MANCHESTER 1824
The University of Manchester

The University of Manchester Research

Mapping Ligand-Shape Space for Protein–Ligand Systems: Distinguishing Key-in-Lock and Hand-in-Glove Proteins

DOI:
10.1021/acs.jcim.1c00089

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Zarnecka, J., Lukac, I., Messham, S. J., Hussin, A., Coppola, F., Enoch, S. J., Dossetter, A. G., Griffen, E. J., & Leach, A. G. (2021). Mapping Ligand-Shape Space for Protein–Ligand Systems: Distinguishing Key-in-Lock and Hand-in-Glove Proteins. *Journal of Chemical Information and Modeling*, *61*(4), 1859–1874. https://doi.org/10.1021/acs.jcim.1c00089

Published in:
Journal of Chemical Information and Modeling

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact openresearch@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:22. May. 2025

Mapping Ligand Shape Space for Protein-Ligand Systems; Distinguishing Key-in-Lock and Hand-in-Glove Proteins

Joanna Zarnecka[†], Iva Lukac[†], Stephen J. Messham[†], Alhusein Hussin[†], Francesco Coppola[§], Steven J. Enoch[†], Alexander G. Dossetter[+], Edward J. Griffen[+], Andrew G. Leach[,†,+,§]*

† School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool, L3 3AF, UK.

[+] MedChemica Limited, Biohub, Mereside, Alderley Park, Macclesfield, SK10 4TG, UK.

§ Division of Pharmacy and Optometry, School of Health Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT

andrew.leach@manchester.ac.uk

Abstract. Many of the recently developed methods to study the shape of molecules permit one conformation of one molecule to be compared to another conformation of the same or a different molecule: a relative shape. Other methods provide an absolute description of the shape of a conformation that does not rely on comparisons or overlays. Any absolute description of shape can be used to generate a self-organizing map (shape map) that places all molecular shapes relative to one another; in the studies reported here, the shape fingerprint and Ultrafast Shape Recognition methods are employed to create such maps. In the shape maps, molecules that are near to one another have similar shapes and the maps for the 102 targets in the DUD-

E set have been generated. By examining the distribution of actives in comparison to their physical-property-matched decoys, we show that proteins of a key-in-lock type (relatively rigid receptor and ligand) can be distinguished from those that are more of a hand-in-glove type (more flexible receptor and ligand). These are linked to known differences in protein flexibility and binding site size.

Introduction

It has long been known that the shape of a molecule has a profound effect on its biological activity. The shape complementarity between an enzyme and its substrate(s) was the first such link to be made and subsequently led to the expectation that any ligand interacting with a protein is likely to require an appropriate shape; the work of Fischer, in the nineteenth century, was followed by that of Pauling and their ideas remain relevant across many areas of science.[1-4]

There have been many computational methods devised that permit the shape of molecules to be used constructively in applications such as virtual screening for drug discovery. Some provide a description that is absolute i.e. it does not require a query conformation to be used as a comparator and does not require two molecules to be aligned. This includes shape fingerprints,[5] shape signatures,[6] the Ultrafast Shape Recognition (USR) method and its subsequent developments,[7-12] the use of spherical harmonic functions,[13] or processing a meshed version of the surface using the mathematics that considers heat flow (thereby capturing details of curvature).[14] Other approaches provide an assessment of relative shape, that is they allow two conformations to be compared and the similarity of their shapes to be scored. A number of methods use the mathematical simplicity of gaussian functions or other mathematical functions that are able to represent atoms as spherical objects (incorporation of surface

electrostatics and other modelling approaches is also possible).[15-23] Other approaches use a molecular surface comparison, for instance by projecting key points to create a graph representation.[24, 25] Alternatively, shape can be encoded less directly using methods that encode discrete distances between features, in a similar fashion to pharmacophore triplets or related methods; those features present in two conformations can be compared to give a relative shape similarity.[26-29] It is also possible to bring in information about the protein binding site to complement the information available from analyzing the shape of ligands and to use shape overlays with a bound ligand as a means to place other ligands in a protein binding site.[30, 31] The developers of the Pubchem database have showcased many of the insights that can be provided by considering molecular shape through their Pubchem3D program that includes making connections by shape (that complement those based on chemical structure) and linking molecules that share biological activity.[32-35] They also studied the diversity of molecular shapes represented by the Pubchem database and reveal a surprisingly limited number of clusters can represent all shapes and that when grouped according to their volume, molecules with higher volumes do not correspond to larger numbers of shape clusters than smaller volumes (for all apart from the smallest molecules).[33] In principle, any of the methods for assessing the relative shape of molecules can be converted into an absolute description of shape by comparison to a complete set of reference shapes, such as that required to create shape fingerprints.[5]

An absolute description of the shapes of a set of molecules that share biological activity should provide a distribution that can be visualized and assessed to deduce the shapes that are preferred for that activity and how tight such shape preferences are. In this way, it should be possible to visualize whether the system obeys the key-in-lock (tight conformational requirement for the ligand, rigid receptor) or hand-in-glove (the ligand can adopt a range of shapes, flexible receptor) principle.[36, 37] This would provide useful insights for those attempting to understand

the action of enzymes and therefore help in the design of new enzymes. It is also important for the design of molecules that bind to a given protein binding site such as occurs in drug discovery and in chemical biology more generally.

In the following, we show how large sets of known actives against a set of protein targets are distributed in the shape space defined by two absolute descriptions of molecular shape: shape fingerprints and USR. These distributions reveal that insights concerning the nature of the protein target can be obtained – in a way that does not need the protein's structure to be known. In order to generate shape fingerprints, we obtain a set of reference shapes (a shape database) and find optimum settings for grouping compounds with shared biological activity. We also show how they can be combined with other fingerprinting methods to achieve scaffold hopping.

Results and Discussion

Two absolute descriptions of molecular shape have been used to visualize the distribution of ligand shapes for all of the Database of Useful Decoys - Enhanced (DUD-E) set of protein targets: shape fingerprints and USR. The DUD-E sets include a large set of decoys accompanying a curated set of known actives for 102 different protein targets.[38] These include 50 decoys for each active with the intention that the distribution of the physical properties (molecular weight, MilogP, number of rotatable bonds, hydrogen bonding group counts and charge) of the decoys should match that of the actives.

The shape fingerprint approach was introduced in 2005 based on two sets of reference shapes: the first derived from the Cambridge Structural Database of small molecule crystal structures and the second from structures of the MDL Drug Data Report transformed into three-

dimensional conformations by Corina.[5, 39, 40] The resulting fingerprints were benchmarked against alternative methods for computing shape similarity for two molecules (ROCS and shape multipoles); the ordering according to shape similarity achieved with the different methods is compared. The ability of the method to correctly group molecules that share biological similarity was not demonstrated directly. As described in full in the supporting information, we optimized a set of shape fingerprints and the settings that determine how they are applied. In brief, the reference shapes were derived from the Ligand Expo database of ligands bound to proteins in the Protein Data Bank.[41] Following filtering and picking of representatives (as described in the original publication), a set of 929 reference shapes was obtained. This is itself a remarkable finding that suggests that all of the shapes adopted by ligands binding to proteins can be described as being similar to one of less than a thousand types. A benchmarking exercise using two datasets of known bioactive conformations gave optimum settings for generating and applying shape fingerprints that use these reference shapes.[42, 43] When these were used to probe recall by active compounds of other actives in preference to decoys for 8 targets from the DUD-E diverse subset, the fingerprints were found to perform poorly compared to other methods of shape comparison and thus we do not recommend shape fingerprints for this application. Some interesting results obtained during the optimization of the fingerprints are noted at the end of this paper. However, the property that is of most interest to us is the ability to visualize the distribution of shapes that are tolerated by proteins and this requires an absolute description of shape that can be mapped. Shape fingerprints provide us with an absolute description (in this case a sequence of 929 binary 0s or 1s). The USR method also summarizes the shape but as a 12-dimensional descriptor and it has also been applied to the same task.[12, 44] The distributions obtained and the insights they provide are described below.

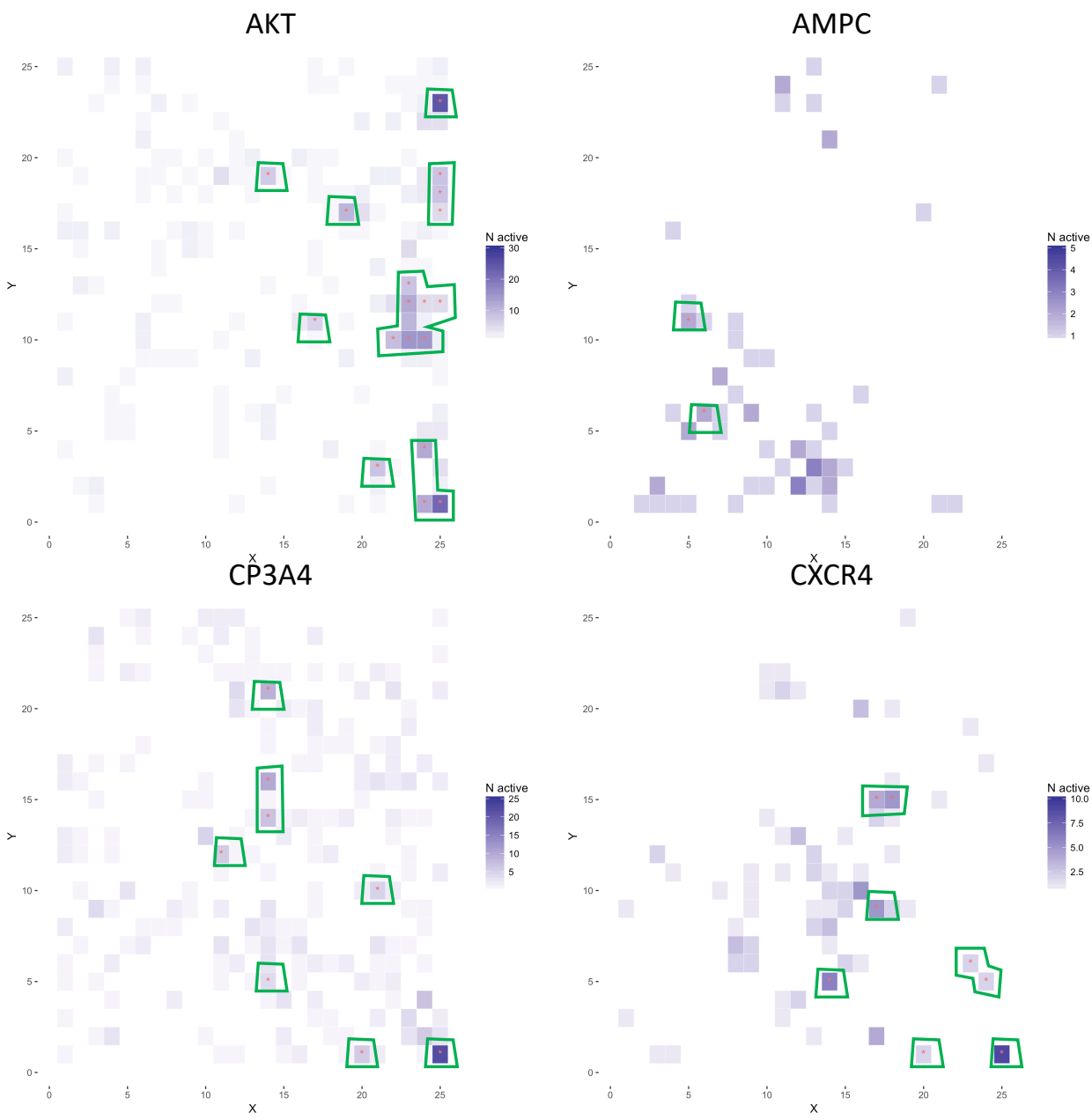
Shape fingerprint-based shape maps.

In the following, we examine the distribution of large sets of compounds with shared biological activity in shape maps provided by the shape fingerprints. These visualize whether the target in question binds ligands with a limited or wide repertoire of shapes and whether each shape is tightly or loosely defined.[36, 37] To enable the distribution of actives and decoys to be compared, the coordinate set provided by the shape fingerprint can be reduced to a two-dimensional map.[45, 46] This required the fingerprint to be folded into a numerical array small enough to be processed by the self-organising map software SOMbrero in R.[47-50] This was achieved by summing every m^{th} bit (m=5, 10, 15, 20, 25, 30). These can then be reduced to a two-dimensional array of n x n coordinates (n = 5, 10, 15, 20, 25) by the self-organising map algorithm, using default settings. The upper limit of m and n was selected to be as high as can be performed on a desktop machine within a working day. When the DUD-E set of actives and decoys is projected into an *n* x *n* map, any coordinates that are statistically enriched in actives can be detected by performing a chi-squared test to compare the number of actives and decoys at each coordinate to the prevalent numbers of actives and decoys for that target. A p-value cutoff of 1% is used to detect enriched coordinates (a Bonferroni correction is used to account for the increased number of tests being performed as the grid size, n, changes).[51] The mapping was created with all compounds for the 8 biological targets in the DUD-E diverse set present such that the coordinates are comparable across the different targets. The values of m and n that maximize the proportion of actives found at enriched coordinates were selected and this is m=20, n=25. The enrichment level at enriched coordinates is large; whereas actives make up 3.2 % of the overall set, at the enriched coordinates in the maps, they make up 23.1 %.

The maps in Figure 1 show the number of active compounds present at each coordinate and those coordinates that are computed to include a statistically significant enrichment of actives

compared to decoys are identified with a red star. Whenever groups of coordinates that are within 3 steps (where one step involves increasing or decreasing x or y by 1) are enriched, these have been highlighted with green borders and are referred to in this section as groups. This paints a picture of the different tolerances of the targets to variation in ligand shape. AKT (serine/threonine kinase AKT1) binders are divided amongst four main shape groups (with four small additional isolated groups), all of which contain a range of different chemotypes (chemical structures are provided in supporting information Section S7). There is a low occurrence of actives outside the enriched regions. AMPC (beta-lactamase) has few actives but two shape groups are enriched in actives and this includes two chemotypes (sulfonamides and a phthalimide) with each group including only one chemotype. Many of the actives for AMPC fall away from an enriched group. CP3A4 (cytochrome P450 3A4) has a few enriched coordinates but these are not generally near one another and there is a background spread of actives across the map. All except one enriched coordinate contain a range of chemotypes. CXCR4 (C-X-C chemokine receptor type 4) has six enriched groups, each of which contains only one chemotype. GCR (glucocorticoid receptor) features six enriched groups each representing more than one chemotype but there is a significant spread of active compounds across the map. HIVPR (HIV type 1 protease) has its actives predominantly in one corner of the map representing many chemotypes; there are few actives away from the enriched groups. HIVRT (HIV type 1 reverse transcriptase) features nine enriched groups, six of which include more than one chemical series type. There are many actives distributed across the map. KIF11 (kinesin-like protein 1) has actives predominantly in two groups in the map near to the six enriched groups that are close together. Some of the active regions for one target in these maps overlap with those from other targets.

These results, that include only information relating to the shape of ligands, were tentatively linked to the known properties of the protein targets with the expectation that, in general, proteins with small and rigid binding sites will be more likely to act like a key-in-lock system while large and flexible binding sites are more like hand-in-glove. Key-in-lock proteins might be expected to have actives at higher proportions at enriched coordinates than do hand-in-glove. This is consistent with the observation of only minor structural variation in the CXCR4 structure that has been implicated in the success of structure-based design in GPCRs.[52] CP3A4 has a binding site that is known to vary greatly in size and shape,[53] HIVRT is often used as an exemplar for a flexible protein,[54] and GCR has a ligand binding domain that has been found to be able to respond to a range of different ligands in such a way that has led to this site being described as inducible by the ligands.[55] It could also be that the electrostatic properties of the surfaces of ligands can override shape preferences for the proteins that behave as hand-in-glove. This is a rather profound insight into the nature of a protein and could be derived solely from screening of a large set molecules for activity against a new target e.g. after high throughput screening. The likely tolerance of the protein to incorrect shapes can be assessed and the drug discovery strategy selected accordingly. These observations have been expanded on using the faster USR method.


AKT
AMPC
CP3A4
CXCR4
N active
X
Y

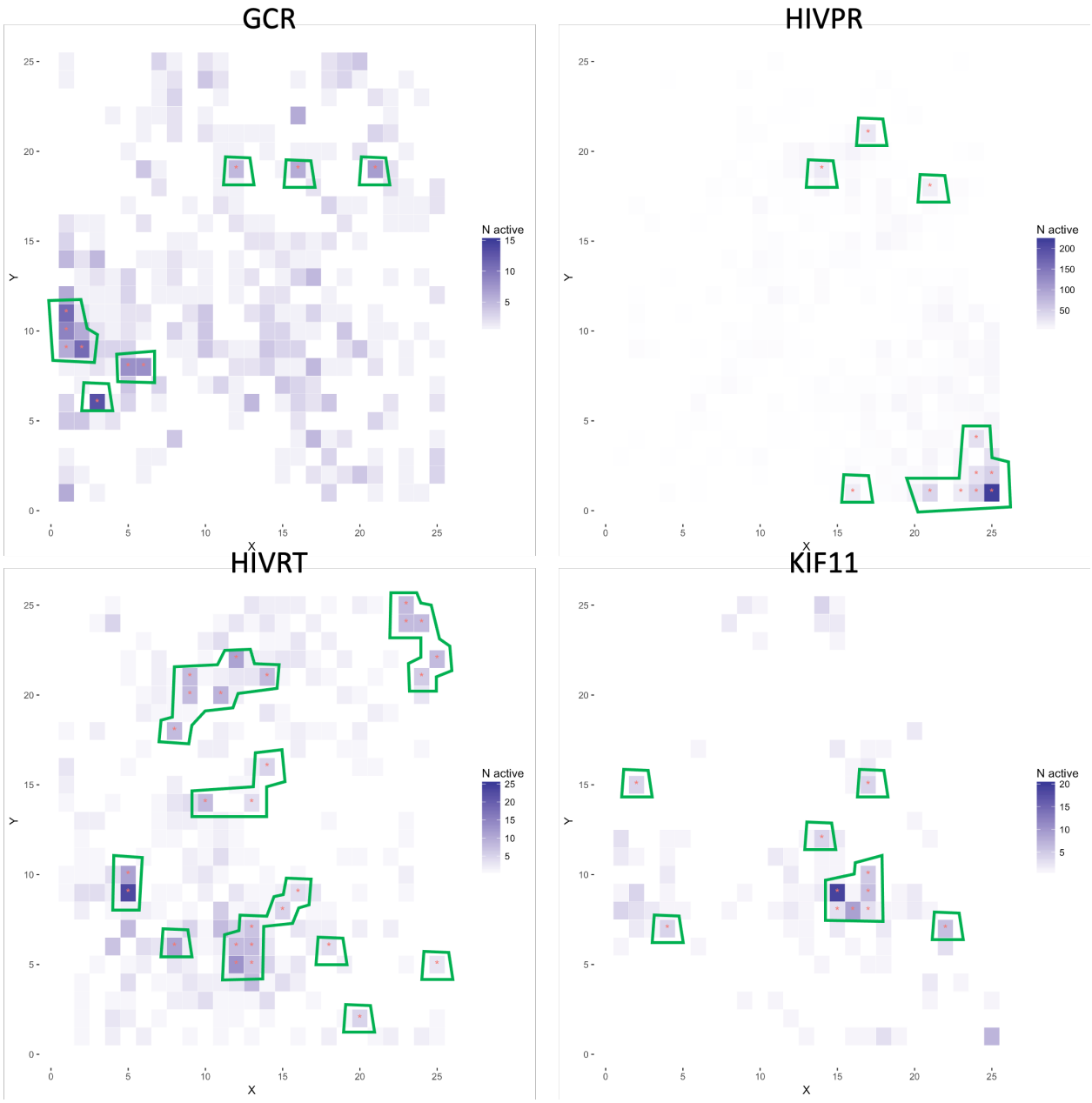


Figure 1. Self-organized maps of the DUD-E sets of actives for the targets indicated. The maps were generated from shape fingerprints. Coordinates are colored shades of blue according to the number of actives at that coordinate, as indicated by the scale. Points found to be enriched in actives are indicated with a red star and such points that are within 3 steps of another enriched point are grouped together inside green lines.

The visualization in Figure 1 can be formalized by performing a clustering using the fingerprints. In this case, the Taylor-Butina cluster algorithm was employed in which clusters are defined by a "seed" molecule that acts as the cluster centroid; the fingerprint with the largest list of neighbors is selected as the first seed.[56, 57] Subsequently, all members of the cluster defined by the first seed are removed and the remaining compound with the largest list of neighbors taken to define the next cluster. This is repeated until all compounds are assigned to a cluster. Such a grouping is greatly facilitated by the external space defined by the reference shapes that allows the placing of each shape relative to every other shape. The clustering employed the shape fingerprints and a range of fingerprint Tanimoto cut-offs spanning 0.2 – 0.7.[57] Each set of actives and its curated set of decoys were clustered and those clusters enriched with actives identified. Enrichment was again detected using a chi-squared test with a p-value cutoff of 0.01 and with a Bonferroni correction depending upon the number of clusters (i.e. if there are 10 clusters for a particular protein target then a p-value cutoff of 0.001 would be used to select likely enriched clusters).[51] This reveals that for all of the targets in the DUD-E diversity set, the best cut-off value for clustering by shared biological activity is 0.55 (Figure 2). This is a good default value to use in clustering molecules when employing these shape fingerprints when the intention is to group together compounds that share biological activity. On average, at this setting, clusters that are statistically enriched in active compounds account for 61 % of the actives suggesting that shape alone is a sufficiently defining feature for a majority of active compounds. These clusterings also indicate one of the reasons that shape might provide low levels of recall of actives from a set of decoys using shape alone: it is perfectly feasible for a target to bind to a range of acceptable shapes (each being an active cluster) and thus any single shape query would be intrinsically incapable of retrieving certain active compounds belonging to a different (but also acceptable) shape cluster.

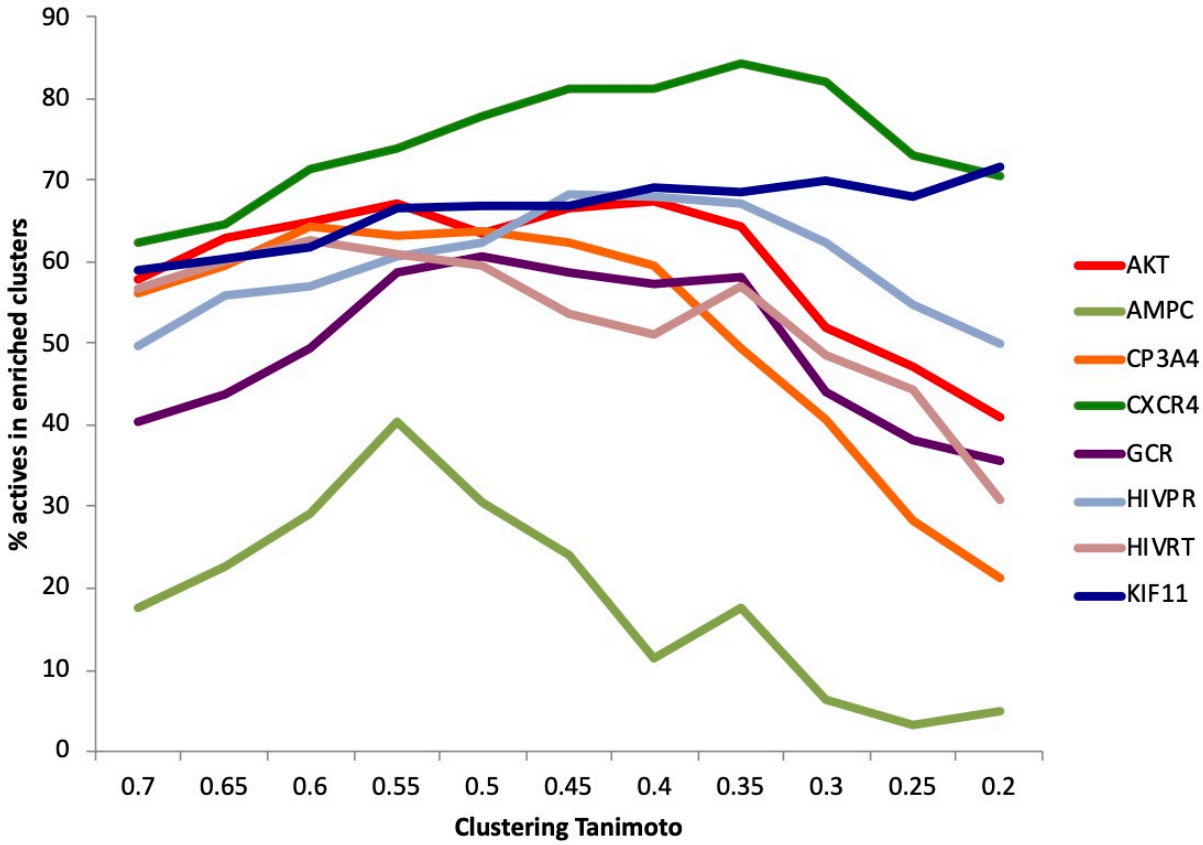


Figure 2. The variation in the percentage of active compounds that are found in enriched clusters as the Tanimoto used to define cluster sizes is varied.

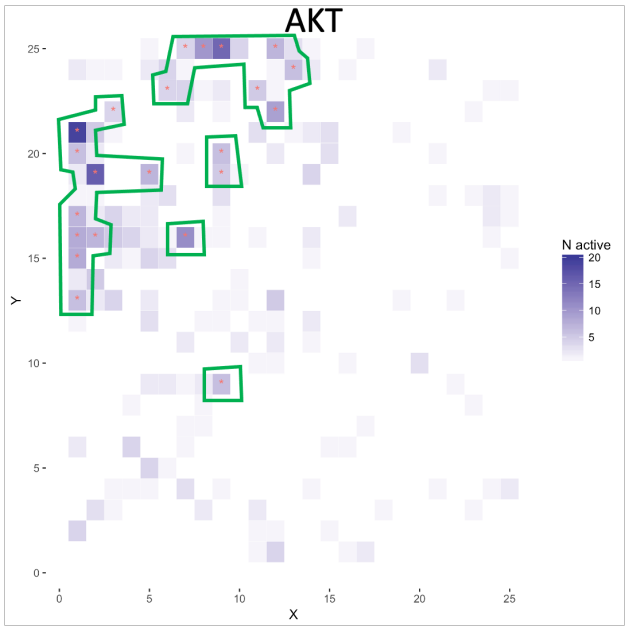
Ultrafast Shape Recognition-based shape distribution maps

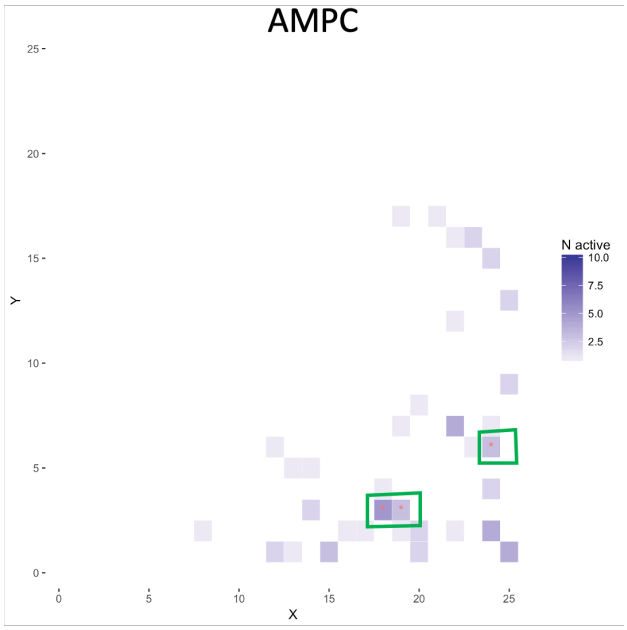
In order to expand the insights available by shape alone, an alternative method for obtaining an absolute description was applied – the Ultrafast Shape Recognition method.[11,12] This method describes each shape as a twelve-element vector and provides an alternative way to generate the maps in Figure 1. The RDkit implementation of the method was used to generate a set of USR descriptors for the same set of ligands (both actives and decoys).[44] These were then projected into two-dimensional 25 x 25 maps using the SOMbrero algorithm. These are shown in Figure 3 and present broadly similar overviews to those in Figure 1 although that for HIVPR

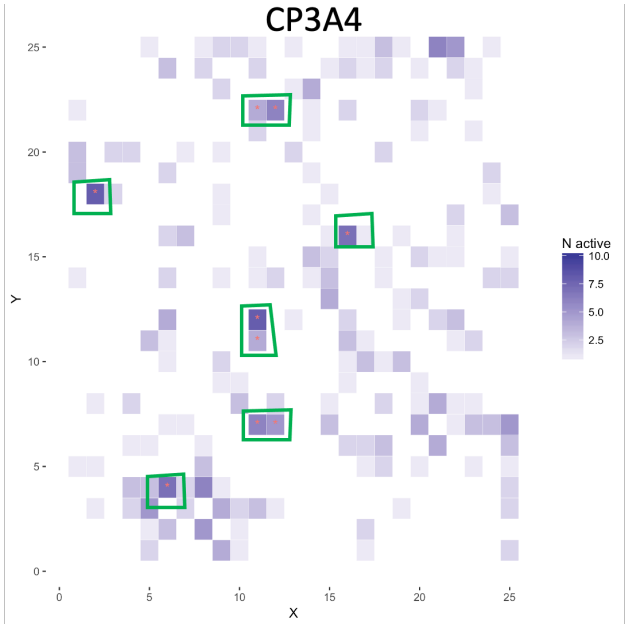
shows a wider distribution of actives across the map and this may be because many of the actives for this target fall outside the molecular weight range for which the shape fingerprints are appropriate (200 – 500 Da). The levels of enrichment achieved in the USR maps is sometimes higher and sometimes lower than that achieved with the fingerprint-derived maps (Table 1).

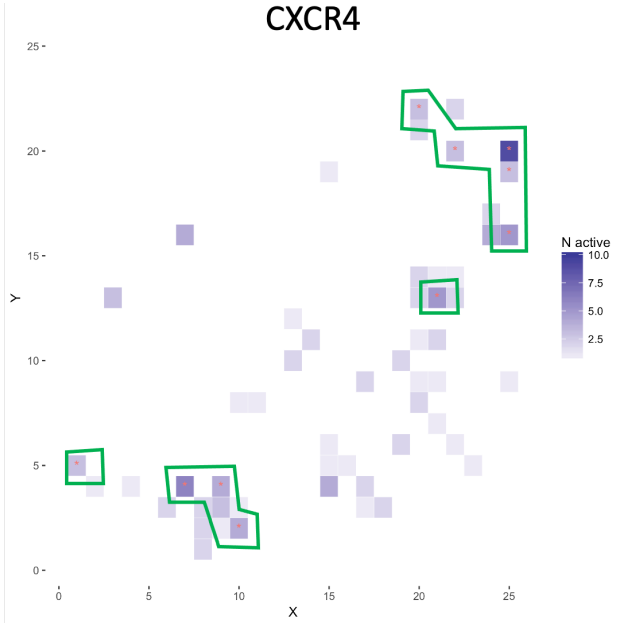
Table 1. The percentage of actives placed at enriched coordinates in shape maps generated by either shape fingerprints or USR are compared.

Target	% of actives at enriched coordinates	
	Shape fingerprints	USR
AKT	45.4	41.6
AMPC	6.5	17.7
CP3A4	19.0	15.2
CXCR4	28.7	36.9
GCR	14.9	30.9
HIVPR	38.3	23.5
HIVRT	33.8	32.6
KIF11	34.0	33.0
Overall	23.1	21.5


AKT
N active
20
15
10
5
X
Y


AMPC
N active
10.0
7.5
5.0
2.5
X
Y


CP3A4
N active
10.0
7.5
5.0
2.5
X
Y


CXCR4
N active
10.0
7.5
5.0
2.5
X
Y

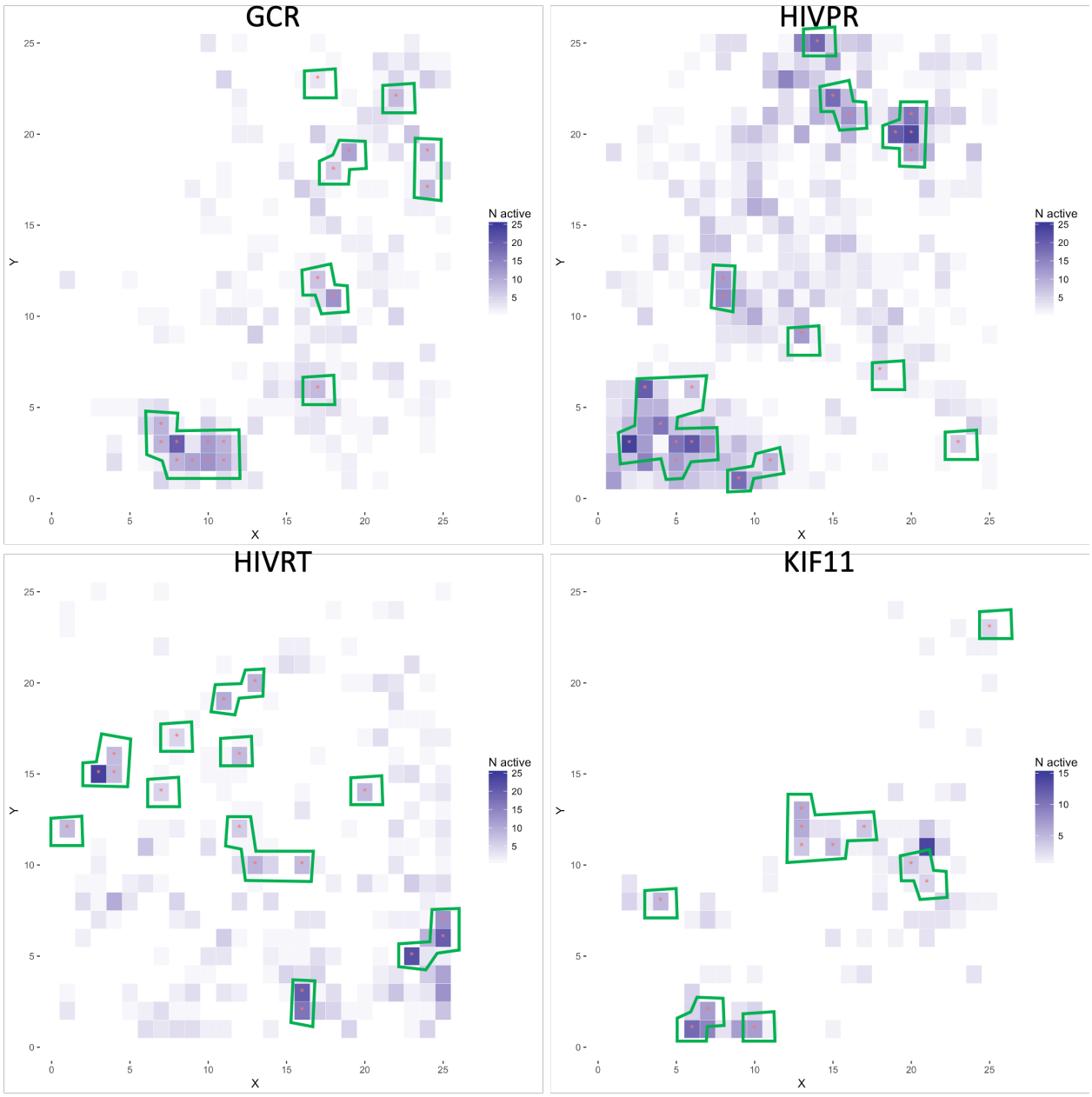


Figure 3. Self-organized maps of the DUD-E sets of actives for the targets indicated. The maps were generated from USR descriptors. Coordinates are colored shades of blue according to the number of actives at that coordinate, as indicated by the scale. Points found to be enriched in actives are indicated with a red star and such points that are within 3 steps of another enriched point are grouped together inside green lines.

The USR-based SOMs can be generated more rapidly than those using shape fingerprints. For instance, generating the shape fingerprints for the DUD-E diverse set took many days even though the tasks were split across multiple processors whereas generating the USRs can be achieved in minutes on a single processor machine. Hence, the USR method was chosen to apply to the full set of 102 DUD-E targets. Whereas in the previous section, a single SOM was created with all compounds for all targets projected at once, here the set of actives for each biological target and its matched set of decoys was processed individually to create a series of 25 x 25-dimensional maps. Two summary values for each map were obtained. The first is the percentage of all active compounds that are found at enriched coordinates. The other is the degree of enrichment that is found across all enriched coordinates (the percentage of actives at enriched coordinates compared to the percentage of actives in the whole set). These are plotted against one another in Figure 4. Targets at the bottom left of this plot include FGFR (Fibroblast growth factor receptor 1, for which there were no enriched coordinates), CDK2 (Cyclin-dependent kinase 2), DPP4 (Dipeptidyl peptidase IV), DHI1 (11-beta-hydroxysteroid dehydrogenase 1), ANDR (Androgen Receptor) and CP3A4 (Cytochrome P450 3A4), the need to incorporate protein flexibility in order to understand and predict ligand binding has been emphasized for each of these proteins in at least one exemplar study.[53, 58-62] Biological targets towards the top right of this plot include COMT (Catechol O-methyltransferase), PA2GA (Phospholipase A2 group IIA), PYGM (Muscle glycogen phosphorylase), THB (Thyroid hormone receptor beta-1) and PPARD (Peroxisome proliferator-activated receptor delta). Similar example papers stressing flexibility of the protein could not be found for these targets. Docking studies support the lack of structural variation for THB (although cast doubt on this for PA2GA).[63] The necessity of creating inhibitors that are pre-organized into the correct conformation has been emphasized for COMT.[64] The conformational specificity of binding to the different PPAR receptor types has been explored with crystallography.[65]

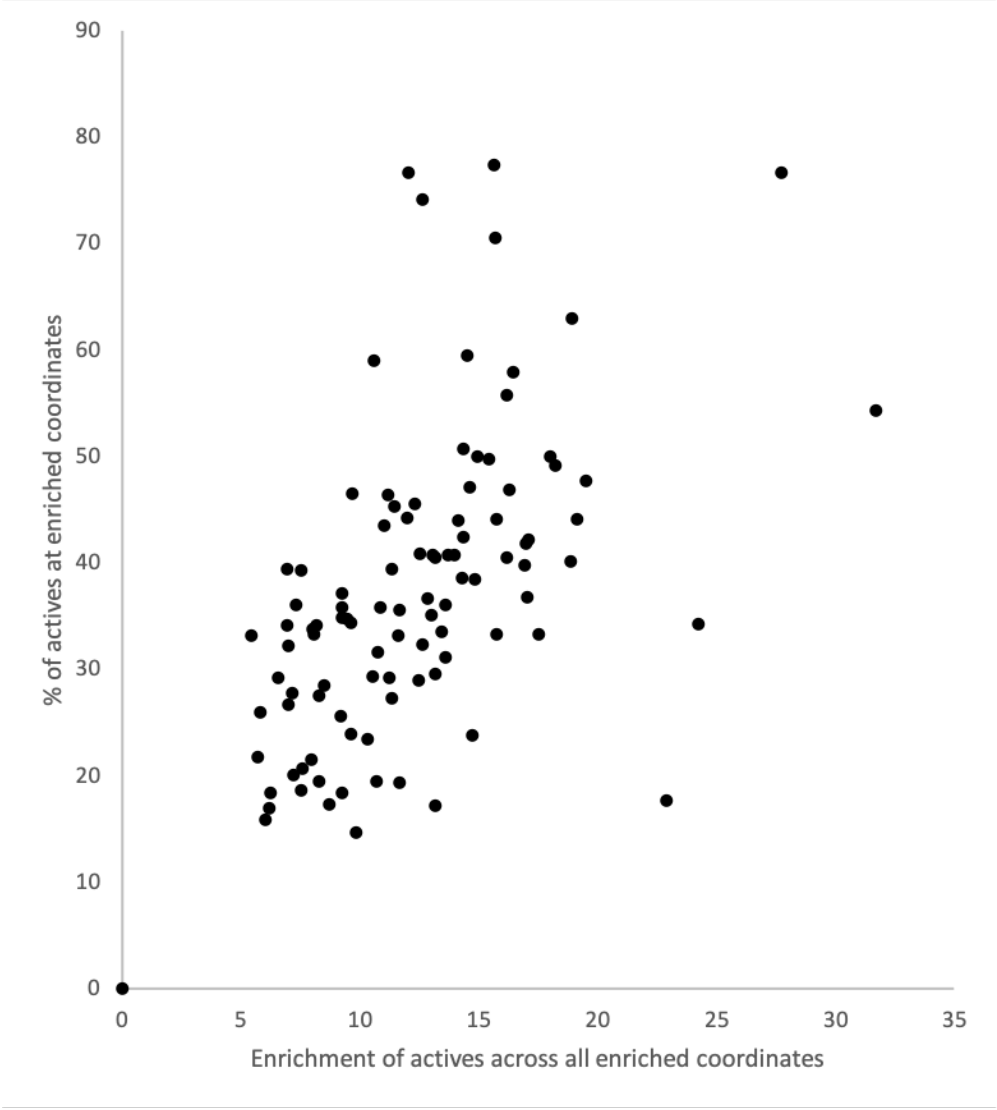


Figure 4. The percentage of all active compounds that are found at enriched coordinates is plotted against the average enrichment level found for those coordinates for the set of targets in the DUD-E database set of biological targets.

In order to establish that the classifications provided by the shape distributions correspond in a more general way to known properties of each biological target, we have surveyed each of the 102 targets in DUD-E using: bibliometric approaches, sequence-based classifications of flexibility and analysis of a representative crystal structure. In particular, these studies have

sought to identify whether certain targets are more flexible than others and whether certain targets have much larger binding sites than others.

For the bibliometric analysis, each target's name was entered as a "Topic" search term in the Web of Science.[66] The total number of entries returned was then recorded as N(papers) in Table 2. Three searches of this set were then performed for the extra terms "induced fit", "flexible" and "rigid". While the number of records returned for each is recorded in Table 2, it was clear than the latter two terms are much more likely to return entries that are not relevant whereas "induced fit" is more specific to whether the protein adapts to its bound ligand. The percentage of the records for each target that contain this term is then computed (Table 2, % Induced Fit). This provides the distribution shown in Figure 5. Unsurprisingly, there is not a simple correlation of the bibliometric values with the ligand shape-derived metric but it is clear that among those 16 targets for which more than 0.2% of their references refer to "induced fit", 15 (94 %) have distributions of ligand shapes that place 50 % or less of active compounds at enriched coordinates. Meanwhile, among those proteins where more than 50 % of active compounds are at enriched coordinates, 10 out of 11 (91 %) have less than 0.2 % of references referring to induced fit.

Table 2. Summary values for each DUD-E target for shape-map-derived values, bibliometry, sequence analysis and structural analysis. Rows above the division in the table are more hand-in-glove-like while those below are more key-in-lock-like. Figures in bold indicate those values that are on the more flexible or larger side of the cutoff values described in the text.

DUD-E code	Description (used for literature search)	Shape map summary		Bibliometry					Sequence analysis (FoldUnfold)				PDB code	Structural analysis			
		Enrichment	Recall %	N(papers)	N(flexible)	N(induced fit)	N(rigid)	% induced fit	Sequence length	N(unfolded regions)	N(unfolded AAs)	% sequence unfolded		Per-residue averaged B-factors	Per-residue average deviation	Active site B factor (sc-PDB)	Binding site volume ($Å^3$)
FGFR1	Fibroblast growth factor receptor 1	0.0	0.0	18910	27	16	16	0.08	290	2	22	7.59	3c4f	32.5	6.35		
PGH1	Cyclooxygenase-1	9.8	14.7	2893	3	5	1	0.17	599	2	35	5.84	2oyu	**52.9**	3.32	50.1	**1225**
CDK2	Cyclin-dependent kinase 2	6.0	15.9	10070	37	14	14	0.14	283	0	0	0.00	1h00	35.5	**11.11**		
DPP4	Dipeptidyl peptidase IV	6.2	17.0	6977	10	3	9	0.04	766	0	0	0.00	2i78	42.2	**8.78**		
AOFB	Monoamine oxidase B	13.2	17.3	5806	27	10	13	0.17	520	1	21	4.04	1s3b	38.2	3.96		
UROK	Urokinase-type plasminogen activator	8.7	17.3	5631	24	3	6	0.05	431	3	43	9.98	1sqt	12.5	5.40	8.1	520
AMPC	Beta-lactamase	22.9	17.7	28348	107	23	41	0.08	377	2	22	5.84	1l2s	23.1	5.14	22.7	722
MK14	MAP kinase p38 alpha	6.2	18.4	3948	15	8	2	**0.20**	285	0	0	0.00	2qd9	36.0	**8.66**	29.6	442
CAH2	Carbonic anhydrase II	9.2	18.5	5736	34	7	32	0.12	260	4	55	**21.15**	1bcd	13.2	5.42		
DHI1	11-beta-hydroxysteroid dehydrogenase 1	7.5	18.7	3648	1	3	0	0.08	292	0	0	0.00	3frj	**58.2**	**10.14**		
CASP3	Caspase-3	11.7	19.4	66793	44	45	12	0.07	277	4	59	**21.30**	2cnk	13.6	3.85		
PTN1	Tyrosine phosphatase 1B	10.7	19.6	2995	20	3	1	0.10	435	4	85	**19.54**	2azr	29.3	7.75		

CP3A4	Cytochrome P450 3A4	8.3	19.6	5000	38	20	9	**0.40**	503	4	47	9.34	3nxu	36.5	6.08	31.2	**2376**
ANDR	Androgen Receptor	7.2	20.1	43082	80	44	20	0.10	920	11	293	**31.85**	2am9	25.4	4.86	21.7	290
HIVPR	Human immunodeficiency virus type 1 protease	7.6	20.7	3611	34	8	9	**0.22**	99	0	0	0.00	1xl2	16.4	5.26	15.4	925
MMP13	Matrix metalloproteinase 13	7.9	21.5	5695	9	4	5	0.07	471	4	62	13.16	830c	13.4	4.08		
ACE	Angiotensin-converting enzyme	5.7	21.8	47417	63	21	23	0.04	1306	6	74	5.67	3bkl	30.0	8.43		
LCK	Kinase LCK	10.3	23.4	2346	8	3	3	0.13	254	0	0	0.00	2of2	18.8	6.74	14.8	500
PLK1	Protein kinase PLK1	14.7	23.9	1281	2	0	0	0.00	253	1	12	4.74	2owb	24.5	6.69	20.2	621
BACE1	Beta-secretase 1	9.6	23.9	2145	24	8	6	**0.37**	501	1	14	2.79	3l5d	18.9	6.28	16.8	716
PRGR	Progesterone receptor	9.2	25.7	42576	51	44	15	0.10	933	15	356	**38.16**	3kba	34.2	2.83		
AA2AR	Adenosine A2a receptor	5.8	25.9	1937	5	1	5	0.05	412	6	92	**22.33**	3eml	**62.1**	8.22		581
PGH2	Cyclooxygenase-2	7.0	26.7	35113	47	30	17	0.09	604	4	65	10.76	3ln1	**55.8**	5.94	45.6	**1218**
ALDR	Aldose reductase	11.3	27.3	7312	32	31	15	**0.42**	316	1	20	6.33	2hv5	17.1	4.48	18.4	425
GCR	Glucocorticoid receptor	8.3	27.5	34880	59	58	15	0.17	777	11	216	**27.80**	3bqd	**73.9**	**16.60**	62.5	486
TRY1	Trypsin I	7.2	27.8	6395	68	18	48	**0.28**	247	3	46	**18.62**	2ayw	8.8	1.30		
EGFR	Epidermal growth factor receptor erbB1	8.5	28.5	68946	133	48	40	0.07	268	1	12	4.48	2rgp	49.5	**12.37**		
MK10	c-Jun N-terminal kinase 3	12.4	29.0	5088	5	1	1	0.02	296	0	0	0.00	2zdt	33.5	7.29	28.8	533

IGF1R	Insulin-like growth factor I receptor	11.2	29.2	14323	9	8	8	0.06	276	1	12	4.35	2oj9	36.3	**10.21**	28.4	716
PARP1	Poly [ADP-ribose] polymerase-1	6.6	29.2	3461	5	4	1	0.12	1014	14	304	**29.98**	3l3m	42.0	9.15	27.8	867
CSF1R	Macrophage colony stimulating factor receptor	10.5	29.4	8916	9	8	3	0.09	329	2	40	12.16	3krj	**57.3**	**16.26**		
GRIK1	Glutamate receptor ionotropic kainate 1	13.2	29.6	571	1	1	1	0.18	918	7	106	11.55	1vso	23.2	4.92	18.3	850
ABL1	Kinase ABL	13.6	31.2	9406	22	10	5	0.11	252	0	0	0.00	2hzi	22.5	5.28	19.8	419
CP2C9	Cytochrome P450 2C9	10.7	31.7	1709	12	7	4	**0.41**	490	0	0	0.00	1r9o	40.3	**12.11**	35.2	675
ADA17	ADAM17	7.0	32.2	1743	2	2	1	0.11	824	13	197	**23.91**	2oi0	31.0	**8.94**		
DEF	Peptide deformylase	12.6	32.3	661	2	4	1	**0.61**	169	0	0	0.00	1lru	29.0	**9.98**	19.3	419
FNTA	Protein farnesyltransferase type I alpha subunit	5.5	33.2	48	n.c.	n.c.	n.c.	n.c.	379	3	75	**19.79**	3E37	22.4	4.46		
MCR	Mineralocorticoid receptor	11.6	33.2	8685	14	15	6	0.17	984	12	351	**35.67**	2aa2	37.0	7.15	27.0	334
ESR2	Estrogen receptor beta	8.1	33.3	29720	54	44	15	0.15	530	6	109	**20.57**	2fsz	39.2	**9.53**	32.8	415
FAK1	Focal adhesion kinase 1	17.5	33.3	7040	8	7	18	0.10	259	0	0	0.00	3bz3	41.0	5.67		
KIT	Stem cell growth factor receptor	15.7	33.3	19205	26	14	7	0.07	349	1	11	3.15	3g0e	19.7	5.46	15.2	422
HDAC2	Histone deacetylase 2	13.5	33.6	10210	17	6	4	0.06	488	3	108	**22.13**	3max	35.7	6.17		
ESR1	Estrogen receptor alpha	8.0	33.8	35986	98	66	28	0.18	595	5	99	**16.64**	1sj0	40.4	**12.72**	35.1	645

HIVRT	Human immunodeficiency virus type 1 reverse transcriptase	8.1	34.1	5751	27	12	12	**0.21**	566	3	42	7.42	3lan	**75.3**	**11.32**	69.1	540
PDE5A	Phosphodiesterase 5A	6.9	34.1	219	0	0	0	0.00	875	10	172	**19.66**	1udt	35.7	7.62	36.6	908
PYRD	Dihydroorotate dehydrogenase	24.3	34.3	1106	6	0	2	0.00	395	6	84	**21.27**	1d3g	30.5	**11.90**	23.1	560
VGFR2	Vascular endothelial growth factor receptor 2	9.6	34.4	17618	28	31	11	0.18	329	1	30	9.12	2p2i	31.6	**12.31**	24.8	651
HMDH	HMG-CoA reductase	9.5	34.8	11057	17	3	4	0.03	888	7	91	10.25	3ccw	**56.4**	5.46		
ADRB1	Beta-1 adrenergic receptor	9.2	34.9	5016	3	12	9	**0.24**	477	5	146	**30.61**	2vt4	38.8	**11.41**	32.0	452
MP2K1	Dual specificity mitogen-activated protein kinase kinase 1	13.0	35.1	518	0	0	0	0.00	294	1	20	6.80	3eqh	**57.9**	**13.08**		
ITAL	Leukocyte adhesion glycoprotein LFA-1 alpha	11.7	35.6	100	0	0	0	0.00	1170	9	154	13.16	2ica	14.1	3.53		
DYR	Dihydrofolate reductase	9.2	35.9	10377	238	51	65	**0.49**	187	0	0	0.00	3nxo	16.0	5.97	13.8	466
FKB1A	FK506-binding protein 1A	10.8	35.9	12	n.c.	n.c.	n.c.	n.c.	108	2	24	**22.22**	1j4h	17.1	2.77	16.4	270
LKHA4	Leukotriene A4 hydrolase	13.6	36.1	313	2	2	2	**0.64**	611	1	13	2.13	3chp	**53.7**	**10.91**		
FA10	Coagulation factor X	7.3	36.1	5433	24	7	14	0.13	488	5	81	**16.60**	3kl6	14.3	4.15		
BRAF	Kinase B-raf	12.9	36.7	349	2	0	1	0.00	261	0	0	0.00	3d4q	**58.1**	**10.74**		

NOS1	Nitric-oxide synthase, brain	17.1	36.8	16753	10	12	8	0.07	1434	11	230	**16.04**	1qw6	43.3	**13.34**		
ADRB2	Beta-2 adrenergic receptor	9.3	37.1	11675	43	41	28	**0.35**	413	4	72	**17.43**	3ny8	48.7	8.24		**1188**
HDAC8	Histone deacetylase 8	14.8	38.5	2631	8	4	0	0.15	377	2	25	6.63	3f07	**69.9**	7.93		
JAK2	Protein kinase JAK2	14.3	38.6	3399	7	3	4	0.09	265	0	0	0.00	3lpb	35.4	**10.13**	31.7	**1141**
SRC	Protein kinase SRC	7.5	39.4	20904	62	22	21	0.11	254	0	0	0.00	3el8	**56.2**	**13.83**	51.4	702
AKT1	Kinase AKT	11.3	39.5	3081	3	0	0	0.00	259	1	17	6.56	3cqw	42.1	7.37	35.2	**1083**
THRB	Thrombin	6.9	39.5	57442	213	81	138	0.14	622	6	92	14.79	1ype	19.2	6.14	16.5	763
TRYB1	Tryptase beta-1	16.9	39.8	76	n.c.	n.c.	n.c.	n.c.	275	2	26	9.45	2zec	30.0	5.96		
MET	Hepatocyte growth factor receptor	18.8	40.2	8888	11	7	6	0.08	268	0	0	0.00	3lq8	34.5	**9.47**		
AKT2	Kinase AKT2	16.2	40.5	167	0	0	0	0.00	258	1	17	6.59	3d0e	23.6	2.70		
FA7	Coagulation factor VII	13.1	40.5	4852	6	1	7	0.02	466	7	114	**24.46**	1w7x	14.4	3.46		
GRIA2	Glutamate receptor ionotropic, AMPA 2	13.7	40.7	985	1	1	2	0.10	883	5	85	9.63	3kgc	16.9	5.06	11.9	**1596**
MAPK2	MAP kinase-activated protein kinase 2	14.0	40.8	283	0	0	0	0.00	262	0	0	0.00	3m2w	28.9	7.45		
HS90A	Heat shock protein HSP 90-alpha	13.1	40.8	14627	70	17	19	0.12	732	7	170	**23.22**	1uyg	48.6	6.01		
ROCK1	Rho-associated protein kinase 1	12.5	40.9	1301	0	1	3	0.08	263	1	12	4.56	2etr	**69.0**	**11.34**	61.2	820
CXCR4	C-X-C chemokine receptor type 4	17.0	41.8	17056	46	15	14	0.09	352	2	40	11.36	3odu	36.3	**11.90**	32.0	**1657**

INHA	Enoylreductase INHA	17.1	42.3	2587	34	9	7	**0.35**	269	1	12	4.46	2h7l				
MK01	MAP kinase ERK2	14.4	42.4	1818	11	3	3	0.17	289	0	0	0.00	2ojg	**53.5**	**13.72**	43.5	432
PPARG	Peroxisome proliferator-activated receptor gamma	11.0	43.6	20228	26	22	7	0.11	505	3	41	8.12	2gtk	31.9	**12.16**	27.8	**1013**
KPCB	"Protein kinase C" beta	14.1	44.0	17725	25	24	5	0.14	259	2	25	9.65	2i0e	**52.2**	7.09	56.7	557
PA2GA	Phospholipase A2 group IIA	19.2	44.1	344	1	0	1	0.00	144	1	12	8.33	1kvo	19.5	5.19	16.1	**1563**
XIAP	Inhibitor of apoptosis protein 3	15.7	44.2	31070	31	28	9	0.09	497	4	72	14.49	3hl5	17.7	3.98		
PNPH	Purine nucleoside phosphorylase	12.0	44.2	2482	20	9	6	**0.36**	289	2	28	9.69	3bgs	**56.8**	**11.69**	49.1	354
TYSY	Thymidylate synthase	11.4	45.3	7914	37	8	16	0.10	313	2	38	12.14	1syn	16.1	6.75	27.0	604
TGFR1	TGF-beta receptor type I	12.3	45.6	4908	8	2	6	0.04	346	1	12	3.47	3hmm	22.1	4.47	20.1	641
DRD3	Dopamine D3 receptor	11.2	46.4	3123	20	3	10	0.10	400	2	40	10.00	3pbl	**73.2**	**18.25**		1782
RENI	Renin	9.7	46.5	61151	38	30	17	0.05	406	0	0	0.00	3g6z	41.1	**9.22**	29.7	**1455**
RXRA	Retinoid X receptor alpha	16.3	46.9	3454	14	3	2	0.09	462	5	96	**20.78**	1mv9	25.6	5.58		
ACES	Acetylcholinesterase	14.6	47.1	37292	130	70	51	0.19	614	3	42	6.84	1E66	35.4	3.96	26.1	665
COMT	Catechol O-methyltransferase	19.5	47.7	5847	20	6	7	0.10	271	2	23	8.49	3bwm	33.2	3.59		
PYGM	Muscle glycogen phosphorylase	18.2	49.1	1892	4	2	3	0.11	842	4	56	6.65	1c8k	35.9	**11.26**	24.0	651

HIVINT	Human immunodeficiency virus type 1 integrase	15.4	49.8	1381	14	2	3	0.14	288	3	40	13.89	3nf7	24.2	**8.50**	26.6	726
SAHH	Adenosylhomocysteinase	14.9	50.0	108	2	0	0	0.00	432	0	0	0.00	1li4	33.5	5.91		
THB	Thyroid hormone receptor beta-1	18.0	50.0	807	2	0	1	0.00	461	4	77	**16.70**	1q4x	**59.6**	**17.23**	57.0	493
KITH	Thymidine kinase	14.4	50.8	14922	41	18	16	0.12	234	1	24	10.26	2b8t	31.7	7.35	29.8	878
FABP4	Fatty acid binding protein adipocyte	31.7	54.4	2423	7	4	4	0.17	132	2	27	**20.45**	2nnq	11.7	2.83	10.7	560
KIF11	Kinesin-like protein 1	16.2	55.8	276	0	0	1	0.00	1056	15	271	**25.66**	3cjo	41.7	**13.96**		
PPARD	Peroxisome proliferator-activated receptor delta	16.5	58.0	3771	5	2	0	0.05	441	3	63	14.29	2znp	48.7	8.12	34.8	975
PPARA	Peroxisome proliferator-activated receptor alpha	10.6	59.0	15712	23	15	5	0.10	468	4	57	12.18	2p54	28.4	8.36	19.7	338
NRAM	Neuraminidase	14.5	59.5	14522	101	14	35	0.10	466	3	49	10.52	1b9v	14.6	8.26		
HXK4	Hexokinase type IV	18.9	63.0	39	n.c.	n.c.	n.c.	n.c.	465	2	22	4.73	3f9m	20.7	4.56	15.6	452
ADA	Adenosine deaminase	15.7	70.6	10490	15	15	9	0.14	363	1	11	3.03	2e1w	23.8	7.08		
PUR2	GAR transformylase	12.6	74.1	102	2	0	0	0.00	1010	7	93	9.21	1njs	32.6	5.36	26.1	631
WEE1	Protein kinase WEE1	27.7	76.6	765	1	1	0	0.13	271	3	38	14.02	3biz	37.0	6.34	33.9	466
GLCM	Beta-glucocerebrosidase	12.1	76.7	601	1	1	1	0.17	536	2	28	5.22	2v3f	16.2	5.81		

FPPS	Farnesyl diphosphate synthase	15.6	77.5	1712	16	5	0	**0.29**	419	3	34	8.11	1zw5	**51.7**	8.48	43.7	635

n.c.: not calculated as total sample size less than 100.

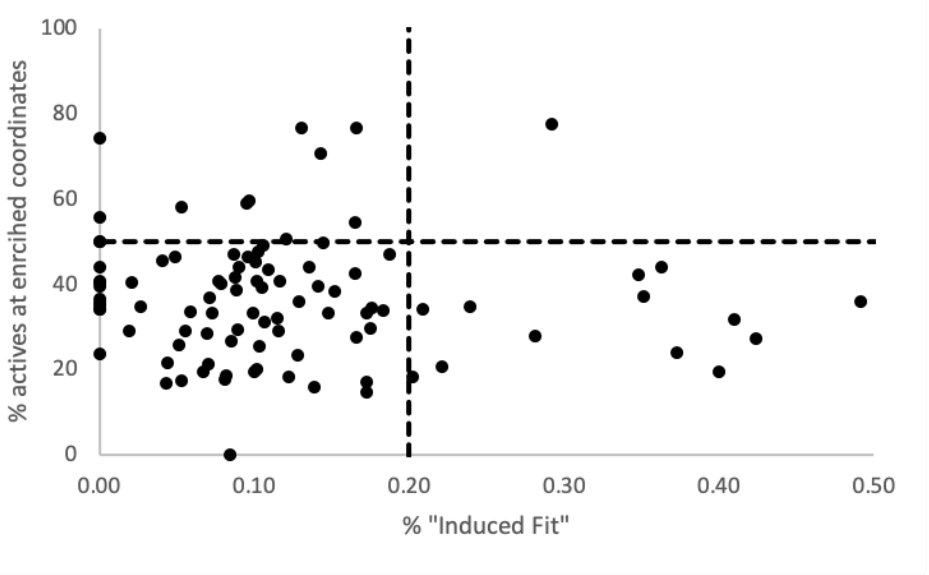


Figure 5. The ligand-shape-derived percentage of actives that are found at enriched coordinates in a shape map for each DUD-E target is plotted against the percentage of references in Web of Science for that target that also include the term "induced fit".

To analyze the targets by sequence, sequences were obtained by searching for the DUD-E target name in the Uniprot system and in all cases the human sequence was selected apart from for: AMPC (beta-lactamse, E. Coli), DEF (peptide deformylase, E. Coli), HIVINT (HIV integrase, HIV), HIVPR (HIV protease, HIV), HIVRT (HIV reverse transcriptase, HIV), INHA (enoyl acyl carrier protein reductase, M. Tuberculosis) and NRAM (neuraminidase, Influenza B).[67, 68] In the case of kinases, to avoid including large sections, such as associated receptors, that are likely not relevant to the measured activity, only the kinase domain was selected. Each sequence was provided to the FoldUnfold server which classifies whole proteins or sections of proteins as being likely to be folded or unfolded/disordered.[69-72] Our assumption is that proteins that have more regions predicted to be disordered are more likely to behave as Hand-in-Glove. In each analysis,

the number of disordered regions (N(unfolded regions)) and the number of amino acids contained in disordered regions (N(unfolded AAs)) were computed and alongside the derived percentage of the sequence computed to be likely to be disordered (% sequence unfolded) recorded in Table 2. The latter is plotted against the ligand shape-derived values in Figure 6. Again, unsurprisingly there is no simple correspondence but among the 28 targets that are computed to include more than 15 % of amino acids in disordered regions, 26 (93 %) place 50 % or less of actives at enriched coordinates.

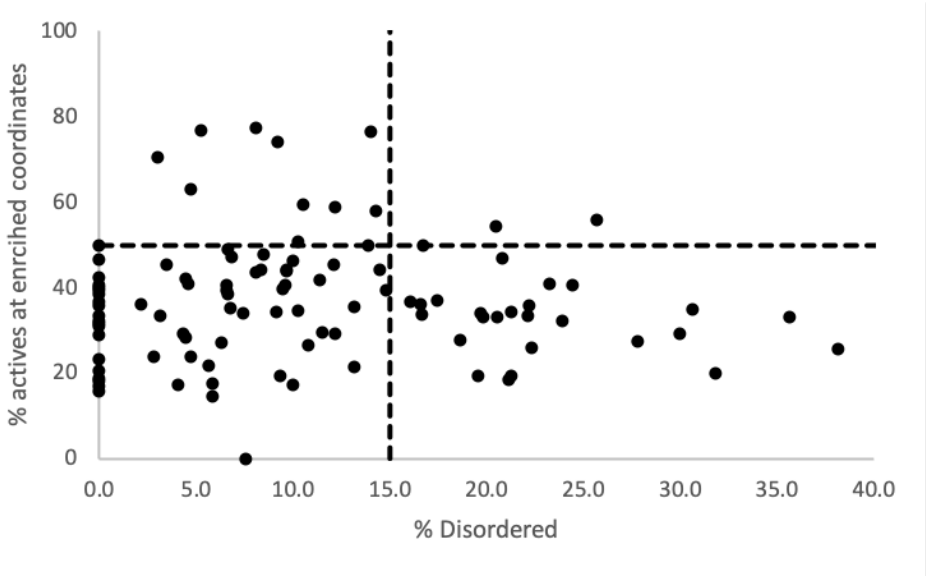
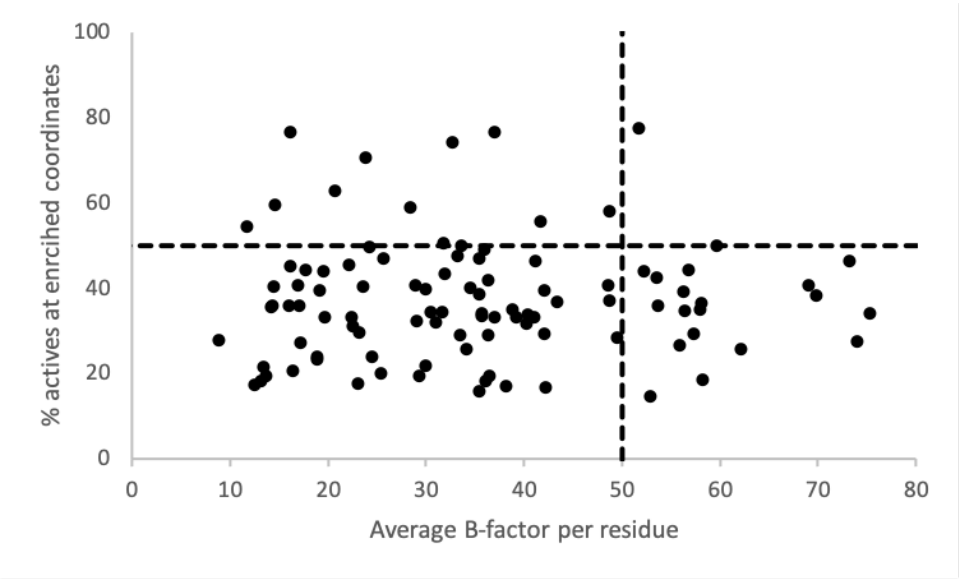


Figure 6. The ligand-shape-derived percentage of actives that are found at enriched coordinates in a shape map for each DUD-E target is plotted against the percentage of amino acids in that target's sequence that are predicted to be in disordered regions by the FoldUnfold method.

Two analyses of a representative structure for each target were also undertaken. Each used the structure registered in the Protein Data Bank with the code provided as an exemplar by the DUD-E system (shown in Table 2). In their selection of these structures, consideration was given to the

resolution, the retrieval of actives in an automated docking campaign and the human structure was preferred.[38] The first analysis of these structures that we performed sought to classify the overall flexibility of the protein while the second focused on the size and localized flexibility of the binding site. The first used the values of the B factors. The B factor has been interpreted as related to flexibility, although this can be a problematic interpretation.[73] A more reliable guide can be provided by the variation in the B factor, which would indicate that certain regions are particularly mobile/disordered compared to others. Both the average B-factor per residue and the average deviation of B-factors per residue were computed with an online service hosted by Radboud University.[74] Of the 20 proteins with an average B-factor above 50, 19 (95 %) see 50 % or less of actives at enriched coordinates (Figure 7, top). More starkly, of the 35 proteins with an average deviation of per residue B factors above 8.5, 34 (97 %) are in this category (Figure 7, bottom)



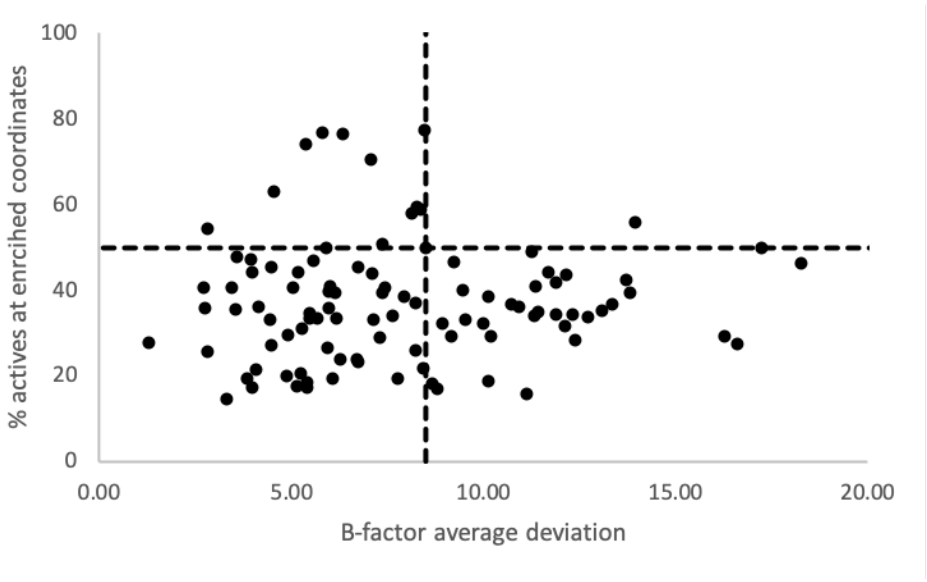


Figure 7. The ligand-shape-derived percentage of actives that are found at enriched coordinates in a shape map for each DUD-E target is plotted against average B-factor per residue (top) and the average deviation for the per residue B-factor (bottom), as computed by the Radboud University service.

The second structural analysis used the sc-PDB service that permits the analysis of the binding site of each structure.[75-77] Unfortunately, not all of the structures have been processed and so an incomplete coverage is available. The binding site volume in each of the representative PDB structures was obtained and is listed in Table 2. As shown in Figure 8, 100 % of the 11 proteins with binding site volumes above 1000 $Å^3$ have less than 50 % of actives at enriched coordinates.

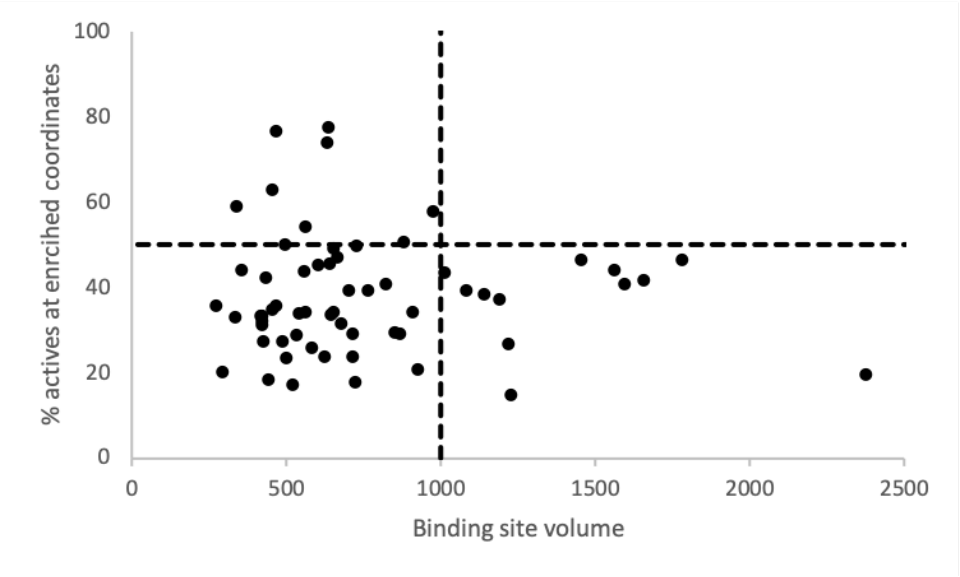


Figure 8. The ligand-shape-derived percentage of actives that are found at enriched coordinates in a shape map for each DUD-E target is plotted against the binding site volume as computed by the sc-PDB service.

All of these analyses point in the same direction and support that targets placing less than 50 % of actives at enriched coordinates are more likely to be flexible and/or have larger binding sites and therefore are predisposed to behave more like hand-in-glove proteins. Those proteins targets achieving greater than 50 % of actives at enriched coordinates are unlikely to be categorized as flexible or to have a large binding site and therefore can be categorized as behaving as a key-in-lock protein. This means that among the set of 102 targets, 90 are categorized as hand-in-glove and 12 as key-in-lock (Table 2 is divided at this cutoff point). This classification can be readily computed given a set of known actives for a target and should be a useful means of identifying those targets where shape is likely to be a very strong determinant of activity and those where its

influence is weaker (although likely still significant as all targets except one provide enrichment at one or more coordinates in the shape map).

Comparison of shape fingerprints with alternative methods and scaffold hopping. Although not our central concern, some comments about the optimized shape fingerprints should be made. Our studies suggest that the fingerprints obtained are effective for molecules in the 200 – 500 Da molecular weight range. The optimized settings entail using the best shape database with a Bit On Value of 0.60. A "balanced" subset of each set of DUD-E decoys was randomly selected to include the same number of compounds as present in the set of actives. This selection was repeated ten times to generate ten random subsets of decoys paired with the actives. The Area Under the receiver operating characteristic Curve (AUC) values for retrieval of actives by actives were computed for each subset (individual curves and values are provided in Figure S17) and are summarized by the mean, standard error in the mean and range for each target in Table 3. The AUC for the full set (not the balanced subset) could only be computed for two of the smaller sets (AMPK and CXCR4) due to the requirement for N x N comparisons to compute this metric; these two targets achieved AUC values of 0.65 and 0.57 respectively, both of which have a 95 % confidence interval range that indicates real enrichment. All of the targets apart from CP3A4 and HIVPR exhibit enrichment by this measure and these are the targets with the highest median molecular weight and also more hand-in-glove-like. Context for these values is provided by comparison to those obtained by the devisors of DUD-E using docking, shown in Figure 9.[28] The performance of the two methods tracks quite well and docking usually performs better than shape fingerprints, as might be expected given the higher information content in a protein structure. GCR achieves better outcomes with shape fingerprints than with docking and this likely reveals that the protein structure used for the docking is inappropriate in this case. Ligand shape methods are

advantageous when no structure or only structures with a weak link to the desired mode of action are available but even then, a number of alternative methods for computing relative shape have superior performance.[13,15,29] An alternative classification in which the shape fingerprints were used to create a decision tree for each of the 10 balanced subsets is detailed in supporting information Section S6 and permits over 80% of compounds to be correctly divided between active and decoy (Table 3).

Table 3. Descriptive statistics for shape fingerprints applied to the DUD-E diverse set.

Target	Average AUC from ten-fold resampling ± standard error [range in brackets]	Average percentage of compounds classified correctly by decision tree	Median molecular weight of actives [range in brackets]
AKT1	0.571 ± 0.001 [0.565-0.577]	82.4	441 [251 – 594]
AMPC	0.622 ± 0.007 [0.597-0.657]	81.2	311 [137 – 426]
CP3A4	0.514 ± 0.005 [0.486-0.527]	80.8	464 [134 – 598]
CXCR4	0.603 ± 0.008 [0.576-0.637]	87.4	391 [211 – 571]
GCR	0.579 ± 0.003 [0.560-0.592]	84.6	430 [294 – 593]
HIVPR	0.468 ± 0.010 [0.457-0.545]	85.6	543 [245 – 600]
HIVRT	0.536 ± 0.003 [0.525-0.552]	83.0	347 [182 – 590]
KIF11	0.699 ± 0.004 [0.685-0.717]	89.2	404 [211 – 595]

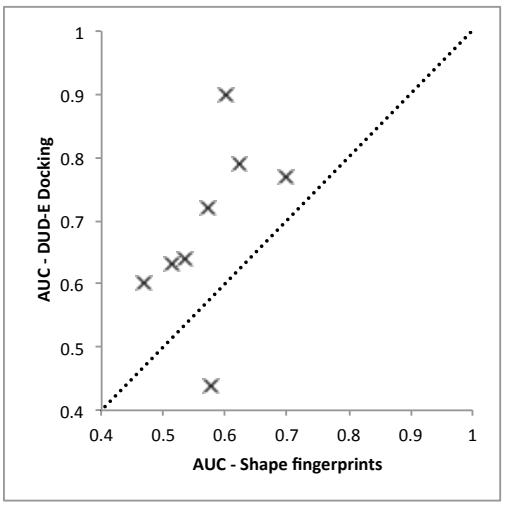
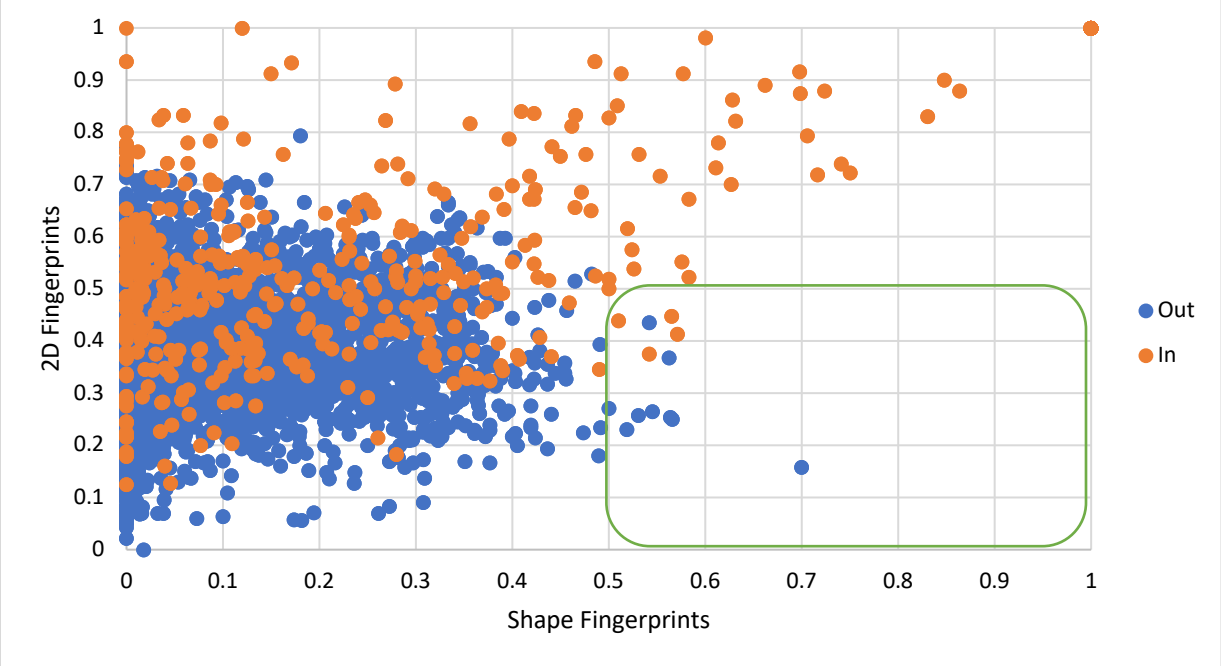


Figure 9. The AUC values obtained by docking are plotted against those obtained by shape fingerprint comparisons.

Given that shape fingerprints neglect the chemical structure of the molecule they should complement the 2D fingerprint methods that are exclusively dependent on the chemical structure (it has been shown that this is a good way of detecting when a method has identified scaffold hops).[78] In order to compare and contrast the two approaches, various types of 2D fingerprints (MACCS keys, path, tree and circular) for both test sets were generated using Openeye's TK and compared using a Similarity Tanimoto.[79] The calculated AUC values are shown in Table 4. The AUC values are higher when using 2D fingerprints for both test sets (full details of the test sets are provided in section S3). However, considering that shape fingerprints do not use any chemical information about the molecules but only their shape, the slightly worse AUC values than for well-established methods is not too surprising.

Table 4. Comparison of the AUC values for different fingerprint methods. In the case of shape fingerprints, values obtained for SD10 with Design Tanimoto = 0.65 and Bit On Value = 0.60 are shown.[5]

	Fingerprint method				
	MACCS166	**Path**	**Tree**	**Circular**	**Shape Fingerprints**
Test Set 1	0.74	0.67	0.69	0.69	0.64
Test Set 2	0.94	0.94	0.94	0.97	0.85



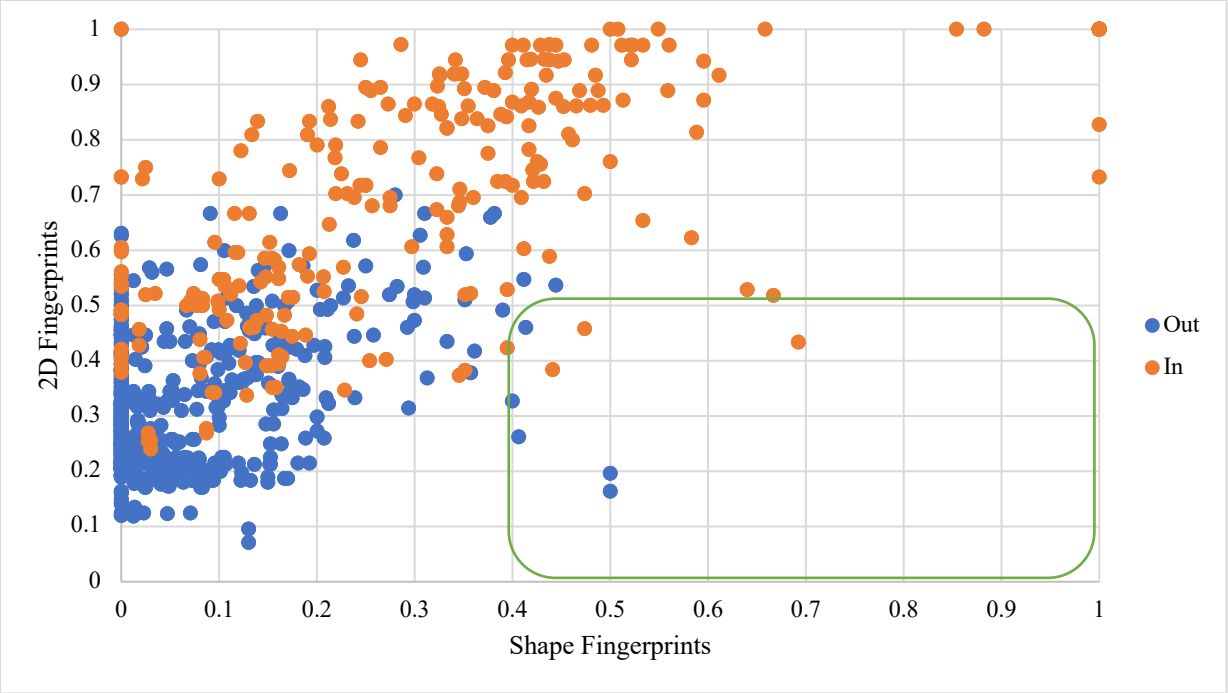


Figure 10. Plot showing Fingerprint Tanimoto values obtained by both Shape Fingerprints (x-axis) and MACCS166 fingerprints (2D fingerprints, y-axis) for each comparison in Test Set 1 (top) and Test Set 2 (bottom). Points in red correspond to compound pairs that share biological activity those in blue do not.

To investigate the complementarity between the two fingerprint types, the Tanimotos between pairs of molecules have been computed with both MACCS166 and shape fingerprints. These are plotted against one another in Figure 10. Many pairs of molecules with shared biological activity (colored red) have high similarity according to both methods, which is unsurprising. There are a small number of examples of molecules with low shape similarity but high 2D fingerprint similarity that share biological activity but, there are also a small number with high shape similarity and relatively low 2D fingerprint similarity. The orange box on each of the plots in Figure 10 highlights these examples. These are connections that represent scaffold hopping.[78]

Two examples of a pair of structures for each of the test sets is shown in Figure 11. Two neuraminidase inhibitors (one that contains an aromatic core and one with a monosaccharide core) provide a very clear example of scaffold hopping between compounds that are likely to have different physical properties while the indole and ortho-substituted phenol pair of tryptophan synthase inhibitors show that ring-opening scaffold hops can also be detected by shape fingerprints.[80-83]

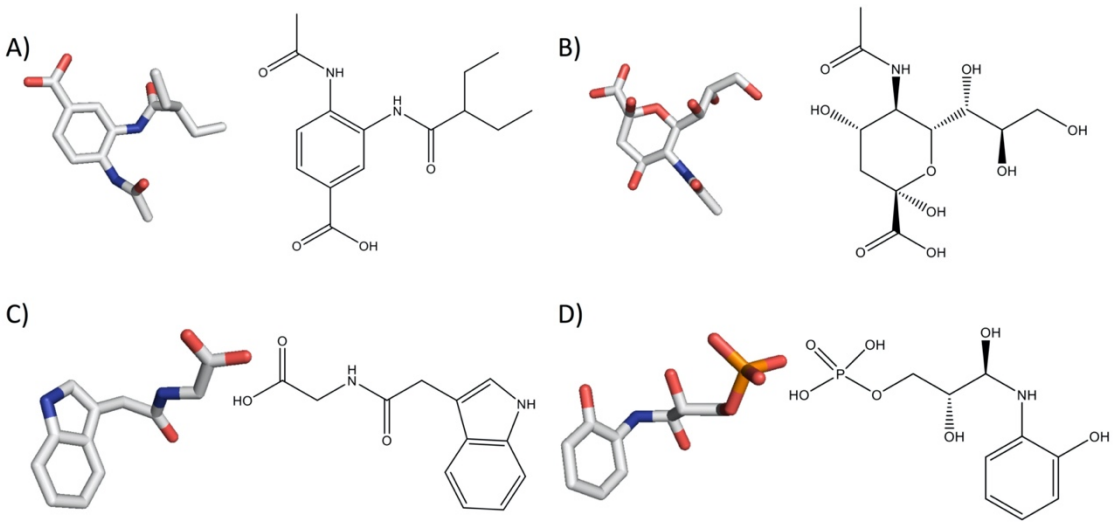


Figure 11. The structures of molecules binding to Neuraminidase with pdb codes: 1b9s (A) and 1nsc (B) and Tryptophan Synthase with pdb codes: 1k7e (C) and 1tjp (D).

Conclusions

Two absolute descriptions of molecular shape were used to create self-organizing maps (shape maps) for the 8 targets in the DUD-E diverse subset. Shape fingerprints and USR are able to provide shape maps that reveal the preferences of each target and provide insights concerning the range of shapes that are tolerated and the strictness of any shape preferences. The USR method is sufficiently fast to be applicable to the full set of 102 DUD-E targets. By placing a set of actives

and the set of physical-property-matched decoys generated by the DUD-E system, coordinates enriched in actives in the shape maps can be identified. Those targets that find more than 50 % of actives at enriched coordinates can be thought of as key-in-lock, while those with 50 % or less are hand-in-glove. The former (key-in-lock) are unlikely to either show high protein flexibility or to have large binding sites. This is a remarkable insight concerning the nature of the protein target that is available from study of a set of active compounds alone and we will be probing this link further.

In the development of a shape fingerprint method, the best Shape Database that was obtained is provided to permit others to apply this method and can be accessed via our GitHub repository: https://github.com/LeachResearchGroup/ShapeFingerprints.

Supporting Information.

A full description of the developments of the shape fingerprints, the methods, the two test sets, AUC values for the two test sets obtained with varying shape databases, analysis of the DUD-E set with the optimized fingerprint method and examples of its use to obtain shape comparators, chemical structures of compounds at enriched coordinates in shape maps, shape maps for the full DUD-E set.

AUTHOR INFORMATION

Corresponding Author

* (A.G.L.) E-mail: andrew.leach@manchester.ac.uk

Funding Sources

This research was financially supported by MedChemica Limited and Liverpool John Moores University.

Data and Software Availability. All biological data on which these studies are based has been drawn from public sources: DUD-E (http://dude.docking.org/) and from published subsets of the Protein Data Bank. Shape fingerprint studies employed python code (provided in the supporting information) and shape comparisons rely upon the Openeye toolkit version OpenEye-toolkits-2014.Jun.6-osx-10.8 (https://www.eyesopen.com/cheminformatics). Self-organizing maps and statistical analysis employed open-source software: R version 4.0.2 including SOMbrero version 1.3.1. The code used to generate and run R scripts is provided in the supporting information. USR descriptors were generated using the code lines provided in the supporting information and employed Python 3.8.5 and rdkit version 2020.03.4. File splitting was performed in Open Babel version 2.4.1.

Acknowledgments.

The authors wish to thank Philip Rowe and Marc Reid for useful discussion and Openeye for an academic license. Plots were made using the Matplotlib for python and ggplot2 module in R.[84, 85]

References.

1. Lauria, A.; Tutone, M.; Almerico, A. M. Virtual Lock-and-key Approach: the In Silico Revival of Fischer Model by Means of Molecular Descriptors. *Eur. J. Med. Chem.* **2011**, *46*, 4274-4280.
2. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985-2993.
3. Pauling, L. Molecular Architecture and Biological Reactions. *Chem. Eng. News* **1946**, *24*, 1375-1377.
4. Ghibaudi, E.; Cerruti, L.; Villani, G. Structure, Shape, Topology: Entangled Concepts in Molecular Chemistry. *Found. Chem.* **2019**, *22*, 279-307.

5. Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673-684.
6. Zauhar, R. J.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. Shape Signatures: a New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design. *J. Med. Chem.* **2003**, *46*, 5674-90.
7. Cannon, E. O.; Nigsch, F.; Mitchell, J. B. A Novel Hybrid Ultrafast Shape Descriptor Method for Use in Virtual Screening *Chem. Cent. J.* **2008**, *2*, 3.
8. Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: Fast Molecular Similarity Calculations Incorporating Shape, Chirality and Electrostatics *J. Comput. Aided Mol. Des.* **2010**, *24*, 789-801.
9. Schreyer, A. M.; Blundell, T. USRCAT: Real-Time Ultrafast Shape Recognition with Pharmacophoric Constraints *J. Cheminform.* **2012**, *4*, 27.
10. Tarko, L. Computation of the Molecular Shapes' Similarity and Diversity Using USR Method and General Shape Index. *J. Math. Chem.* **2015**, *53*, 1576-1591.
11. Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition for Similarity Search in Molecular Databases. *Proc. R. Soc. A* **2007**, *463*, 1307-1321.
12. Ballester, P. J. Ultrafast Shape Recognition: Method and Applications. *Future Med. Chem.* **2011**, *3*, 65-78.
13. Cai, C.; Gong, J.; Liu, X.; Jiang, H.; Gao, D.; Li, H. A Novel, Customizable And Optimizable Parameter Method Using Spherical Harmonics For Molecular Shape Similarity Comparisons. *J. Mol. Model.* **2012**, *18*, 1597-1610.
14. Seddon, M. P.; Cosgrove, D. A.; Packer, M. J.; Gillet, V. J. Alignment-Free Molecular Shape Comparison Using Spectral Geometry: The Framework. *J. Chem. Inf. Model.* **2019**, *59*, 98-116.
15. Roy, A.; Skolnick, J. LIGSIFT: An Open-Source Tool For Ligand Structural Alignment and Virtual Screening. *Bioinformatics* **2015**, *31*, 539-44.
16. Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Model.* **1993**, *33*, 112-116.
17. Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method And Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489-95.
18. AbdulHameed, M. D.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring Polypharmacology Using a ROCS-Based Target Fishing Approach. *J. Chem. Inf. Model.* **2012**, *52*, 492-505.
19. Sato, T.; Yuki, H.; Takaya, D.; Sasaki, S.; Tanaka, A.; Honma, T. Application Of Support Vector Machine To Three-Dimensional Shape-Based Virtual Screening Using Comprehensive Three-Dimensional Molecular Shape Overlay With Known Inhibitors. *J. Chem. Inf. Model.* **2012**, *52*, 1015-26.
20. Leherte, L.; Vercauteren, D. P., Smoothed Gaussian Molecular Fields: An Evaluation Of Molecular Alignment Problems. *Theor. Chem. Acc.* **2012**, *131*, 1259.
21. Vaz de Lima, L. A.; Nascimento, A. S. Molshacs: A Free And Open Source Tool For Ligand Similarity Identification Based On Gaussian Descriptors. *Eur. J. Med. Chem.* **2013**, *59*, 296-303.
22. Yan, X.; Li, J.; Liu, Z.; Zheng, M.; Ge, H.; Xu, J. Enhancing Molecular Shape Comparison By Weighted Gaussian Functions. *J. Chem. Inf. Model.* **2013**, *53*, 1967-78.
23. Kalaszi, A.; Szisz, D.; Imre, G.; Polgar, T. Screen3D: A Novel Fully Flexible High-Throughput Shape-Similarity Search Method. *J. Chem. Inf. Model.* **2014**, *54*, 1036-1049.

24. Vainio, M. J.; Puranen, J. S.; Johnson, M. S. Shaep: Molecular Overlay Based On Shape and Electrostatic Potential. *J. Chem. Inf. Model.* **2009**, *49*, 492-502.
25. Heal, G. A.; Walker, P. D.; Mezey, P. G.; Ramek, M. Shape-Similarity Analysis of 20 Stable Conformations of Neutral B-Alanine. *Can. J. Chem.* **1996**, *74*, 1660-1670.
26. Liu, X.; Jiang, H.; Li, H. SHAFTS: A Hybrid Approach For 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 2372-2385.
27. Awale, M.; Jin, X.; Reymond, J. L. Stereoselective Virtual Screening Of The ZINC Database Using Atom Pair 3D-Fingerprints. *J. Cheminform.* **2015**, *7*, 3.
28. von Behren, M. M.; Bietz, S.; Nittinger, E.; Rarey, M. mRaise: An Alternative Algorithmic Approach To Ligand-Based Virtual Screening. *J. Comput. Aided Mol. Des.* **2016**, *30*, 583-594.
29. von Behren, M. M.; Rarey, M. Ligand-Based Virtual Screening Under Partial Shape Constraints. *J. Comput. Aided Mol. Des.* **2017**, *31*, 335-347.
30. Wei, N. N.; Hamza, A. SABRE: Ligand/Structure-Based Virtual Screening Approach Using Consensus Molecular-Shape Pattern Recognition. *J. Chem. Inf. Model.* **2014**, *54*, 338-346.
31. Kumar, A.; Zhang, K. Y. Prospective Evaluation Of Shape Similarity Based Pose Prediction Method In D3R Grand Challenge 2015. *J. Comput. Aided Mol. Des.* **2016**, *30*, 685-693.
32. Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. Pubchem3D: A New Resource For Scientists. *J Cheminform.* **2011**, *3*, 32.
33. Bolton, E. E.; Kim, S.; Bryant, S. H. Pubchem3D: Diversity Of Shape. *J. Cheminform.* **2011**, *3*, 9.
34. Bolton, E. E.; Kim, S.; Bryant, S. H. PubChem3D: Similar Conformers. *J. Cheminform.* **2011**, *3*, 13.
35. Kim, S.; Bolton, E. E.; Bryant, S. H., PubChem3D: Biologically Relevant 3-D Similarity. *J. Cheminform.* **2011**, *3*, 26.
36. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98-104.
37. Koshland, D. E. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chem., Int. Ed. Engl.* **1995**, *33*, 2375-2378.
38. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory Of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys For Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582-6594.
39. Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development Of Versions 3 And 4 Of The Cambridge Structural Database System. *J. Chem. Inf. Model.* **1991**, *31*, 187-204.
40. Allen, F. H. The Cambridge Structural Database: A Quarter Of A Million Crystal Structures And Rising. *Acta Crystallogr. B* **2002**, *58*, 380-8.
41. Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse For Ligands Bound To Macromolecules. *Bioinformatics* **2004**, *20*, 2153-2155.
42. Taylor, R.; Cole, J. C.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Korb, O. Development And Validation Of An Improved Algorithm For Overlaying Flexible Molecules. *J. Comput. Aided Mol. Des.* **2012**, *26*, 451-472.

43. Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-Ligand Docking Against Non-Native Protein Conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214-2225.
44. *RDKit: Open-source cheminformatics*, http://www.rdkit.org (accessed on 21st January 2021).
45. Covell, D. G.; Wallqvist, A.; Rabow, A. A.; Thanki, N. Molecular Classification Of Cancer: Unsupervised Self-Organizing Map Analysis Of Gene Expression Microarray Data. *Mol. Cancer Ther.* **2003**, *2*, 317-332.
46. Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Mining The National Cancer Institute's Tumor-Screening Database: Identification Of Compounds With Similar Cellular Activities. *J. Med. Chem.* **2002**, *45*, 818-840.
47. Olteanu, M.; Villa-Vialaneix, N.; Cottrell, M. On-Line Relational SOM for Dissimilarity Data. In *Advances in Self-Organizing Maps*; 2013; Chapter 2, pp 13-22.
48. Olteanu, M.; Villa-Vialaneix, N. On-Line Relational And Multiple Relational SOM. *Neurocomputing* **2015**, *147*, 15-30.
49. Olteanu, M.; Villa-Vialaneix, N. Using Sombrero For Clustering And Visualizing Graphs. *J. Soc. Fr. Statistique* **2015**, *156*, 95-119.
50. Vialaneix, N. M., E.; Mariette, J.; Olteanu, M.; Rossi, F.; Bendhaibi, L.; Bolaert, J. *SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs*, R package version 1.3-1; 2020.
51. Bonferroni, C. Teoria Statistica Delle Classi E Calcolo Delle Probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* **1936**, *8*, 3-62.
52. Katritch, V.; Cherezov, V.; Stevens, R. C. Structure-Function Of The G Protein-Coupled Receptor Superfamily. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 531-56.
53. Ekroos, M.; Sjogren, T. Structural Basis For Ligand Promiscuity In Cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 13682-13687.
54. Teague, S. J. Implications Of Protein Flexibility For Drug Discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 527-541.
55. Huang, P.; Chandra, V.; Rastinejad, F. Structural Overview Of The Nuclear Receptor Superfamily: Insights Into Physiology And Therapeutics. *Annu. Rev. Physiol.* **2010**, *72*, 247-272.
56. Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Model.* **1995**, *35*, 59-67.
57. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747-750.
58. Osguthorpe, D. J.; Sherman, W.; Hagler, A. T., Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols. *J. Phys. Chem. B* **2012**, *116*, 6952-6959.
59. Longenecker, K. L.; Stewart, K. D.; Madar, D. J.; Jakob, C. G.; Fry, E. H.; Wilk, S.; Lin, C. W.; Ballaron, S. J.; Stashko, M. A.; Lubben, T. H.; Yong, H.; Pireh, D.; Pei, Z. H.; Basha, F.; Wiedeman, P. E.; von Geldern, T. W.; Trevillyan, J. M.; Stoll, V. S. Crystal Structures Of DPP-IV (CD26) From Rat Kidney Exhibit Flexible Accommodation Of Peptidase-Selective Inhibitors. *Biochemistry* **2006**, *45*, 7474-7482.
60. Favia, A. D.; Masetti, M.; Recanatini, M.; Cavalli, A. Substrate Binding Process And Mechanistic Functioning Of Type 1 11-Beta-Hydroxysteroid Dehydrogenase From Enhanced Sampling Methods. *PLoS One* **2011**, *6*, e25375.

61. Hosfield, D. J.; Wu, Y.; Skene, R. J.; Hilgers, M.; Jennings, A.; Snell, G. P.; Aertgeerts, K. Conformational Flexibility In Crystal Structures Of Human 11-Beta-Hydroxysteroid Dehydrogenase Type I Provide Insights Into Glucocorticoid Interconversion And Enzyme Regulation. *J. Biol. Chem.* **2005**, *280*, 4639-4648.
62. Klein, T.; Vajpai, N.; Phillips, J. J.; Davies, G.; Holdgate, G. A.; Phillips, C.; Tucker, J. A.; Norman, R. A.; Scott, A. D.; Higazi, D. R.; Lowe, D.; Thompson, G. S.; Breeze, A. L. Structural And Dynamic Insights Into The Energetics Of Activation Loop Rearrangement In FGFR1 Kinase. *Nat. Commun.* **2015**, *6*, 12.
63. Huang, Z.; Wong, C. F. Inexpensive Method for Selecting Receptor Structures for Virtual Screening. *J. Chem. Inf. Model.* **2016**, *56*, 21-34.
64. Lerner, C.; Masjost, B.; Ruf, A.; Gramlich, V.; Jakob-Roetne, R.; Zurcher, G.; Borroni, E.; Diederich, F. Bisubstrate Inhibitors For The Enzyme Catechol-O-Methyltransferase (COMT): Influence Of Inhibitor Preorganisation And Linker Length Between The Two Substrate Moieties On Binding Affinity. *Org. Biomol. Chem.* **2003**, *1*, 42-49.
65. Kuwabara, N.; Oyama, T.; Tomioka, D.; Ohashi, M.; Yanagisawa, J.; Shimizu, T.; Miyachi, H. Peroxisome Proliferator-Activated Receptors (PPARs) Have Multiple Binding Points That Accommodate Ligands in Various Conformations: Phenylpropanoic Acid-Type PPAR Ligands Bind to PPAR in Different Conformations, Depending on the Subtype. *J. Med. Chem.* **2012**, *55*, 893-902.
66. Web of Science. https://apps.webofknowledge.com/ (accessed 21st January 2021).
67. Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115-D119.
68. UniProt Consortium UniProt: A Worldwide Hub Of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506-D515.
69. Galzitskaya, O. V.; Garbuzynskiy, S. O.; Lobanov, M. Y. FoldUnfold: Web Server for the Prediction of Disordered Regions in Protein Chain. *Bioinformatics* **2006**, *22*, 2948-2949.
70. Galzitskaya, O. V.; Garbuzynskiy, S. O.; Lobanov, M. Y. Prediction of Amyloidogenic and Disordered Regions in Protein Chains. *PLoS Comput. Biol.* **2006**, *2*, e177.
71. Galzitskaya, O. V.; Garbuzynskiy, S. O.; Lobanov, M. Y. Prediction of Natively Unfolded Regions in Protein Chains. *Mol. Bio.* **2006**, *40*, 298-304.
72. Garbuzynskiy, S. O.; Lobanov, M. Y.; Galzitskaya, O. V. To Be Folded Or To Be Unfolded? *Protein Sci.* **2004**, *13*, 2871-7.
73. Sun, Z.; Liu, Q.; Qu, G.; Feng, Y.; Reetz, M. T., Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* **2019**, *119*, 1626-1665.
74. Krause, R. H.; Hekkelman, M. L.; Nielsen, J. E.; Vriend, G. Average B factors. https://swift.cmbi.umcn.nl/servers/html/listavb.html (accessed January 21st 2021)
75. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-Database Of Ligandable Binding Sites—10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399-D404.
76. Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An Annotated Database Of Druggable Binding Sites From The Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717-727.

77. Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: A Database For Identifying Variations And Multiplicity Of 'Druggable' Binding Sites In Proteins. *Bioinformatics* **2011**, *27*, 1324-1326.
78. Lovrics, A.; Pape, V. F. S.; Szisz, D.; Kalászi, A.; Heffeter, P.; Magyar, C.; Szakács, G. Identifying New Topoisomerase II Poison Scaffolds By Combining Publicly Available Toxicity Data And 2D/3D-Based Virtual Screening. *J. Cheminf.* **2019**, *11*, 67.
79. *OpenEye Toolkits*, OpenEye Scientific Software: Santa Fe, NM, 2014.
80. Finley, J. B.; Atigadda, V. R.; Duarte, F.; Zhao, J. J.; Brouillette, W. J.; Air, G. M.; Luo, M. Novel Aromatic Inhibitors Of Influenza Virus Neuraminidase Make Selective Interactions With Conserved Residues And Water Molecules In The Active Site. *J. Mol. Bio.* **1999**, *293*, 1107-1119.
81. Burmeister, W. P.; Henrissat, B.; Bosso, C.; Cusack, S.; Ruigrok, R. W. H. Influenza B Virus Neuraminidase Can Synthesize Its Own Inhibitor. *Structure* **1993**, *1*, 19-26.
82. Weyand, M.; Schlichting, I.; Marabotti, A.; Mozzarelli, A. Crystal Structures Of A New Class Of Allosteric Effectors Complexed to Tryptophan Synthase. *J. Biol. Chem.* **2002**, *277*, 10647-10652.
83. Kulik, V.; Hartmann, E.; Weyand, M.; Frey, M.; Gierl, A.; Niks, D.; Dunn, M. F.; Schlichting, I. On The Structural Basis Of The Catalytic Mechanism And The Regulation Of The Alpha Subunit Of Tryptophan Synthase From Salmonella Typhimurium And BX1 From Maize, Two Evolutionarily Related Enzymes. *J. Mol. Biol.* **2005**, *352*, 608-620.
84. Hunter, J. D., Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90-95.
85. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag: New York, NY, https://ggplot2.tidyverse.org, 2016.

Table of Contents graphic