Generative modeling of RNA switches with RBMs

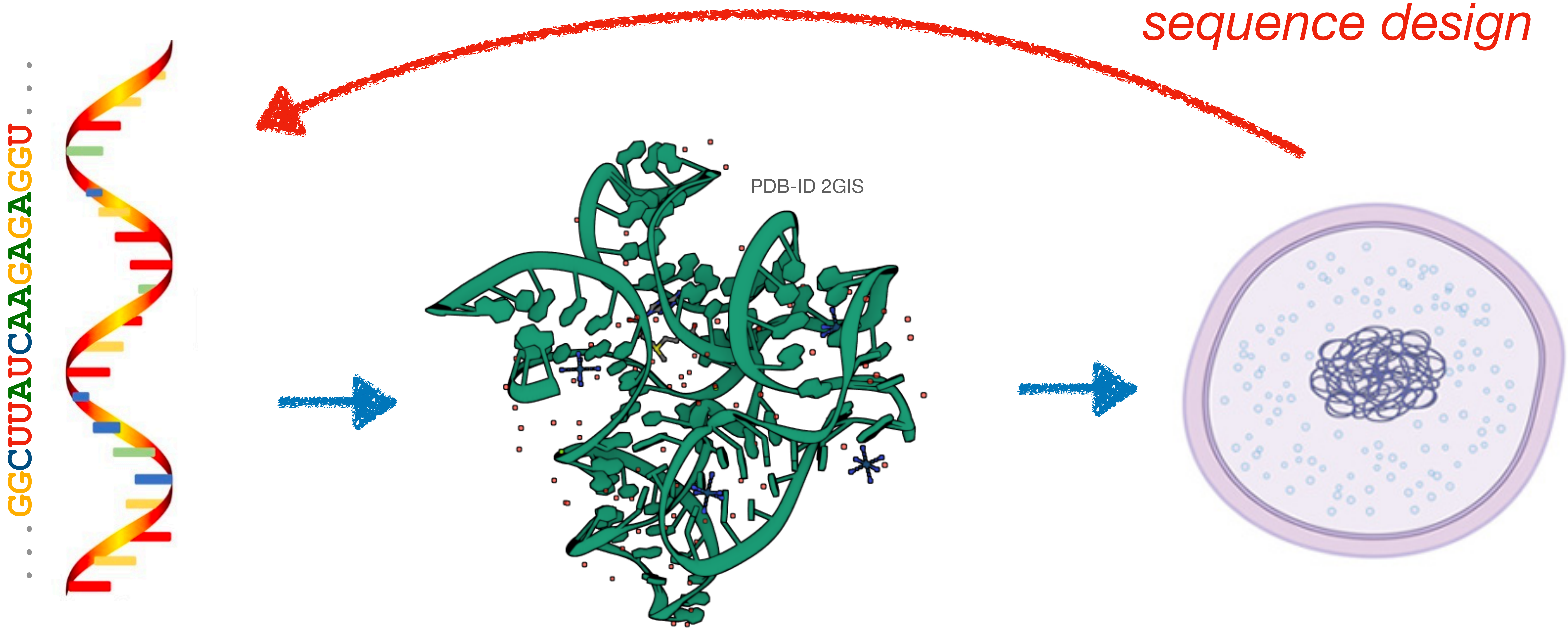
Jorge Fernandez-de-Cossio-Diaz

Paris, ENS

In collaboration with: P. Hardouin, F.-X. Lyonnet du Moutier, A. Di Gioacchino, B. Marchand, Y. Ponty, B. Sargueil, R. Monasson, S. Cocco

Computational approaches to RNA structure and function
Centro de Ciencias de Benasque
2024-07-24

Structure and function of non-coding RNA molecules



sequence

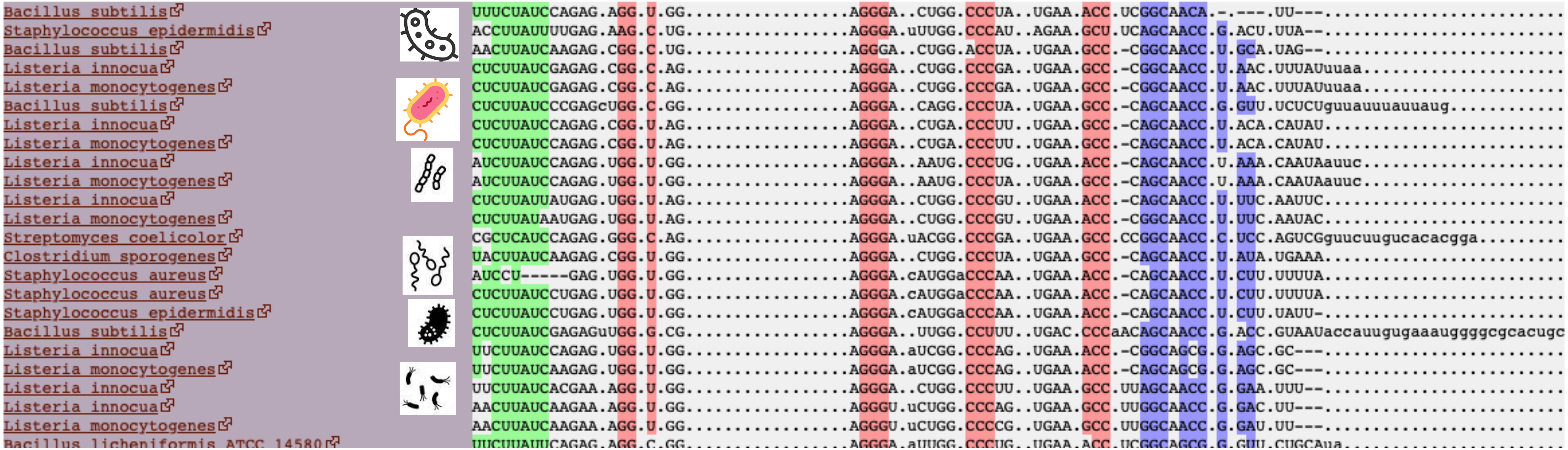
3d-structure

Biological function in the cell

Problem: sequence ↔ function mapping

(Sequence) data-driven approach

Multiple sequence alignment (MSA) of homologs are rich in evolutionary information:
Conservation, covariation, …



```
Bacillus subtilis                 UUUCUAUCCAGAG.AGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.ACC.UCGGCAACA.-.---.UU---........................
Staphylococcus epidermidis        ACCUUAUUUUGAG.AAG.C.UG.................AGGGA.uUUGG.CCCAU..AGAA.GCU.UCAGCAACC.G.ACU.UUA--........................
Bacillus subtilis                 AACUUAUCAAGAG.CGG.C.UG.................AGGGA..CUGG.ACCUA..UGAA.GCC.-CGGCAACC.U.GCA.UAG--........................
Listeria innocua                  CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa....................
Listeria monocytogenes            CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa....................
Bacillus subtilis                 CUCUUAUCCCGAGcUGG.C.GG.................AGGGA..CAGG.CCCUA..UGAA.GCC.-CAGCAACC.G.GUU.UCUCUguuauuuauuaug...........
Listeria innocua                  CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU........................
Listeria monocytogenes            CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU........................
Listeria innocua                  AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUG..UGAA.ACC.-CAGCAACC.U.AAA.CAAUAauuc....................
Listeria monocytogenes            AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUA..UGAA.GCC.-CAGCAACC.U.AAA.CAAUAauuc....................
Listeria innocua                  CUCUUAUUAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CAGCAACC.U.UUC.AAUUC........................
Listeria monocytogenes            CUCUUAUAAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CGGCAACC.U.UUC.AAUAC........................
Streptomyces coelicolor           CGCUCAUCCAGAG.GGG.C.AG.................AGGGA.uACGG.CCCGA..UGAA.GCC.CCGGCAACC.C.UCC.AGUCGguucuugucacacgga........
Clostridium sporogenes            UACUUAUCAAGAG.CGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.GCC.-CAGCAACC.U.AUA.UGAAA........................
Staphylococcus aureus             AUCCU-----GAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA........................
Staphylococcus aureus             CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA........................
Staphylococcus epidermidis        CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UAUU-........................
Bacillus subtilis                 CUCUUAUCGAGAGuUGG.G.CG.................AGGGA..UUGG.CCUUU..UGAC.CCCaACAGCAACC.G.ACC.GUAAUaccauugugaaaugggcgcacugc
Listeria innocua                  UUCUUAUCCAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CGGCAGCG.G.AGC.GC---........................
Listeria monocytogenes            UUCUUAUCAAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CAGCAGCG.G.AGC.GC---........................
Listeria innocua                  UUCUUAUCACGAA.AGG.U.GG.................AGGGA..CUGG.CCCUU..UGAA.GCC.UUAGCAACC.G.GAA.UUU--........................
Listeria innocua                  AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCAG..UGAA.GCC.UUGGCAACC.G.GAC.UU---........................
Listeria monocytogenes            AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCCG..UGAA.GCC.UUGGCAACC.G.GAU.UU---........................
Bacillus licheniformis ATCC 14580 UUCUUAUUCAGAG.AGG.C.GG.................AGGGA.aUUGG.CCCUG..UGAA.ACC.UCGGCAGCG.G.GUU.CUGCAua......................
```

Models trained on MSA data can infer constraints necessary for function.
*We can then test (experimentally) their **generative** capabilities.*

S. Cocco et al 2018 Rep. Prog. Phys. 81 032601; W. P. Russ, et al. Science 369.6502 (2020): 440-445.

SAM-I riboswitch aptamer domain

Upstream of coding part of mRNA.

Recognizes a metabolite (SAM) specifically

SAM = S-adenosyl methionine

Regulates methionine metabolism in bacteria

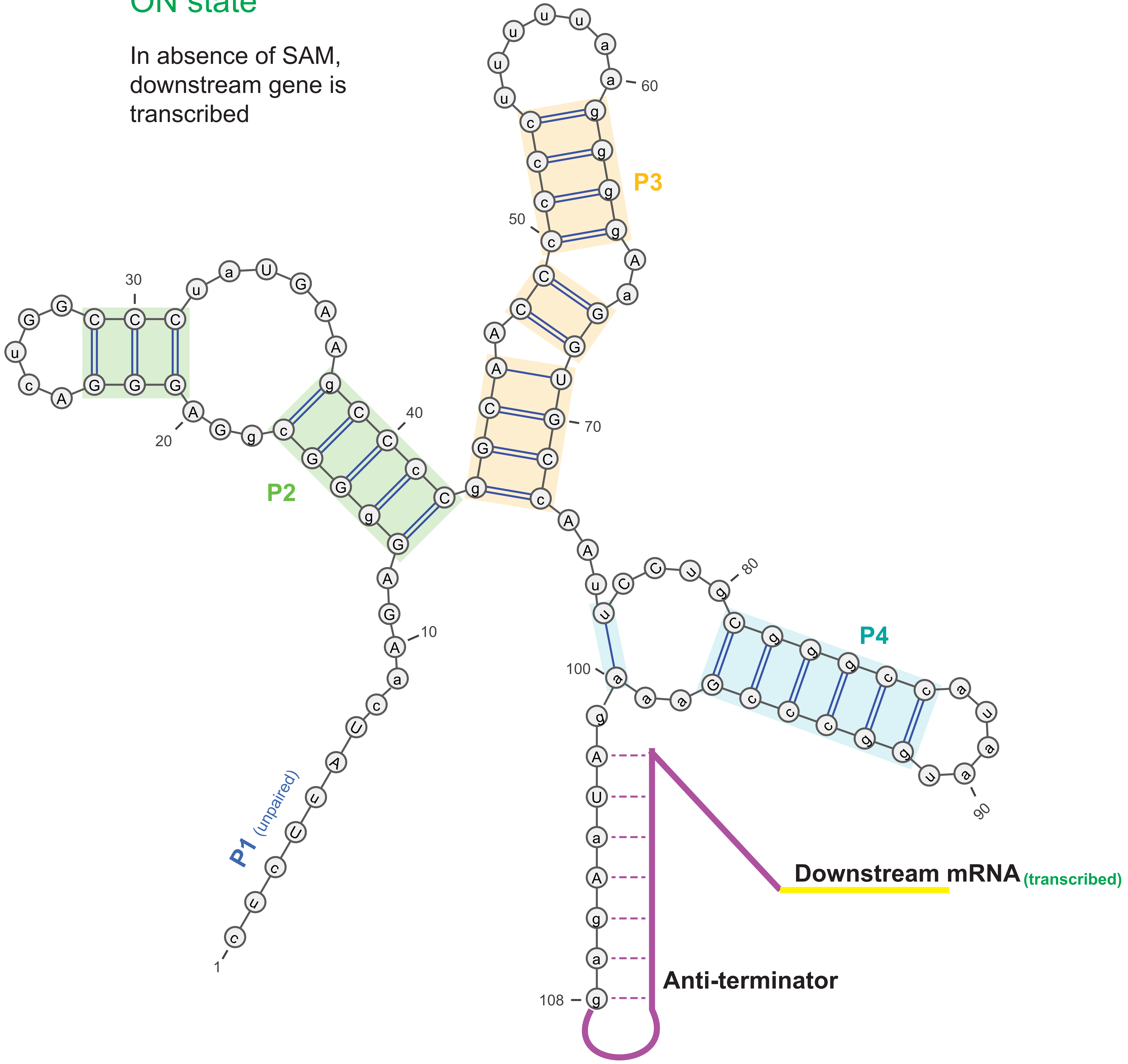
Model system for *riboswitches*, which regulate gene expression in bacteria and eukaryotes, in response to specific metabolites

SAM-I riboswitch aptamer

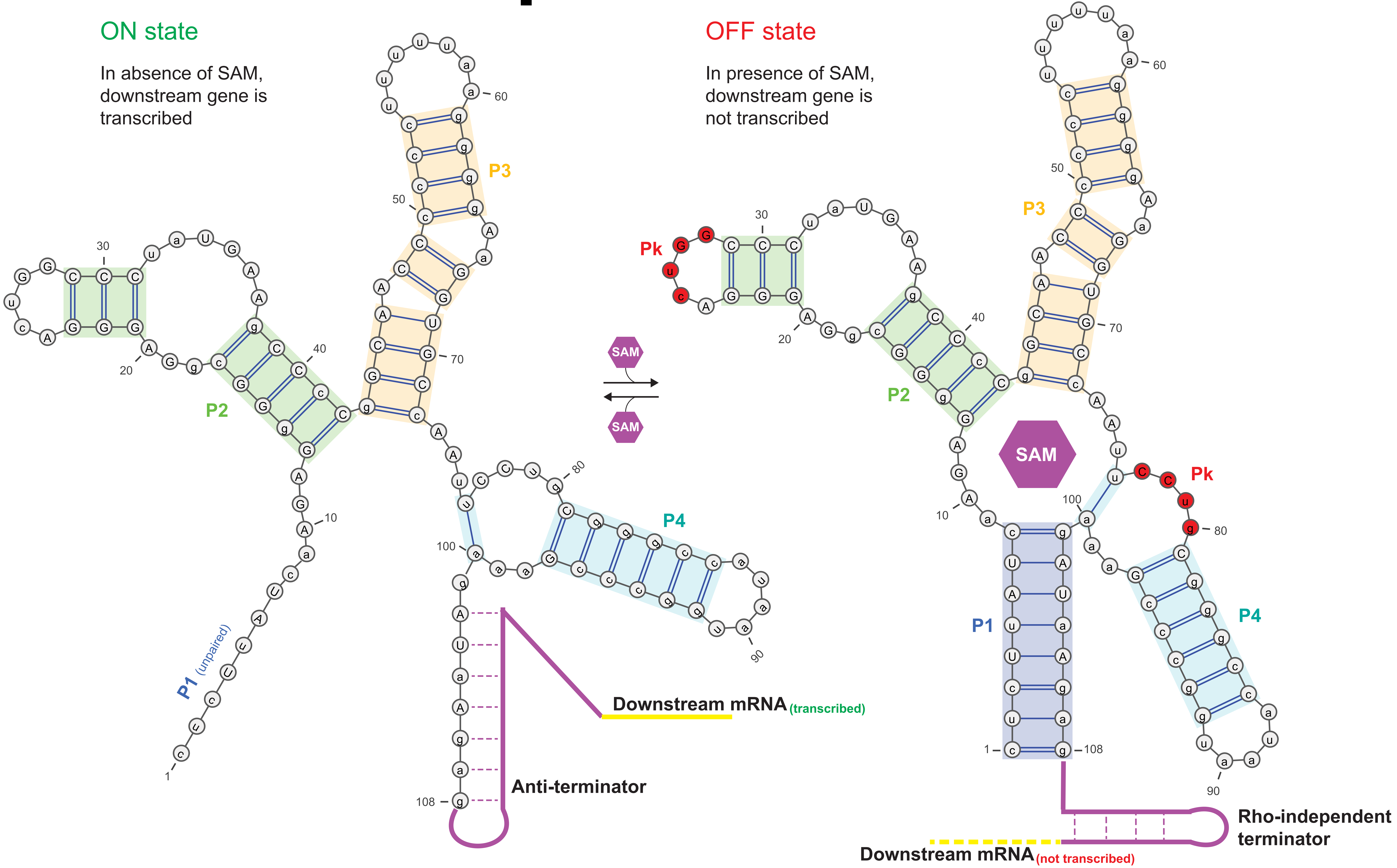
SAM-I riboswitch aptamer domain

ON state

In absence of SAM,
downstream gene is
transcribed



SAM-I riboswitch aptamer domain



ON state

In absence of SAM, downstream gene is transcribed

OFF state

In presence of SAM, downstream gene is not transcribed

SAM-I riboswitch aptamer domain

Challenges for successful riboswitch regulatory function:

- network of **tertiary contacts**
- specific **ligand** binding
- **switch** between competing conformations in response to the ligand
- combinatorially large space of possible sequences $(5^{108} \sim 10^{75})$ and structures ⇒ **difficult to design**

SAM-I riboswitch homologs are found in many different organisms

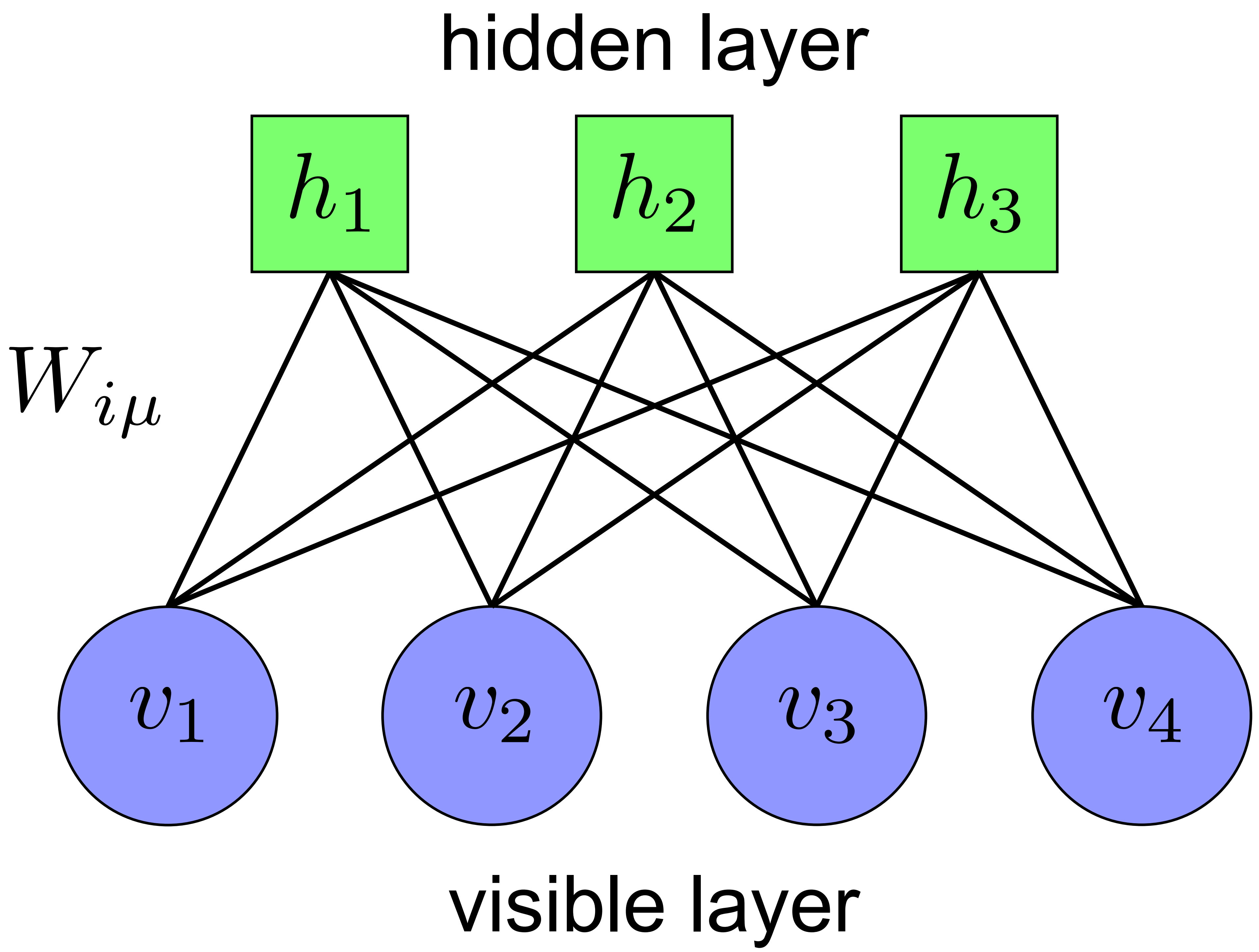
https://rfam.org

RF00162 aptamer domain family

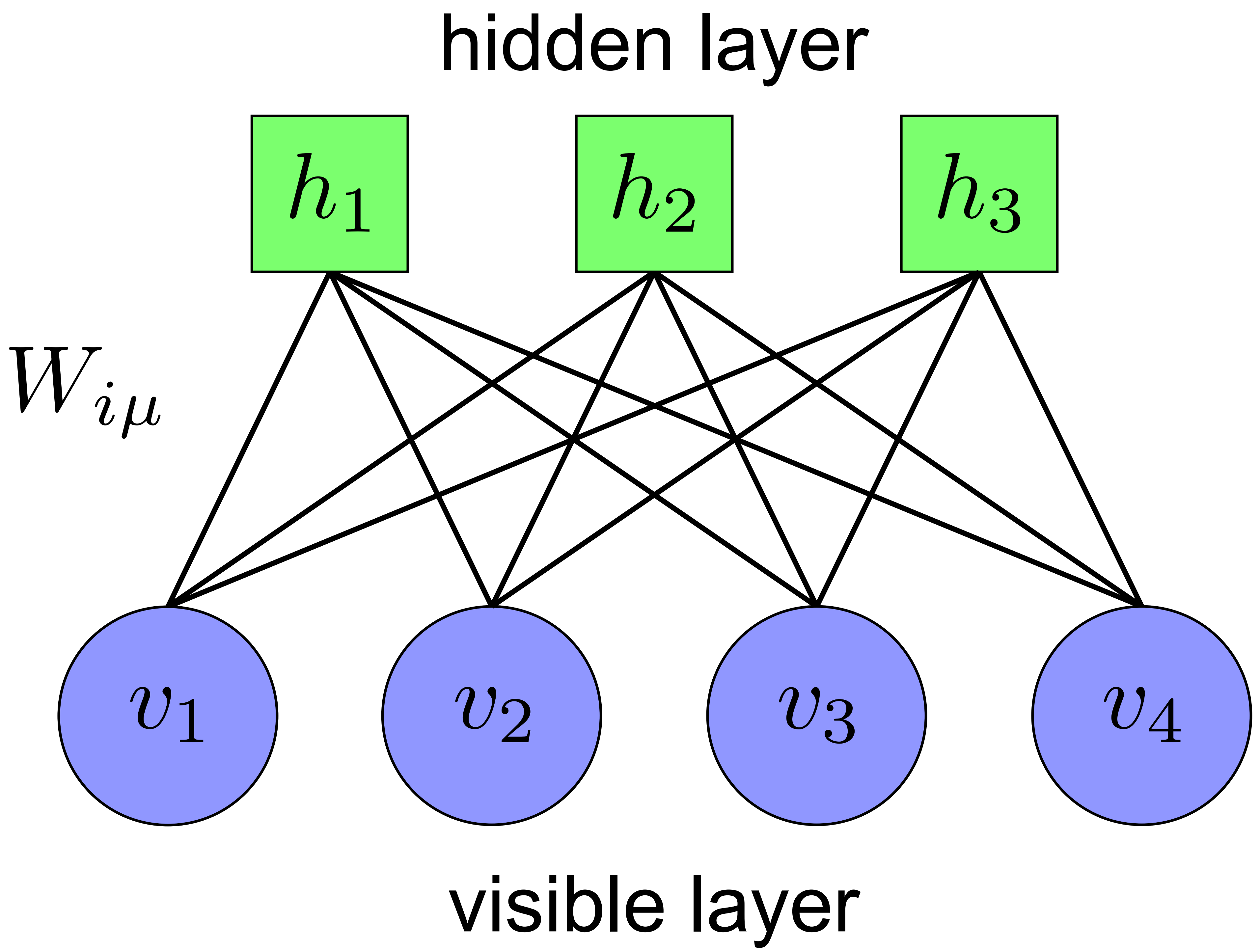
Multiple Sequence Alignment (MSA), consisting of:

6161 sequences, 108 aligned positions

Restricted Boltzmann Machines (RBM)



Restricted Boltzmann Machines (RBM)



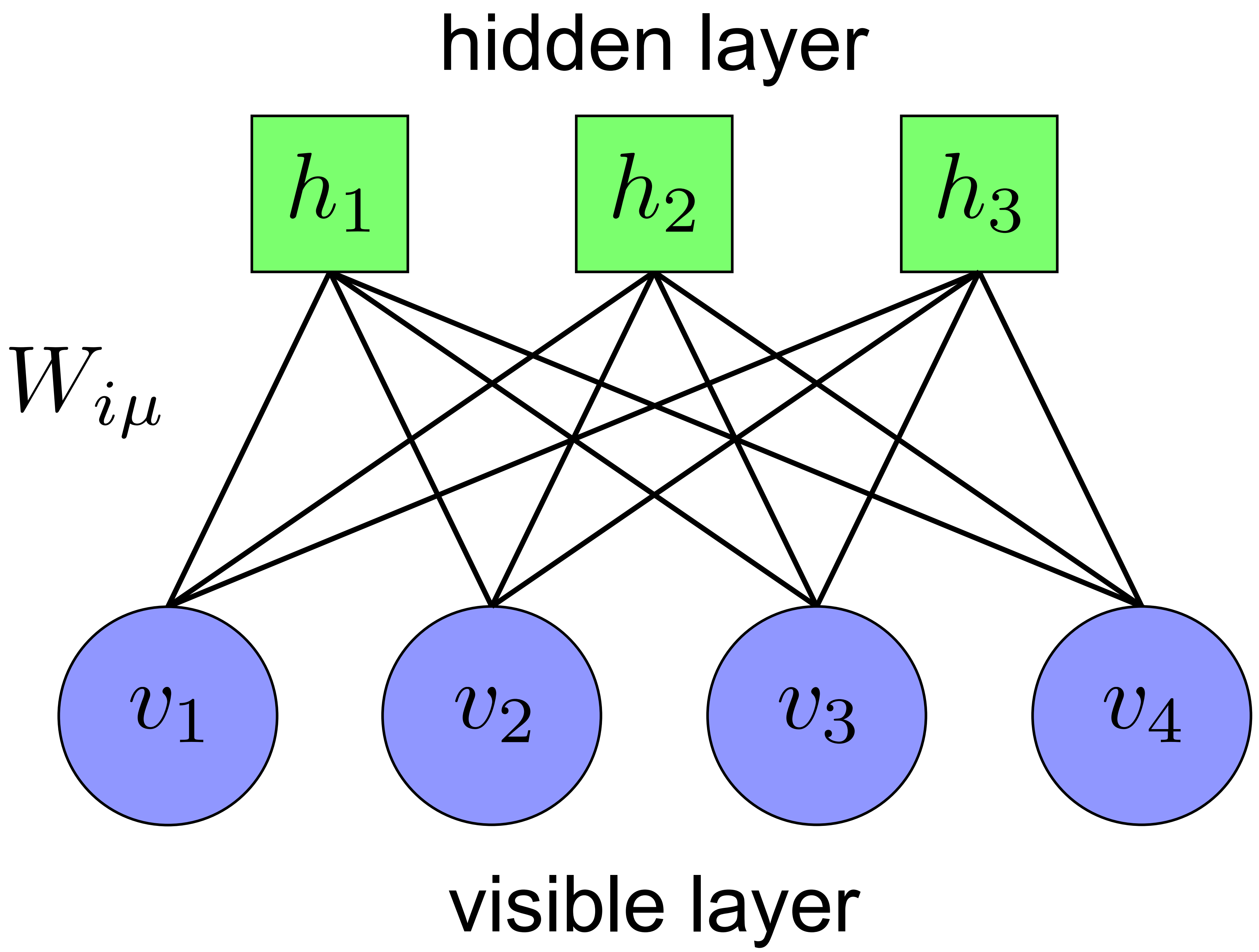
$\mathbf{v} = (v_1, \ldots, v_N)$ = the RNA sequence

$\mathbf{h} = (h_1, \ldots, h_M)$ = latent variables (features)

Energy function:

$$E(\mathbf{v}, \mathbf{h}) = \sum_i \mathscr{V}_i(v_i) + \sum_\mu \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} W_{i\mu} v_i h_\mu$$

Restricted Boltzmann Machines (RBM)



$\mathbf{v} = (v_1, \ldots, v_N)$ = the RNA sequence

$\mathbf{h} = (h_1, \ldots, h_M)$ = latent variables (features)

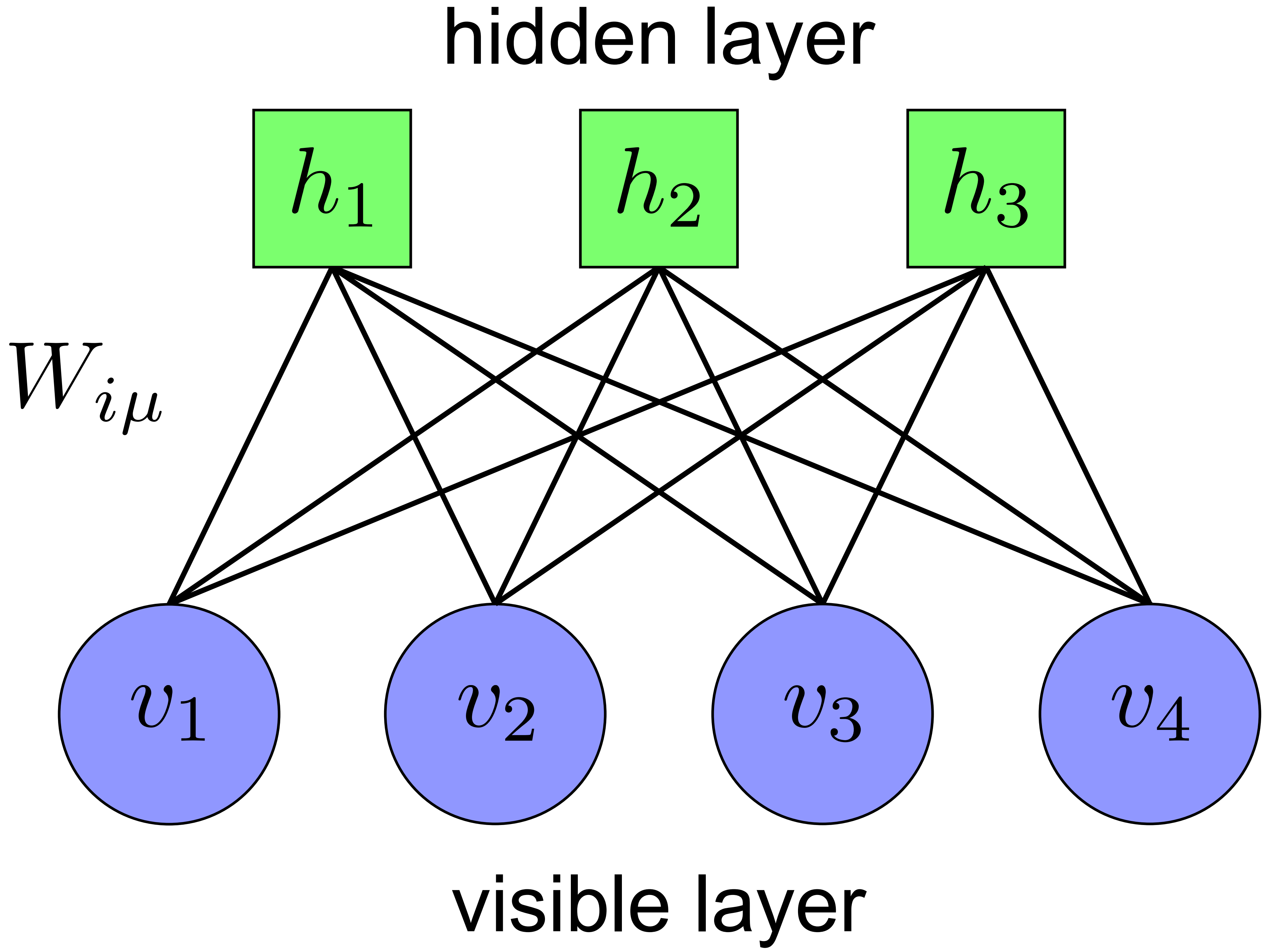
Energy function:

$$E(\mathbf{v}, \mathbf{h}) = \sum_i \mathcal{V}_i(v_i) + \sum_\mu \mathcal{U}_\mu(h_\mu) - \sum_{i\mu} W_{i\mu} v_i h_\mu$$

Boltzmann law:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}$$

Restricted Boltzmann Machines (RBM)



$\mathbf{v} = (v_1, \ldots, v_N)$ = the RNA sequence

$\mathbf{h} = (h_1, \ldots, h_M)$ = latent variables (features)

Energy function:

$$E(\mathbf{v}, \mathbf{h}) = \sum_i \mathscr{V}_i(v_i) + \sum_\mu \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} W_{i\mu} v_i h_\mu$$

Boltzmann law:

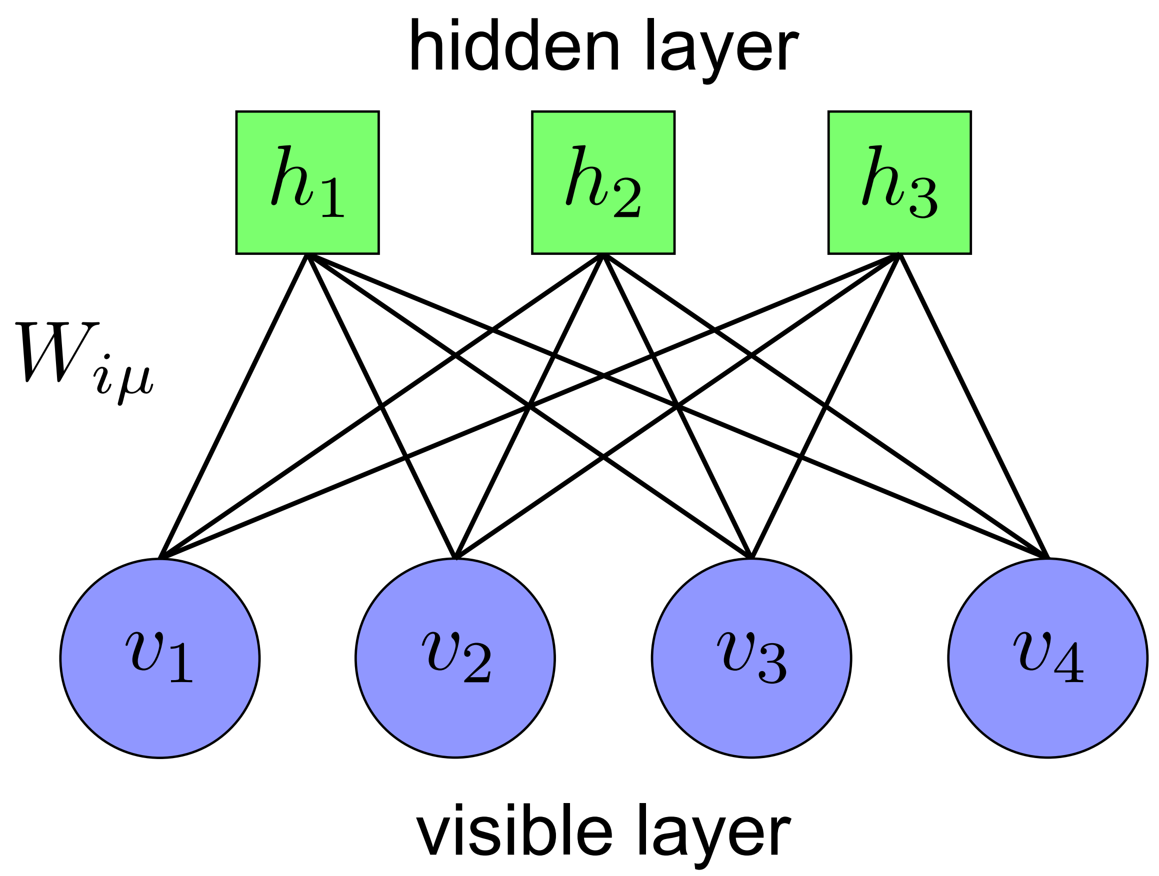
$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}$$

Partition function (normalization constant)

$$Z = \sum_{v_1,\ldots,v_N} \int dh_1 \ldots dh_M \, e^{-E(\mathbf{v},\mathbf{h})}$$

Restricted Boltzmann Machines (RBM)

Sampling



Energy

$$E(\mathbf{v},\mathbf{h}) = \sum_i \mathscr{V}_i(v_i) + \sum_\mu \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} W_{i\mu} v_i h_\mu$$

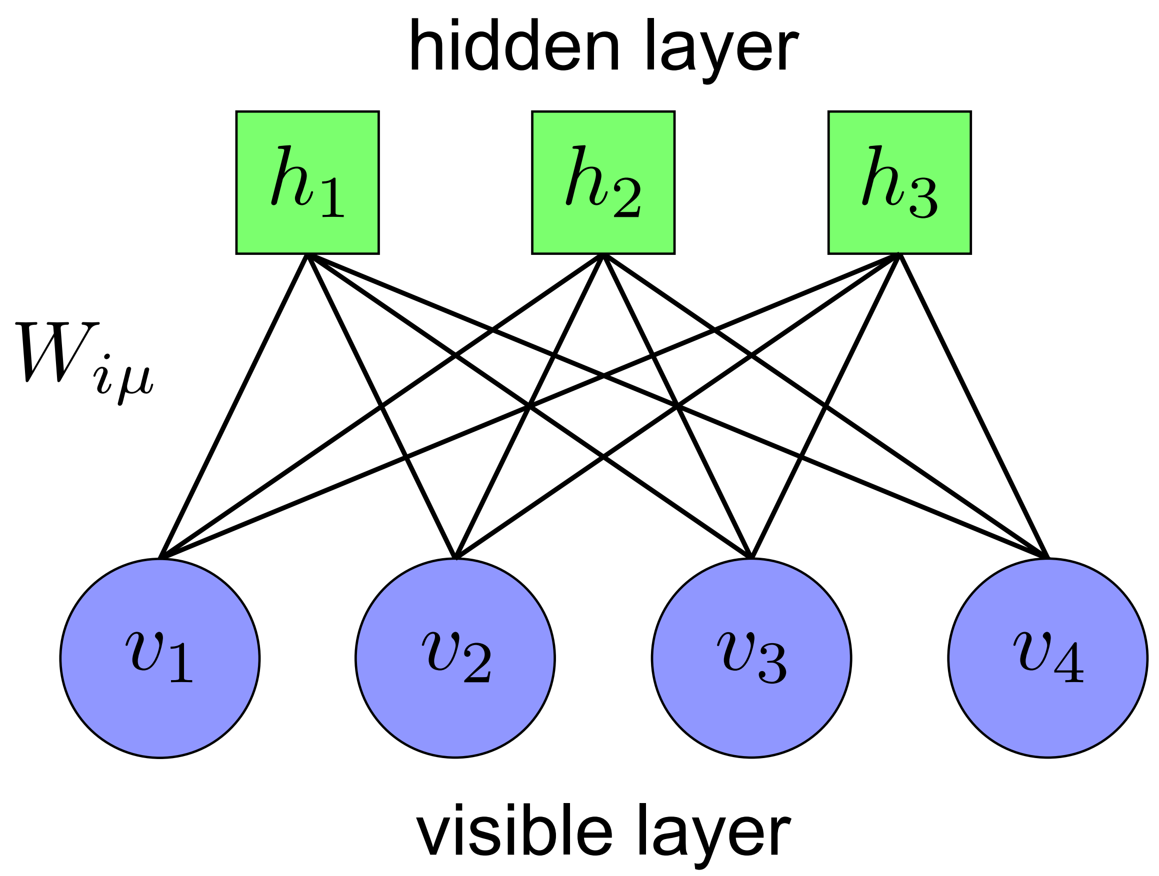
Conditional probabilities factorize

$$P(\mathbf{v}\,|\,\mathbf{h}) = \prod_i P(v_i\,|\,\mathbf{h})$$

$$P(\mathbf{h}\,|\,\mathbf{v}) = \prod_\mu P(h_\mu\,|\,\mathbf{v})$$

Restricted Boltzmann Machines (RBM)

Sampling



Energy

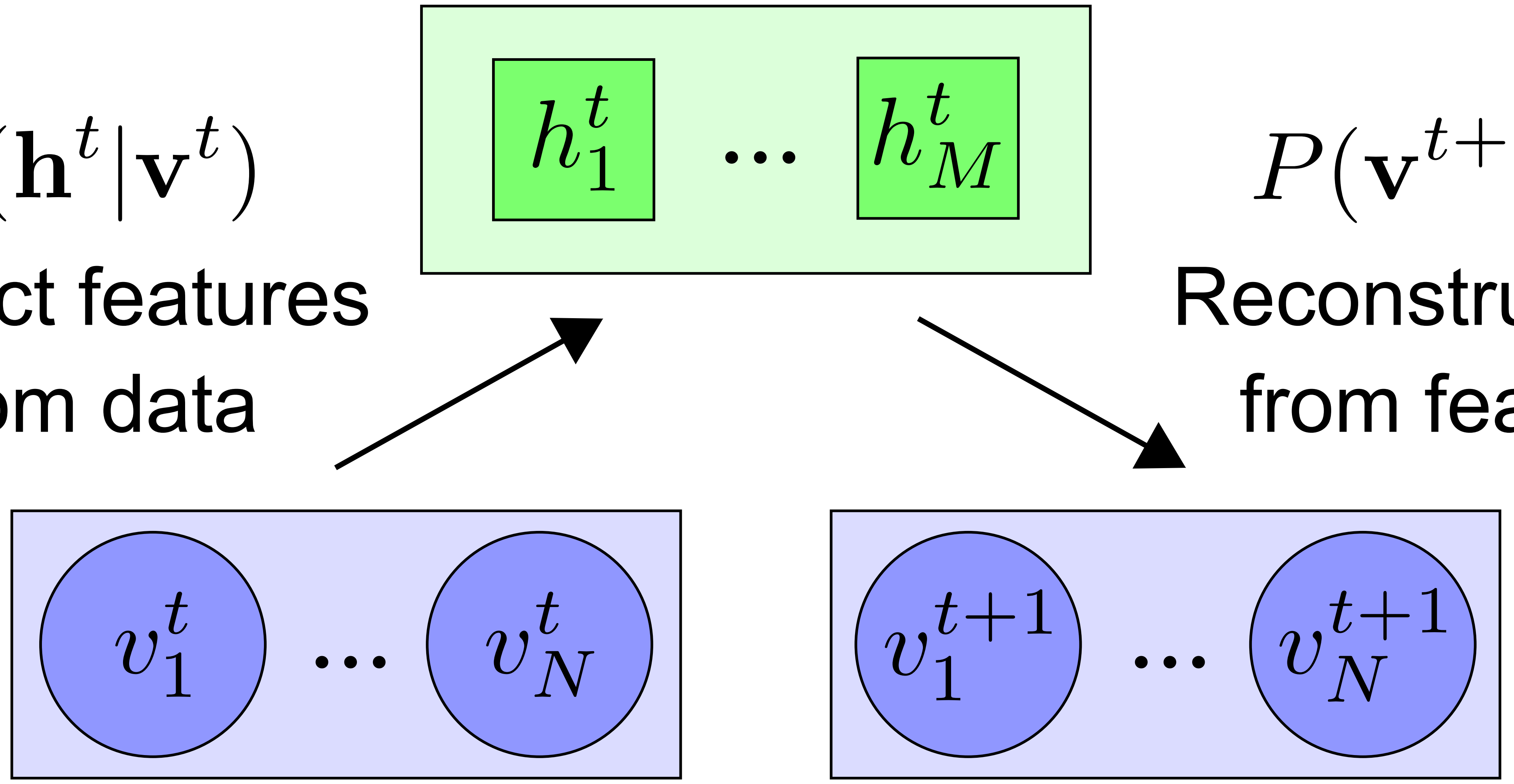
$$E(\mathbf{v},\mathbf{h}) = \sum_i \mathscr{V}_i(v_i) + \sum_\mu \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} W_{i\mu} v_i h_\mu$$

Conditional probabilities factorize

$$P(\mathbf{v}\,|\,\mathbf{h}) = \prod_i P(v_i\,|\,\mathbf{h})$$

$$P(\mathbf{h}\,|\,\mathbf{v}) = \prod_\mu P(h_\mu\,|\,\mathbf{v})$$

Gibbs sampling



$P(\mathbf{h}^t|\mathbf{v}^t)$

Extract features from data

$P(\mathbf{v}^{t+1}|\mathbf{h}^t)$

Reconstruct data from features

Converges to equilibrium samples from the model

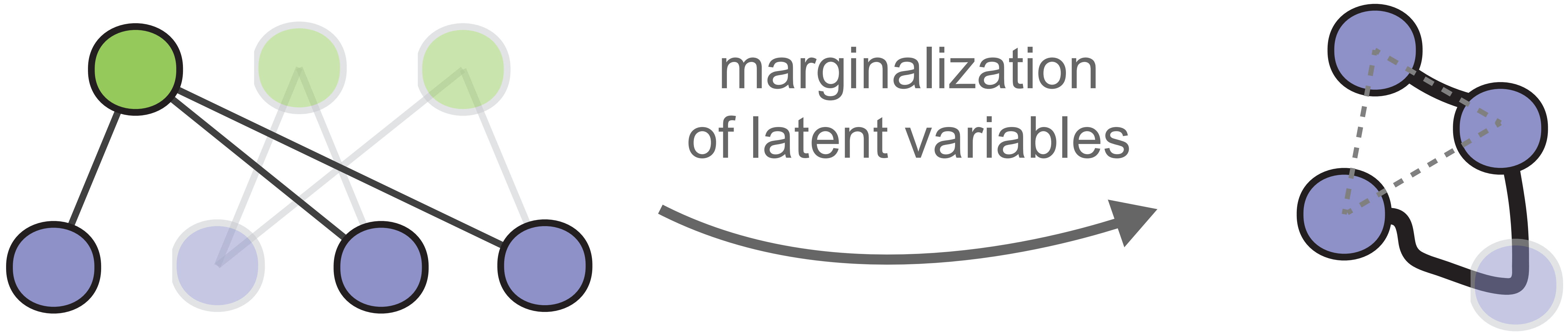
Restricted Boltzmann Machines (RBM)

Likelihood:

$$P_v(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \frac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})}$$

where:

$$E_{\text{eff}}(\mathbf{v}) = \sum_{i} \mathscr{V}_i(v_i) - \sum_{\mu} \ln \int e^{\sum_i W_{i\mu} v_i h_\mu - \mathscr{U}_\mu(h_\mu)} dh_\mu$$



Full RBM

visible units interact
with hidden units

marginalization
of latent variables

Effective model

effective interactions
between visible units

Restricted Boltzmann Machines (RBM)

Likelihood:

$$P_v(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \frac{1}{Z}e^{-E_{\mathrm{eff}}(\mathbf{v})}$$

where:

$$E_{\mathrm{eff}}(\mathbf{v}) = \sum_i \mathscr{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i W_{i\mu} v_i h_\mu - \mathscr{U}_\mu(h_\mu)} dh_\mu$$

RBM are **universal approximators**:

They can model arbitrarily complex functions, provided they have enough hidden units

Le Roux, Nicolas, and Yoshua Bengio. Neural computation 20.6 (2008): 1631-1649.

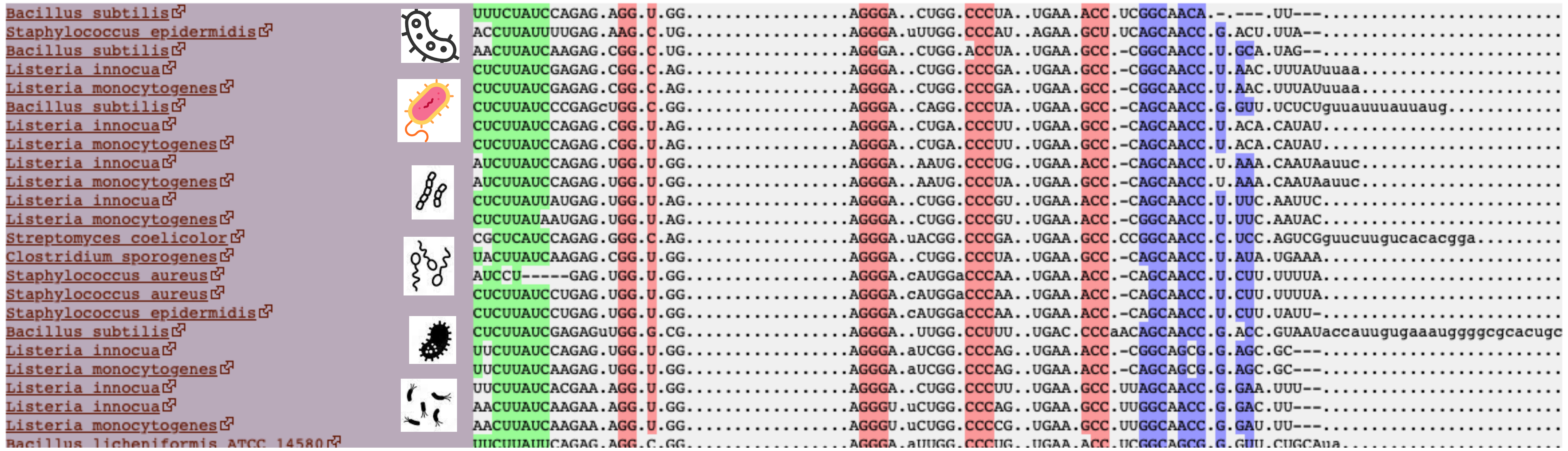
Montufar, Guido, and Nihat Ay. Neural computation 23.5 (2011): 1306-1319.

visible units interact with hidden units

effective interactions between visible units

Restricted Boltzmann Machines (RBM)

MSA (training data):



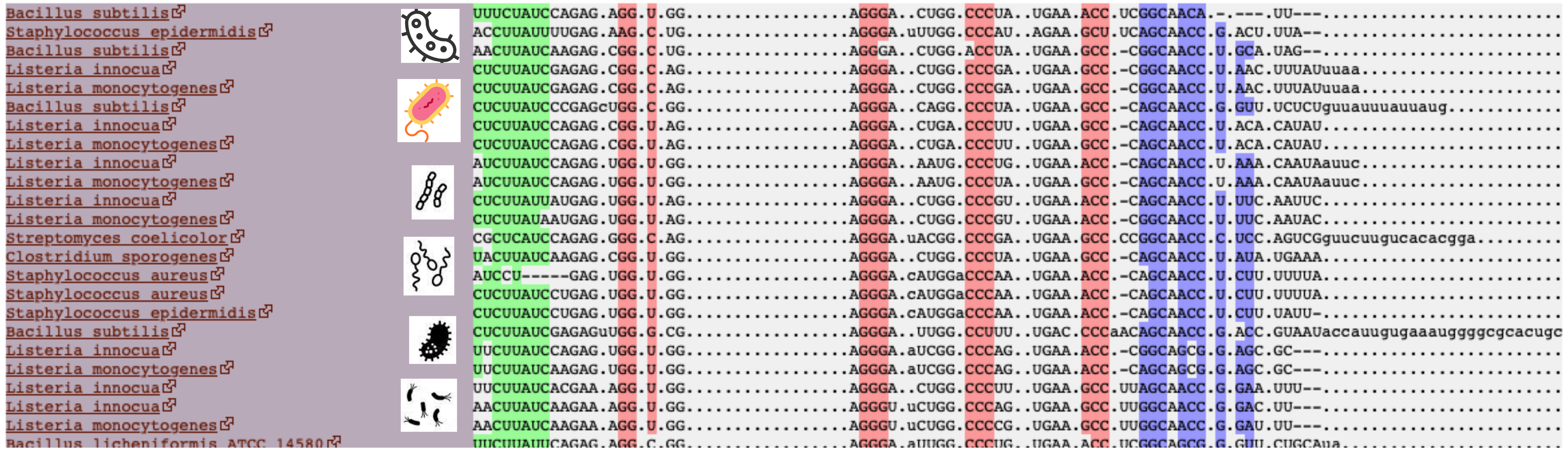
```
Bacillus subtilis                   UUUCUAUCCAGAG.AGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.ACC.UCGGCAACA.-.---.UU---.........................
Staphylococcus epidermidis          ACCUUAUUUUGAG.AAG.C.UG.................AGGGA.uUUGG.CCCAU..AGAA.GCU.UCAGCAACC.G.ACU.UUA--.........................
Bacillus subtilis                   AACUUAUCAAGAG.CGG.C.UG.................AGGGA..CUGG.ACCUA..UGAA.GCC.-CGGCAACC.U.GCA.UAG--.........................
Listeria innocua                    CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa.....................
Listeria monocytogenes              CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa.....................
Bacillus subtilis                   CUCUUAUCCCGAGcUGG.C.GG.................AGGGA..CAGG.CCCUA..UGAA.GCC.-CAGCAACC.G.GUU.UCUCUguuauuuauuaug............
Listeria innocua                    CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU.........................
Listeria monocytogenes              CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU.........................
Listeria innocua                    AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUG..UGAA.ACC.-CAGCAACC.U.AAA.CAAUAauuc.....................
Listeria monocytogenes              AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUA..UGAA.GCC.-CAGCAACC.U.AAA.CAAUAauuc.....................
Listeria innocua                    CUCUUAUUAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CAGCAACC.U.UUC.AAUUC.........................
Listeria monocytogenes              CUCUUAUAAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CGGCAACC.U.UUC.AAUAC.........................
Streptomyces coelicolor             CGCUCAUCCAGAG.GGG.C.AG.................AGGGA.uACGG.CCCGA..UGAA.GCC.CCGGCAACC.C.UCC.AGUCGguucuugucacacgga.........
Clostridium sporogenes              UACUUAUCAAGAG.CGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.GCC.-CAGCAACC.U.AUA.UGAAA.........................
Staphylococcus aureus               AUCCU-----GAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA.........................
Staphylococcus aureus               CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA.........................
Staphylococcus epidermidis          CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UAUU-.........................
Bacillus subtilis                   CUCUUAUCGAGAGuUGG.G.CG.................AGGGA..UUGG.CCUUU..UGAC.CCCaACAGCAACC.G.ACC.GUAAUaccauugugaaauggggcgcacugc
Listeria innocua                    UUCUUAUCCAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CGGCAGCG.G.AGC.GC---.........................
Listeria monocytogenes              UUCUUAUCAAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CAGCAGCG.G.AGC.GC---.........................
Listeria innocua                    UUCUUAUCACGAA.AGG.U.GG.................AGGGA..CUGG.CCCUU..UGAA.GCC.UUAGCAACC.G.GAA.UUU--.........................
Listeria innocua                    AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCAG..UGAA.GCC.UUGGCAACC.G.GAC.UU---.........................
Listeria monocytogenes              AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCCG..UGAA.GCC.UUGGCAACC.G.GAU.UU---.........................
Bacillus licheniformis ATCC 14580   UUCUUAUUCAGAG.AGG.C.GG.................AGGGA.aUUGG.CCCUG..UGAA.ACC.UCGGCAGCG.G.GUU.CUGCAua.......................
```

Likelihood of training data:

$$\mathscr{L} = \frac{1}{\#\mathrm{MSA}} \sum_{s \in \mathrm{MSA}} P_v(\mathbf{v}^{(n)})$$

Restricted Boltzmann Machines (RBM)

MSA (training data):



```
Bacillus subtilis                  UUUCUAUCCAGAG.AGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.ACC.UCGGCAACA.-.---.UU---........................
Staphylococcus epidermidis         ACCUUAUUUUGAG.AAG.C.UG.................AGGGA.uUUGG.CCCAU..AGAA.GCU.UCAGCAACC.G.ACU.UUA--........................
Bacillus subtilis                  AACUUAUCAAGAG.CGG.C.UG.................AGGGA..CUGG.ACCUA..UGAA.GCC.-CGGCAACC.U.GCA.UAG--........................
Listeria innocua                   CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa....................
Listeria monocytogenes             CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa....................
Bacillus subtilis                  CUCUUAUCCCGAGcUGG.C.GG.................AGGGA..CAGG.CCCUA..UGAA.GCC.-CAGCAACC.G.GUU.UCUCUguuauuuauuaug...........
Listeria innocua                   CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU........................
Listeria monocytogenes             CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU........................
Listeria innocua                   AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUG..UGAA.ACC.-CAGCAACC.U.AAA.CAAUAauuc....................
Listeria monocytogenes             AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUA..UGAA.GCC.-CAGCAACC.U.AAA.CAAUAauuc....................
Listeria innocua                   CUCUUAUUAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CAGCAACC.U.UUC.AAUUC........................
Listeria monocytogenes             CUCUUAUAAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CGGCAACC.U.UUC.AAUAC........................
Streptomyces coelicolor            CGCUCAUCCAGAG.GGG.C.AG.................AGGGA.uACGG.CCCGA..UGAA.GCC.CCGGCAACC.C.UCC.AGUCGguucuugucacacgga........
Clostridium sporogenes             UACUUAUCAAGAG.CGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.GCC.-CAGCAACC.U.AUA.UGAAA........................
Staphylococcus aureus              AUCCU-----GAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA........................
Staphylococcus aureus              CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA........................
Staphylococcus epidermidis         CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UAUU-........................
Bacillus subtilis                  CUCUUAUCGAGAGuUGG.G.CG.................AGGGA..UUGG.CCUUU..UGAC.CCCaACAGCAACC.G.ACC.GUAAUaccauugugaaaugggcgcacugc
Listeria innocua                   UUCUUAUCCAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CGGCAGCG.G.AGC.GC---........................
Listeria monocytogenes             UUCUUAUCAAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CAGCAGCG.G.AGC.GC---........................
Listeria innocua                   UUCUUAUCACGAA.AGG.U.GG.................AGGGA..CUGG.CCCUU..UGAA.GCC.UUAGCAACC.G.GAA.UUU--........................
Listeria innocua                   AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCAG..UGAA.GCC.UUGGCAACC.G.GAC.UU---........................
Listeria monocytogenes             AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCCG..UGAA.GCC.UUGGCAACC.G.GAU.UU---........................
Bacillus licheniformis ATCC 14580  UUCUUAUUCAGAG.AGG.C.GG.................AGGGA.aUUGG.CCCUG..UGAA.ACC.UCGGCAGCG.G.GUU.CUGCAua......................
```

Likelihood of training data:

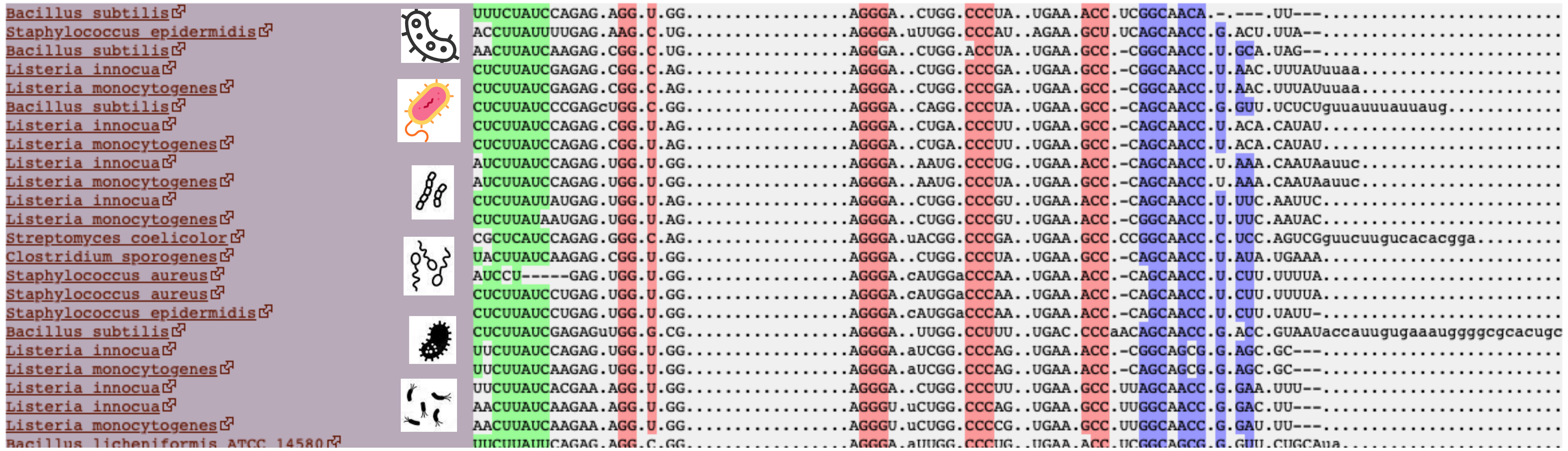
$$\mathscr{L} = \frac{1}{\#\mathrm{MSA}} \sum_{s \in \mathrm{MSA}} P_v(\mathbf{v}^{(n)})$$

Model is trained, by finding parameters that maximise the likelihood of the data:

$$\frac{\partial \mathcal{L}}{\partial \omega} = \underbrace{\left\langle \frac{\partial(-E_{\mathrm{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\mathrm{MSA}}}_{\text{positive gradient}} - \underbrace{\left\langle \frac{\partial(-E_{\mathrm{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\mathrm{RBM}}}_{\text{negative gradient}}$$

Restricted Boltzmann Machines (RBM)

MSA (training data):



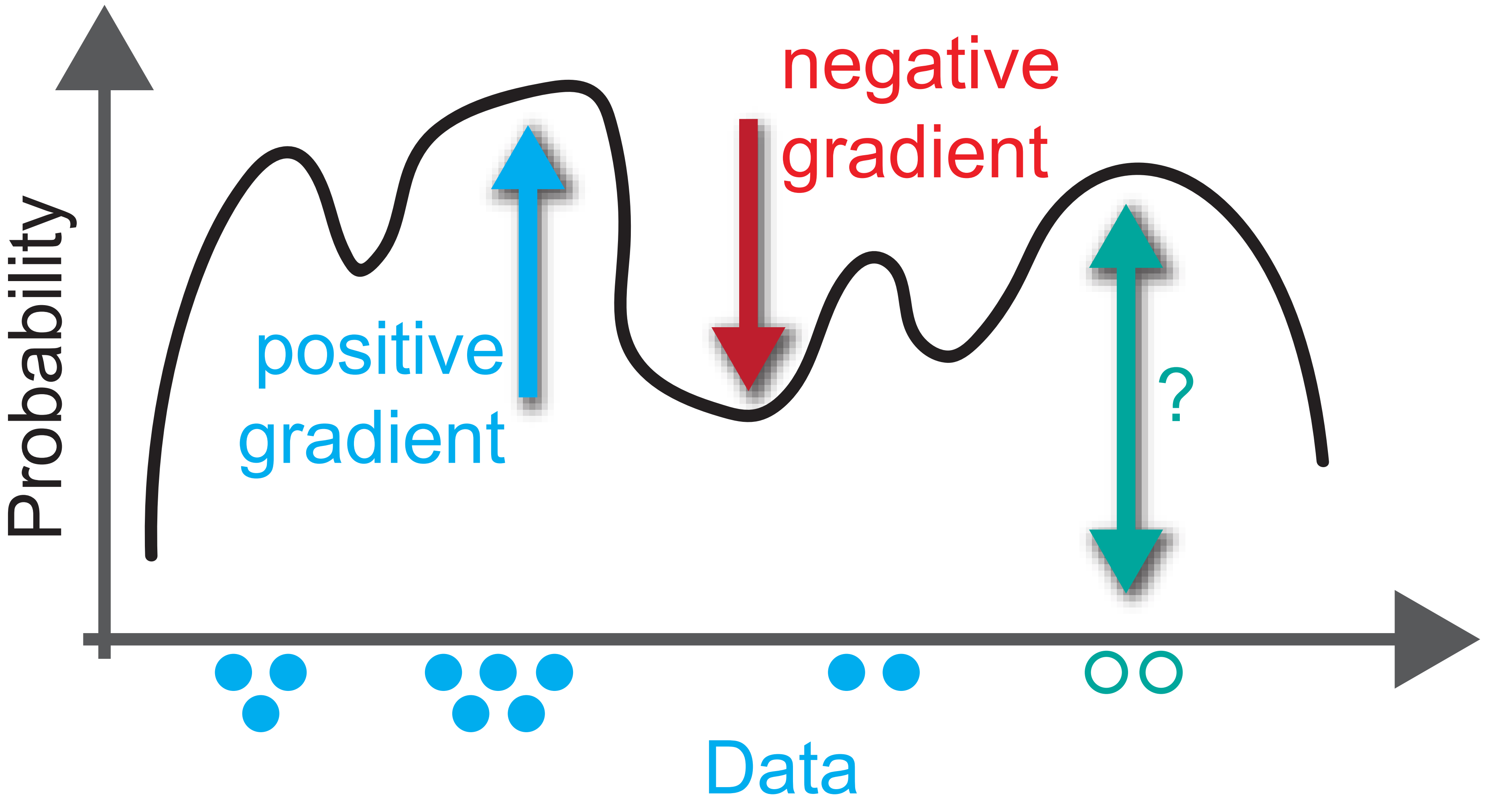
```
Bacillus subtilis                    UUUCUAUCCAGAG.AGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.ACC.UCGGCAACA.-.---.UU---.........................
Staphylococcus epidermidis           ACCUUAUUUUGAG.AAG.C.UG.................AGGGA.uUUGG.CCCAU..AGAA.GCU.UCAGCAACC.G.ACU.UUA--.........................
Bacillus subtilis                    AACUUAUCAAGAG.CGG.C.UG.................AGGGA..CUGG.ACCUA..UGAA.GCC.-CGGCAACC.U.GCA.UAG--.........................
Listeria innocua                     CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa.....................
Listeria monocytogenes               CUCUUAUCGAGAG.CGG.C.AG.................AGGGA..CUGG.CCCGA..UGAA.GCC.-CGGCAACC.U.AAC.UUUAUuuaa.....................
Bacillus subtilis                    CUCUUAUCCCGAGcUGG.C.GG.................AGGGA..CAGG.CCCUA..UGAA.GCC.-CAGCAACC.G.GUU.UCUCUguuauuuauuaug............
Listeria innocua                     CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU.........................
Listeria monocytogenes               CUCUUAUCCAGAG.CGG.U.AG.................AGGGA..CUGA.CCCUU..UGAA.GCC.-CAGCAACC.U.ACA.CAUAU.........................
Listeria innocua                     AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUG..UGAA.ACC.-CAGCAACC.U.AAA.CAAUAauuc.....................
Listeria monocytogenes               AUCUUAUCCAGAG.UGG.U.GG.................AGGGA..AAUG.CCCUA..UGAA.GCC.-CAGCAACC.U.AAA.CAAUAauuc.....................
Listeria innocua                     CUCUUAUUAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CAGCAACC.U.UUC.AAUUC.........................
Listeria monocytogenes               CUCUUAUAAUGAG.UGG.U.AG.................AGGGA..CUGG.CCCGU..UGAA.ACC.-CGGCAACC.U.UUC.AAUAC.........................
Streptomyces coelicolor              CGCUCAUCCAGAG.GGG.C.AG.................AGGGA.uACGG.CCCGA..UGAA.GCC.CCGGCAACC.C.UCC.AGUCGguucuugucacacgga.........
Clostridium sporogenes               UACUUAUCAAGAG.CGG.U.GG.................AGGGA..CUGG.CCCUA..UGAA.GCC.-CAGCAACC.U.AUA.UGAAA.........................
Staphylococcus aureus                AUCCU-----GAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA.........................
Staphylococcus aureus                CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UUUUA.........................
Staphylococcus epidermidis           CUCUUAUCCUGAG.UGG.U.GG.................AGGGA.cAUGGaCCCAA..UGAA.ACC.-CAGCAACC.U.CUU.UAUU-.........................
Bacillus subtilis                    CUCUUAUCGAGAGuUGG.G.CG.................AGGGA..UUGG.CCUUU..UGAC.CCCaACAGCAACC.G.ACC.GUAAUaccauugugaaauggggcgcacugc
Listeria innocua                     UUCUUAUCCAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CGGCAGCG.G.AGC.GC---.........................
Listeria monocytogenes               UUCUUAUCAAGAG.UGG.U.GG.................AGGGA.aUCGG.CCCAG..UGAA.ACC.-CAGCAGCG.G.AGC.GC---.........................
Listeria innocua                     UUCUUAUCACGAA.AGG.U.GG.................AGGGA..CUGG.CCCUU..UGAA.GCC.UUAGCAACC.G.GAA.UUU--.........................
Listeria innocua                     AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCAG..UGAA.GCC.UUGGCAACC.G.GAC.UU---.........................
Listeria monocytogenes               AACUUAUCAAGAA.AGG.U.GG.................AGGGU.uCUGG.CCCCG..UGAA.GCC.UUGGCAACC.G.GAU.UU---.........................
Bacillus licheniformis ATCC 14580   UUCUUAUUCAGAG.AGG.C.GG.................AGGGA.aUUGG.CCCUG..UGAA.ACC.UCGGCAGCG.G.GUU.CUGCAua.......................
```

Likelihood of training data:

$$\mathscr{L} = \frac{1}{\#\mathrm{MSA}} \sum_{s\in\mathrm{MSA}} P_v(\mathbf{v}^{(n)})$$

Model is trained, by finding parameters that maximise the likelihood of the data:

$$\frac{\partial \mathcal{L}}{\partial \omega} = \underbrace{\left\langle \frac{\partial(-E_{\mathrm{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\mathrm{MSA}}}_{\text{positive gradient}} - \underbrace{\left\langle \frac{\partial(-E_{\mathrm{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\mathrm{RBM}}}_{\text{negative gradient}}$$



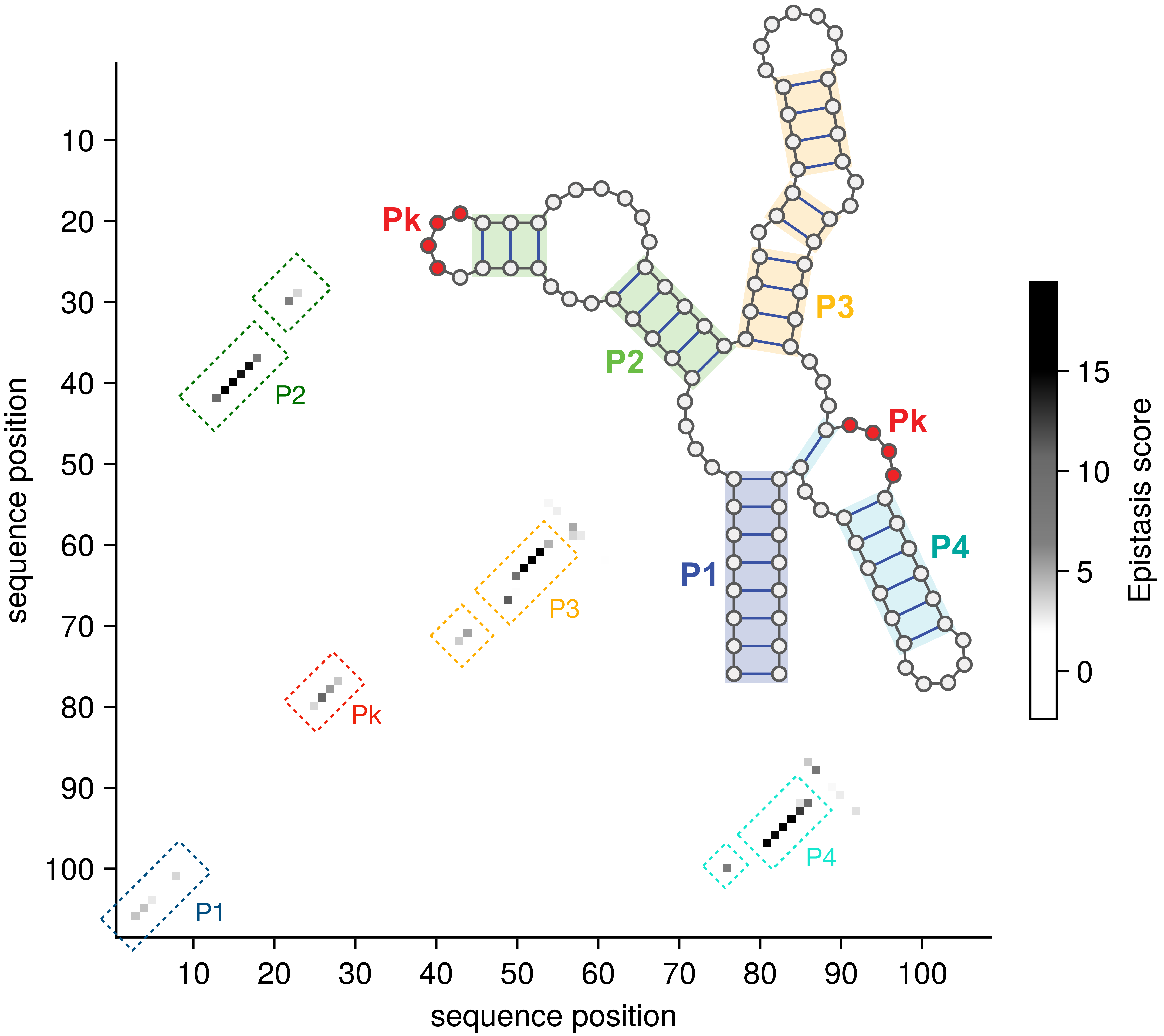
RBM is compatible with structural constraints

	seq.	prob.
wild type	GAUAAG	$P(.,.)$
single & double mutants	GGUAAG	$P(G,.)$
	GAUACG	$P(.,C)$
	GGUACG	$P(G,C)$

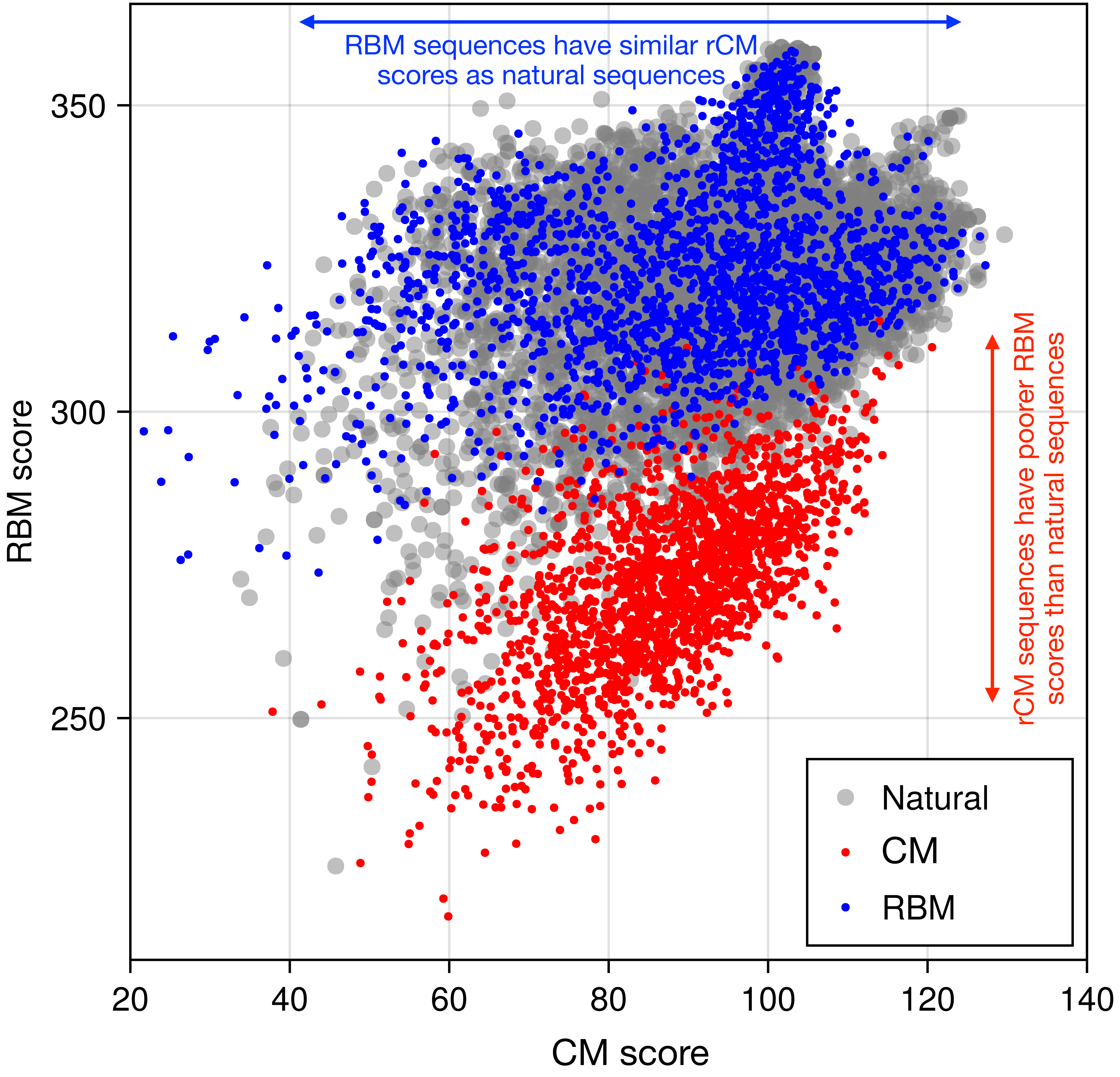
epistatic score

$$\frac{P(.,.)P(G,C)}{P(G,.)P(.,C)}$$

- De Leonardis, et al. "DCA of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction." NAR 43.21 (2015): 10444-10455.
- Weinreb, et al. "3D RNA and functional interactions from evolutionary couplings." Cell 165.4 (2016): 963-975.
- Tubiana, et al. "Learning protein constitutive motifs from sequence data." Elife 8 (2019): e39397.



RBM is compatible with structural constraints



RBM “score”:

$$\ln P_v(\mathbf{v}) = -E_{\text{eff}}(\mathbf{v}) + \text{const.}$$

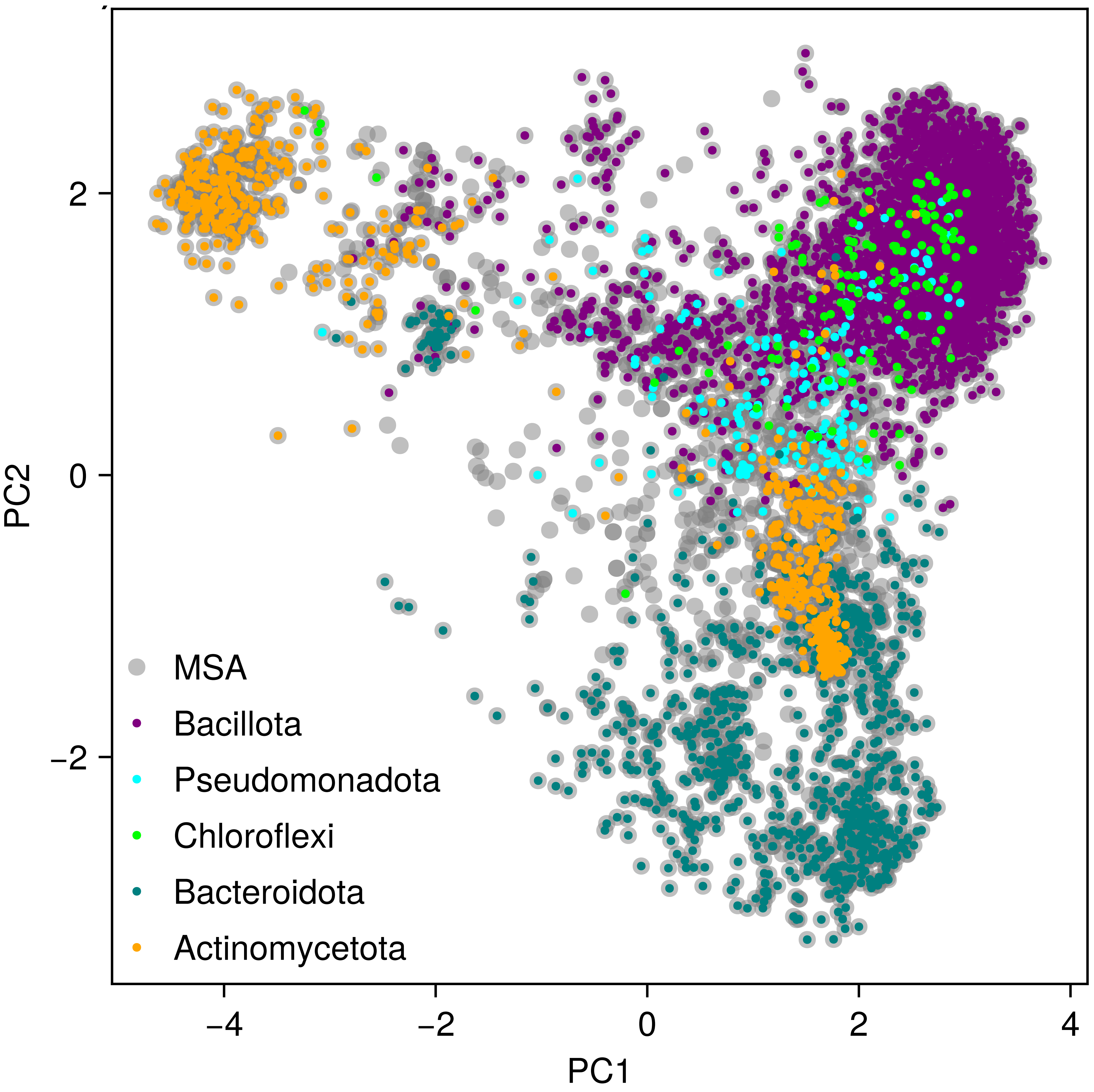
CM score:

log-prob. of sequence under covariance model (consensus secondary structure)

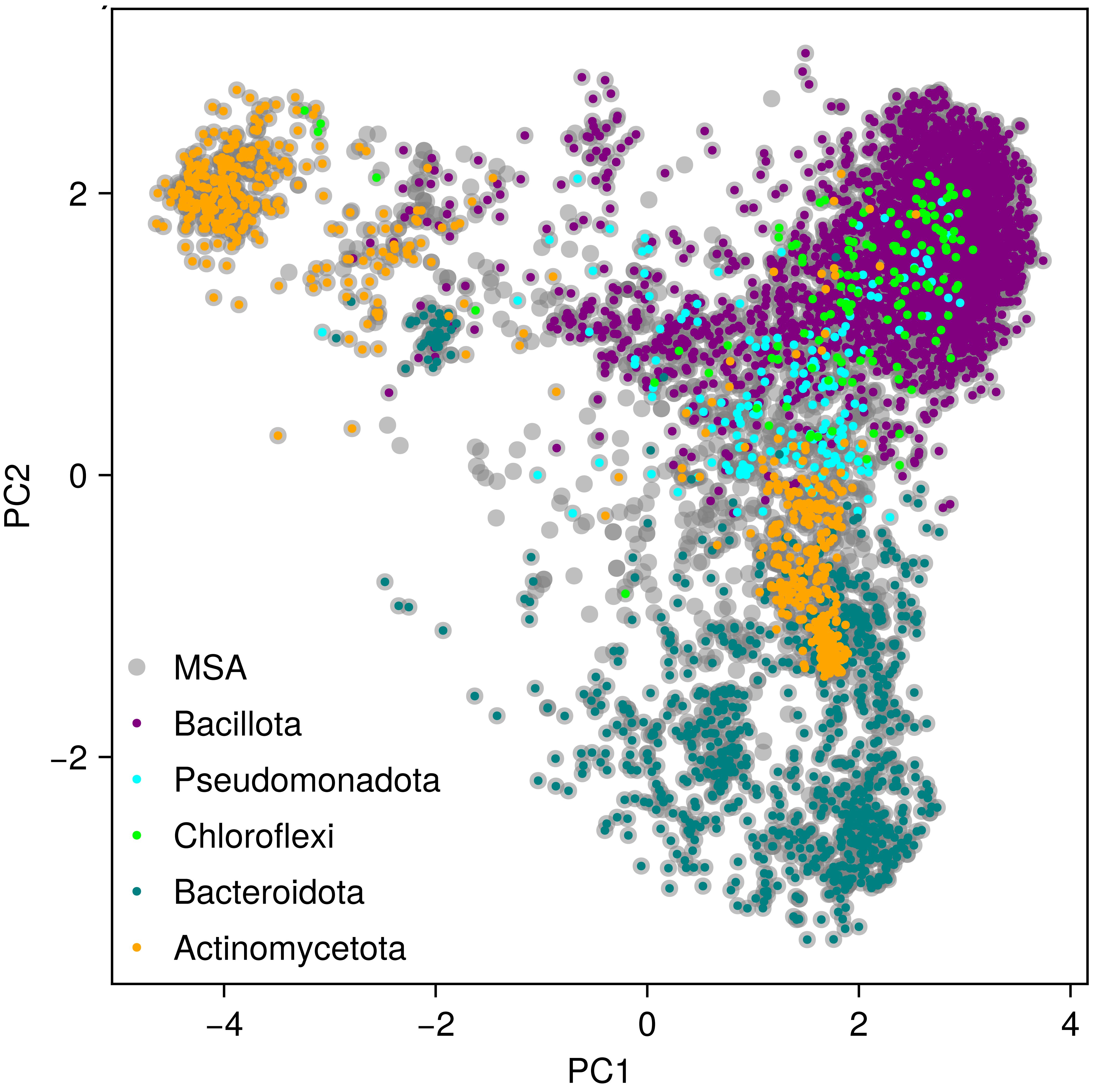
Nawrocki, Eric P., Diana L. Kolbe, and Sean R. Eddy. "Infernal 1.0: inference of RNA alignments." Bioinformatics 25.10 (2009): 1335-1337.

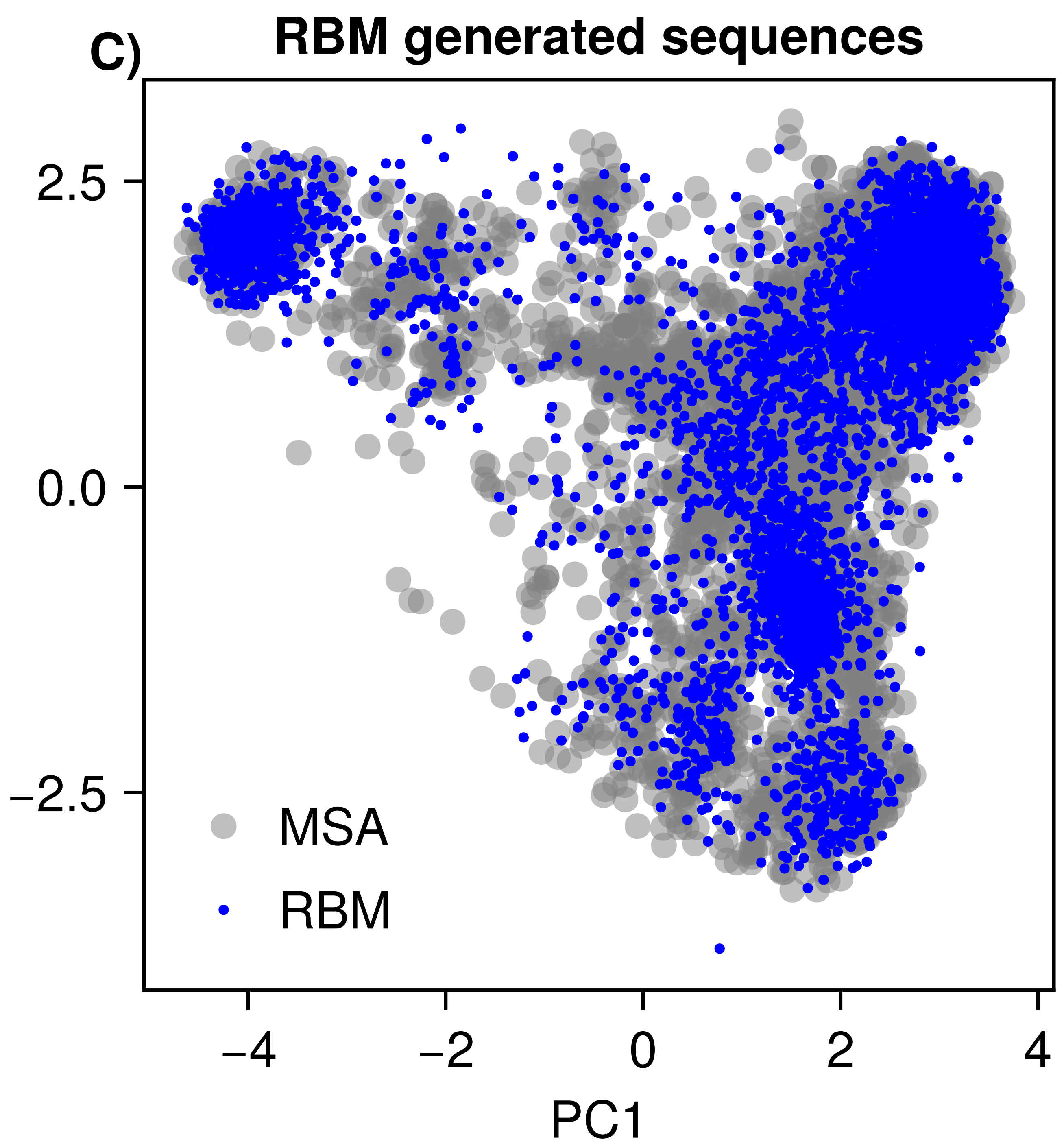
Eddy, Sean R., and Richard Durbin. "RNA sequence analysis using covariance models." Nucleic acids research 22.11 (1994): 2079-2088.

RBM generates diverse novel sequences

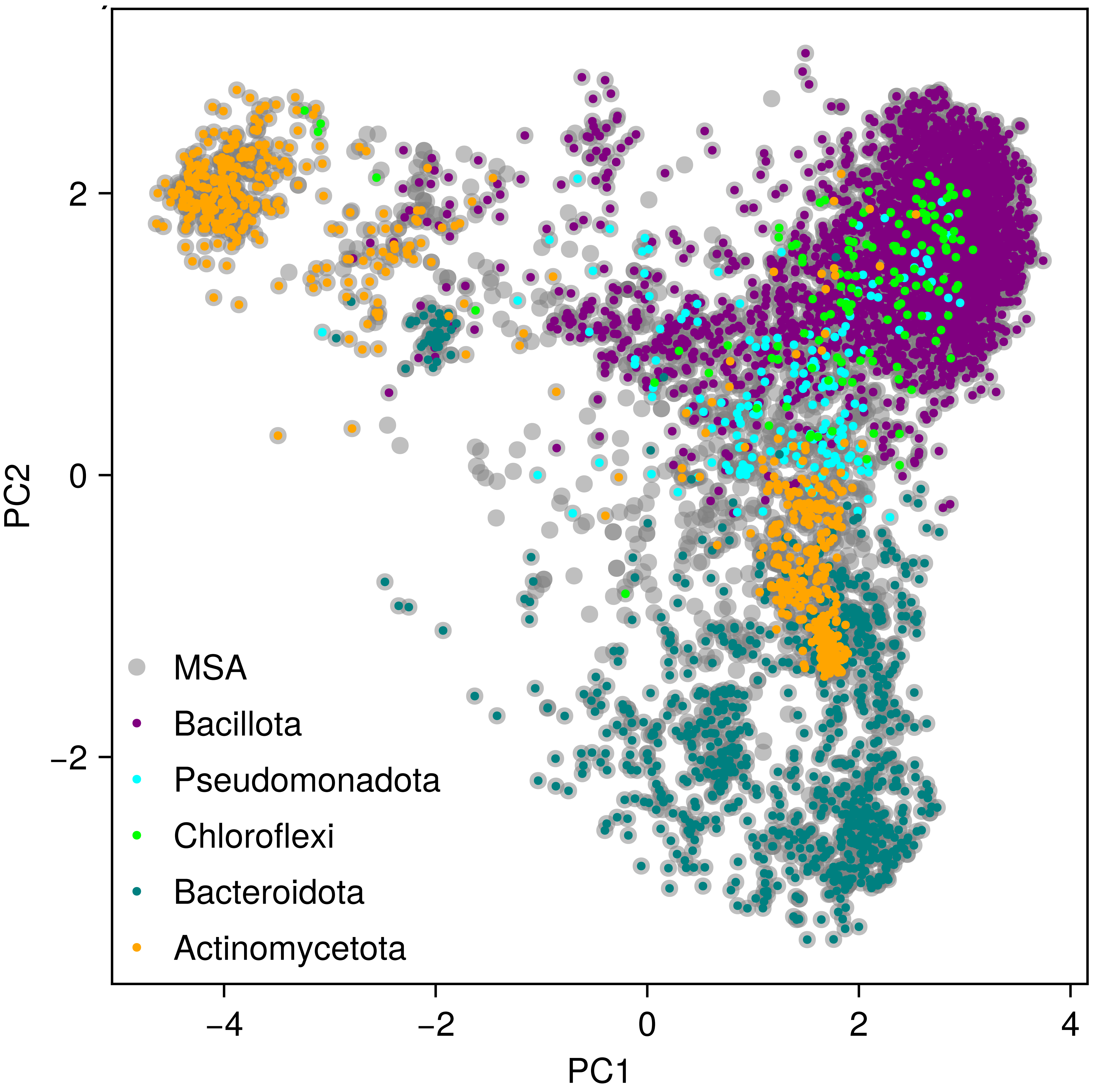


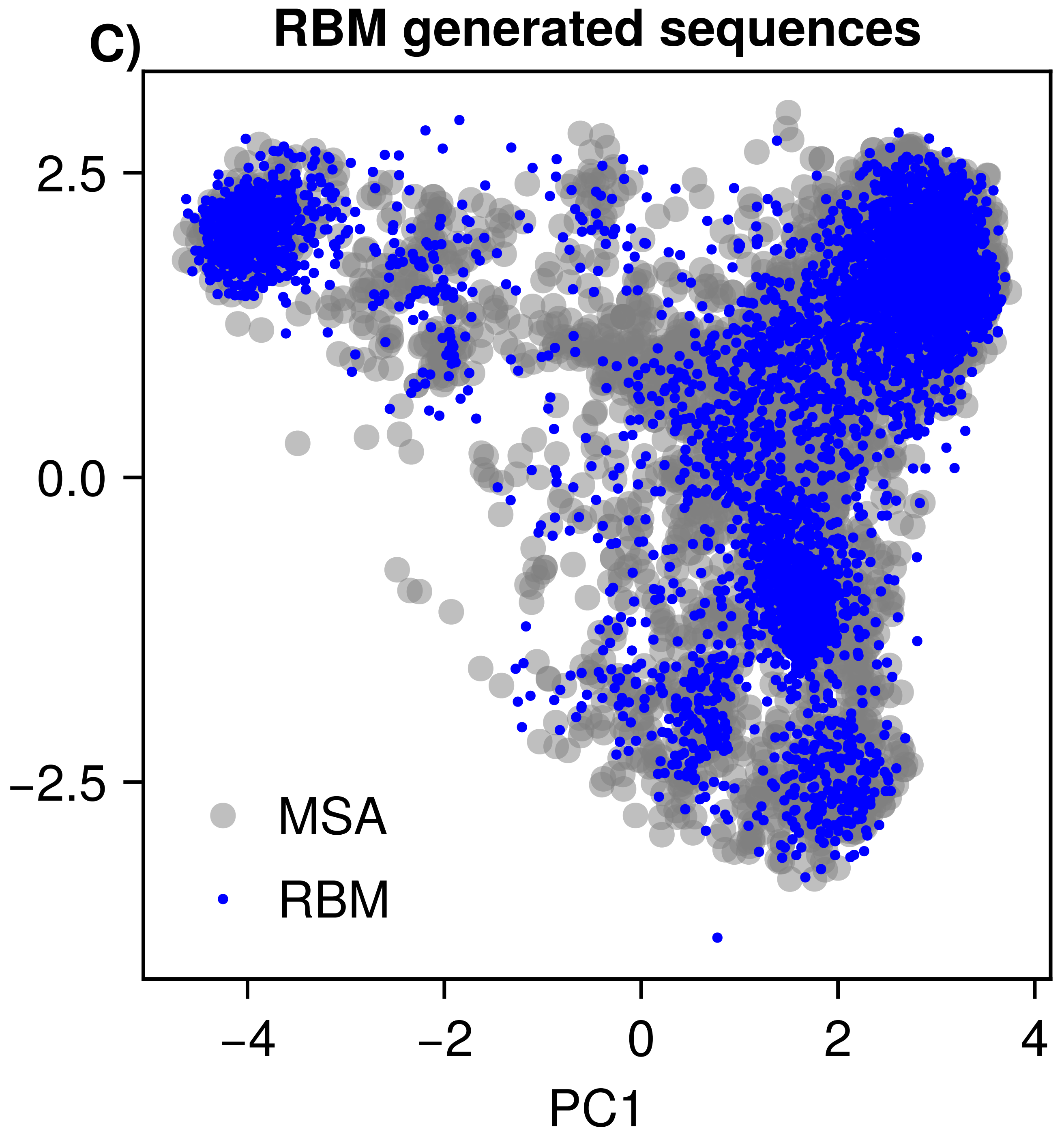
RBM generates diverse novel sequences

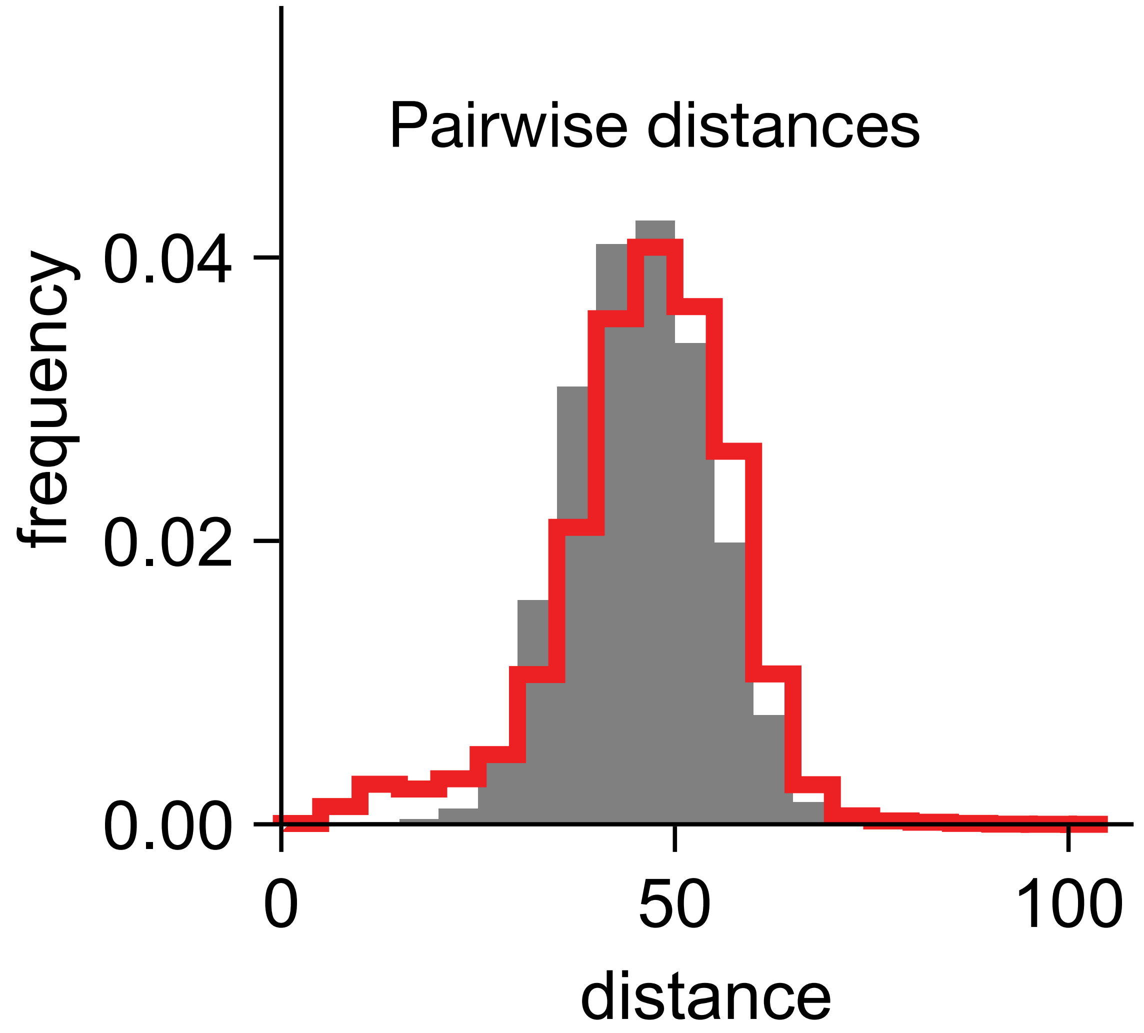


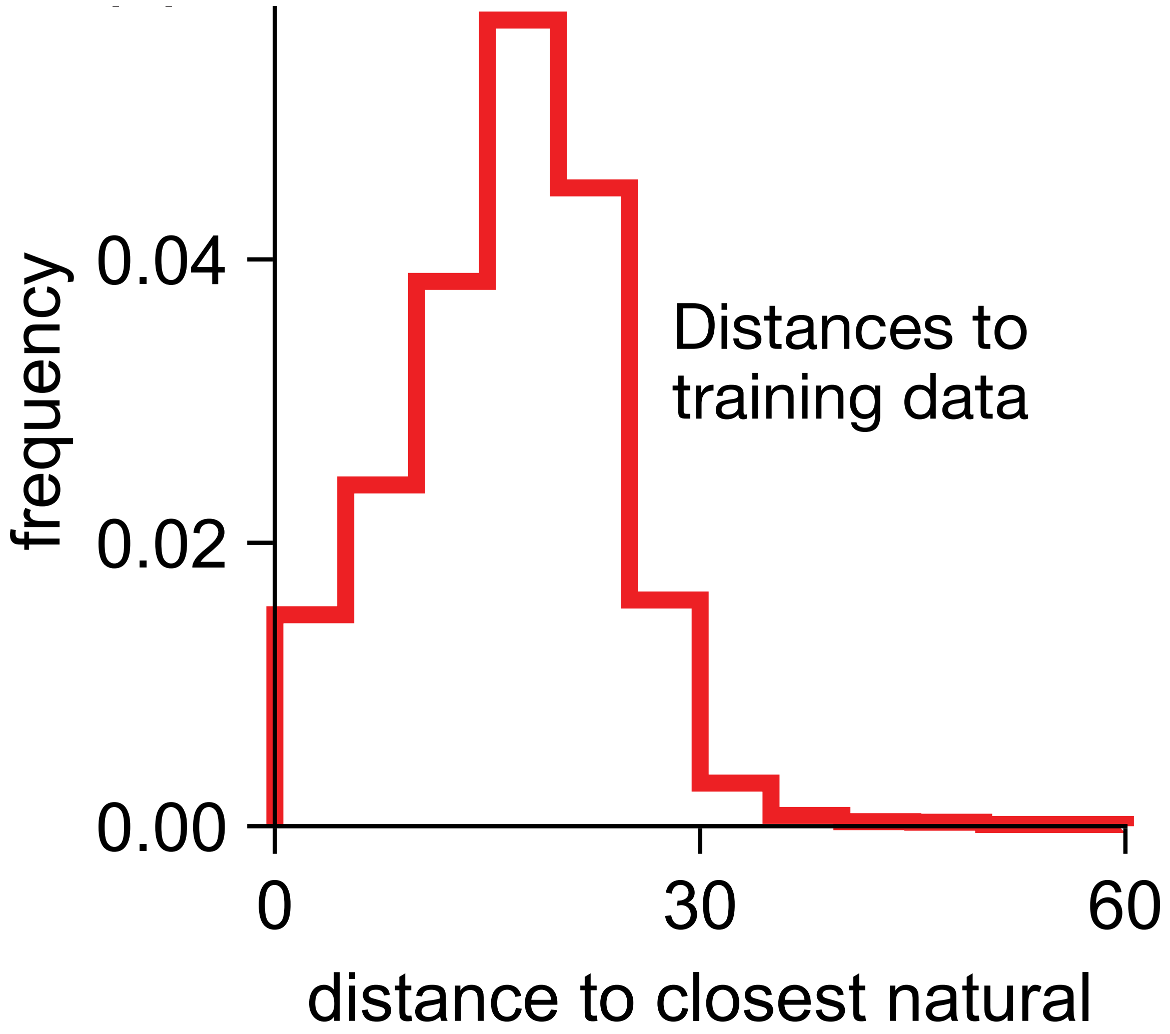


RBM generates diverse novel sequences

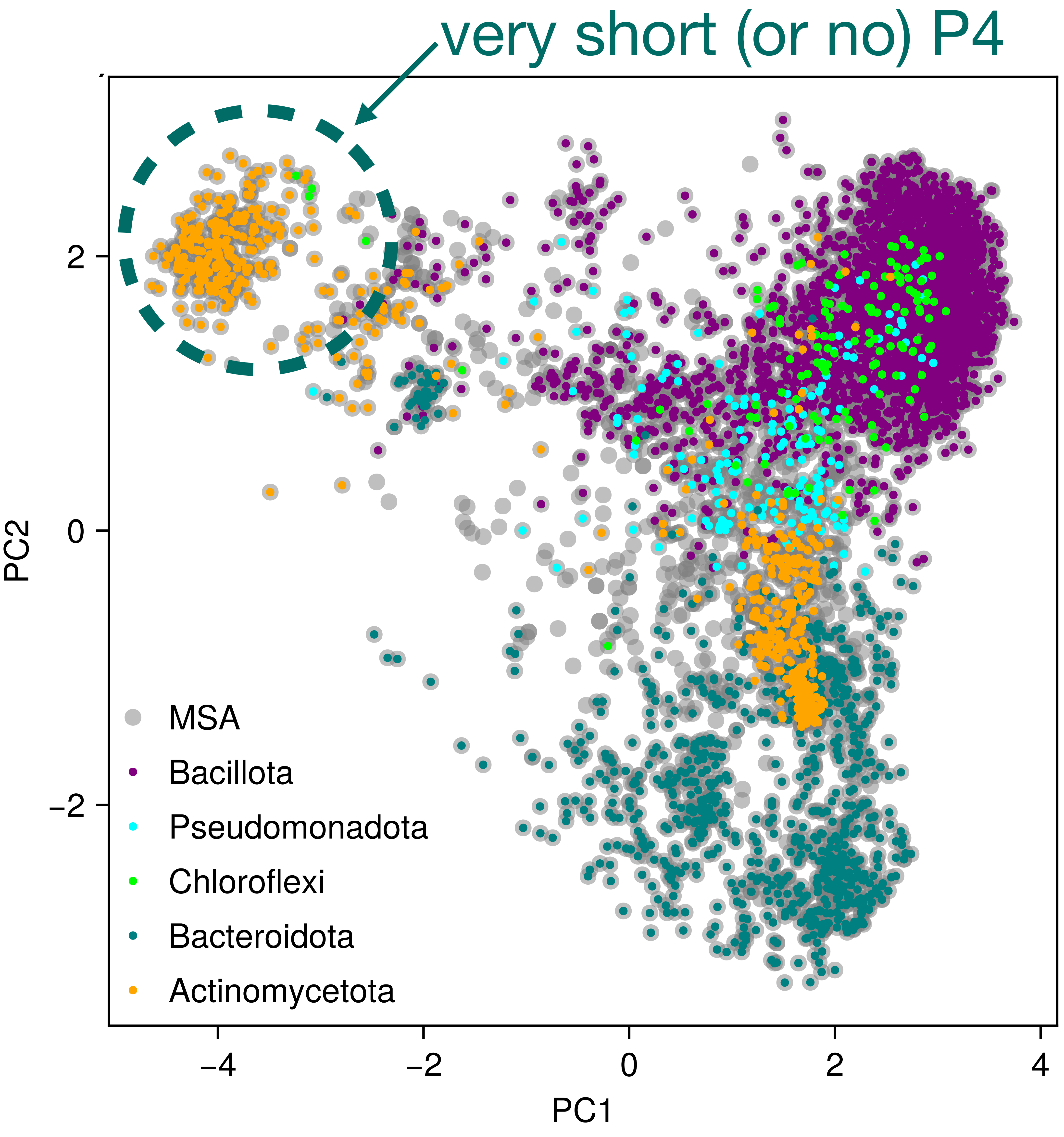


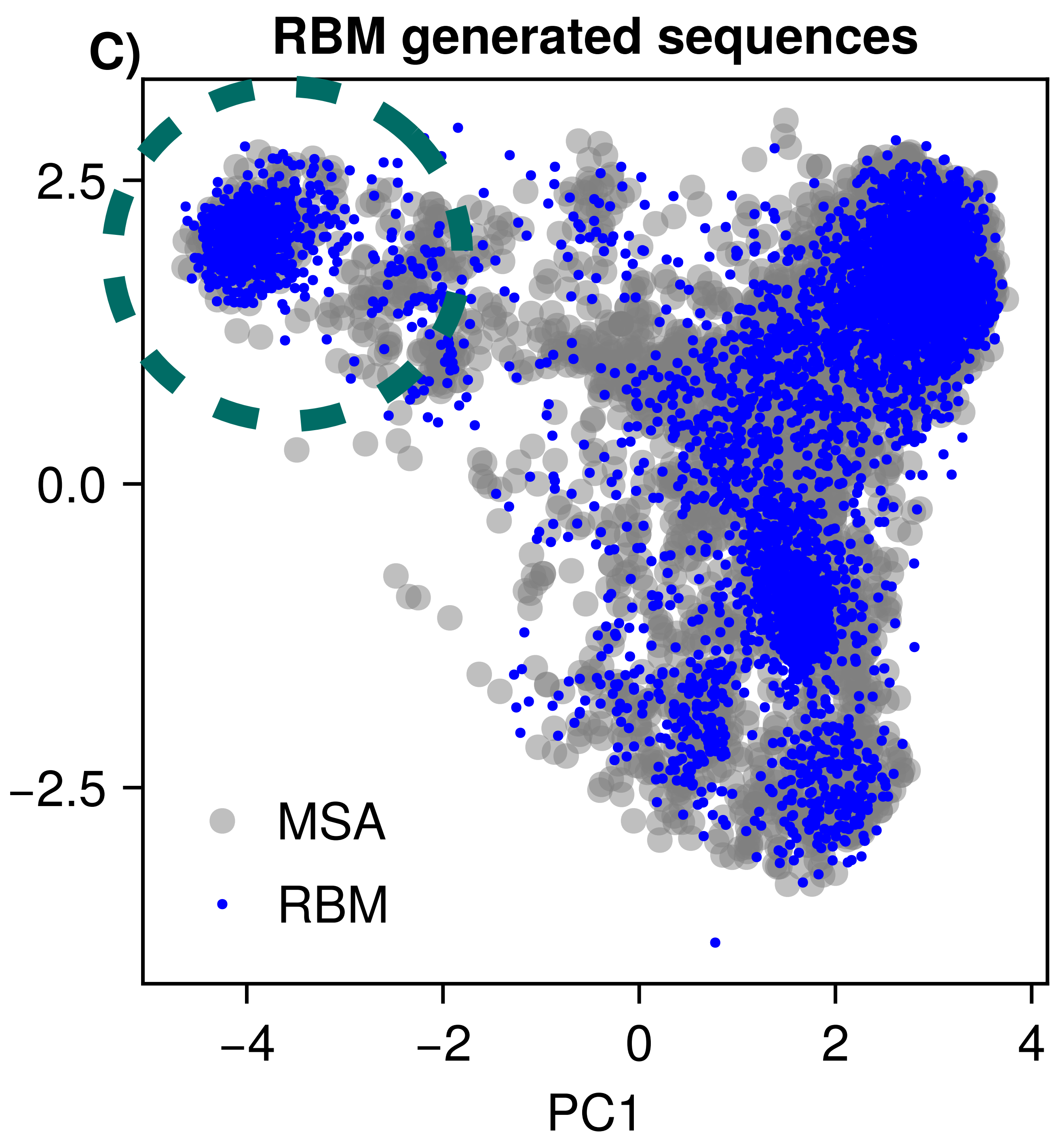

RBM generated sequences

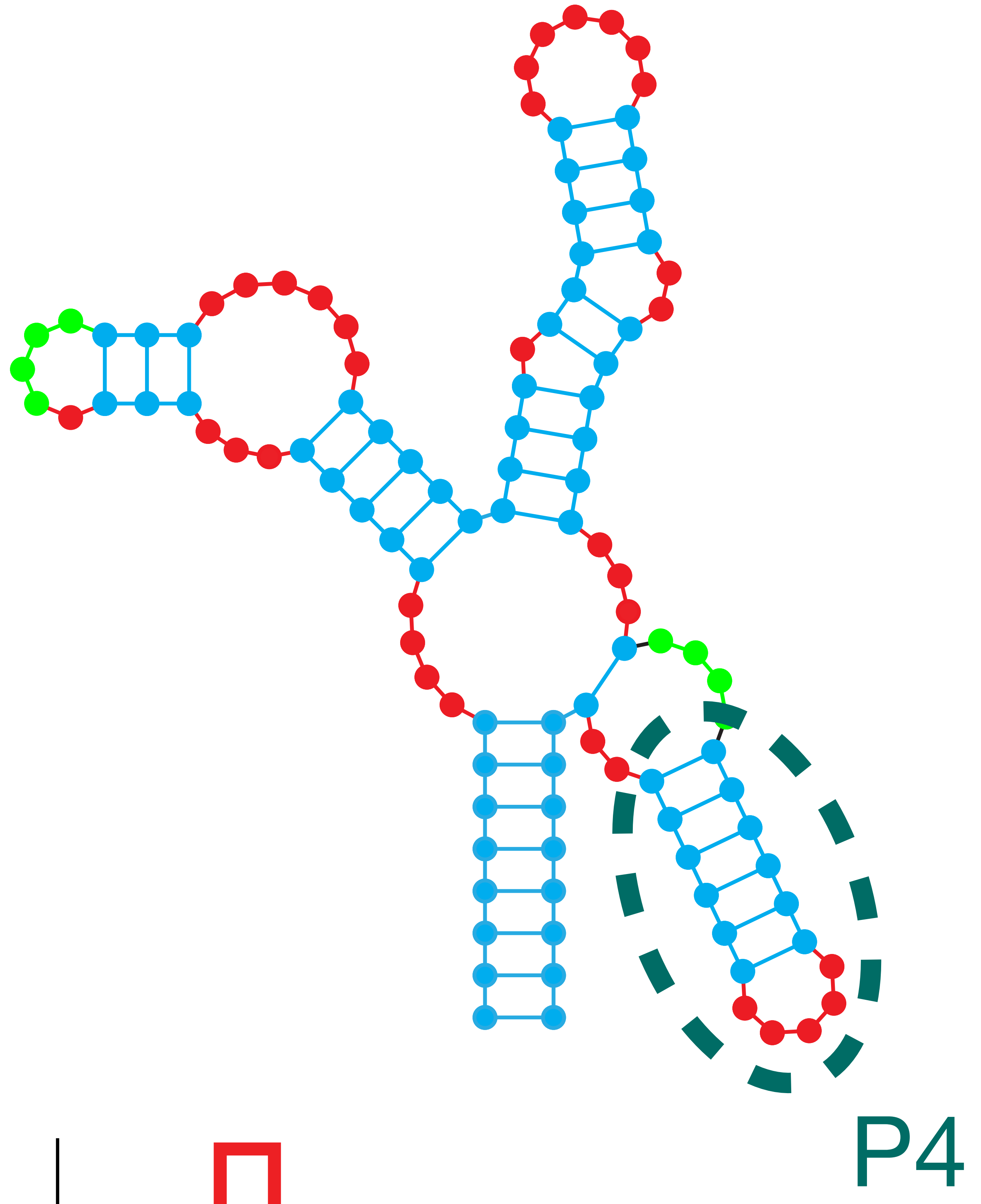


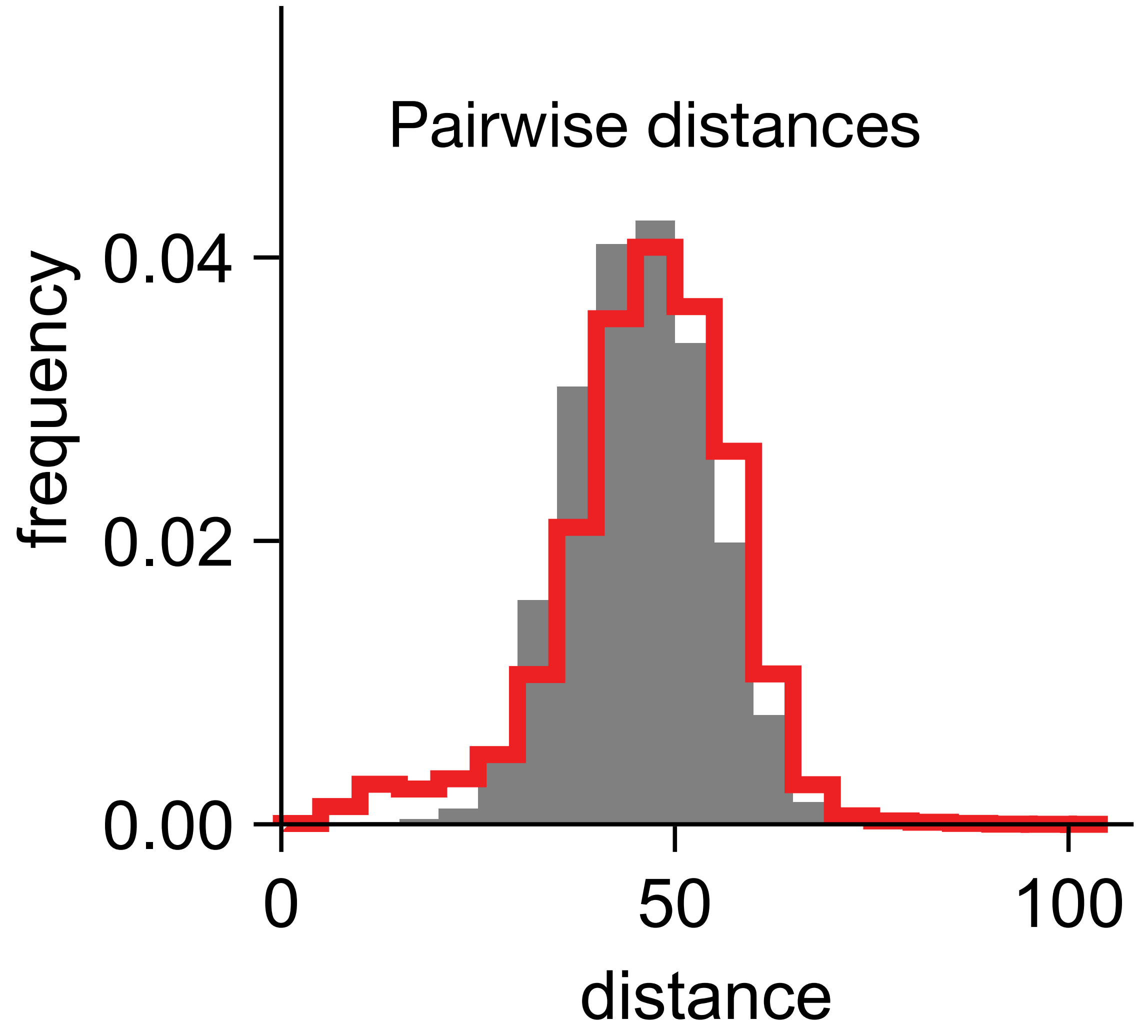


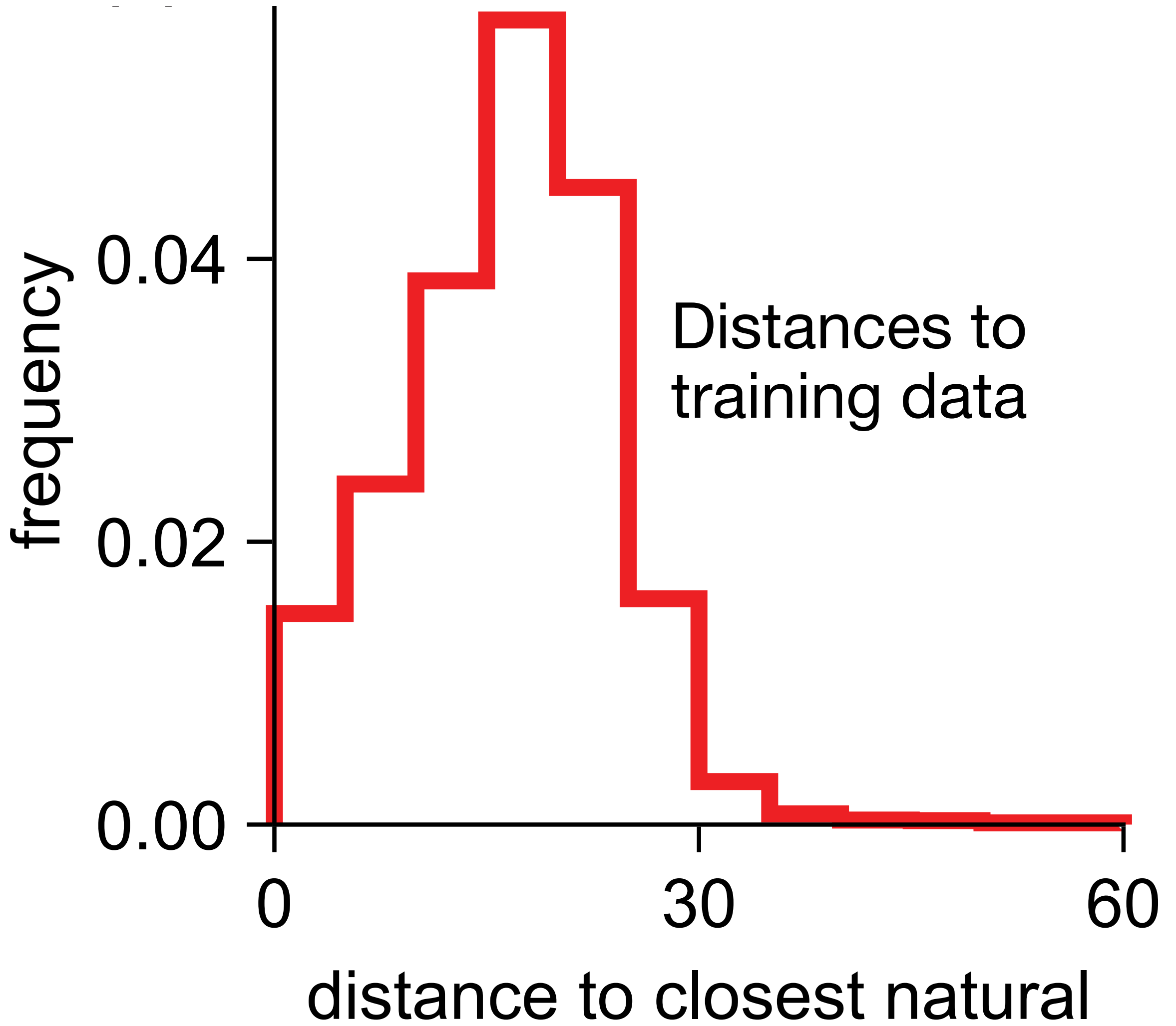
RBM generates diverse novel sequences











Trausch, Jeremiah J., et al. PNAS 111.18 (2014): 6624-6629.

Experimental validation with SHAPE-MaP

Probed ~500 sequences

UNIVERSITÉ
PARIS
DESCARTES

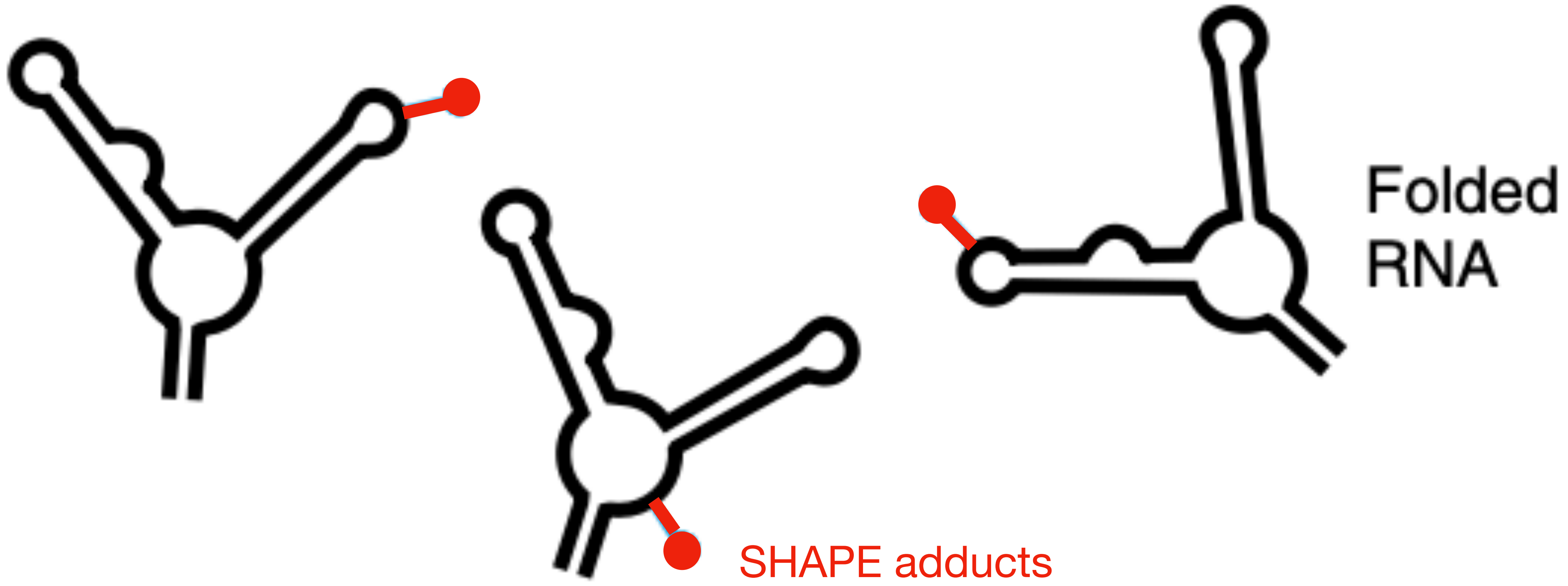
with P. Hardouin, F.-X. Lyonnet, B. Sargueil

Siegfried, Nathan A., et al. Nat. Meth. 11.9 (2014): 959-965.

Eddy, Sean R. Ann. Rev. Biophys. 43.1 (2014): 433-456.

Experimental validation with SHAPE-MaP

Probed ~500 sequences



UNIVERSITÉ
PARIS
DESCARTES

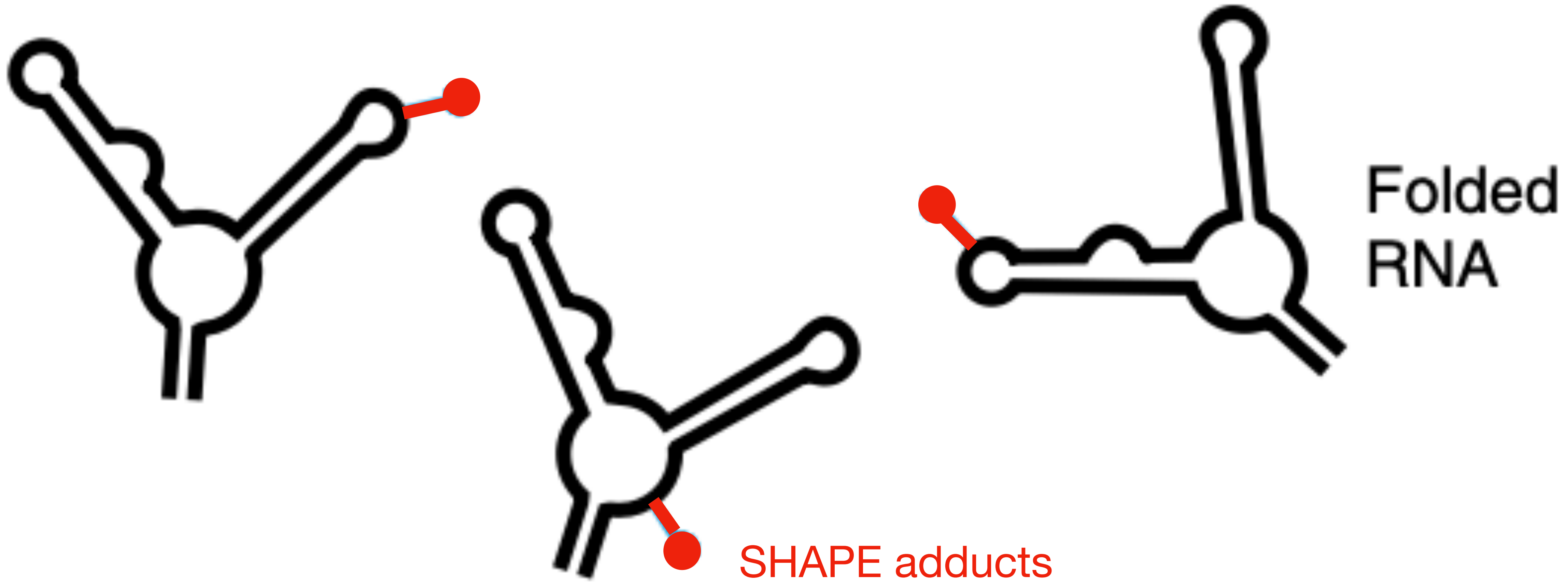
with P. Hardouin, F.-X. Lyonnet, B. Sargueil

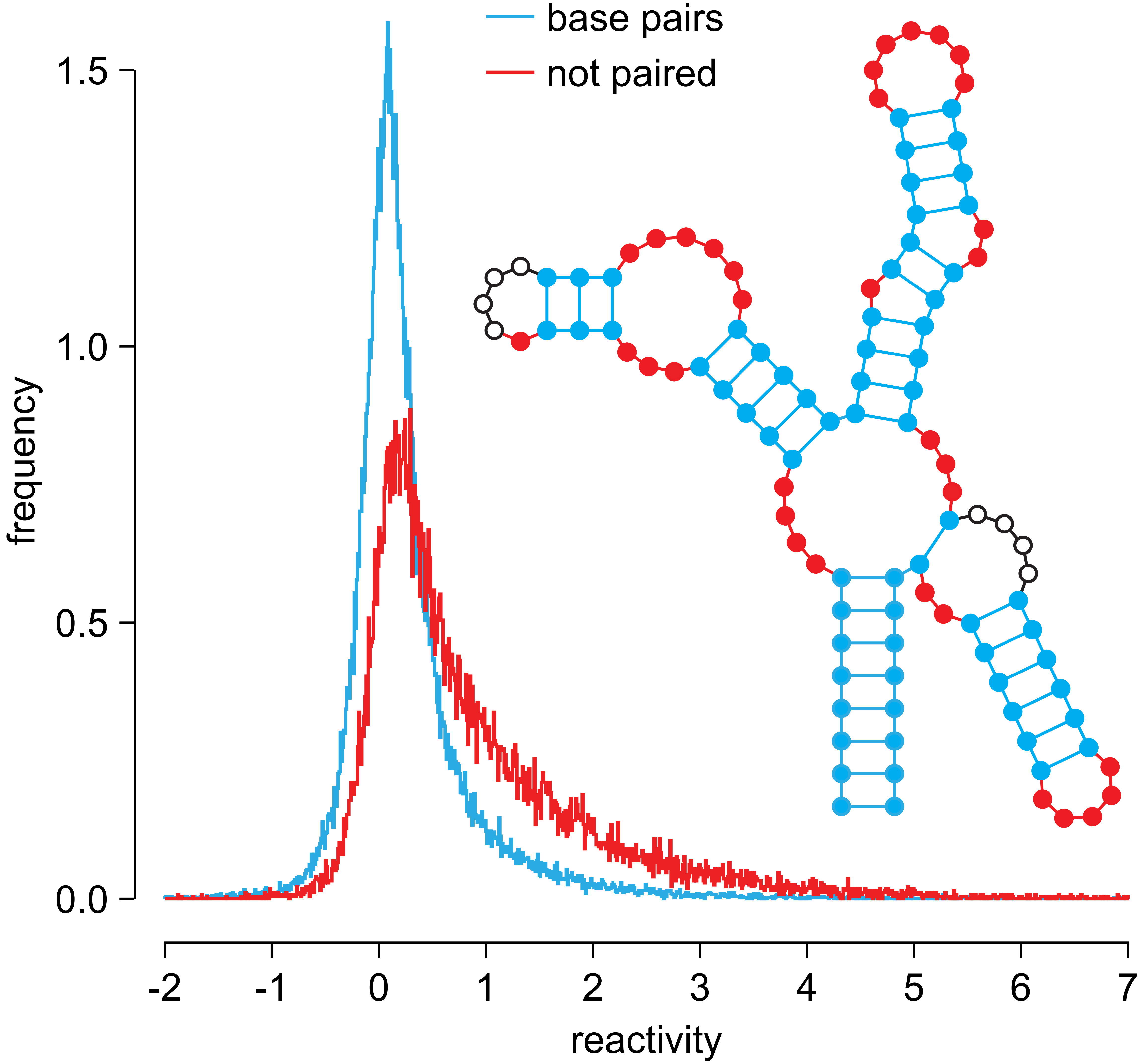
Siegfried, Nathan A., et al. Nat. Meth. 11.9 (2014): 959-965.

Eddy, Sean R. Ann. Rev. Biophys. 43.1 (2014): 433-456.

Experimental validation with SHAPE-MaP

Probed ~500 sequences



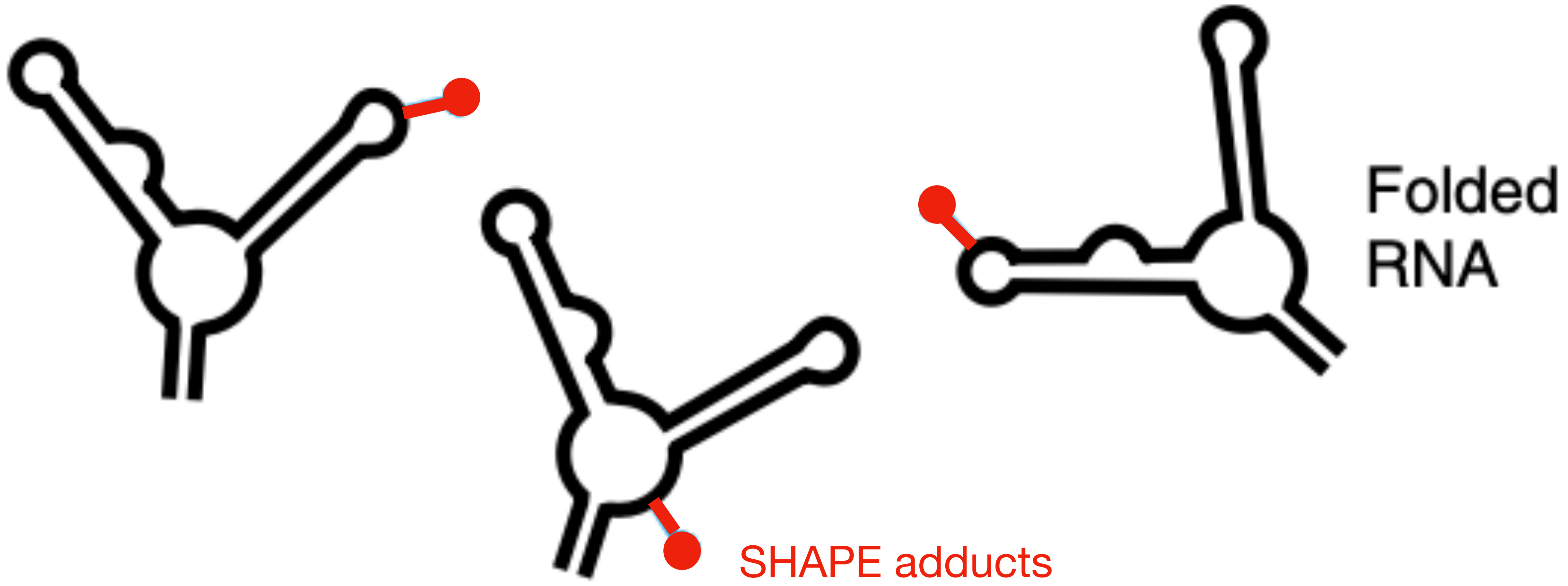


UNIVERSITÉ PARIS DESCARTES

with P. Hardouin, F.-X. Lyonnet, B. Sargueil

Experimental validation with SHAPE-MaP

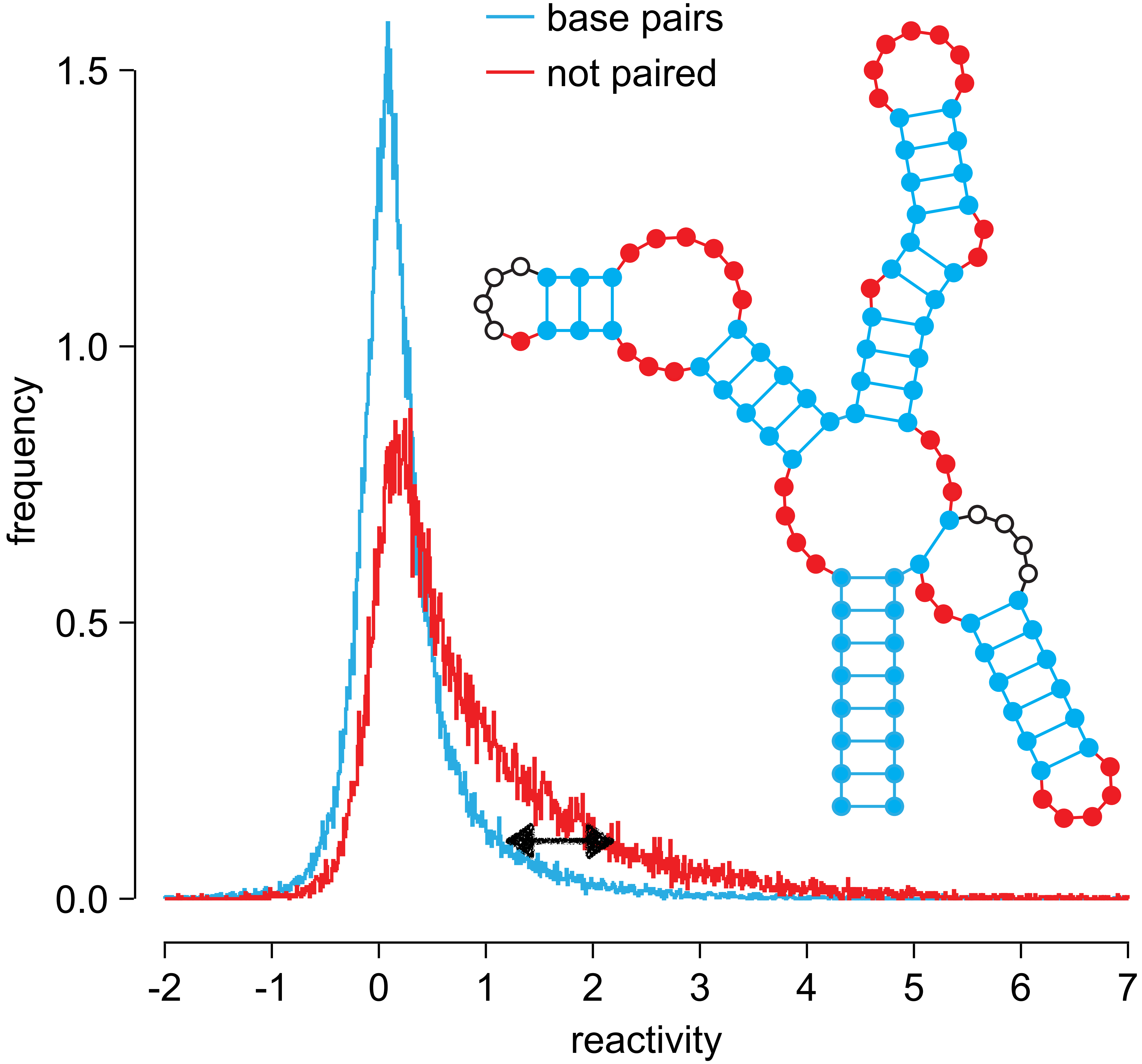
Probed ~500 sequences



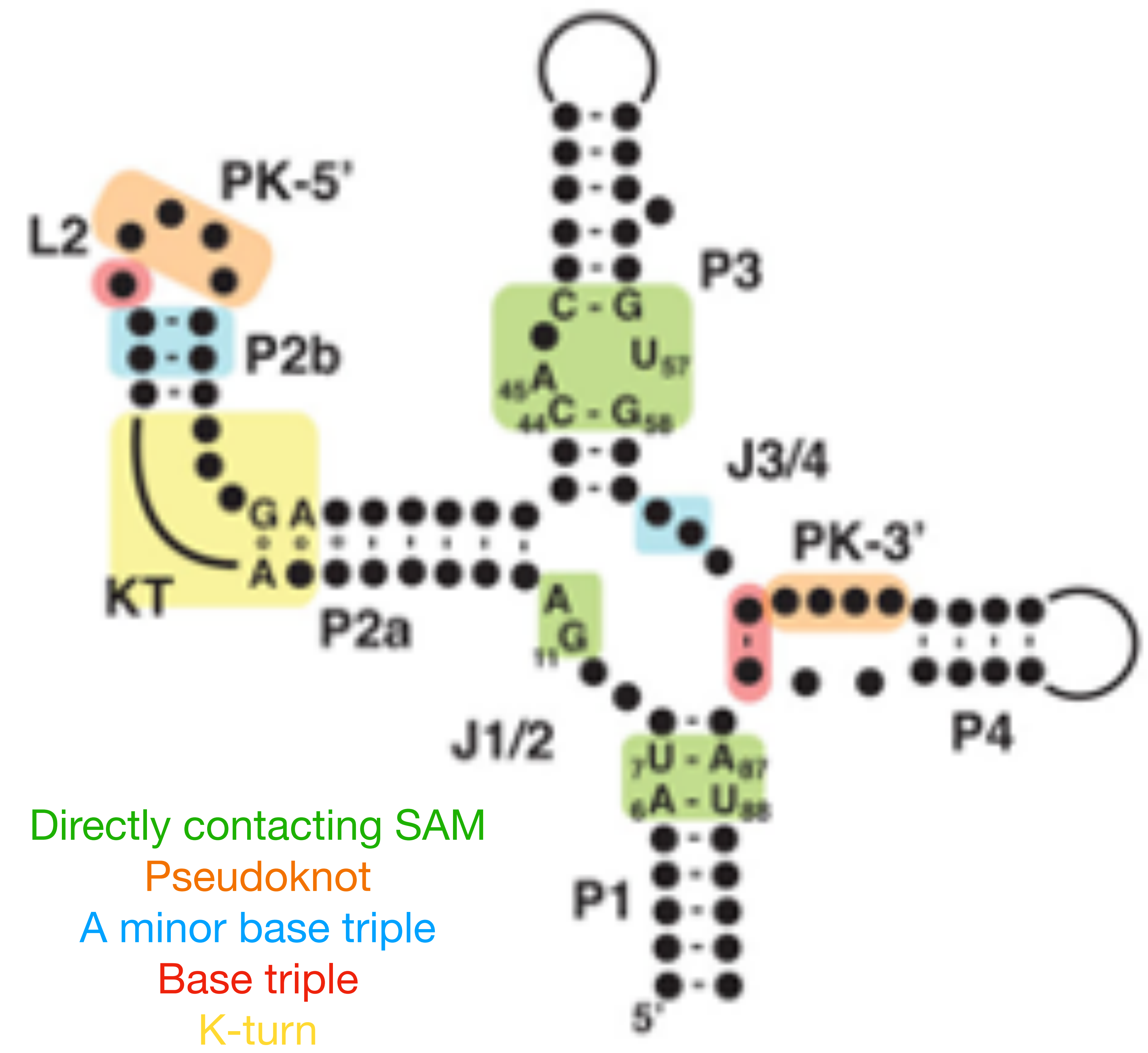
SHAPE reagent reacts preferentially with flexible (unpaired) RNA sites

UNIVERSITÉ
PARIS
DESCARTES

with P. Hardouin, F.-X. Lyonnet, B. Sargueil



SHAPE reactivities reflect expected structural changes in response to SAM



- Reactivity changes (SAM vs. no SAM) indicate SAM binding
- Consistent with previously reported structural responses in natural sequences

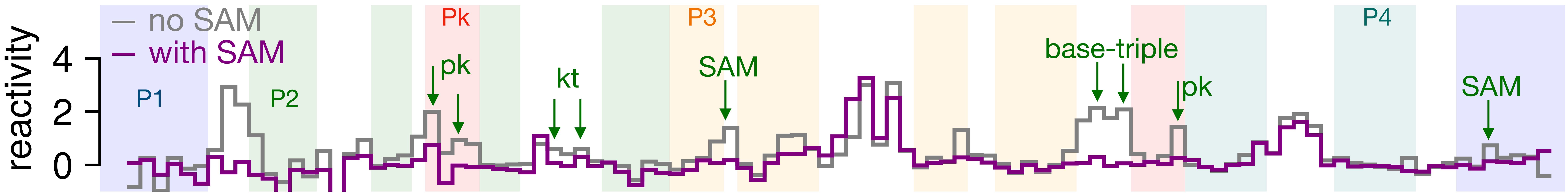
Trausch, Jeremiah J., et al. PNAS 111.18 (2014): 6624-6629

Hennelly, Scott P., Irina V. Novikova, and Karissa Y. Sanbonmatsu. NAR 41.3 (2013): 1922-1935.

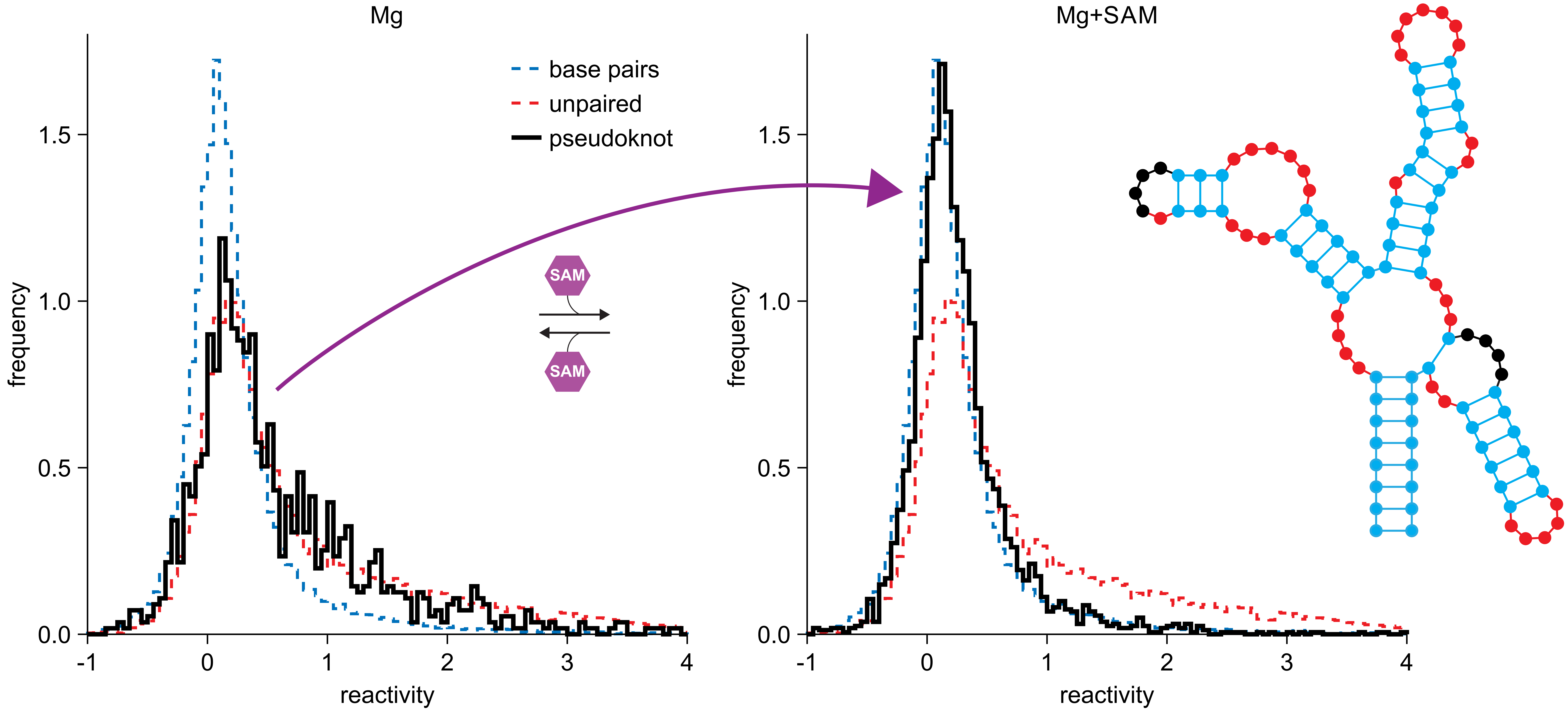
Heppell, Benoit, et al. "Molecular insights into the ligand-controlled organization of the SAM-I riboswitch." Nat. Chem. Biol. 7.6 (2011): 384-392.

IPANEMAP: *Saaidi, Afaf, et al. NAR 48.15 (2020): 8276-8289.*

Example:

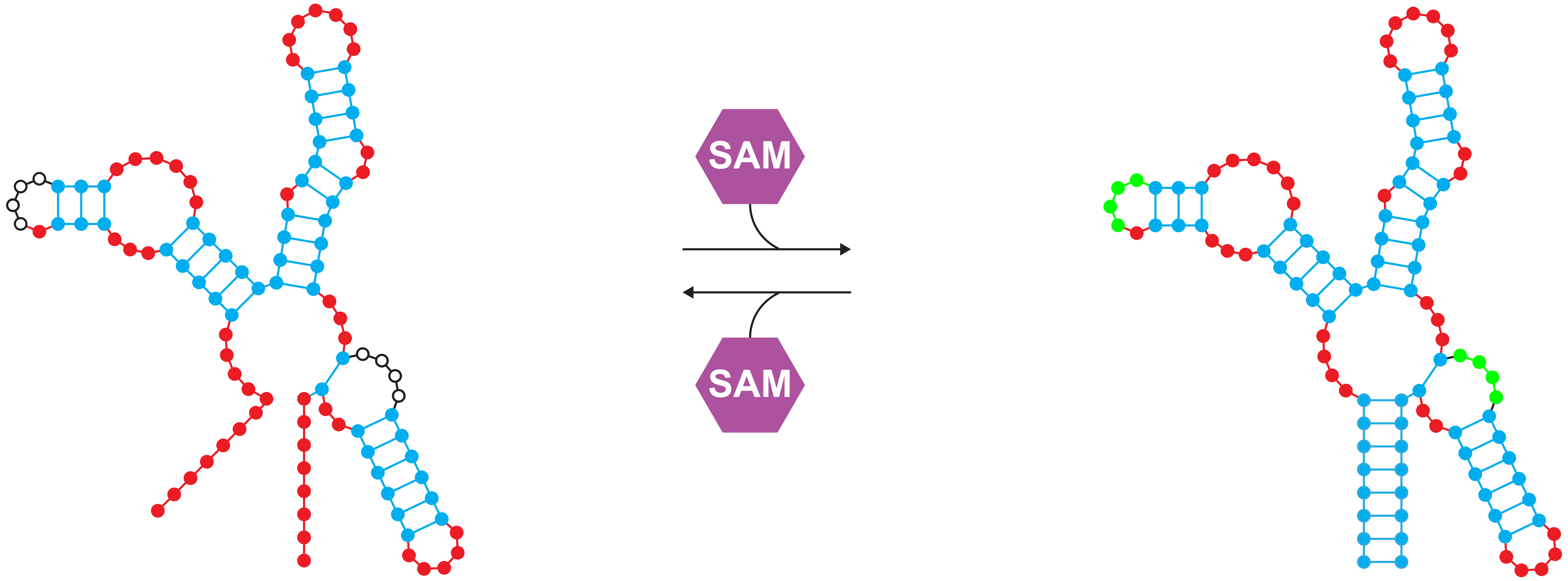


SHAPE reactivities reflect expected structural changes in response to SAM



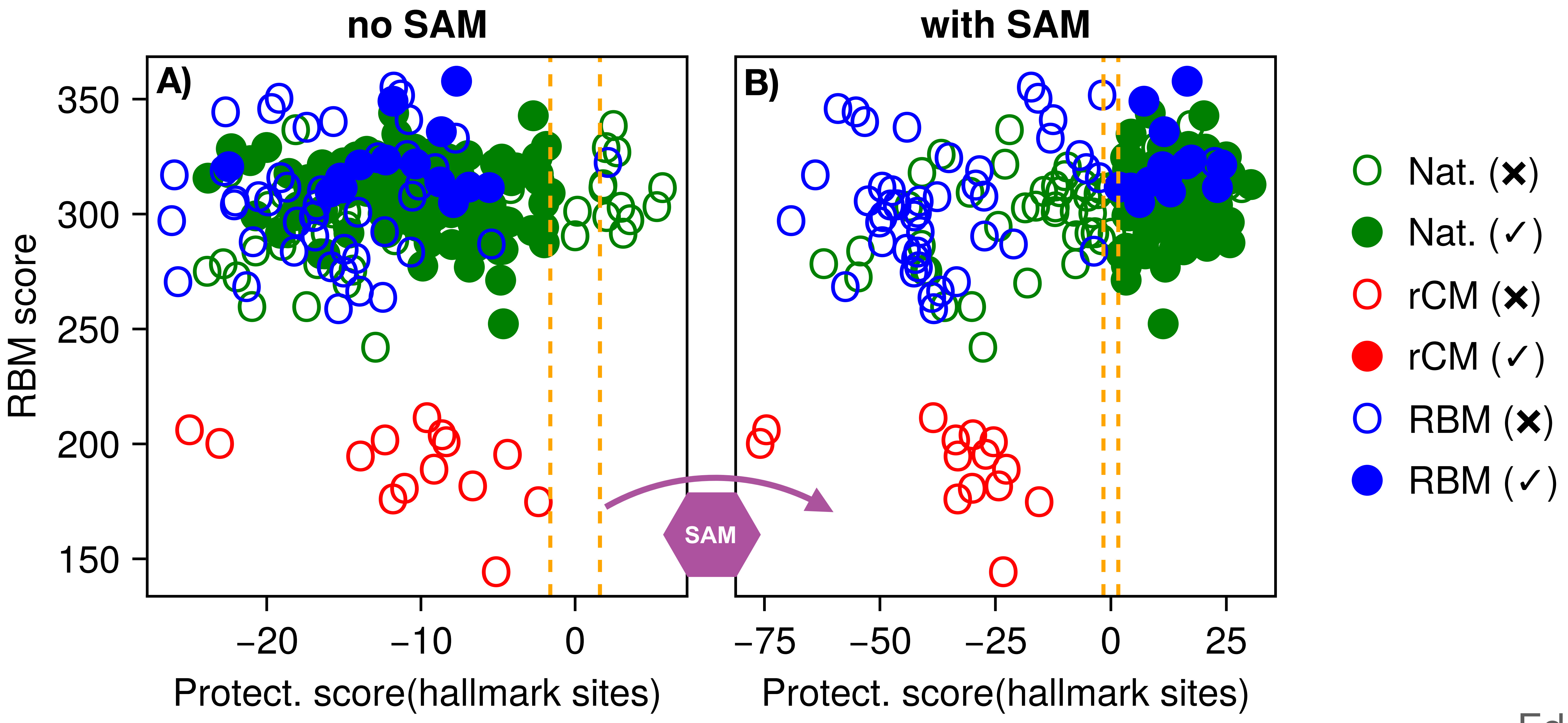
Eddy, Sean R. Ann. Rev. Biophys. 43.1 (2014): 433-456.;
IPANEMAP: Saaidi, Afaf, et al. NAR 48.15 (2020): 8276-8289.

SHAPE reactivity response to SAM in generated sequences



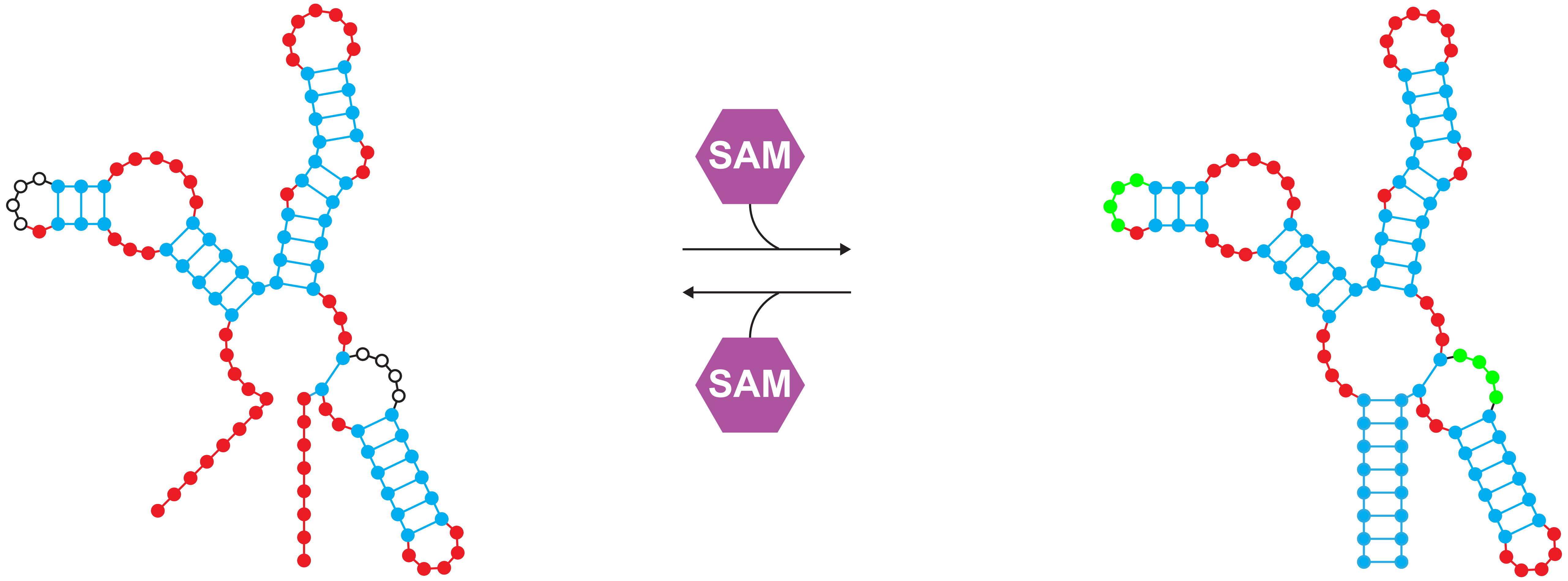
Sequences probed

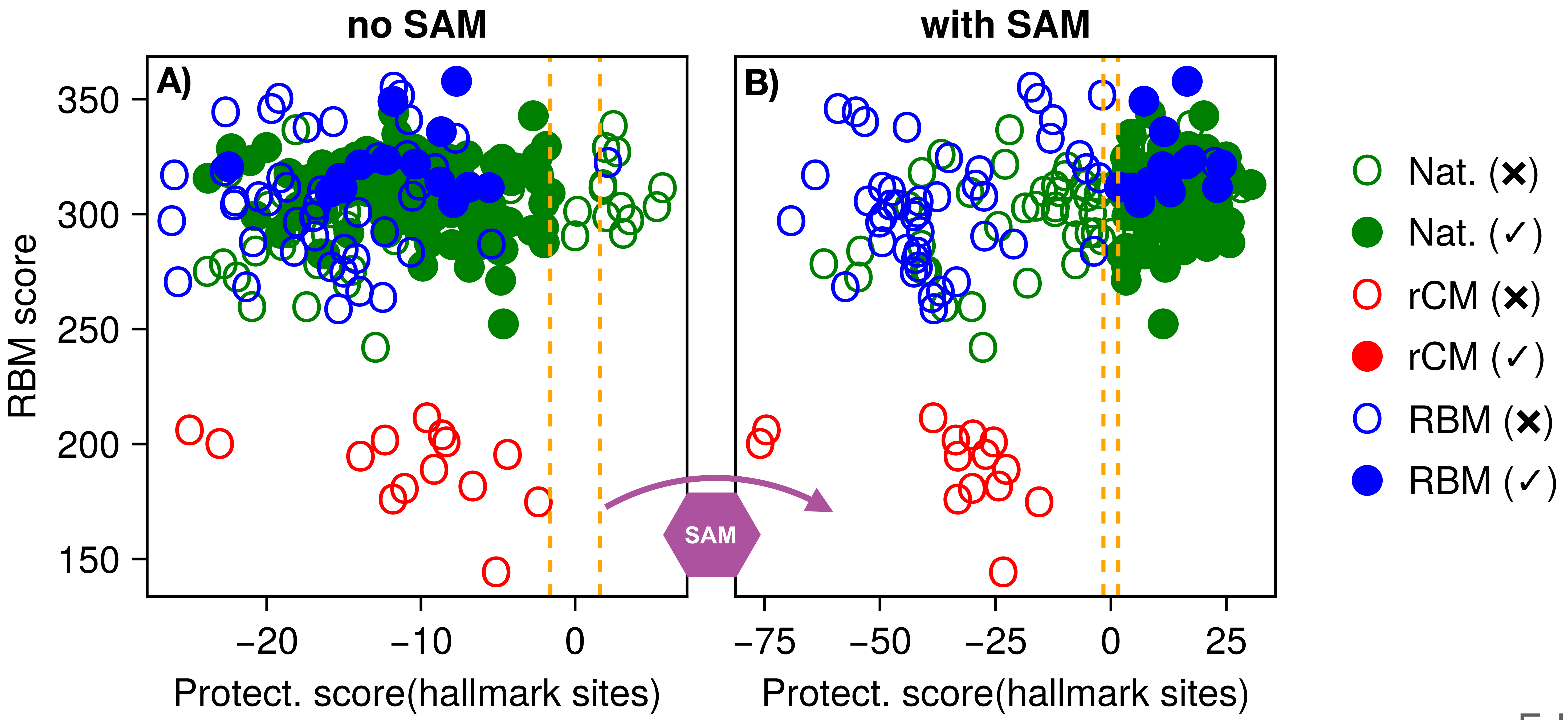
476 RBM, 206 natural



Eddy, Sean R. Ann. Rev. Biophys. 43.1 (2014): 433-456.;
IPANEMAP: Saaidi, Afaf, et al. NAR 48.15 (2020): 8276-8289.

SHAPE reactivity response to SAM in generated sequences





Sequences probed

476 RBM, 206 natural

Results:

$22\,\%$ to $38\,\%$ of RBM generated sequences exhibit expected structural responses to SAM

with $10 - 40\,\%$ divergence from natural sequences and high diversity (e.g. no P4)

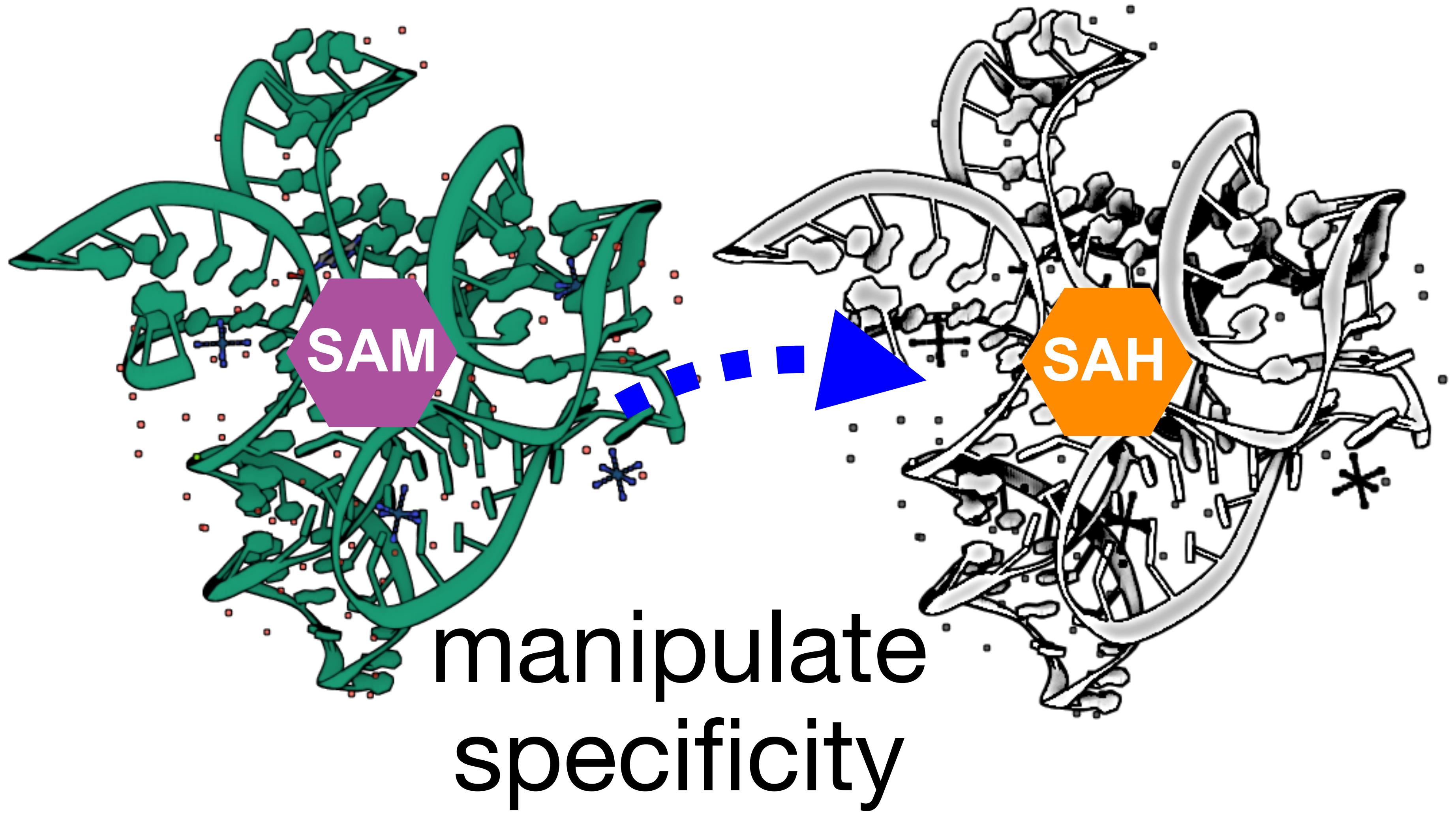
These results support RBM as generative model

Eddy, Sean R. Ann. Rev. Biophys. 43.1 (2014): 433-456.;
IPANEMAP: Saaidi, Afaf, et al. NAR 48.15 (2020): 8276-8289.

Next: Manipulate properties of generated sequences with representation learning

Different properties of natural riboswitches:

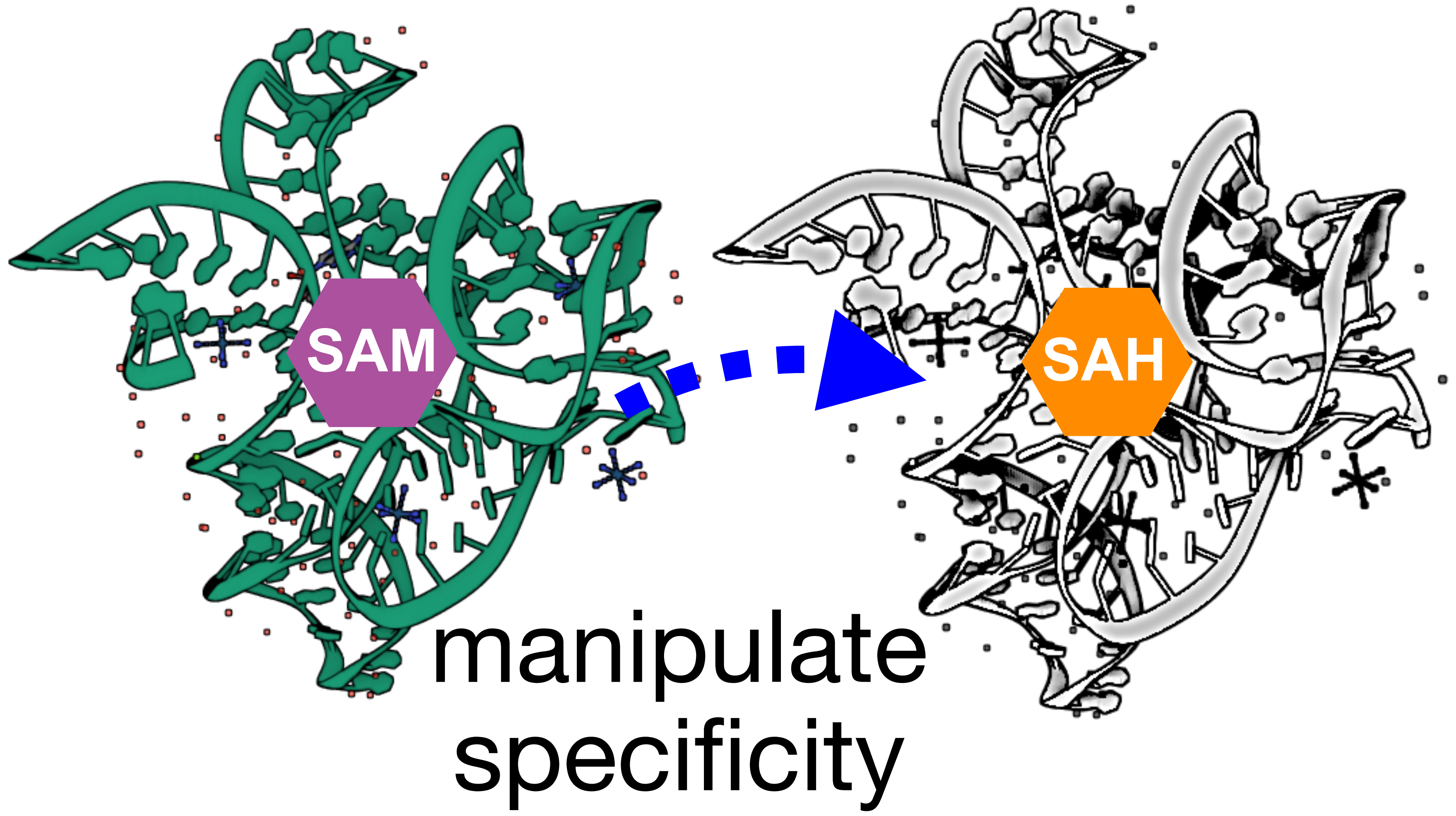
- specificities (SAM vs. SAH)
- ligand affinity ($0.13 - 20$µM)
- thermostabilities

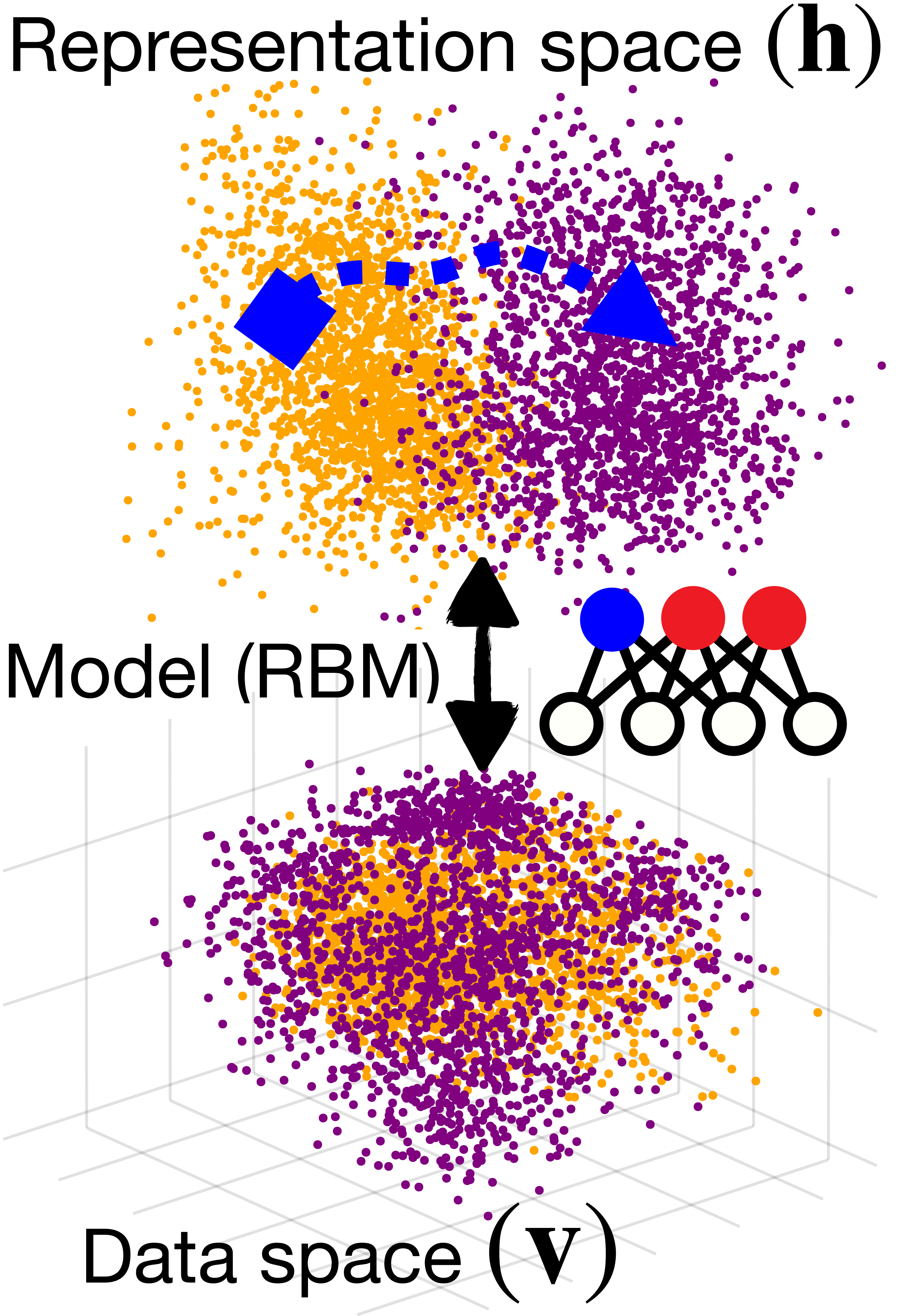


Next: Manipulate properties of generated sequences with representation learning

Different properties of natural riboswitches:

- specificities (SAM vs. SAH)
- ligand affinity $(0.13 - 20\mu\text{M})$
- thermostabilities

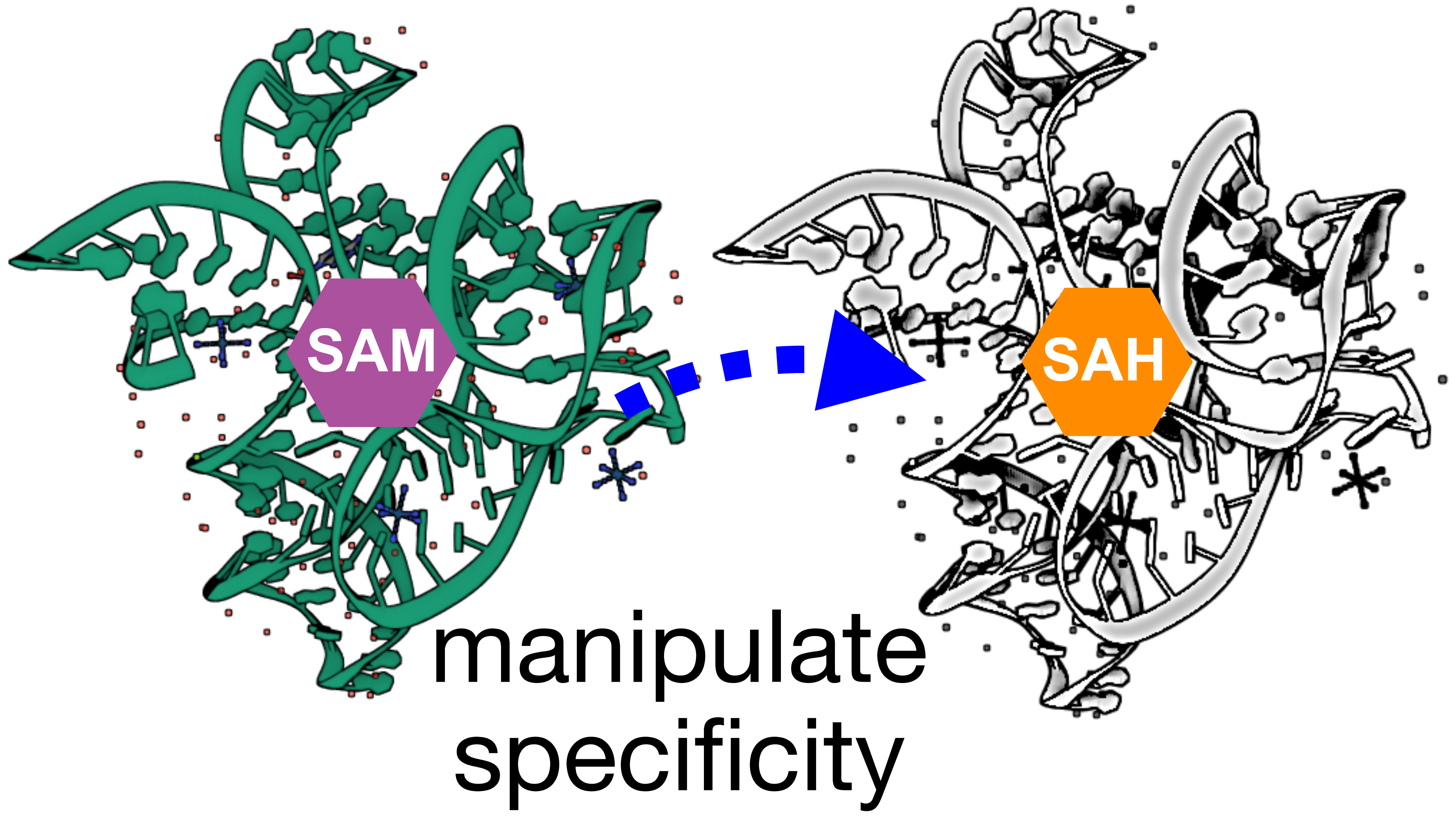




Next: Manipulate properties of generated sequences with representation learning

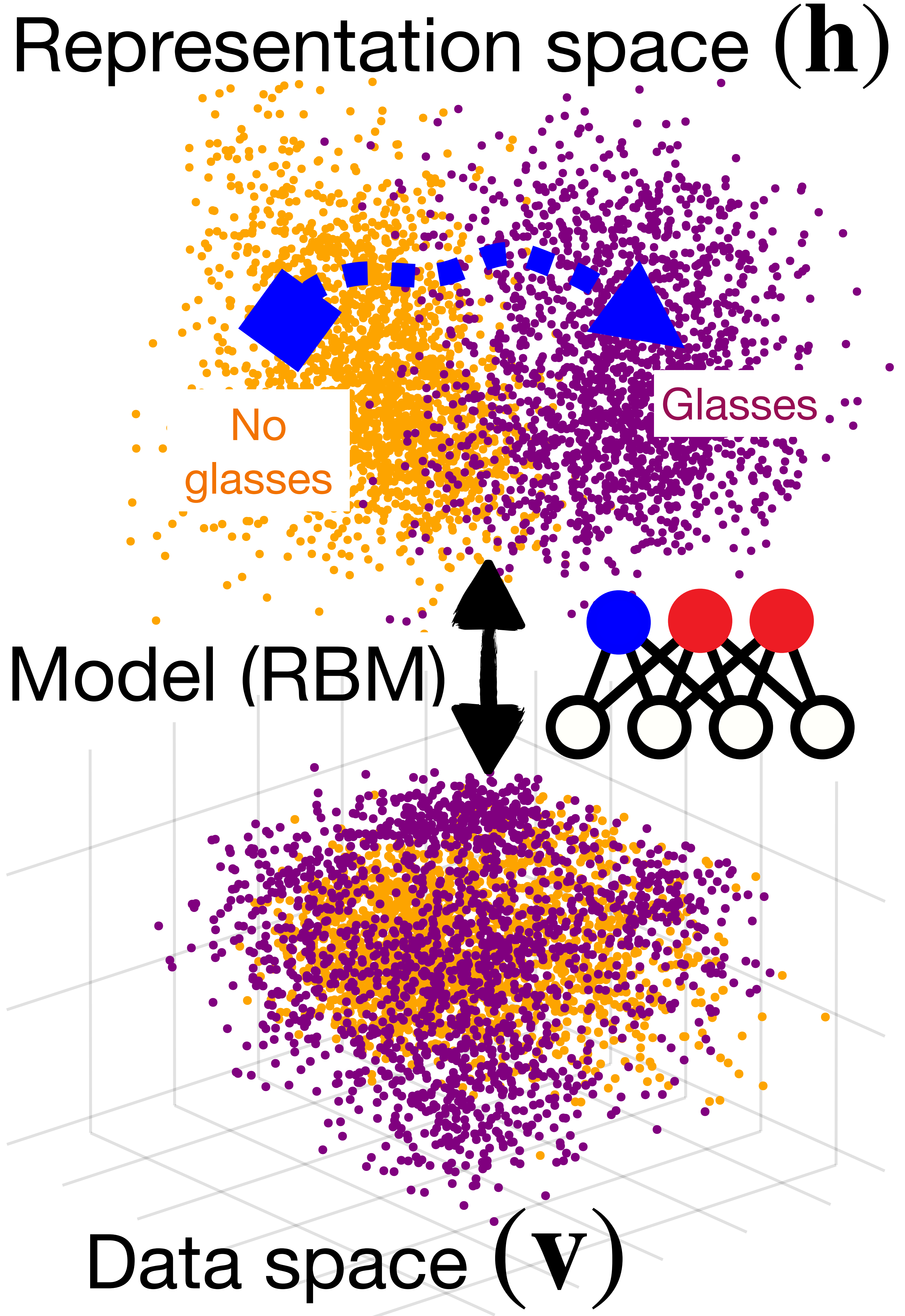
Different properties of natural riboswitches:

- specificities (SAM vs. SAH)
- ligand affinity $(0.13 - 20\mu\text{M})$
- thermostabilities



manipulate specificity

[**JFdCD** et al PRX’2023]



Thank you

ENS, Paris

Andrea Di Gioacchino
Rémi Monasson
Simona Cocco

CitCOM, Université Paris Cité

Pierre Hardouin
Francois-Xavier Lyonnet
Bruno Sargueil

LIX Ecole Polytechnique, Palaiseau

Bertrand Marchand
Yann Ponty

10.1101/2023.05.10.5

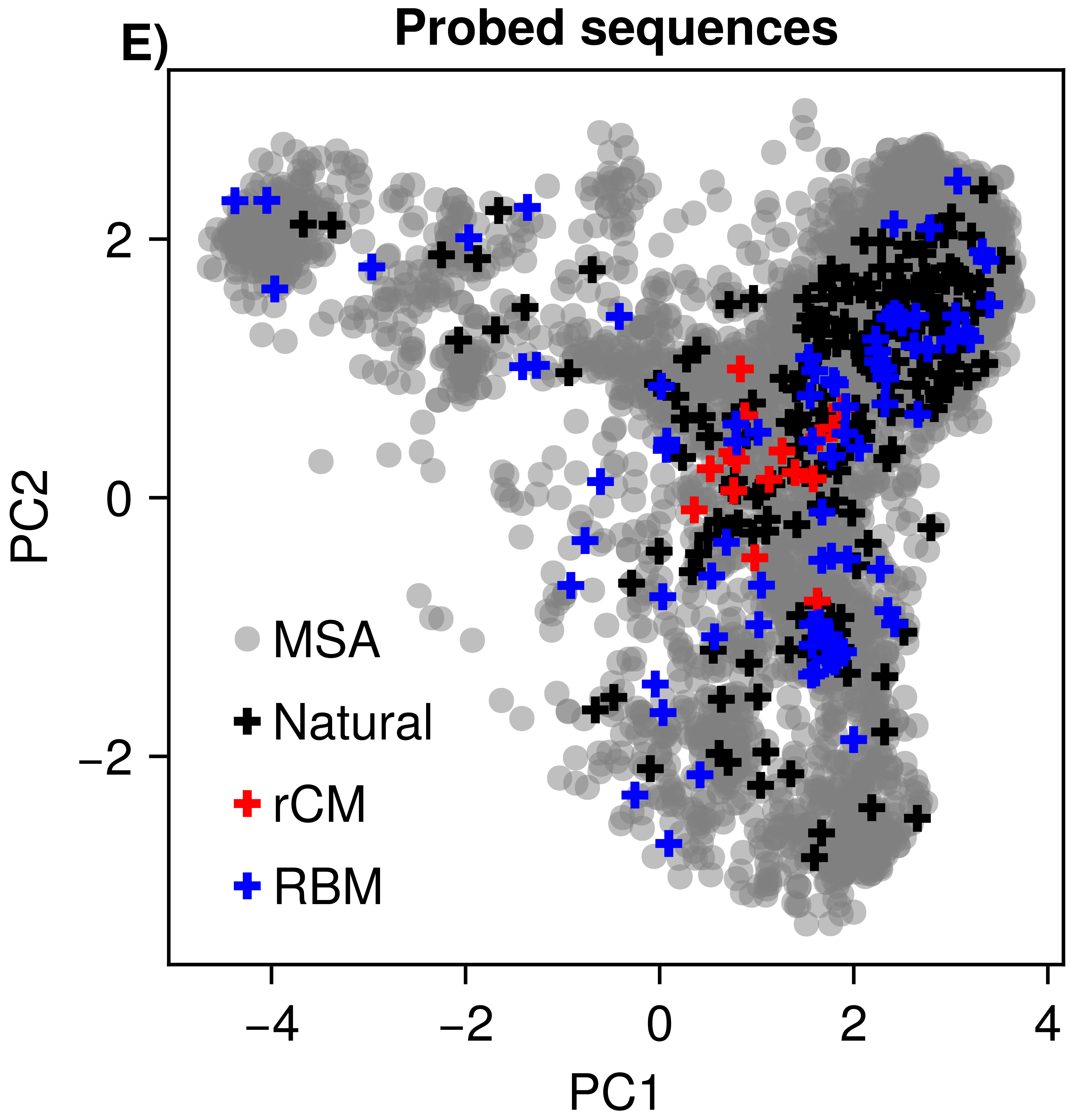
bioRxiv

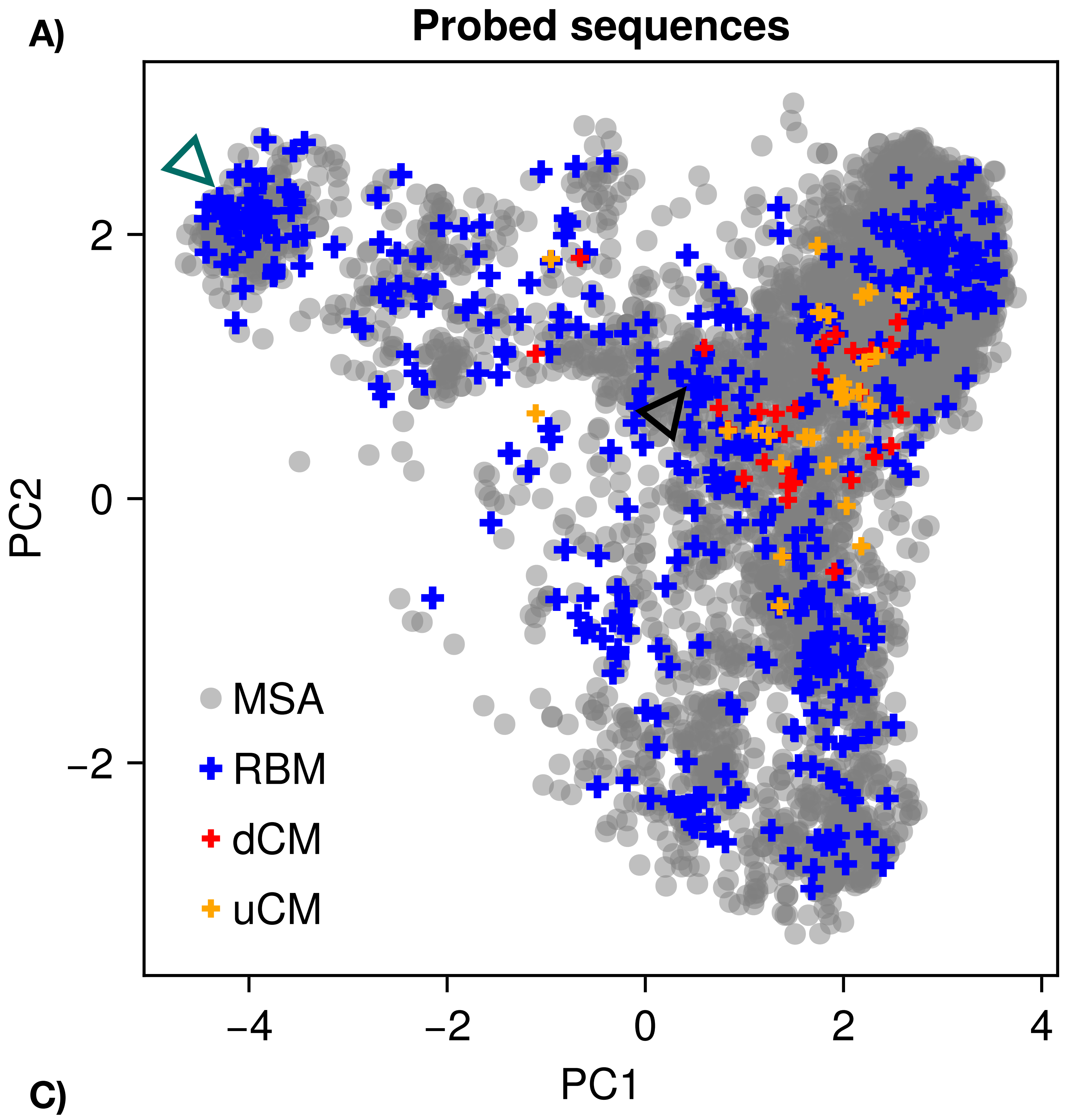
THE PREPRINT SERVER FOR BIOLOGY

JFdCD et al bioRxiv:2023.05.10.540155

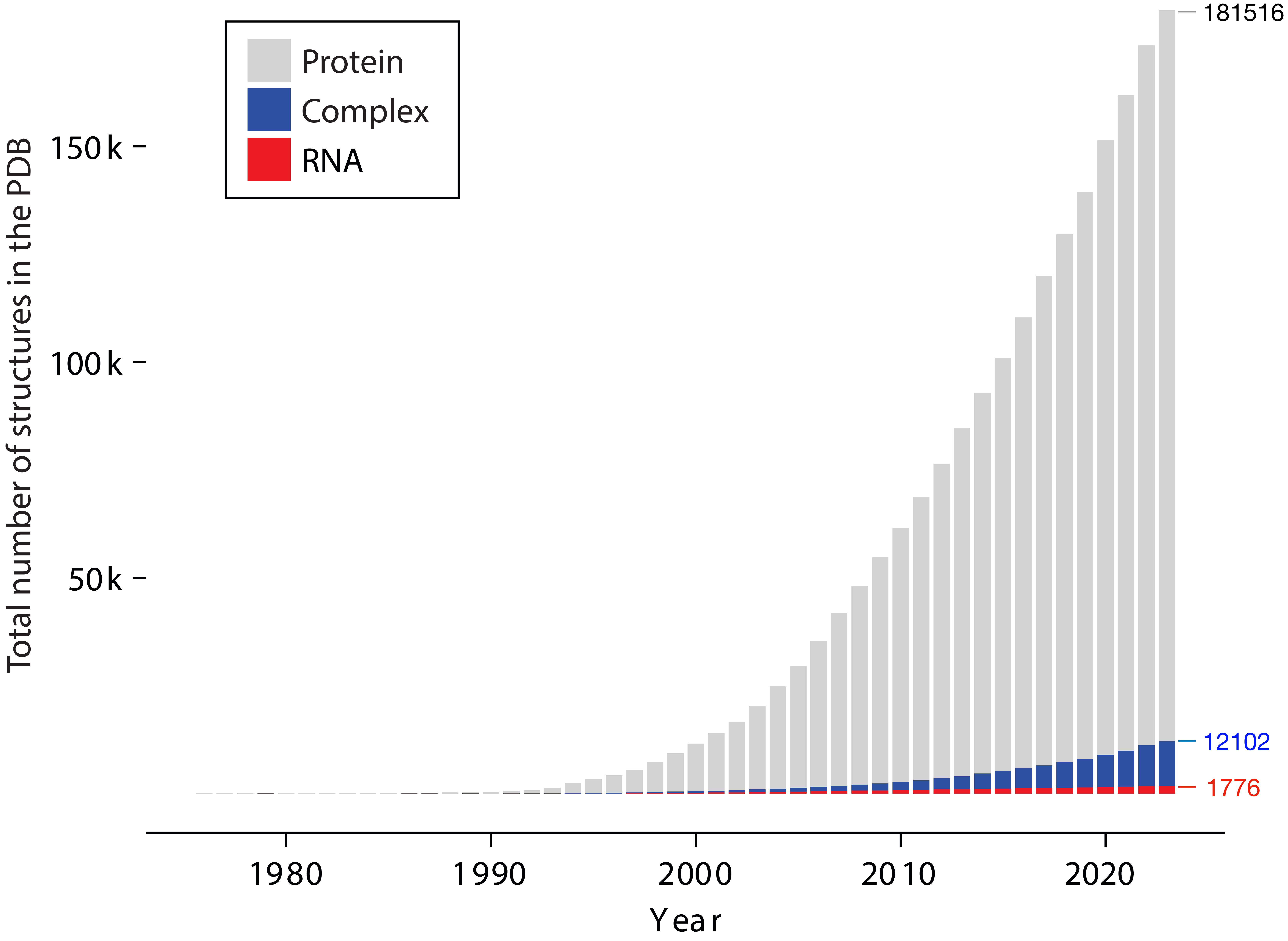
Backup

PCA projections of probed sequences





Structure and function of non-coding RNA molecules



Less RNA structural data

Structure and function of non-coding RNA molecules

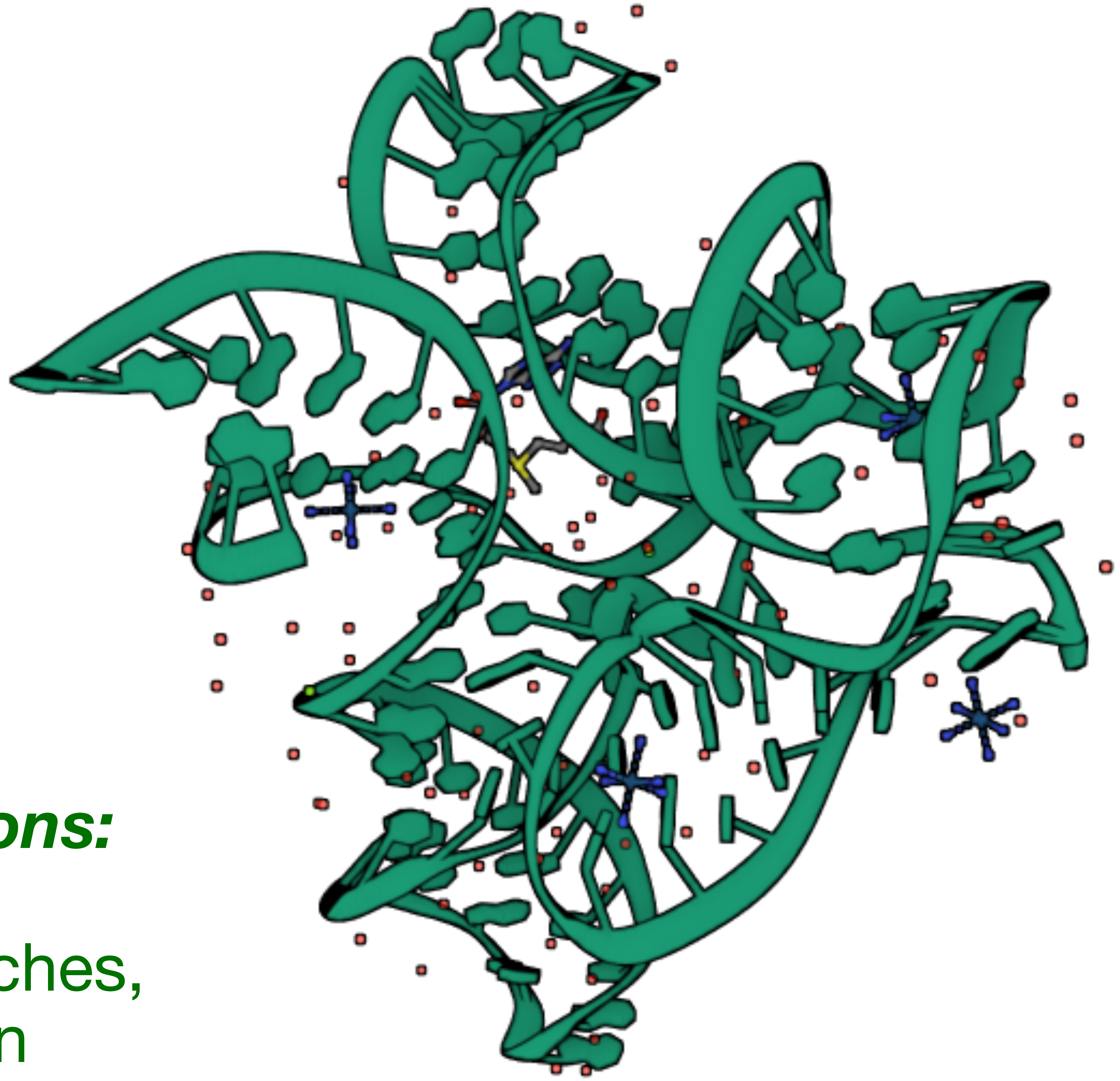
RNA: linear chain of nucleotides (four types: A,U,C,G)

. . . GGCUUAUCAAGAGAGGGU . . .

Central dogma: DNA → mRNA → protein

role as information carrier

Tinoco, Bustamante. How RNA folds. *J Mol Biol.* 1999



SAM riboswitch

Large fraction of RNA is non-coding

rRNA 80-90% of cell RNA content

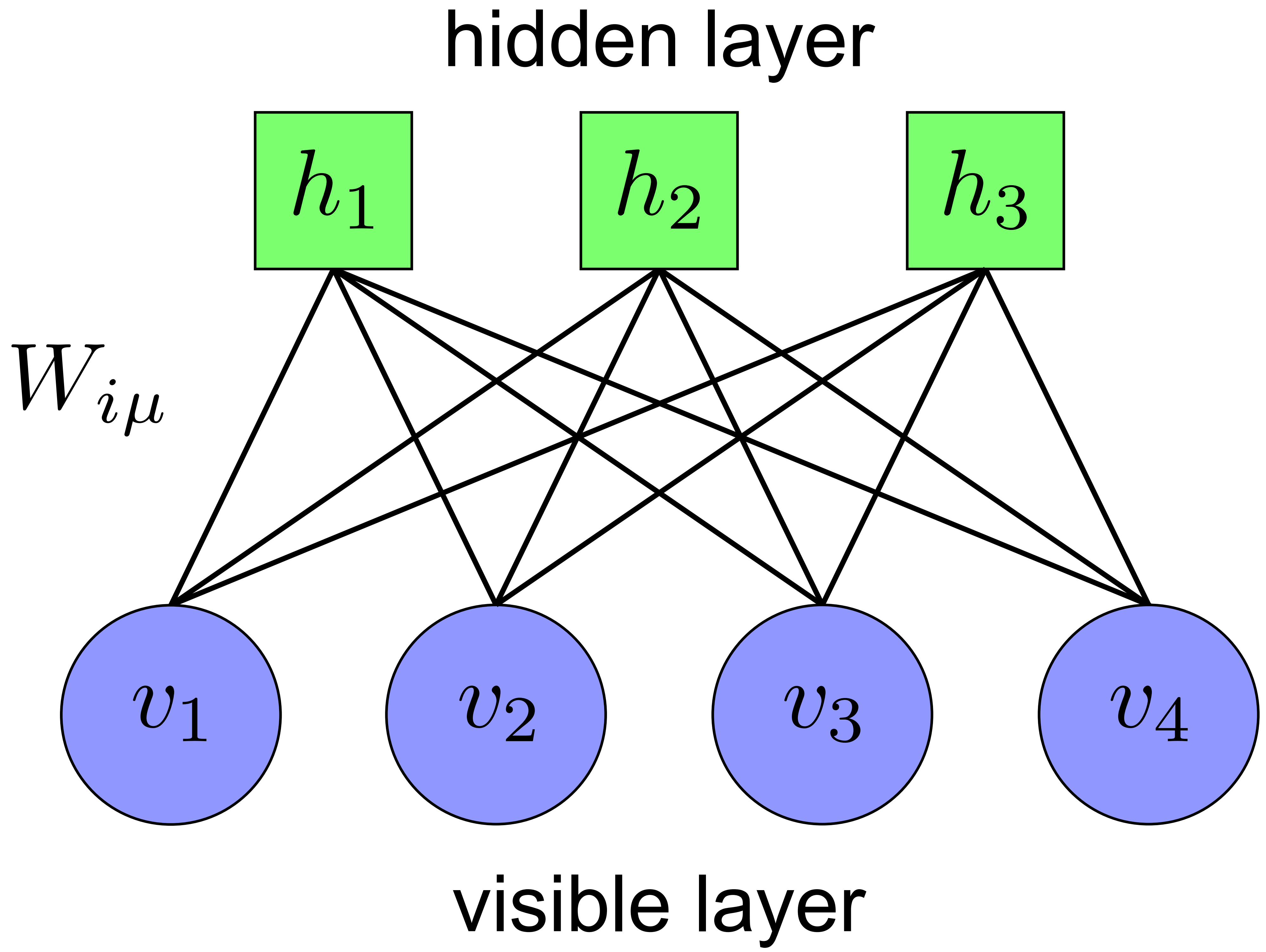
tRNA 10-15%, mRNA 3-7%

Many ncRNA functions:

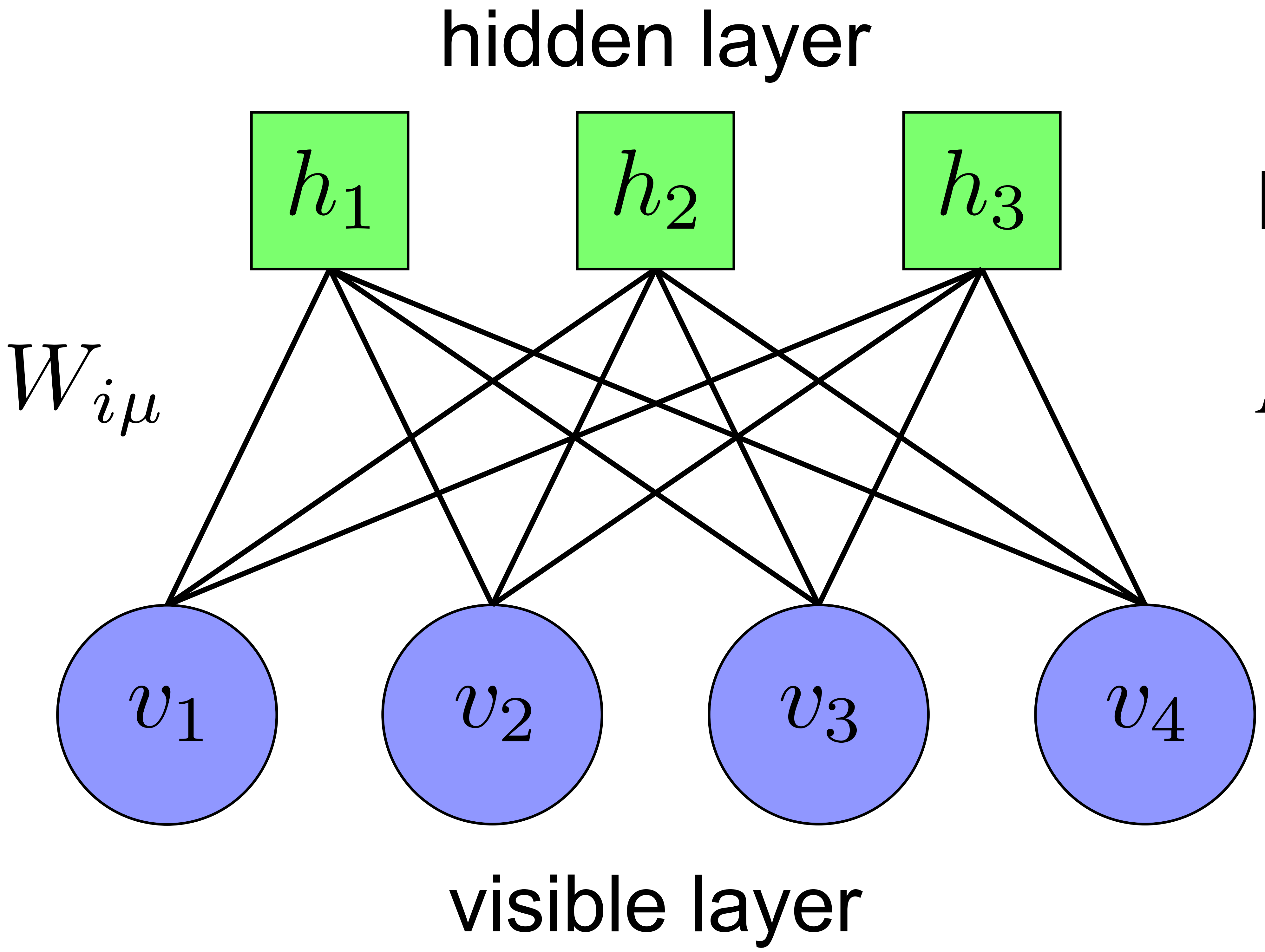
Rybozymes, riboswitches, T box, potential role in origin of life, …

https://assets.thermofisher.com/TFS-Assets/BID/Technical-Notes/collibri-stranded-rna-library-prep-kit-total-rna-seq-mrna-seq-technical-note.pdf

Restricted Boltzmann Machines (RBM)



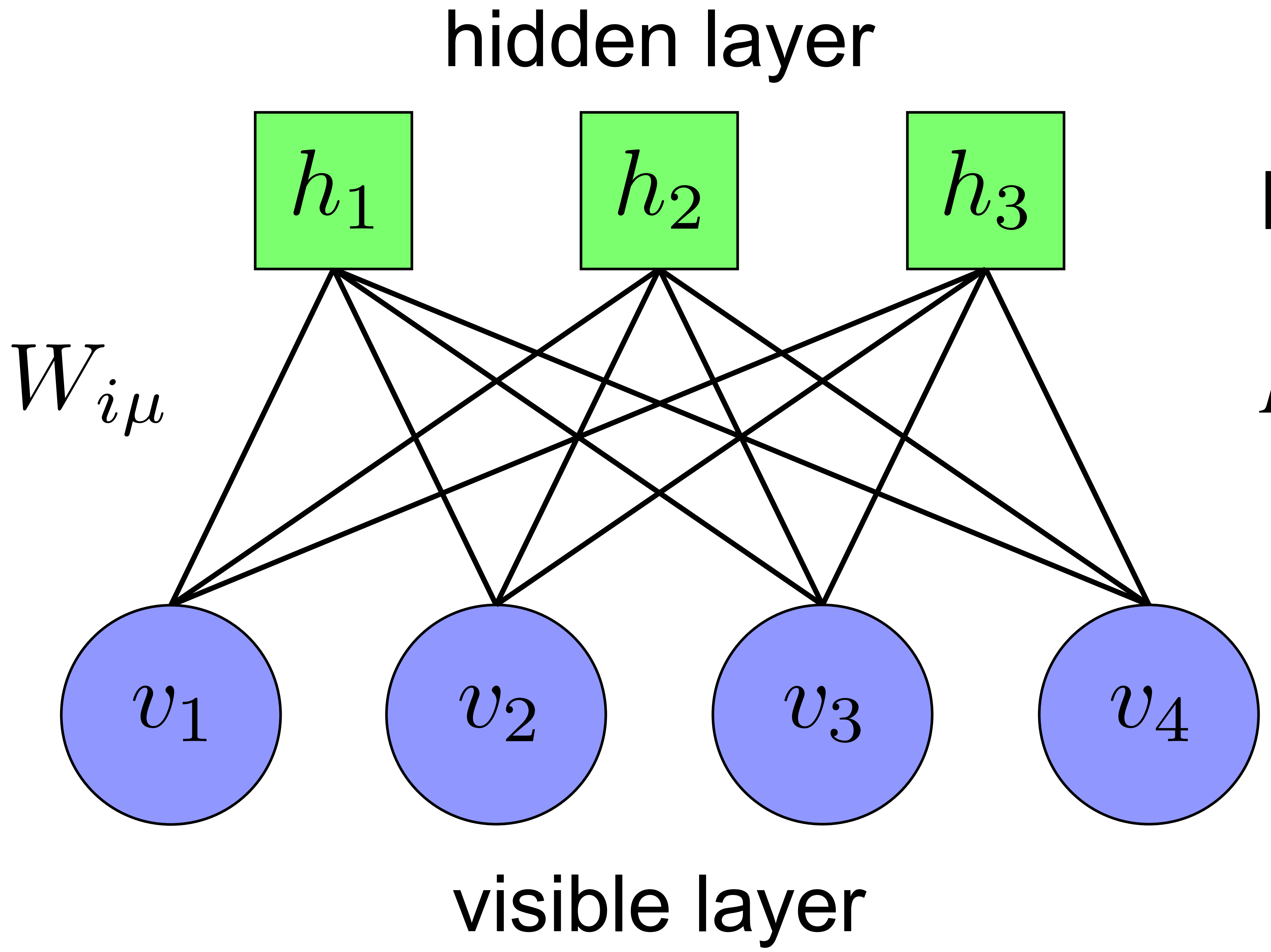
Restricted Boltzmann Machines (RBM)



Energy

$$E(\mathbf{v},\mathbf{h}) = \sum_{\mu} \mathscr{V}_i(v_i) + \sum_{\mu} \mathscr{U}_{\mu}(h_{\mu}) - \sum_{i\mu} w_{i\mu} v_i h_{\mu}$$

Restricted Boltzmann Machines (RBM)



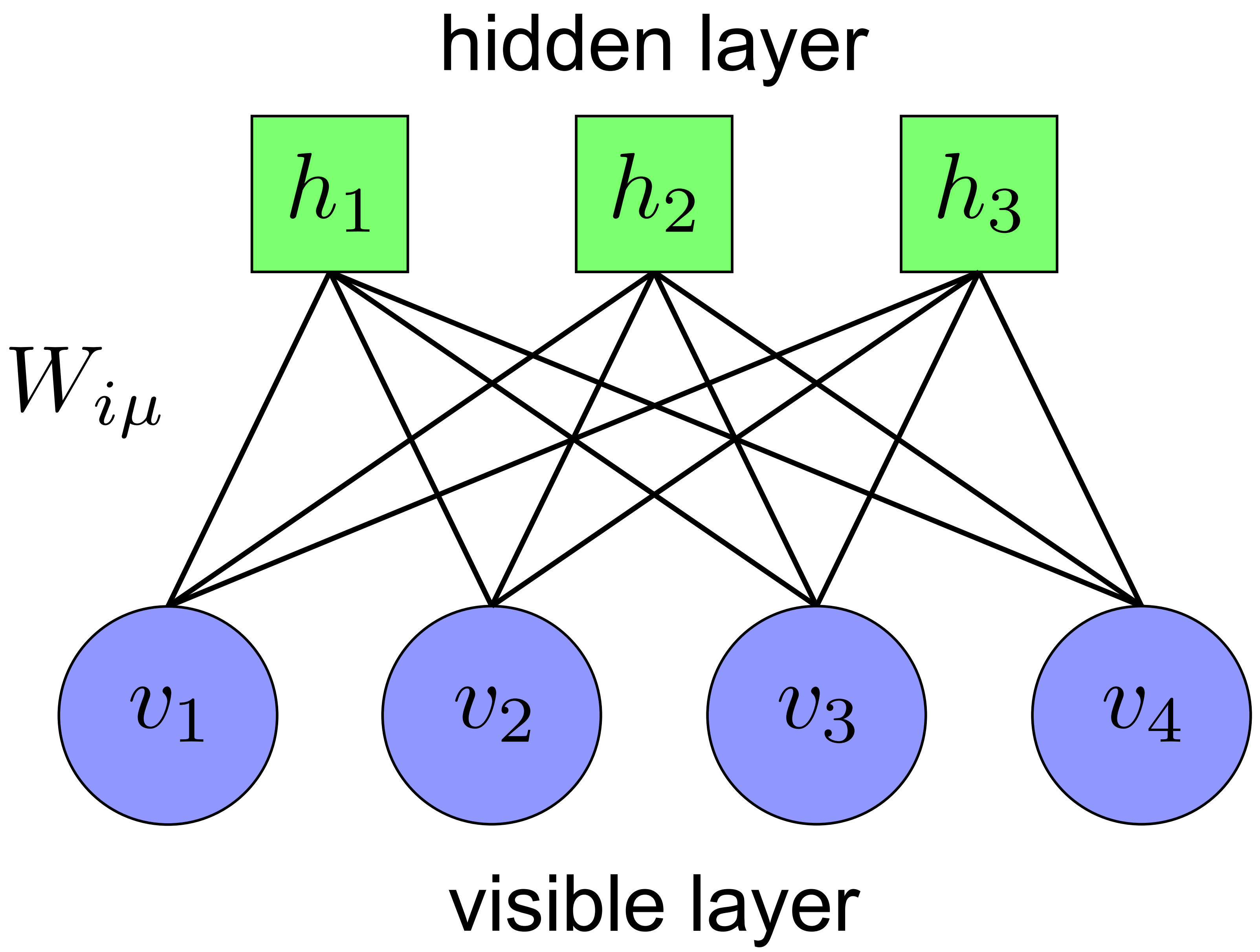
Energy

$$E(\mathbf{v},\mathbf{h}) = \sum_{\mu} \mathscr{V}_i(v_i) + \sum_{\mu} \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} w_{i\mu} v_i h_\mu$$

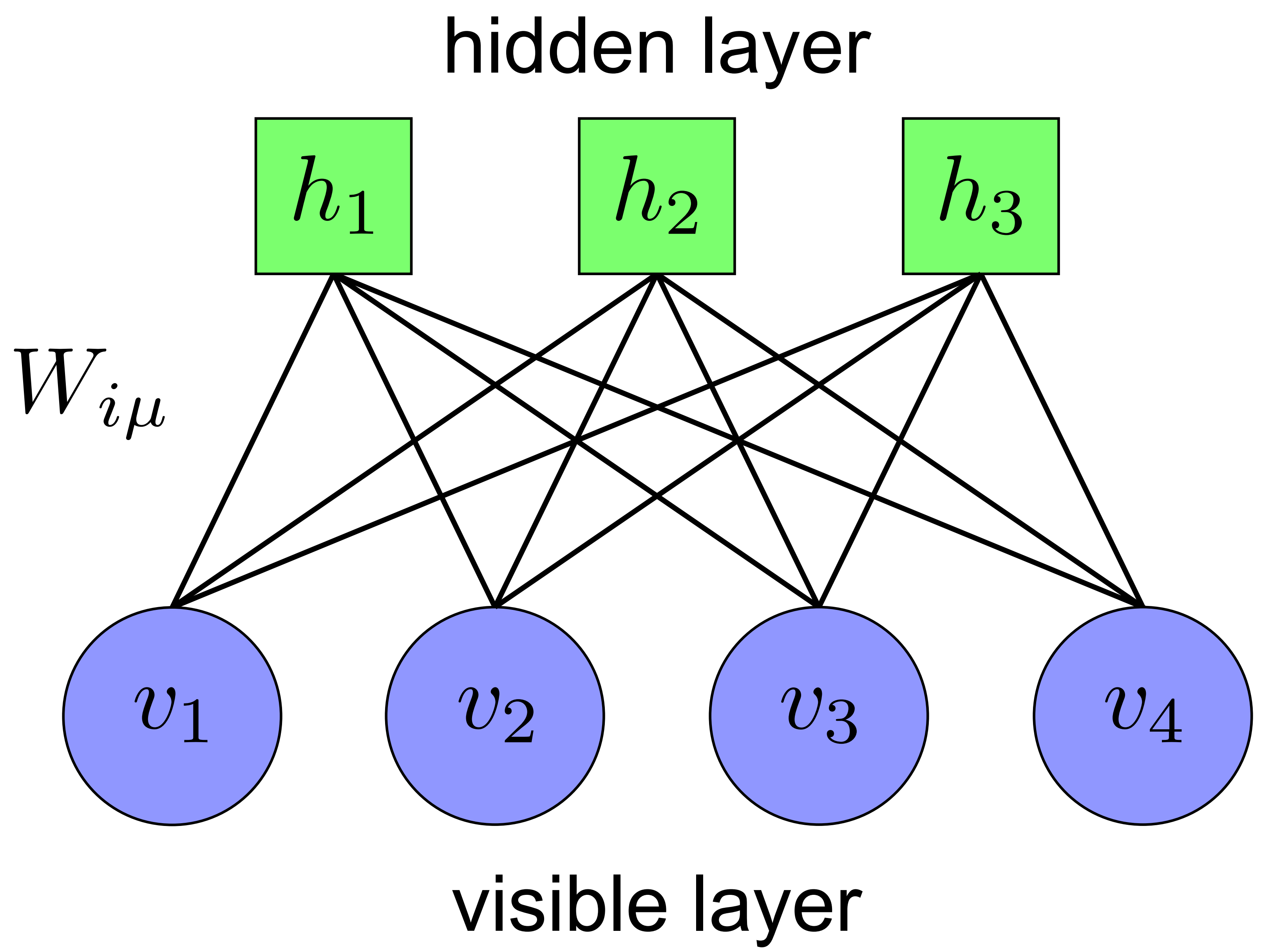
Boltzmann probability law

$P(\mathbf{v},\mathbf{h}) = \dfrac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}$, where $Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$ is the partition function.

Restricted Boltzmann Machines (RBM)



Restricted Boltzmann Machines (RBM)

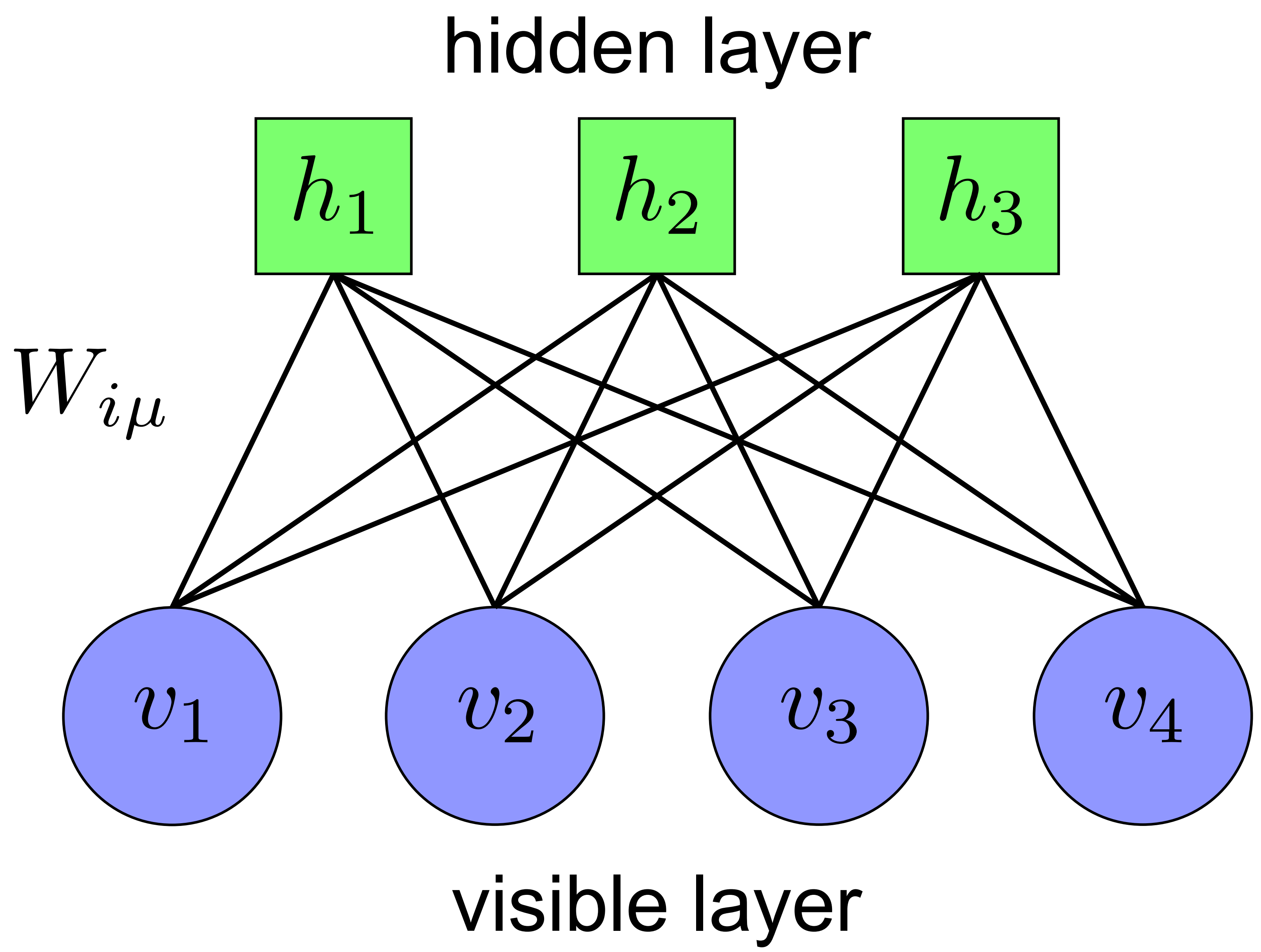


$\mathbf{v}$ = the RNA sequence; $\mathbf{h}$ = latent variables (features)

Likelihood: $P_v(\mathbf{v}) = \dfrac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \dfrac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})}$

where: $E_{\text{eff}}(\mathbf{v}) = \sum_i \mathscr{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathscr{U}_\mu(h_\mu)}$

Restricted Boltzmann Machines (RBM)

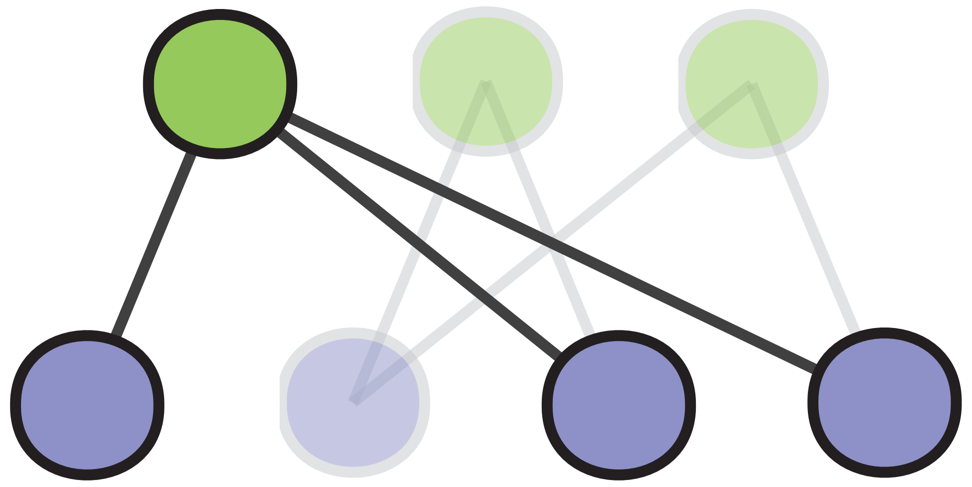


$\mathbf{v}$ = the RNA sequence; $\mathbf{h}$ = latent variables (features)

Likelihood: $P_v(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \frac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})}$

where: $E_{\text{eff}}(\mathbf{v}) = \sum_i \mathscr{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathscr{U}_\mu(h_\mu)}$

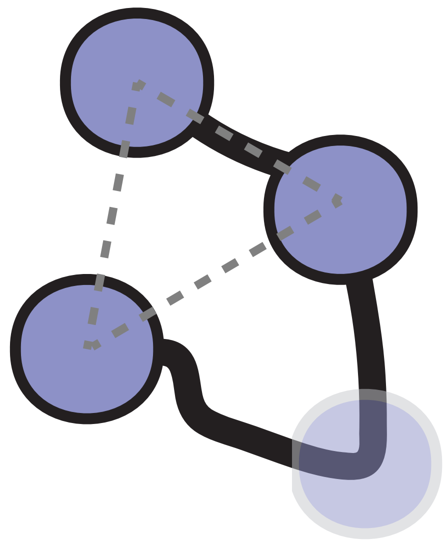
Full RBM



visible units interact
with hidden units

marginalization
of latent variables

Effective model



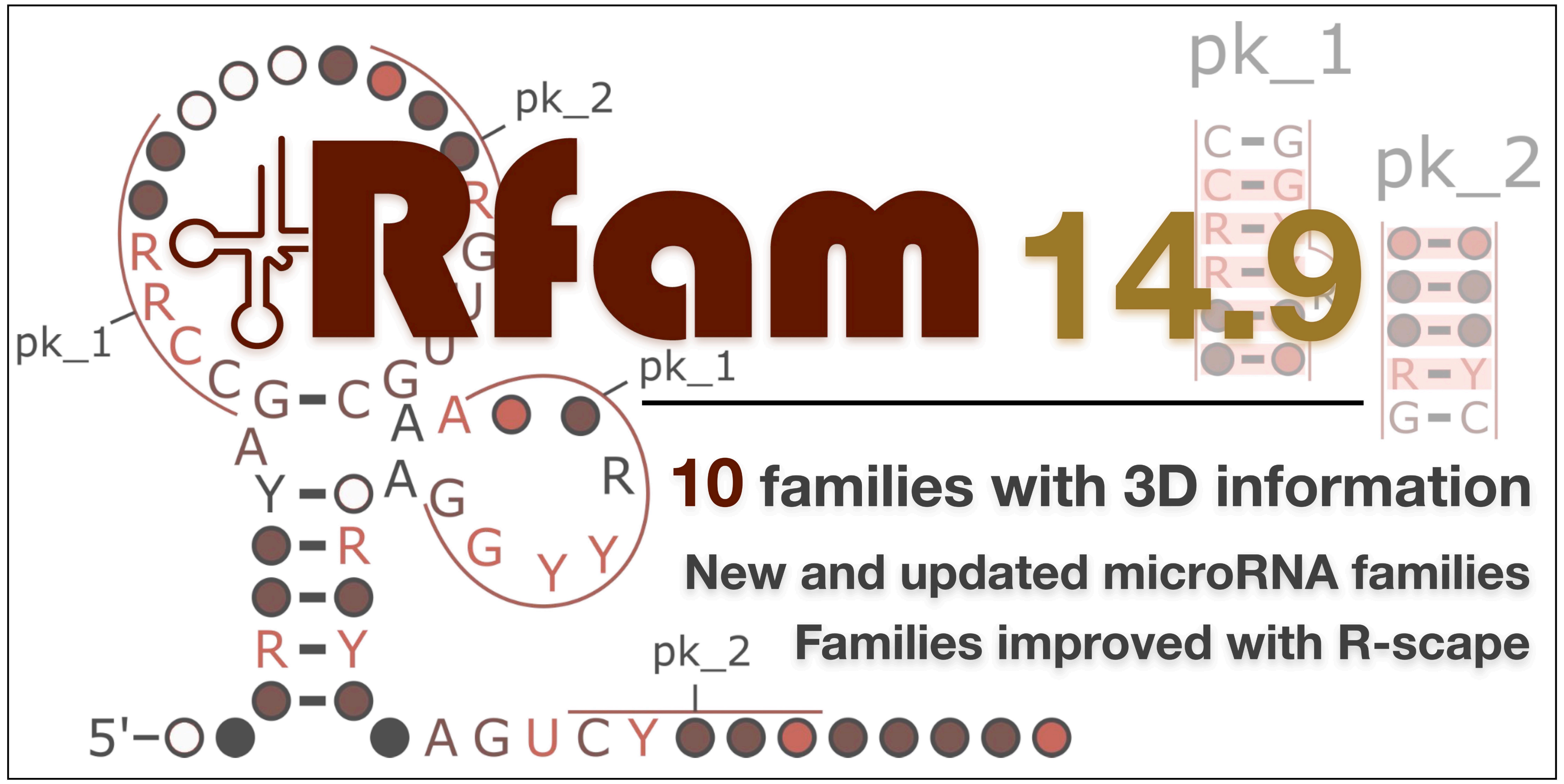
effective interactions
between visible units

The SAM-I aptamer domain family

SAM-I riboswitch aptamer domain RNA family (RF00162)

6161 sequences

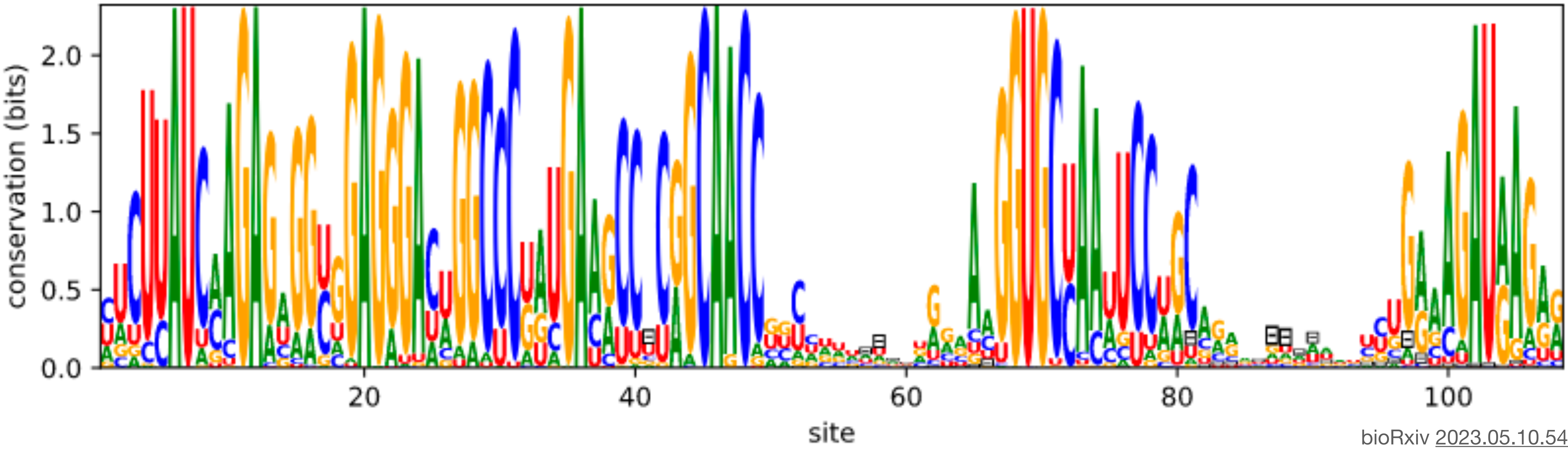
108 aligned positions



10 families with 3D information

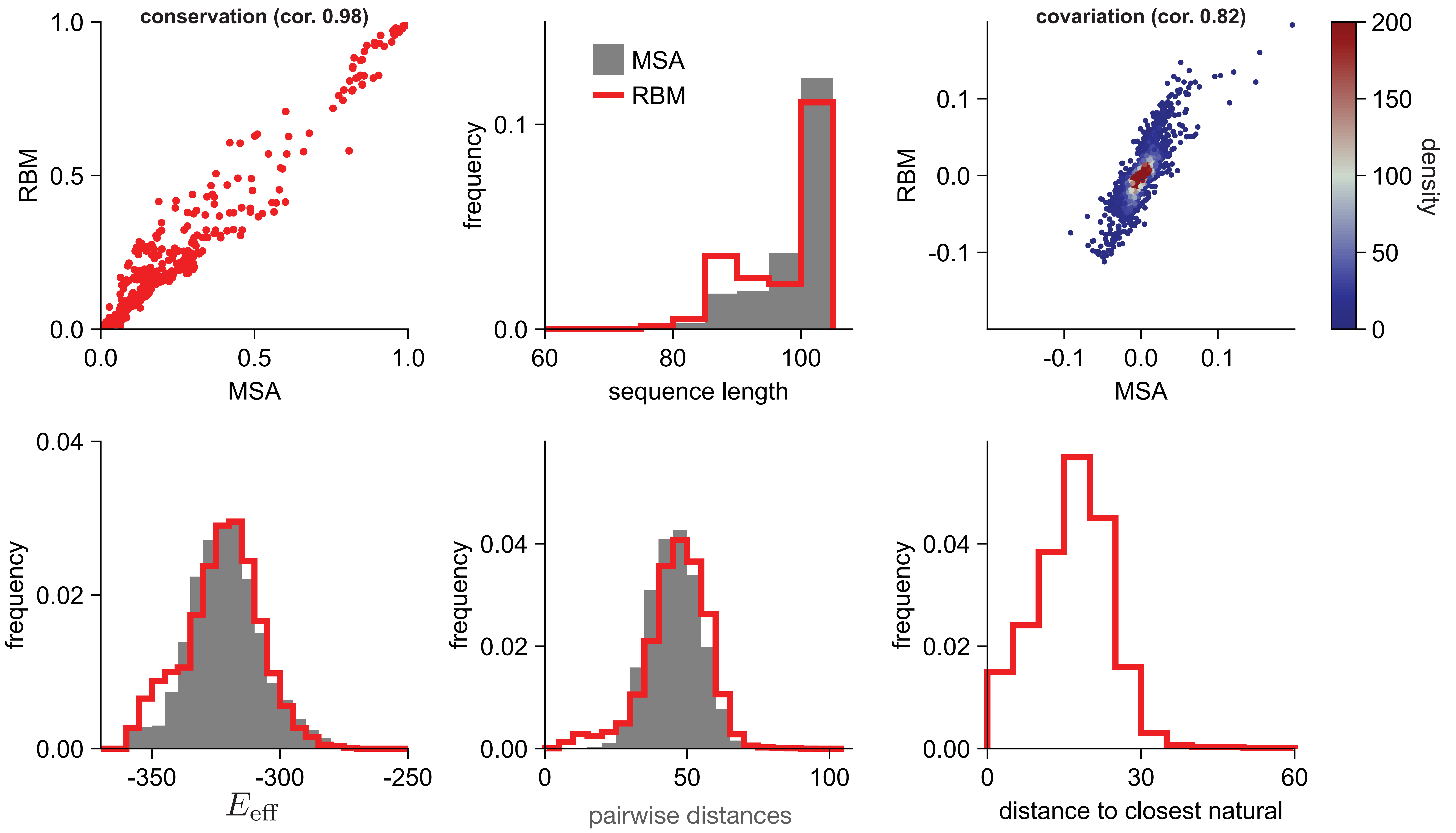
New and updated microRNA families

Families improved with R-scape



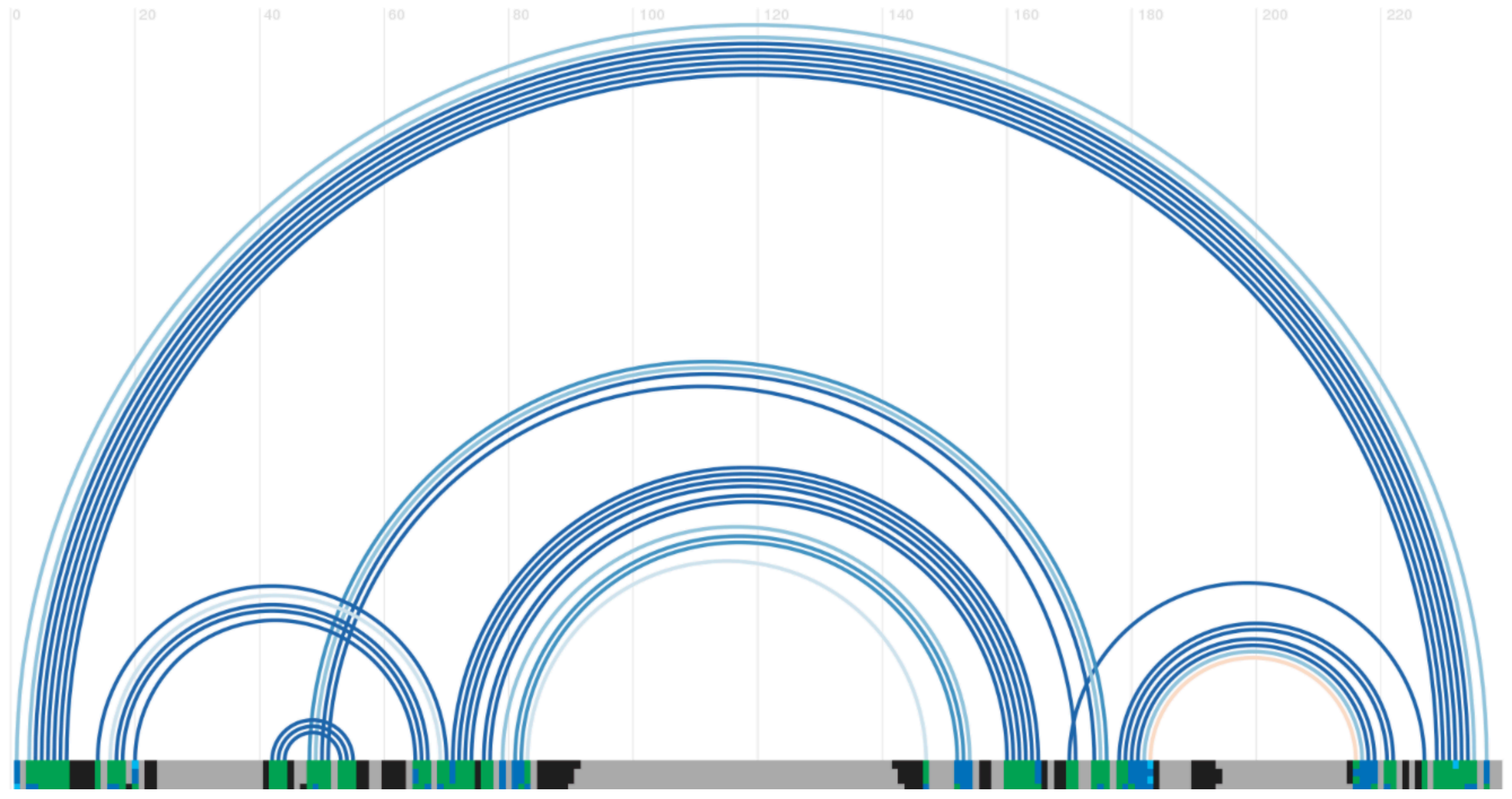
bioRxiv 2023.05.10.540155

RBM reproduces statistics of SAM family

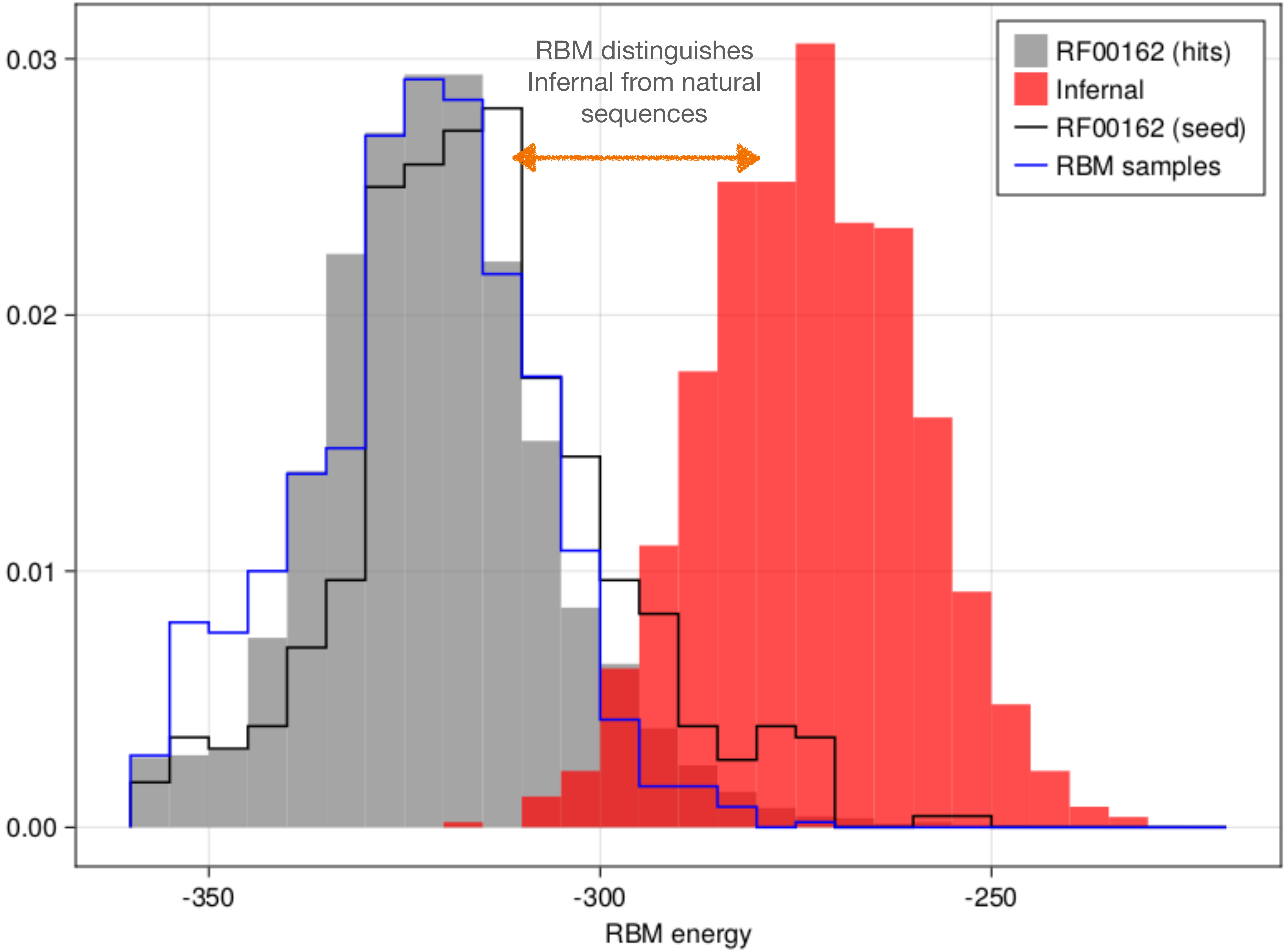


Secondary structure is not sufficient

Compatibility with secondary structure



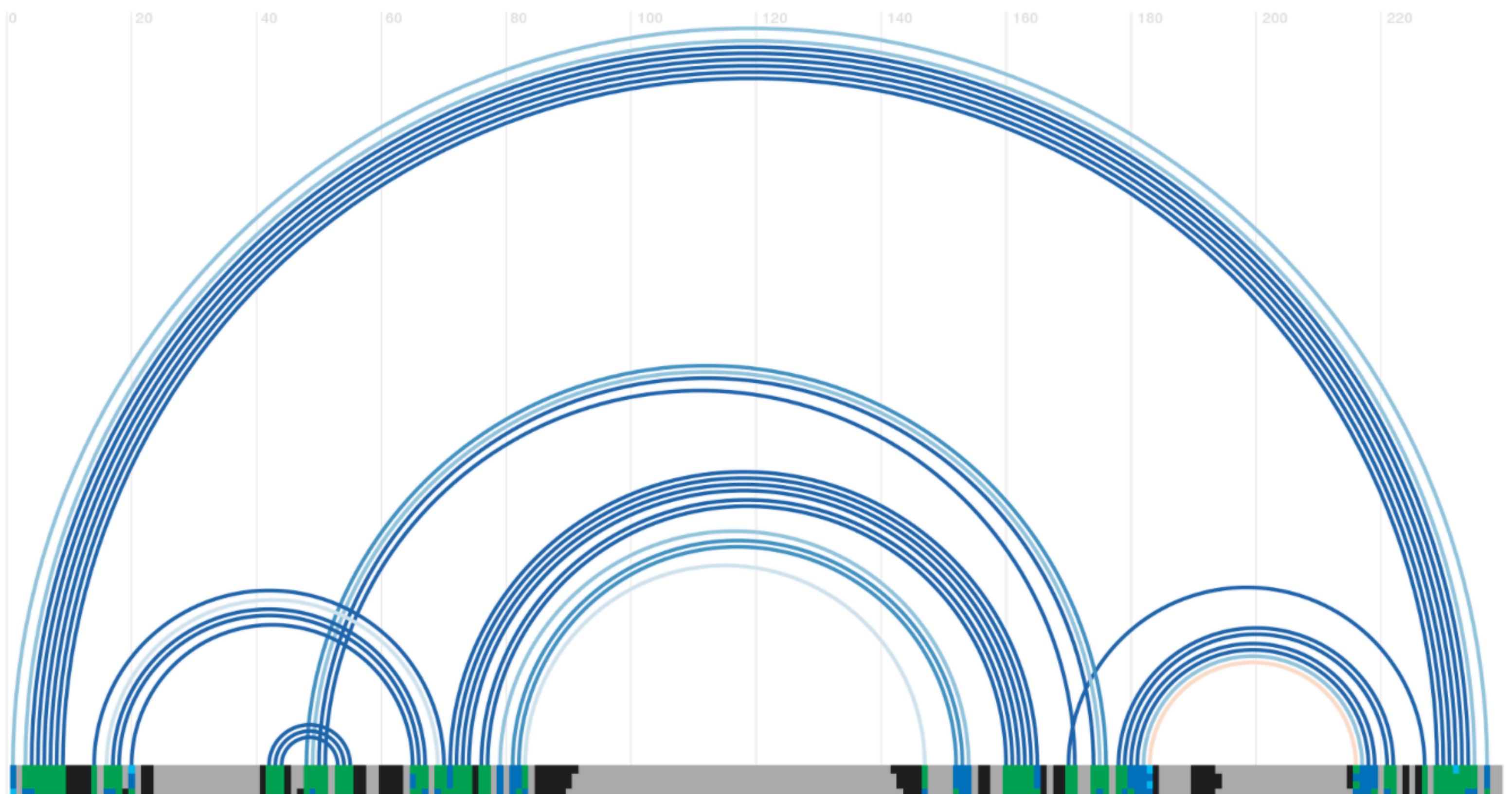
Infernal (Covariance model)



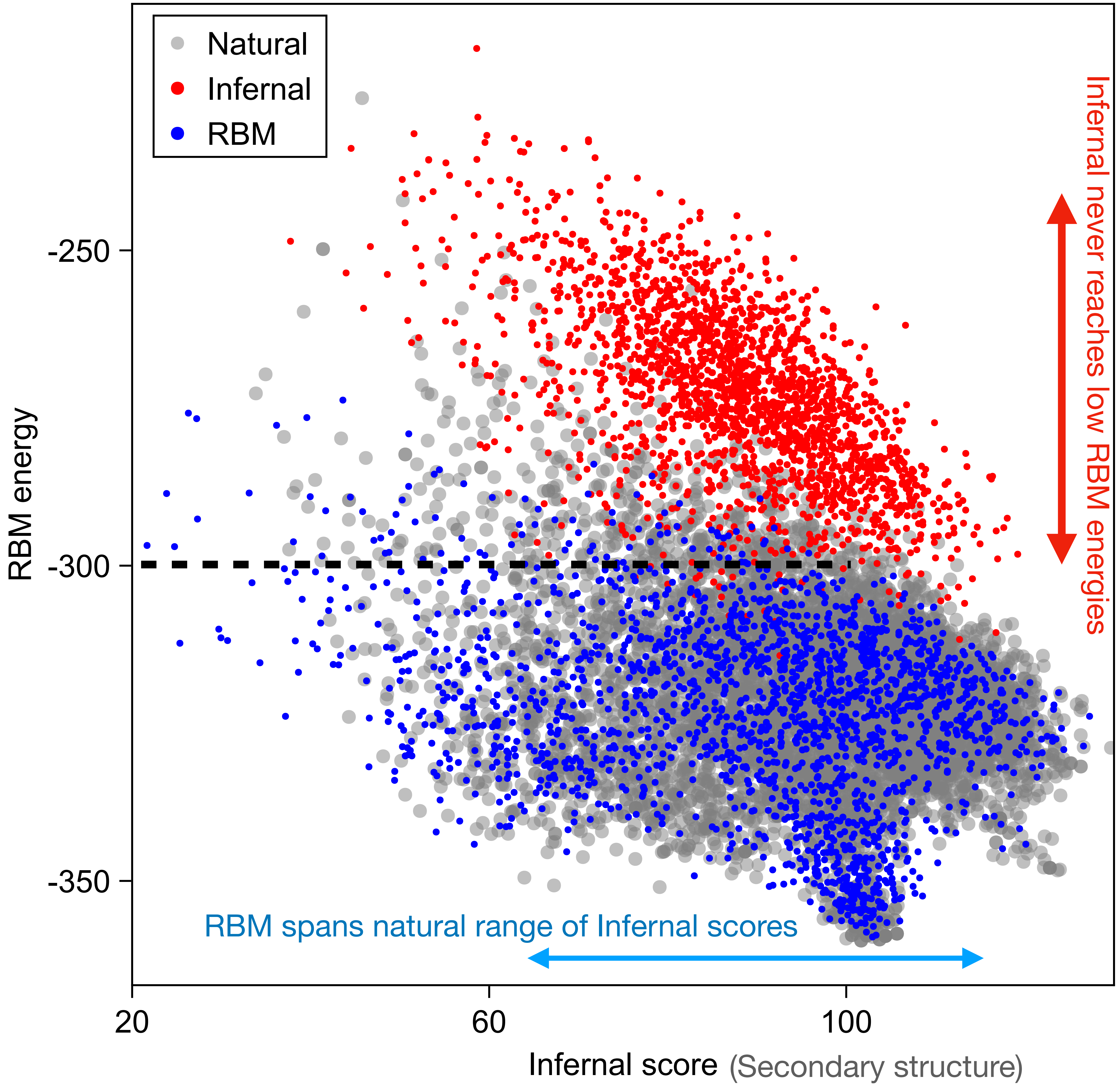
RBM distinguishes Infernal from natural sequences

RBM samples are compatible with secondary structure

Compatibility with secondary structure



Infernal (Covariance model)



Infernal never reaches low RBM energies

RBM spans natural range of Infernal scores

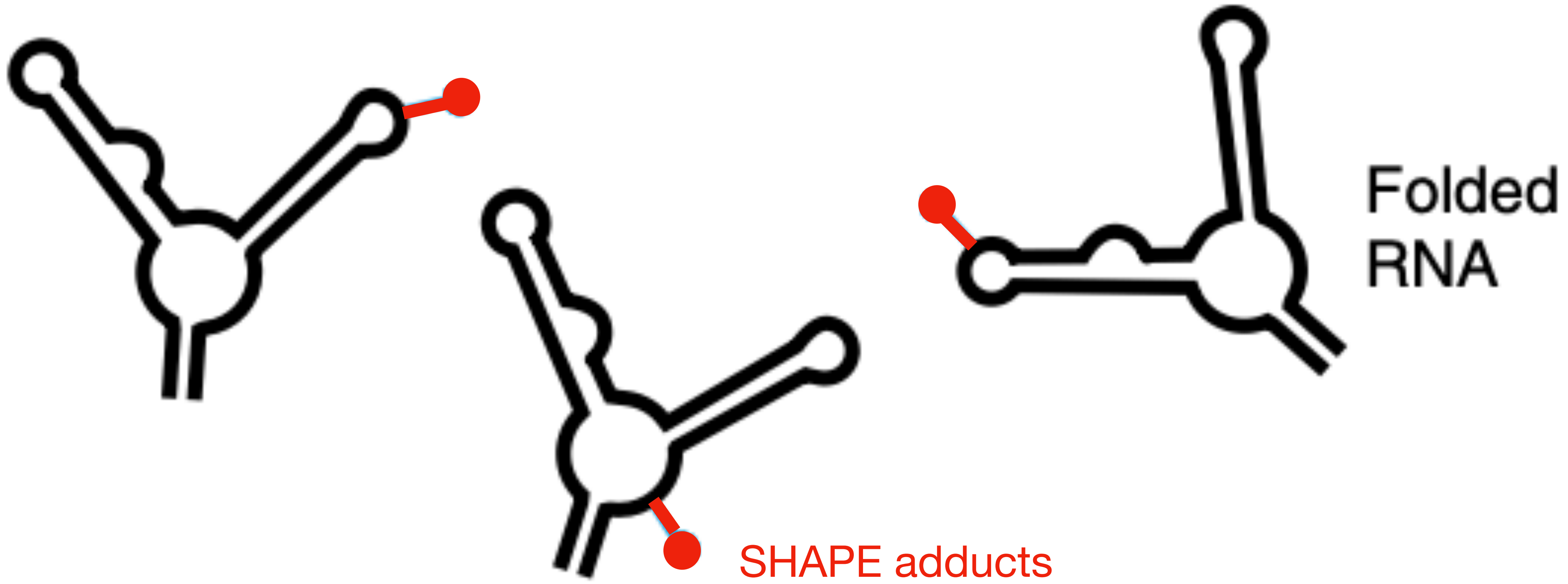
Experimental validation with SHAPE-MaP

UNIVERSITÉ
PARIS
DESCARTES

with P. Hardouin, F.-X. Lyonnet, B. Sargueil

Wilkinson, KA et al. "SHAPE: quantitative RNA structure analysis at single nucleotide resolution." Nature Protocols 1.3 (2006): 1610-1616.

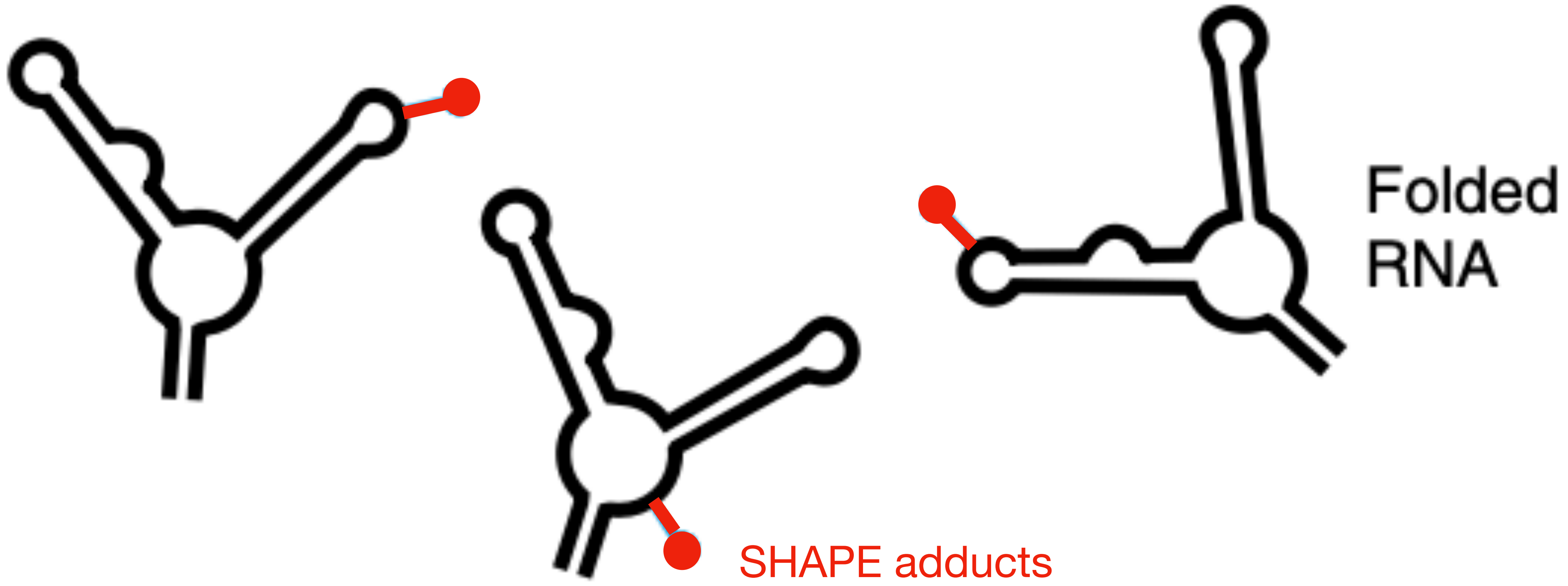
Experimental validation with SHAPE-MaP



with P. Hardouin, F.-X. Lyonnet, B. Sargueil

Wilkinson, KA et al. "SHAPE: quantitative RNA structure analysis at single nucleotide resolution." Nature Protocols 1.3 (2006): 1610-1616.

Experimental validation with SHAPE-MaP



After adding the probe, the RNA molecules are sequenced. The probe results in sequencing errors.

Reactivity: $r = \dfrac{m - u}{d}$, where

$m =$ mutation rate with probe

$u =$ mutation rate without probe

(cancel seq. error biases)

$d =$ mutation rate in denatured state

(cancel site-dependent biases)

UNIVERSITÉ
PARIS
DESCARTES

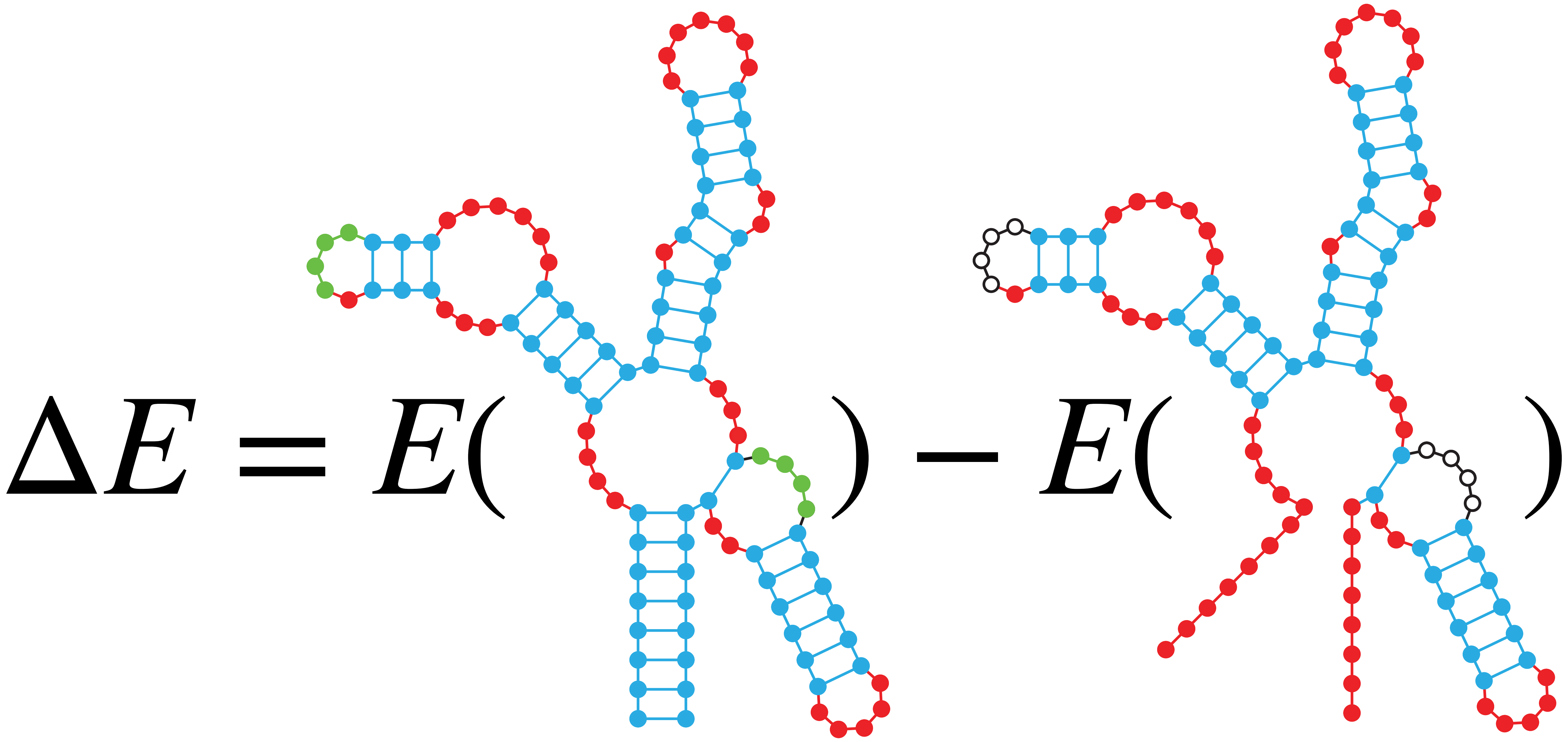
with P. Hardouin, F.-X. Lyonnet, B. Sargueil

Wilkinson, KA et al. "SHAPE: quantitative RNA structure analysis at single nucleotide resolution." Nature Protocols 1.3 (2006): 1610-1616.

Structural flexibility requires moderate binding energies

ViennaRNA

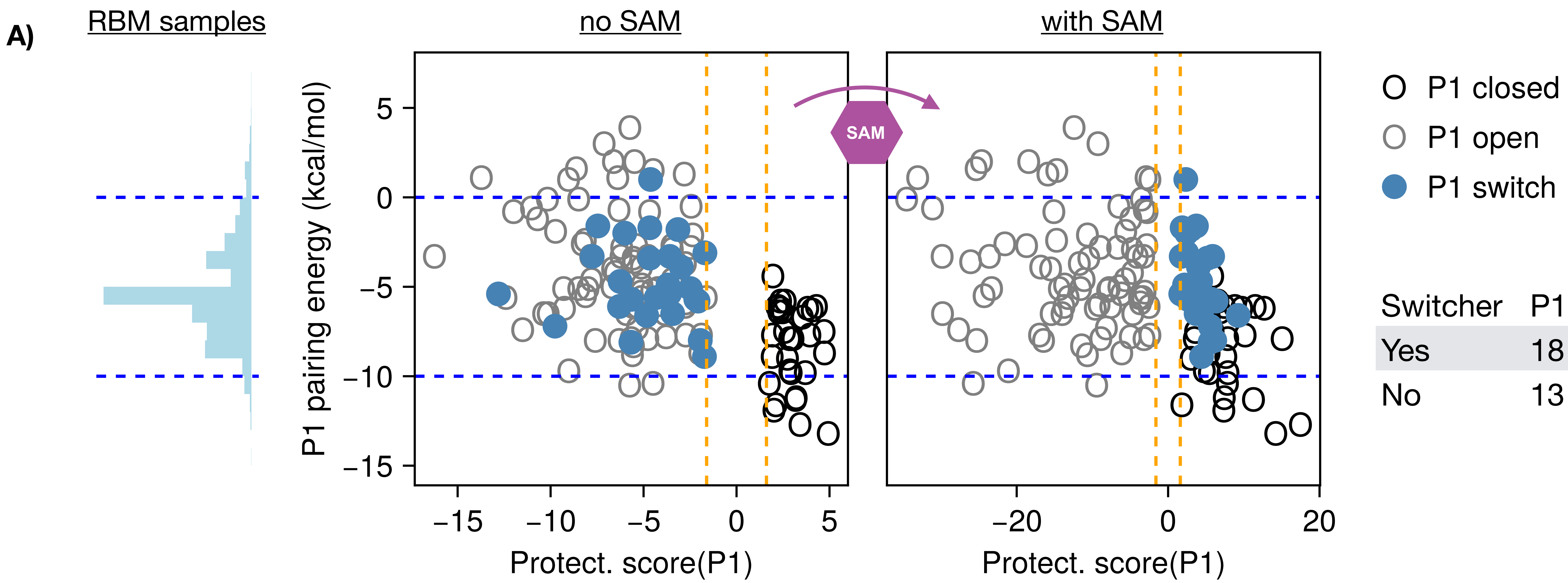
ViennaRNA: Energy calculations of RNA secondary structures, based on the Turner model.



$$\Delta E = E(\quad) - E(\quad)$$

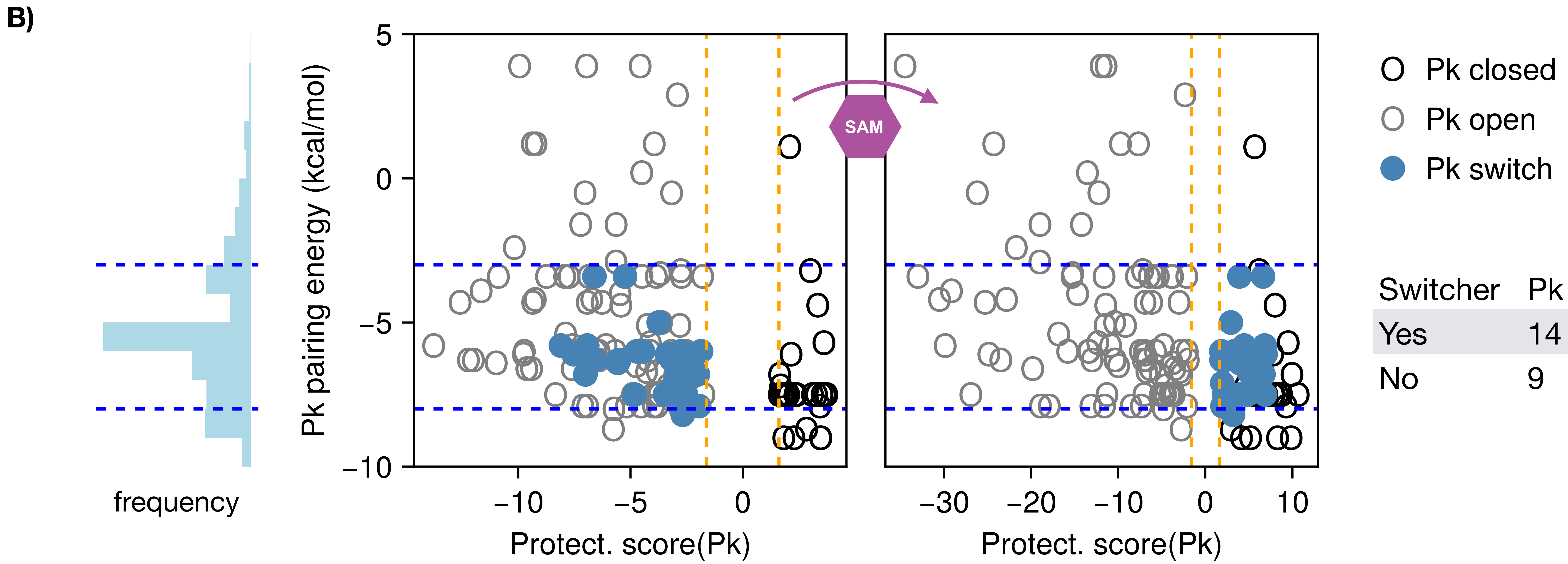
- Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011). https://doi.org/10.1186/1748-7188-6-26
- Mathews, DH., et al. "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure." *PNAS* 101.19 (2004): 7287-7292.

Experimental validation with SHAPE-MaP



Switcher	P1 closed	P1 open	P1 switch
Yes	18	19	25
No	13	56	2

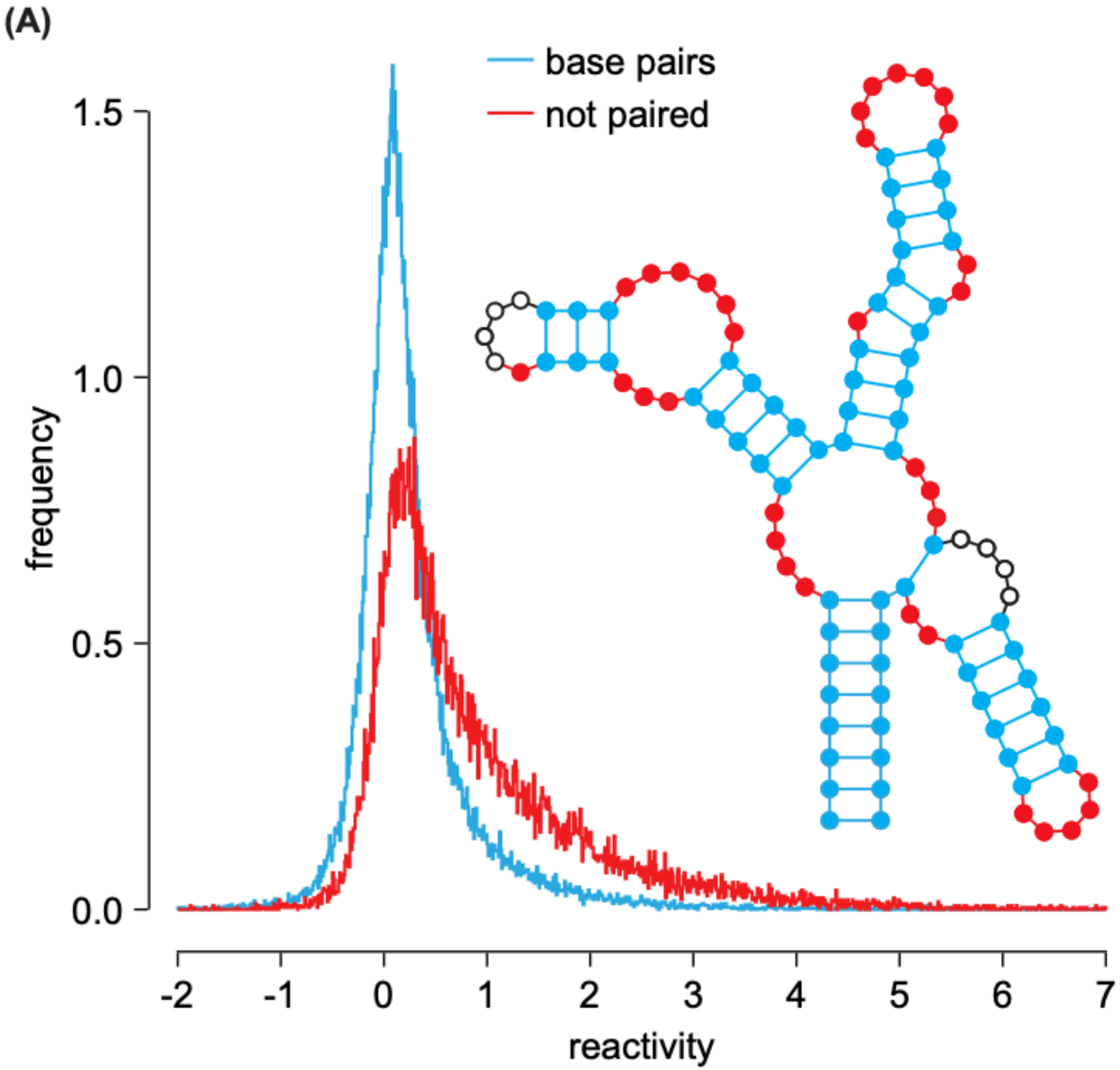
Total P1 conclusive: 133



Switcher	Pk closed	Pk open	Pk switch
Yes	14	12	29
No	9	74	3

Total Pk conclusive: 141

Statistical inference of base-pairs from SHAPE data



SHAPE log-odds score

$$\mathcal{S}_n(i) = \ln \frac{P_{ni}(\tilde{r}_{ni}\,|\,\text{bp})}{P_{ni}(\tilde{r}_{ni}\,|\,\text{np})}$$

$$= \ln \frac{\int P(r\,|\,\text{bp}) P_{ni}(\tilde{r}_{ni}\,|\,r)dr}{\int P(r\,|\,\text{np}) P_{ni}(\tilde{r}_{ni}\,|\,r)dr}$$

$P_{ni}(\tilde{r}_{ni}\,|\,r)$ prob. of measuring $\tilde{r}_{ni}$ given underlying reactivity r (account for sampling error during sequencing)

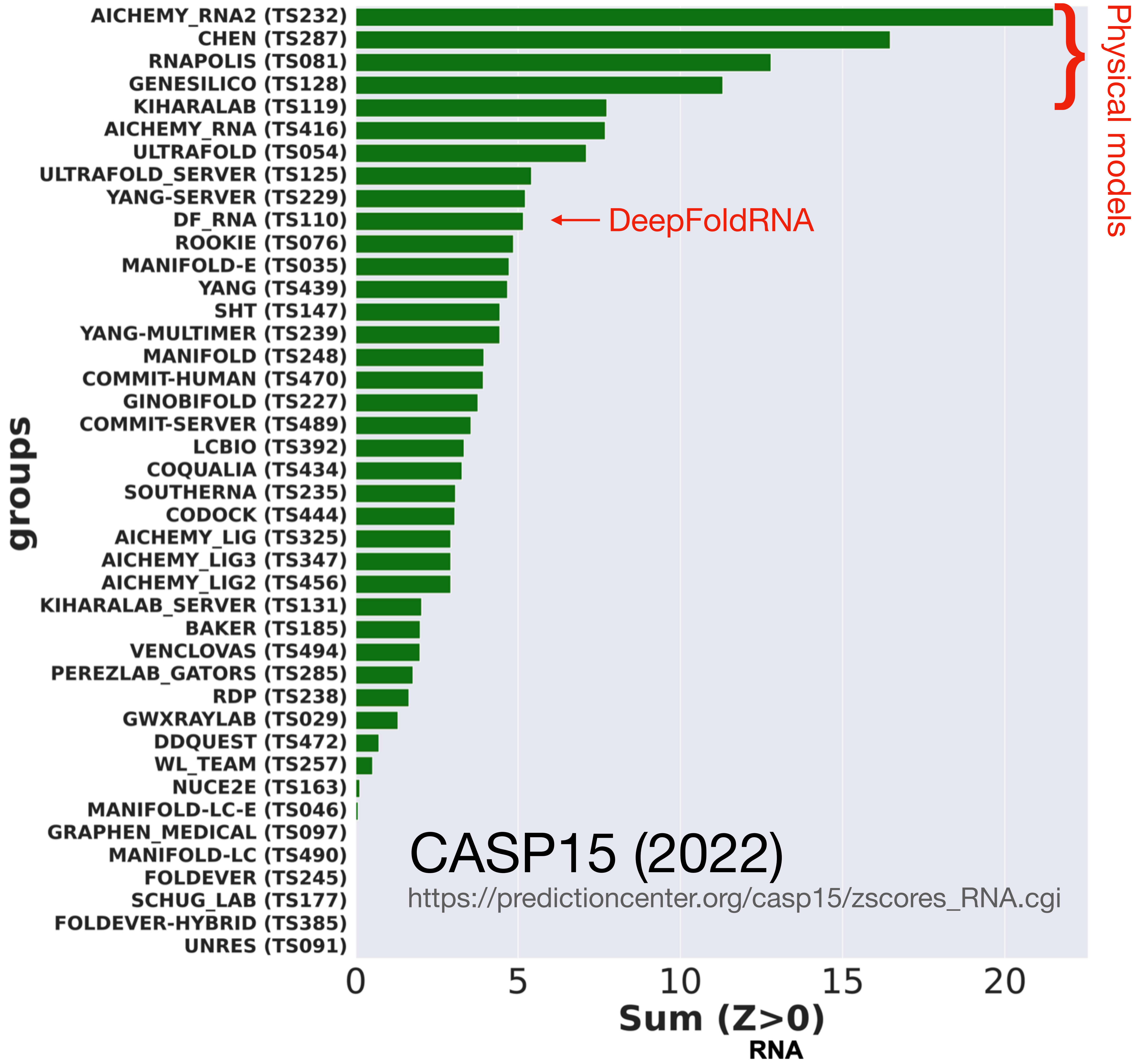
$P(r\,|\,\text{bp})$ prob. of reactivities of base-paired sites

Structure and function of non-coding RNA molecules

Top model performers in CASP15 (2022) RNA structure prediction competition are “physical models”

AlphaFold-derived attempts (e.g., DeepFoldRNA) generalize poorly to new sequences

Qiu, PLoS Comp. Biol (2023)



Physical models

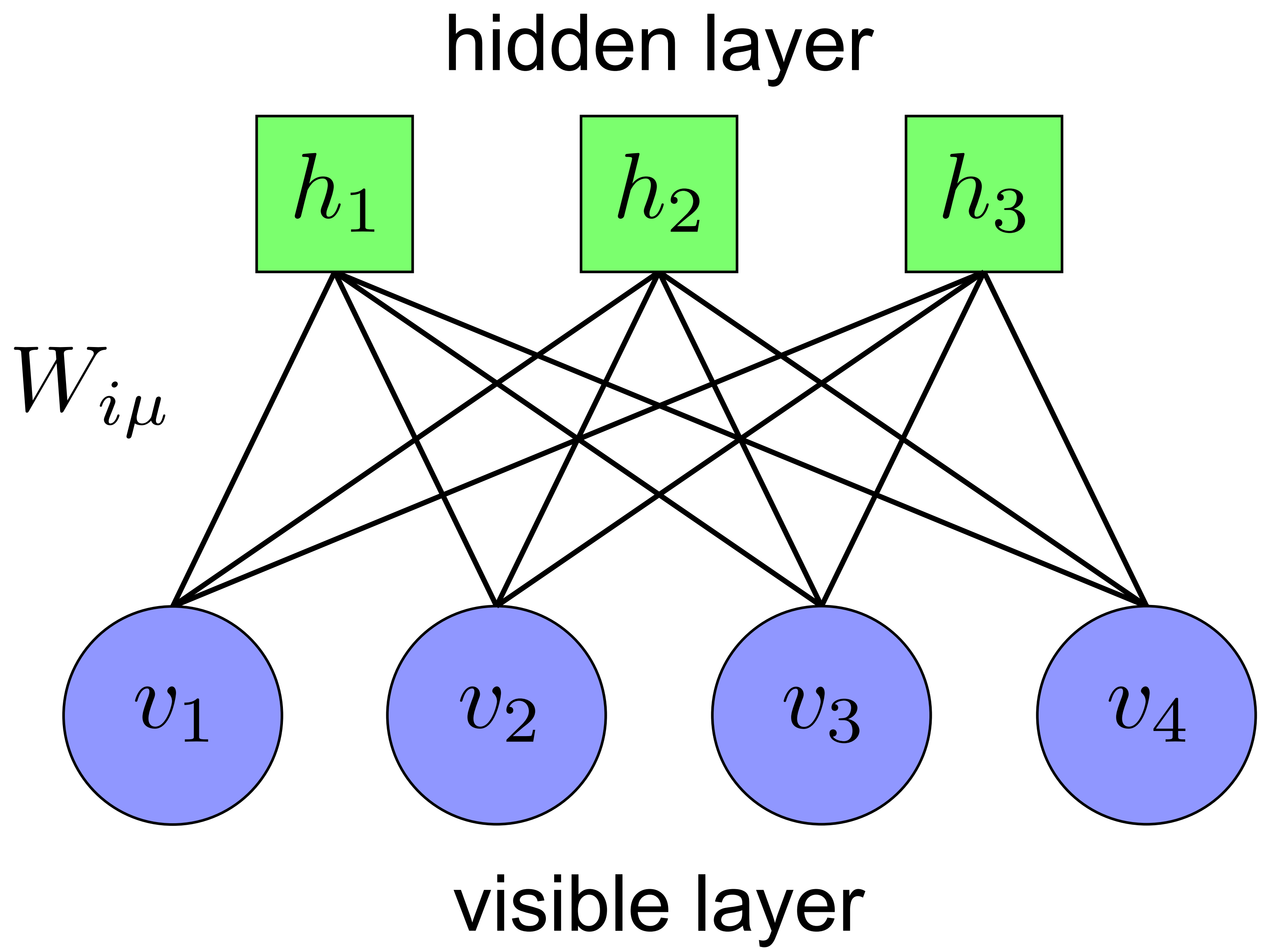
← DeepFoldRNA

CASP15 (2022)

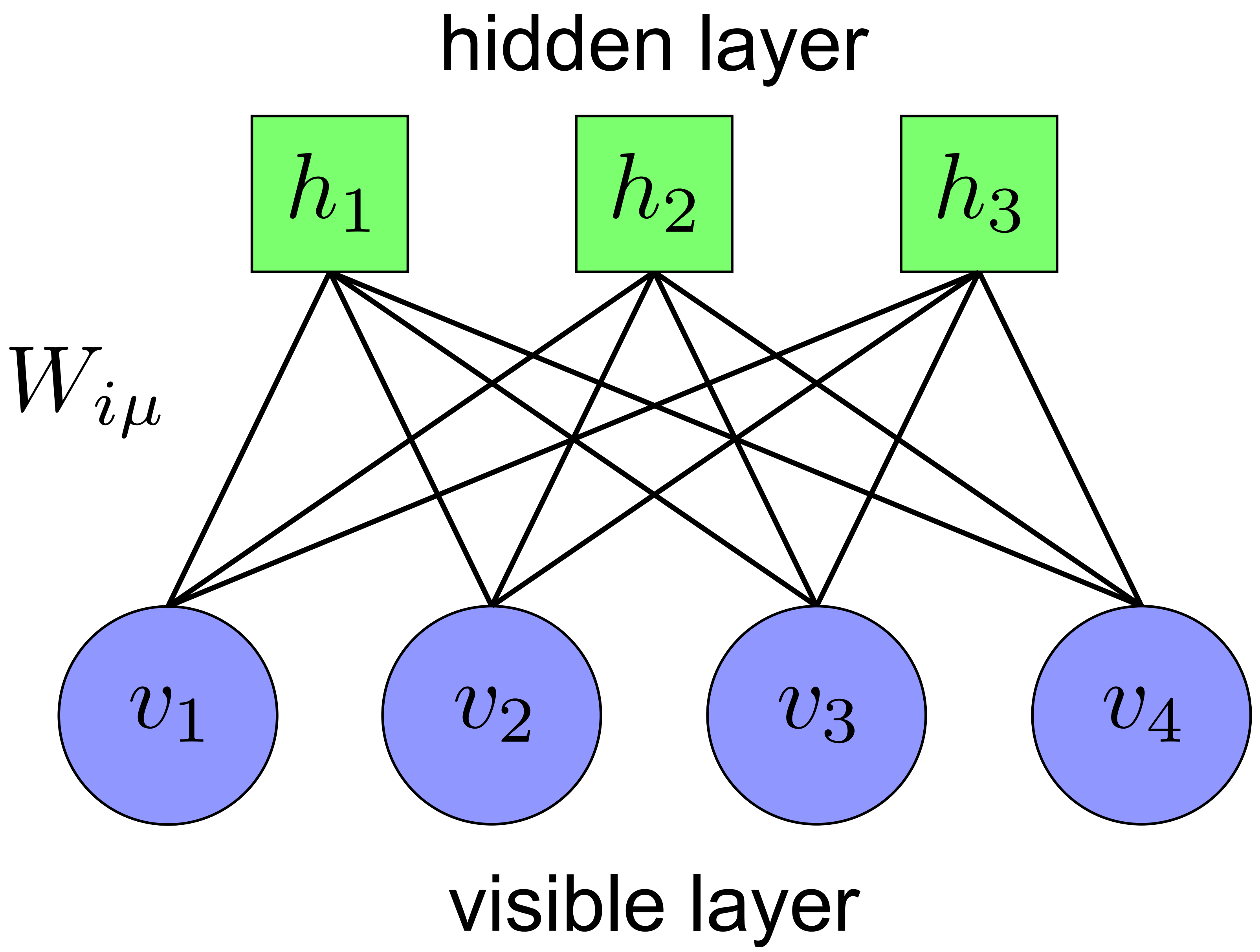
https://predictioncenter.org/casp15/zscores_RNA.cgi

Statistical inference of base-pairs from SHAPE data

Restricted Boltzmann Machines (RBM)



Restricted Boltzmann Machines (RBM)

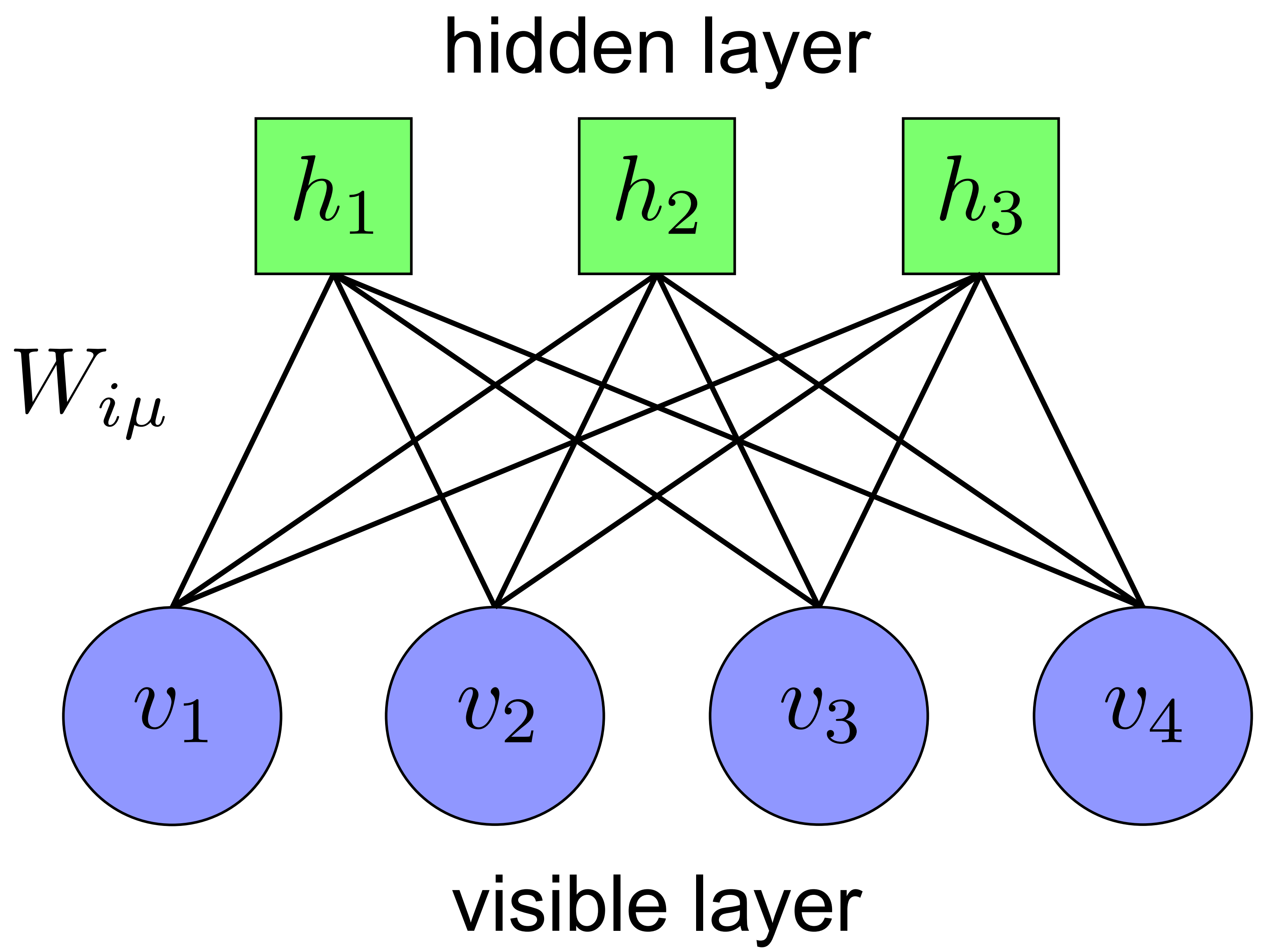


Likelihood: $P_v(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \frac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})}$

where $E_{\text{eff}}(\mathbf{v}) = \sum_i \mathcal{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathcal{U}_\mu(h_\mu)}$

$\mathbf{v}$ = the RNA sequence

Restricted Boltzmann Machines (RBM)



Likelihood: $P_v(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \frac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})}$

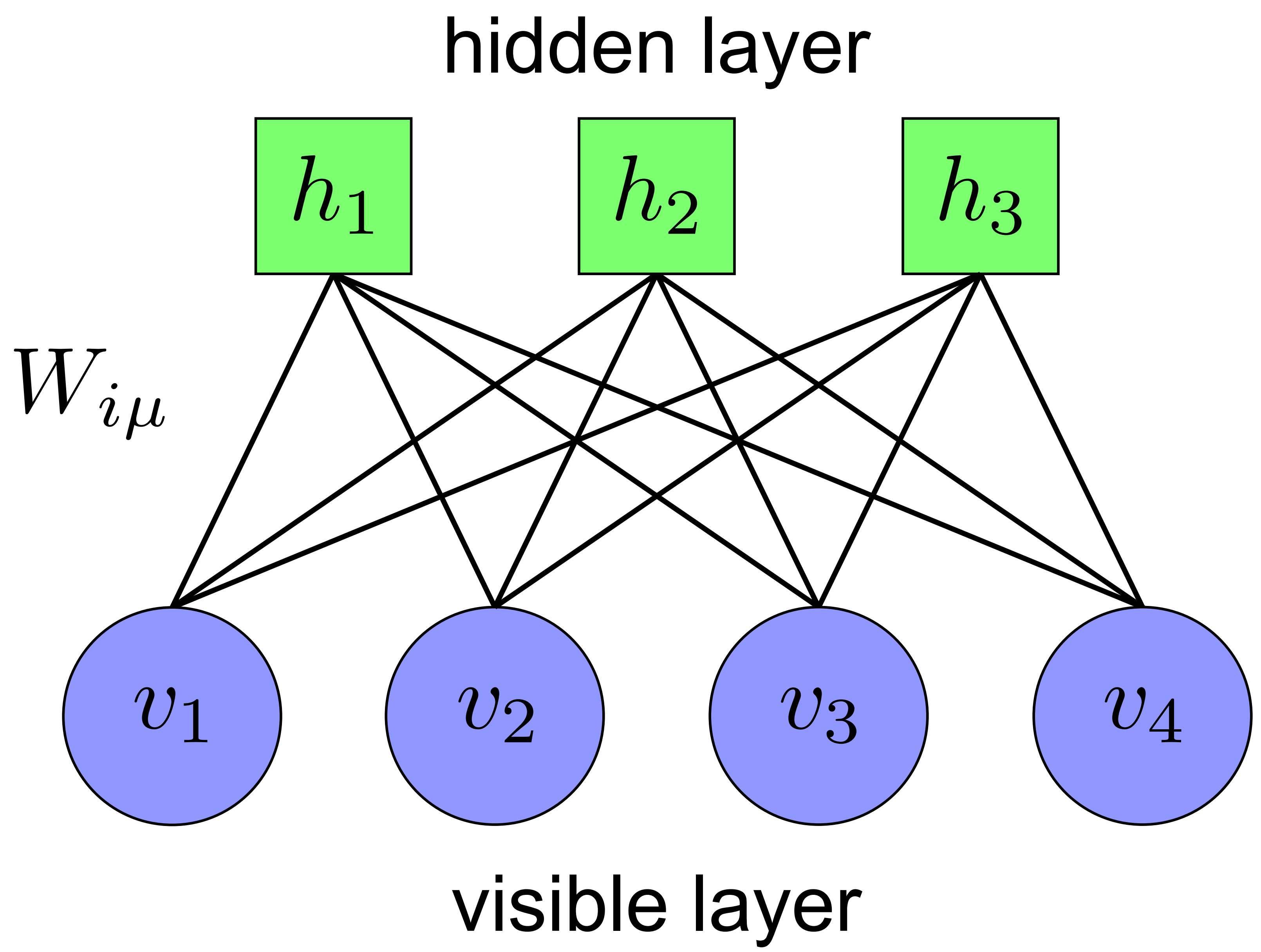
where $E_{\text{eff}}(\mathbf{v}) = \sum_i \mathscr{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathscr{U}_\mu(h_\mu)}$

$\mathbf{v}$ = the RNA sequence

Model is trained, by finding parameters that maximise the likelihood of the data MSA

$$\frac{\partial \mathcal{L}}{\partial \omega} = \underbrace{\left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{MSA}}}_{\text{positive gradient}} - \underbrace{\left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{RBM}}}_{\text{negative gradient}}$$

Restricted Boltzmann Machines (RBM)



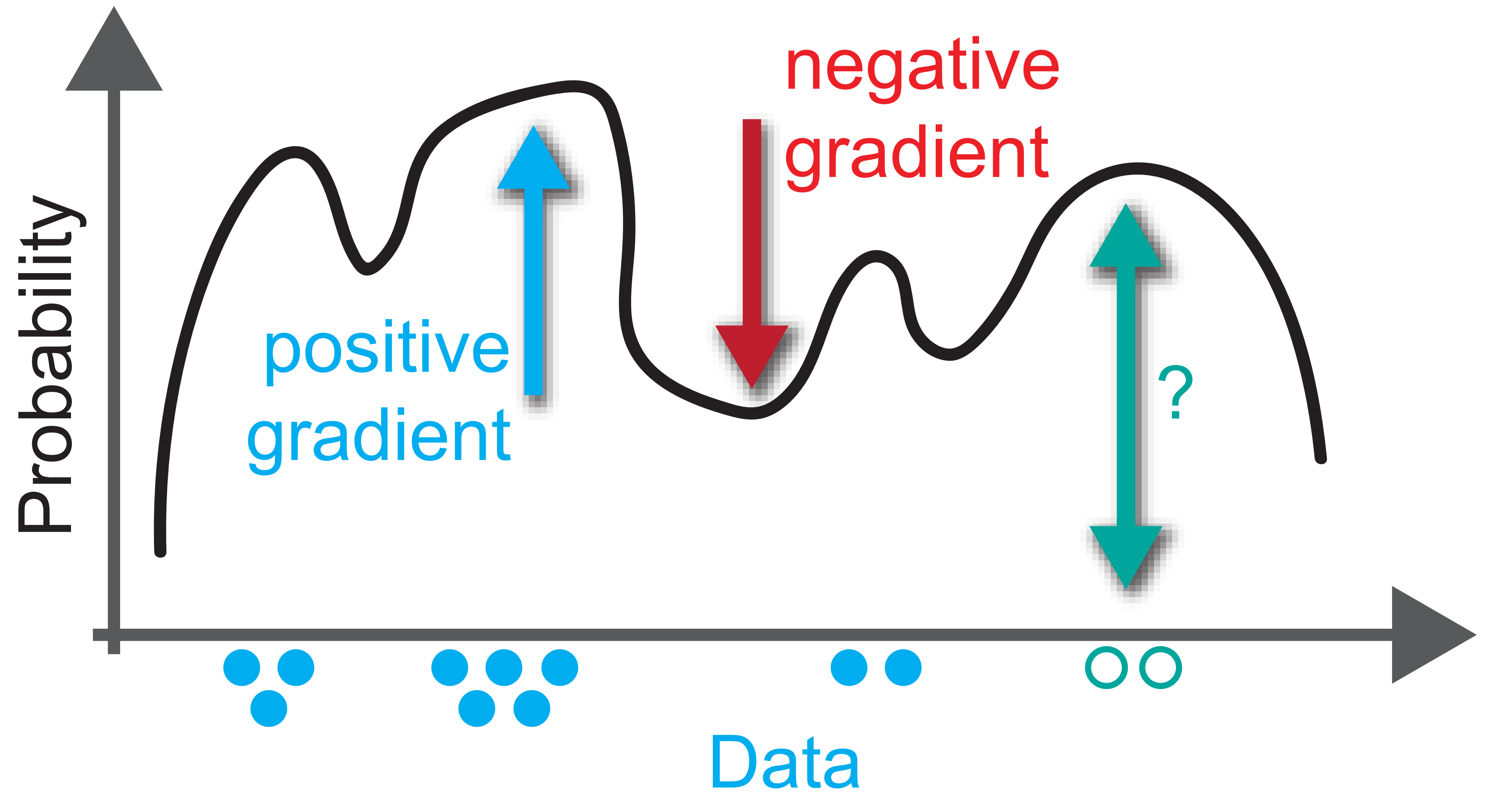
Likelihood: $P_v(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = \frac{1}{Z} e^{-E_{\text{eff}}(\mathbf{v})}$

where $E_{\text{eff}}(\mathbf{v}) = \sum_i \mathcal{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathcal{U}_\mu(h_\mu)}$

$\mathbf{v}$ = the RNA sequence

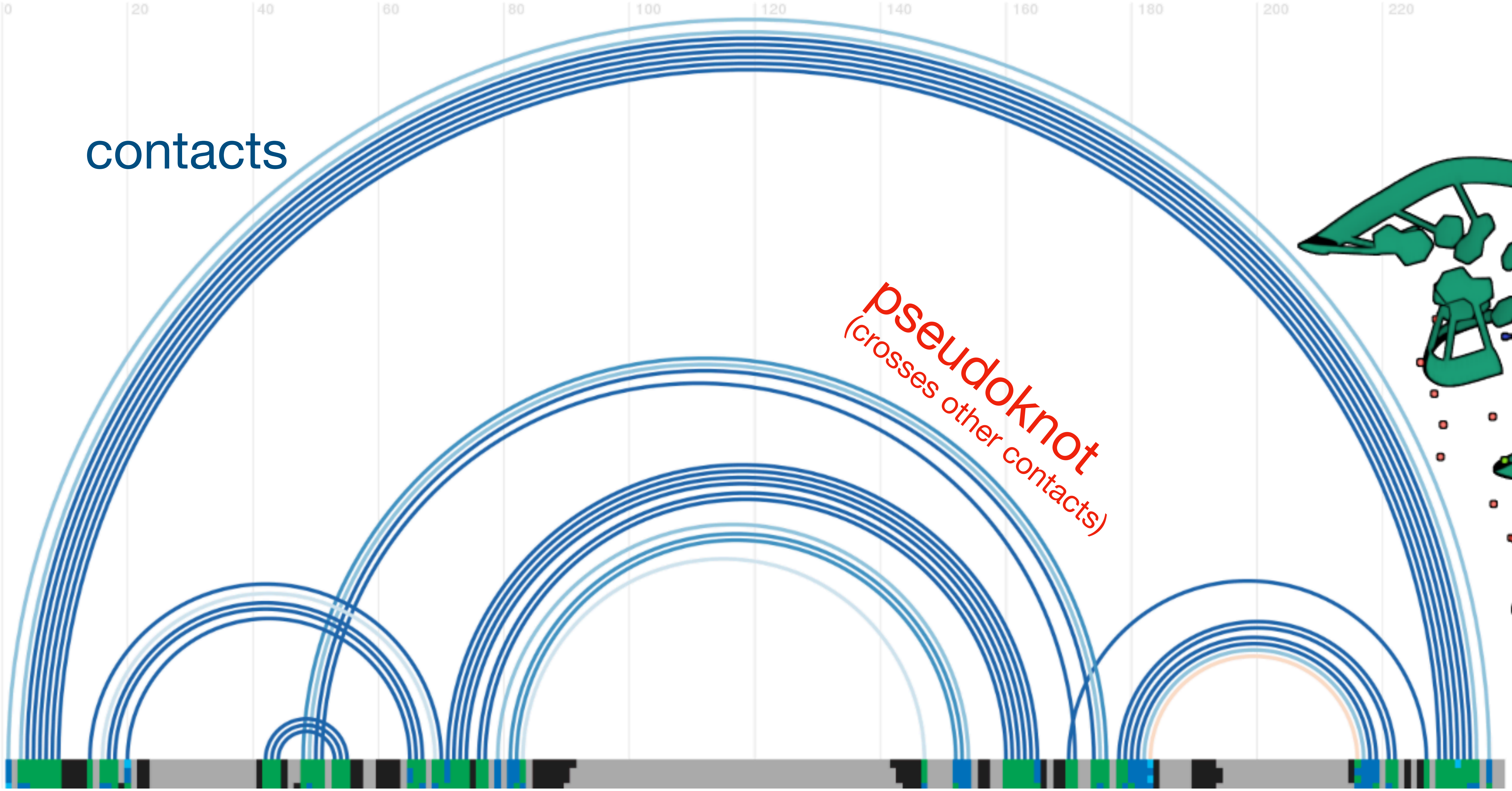
Model is trained, by finding parameters that maximise the likelihood of the data MSA

$$\frac{\partial \mathcal{L}}{\partial \omega} = \underbrace{\left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{MSA}}}_{\text{positive gradient}} - \underbrace{\left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{RBM}}}_{\text{negative gradient}}$$



SAM riboswitches

Arc diagram



contacts

pseudoknot
(crosses other contacts)

linear sequence

3D structure

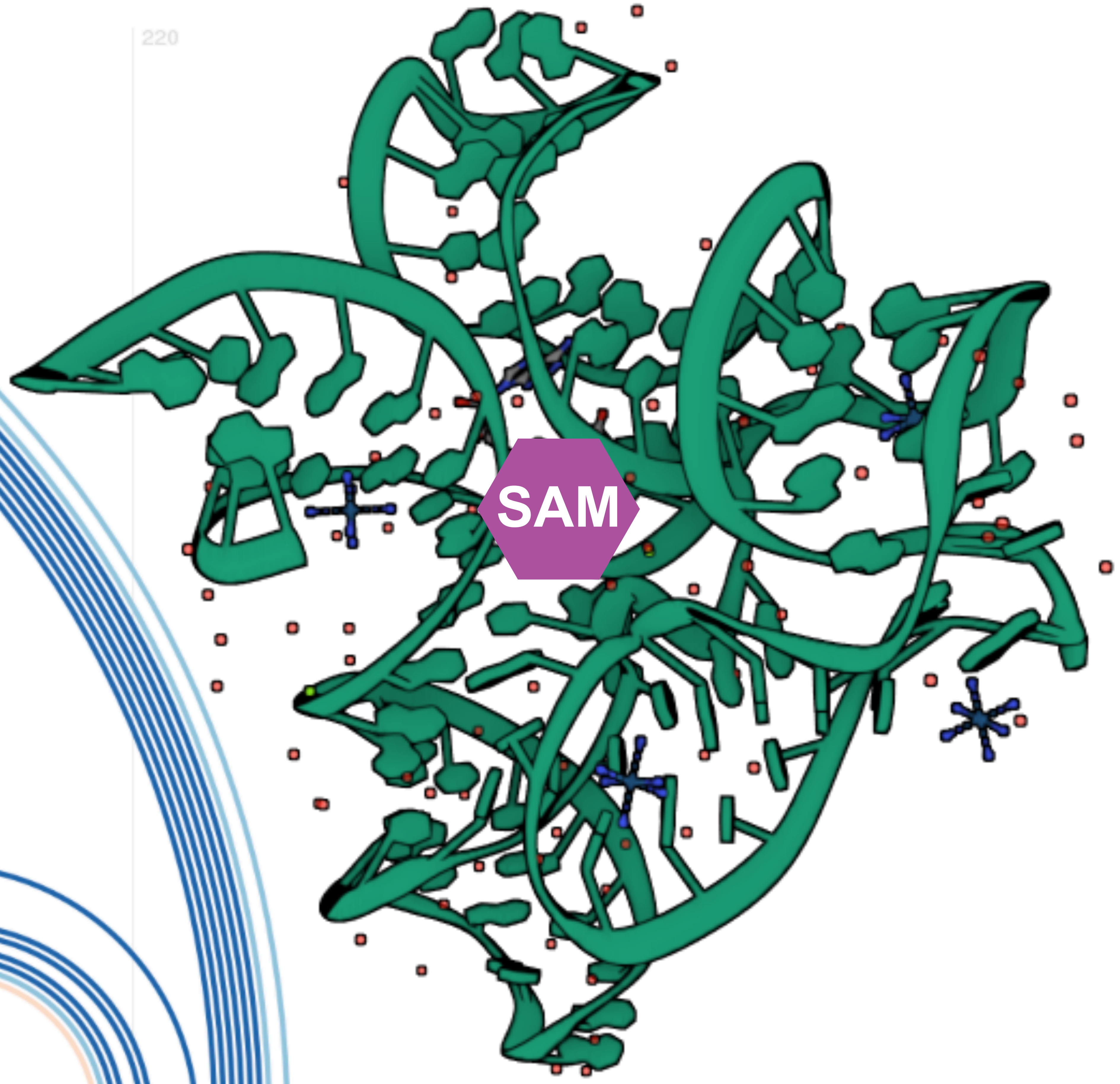


Figure from Rfam: https://rfam.org/family/RF00162#tabview=tab3

PCA of RF00162 MSA