∞	−	−	−
COMPACK PXR	1.00	1.00	1.00	2.96	−	−	−

S7.2 Variation between determinations of the same form

Table S17: Category and overall similarities of pairs of sulfathiazole structures of the same form (independent determinations, including the DFT-D structure of the form) and of different forms (47 structures).

Category	Same form (210 pairs)		Different forms (871 pairs)	
	Mean	Range	Mean	Range
Crystal packing	0.993	0.98–1.00	0.836	0.75–0.95
Interaction networks	0.853	0.52–0.97	0.699	0.46–0.87
Conformation	0.909	0.69–0.99	0.781	0.50–0.96
Fragment orientation	0.826	0.65–0.98	0.484	0.34–0.66
Spatial geometry	0.918	0.77–0.99	0.814	0.68–0.88
Overall	0.900	0.78–0.98	0.723	0.61–0.82

Table S18: Category and overall similarities of pairs of aripiprazole determinations of the same form and of different forms (21 determinations).

Category	Same form (33 pairs)		Different forms (177 pairs)	
	Mean	Minimum	Mean	Maximum
Crystal packing	0.996	0.983	0.774	0.983
Interaction networks	0.932	0.818	0.779	0.870
Conformation	0.964	0.925	0.752	0.914
Fragment orientation	0.912	0.745	0.340	0.657
Spatial geometry	0.939	0.854	0.563	0.755
Overall	0.949	0.871	0.641	0.792

S7.3 Agreement with COMPACK and category response

Table S19: Spearman rank correlations with COMPACK. For each method the similarity (negative distance) is correlated with N_{match} , with $-\text{RMSD}$ and with the simulated PXR D similarity, so that positive values mean agreement. Left and centre: distinct polymorphs (sulfathiazole DFT-D structures, galunisertib Polymorphs I–X). Right: all 1035 pairs of the 46 sulfathiazole structures with COMPACK results. For galunisertib, pair-bootstrap 95% intervals (5000 resamples) of the difference between the N_{match} correlation of the overall similarity and that of each descriptor are $[-0.05, 0.68]$ for AMD, $[-0.08, 0.64]$ for PDD and $[-0.02, 0.71]$ for SOAP.

Method	Sulfathiazole, DFT (10 pairs)			Galunisertib (45 pairs)			All sulfathiazole PXR D (1035)
	N_{match}	RMSD	PXR D	N_{match}	RMSD	PXR D	
This work, overall S	+0.73	+0.89	+0.85	+0.44	+0.33	+0.38	+0.90
Crystal packing	−0.38	−0.08	−0.19	+0.02	−0.01	+0.14	+0.32
Interaction networks	+0.73	+0.88	+0.81	+0.17	+0.10	−0.06	+0.87
Conformation	+0.95	+0.84	+0.76	+0.41	+0.62	+0.18	+0.81
Fragment orientation	+0.69	+0.72	+0.76	+0.50	+0.06	+0.45	+0.87
Spatial geometry	−0.02	+0.24	+0.26	+0.32	+0.37	+0.47	+0.54
AMD ($k = 100$)	+0.12	+0.44	+0.58	+0.13	+0.07	+0.18	+0.78
PDD ($k = 100$)	+0.51	+0.66	+0.78	+0.16	+0.11	+0.39	+0.87
SOAP/REMatch	+0.42	+0.62	+0.59	+0.08	−0.00	+0.55	+0.84
Reduced cell	−0.54	−0.20	−0.26	+0.01	−0.05	+0.04	+0.26
Deposited cell	−0.31	−0.15	+0.02	+0.30	+0.25	−0.17	+0.32
Density	+0.24	+0.47	+0.70	+0.04	+0.02	+0.13	+0.70

Table S20: Category response: Spearman correlation between the similarity given by each single-number method (negative distance; PXRD similarity for COMPACK) and each category similarity of the fingerprint. CP, crystal packing; IN, interaction networks; CF, conformation; FO, fragment orientation; SG, spatial geometry. Distinct forms: the 10 pairs of sulfathiazole DFT-D structures, the 45 galunisertib pairs and the 36 pairs of aripiprazole representatives; in none of the three systems do AMD, PDD and SOAP follow crystal packing. Over all 1081 sulfathiazole pairs (^a1035 pairs for COMPACK PXRD), which are dominated by the difference between same-form and different-form pairs, all categories except crystal packing correlate with AMD, PDD, SOAP, the density and the PXRD similarity, whereas the reduced cell follows crystal packing only.

Pairs	Method	CP	IN	CF	FO	SG	Overall
Sulfathiazole, DFT (10)	AMD ($k = 100$)	−0.01	+0.41	+0.14	+0.37	+0.81	+0.54
	PDD ($k = 100$)	−0.16	+0.62	+0.50	+0.60	+0.65	+0.73
	SOAP/REMatch	−0.15	+0.65	+0.44	+0.62	+0.64	+0.73
	Reduced cell	+0.87	−0.22	−0.43	−0.38	+0.08	−0.12
	Density	−0.09	+0.45	+0.20	+0.45	+0.67	+0.55
	COMPACK PXRD	−0.19	+0.81	+0.76	+0.76	+0.26	+0.85
Galunisertib (45)	AMD ($k = 100$)	+0.08	+0.16	−0.11	+0.19	+0.28	+0.18
	PDD ($k = 100$)	+0.07	+0.13	−0.06	+0.27	+0.38	+0.28
	SOAP/REMatch	−0.08	+0.10	−0.01	+0.25	+0.41	+0.16
	Reduced cell	+0.71	−0.18	+0.01	−0.21	−0.17	+0.21
	Density	+0.06	+0.13	−0.06	+0.23	+0.27	+0.17
	COMPACK PXRD	+0.14	−0.06	+0.18	+0.45	+0.47	+0.38
Aripiprazole (36)	AMD ($k = 100$)	−0.20	+0.04	+0.35	+0.28	−0.21	−0.05
	PDD ($k = 100$)	+0.04	+0.22	+0.41	+0.22	+0.10	+0.20
	SOAP/REMatch	+0.02	+0.40	+0.22	+0.31	+0.06	+0.17
	Reduced cell	+0.36	−0.11	−0.28	−0.01	+0.15	+0.11
	Density	+0.03	−0.14	+0.30	+0.22	+0.06	+0.11
	COMPACK PXRD	+0.20	+0.08	+0.01	−0.13	+0.36	+0.19
All sulfathiazole (1081) ^a	AMD ($k = 100$)	+0.37	+0.62	+0.50	+0.61	+0.63	+0.68
	PDD ($k = 100$)	+0.35	+0.74	+0.67	+0.76	+0.61	+0.81
	SOAP/REMatch	+0.36	+0.79	+0.69	+0.85	+0.64	+0.88
	Reduced cell	+0.93	+0.15	+0.13	+0.18	+0.28	+0.32
	Density	+0.31	+0.53	+0.44	+0.49	+0.50	+0.56
	COMPACK PXRD	+0.32	+0.87	+0.81	+0.87	+0.54	+0.90

S7.4 Baseline parameters and invariance

Table S21: Sensitivity of the baseline results to their parameters: inclusion of hydrogen atoms, SOAP cutoff radius, REMatch regularization α and the number of neighbours k of AMD and PDD. Only the parameter named in the first column differs from the reference settings. Including hydrogen atoms degrades all three descriptors; no variant raises the agreement with COMPACT N_{match} for galunisertib above $\rho = 0.34$ (fingerprint: 0.44).

Variant	Method	mAP	AUC	ARI	Dup. ratio	$\rho(N_{\text{match}})$, galunisertib
Reference	AMD	0.85	0.94	0.64	0.89	+0.13
	PDD	1.00	1.00	1.00	0.38	+0.16
	SOAP/REMatch	1.00	1.00	1.00	0.14	+0.08
With H atoms	AMD	0.58	0.79	0.35	1.01	+0.14
	PDD	0.95	0.97	0.72	0.61	+0.07
	SOAP/REMatch	0.94	0.97	0.72	0.29	+0.27
SOAP $r_{\text{cut}} = 4 \text{ \AA}$	AMD	0.85	0.94	0.64	0.89	+0.13
	PDD	1.00	1.00	1.00	0.38	+0.16
	SOAP/REMatch	1.00	1.00	1.00	0.16	+0.24
SOAP $r_{\text{cut}} = 8 \text{ \AA}$	AMD	0.85	0.94	0.64	0.89	+0.13
	PDD	1.00	1.00	1.00	0.38	+0.16
	SOAP/REMatch	1.00	1.00	1.00	0.15	+0.08
REMatch $\alpha = 0.5$	AMD	0.85	0.94	0.64	0.89	+0.13
	PDD	1.00	1.00	1.00	0.38	+0.16
	SOAP/REMatch	1.00	1.00	1.00	0.17	+0.16
AMD/PDD $k = 50$	AMD	0.88	0.95	0.84	0.89	+0.34
	PDD	1.00	0.99	0.72	0.33	+0.22
	SOAP/REMatch	1.00	1.00	1.00	0.14	+0.08
AMD/PDD $k = 200$	AMD	0.80	0.92	0.56	0.96	+0.10
	PDD	1.00	1.00	1.00	0.41	+0.15
	SOAP/REMatch	1.00	1.00	1.00	0.14	+0.08

Table S22: Invariance of all methods, tested as in Section 2.4 on all 79 structures (47 sulfathiazole, 11 galunisertib, 21 aripiprazole). For the fingerprint, the maximum absolute change in the overall similarity (largest change in any category: 7.8×10^{-9} , fragment orientation); for the other methods, the maximum absolute change in any pairwise distance divided by the median pairwise distance. Only the comparison of cells as deposited is not invariant.

Method	$P2_1/c \leftrightarrow P2_1/n$	Frame rotation	Triclinic re-cell
This work, overall S	1.6×10^{-9}	1.5×10^{-9}	4.0×10^{-10}
AMD ($k = 100$)	3.2×10^{-14}	7.9×10^{-14}	1.1×10^{-14}
PDD ($k = 100$)	3.3×10^{-16}	3.3×10^{-16}	0
SOAP/REMatch	5.7×10^{-13}	1.2×10^{-12}	2.0×10^{-13}
Reduced cell	0	1.0×10^{-15}	0
Deposited cell	1.5×10^0	8.6×10^{-16}	5.2×10^{-1}
Density	0	0	0

S8. Robustness, sensitivity and redundancy

Results are discussed in Section 3.7 of the main text and summarized in Figure S1. All analyses use the fifteen sulfathiazole and galunisertib structures of the main analysis and the nine aripiprazole representatives, with the reference parameters unless stated otherwise.

S8.1 Scale parameters and weights

Table S23: Mean category and overall similarities when all scale parameters σ of Eq. 8 are halved or doubled, and Spearman rank correlation of the overall similarity of all pairs with the reference values (σ). In crystal packing only the reduced-cell obliquity depends on σ . Absolute values shift; the ranking of pairs is essentially unchanged.

Category	Sulfathiazole			Galunisertib			Aripiprazole		
	$\sigma/2$	σ	2σ	$\sigma/2$	σ	2σ	$\sigma/2$	σ	2σ
Crystal packing	0.84	0.84	0.85	0.73	0.76	0.78	0.72	0.74	0.76
Interaction networks	0.54	0.70	0.82	0.46	0.62	0.76	0.62	0.77	0.87
Conformation	0.66	0.80	0.89	0.56	0.71	0.83	0.62	0.76	0.86
Fragment orientation	0.31	0.49	0.66	0.19	0.35	0.55	0.15	0.26	0.42
Spatial geometry	0.75	0.82	0.88	0.64	0.74	0.82	0.47	0.56	0.67
Overall	0.62	0.73	0.82	0.52	0.64	0.75	0.52	0.62	0.72
Rank ρ (overall)	0.99	1	1.00	0.98	1	0.98	0.97	1	0.98

Table S24: Top: effect of weighting the sub-metrics unequally (conformation 50/30/20%, fragment orientation 30/30/40%, with the fragment-shape sub-metric scored) instead of using the unweighted mean of Eq. 9. Bottom: Spearman rank correlation of the overall similarity of all pairs with the five-category value when one category is omitted and the remaining four are weighted equally.

Test	Sulfathiazole	Galunisertib	Aripiprazole
<i>Unequal sub-metric weights</i>			
Spearman ρ with Eq. (9)	0.976	0.970	0.945
max $ \Delta S $	0.040	0.038	0.047
<i>Category omitted (Spearman ρ with five-category S)</i>			
Crystal packing	0.952	0.793	0.879
Interaction networks	0.976	0.877	0.963
Conformation	1.000	0.958	0.925
Fragment orientation	0.976	0.916	0.897
Spatial geometry	0.976	0.945	0.819

S8.2 Redundancy, coordinate perturbation and cutoffs

Table S25: Spearman rank correlations between the five category similarities over the pairs of each system. CP, crystal packing; IN, interaction networks; CF, conformation; FO, fragment orientation; SG, spatial geometry.

		CP	IN	CF	FO	SG
Sulfathiazole (10 pairs)	CP	–	–0.18	–0.25	–0.32	–0.01
	IN	–0.18	–	+0.68	+0.94	+0.13
	CF	–0.25	+0.68	–	+0.55	+0.10
	FO	–0.32	+0.94	+0.55	–	–0.01
	SG	–0.01	+0.13	+0.10	–0.01	–
Galunisertib (45 pairs)	CP	–	–0.13	–0.01	–0.15	–0.12
	IN	–0.13	–	+0.09	+0.36	+0.04
	CF	–0.01	+0.09	–	+0.22	+0.25
	FO	–0.15	+0.36	+0.22	–	+0.40
	SG	–0.12	+0.04	+0.25	+0.40	–
Aripiprazole (36 pairs)	CP	–	+0.21	+0.04	+0.12	+0.77
	IN	+0.21	–	–0.14	–0.04	+0.16
	CF	+0.04	–0.14	–	+0.07	+0.02
	FO	+0.12	–0.04	+0.07	–	–0.08
	SG	+0.77	+0.16	+0.02	–0.08	–

Table S26: Coordinate perturbation. Every atom was displaced by an isotropic Gaussian vector with the given root-mean-square length (10 realizations per level, cell fixed) and all coordinate-dependent quantities were recomputed; the lists of contacts and H-bond candidates were kept fixed. Crystal packing does not change because the cell, density and packing coefficient are fixed. (a) Similarity of each structure to its perturbed copy, averaged over structures and realizations. (b) Standard deviation of each pair similarity over the realizations, averaged over pairs (for comparison, the standard deviation of the overall similarity across pairs is 0.055 for sulfathiazole, 0.032 for galunisertib and 0.051 for aripiprazole), and the Spearman correlation between the perturbed and reference overall similarities of all pairs. ^aAripiprazole without the eight pairs that involve Form VII, which has an N–H···N contact at the H-bond criterion ($H\cdots N = 2.75 \text{ \AA}$).

	rms (Å)	CP	IN	CF	FO	SG	Overall	ρ_S
<i>(a) Similarity to the perturbed copy</i>								
Sulfathiazole	0.01	1.000	0.976	0.971	0.980	0.995	0.984	
	0.02	1.000	0.957	0.944	0.961	0.991	0.970	
	0.05	1.000	0.904	0.862	0.879	0.979	0.925	
	0.10	1.000	0.851	0.744	0.796	0.967	0.872	
Galunisertib	0.01	1.000	0.975	0.968	0.986	0.997	0.985	
	0.02	1.000	0.955	0.938	0.972	0.993	0.972	
	0.05	1.000	0.899	0.855	0.928	0.983	0.933	
	0.10	1.000	0.838	0.735	0.868	0.970	0.882	
Aripiprazole	0.01	1.000	0.953	0.971	0.939	0.997	0.972	
	0.02	1.000	0.930	0.944	0.905	0.992	0.954	
	0.05	1.000	0.866	0.868	0.844	0.978	0.911	
	0.10	1.000	0.799	0.756	0.784	0.961	0.860	
<i>(b) Standard deviation of the pair similarities; ρ_S (mean; minimum)</i>								
Sulfathiazole	0.01	0.000	0.004	0.006	0.003	0.001	0.001	1.00; 1.00
	0.02	0.000	0.007	0.010	0.005	0.001	0.003	1.00; 0.99
	0.05	0.000	0.022	0.021	0.017	0.003	0.008	0.98; 0.94
	0.10	0.000	0.028	0.033	0.022	0.004	0.010	0.95; 0.90
Galunisertib	0.01	0.000	0.008	0.005	0.001	0.001	0.002	0.99; 0.98
	0.02	0.000	0.013	0.008	0.002	0.001	0.003	0.99; 0.97
	0.05	0.000	0.031	0.018	0.007	0.003	0.008	0.96; 0.91
	0.10	0.000	0.043	0.030	0.012	0.005	0.011	0.93; 0.89
Aripiprazole	0.01	0.000	0.036	0.004	0.010	0.001	0.008	0.89; 0.74
	0.02	0.000	0.043	0.007	0.015	0.002	0.010	0.88; 0.74
	0.05	0.000	0.064	0.014	0.022	0.003	0.014	0.86; 0.72
	0.10	0.000	0.069	0.023	0.025	0.005	0.016	0.85; 0.70
without Form VII ^a	0.01							0.99; 0.98
	0.02							0.98; 0.95
	0.05							0.94; 0.92
	0.10							0.92; 0.85

Table S27: Sensitivity to the cutoffs of the method. One parameter at a time was changed from its reference value. ρ_{cat} : Spearman correlation between the new and reference similarities of the affected category; $\max|\Delta s|$: largest change of that category similarity; ρ_S : Spearman correlation of the overall similarity. The largest change of the overall similarity over all variants is 0.014 for sulfathiazole, 0.041 for galunisertib and 0.081 for aripiprazole (aromatic-contact cutoff 5 Å).

Parameter	Ref.	Value	Cat.	Sulfathiazole			Galunisertib			Aripiprazole		
				ρ_{cat}	$\max \Delta s $	ρ_S	ρ_{cat}	$\max \Delta s $	ρ_S	ρ_{cat}	$\max \Delta s $	ρ_S
Aromatic-contact cutoff (Å)	6	5	FO	0.79	0.045	0.98	0.81	0.115	0.95	0.45	0.405	0.75
Aromatic-contact cutoff (Å)	6	7	FO	0.89	0.044	0.99	0.85	0.078	0.97	0.75	0.243	0.90
Orientation cutoff (Å)	8	6	FO	0.92	0.072	1.00	0.72	0.130	0.95	0.70	0.170	0.92
Orientation cutoff (Å)	8	10	FO	0.90	0.067	0.99	0.89	0.095	0.98	0.84	0.071	0.97
Environment cutoff (Å)	12	10	SG	0.96	0.019	1.00	0.82	0.069	0.98	0.95	0.077	0.98
Environment cutoff (Å)	12	15	SG	0.87	0.043	0.99	0.82	0.060	0.98	0.98	0.057	0.99
Min. H-bond angle (°)	120	110	IN	1.00	0.000	1.00	1.00	0.000	1.00	1.00	0.000	1.00
Min. H-bond angle (°)	120	130	IN	0.95	0.068	0.99	0.78	0.205	0.95	1.00	0.000	1.00
$g(r)$ smoothing (Å)	0.2	0.1	SG	0.98	0.027	1.00	0.99	0.037	1.00	0.99	0.038	1.00
$g(r)$ smoothing (Å)	0.2	0.3	SG	0.98	0.011	1.00	1.00	0.019	1.00	1.00	0.024	1.00

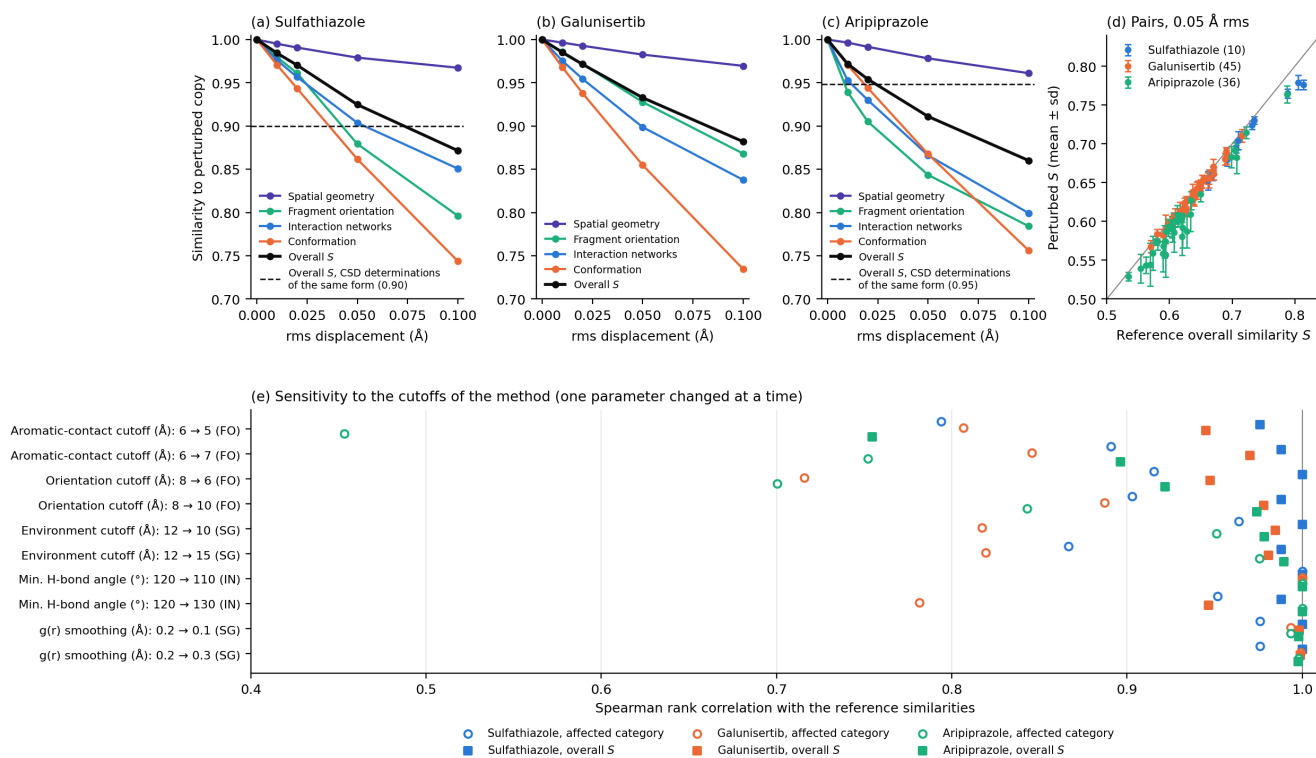


Figure S1: Robustness of the similarities. (a–c) Similarity of each structure to its own perturbed copy as a function of the rms displacement of the atoms (mean over structures and 10 realizations; crystal packing is 1 at all levels because the cell is fixed). The dashed lines in (a) and (c) are the mean overall similarities of independent CSD determinations of the same form. (d) Overall similarity of every pair of distinct polymorphs after perturbation by 0.05 Å rms (mean $\pm$ standard deviation over 10 realizations) against its reference value; perturbation lowers all similarities slightly but preserves their order, except for the aripiprazole pairs that involve Form VII (see Table S26). (e) Spearman rank correlation between the similarities computed with one cutoff changed and the reference similarities, for the affected category (open circles) and the overall similarity (filled squares).