Supporting Information

Alignment-independent comparison of binding sites based on DrugScore potential fields encoded by 3D Zernike descriptors

Britta Nisius and Holger Gohlke*

Department of Mathematics and Natural Sciences,

Institute of Pharmaceutical and Medicinal Chemistry,

Heinrich-Heine University Düsseldorf,

Universitätsstraße 1, 40225 Düsseldorf, Germany.

Running title head: DrugScore-based binding site descriptors

*To whom correspondence should be addressed: Tel: +49-211-81-13662, Fax: +49-221-81-13847, E-mail: gohlke@uni-duesseldorf.de

Table S1. AUC values for the Hoffmann dataset using EasyMIF potential fields[a]

Ligand ID	CR1	CH3	O	OA	N	Combined probe atoms
LDA	0.740	0.750	0.791	0.801	0.760	0.779
PMP	0.947	0.954	0.942	0.929	0.938	0.950
SAM	0.723	0.746	0.723	0.698	0.732	0.730
LLP	0.925	0.932	0.904	0.908	0.910	0.925
U5P	0.715	0.719	0.719	0.700	0.714	0.720
1PE	0.595	0.611	0.625	0.650	0.627	0.623
SUC	0.777	0.769	0.773	0.765	0.772	0.776
GSH	0.769	0.767	0.747	0.773	0.753	0.765
PLM	0.628	0.615	0.636	0.641	0.627	0.632
BOG	0.576	0.560	0.555	0.563	0.538	0.562
average	0.739	0.742	0.741	0.743	0.737	0.746

[a] Average AUC values for each set of proteins binding identical ligands are shown for Zernike descriptors encoding EasyMIF potential fields for five individual probe atoms (CR1: aromatic carbon, CH3: aliphatic carbon, O: carbonyl oxygen, OA: hydroxyl oxygen, N: peptide nitrogen), as well as the combination of all probe atoms.

Table S2. Flexibility dataset[a]

Drug target	Abbreviation	PDB ID	$RMSD_{ave}$ [b]	$RMSD_{max}$ [b]	Zernike distance[c]
Angiotensin-converting enzyme	ACE	1uze			
		1o86	0.17	0.37	21330.64
		1uzf	0.35	0.79	41197.00
Aldose reductase	AR	1ah0			
		1ah3	1.06	3.19	111021.46
		2acr*	0.88	1.72	86527.13
Cyclin-dependent kinase 2	CDK2	1aq1			
		1buh*	1.77	3.20	155848.87
		1dm2	1.75	4.49	100957.52
Cyclooxygenase 2	COX2	1cvu			
		1cx2	1.24	3.78	84748.22
		3pgh	1.11	3.96	79374.00
Dihydrofolate reductase	DHFR	3dfr			
		6dfr*	1.47	1.96	134312.93
Estrogen receptor	ER	1l2i			
		3ert	2.61	4.47	163801.75
		1err	2.01	4.39	161791.85
Factor Xa	FXa	1f0r			
		1fjs	1.09	2.57	74442.37
		1ksn	0.67	1.65	33929.45
		1xka	1.27	2.46	97423.25
HIV reverse transcriptase	HIVRT	1vrt			
		1rt1	1.51	2.45	139005.08
		1c1c	1.88	3.12	147504.77

		1rth	1.62	2.28	115346.27
Hydroxymethylglutaryl-CoA reductase	HMGR	1hw8			
		1hwk	0.61	1.49	26031.82
Neuraminidase	NA	1a4g			
		1a4q	0.48	2.11	51189.42
		1nsc	0.34	1.49	26188.83
P38 mitogen activated protein kinase	P38	1a9u			
		1kv1	3.84	10.41	198962,60
		1kv2	3.54	11.26	183360.77
Phosphodiesterase 5	PDE5	1xoz			
		1xp0	0.79	2.23	54872.88
Peroxisome proliferator activated receptor gamma	PPARg	1fm6			
		1fm9	1.47	4.64	129794.50
		2prg	0.71	1.27	99456.17
Thrombin	THR	1ba8			
		1hgt*	0.69	1.85	58977.18
Thymidine kinase	TK	1kim			
		1ki4	1.78	2.90	56462.93

[a] *Apo* structures are marked by an asterisk. $RMSD_{ave}$ is the average conformational variability of all binding site residues. $RMSD_{max}$ is the maximum side chain movement among all binding site residues. Both values were taken from Huang and Jacobsen[36].

[b] In Å.

[c] Manhattan distance of Zernike descriptors encoding all five utilized DrugScore probe atoms.

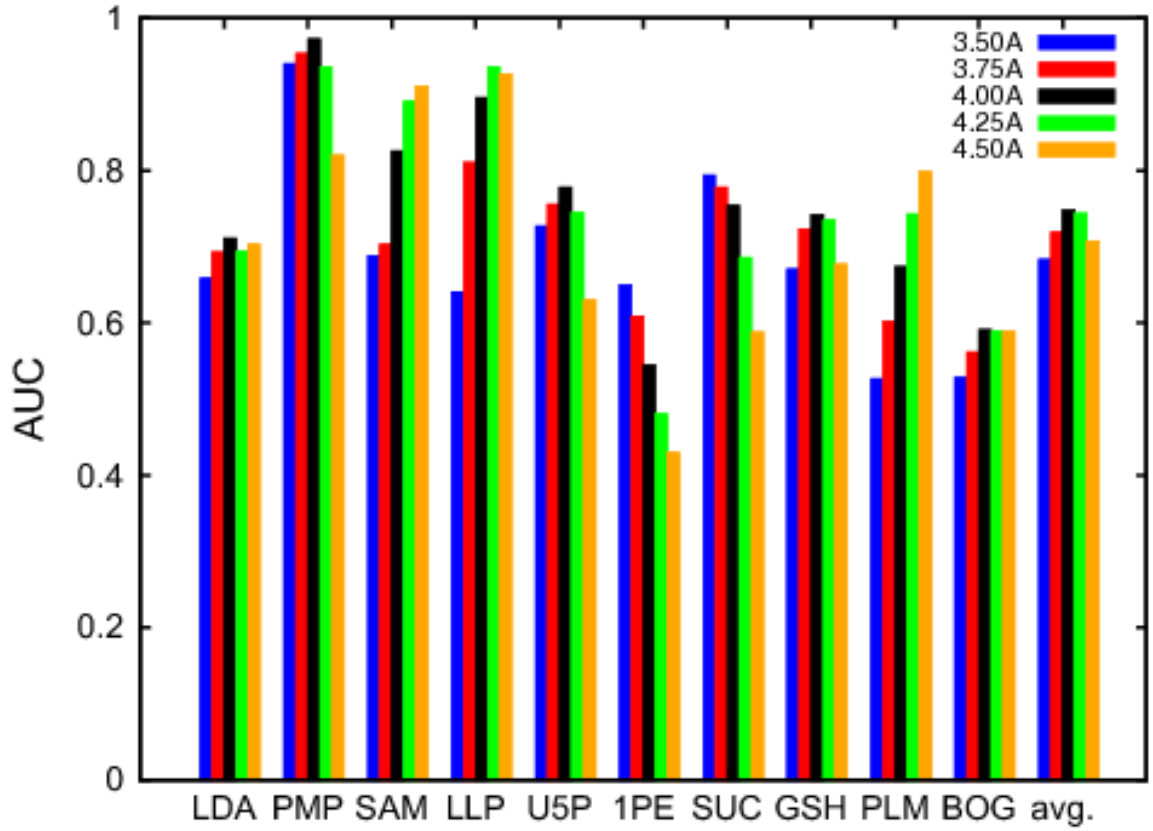

Figure S1. Influence of varying DrugScore grid box sizes on Zernike descriptor performance using the Hoffmann dataset

DrugScore potential fields for grid boxes of varying size, i.e. extending by 3.5 Å, 3.75 Å, 4.0 Å, 4.25 Å and 4.5 Å over the ligand in the crystal structure, were computed and transformed to Zernike descriptors. For each protein five individual trials were performed for which the size of the DrugScore grid box of the query protein was gradually varied according to the above values whereas the size of the DrugScore grid boxes of all other proteins were kept constant. Resulting averaged AUC values for each ligand subset (LDA, PMP, SAM, LLP, U5P, 1PE, SUC, GSH, PLM, and BOG) as well as for the entire Hoffmann dataset (avg.) are shown.

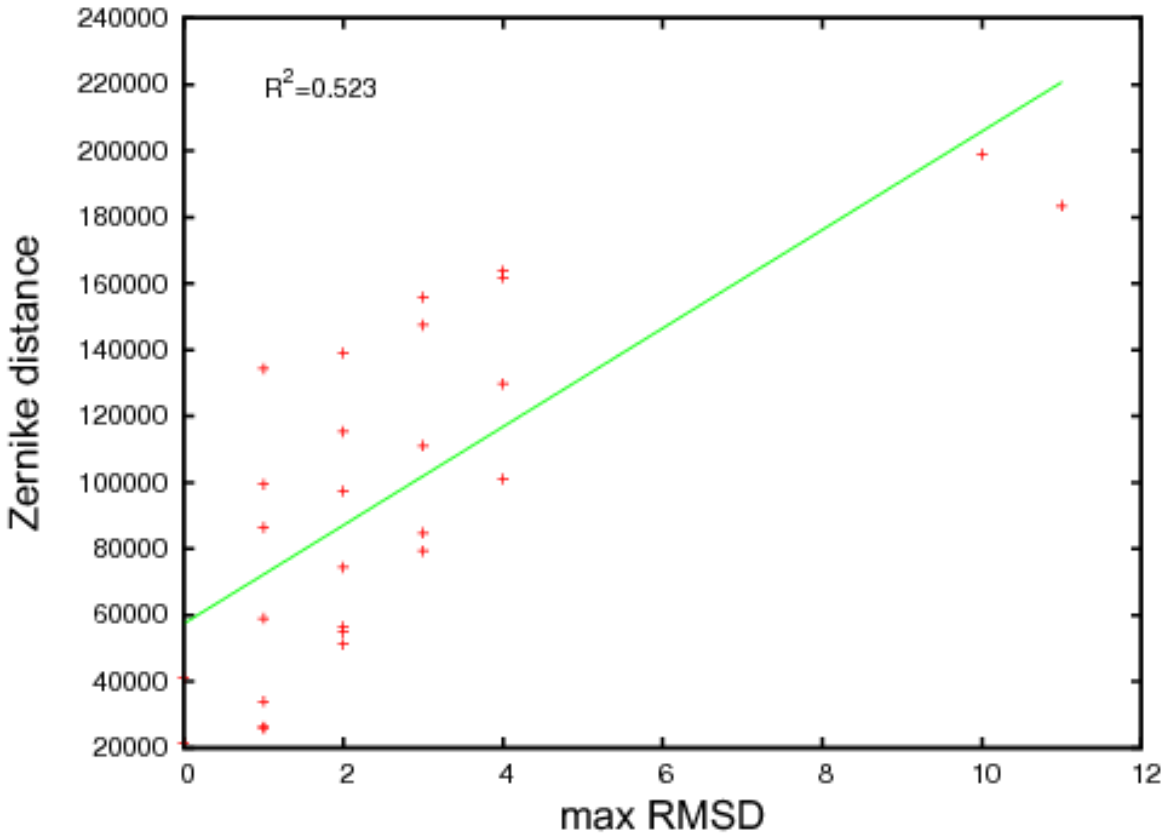

Figure S2. Structural differences between binding sites and Zernike distances using the flexibility dataset

For pairs of crystal structures of the same protein (compare Table S2) maximal structural differences (max RMSD; in Å) between corresponding side chains or loops are plotted against binding site distances in terms of Zernike descriptors (Zernike distances). R^2 = 0.523 (0.434) is obtained if all crystal structures are taken into account (if the p38 structures showing the largest structural differences (max RMSD > 10 Å) are discarded).