Annotation, Validation, Refinement, and Modeling of Nucleic Acid Structures

/in/cernylab

Jiří Černý, DNATCO & NA-VAL teams

Laboratory of Structural Bioinformatics, jiri.cerny@ibt.cas.cz

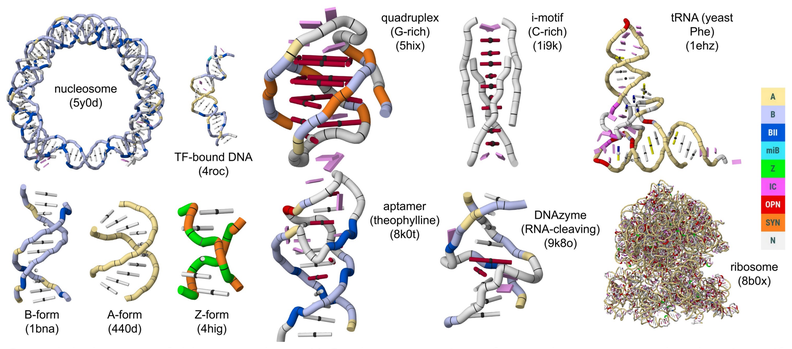
Institute of Biotechnology of the Czech Academy of Sciences, Průmyslová 595, 252 50 Vestec, Czech Republic

Recent advances in the DNATCO web server (dnatco.datmos.org) bring intuitive annotation, validation, modeling and refinement of nucleic acid structures, built on a universal structural alphabet that describes DNA and RNA backbone conformations. Its computational pipeline is now a C++ library that powers both the web server and standalone tools, including integration into structural biology software such as CCP4 (release 10). We also present the latest recommendations of the Nucleic Acid Valence Geometry Working Group, which is establishing standardized parameters for valence-geometry validation. Finally, we provide an overview of the challenges hindering current AI-driven prediction of nucleic acid structures and discuss how validated reference data sets and integrated computational tools can help address them.

1 · DNATCO v5.0 Web Platform

A universal DNA/RNA structural alphabet

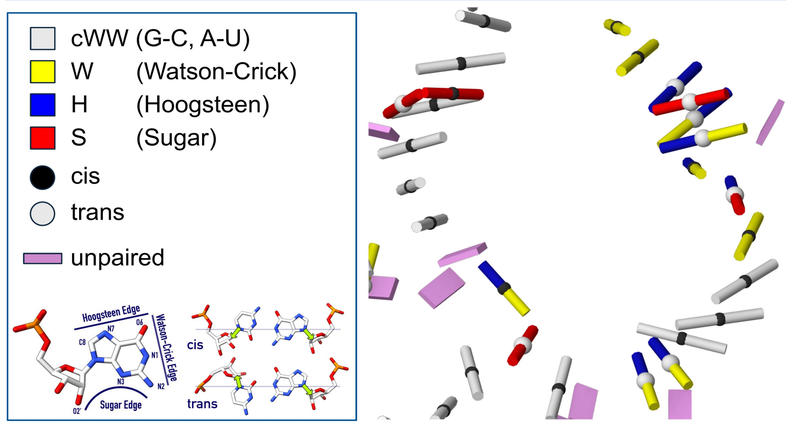
DNATCO (dnatco.datmos.org) integrates the *NtC* dinucleotide conformational classes and the *CANA* structural alphabet of nucleic acids, *i.e.* 96 conformers organized into 14 classes, to give a complete, geometrically defined description of local backbone and base orientation for both DNA and RNA. A *confal* score and an RMSD to the nearest NtC class quantify how closely each step in a deposited structure matches this reference alphabet.



One alphabet, any fold or topology: NtC-tube annotation of a nucleosome, quadruplex, i-motif, tRNA, aptamer, DNAzyme and the ribosome, alongside representative A-, B- and Z-form duplexes.

Visualization of Leontis-Westhof base pairing motifs

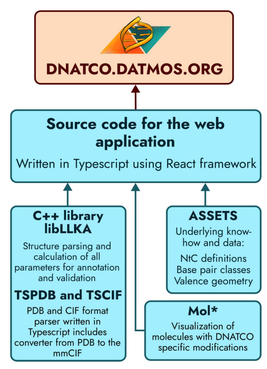
Every base pair identified by DNATCO is classified by its interacting edges, Watson–Crick, Hoogsteen or Sugar, and by the cis or trans orientation of the glycosidic bonds, following the Leontis–Westhof nomenclature; unpaired bases are flagged separately. Base pairs can be rendered as color-coded segments together with the 3D structure.



Base-pairing motifs across a fragment of the 8b0x ribosome (1.55 Å cryo-EM): edges color-coded Watson–Crick (yellow), Hoogsteen (blue) and Sugar (red); canonical cWW pairs in grey; unpaired bases in pink. Black and white open spheres mark cis/trans glycosidic-bond orientation.

Client-side, fast, and privacy-preserving

The structural-processing engine has been re-implemented from scratch as libLLKA, an open-source C++ library compiled to WebAssembly. All analysis now runs directly in the user's browser, over 50× faster than DNATCO v4.1, with no structural data ever leaving the user's computer. The same C++ core also drives dnatco.js, a standalone Node.js tool that runs the full analysis without the web GUI, writing the same validation reports and refinement restraints from the command line, the version bundled with CCP4 10, and the one wrapped by the containerized dnatco-nf pipeline for batch runs.



Open source (MIT)

github.com/cernylab/...

dnatco	the web application (and dnatco.js CLI)
libLLKA	C++ processing core
jsllka	WebAssembly wrapper
libnaval	NA-VAL geometry library
tspdb / tscif	PDB & mmCIF parsers
molstar	DNATCO fork of Mol*
dnatco-nf	Nextflow batch pipeline

```
nextflow run cernylab/dnatco-nf --input '*.cif.gz'
```

maxit.datmos.org companion service standardizing PDB/mmCIF files via the RCSB MAXIT suite; used only for files DNATCO cannot correct locally.

DNATCO v5.0 application architecture: the TypeScript/React web app and Mol viewer sit on the libLLKA C++/WASM core.*

Six analysis modules, one interface

- ***Annotation*** of backbone conformation and Leontis–Westhof base pairing, with an interactive base-pair ladder and 3D Mol* view
- ***Validation*** of conformational similarity and covalent geometry (see column 2)
- ***Refinement***: one-click torsion restraints for Refmac/Servalcat, Phenix, Coot and BUSTER
- ***Downloads***: NtC and base pairing-annotated mmCIF, CSV/JSON tables and full PDF validation reports
- ***Browse*** and ***Help***: searchable reference examples for every NtC class

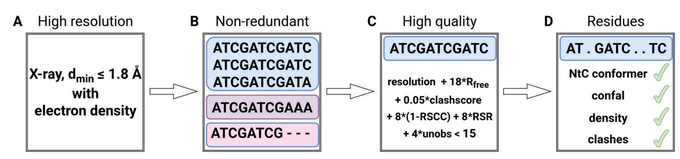
2 · New Standards for Valence Geometry

Why the 1996 standards fall short

PDB validation of nucleic-acid bond lengths and angles has, since 1996, relied on targets from a few hundred small-molecule Cambridge Structural Database (CSD) fragments, scored with a Z-score that assumes a Gaussian distribution. Re-examining a much larger, quality-filtered set of PDB structures, the Nucleic Acid Valence Geometry (NA-VAL) Working Group found that many parameters are in fact highly non-Gaussian or even multimodal, often simply because different refinement programs pull structures toward slightly different target values. However, genuine chemistry contributes as well: structural context, tautomeric and (de)protonation states, and the charge on the phosphate group each have their own characteristic bond lengths and angles, so some of the observed multimodality is real rather than an artifact.

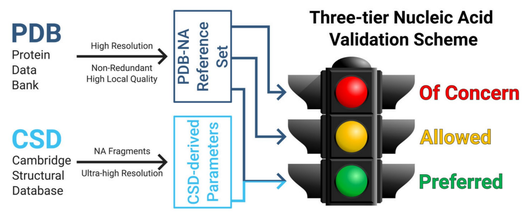
An empirical reference set

The Working Group curated the ***PDB-NA Reference Set***: sequence-non-redundant, high-resolution (≤1.8 Å) chains filtered for per-residue NtC conformational quality, fit to the electron density and absence of steric clashes - 3202 DNA and 2544 RNA residues in total.



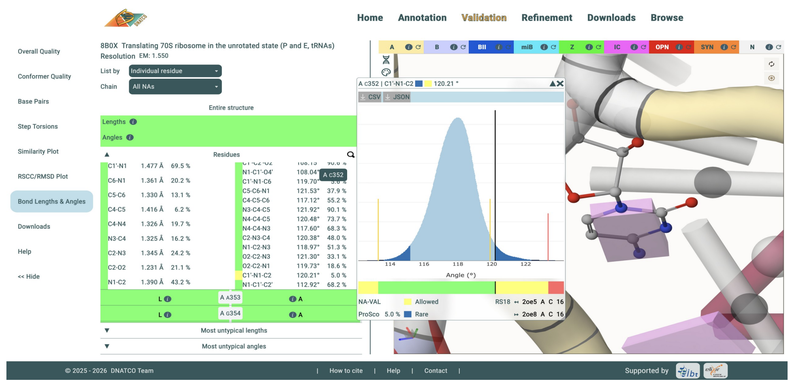
A three-tier validation scheme

Every bond length and angle is scored with a probability percentile (*ProSco*, 0 to 100) against the reference distribution and classified into three tiers, replacing the old, assumption-laden Z-score with a scheme robust to skewed and multimodal data. libnaval, the open-source C++ classification library, and the standalone DNATCO tool are the reference implementations of this scheme.



Live at dnatco.datmos.org

DNATCO v5.0 implements this protocol directly: every bond and angle in an uploaded structure is checked against the reference distributions and color-coded ***Preferred / Allowed / Of Concern***, with an interactive percentile plot for each parameter. Alongside the three tiers, an annotation grades every value as ***Common, Rare, Ambiguous or Unique***, shaded ***light to dark blue, reporting*** not whether the geometry is acceptable but ***how much reference data actually stands behind it***. Especially for parameters flagged as Of Concern, links to comparable cases in the PDB-NA Reference Set are provided, so a value arising in a similar structural context can be judged genuine rather than refined away.



How the whole PDB scores

Applying the new scheme across every DNA and RNA chain in the PDB, only a small minority of bonds and angles fall into the Of Concern tier, close to the ~0.1% expected for a well-refined structure. NMR ensembles are the striking exception, at several times the Of Concern rate of X-ray or cryo-EM, their valence geometry is determined largely by the force field, with only sparse experimental restraints.

NA	Method	Bonds (n checked)	Bonds Of Concern	Angles (n checked)	Angles Of Concern
DNA	X-ray	8.1M	**0.5%**	12.5M	**0.7%**
DNA	cryo-EM	9.0M	**0.5%**	13.9M	**0.8%**
DNA	NMR	5.0M	**2.2%**	7.6M	**4.7%**
RNA	X-ray	92.0M	**0.2%**	143.5M	**0.5%**
RNA	cryo-EM	132.5M	**0.6%**	206.6M	**0.6%**
RNA	NMR	6.2M	**0.9%**	9.6M	**2.3%**

Of-Concern rate for bond lengths and angles, by nucleic-acid type and experimental method, across the full PDB.

3 · AI-Driven Structure Prediction

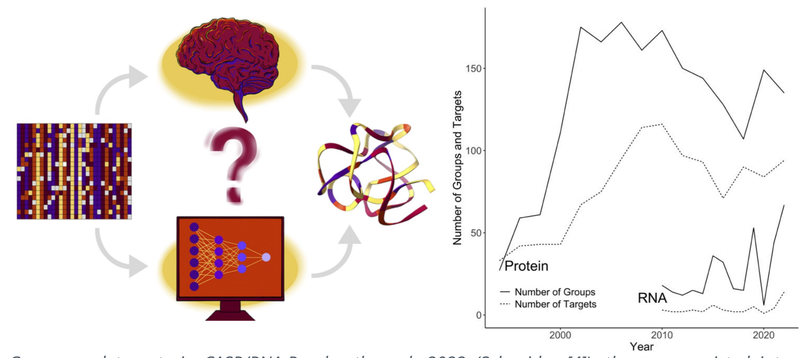
RNA hasn't had its "AlphaFold moment", yet

AlphaFold solved protein structure prediction by training on over hundred of thousands of high-resolution structures and deep sequence alignments. RNA has seen no comparable breakthrough: across CASP16's 42 targets (65 groups, 46 labs) and RNA-Puzzles Round V's 23 targets (18 groups), no blind prediction of a previously unseen natural RNA structure reached a TM-score above 0.8, and every ***top-ranked group, Vfold, GuangzhouRNA-human, KiharaLab***, was a human-expert team, not an automated pipeline.

"... prediction accuracy for nucleic acid complexes was generally poor unless 3D templates were available. Accounting for template availability, there has not been a notable increase in nucleic acid modeling accuracy between previous blind challenges and CASP16." Kretsch [5]

F1 < 0.5 non-canonical base pairs, best CASP16 (2024) groups. F1 runs 0 to 1; below 0.5, most predicted pairs are wrong or missed. The score counts only which nucleotides pair, not the specific edges involved.

8th → 19th AlphaFold 3 server's CASP16 rank on RNA targets with deep vs. shallow sequence alignments, behind human experts either way.



Groups and targets in CASP/RNA-Puzzles through 2022 (Schneider [4]); the gap persisted into CASP16 (2024), which still drew only 65 nucleic-acid predictor groups.

The data bottleneck

Data-hungry deep-learning methods are starved of RNA training data: high-resolution (≤2.0 Å) PDB structures contain roughly 100× more protein than RNA residues, and RNA sequence alignments (Rfam) average far fewer and shorter members than protein alignments (Pfam). RNA architecture also depends on non-Watson–Crick and tertiary contacts that are inconsistently or incompletely annotated across structures and software.

~25× more protein than RNA structures deposited in the PDB, a ratio that has stayed roughly constant for decades

~100× more high-resolution (≤2.0 Å) protein than RNA structures, the data richest in reliable 3D detail

Better ground truth, better models

DNATCO and the NA-VAL reference sets target exactly these gaps: consistent, automated base-pair and backbone annotation across the whole PDB; curated, non-redundant, quality-filtered structure sets suitable for benchmarking or training; and standardized valence-geometry targets that remove a major source of inconsistency between refinement programs, a common, reliable geometric ground truth for both machine-learning developers and structural biologists refining AI-generated models.

Reference sets ready for training and benchmarking

Beyond the strict ≤1.8 Å PDB-NA Reference Set used for valence-geometry validation, DNATCO publishes two general-purpose sets built for modeling and machine learning: RS25 (≤2.5 Å, 2316 DNA / 831 RNA chains) and RS35 (≤3.5 Å, 4329 DNA / 1678 RNA chains). Both keep only complete X-ray chains scoring CQS < 15, a composite of resolution, R_free, clashscore, RSCC, RSR and model completeness, taking up to two best chains per 90% sequence-identity cluster and treating naked and protein-bound nucleic acids as separate structural contexts. Companion "with contacts" versions add every nucleic-acid chain within 4 Å, preserving realistic interaction environments. All downloadable as CSV at the dnatco.datmos.org/app/browse/reference-sets address.

WANTED

MORE NUCLEIC ACID STRUCTURES

and the structural biologists who determine them

- **Annotate** backbone conformation with NtC / CANA and base pairing with the Leontis–Westhof scheme
- **Validate** conformation, base pairing and valence geometry
- **Refine** with NtC torsion restraints and NA-VAL valence geometry targets
- **Deposit** high-quality DNA and RNA structures; they join the curated reference sets that future models learn from

REWARD: an AlphaFold moment for RNA

We are offering collaboration and assistance with integrating DNATCO and NA-VAL into your tools and workflows.

Talk to us: jiri.cerny@ibt.cas.cz · dnatco.datmos.org

"We believe that it is possible to create an accurate RNA structure prediction method, but it will require solving several data quality and volume issues." Schneider [4]

References

1. Černý J, Malý M, Božíková P, Prchalová T, Svoboda J, Biedermannová L, Schneider B. DNATCO v5.0: integrated web platform for 3D nucleic acid structure analysis. Nucleic Acids Res. 2026. doi:10.1093/nar/gkaf1491
2. Černý J, Božíková P, Svoboda J, Schneider B. A unified dinucleotide alphabet describing both RNA and DNA structures. Nucleic Acids Res. 2020. doi:10.1093/nar/gkaa383
3. Černý J, Nicholls RA, Brzezinski D, Berman HM, Gilski M, Joosten RP, Kowiel M, Lawson CL, Moriarty NW, Richardson JS, Schneider B, Vonrhein C, Williams CJ, Jaskólski M, Egli M. New targets and procedures for validating the valence geometry of nucleic acid structures. Nucleic Acids Res. 2026. doi:10.1093/nar/gkaf1335
4. Schneider B, Sweeney BA, Bateman A, Černý J, Žok T, Szachniuk M. When will RNA get its AlphaFold moment? Nucleic Acids Res. 2023. doi:10.1093/nar/gkad726
5. Kretsch RC, Hummer AM, He S, Yuan R, Zhang J, Karagianes T, Cong Q, Kryshtafovych A, Das R. Assessment of nucleic acid structure prediction in CASP16. Proteins. 2026;94(1):192–217. doi:10.1002/prot.70072
6. Bu F, Adam Y, Adamiak RW, et al., Westhof E, Miao Z. RNA-Puzzles Round V: blind predictions of 23 RNA structures. Nat Methods. 2024;22:399–411. doi:10.1038/s41592-024-02543-9

Acknowledgments

The DNATCO team: Jiří Černý, Paulína Božíková, Michal Malý, Jakub Svoboda, Terezie Prchalová, Daniel Šrom, Lada Biedermannová, Bohdan Schneider.

The Nucleic Acid Valence Geometry Working Group: Jiří Černý, Robert A Nicholls, Dariusz Brzezinski, Helen M Berman, Miroslaw Gilski, Robbie P Joosten, Marcin Kowiel, Catherine L Lawson, Nigel W Moriarty, Jane S Richardson, Bohdan Schneider, Clemens Vonrhein, Christopher J Williams, Mariusz Jaskólski, Martin Egli

Supported by the ***ELIXIR CZ Research Infrastructure*** (ID LM2023055, MEYS CR) and by the ***Institute of Biotechnology of the Czech Academy of Sciences*** (RVO 86652036).