



Full wwPDB X-ray Structure Validation Report ⓘ

Dec 12, 2023 – 11:10 am GMT

PDB ID : 8BX2
Title : Oct4/Sox2 protein:DNA complex
Authors : Chojnowski, G.; Wilmanns, M.; Round, E.
Deposited on : 2022-12-07
Resolution : 3.14 Å(reported)

This is a Full wwPDB X-ray Structure Validation Report for a publicly released PDB entry.

We welcome your comments at validation@mail.wwpdb.org

A user guide is available at

<https://www.wwpdb.org/validation/2017/XrayValidationReportHelp>

with specific help available everywhere you see the ⓘ symbol.

The types of validation reports are described at

<http://www.wwpdb.org/validation/2017/FAQs#types>.

The following versions of software and data (see [references ⓘ](#)) were used in the production of this report:

MolProbity : 4.02b-467
Xtriage (Phenix) : 1.13
EDS : 2.36
Percentile statistics : 20191225.v01 (using entries in the PDB archive December 25th 2019)
Refmac : 5.8.0158
CCP4 : 7.0.044 (Gargrove)
Ideal geometry (proteins) : Engh & Huber (2001)
Ideal geometry (DNA, RNA) : Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP) : 2.36

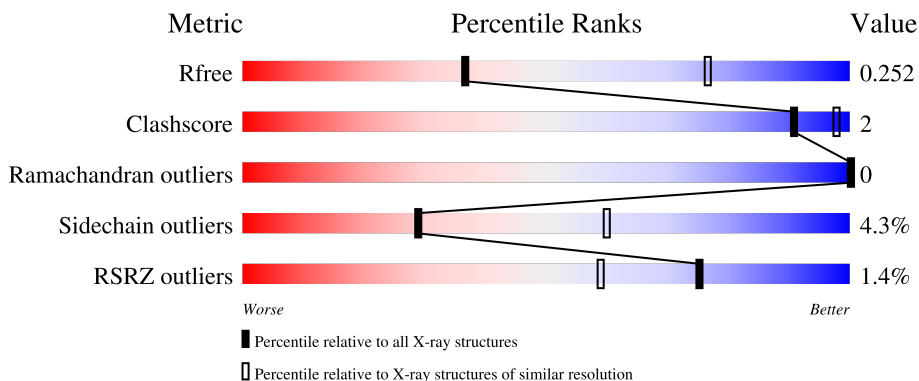
1 Overall quality at a glance i

The following experimental techniques were used to determine the structure:

X-RAY DIFFRACTION

The reported resolution of this entry is 3.14 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	130704	1626 (3.18-3.10)
Clashscore	141614	1735 (3.18-3.10)
Ramachandran outliers	138981	1677 (3.18-3.10)
Sidechain outliers	138945	1677 (3.18-3.10)
RSRZ outliers	127900	1588 (3.18-3.10)

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments of the lower bar indicate the fraction of residues that contain outliers for ≥ 3 , 2, 1 and 0 types of geometric quality criteria respectively. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions $\leq 5\%$. The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

Mol	Chain	Length	Quality of chain
1	E	83	
2	A	155	
3	C	22	
4	D	22	

2 Entry composition i

There are 4 unique types of molecules in this entry. The entry contains 2298 atoms, of which 0 are hydrogens and 0 are deuteriums.

In the tables below, the ZeroOcc column contains the number of atoms modelled with zero occupancy, the AltConf column contains the number of residues with at least one atom in alternate conformation and the Trace column contains the number of residues modelled with at most 2 atoms.

- Molecule 1 is a protein called Transcription factor SOX-2.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
			Total	C	N	O	S			
1	E	76	640	399	132	104	5	0	0	0

There are 3 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
E	-2	GLY	-	expression tag	UNP P48432
E	-1	ALA	-	expression tag	UNP P48432
E	0	MET	-	expression tag	UNP P48432

- Molecule 2 is a protein called POU domain, class 5, transcription factor 1.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
			Total	C	N	O	S			
2	A	100	756	480	128	146	2	0	0	0

There are 8 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	-2	GLY	-	expression tag	UNP P20263
A	-1	ALA	-	expression tag	UNP P20263
A	0	MET	-	expression tag	UNP P20263
A	48	SER	CYS	conflict	UNP P20263
A	61	SER	CYS	conflict	UNP P20263
A	84	SER	CYS	conflict	UNP P20263
A	115	SER	CYS	conflict	UNP P20263
A	142	SER	CYS	conflict	UNP P20263

- Molecule 3 is a DNA chain called UTF1_r.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
			Total	C	N	O	P			
3	C	22	453	217	80	134	22	0	0	0

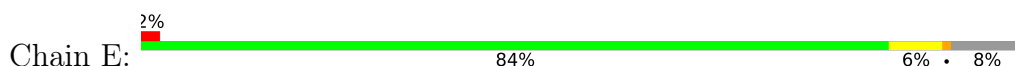
- Molecule 4 is a DNA chain called UTF1_f.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
			Total	C	N	O	P			
4	D	22	449	215	82	130	22	0	0	0

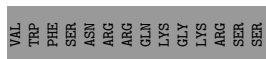
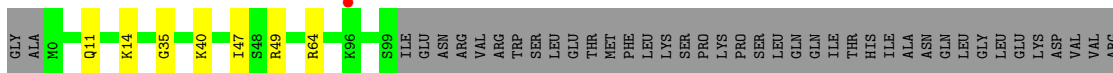
3 Residue-property plots [i](#)

These plots are drawn for all protein, RNA, DNA and oligosaccharide chains in the entry. The first graphic for a chain summarises the proportions of the various outlier classes displayed in the second graphic. The second graphic shows the sequence view annotated by issues in geometry and electron density. Residues are color-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. A red dot above a residue indicates a poor fit to the electron density ($RSRZ > 2$). Stretches of 2 or more consecutive residues without any outlier are shown as a green connector. Residues present in the sample, but not in the model, are shown in grey.

- Molecule 1: Transcription factor SOX-2



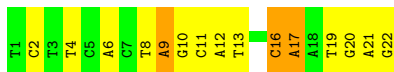
- Molecule 2: POU domain, class 5, transcription factor 1



- Molecule 3: UTF1_r



- Molecule 4: UTF1_f



4 Data and refinement statistics

Property	Value	Source
Space group	P 63	Depositor
Cell constants a, b, c, α , β , γ	163.17Å 163.17Å 39.79Å 90.00° 90.00° 120.00°	Depositor
Resolution (Å)	81.72 – 3.14 81.58 – 3.14	Depositor EDS
% Data completeness (in resolution range)	77.5 (81.72-3.14) 77.7 (81.58-3.14)	Depositor EDS
R_{merge}	0.11	Depositor
R_{sym}	(Not available)	Depositor
$\langle I/\sigma(I) \rangle$ ¹	1.25 (at 3.13Å)	Xtrriage
Refinement program	REFMAC 5.8.0267	Depositor
R, R_{free}	0.199 , 0.252 0.198 , 0.252	Depositor DCC
R_{free} test set	368 reflections (4.33%)	wwPDB-VP
Wilson B-factor (Å ²)	127.0	Xtrriage
Anisotropy	0.265	Xtrriage
Bulk solvent k_{sol} (e/Å ³), B_{sol} (Å ²)	0.28 , 99.8	EDS
L-test for twinning ²	$\langle L \rangle = 0.47$, $\langle L^2 \rangle = 0.30$	Xtrriage
Estimated twinning fraction	0.048 for h,-h-k,-l	Xtrriage
F_o, F_c correlation	0.96	EDS
Total number of atoms	2298	wwPDB-VP
Average B, all atoms (Å ²)	149.0	wwPDB-VP

Xtrriage's analysis on translational NCS is as follows: *The largest off-origin peak in the Patterson function is 4.72% of the height of the origin peak. No significant pseudotranslation is detected.*

¹Intensities estimated from amplitudes.

²Theoretical values of $\langle |L| \rangle$, $\langle L^2 \rangle$ for acentric reflections are 0.5, 0.333 respectively for untwinned datasets, and 0.375, 0.2 for perfectly twinned datasets.

5 Model quality i

5.1 Standard geometry i

The Z score for a bond length (or angle) is the number of standard deviations the observed value is removed from the expected value. A bond length (or angle) with $|Z| > 5$ is considered an outlier worth inspection. RMSZ is the root-mean-square of all Z scores of the bond lengths (or angles).

Mol	Chain	Bond lengths		Bond angles	
		RMSZ	# Z >5	RMSZ	# Z >5
1	E	0.63	0/655	1.37	3/874 (0.3%)
2	A	0.46	0/763	1.25	2/1031 (0.2%)
3	C	1.13	0/507	1.72	6/781 (0.8%)
4	D	1.10	0/503	1.79	13/773 (1.7%)
All	All	0.83	0/2428	1.52	24/3459 (0.7%)

There are no bond length outliers.

All (24) bond angle outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(°)	Ideal(°)
4	D	8	DT	P-O3'-C3'	-10.25	107.40	119.70
3	C	20	DA	P-O3'-C3'	9.44	131.03	119.70
3	C	16	DA	P-O3'-C3'	8.19	129.53	119.70
1	E	75	ARG	CG-CD-NE	7.43	127.41	111.80
4	D	17	DA	P-O3'-C3'	-7.38	110.84	119.70
4	D	20	DG	O5'-P-OP2	-7.17	99.24	105.70
4	D	6	DA	O4'-C1'-N9	6.67	112.67	108.00
1	E	77	LYS	CB-CA-C	6.52	123.44	110.40
3	C	1	DA	O4'-C1'-N9	6.35	112.45	108.00
3	C	5	DA	P-O3'-C3'	-5.98	112.52	119.70
4	D	13	DT	O4'-C1'-N1	5.93	112.15	108.00
4	D	2	DC	P-O3'-C3'	5.87	126.75	119.70
1	E	50	ARG	NE-CZ-NH1	5.64	123.12	120.30
4	D	4	DT	O4'-C1'-N1	5.56	111.89	108.00
2	A	11	GLN	CB-CA-C	5.49	121.38	110.40
4	D	12	DA	P-O3'-C3'	-5.46	113.15	119.70
2	A	49	ARG	NE-CZ-NH1	5.35	122.97	120.30
4	D	9	DA	OP2-P-O3'	5.33	116.92	105.20
4	D	16	DC	O5'-P-OP2	-5.30	100.93	105.70
3	C	2	DC	O4'-C1'-N1	5.27	111.69	108.00
4	D	2	DC	O4'-C1'-N1	5.26	111.68	108.00
4	D	13	DT	P-O3'-C3'	-5.20	113.46	119.70
4	D	11	DC	O3'-P-O5'	-5.05	94.40	104.00

Continued on next page...

Continued from previous page...

Mol	Chain	Res	Type	Atoms	Z	Observed(°)	Ideal(°)
3	C	1	DA	C1'-O4'-C4'	-5.03	105.07	110.10

There are no chirality outliers.

There are no planarity outliers.

5.2 Too-close contacts [i](#)

In the following table, the Non-H and H(model) columns list the number of non-hydrogen atoms and hydrogen atoms in the chain respectively. The H(added) column lists the number of hydrogen atoms added and optimized by MolProbity. The Clashes column lists the number of clashes within the asymmetric unit, whereas Symm-Clashes lists symmetry-related clashes.

Mol	Chain	Non-H	H(model)	H(added)	Clashes	Symm-Clashes
1	E	640	0	642	1	0
2	A	756	0	736	1	0
3	C	453	0	251	2	0
4	D	449	0	249	4	0
All	All	2298	0	1878	7	0

The all-atom clashscore is defined as the number of clashes found per 1000 atoms (including hydrogen atoms). The all-atom clashscore for this structure is 2.

All (7) close contacts within the same asymmetric unit are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Interatomic distance (Å)	Clash overlap (Å)
3:C:10:DT:H2''	3:C:11:DA:C8	2.46	0.51
4:D:16:DC:H2''	4:D:17:DA:C8	2.49	0.47
3:C:2:DC:C6	3:C:3:DT:H72	2.49	0.46
1:E:15:ARG:HD3	4:D:19:DT:OP1	2.16	0.46
4:D:9:DA:H1'	4:D:10:DG:H5'	2.00	0.43
2:A:35:GLY:HA2	2:A:40:LYS:O	2.20	0.41
4:D:21:DA:H2''	4:D:22:DG:C8	2.55	0.41

There are no symmetry-related clashes.

5.3 Torsion angles [i](#)

5.3.1 Protein backbone [i](#)

In the following table, the Percentiles column shows the percent Ramachandran outliers of the chain as a percentile score with respect to all X-ray entries followed by that with respect to entries of similar resolution.

The Analysed column shows the number of residues for which the backbone conformation was analysed, and the total number of residues.

Mol	Chain	Analysed	Favoured	Allowed	Outliers	Percentiles	
1	E	74/83 (89%)	73 (99%)	1 (1%)	0	100	100
2	A	98/155 (63%)	93 (95%)	5 (5%)	0	100	100
All	All	172/238 (72%)	166 (96%)	6 (4%)	0	100	100

There are no Ramachandran outliers to report.

5.3.2 Protein sidechains [i](#)

In the following table, the Percentiles column shows the percent sidechain outliers of the chain as a percentile score with respect to all X-ray entries followed by that with respect to entries of similar resolution.

The Analysed column shows the number of residues for which the sidechain conformation was analysed, and the total number of residues.

Mol	Chain	Analysed	Rotameric	Outliers	Percentiles	
1	E	65/74 (88%)	62 (95%)	3 (5%)	27	58
2	A	76/140 (54%)	73 (96%)	3 (4%)	32	63
All	All	141/214 (66%)	135 (96%)	6 (4%)	29	60

All (6) residues with a non-rotameric sidechain are listed below:

Mol	Chain	Res	Type
1	E	36	ARG
1	E	62	LEU
1	E	77	LYS
2	A	14	LYS
2	A	47	ILE
2	A	64	ARG

Sometimes sidechains can be flipped to improve hydrogen bonding and reduce clashes. There are

no such sidechains identified.

5.3.3 RNA [i](#)

There are no RNA molecules in this entry.

5.4 Non-standard residues in protein, DNA, RNA chains [i](#)

There are no non-standard protein/DNA/RNA residues in this entry.

5.5 Carbohydrates [i](#)

There are no monosaccharides in this entry.

5.6 Ligand geometry [i](#)

There are no ligands in this entry.

5.7 Other polymers [i](#)

There are no such residues in this entry.

5.8 Polymer linkage issues [i](#)

There are no chain breaks in this entry.

6 Fit of model and data [i](#)

6.1 Protein, DNA and RNA chains [i](#)

In the following table, the column labelled ‘#RSRZ> 2’ contains the number (and percentage) of RSRZ outliers, followed by percent RSRZ outliers for the chain as percentile scores relative to all X-ray entries and entries of similar resolution. The OWAB column contains the minimum, median, 95th percentile and maximum values of the occupancy-weighted average B-factor per residue. The column labelled ‘Q< 0.9’ lists the number of (and percentage) of residues with an average occupancy less than 0.9.

Mol	Chain	Analysed	<RSRZ>	#RSRZ>2	OWAB(Å ²)	Q<0.9
1	E	76/83 (91%)	0.21	2 (2%) 56 35	99, 129, 175, 199	0
2	A	100/155 (64%)	-0.19	1 (1%) 82 70	94, 127, 213, 263	0
3	C	22/22 (100%)	-0.90	0 100 100	97, 162, 240, 285	0
4	D	22/22 (100%)	-0.86	0 100 100	107, 156, 252, 254	0
All	All	220/282 (78%)	-0.19	3 (1%) 75 59	94, 132, 229, 285	0

All (3) RSRZ outliers are listed below:

Mol	Chain	Res	Type	RSRZ
2	A	96	LYS	2.4
1	E	9	ALA	2.3
1	E	17	GLN	2.3

6.2 Non-standard residues in protein, DNA, RNA chains [i](#)

There are no non-standard protein/DNA/RNA residues in this entry.

6.3 Carbohydrates [i](#)

There are no monosaccharides in this entry.

6.4 Ligands [i](#)

There are no ligands in this entry.

6.5 Other polymers [i](#)

There are no such residues in this entry.