Annotation and validation of nucleic acid structures.

Increasing quality and interoperability of the data

Bohdan Schneider
bs.structbio.org
Institute of Biotechnology of the Czech Academy of Sciences

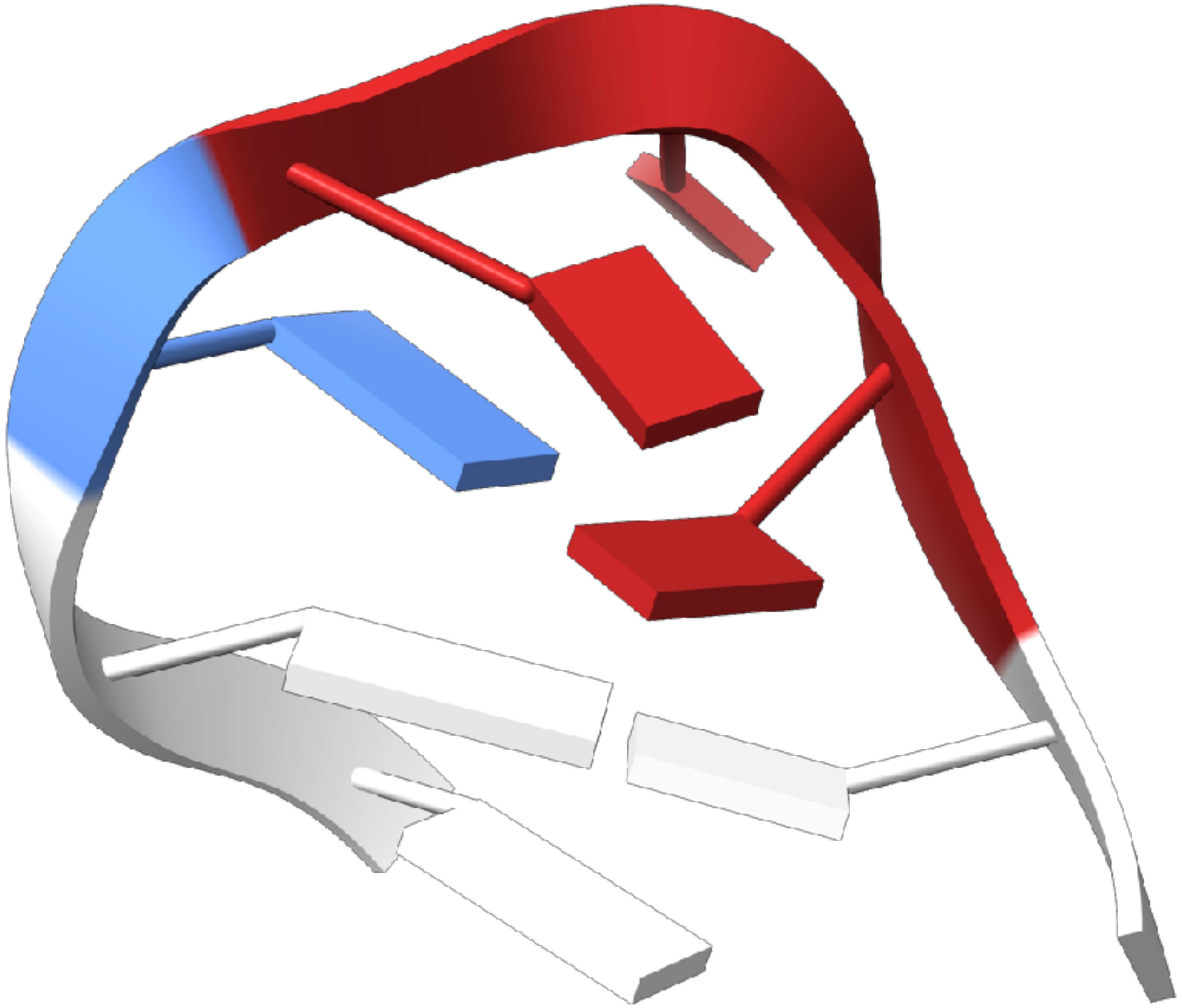
Computational Approaches to RNA Structure and Function

2024-08-02

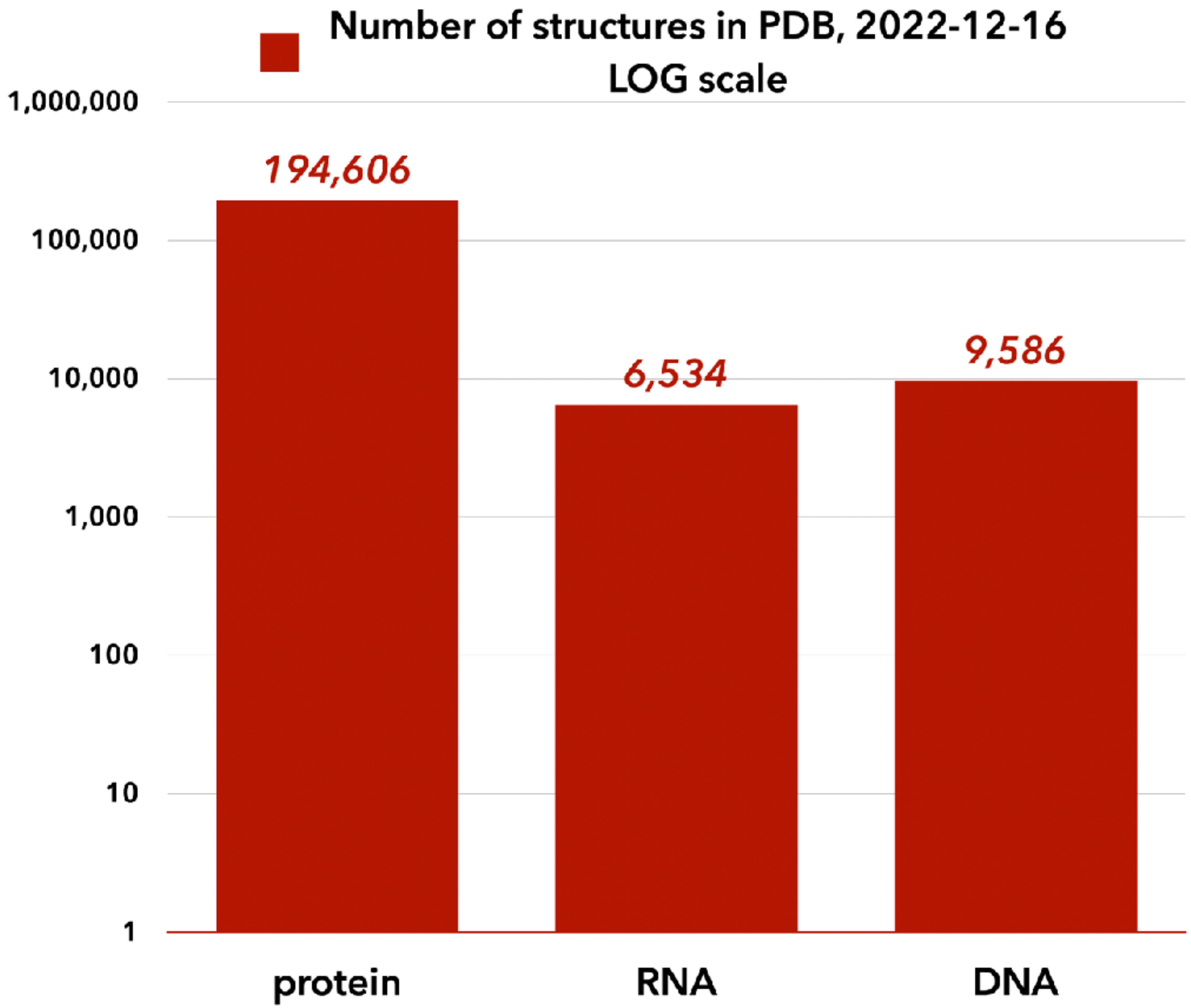
Why discuss the quality of structural data

- Public archive should be reliable source of ***validated*** information
    - PDB data are often used by non-experts
    - mmCIF does not capture all features of biomolecules

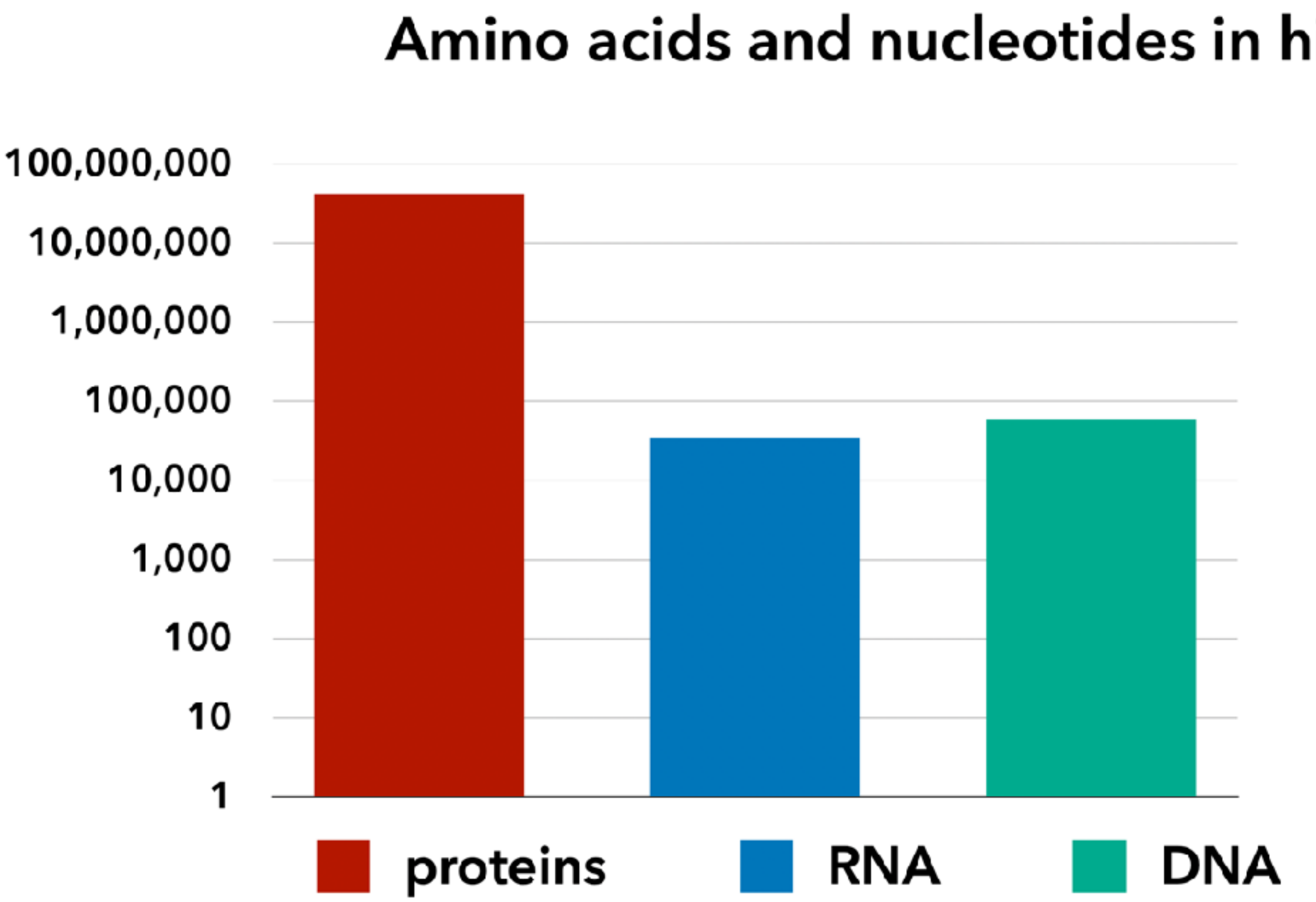
- Nucleic acids are of special concern …



PDB archive: 7.6% of NA structures

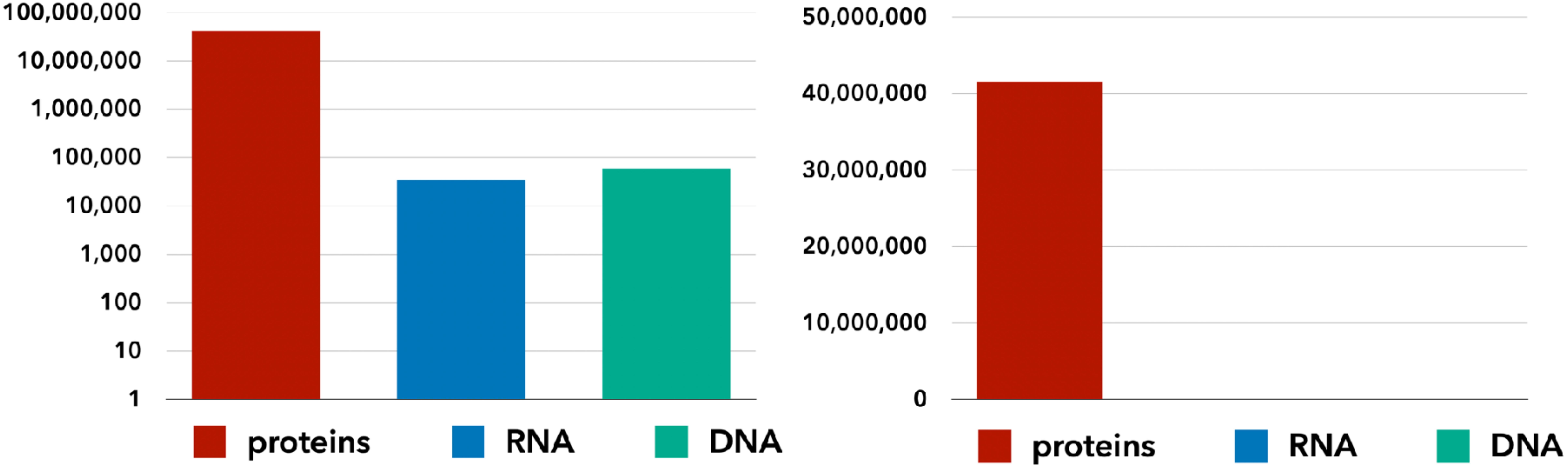


Small number of high quality NA models: need for sophisticated tools



Small number of high quality NA models: need for sophisticated tools

Amino acids and nucleotides in high-resolution structures (≤2.0 Å)

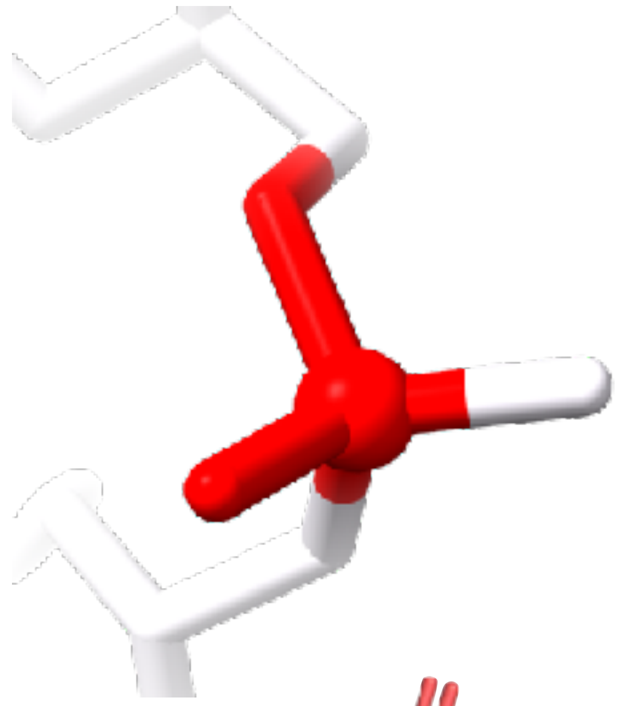


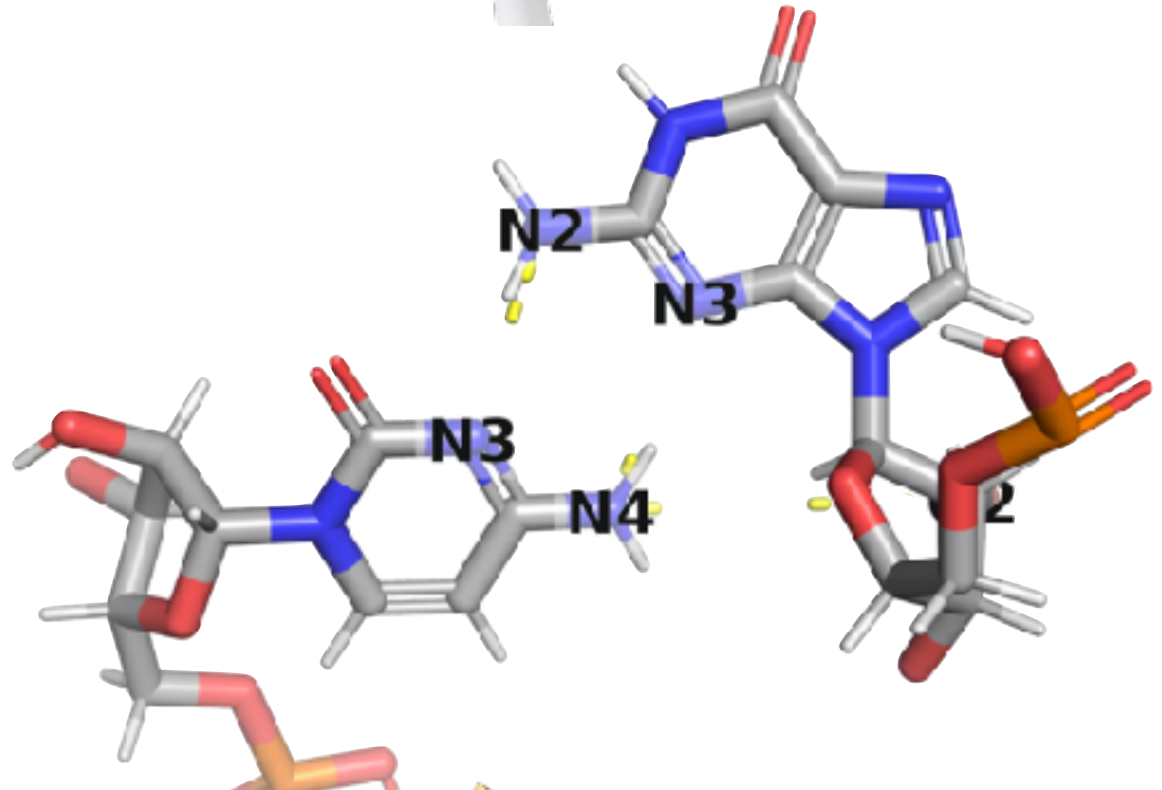
Three main sources of errors in NA structures

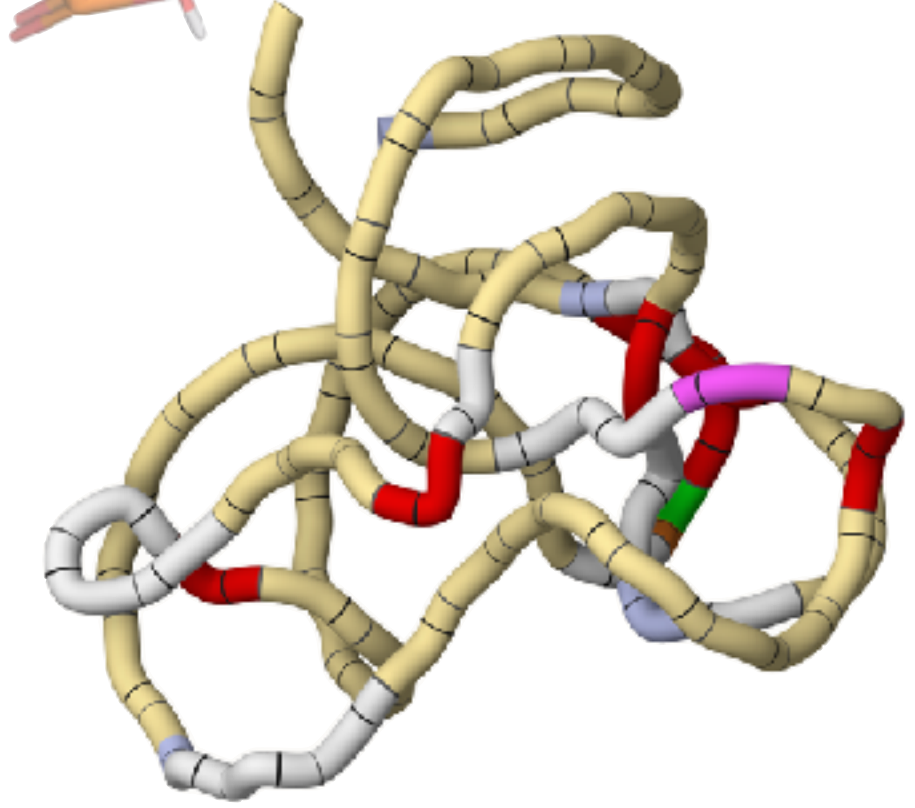
1. Valence geometry

2. Base pairing

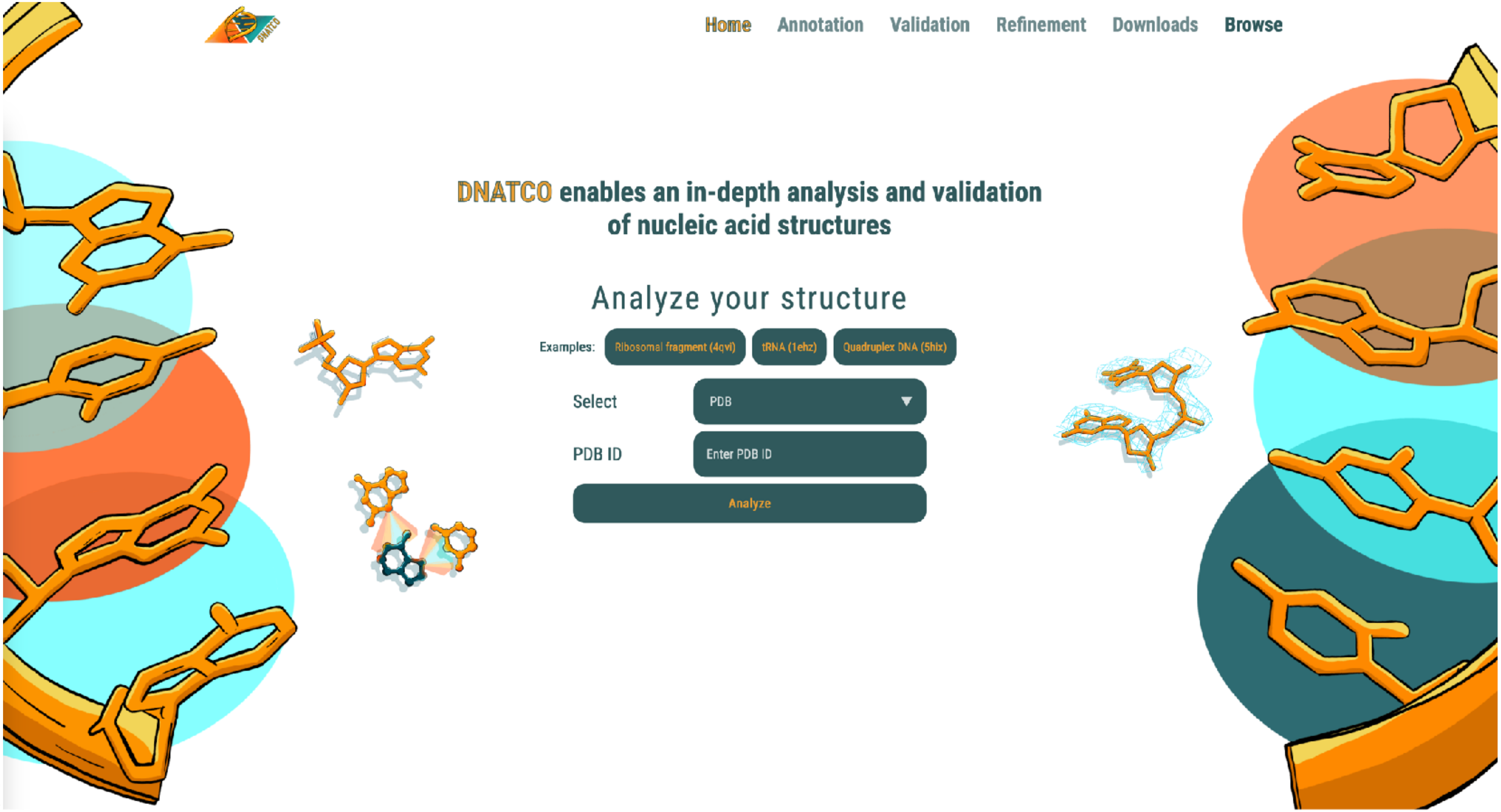
3. Backbone geometries







Remediation at *dnatco.datmos.org*

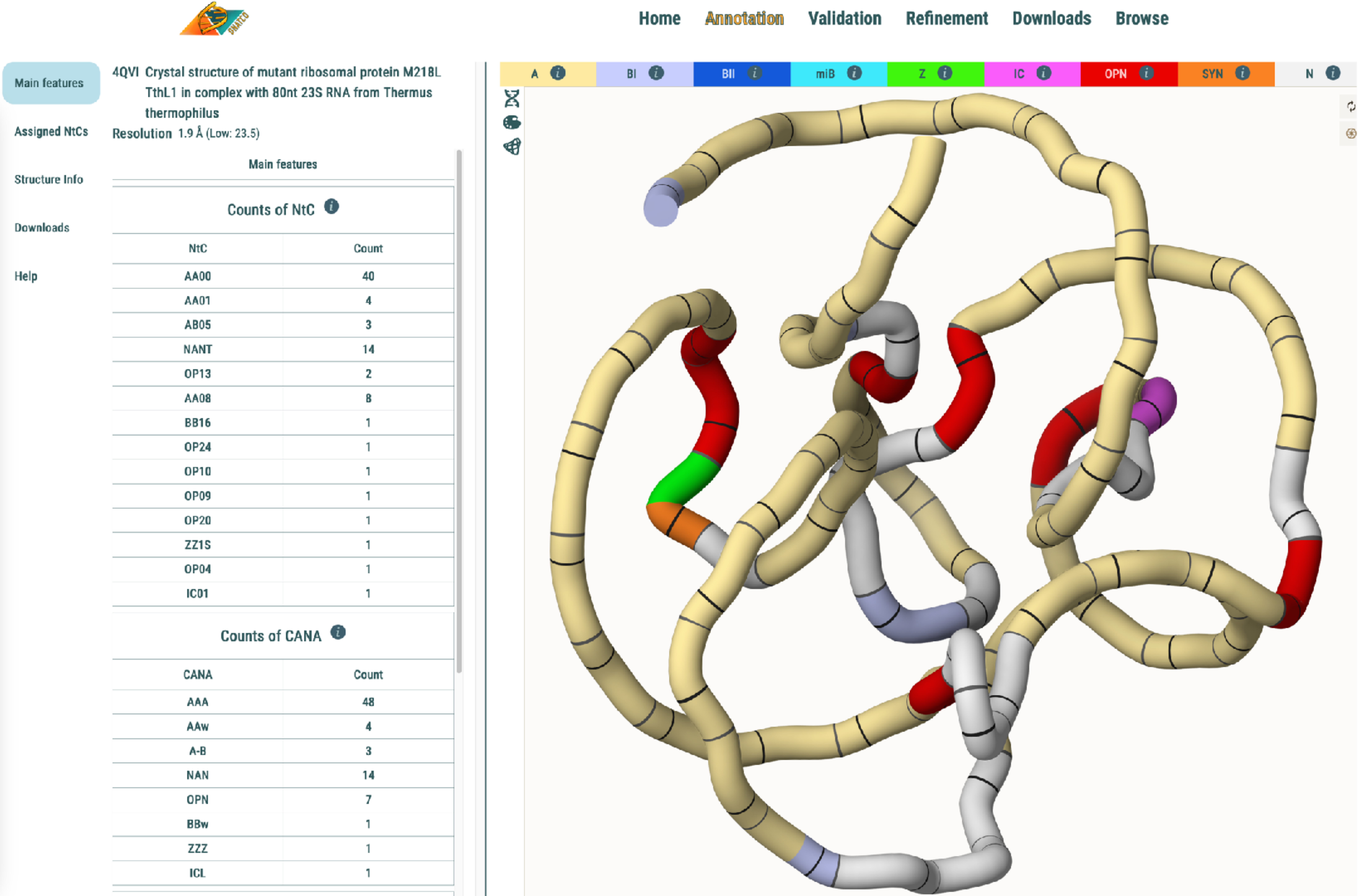


DNATCO enables an in-depth analysis and validation of nucleic acid structures

Analyze your structure

Select: PDB

PDB ID: Enter PDB ID



4QVI Crystal structure of mutant ribosomal protein M218L TthL1 in complex with 80nt 23S RNA from Thermus thermophilus

Resolution 1.9 Å (Low: 23.5)

Main features

Counts of NtC

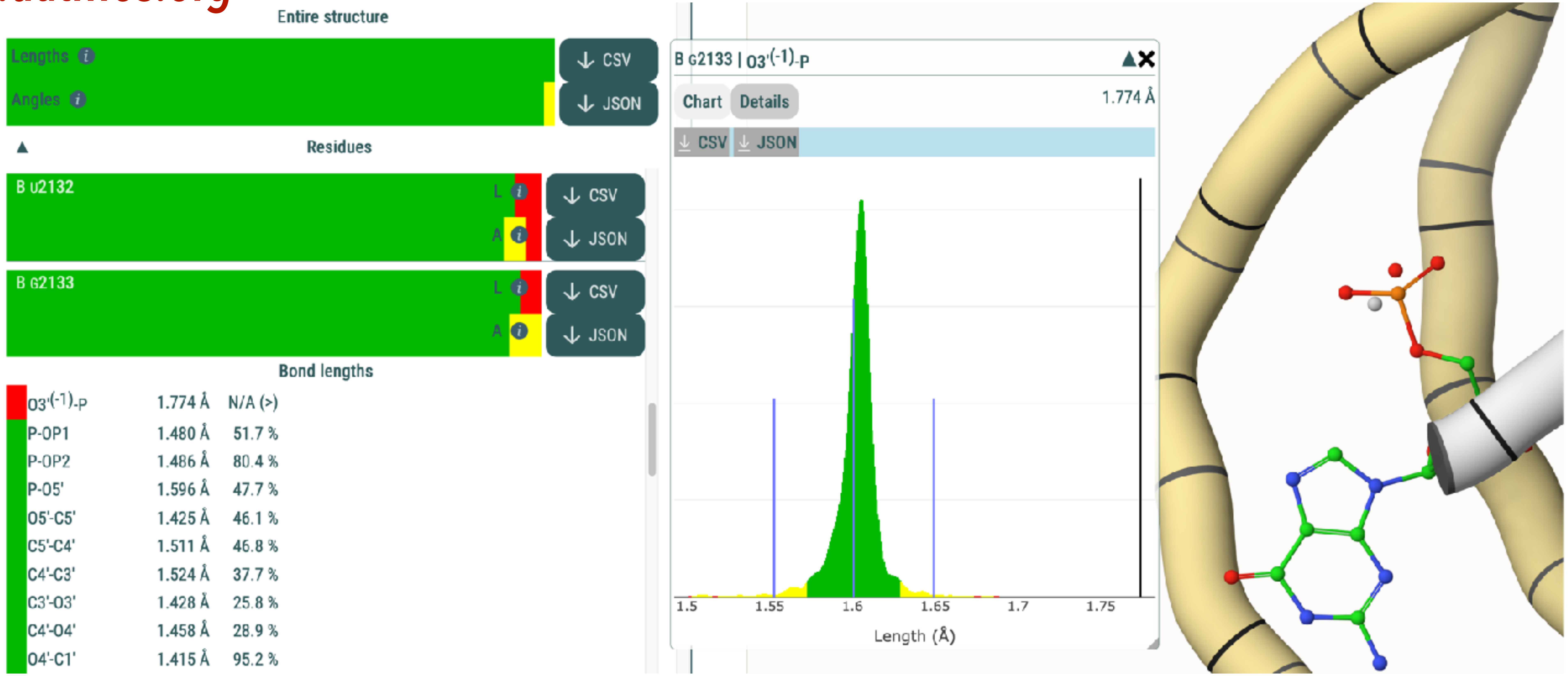
NtC	Count
AA00	40
AA01	4
AB05	3
NANT	14
OP13	2
AA08	8
BB16	1
OP24	1
OP10	1
OP09	1
OP20	1
ZZ1S	1
OP04	1
IC01	1

Counts of CANA

CANA	Count
AAA	48
AAw	4
A-B	3
NAN	14
OPN	7
BBw	1
ZZZ	1
ICL	1

1. Valence geometry

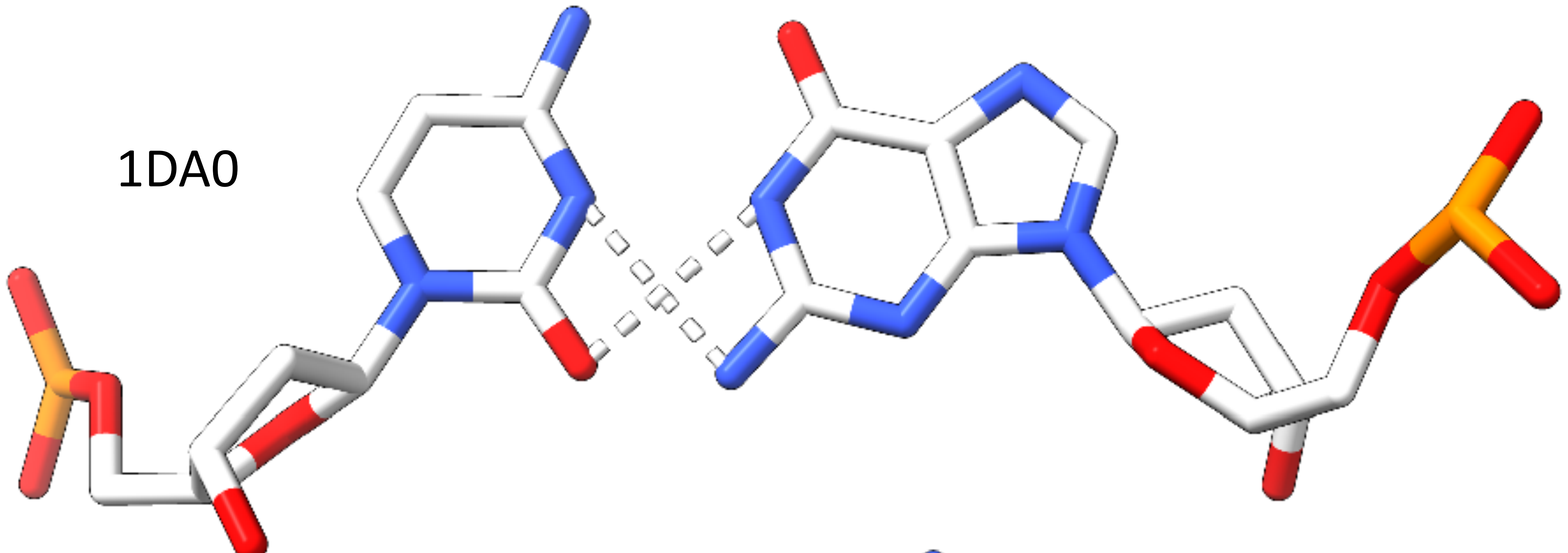
- Refinement programs use different restraints
- Validation by PDB and other tools is inconsistent and statistically naive
- NAVAL - an expert group to suggest clarification
    - results at ***dnatco.datmos.org***



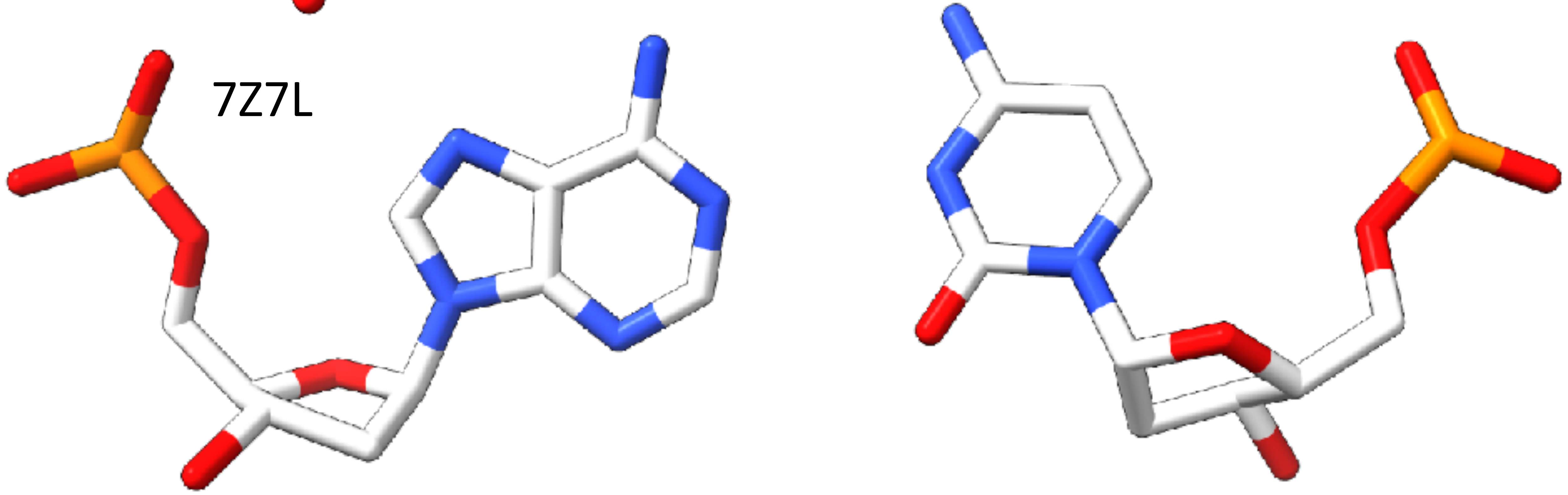
🟥	$O3'^{(-1)}$-P	1.774 Å	N/A (>)
🟩	P-OP1	1.480 Å	51.7 %
🟩	P-OP2	1.486 Å	80.4 %
🟩	P-O5'	1.596 Å	47.7 %
🟩	O5'-C5'	1.425 Å	46.1 %
🟩	C5'-C4'	1.511 Å	46.8 %
🟩	C4'-C3'	1.524 Å	37.7 %
🟩	C3'-O3'	1.428 Å	25.8 %
🟩	C4'-O4'	1.458 Å	28.9 %
🟩	O4'-C1'	1.415 Å	95.2 %

2. Base pairing classification is essential for prediction of RNA 2D and 3D

- The PDB archive is not reliable
- Available programs provide incomplete & conflicting results

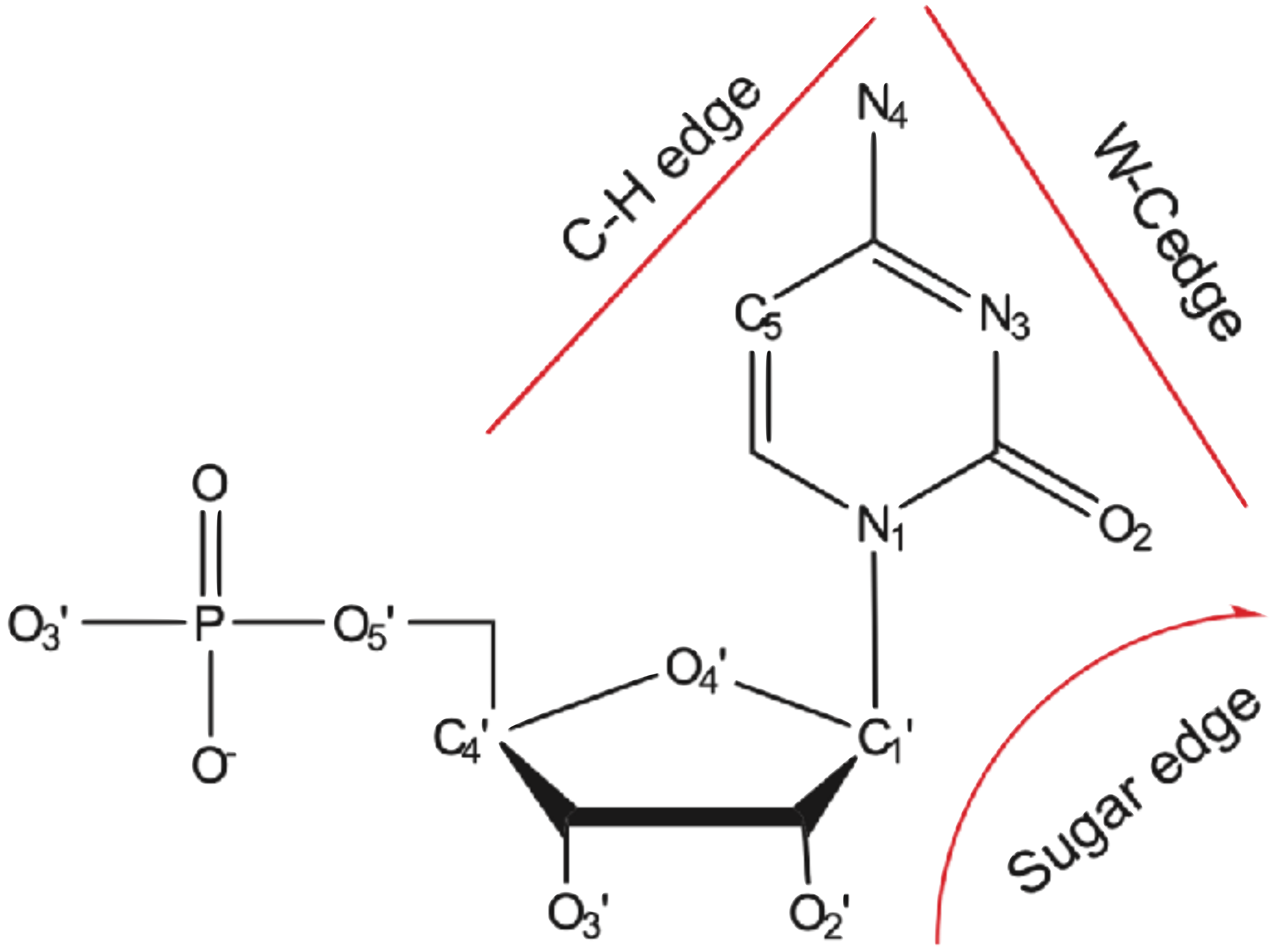


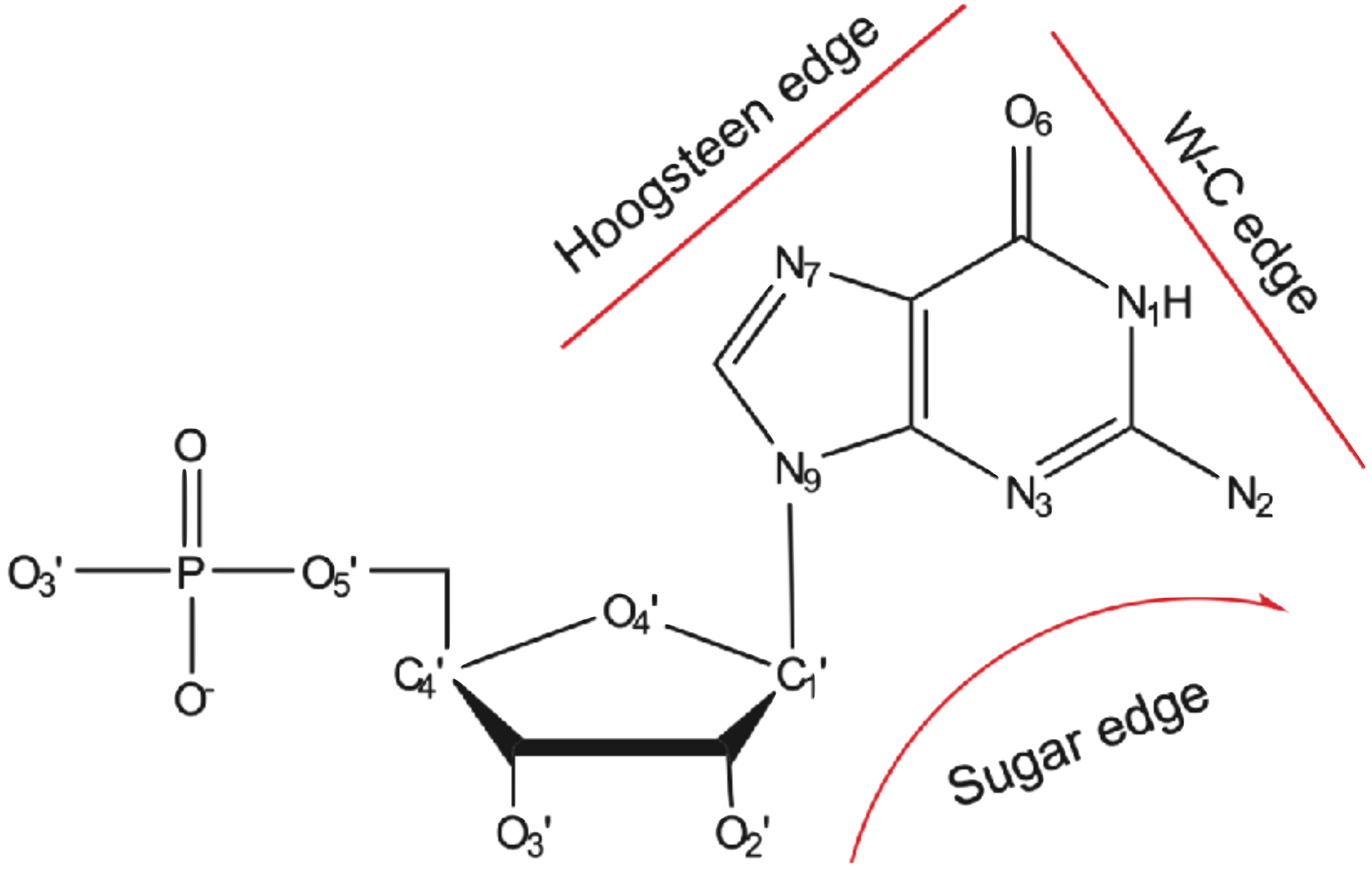
1DA0



7Z7L

2. New Base Pair assignment based on the Leontis-Westhof classes





2. New Base Pair assignment at *basepairs.datmos.org*

Home | cWW | cWWa | tWW | tWWa | cWH | tWH | cWS | **tWS** | cHH | tHH | cHS | tHS | cSS | tSS

A-A (39) | **A-C** (6) | **A-G** (632) | **A-U** (0) | **C-A** (27) | **C-C** (18) | **C-G** (35) | **C-U** (2) | **G-A** | **G-C** (11) | **G-G** | **G-U** (186) | **U-A** (79) | **U-C** (8)

U-G (30) | **U-U** (12)

6: trans Watson-Crick / Sugar

	A	G	C	U
A	**3ccq** 0-A**2018** · · · 0-A**1829** show statistics + exemplars	not defined	**3ccu** 0-C**1983** · · · 0-A**1981** show statistics + exemplars	**1s72** 0-U**779** · · · 0-A**776** show statistics + exemplars

● Basic ○ Parameter ranges ○ SQL

Data source

All Polar Contacts — Reference Set

○ RNA ○ DNA ● Both

Order by

Smallest Edge RMSD

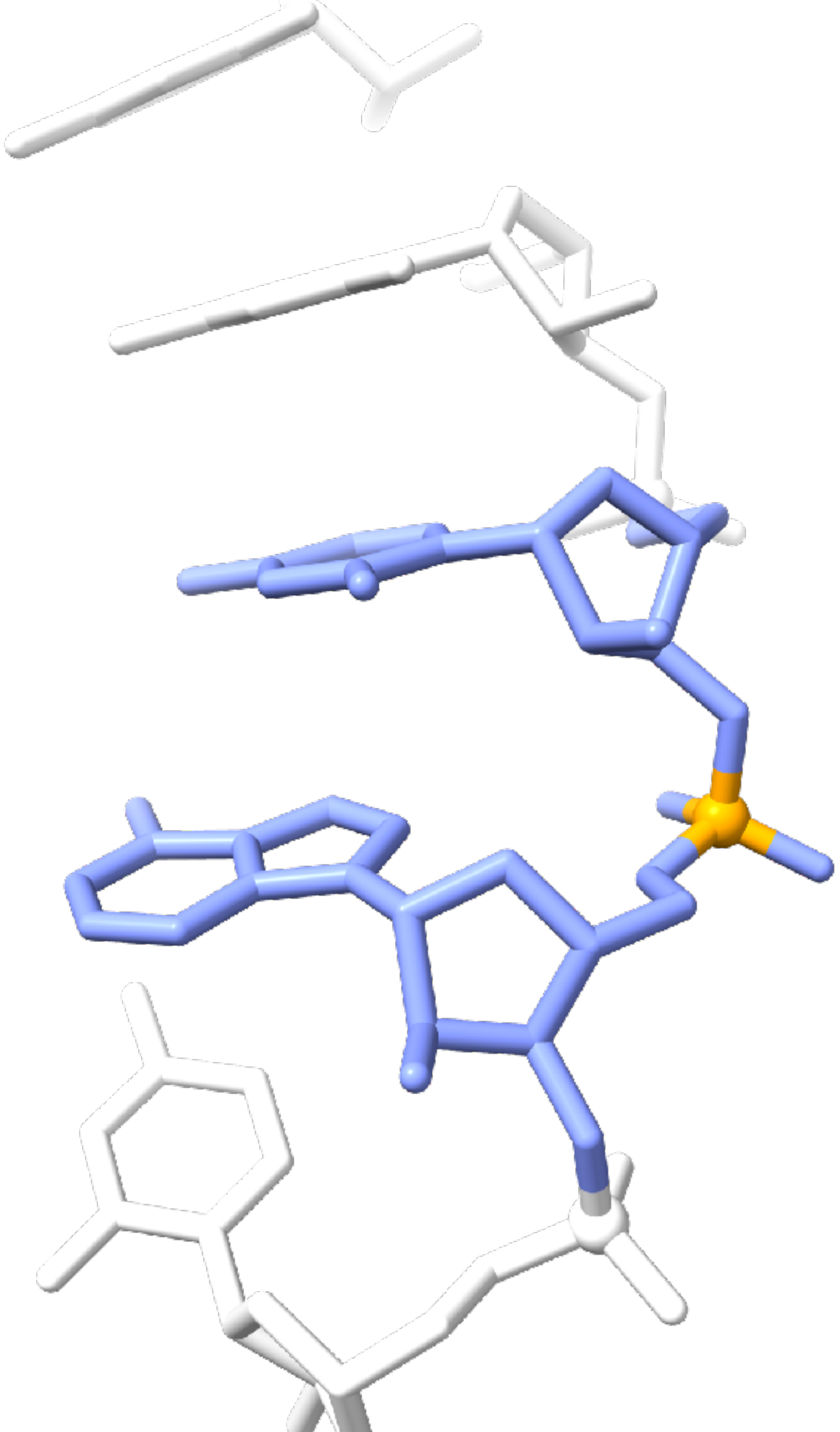
☐ Rotate images

Reference tWS C-G basepair

58462 × tWS-C-G from **3954** PDB structures

▽ expand plots ▽

3. The NA backbone geometry: often ignored … to the detriment of the field

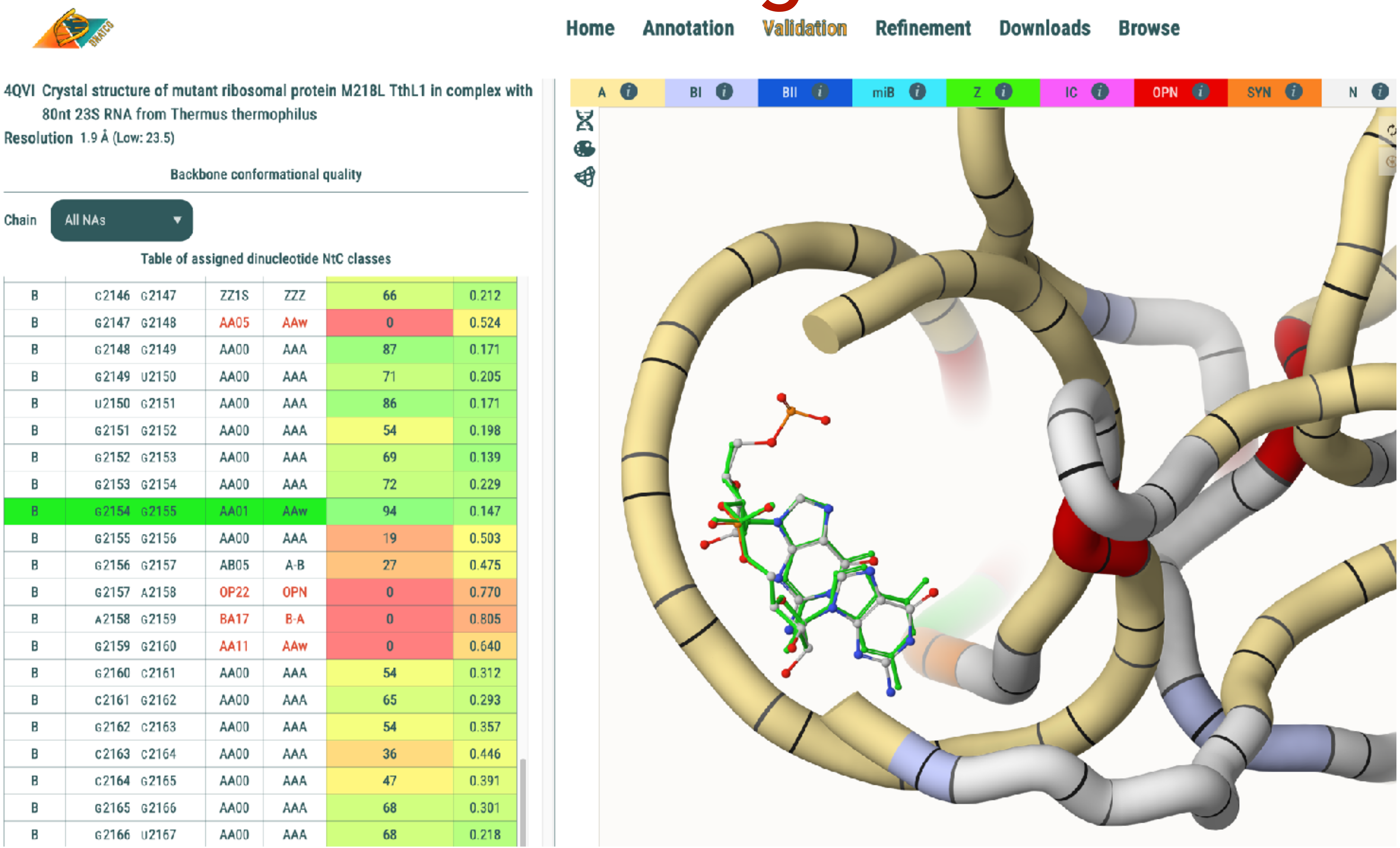


- NA backbone is conformationally complex:
  - the smallest unit for analysis is ***dinucleotide***
  - torsion angles correlate, 2- or 3-D correlations are misleading
- Complex analysis of the PDB data provided
  96 NtC classes

3. NtC: dinucleotide conformational classes

dnatco.datmos.org

- annotate, validate, and refine



4QVI Crystal structure of mutant ribosomal protein M218L TthL1 in complex with 80nt 23S RNA from Thermus thermophilus

Resolution 1.9 Å (Low: 23.5)

Backbone conformational quality

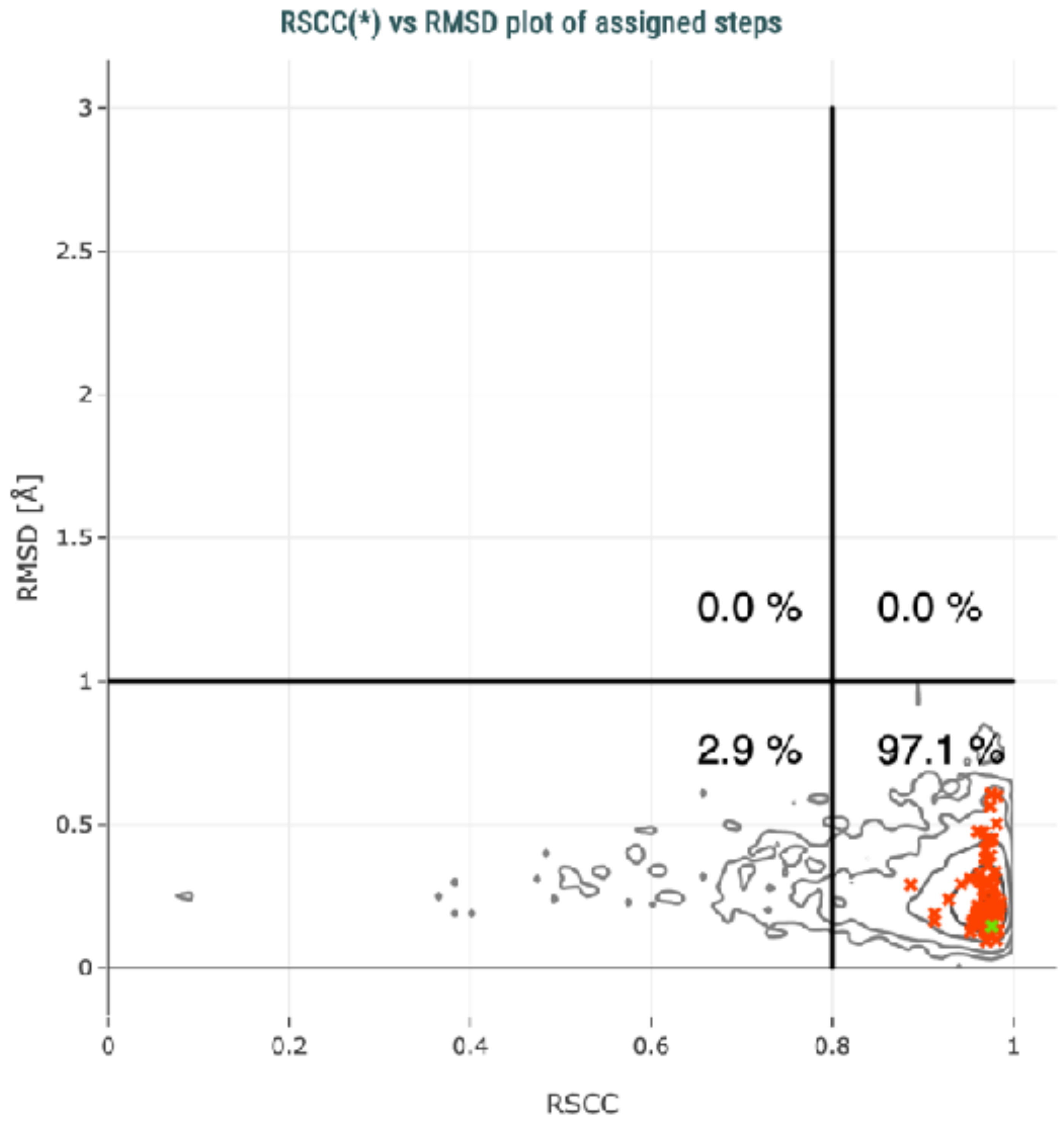
Chain All NAs

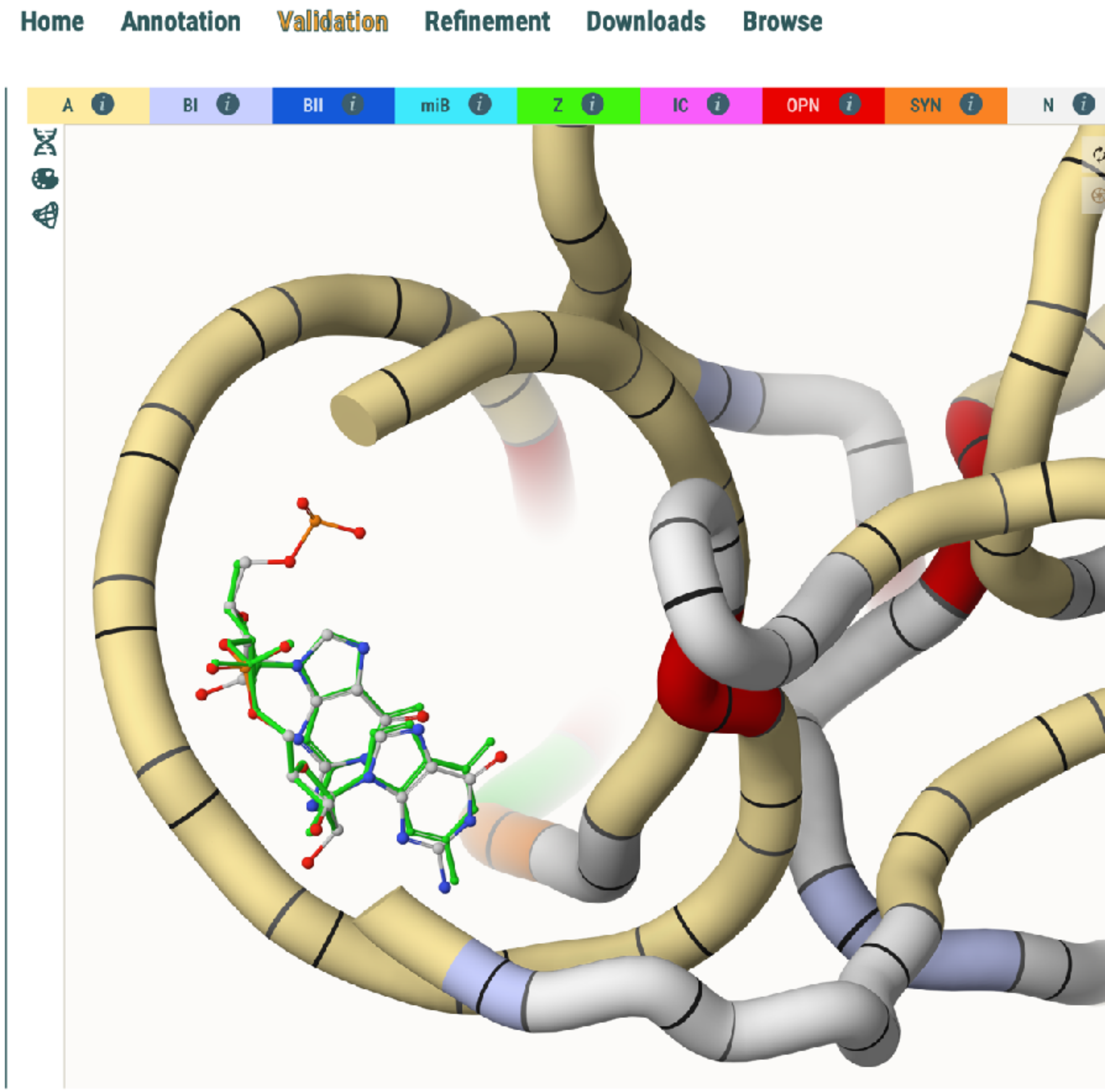
Table of assigned dinucleotide NtC classes

B	C2146 G2147	ZZ1S	ZZZ	66	0.212
B	G2147 G2148	AA05	AAw	0	0.524
B	G2148 G2149	AA00	AAA	87	0.171
B	G2149 U2150	AA00	AAA	71	0.205
B	U2150 G2151	AA00	AAA	86	0.171
B	G2151 G2152	AA00	AAA	54	0.198
B	G2152 G2153	AA00	AAA	69	0.139
B	G2153 G2154	AA00	AAA	72	0.229
B	G2154 G2155	AA01	AAw	94	0.147
B	G2155 G2156	AA00	AAA	19	0.503
B	G2156 G2157	AB05	A-B	27	0.475
B	G2157 A2158	OP22	OPN	0	0.770
B	A2158 G2159	BA17	B-A	0	0.805
B	G2159 G2160	AA11	AAw	0	0.640
B	G2160 C2161	AA00	AAA	54	0.312
B	C2161 G2162	AA00	AAA	65	0.293
B	G2162 C2163	AA00	AAA	54	0.357
B	C2163 C2164	AA00	AAA	36	0.446
B	C2164 G2165	AA00	AAA	47	0.391
B	G2165 G2166	AA00	AAA	68	0.301
B	G2166 U2167	AA00	AAA	68	0.218

3. New validation tool: rmsd/RSCC diagrams

- … fit between electron density (RSCC) & the closest NtC (rmsd)





Thank you for your time!

The dnatco team

Bohdan Schneider
bohdan.schneider@ibt.cas.cz

Lada Biedermannová
Paulína Božíková
Jiří Černý
Stanislav Lukeš
Michal Malý
Tereza Prchalová
Jakub Svoboda

Institute of Biotechnology of the Czech Academy of Sciences

The Czech Academy of Sciences