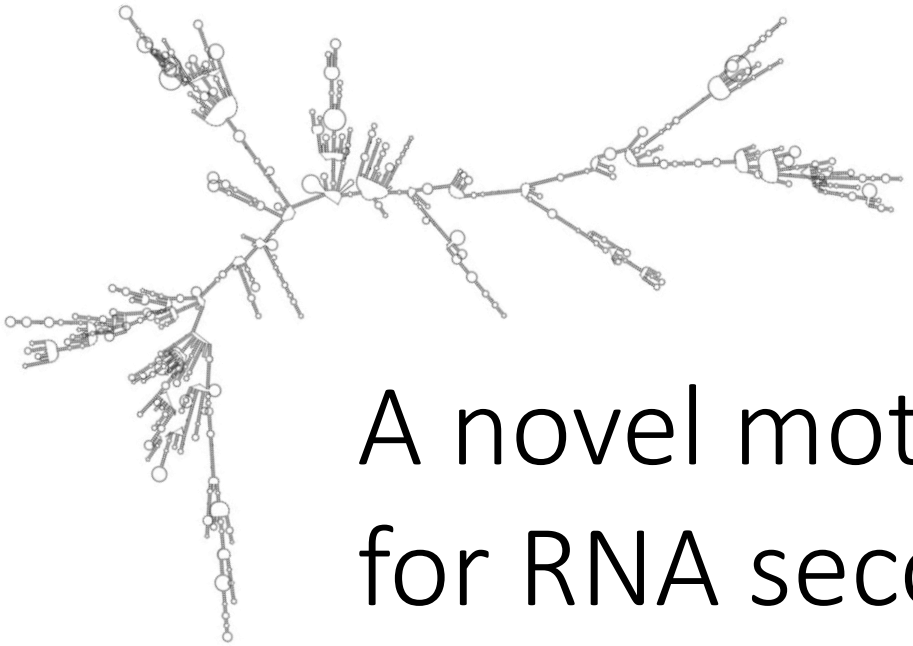
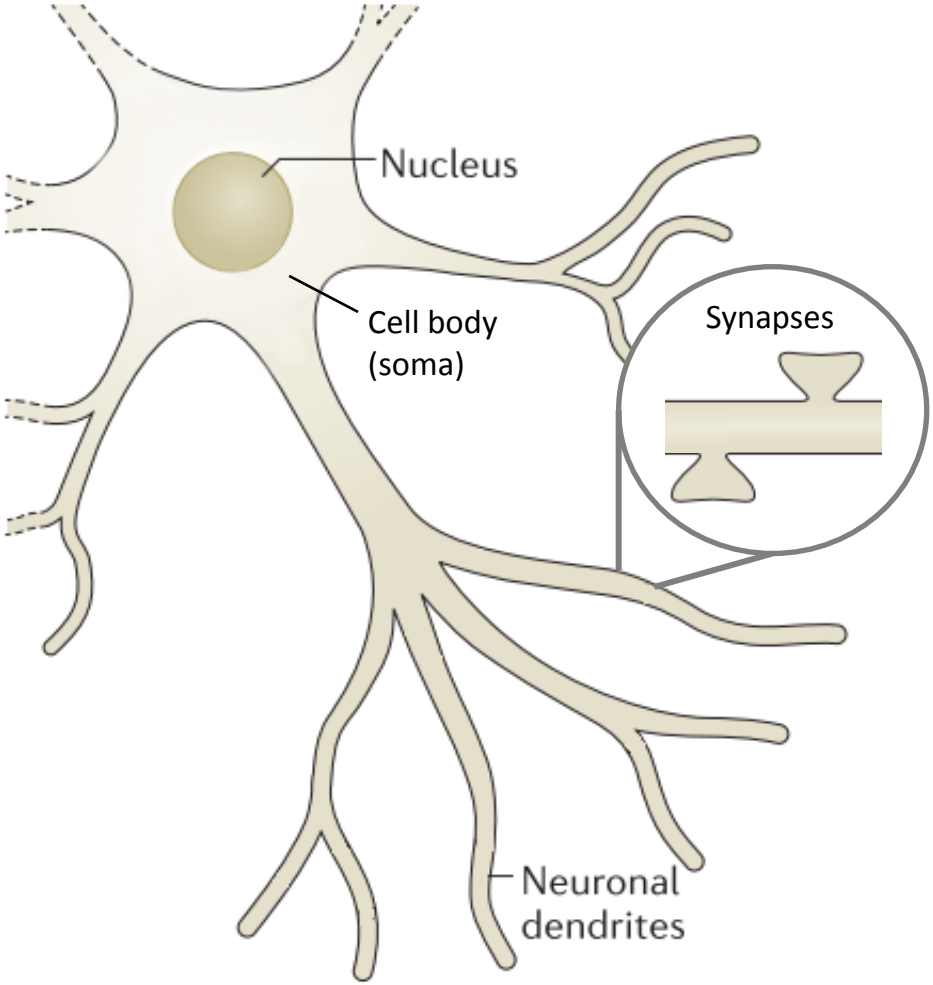


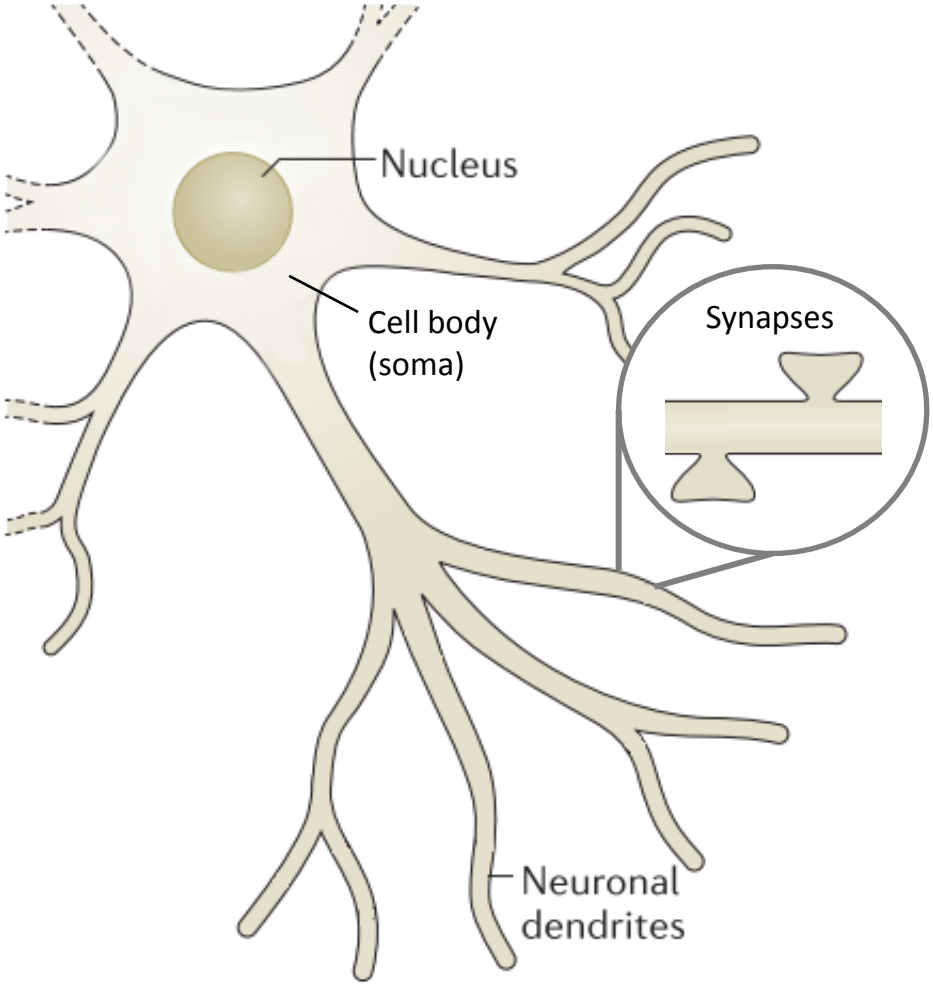
A novel motif discovery tool for RNA secondary structures

Sarah Middleton
Genomics & Computational Biology
University of Pennsylvania



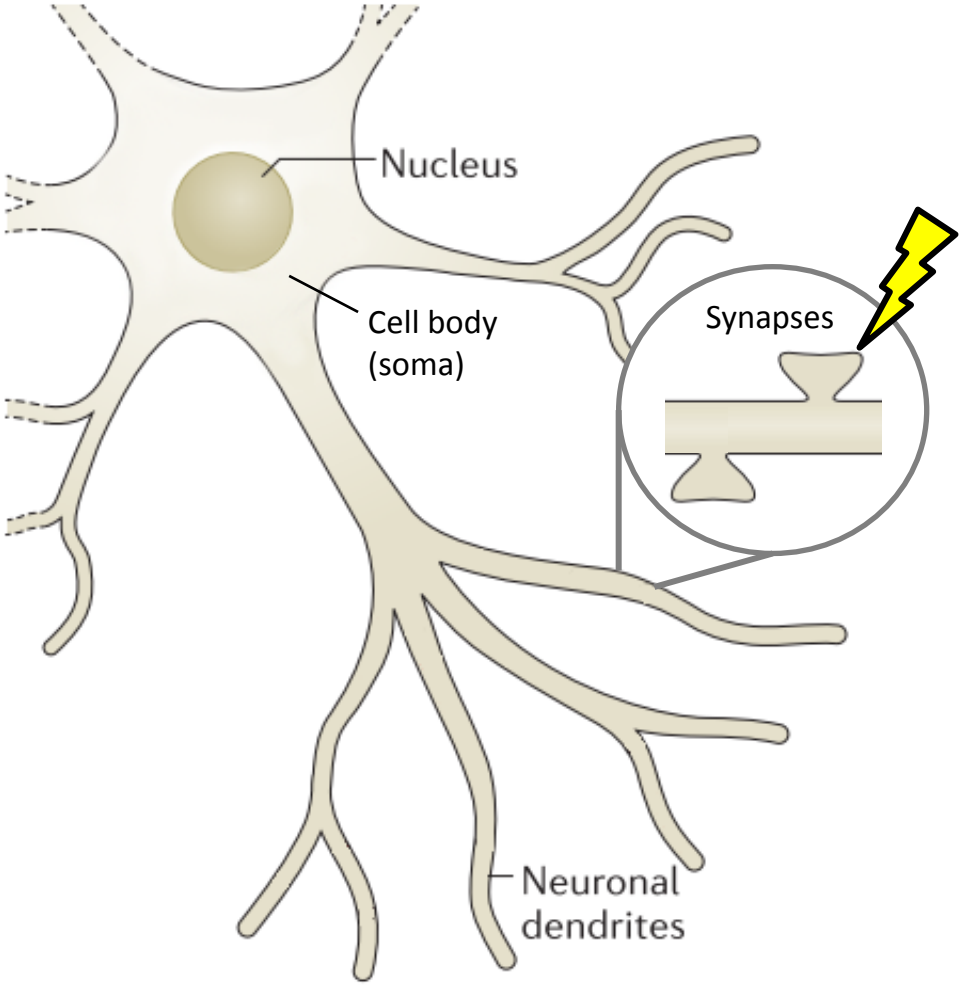
Edited from Buxbaum et al. 2015

Long-term potentiation (LTP)



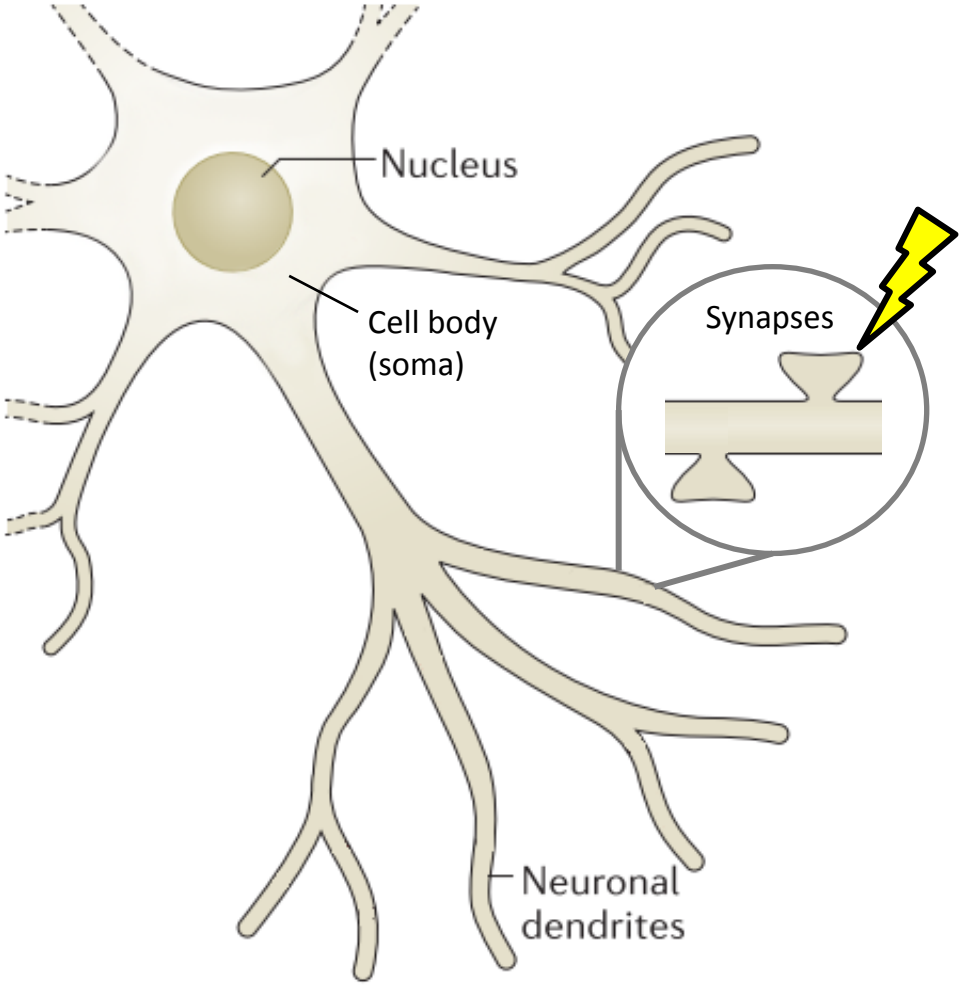
Edited from Buxbaum et al. 2015

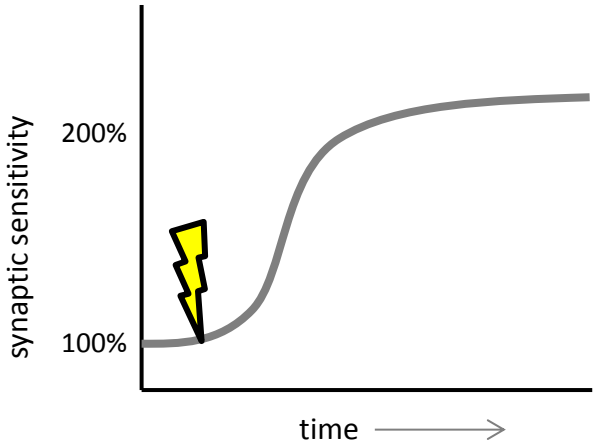
Long-term potentiation (LTP)



Edited from Buxbaum et al. 2015

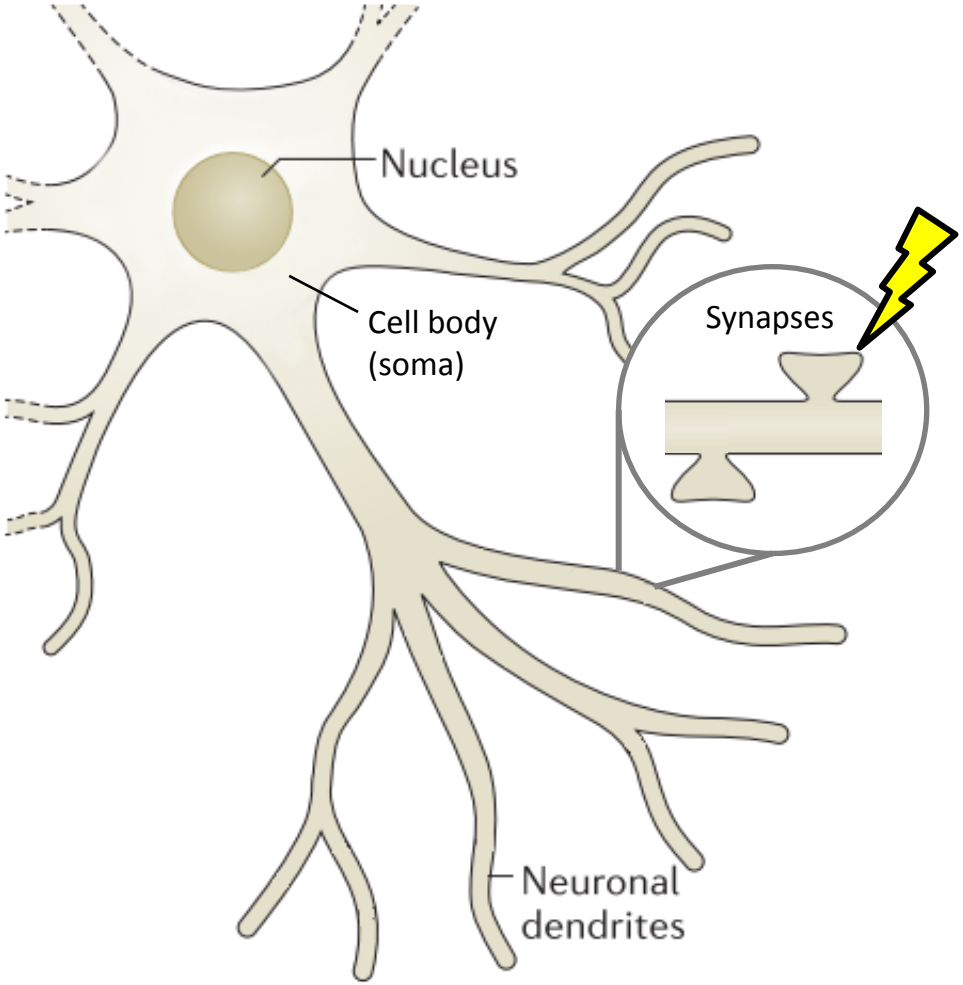
Long-term potentiation (LTP)

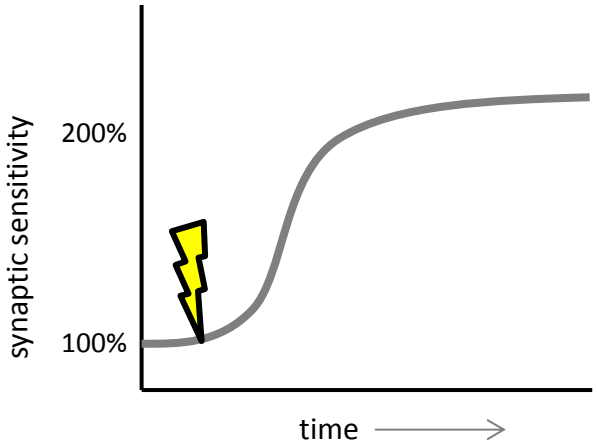




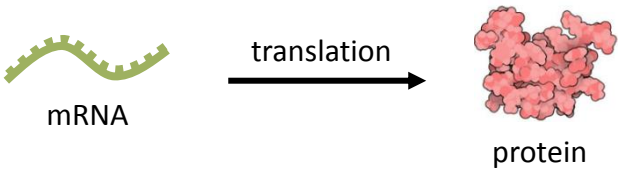
Edited from Buxbaum et al. 2015

Long-term potentiation (LTP)



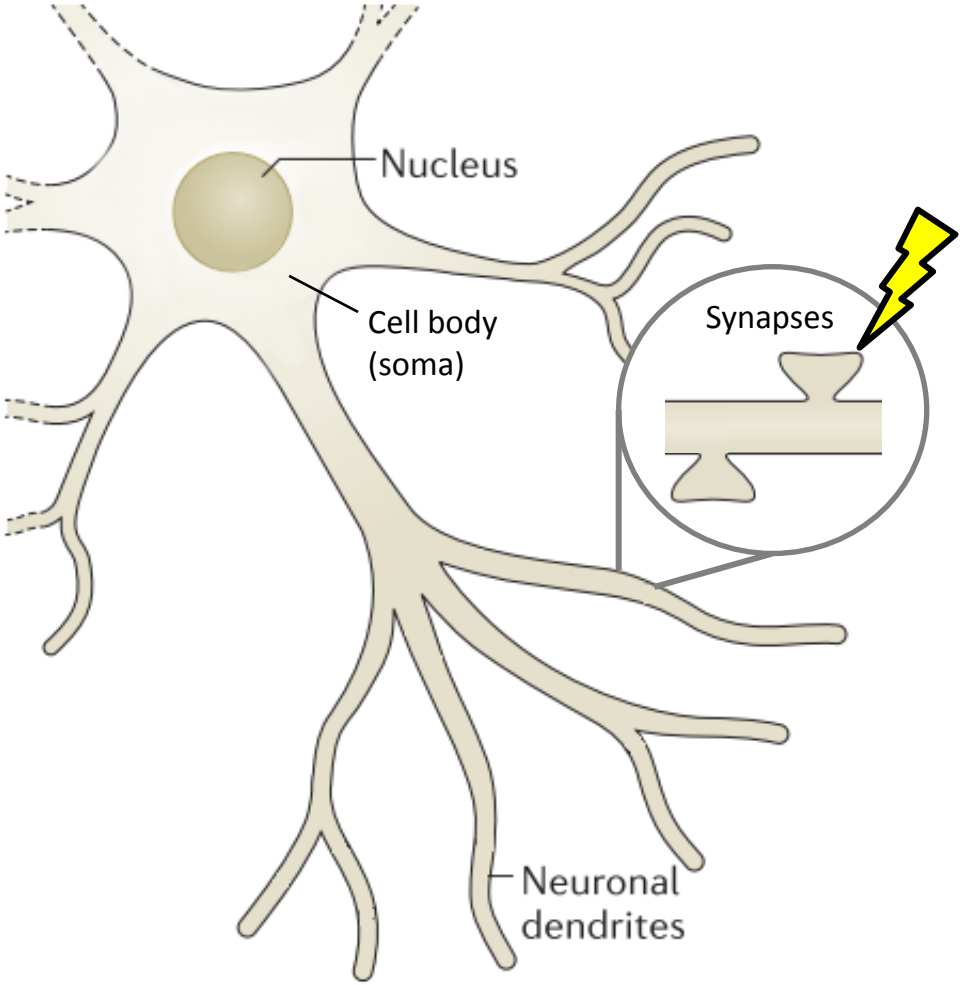


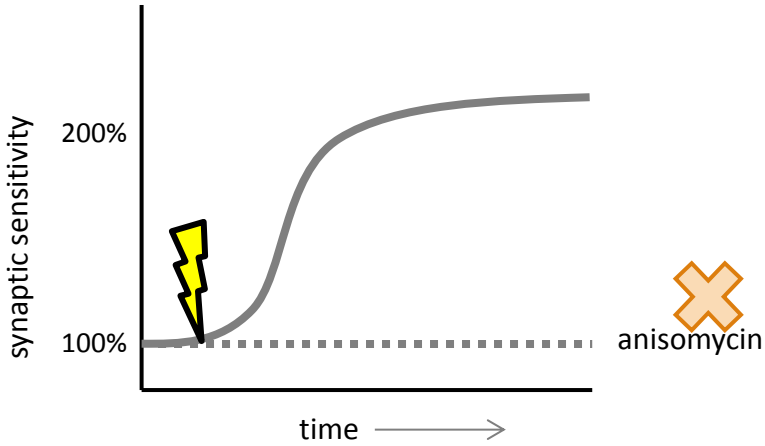
Protein synthesis is required!



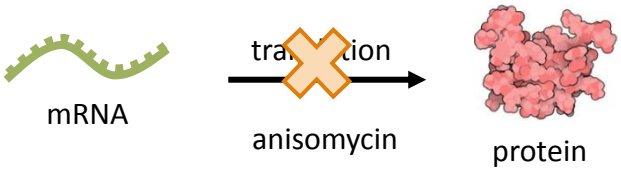
Edited from Buxbaum et al. 2015

Long-term potentiation (LTP)





Protein synthesis is required!

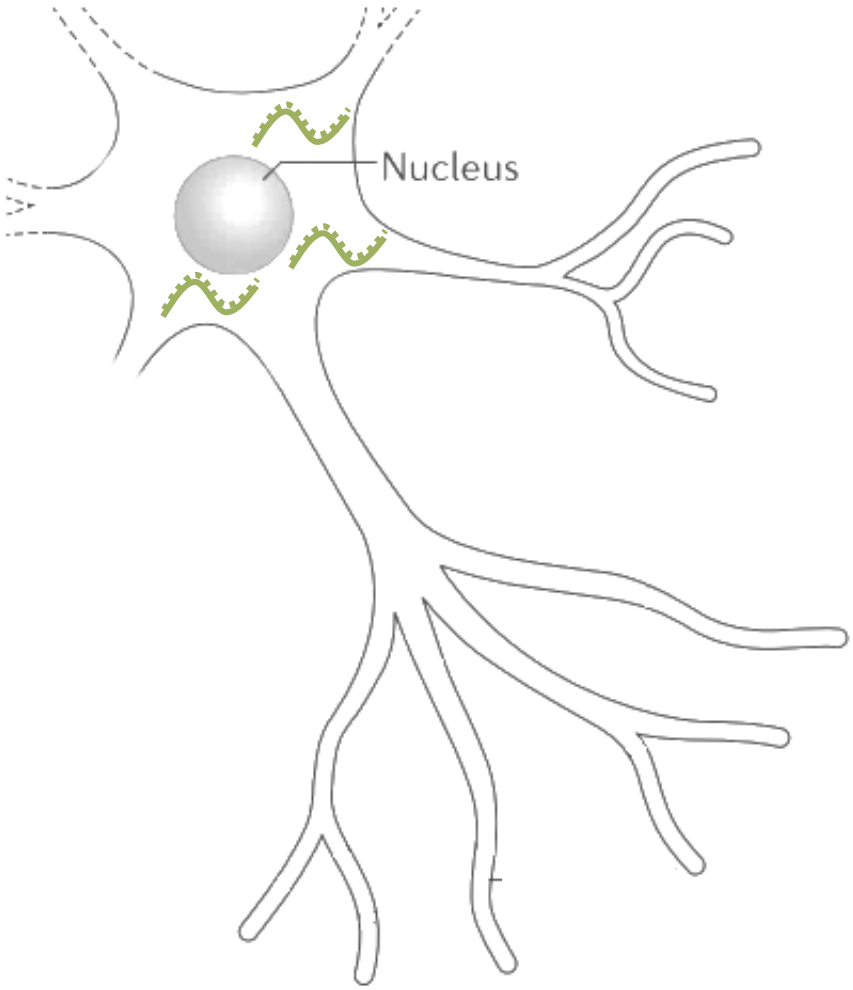


Edited from Buxbaum et al. 2015

Active mRNA

Repressed mRNA

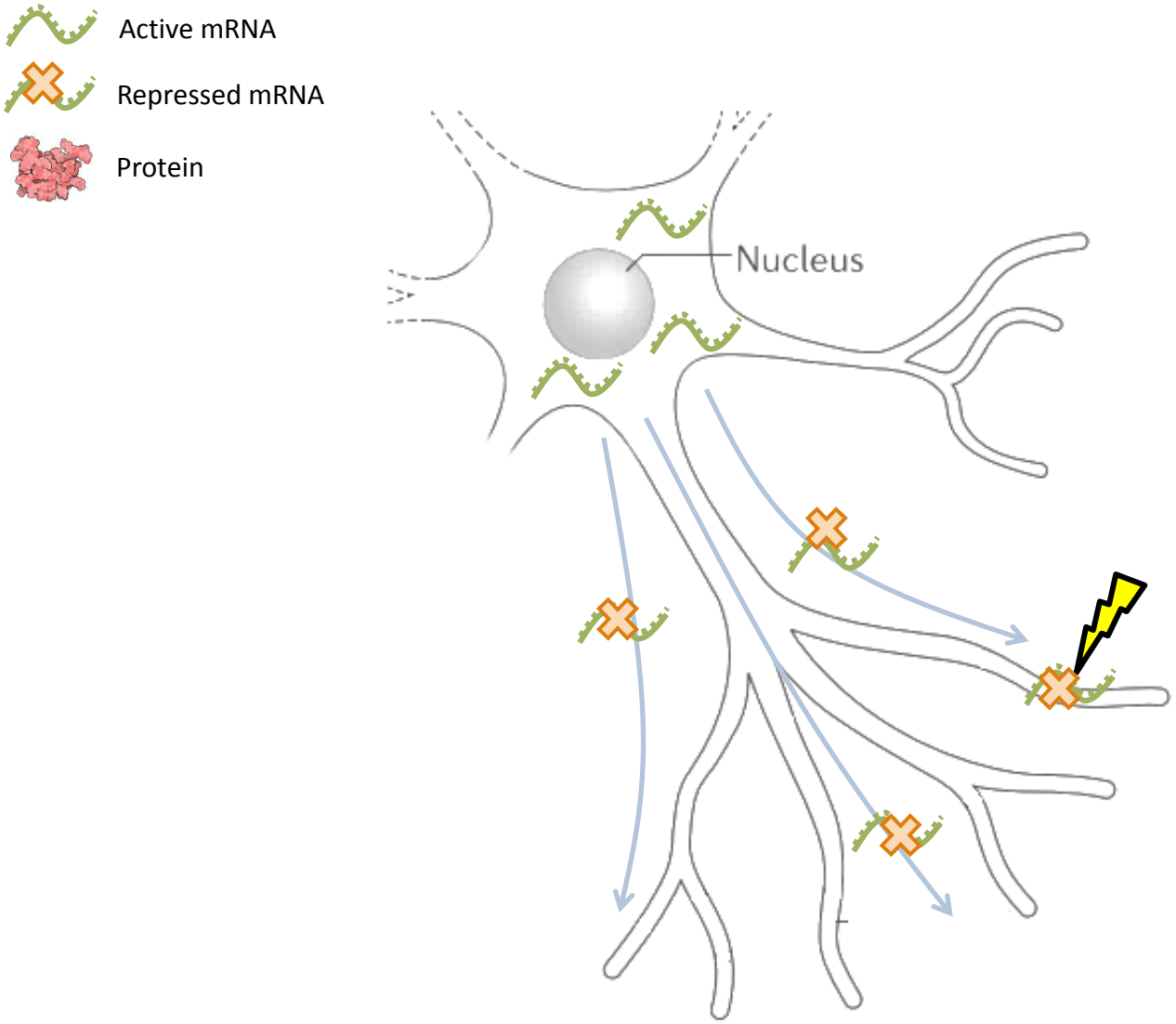
Protein



Distributed production

1. mRNA already present in dendrites (repressed state)
2. Activated synapse triggers translation of nearby mRNAs into protein

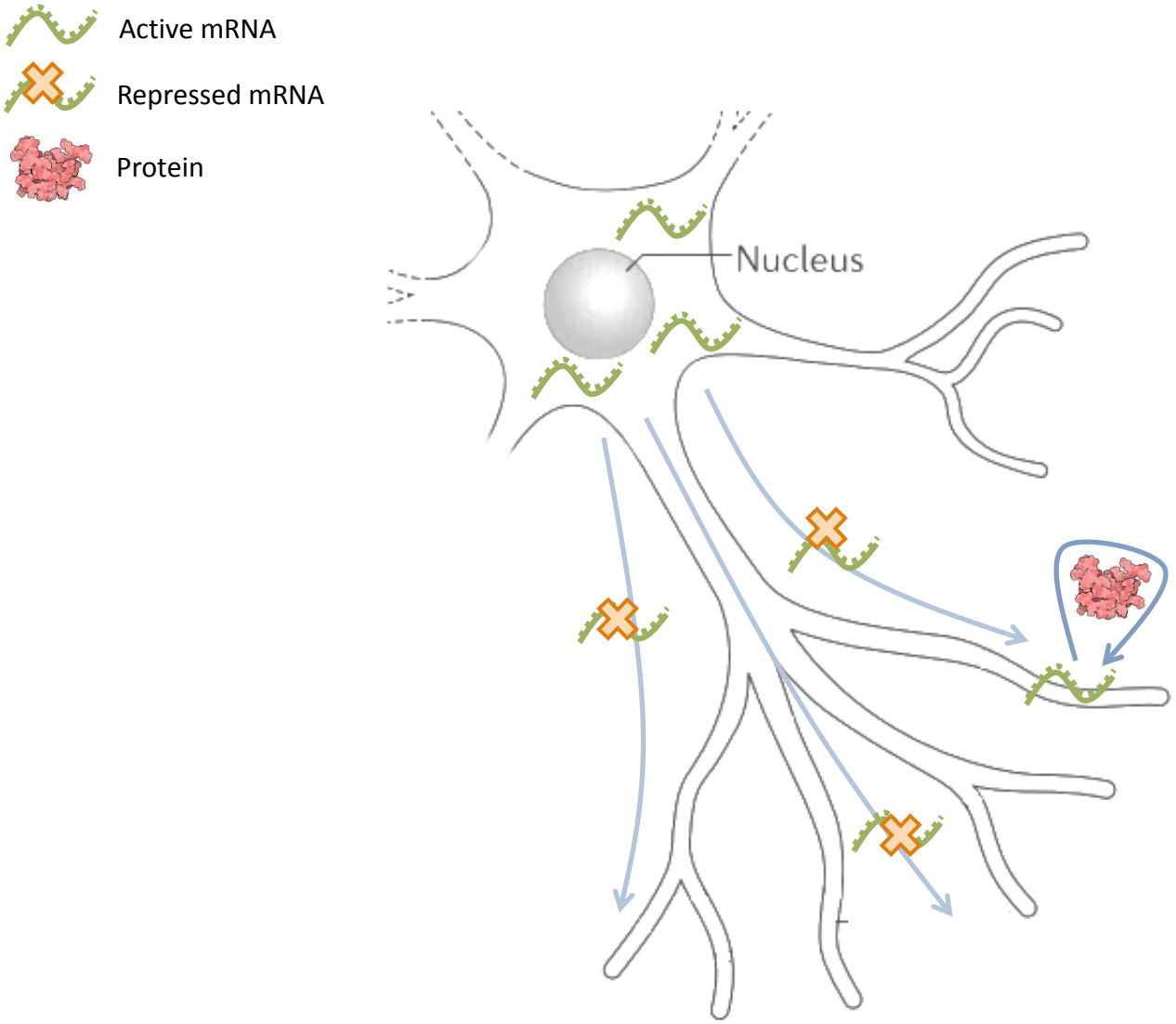
Edited from Buxbaum et al. 2015



Distributed production

1. mRNA already present in dendrites (repressed state)
2. Activated synapse triggers translation of nearby mRNAs into protein

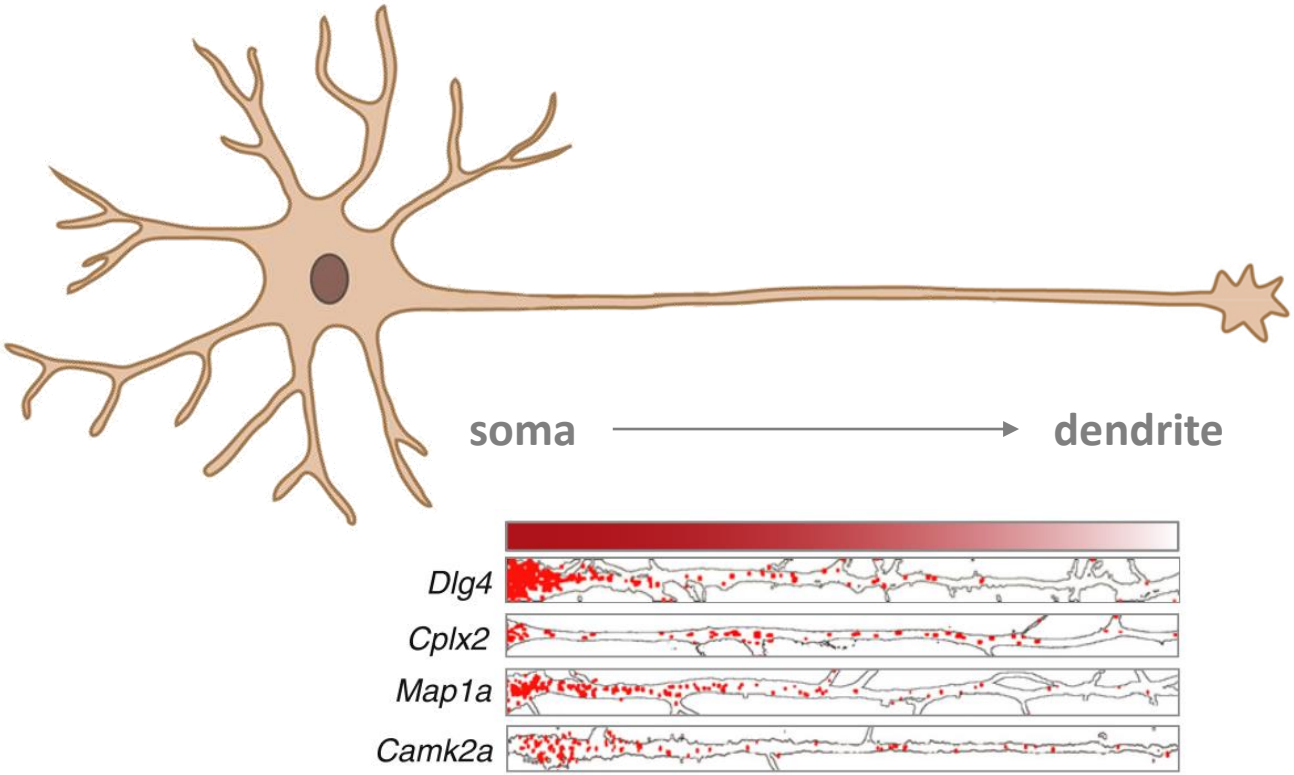
Edited from Buxbaum et al. 2015



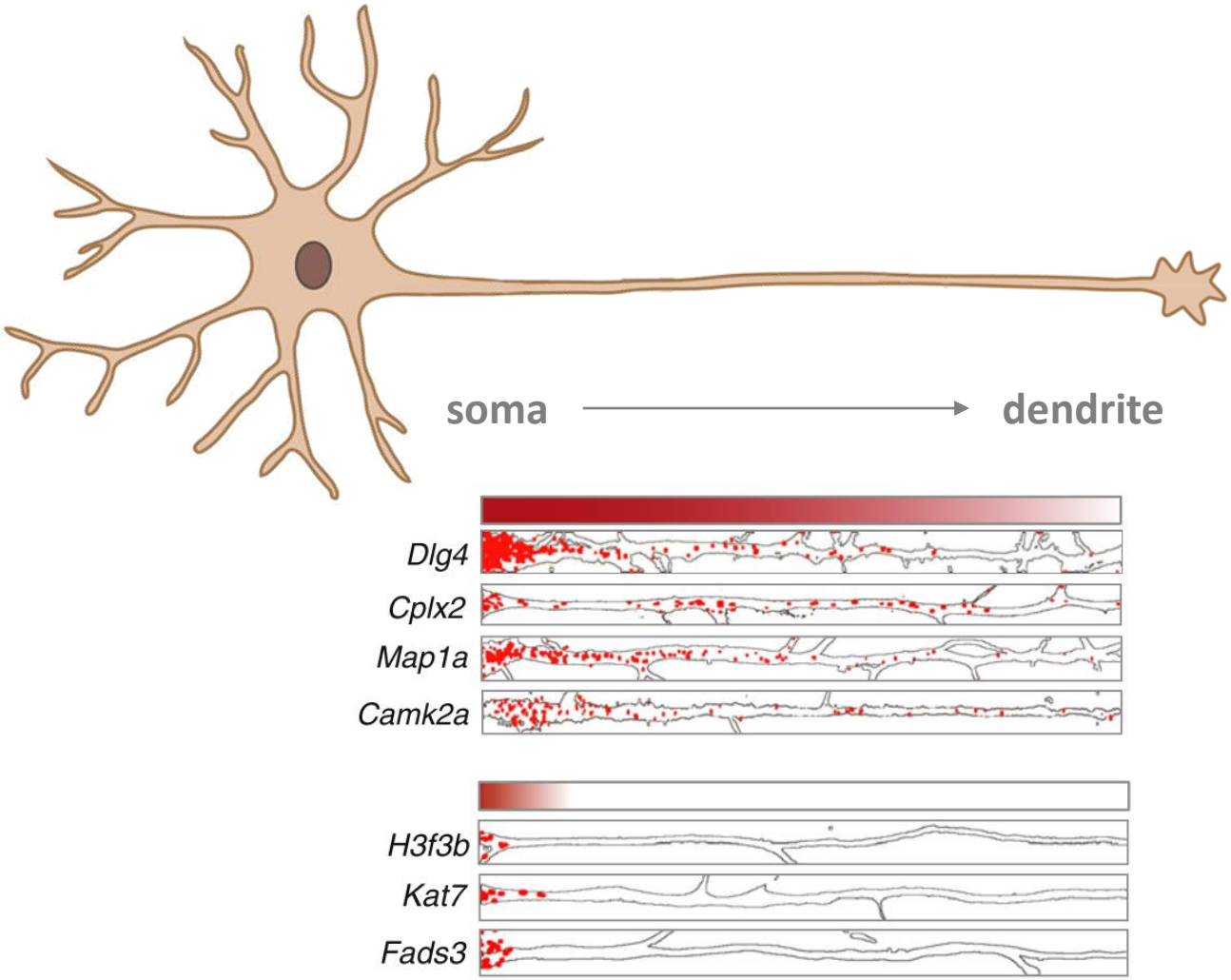
Distributed production

1. mRNA already present in dendrites (repressed state)
2. Activated synapse triggers translation of nearby mRNAs into protein

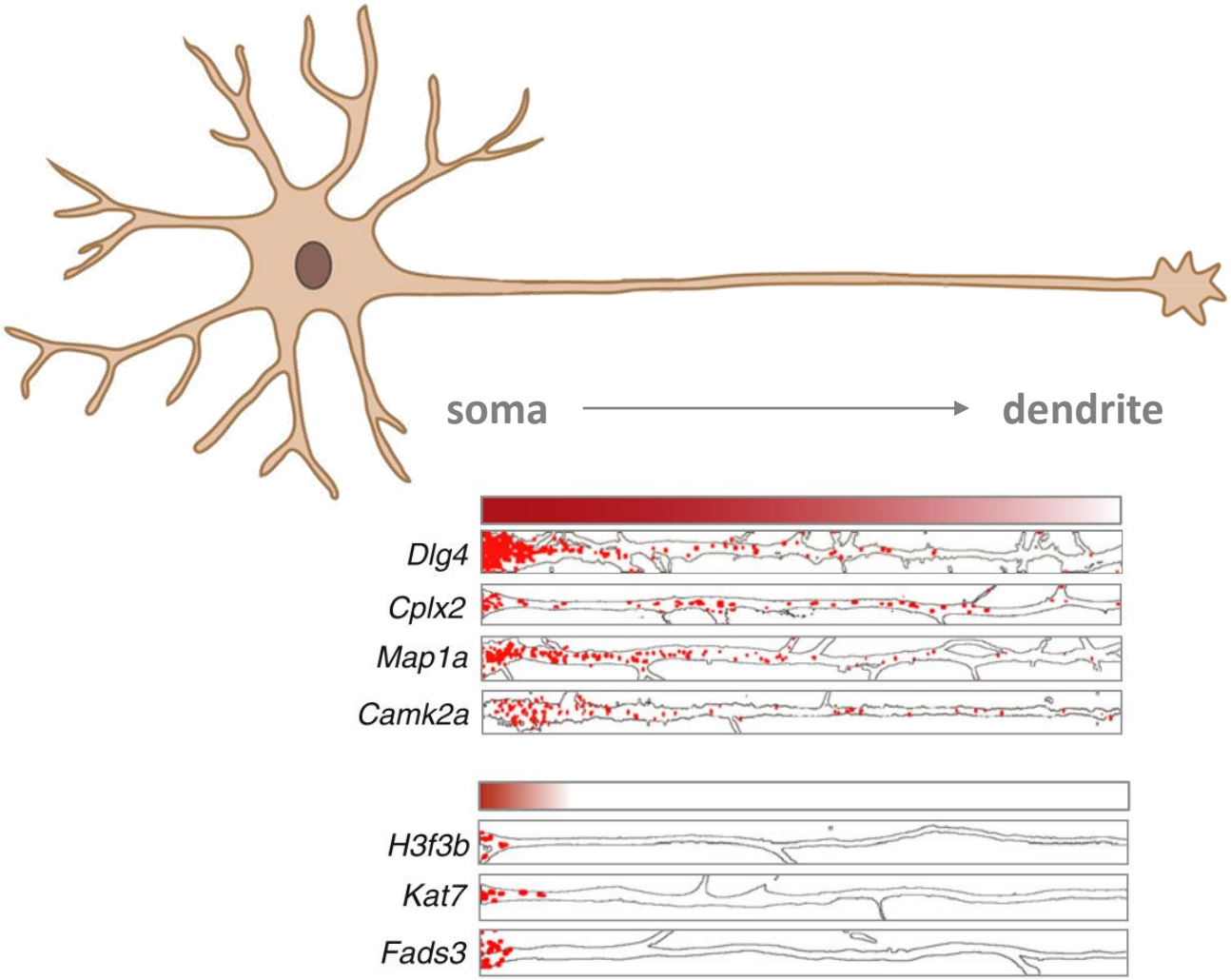
Edited from Buxbaum et al. 2015



Kevenaar et al. 2015; Cajigas et al. 2012



Kevenaar et al. 2015; Cajigas et al. 2012

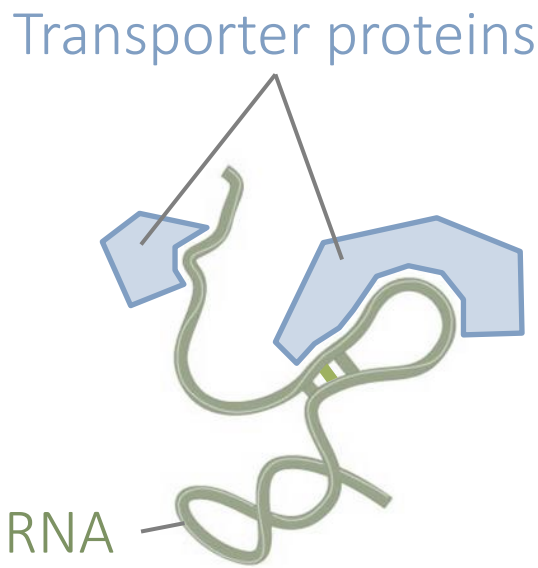


How is dendritic localization regulated?

(How does the cell control which RNAs are localized, and when?)

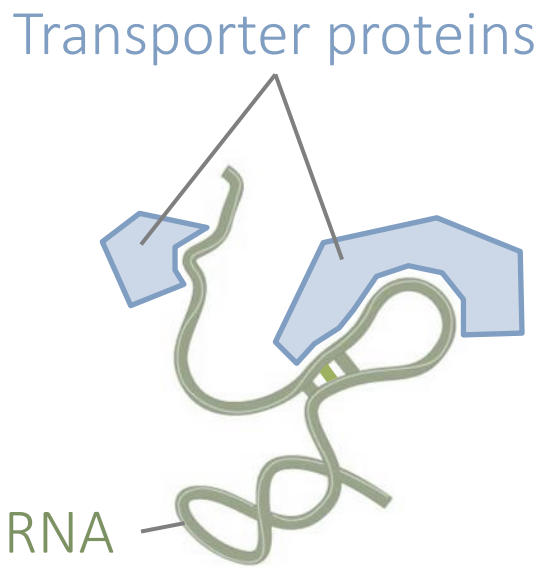
Kevenaar et al. 2015; Cajigas et al. 2012

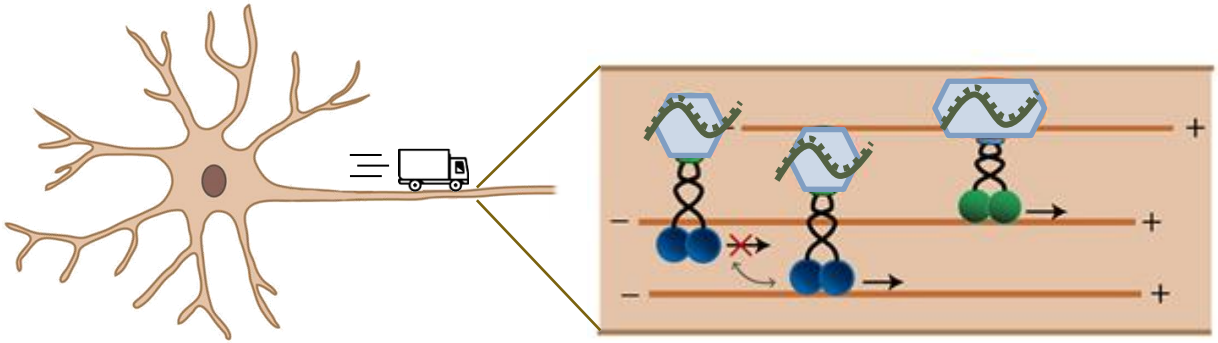
What we know so far



Broad Inst.; RSCB; Kevenaar et al. 2015

What we know so far

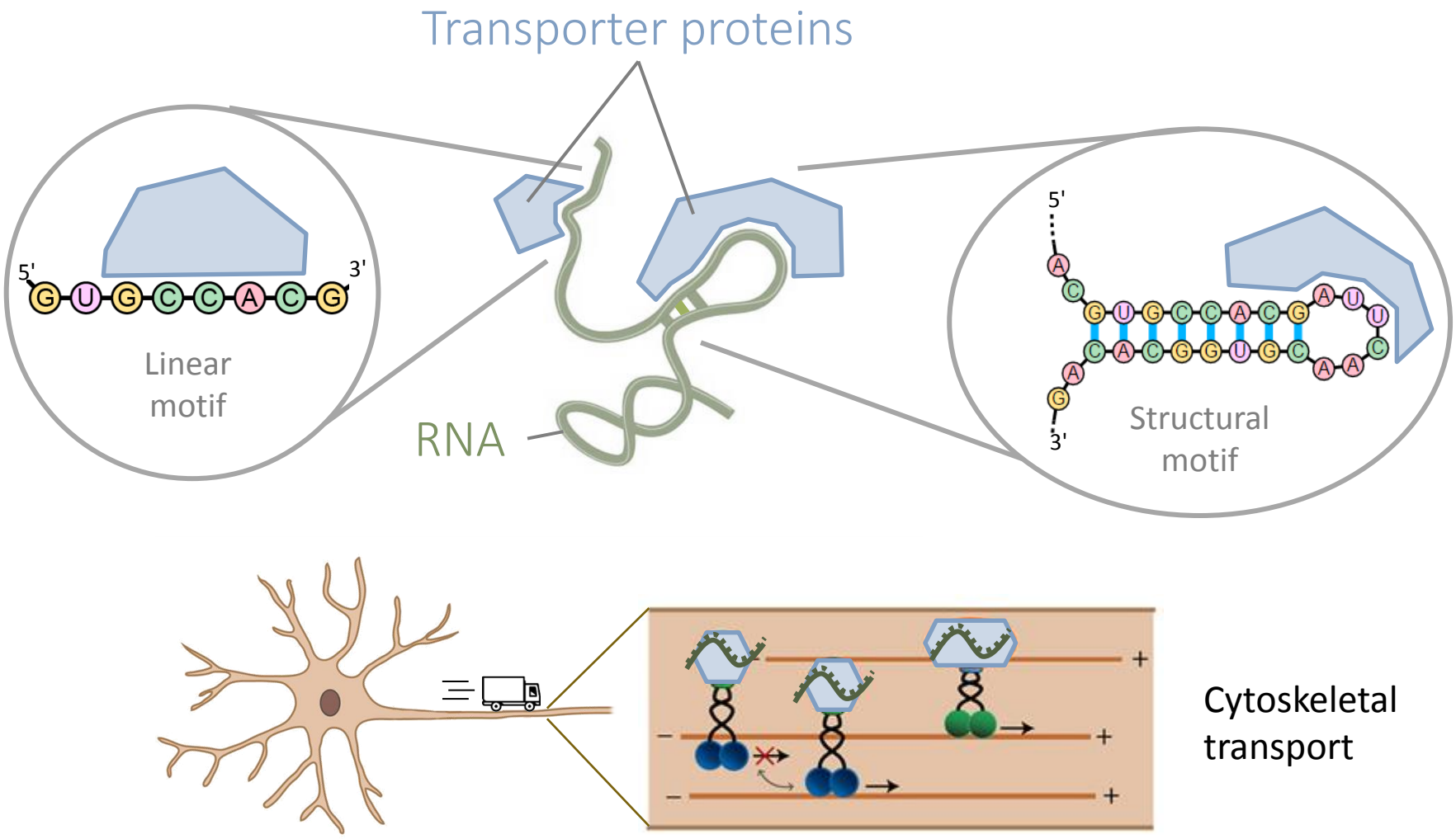




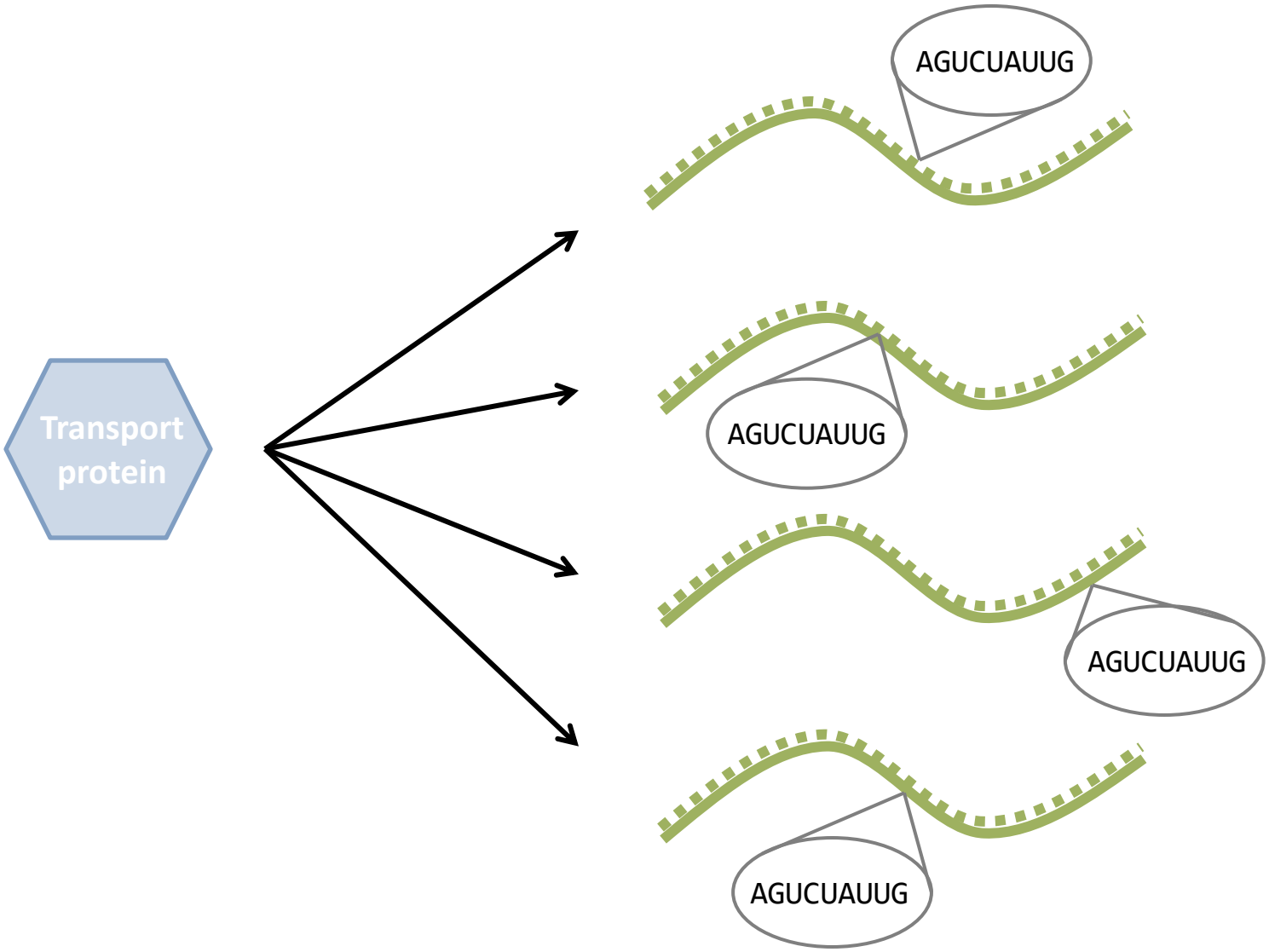
Cytoskeletal transport

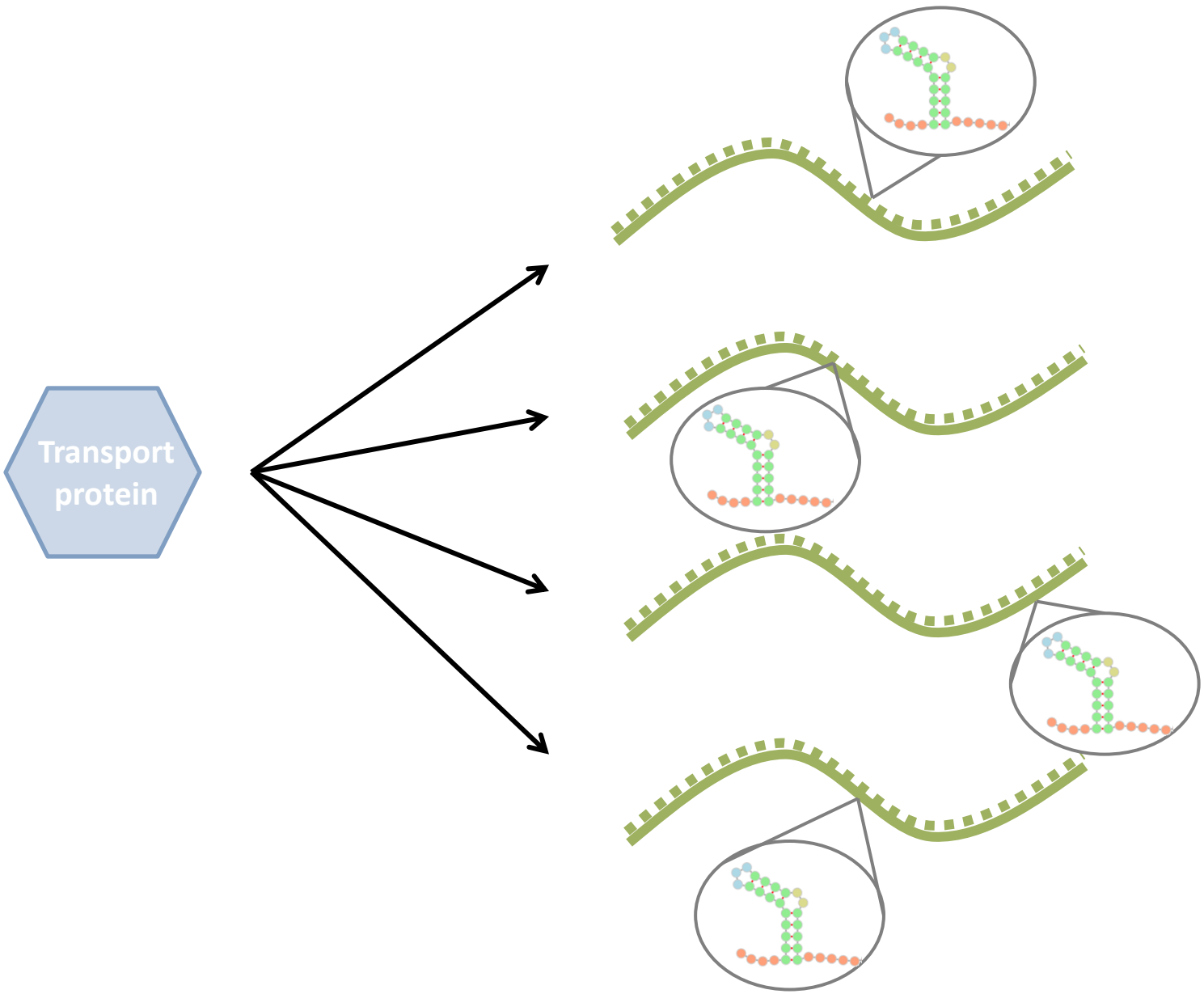
Broad Inst.; RSCB; Kevenaar et al. 2015

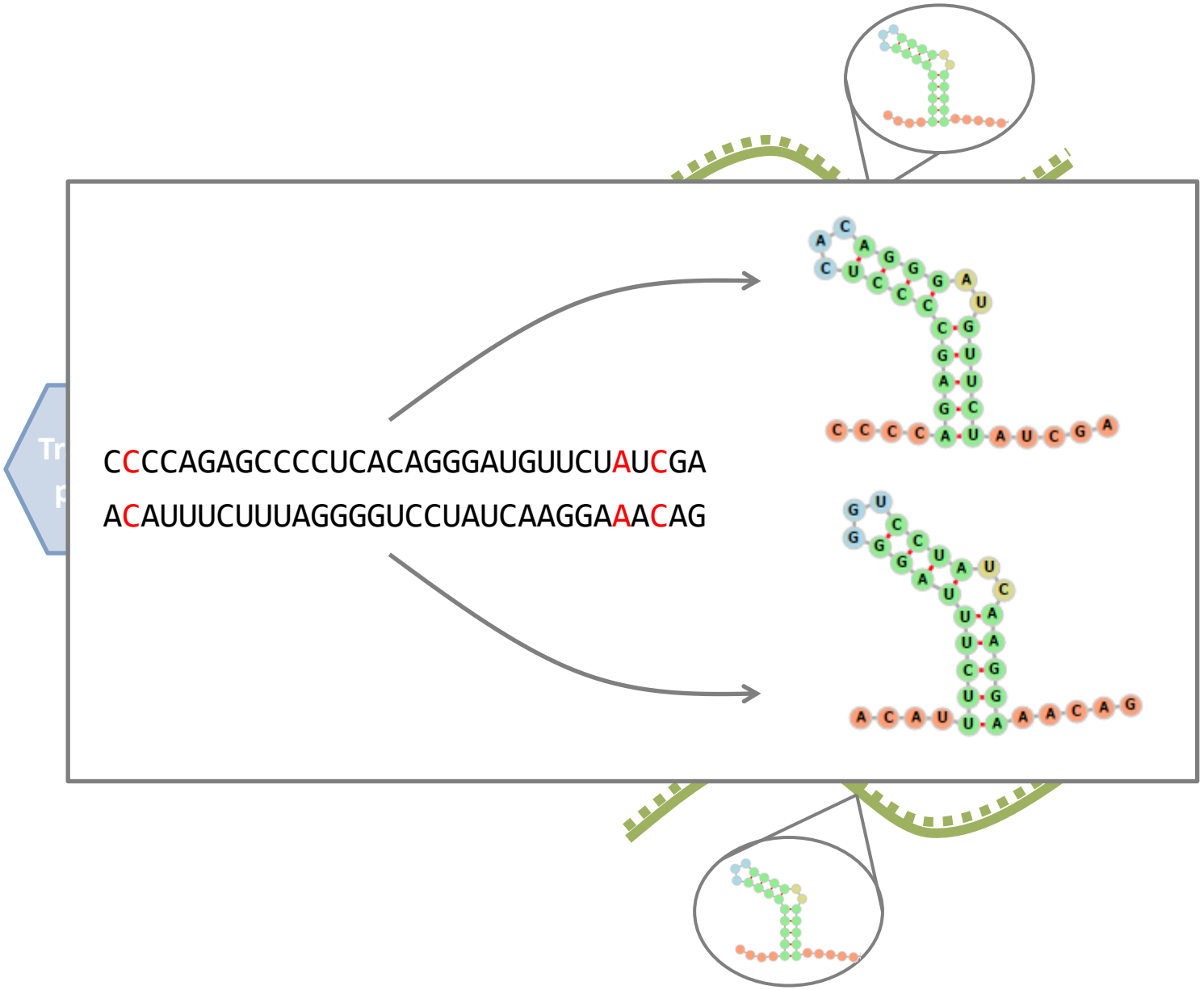
What we know so far



Broad Inst.; RSCB; Kevenaar et al. 2015


Transport protein
AGUCUAUUG
AGUCUAUUG
AGUCUAUUG
AGUCUAUUG

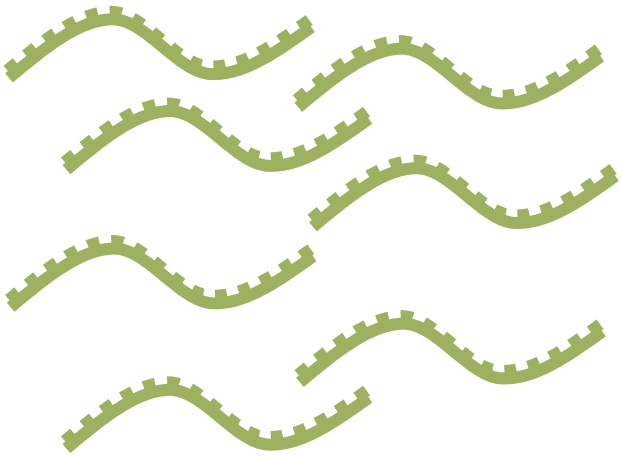

Transport protein


CCCCAGAGCCCCUCACAGGGAUGUUCUAUCGA
ACAUUUCUUUAGGGGUCCUAUCAAGGAAACAG

How can we identify RNA structure motifs?

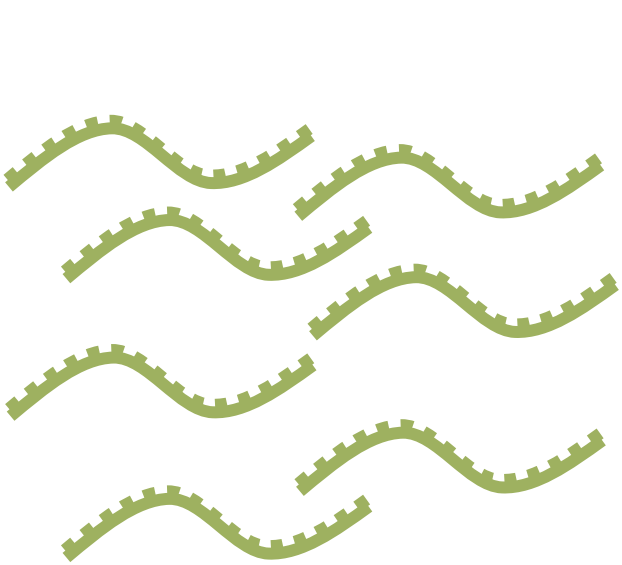
Main task:

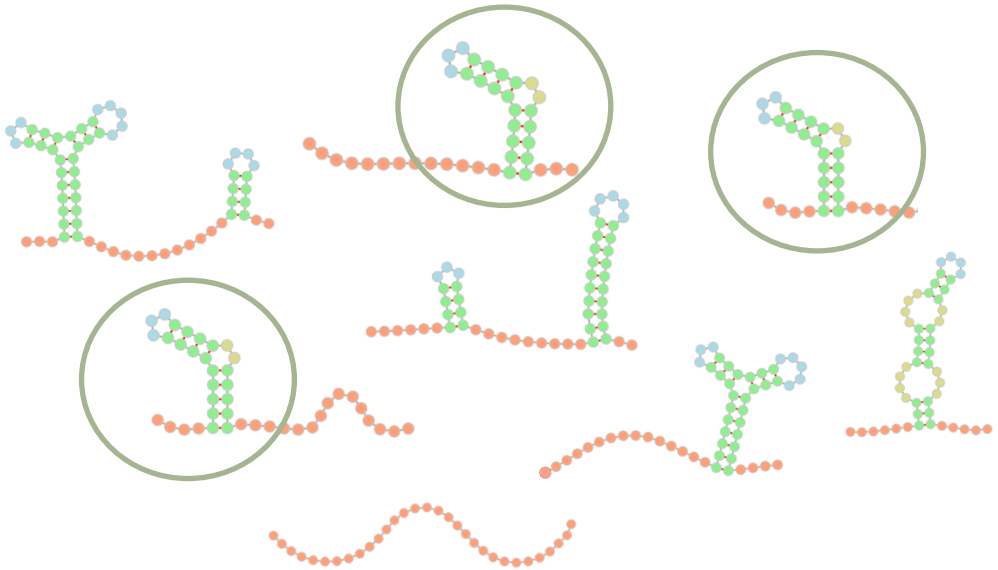
Given a set of sequences with unknown structure,
predict sub-regions that share a common structure (motif).



Main task:

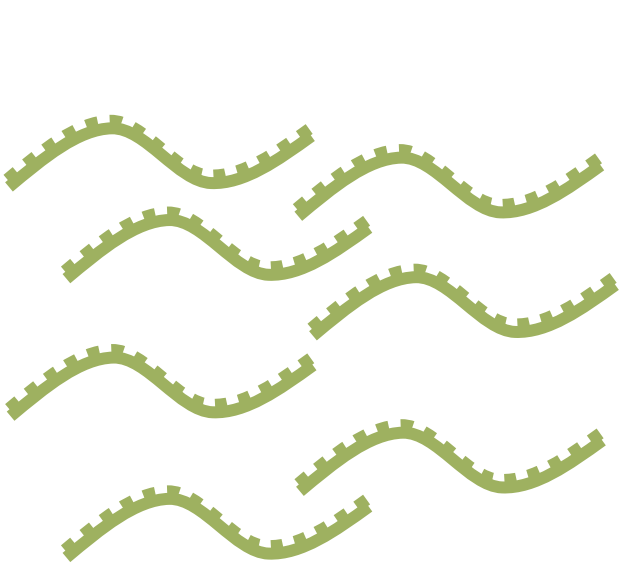
Given a set of sequences with unknown structure,
predict sub-regions that share a common structure (motif).

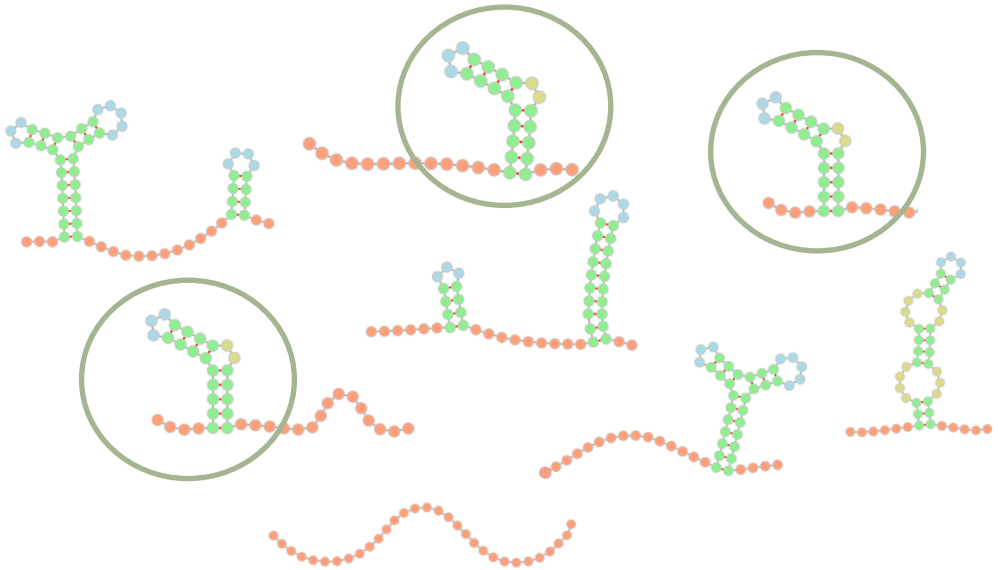




Main task:

Given a set of sequences with unknown structure,
predict sub-regions that share a common structure (motif).

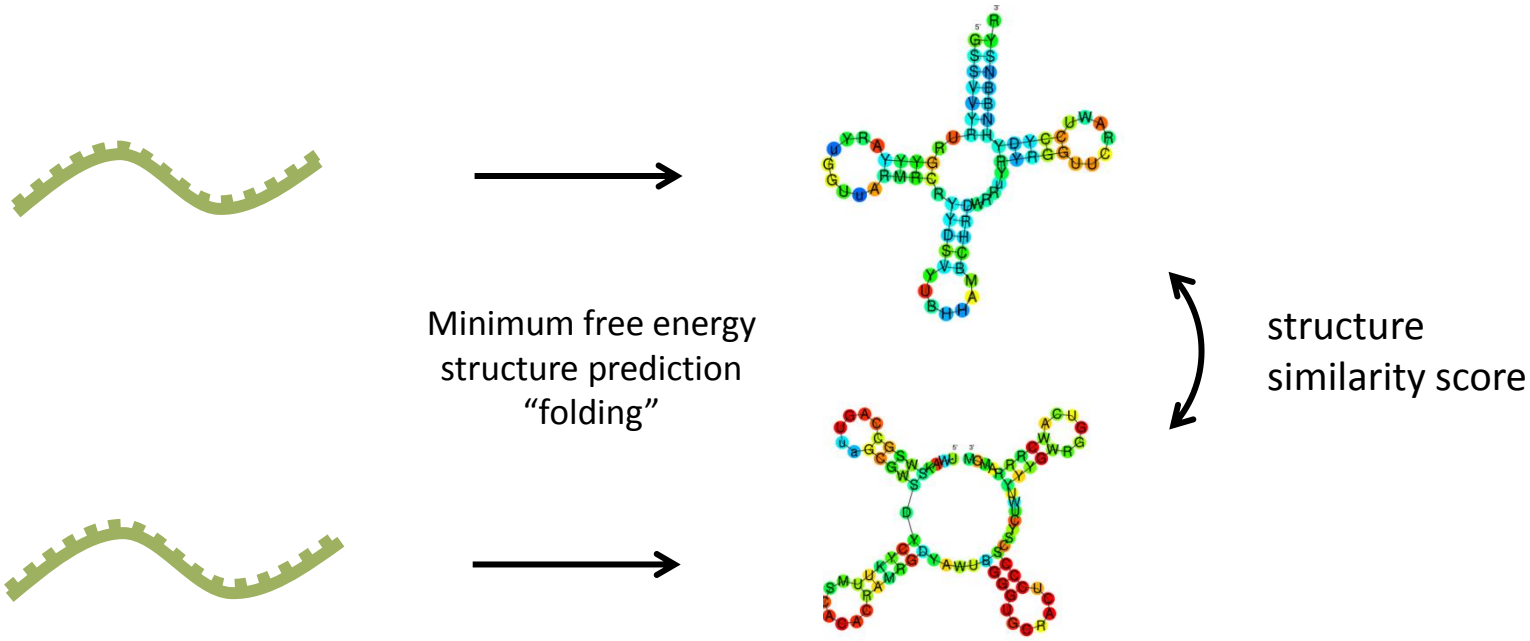




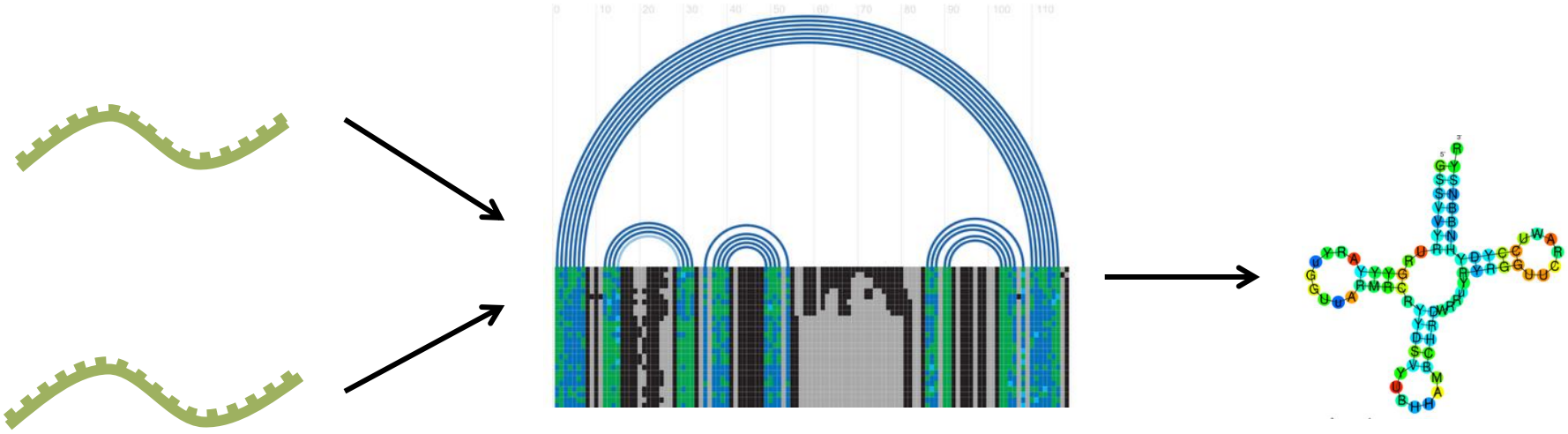
Complications:

- A sequence may have no motifs
- Boundary of individual motifs is unknown

Method 1: **fold-and-compare**



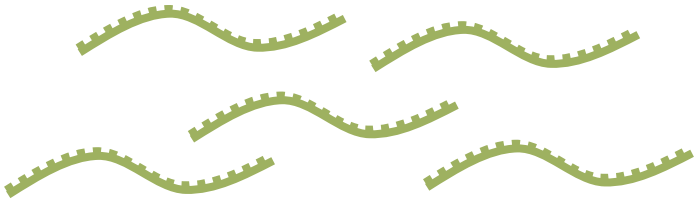
Method 2: **align-and-fold**



Covariance + MFE

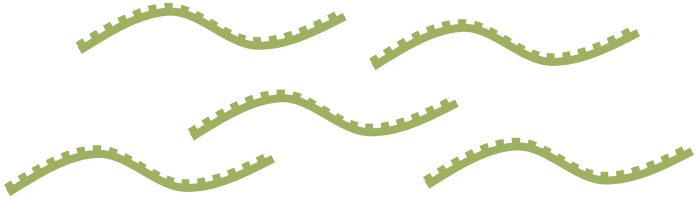
alignment score + structure conservation

NoFold: RNA structure clustering without folding or alignment



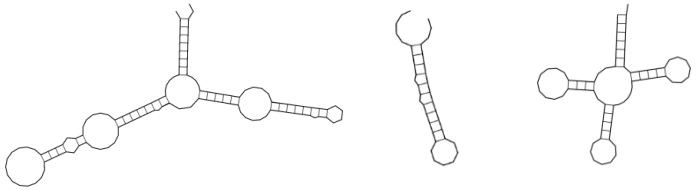
RNA sequences

NoFold: RNA structure clustering without folding or alignment



RNA sequences

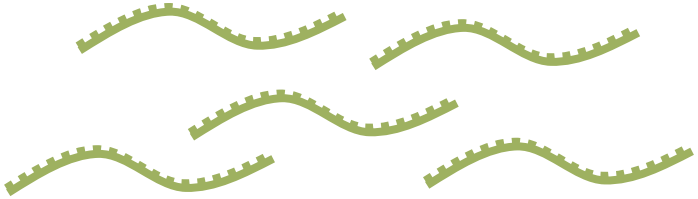
score against



~2000 Known structures

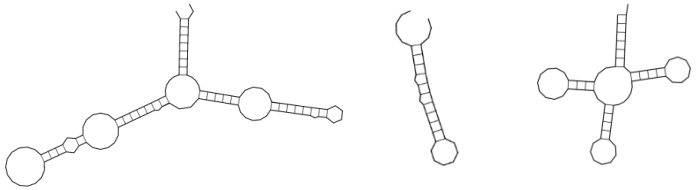
Represented as covariance Models (CMs)

NoFold: RNA structure clustering without folding or alignment



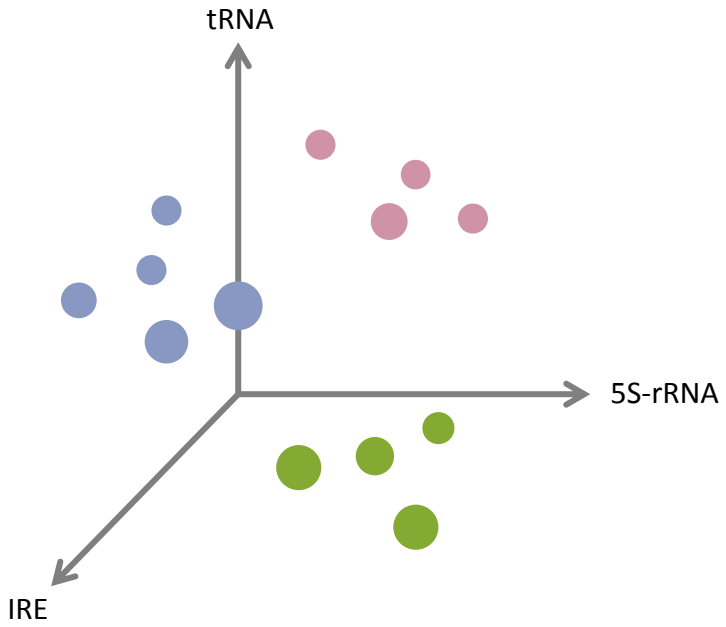
RNA sequences

score against



~2000 Known structures

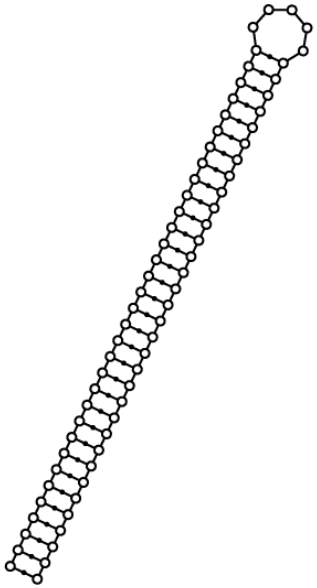
Represented as covariance Models (CMs)



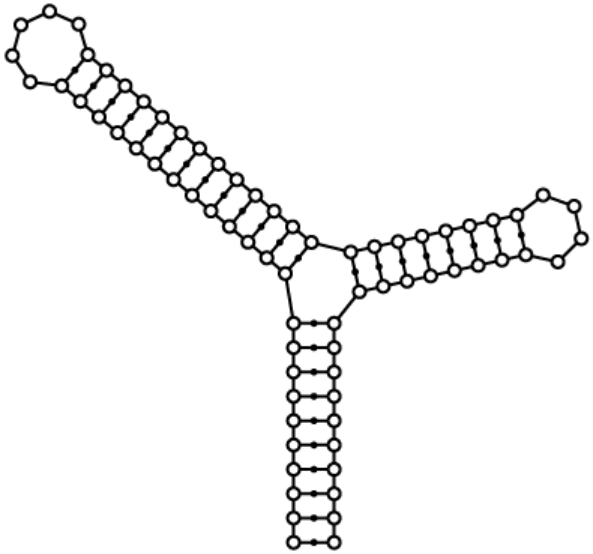
Structural feature space

(not shown: additional 1,970 dimensions)

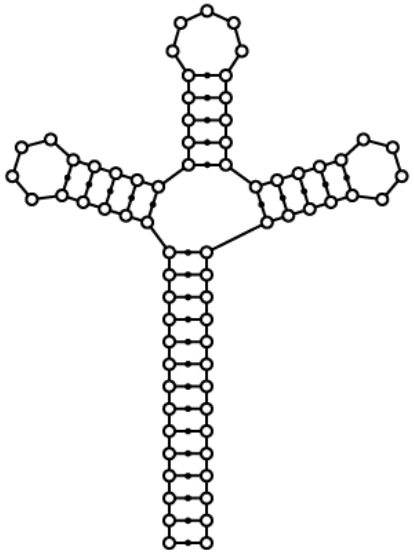
Benchmark: Randomized sequences



1-hp

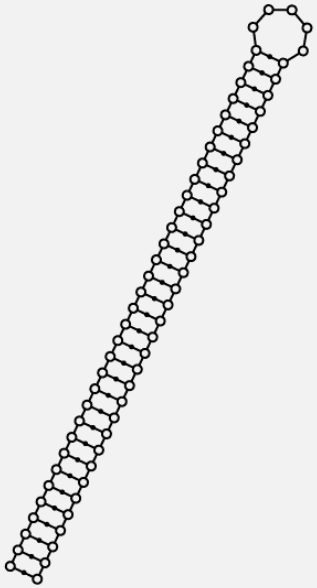


2-hp

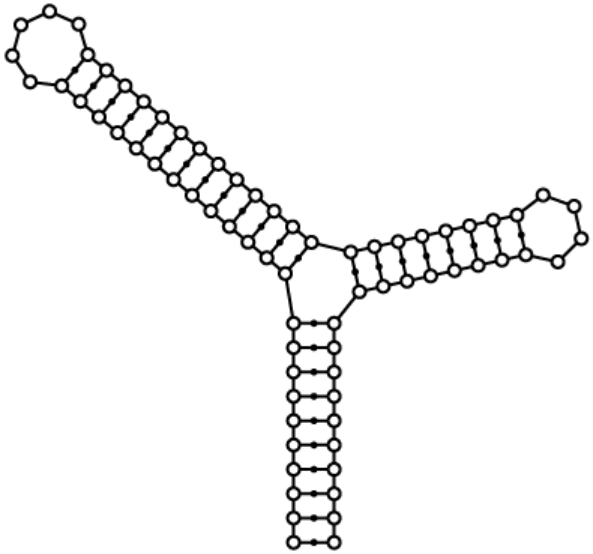


3-hp

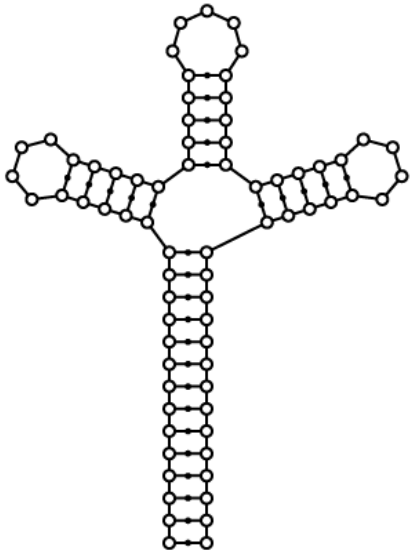
Benchmark: Randomized sequences



1-hp



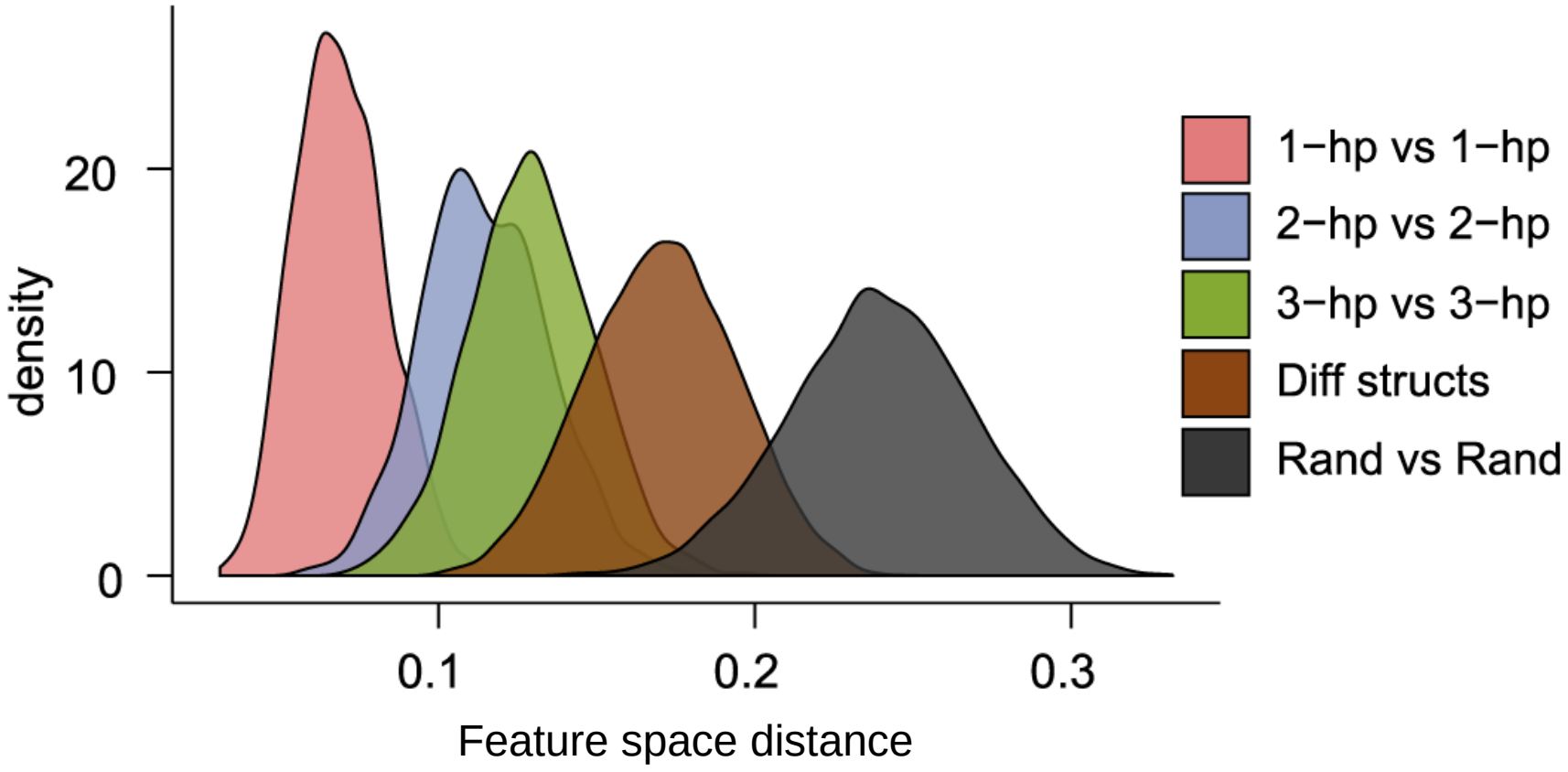
2-hp



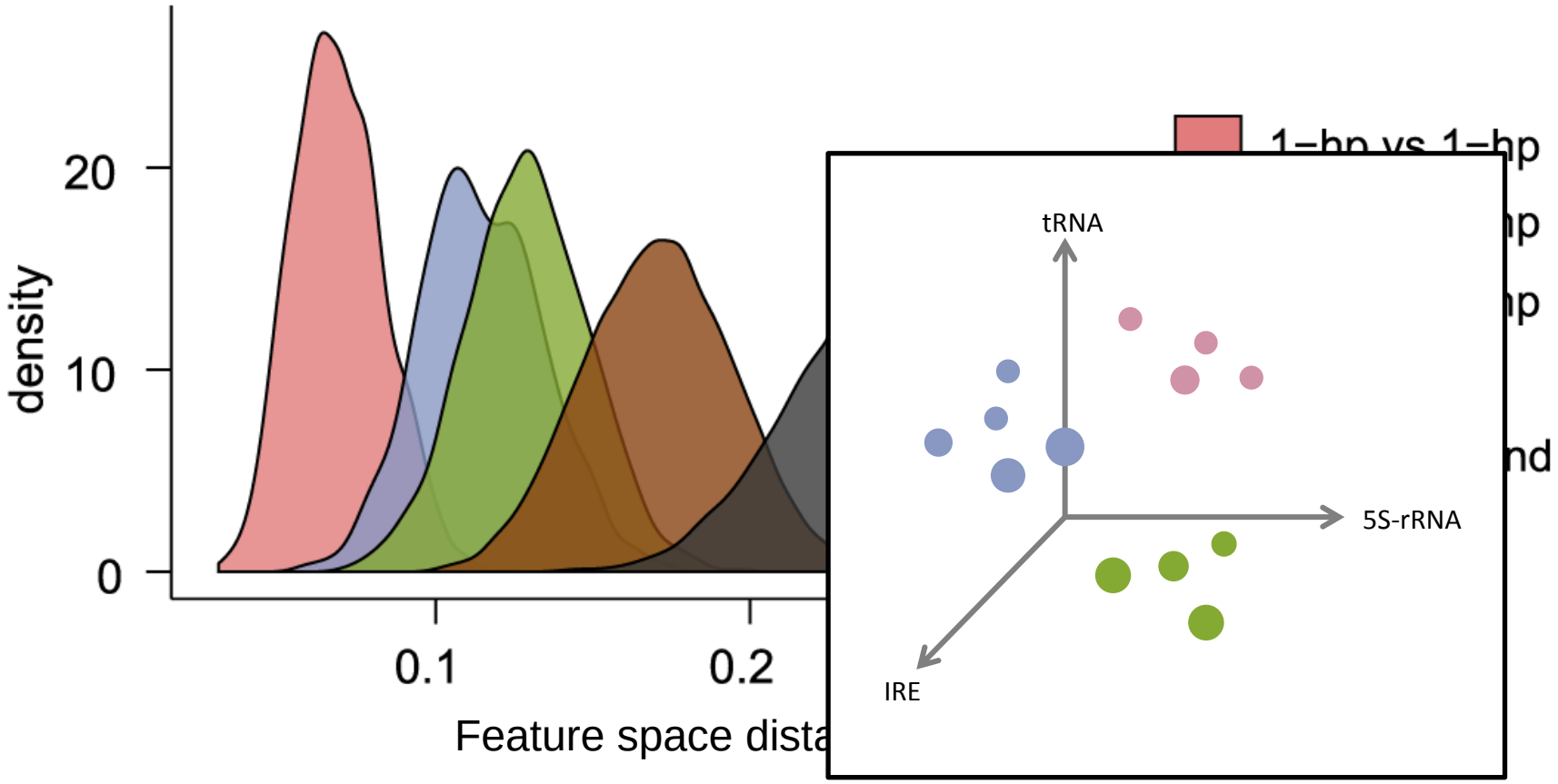
3-hp

GUGGAUCACAAUUACAAUCCCCAGGCUGUCUAGGCUAUCUAGACAGCCUGGGGAUUGUAAUUGUGAUCCAC
AGUGCCAGUUGAAGCUCUCUUACAUCCUAUAUGUUGCACAUAUAGGAUGUAAGAGAGCUUCAACUGGCACU
...

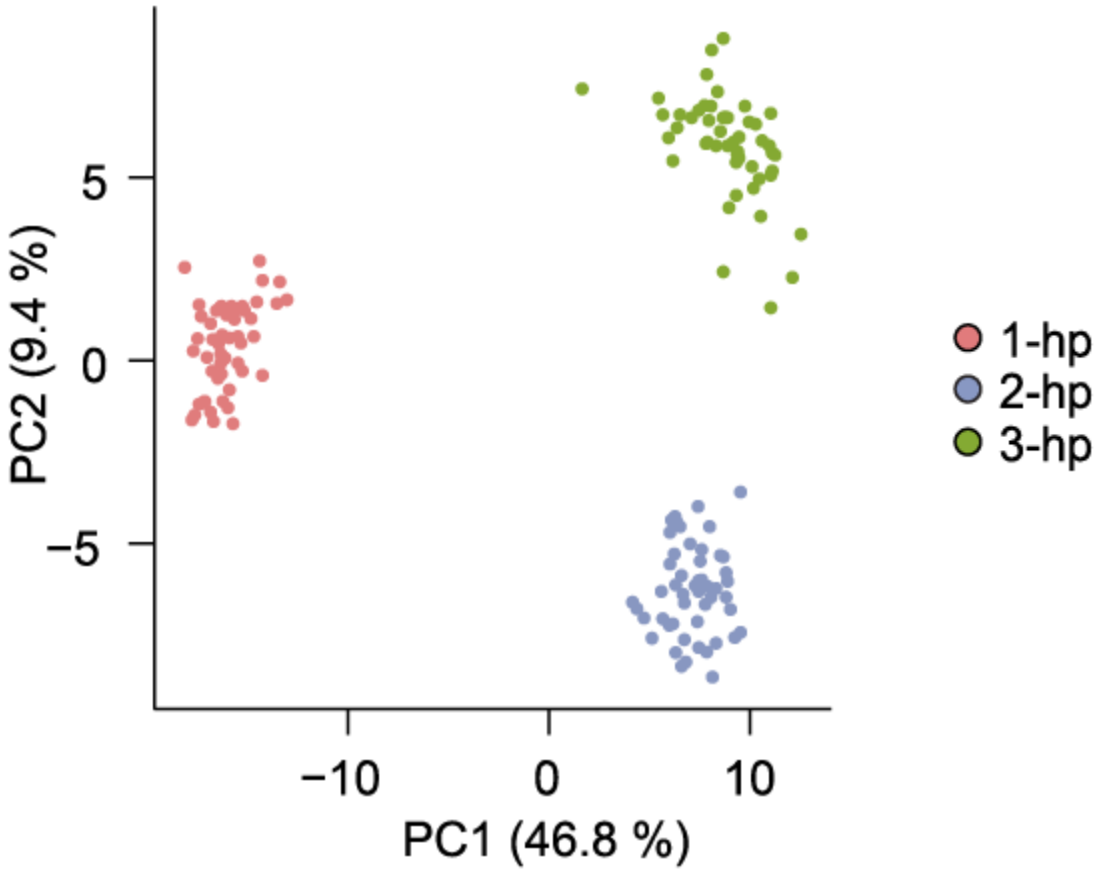
Motifs cluster within the feature space



Motifs cluster within the feature space



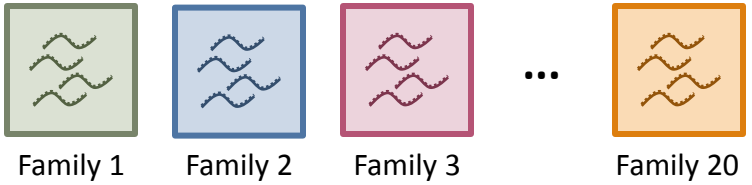
Motifs cluster within the feature space



Does this work with real structures?

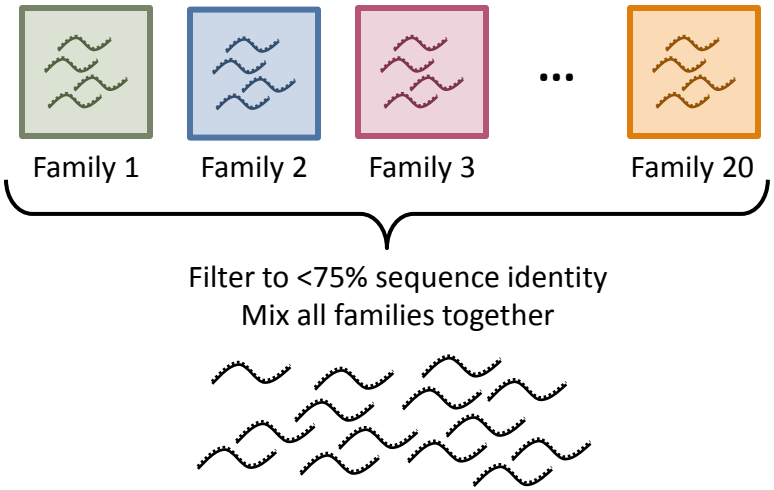
Benchmark: 20 Known Structure Families

Family	# Seqs	Avg % identity
5S_rRNA	100	49%
5_8S_rRNA	22	54%
U1	20	48%
U2	70	47%
tRNA	100	40%
Vault	52	50%
U12	27	46%
Hammerhead_3	13	45%
RNaseP_nuc	68	32%
RNaseP_bact_a	100	49%
RNaseP_bact_b	41	53%
U3	38	41%
6S	86	38%
U4	61	45%
SNORD14	7	44%
Metazoa_SRP	17	45%
CsrB	7	53%
Y_RNA	24	47%
U5	82	44%
Histone3	43	45%



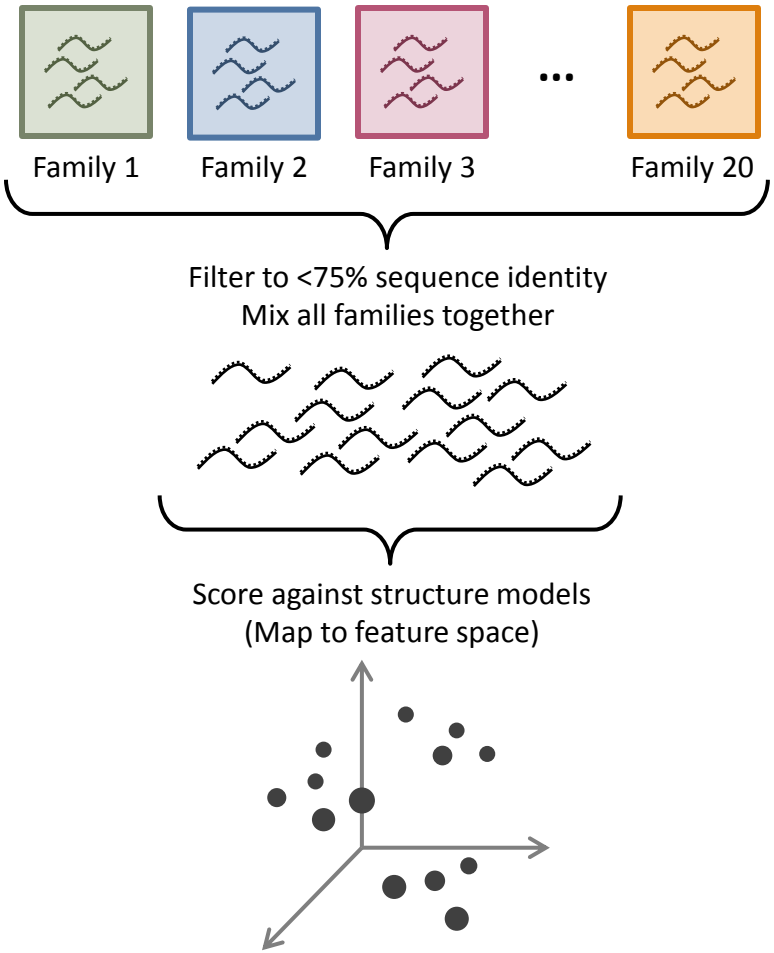
Benchmark: 20 Known Structure Families

Family	# Seqs	Avg % identity
5S_rRNA	100	49%
5_8S_rRNA	22	54%
U1	20	48%
U2	70	47%
tRNA	100	40%
Vault	52	50%
U12	27	46%
Hammerhead_3	13	45%
RNaseP_nuc	68	32%
RNaseP_bact_a	100	49%
RNaseP_bact_b	41	53%
U3	38	41%
6S	86	38%
U4	61	45%
SNORD14	7	44%
Metazoa_SRP	17	45%
CsrB	7	53%
Y_RNA	24	47%
U5	82	44%
Histone3	43	45%



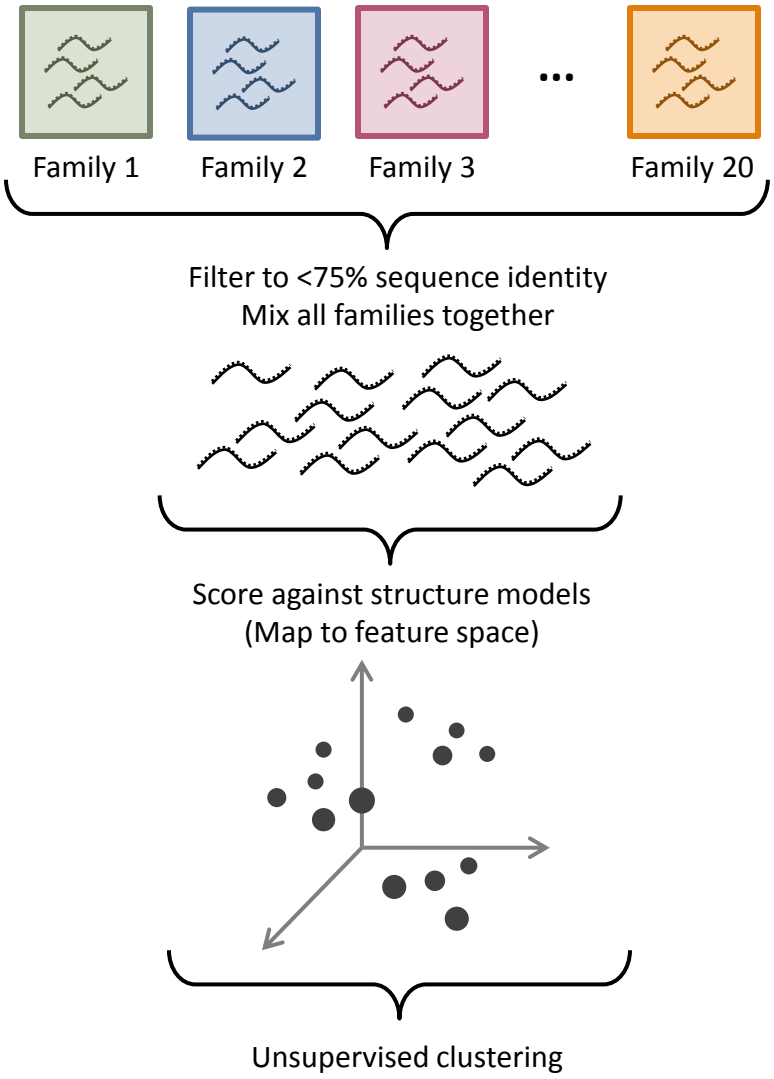
Benchmark: 20 Known Structure Families

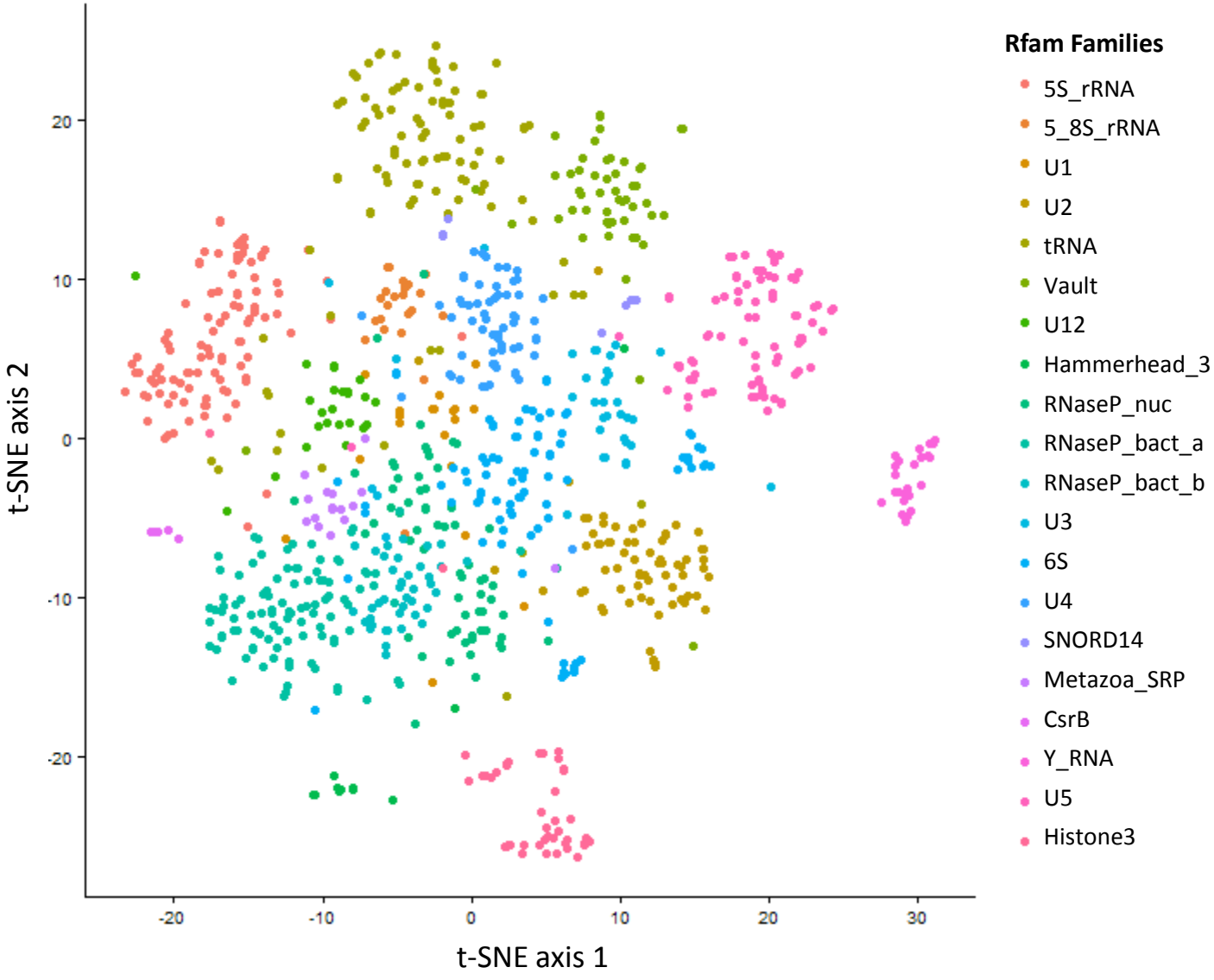
Family	# Seqs	Avg % identity
5S_rRNA	100	49%
5_8S_rRNA	22	54%
U1	20	48%
U2	70	47%
tRNA	100	40%
Vault	52	50%
U12	27	46%
Hammerhead_3	13	45%
RNaseP_nuc	68	32%
RNaseP_bact_a	100	49%
RNaseP_bact_b	41	53%
U3	38	41%
6S	86	38%
U4	61	45%
SNORD14	7	44%
Metazoa_SRP	17	45%
CsrB	7	53%
Y_RNA	24	47%
U5	82	44%
Histone3	43	45%



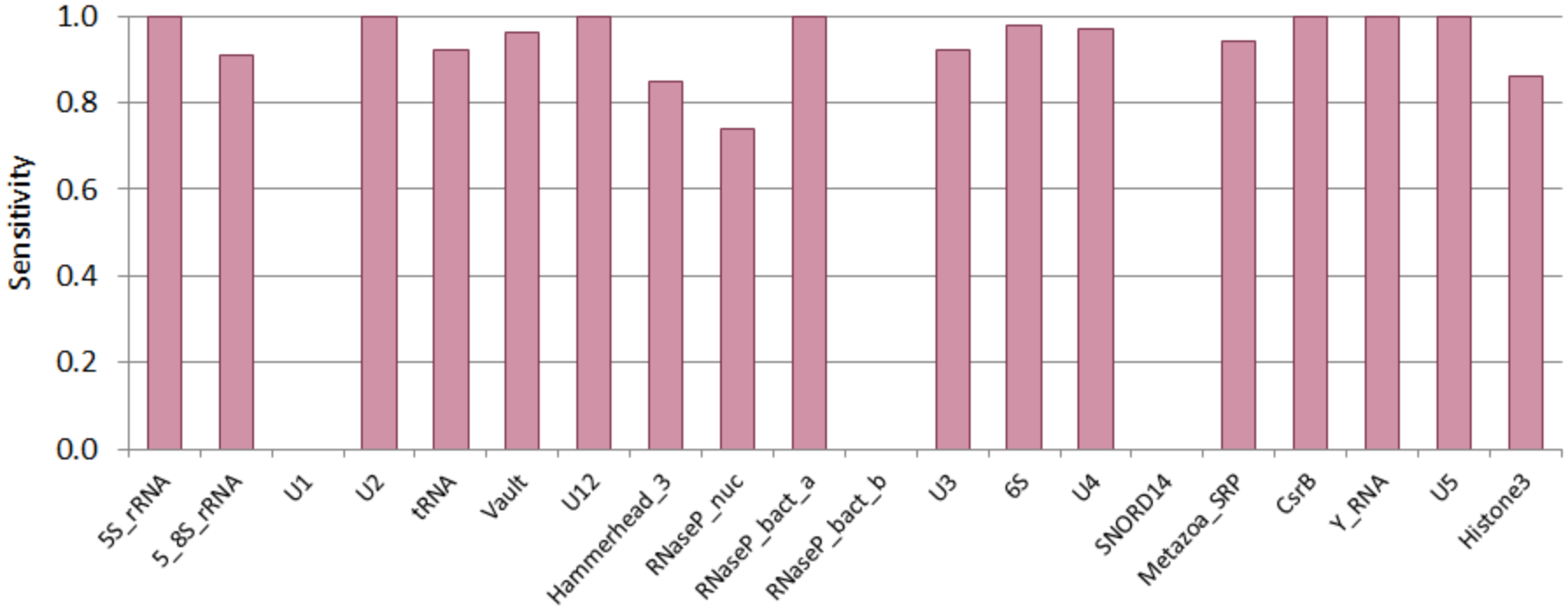
Benchmark: 20 Known Structure Families

Family	# Seqs	Avg % identity
5S_rRNA	100	49%
5_8S_rRNA	22	54%
U1	20	48%
U2	70	47%
tRNA	100	40%
Vault	52	50%
U12	27	46%
Hammerhead_3	13	45%
RNaseP_nuc	68	32%
RNaseP_bact_a	100	49%
RNaseP_bact_b	41	53%
U3	38	41%
6S	86	38%
U4	61	45%
SNORD14	7	44%
Metazoa_SRP	17	45%
CsrB	7	53%
Y_RNA	24	47%
U5	82	44%
Histone3	43	45%




t-SNE axis 2
20
10
0
-10
-20
-20
-10
0
10
20
30
t-SNE axis 1
Rfam Families
5S_rRNA
5_8S_rRNA
U1
U2
tRNA
Vault
U12
Hammerhead_3
RNaseP_nuc
RNaseP_bact_a
RNaseP_bact_b
U3
6S
U4
SNORD14
Metazoa_SRP
CsrB
Y_RNA
U5
Histone3

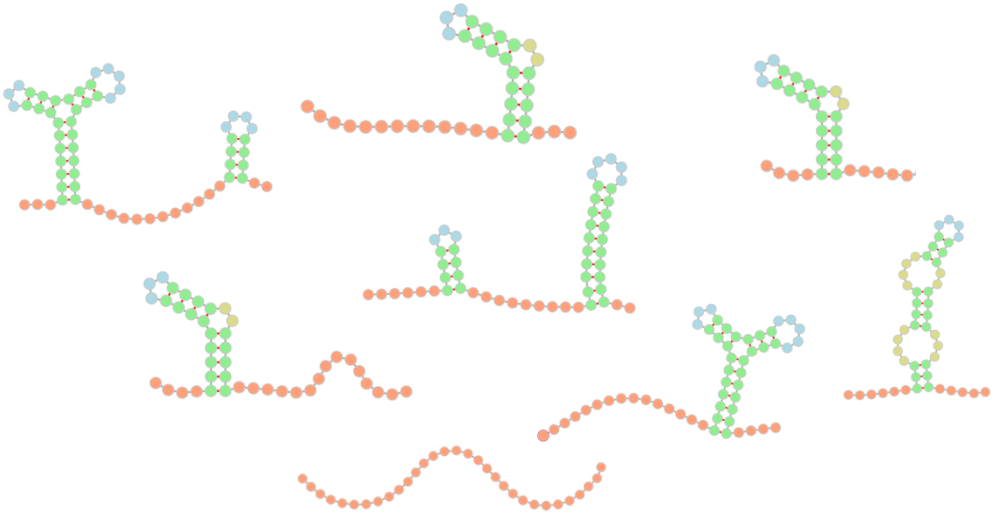
Clustering performance



Average sensitivity: 0.80
Average specificity: 0.98

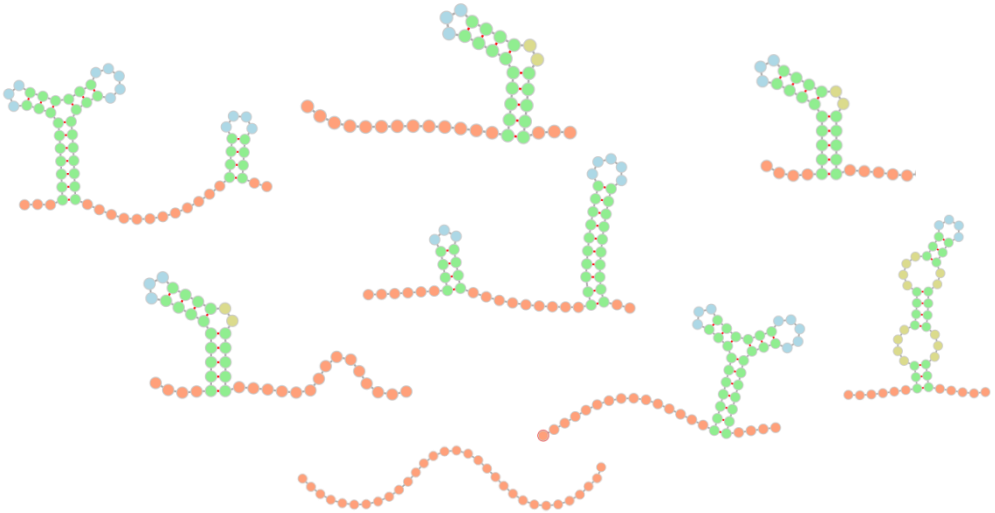
Complications in real datasets:

- A sequence may have no motifs
- Boundary of individual motifs is unknown

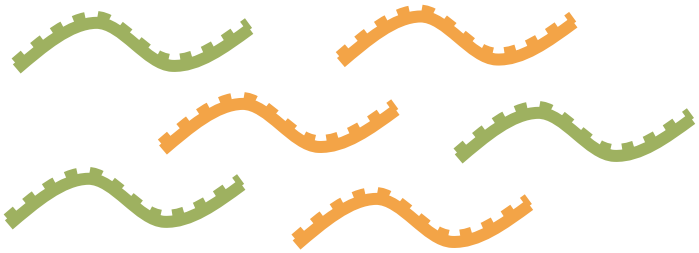


Complications in real datasets:

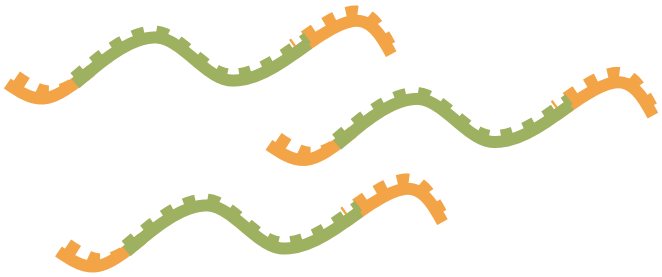
- A sequence may have no motifs
- Boundary of individual motifs is unknown



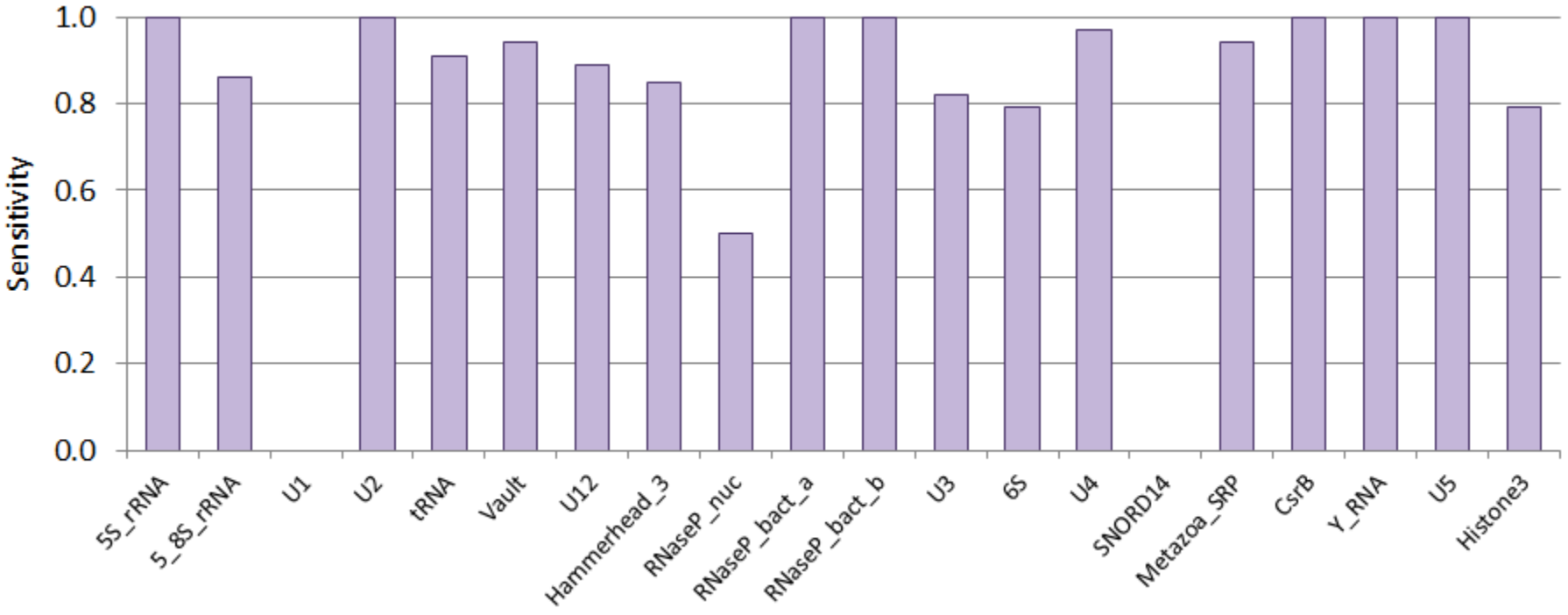
Added decoys

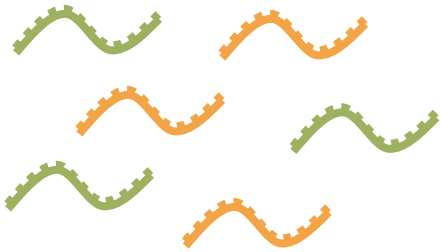


Embedded sequences



Performance: With decoys

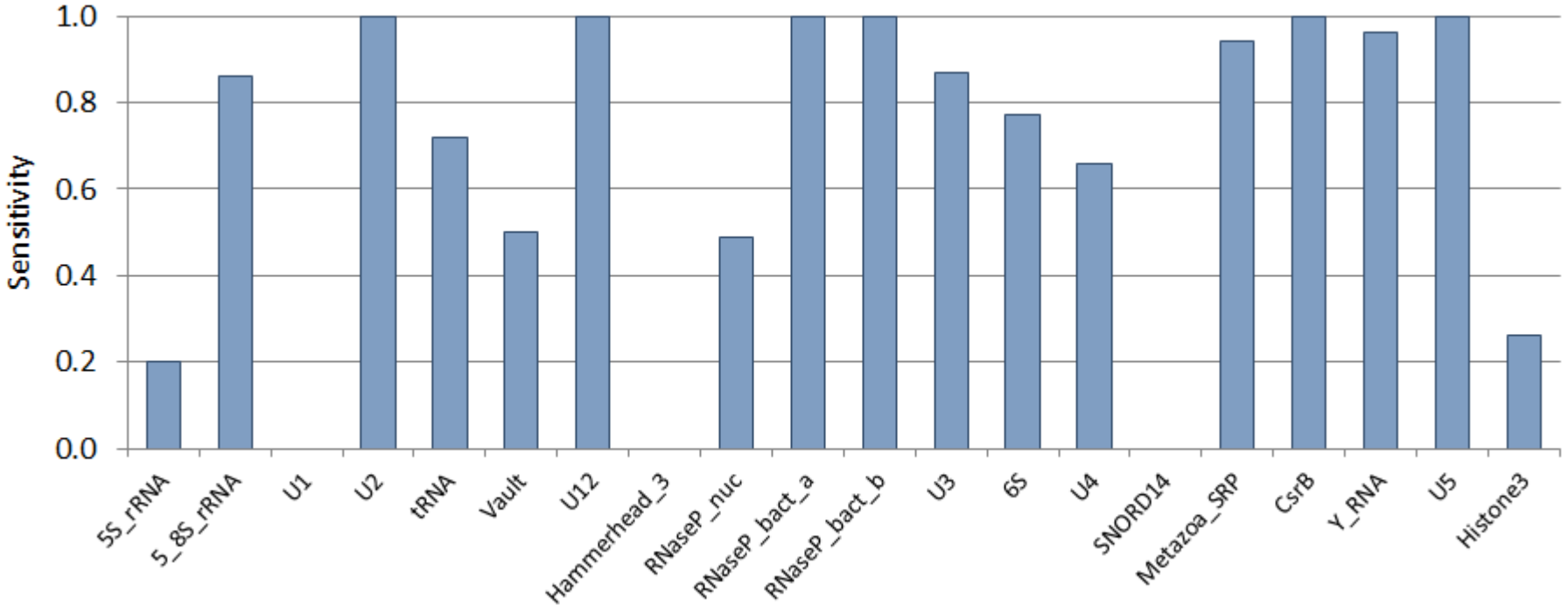


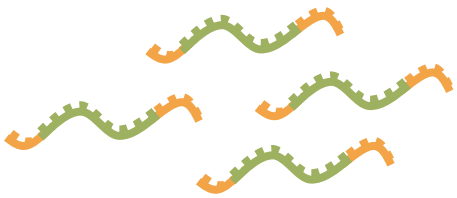


Average sensitivity: 0.81
Average specificity: 0.99

0 decoy clusters

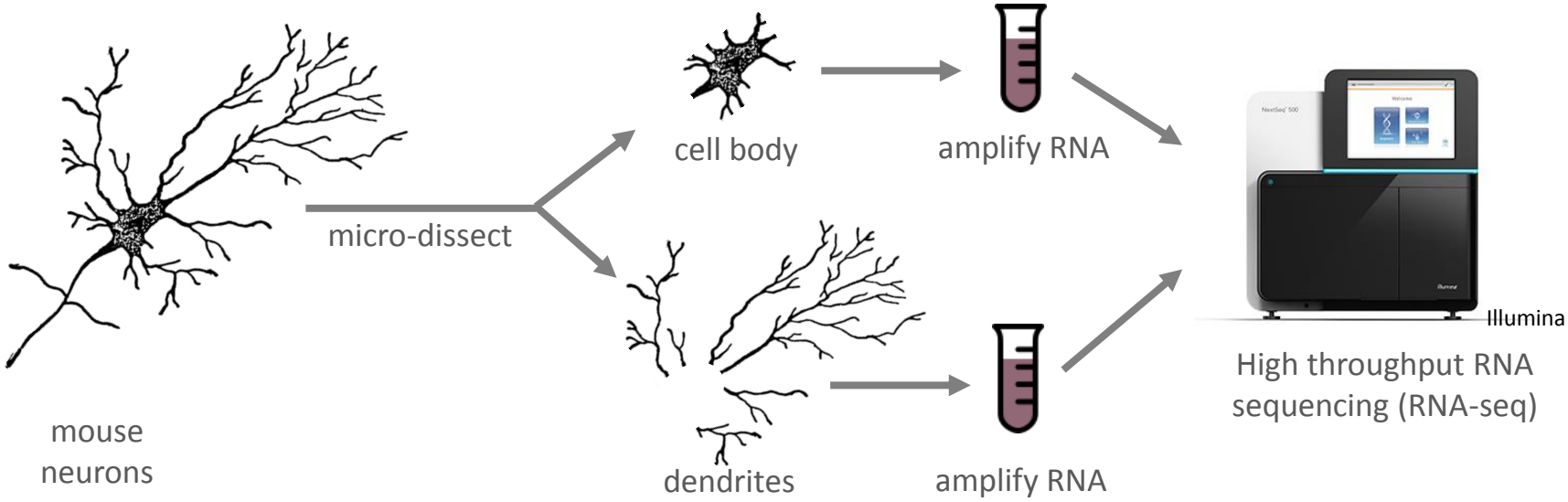
Performance: Embedded sequences

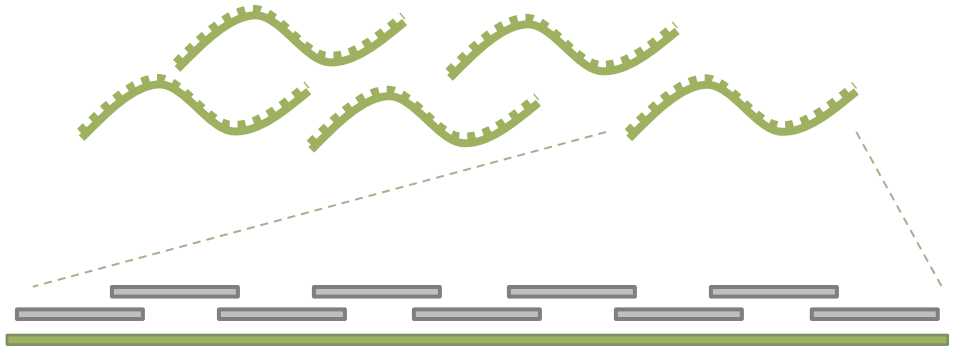




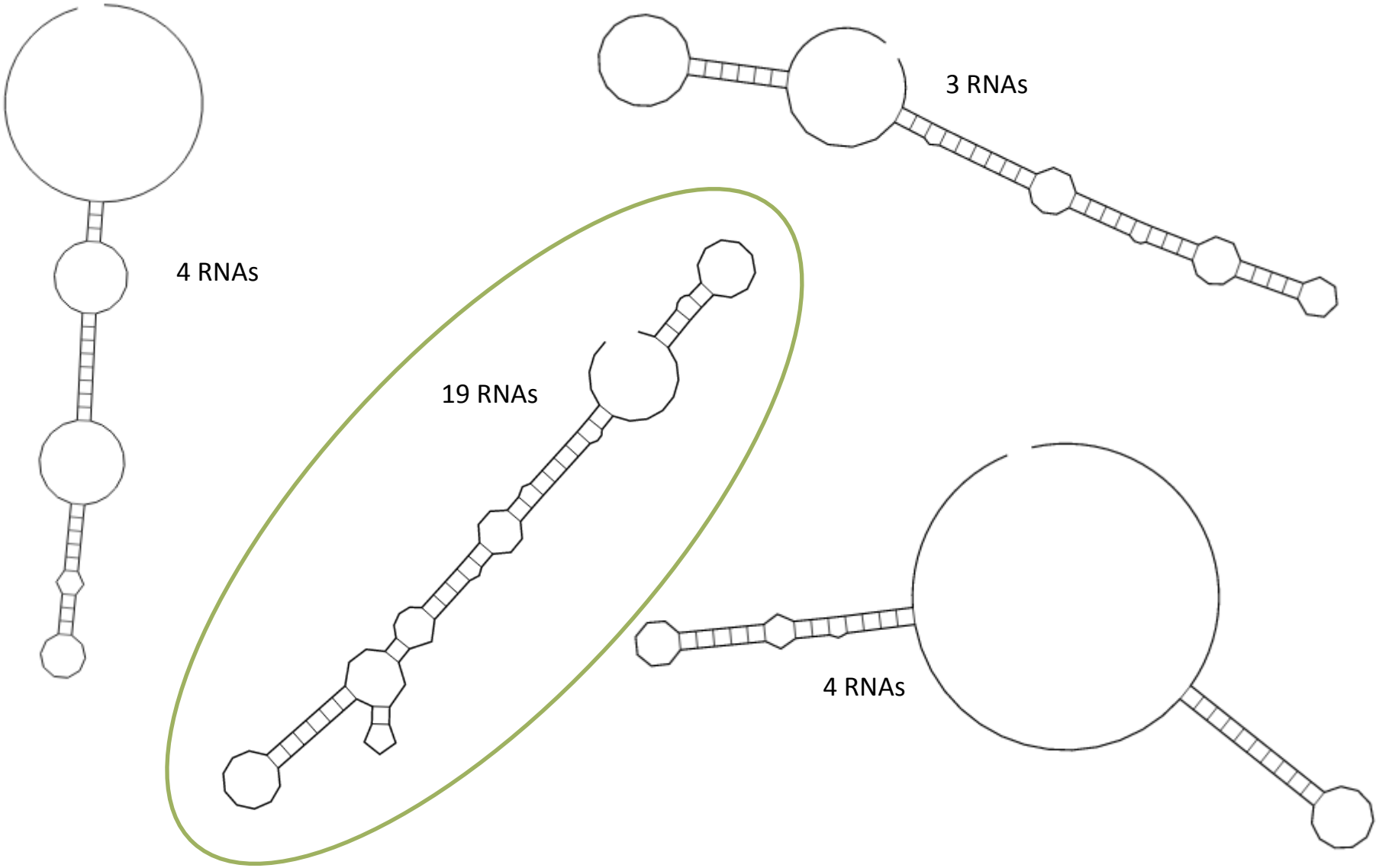
Average sensitivity: 0.66
Average specificity: 0.99

Can we use this method to find dendritic localization motifs?


micro-dissect
cell body
amplify RNA
dendrites
amplify RNA
mouse
neurons
Illumina
High throughput RNA
sequencing (RNA-seq)



Discovered motifs



Acknowledgments

Kim Lab

Junhyong Kim
Derek Stefanik
Gehoon Chung
Hannah Dueck
Jamie Shallcross
Jean Rosario
Qin Zhu
Stephen Fisher
Syung-Hun Han
Youngji Na

Chantal Francis
Hoa Giang
Mugdha Khaladkar

Eberwine Lab

Jim Eberwine
Jenn Singh
Jinhui Wang

Special thanks to CSGF program & Krell for this opportunity!

DOE CSGF

Structure pictures generated using various services at http://rna.tbi.univie.ac.at/