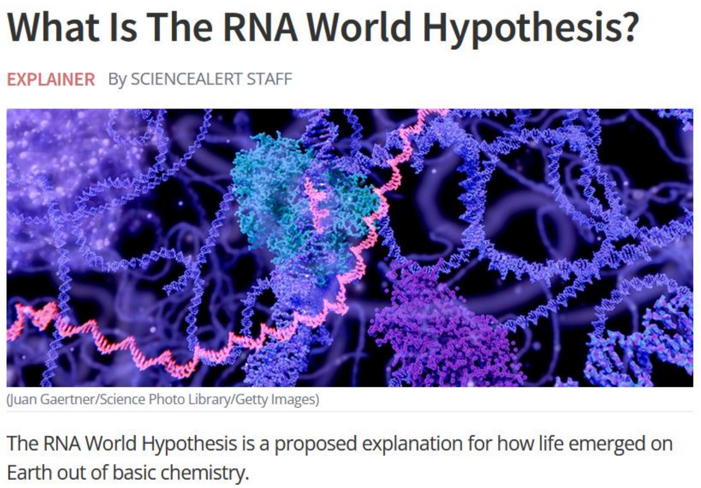
Springboard Capstone 3: RNA Structure Prediction

Diana Koulechova | May 2024

RNA is a fundamental building block of our world



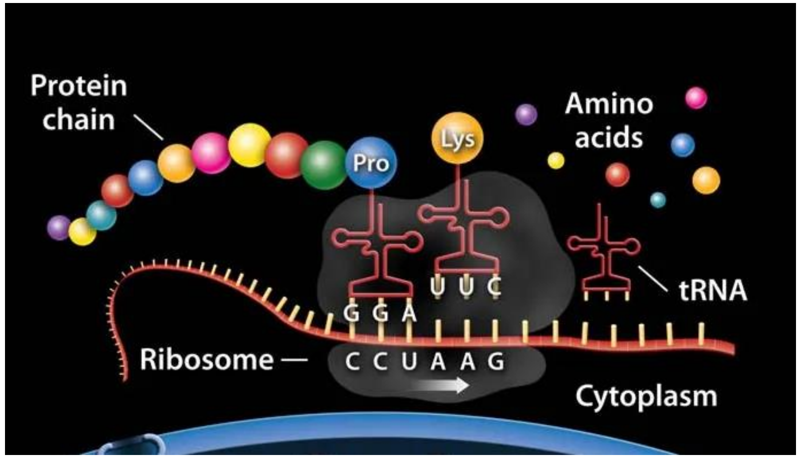
What Is The RNA World Hypothesis?

EXPLAINER By SCIENCEALERT STAFF

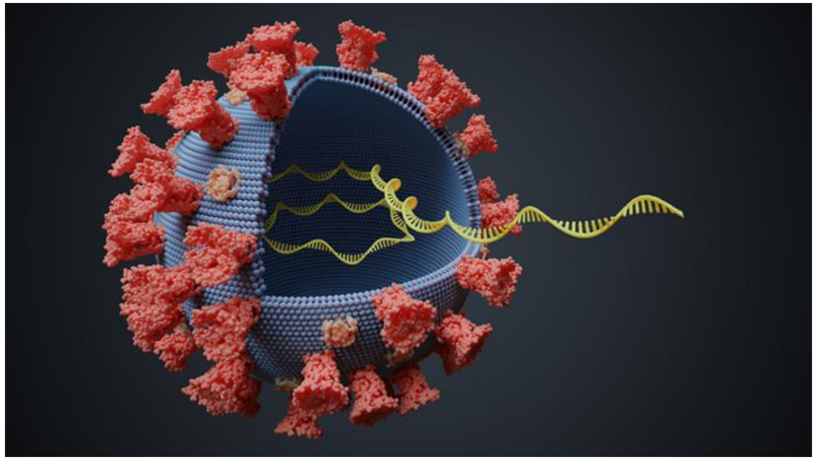
(Juan Gaertner/Science Photo Library/Getty Images)

The RNA World Hypothesis is a proposed explanation for how life emerged on Earth out of basic chemistry.

https://www.salon.com/2020/12/05/mrna-history-vaccines-coronavirus-moderna-immunology-lipid-nanoparticles/



https://www.britannica.com/science/ribosomal-RNA



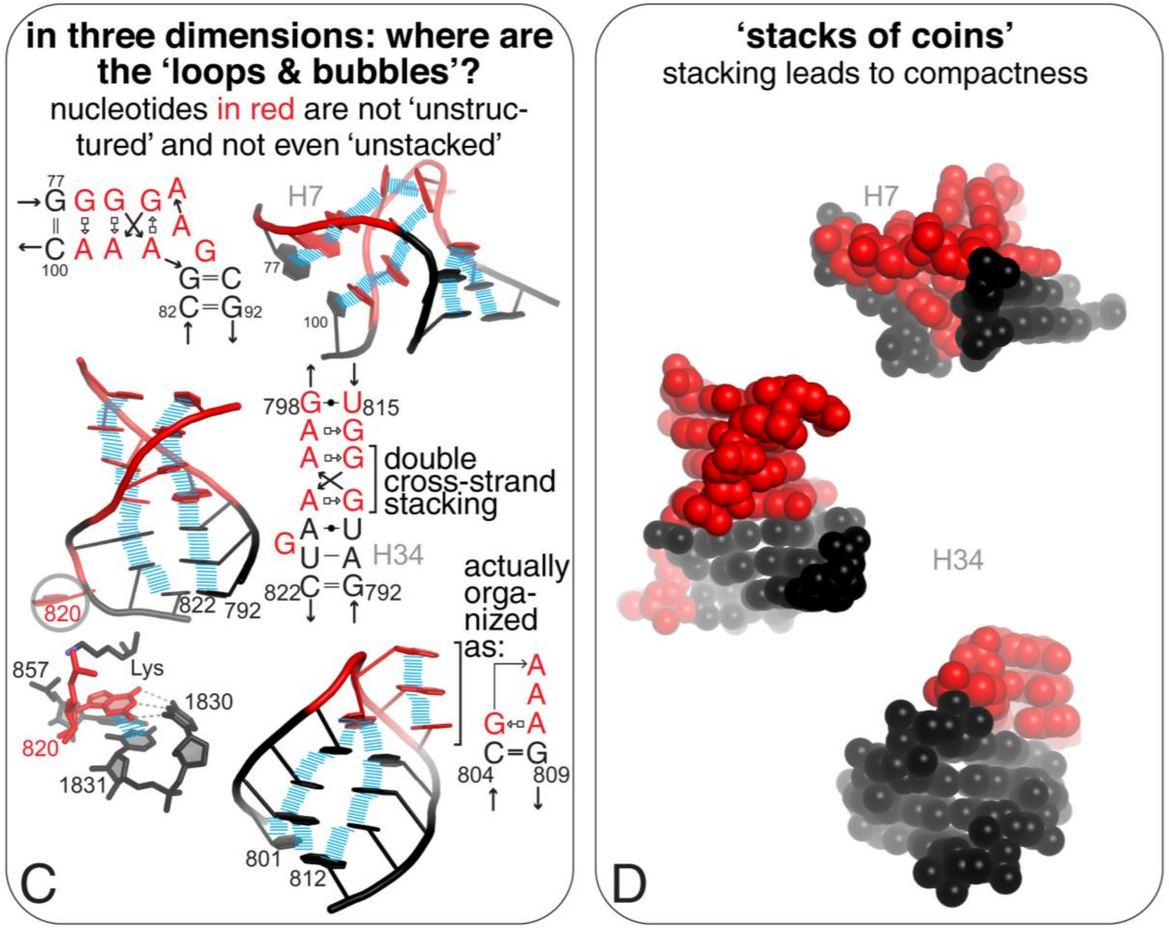
https://directorsblog.nih.gov/2020/07/21/genome-data-helps-track-community-spread-of-covid-19/

Determinants of RNA Structure

- composed of (-)charged ribose backbone + planar aromatic rings studded with polar bonds

- bases stack like coins

- both Watson-Crick and non-Watson Crick base pairing



in three dimensions: where are the 'loops & bubbles'?

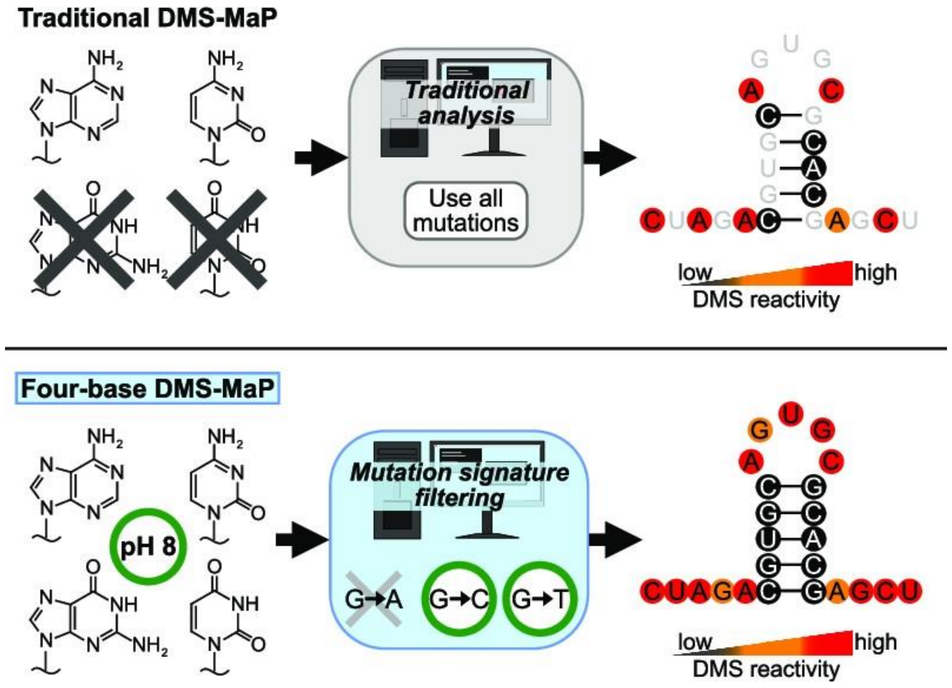
nucleotides in red are not 'unstructured' and not even 'unstacked'

'stacks of coins'

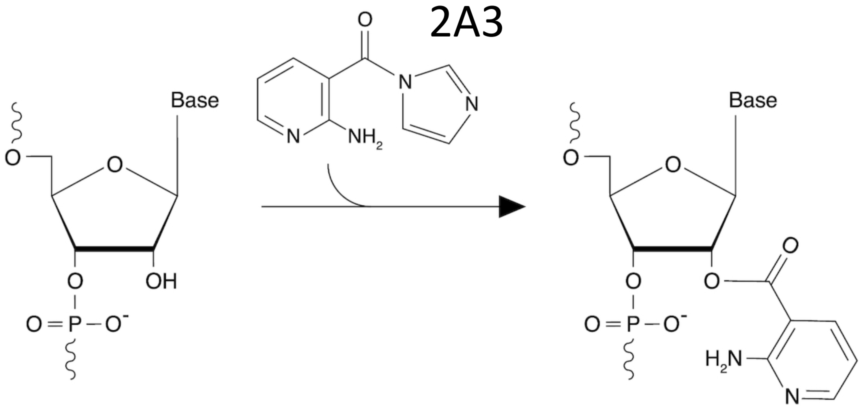
stacking leads to compactness

https://www.pnas.org/doi/10.1073/pnas.2112677119

DMS and 2A3 are RNA structural probes with differing mechanisms of action



https://doi.org/10.1093/nar/gkad522


2A3

https://doi.org/10.1093/nar/gkaa1255

What the starting data look like

COLUMN NAME	EXAMPLE OF CONTENTS	ADDITIONAL NOTES
sequence_id	00026ef17e1b	2 rows for each ID; 167,808 unique IDs
sequence	GGGAACGACUCGAGUAGAGUCGAAAAGGAGAU	11 – 206 characters
experiment_type	either DMS_MaP or 2A3_MaP	
dataset_name	OpenKnot1_Twist_2A3_EternaPlayers	origin of experimental data
reactivity_001	[null]	measured probe reactivity for each position in sequence
…	0.725	normalized so 90th percentile value is 1; theoretically ≥ 0
reactivity_206	[null]	start and end of sequences cannot be probed (all null values)
reactivity_error_001	[null]	
…	0.256	error associated with each measurement
reactivity_error_206	[null]	

What the starting data look like

COLUMN NAME	EXAMPLE OF CONTENTS	ADDITIONAL NOTES
sequence_id	00026ef17e1b	2 rows for each ID; 167,808 unique IDs
sequence	GGGAACGACUCGAGUAGAGUCGAAAAGGAGAU	11 – 206 characters
experiment_type	either DMS_MaP or 2A3_MaP	
dataset_name	OpenKnot1_Twist_2A3_EternaPlayers	origin of experimental data
reactivity_001	[null]	measured probe reactivity for each position in sequence
…	0.725	normalized so 90th percentile value is 1; theoretically ≥ 0
reactivity_206	[null]	start and end of sequences cannot be probed (all null values)
reactivity_error_001	[null]	
…	0.256	error associated with each measurement
reactivity_error_206	[null]	

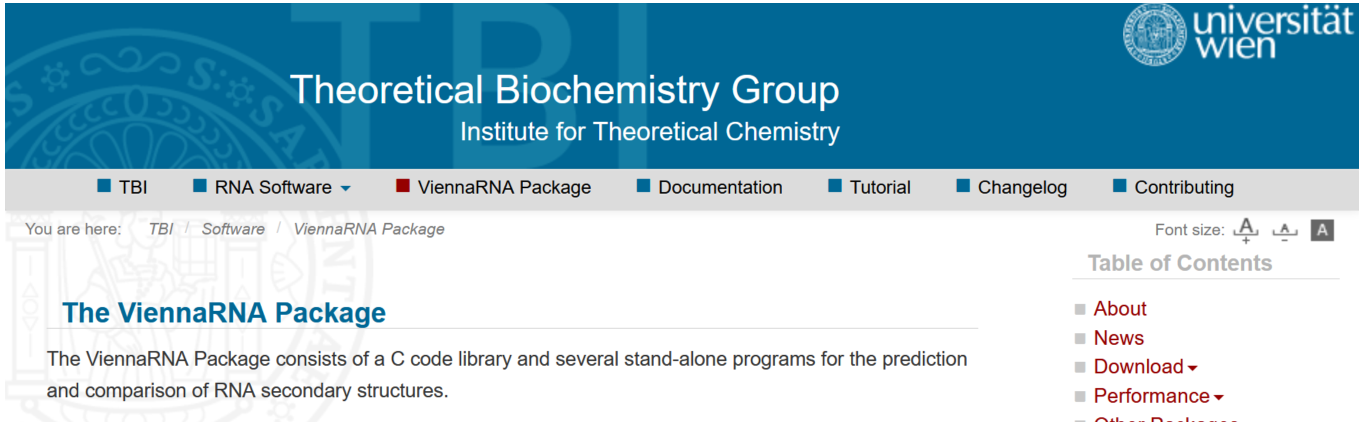
what we're trying to predict

What the starting data look like

COLUMN NAME	EXAMPLE OF CONTENTS	ADDITIONAL NOTES
sequence_id	00026ef17e1b	2 rows for each ID; 167,808 unique IDs
sequence	GGGAACGACUCGAGUAGAGUCGAAAAGGAGAU	11 – 206 characters
experiment_type	either DMS_MaP or 2A3_MaP	
dataset_name	OpenKnot1_Twist_2A3_EternaPlayers	origin of experimental data
reactivity_001	[null]	measured probe reactivity for each position in sequence
…	0.725	normalized so 90th percentile value is 1; theoretically ≥ 0
reactivity_206	[null]	start and end of sequences cannot be probed (all null values)
reactivity_error_001	[null]	
…	0.256	error associated with each measurement
reactivity_error_206	[null]	

what we're trying to predict

Vienna Prediction Added as Feature



universität wien

Theoretical Biochemistry Group

Institute for Theoretical Chemistry

- TBI
- RNA Software
- ViennaRNA Package
- Documentation
- Tutorial
- Changelog
- Contributing

You are here: TBI / Software / ViennaRNA Package

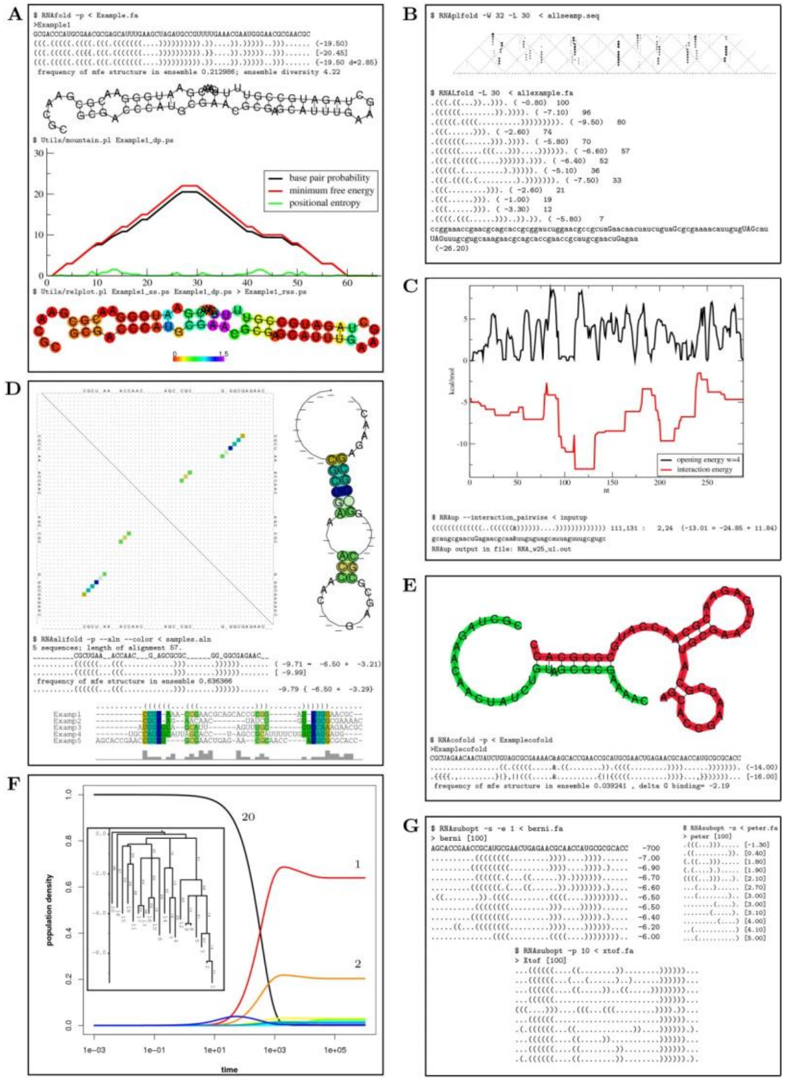
Font size:

Table of Contents

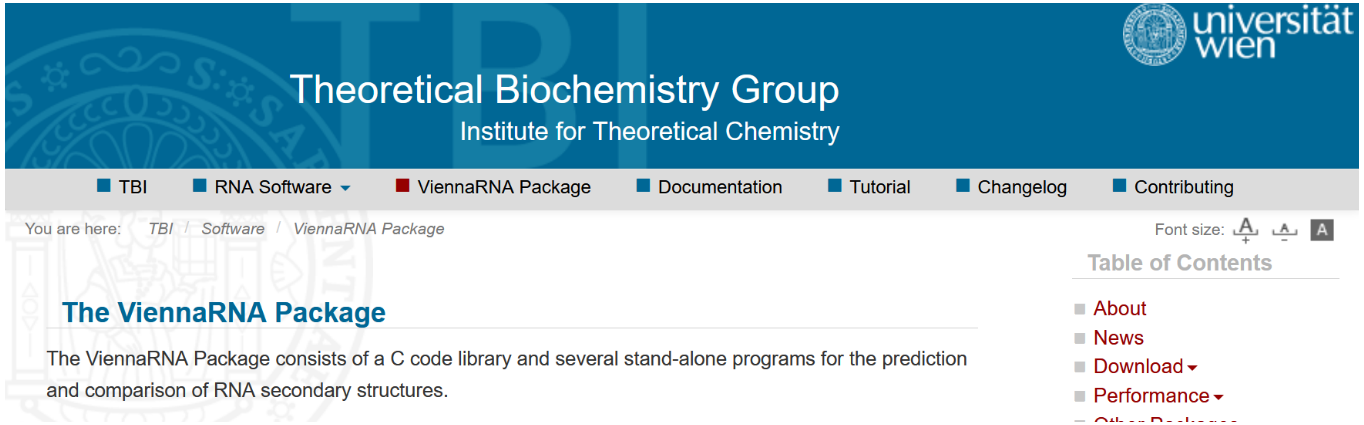
- About
- News
- Download
- Performance

The ViennaRNA Package

The ViennaRNA Package consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures.



Vienna Prediction Added as Feature



universität wien

Theoretical Biochemistry Group

Institute for Theoretical Chemistry

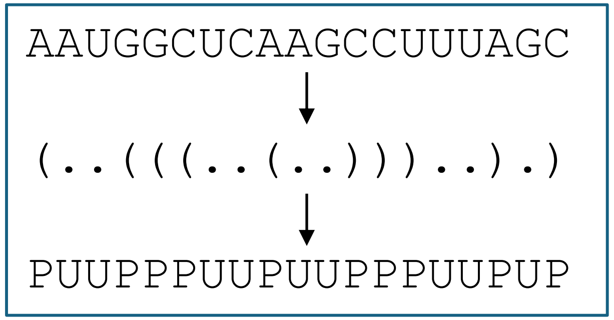
You are here: TBI / Software / ViennaRNA Package

Table of Contents

- About
- News
- Download ▾
- Performance ▾

The ViennaRNA Package

The ViennaRNA Package consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures.



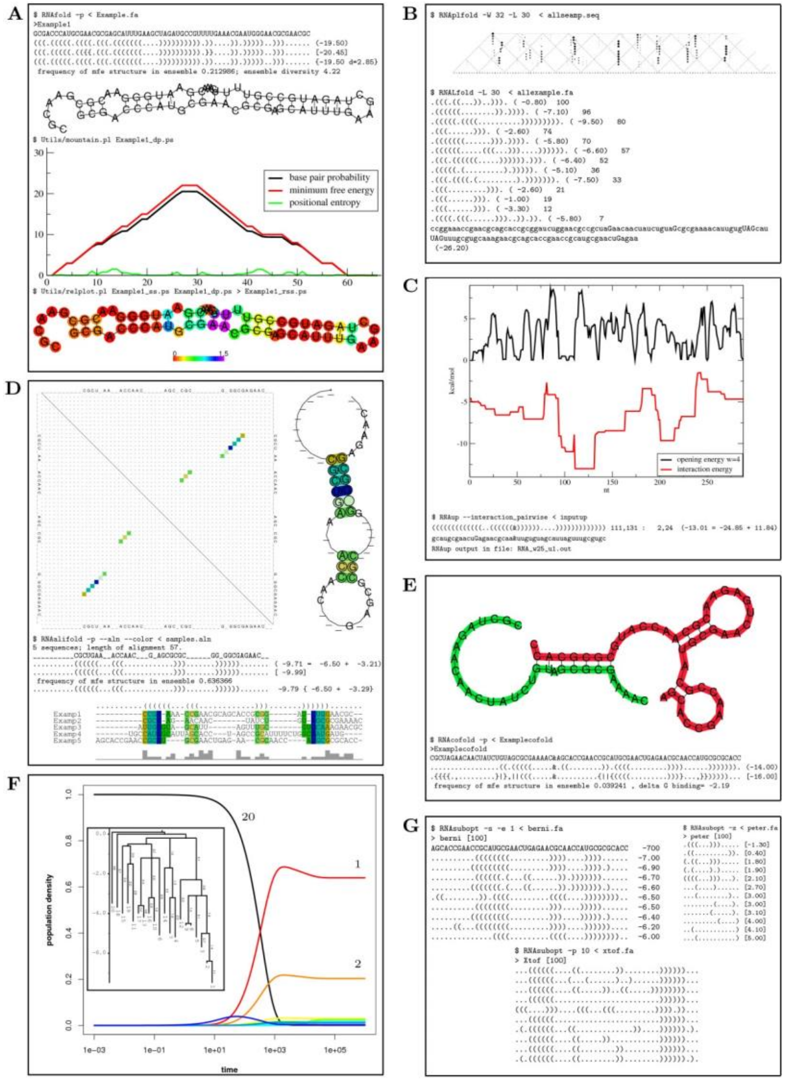
AAUGGCUCAAGCCUUUAGC

↓

(..(((..(..)))..).)

↓

PUUPPPUUPUUPPPUUPUP

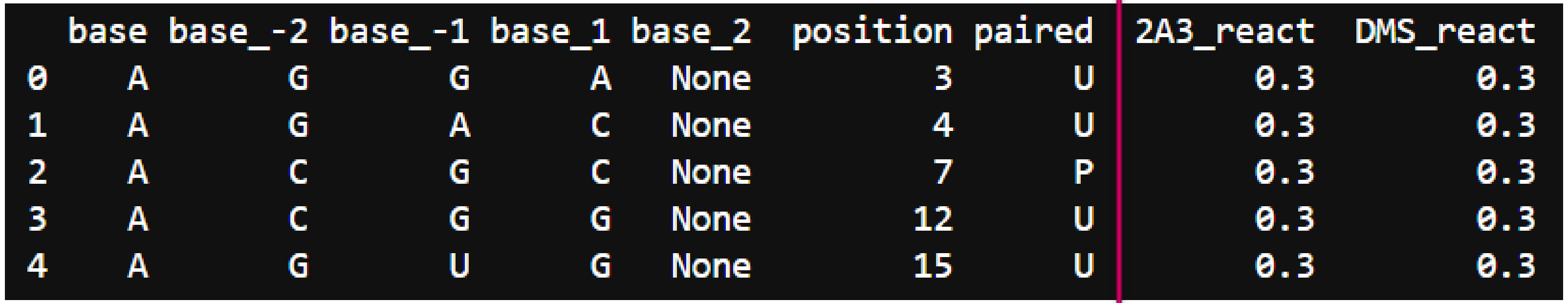


Final Processed Data

neighboring nucleotides

(in sequence)

Vienna



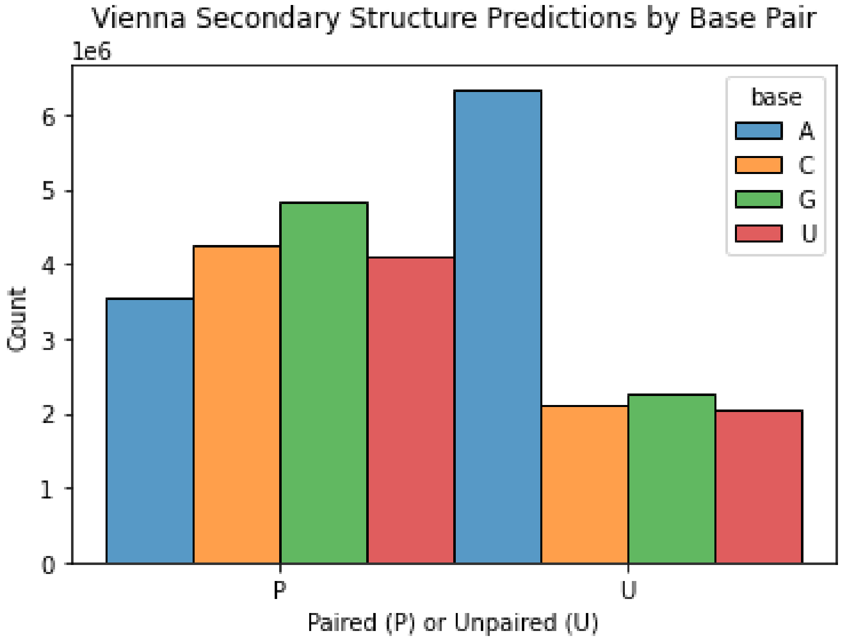
```
   base base_-2 base_-1 base_1 base_2  position paired  2A3_react  DMS_react
0     A       G       G      A   None         3      U        0.3        0.3
1     A       G       A      C   None         4      U        0.3        0.3
2     A       C       G      C   None         7      P        0.3        0.3
3     A       C       G      G   None        12      U        0.3        0.3
4     A       G       U      G   None        15      U        0.3        0.3
```

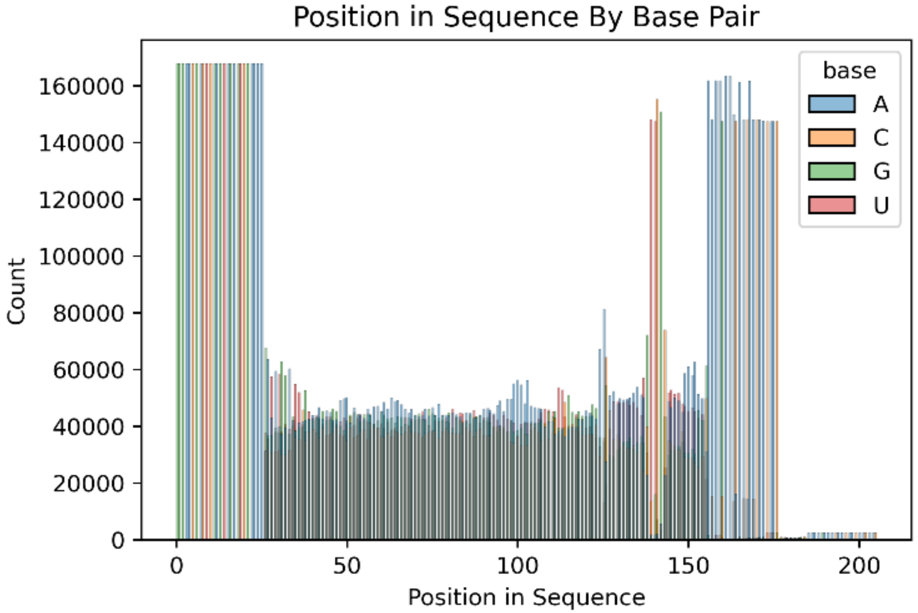
features (X)

targets
(“y”s to be predicted)

22,107,697 rows in the training set

Feature trends by base pair

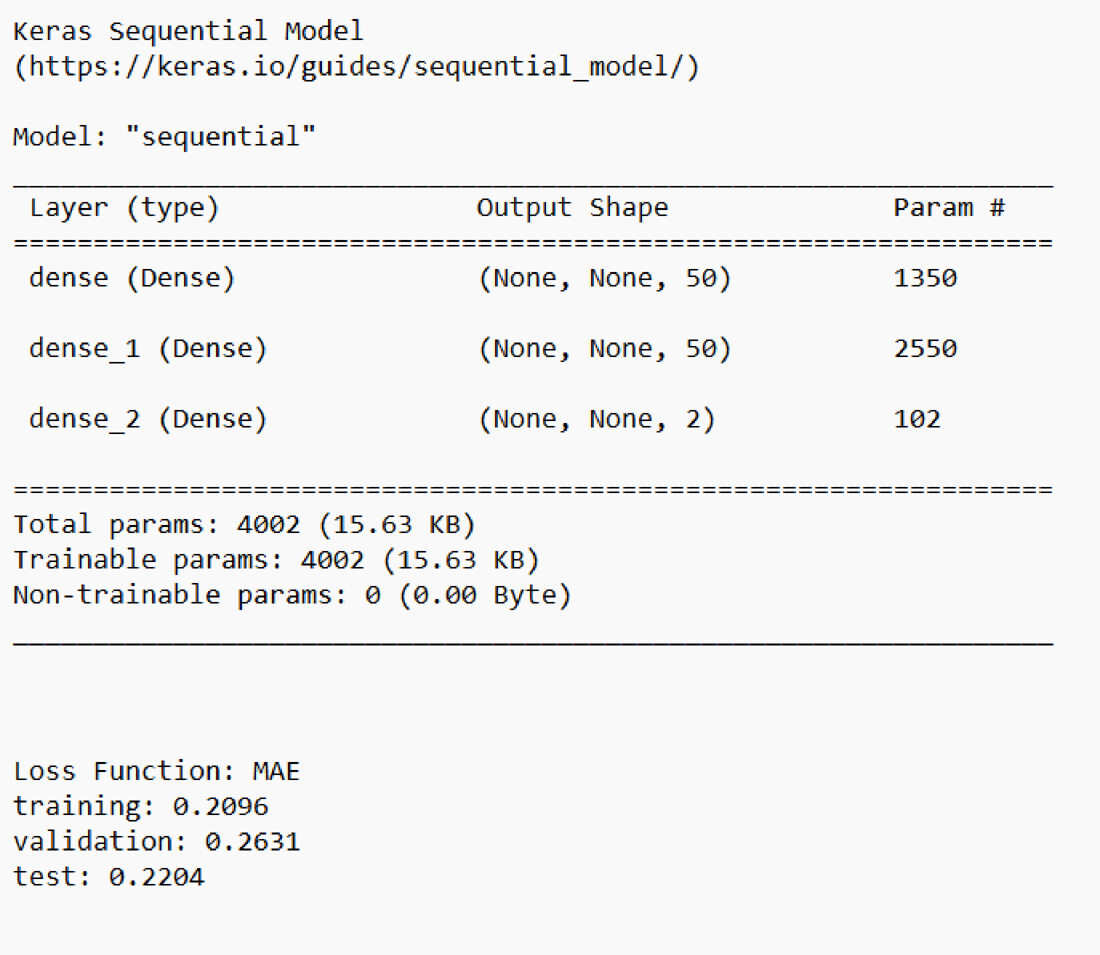




Models Evaluated

- LSTM
- Sequential neural network with dense layers. Iterations evaluated:
  - # nodes
  - # layers
  - batch size

Final Model

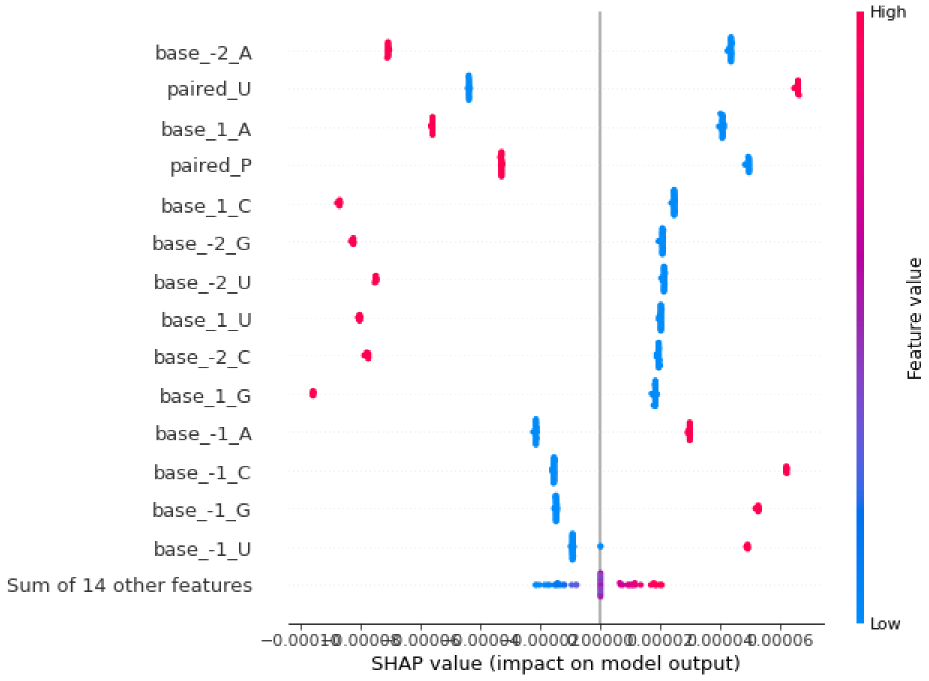


```
Keras Sequential Model
(https://keras.io/guides/sequential_model/)

Model: "sequential"

_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, None, 50)          1350

 dense_1 (Dense)             (None, None, 50)          2550

 dense_2 (Dense)             (None, None, 2)           102

=================================================================
Total params: 4002 (15.63 KB)
Trainable params: 4002 (15.63 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________



Loss Function: MAE
training: 0.2096
validation: 0.2631
test: 0.2204
```

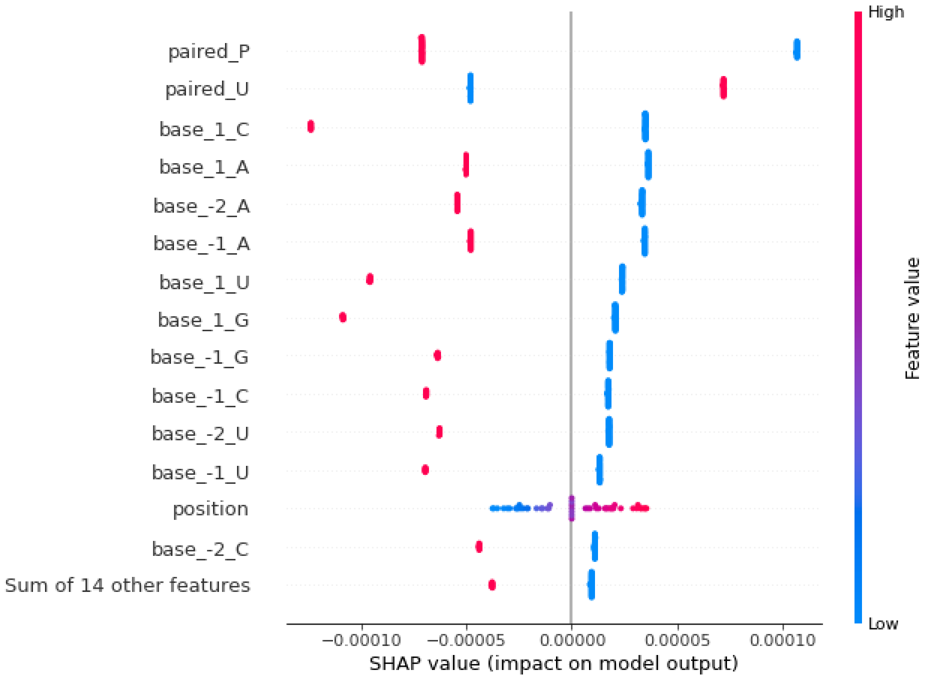
mean (experimental) reactivity
error in dataset: **0.134**

SHAP Beeswarm Plots Highlight Most Important Features for Predictions

2A3

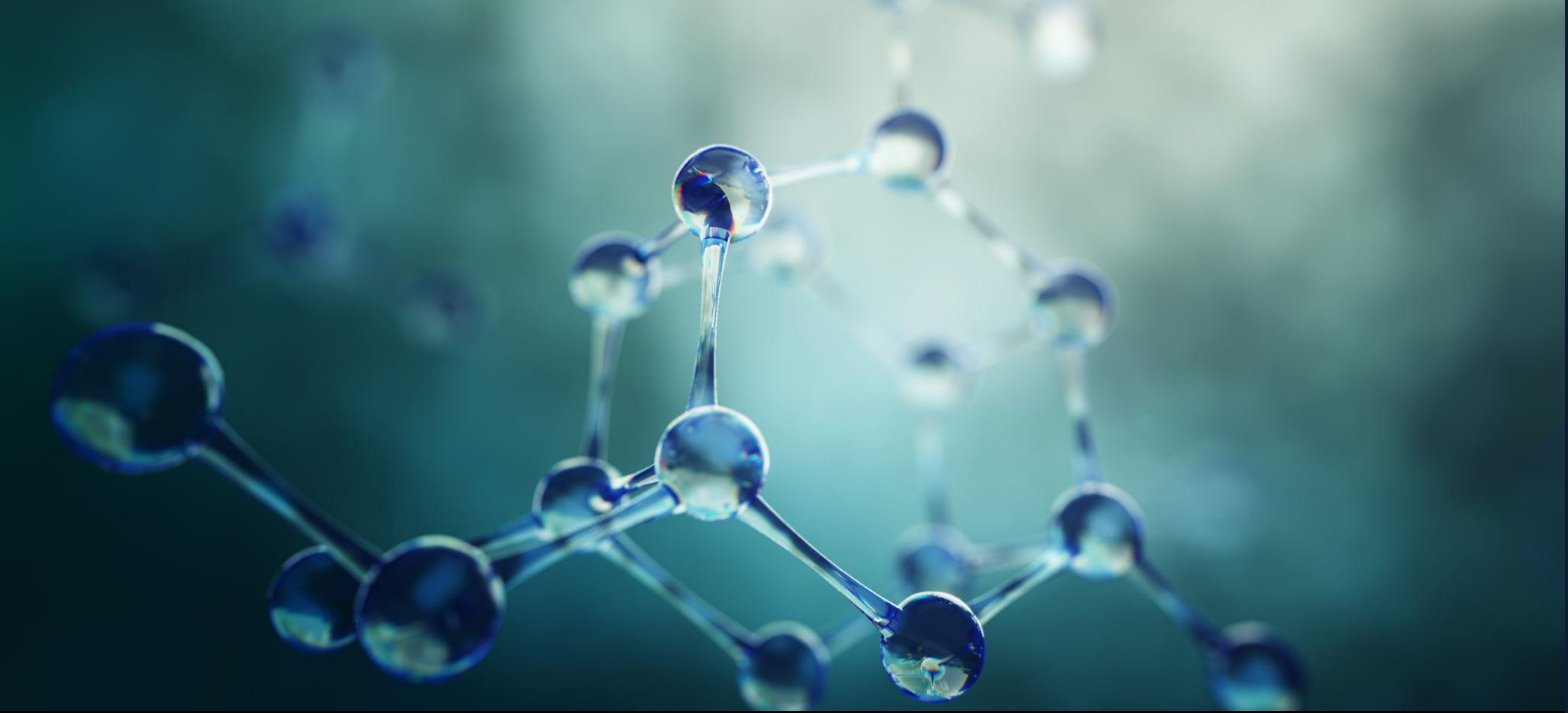


DMS



Next Steps

- How does the number of neighbor nucleotide identities impact the model predictions?
- The overall length of each sequence was not included as a feature. Would that have impacted the predictions?
- Is the prediction accuracy agnostic to probe type? RNA functional and/or structural families?
- RNA structure is inherently highly dynamic, something that is not captured in a dataset/model like this. Additionally, the model is not able to comment on how changes in conditions (e.g. pH, temperature) would impact the structure. Being able to incorporate these types of nuances into the model would be hugely beneficial.



Thank you!