



# wwPDB X-ray Structure Validation Summary Report ⓘ

Mar 29, 2016 – 01:54 PM EDT

PDB ID : 5DS6  
Title : Crystal structure the Escherichia coli Cas1-Cas2 complex bound to protospacer DNA with splayed ends  
Authors : Nunez, J.K.; Harrington, L.B.; Kranzusch, P.J.; Engelman, A.N.; Doudna, J.A.  
Deposited on : 2015-09-16  
Resolution : 3.35 Å(reported)

This is a wwPDB X-ray Structure Validation Summary 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  
<http://wwpdb.org/validation/2016/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  
Mogul : unknown  
Xtriage (Phenix) : 1.9-1692  
EDS : rb-20027107  
Percentile statistics : 20151230.v01 (using entries in the PDB archive December 30th 2015)  
Refmac : 5.8.0135  
CCP4 : 6.5.0  
Ideal geometry (proteins) : Engh & Huber (2001)  
Ideal geometry (DNA, RNA) : Parkinson et al. (1996)  
Validation Pipeline (wwPDB-VP) : rb-20027107

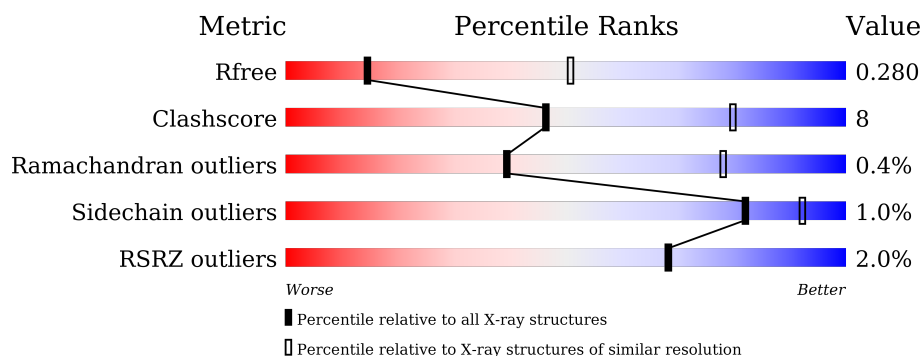
# 1 Overall quality at a glance

The following experimental techniques were used to determine the structure:

*X-RAY DIFFRACTION*

The reported resolution of this entry is 3.35 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))
$R_{free}$	91344	1005 (3.42-3.30)
Clashscore	102246	1076 (3.42-3.30)
Ramachandran outliers	100387	1059 (3.42-3.30)
Sidechain outliers	100360	1058 (3.42-3.30)
RSRZ outliers	91569	1010 (3.42-3.30)

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	306	<div> <div>3%</div> <div>70% 13% . 16%</div> </div>
1	B	306	<div> <div>67% 20% . 12%</div> </div>
1	C	306	<div> <div>3%</div> <div>67% 15% . 17%</div> </div>
1	D	306	<div> <div>3%</div> <div>66% 17% 17%</div> </div>
2	E	104	<div> <div>75% 14% 11%</div> </div>
2	F	104	<div> <div>66% 24% 10%</div> </div>

*Continued on next page...*

Continued from previous page...

Mol	Chain	Length	Quality of chain
3	G	33	<div><div><div></div><div></div><div></div></div><div><div>3%</div><div>52%</div><div>36%</div><div>12%</div></div></div>
4	H	33	<div><div><div></div><div></div><div></div></div><div><div>3%</div><div>76%</div><div>9%</div><div>15%</div></div></div>

## 2 Entry composition

There are 4 unique types of molecules in this entry. The entry contains 10540 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 CRISPR-associated endonuclease Cas1.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
1	A	256	Total	C	N	O	S	0	0	0
			1953	1249	347	350	7			
1	B	268	Total	C	N	O	S	0	0	0
			2061	1319	367	368	7			
1	C	254	Total	C	N	O	S	0	0	0
			1941	1241	345	348	7			
1	D	253	Total	C	N	O	S	0	0	0
			1949	1253	344	345	7			

There are 4 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	0	SER	-	expression tag	UNP Q46896
B	0	SER	-	expression tag	UNP Q46896
C	0	SER	-	expression tag	UNP Q46896
D	0	SER	-	expression tag	UNP Q46896

- Molecule 2 is a protein called CRISPR-associated endoribonuclease Cas2.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
2	E	93	Total	C	N	O	S	0	0	0
			732	470	127	131	4			
2	F	94	Total	C	N	O	S	0	0	0
			739	475	128	132	4			

There are 20 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
E	0	MET	-	initiating methionine	UNP P45956
E	95	GLY	-	expression tag	UNP P45956
E	96	SER	-	expression tag	UNP P45956
E	97	SER	-	expression tag	UNP P45956

*Continued on next page...*

*Continued from previous page...*

Chain	Residue	Modelled	Actual	Comment	Reference
E	98	GLU	-	expression tag	UNP P45956
E	99	ASN	-	expression tag	UNP P45956
E	100	LEU	-	expression tag	UNP P45956
E	101	TYR	-	expression tag	UNP P45956
E	102	PHE	-	expression tag	UNP P45956
E	103	GLN	-	expression tag	UNP P45956
F	0	MET	-	initiating methionine	UNP P45956
F	95	GLY	-	expression tag	UNP P45956
F	96	SER	-	expression tag	UNP P45956
F	97	SER	-	expression tag	UNP P45956
F	98	GLU	-	expression tag	UNP P45956
F	99	ASN	-	expression tag	UNP P45956
F	100	LEU	-	expression tag	UNP P45956
F	101	TYR	-	expression tag	UNP P45956
F	102	PHE	-	expression tag	UNP P45956
F	103	GLN	-	expression tag	UNP P45956

- Molecule 3 is a DNA chain called DNA (29-MER).

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
3	G	29	Total	C	N	O	P	0	0	0
			601	285	120	167	29			

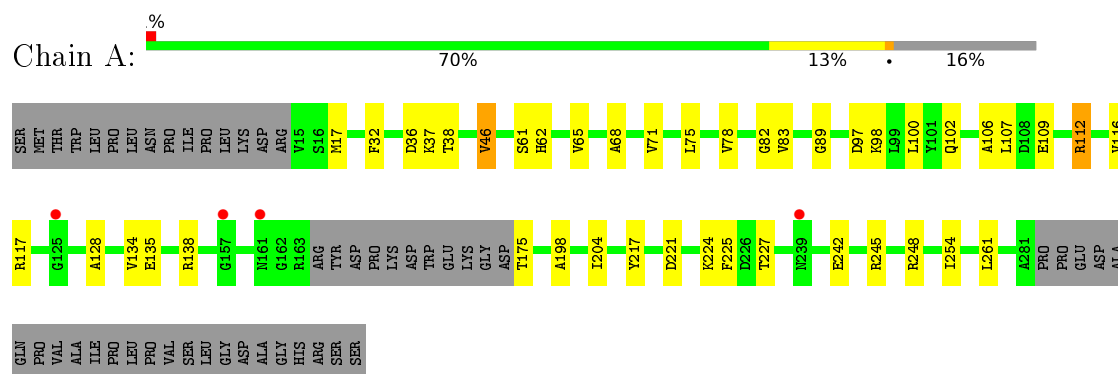
- Molecule 4 is a DNA chain called DNA (28-MER).

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
4	H	28	Total	C	N	O	P	0	0	0
			564	274	86	177	27			

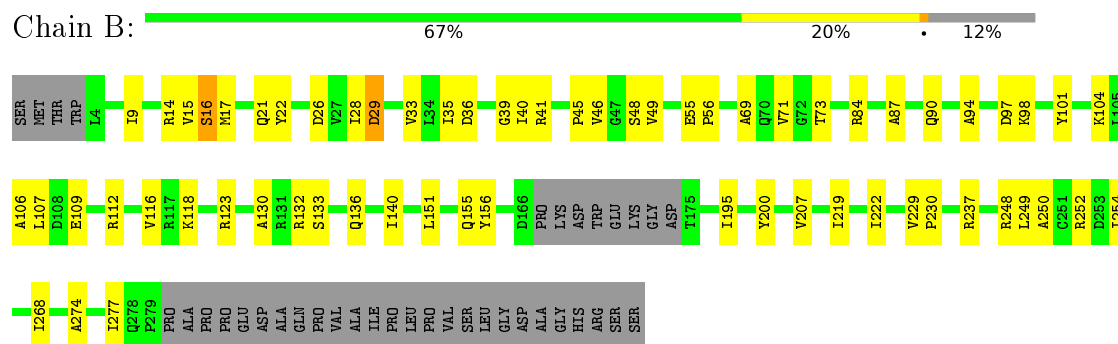
### 3 Residue-property plots [i](#)

These plots are drawn for all protein, RNA and DNA chains in the entry. The first graphic for a chain summarises the proportions of errors 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.

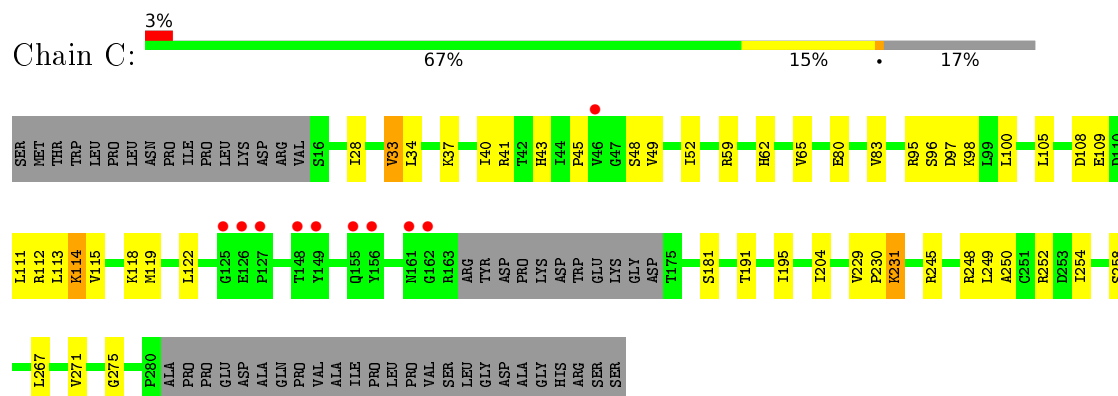
#### • Molecule 1: CRISPR-associated endonuclease Cas1



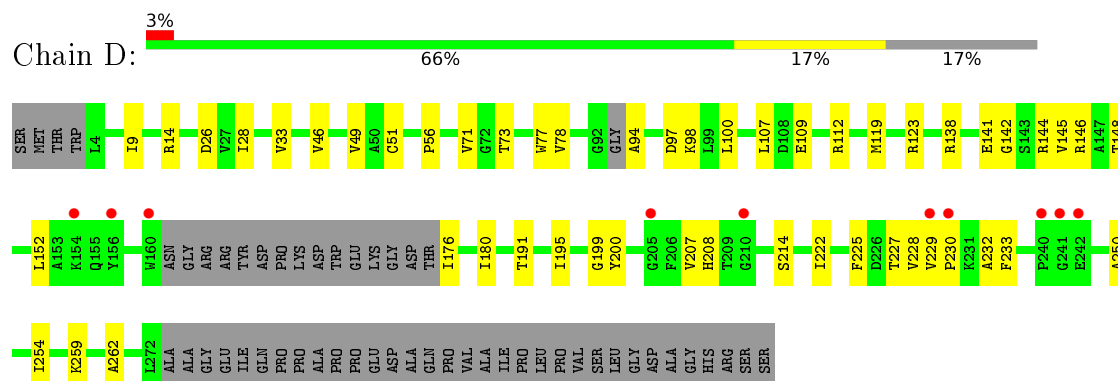
#### • Molecule 1: CRISPR-associated endonuclease Cas1



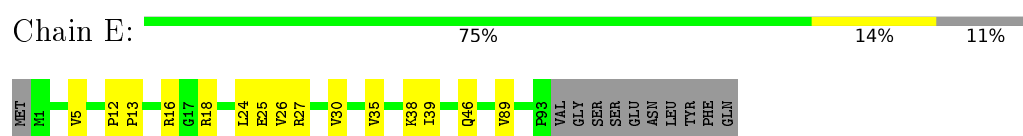
#### • Molecule 1: CRISPR-associated endonuclease Cas1



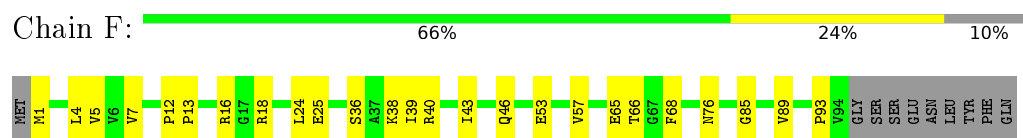
- Molecule 1: CRISPR-associated endonuclease Cas1



- Molecule 2: CRISPR-associated endoribonuclease Cas2



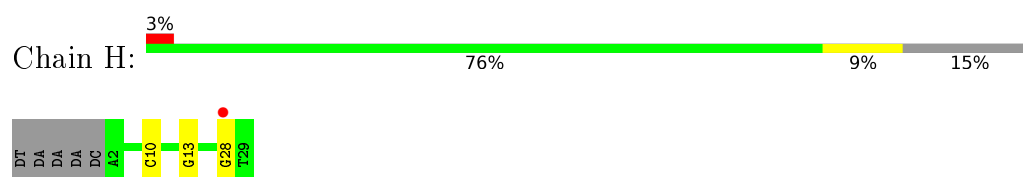
- Molecule 2: CRISPR-associated endoribonuclease Cas2



- Molecule 3: DNA (29-MER)



- Molecule 4: DNA (28-MER)



## 4 Data and refinement statistics

Property	Value	Source
Space group	P 21 21 21	Depositor
Cell constants a, b, c, $\alpha$ , $\beta$ , $\gamma$	88.02Å 123.01Å 196.01Å 90.00° 90.00° 90.00°	Depositor
Resolution (Å)	49.00 – 3.35 49.00 – 3.35	Depositor EDS
% Data completeness (in resolution range)	99.1 (49.00-3.35) 99.3 (49.00-3.35)	Depositor EDS
$R_{merge}$	0.22	Depositor
$R_{sym}$	(Not available)	Depositor
$\langle I/\sigma(I) \rangle$ <sup>1</sup>	1.25 (at 3.33Å)	Xtriage
Refinement program	PHENIX (phenix.refine: 1.9_1692)	Depositor
R, $R_{free}$	0.231 , 0.275 0.241 , 0.280	Depositor DCC
$R_{free}$ test set	3241 reflections (5.56%)	DCC
Wilson B-factor (Å <sup>2</sup> )	74.6	Xtriage
Anisotropy	0.627	Xtriage
Bulk solvent $k_{sol}$ (e/Å <sup>3</sup> ), $B_{sol}$ (Å <sup>2</sup> )	0.30 , 36.2	EDS
Estimated twinning fraction	No twinning to report.	Xtriage
L-test for twinning <sup>2</sup>	$\langle  L  \rangle = 0.40$ , $\langle L^2 \rangle = 0.22$	Xtriage
Outliers	1 of 31049 reflections (0.003%)	Xtriage
$F_o, F_c$ correlation	0.91	EDS
Total number of atoms	10540	wwPDB-VP
Average B, all atoms (Å <sup>2</sup> )	89.0	wwPDB-VP

Xtriage's analysis on translational NCS is as follows: *The largest off-origin peak in the Patterson function is 4.27% 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, 0.375 respectively for untwinned datasets, and 0.333, 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	A	0.22	0/1988	0.43	0/2695
1	B	0.22	0/2099	0.44	0/2846
1	C	0.22	0/1976	0.46	0/2678
1	D	0.22	0/1984	0.46	0/2689
2	E	0.20	0/746	0.39	0/1014
2	F	0.20	0/753	0.38	0/1024
3	G	0.51	0/677	0.82	0/1043
4	H	0.51	0/627	1.02	0/966
All	All	0.27	0/10850	0.53	0/14955

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 [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	A	1953	0	2019	29	0
1	B	2061	0	2134	39	0
1	C	1941	0	2005	33	0
1	D	1949	0	2030	39	0
2	E	732	0	747	11	0
2	F	739	0	756	17	0
3	G	601	0	325	9	0
4	H	564	0	324	3	0

*Continued on next page...*

*Continued from previous page...*

Mol	Chain	Non-H	H(model)	H(added)	Clashes	Symm-Clashes
All	All	10540	0	10340	158	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 8.

The worst 5 of 158 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:D:148:THR:HG22	1:D:152:LEU:HD11	1.50	0.91
2:E:5:VAL:HG21	2:F:5:VAL:HG21	1.57	0.85
1:A:135:GLU:HA	1:A:138:ARG:HD3	1.73	0.71
1:D:109:GLU:HA	1:D:112:ARG:HB3	1.71	0.71
1:D:138:ARG:NH1	1:D:207:VAL:O	2.19	0.71

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	A	252/306 (82%)	244 (97%)	8 (3%)	0	100	100
1	B	264/306 (86%)	251 (95%)	10 (4%)	3 (1%)	17	58
1	C	250/306 (82%)	239 (96%)	9 (4%)	2 (1%)	24	65
1	D	247/306 (81%)	240 (97%)	7 (3%)	0	100	100
2	E	91/104 (88%)	87 (96%)	4 (4%)	0	100	100
2	F	92/104 (88%)	89 (97%)	3 (3%)	0	100	100
All	All	1196/1432 (84%)	1150 (96%)	41 (3%)	5 (0%)	39	78

All (5) Ramachandran outliers are listed below:

Mol	Chain	Res	Type
1	B	16	SER
1	C	40	ILE
1	C	114	LYS
1	B	29	ASP
1	B	94	ALA

### 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	
1	A	202/246 (82%)	199 (98%)	3 (2%)	72	89
1	B	215/246 (87%)	214 (100%)	1 (0%)	92	96
1	C	201/246 (82%)	197 (98%)	4 (2%)	63	86
1	D	205/246 (83%)	204 (100%)	1 (0%)	92	96
2	E	78/88 (89%)	78 (100%)	0	100	100
2	F	79/88 (90%)	78 (99%)	1 (1%)	76	90
All	All	980/1160 (84%)	970 (99%)	10 (1%)	82	92

5 of 10 residues with a non-rotameric sidechain are listed below:

Mol	Chain	Res	Type
1	C	33	VAL
1	C	43	HIS
1	C	231	LYS
1	B	16	SER
1	C	113	LEU

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 [i](#)

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](#)

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 ⓘ

### 6.1 Protein, DNA and RNA chains ⓘ

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	256/306 (83%)	0.06	4 (1%) 74 75	49, 87, 120, 140	0
1	B	268/306 (87%)	-0.08	0 100 100	45, 78, 117, 132	0
1	C	254/306 (83%)	0.23	10 (3%) 43 42	49, 91, 127, 140	0
1	D	253/306 (82%)	0.23	10 (3%) 42 41	52, 97, 136, 162	0
2	E	93/104 (89%)	-0.28	0 100 100	39, 66, 93, 104	0
2	F	94/104 (90%)	-0.23	0 100 100	46, 67, 92, 101	0
3	G	29/33 (87%)	0.01	1 (3%) 49 49	65, 102, 149, 158	0
4	H	28/33 (84%)	-0.10	1 (3%) 46 46	69, 95, 148, 176	0
All	All	1275/1498 (85%)	0.05	26 (2%) 68 68	39, 85, 127, 176	0

The worst 5 of 26 RSRZ outliers are listed below:

Mol	Chain	Res	Type	RSRZ
1	D	230	PRO	5.5
4	H	28	DG	3.8
1	C	126	GLU	3.4
3	G	28	DG	3.4
1	D	240	PRO	2.9

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

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

### 6.3 Carbohydrates ⓘ

There are no carbohydrates in this entry.

## 6.4 Ligands

There are no ligands in this entry.

## 6.5 Other polymers

There are no such residues in this entry.