



# Full wwPDB X-ray Structure Validation Report ⓘ

Mar 8, 2018 – 07:39 pm GMT

PDB ID : 5K5L  
Title : Homo sapiens CCCTC-binding factor (CTCF) ZnF6-8 and H19 sequence DNA complex structure  
Authors : Hashimoto, H.; Cheng, X.  
Deposited on : 2016-05-23  
Resolution : 3.12 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](mailto: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 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 : trunk30967  
Percentile statistics : 20171227.v01 (using entries in the PDB archive December 27th 2017)  
Refmac : 5.8.0158  
CCP4 : 7.0 (Gargrove)  
Ideal geometry (proteins) : Engh & Huber (2001)  
Ideal geometry (DNA, RNA) : Parkinson et al. (1996)  
Validation Pipeline (wwPDB-VP) : trunk30967

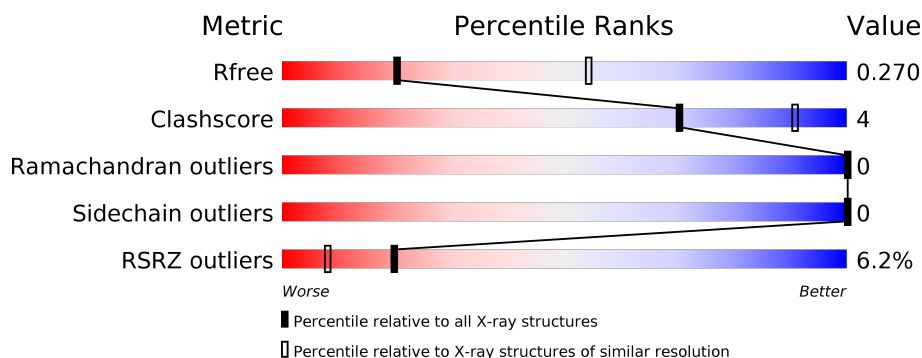
# 1 Overall quality at a glance

The following experimental techniques were used to determine the structure:

*X-RAY DIFFRACTION*

The reported resolution of this entry is 3.12 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))
$R_{free}$	111664	1111 (3.14-3.10)
Clashscore	122126	1202 (3.14-3.10)
Ramachandran outliers	120053	1163 (3.14-3.10)
Sidechain outliers	120020	1163 (3.14-3.10)
RSRZ outliers	108989	1084 (3.14-3.10)

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 on the lower bar indicate the fraction of residues that contain outliers for  $\geq 3$ , 2, 1 and 0 types of geometric quality criteria. 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	A	11	<div><div></div><div>55%45%</div></div>
1	C	11	<div><div></div><div>36%64%</div></div>
2	B	11	<div><div></div><div>73%18%9%</div></div>
2	D	11	<div><div></div><div>100%</div></div>
3	E	93	<div><div>5%</div><div>57%40%</div><div>.</div></div>
3	F	93	<div><div>6%</div><div>54%8%39%</div></div>

Continued on next page...

Continued from previous page...

Mol	Chain	Length	Quality of chain
3	G	93	<div><div></div><div>4%</div><div>86%</div><div>•</div><div>10%</div></div>

## 2 Entry composition [i](#)

There are 5 unique types of molecules in this entry. The entry contains 2507 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 DNA chain called DNA (5'-D(\*GP\*TP\*TP\*GP\*CP\*CP\*GP\*CP\*GP\*TP\*G)-3').

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
1	A	11	Total	C	N	O	P	0	0	0
			206	97	35	64	10			
1	C	11	Total	C	N	O	P	0	0	0
			224	107	40	67	10			

- Molecule 2 is a DNA chain called DNA (5'-D(P\*AP\*CP\*GP\*CP\*GP\*GP\*CP\*AP\*AP\*C)-3').

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
2	B	10	Total	C	N	O	P	0	0	0
			205	96	42	57	10			
2	D	11	Total	C	N	O	P	0	0	0
			221	105	45	61	10			

- Molecule 3 is a protein called Transcriptional repressor CTCF.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
3	E	56	Total	C	N	O	S	0	0	0
			469	293	94	78	4			
3	F	57	Total	C	N	O	S	0	0	0
			478	298	95	81	4			
3	G	84	Total	C	N	O	S	0	0	0
			695	433	136	118	8			

There are 15 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
E	400	GLY	-	expression tag	UNP P49711
E	401	PRO	-	expression tag	UNP P49711
E	402	LEU	-	expression tag	UNP P49711
E	403	GLY	-	expression tag	UNP P49711

*Continued on next page...*

*Continued from previous page...*

Chain	Residue	Modelled	Actual	Comment	Reference
E	404	SER	-	expression tag	UNP P49711
F	400	GLY	-	expression tag	UNP P49711
F	401	PRO	-	expression tag	UNP P49711
F	402	LEU	-	expression tag	UNP P49711
F	403	GLY	-	expression tag	UNP P49711
F	404	SER	-	expression tag	UNP P49711
G	400	GLY	-	expression tag	UNP P49711
G	401	PRO	-	expression tag	UNP P49711
G	402	LEU	-	expression tag	UNP P49711
G	403	GLY	-	expression tag	UNP P49711
G	404	SER	-	expression tag	UNP P49711

- Molecule 4 is ZINC ION (three-letter code: ZN) (formula: Zn).

Mol	Chain	Residues	Atoms	ZeroOcc	AltConf
4	G	3	Total Zn 3 3	0	0
4	F	2	Total Zn 2 2	0	0
4	E	2	Total Zn 2 2	0	0

- Molecule 5 is water.

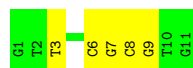
Mol	Chain	Residues	Atoms	ZeroOcc	AltConf
5	E	1	Total O 1 1	0	0
5	G	1	Total O 1 1	0	0

### 3 Residue-property plots

These plots are drawn for all protein, RNA and DNA 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.

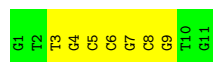
- Molecule 1: DNA (5'-D(\*GP\*TP\*TP\*GP\*CP\*CP\*GP\*CP\*GP\*TP\*G)-3')

Chain A: 



- Molecule 1: DNA (5'-D(\*GP\*TP\*TP\*GP\*CP\*CP\*GP\*CP\*GP\*TP\*G)-3')

Chain C: 



- Molecule 2: DNA (5'-D(P\*AP\*CP\*GP\*CP\*GP\*GP\*CP\*AP\*AP\*C)-3')

Chain B: 



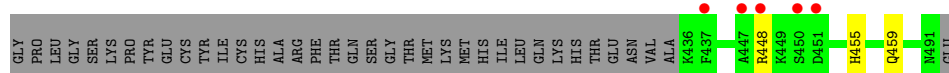
- Molecule 2: DNA (5'-D(P\*AP\*CP\*GP\*CP\*GP\*GP\*CP\*AP\*AP\*C)-3')

Chain D: 

There are no outlier residues recorded for this chain.

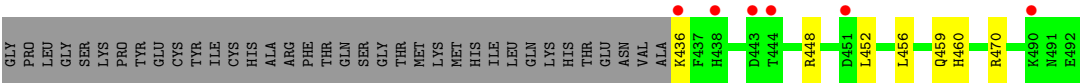
- Molecule 3: Transcriptional repressor CTCF

Chain E: 

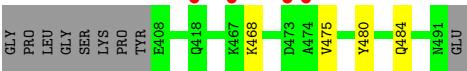
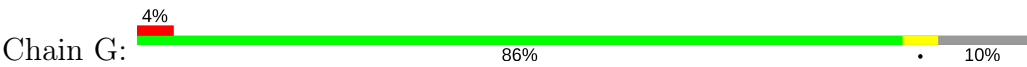


- Molecule 3: Transcriptional repressor CTCF

Chain F: 



● Molecule 3: Transcriptional repressor CTCF



## 4 Data and refinement statistics

Property	Value	Source
Space group	P 21 21 2	Depositor
Cell constants a, b, c, $\alpha$ , $\beta$ , $\gamma$	126.79Å 52.40Å 69.14Å 90.00° 90.00° 90.00°	Depositor
Resolution (Å)	46.73 – 3.12 46.73 – 3.13	Depositor EDS
% Data completeness (in resolution range)	96.0 (46.73-3.12) 93.1 (46.73-3.13)	Depositor EDS
$R_{merge}$	0.11	Depositor
$R_{sym}$	(Not available)	Depositor
$\langle I/\sigma(I) \rangle$ <sup>1</sup>	2.18 (at 3.12Å)	Xtriage
Refinement program	PHENIX (dev_2313: ???)	Depositor
R, $R_{free}$	0.236 , 0.270 0.236 , 0.270	Depositor DCC
$R_{free}$ test set	421 reflections (5.04%)	wwPDB-VP
Wilson B-factor (Å <sup>2</sup> )	66.9	Xtriage
Anisotropy	0.580	Xtriage
Bulk solvent $k_{sol}$ (e/Å <sup>3</sup> ), $B_{sol}$ (Å <sup>2</sup> )	0.30 , 37.6	EDS
L-test for twinning <sup>2</sup>	$\langle  L  \rangle = 0.49$ , $\langle L^2 \rangle = 0.33$	Xtriage
Estimated twinning fraction	No twinning to report.	Xtriage
$F_o, F_c$ correlation	0.91	EDS
Total number of atoms	2507	wwPDB-VP
Average B, all atoms (Å <sup>2</sup> )	86.0	wwPDB-VP

Xtriage's analysis on translational NCS is as follows: *The largest off-origin peak in the Patterson function is 7.00% of the height of the origin peak. No significant pseudotranslation is detected.*

<sup>1</sup>Intensities estimated from amplitudes.

<sup>2</sup>Theoretical 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

### 5.1 Standard geometry

Bond lengths and bond angles in the following residue types are not validated in this section: ZN

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	A	0.55	0/229	0.98	0/353
1	C	0.58	0/250	0.94	0/385
2	B	0.58	0/230	0.76	0/352
2	D	0.57	0/248	0.72	0/380
3	E	0.24	0/481	0.38	0/642
3	F	0.25	0/490	0.37	0/654
3	G	0.25	0/712	0.39	0/952
All	All	0.39	0/2640	0.62	0/3718

There are no bond length outliers.

There are no bond angle outliers.

There are no chirality outliers.

There are no planarity outliers.

### 5.2 Too-close contacts

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	A	206	0	114	4	0
1	C	224	0	126	6	0
2	B	205	0	111	1	0
2	D	221	0	123	0	0
3	E	469	0	457	2	0
3	F	478	0	463	5	0

*Continued on next page...*

*Continued from previous page...*

Mol	Chain	Non-H	H(model)	H(added)	Clashes	Symm-Clashes
3	G	695	0	676	2	0
4	E	2	0	0	0	0
4	F	2	0	0	0	0
4	G	3	0	0	0	0
5	E	1	0	0	0	0
5	G	1	0	0	0	0
All	All	2507	0	2070	16	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 4.

All (16) 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)
1:A:3:DT:OP1	3:F:470:ARG:NH1	2.28	0.67
3:F:436:LYS:HD3	3:F:452:LEU:HD23	1.84	0.58
1:A:8:DC:H2''	1:A:9:DG:C8	2.43	0.54
3:G:468:LYS:HG2	3:G:475:VAL:HG22	1.88	0.54
1:A:9:DG:N7	3:E:448:ARG:NH1	2.59	0.51
3:E:455:HIS:CE1	3:E:459:GLN:HG3	2.47	0.49
2:B:9:DA:H2''	2:B:10:DA:C8	2.48	0.49
1:A:6:DC:H2''	1:A:7:DG:C8	2.48	0.48
1:C:7:DG:N7	3:F:448:ARG:NH2	2.60	0.48
1:C:5:DC:H2''	1:C:6:DC:O5'	2.14	0.47
3:F:456:LEU:HD23	3:F:460:HIS:HD2	1.79	0.47
3:G:480:TYR:O	3:G:484:GLN:HG2	2.15	0.47
1:C:3:DT:H2''	1:C:4:DG:C8	2.50	0.47
1:C:6:DC:H2''	1:C:7:DG:C8	2.50	0.46
1:C:8:DC:H2''	1:C:9:DG:C8	2.52	0.44
1:C:3:DT:P	3:F:459:GLN:HE22	2.42	0.43

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	
3	E	54/93 (58%)	54 (100%)	0	0	100	100
3	F	55/93 (59%)	53 (96%)	2 (4%)	0	100	100
3	G	82/93 (88%)	77 (94%)	5 (6%)	0	100	100
All	All	191/279 (68%)	184 (96%)	7 (4%)	0	100	100

There are no Ramachandran outliers to report.

### 5.3.2 Protein sidechains ⓘ

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	
3	E	51/83 (61%)	51 (100%)	0	100	100
3	F	52/83 (63%)	52 (100%)	0	100	100
3	G	76/83 (92%)	76 (100%)	0	100	100
All	All	179/249 (72%)	179 (100%)	0	100	100

There are no protein residues with a non-rotameric sidechain to report.

Some sidechains can be flipped to improve hydrogen bonding and reduce clashes. There are no such sidechains identified.

### 5.3.3 RNA ⓘ

There are no RNA molecules in this entry.

## 5.4 Non-standard residues in protein, DNA, RNA chains ⓘ

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

## 5.5 Carbohydrates [i](#)

There are no carbohydrates in this entry.

## 5.6 Ligand geometry [i](#)

Of 7 ligands modelled in this entry, 7 are monoatomic - leaving 0 for Mogul analysis.

There are no bond length outliers.

There are no bond angle outliers.

There are no chirality outliers.

There are no torsion outliers.

There are no ring outliers.

No monomer is involved in short contacts.

## 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, 95<sup>th</sup> 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 <sup>2</sup> )	Q<0.9
1	A	11/11 (100%)	0.38	0	100 100	80, 95, 149, 174	0
1	C	11/11 (100%)	0.29	0	100 100	71, 80, 138, 150	0
2	B	10/11 (90%)	0.46	0	100 100	74, 97, 132, 143	0
2	D	11/11 (100%)	0.33	0	100 100	57, 79, 125, 139	0
3	E	56/93 (60%)	0.32	5 (8%)	9 4	46, 77, 110, 119	0
3	F	57/93 (61%)	0.52	6 (10%)	6 2	61, 87, 121, 127	0
3	G	84/93 (90%)	0.37	4 (4%)	30 14	43, 75, 109, 133	0
All	All	240/323 (74%)	0.39	15 (6%)	20 8	43, 82, 125, 174	0

All (15) RSRZ outliers are listed below:

Mol	Chain	Res	Type	RSRZ
3	F	436	LYS	8.7
3	G	474	ALA	4.1
3	F	443	ASP	3.6
3	E	450	SER	3.6
3	F	438	HIS	3.6
3	E	448	ARG	2.8
3	E	437	PHE	2.6
3	G	473	ASP	2.6
3	G	467	LYS	2.4
3	E	451	ASP	2.3
3	F	490	LYS	2.3
3	G	418	GLN	2.2
3	E	447	ALA	2.2
3	F	451	ASP	2.1
3	F	444	THR	2.0

## 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 carbohydrates in this entry.

## 6.4 Ligands [i](#)

In the following table, the Atoms column lists the number of modelled atoms in the group and the number defined in the chemical component dictionary. The B-factors column lists the minimum, median, 95<sup>th</sup> percentile and maximum values of B factors of atoms in the group. The column labelled 'Q< 0.9' lists the number of atoms with occupancy less than 0.9.

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors( $\text{\AA}^2$ )	Q<0.9
4	ZN	F	502	1/1	0.79	0.14	103,103,103,103	0
4	ZN	G	503	1/1	0.94	0.10	84,84,84,84	0
4	ZN	G	502	1/1	0.94	0.04	100,100,100,100	0
4	ZN	E	502	1/1	0.95	0.09	66,66,66,66	0
4	ZN	F	501	1/1	0.95	0.09	89,89,89,89	0
4	ZN	E	501	1/1	0.98	0.07	92,92,92,92	0
4	ZN	G	501	1/1	0.99	0.12	54,54,54,54	0

## 6.5 Other polymers [i](#)

There are no such residues in this entry.