RNA STRUCTURE PREDICTION WITH AI

BARNACLE: RNA contact prediction by data efficient deep learning

05.12 ALINA BAZAROVA

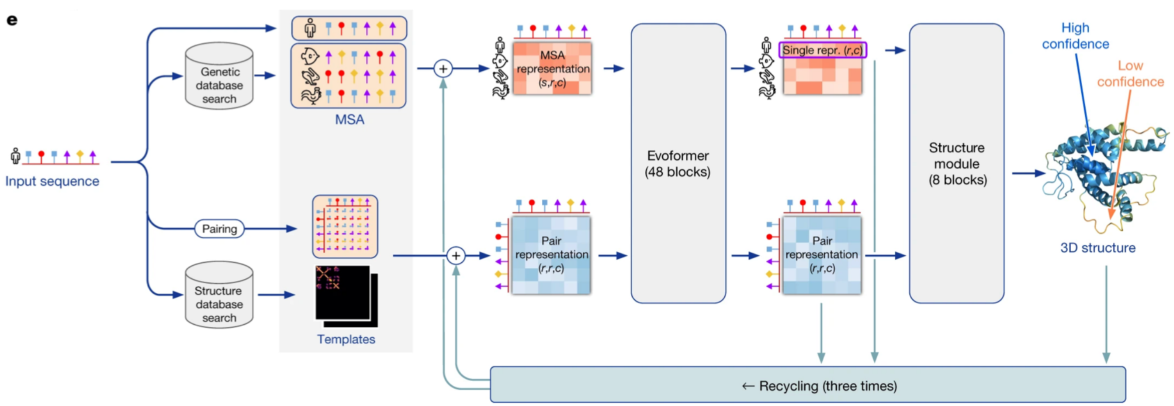
JÜLICH
Forschungszentrum
Wandel gestalten

WHAT IS RNA AND WHY IS IT IMPORTANT?

- Ribonucleic acids (RNA) are a **crucial class of biomolecules** which play essential role in many biological processes

- RNA is of great **biological and medical** importance: vaccine development, regulatory functions etc.

- RNA interact a lot with **proteins**. Yet, there is much more research on the latter than on RNA itself.



Alphafold 2

WHAT ARE THE CHALLENGES WHEN WORKING WITH RNA?

- Unlike protein structure, knowledge on RNA structure is **rather limited** (less state of the art knowledge, less databases)

- Available labeled data for RNA is **orders of magnitude smaller** than for proteins (~100 MSAs in our case)

- Transfer of methods developed for proteins is **not feasible** due to a huge data gap.

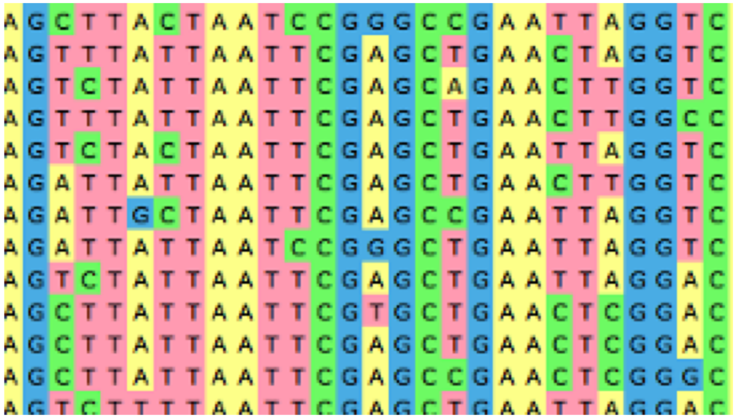
- Obvious case for the **self-supervised learning:**

  - Develop an **upstream model** based on augmentation or transformation of the **unlabeled data**

  - Fine-tune the upstream model with a **downstream** task using the **labeled data**

WHAT ARE THE INPUTS? AND WHAT ARE THE LABELS?

Multiple Sequence Alignments (MSAs)

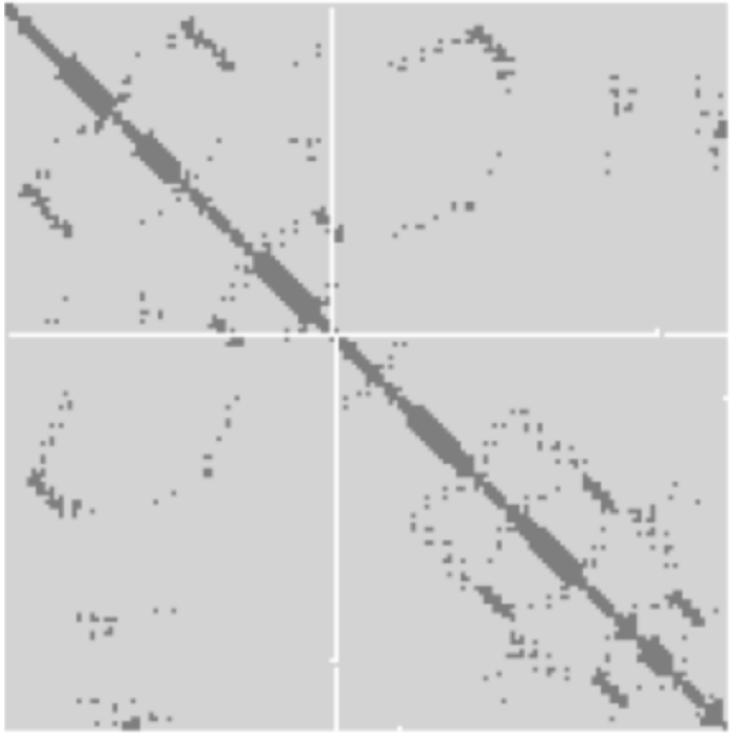


a set of sequences with evolutionary relationship between them

2D: somewhat similar to an image

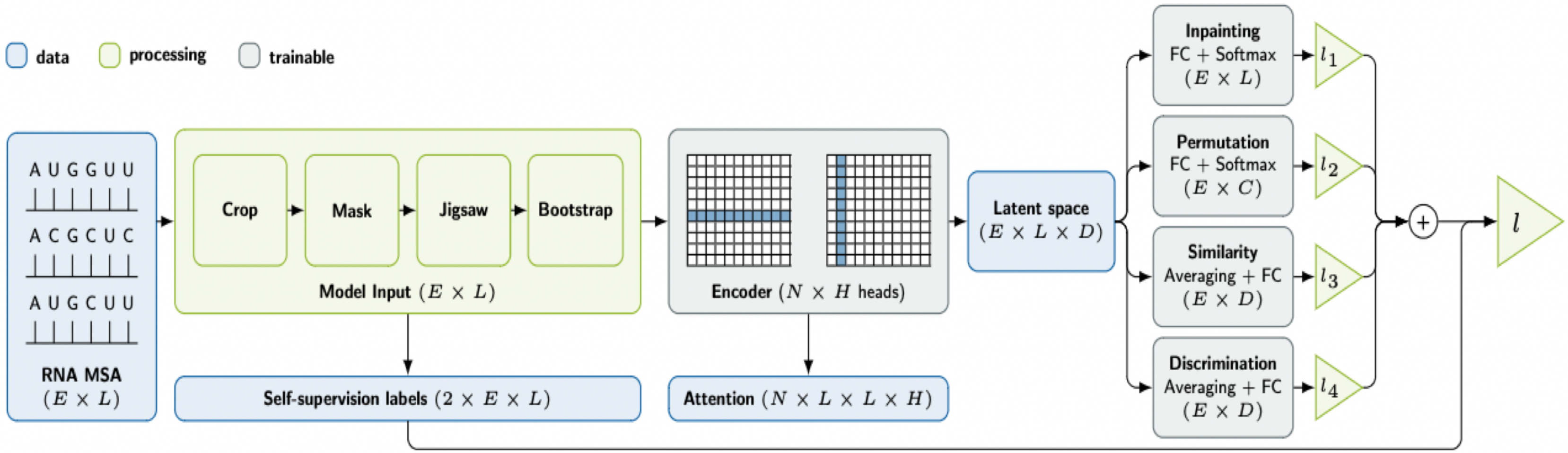
~4000 MSAs for upstream training

Contact maps



The actual sequence along x and y axis.
Which bases are in contact with each other

PROPOSED APPROACH. UPSTREAM TASKS (BACKBONE)



Self-supervised multi-task pre-training:

- **inpainting**: randomly omit nucleotides from the MSA

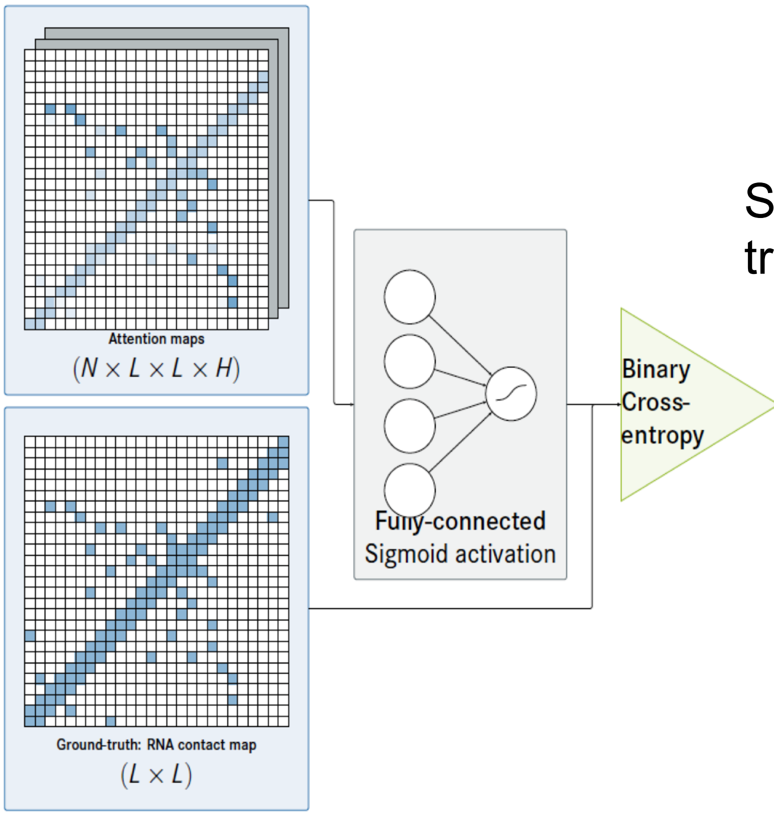
- **jigsaw**: permutes the sequences of the MSA -> good upstream, terrible downstream *(clear shortcut)*

- **bootstrapping**: replaces some sequences by sampling synthetic ones from the *columnwise distribution*

- **contrastive**: minimizes the distance between the sequences from the same MSA *(no augmentation)*

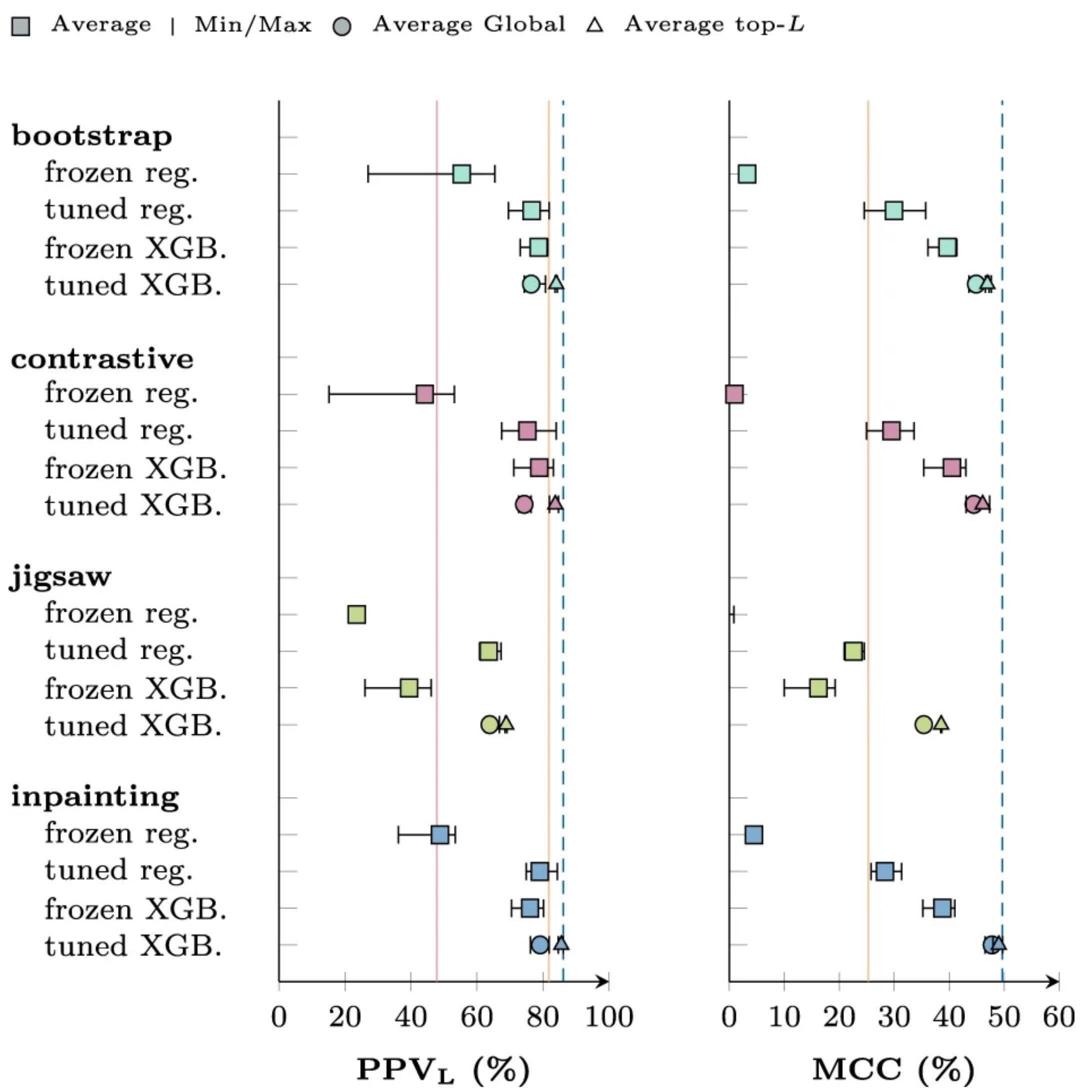
PROPOSED APPROACH. DOWNSTREAM TASKS

Concatenated latent attention maps are fed into **Logistic regression**, **XGBoost** using **frozen/fine-tuned** backbone



Schematic for logistic regression model trained with binary cross-entropy

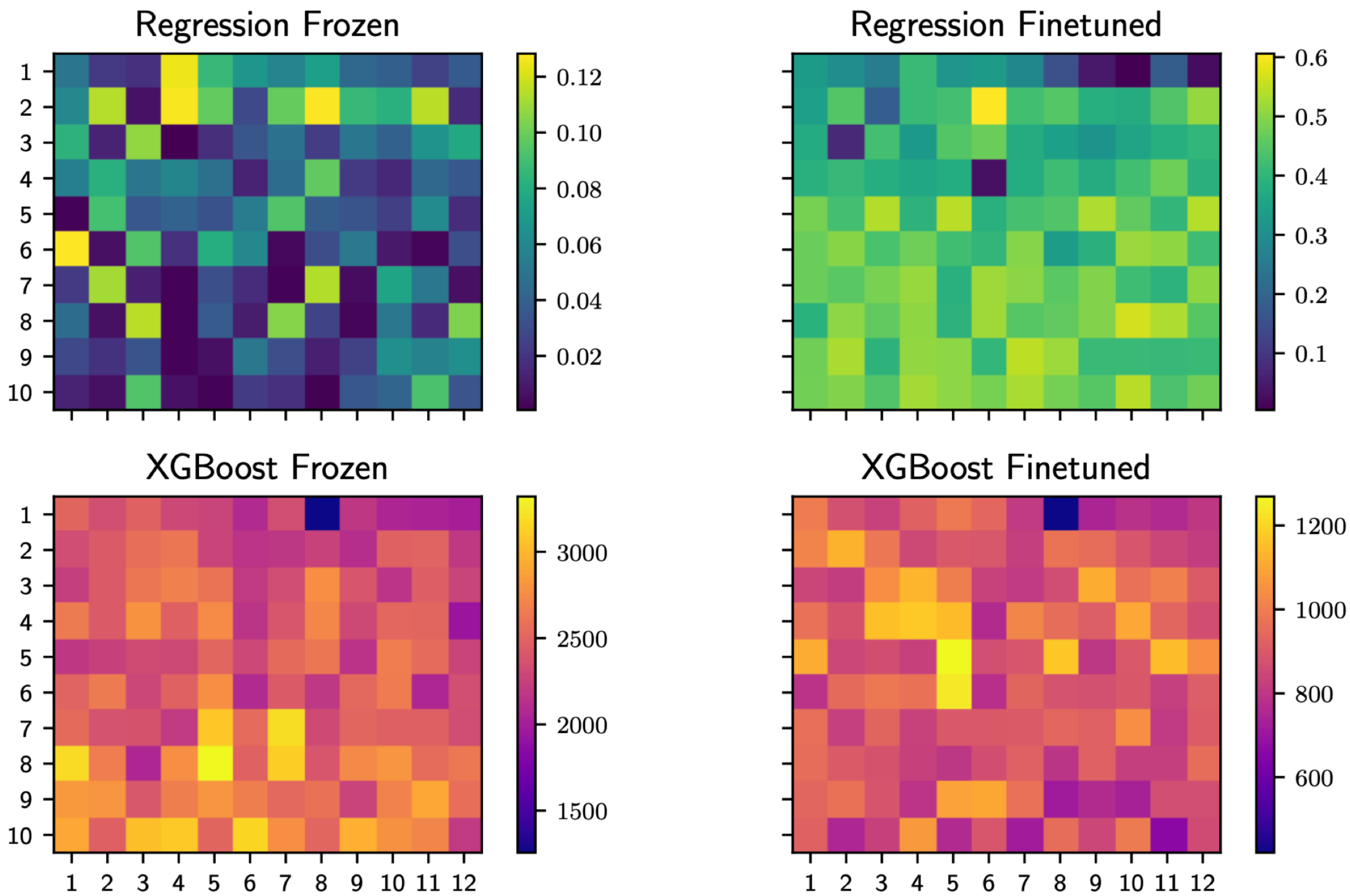
RESULTS



Red line is the DCA baseline performance for PPV, orange line the shallow neural network CoCoNet, and the dotted blue line the best-trained model performance.

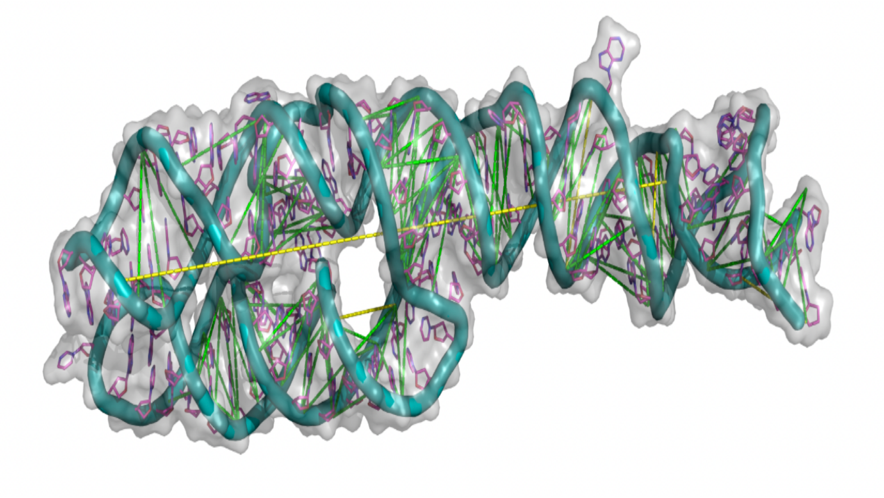
JÜLICH Forschungszentrum

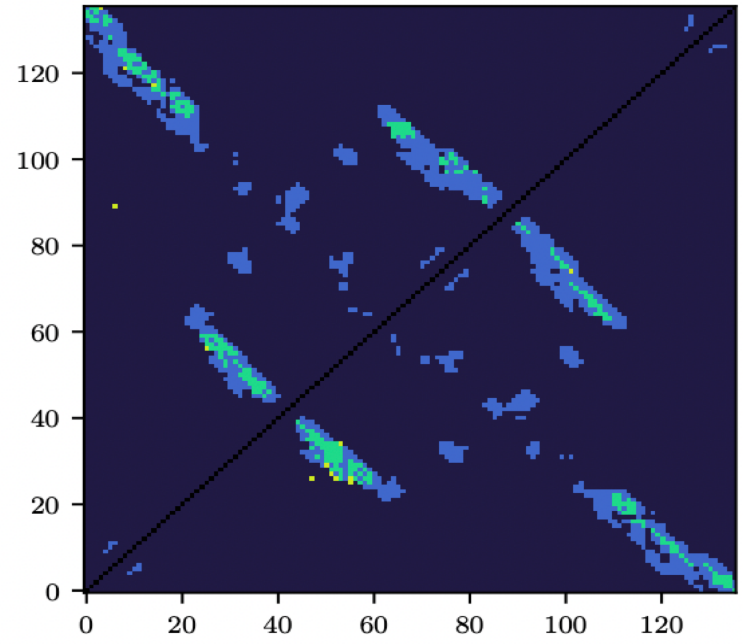
FEATURE IMPORTANCE OF THE DOWNSTREAM TASKS



Attention heads along x axis and the block number along y

VISUALISATION OF THE RESULTS





RNA molecule, where green are the TP contacts and yellow are the FP contacts.

Contact map: TP green, FP yellow
FN light blue, TN dark blue

GENERALISIBILITY

- Apply a pre-trained backbone to **solvent accessibility surface area** (ASA) prediction
- **Regress** the outcome **per token** -> a different downstream task
- Instead of **latent attention maps** use **latent output of the upstream model** as input to downstream task
- RNA temperature adaptation dataset (172 MSAs)

	Frozen NN	Tuned NN	Frozen XGB	Tuned XGB
Jigsaw	0.1731	0.4849	0.4877	0.6740
Contrastive	0.1832	0.5292	0.4930	0.7194
Bootstrap	0.1115	0.4938	0.4927	0.6984
Inpainting	0.1934	0.5222	0.4905	0.7443

Pearson correlation coefficient for ASA prediction. The baseline performance is 0.63

Thank you very much for your attention!

https://go.fzj.de/barnacle2

Article | Open access | Published: 06 September 2023

RNA contact prediction by data efficient deep learning

Oskar Taubert, Fabrice von der Lehr, Alina Bazarova, Christian Faber, Philipp Knechtges, Marie Weiel, Charlotte Debus, Daniel Coquelin, Achim Basermann, Achim Streit, Stefan Kesselheim, Markus Götz & Alexander Schug

Karlsruhe Institute of Technology

DLR