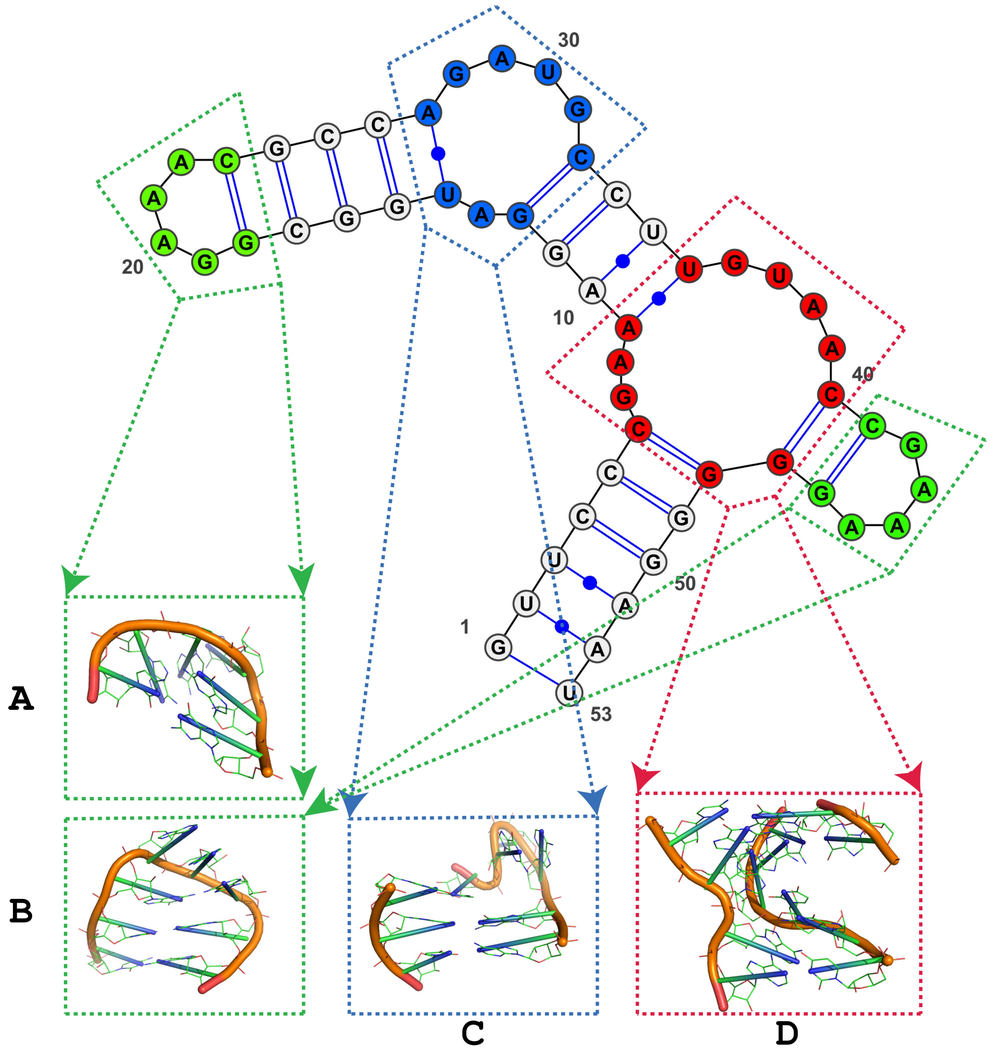
All basepairs interactions RNA 3D modules, variants, and prediction

Vladimir Reinharz (UQAM)

Wilfried Agbeto, Théo Boury, Roman Sarrazin-Gendron, Maëva Burillo

Loops are organized



RNA motifs

Current Opinion in Structural Biology
Volume 13, Issue 3, June 2003, Pages 300-308

Analysis of RNA motifs

Neocles B Leontis * ✉, Eric Westhof † ✉

RNA motifs mediate the specific interactions that induce the compact folding of complex RNAs

RNA motifs also constitute specific protein or ligand binding sites

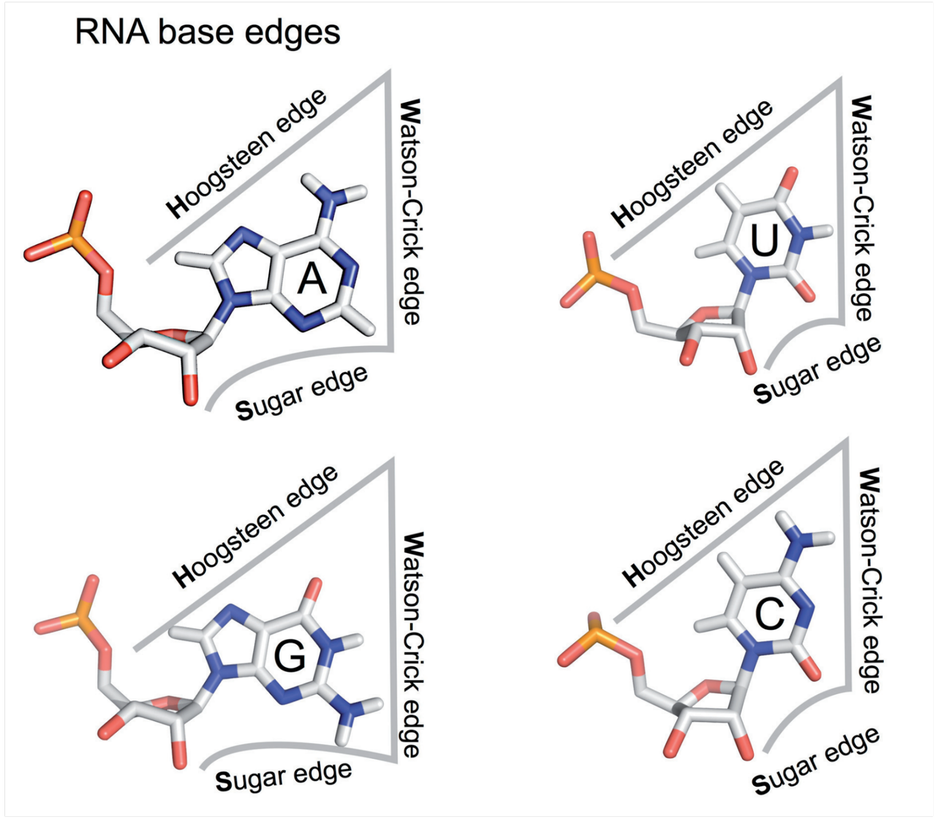
RNA motifs are directed and ordered stacked arrays of non-Watson–Crick base pairs forming distinctive foldings of the phosphodiester backbones of the interacting RNA strands. [...]

A given motif is characterized by all the sequences that fold into essentially identical three-dimensional structures with the same ordered array of isosteric non-Watson–Crick base pairs

RNA modules

RNA ~~motifs~~ modules are directed and ordered stacked arrays of non-Watson–Crick base pairs forming distinctive foldings of the phosphodiester backbones of the interacting RNA strands

Non-canonical base pairs



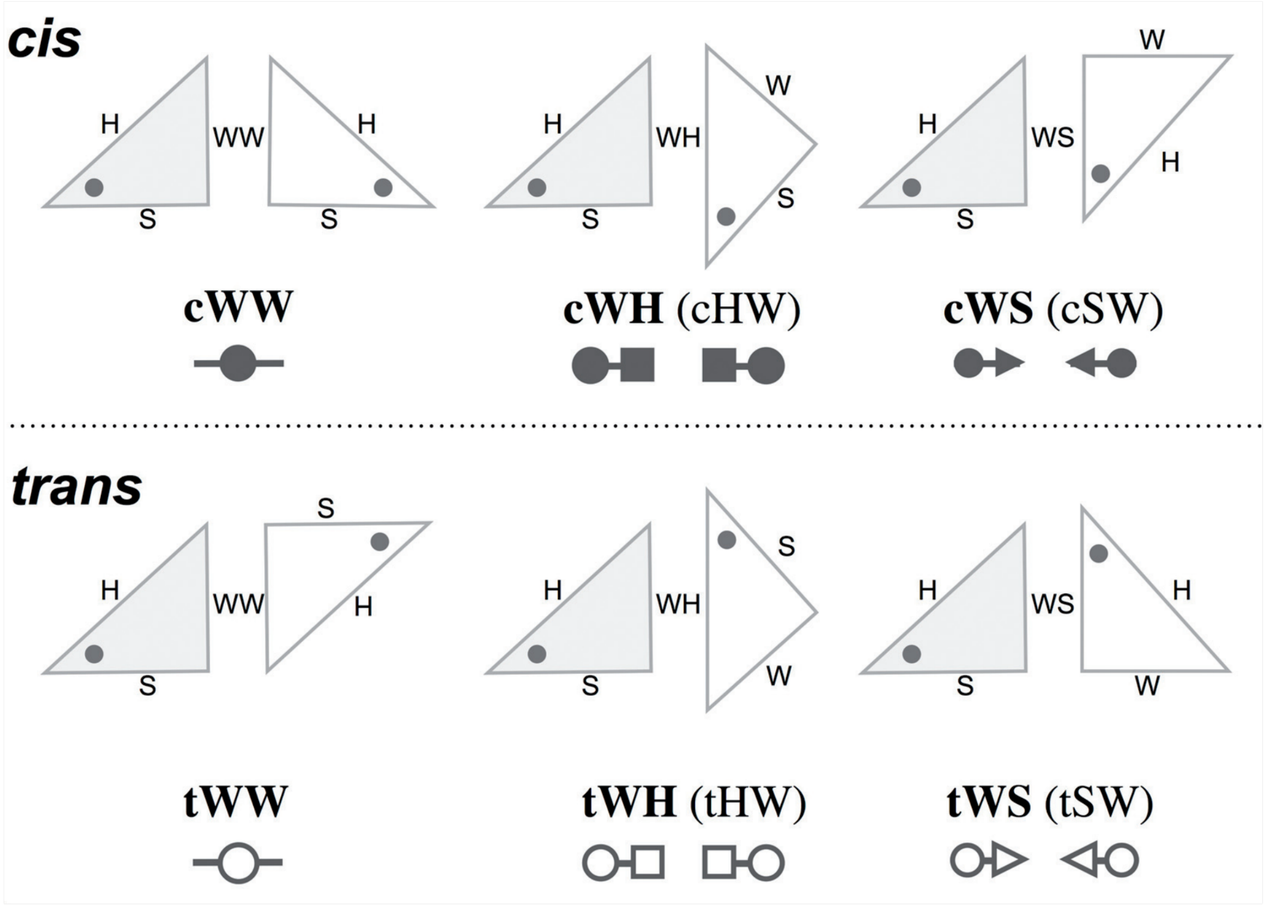
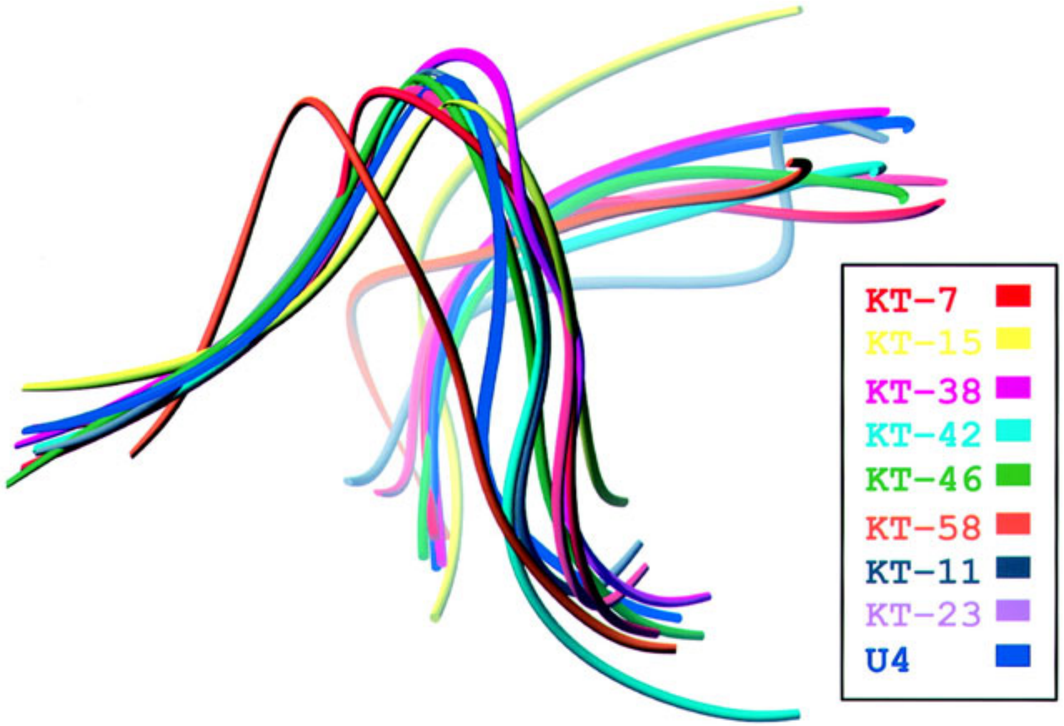
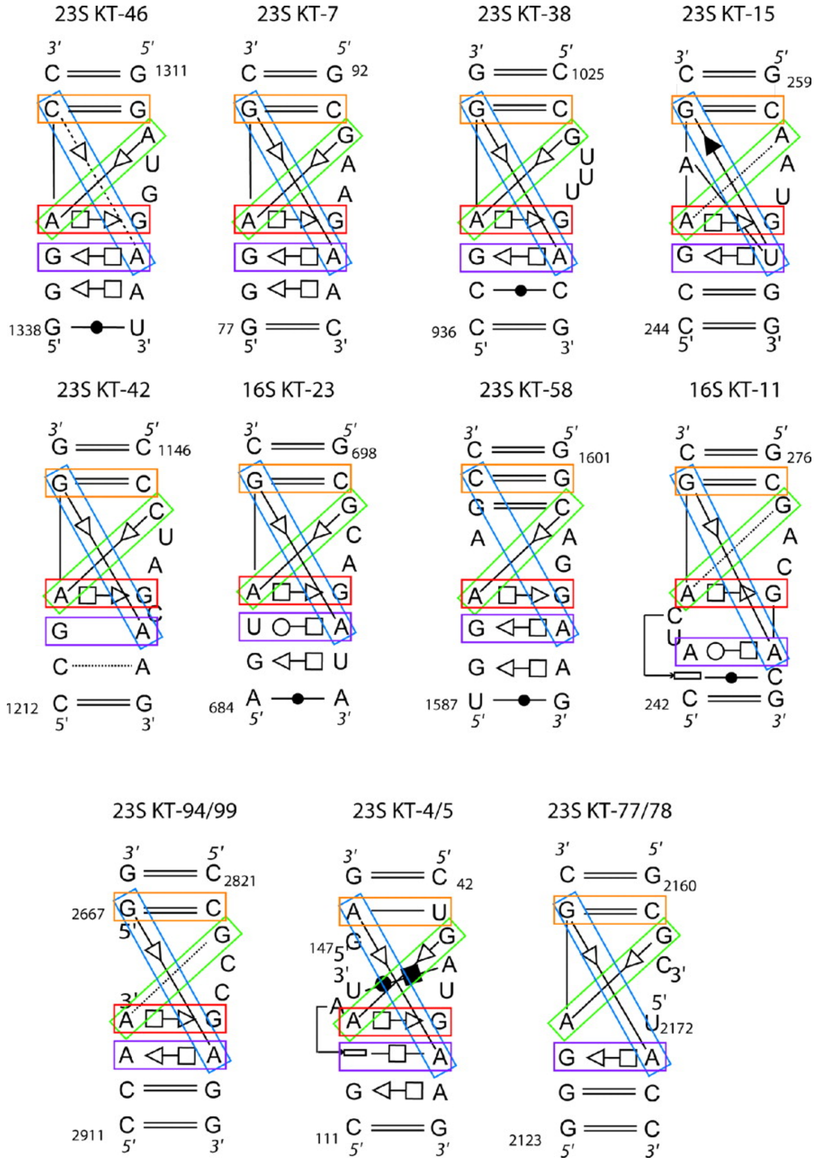


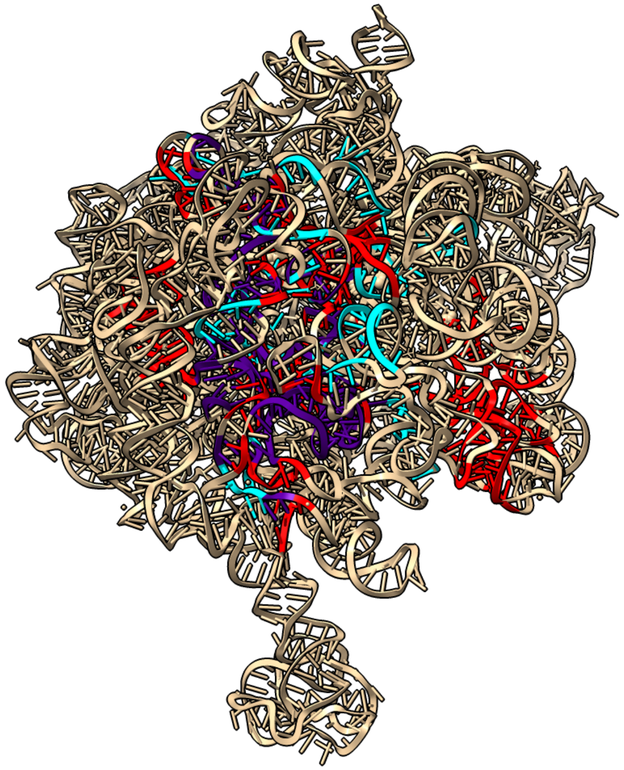
Figure adapted from
Almakarem et al, 2011

Conserved basepairs interaction networks

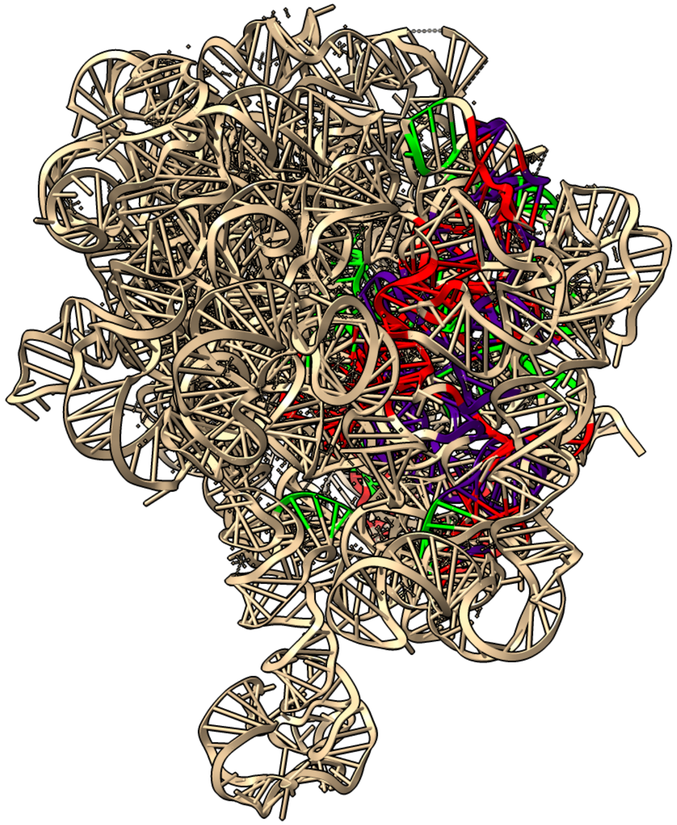




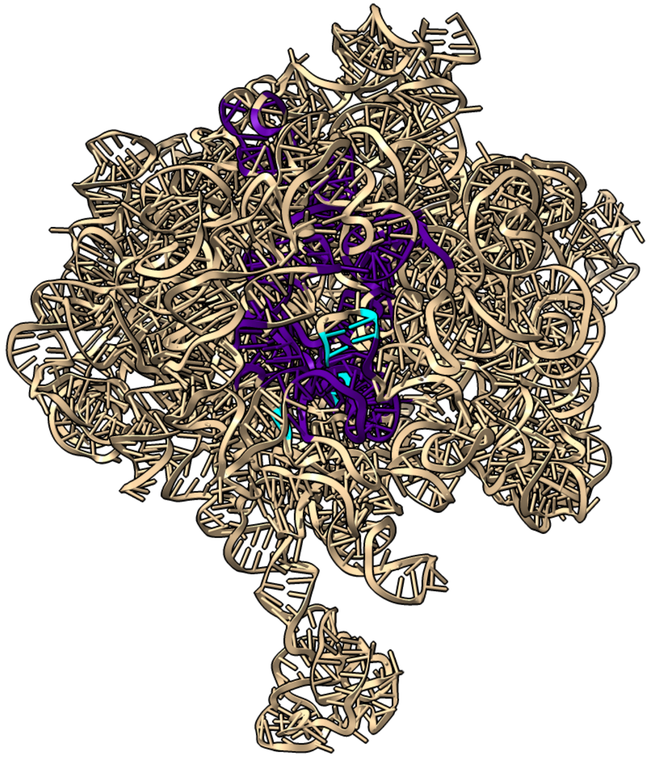
Large interaction networks span many loops



5J7L



4Y4O



6SPB

Given arbitrary graphs, what are all conserved subgraphs? (enumeration of maximal subgraph isomorphism)

- NP-Hard
- Can be exponential in nodes

But

- Biological structures have low degree
- We have lots of computers

Subgraphs isomorphisms in brief

- Non-canonical basepairs modules between pairs of loops (2018)
  337 modules

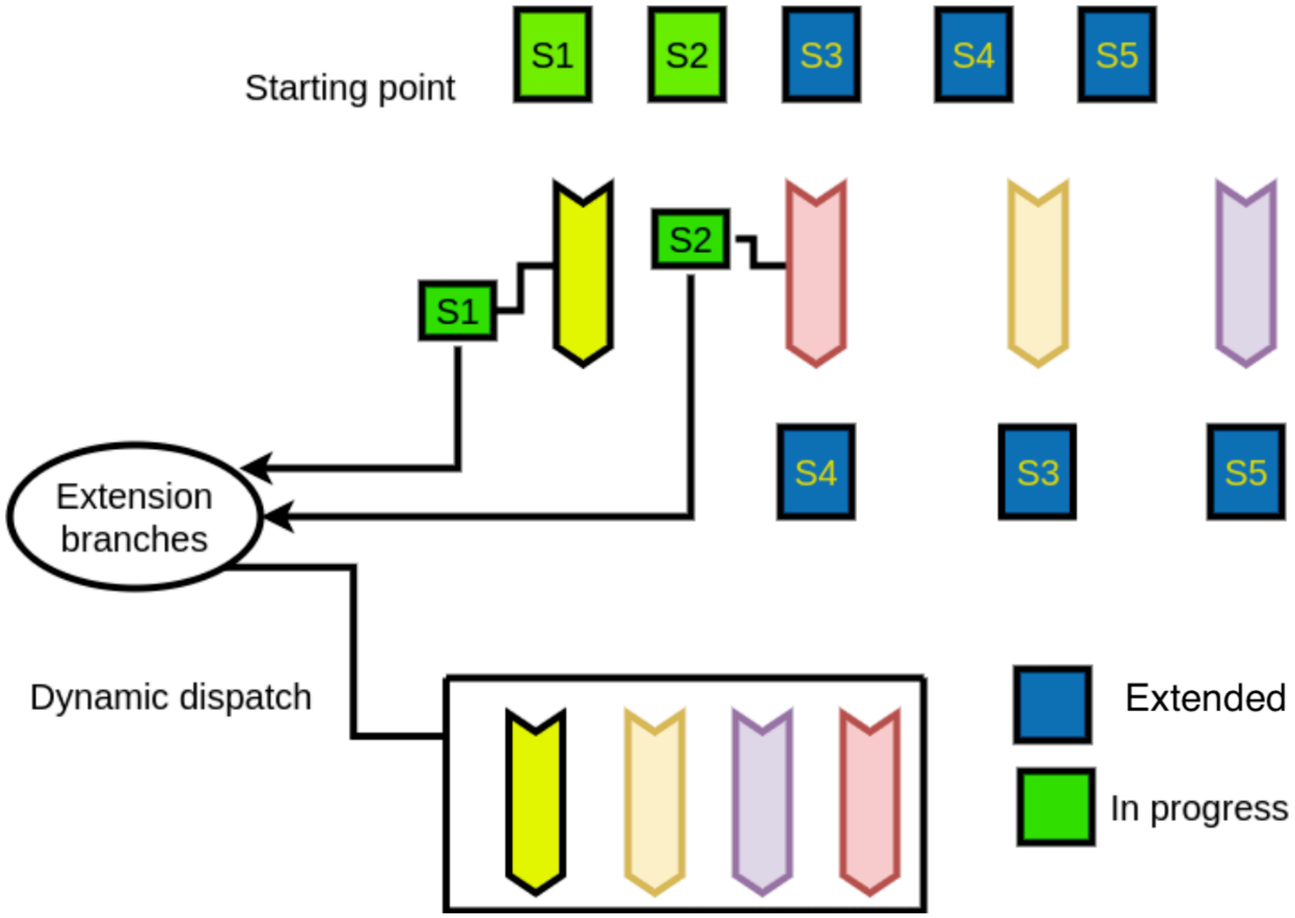
- Any non-canonical basepairs modules between RNAs (2021)
  3300 modules

Wilfried Agbeto

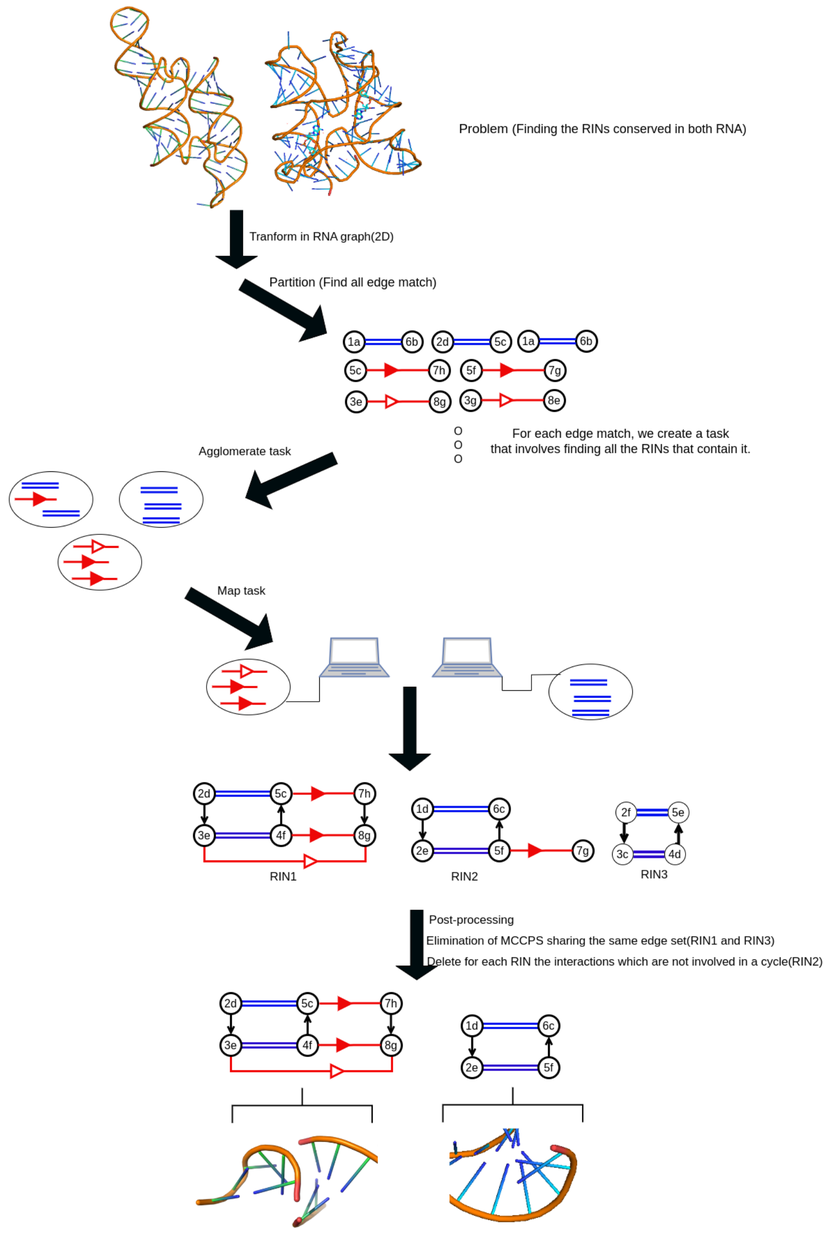
- Non-canonical basepairs and stacking modules
  (2024: PasiGraph on github to appear in PPAM2024)
  157 344 modules

Distributed system and memory subgraph isomorphism

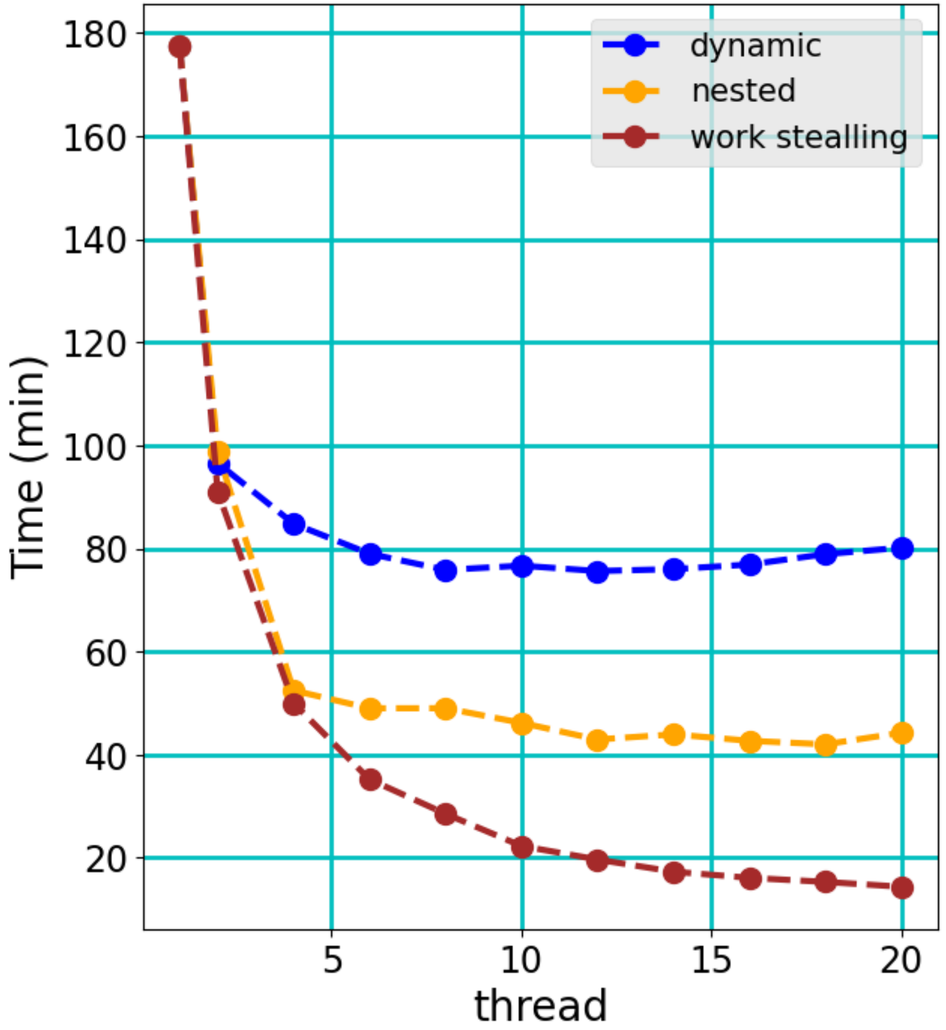
- Threads receive tasks (starting points)
- Threads steal sub-tasks (extension branches) when finished
- Keys for each subgraphs avoid doubling tasks

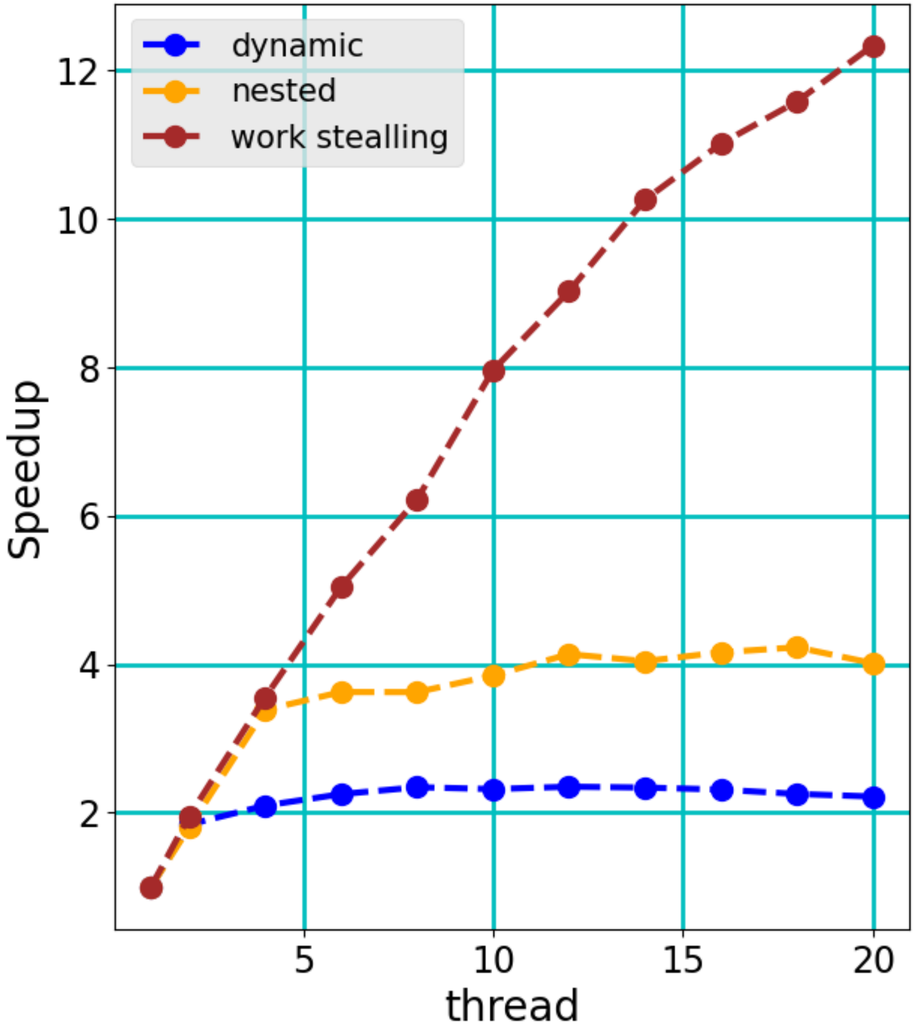


PasiGraph : https://gitlab.info.uqam.ca/cbe/pasigraph

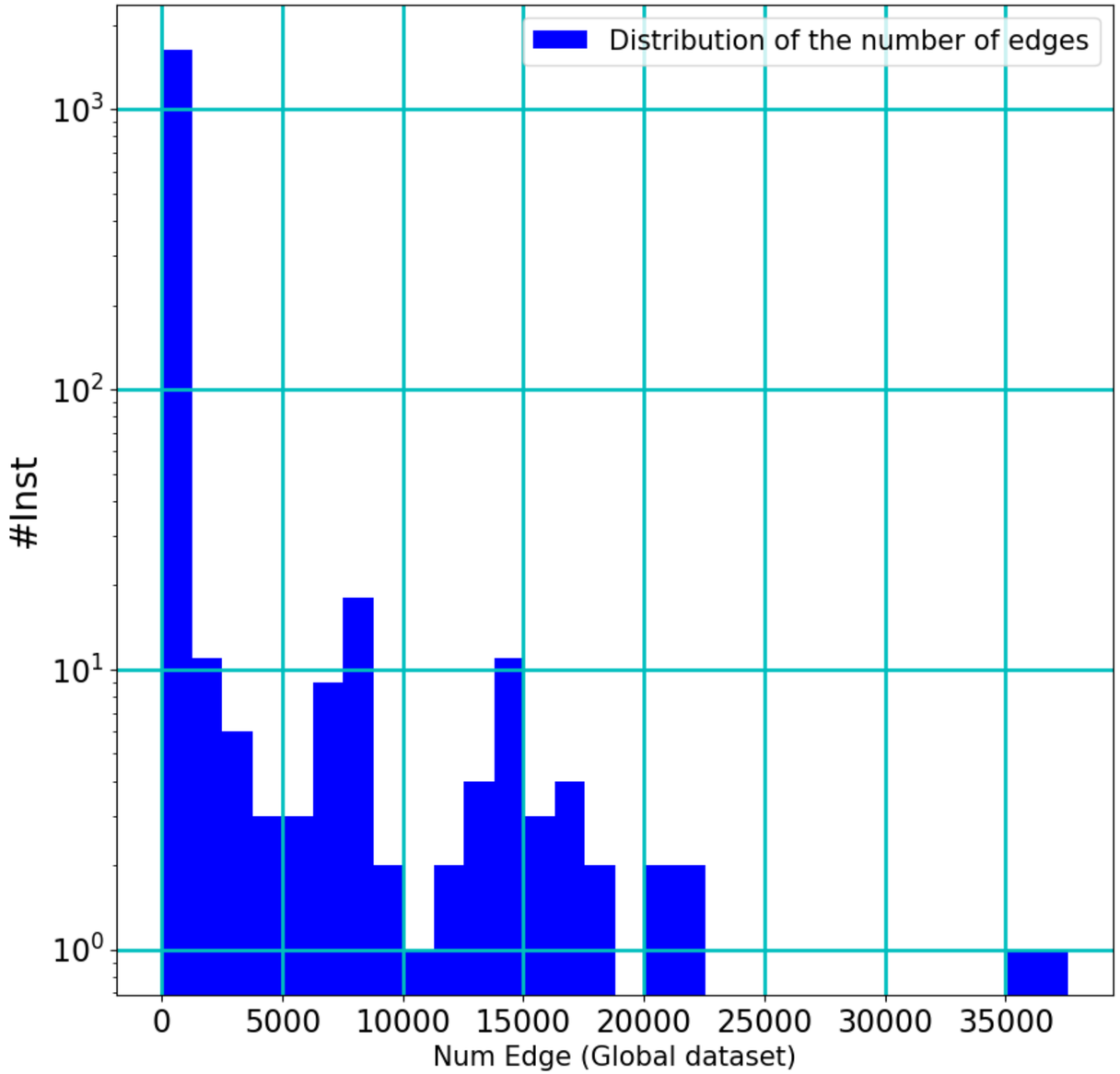


Distributed computing implementation matters





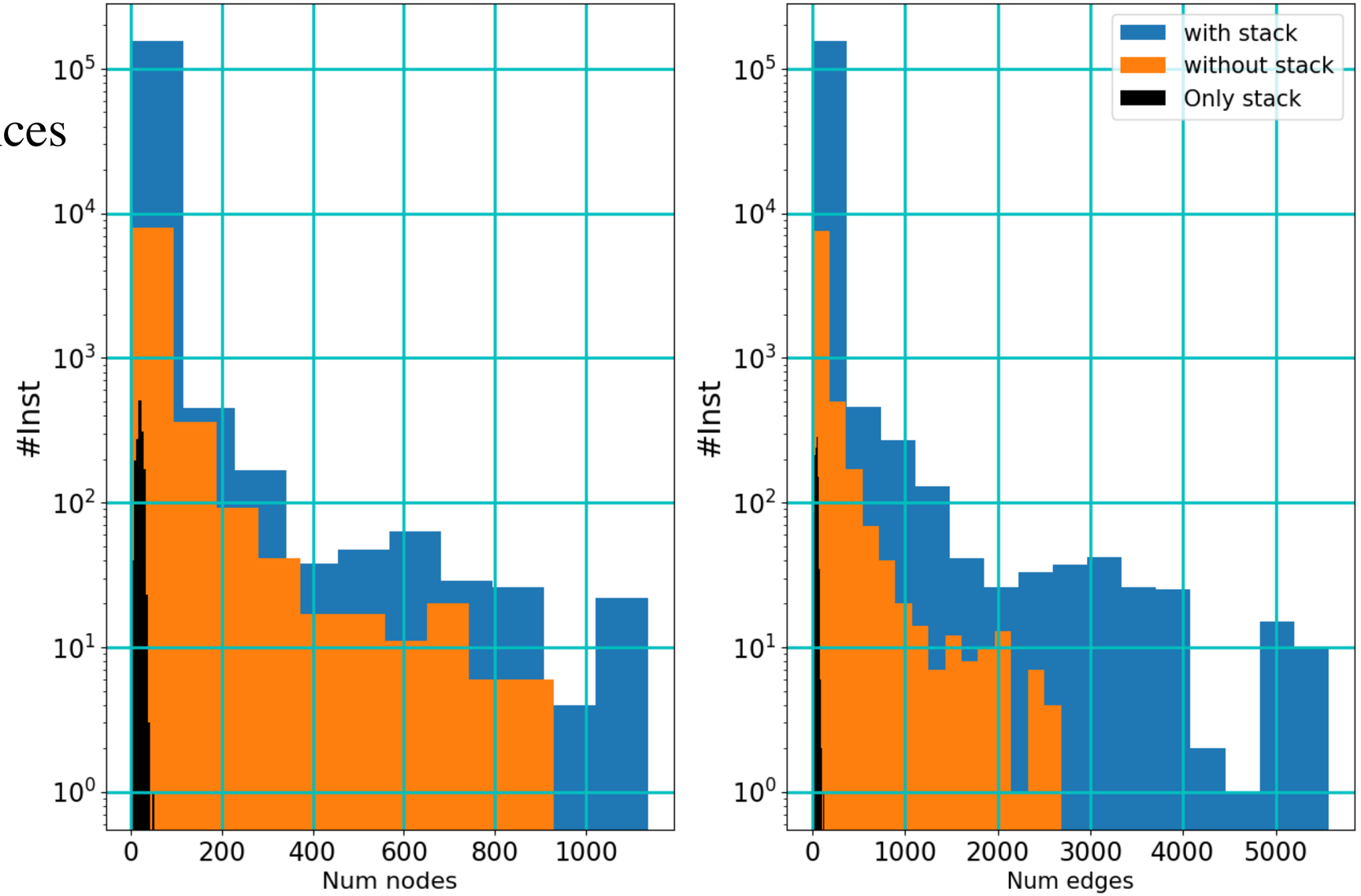
The 1 722 non-redundant structures RNA3DAtlas dataset 3.269



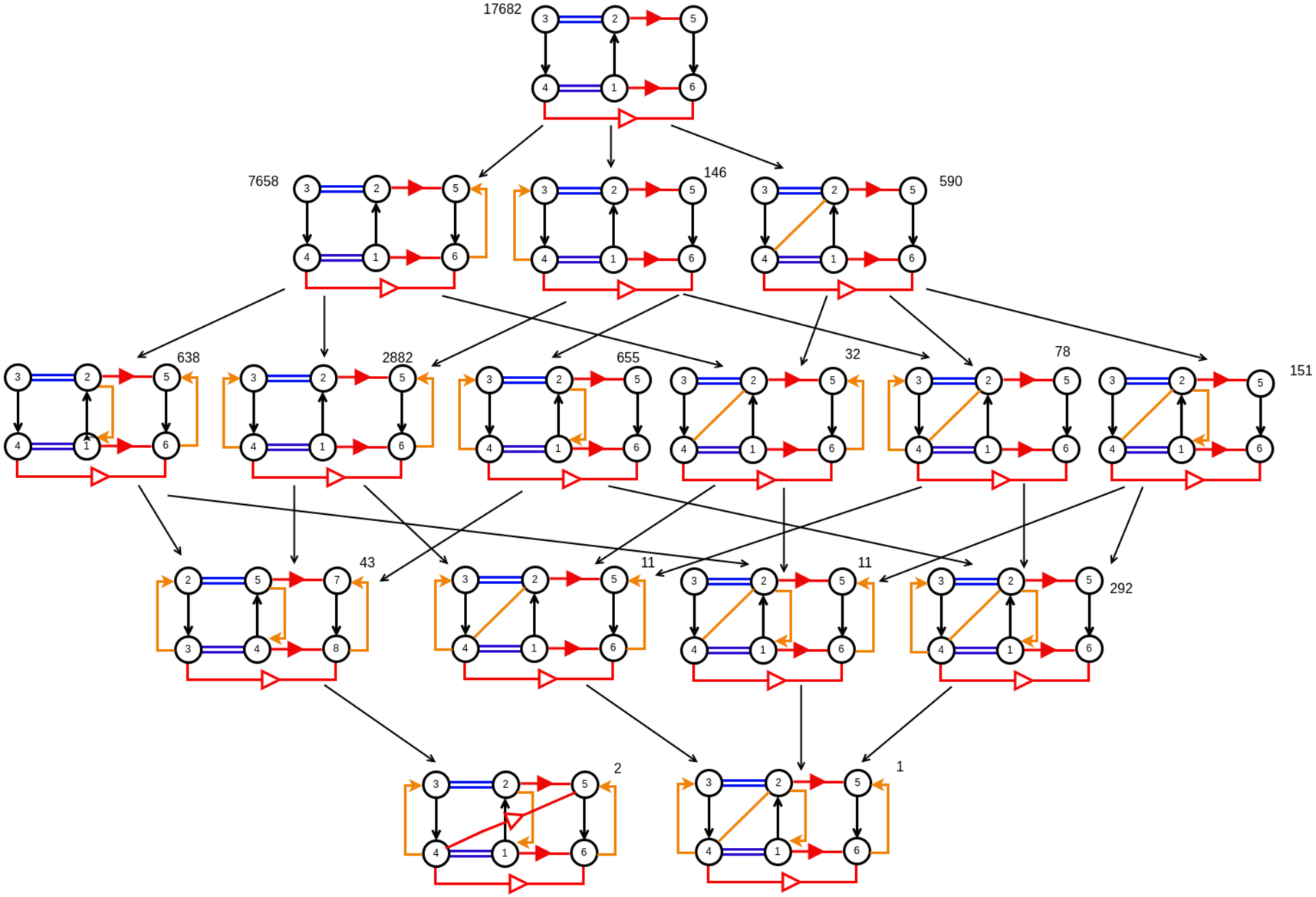
After some time and some C02 (but no GPU)

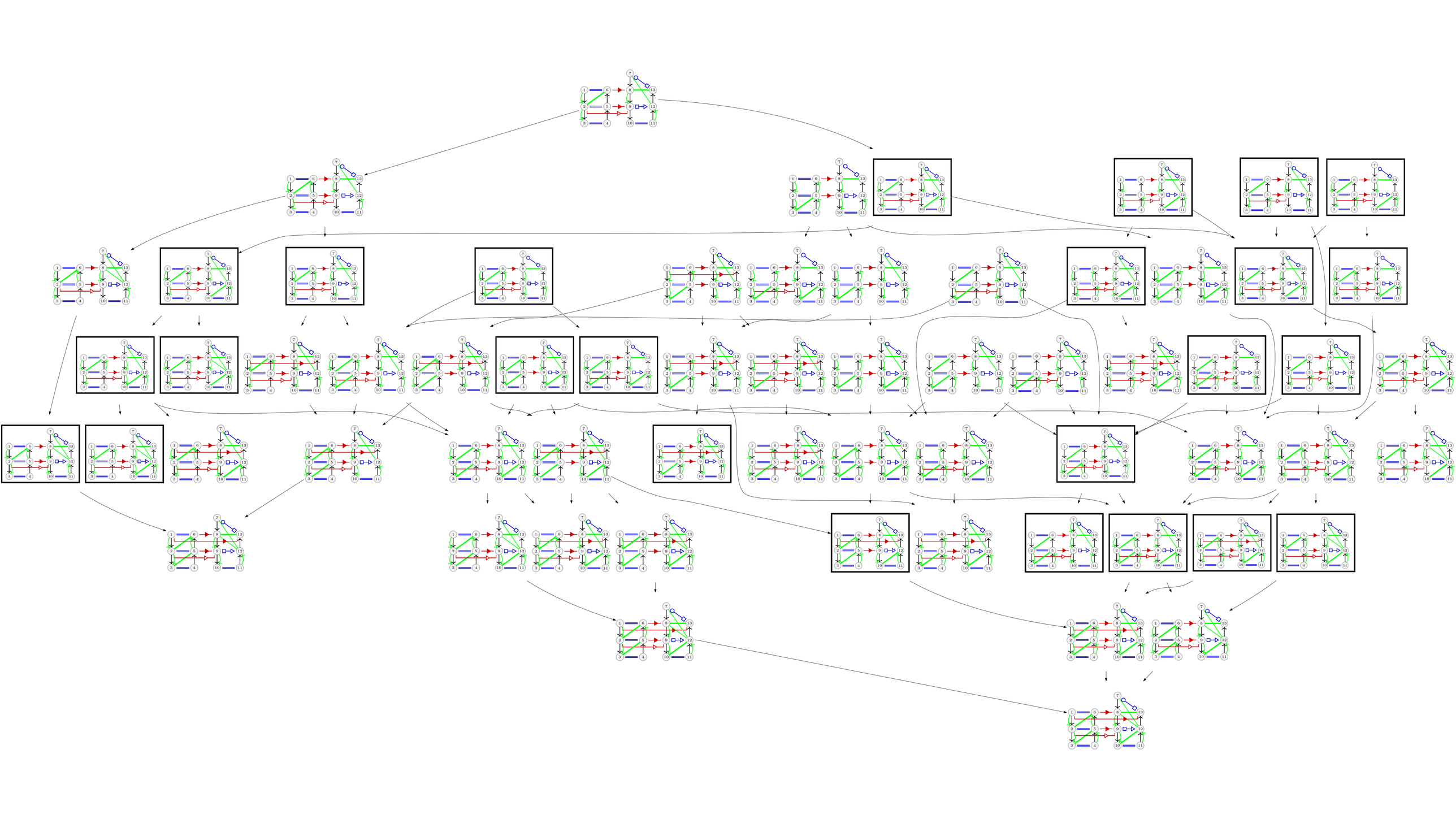
- ComputeCanada cluster -> 3200 processors on 70 nodes each with 128GB RAM for 4 weeks

- Found 157 344 RINs
  209 750 474 occurrences



A-minor with stackings variations

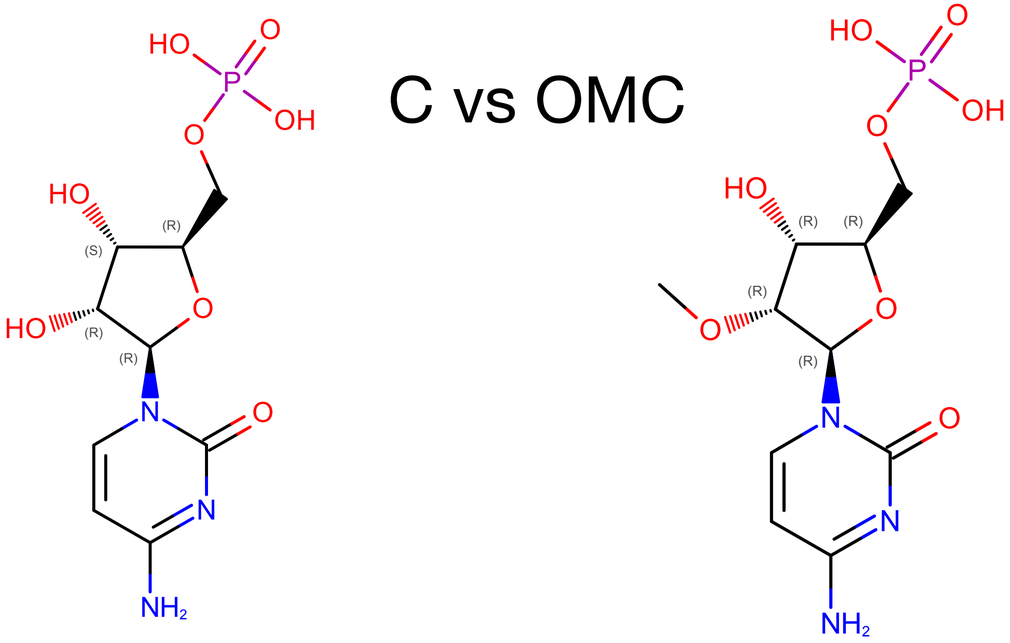




340 chemically modified nucleotides

fr3D-python annotates non-canonical for chemically modified

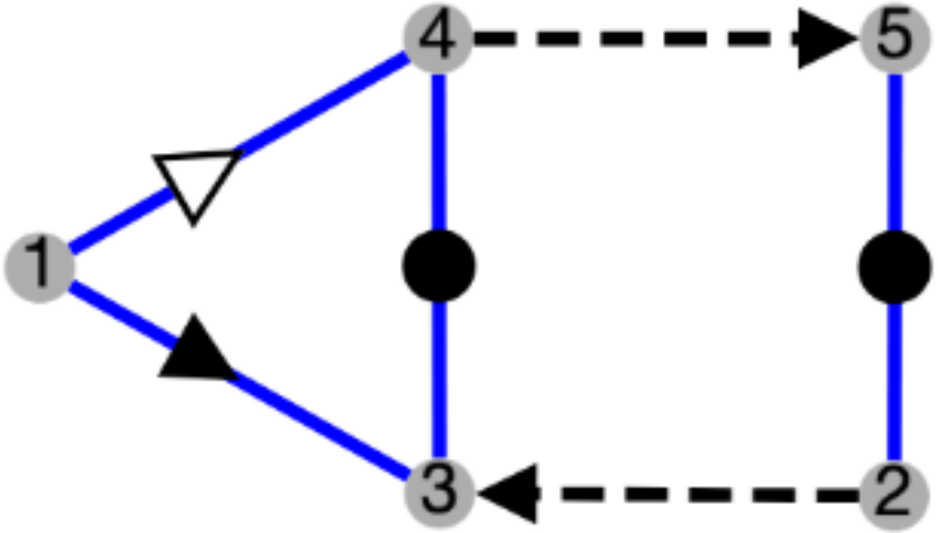
github.com/BGSU-RNA/fr3d-python

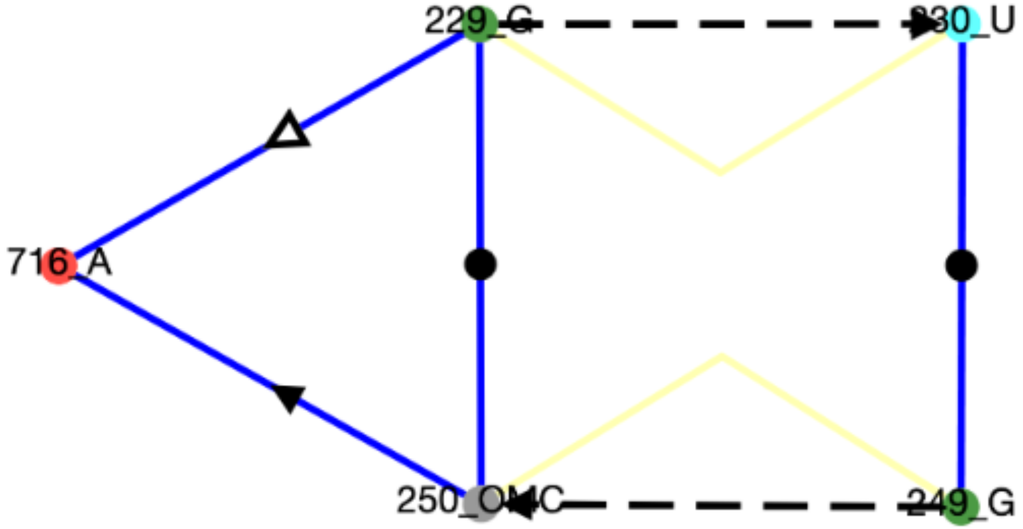


C vs OMC

Thanks Craig!

Going back to sequence: 140 chemically modified nucleotides annotated in RINs

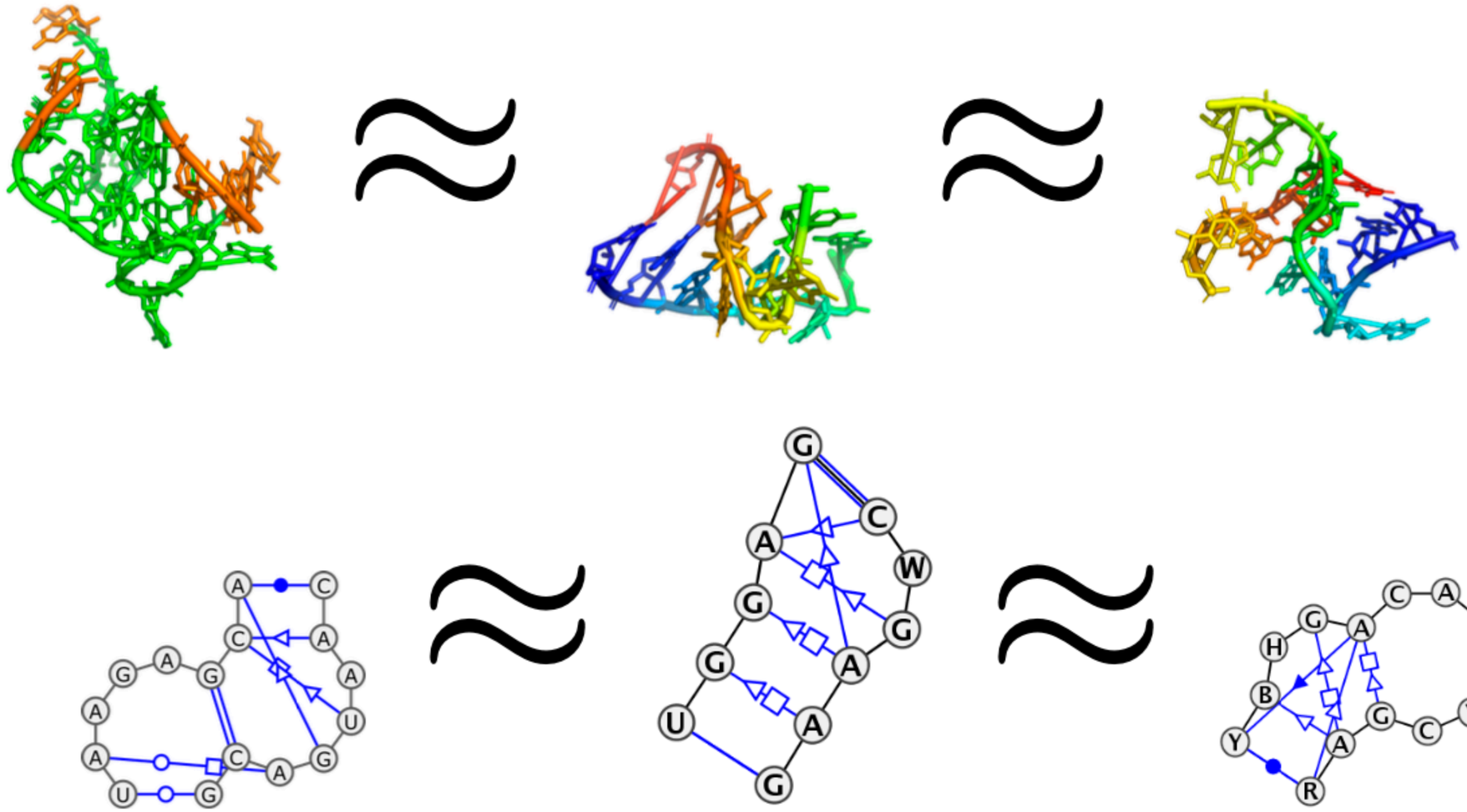




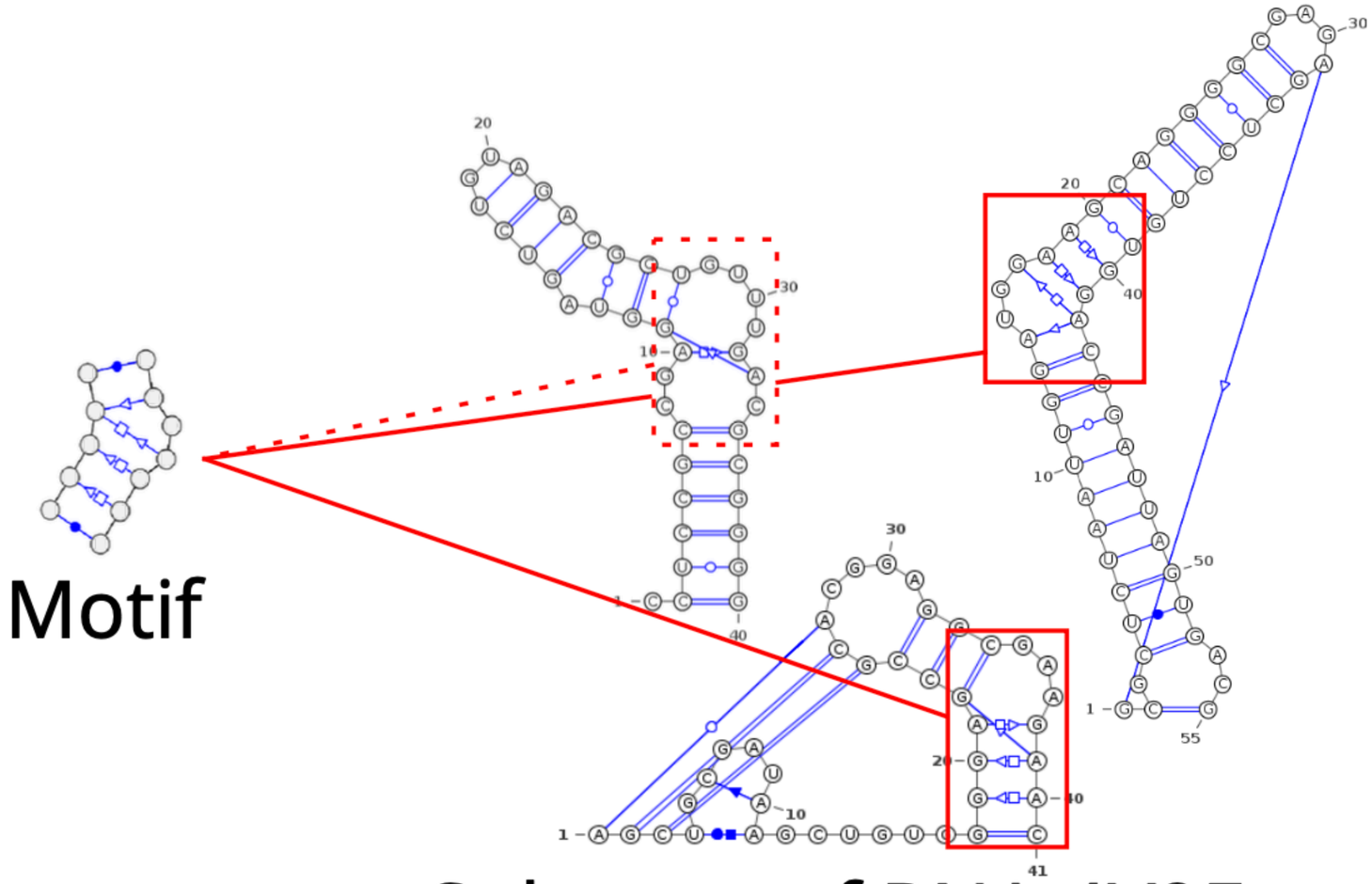
Explore the data
carnaval.cbe.uqam.ca
(alpha version no stacks)

- Quickly presented in visualisation session (in 6h)
- All Recurrent Interaction Networks, sequences found in PDB
- Search any basepairs pattern
- Find chemical modifications and associated Recurrent Interaction Networks

Back to 3D reality (which is non isomorphic)
3 kink-turn from RNA3DHub

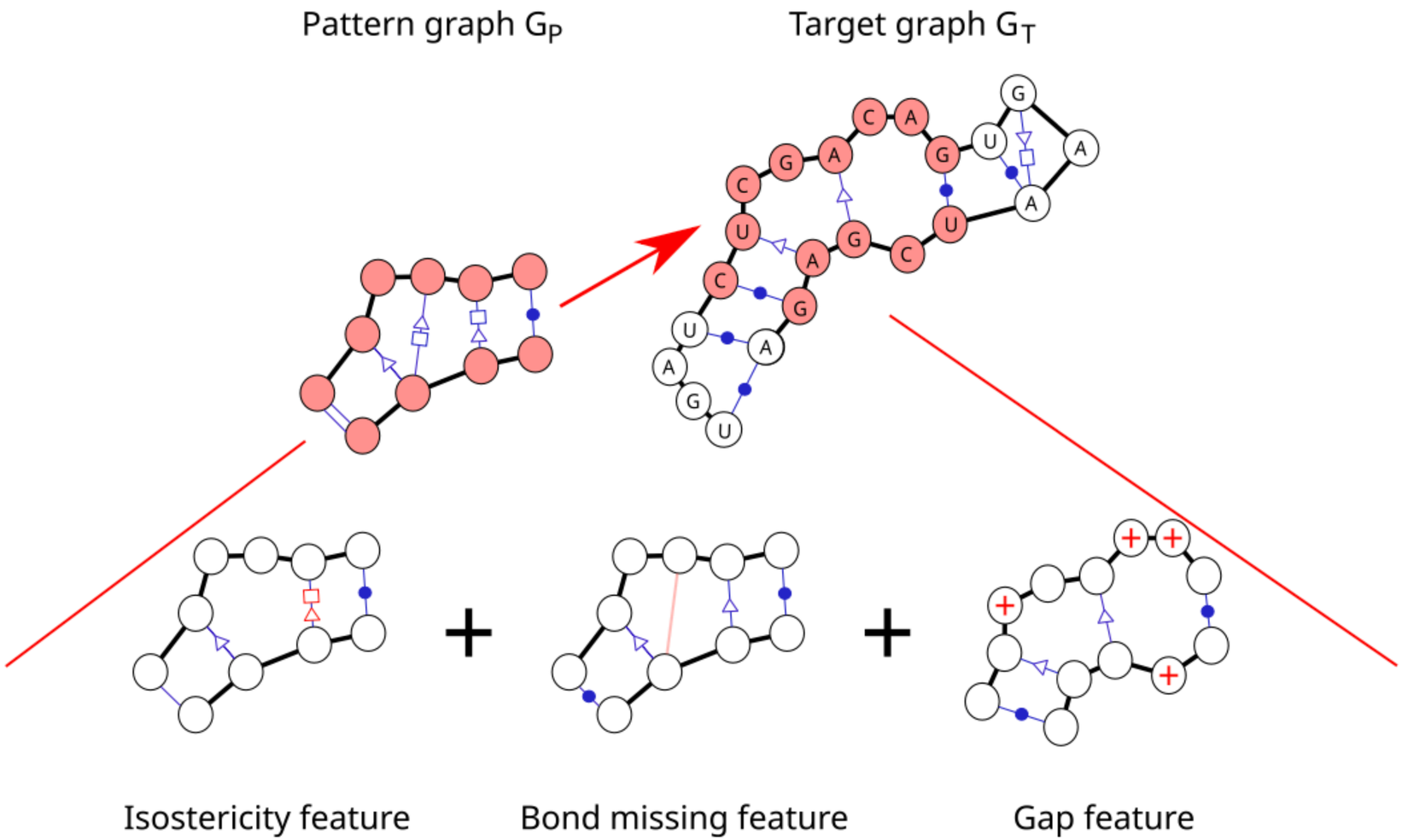


Can we do fuzzy graph matching?



FuzzTree

github.com/theoboury/FuzzTree (wabi 2024)



Théo Boury

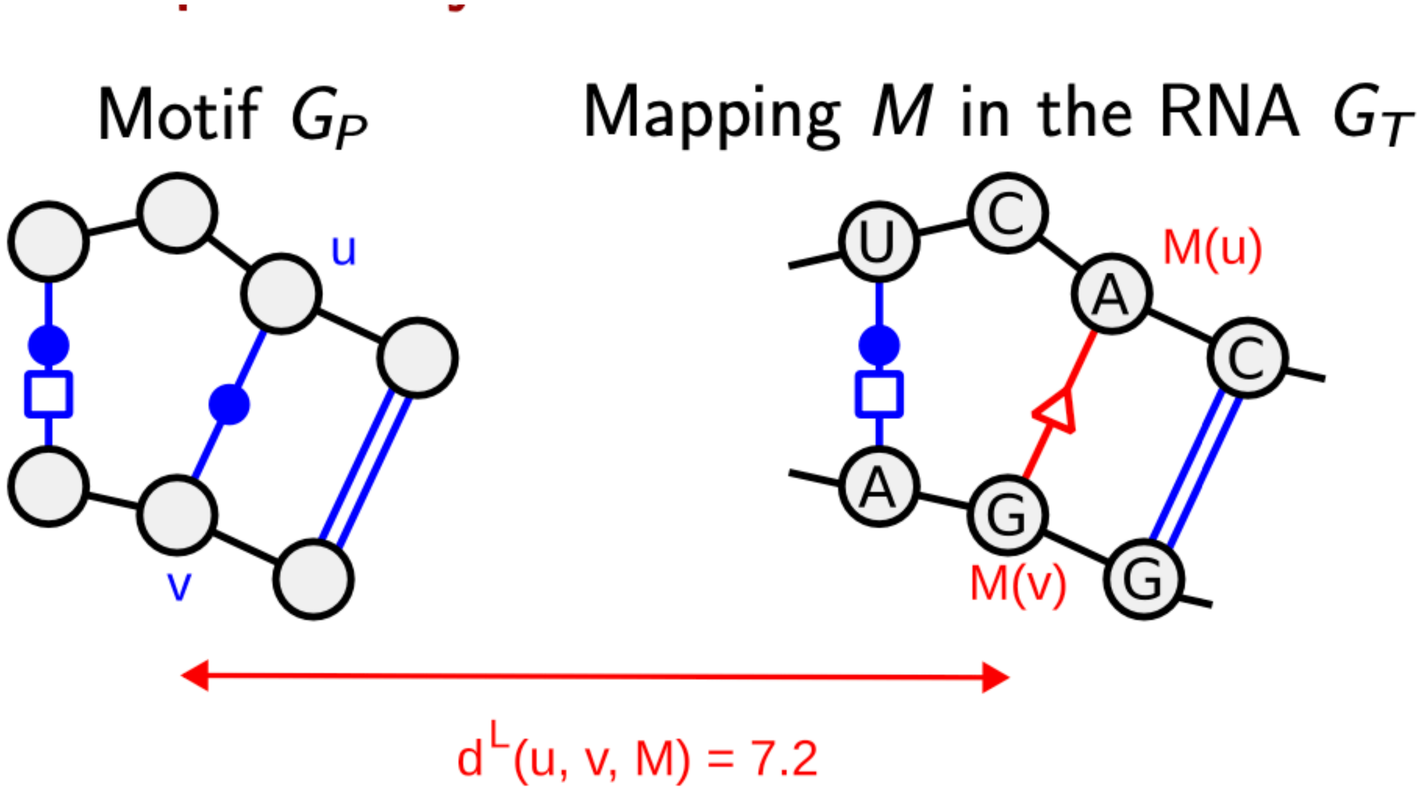
Boltzmann sampling of graphs

For each graph mapping define an energy:

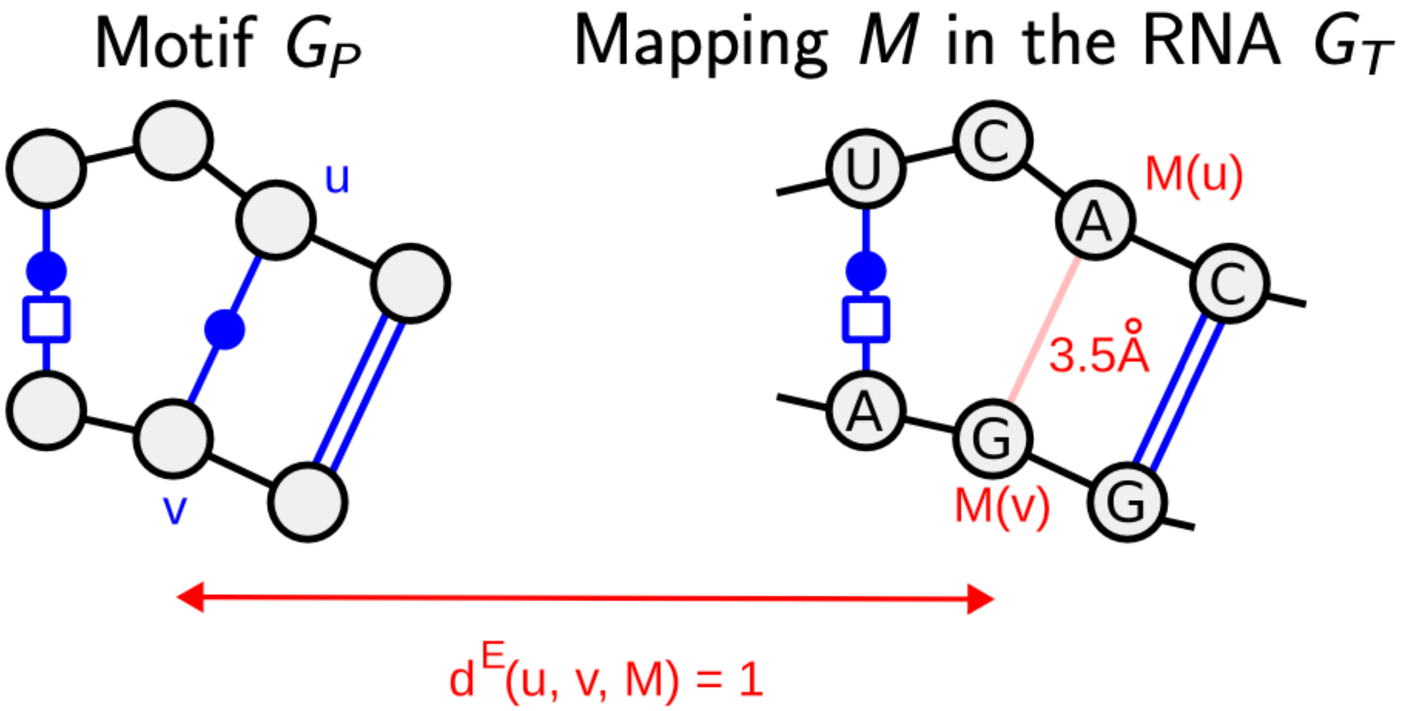
$$E(M) = \sum_{(u,v)\in E_P} \underbrace{w_L \times d^L(u,v,M)}_{\text{BP isostericity}} + \underbrace{w_E \times d^E(u,v,M)}_{\text{Missing edges}} + \underbrace{w_G \times d^G(u,v,M)}_{\text{Inserted nodes}}$$

Probability of sampling a mapping: $\mathbb{P}(M) = \frac{e^{-E(M)}}{\mathcal{Z}}$ where $\mathcal{Z} = \sum_{M'} e^{-E(M')}$

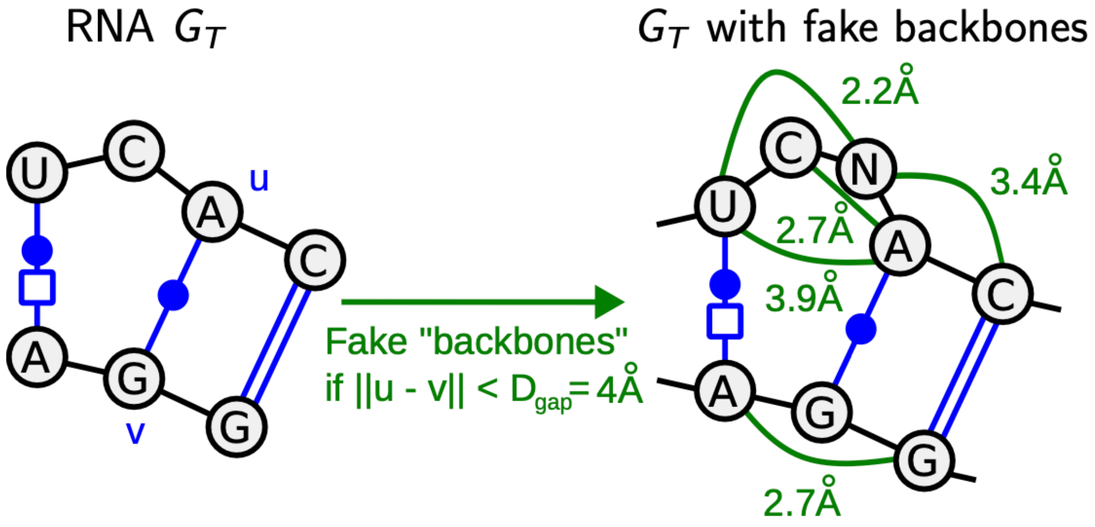
Isostericity distance (1/3)

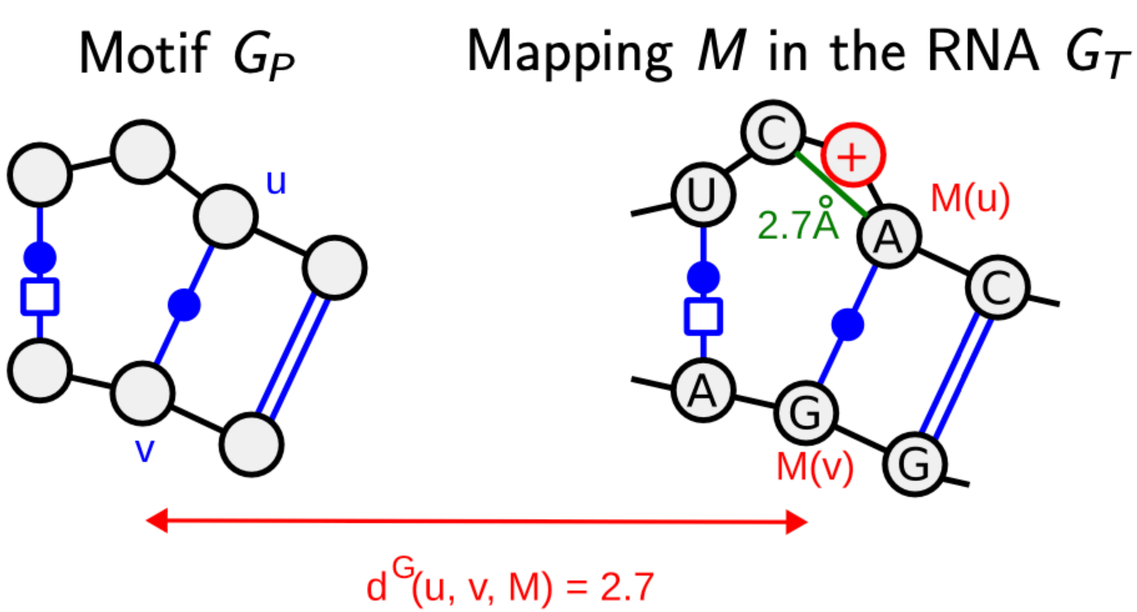


Missing edges (2/3)

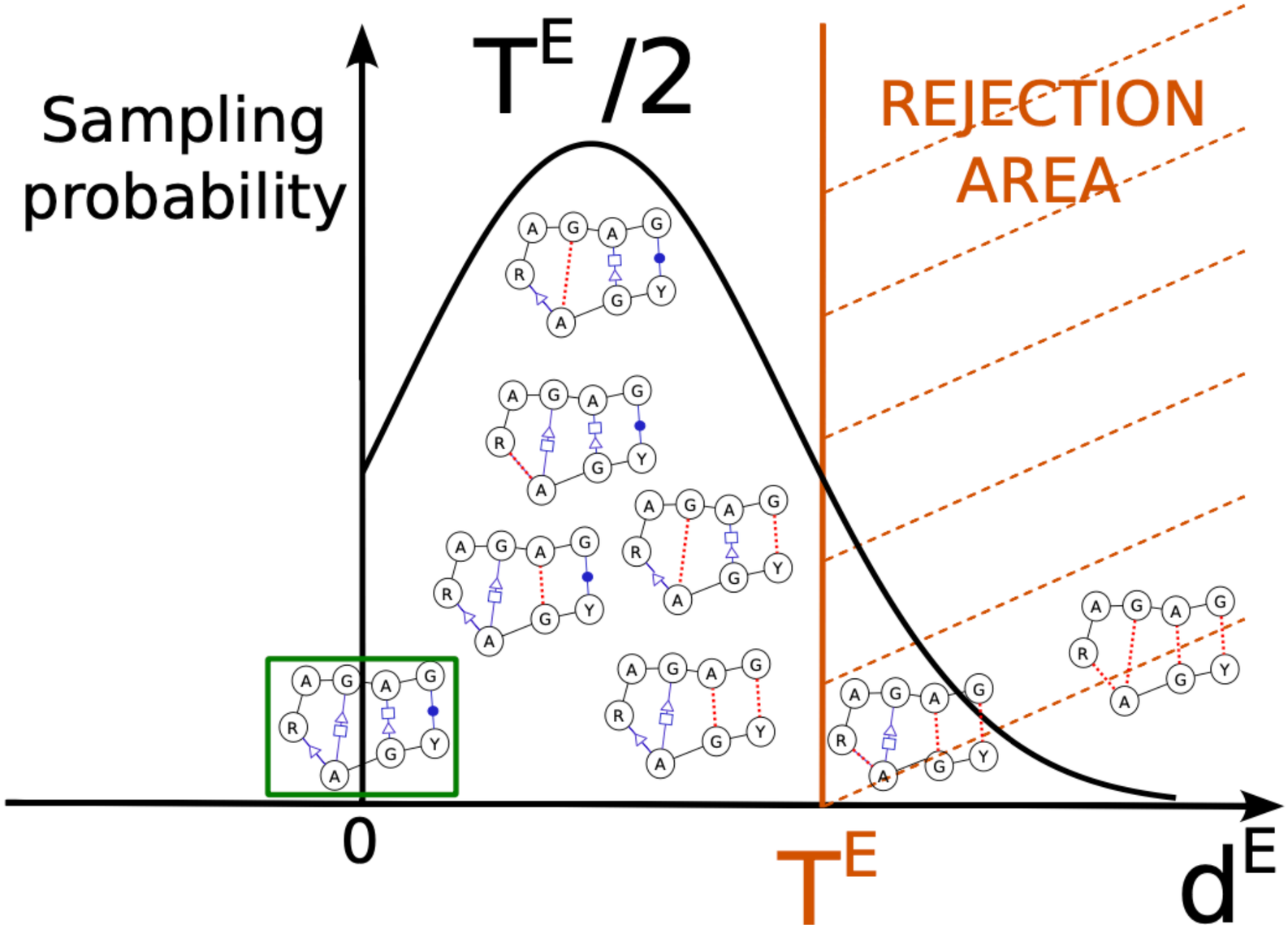


Inserted nodes (3/3)





Parametrizing the sampling



Complexity exponential in tree width of pattern

Computing partition function: $O(kn^{\phi+1})$

Sampling: $O(knt)$

n: nodes target

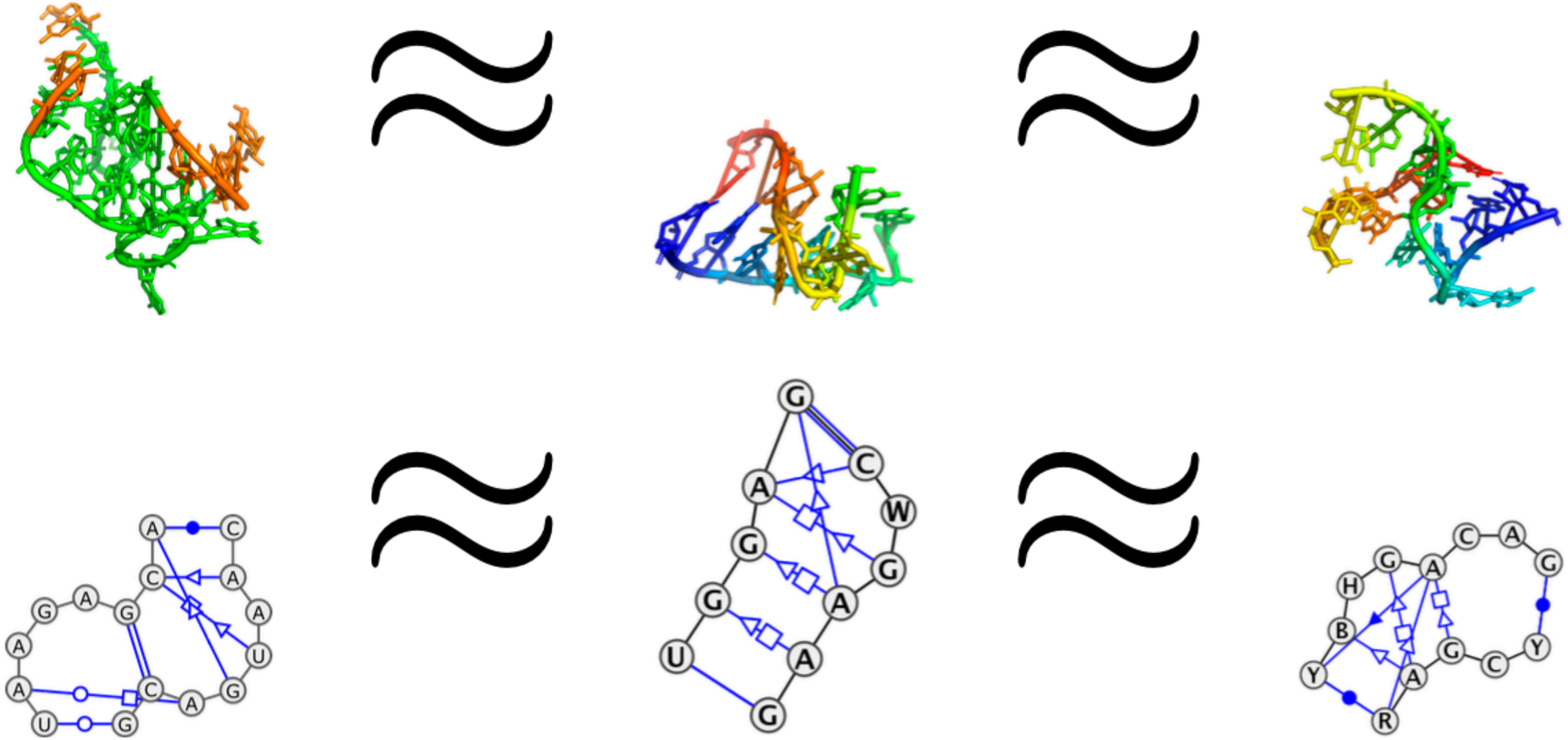
k: nodes pattern

ϕ: treewidth pattern

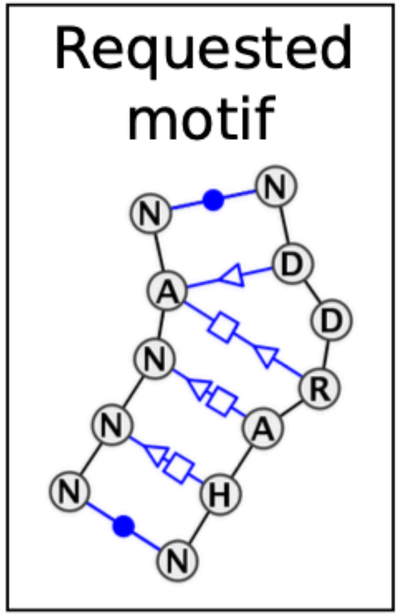
t: number samples

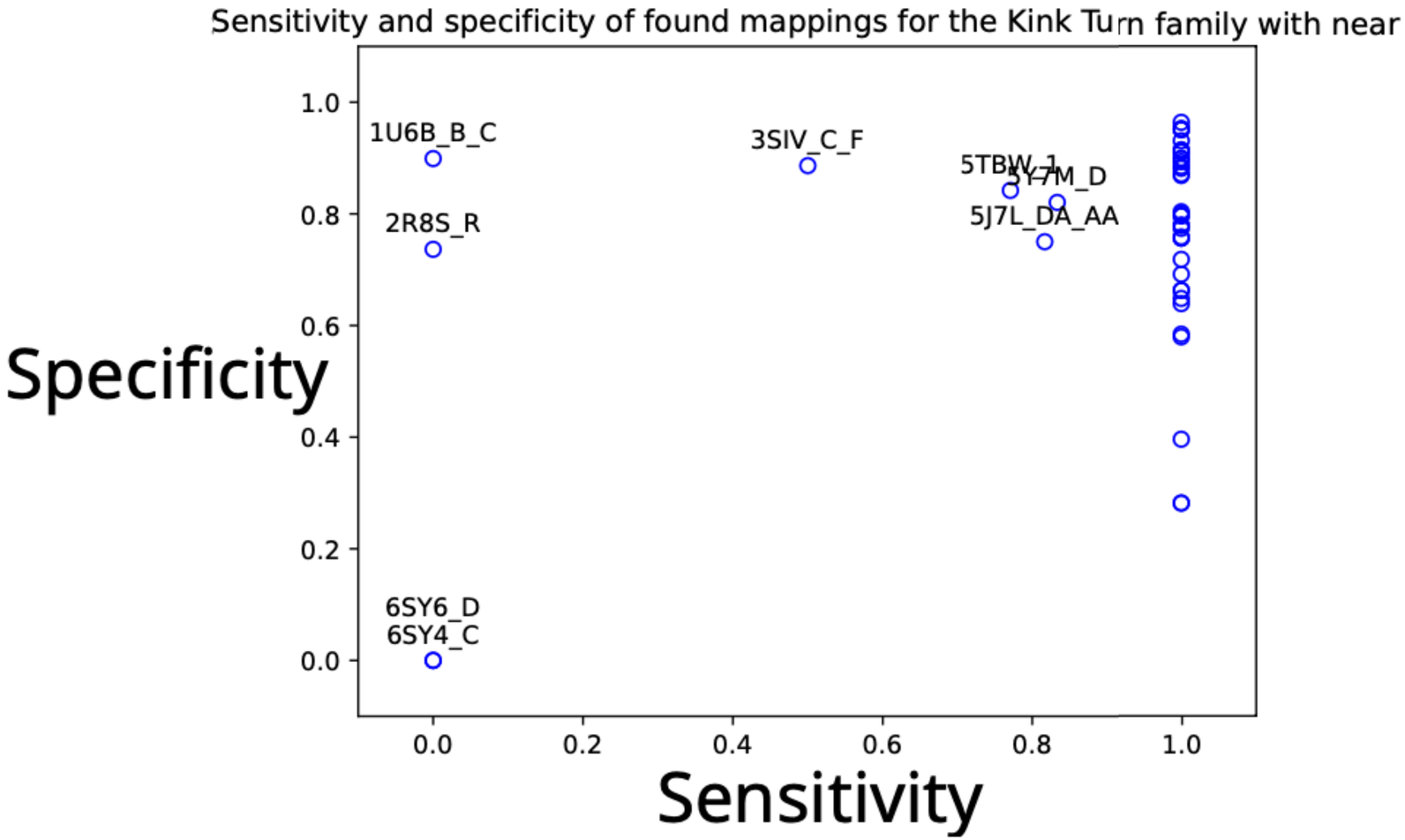
Back to kink-turn from RNA3DHub

- A biological family that contains 72 known motifs over more than 25 different RNAs.
- Kink-Turns are clustered in 18 different families according to atomic cristallography. [4]

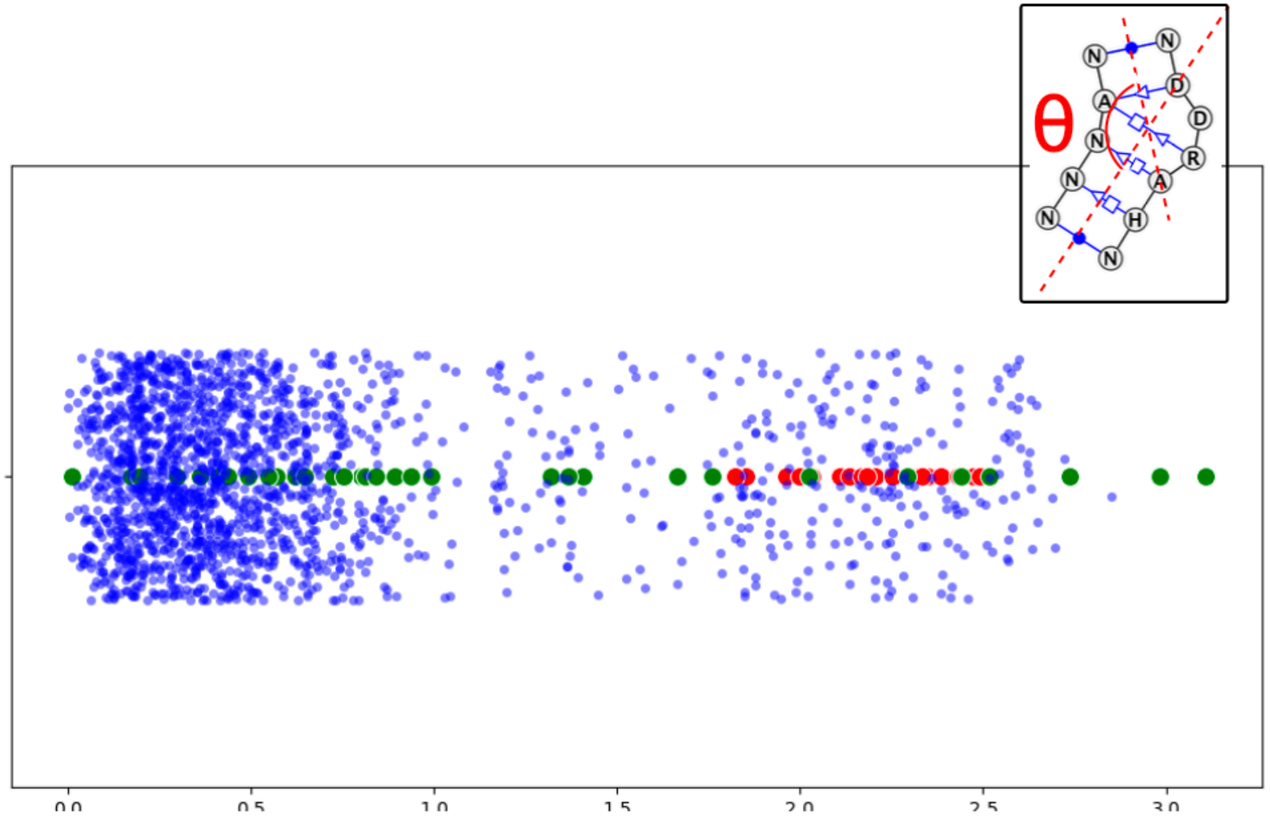


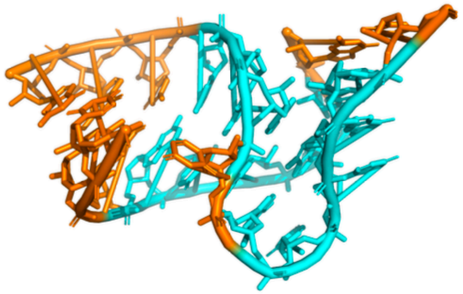
From one pattern we find most occurrences

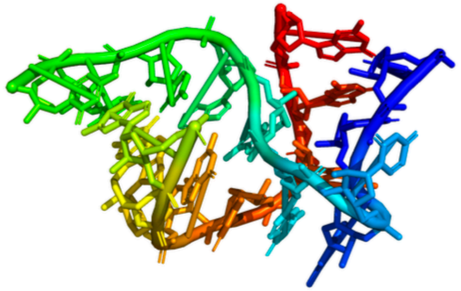


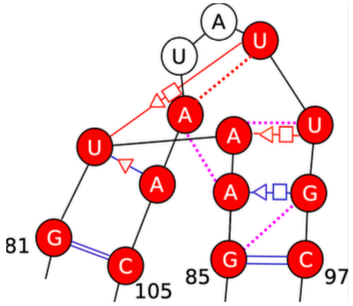


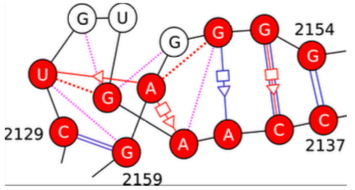
New candidates appear



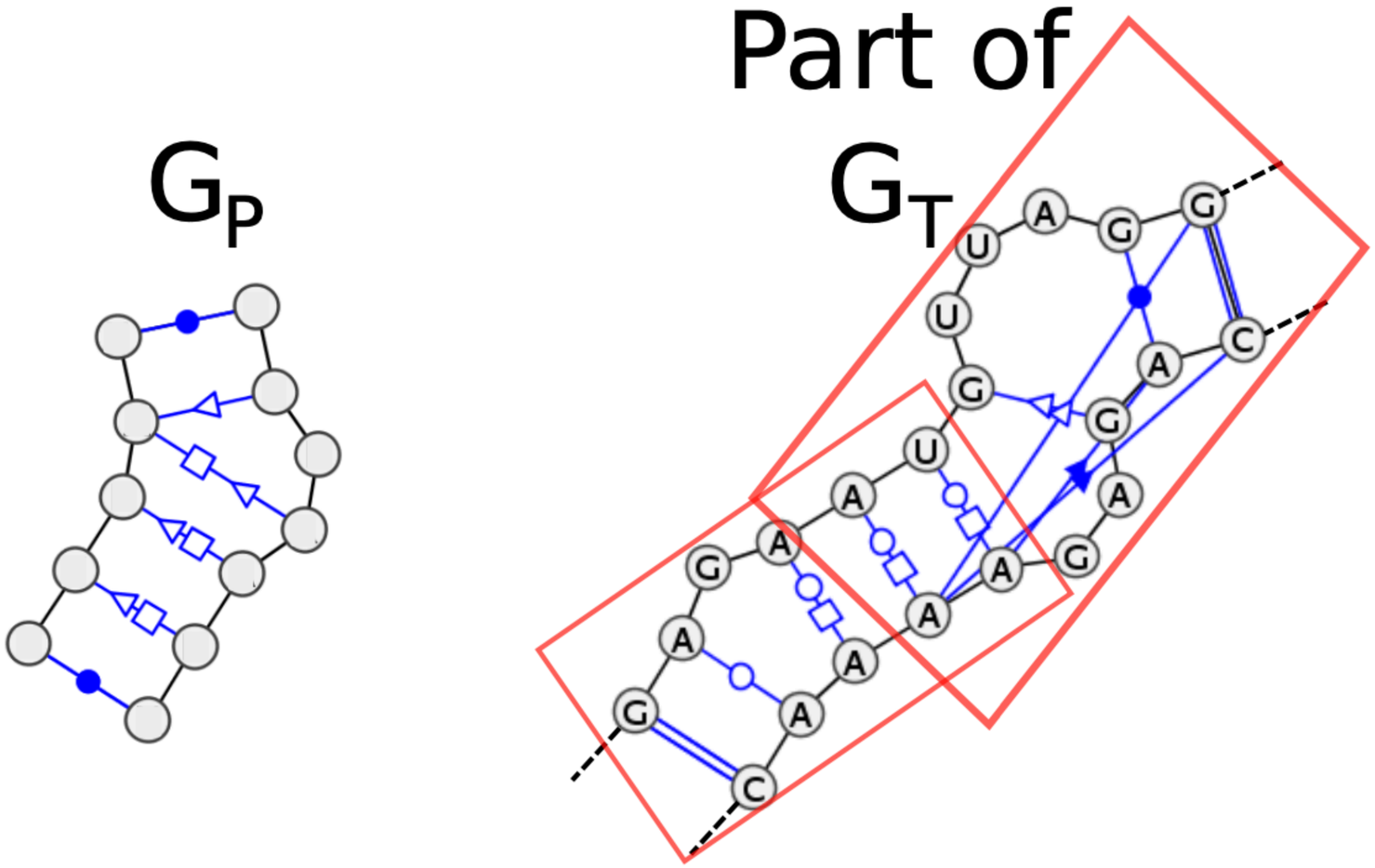




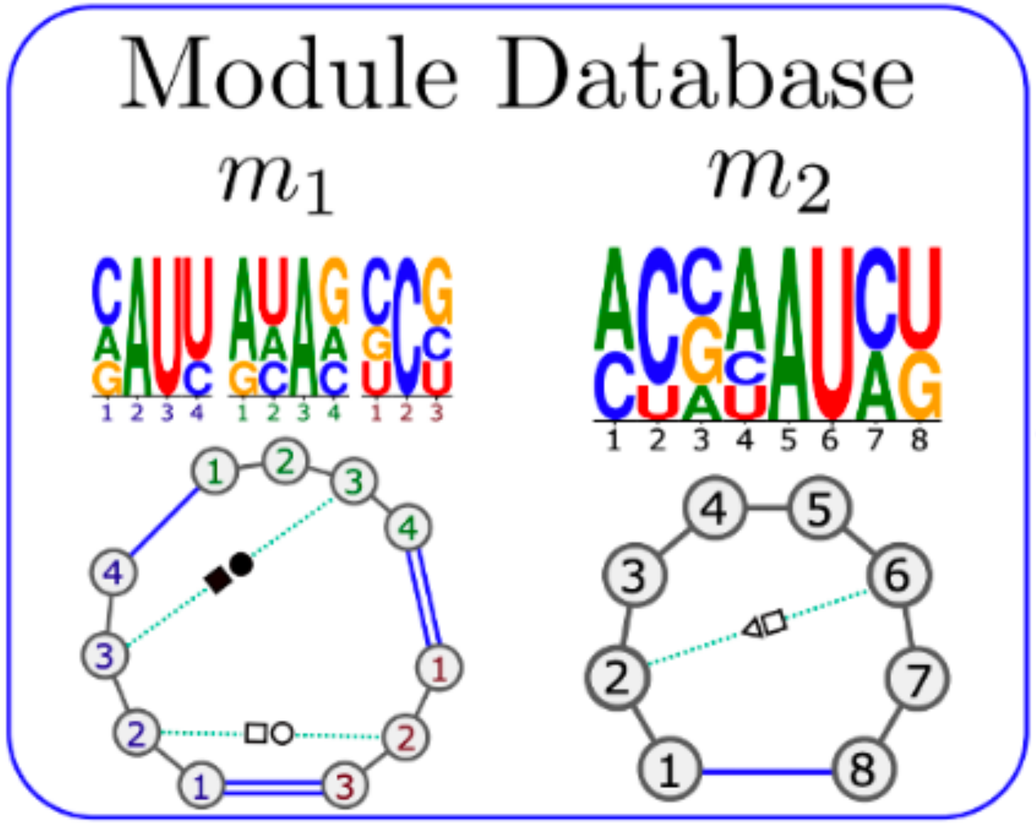




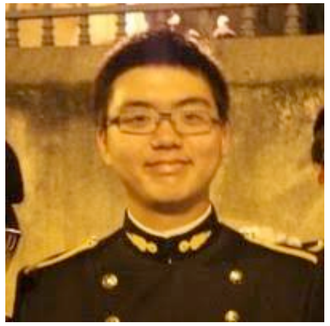
Designing core patterns a challenge
A kink-turn that can not be found



Integrating sequences to predict modules
BayesPairing2
jwgitlab.cs.mcgill.ca/sarrazin/rnabayespairing2 RECOMB2020

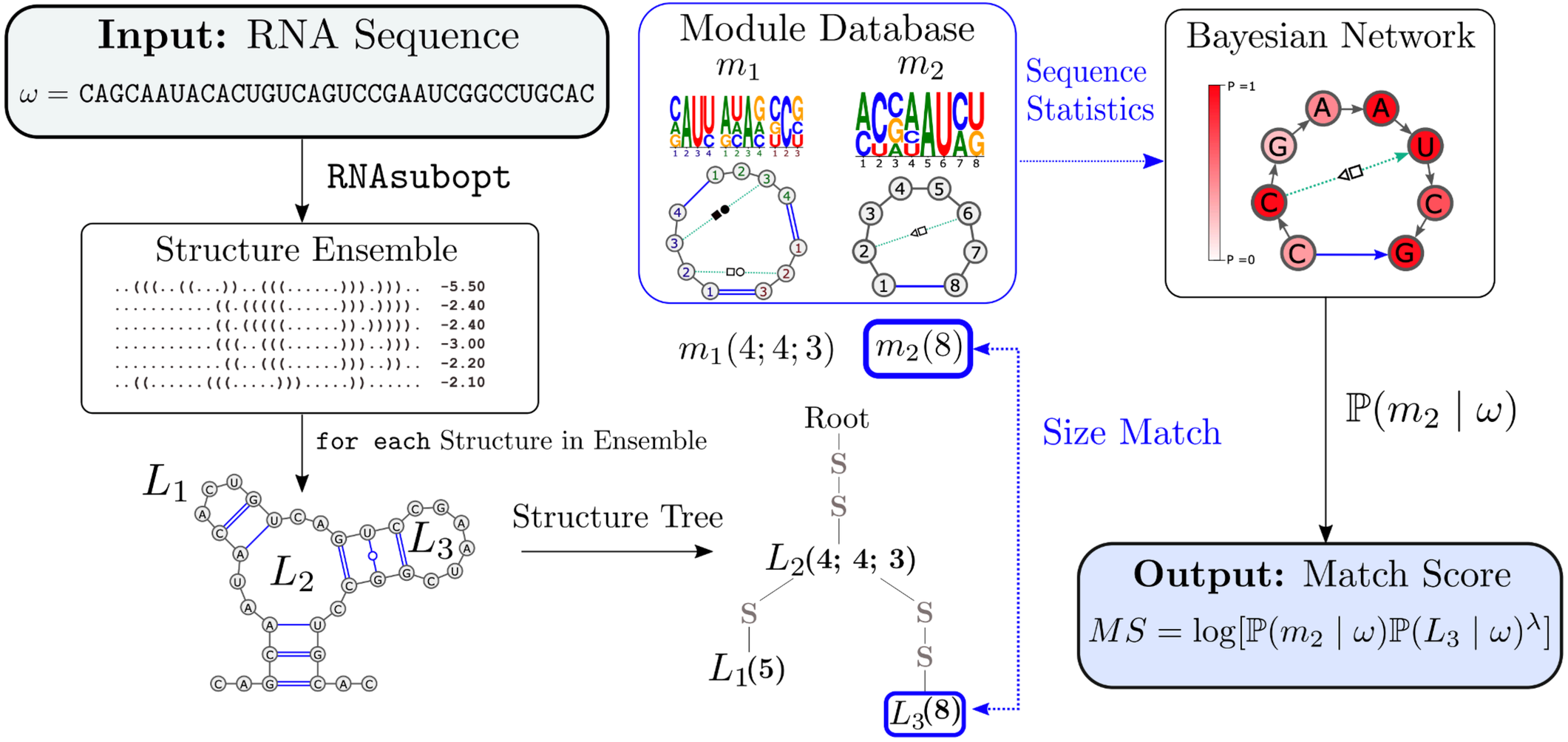


Roman Sarrazin-Gendron



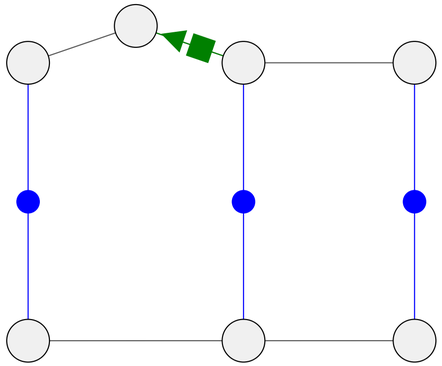
Hua-Ting Yao
(in the room, not in uniform)

Reality hits again, modules must be combined with loops

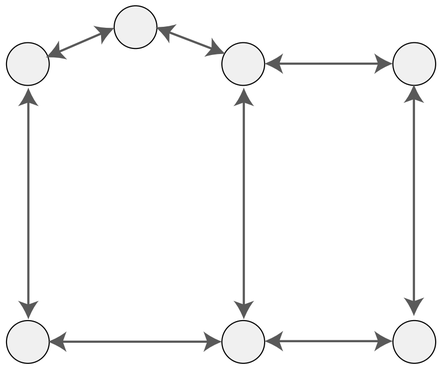


CantaLOOPS (work in progress)

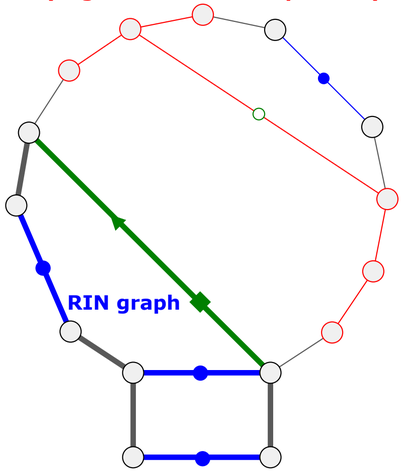
Recurrent Interaction Network



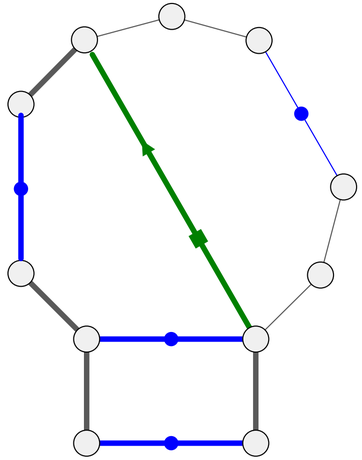
RIN graph representation



complementary region
(edges in red are compressed)



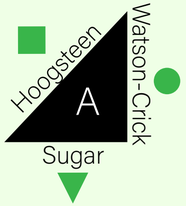
RNA structure loop that includes the RIN



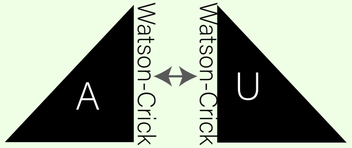
Simplified model of the loop

Leontis-Westhof base pair nomenclature

RNA bases have three edges



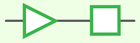
Base pair geometry can be summarized by interacting edges and orientation



cis Watson-Crick - Watson-Crick base pair (cWW)

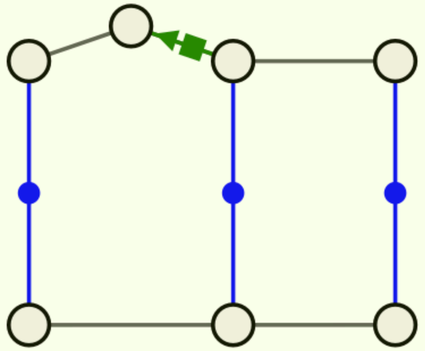
Twelve types of base pairs exist
Symbols show edge type, and fill shows orientation

trans Sugar-Hoogsteen base pair (tSH - tHS)



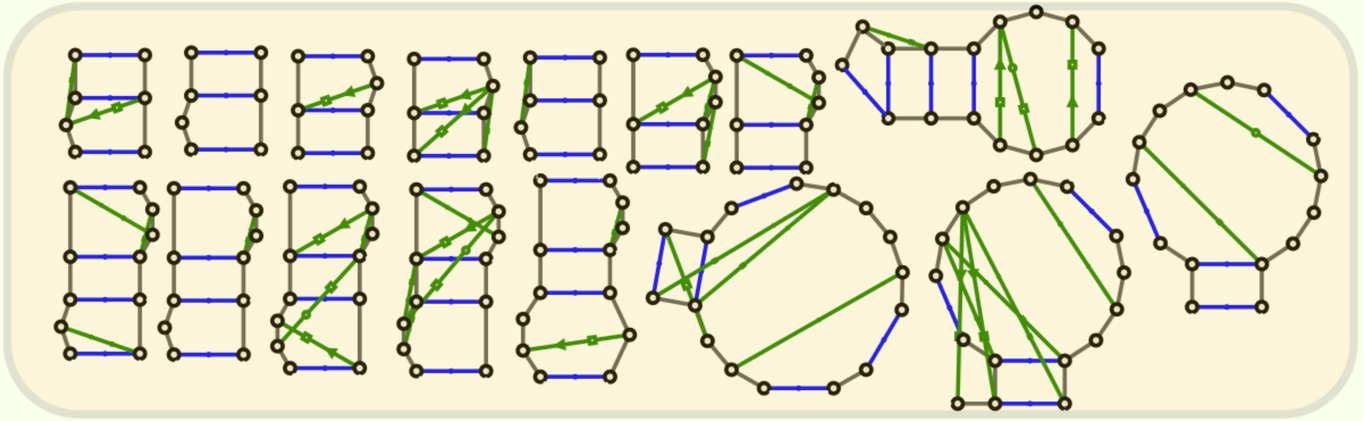
From the Recurrent Interaction Networks to the loops

RIN # 205

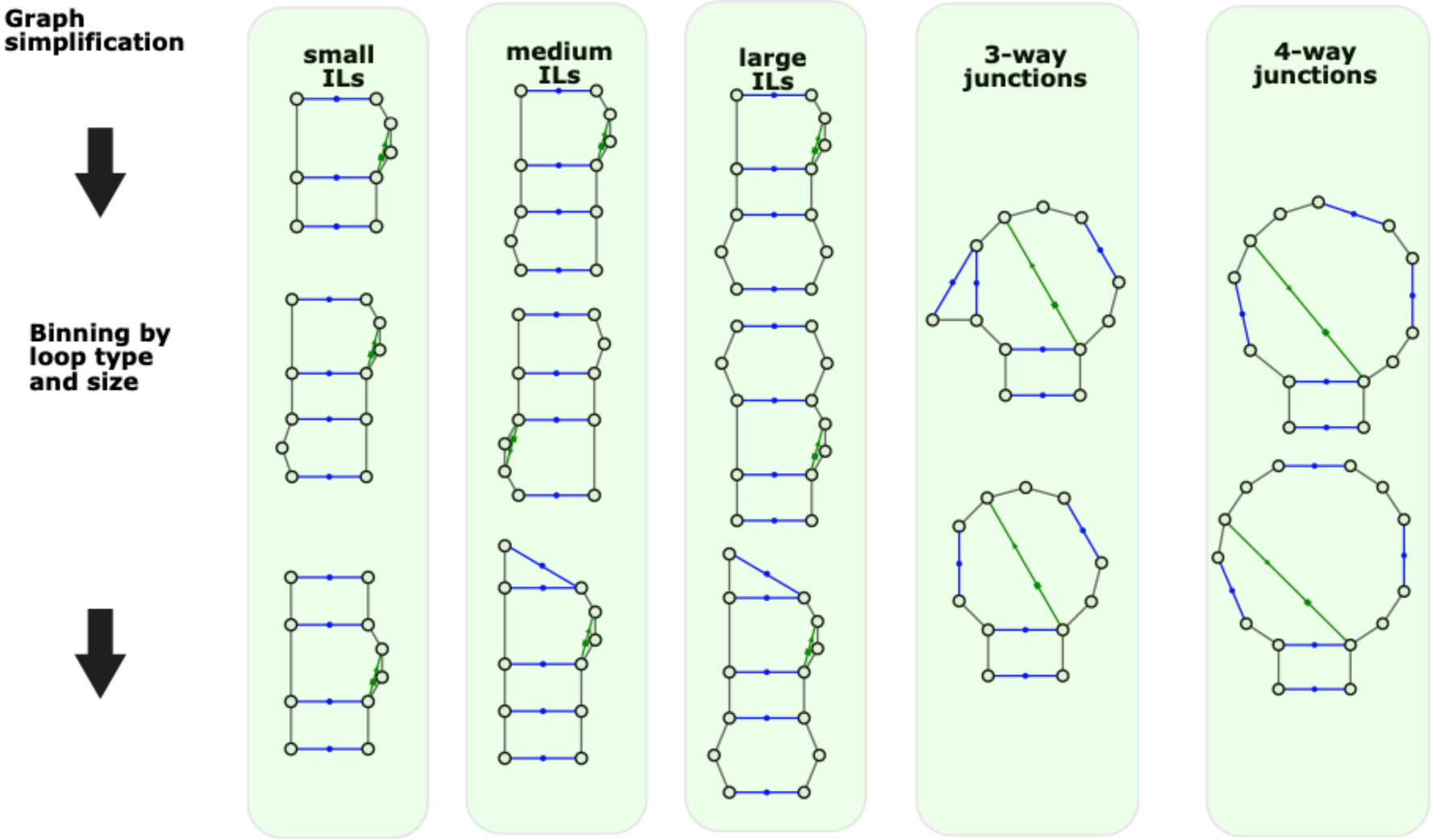


recurrent network

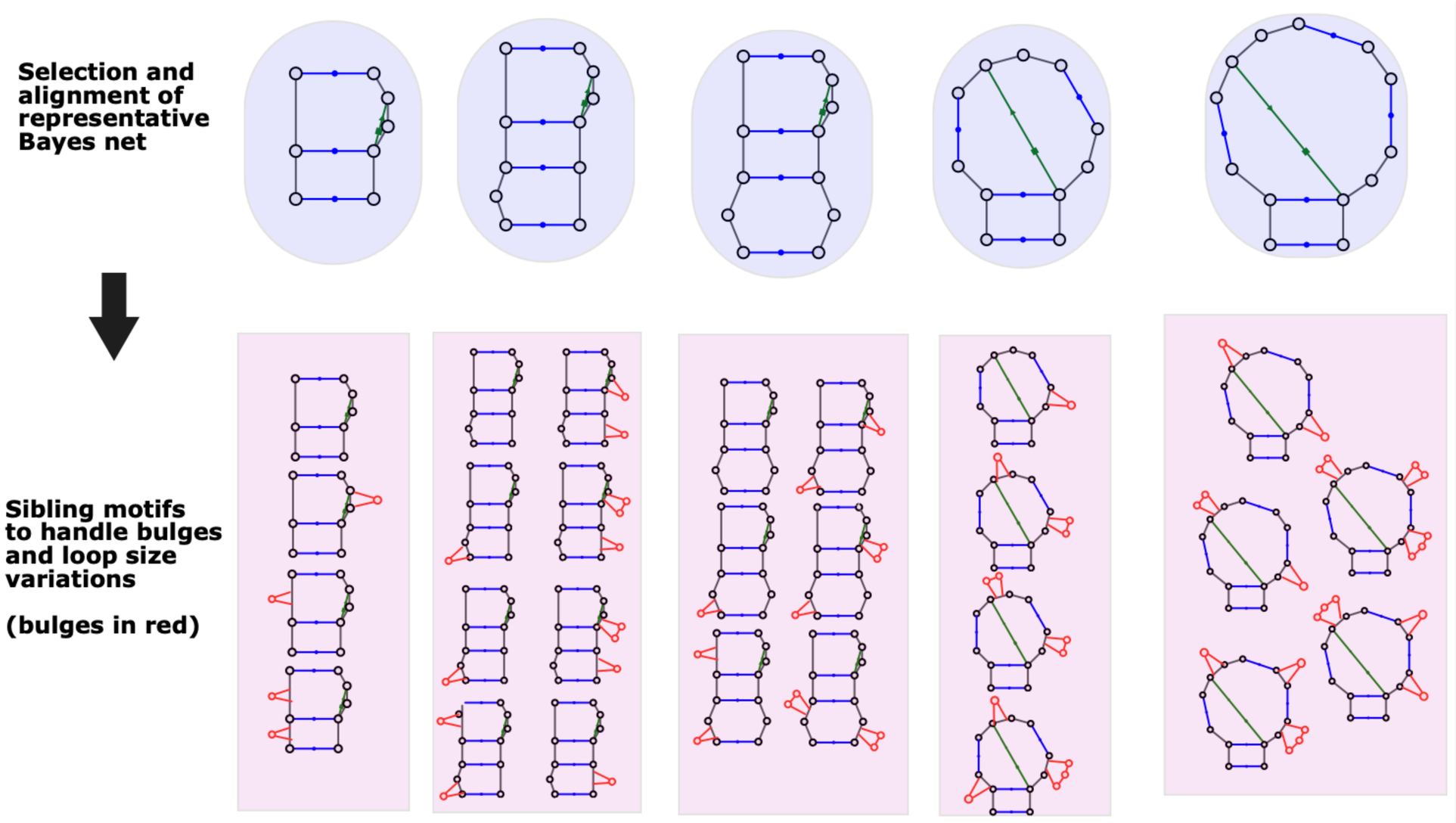
sample of loops containing network



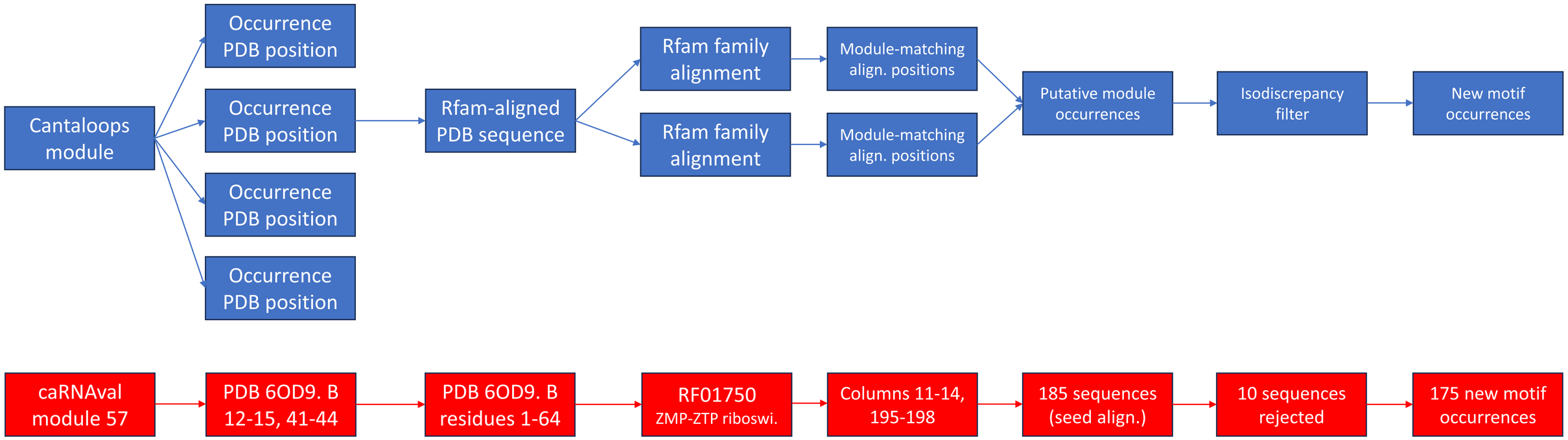
Clustering the real loops and their sequences



An ensemble of bayesian networks can be used to model a module

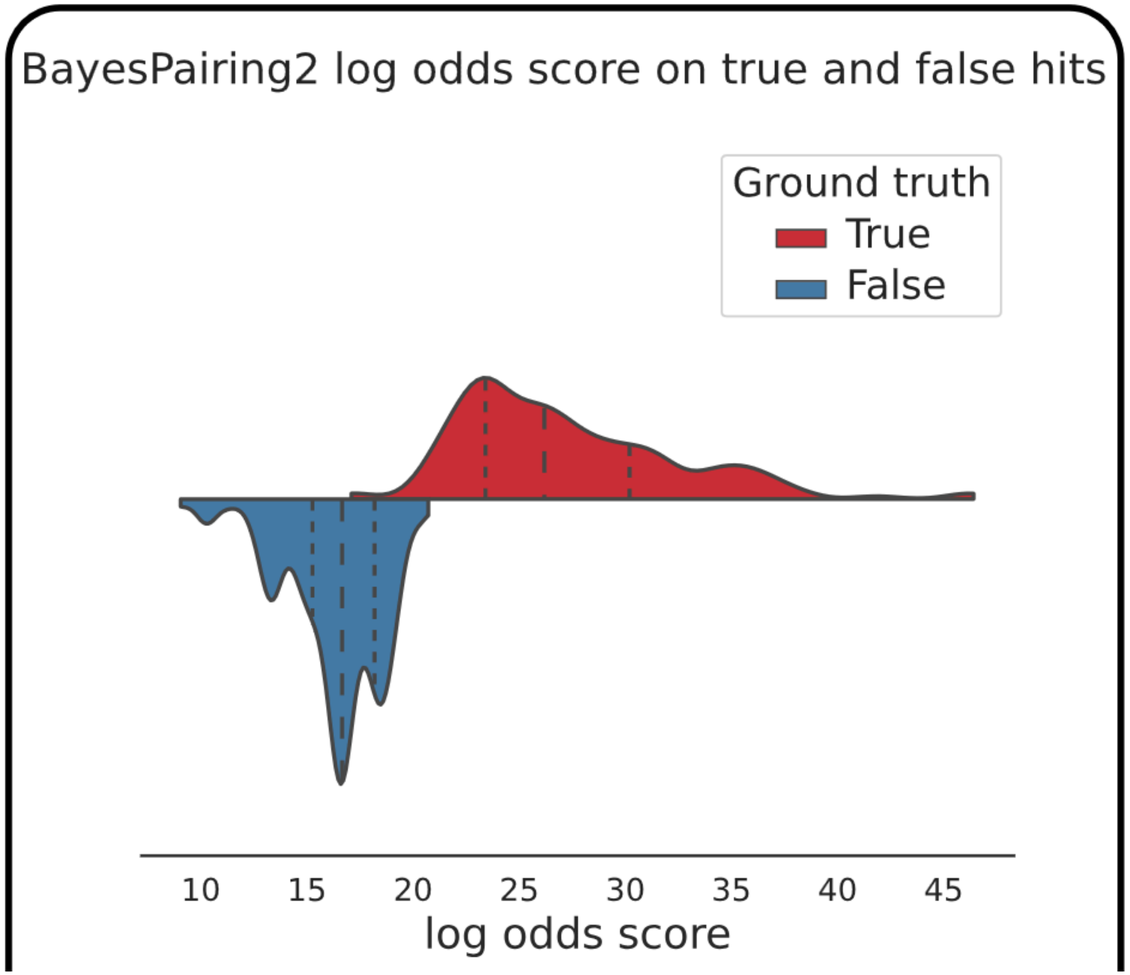


Populating with additional sequences (thanks Rfam!)

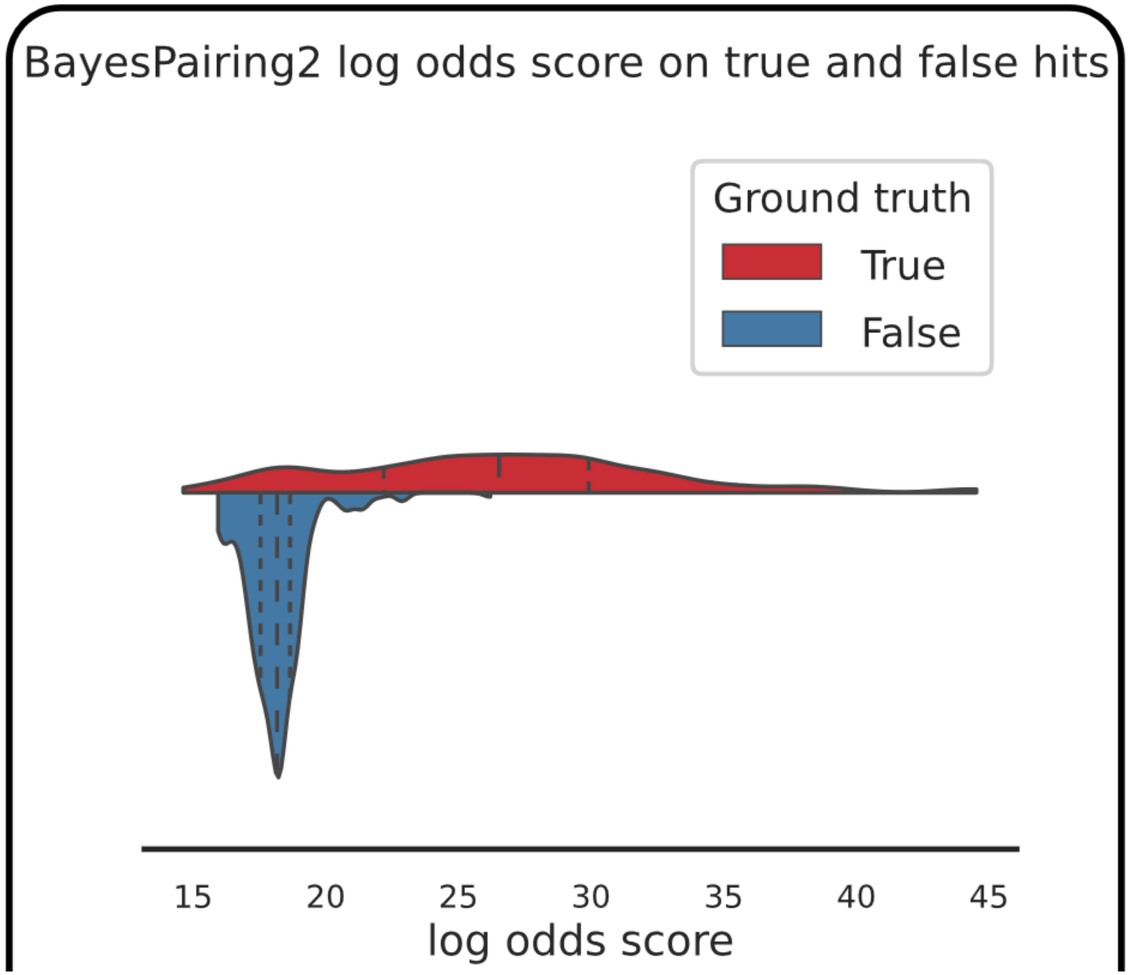


We can train and compare with the RNA3Dmotif atlas

3D motif atlas

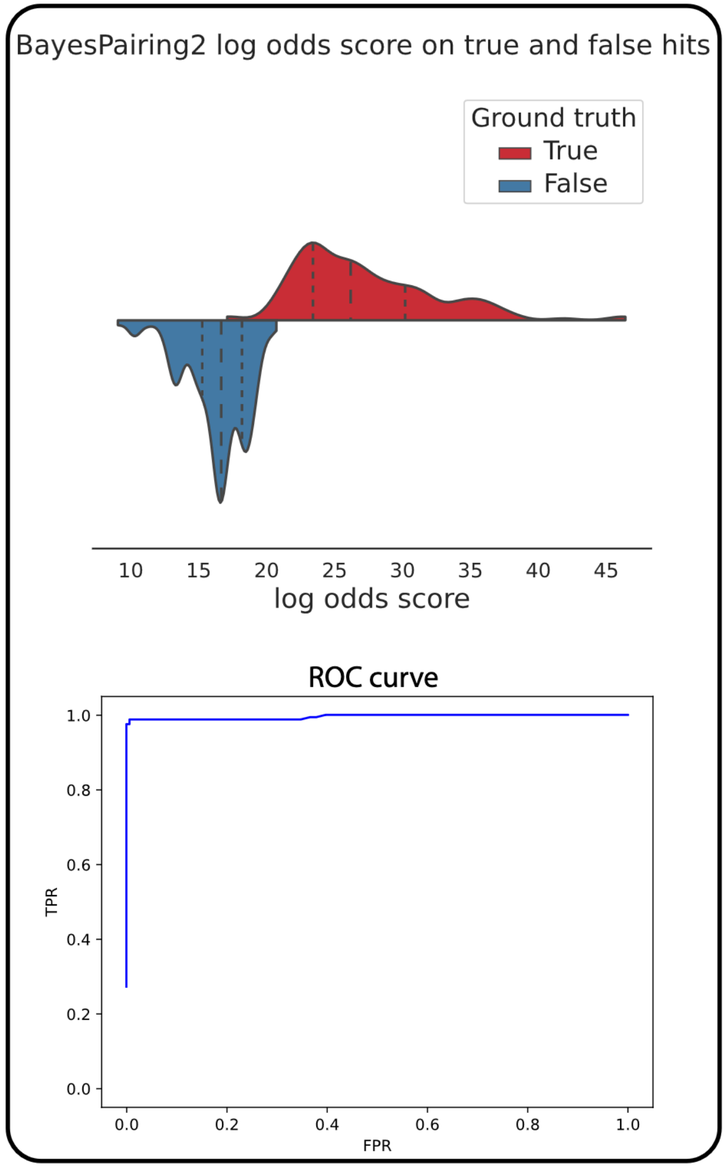


caRNAval

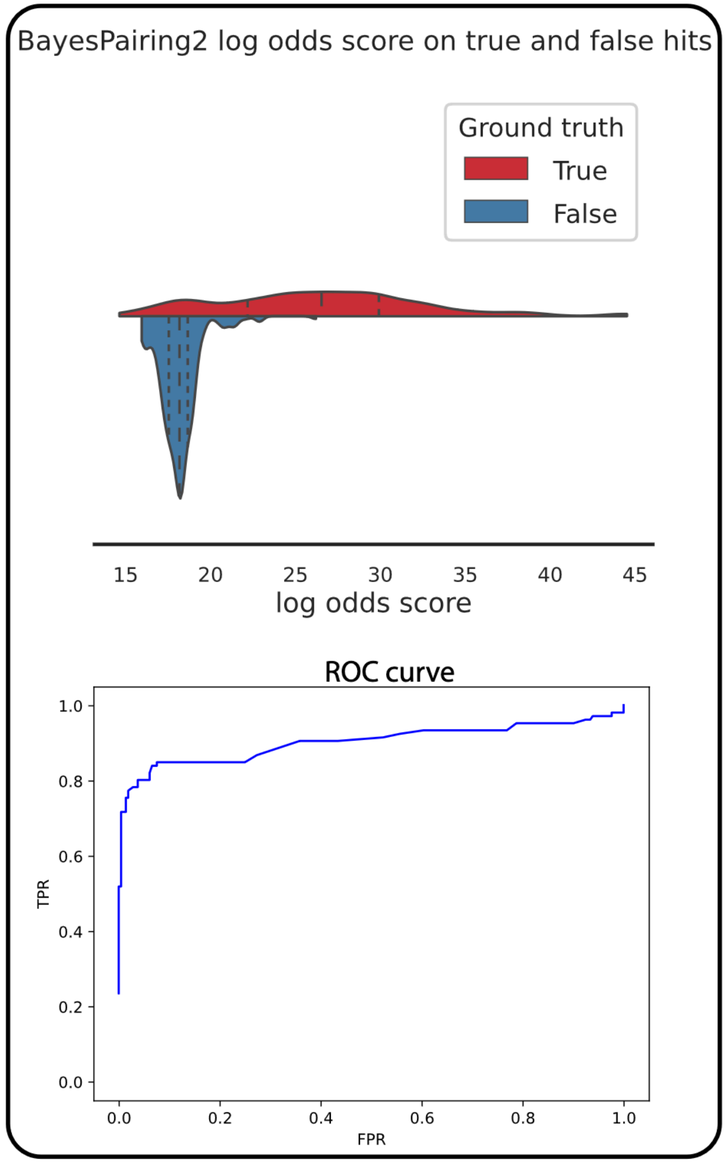


ROC curves

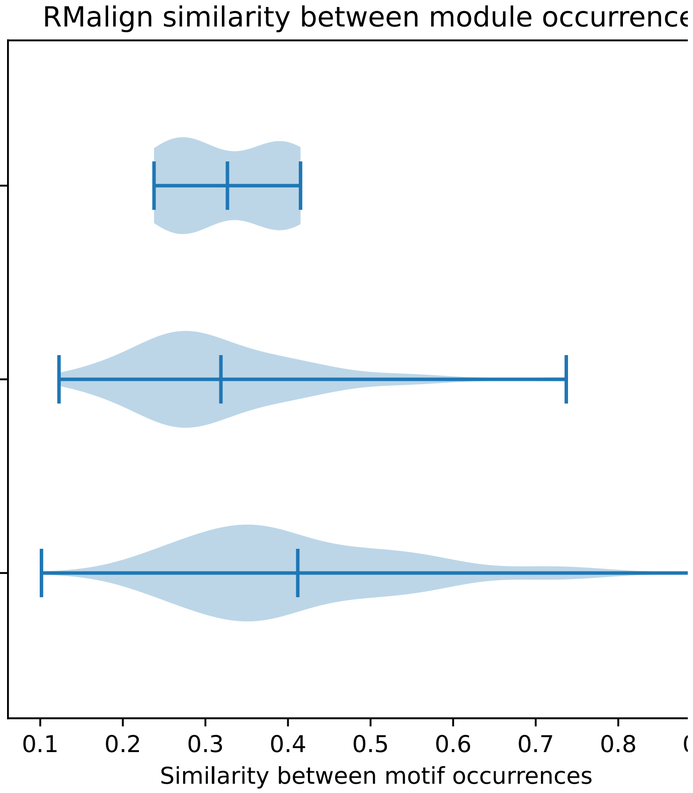
3D motif atlas

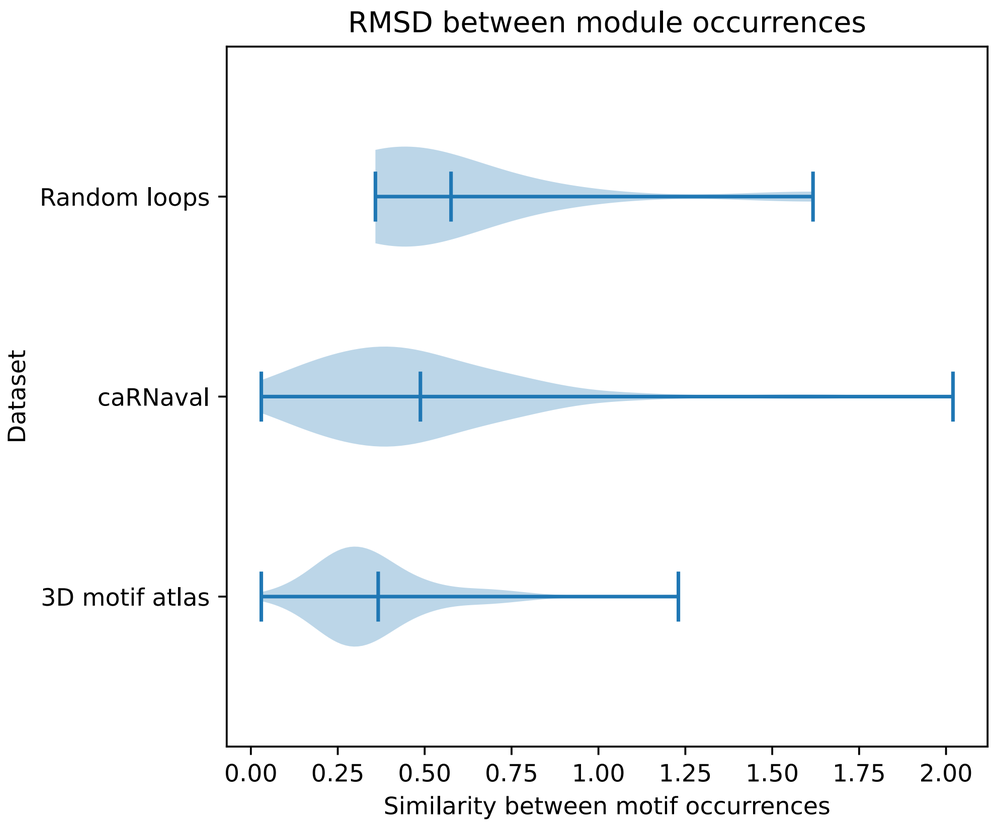


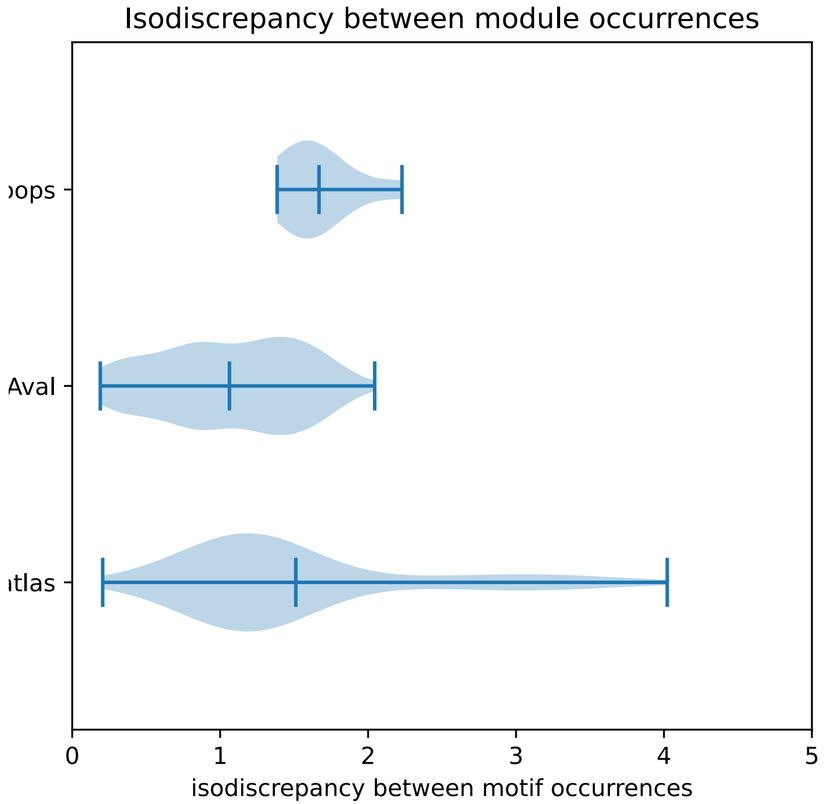
caRNAval



How consistant are structures inside the same module group?







Preliminary numbers
2-fold cross validation

	CaRNAval	RNA3DMotifs Atlas
MCC	0.777	0.981
FPR	0.066	0.006
FDR	0.073	0.006
F1	0.881	0.991
Average sensitivity (module occurrences correctly identified above threshold)	0.777	0.867

From modules to variants to predictions

Pasigraph: Finding all subgraphs isomorphims
https://gitlab.info.uqam.ca/cbe/pasigraph

FuzzTree: Exploring module variants with explicit control of distance
https://github.com/theoboury/FuzzTree

Cantaloops: Building motifs families from modules (Recurrent Interaction Networks)

RINs (without stacks for now) and annotations hosted on carnaval.cbe.uqam.ca
short talk in the visualisation session this afternoon

Thanks to all

Wilfried Agbeto

Théo Boury

Roman Sarrazin-Gendron

Hua-Ting Yao

Maëva Burillo

Camille Coti

Jérôme Waldispühl

Yann Ponty

Alvar Nuñez Cabeza de Vaca

FSM

NSERC CRSNG

Fonds de recherche Santé Québec

INNOVATION
Fondation canadienne pour l'innovation
Canada Foundation for Innovation

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

From modules to variants to predictions

Pasigraph: Finding all subgraphs isomorphims
https://gitlab.info.uqam.ca/cbe/pasigraph

FuzzTree: Exploring module variants with explicit control of distance
https://github.com/theoboury/FuzzTree

Cantaloops: Building motifs families from modules (Recurrent Interaction Networks)

RINs (without stacks for now) and annotations hosted on carnaval.cbe.uqam.ca
short talk in the visualisation session this afternoon