wwPDB NMR Structure Validation Summary Report

Jun 4, 2023 – 05:19 PM EDT

PDB ID	:	2LE4
BMRB ID	:	17691
Title	:	Solution structure of the HMG box DNA-binding domain of human stem cell transcription factor Sox2
Authors	:	Sahu, S.C.; Markley, J.L.; Tonelli, M.; Bahrami, A.; Eghbalnia, H.R.; Center for Eukaryotic Structural Genomics (CESG)
Deposited on	:	2011-06-06

This is a wwPDB NMR Structure Validation Summary 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/NMRValidationReportHelp
with specific help available everywhere you see the (i) 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 (1)) were used in the production of this report:

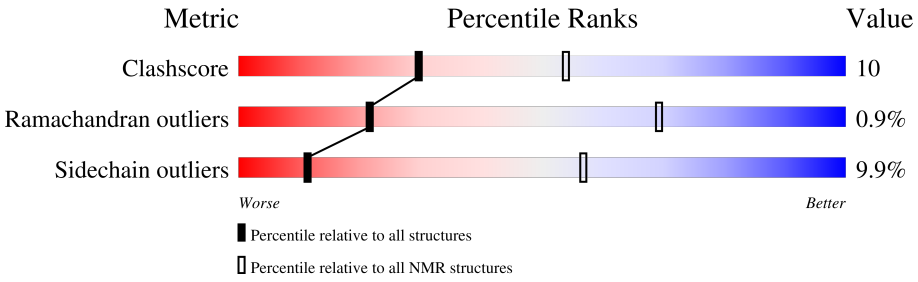
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
wwPDB-RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
wwPDB-ShiftChecker	:	v1.2
BMRB Restraints Analysis	:	v1.2
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.33

1 Overall quality at a glance (i)

The following experimental techniques were used to determine the structure:
SOLUTION NMR

The overall completeness of chemical shifts assignment is 82%.

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)	NMR archive (#Entries)
Clashscore	158937	12864
Ramachandran outliers	154571	11451
Sidechain outliers	154315	11428

The table below summarises the geometric issues observed across the polymeric chains and their fit to the experimental data. The red, orange, yellow and green segments indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A cyan segment indicates the fraction of residues that are not part of the well-defined cores, and 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%

Mol	Chain	Length	Quality of chain
1	A	81	52% 20% 28%

2 Ensemble composition and analysis

This entry contains 20 models. Model 16 is the overall representative, medoid model (most similar to other models). The authors have identified model 1 as representative, based on the following criterion: *fewest violations*.

The following residues are included in the computation of the global validation metrics.

Well-defined (core) protein residues			
Well-defined core	**Residue range (total)**	**Backbone RMSD (Å)**	**Medoid model**
1	A:7-A:64 (58)	0.98	16

Ill-defined regions of proteins are excluded from the global statistics.

Ligands and non-protein polymers are included in the analysis.

The models can be grouped into 3 clusters and 1 single-model cluster was found.

Cluster number	Models
1	1, 3, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20
2	5, 6, 7
3	2, 4
Single-model clusters	11

3 Entry composition

There is only 1 type of molecule in this entry. The entry contains 1415 atoms, of which 721 are hydrogens and 0 are deuteriums.

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

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	81	1415	431	721	141	117	5	0

There is a discrepancy between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	1	SER	-	expression tag	UNP P48431

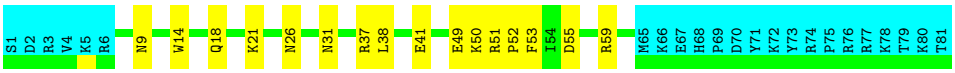
4 Residue-property plots (i)

4.1 Average score per residue in the NMR ensemble

These plots are provided for all protein, RNA, DNA and oligosaccharide chains in the entry. The first graphic is the same as shown in the summary in section 1 of this report. The second graphic shows the sequence where residues are colour-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. Stretches of 2 or more consecutive residues without any outliers are shown as green connectors. Residues which are classified as ill-defined in the NMR ensemble, are shown in cyan with an underline colour-coded according to the previous scheme. Residues which were present in the experimental sample, but not modelled in the final structure are shown in grey.

- Molecule 1: Transcription factor SOX-2

Chain A:



4.2 Residue scores for the representative (medoid) model from the NMR ensemble

The representative model is number 16. Colouring as in section 4.1 above.

- Molecule 1: Transcription factor SOX-2

Chain A:

5 Refinement protocol and experimental data overview

The models were refined using the following method: *molecular dynamics*.

Of the 100 calculated structures, 20 were deposited, based on the following criterion: *target function*.

The following table shows the software used for structure solution, optimisation and refinement.

Software name	Classification	Version
CNS	refinement	

The following table shows chemical shift validation statistics as aggregates over all chemical shift files. Detailed validation can be found in section 7 of this report.

Chemical shift file(s)	working_cs.cif
Number of chemical shift lists	1
Total number of shifts	982
Number of shifts mapped to atoms	982
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	82%

6 Model quality

6.1 Standard geometry

There are no covalent bond-length or bond-angle outliers.

There are no bond-length outliers.

There are no bond-angle outliers.

There are no chirality outliers.

There are no planarity outliers.

6.2 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 each 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 averaged over the ensemble.

Mol	Chain	Non-H	H(model)	H(added)	Clashes
1	A	485	498	496	10±3
All	All	9700	9960	9920	198

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 10.

5 of 92 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:A:51:ARG:HG3	1:A:52:PRO:HD3	0.97	1.36	1	1
1:A:8:MET:HG3	1:A:12:MET:HB2	0.71	1.61	6	2
1:A:47:GLU:HG3	1:A:50:LYS:HD2	0.69	1.63	10	1
1:A:47:GLU:O	1:A:50:LYS:HG3	0.68	1.87	20	1
1:A:10:ALA:HB1	1:A:53:PHE:HB3	0.64	1.69	16	7

6.3 Torsion angles

6.3.1 Protein backbone

In the following table, the Percentiles column shows the percent Ramachandran outliers of the chain as a percentile score with respect to all PDB entries followed by that with respect to all NMR entries. 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	58/81 (72%)	56±1 (96±2%)	2±1 (3±2%)	0±0 (1±1%)	21 69
All	All	1160/1620 (72%)	1110 (96%)	40 (3%)	10 (1%)	21 69

All 2 unique Ramachandran outliers are listed below. They are sorted by the frequency of occurrence in the ensemble.

Mol	Chain	Res	Type	Models (Total)
1	A	31	ASN	9
1	A	30	HIS	1

6.3.2 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 PDB entries followed by that with respect to all NMR entries. 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	51/74 (69%)	46±2 (90±4%)	5±2 (10±4%)	11 57
All	All	1020/1480 (69%)	919 (90%)	101 (10%)	11 57

5 of 26 unique residues with a non-rotameric sidechain are listed below. They are sorted by the frequency of occurrence in the ensemble.

Mol	Chain	Res	Type	Models (Total)
1	A	26	ASN	20
1	A	14	TRP	9
1	A	9	ASN	8
1	A	21	LYS	8
1	A	29	MET	8

6.3.3 RNA (i)

There are no RNA molecules in this entry.

6.4 Non-standard residues in protein, DNA, RNA chains (i)

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

6.5 Carbohydrates (i)

There are no monosaccharides in this entry.

6.6 Ligand geometry (i)

There are no ligands in this entry.

6.7 Other polymers (i)

There are no such molecules in this entry.

6.8 Polymer linkage issues (i)

There are no chain breaks in this entry.

7 Chemical shift validation

The completeness of assignment taking into account all chemical shift lists is 82% for the well-defined parts and 80% for the entire structure.

7.1 Chemical shift list 1

File name: working_cs.cif

Chemical shift list name: *sox2Asn*

7.1.1 Bookkeeping

The following table shows the results of parsing the chemical shift list and reports the number of nuclei with statistically unusual chemical shifts.

Total number of shifts	982
Number of shifts mapped to atoms	982
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	4

7.1.2 Chemical shift referencing

The following table shows the suggested chemical shift referencing corrections.

Nucleus	# values	Correction ± precision, *ppm*	Suggested action
$^{13}C_{\alpha}$	81	-0.22 ± 0.13	None needed (< 0.5 ppm)
$^{13}C_{\beta}$	79	0.18 ± 0.13	None needed (< 0.5 ppm)
$^{13}C'$	66	-0.47 ± 0.09	None needed (< 0.5 ppm)
^{15}N	75	-0.34 ± 0.33	None needed (< 0.5 ppm)

7.1.3 Completeness of resonance assignments

The following table shows the completeness of the chemical shift assignments for the well-defined regions of the structure. The overall completeness is 82%, i.e. 702 atoms were assigned a chemical shift out of a possible 861. 0 out of 6 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	221/286 (77%)	58/115 (50%)	108/116 (93%)	55/55 (100%)
Sidechain	435/515 (84%)	296/330 (90%)	134/153 (88%)	5/32 (16%)

Continued on next page...

Continued from previous page...

	Total	^{1}H	^{13}C	^{15}N
Aromatic	46/60 (77%)	24/30 (80%)	22/24 (92%)	0/6 (0%)
Overall	702/861 (82%)	378/475 (80%)	264/293 (90%)	60/93 (65%)

7.1.4 Statistically unusual chemical shifts (i)

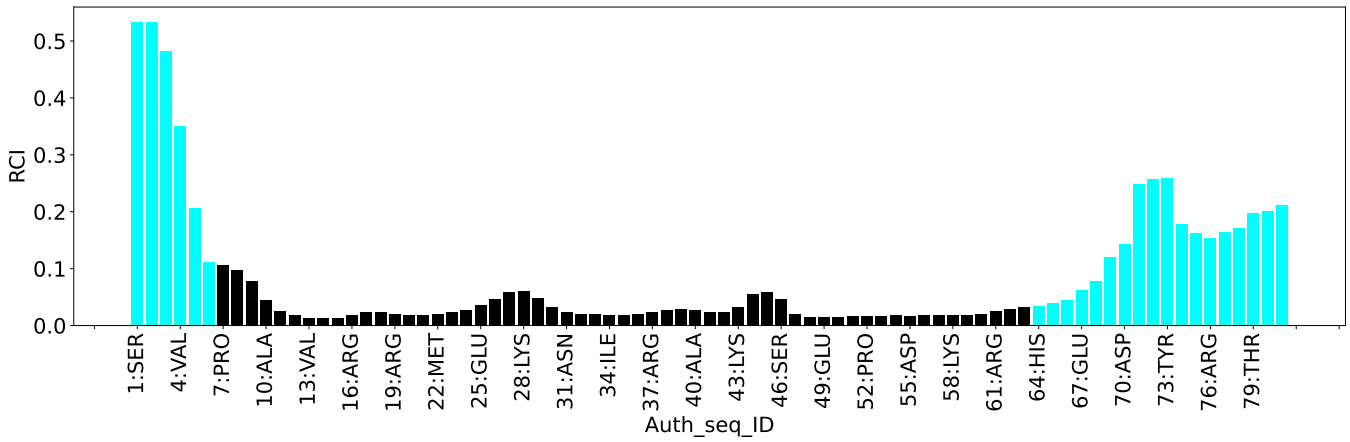
The following table lists the statistically unusual chemical shifts. These are statistical measures, and large deviations from the mean do not necessarily imply incorrect assignments. Molecules containing paramagnetic centres or hemes are expected to give rise to anomalous chemical shifts.

List Id	Chain	Res	Type	Atom	Shift, *ppm*	Expected range, *ppm*	Z-score
1	A	50	LYS	HE3	1.67	1.92 – 3.89	-6.3
1	A	50	LYS	HE2	1.72	1.95 – 3.88	-6.2
1	A	50	LYS	HG3	-0.06	0.04 – 2.67	-5.4
1	A	18	GLN	HG2	0.98	1.01 – 3.62	-5.1

7.1.5 Random Coil Index (RCI) plots (i)

The image below reports *random coil index* values for the protein chains in the structure. The height of each bar gives a probability of a given residue to be disordered, as predicted from the available chemical shifts and the amino acid sequence. A value above 0.2 is an indication of significant predicted disorder. The colour of the bar shows whether the residue is in the well-defined core (black) or in the ill-defined residue ranges (cyan), as described in section 2 on ensemble composition.If well-defined core and ill-defined regions are not identified then it is shown as gray bars.

Random coil index (RCI) for chain A:



8 NMR restraints analysis

8.1 Conformationally restricting restraints

The following table provides the summary of experimentally observed NMR restraints in different categories. Restraints are classified into different categories based on the sequence separation of the atoms involved.

Description	Value
Total distance restraints	1141
Intra-residue (|i-j|=0)	185
Sequential (|i-j|=1)	327
Medium range (|i-j|>1 and |i-j|<5)	393
Long range (|i-j|≥5)	176
Inter-chain	0
Hydrogen bond restraints	60
Disulfide bond restraints	0
Total dihedral-angle restraints	0
Number of unmapped restraints	0
Number of restraints per residue	14.1
Number of long range restraints per residue[1]	2.2

[1]Long range hydrogen bonds and disulfide bonds are counted as long range restraints while calculating the number of long range restraints per residue

8.2 Residual restraint violations

This section provides the overview of the restraint violations analysis. The violations are binned as small, medium and large violations based on its absolute value. Average number of violations per model is calculated by dividing the total number of violations in each bin by the size of the ensemble.

8.2.1 Average number of distance violations per model

Distance violations less than 0.1 Å are not included in the calculation.

Bins (Å)	Average number of violations per model	Max (Å)
0.1-0.2 (Small)	13.6	0.2
0.2-0.5 (Medium)	29.2	0.5
>0.5 (Large)	6.5	4.67

8.2.2 Average number of dihedral-angle violations per model (i)

Dihedral-angle violations less than 1° are not included in the calculation.There are no dihedral-angle violations

9 Distance violation analysis (i)

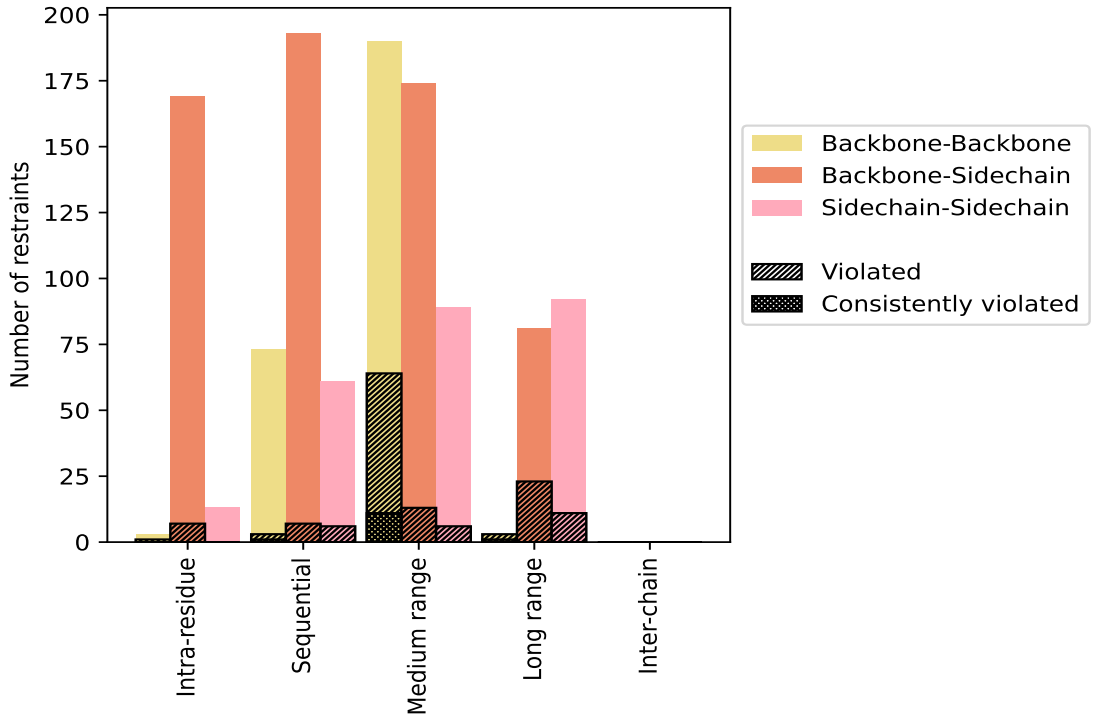
9.1 Summary of distance violations (i)

The following table shows the summary of distance violations in different restraint categories based on the sequence separation of the atoms involved. Each category is further sub-divided into three sub-categories based on the atoms involved. Violations less than 0.1 Å are not included in the statistics.

Restraints type	Count	%[1]	Violated[3]			Consistently Violated[4]		
			Count	%[2]	%[1]	Count	%[2]	%[1]
Intra-residue (|i-j|=0)	185	16.2	8	4.3	0.7	0	0.0	0.0
Backbone-Backbone	3	0.3	1	33.3	0.1	0	0.0	0.0
Backbone-Sidechain	169	14.8	7	4.1	0.6	0	0.0	0.0
Sidechain-Sidechain	13	1.1	0	0.0	0.0	0	0.0	0.0
Sequential (|i-j|=1)	327	28.7	16	4.9	1.4	1	0.3	0.1
Backbone-Backbone	73	6.4	3	4.1	0.3	1	1.4	0.1
Backbone-Sidechain	193	16.9	7	3.6	0.6	0	0.0	0.0
Sidechain-Sidechain	61	5.3	6	9.8	0.5	0	0.0	0.0
Medium range (|i-j|>1 & |i-j|<5)	393	34.4	27	6.9	2.4	1	0.3	0.1
Backbone-Backbone	130	11.4	8	6.2	0.7	1	0.8	0.1
Backbone-Sidechain	174	15.2	13	7.5	1.1	0	0.0	0.0
Sidechain-Sidechain	89	7.8	6	6.7	0.5	0	0.0	0.0
Long range (|i-j|≥5)	176	15.4	37	21.0	3.2	1	0.6	0.1
Backbone-Backbone	3	0.3	3	100.0	0.3	1	33.3	0.1
Backbone-Sidechain	81	7.1	23	28.4	2.0	0	0.0	0.0
Sidechain-Sidechain	92	8.1	11	12.0	1.0	0	0.0	0.0
Inter-chain	0	0.0	0	0.0	0.0	0	0.0	0.0
Backbone-Backbone	0	0.0	0	0.0	0.0	0	0.0	0.0
Backbone-Sidechain	0	0.0	0	0.0	0.0	0	0.0	0.0
Sidechain-Sidechain	0	0.0	0	0.0	0.0	0	0.0	0.0
Hydrogen bond	60	5.3	56	93.3	4.9	10	16.7	0.9
Disulfide bond	0	0.0	0	0.0	0.0	0	0.0	0.0
Total	1141	100.0	144	12.6	12.6	13	1.1	1.1
Backbone-Backbone	269	23.6	71	26.4	6.2	13	4.8	1.1
Backbone-Sidechain	617	54.1	50	8.1	4.4	0	0.0	0.0
Sidechain-Sidechain	255	22.3	23	9.0	2.0	0	0.0	0.0

[1] percentage calculated with respect to the total number of distance restraints, [2] percentage calculated with respect to the number of restraints in a particular restraint category, [3] violated in at least one model, [4] violated in all the models

9.1.1 Bar chart : Distribution of distance restraints and violations



Violated and consistently violated restraints are shown using different hatch patterns in their respective categories. The hydrogen bonds and disulfied bonds are counted in their appropriate category on the x-axis

9.2 Distance violation statistics for each model

The following table provides the distance violation statistics for each model in the ensemble. Violations less than 0.1 Å are not included in the statistics.

Model ID	Number of violations						Mean (Å)	Max (Å)	SD[6] (Å)	Median (Å)
	IR[1]	SQ[2]	MR[3]	LR[4]	IC[5]	Total				
1	1	5	35	15	0	56	0.62	3.86	0.8	0.34
2	1	4	38	11	0	54	0.35	1.48	0.26	0.3
3	3	2	30	9	0	44	0.37	1.04	0.23	0.32
4	1	4	43	12	0	60	0.32	1.06	0.2	0.27
5	3	1	31	12	0	47	0.34	1.05	0.2	0.31
6	1	2	30	10	0	43	0.33	1.8	0.29	0.27
7	1	4	37	16	0	58	0.38	1.69	0.3	0.3
8	3	3	33	7	0	46	0.34	2.25	0.32	0.3
9	1	4	29	10	0	44	0.37	3.09	0.45	0.26
10	3	3	27	11	0	44	0.43	2.42	0.39	0.33
11	0	2	38	13	0	53	0.54	4.67	0.83	0.29

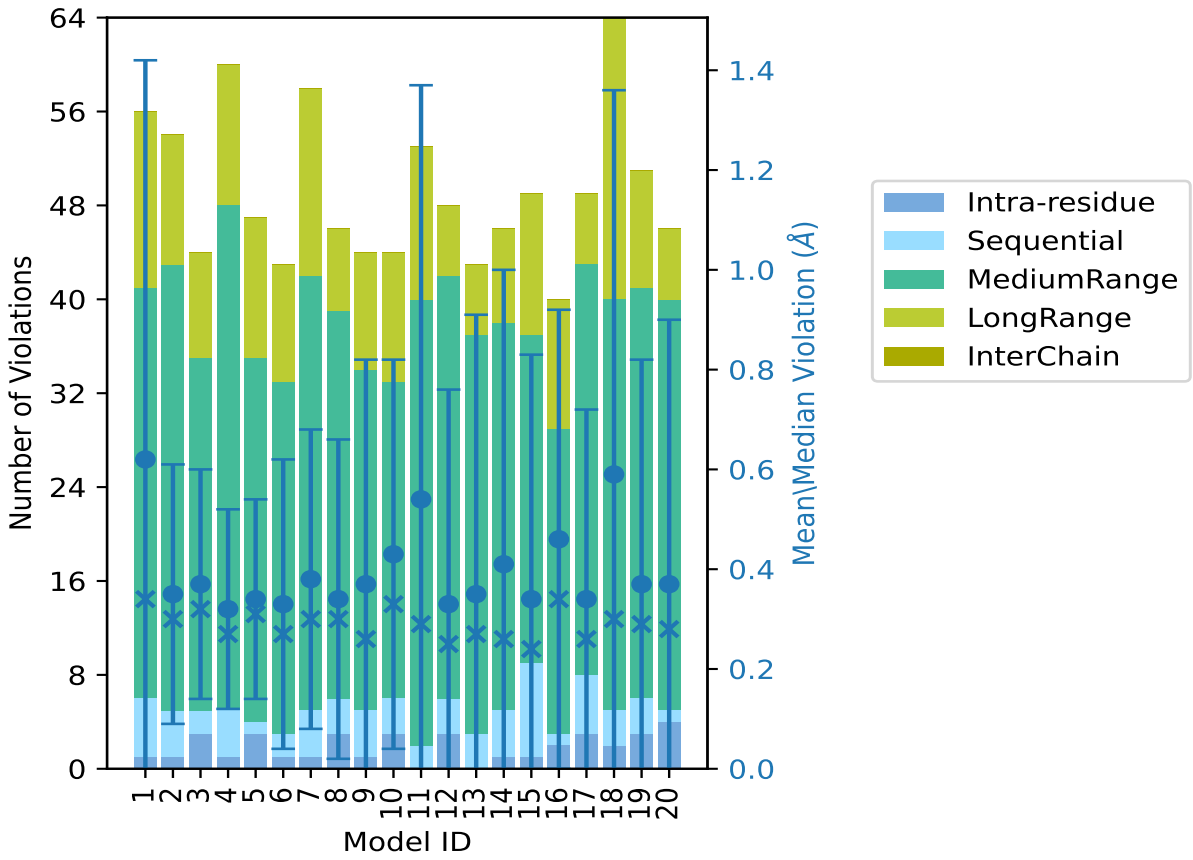
Continued on next page...

Continued from previous page...

Model ID	Number of violations						Mean (Å)	Max (Å)	SD[6] (Å)	Median (Å)
	IR[1]	SQ[2]	MR[3]	LR[4]	IC[5]	Total				
12	3	3	36	6	0	48	0.33	3.11	0.43	0.25
13	0	3	34	6	0	43	0.35	3.92	0.56	0.27
14	1	4	33	8	0	46	0.41	4.13	0.59	0.26
15	1	8	28	12	0	49	0.34	3.63	0.49	0.24
16	2	1	26	11	0	40	0.46	2.89	0.46	0.34
17	3	5	35	6	0	49	0.34	2.58	0.38	0.26
18	2	3	35	24	0	64	0.59	4.11	0.77	0.3
19	3	3	35	10	0	51	0.37	3.2	0.45	0.29
20	4	1	35	6	0	46	0.37	3.75	0.53	0.28

[1]Intra-residue restraints, [2]Sequential restraints, [3]Medium range restraints, [4]Long range restraints, [5]Inter-chain restraints, [6]Standard deviation

9.2.1 Bar graph : Distance Violation statistics for each model



The mean(dot),median(x) and the standard deviation are shown in blue with respect to the y axis on the right

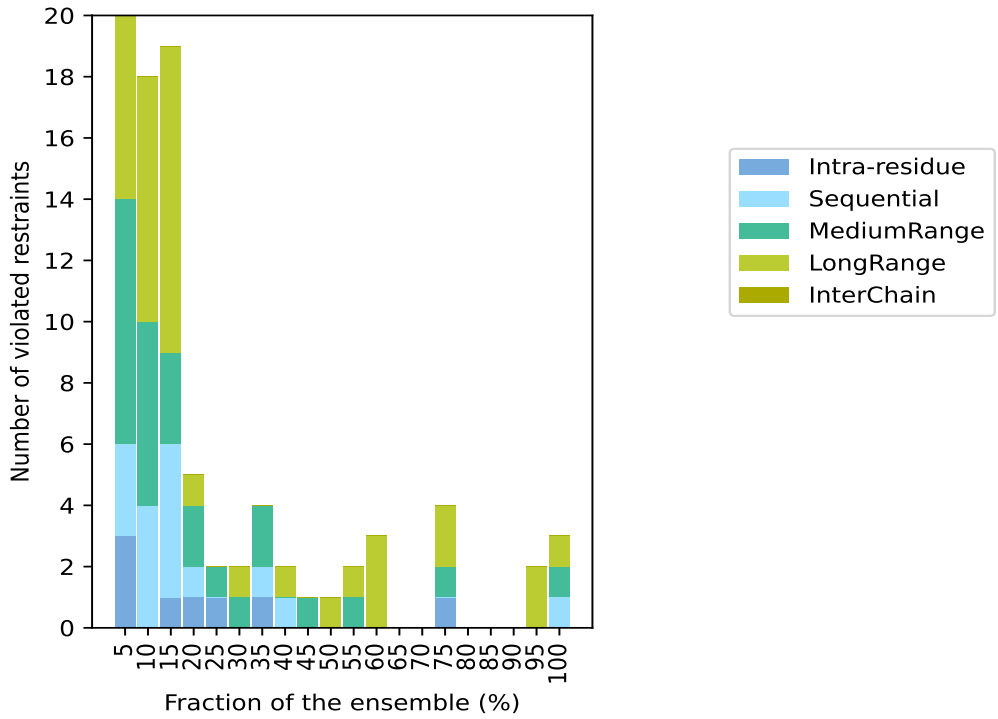
9.3 Distance violation statistics for the ensemble (i)

Violation analysis may find that some restraints are violated in few models and some are violated in most of models. The following table provides this information as number of violated restraints for a given fraction of the ensemble. In total, 993(IR:177, SQ:311, MR:366, LR:139, IC:0) restraints are not violated in the ensemble.

Number of violated restraints						Fraction of the ensemble	
IR[1]	SQ[2]	MR[3]	LR[4]	IC[5]	Total	Count[6]	%
3	3	8	6	0	20	1	5.0
0	4	6	8	0	18	2	10.0
1	5	3	10	0	19	3	15.0
1	1	2	1	0	5	4	20.0
1	0	1	0	0	2	5	25.0
0	0	1	1	0	2	6	30.0
1	1	2	0	0	4	7	35.0
0	1	0	1	0	2	8	40.0
0	0	1	0	0	1	9	45.0
0	0	0	1	0	1	10	50.0
0	0	1	1	0	2	11	55.0
0	0	0	3	0	3	12	60.0
0	0	0	0	0	0	13	65.0
0	0	0	0	0	0	14	70.0
1	0	1	2	0	4	15	75.0
0	0	0	0	0	0	16	80.0
0	0	0	0	0	0	17	85.0
0	0	0	0	0	0	18	90.0
0	0	0	2	0	2	19	95.0
0	1	1	1	0	3	20	100.0

[1]Intra-residue restraints, [2]Sequential restraints, [3]Medium range restraints, [4]Long range restraints, [5]Inter-chain restraints, [6] Number of models with violations

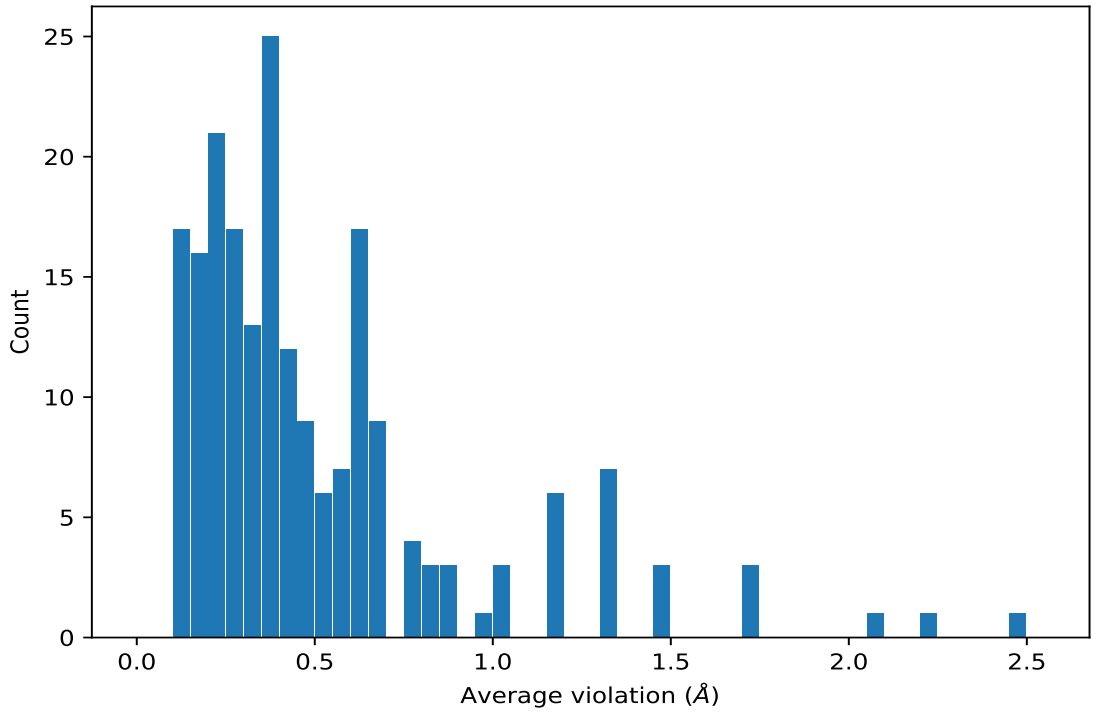
9.3.1 Bar graph : Distance violation statistics for the ensemble



9.4 Most violated distance restraints in the ensemble

9.4.1 Histogram : Distribution of mean distance violations

The following histogram shows the distribution of the average value of the violation. The average is calculated for each restraint that is violated in more than one model over all the violated models in the ensemble



9.4.2 Table: Most violated distance restraints

The following table provides the mean and the standard deviation of the violations for the 10 worst performing restraints, sorted by number of violated models and the mean violation value. The Key (restraint list ID, restraint ID) is the unique identifier for a given restraint. Rows with same key represent combinatorial or ambiguous restraints and are counted as a single restraint.

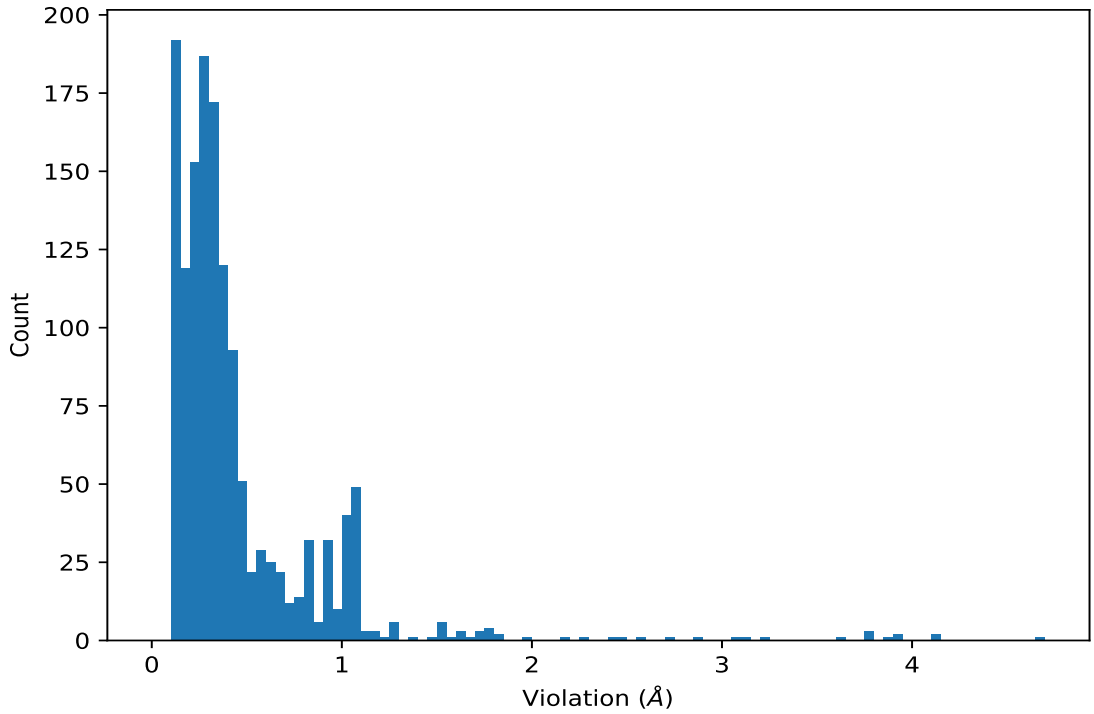
Key	Atom-1	Atom-2	Models[1]	Mean (Å)	SD[1] (Å)	Median (Å)
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	20	2.47	1.5	2.99
(3,11)	1:A:15:SER:O	1:A:19:ARG:H	20	0.43	0.04	0.43
(3,5)	1:A:12:MET:O	1:A:15:SER:H	20	0.42	0.02	0.42
(3,25)	1:A:33:GLU:O	1:A:36:LYS:H	20	0.39	0.03	0.39
(3,23)	1:A:32:SER:O	1:A:35:SER:H	20	0.38	0.03	0.38
(3,41)	1:A:53:PHE:O	1:A:57:ALA:H	20	0.36	0.06	0.38
(3,3)	1:A:11:PHE:O	1:A:15:SER:H	20	0.35	0.05	0.36
(3,39)	1:A:39:GLY:O	1:A:43:LYS:H	20	0.35	0.05	0.35
(3,31)	1:A:35:SER:O	1:A:39:GLY:H	20	0.34	0.05	0.35
(3,12)	1:A:15:SER:O	1:A:19:ARG:N	20	0.33	0.05	0.33

[1]Number of violated models, [2]Standard deviation

9.5 All violated distance restraints (i)

9.5.1 Histogram : Distribution of distance violations (i)

The following histogram shows the distribution of the absolute value of the violation for all violated restraints in the ensemble.



9.5.2 Table : All distance violations (i)

The following table provides the 10 worst performing restraints, sorted by the violation value. The Key (restraint list ID, restraint ID) is the unique identifier for a given restraint. Rows with same key represent combinatorial or ambiguous restraints and are counted as a single restraint.

Key	Atom-1	Atom-2	Model ID	Violation (Å)
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	11	4.67
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	14	4.13
(1,369)	1:A:13:VAL:HB	1:A:53:PHE:HD2	18	4.11
(1,369)	1:A:13:VAL:HB	1:A:53:PHE:HD2	11	3.93
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	13	3.92
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	1	3.86
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	18	3.8
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	20	3.75
(1,369)	1:A:13:VAL:HB	1:A:53:PHE:HD2	1	3.75
(1,82)	1:A:21:LYS:H	1:A:56:GLU:HA	15	3.63

10 Dihedral-angle violation analysis (i)

Dihedral angle analysis failed due to data error in the dihedral angle restraints, possibly missing target value