Supporting Information

for

A diverse, high-quality test set for the validation of protein-ligand docking performance

Michael J. Hartshorn, Marcel L. Verdonk, Gianni Chessari, Suzanne C. Brewerton, Wijnand T.M. Mooij, Paul N. Mortenson and Christopher W. Murray*

Astex Therapeutics Ltd, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, UK

Table S1. HET groups that were excluded in this study.

A	C	G	K	T	U	BR	CA	CD	CL	CO	CS
CU	FE	HG	MG	MN	NA	NI	NO	O2	OH	PI	PT
SR	TL	XE	ZN	1CP	1MA	1PE	1PG	2CP	2GP	2GP	2MG
3CP	3DR	3HC	4CA	4CO	5GP	5MC	5MU	6FA	7MG	8OG	A3P
ABA	ACE	ACO	ACO	ACP	ACT	ACY	ADE	ADJ	ADN	ADP	ADP
AF3	AIB	AKG	AMP	AMP	AMX	ANP	ANP	APC	APR	ATP	ATP
AZI	B12	BCA	BCB	BCL	BCL	BEN	BEZ	BGC	BH4	BLA	BLV
BMA	BMA	BME	BOG	BOG	BPB	BPH	BPH	BRO	BTN	C8E	C8E
CA3	CA5	CAA	CAC	CAN	CAO	CCH	CDL	CDP	CEA	CGU	CH2
CH3	CIC	CIT	CIT	CL1	CL2	CLA	CLN	CMC	CME	CMO	CMP
CMP	CMX	CNC	CND	CO3	CO8	COA	COA	COB	COD	COF	COH
COJ	CON	COS	COT	COY	CP3	CPS	CRO	CRY	CRY	CS8	CSD
CSO	CSS	CSW	CTP	CU1	CXM	CYN	DAK	DAL	DAN	DCA	DCC
DDH	DEU	DHE	DHE	DLE	DMS	DND	DOX	DPN	DTT	DTT	DVA
EDO	EGL	EOH	EPE	EPE	ETA	F3S	F43	F6P	F6P	FAA	FAB
FAD	FAD	FAE	FAM	FAS	FCX	FDA	FE2	FEC	FES	FLC	FMA
FME	FMN	FMN	FMT	FNE	FNS	FOR	FPP	FS3	FS4	FUC	FUC
GAL	GAL	GDP	GDP	GLC	GLC	GLU	GLY	GMP	GNP	GNP	GOL
GOL	GOL	GSH	GTP	GTP	GTT	H2U	H4B	HAS	HAX	HC4	HDD
HDM	HE6	HEA	HEB	HEC	HEC	HED	HED	HEG	HEM	HEM	HEO
HES	HEV	HIF	HMG	HNI	HPA	HXC	HYP	IMD	IMP	IOD	IOH
IPA	IPH	IUM	KCX	LAT	LDA	LDA	LI1	LI1	LLP	LYX	M2G
MAL	MAL	MAN	MAN	MCA	MCD	MDE	MES	MES	MET	MGD	MGD
MHM	MK1	MLC	MLY	MMP	MO2	MO3	MO4	MO5	MO6	MOH	MP1
MPD	MPD	MPO	MSE	MTE	MTO	MVA	MYA	MYR	MYR	NAC	NAD
NAD	NAE	NAG	NAG	NAH	NAI	NAI	NAJ	NAP	NAP	NAQ	NAX
NBD	NBP	NCO	NDA	NDC	NDO	NDP	NDP	NGA	NGA	NH2	NH4
NHD	NHM	NHO	NLE	NMO	NMX	NO3	OCS	ODP	OMC	OMG	ORN
OXY	P6G	PAD	PC3	PCA	PCU	PEG	PEG	PEO	PEP	PG4	PG4
PGA	PGD	PGE	PHB	PLM	PLP	PNI	PO4	POP	POP	POR	PP9
PSU	PSU	PTR	PXZ	PYR	PYR	RET	RFL	SAH	SAM	SAR	SCA
SCD	SCN	SCO	SEO	SEP	SFP	SIA	SIN	SND	SO4	SPM	SPM
SRM	SUC	SUL	TAP	TAR	TDP	TDP	TMN	TMP	TMP	TPO	TPQ
TRS	TRS	TTP	TTP	TYS	U10	UDP	UDP	UMP	UMP	UNL	UNX
UPG	UTP	VO4	WO4	XYP	XYS	XYS	ZEM	ZID	ZN2	ZN3	ZNH

Table S2. Docking performance for the protocols tested in this study. The success rates listed are the percentage of complexes for which the top-ranked GOLD solution is within 2Å of the experimental binding mode, averaged over 20 runs. Errors in the mean are given in parentheses. Results are given for the Goldscore function and for the Chemscore function, and for three GA settings that are described in *Proteins* **52** (2003), 609-623.

	Goldscore			*Chemscore*		
	default 4	default 3	default 1	default 4	default 3	default 1
Standard (6Å)	77.6(0.8)	79.9(0.6)	80.5(0.5)	75.3(0.6)	78.9(0.6)	80.9(0.5)
Smaller site (4Å)	85.0(0.7)	86.5(0.4)	86.5(0.4)	79.8(0.6)	82.9(0.5)	84.1(0.3)
Larger site (10Å)	70.8(0.8)	78.2(0.5)	80.4(0.5)	68.5(0.5)	71.5(0.4)	76.7(0.4)
Pre-optimised polar Hs	82.6(0.5)	84.5(0.5)	86.9(0.3)	77.0(0.4)	80.5(0.6)	82.2(0.5)
X-ray waters present	96.4(0.2)	98.2(0.2)	98.6(0.1)	88.5(0.5)	93.1(0.5)	96.4(0.4)
Corina ligand geometry	70.6(0.5)	72.8(0.7)	75.2(0.4)	70.6(0.5)	73.4(0.6)	75.2(0.5)

Table S3. Docking times for the standard protocol, averaged over the 85 complexes in the Astex Diverse Set. Timings are in minutes and do not include protein initialisation time.

	Goldscore			*Chemscore*		
	default 4	default 3	default 1	default 4	default 3	default 1
Standard (6Å)	0.62	1.74	4.71	0.23	0.77	2.35

Table S4. Number of runs out of 20 for which top-ranked GOLD solution was within 0.5Å, 1.0Å, 1.5Å or 2.0Å of the experimental binding mode, using the standard protocol and the *Goldscore* function. Results are shown for all 85 complexes in the Astex Diverse Set.

Entry	RMSD<0.5Å	RMSD<1.0Å	RMSD<1.5Å	RMSD<2.0Å
1g9v	0	0	1	1
1gkc	0	14	20	20
1gm8	0	0	0	0
1gpk	7	7	7	7
1jla	0	19	20	20
1hnn	0	0	20	20
1hp0	0	10	10	10
1hq2	20	20	20	20
1hvy	0	0	0	0
1hwi	14	20	20	20
1hww	20	20	20	20
1ia1	18	20	20	20
1ig3	0	2	20	20
1j3j	7	20	20	20
1jd0	0	0	0	20
1jje	0	10	10	12
1ke5	8	20	20	20
1k3u	20	20	20	20
1kzk	10	16	20	20
1l2s	0	0	0	3
1l7f	14	20	20	20
1lpz	0	15	20	20
1lrh	0	0	20	20
1m2z	10	20	20	20
1meh	0	13	20	20
1mmv	3	14	20	20

1mzc	0	19	19	20
1n1m	0	15	20	20
1n2j	1	5	5	5
1n2v	0	0	0	0
1n46	0	20	20	20
1nav	15	18	18	18
1of1	20	20	20	20
1of6	1	20	20	20
1opk	2	6	6	20
1oq5	10	20	20	20
1owe	0	0	0	4
1oyt	0	9	20	20
1p2y	0	0	5	8
1p62	0	20	20	20
1pmn	0	0	0	6
1q1g	0	4	20	20
1q41	11	20	20	20
1q4g	20	20	20	20
1r1h	0	6	20	20
1r55	18	20	20	20
1r58	0	14	14	14
1r9o	0	0	0	2
1s19	12	20	20	20
1s3v	0	20	20	20
1sg0	0	20	20	20
1sj0	1	20	20	20
1sq5	0	0	0	0
1sqn	17	20	20	20
1t40	2	20	20	20
1t46	2	20	20	20

1t9b	0	20	20	20
1tow	0	0	0	0
1tt1	0	20	20	20
1tz8	0	0	20	20
1u1c	0	18	20	20
1u4d	0	20	20	20
1uml	0	0	3	4
1unl	0	20	20	20
1uou	0	0	0	20
1v0p	0	16	17	20
1v48	7	20	20	20
1v4s	20	20	20	20
1vcj	0	12	20	20
1w1p	20	20	20	20
1w2g	0	4	20	20
1x8x	0	20	20	20
1xm6	0	0	0	0
1xoq	0	0	11	20
1xoz	7	20	20	20
1y6b	11	20	20	20
1ygc	0	1	1	1
1yqy	4	20	20	20
1yv3	19	20	20	20
1yvf	0	0	0	0
1ywr	4	19	19	19
1z95	0	20	20	20
2bm2	0	17	17	17
2br1	0	0	1	17
2bsm	0	20	20	20

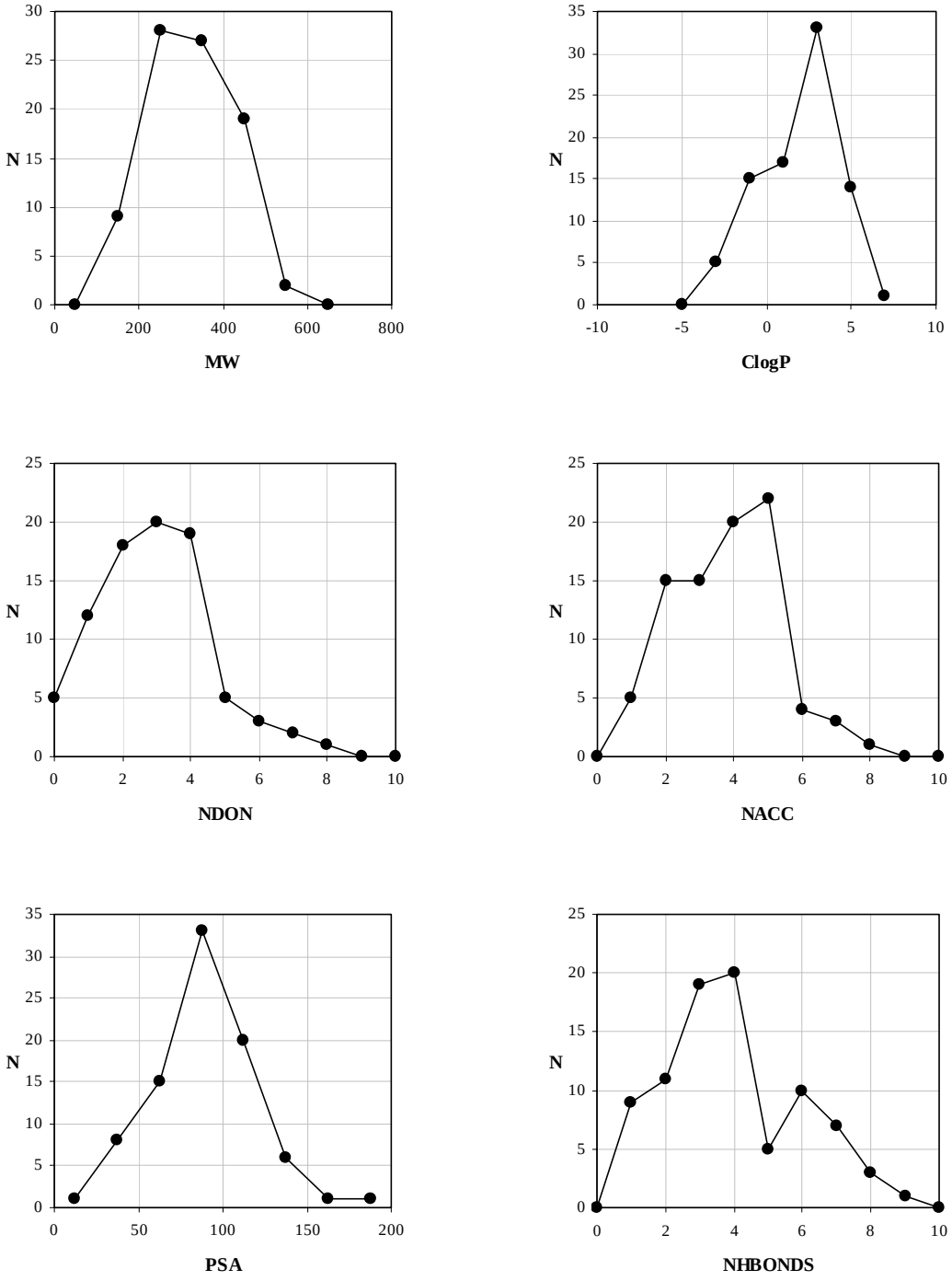

Figure S1. Distributions for ligand molecular weight (MW), ligand calculated logP (CLogP), number of donors in ligand (NDON), number of acceptors in ligand (NACC), ligand polar surface area (PSA) in $Å^2$, and number of direct hydrogen bonds between protein and ligand (NHBONDS) for the Astex Diverse Set.