Modified Nucleotides in RNA structure prediction

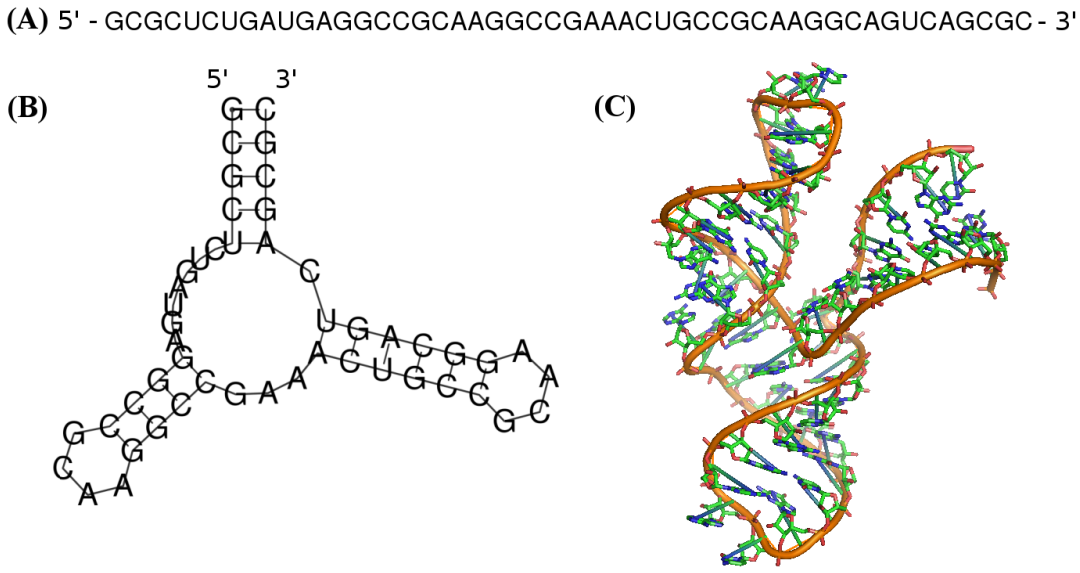
Ronny Lorenz, Thomas Spicher, Ivo L. Hofacker
ronny@tbi.univie.ac.at

University of Vienna, Theoretical Biochemistry Group (TBI)

Mondsee, Austria, September 8th 2020

RNA Secondary Structures

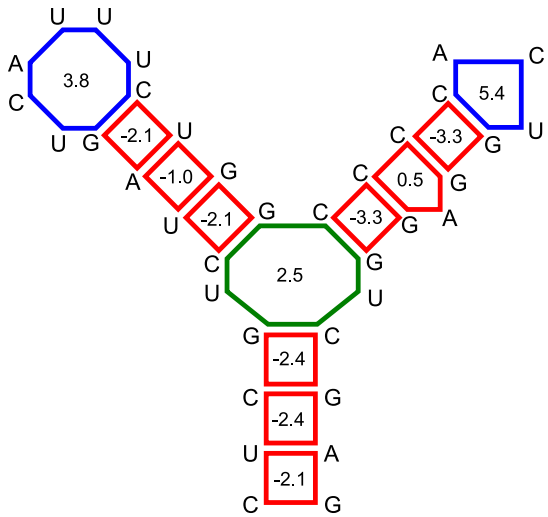
RNAs fold hierarchical (A) $\rightarrow$ (B) $\rightarrow$ (C)

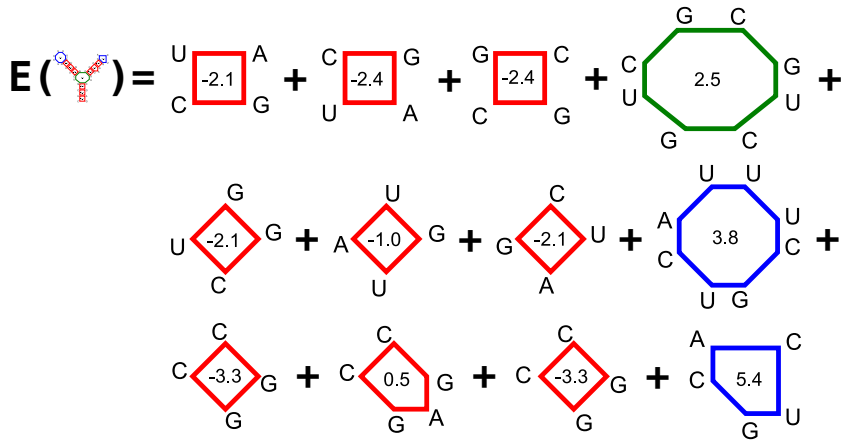


Secondary Structure (B):

- Set of nested base pairs
- Captures majority of stabilizing interactions
- Many thermodynamic properties can be predicted efficiently
- Very good prediction accuracy for small RNAs
- Starting point for 3D structure modeling
- Good abstraction to perform folding kinetics simulation

RNA Secondary Structures - Nearest Neighbor Energy Model





- Secondary structures s can be uniquely decomposed into loops L
- Contributions of a base pair only depends on neighboring pairs
- Each loop L is assigned a free energy contribution E_L[1]

$$E(s) \approx \sum_{L \in s} E_L$$

[1] Turner et al., “NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure.”, 2009, NAR 38, D280–D282

RNA Secondary Structures - Statistical Thermodynamics

$$
\begin{aligned}
p(F) &\propto e^{-\beta E(F)}, \quad \text{with} \quad \beta = \frac{1}{RT} \\
&= \frac{1}{Q} \sum_{s|F \in s} e^{-\beta E(s)}, \quad \text{with} \quad Q = \sum_{s} e^{-\beta E(s)}
\end{aligned}
$$

Efficient dynamic programming algorithms are available to compute

- Minimum free energy $MFE = \min_s E(s)$
- Partition function Q
- Probabilities p_{kl} for base pairs (k,l), i.e. $p_{kl} = p(F)$ with $F = (k,l)$
- Suboptimal and locally stable structures, ligand binding, ...

The `ViennaRNA Package`[2] provides fast implementations for

- single sequences
- (multiple) interacting sequences
- sequence alignments (consensus structure)

[2] Lorenz et al., "ViennaRNA Package 2.0.", 2011, ALMOB 6.1, 26

RNA Secondary Structures - Statistical Thermodynamics

$$
\begin{aligned}
p(F) &\propto e^{-\beta E(F)}, \quad \text{with} \quad \beta = \frac{1}{RT} \\
&= \frac{1}{Q} \sum_{s|F \in s} e^{-\beta E(s)}, \quad \text{with} \quad Q = \sum_{s} e^{-\beta E(s)}
\end{aligned}
$$

Efficient dynamic programming algorithms are available to compute

- Minimum free energy $MFE = \min_s E(s)$
- Partition function Q
- Probabilities p_{kl} for base pairs (k,l), i.e. $p_{kl} = p(F)$ with $F = (k,l)$
- Suboptimal and locally stable structures, ligand binding, ...

The `ViennaRNA Package`[2] provides fast implementations for

- single sequences
- (multiple) interacting sequences
- sequence alignments (consensus structure)

Implementations are restricted to unmodified bases!

[2] Lorenz et al., "ViennaRNA Package 2.0.", 2011, ALMOB 6.1, 26

RNA Secondary Structure - Modified Bases

Many RNAs require modified bases to fulfill their function!

- Modifications may change pairing partner preference
- Modifications may (de-)stabilize loop formation
- `Modomics` Database[3] lists 172 different modified bases
- Only a few energy parameters involving modified bases available, e.g. stacking with $\Psi - A$[4] and $I - U$[5] pairs

How to incorporate modified bases in predictions?

- Use constraints
  1. Mask modified bases, e.g. to stay unpaired using letter “N”
  2. Write wrapper (Python/Perl/C/C++) using the *constraints framework* of the `ViennaRNA Package`[6] to supplement energy parameters
- Re-implement algorithms with
  1. enhanced nucleotide alphabet
  2. additional pairing rules
  3. more energy parameters

---

[3] Boccaletto et al., “MODOMICS: a database of RNA modification pathways. 2017 update.”, 2018, NAR 46.D1, D303–D307

[4] Hudson et al., “Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides.”, 2013, RNA 19.11, 1474–1482

[5] Wright et al., “Nearest Neighbor Parameters for Inosine-Uridine Pairs in RNA Duplexes.”, 2007, Biochemistry 46, 4625–4634

[6] Lorenz et al., “RNA folding with hard and soft constraints.”, 2016, ALMOB 11, 8

RNA Secondary Structure - Modified Bases

Many RNAs require modified bases to fulfill their function!

- Modifications may change pairing partner preference
- Modifications may (de-)stabilize loop formation
- `Modomics` Database[3] lists 172 different modified bases
- Only a few energy parameters involving modified bases available, e.g. stacking with $\Psi - A$[4] and $I - U$[5] pairs

How to incorporate modified bases in predictions?

- **Use constraints**
  1. Mask modified bases, e.g. to stay unpaired using letter “N”
  2. Write wrapper (Python/Perl/C/C++) using the *constraints framework* of the `ViennaRNA Package`[6] to supplement energy parameters
- Re-implement algorithms with
  1. enhanced nucleotide alphabet
  2. additional pairing rules
  3. more energy parameters

---

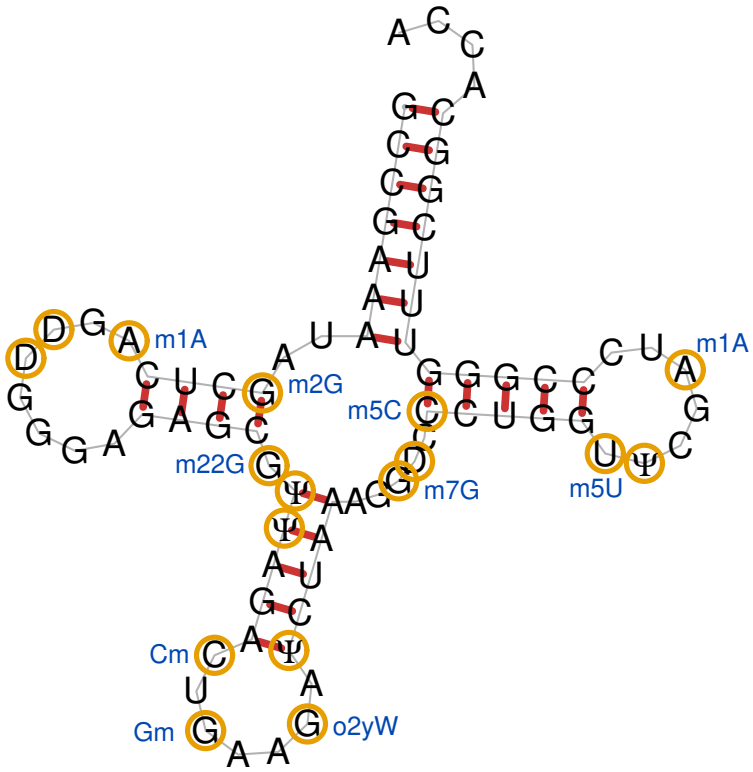
[3] Boccaletto et al., “MODOMICS: a database of RNA modification pathways. 2017 update.”, 2018, NAR 46.D1, D303–D307

[4] Hudson et al., “Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides.”, 2013, RNA 19.11, 1474–1482

[5] Wright et al., “Nearest Neighbor Parameters for Inosine-Uridine Pairs in RNA Duplexes.”, 2007, Biochemistry 46, 4625–4634

[6] Lorenz et al., “RNA folding with hard and soft constraints.”, 2016, ALMOB 11, 8

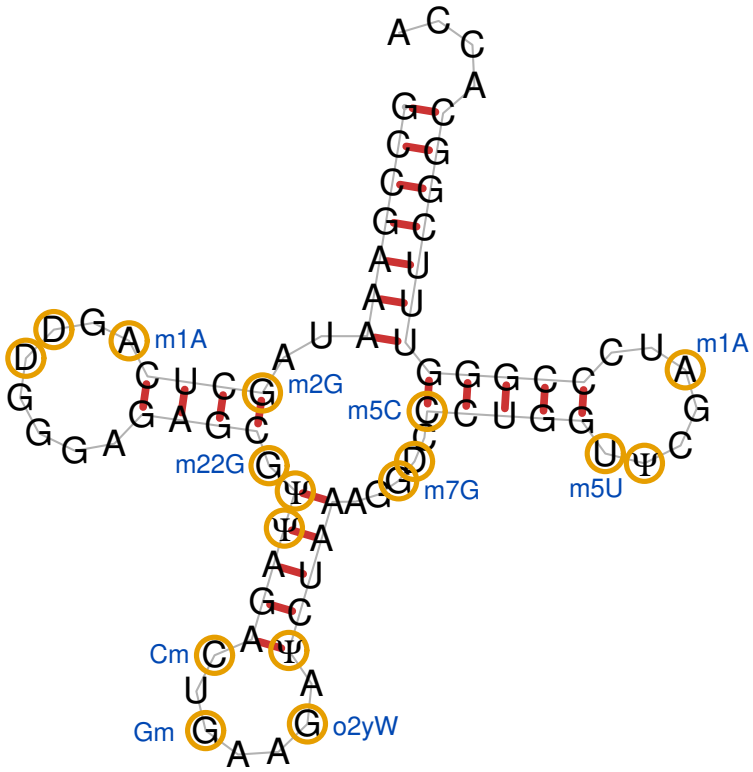
Example: human tRNA Phe (tdbR00000103[7])



- 17 out of 76 nucleotides are modified
- Predicted ground state (MFE) is **not** cloverleaf structure

[7] Jühling et al., "tRNAdb 2009: compilation of tRNA sequences and tRNA genes.", 2009, NAR 37, D159–D162

Example: human tRNA Phe (tdbR00000103[7])



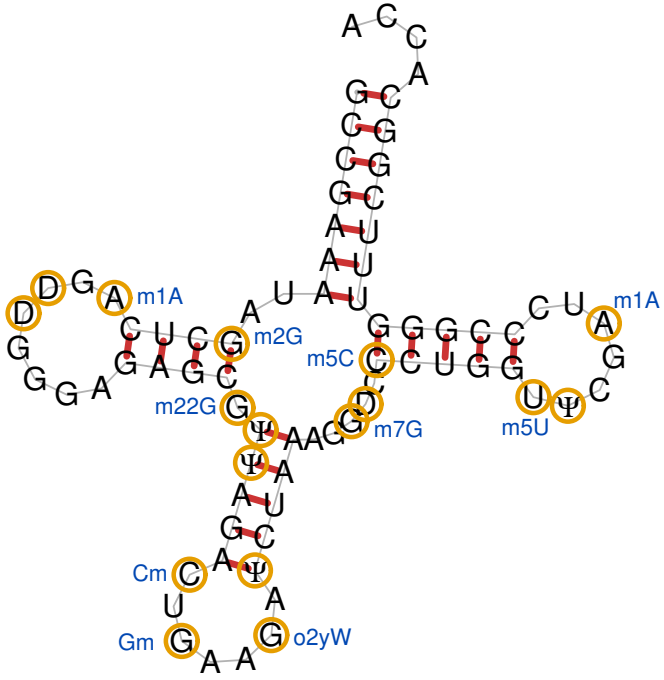
- 17 out of 76 nucleotides are modified
- Predicted ground state (MFE) is **not** cloverleaf structure

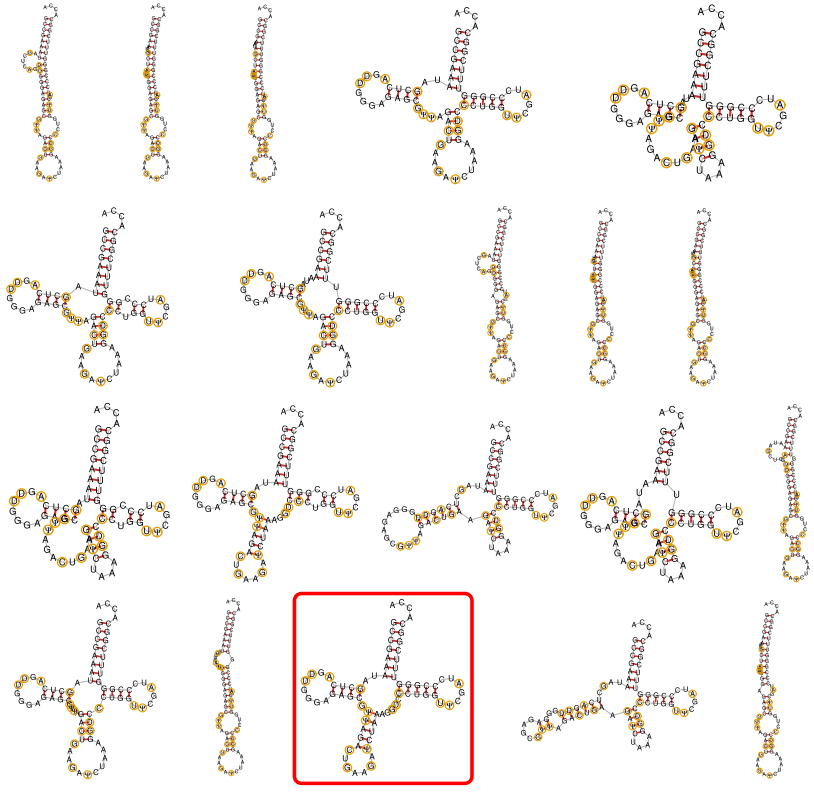
What about suboptimal structures $S^{\delta} = \{s \in \Omega \mid E(s) \leq E_{\text{MFE}} + \delta\}$?

[7] Jühling et al., "tRNAdb 2009: compilation of tRNA sequences and tRNA genes.", 2009, NAR 37, D159–D162

Example: human tRNA Phe (tdbR00000103)

1. Suboptimals for *unmodified sequence*

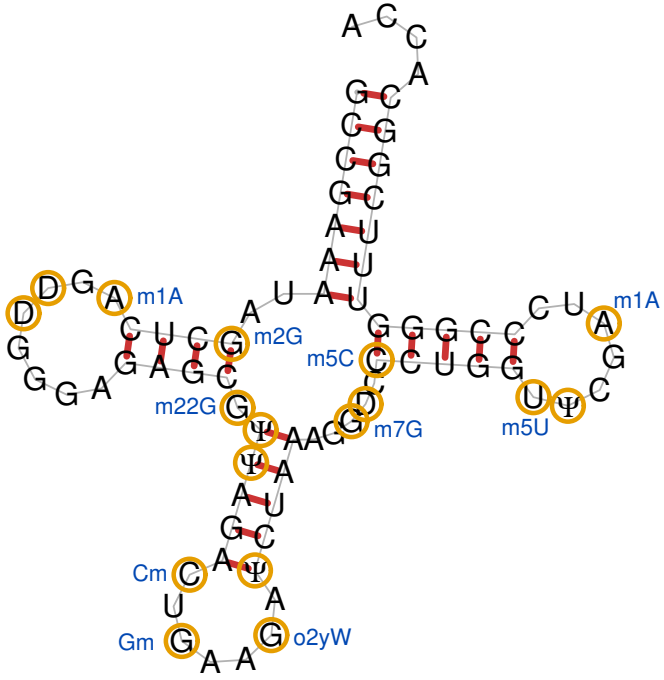


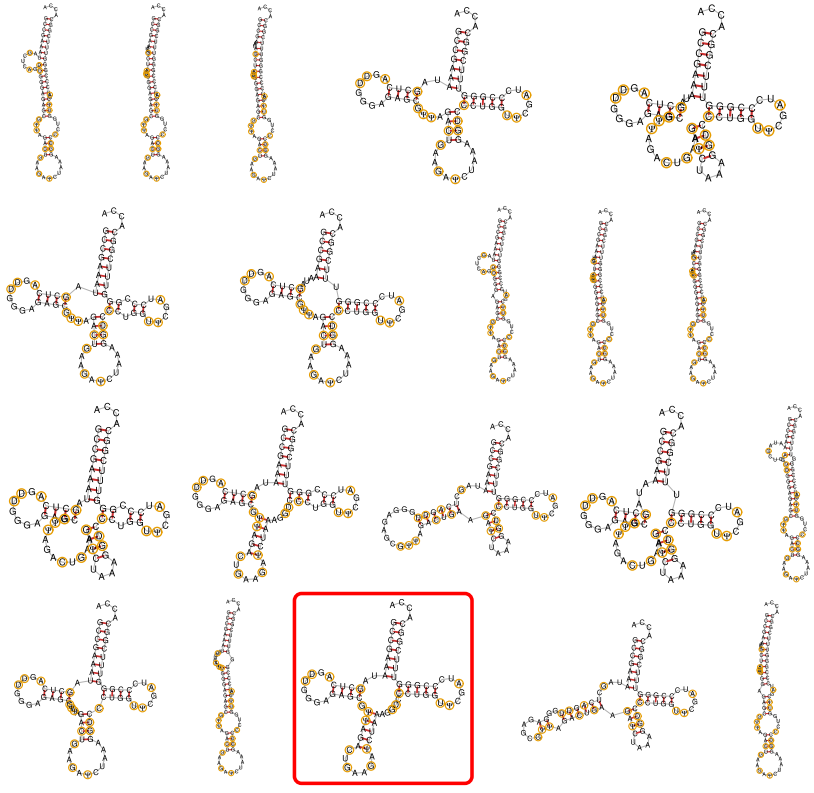


- Correct cloverleaf structure at position 18
- $\bar{d}^{3}_{\mathrm{BP}} = \frac{1}{|S^3|}\sum_{s\in S^3} d_{\mathrm{BP}}(s_{\mathrm{ref}}, s) = 21.55$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 1.7$ kcal/mol
- Most predictions are not cloverleaf structure

Example: human tRNA Phe (tdbR00000103)

1. Suboptimals for *unmodified sequence*



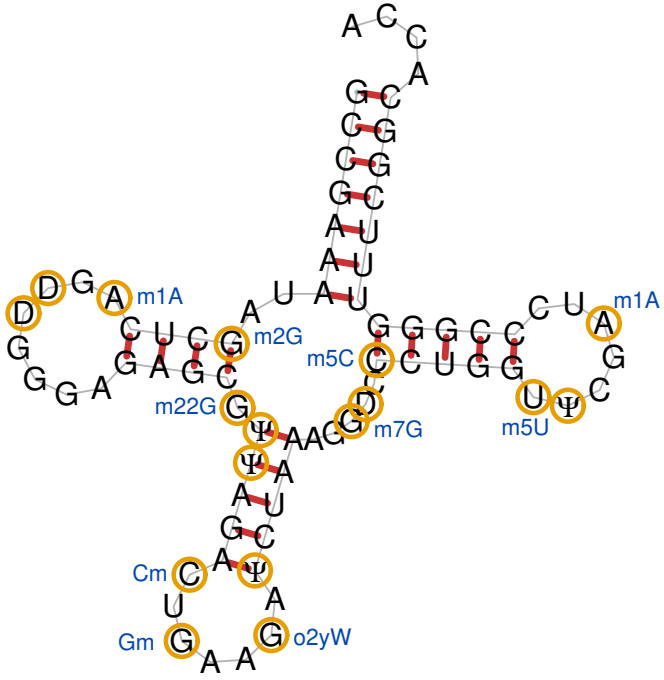


- Correct cloverleaf structure at position 18
- $\bar{d}^{3}_{\mathrm{BP}} = \frac{1}{|S^3|}\sum_{s\in S^3} d_{\mathrm{BP}}(s_{\mathrm{ref}}, s) = 21.55$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 1.7$ kcal/mol
- Most predictions are not cloverleaf structure

How about masking modified bases?

Example: human tRNA Phe (tdbR00000103)

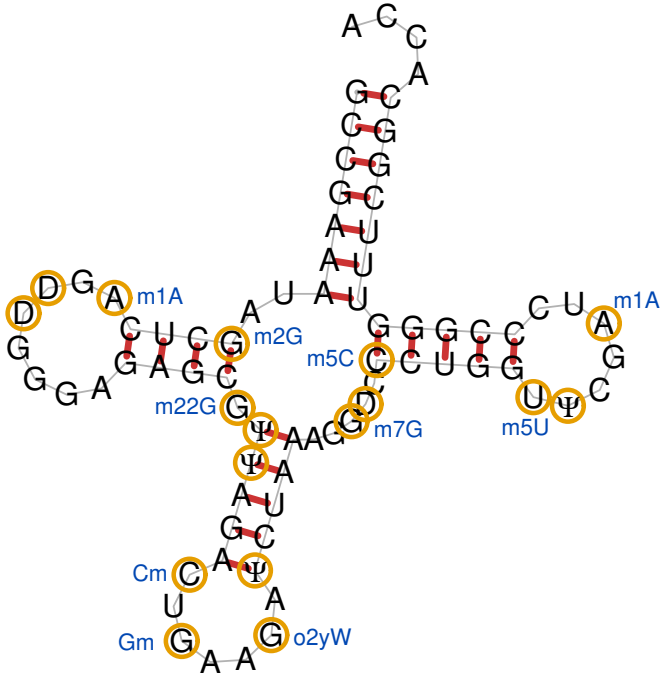
2. Suboptimals for sequence with *all modified bases masked*



- Modified bases are excluded from pairing
- Correct cloverleaf structure not among the solution set
- $\bar{d}^{3}_{\mathrm{BP}} = 16.7$ bp, anticodon arm is missing

Example: human tRNA Phe (tdbR00000103)

2. Suboptimals for sequence with *all modified bases masked*

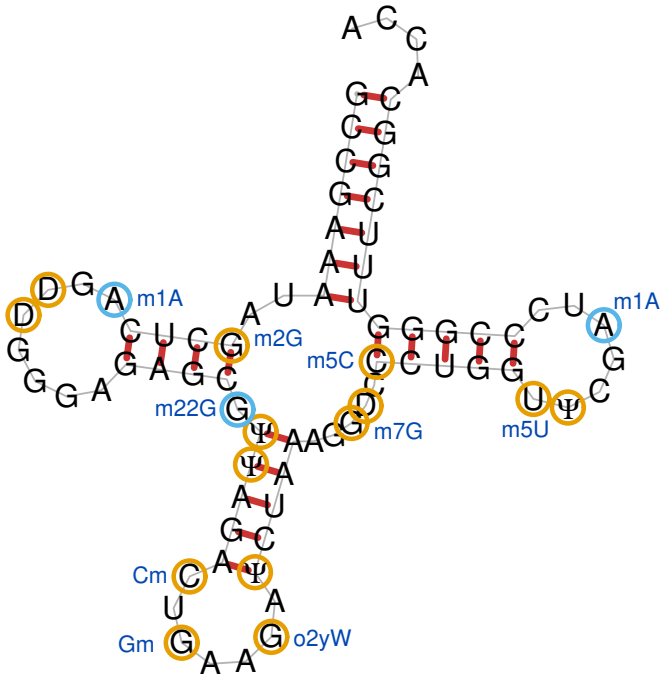


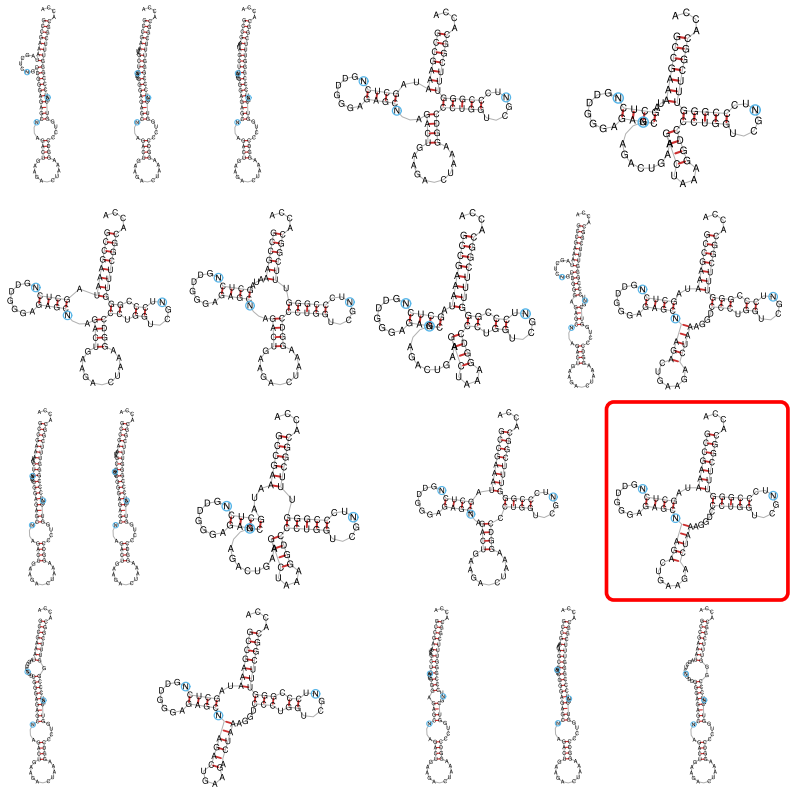
- Modified bases are excluded from pairing
- Correct cloverleaf structure not among the solution set
- $\bar{d}^3_{\mathrm{BP}} = 16.7$ bp, anticodon arm is missing

Some modified bases are required to pair!

Example: human tRNA Phe (tdbR00000103)

3. Suboptimals with *masked bases* for RT blocking modifications[8]



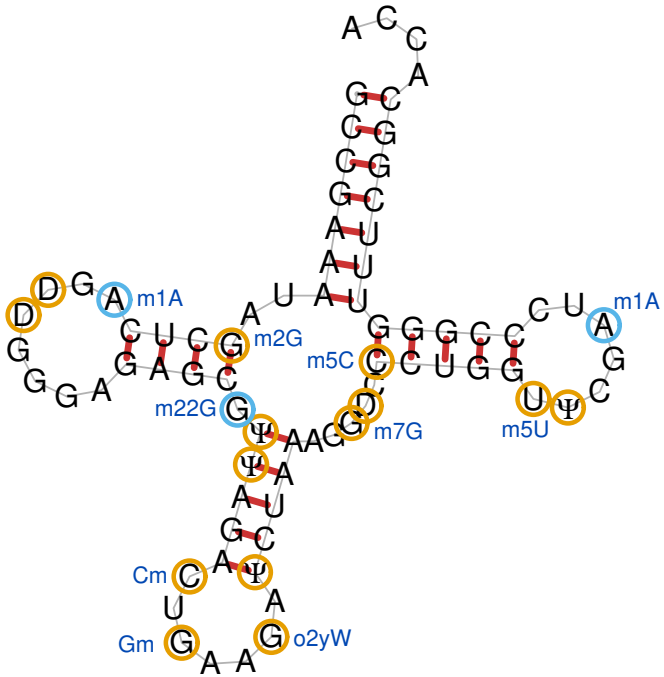


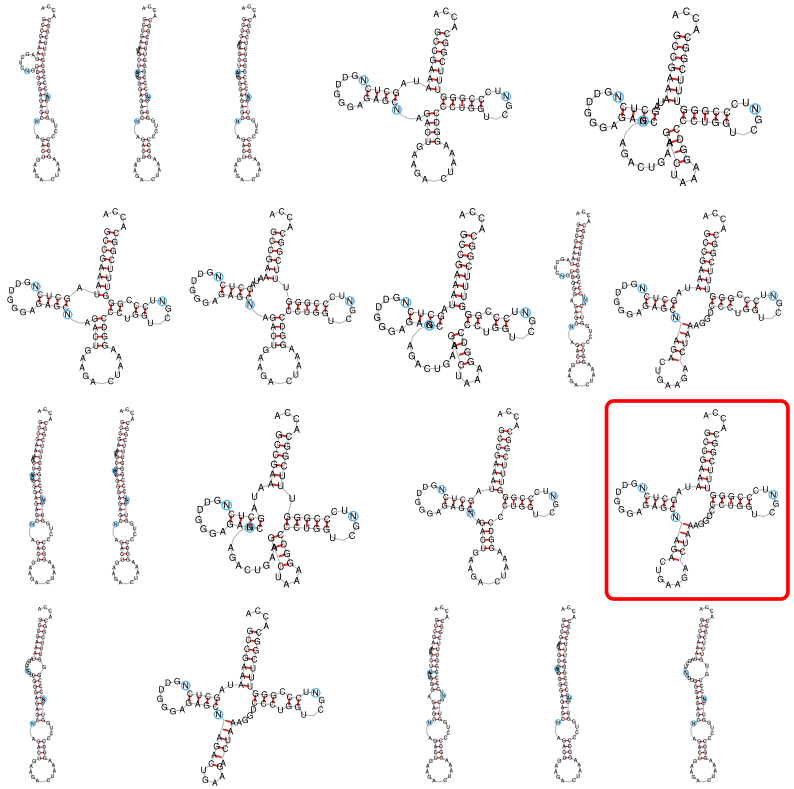
- Correct cloverleaf structure at position 15
- $\bar{d}^{3}_{\mathrm{BP}} = 19.7$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 1.6$ kcal/mol
- Still, most predicted structure are not tRNA-like!

[8] Motorin et al., "Identification of modified residues in RNAs by reverse transcription-based methods.", 2007, Methods in Enzymology 425, 21–53

Example: human tRNA Phe (tdbR00000103)

3. Suboptimals with *masked bases* for RT blocking modifications[8]





- Correct cloverleaf structure at position 15
- $\bar{d}^3_{\mathrm{BP}} = 19.7$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 1.6$ kcal/mol
- Still, most predicted structure are not tRNA-like!

Ψ-A stacking energies are available in literature!

[8] Motorin et al., "Identification of modified residues in RNAs by reverse transcription-based methods.", 2007, Methods in Enzymology 425, 21–53

Example: human tRNA Phe (tdbR00000103)

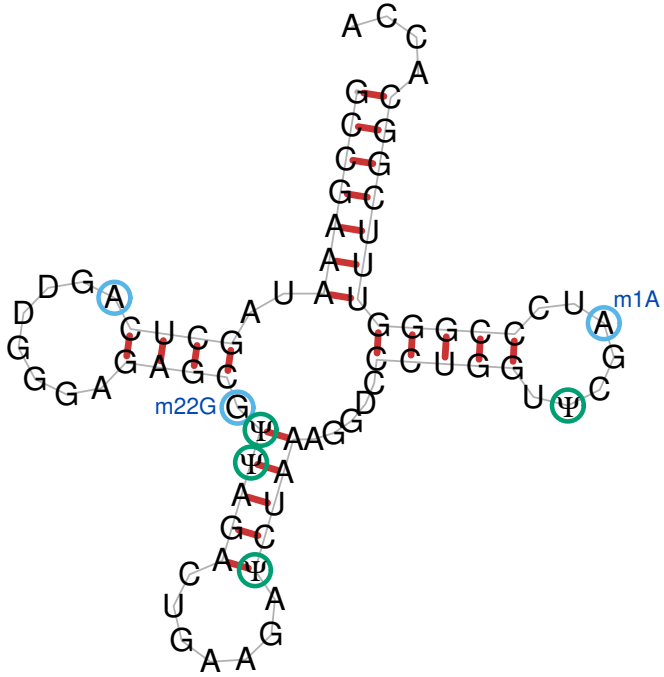
Python overhead to add energies for stacks with Ψ-A[9]:

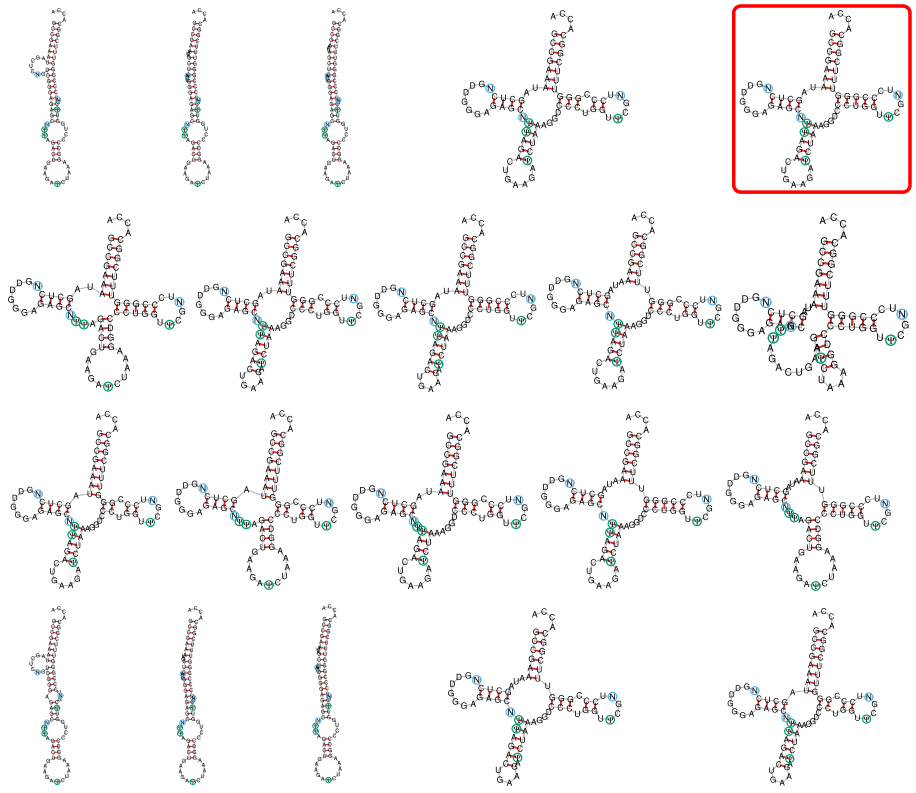
```
# Energy parameters from Hudson et al., 2013
stacking_psi_A = [
  # N   A     C   G     U     P
  [ 0,  0,    0,  0,    0,    0 ], # N
  [ 0,  0,    0,  0,    -77,  0 ], # A
  [ 0,  0,    0,  -14,  0,    0 ], # C
  [ 0,  0,    -9, 0,    0,    0 ], # G
  [ 0,  -181, 0,  0,    0,    0 ], # U
  [ 0,  0,    0,  0,    0,    0 ]  # P
]

stacking_A_psi = [
  # N   A     C     G     U     P
  [ 0,  0,    0,    0,    0,    0 ], # N
  [ 0,  0,    0,    0,    -69,  0 ], # A
  [ 0,  0,    0,    -105, 0,    0 ], # C
  [ 0,  0,    -69,  0,    0,    0 ], # G
  [ 0,  -170, 0,    0,    0,    0 ], # U
  [ 0,  0,    0,    0,    0,    0 ]  # P
]

# Energy correction callback passed to ViennaRNA library
def correction(i, j, k, l, step, encoding):
    if (step == RNA.DECOMP_PAIR_IL): # an internal loop is evaluated
        if  i + 1 == k and j - 1 == l:  # (i,j) and (k,l) are stacking
            if (encoding[i] == 5 and encoding[j] == 1):
                return stacking_psi_A[encoding[k]][encoding[l]]
            elif (encoding[k] == 1 and encoding[l] == 5):
                return stacking_psi_A[encoding[j]][encoding[i]]
            elif (encoding[i] == 1 and encoding[j] == 5):
                return stacking_A_psi[encoding[k]][encoding[l]]
            elif (encoding[k] == 5 and encoding[l] == 1):
                return stacking_A_psi[encoding[j]][encoding[i]]

    return 0
```

[9] Hudson et al., “Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides.”, 2013, RNA 19.11, 1474–1482

Example: human tRNA Phe (tdbR00000103)

4. Suboptimals with RT masking and stacking energies for Ψ-A[10]



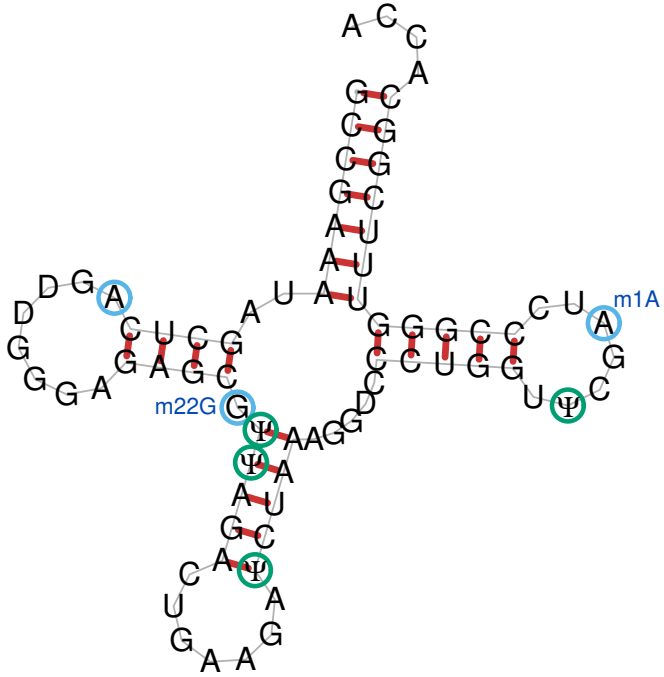


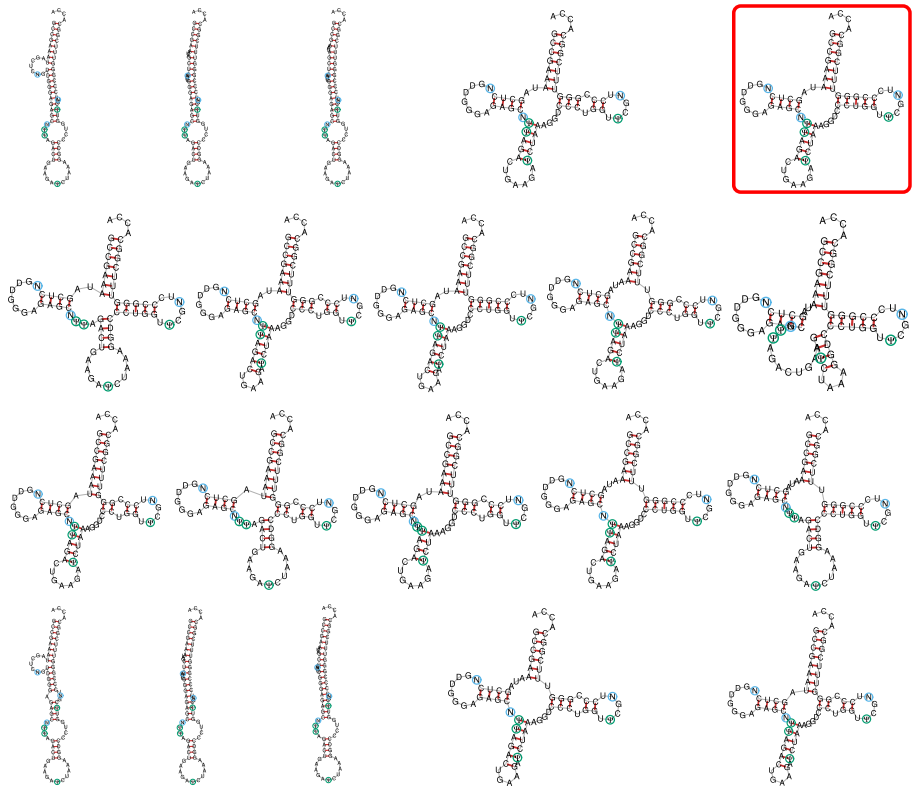
- Correct cloverleaf structure at position 5
- $\bar{d}^{3}_{\mathrm{BP}} = 15.5$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 0.83$ kcal/mol
- Can we do better?

[10] Hudson et al., "Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides.", 2013, RNA 19.11, 1474–1482

Example: human tRNA Phe (tdbR00000103)

4. Suboptimals with RT masking and stacking energies for Ψ-A[10]





- Correct cloverleaf structure at position 5
- $\bar{d}^{3}_{\mathrm{BP}} = 15.5$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 0.83$ kcal/mol
- Can we do better?

Dihydrouridine (D) adds flexibility to the D-loop!

[10] Hudson et al., "Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides.", 2013, RNA 19.11, 1474–1482

Example: human tRNA Phe (tdbR00000103)

No (stacking) energies with dihydrouridine (D) are available

Structural effects of dihydrouridine[11]

- C3'-endo sugar conformation is destabilized in favor of C2'-endo
- more flexibility
- promotes destacking
- destabilization of 1.5 kcal/mol (mono), up to 5.3 kcal/mol (oligo)

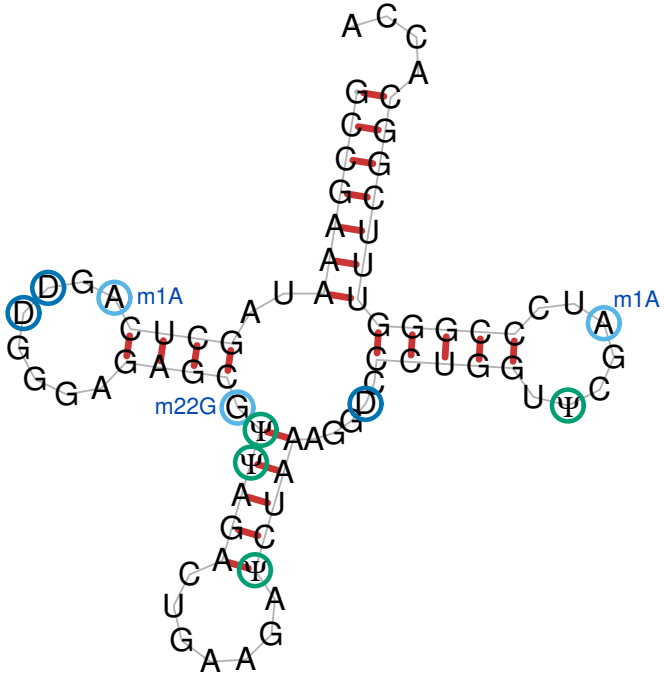
Required changes for `correction` function:

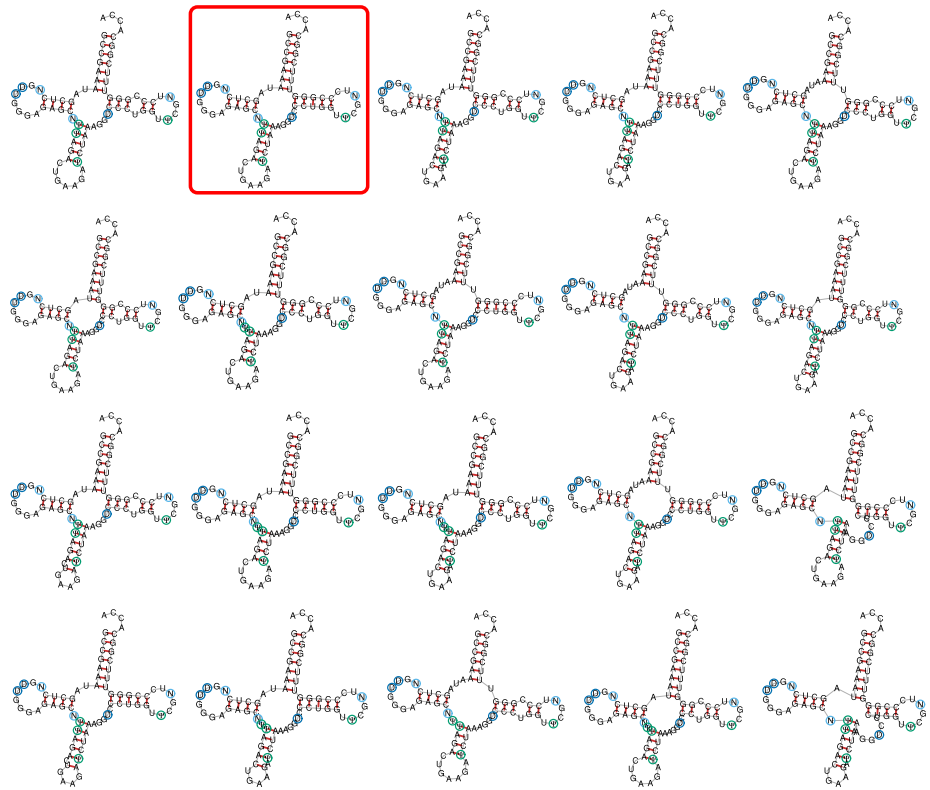
```
def correction(i, j, k, l, step, encoding):
    e = 0 # default correction

    if (step == RNA.DECOMP_PAIR_IL): # an internal loop is evaluated
        if  i + 1 == k and j - 1 == l:  # (i,j) and (k,l) are stacking
            if (encoding[i] == 5 and encoding[j] == 1):
                e += stacking_psi_A[encoding[k]][encoding[l]]
            elif (encoding[k] == 1 and encoding[l] == 5):
                e += stacking_psi_A[encoding[j]][encoding[i]]
            elif (encoding[i] == 1 and encoding[j] == 5):
                e += stacking_A_psi[encoding[k]][encoding[l]]
            elif (encoding[k] == 5 and encoding[l] == 1):
                e += stacking_A_psi[encoding[j]][encoding[i]]

            if  encoding[i] == 6 or \
                encoding[j] == 6 or \
                encoding[k] == 6 or \
                encoding[l] == 6:
                e += 150 # D destabilizes stacking by at least 1.5 kcal/mol

    return e
```

[11] Dalluge et al., "Conformational flexibility in RNA: the role of dihydrouridine.", 1996, NAR 24.6, 1073–1079

Example: human tRNA Phe (tdbR00000103)

5. Suboptimals with RT masking and energies for Ψ-A and D[12]



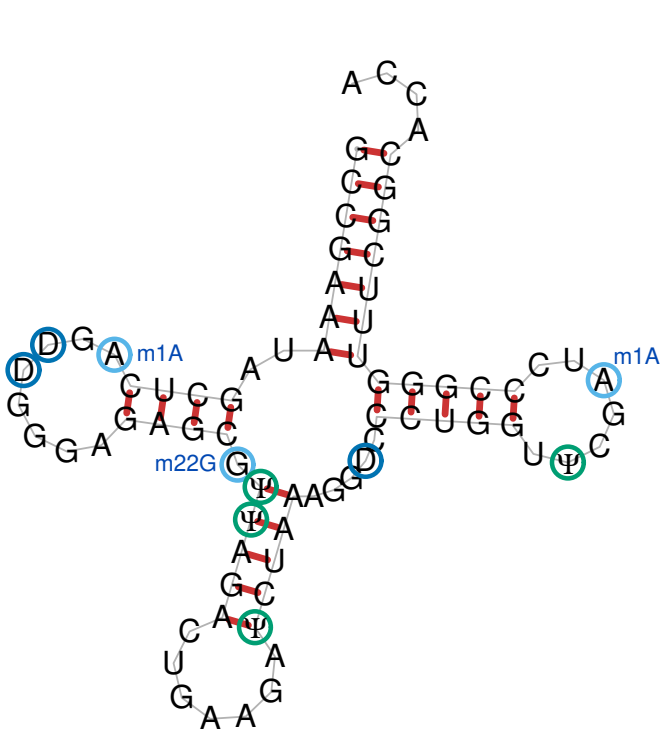


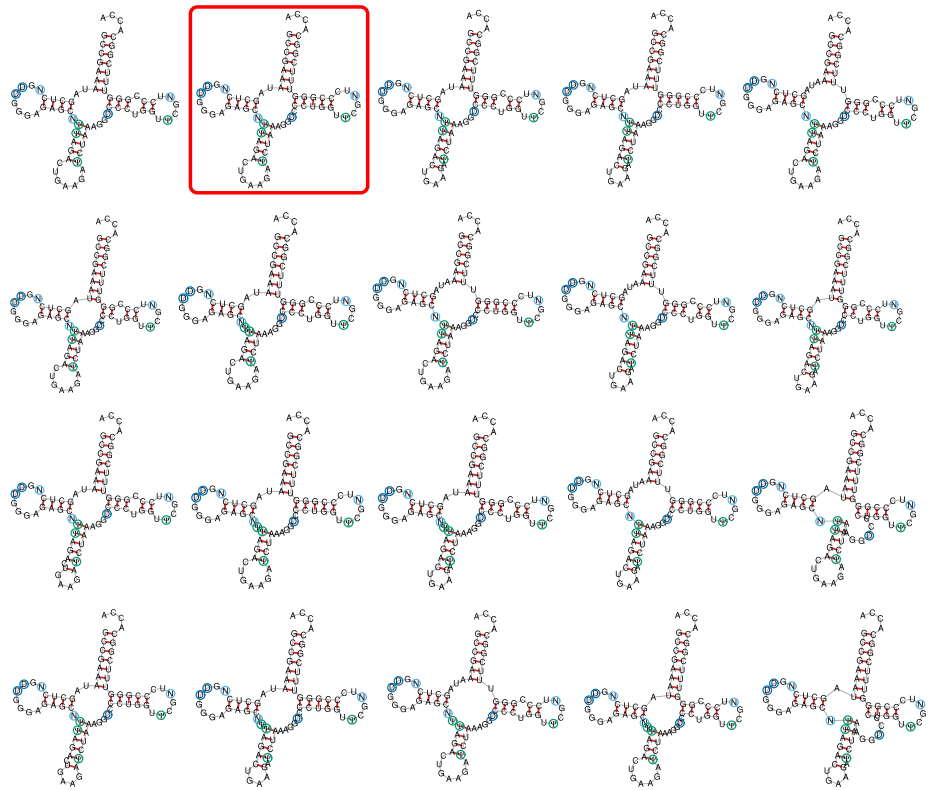
- Correct cloverleaf structure at position 2
- $\bar{d}^{3}_{\mathrm{BP}} = 6.92$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 0.2$ kcal/mol

[12] Dalluge et al., “Conformational flexibility in RNA: the role of dihydrouridine.”, 1996, NAR 24.6, 1073–1079

Example: human tRNA Phe (tdbR00000103)

5. Suboptimals with RT masking and energies for Ψ-A and D[12]



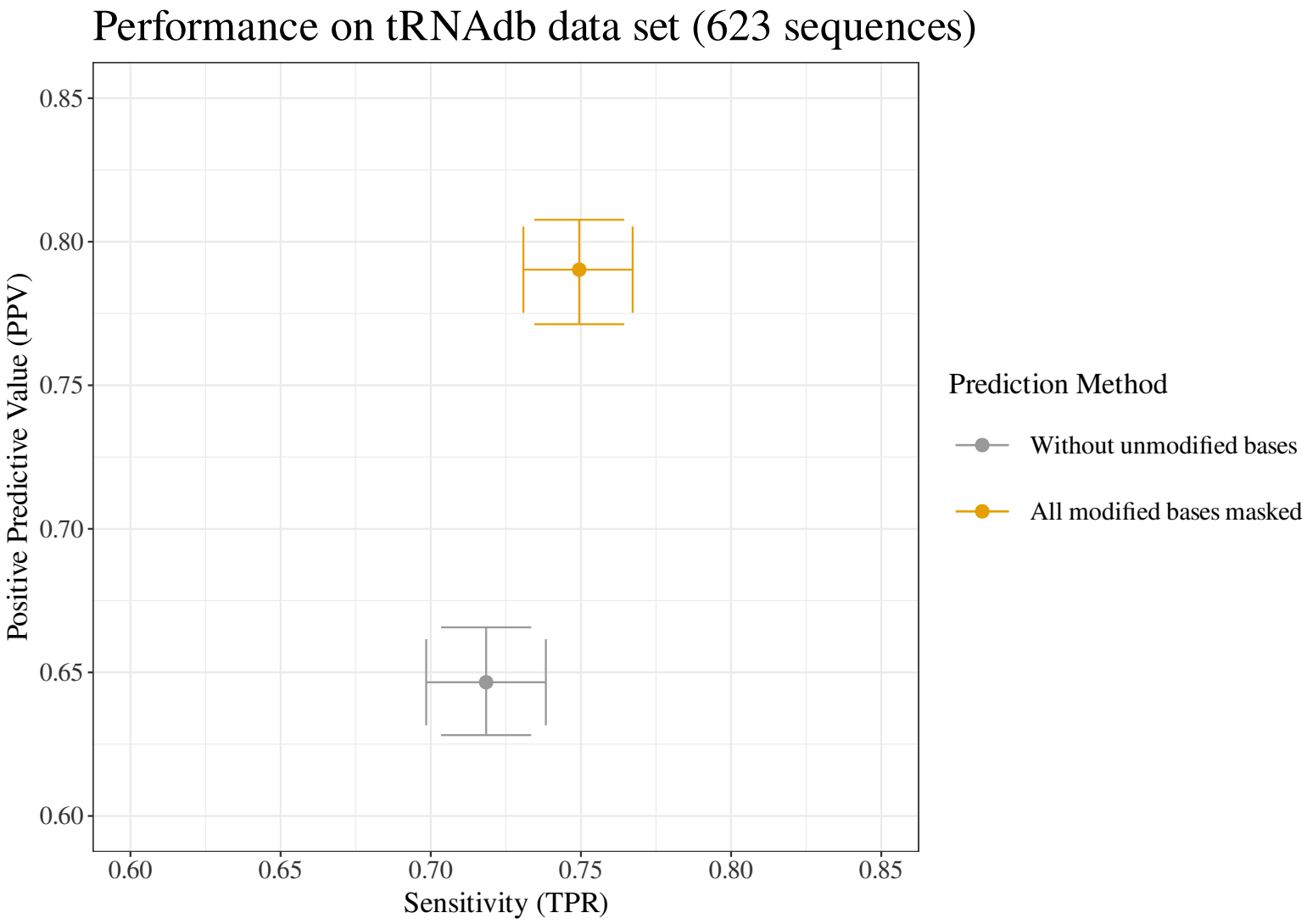


- Correct cloverleaf structure at position 2
- $\bar{d}^{3}_{\mathrm{BP}} = 6.92$ bp, $E_{\mathrm{ref}} - E_{\mathrm{MFE}} = 0.2$ kcal/mol

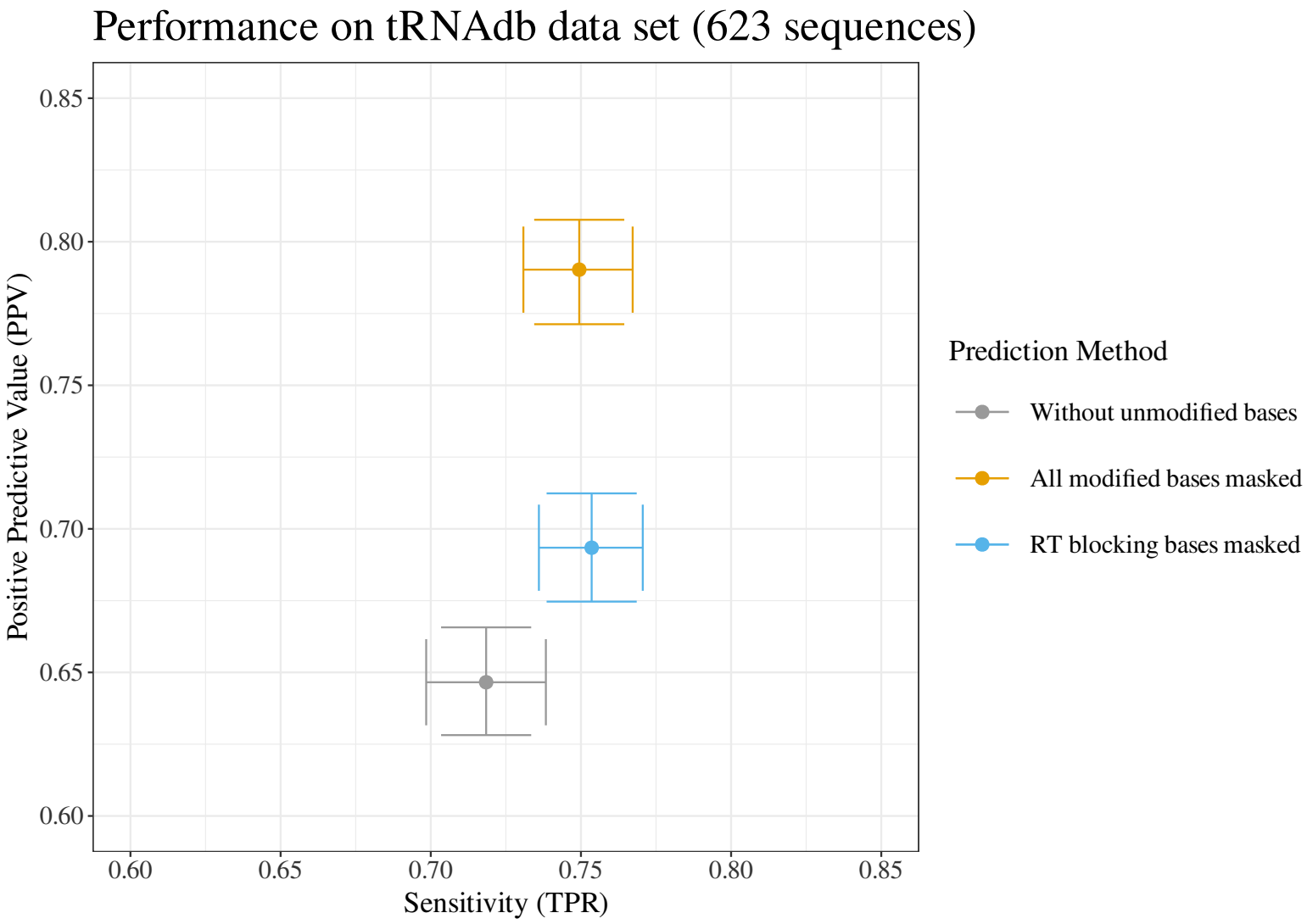
How do these constraints affect prediction for other tRNAs?

[12] Dalluge et al., "Conformational flexibility in RNA: the role of dihydrouridine.", 1996, NAR 24.6, 1073–1079

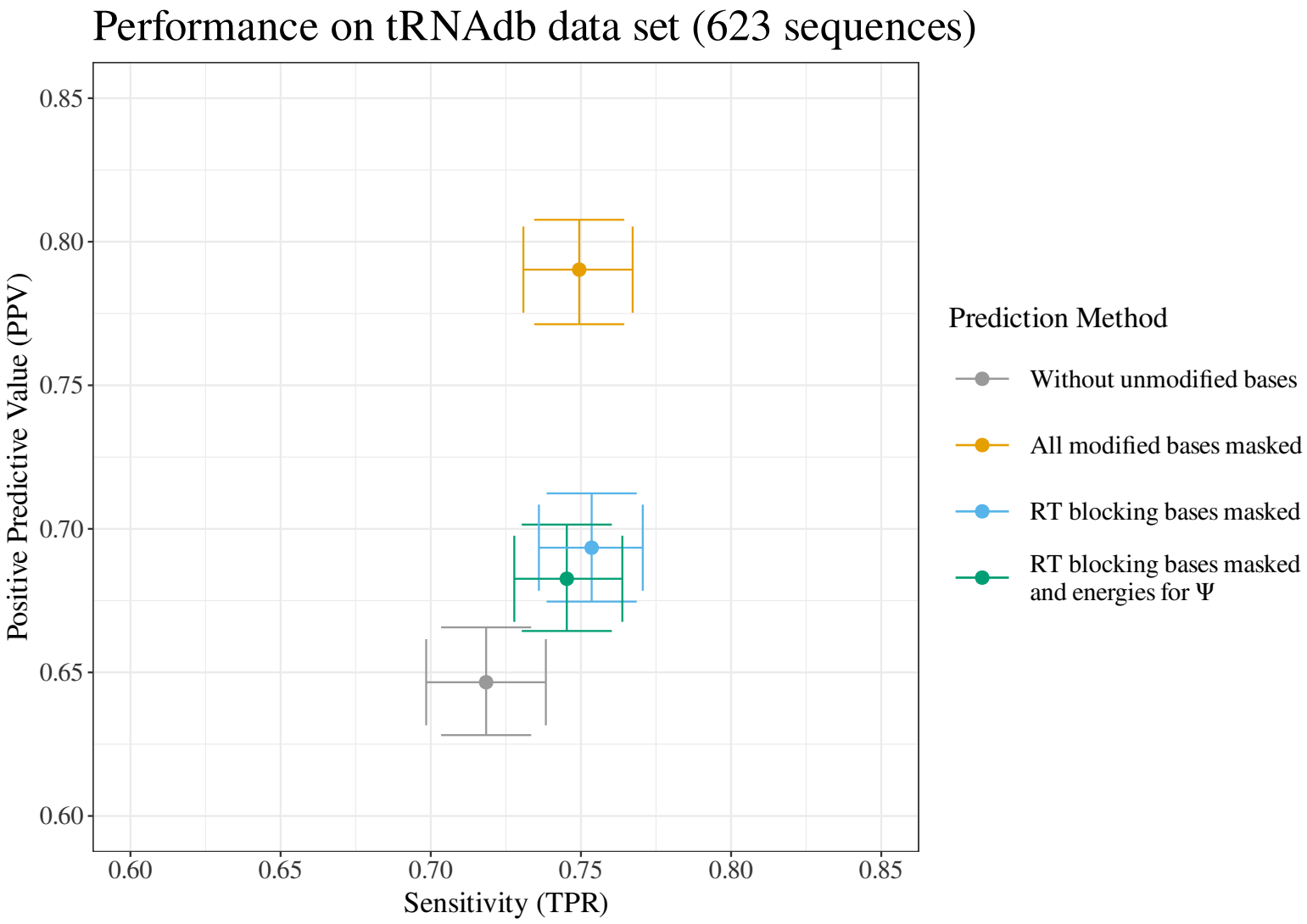
Overall prediction improvement



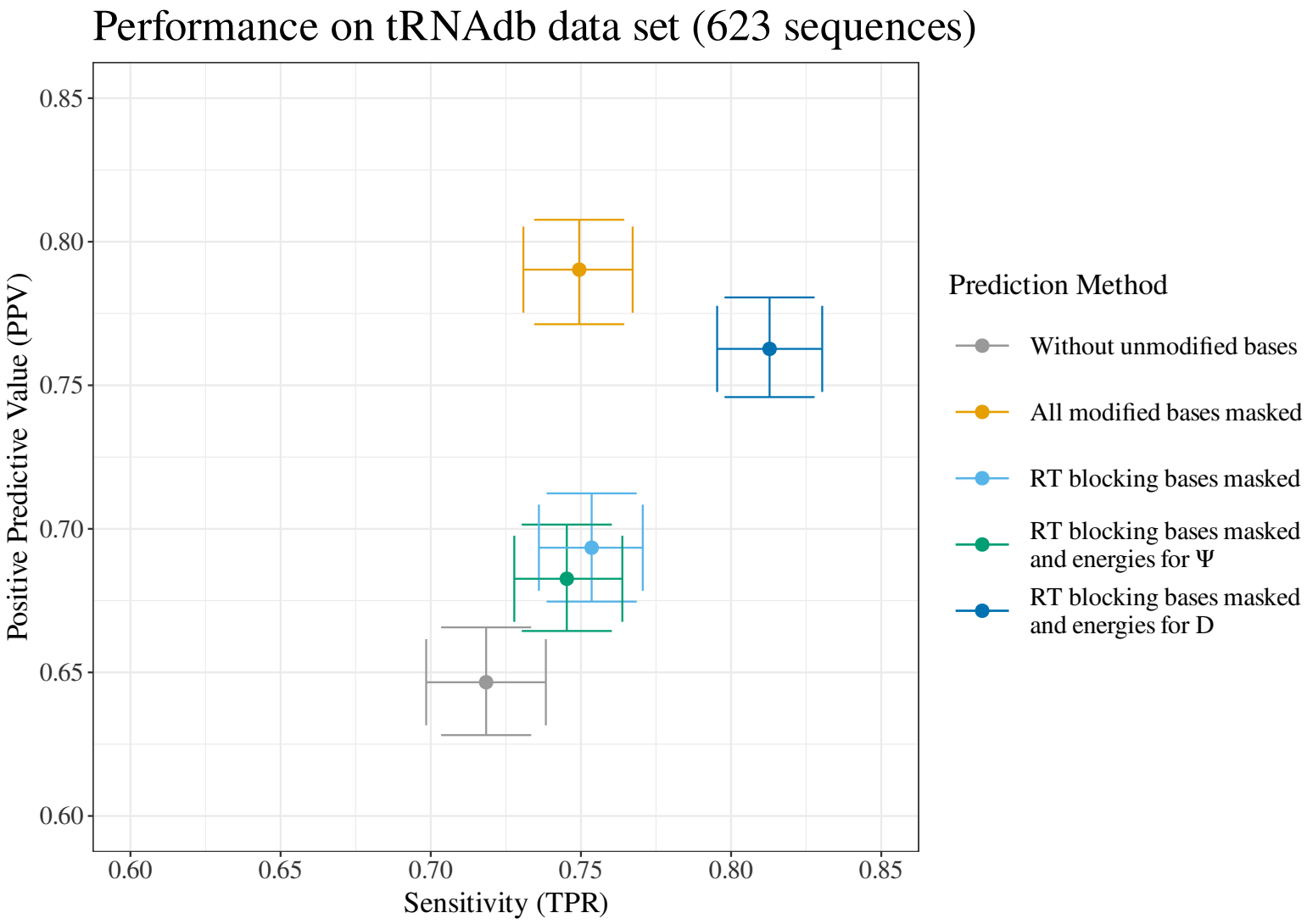
Overall prediction improvement



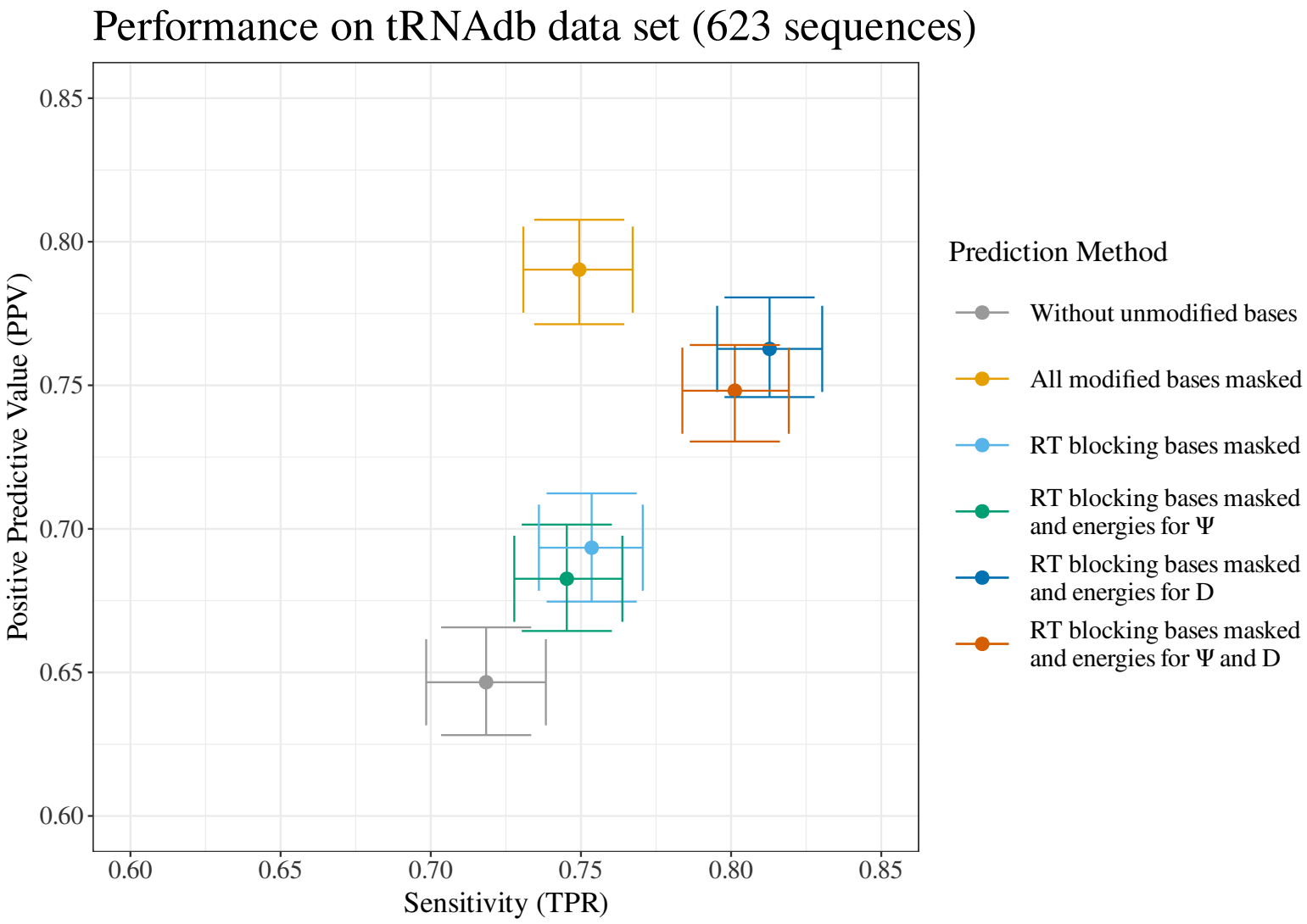
Overall prediction improvement



Overall prediction improvement



Overall prediction improvement



Conclusion

Modified bases may heavily influence structure space

The good news

- Some data on structural effects of modifications available
- Better prediction performance through constraints
- Constraints available for many prediction algorithms

The not so good news

- Additional parameters do not necessarily increase performance
- Constraints become complex for more modifications and contexts
- Unrealistic to include full parameters with many modified bases
- No unique base annotation (`tRNAdb`[13], `RNAmod`[14], `MODOMICS`[15])

Outlook

- Gather more data on structural effects of modified bases
- Rule and energy parameter set for pairs with modified bases
- Integration of modified bases in `ViennaRNA Package`

---

[13] Jühling et al., "tRNAdb 2009: compilation of tRNA sequences and tRNA genes.", 2009, NAR 37, D159–D162

[14] Liu et al., "RNAmod: an integrated system for the annotation of mRNA modifications", 2019, NAR 47, W548-W555

[15] Boccaletto et al., "MODOMICS: a database of RNA modification pathways. 2017 update", 2018, NAR 46, D303-D307

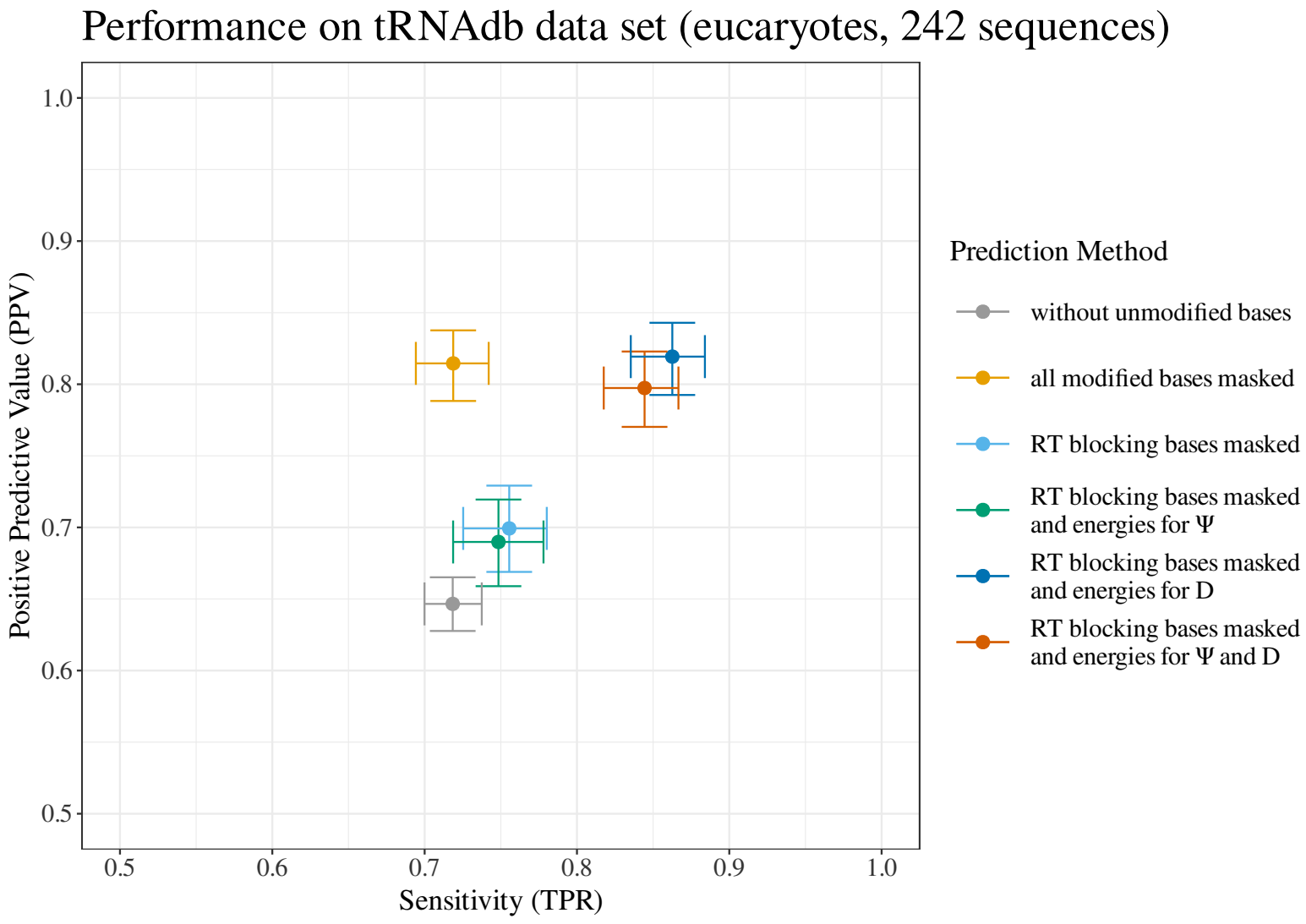
Acknowledgements

- Ivo L. Hofacker
- Thomas Spicher
- Peter F. Stadler
- Yuliia Varenyk
- The remaining TBI group

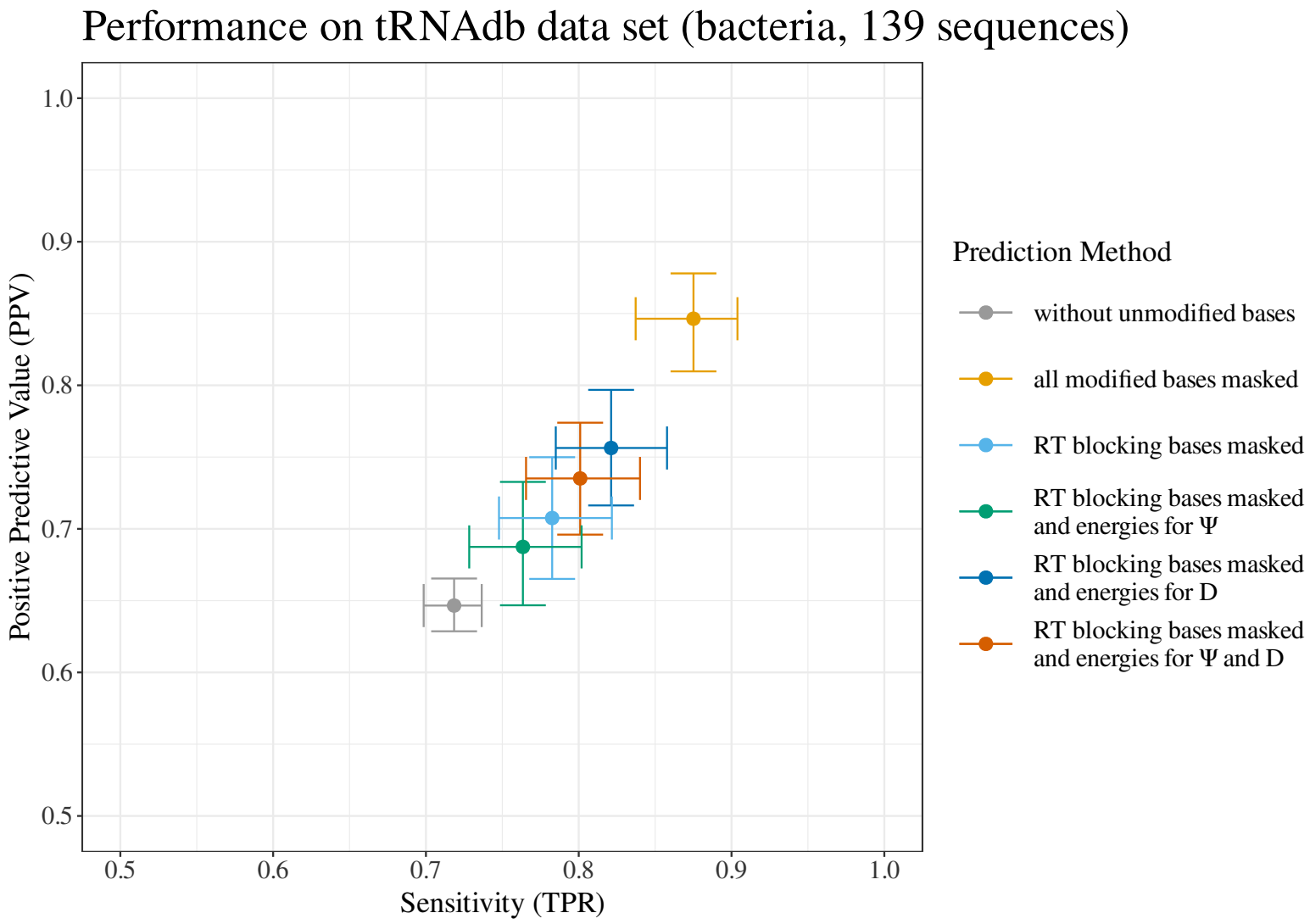
Thank You for your attention!

FWF Der Wissenschaftsfonds.

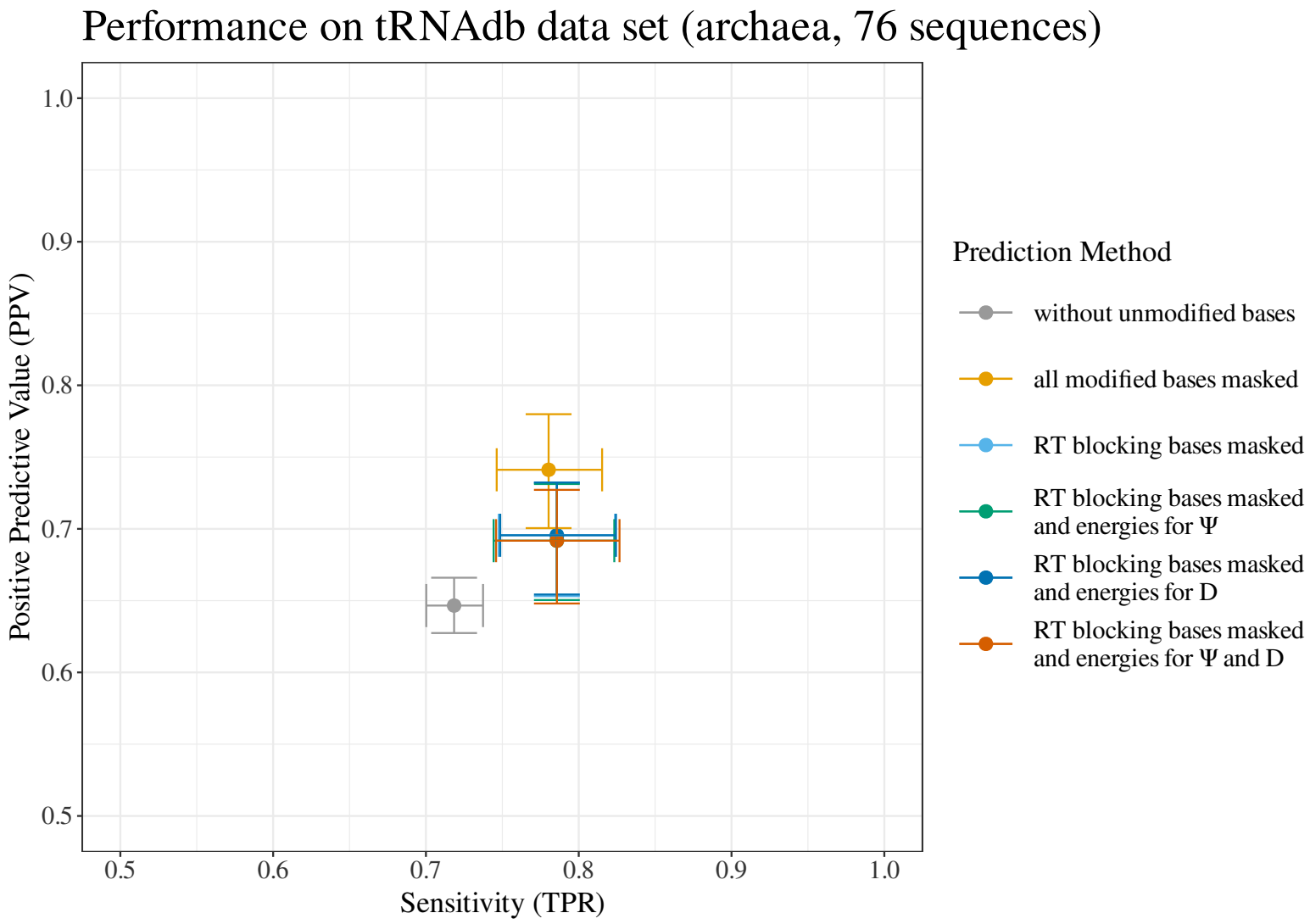
Overall prediction improvement



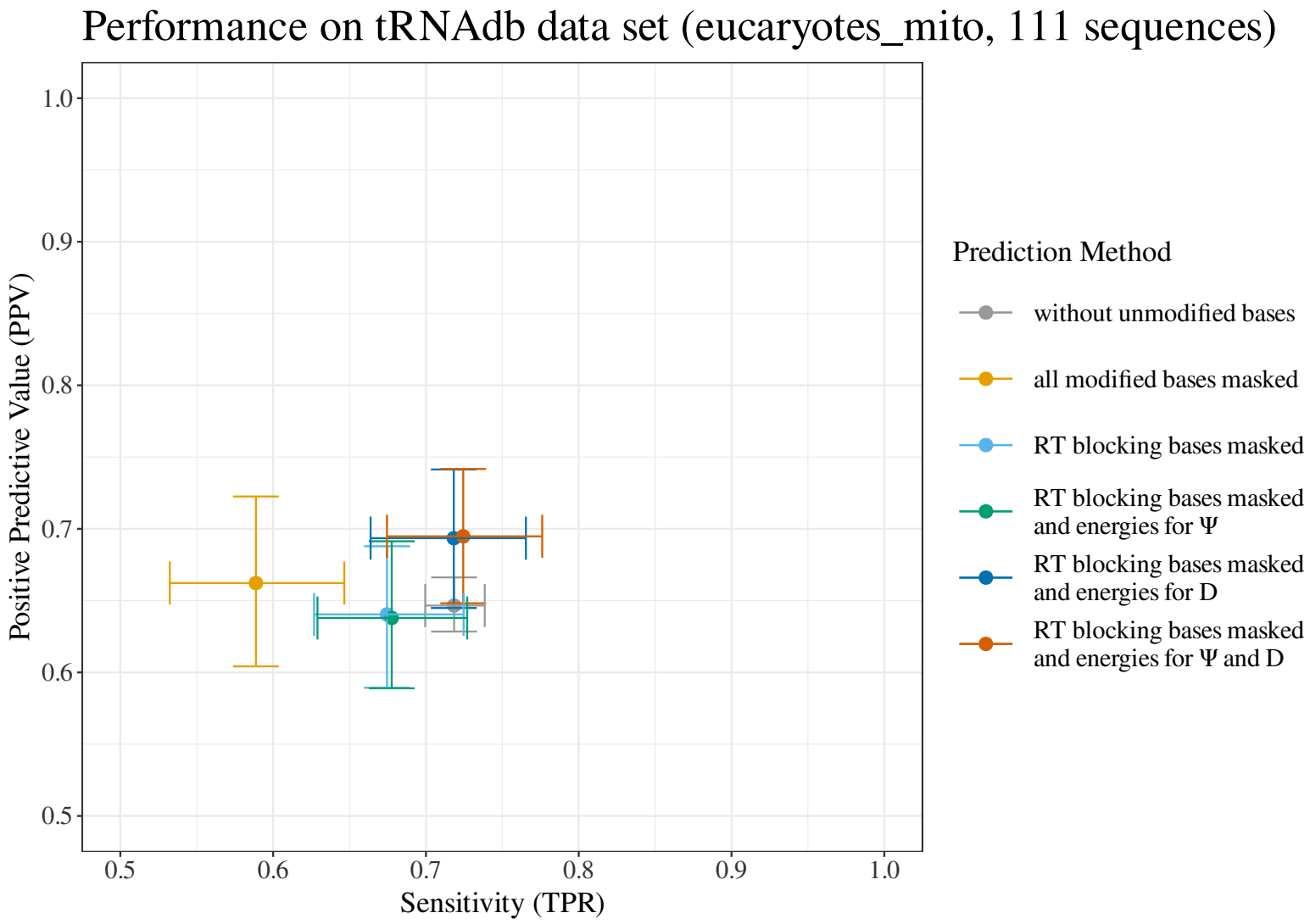
Overall prediction improvement



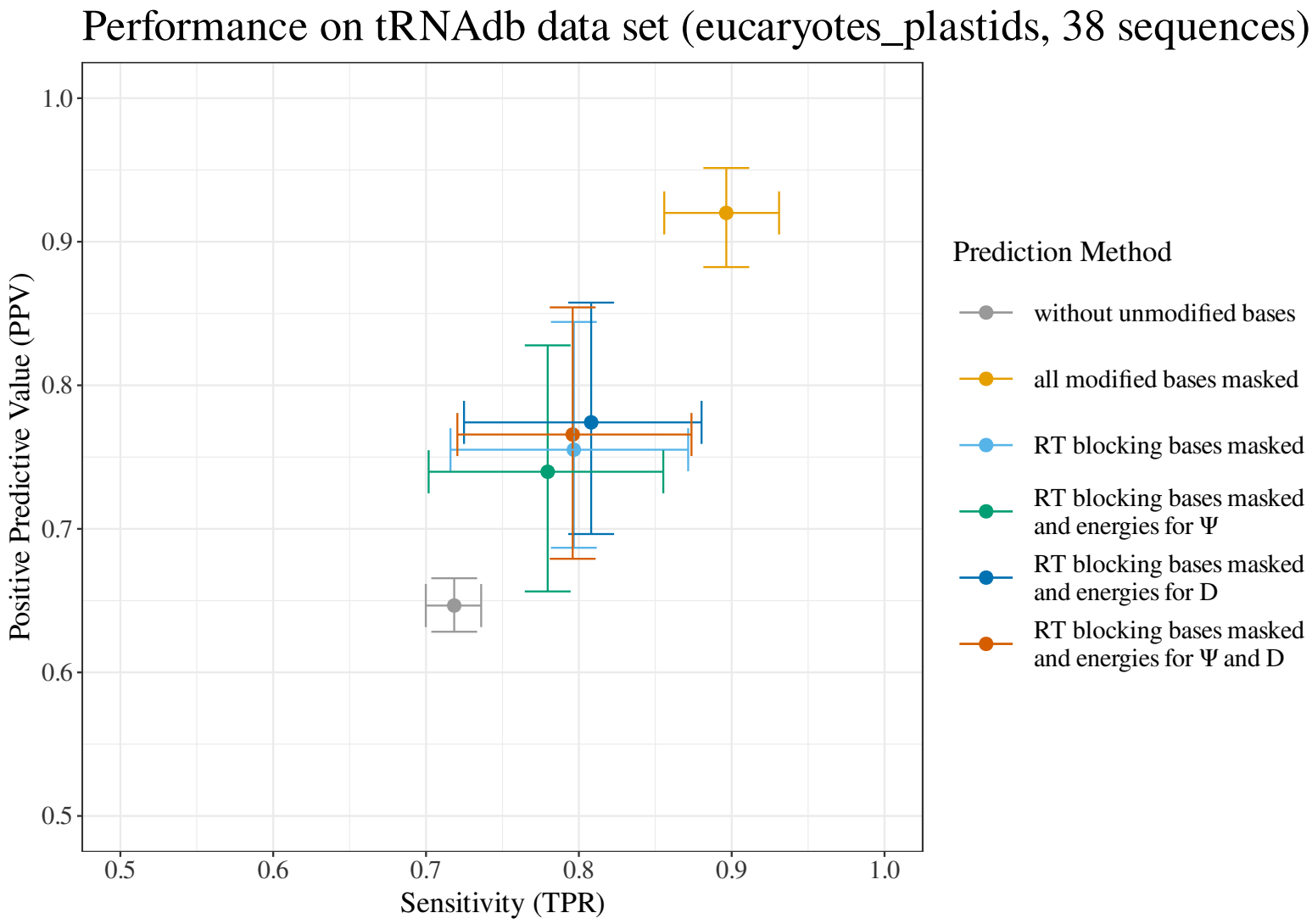
Overall prediction improvement



Overall prediction improvement



Overall prediction improvement



Overall prediction improvement