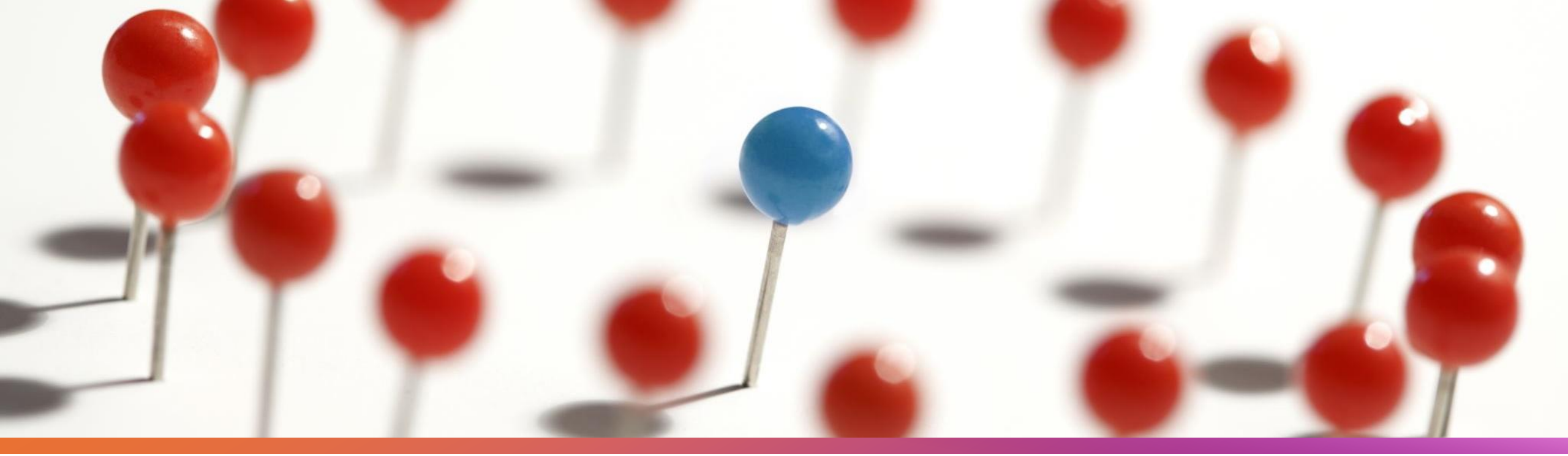


Finding conserved disruptive base pairs in Multi-Sequence Alignments

… and insights for RNAPOND Strategies

By Patrick-Pascal Koller

Supervised by Hua-Ting Yao and Ivo Hofacker

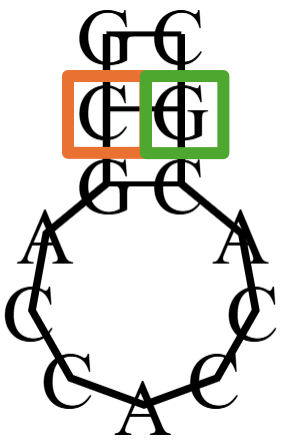
What Are Conserved Base Pairs?

- They are pairs of nucleotides that will always pair across different species
- Instead of exact nucleotide conservation we consider covariation

```
>Organism_1
GCGACCACCACGC
>Organism_2
CGCCAACAACGCG
>Organism_3
GCGACCACCACGC
```

2/3 C
1/3 G

1/3 C
2/3 G

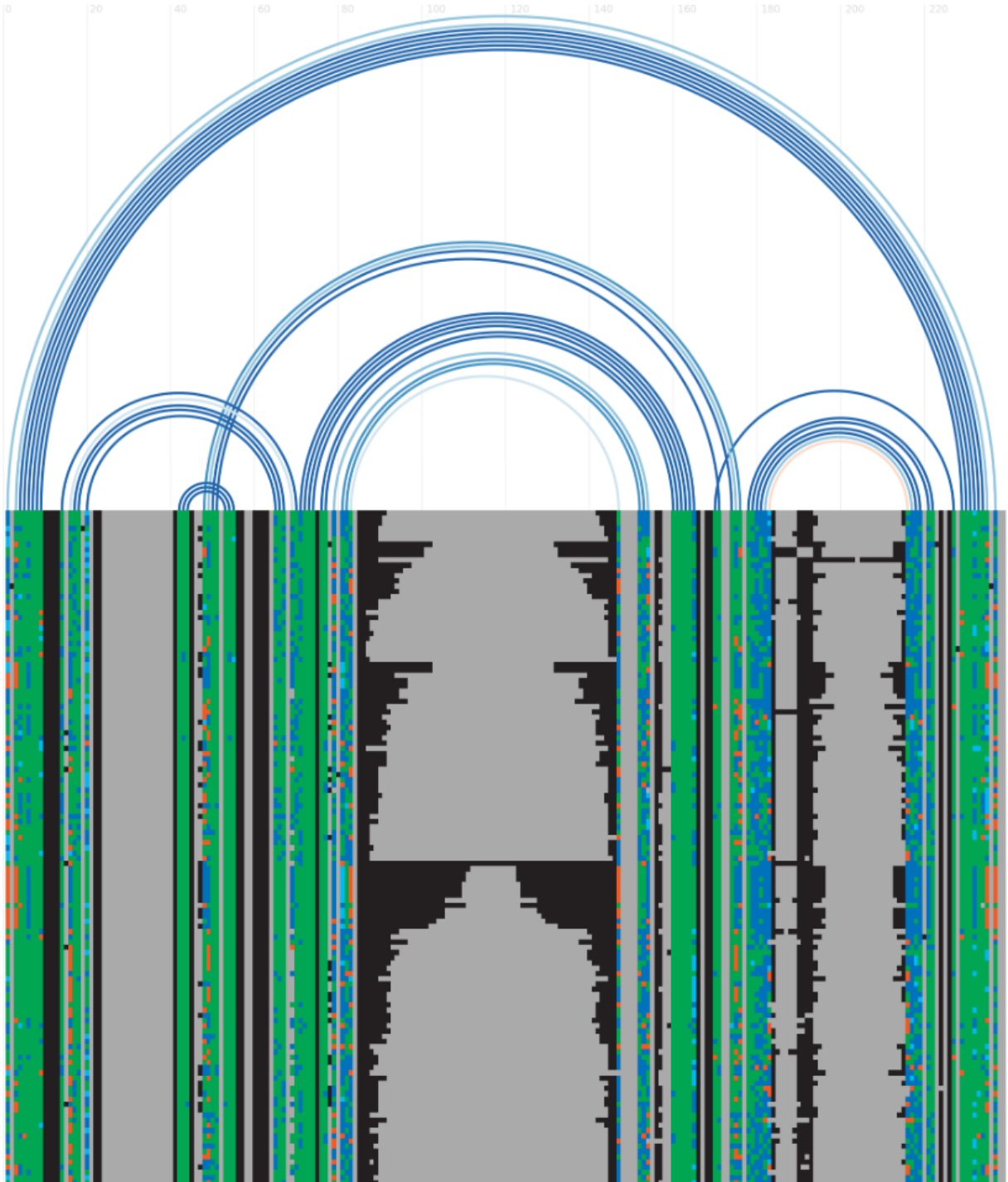


Arc colours

- 100% canonical basepair
- 50%
- 0%

Nucleotide colours

- Valid basepairing
- Two-sided covariation
- One-sided covariation
- Invalid
- Unpaired
- Gap
- Ambiguous



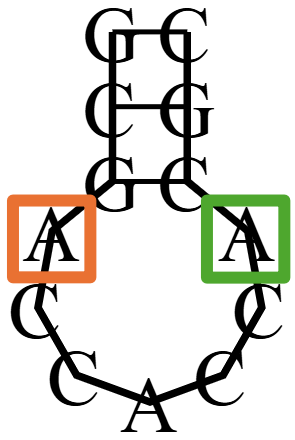
Covariation Family For SAM-Aptamer RF00162

What Are Conserved Disruptive Base Pairs (DBPs)?

- They are pairs of nucleotides that are never allowed to form pairs during evolution

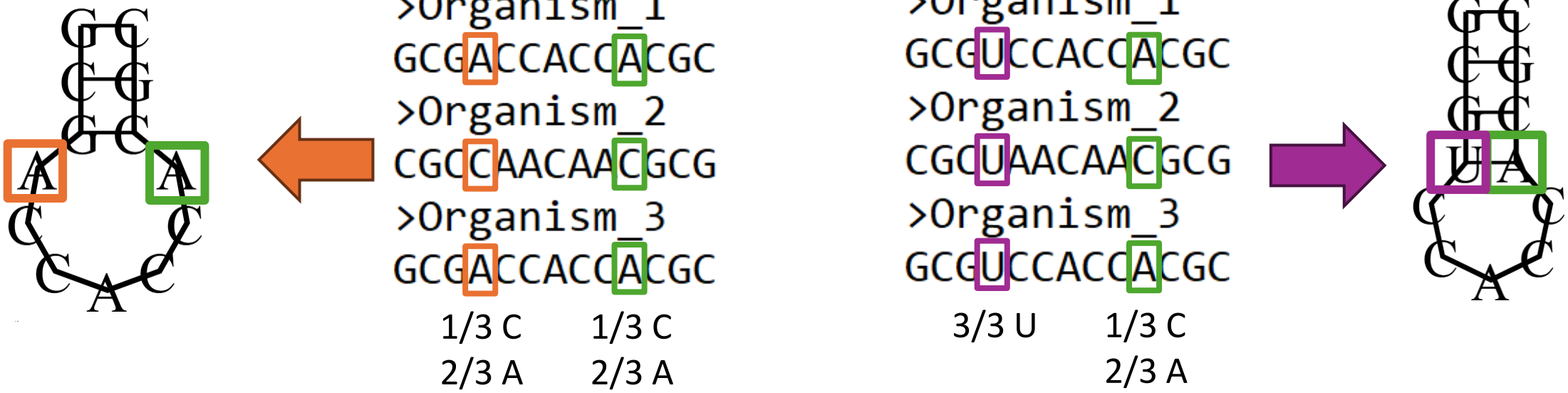
```
>Organism_1
GCGACCACCACGC
>Organism_2
CGCCAACAACGCG
>Organism_3
GCGACCACCACGC
```

1/3 C 1/3 C
2/3 A 2/3 A



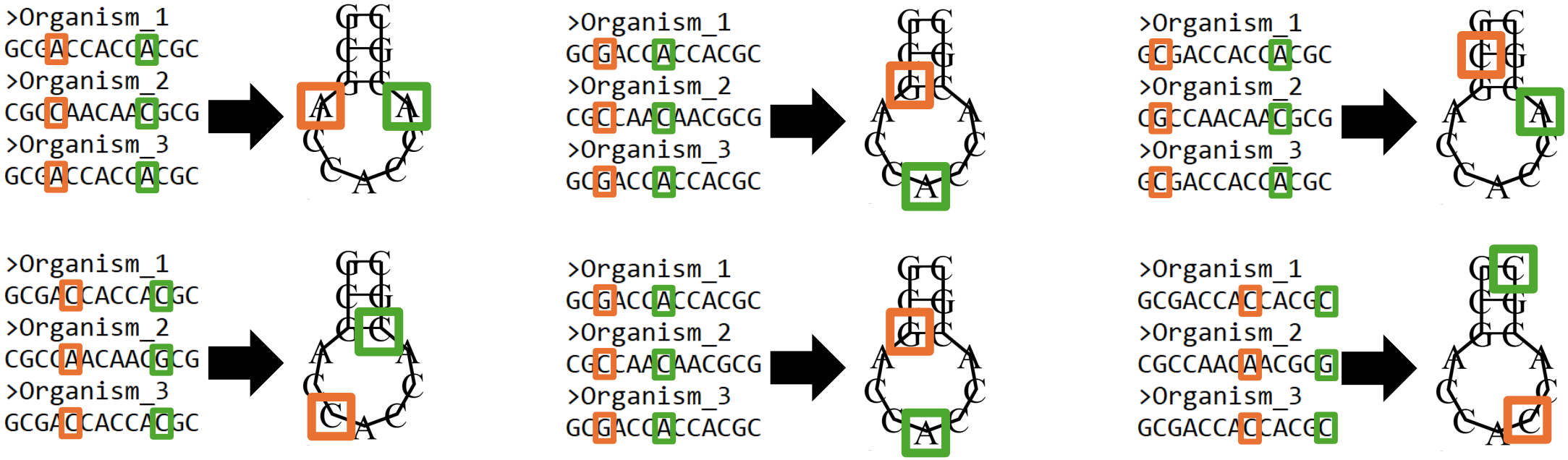
What Are Conserved Disruptive Base Pairs (DBPs)?

- If they were to be changed to form pairs, they can change the secondary structure significantly

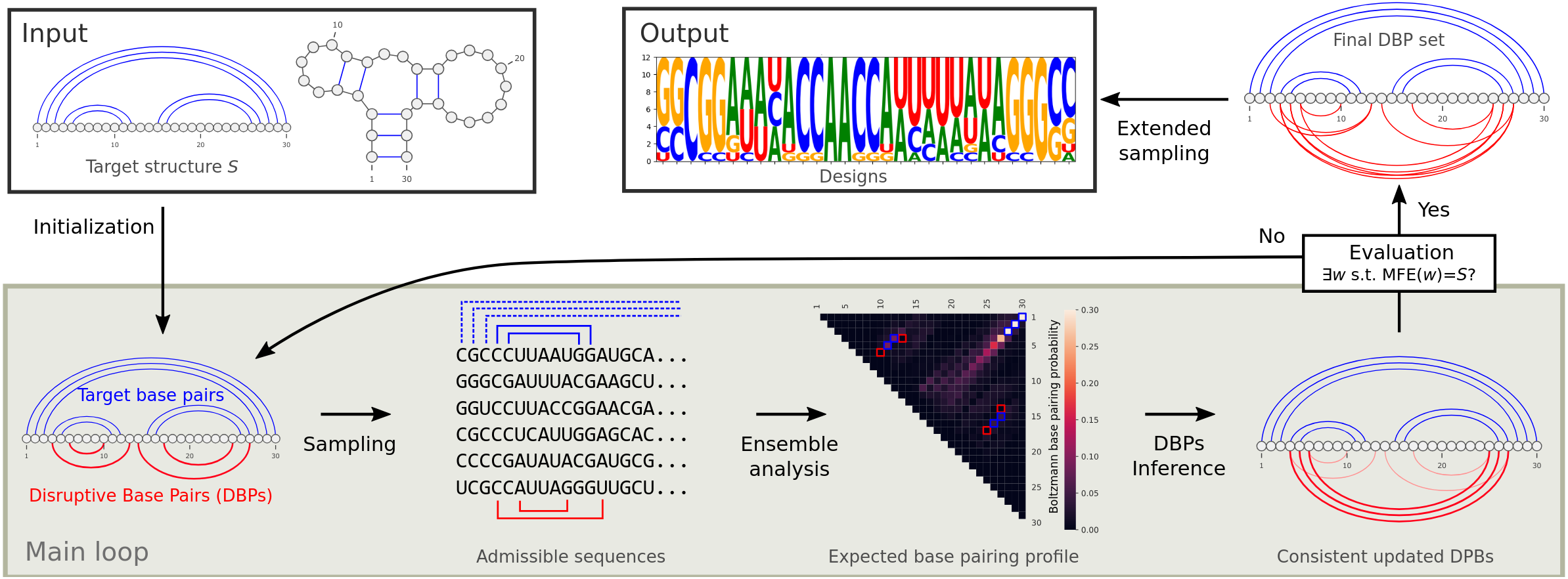


What Are Conserved Disruptive Base Pairs (DBPs)?

- There can be many different interactions between DBPs that contribute to the structure



Disruptive Base Pairs In RNAPOND



Disruptive Base Pair (DBP) Detection Tool

- **Input:** MSA (seed or full alignment)
- **Output:** A list of ranked disruptive base pairs.

- **Method:**
  - Find covarying pairs of columns
    - ➢ **Mutual Information (MI)**

  - Differentiate between disruptive and conserved base pairs
    - ➢ **Log-Scoring function**

Important: This is just one method to find DBPs

Mutual Information

$$H = -\sum_{i=1} p_i * \log_2(p_i)$$

$$\mathrm{I}(X, Y) = H(X) + H(Y) - H(X, Y)$$

Example calculation:

```
>Organism 1
AAAAGAAAACAAAACAAAA
>Organism 2
AAAAUAAAAGAAAAGAAAA
>Organism 3
AAAAGAAAACAAAAUAAAA
>Organism 4
AAAAGAAAACAAAACAAAA
>Organism 5
AAAAUAAAAGAAAAGAAAA
>Organism 6
AAAAGAAAACAAAAUAAAA
```

Marginal Probabilities

Column 1: [4/6 G; 2/6 U]
Column 2: [2/6 G; 4/6 C]
Column 3: [2/6 G; 2/6 C; 2/6 U]

Joint Probabilities

Column 1 x 2: [4/6 G-C; 2/6 G-U]
Column 2 x 3: [2/6 C-C; 2/6 G-G; 2/6 C-U]
Column 1 x 3: [2/6 G-C; 4/6 G-U]

$$\boldsymbol{I(1,2) = H(1) + H(2) - H(1,2) \approx 0.92}$$
$$\boldsymbol{I(2,3) \approx 0.92}$$
$$\boldsymbol{I(1,3) \approx 1.58}$$

Log-Score

$$L(X,Y) = log_2\left(\frac{\sum_{i,j=1}^{WC\ Pairs} xy_{ij}}{\frac{3}{8} * \sum_{ij=1}^{Pairs} xy_{ij}}\right)$$

Column 1 x 2: [6 WC]

$$\mathrm{L}(1,2) = \log_2\left(\frac{6}{\frac{3}{8}*6}\right) \approx 1.42$$

Column 2 x 3: [0 WC, 6 NWC]*

$$\mathrm{L}(2,\ 3) = \log_2\left(\frac{0.9}{\frac{3}{8}*6}\right) \approx -1.32$$

Column 1 x 3: [6 WC]

$$\mathrm{L}(1\ ,\ 3) = \log_2\left(\frac{6}{\frac{3}{8}*6}\right) \approx 1.42$$

Overall Score:

$$Score = I(X,Y) * L(X,Y)$$

$$\boldsymbol{L(1,2) * I(1,2) \approx 1.31}$$

$$\boldsymbol{L(2,\ 3)\ *\ I(2,3) \approx -1,21}$$

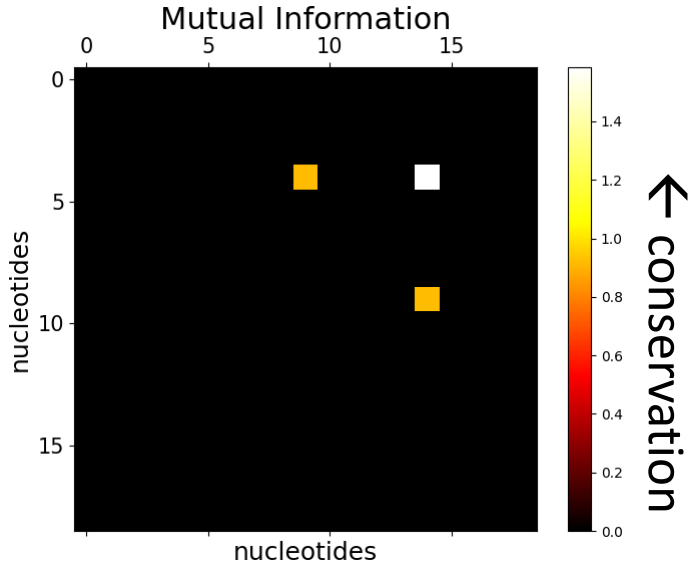
$$\boldsymbol{L(1\ ,\ 3) * I(1,3) \approx 2,24}$$

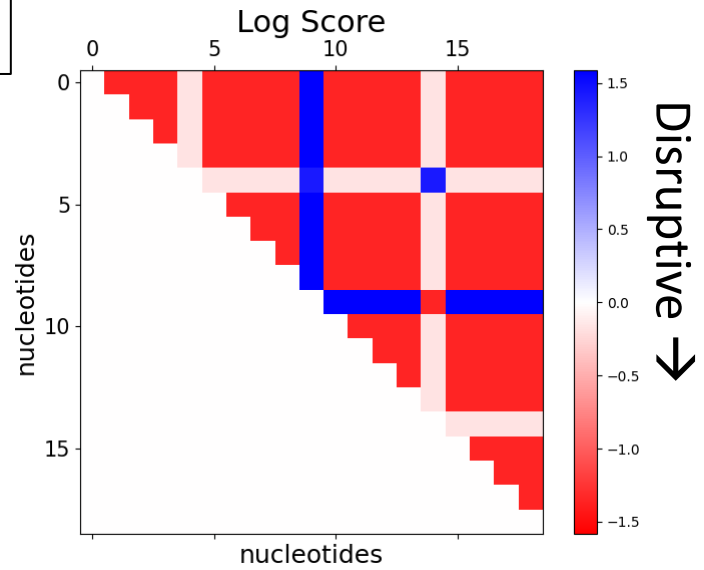
*We add a small pseudocount of 0.9 (We don't do this in production))

Example: Test-Set

```
>Organism 1
AAAAGAAAACAAAACAAAA
>Organism 2
AAAAUAAAAGAAAAGAAAA
>Organism 3
AAAAGAAAACAAAAUAAAA
>Organism 4
AAAAGAAAACAAAACAAAA
>Organism 5
AAAAUAAAAGAAAAGAAAA
>Organism 6
AAAAGAAAACAAAAUAAAA
```

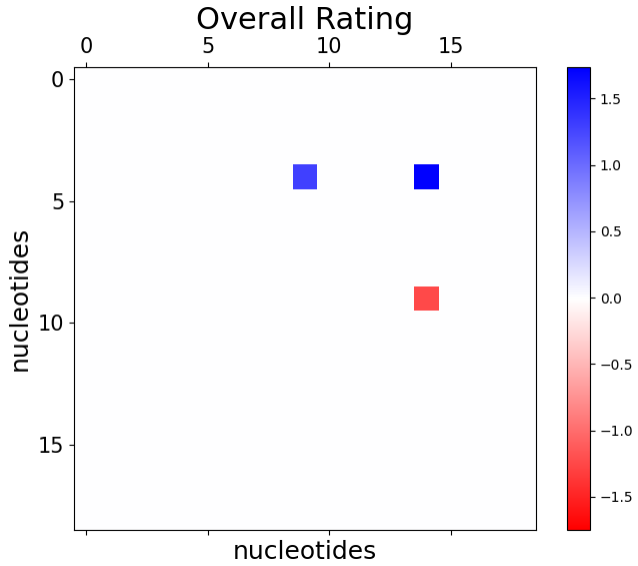
Dataframe

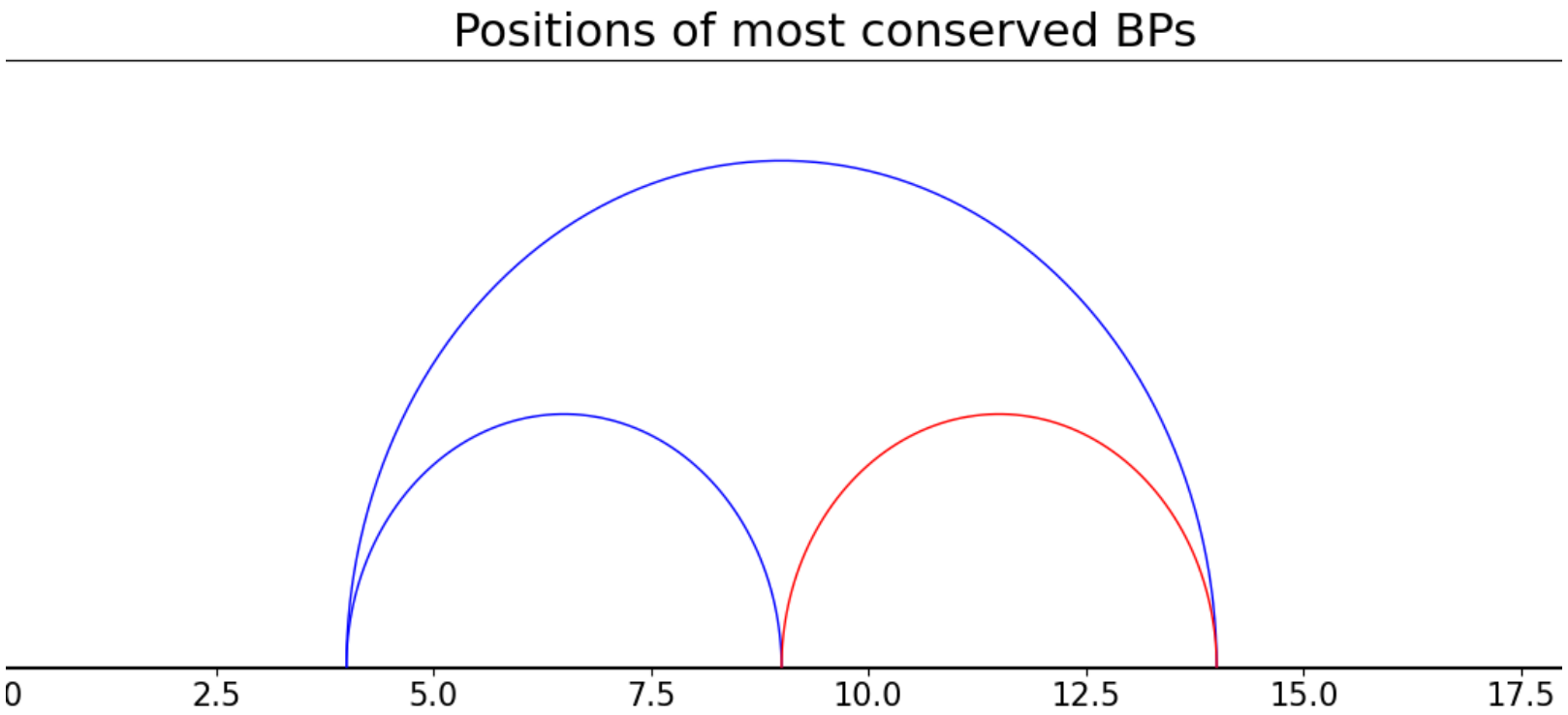




← conservation

Disruptive →



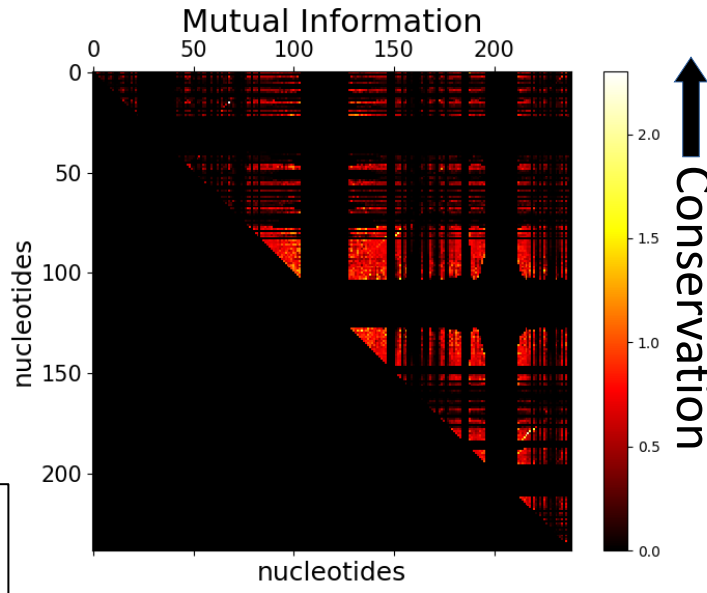



```
>Organism 1
AAAAGAAAACAAAACAAAA
>Organism 2
AAAAUAAAAGAAAAGAAAA
>Organism 3
AAAAGAAAACAAAAUAAAA
>Organism 4
AAAAGAAAACAAAACAAAA
>Organism 5
AAAAUAAAAGAAAAGAAAA
>Organism 6
AAAAGAAAACAAAAUAAAA
```

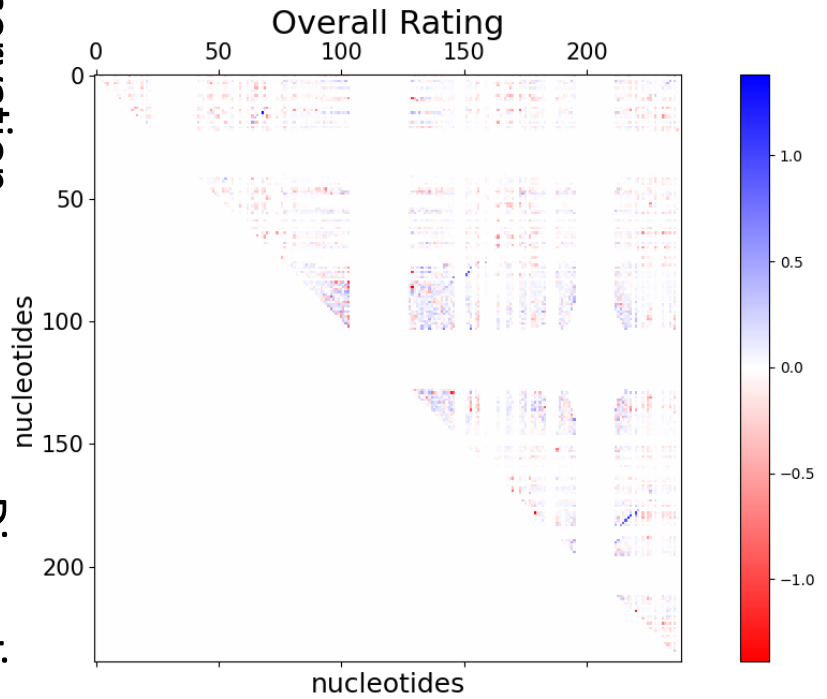
Example: SAM-Riboswitch Aptamer RF00162

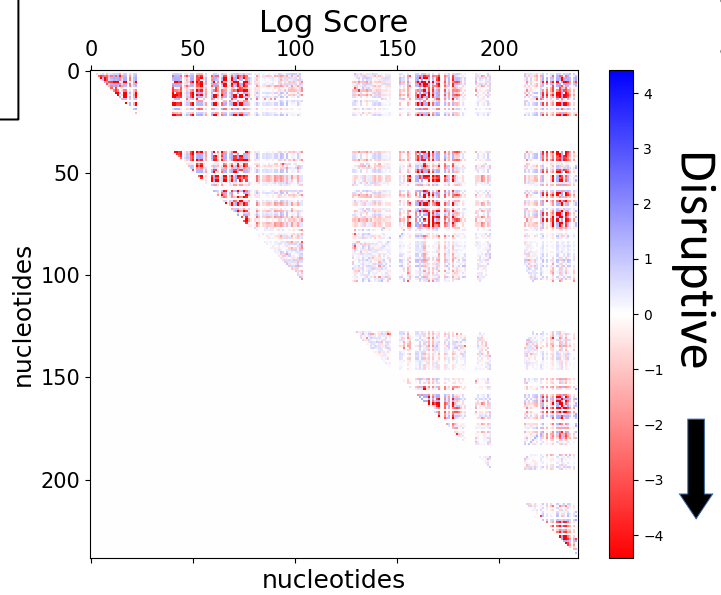
```
>AF027868.1/5245-5154
U-UUCUAUCCAGAG-AGG-U-GG-----------------
GAA-ACC-UCGGCAACA-------UU--------------
------------------------AU------------U
GC-------GCUA-----------------------GCU-
>AF269983.1/571-671
A-CCUUAUUUUGAG-AAG-C-UG-----------------
GAA-GCU-UCAGCAACC-G-ACU-UUA-------------
-----------------------AAU----AGC-AC--G
G---------CAA--------------------------CU-
>AJ000974.1/281-386
A-ACUUAUCAAGAG-CGG-C-UG-----------------
GAA-GCC--CGGCAACC-U-GCA-UAG-------------
----------------------UUU----GUA-A---G
AAUG-----AAUUC--------------------CAUUU-
>AL596164.1/95936-95817
C-UCUUAUCGAGAG-CGG-C-AG-----------------
GAA-GCC--CGGCAACC-U-AAC-UUUAUUUAA-------
```

Dataframe



Conservation

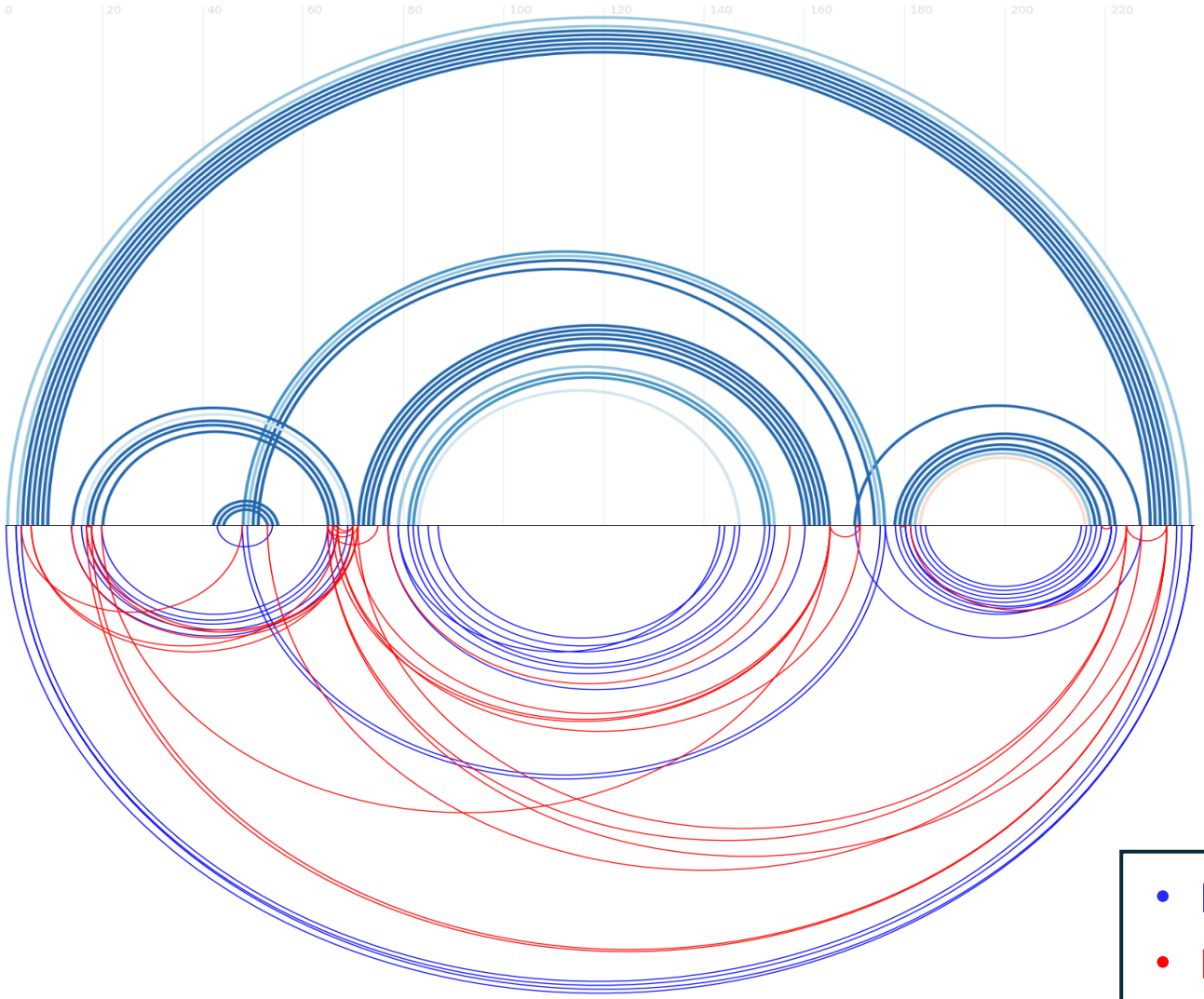




Disruptive

Example: SAM-Riboswitch Aptamer RF00162

Conserved covariance family structure

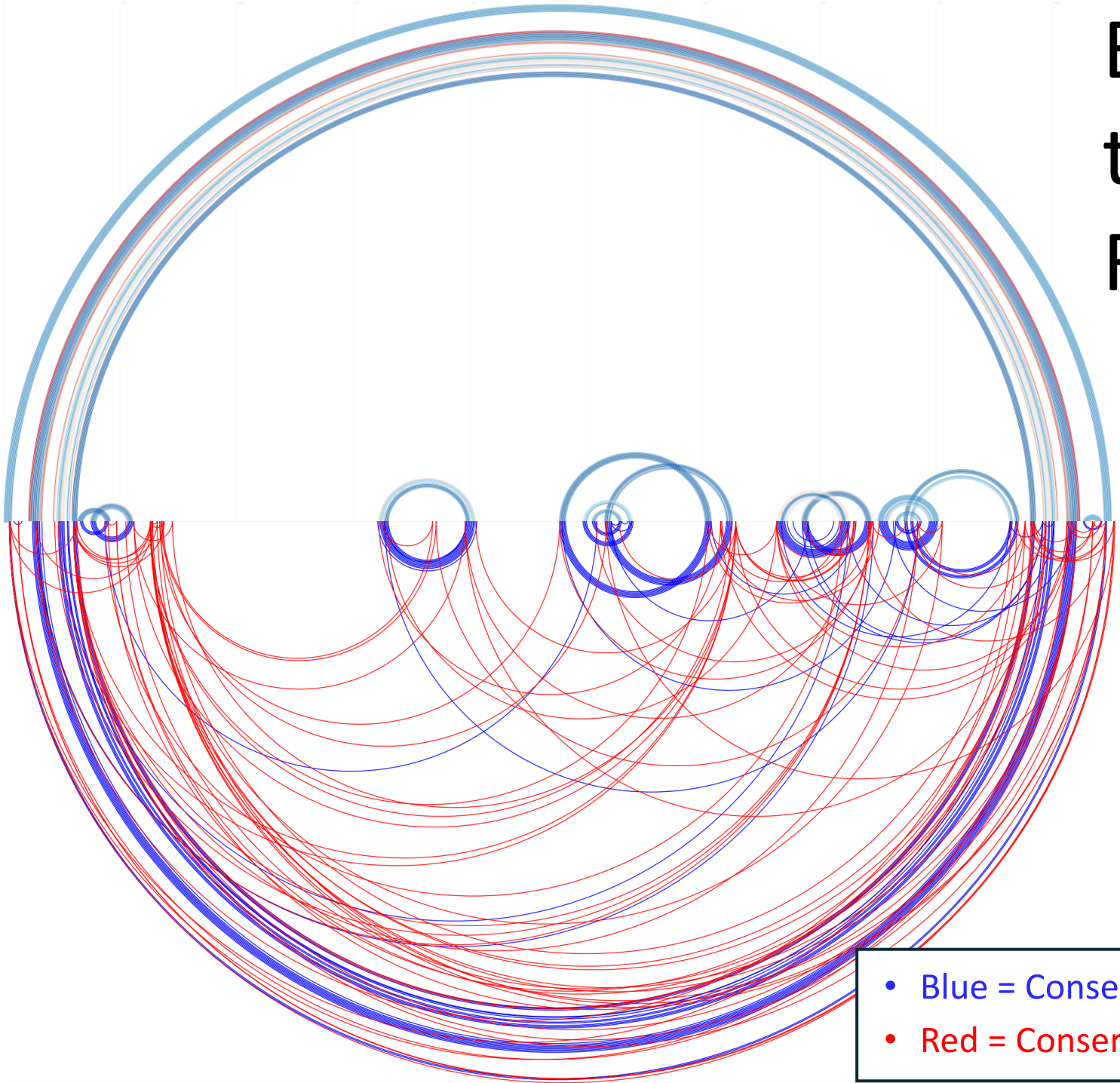


Our covariance family structure

- Blue = Conserved
- Red = Conserved and disruptive

Example: tmRNA RF00023

Conserved covariance family structure

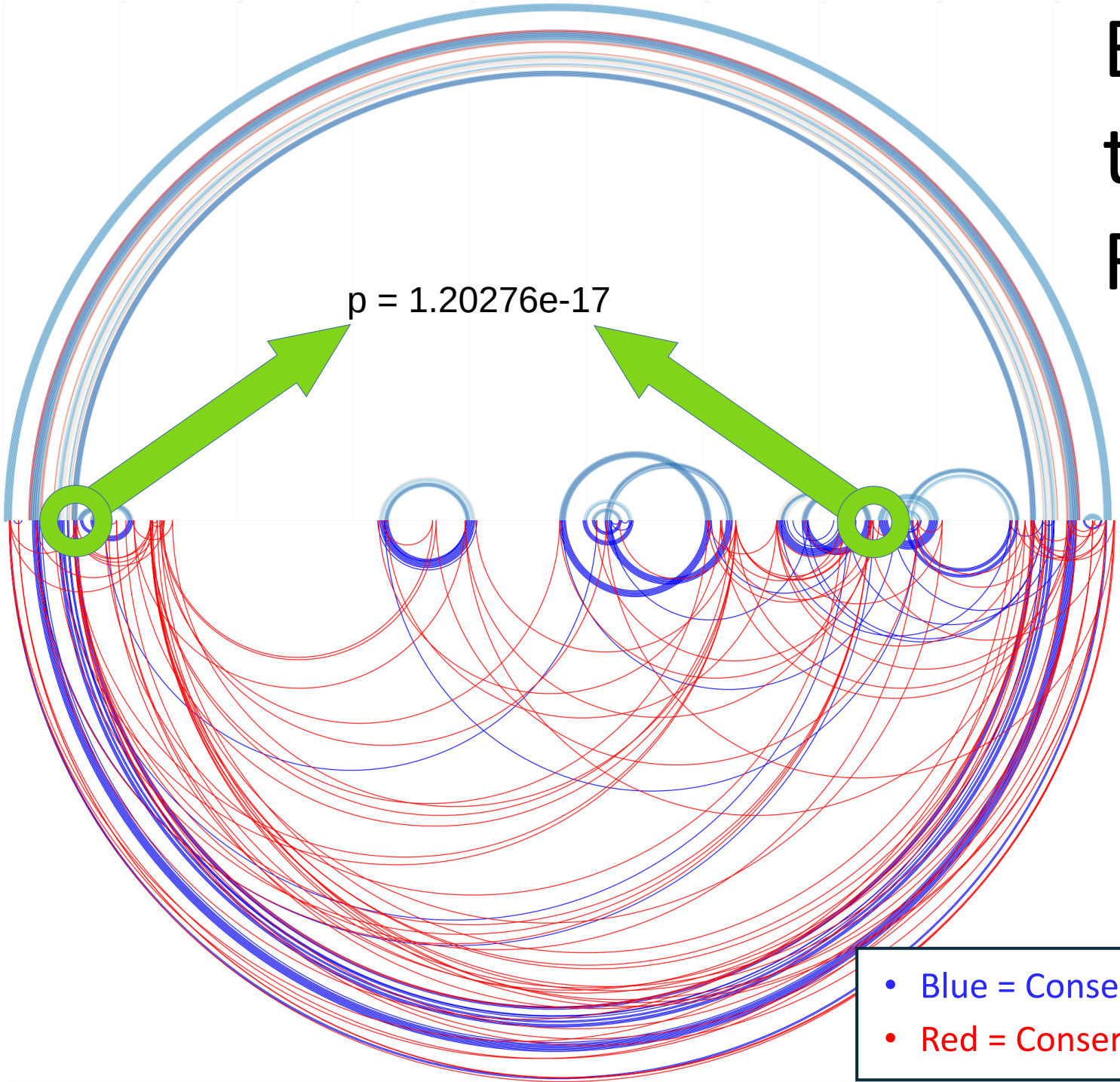


Our covariance family structure

- Blue = Conserved
- Red = Conserved and disruptive

Example: tmRNA RF00023

Conserved covariance family structure

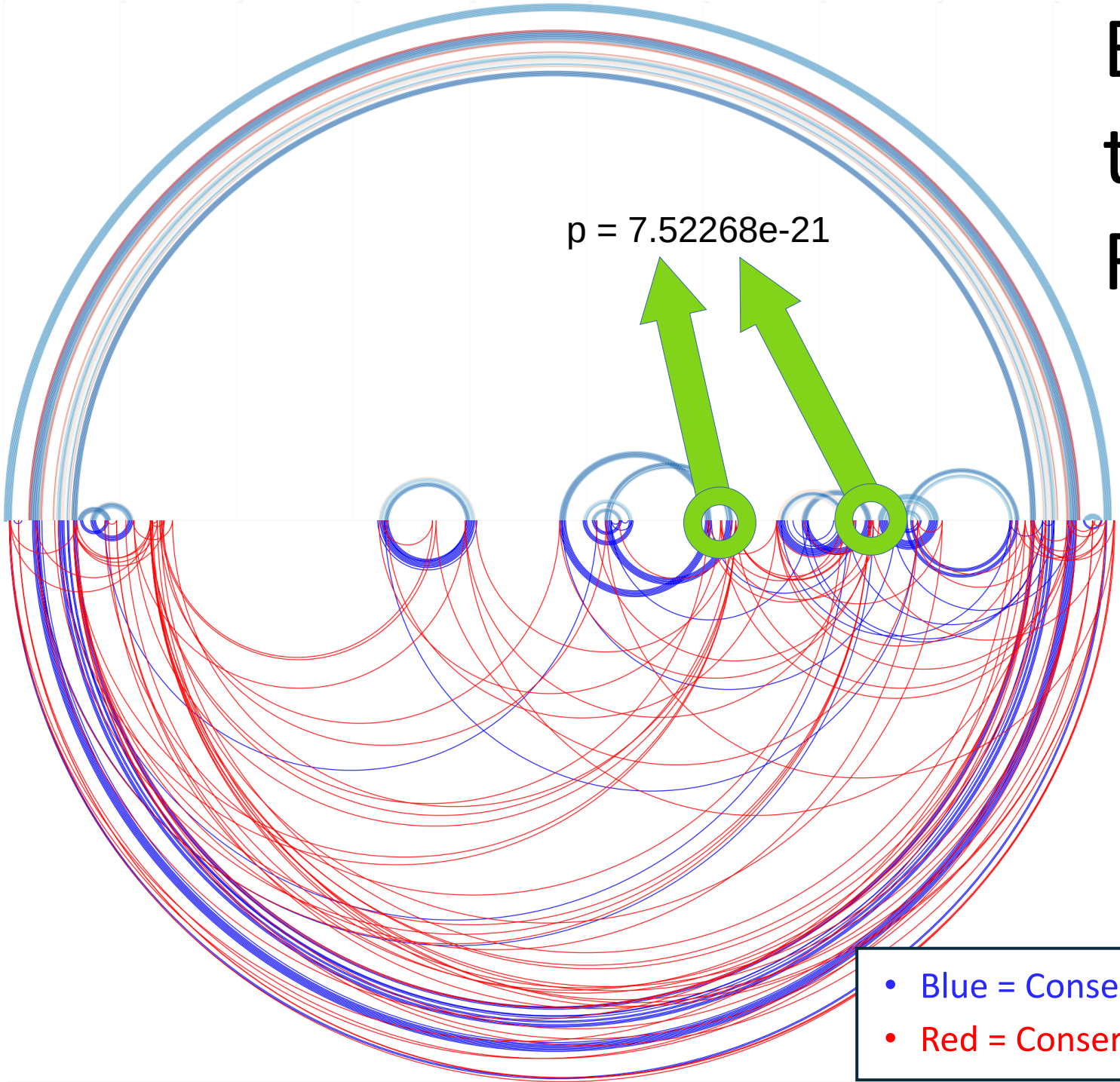


Our covariance family structure

- Blue = Conserved
- Red = Conserved and disruptive

Example: tmRNA RF00023

Conserved covariance family structure



Our covariance family structure

- Blue = Conserved
- Red = Conserved and disruptive

Perspective

- Integration of structural inhibitiory constraints from a nature-learned perspective in RNA design
- Contributions to RNA stability
- Deeper understanding of selective pressure in RNA evolution

Thank you for your Attention!

Questions?