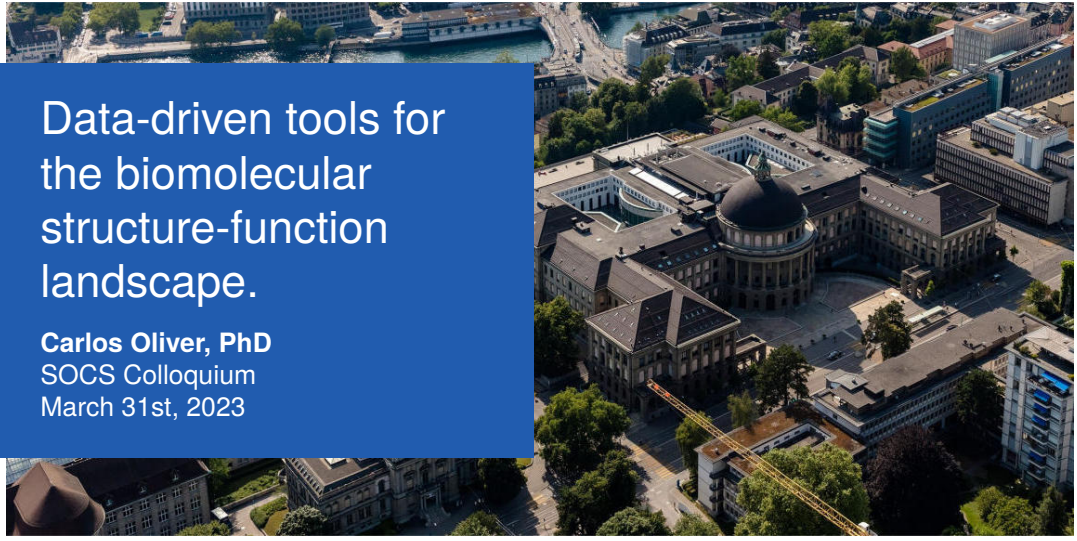
ETH zürich

D-BSSE
Machine Learning & Computational Biology Lab



Data-driven tools for the biomolecular structure-function landscape.

Carlos Oliver, PhD

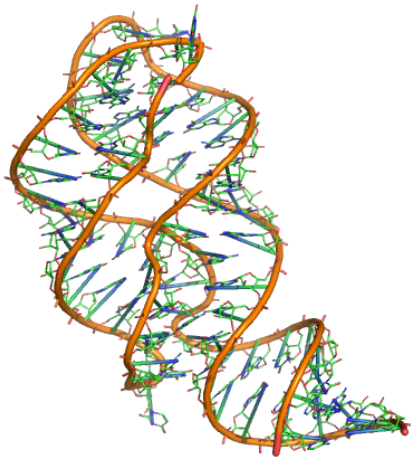
SOCS Colloquium
March 31st, 2023

Outline

1. Biological Background
2. Tools for RNA
3. Tools for Proteins
4. Conclusions

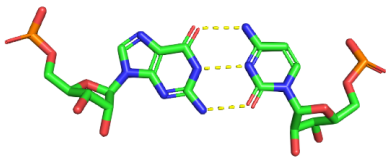
Outline

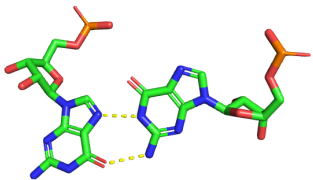
What is an RNA?



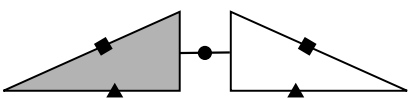
RNA Structure Basics

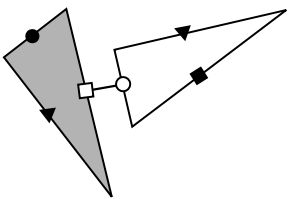
3D





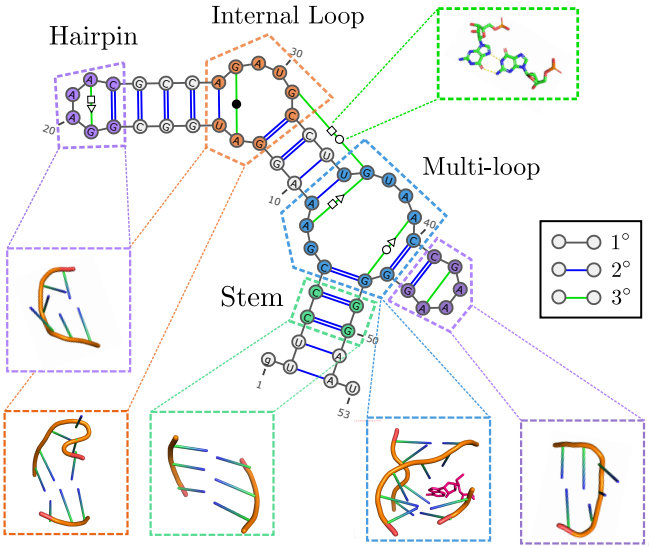
Leontis-Westhof



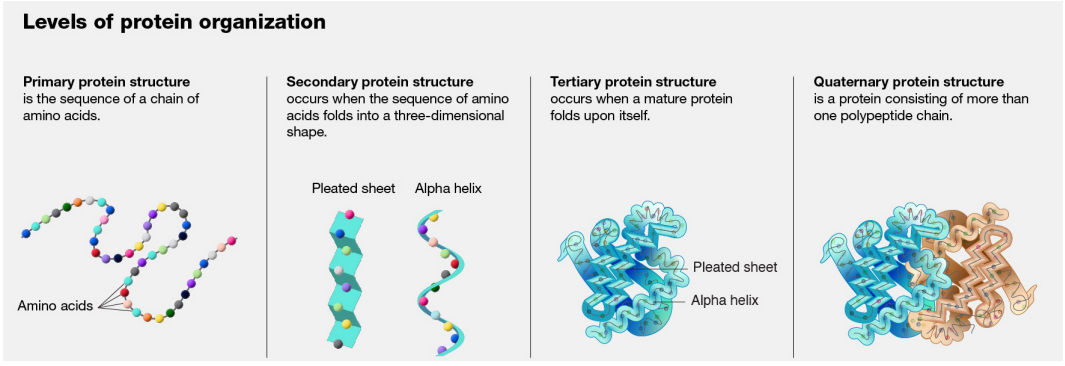


● Watson-Crick (WC) ▲ Sugar (S) ■ Hoogsten (H) ●/○ *cis/trans*

RNA Structure Basics



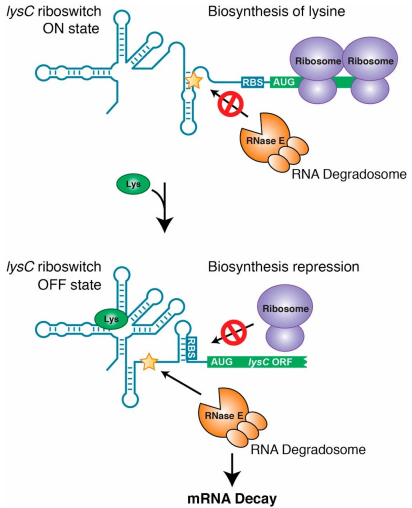
What is a protein?



1

[1] https://www.genome.gov/genetics-glossary/Protein

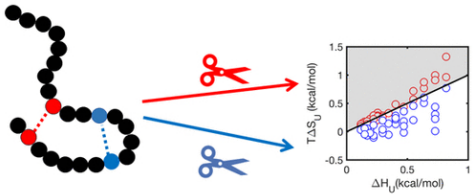
Case study: riboswitches



- Structure (conformation) mediates physical/chemical interactions.
- Interactions drive biological function

[a]Caron, Marie-Pier, et al. 2012.

Question 1: How does structure map to function?



$$\Delta G^{\text{Mutant}} > \Delta G^{\text{Mutant}} > \Delta G^{\text{Wild-type}}$$

a

[a] Bigman, Lavi S., and Yaakov Levy., 2018

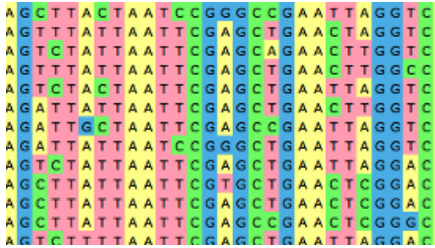
Pros

- Directly explainable

Cons

- Slow & costly
- Lack of generalization

Question 2: How do we discover new functional structures?



[a] http://ugene.net/multiple-sequence-alignment-with-muscle/

Pros

- DNA sequences are cheap
- Strong signals
- Computationally efficient

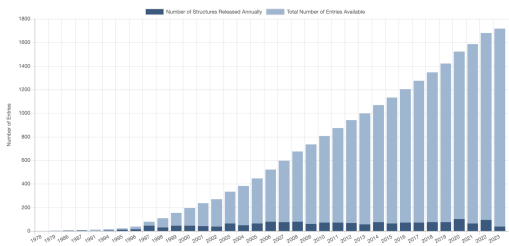
Cons

- Requires known homologs
- Only works for 'natural' molecules
- Structure more conserved than sequence.

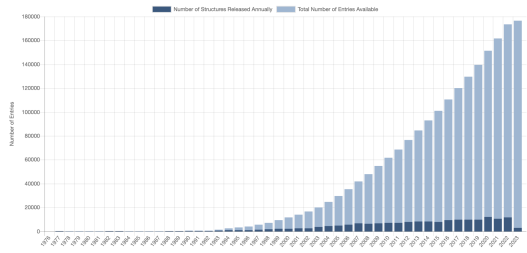
The data-driven approach on 3D

Idea

Experimental and computational 3D data is growing rapidly. How do we take advantage of it to answer these questions?



RNA

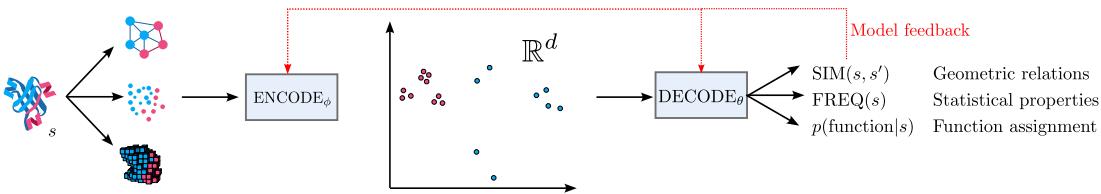


Protein

Basic Computational Framework

Idea

Learn **differentiable** models from data to map geometric objects into Euclidean spaces.



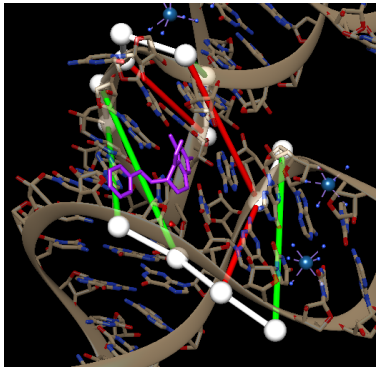
Objective

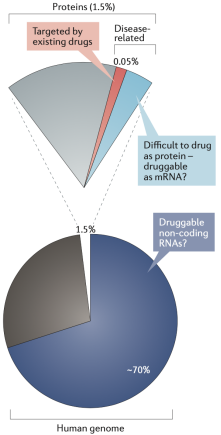
To convince you that **data-driven** analysis of RNA and proteins in **3D** can unlock new insights into **structure-function relationships**.

Outline

RNAmigos: Local geometry informs small molecule binding

- RNA is quickly becoming a promising new class of drug targets.

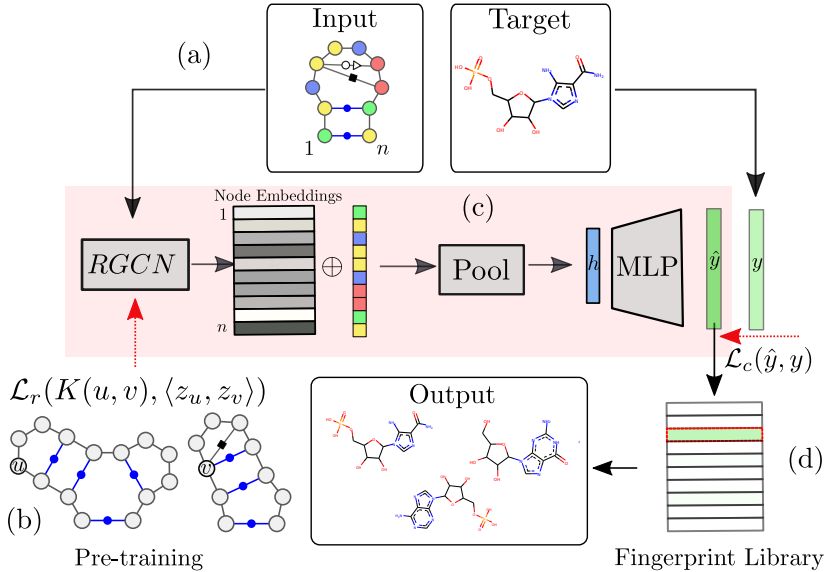




a

[a] Warner, KD, et al. 2018

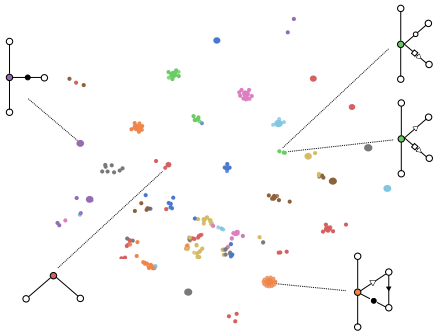
Approach


(a)
Input
Target
1
n
Node Embeddings
1
n
RGCN
(c)
Pool
h
MLP
$\hat{y}$
y
$\mathcal{L}_c(\hat{y}, y)$
$\mathcal{L}_r(K(u,v), \langle z_u, z_v \rangle)$
u
v
(b)
Pre-training
Output
(d)
Fingerprint Library

Encoding structural similarity

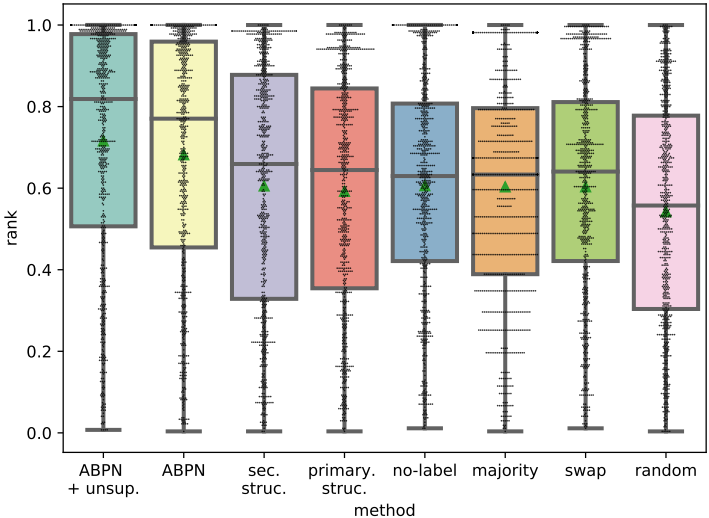
Encoder Loss: Similar structures → proximal embeddings

$$\mathcal{L} = \|\langle \phi(g_u), \phi(g_{u'}) \rangle - s_G(g_u, g_{u'})\|_2^2, \quad (1)$$



- Fast similarity search (vs graph space)
- Models structural variability

Does structure improve the chance of finding the ligand?



Wrap-up

Takeaways

- First systematic 2.5D drug discovery approach.

Open Questions

- Are other representations (point cloud, voxel) sufficient?
- Can we incorporate simulation data?

veRNA1: Discovering functional elements

Idea

High frequency (sub)structures (aka motifs) are functional.

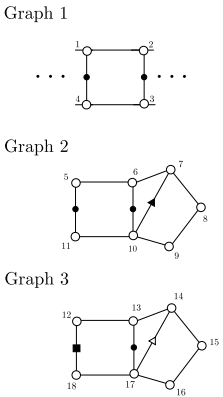
We say a pattern (small connected graph) is a motif given a set of graphs $\mathbb{G}$ if it is isomorphic to more subgraphs of $\mathbb{G}$ than expected.

Challenges in motif discovery:

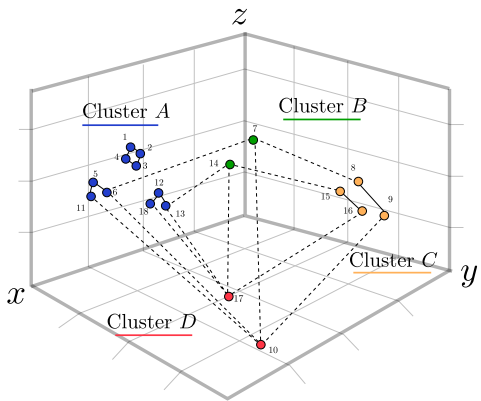
- Approximate (non-isomorphic) motifs occurrences.
- Scaling to large motifs.
- Motifs with continuous node/edge attributes.
- Connecting motifs with function.

1. Prune search space using learned embeddings and connectivity.
2. Iterative clustering to grow motifs.

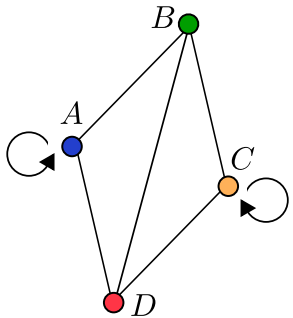
RNA Graphs



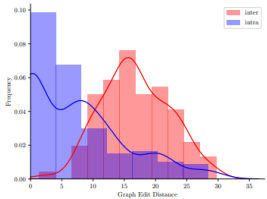
Euclidean Space

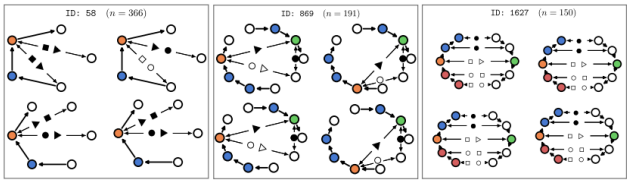


Meta-Graph



VERNAL Motifs





1. We show that the motifs we identify are structurally consistent.
2. Found new occurrences of known motifs
3. Identified 100+ new motifs.

Wrap-up

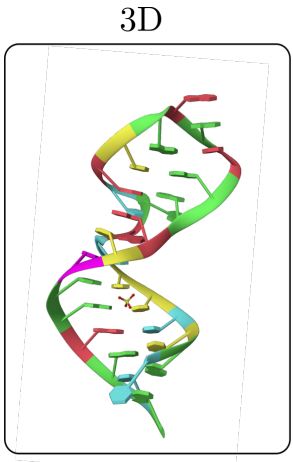
Takeaways

- First RNA 2.5D miner to extract fuzzy patterns.
- Show that large ensembles of possibly functional motifs exist.

Open Questions

- Can we make the mining process fully differentiable?
- Can we systematically link the motifs to function?

`rnaglib`: Easy access to ML-ready RNA structures



Features

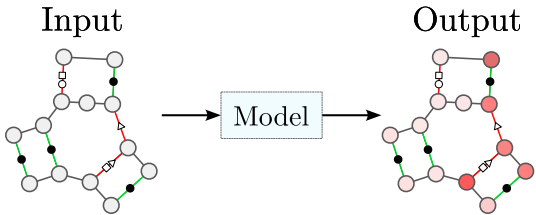
- Easy dataset imports
- 5759 RNAs, 1176 non-redundant RNAs.
- Convert to graphs, voxels, and point clouds
- Standardized splits
- Prediction tasks and evaluators


```python
from rnaglib.dataset import RNADataset
from rnaglib.representations import
    GraphRepresentation

nt_features = ['nt_code'] # residue type

graph_rep = GraphRepresentation(framework='pyg')
dataset = RNADataset(nt_features=nt_features,
    representations=[graph_rep])
```

Benchmarks: Binding prediction



	Task	Dummy Accuracy	Accuracy	Recall	Precision	F1
FR3D	Protein	0.663	0.490	0.507	0.646	0.568
	Ion	0.468	0.609	0.649	0.670	0.659
	Small-Mol.	0.501	0.670	0.743	0.533	0.621
DSSR	Protein	0.537	0.557	0.779	0.512	0.618
	Ion	0.468	0.809	0.849	0.770	**0.808**
	Small-Mol.	0.480	0.652	0.724	0.690	**0.707**

Wrap-up

Takeaways

- Library to broaden adoption of RNA 3D ML tasks

Open Questions

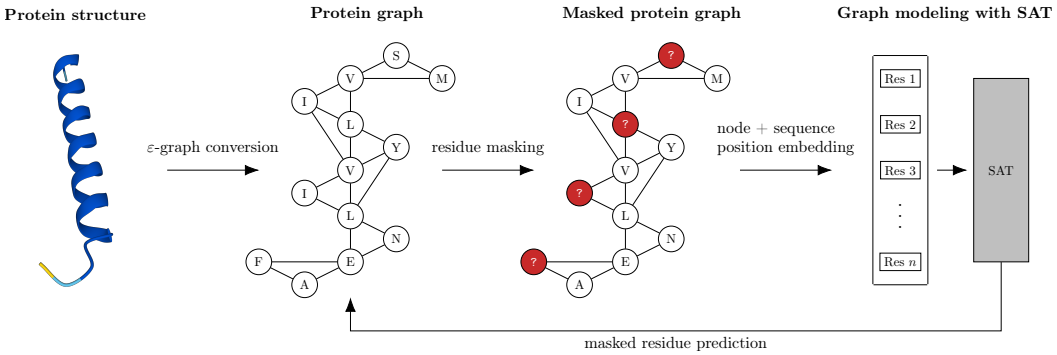
- Will we learn new ways of extracting signal from RNA datasets?
- Can we design tasks at the whole-RNA level?

Outline

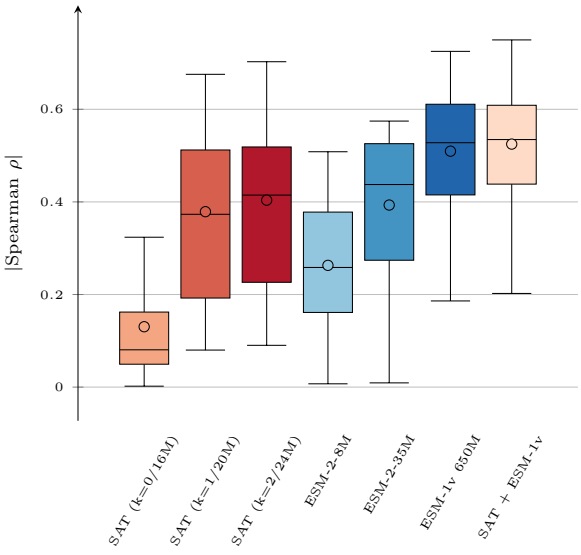
Can we predict the functional outcome of point mutations (MEF)?

Idea

Leverage large AlphaFold 3D dataset as pre-training for mutation effect prediction.



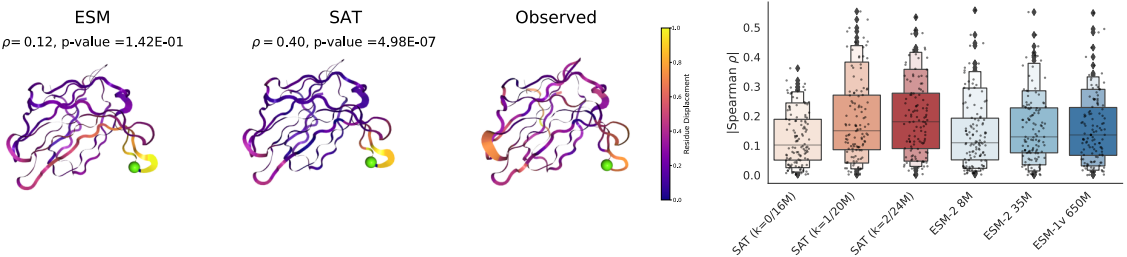
Structure signal is necessary for MEF



- Trained on **500k** AF structures. Compare to ESM trained on **200M+** sequences.
- Predict on deep experimental mutational probes.
- Increasing structural information improves prediction.

Mapping model effects on structure

- Structure-aware model quantitatively maps predictions more closely to observed structural rearrangements.



Wrap-up

Takeaways

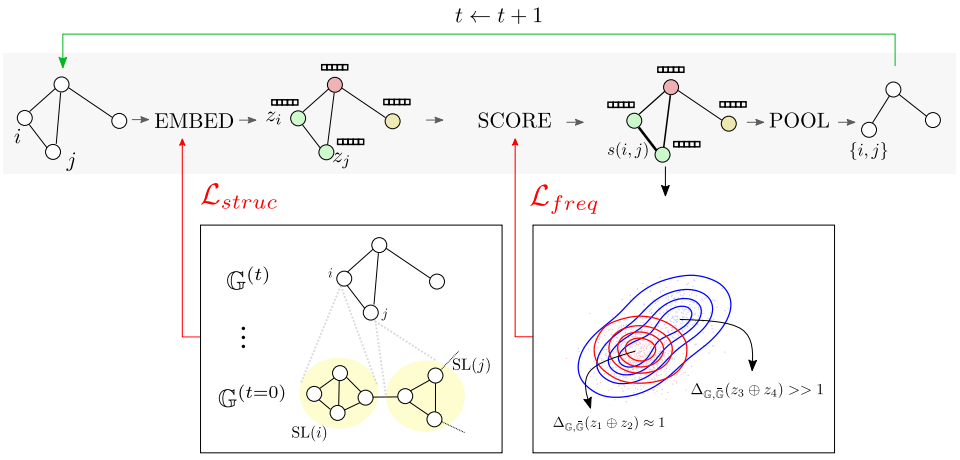
- Predicted structures are rich feature extractors.
- Sequence alone may not be enough.

Open questions

- Can larger structure datasets bring even more gains?
- Can different pre-training strategies unlock fine-grain (displacement) effects?

`MotiFiesta`: Large-scale functional element discovery

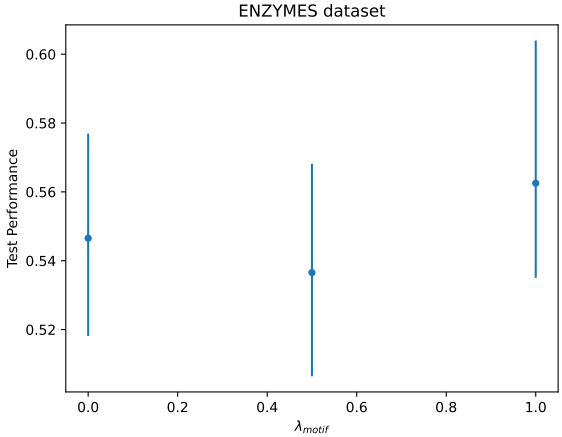
- Protein datasets much larger than RNA, requires more efficient motif extraction.
- Need a way to directly link motifs to function.



Mining & Predicting

- One model to predict **and** mine subgraphs.

$$\mathcal{L} = \lambda_{motif}[\mathcal{L}_{struc} + \mathcal{L}_{freq}] + \mathcal{L}_{clf} \tag{2}$$



Wrap-up

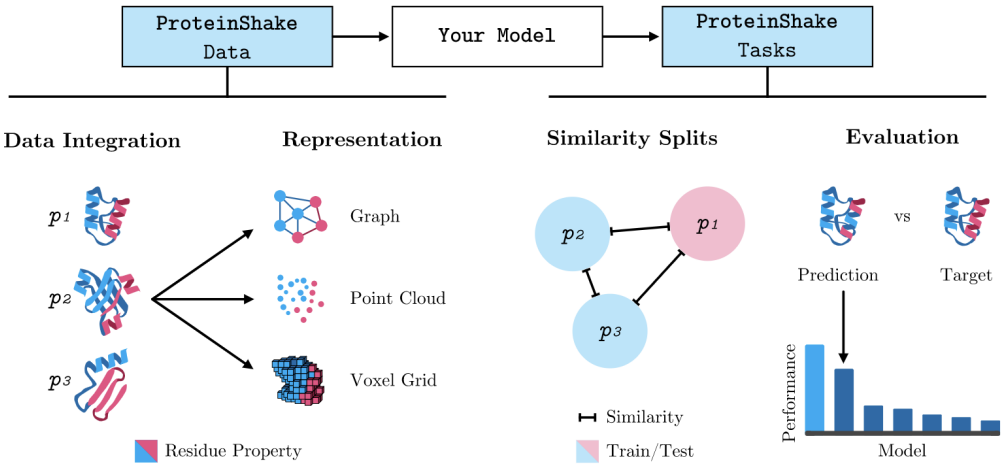
Takeaways

- First fully differentiable fuzzy motif miner.
- Potential for pushing models towards automatic functional substructures.

Open Questions

- Which datasets and contexts rely on motifs for function?
- Can we use this to characterize large ensembles of structures? (e.g. AF vs RCSB)

`proteinshake`: Easy access to ML-ready protein structures



Biologically Relevant Tasks

- For all tasks we provide **structure-aware** splits and evaluators.

Task	Biological Relevance	Input	Output	Task Type
`GeneOntologyTask`	Structure-function	Protein	Function label	Multi-class
`EnzymeComissionTask`		Protein	Enzyme class	Multi-class
`ProteinFamilyTask`		Protein	Protein family	Multi-class
`StructuralClassTask`	Geometric reasoning	Protein	Structure class	Multi-class
`StructureSimilarityTask`		Protein pair	Structural similarity	Regression
`StructureSearchTask`		Protein	Similar proteins	Retrieval
`LigandAffinityTask`	Physical modeling	Protein & Small mol.	Binding affinity	Regression
`ProteinProteinInterfaceTask`		Protein pair	Binding interface	Binary
`BindingSiteDetectionTask`		Residue	Binding pocket	Binary
`VirtualScreenTask`		Protein & small mol. library	Active binders	Regression

Very simple API


```python
from proteinshake.tasks import GeneOntologyTask
from torch_geometric.loader import DataLoader

task = GeneOntologyTask().to_graph(eps=8.0).pyg()

for batch in DataLoader(task.train):
    model.train_step(batch) # your model training goes here

prediction = model.inference(task.test)
metrics = task.evaluate(prediction)
```

Outline

This is just the beginning...

RNA & Protein 3D datasets contain useful signals but we have just scratched the surface. Successful structure models will make breakthroughs in:

- Drug discovery
- Protein engineering
- Gene editing (RNA & CRISPR)
- Antibiotic resistance
- ...

Some large challenges we still need to face:

- Non-static conformations
- Structure generation

Acknowledgements

- Karsten Borgwardt (ETH Zürich & Max Planck Institute)
- Jérôme Waldispühl (McGill)
- Dexiong Chen (ETH Zürich)
- Tim Kucera (ETH Zürich)
- Philip Hartout (Max Planck Institute)
- Vincent Mallet (McGill & Polytechnique Paris)
- Juan Guillermo Carvajal Patiño(McGill)
- Roman Sarrazin Gendron (McGill)
- Pericles Philippopoulos (McGill)
- Jonathan Broadbent (McGill)

Resources & Contact

- Resources
  - `rnaglib`
  - `rnamigos`
  - `vernal`
  - `proteinshake`
  - `motifiesta`
- Contact
  - Email: carlos.oliver@bsse.ethz.ch
  - Webpage: https://carlosoliver.co
  - GitHub: https://github.com/cgoliver

Open Positions!

Looking for PhD students for new group at Max Planck Institute. Contact me if you are interested!