RNA structure determination via chemical probing

Christine Heitsch, Georgia Tech Math (+ SCMB Director)

Considering the next step in your career? Or know someone who is? Let's chat!

BURROUGHS WELLCOME FUND

NIGMS

NSF

SIMONS FOUNDATION

A long time ago (2004), in a journal far, far away (PNAS) . . .

Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure

David H. Mathews[†], Matthew D. Disney[†‡], Jessica L. Childs[†‡], Susan J. Schroeder[‡], Michael Zuker[§], and Douglas H. Turner[†‡¶]

"Chemical modification is a technique that reveals solvent accessible nucleotides . . . and an algorithm allowing constraints from such chemical modification has not been reported."

"In this study, a dynamic programming algorithm for prediction of RNA secondary structure has been revised to use experimentally determined chemical modification constraints. These constraints dramatically improve the accuracy of structure prediction when free energy minimization alone predicts $<40\%$ of known base pairs."

Chemical probing: our only hope?

1. ~~Reactive if accessible.~~ If reactive, then accessible!
2. ~~Accessible if unpaired~~[1]. If accessible, then unpaired[2]!
3. Ergo, forbid[3] highly[4] reactive positions from pairing.

Voila! Minimum free energy prediction accuracy increases (significantly) under chemical modification constraints[5].

[1] or "in A-U or G-C pairs at the ends of helices, G-U pairs anywhere, or adjacent to G-U pairs"
[2] almost always
[3] "conformations inconsistent with the data", i.e.
[4] but maybe not moderately and definitely not weakly (as if adjectives were well-defined)
[5] which were both "hard" (must be satisfied) and "negative" (always prohibit, never enforce)

“Never tell me the odds!”

Feeding an optimization algorithm high quality information helps.

A lot.

But. . .

Can the process be made more automatic?

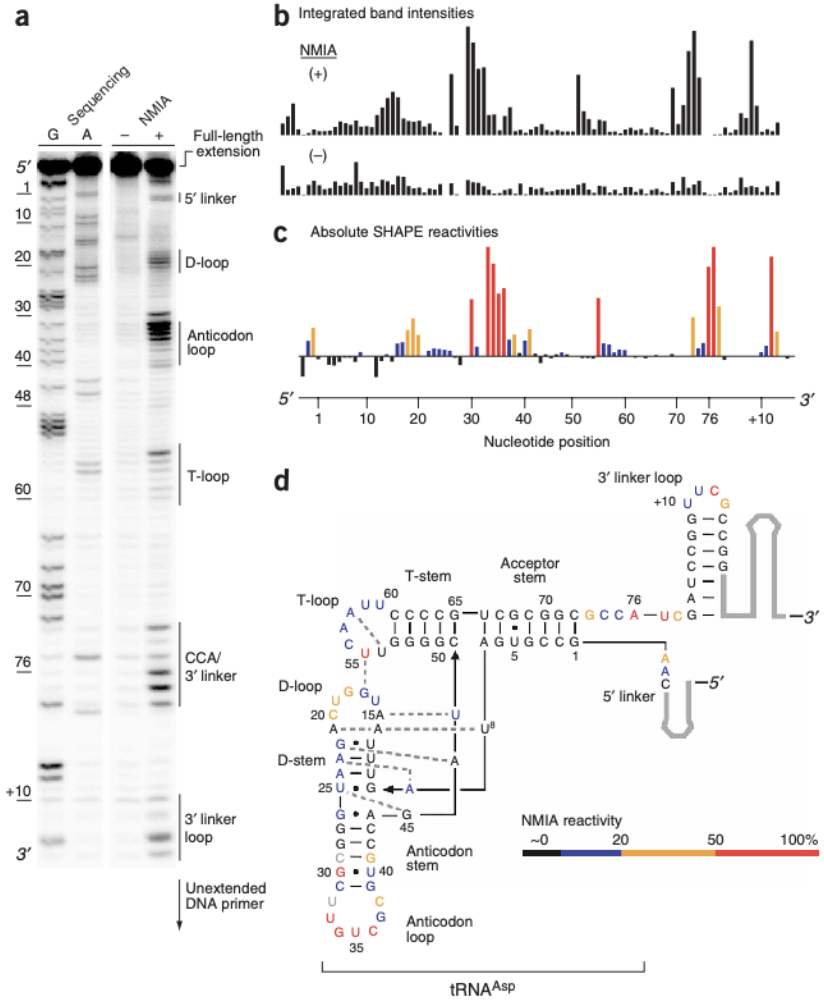
Systematic?

Comprehensive?

Etc?

May the force be with you

Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution, KA Wilkinson, EJ Merino & KM Weeks, Nature Protoc 2006.



SHAPE strategy:

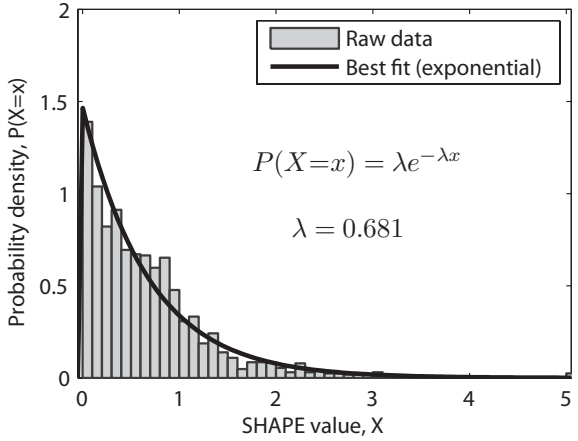
1. Chemical modifications
2. Read-out changes
3. Infer structure

Predict secondary structure using data as "soft" constraints (i.e. "restraints") in new reward/penalty function under NNTM optimization.

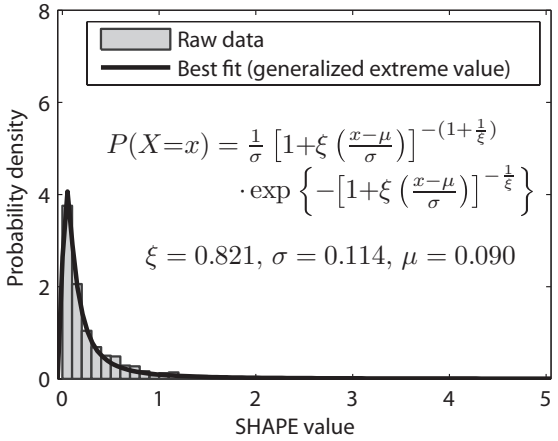
Understanding SHAPE-directed MFE predictions

Sükösd, Swenson, Kjems, & Heitsch, Nucleic Acids Res, 2013.

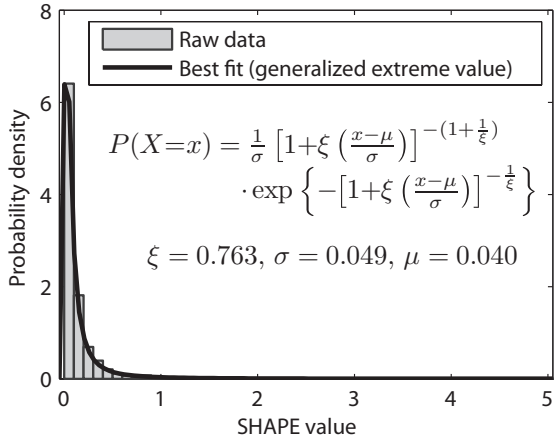
A probabilistic model for SHAPE data:



Unpaired (4x mag)



Helix-end

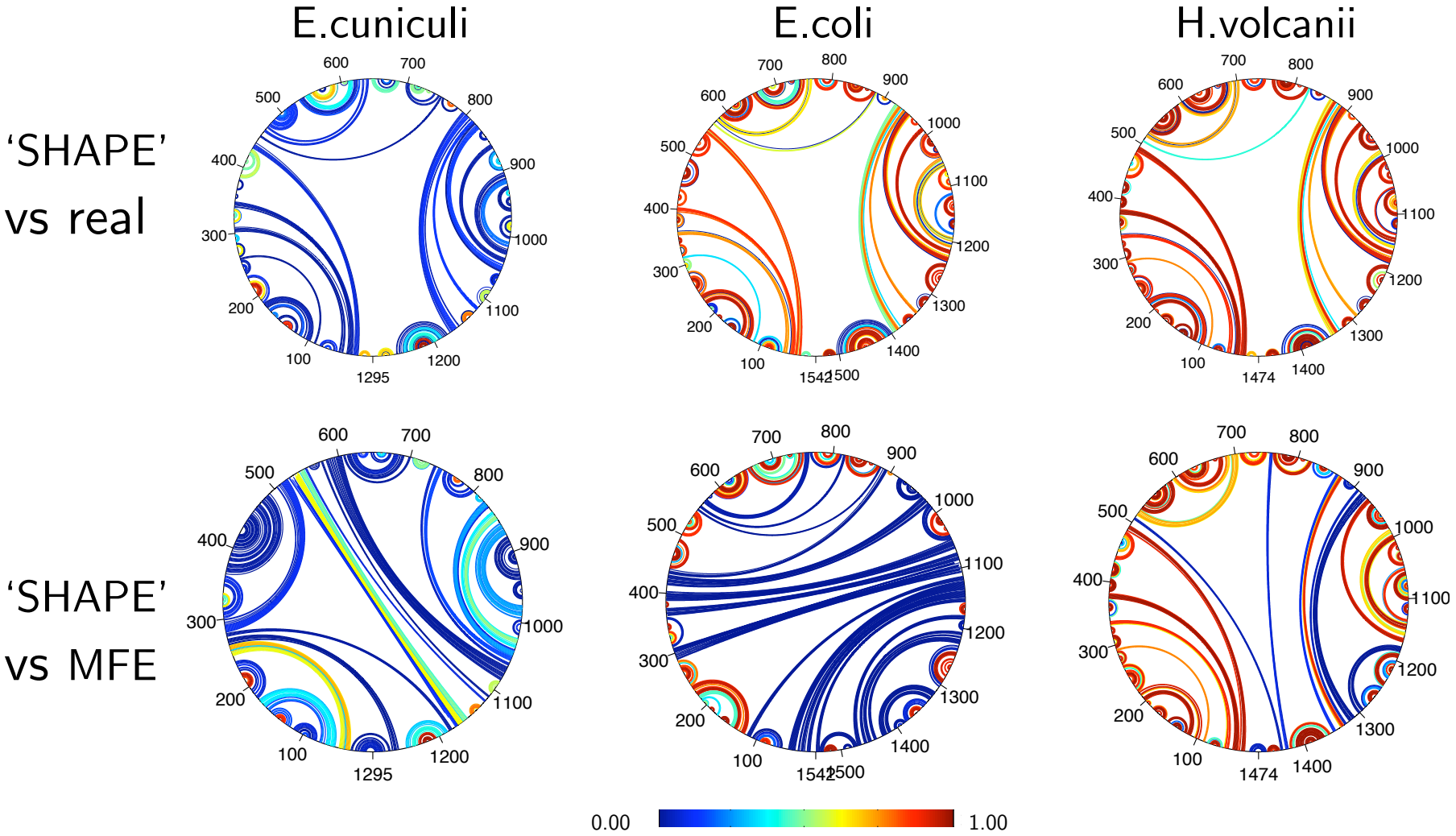


Stacked

Raw data from Deigan *et al*, PNAS 2009; K Weeks, personal communication.

Use model to get **statistics** on accuracy improvement over 1000 trials per sequence for a diverse set of 16S ribosomal RNA.

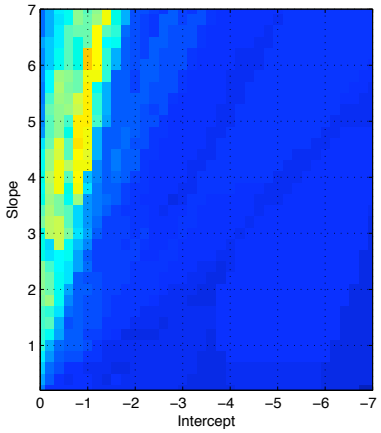
Predictions can vary (a lot!) in accuracy, but generally preserve accurate MFE pairs

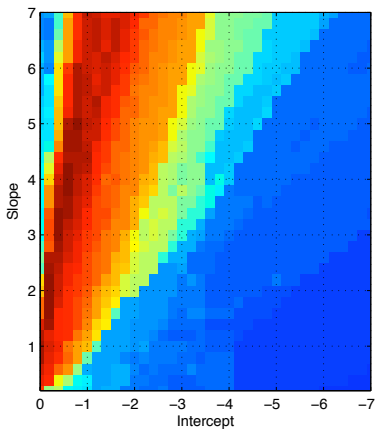


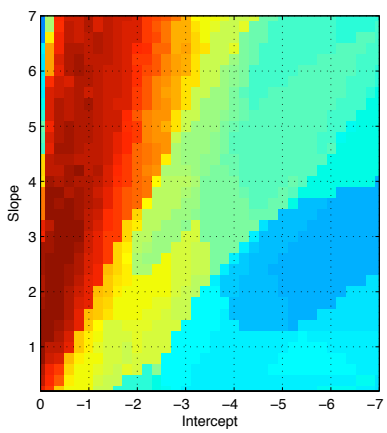
How "directable" are MFE predictions?
It depends on the sequence!

E.cuniculi | E.coli | H.volcanii

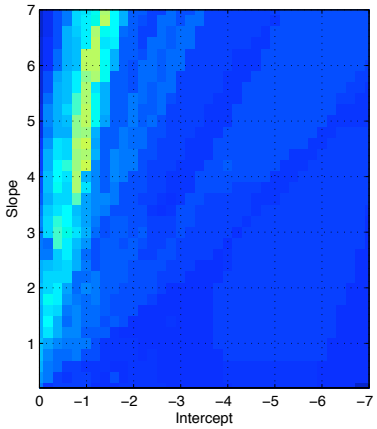
Precision $\left(\frac{tp}{tp+fp}\right)$

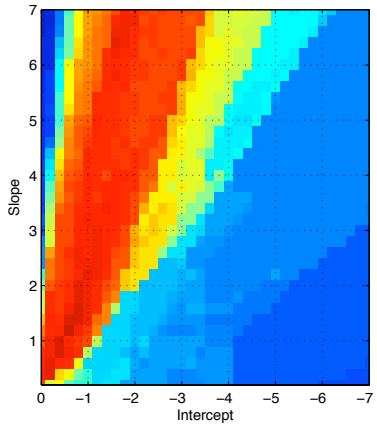


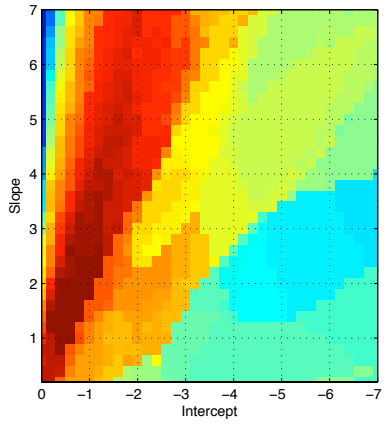




Recall $\left(\frac{tp}{tp+fn}\right)$





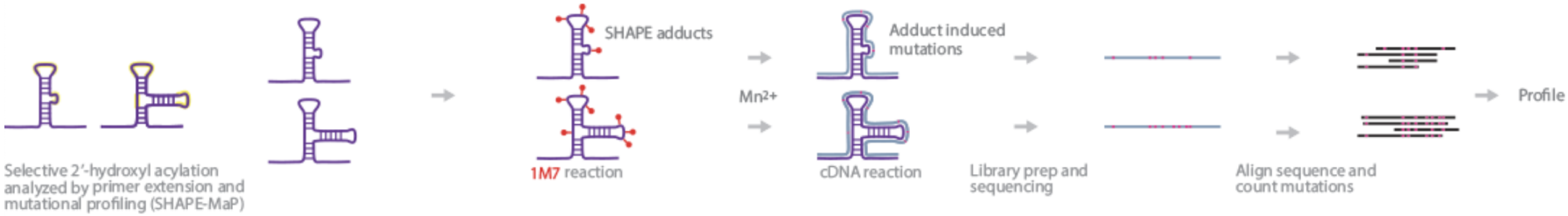


Parameterize $\Delta G_{\mathsf{S}}(i) = m \ln(\mathsf{S}(i) + 1) + b$, accuracy: 0  1.

State-of-the-art? (According to Illumina. . .)

SHAPE-Seq (Lucks et al, PNAS 2011):

SHAPE-MaP (Siegfied et al, Nat Methods 2014):



https://emea.illumina.com/science/sequencing-method-explorer/kits-and-arrays/shape-{seq,map}.html

According to J White Bear's talk: SHAPE-nanopore/nanoSHAPE (Leger et al, Nat Commun 2021; Stephenson et al, Cell Genom 2022)

DMS-MaP: different chemistry, same story

1. New procedure for introducing chemical modifications

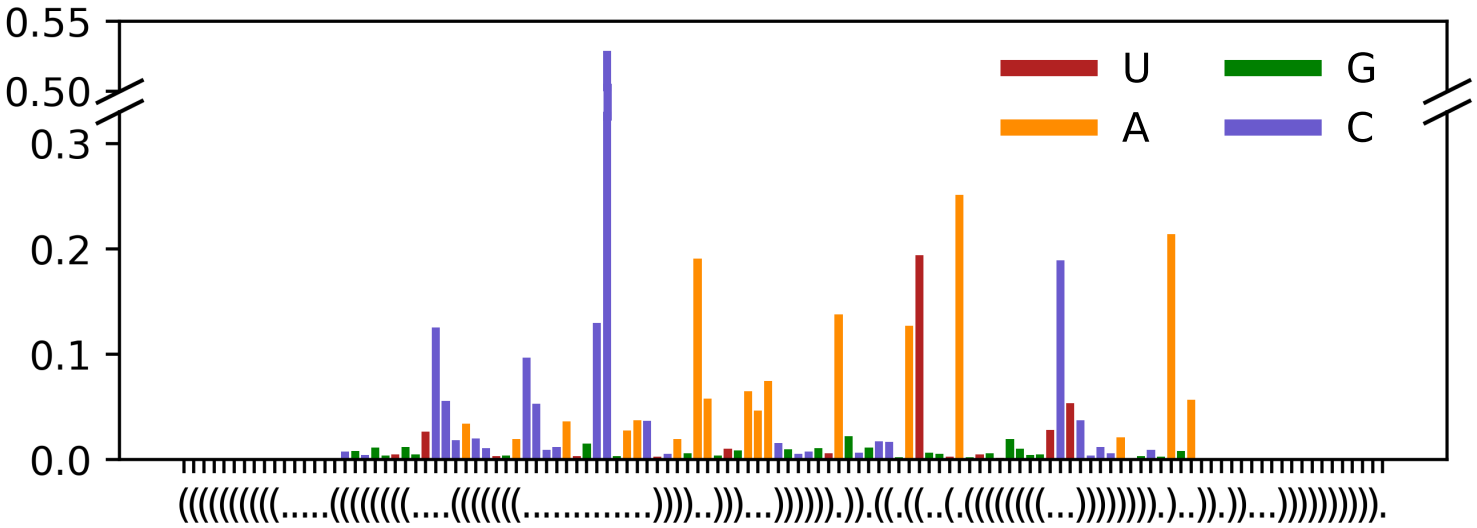
2. High-throughput sequencing read-out of alterations

3. "Black box" bioinformatics to infer structure

Relative difference of maximal helices (ReDMaxH)

with Alfie Brownless, Afaf Saaidi & Alain Laederach

Goal: Identify base pairing signal in DMS-MaP data.

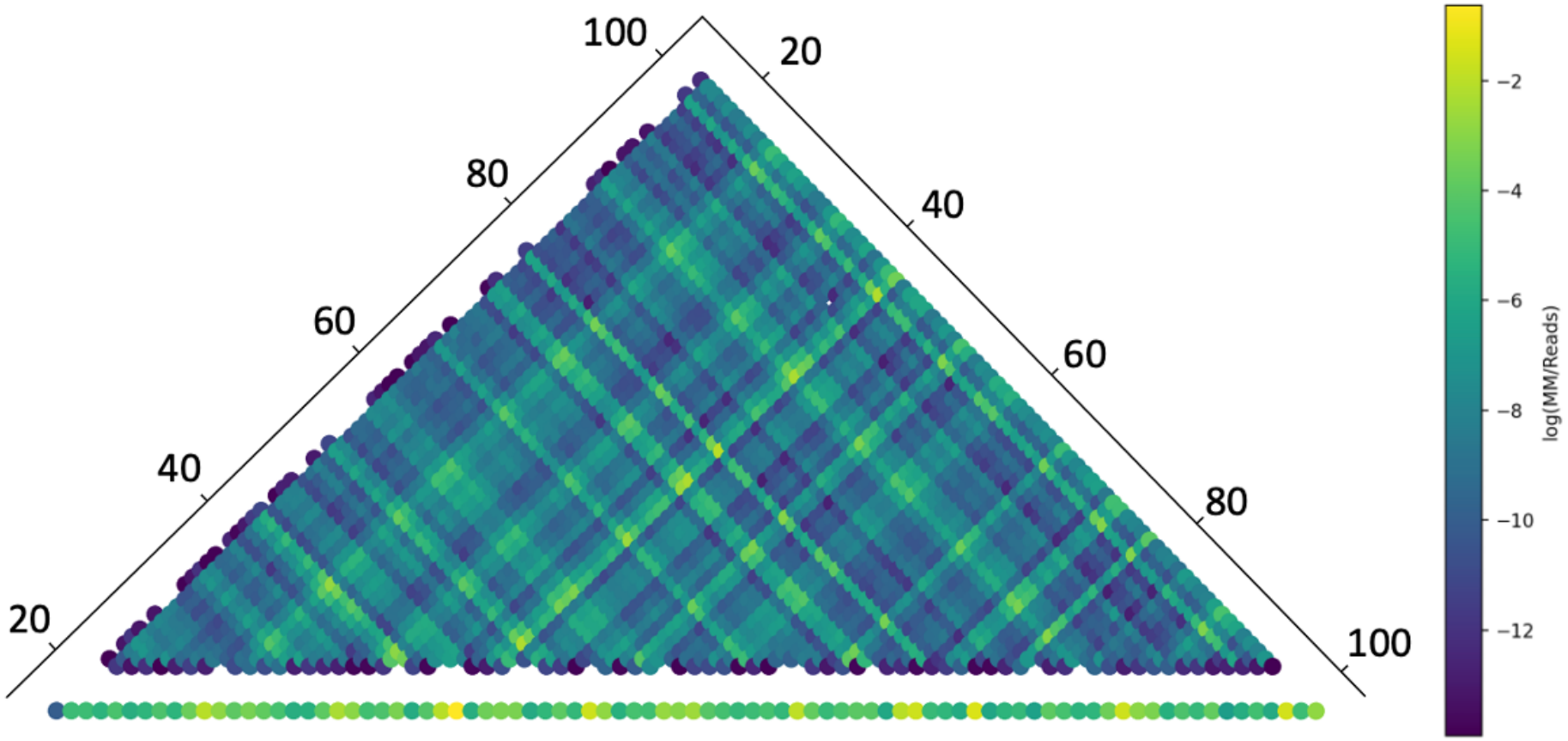


E. coli 5S rRNA (Mustoe et al, PNAS 2019)

Theory: Signal in co-occurring modifications.

Practice: Complex dependencies.

Binary mutation distribution



Mutation counts/reads vary over multiple orders of magnitude.

Relative difference computation

Input: Matrices M and N where $m_{ij} = \#$ of reads where both i and j are mutated and $n_{ij} =$ total $\#$ of reads covering both.

Estimate binary/joint and unary/independent probabilities:

$$\hat{p}_{ij} = \frac{m_{ij}}{n_{ij}} \text{ and } \hat{p}_k = \frac{m_{kk}}{n_{kk}} \text{ for } k = i, j.$$

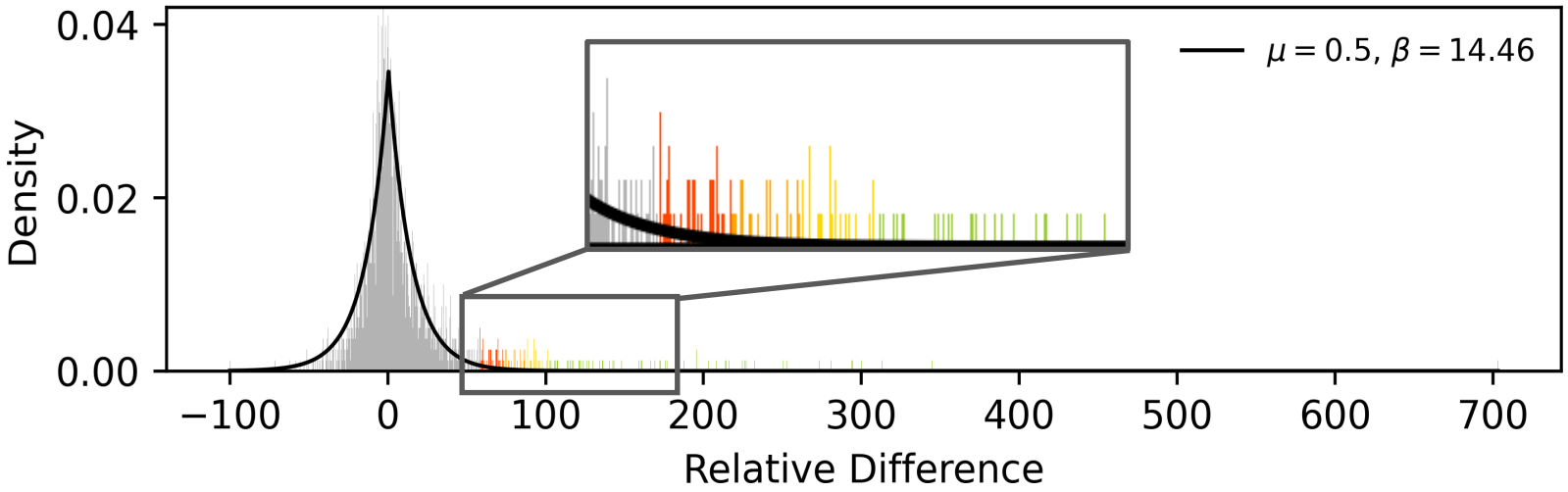
Define:

$$Rd_{ij} = \frac{\hat{p}_{ij} - \hat{p}_i \hat{p}_j}{\hat{p}_i \hat{p}_j} = \frac{\hat{p}_{ij}}{\hat{p}_i \hat{p}_j} - 1$$

Actually work with percentage ($\times 100$ value)

Rd follow Laplace's first law of errors (1774)

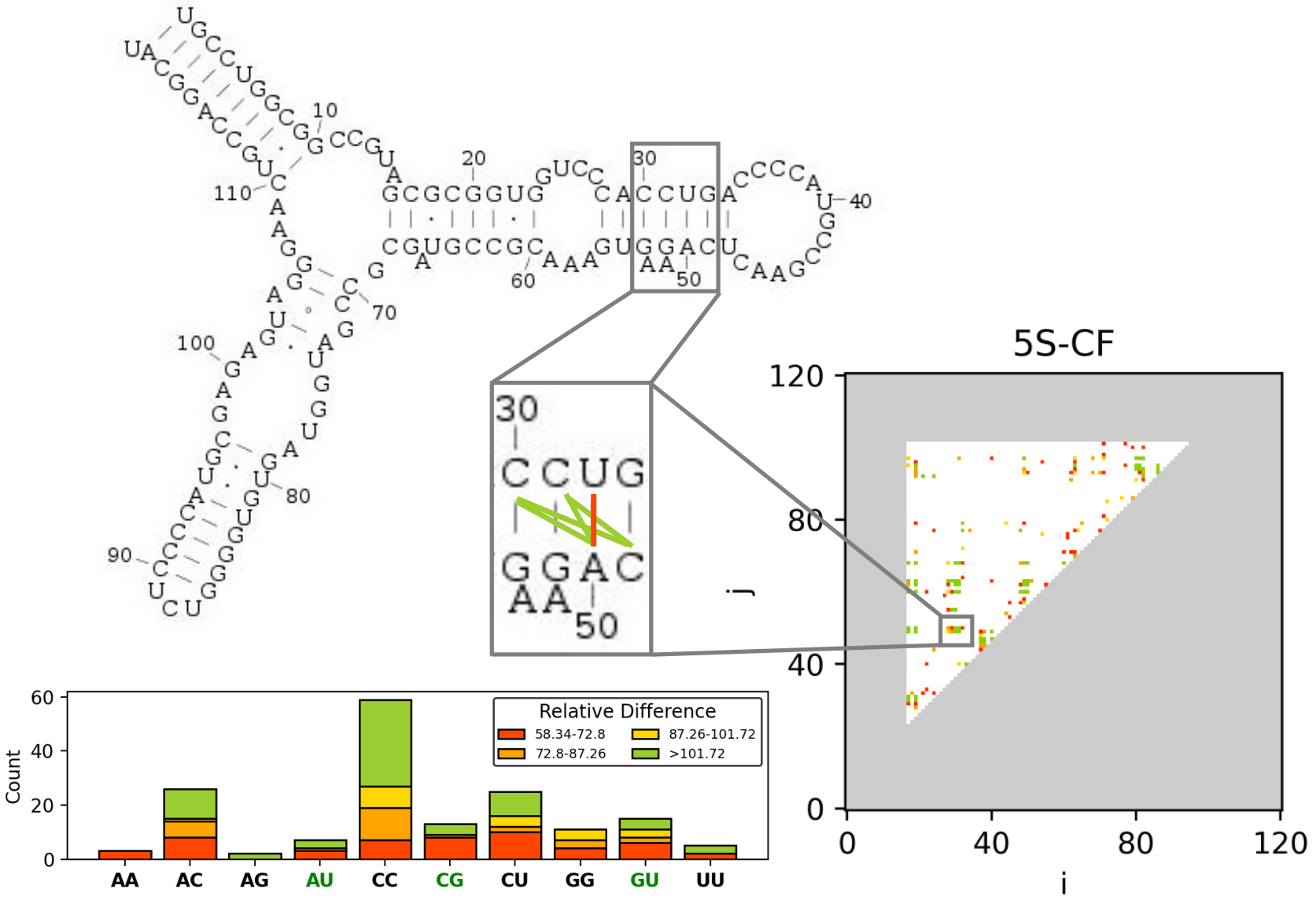
$$f(x) = \frac{1}{2\beta} \exp -\frac{|x-\mu|}{\beta} \text{ with median } \mu \text{ and spread } \beta$$



Not as nice analytically as second law (1778, aka Gaussian), but often a much better error model[6] (Wilson, J Am Stat Assoc 1923).

[6] "No phenomenon is better known perhaps, as a plain matter of fact, than that the frequencies which we actually meet in everyday work in economics, in biometrics, or in vital statistics, very frequently fail to conform at all closely to the so-called normal distribution."

Strongest signal AROUND base pairs



Tuple census under Rd classification over Moore neighborhoods for maximal helices

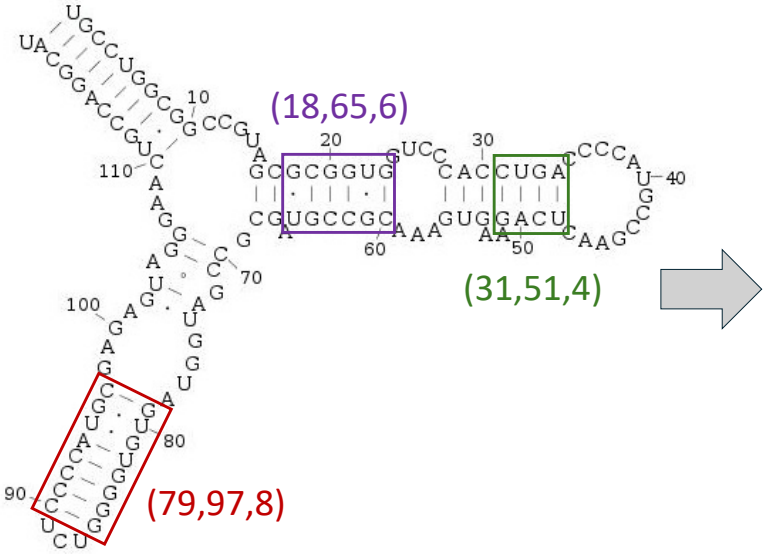
A tuple $[i, j]$ is 'high' if $Rd_{ij} - \mu > \lambda\beta$ and 'extreme' if $> (\lambda + 3)\beta$. Default $\lambda = 6$, but can change to alter precision/recall trade-offs.

The Moore neighborhood of base pair (i, j) is the set of tuples $\{[i + a, j + b] \mid a, b \in \{-1, 0, 1\}\}$.

A helix (i, j, k) is maximal if $(i + m, j - m)$ are canonical for $0 \leq m < k$, neither $(i - 1, j + 1)$ nor $(i + k, j - k)$ are, and $j - i - 2k \geq 2$.

For each maximal helix of length ≥ 4, count the number of extreme and of high tuples over its Moore neighborhood.

Build structure(s) greedily but without conflict



Rate Nucleotide Pairings

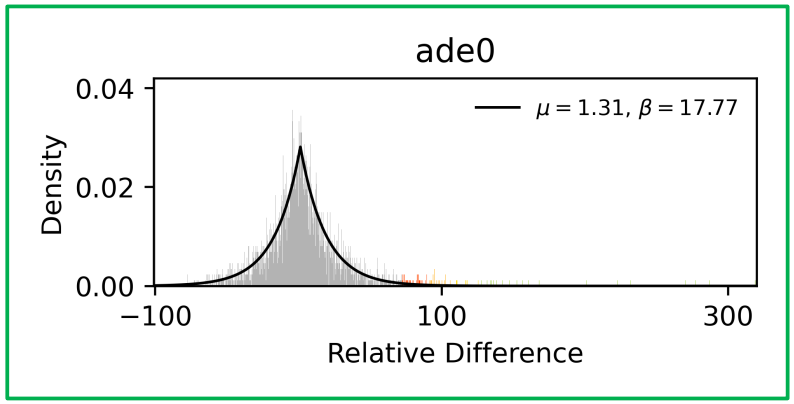
Nucleotide pair	Relative Difference	Rating
(81,94)	703.2	ET
(17,68)	535.9	ET
(31,63)	344.5	ET
⋮	⋮	⋮
(30,50)	123.8	ET
⋮	⋮	⋮
(82,93)	84.8	HT
⋮	⋮	⋮

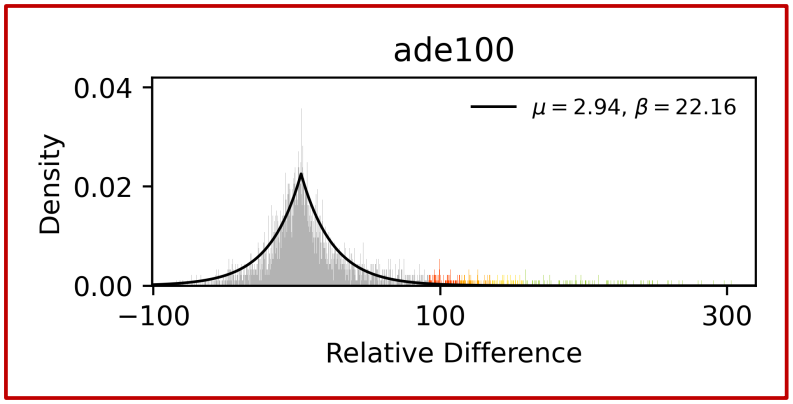
Rank Maximal Helices

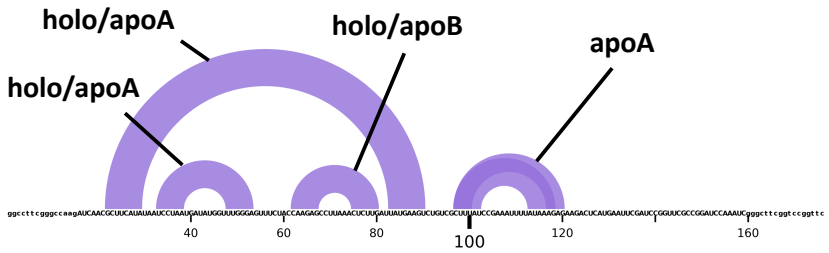
Helix	(ET,HT)	Accepted?
(79,97,8)	(8,2)	✓
(80,98,4)	(5,0)	✓
(31,51,4)	(3,1)	✓
(18,65,6)	(2,1)	✓
(20,31,4)	(2,1)	✗
⋮	⋮	⋮
(1,119,10)	(0,0)	✗
⋮	⋮	⋮

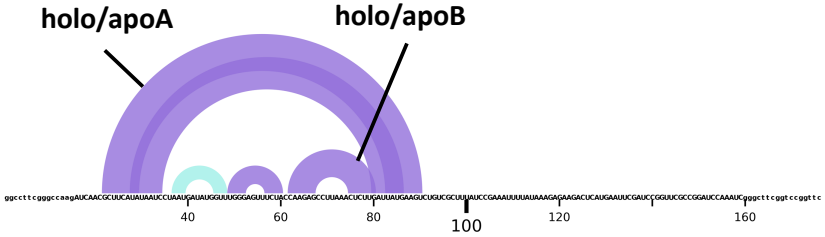
Generate Summary Structures

Many advantages to ReDMaxH approach





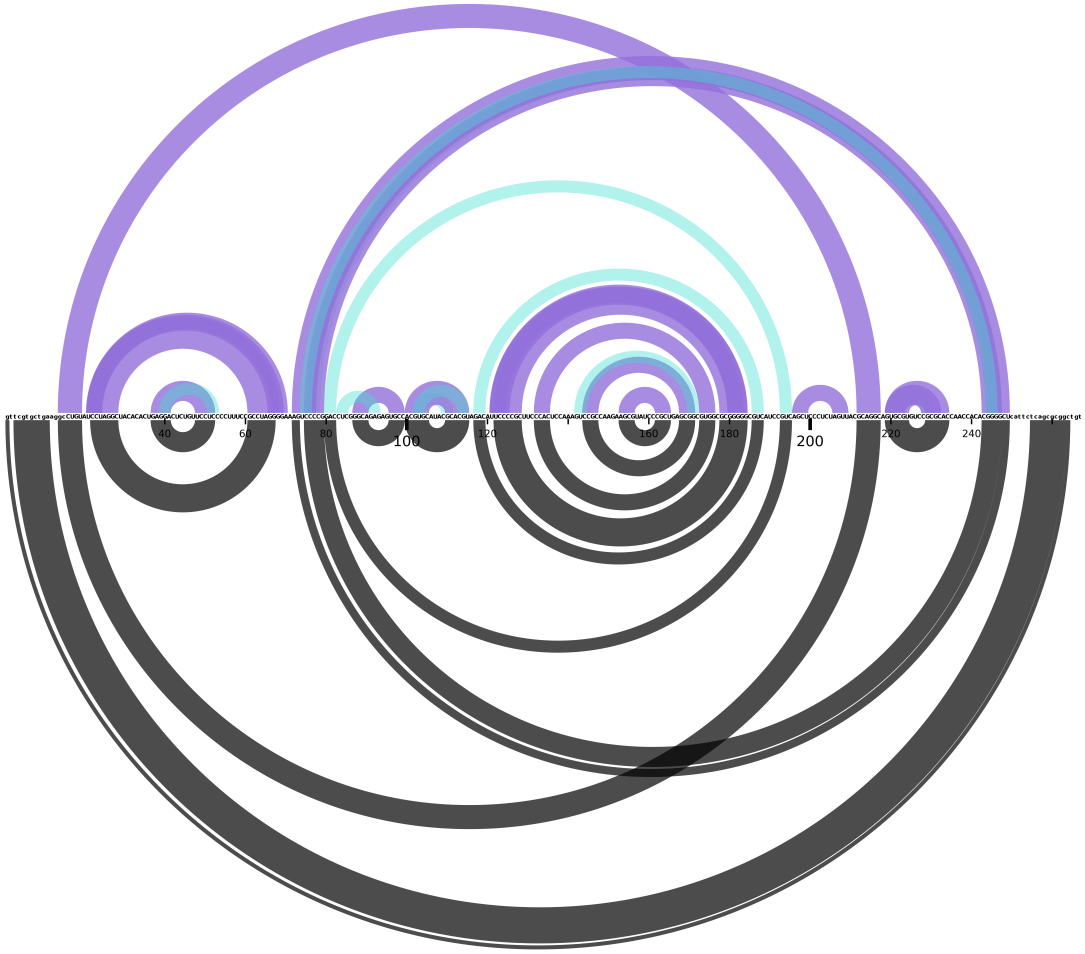




1. Confirm read depth (= data quality) with Laplace distribution.
2. Directly compare/contrast helix support in different conditions.
3. Structure prediction independent of NNTM optimization. Etc.

In conclusion: keep it simple

(Robust too. Comprehensible is good. Interpretable as well. And. . .)



ReDMaxH structure for human RMRP in vitro DMS-MaP data