Supporting Information

First Global Analysis of the GSK Database of Small Molecule Crystal Structures

Leen N. Kalash,[1] Jason C. Cole,[2] Royston C. B Copley,[1] Colin M. Edge,[1] Alexandru A. Moldovan,[2] Ghazala Sadiq,[2] Cheryl L. Doherty[1*]

1. Medicinal Science & Technology, GlaxoSmithKline, GSK Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK.
2. The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK.

Tables

	T (K)	Mean	Standard Deviation	Median	Frequency
GSK	100	0.685	0.046	0.695	72
	110	0.701	0.027	0.705	132
	150	0.689	0.036	0.691	757
	290-300	0.669	0.034	0.671	278
	Not reported	N/A	N/A	N/A	1037
	Other	N/A	N/A	N/A	141
	All data	0.683	0.038	0.686	2417
CSD-DS	All data	0.691	0.039	0.693	4885

Table S1: Packing coefficients at different temperatures

	Mann-Whitney U Statistics	p-value
R-factor	4620834.500	0.458
Packing Coefficient	5119669.000	0.000
Percenatge Void Volume	5544330.500	0.000
HB Count	3110457.500	0.000
Logp	567313.000	0.000
Molecular Weight	522987.000	0.000
Flexibility	804313.000,	0.474
Branching	522008.000	0.000

Table S2: Mann-Whitney U Statistics and p values

Donor	Acceptor	Competition	Donor steric density	Acceptor steric density	Donor aromaticity	Acceptor aromaticity	**Propensity**	Lower bound	Upper bound	Frequency	Observed Inter-?
N1 of al_prim_amine	O3 of al_carbonyl_1	2.75	66.42	66.39	0.36	0.36	**0.30**	0.09	0.63		
N1 of al_prim_amine	O1 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.15**	0.04	0.43		observed
N1 of al_prim_amine	O2 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.15**	0.04	0.43		observed
N1 of al_prim_amine	N3 of cyano	2.75	66.42	34.65	0.36	0.36	**0.08**	0.02	0.25		
N1 of al_prim_amine	O4 of ar_methoxy	3.67	66.42	39.53	0.36	0.36	**0.04**	0.01	0.15		
N1 of al_prim_amine	N1 of al_prim_amine	5.50	66.42	66.42	0.36	0.36	**0.04**	0.02	0.09		

Intra-molecular H-bond propensities:

Donor	Acceptor	Donor sybyl atom type	Acceptor sybyl atom type	DA Pair constrained connectivity	DA Pair path string	Donor count	**Propensity**	Lower bound	Upper bound	Observed Intra-?
N1	O3	N.3	O.2	0	0	N/A	**0.57**	0.57	0.57	
N1	O1	N.3	O.2	0	0	2	**0.55**	0.55	0.55	
N1	O2	N.3	O.2	0	0	2	**0.55**	0.55	0.55	

Table S3: Hydrogen bond propensity values using GSK training set for experimentally less stable form of GW825964X (X_2915A1)

Donor	Acceptor	Competition	Donor steric density	Acceptor steric density	Donor aromaticity	Acceptor aromaticity	**Propensity**	Lower bound	Upper bound	Frequency	Observed Inter-?
N1 of al_prim_amine	O1 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.18**	0.05	0.47		observed
N1 of al_prim_amine	O2 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.18**	0.05	0.47		observed
N1 of al_prim_amine	N1 of al_prim_amine	5.50	66.42	66.42	0.36	0.36	**0.09**	0.03	0.22		
N1 of al_prim_amine	O3 of al_carbonyl_1	2.75	66.42	66.39	0.36	0.36	**0.07**	0.02	0.26		
N1 of al_prim_amine	N3 of cyano	2.75	66.42	34.65	0.36	0.36	**0.06**	0.02	0.19		

N1 of al_prim_amine	O4 of ar_methoxy	3.67	66.42	39.53	0.36	0.36	**0.04**	0.01	0.11		

Intra-molecular H-bond propensities:

Donor	Acceptor	Donor sybyl atom type	Acceptor sybyl atom type	DA Pair constrained connectivity	DA Pair path string	Donor count	**Propensity**	Lower bound	Upper bound	Observed Intra-?
N1	O3	N.3	O.2	0	0	N/A	**0.57**	0.57	0.57	
N1	O1	N.3	O.2	0	0	2	**0.55**	0.55	0.55	
N1	O2	N.3	O.2	0	0	2	**0.55**	0.55	0.55	

Table S4: Hydrogen bond propensity values using CSD DS training set for experimentally less stable form of GW825964X (X_2915A1)

Donor	Acceptor	Competition	Donor steric density	Acceptor steric density	Donor aromaticity	Acceptor aromaticity	**Propensity**	Lower bound	Upper bound	Frequency	Observed Inter-?
N1 of al_prim_amine	O3 of al_carbonyl_1	2.75	66.42	66.39	0.36	0.36	**0.17**	0.08	0.34		
N1 of al_prim_amine	O1 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.12**	0.05	0.25		observed
N1 of al_prim_amine	O2 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.12**	0.05	0.25		observed
N1 of al_prim_amine	N1 of al_prim_amine	5.50	66.42	66.42	0.36	0.36	**0.06**	0.03	0.10		
N1 of al_prim_amine	N3 of cyano	2.75	66.42	34.65	0.36	0.36	**0.05**	0.02	0.12		
N1 of al_prim_amine	O4 of ar_methoxy	3.67	66.42	39.53	0.36	0.36	**0.04**	0.02	0.09		

Intra-molecular H-bond propensities:

Donor	Acceptor	Donor sybyl atom type	Acceptor sybyl atom type	DA Pair constrained connectivity	DA Pair path string	Donor count	**Propensity**	Lower bound	Upper bound	Observed Intra-?
N1	O3	N.3	O.2	0	0	N/A	**0.57**	0.57	0.57	
N1	O1	N.3	O.2	0	0	2	**0.55**	0.55	0.55	
N1	O2	N.3	O.2	0	0	2	**0.55**	0.55	0.55	

Table S5: Hydrogen bond propensity values using GSK + CSD DS training set for experimentally less stable form of GW825964X (X_2915A1)

Donor	Acceptor	Competition	Donor steric density	Acceptor steric density	Donor aromaticity	Acceptor aromaticity	**Propensity**	Lower bound	Upper bound	Frequency	Observed Inter-?

N1 of al_prim_amine	O3 of al_carbonyl_1	2.75	66.42	66.39	0.36	0.36	**0.30**	0.10	0.64		observed
N1 of al_prim_amine	O1 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.15**	0.04	0.42		
N1 of al_prim_amine	O2 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.15**	0.04	0.42		
N1 of al_prim_amine	N3 of cyano	2.75	66.42	34.65	0.36	0.36	**0.08**	0.02	0.26		
N1 of al_prim_amine	N1 of al_prim_amine	5.50	66.42	66.42	0.36	0.36	**0.04**	0.02	0.09		
N1 of al_prim_amine	O4 of ar_methoxy	3.67	66.42	39.53	0.36	0.36	**0.04**	0.01	0.14		

Intra-molecular H-bond propensities:

Donor	Acceptor	Donor sybyl atom type	Acceptor sybyl atom type	DA Pair constrained connectivity	DA Pair path string	Donor count	**Propensity**	Lower bound	Upper bound	Observed Intra-?
N1	O3	N.3	O.2	0	0	N/A	**0.57**	0.57	0.57	
N1	O1	N.3	O.2	0	0	2	**0.55**	0.55	0.55	
N1	O2	N.3	O.2	0	0	2	**0.55**	0.55	0.55	

Table S6: Hydrogen bond propensity values using GSK training set for experimentally more stable form of GW825964X (X_2947A1)

Donor	Acceptor	Competition	Donor steric density	Acceptor steric density	Donor aromaticity	Acceptor aromaticity	**Propensity**	Lower bound	Upper bound	Frequency	Observed Inter-?
N1 of al_prim_amine	O1 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.16**	0.05	0.44		
N1 of al_prim_amine	O2 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.16**	0.05	0.44		
N1 of al_prim_amine	N1 of al_prim_amine	5.50	66.42	66.42	0.36	0.36	**0.07**	0.03	0.19		
N1 of al_prim_amine	O3 of al_carbonyl_1	2.75	66.42	66.39	0.36	0.36	**0.06**	0.01	0.25		observed
N1 of al_prim_amine	N3 of cyano	2.75	66.42	34.65	0.36	0.36	**0.05**	0.01	0.17		
N1 of al_prim_amine	O4 of ar_methoxy	3.67	66.42	39.53	0.36	0.36	**0.04**	0.01	0.11		

Intra-molecular H-bond propensities:

Donor	Acceptor	Donor sybyl atom type	Acceptor sybyl atom type	DA Pair constrained connectivity	DA Pair path string	Donor count	**Propensity**	Lower bound	Upper bound	Observed Intra-?
N1	O3	N.3	O.2	0	0	N/A	**0.57**	0.57	0.57	
N1	O1	N.3	O.2	0	0	2	**0.55**	0.55	0.55	
N1	O2	N.3	O.2	0	0	2	**0.55**	0.55	0.55	

Table S7: Hydrogen bond propensity values using CSD DS training set for experimentally more stable form of GW825964X (X_2947A1)

Donor	Acceptor	Competition	Donor steric density	Acceptor steric density	Donor aromaticity	Acceptor aromaticity	**Propensity**	Lower bound	Upper bound	Frequency	Observed Inter-?
N1 of al_prim_amine	O3 of al_carbonyl_1	2.75	66.42	66.39	0.36	0.36	**0.17**	0.08	0.34		observed
N1 of al_prim_amine	O1 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.12**	0.05	0.24		
N1 of al_prim_amine	O2 of sulfone	2.75	66.42	63.58	0.36	0.36	**0.12**	0.05	0.24		
N1 of al_prim_amine	N1 of al_prim_amine	5.50	66.42	66.42	0.36	0.36	**0.06**	0.03	0.10		
N1 of al_prim_amine	N3 of cyano	2.75	66.42	34.65	0.36	0.36	**0.05**	0.02	0.12		
N1 of al_prim_amine	O4 of ar_methoxy	3.67	66.42	39.53	0.36	0.36	**0.04**	0.02	0.09		

Intra-molecular H-bond propensities:

Donor	Acceptor	Donor sybyl atom type	Acceptor sybyl atom type	DA Pair constrained connectivity	DA Pair path string	Donor count	**Propensity**	Lower bound	Upper bound	Observed Intra-?
N1	O3	N.3	O.2	0	0	N/A	**0.57**	0.57	0.57	
N1	O1	N.3	O.2	0	0	2	**0.55**	0.55	0.55	
N1	O2	N.3	O.2	0	0	2	**0.55**	0.55	0.55	

Table S8: Hydrogen bond propensity values using GSK + CSD DS training set for experimentally more stable form of GW825964X (X_2947A1)

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Significance code	Lower Bound	Upper Bound
(Intercept)	**1.436**	0.689	2.085	0.0371126	*	0.016	2.742
Donorother	**0.628**	0.439	1.432	0.152161		-0.243	1.481
Acceptoratom_1_of_ar_methoxy	**-1.197**	0.629	-1.902	0.0571287	.	-2.405	0.090
Acceptoratom_2_of_al_carbonyl_1	**2.398**	0.651	3.684	0.000229766	***	1.154	3.731
Acceptoratom_2_of_cyano	**-0.782**	0.651	-1.202	0.229549		-2.031	0.547
Acceptoratom_3(4)_of_sulfone	**1.397**	0.636	2.198	0.0279347	*	0.182	2.701
Acceptorother	**3.328**	0.605	5.500	3.80269e-08	***	2.177	4.579
Competition	**0.049**	0.017	2.844	0.00445772	**	0.016	0.084

Donor_steric_density	**-0.018**	0.003	-6.024	1.7043e-09	***	-0.024	-0.012
Acceptor_steric_density	**-0.049**	0.005	-10.019	1.25875e-23	***	-0.058	-0.039
Donor_aromaticity	**-0.055**	0.585	-0.094	0.925375		-1.204	1.091
Acceptor_aromaticity	**-1.062**	0.505	-2.104	0.0353536	*	-2.057	-0.078
Donoratom_0_of_al_prim_amine	**0.000**	N/A	N/A	N/A	N/A	N/A	N/A
Acceptoratom_0_of_al_prim_amine	**0.000**	N/A	N/A	N/A	N/A	N/A	N/A

Area under ROC curve = 0.933076 (outstanding discrimination)

Table S9: HBP model using GSK training set. Functional group coefficients are considered well represented or significant when they report a significance code of *, ** or ***.

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Significance code	Lower Bound	Upper Bound
(Intercept)	**1.063**	0.546	1.946	0.0516484	.	-0.015	2.133
Donorother	**1.603**	0.415	3.860	0.000113542	***	0.760	2.394
Acceptoratom_1_of_ar_methoxy	**-1.445**	0.491	-2.942	0.00326457	**	-2.431	-0.499
Acceptoratom_2_of_al_carbonyl_1	**-0.159**	0.659	-0.241	0.809251		-1.463	1.131
Acceptoratom_2_of_cyano	**-1.076**	0.540	-1.993	0.046262	*	-2.156	-0.033
Acceptoratom_3(4)_of_sulfone	**0.891**	0.587	1.517	0.129163		-0.267	2.042
Acceptorother	**2.411**	0.476	5.068	4.02378e-07	***	1.460	3.333
Competition	**0.040**	0.021	1.916	0.0553994	.	0.000	0.082
Donor_steric_density	**-0.027**	0.004	-6.691	2.22171e-11	***	-0.035	-0.019
Acceptor_steric_density	**-0.024**	0.007	-3.578	0.000346814	***	-0.037	-0.011
Donor_aromaticity	**-0.266**	0.428	-0.623	0.533579		-1.114	0.565
Acceptor_aromaticity	**-0.473**	0.340	-1.389	0.164858		-1.144	0.191
Donoratom_0_of_al_prim_amine	**0.000**	N/A	N/A	N/A	N/A	N/A	N/A
Acceptoratom_0_of_al_prim_amine	**0.000**	N/A	N/A	N/A	N/A	N/A	N/A

Area under ROC curve = 0.900089 (outstanding discrimination)

Table S10: HBP model using CSD DS training set

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Significance code	Lower Bound	Upper Bound

(Intercept)	**1.301**	0.405	3.212	0.00131674	**	0.498	2.090
Donorother	**1.192**	0.310	3.840	0.000122849	***	0.573	1.791
Acceptoratom_1_of_ar_methoxy	**-1.368**	0.366	-3.734	0.000188754	***	-2.094	-0.654
Acceptoratom_2_of_al_carbonyl_1	**1.315**	0.417	3.153	0.00161586	**	0.494	2.133
Acceptoratom_2_of_cyano	**-1.221**	0.393	-3.111	0.0018634	**	-1.998	-0.456
Acceptoratom_3(4)_of_sulfone	**0.796**	0.391	2.037	0.0416495	*	0.025	1.560
Acceptorother	**2.589**	0.353	7.335	2.21329e-13	***	1.891	3.278
Competition	**0.054**	0.013	4.146	3.38228e-05	***	0.029	0.079
Donor_steric_density	**-0.024**	0.002	-10.913	9.98694e-28	***	-0.029	-0.020
Acceptor_steric_density	**-0.039**	0.004	-11.120	1.00271e-28	***	-0.046	-0.032
Donor_aromaticity	**0.329**	0.322	1.022	0.306908		-0.303	0.961
Acceptor_aromaticity	**-0.646**	0.273	-2.370	0.0178008	*	-1.181	-0.113
Donoratom_0_of_al_prim_amine	**0.000**	N/A	N/A	N/A	N/A	N/A	N/A
Acceptoratom_0_of_al_prim_amine	**0.000**	N/A	N/A	N/A	N/A	N/A	N/A

Area under ROC curve = 0.91745 (outstanding discrimination)

Table S11: HBP model using GSK and CSD DS training sets

X_2947A1

Atom (D/A)	**= 0**	**= 1**	**= 2**	**= 3**
N1 of al_prim_amine (d)	**0.460267**	**0.426638**	**0.104458**	**0.00863675**
N1 of al_prim_amine (a)	**0.956792**	**0.0432076**	**0**	**0**
N3 of nitrile (a)	**0.846256**	**0.151614**	**0.00212938**	**0**
O1 of sulfonyl (a)	**0.917308**	**0.0799743**	**0.00271749**	**0**
O2 of sulfonyl (a)	**0.912385**	**0.084898**	**0.00271749**	**0**
O3 of acyclic_tert_amide (a)	**0.697446**	**0.297075**	**0.00547881**	**0**
O4 of acyclic_ar_ether (a)	**0.934878**	**0.0651223**	**0**	**0**

X_2915A1

Atom (D/A)	= 0	= 1	= 2	= 3

N1 of al_prim_amine (d)	0.472525	0.418655	**0.100988**	0.00783146
N1 of al_prim_amine (a)	**0.966002**	0.0339976	0	0
N3 of nitrile (a)	**0.846256**	0.151614	0.00212938	0
O1 of sulfonyl (a)	0.915777	**0.0815056**	0.00271749	0
O2 of sulfonyl (a)	0.909565	**0.0877177**	0.00271749	0
O3 of acyclic_tert_amide (a)	**0.696703**	0.297786	0.00551186	0
O4 of acyclic_ar_ether (a)	**0.933464**	0.0665365	0	0

Table S12: Coordination Score tables of the two GW825964X polymorphs

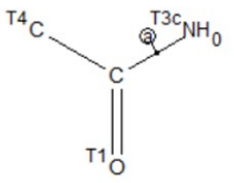
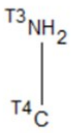
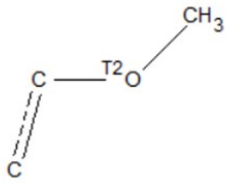
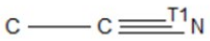
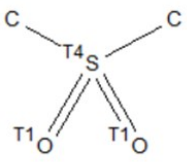
	GW825964X : X_2915A1 and X_2947A1
Search fragments	 al_carbonyl_1  al_prim_amine  ar_methoxy  cyano  sulfone

Table S13: Search fragments used in conquest searches for GW825964X (with two of its polymorphic forms)

Protocol

Classification of hydrogen bonding interactions in polymorphs

For each pair or group of polymorphs the interactions were inspected if the polymorphs have different interacting atoms, they were directly categorized as different. Those that do not have H-bonding interactions were listed as no-interactions, as for those that had the same interacting atoms were inspected in Mercury visually and for their quick view of the H-bond coordination scores (using CSD-Materials > Polymorph Assessment > H-Bond Coordination quick view). Polymorphs that had different H-bonding interactions and/or coordination scores were listed as different, whereas those that had same interacting donor and acceptor atoms and same coordination scores are listed as same interactions.

Assessing crystal morphology in the GSK database

The morphology of the small molecule crystal structures was categorized as 1D, 2D, and 3D according to the following criteria:

1D: rod, needle, flat needle, and fiber
2D: plate, blade, lath, shard, slab, tablet or tabular, thick plate, blocky plate, Triangular plate, irregular plate, platy needle
3D: block, column or columnar, prism, polyhedron, trapezoid, truncated bipyramid, truncated octahedron, triangular prism, prismatic block, irregular block, Block cut from needle, brick, Cut pyramidal block, heart-shaped prism, hexagonal prism,irregular triangular prism, Multi-faceted block, and wedge
In instances where the description was not decisive, the relative dimensions of the crystal in the CIF file were used to determine whether the morphology is 1D (one dimension is much higher than the other two), 2D (two dimensions are much larger than the third), and 3D (three dimensions are approximately similar).
On this basis, two entries (Habit description: Grain) one is classified as 1D and other as 2D.
Two entries one (habit description: Irregular) and other (habit description: colourless) are not considered since the description doesn't not fall into any of the categories: 1, 2, or 3D.