# Identification of Basepairs in 3D Structures

Bohdan Schneider & Base Pairing Working Group

*Computational Approaches to RNA Structure and Function*

*2024-07-30*

Czech Academy of Sciences

# Program

- Basepairing provides fundamental information about RNA
- Determining basepairs from 3D structures
- New algorithm to assign pairs
- Validation of basepair geometries

# Basepairing provides fundamental information

- Watson-Crick, or canonical G-C and A-U pairs form the core of RNA structures.
- Basepairs critical for RNA fold are non-canonical, often sequentially distant.
- Lack of correctly assigned non-W-C pairs hamper ability to:

  A. predict  2D & 3D structures.

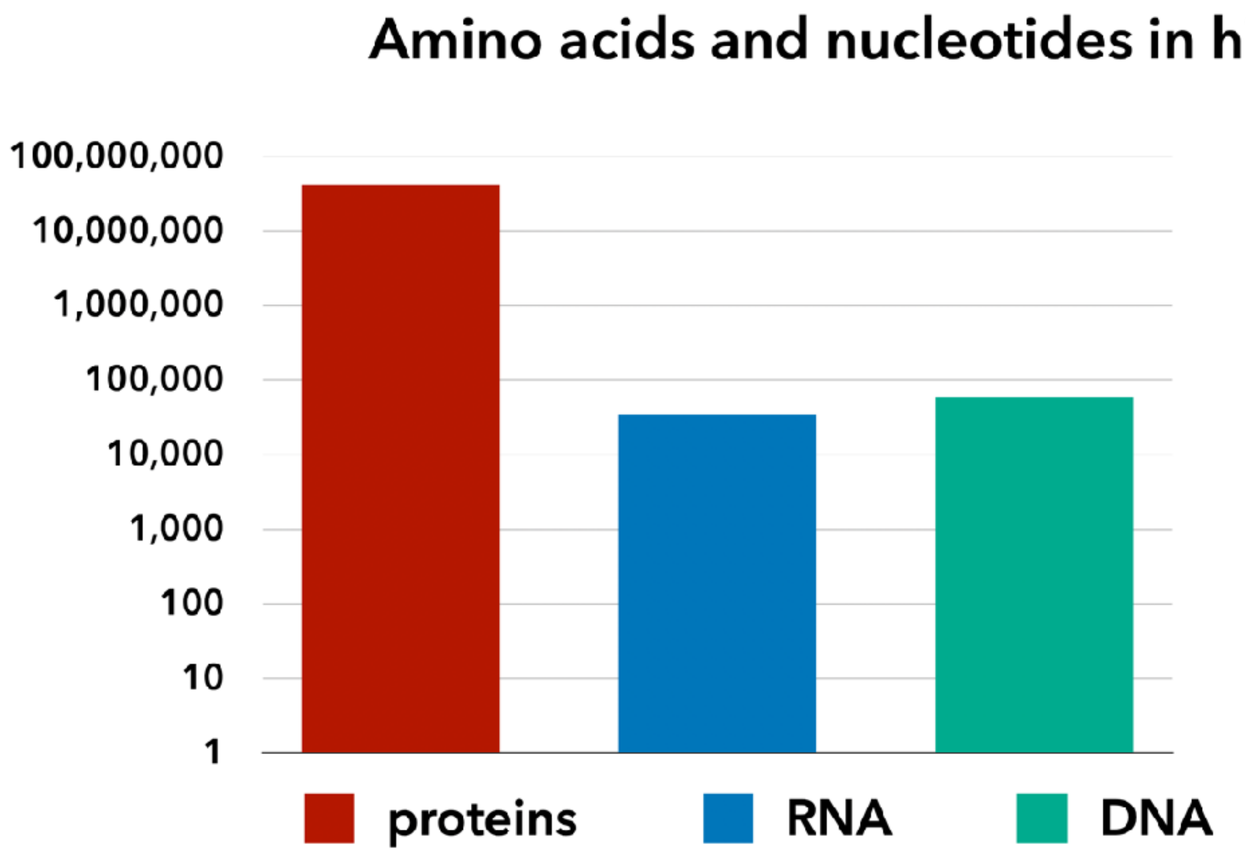
  B. run more robust sequence alignments.

# Problems with Basepair Assignment in 3D Structures

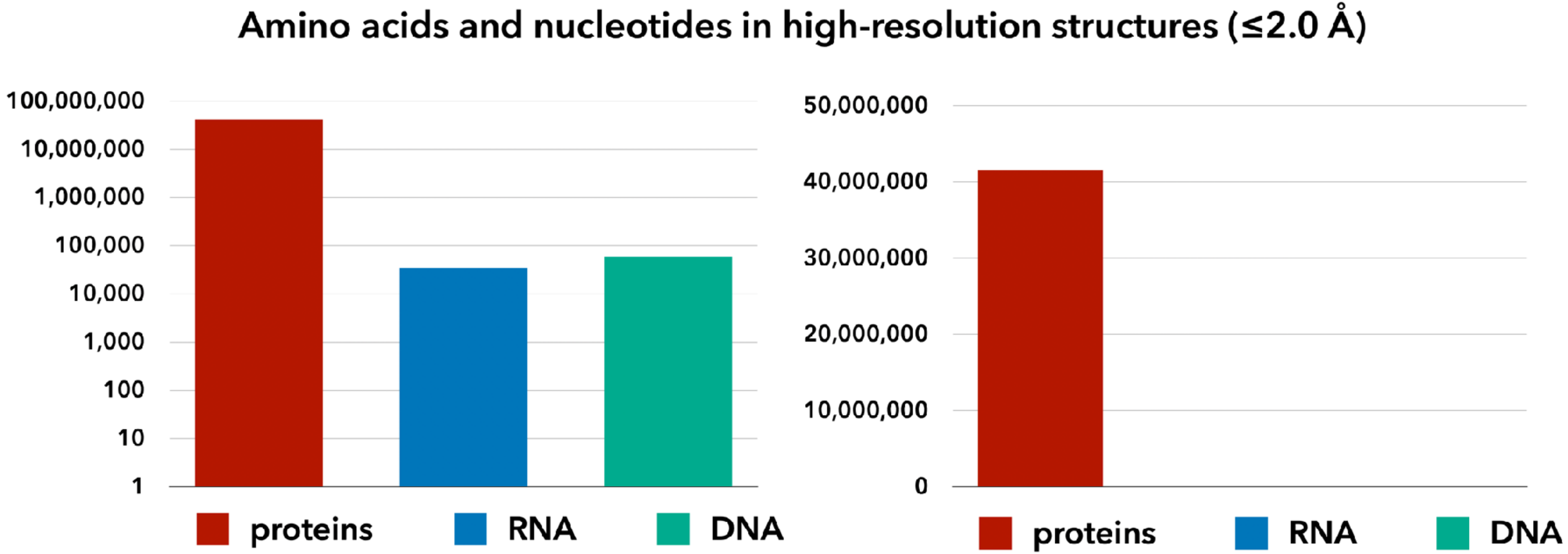
- There is not enough high-quality data.

- PDB-annotated basepairs.
- Free programs assign basepairs inconsistently.

# There Are not Enough High Quality NA Models

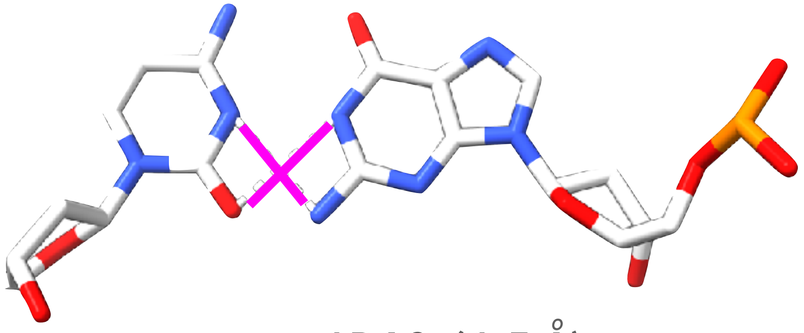


# There Are not Enough High Quality NA Models

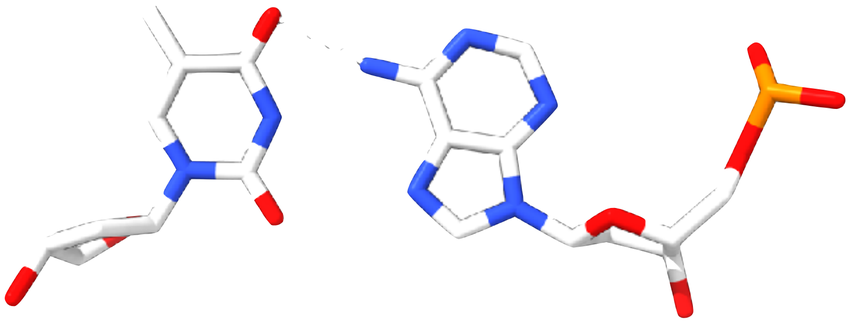


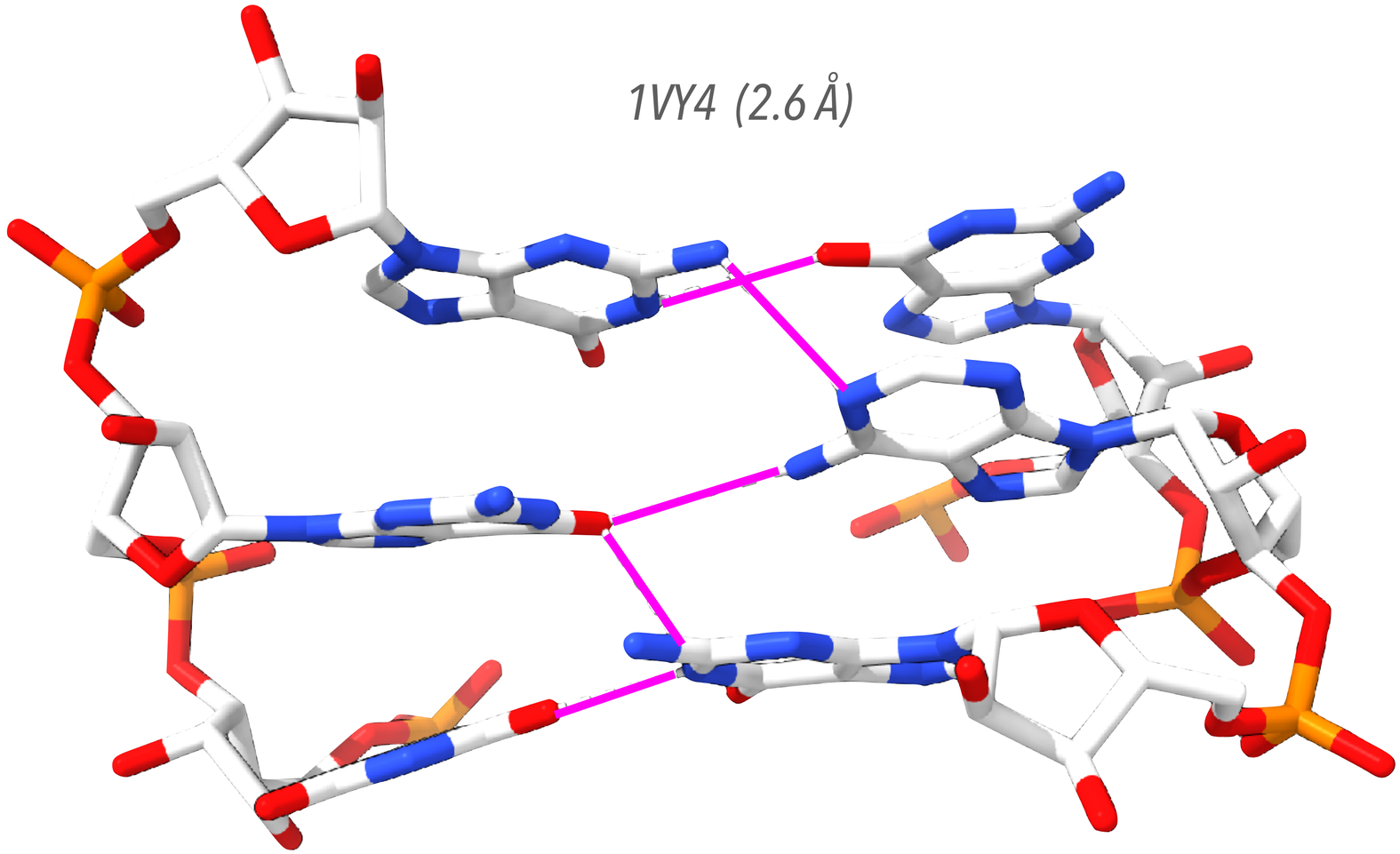
# PDB-Annotated Basepairs Are often Incomplete & Incorrect

mmCIF category *ndb_struct_na_base_pair* sometimes missing; order of bases not defined



*1DA0 (1.5 Å)*
*2PIS (2.8 Å)*





*1VY4 (2.6 Å)*

# Free programs assign basepairs inconsistently

- FR3D
  - Sarver, M., C. L. Zirbel, J. Stombaugh, A. Mokdad and N. B. Leontis (2008): FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol 56: 215-252.
- PDB-annotated pairs
  - program ***maxit***, mmCIF category ***ndb_struct_na_base_pair***
- ClaRNA
  - Walen, T., G. Chojnowski, P. Gierski and J. M. Bujnicki (2014): ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. Nucleic Acids Res 42: e151.
- MC-annotate
  - Gendron, P., S. Lemieux and F. Major (2001): Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol 308: 919-936
- DSSR
  - Li, S., W. K. Olson and X. J. Lu (2019): Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. Nucleic Acids Res 47: W26-W34.
- RNAView
  - Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman HM, Westhof E (2003): Tools for the automatic identification and classification of RNA base pairs. Nucleic Acids Res 31: 3450-3460.

# Benchmarking Identified Various Levels of Problems

| | DNA | RNA | mmCIF | symmetry | models | alt | ins | scores |
|---|---|---|---|---|---|---|---|---|
| PDB | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ |
| DSSR | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ |
| FR3D | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ |
| contacts (*) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| clarna | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ~ | ✓* |
| MC-Annotate | ✗ | ✓ | ✗ | ✗ | ✓* | ~ | ✓ | ✗ |
| rnaview | ✓* | ✓ | ✗ | ✗ | ~ | ~ | ~ | ✗ |

*(*) Contacts is being developed at IBT, not published*

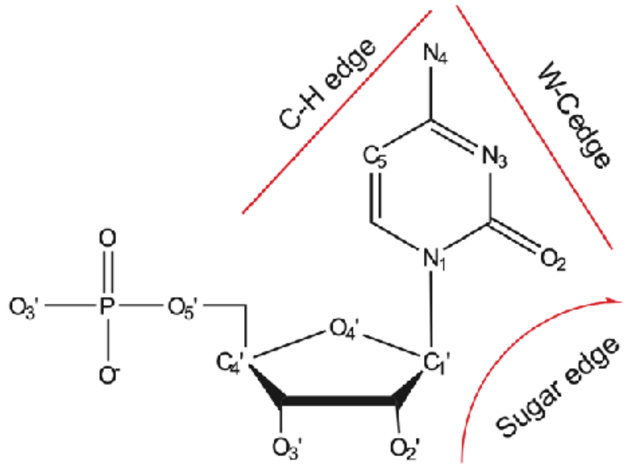
ibt

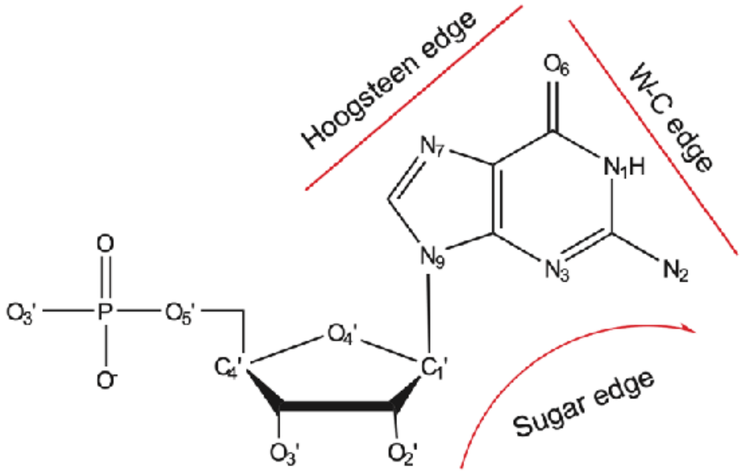
# Going beyond the Benchmarking: Develop a New Method of Basepair Assignment

- Goal:

  ***to define universal set of geometric parameters applicable to all types of basepairs***

- The parameters must describe all basepair classes as described in the Leontis-Westhof classification
  - twelve L-W classes plus combinations of A/G/U/C bases in the pairs generate ~150 types of which 127 are observed
- Describe both DNA and RNA




| | Glycosidic bond | Interacting edges | Local strand |
|---|---|---|---|
| 1 | Cis | Watson–Crick/Watson–Crick | Antiparallel |
| 2 | Trans | Watson–Crick/Watson–Crick | Parallel |
| 3 | Cis | Watson–Crick/Hoogsteen | Parallel |
| 4 | Trans | Watson–Crick/Hoogsteen | Antiparallel |
| 5 | Cis | Watson–Crick/Sugar Edge | Antiparallel |
| 6 | Trans | Watson–Crick/Sugar Edge | Parallel |
| 7 | Cis | Hoogsteen/Hoogsteen | Antiparallel |
| 8 | Trans | Hoogsteen/Hoogsteen | Parallel |
| 9 | Cis | Hoogsteen/Sugar Edge | Parallel |
| 10 | Trans | Hoogsteen/Sugar Edge | Antiparallel |
| 11 | Cis | Sugar Edge/Sugar Edge | Antiparallel |
| 12 | Trans | Sugar Edge/Sugar Edge | Parallel |

# New Transparent Set of Parameters Used to Assign Basepairs

1. Hydrogen bonds
   - H-bonds –-> atom-atom contacts
2. Co-planarity of the bases

- The redundant set of parameters contains 26 parameters
  - distribution minima + maxima are assigned
  - the limits have distinct values for different basepair classes

# Simplified Protocol of Basepair Assignment

- Select the ***Reference Set*** of structures/nucleotides.
- Calculate all base-base contacts ≤ 4.0 Å.
- Sort the contacts by sequences.
  - A-A, A-C, A-U, A-G, C-U, …
- These pairs represent ***potential basepairs***.
  - … in the PDB, a few millions in the Reference Set hundreds of thousands …
- For all ***potential basepairs***:

  Calculate the parameter values.

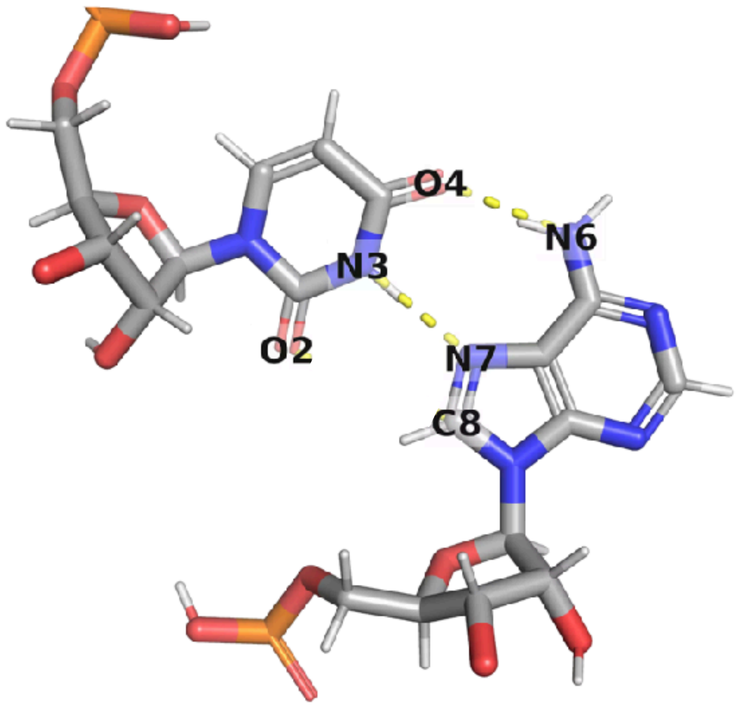
  Sort the parameter values by values expected in individual L-W classes.

  Assign the class.

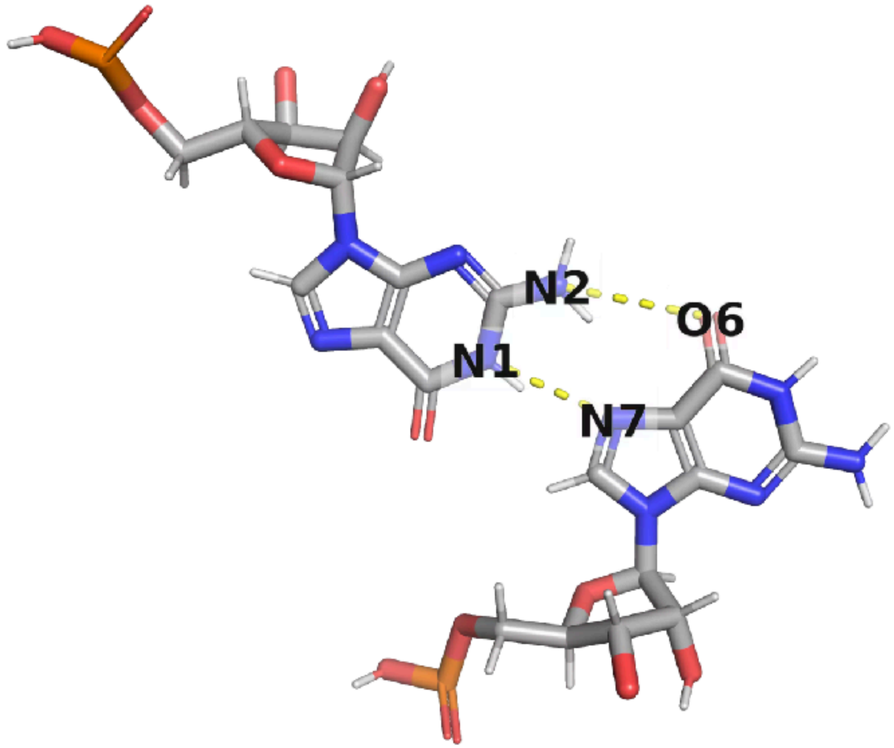
  *Validate quality of the assigned pair (in progress).*

# 1. Hydrogen Bonds –> Atom-Atom Contacts

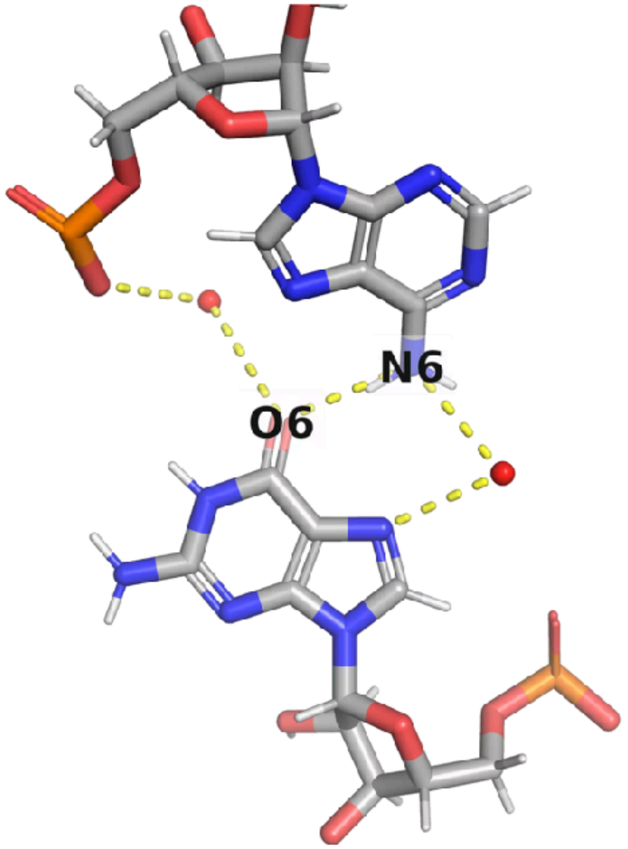
*cWH U-A*



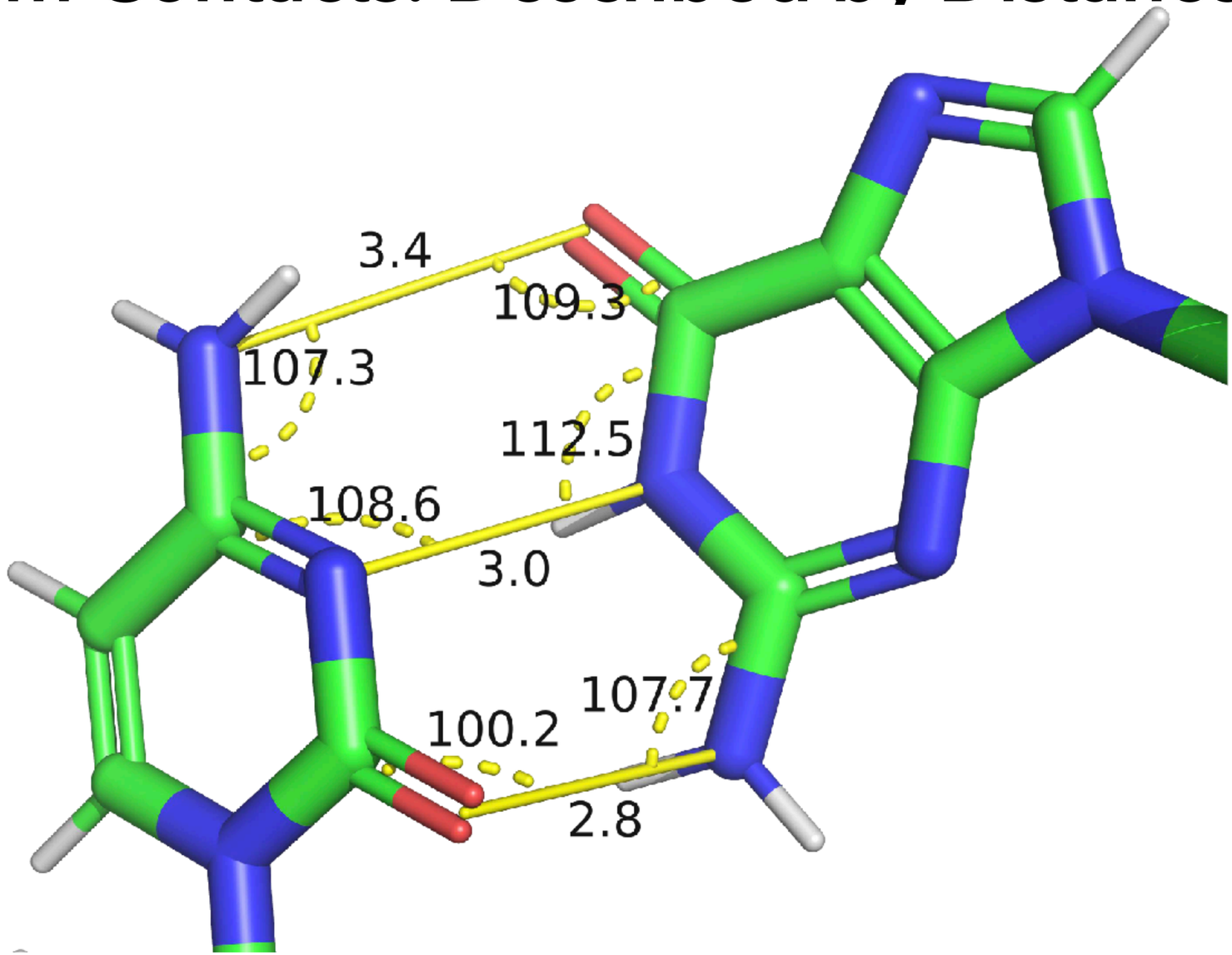

*tWH G-G*




*tHH A-G*



# Atom-Atom Contacts: Described by Distances and Angles




*3 Atom-Atom Contacts … 9 Parameters*

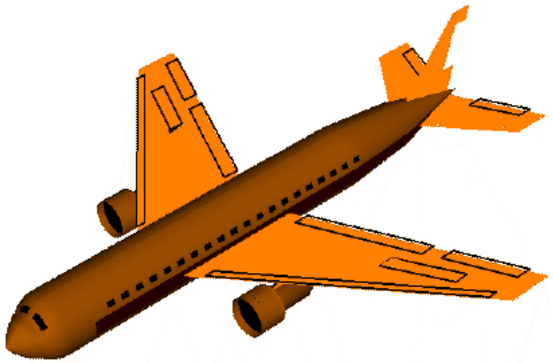
# 2. Co-planarity of the Bases: a Complicated Geometric Problem …

- … solved by defining a redundant set of angles and distances between base parts.

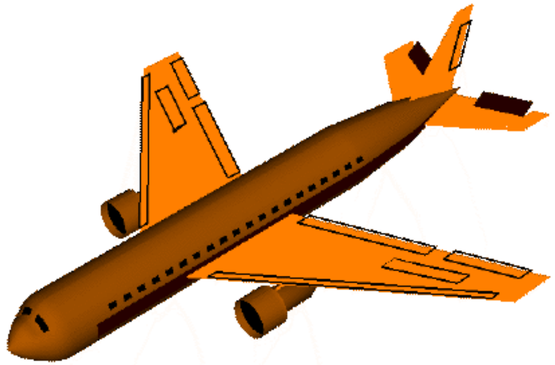
- Aircraft angles
  - yaw-pitch-roll
- Angles and distances between the edges and the bases.
  - … calculated between the edge of ***BASE1*** to the plane of ***BASE2.***
- Angles between the hydrogen bond vectors and the base planes.

# Aircraft angles

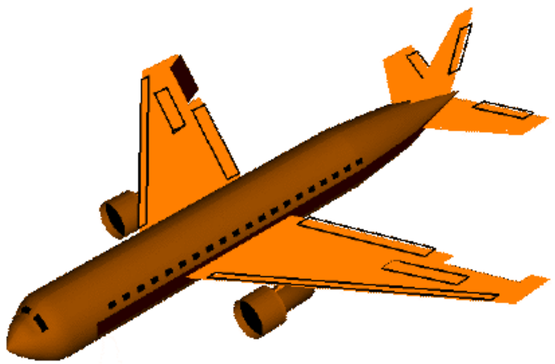
Yaw



Pitch

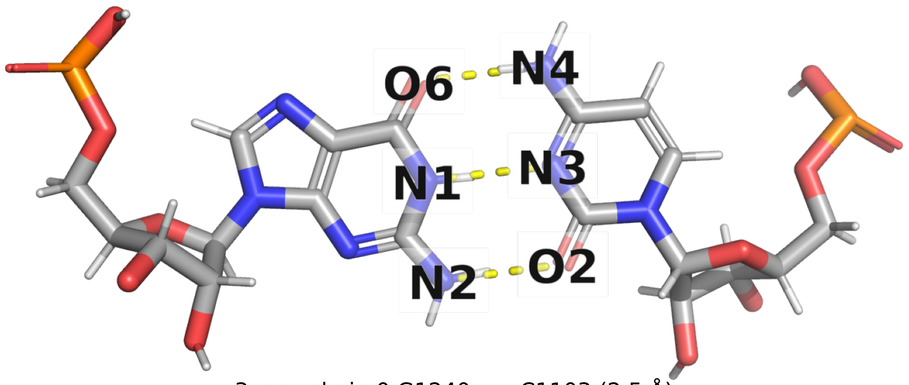


Roll

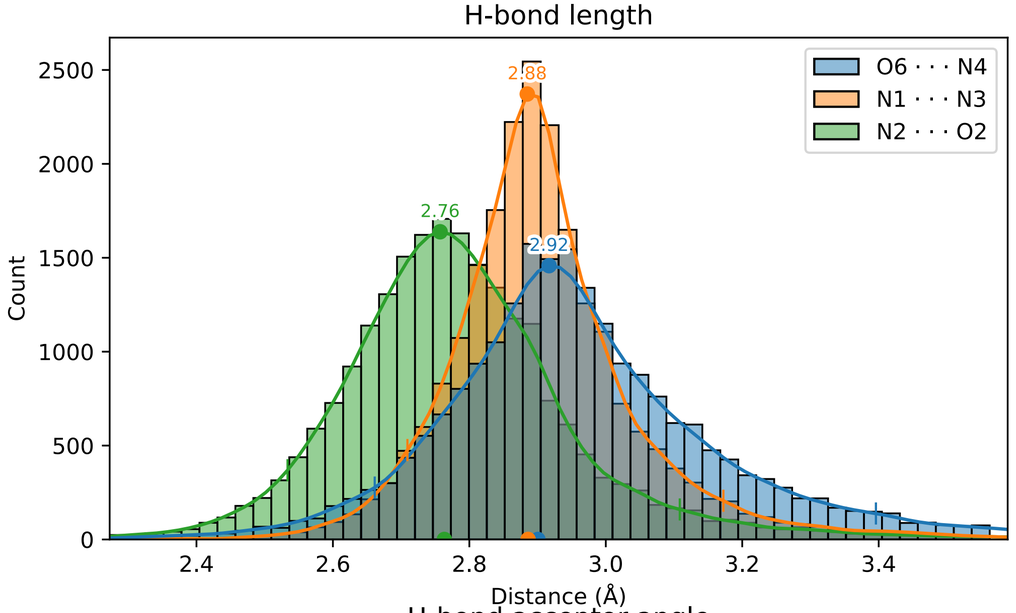


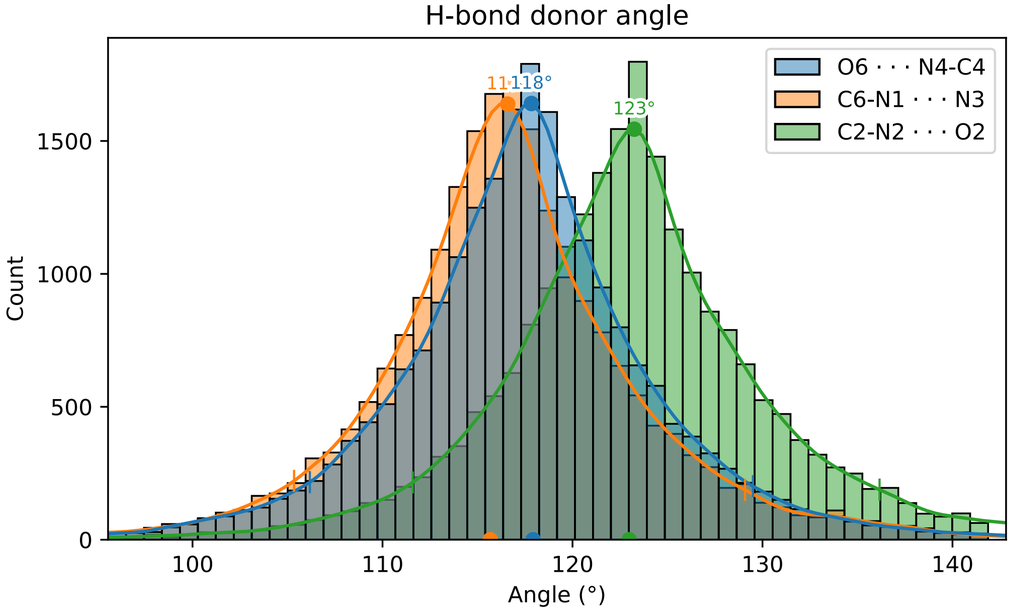
*https://en.wikipedia.org/wiki/Aircraft_principal_axes*

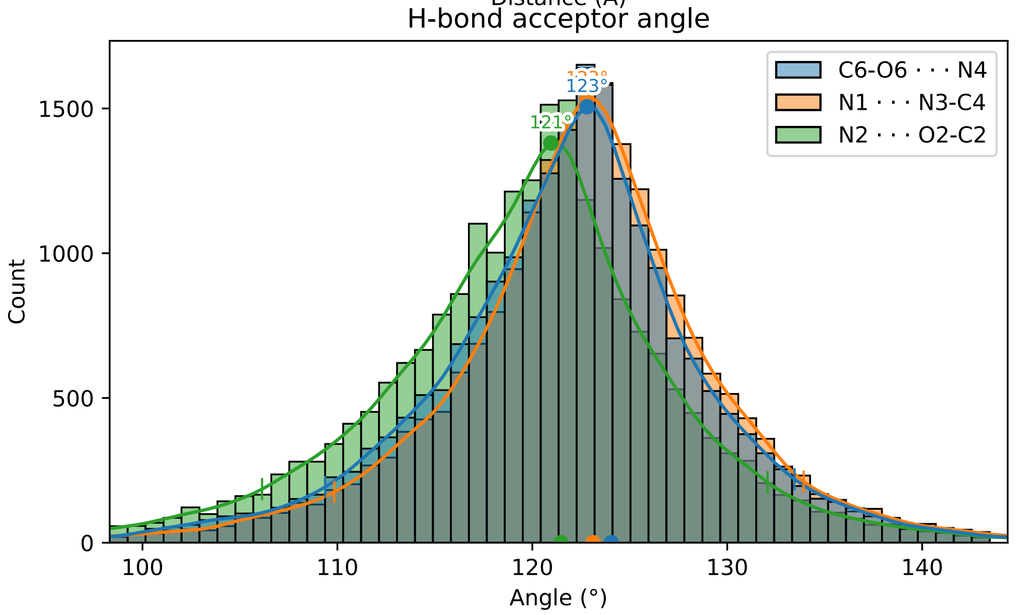
cWW GC RNA ≤3 Å - H-bonds (21893 observations)



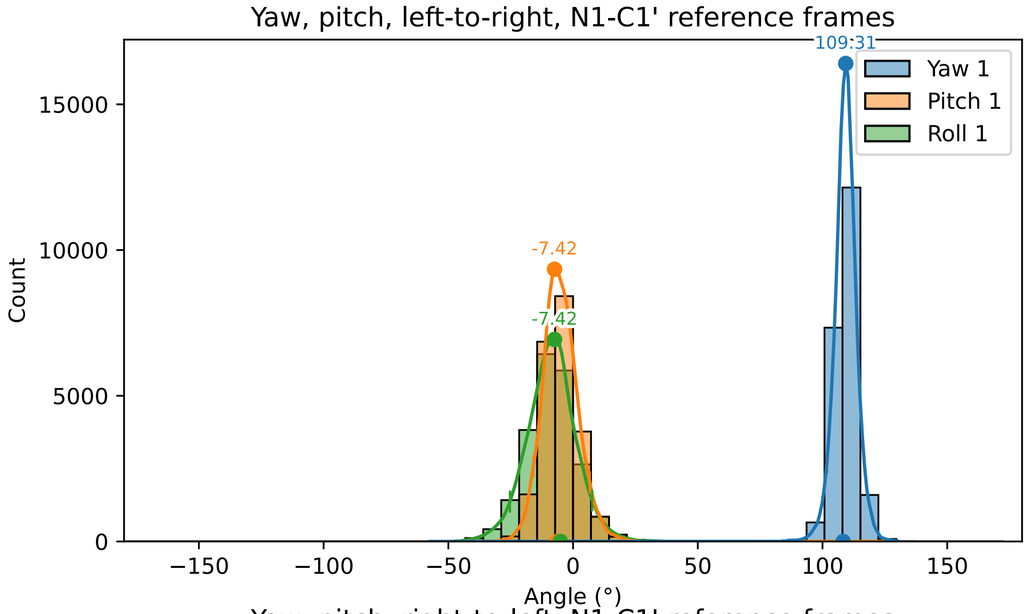

3ccm: chain 0 G1240 · · · C1103 (2.5 Å)

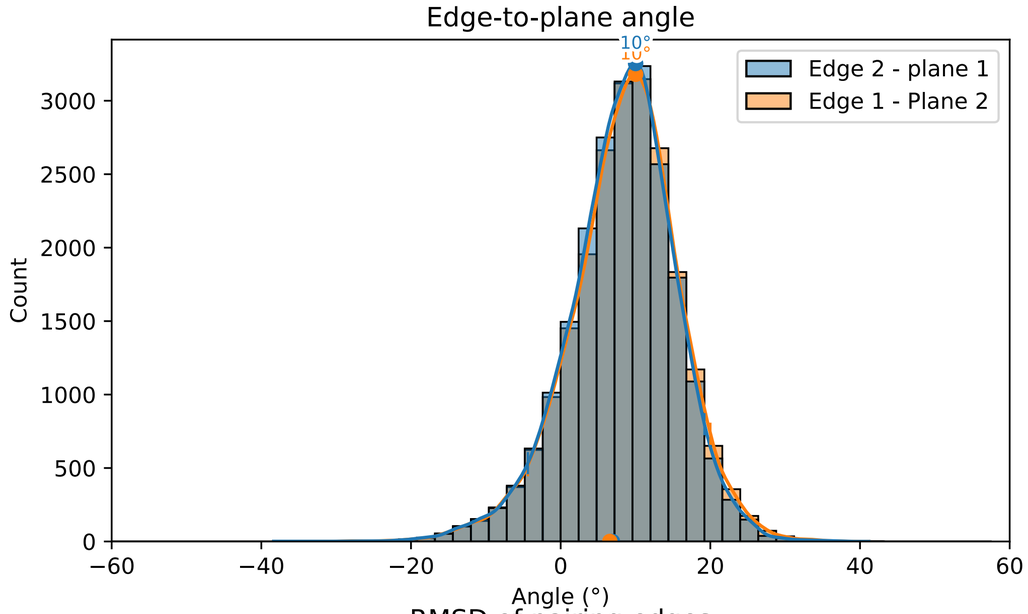


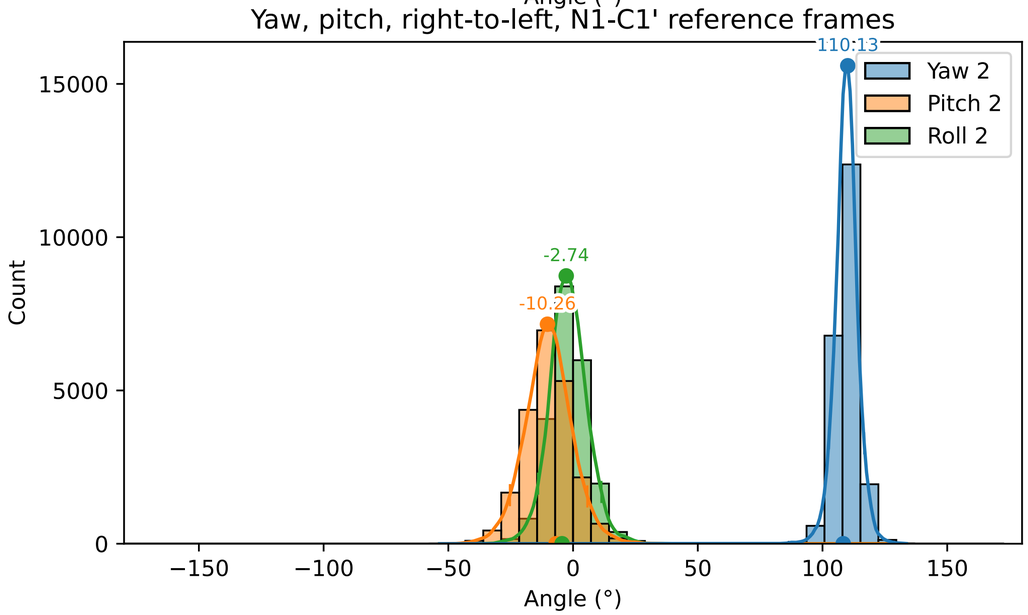



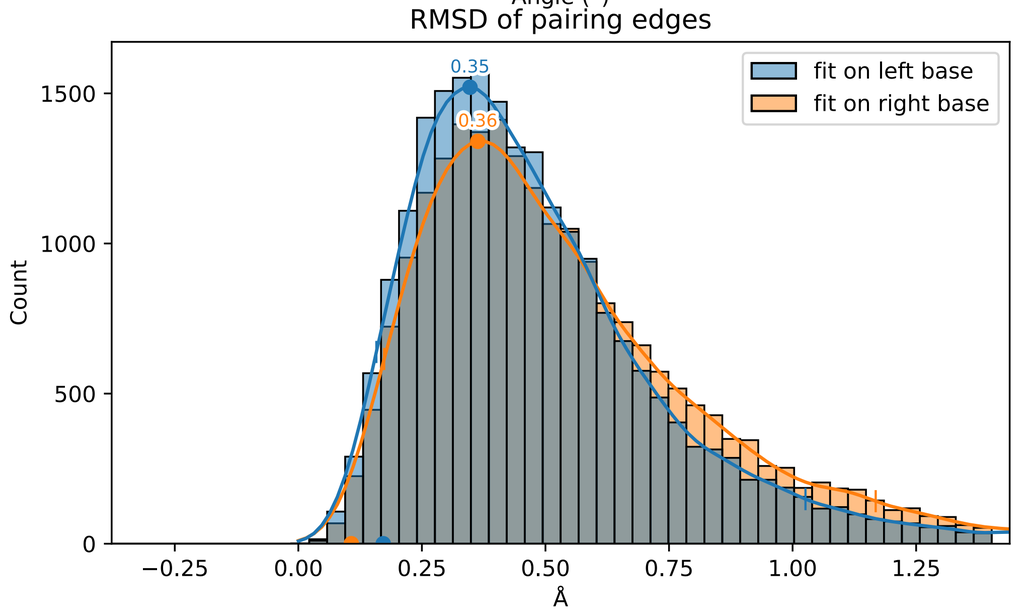



cWW GC RNA ≤3 Å - Coplanarity (21889 observations)



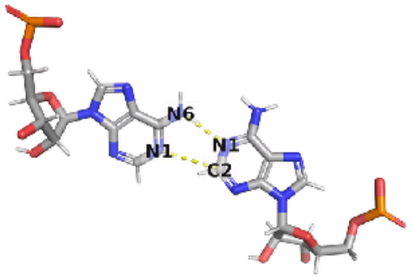
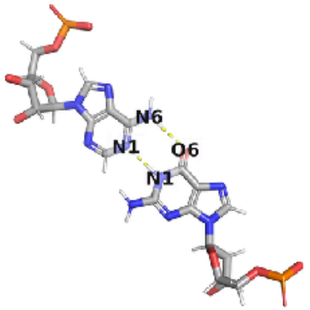
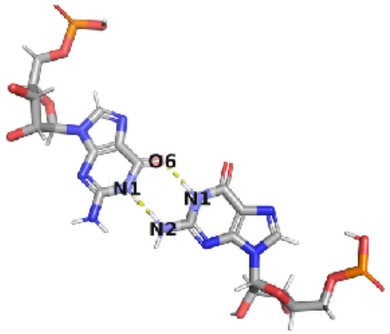



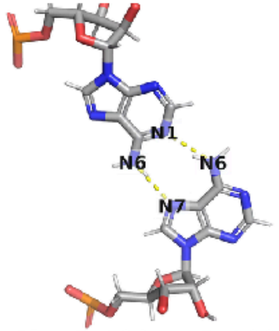
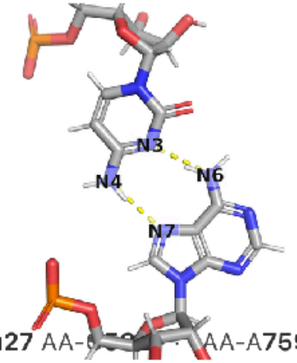
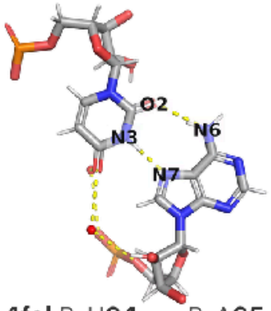
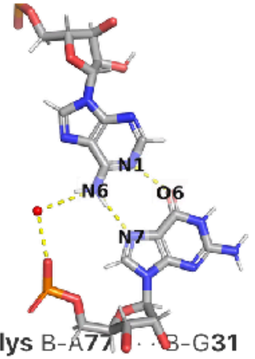
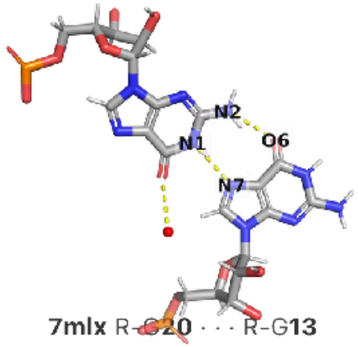
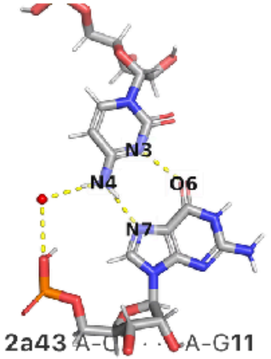
# The New Assignment Is at *basepairs.datmos.org*

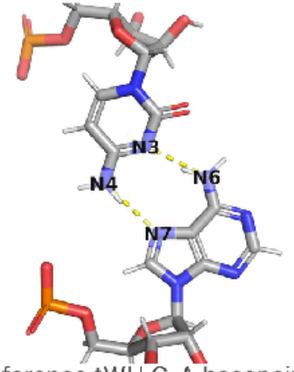
- The web allows interactive play with the assignment by changing the structure ensemble and limiting values of all parameters
- Development version!!!

# 1: cis Watson-Crick / Watson-Crick

| | A | G | C | U |
|---|---|---|---|---|
| A |  **6dme** A-A**21** · · · A-A**34** show statistics + exemplars | same as **A-G** | same as **A-C** | same as **A-U** |
| G |  **4v9h** BA-A**2019** · · · BA-G**2035** show statistics + exemplars |  **4y4o** 1a-G**741** · · · 1a-G**664** show statistics + exemplars | same as **G-C** | same as **G-U** |

# 4: trans Watson-Crick / Hoogsteen

| | A | G | C | U |
|---|---|---|---|---|
| A |  **4feo** B-A**33** · · · B-A**66** show statistics + exemplars | not defined |  **4u27** AA-[illegible] · · · AA-A**759** show statistics + exemplars |  **4fel** B-U**34** · · · B-A**65** show statistics + exemplars |
| G |  **5lys** B-A**77** · · · B-G**31** show statistics + exemplars |  **7mlx** R-G**20** · · · R-G**13** show statistics + exemplars |  **2a43** A-[illegible] · · · A-G**11** show statistics + exemplars | not defined |



Reference tWH C-A basepair

◉ Basic ○ Parameter ranges ○ SQL

**Data source**

All Polar Contacts — Reference Set

**Order by**

Largest Edge RMSD

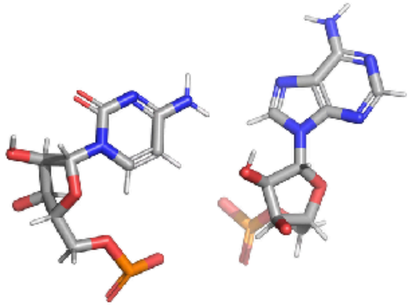
○ RNA ○ DNA ◉ Both

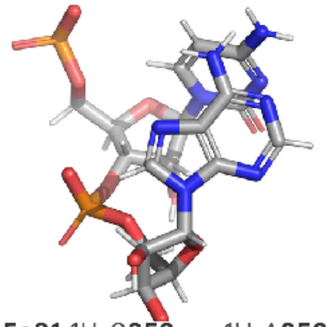
☐ Rotate images

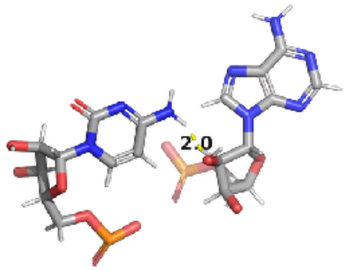
Reset filters

Enable FR3D comparison

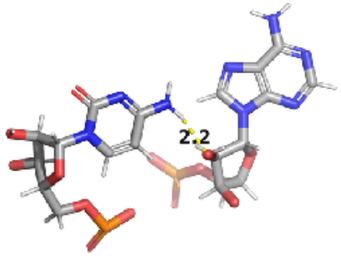
**1994** × tWH-C-A from **559** PDB structures ▽ expand plots ▽


**4woi** DW-C**11** · · · DW-A**9**

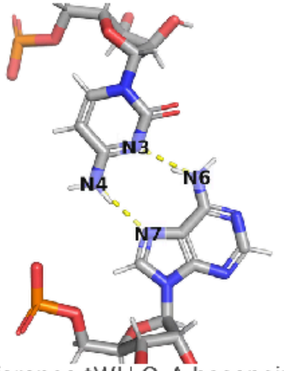

**5e81** 1H-C**652** · · · 1H-A**653**



**6ufm** A-C**11** · · · A-A**9**



**6cu1** A-C**56** · · · A-A**54**



Reference tWH C-A basepair

Basic | Parameter ranges | SQL

**Data source**

Pairs Selected by New Parameters

**Order by**

Smallest Edge RMSD

RNA | DNA | Both

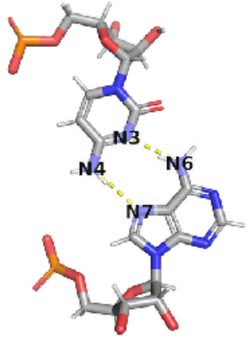
Rotate images

Reset filters

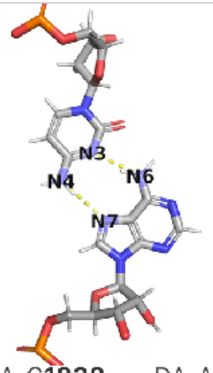
Enable FR3D comparison

**387** × tWH-C-A from **107** PDB structures

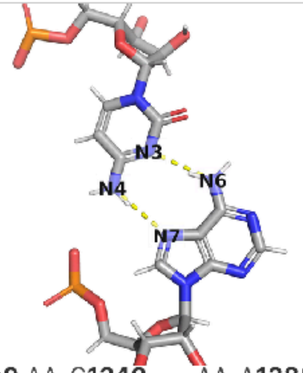
▽ expand plots ▽



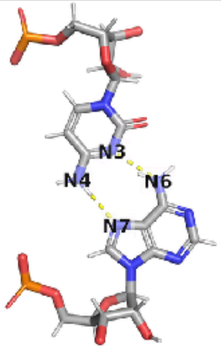
**3l3c** Q-C**122** · · · Q-A**99**



**5j5b** DA-C**1838** · · · DA-A**1901**



**5jc9** AA-C**1249** · · · AA-A**1288**



**3ccm** 0-C**1426** · · · 0-A**1437**

Basic | Parameter ranges | SQL

| | N4 · · · N7 | N3 · · · N6 |
|---|---|---|
| LENGTH | Min Max | Min Max |
| ACCEPTOR ANGLE | Min Max | Min Max |
| DONOR ANGLE | Min Max | Min Max |

RNA | DNA | Both

**Data source**

Pairs Selected by New Parameters, RS

Edit the selection boundaries

**Resolution ≤** 3.5 Å

**Order by**

Smallest Edge RMSD

**Comparison baseline**

FR3D | Set to this

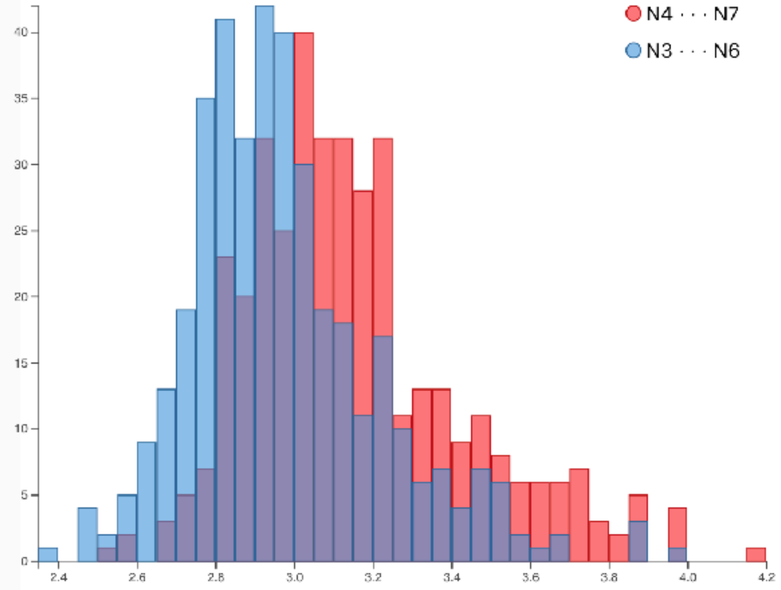
Rotate images

Download ▾

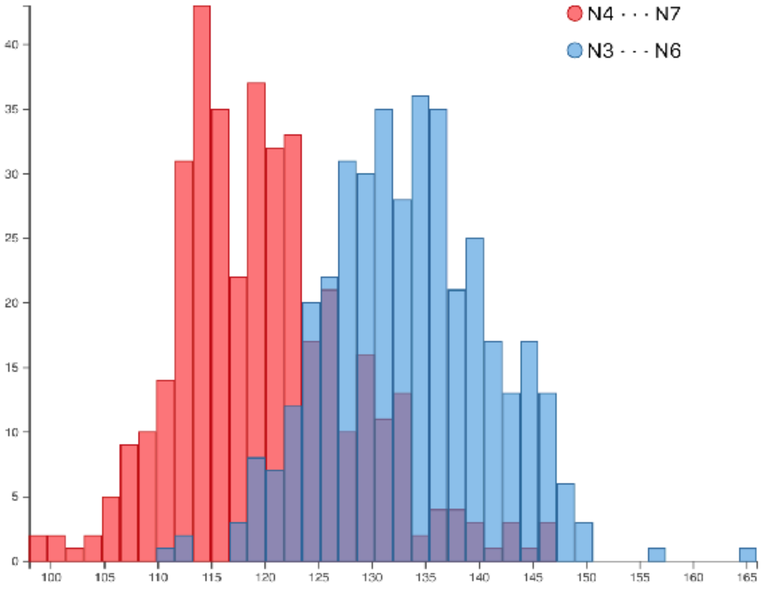
**387** × tWH-C-A from **107** PDB structures

▲ collapse plots ▲

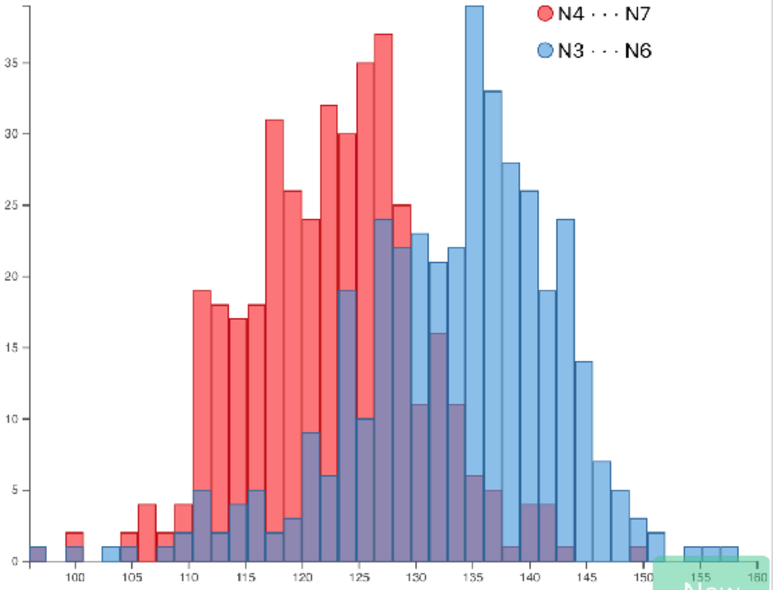
**H-bond length (Å)**




**H-bond donor angle (°)**




**H-bond acceptor angle (°)**




New

# The Next Step: Find the Core Regions of the Parameters Defining “Reference Basepairs”

- The Reference Basepairs will be selected from parameter distributions by Kernel Density Estimation
  - the limits of the KDE values will be determined so that numbers of Reference Basepairs will be:
  - for large classes (cWW GU and the like) ~ a thousand
  - for midsize classes (tWH UA) ~hundreds
  - for small classes with less than a hundred cases, KDE will be expert-adjusted
- rmsd between the validated and the Reference Basepairs will be the validation measure.

# To Be Done

- Be able to assign basepairs on the fly for any uploaded structure.
- Finish development of validation protocol.
  - necessary to have a quantitative measure of basepair quality.
- Mature the web *basepairs.datmos.org*.
- Integrate new basepair assignment into the annotation web *dnatco.datmos.org*.

# Thank you for your time!

- Initiated by ELIXIR 3D-Bioinfo Community
- Supported by enthusiasm of the Base Pairing Working Group:

Helen Berman

Lada Biedermannová

*Jiří Černý*

Robbie Joosten

Catherine Lawson

***Stanislav Lukeš***

Marcin Magnus

Brinda Vallat

Eric Westhof

*Craig Zirbel*

Bohdan Schneider