Full wwPDB X-ray Structure Validation Report (i)

Jun 20, 2022 – 01:13 pm BST

PDB ID	:	7OZ8
Title	:	Sulfated host glycan recognition by carbohydrate sulfatases of the human gut microbiota (BT1918_S1_46)
Authors	:	Cartmell, A.
Deposited on	:	2021-06-27
Resolution	:	1.95 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 (i) symbol.

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

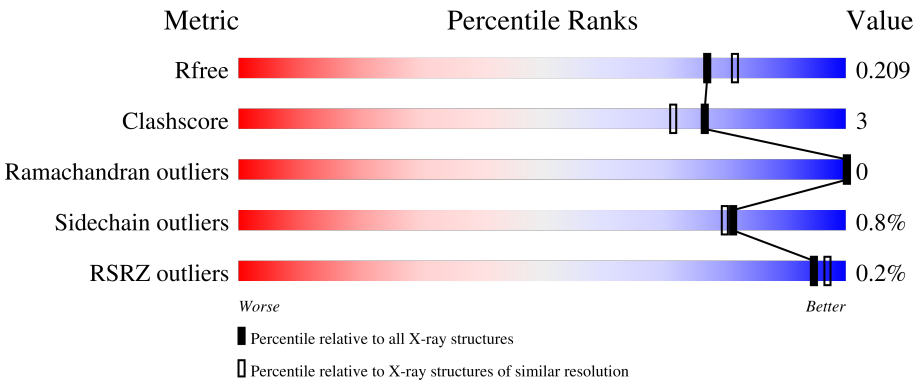
MolProbity	:	4.02b-467
Mogul	:	1.8.4, CSD as541be (2020)
Xtriage (Phenix)	:	1.13
EDS	:	2.28.1
buster-report	:	1.1.7 (2018)
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
Refmac	:	5.8.0267
CCP4	:	7.1.010 (Gargrove)
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.28.1

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 1.95 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	2580 (1.96-1.96)
Clashscore	141614	2705 (1.96-1.96)
Ramachandran outliers	138981	2678 (1.96-1.96)
Sidechain outliers	138945	2678 (1.96-1.96)
RSRZ outliers	127900	2539 (1.96-1.96)

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 <=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	AAA	502	90% · 6%

The following table lists non-polymeric compounds, carbohydrate monomers and non-standard residues in protein, DNA, RNA chains that are outliers for geometric or electron-density-fit criteria:

Mol	Type	Chain	Res	Chirality	Geometry	Clashes	Electron density
3	ACT	AAA	602	-	-	X	-

2 Entry composition (i)

There are 4 unique types of molecules in this entry. The entry contains 3991 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.

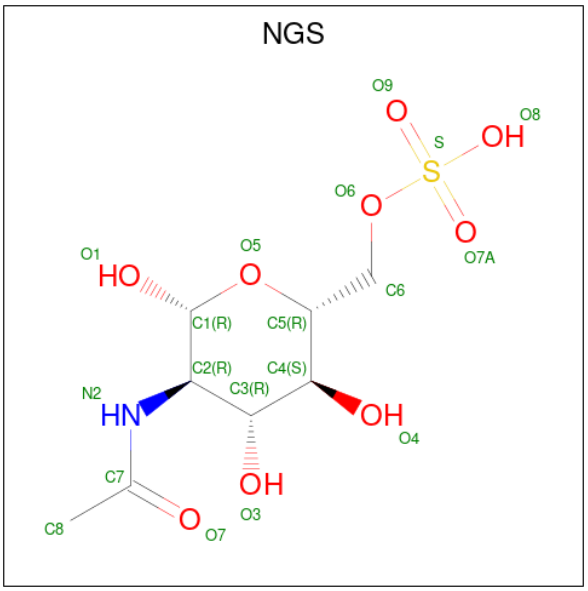
- Molecule 1 is a protein called Choline-sulfatase.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
			Total	C	N	O	S			
1	AAA	474	3801	2428	643	708	22	0	2	0

There are 23 discrepancies between the modelled and reference sequences:

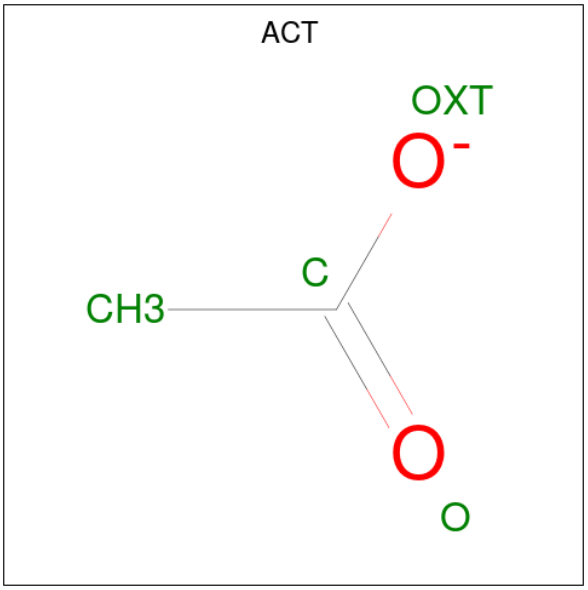
Chain	Residue	Modelled	Actual	Comment	Reference
AAA	1	MET	-	initiating methionine	UNP Q8A6G6
AAA	2	GLY	-	expression tag	UNP Q8A6G6
AAA	3	SER	-	expression tag	UNP Q8A6G6
AAA	4	SER	-	expression tag	UNP Q8A6G6
AAA	5	HIS	-	expression tag	UNP Q8A6G6
AAA	6	HIS	-	expression tag	UNP Q8A6G6
AAA	7	HIS	-	expression tag	UNP Q8A6G6
AAA	8	HIS	-	expression tag	UNP Q8A6G6
AAA	9	HIS	-	expression tag	UNP Q8A6G6
AAA	10	HIS	-	expression tag	UNP Q8A6G6
AAA	11	SER	-	expression tag	UNP Q8A6G6
AAA	12	SER	-	expression tag	UNP Q8A6G6
AAA	13	GLY	-	expression tag	UNP Q8A6G6
AAA	14	LEU	-	expression tag	UNP Q8A6G6
AAA	15	VAL	-	expression tag	UNP Q8A6G6
AAA	16	PRO	-	expression tag	UNP Q8A6G6
AAA	17	ARG	-	expression tag	UNP Q8A6G6
AAA	18	GLY	-	expression tag	UNP Q8A6G6
AAA	19	SER	-	expression tag	UNP Q8A6G6
AAA	20	HIS	-	expression tag	UNP Q8A6G6
AAA	21	MET	-	expression tag	UNP Q8A6G6
AAA	22	ALA	-	expression tag	UNP Q8A6G6
AAA	23	SER	-	expression tag	UNP Q8A6G6

- Molecule 2 is 2-acetamido-2-deoxy-6-O-sulfo-beta-D-glucopyranose (three-letter code: NGS) (formula: $C_8H_{15}NO_9S$) (labeled as "Ligand of Interest" by depositor).



Mol	Chain	Residues	Atoms					ZeroOcc	AltConf
			Total	C	N	O	S		
2	AAA	1	19	8	1	9	1	0	0

- Molecule 3 is ACETATE ION (three-letter code: ACT) (formula: $C_2H_3O_2$).



Mol	Chain	Residues	Atoms			ZeroOcc	AltConf
			Total	C	O		
3	AAA	1	4	2	2	0	0

- Molecule 4 is water.

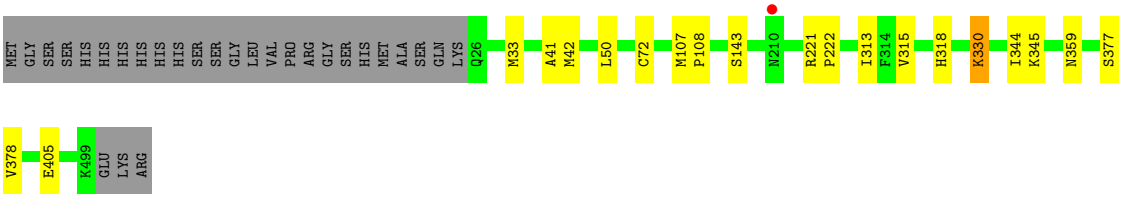
Mol	Chain	Residues	Atoms		ZeroOcc	AltConf
			Total	O		
4	AAA	167	167	167	0	0

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: Choline-sulfatase

Chain AAA:



4 Data and refinement statistics (i)

Property	Value	Source
Space group	P 32 2 1	Depositor
Cell constants a, b, c, α, β, γ	56.39Å 56.39Å 279.76Å 90.00° 90.00° 120.00°	Depositor
Resolution (Å)	48.15 – 1.95 48.11 – 1.95	Depositor EDS
% Data completeness (in resolution range)	99.9 (48.15-1.95) 99.9 (48.11-1.95)	Depositor EDS
R_{merge}	(Not available)	Depositor
R_{sym}	(Not available)	Depositor
$< I/\sigma(I) >$ [1]	1.66 (at 1.95Å)	Xtriage
Refinement program	REFMAC 5.8.0267	Depositor
R, R_{free}	0.188 , 0.204 0.197 , 0.209	Depositor DCC
R_{free} test set	2030 reflections (5.21%)	wwPDB-VP
Wilson B-factor (Å^2)	26.7	Xtriage
Anisotropy	0.633	Xtriage
Bulk solvent k_{sol}(e/Å^3), B_{sol}(Å^2)	(Not available) , (Not available)	EDS
L-test for twinning[2]	$< \|L\| > = 0.45$, $< L^2 > = 0.27$	Xtriage
Estimated twinning fraction	0.097 for -h,-k,l	Xtriage
F_o,F_c correlation	0.95	EDS
Total number of atoms	3991	wwPDB-VP
Average B, all atoms (Å^2)	31.0	wwPDB-VP

Xtriage's analysis on translational NCS is as follows: *The largest off-origin peak in the Patterson function is 4.83% of the height of the origin peak. No significant pseudotranslation is detected.*

[1]Intensities estimated from amplitudes.

[2]Theoretical values of $< |L| >$, $< L^2 >$ 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)

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

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	AAA	0.69	0/3918	0.81	0/5324

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	AAA	3801	0	3576	21	0
2	AAA	19	0	9	0	0
3	AAA	4	0	3	2	0
4	AAA	167	0	0	4	0
All	All	3991	0	3588	22	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 3.

All (22) 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:AAA:602:ACT:H1	4:AAA:847:HOH:O	1.75	0.85
1:AAA:143:SER:OG	3:AAA:602:ACT:H2	1.77	0.84
1:AAA:41:ALA:C	1:AAA:50[B]:LEU:CD2	2.49	0.81
1:AAA:41:ALA:O	1:AAA:50[B]:LEU:HD21	1.94	0.68
1:AAA:42:MET:CA	1:AAA:50[B]:LEU:HD23	2.29	0.62
1:AAA:318:HIS:CE1	1:AAA:330:LYS:HG2	2.35	0.62
1:AAA:378:VAL:CG2	4:AAA:732:HOH:O	2.47	0.62
1:AAA:107:MET:HB3	1:AAA:108:PRO:HD3	1.81	0.61
1:AAA:378:VAL:HG21	4:AAA:732:HOH:O	2.00	0.60
1:AAA:41:ALA:C	1:AAA:50[B]:LEU:HD22	2.23	0.57
1:AAA:42:MET:HA	1:AAA:50[A]:LEU:HD22	1.89	0.54
1:AAA:41:ALA:C	1:AAA:50[B]:LEU:HD21	2.27	0.52
1:AAA:42:MET:N	1:AAA:50[B]:LEU:CD2	2.75	0.49
1:AAA:42:MET:HA	1:AAA:50[B]:LEU:HD23	1.97	0.46
1:AAA:42:MET:N	1:AAA:50[B]:LEU:HD23	2.32	0.44
1:AAA:41:ALA:CA	1:AAA:50[B]:LEU:HD22	2.48	0.43
1:AAA:41:ALA:HA	1:AAA:50[B]:LEU:HD22	2.00	0.42
1:AAA:33:MET:HA	1:AAA:315:VAL:O	2.20	0.42
1:AAA:221:ARG:HB3	1:AAA:222:PRO:HA	2.03	0.41
1:AAA:313:ILE:HD12	1:AAA:344:ILE:HG12	2.01	0.41
1:AAA:72:CYS:HA	1:AAA:405:GLU:O	2.21	0.41
1:AAA:359:ASN:ND2	4:AAA:703:HOH:O	2.38	0.40

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	AAA	474/502 (94%)	457 (96%)	17 (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	AAA	394/429 (92%)	391 (99%)	3 (1%)	81 80

All (3) residues with a non-rotameric sidechain are listed below:

Mol	Chain	Res	Type
1	AAA	330	LYS
1	AAA	345	LYS
1	AAA	377	SER

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)

2 ligands are modelled in this entry.

In the following table, the Counts columns list the number of bonds (or angles) for which Mogul statistics could be retrieved, the number of bonds (or angles) that are observed in the model and the number of bonds (or angles) that are defined in the Chemical Component Dictionary. The Link column lists molecule types, if any, to which the group is linked. 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 $|Z| > 2$ is considered an outlier worth inspection. RMSZ is the root-mean-square of all Z scores of the bond lengths (or angles).

Mol	Type	Chain	Res	Link	Bond lengths			Bond angles		
					Counts	RMSZ	$\#\|Z\| > 2$	Counts	RMSZ	$\#\|Z\| > 2$
2	NGS	AAA	601	-	19,19,19	0.53	0	26,28,28	1.04	1 (3%)
3	ACT	AAA	602	-	1,3,3	2.08	1 (100%)	0,3,3	-	-

In the following table, the Chirals column lists the number of chiral outliers, the number of chiral centers analysed, the number of these observed in the model and the number defined in the Chemical Component Dictionary. Similar counts are reported in the Torsion and Rings columns. '-' means no outliers of that kind were identified.

Mol	Type	Chain	Res	Link	Chirals	Torsions	Rings
2	NGS	AAA	601	-	-	0/10/30/30	0/1/1/1

All (1) bond length outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(Å)	Ideal(Å)
3	AAA	602	ACT	CH3-C	-2.08	1.46	1.48

All (1) bond angle outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(^{o})	Ideal(^{o})
2	AAA	601	NGS	O1-C1-C2	-2.48	104.07	109.22

There are no chirality outliers.

There are no torsion outliers.

There are no ring outliers.

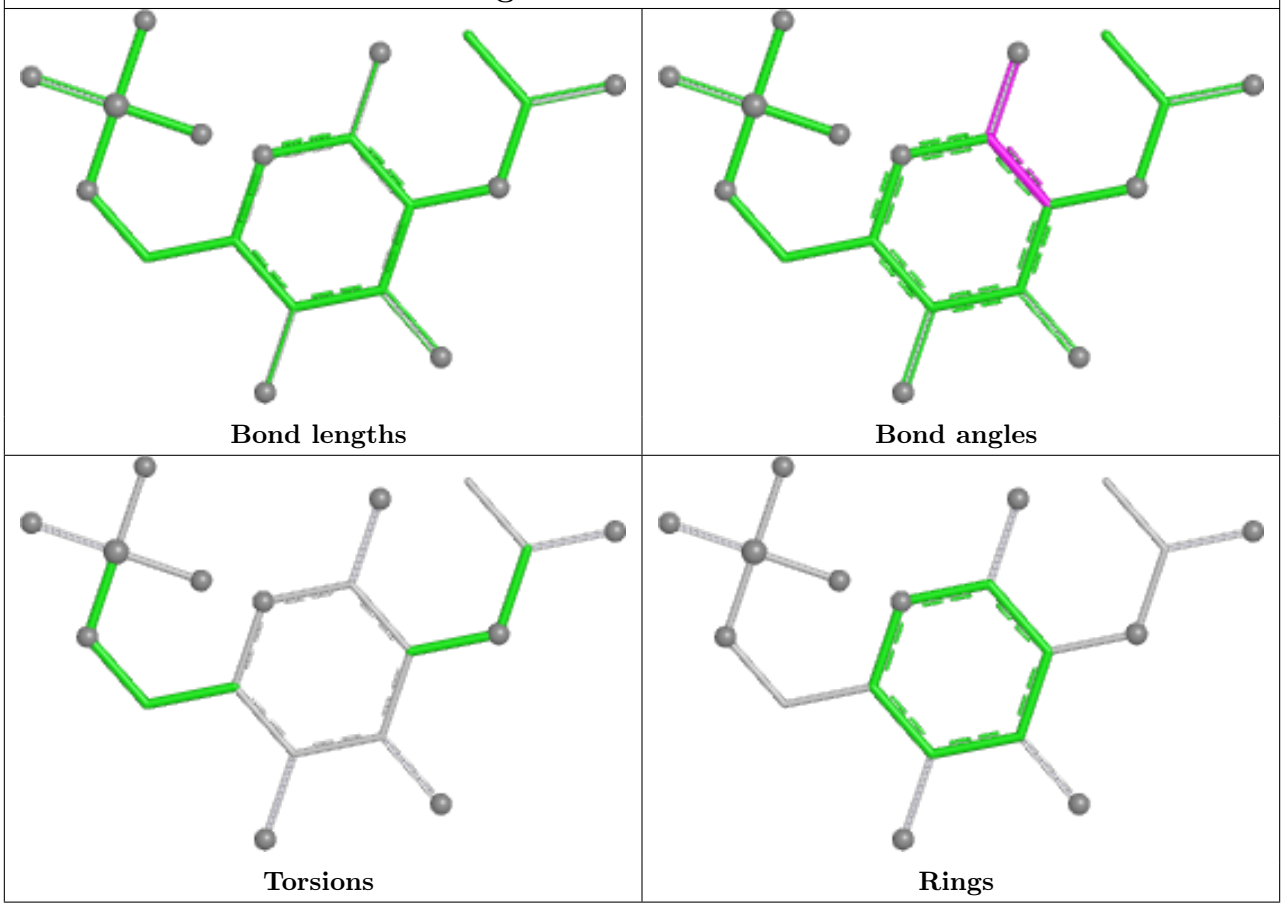
1 monomer is involved in 2 short contacts:

Mol	Chain	Res	Type	Clashes	Symm-Clashes
3	AAA	602	ACT	2	0

The following is a two-dimensional graphical depiction of Mogul quality analysis of bond lengths, bond angles, torsion angles, and ring geometry for all instances of the Ligand of Interest. In addition, ligands with molecular weight > 250 and outliers as shown on the validation Tables will also be included. For torsion angles, if less then 5% of the Mogul distribution of torsion angles is within 10 degrees of the torsion angle in question, then that torsion angle is considered an outlier. Any bond that is central to one or more torsion angles identified as an outlier by Mogul will be highlighted in the graph. For rings, the root-mean-square deviation (RMSD) between the ring in question and similar rings identified by Mogul is calculated over all ring torsion angles. If the

average RMSD is greater than 60 degrees and the minimal RMSD between the ring in question and any Mogul-identified rings is also greater than 60 degrees, then that ring is considered an outlier. The outliers are highlighted in purple. The color gray indicates Mogul did not find sufficient equivalents in the CSD to analyse the geometry.

Ligand NGS AAA 601



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	AAA	474/502 (94%)	-0.20	1 (0%) 95 97	20, 30, 48, 61	0

All (1) RSRZ outliers are listed below:

Mol	Chain	Res	Type	RSRZ
1	AAA	210	ASN	2.6

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)

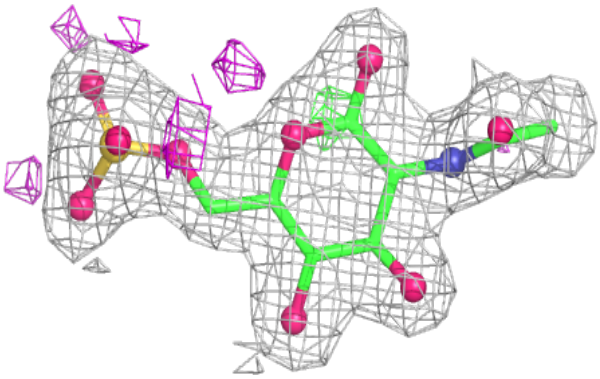
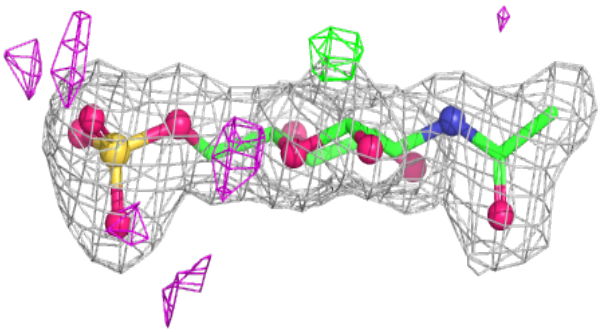
There are no monosaccharides in this entry.

6.4 Ligands (i)

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, 95th 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(Å^2)	Q<0.9
3	ACT	AAA	602	4/4	0.96	0.10	18,18,20,24	0
2	NGS	AAA	601	19/19	0.98	0.10	22,25,31,31	0

The following is a graphical depiction of the model fit to experimental electron density of all instances of the Ligand of Interest. In addition, ligands with molecular weight > 250 and outliers as shown on the geometry validation Tables will also be included. Each fit is shown from different orientation to approximate a three-dimensional view.

Electron density around NGS AAA 601: 2mF$_o$-DF$_c$ (at 0.7 rmsd) in gray mF$_o$-DF$_c$ (at 3 rmsd) in purple (negative) and green (positive)	
	

6.5 Other polymers (i)

There are no such residues in this entry.