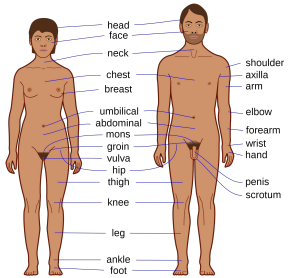
Modeling structural biology with geometric deep learning

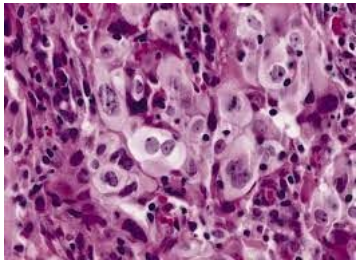
Vincent Mallet - Ecole Polytechnique, CNRS - Maks Ovsjanikov

Fast molecular biology : zooming in !

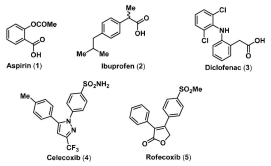


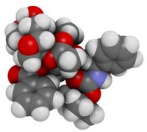
zoom

zoom

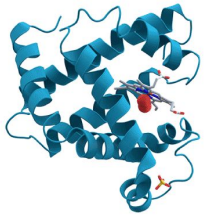


zoom

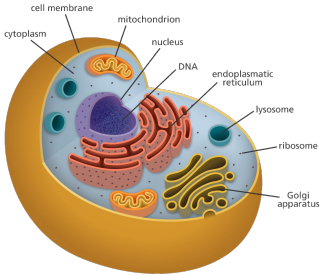




zoom



zoom

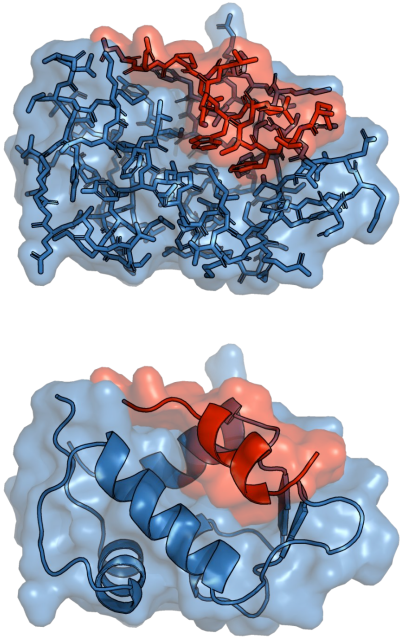


Biomolecules

Cells

Biomolecules structure and function

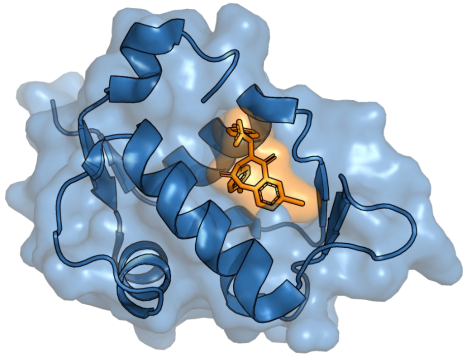
- Biomolecules are the building bricks of living systems and they interact
- *Structure* denotes the relative positions of the atoms of a molecule
- Physics (hence *function*) depend on relative positions



Example structure of MDM2 - p53 complex (PDB code 1ycr)

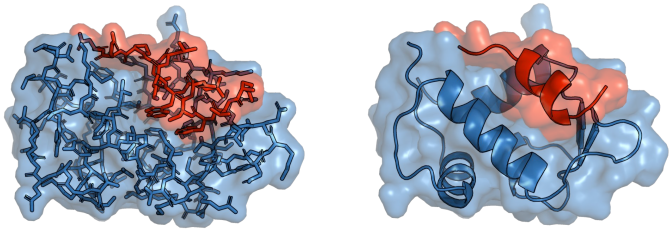
Target-centric drug discovery

- Diseases are induced by pathological function of biomolecules
- Drugs disrupt the pathological function of a target biomolecule
- Uses the target **structure** to simulate its interaction with potential binders



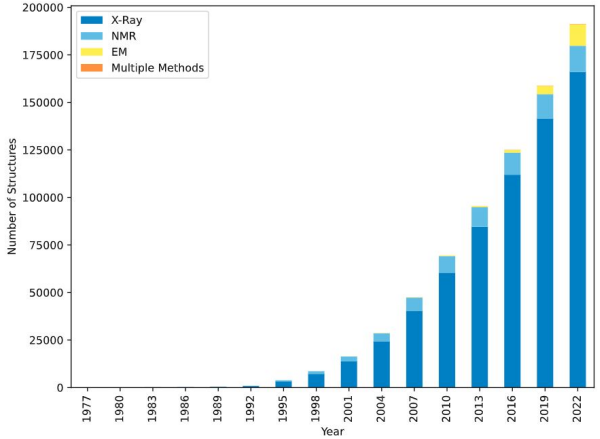
Example structure of MDM2 - Benzodiazepine complex (PDB 1t4e)

Structural data is available



- Structural data can be obtained from experimental and computational methods
  - X-Ray, Cryo-EM, NMR…
  - Gathered in a database

  … and in-silico approaches
  - Alphafold-{1,2,multimer}, ESMFold, OmegaFold…



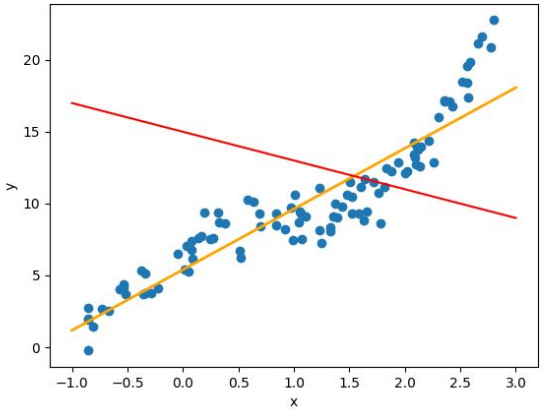
Evolution of the number of available biomolecules structures (Source : RCSPDB)

Machine learning (quick)

- Algorithms for which **performance increases with data**

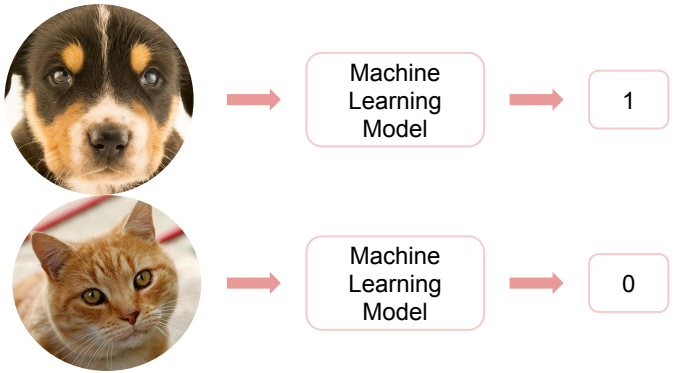
$$f : \mathbb{R} \to \mathbb{R}$$
$$x \to f(x) \sim f_\theta(x) = \theta_1 x + \theta_0$$

- Let's use machine learning to solve an example task $\mathcal{T}_1$: classify pictures of dog (1) vs cat (0), get a metric on a test set (accuracy for instance)



Toy example : blue is data points, orange is the best model, red is another random one

$\mathcal{T}_1$



Representation

- Our object is a vector / list of numbers (pixels values)
- Perform linear regression

=

12	98	231	101	253	57	0	132
251	78	43	23	156	99	109	126
136	44	208	122	237	19	252	211
99	133	4	146	135	231	13	134
225	233	137	68	127	131	93	254
241	129	178	234	14	250	6	237
185	255	196	118	0	198	34	235
0	213	251	11	129	192	118	212

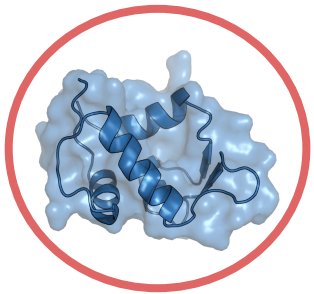
$\mathcal{T}_1$

$(x_1, x_2, \ldots x_n) \in \mathbb{R}^k$ → Linear regression → Metric : 0.84

Representation

- Turn object into a (*feature*) vector :
  - Weight, size, number of amino-acids…
- It's a bottleneck to **represent** complex objects as vectors

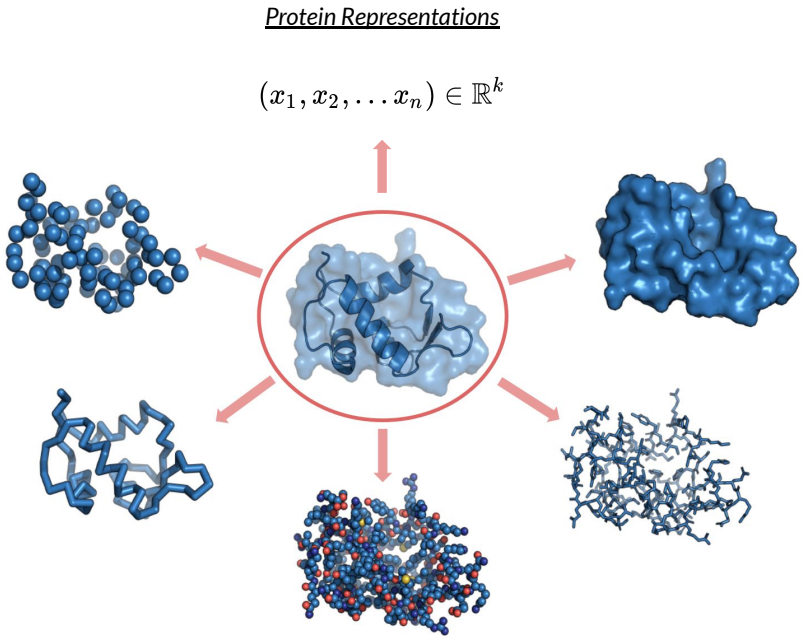
$\mathcal{T}_1$



Representation → $(x_1, x_2, \ldots x_n) \in \mathbb{R}^k$ → Linear regression → Metric : 0.83

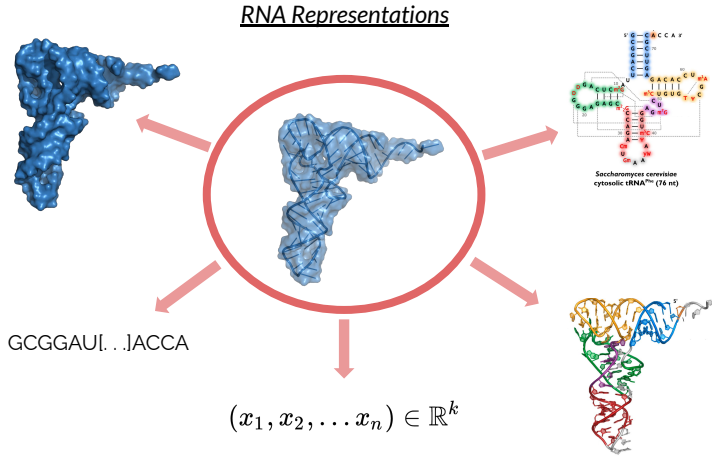
Modeling of a biomolecule

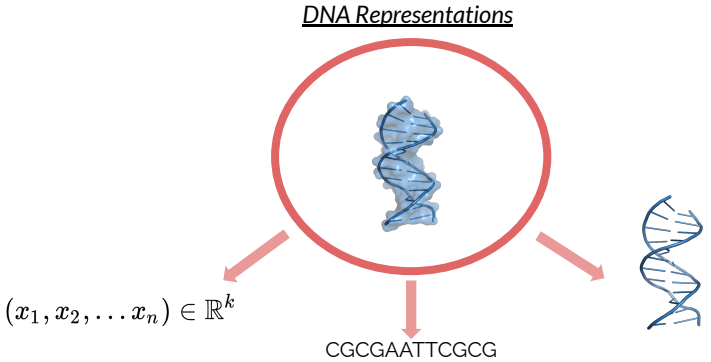
- We can model our biomolecule with more than a vector :
  - Point cloud, graph, surface…



Modeling of a biomolecule

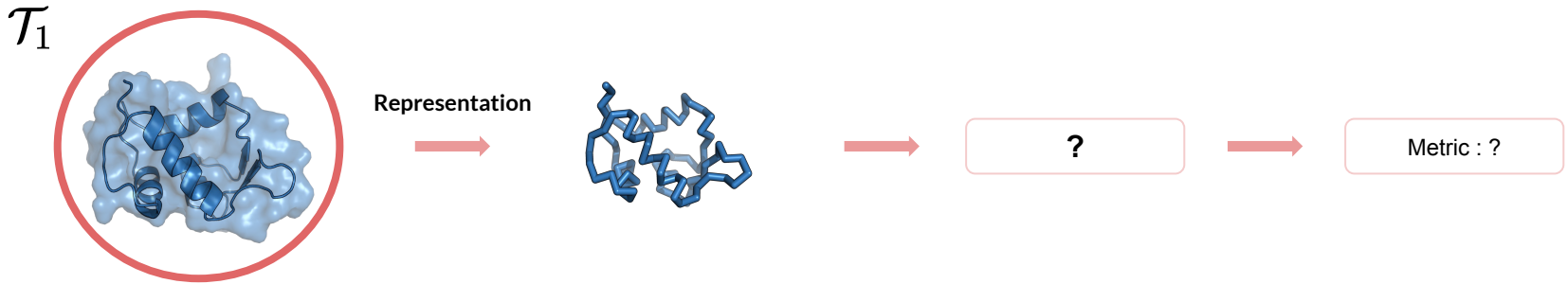
- Different models are relevant for proteins, RNA or DNA
- We only **model** our object, as a mathematical, numerical object





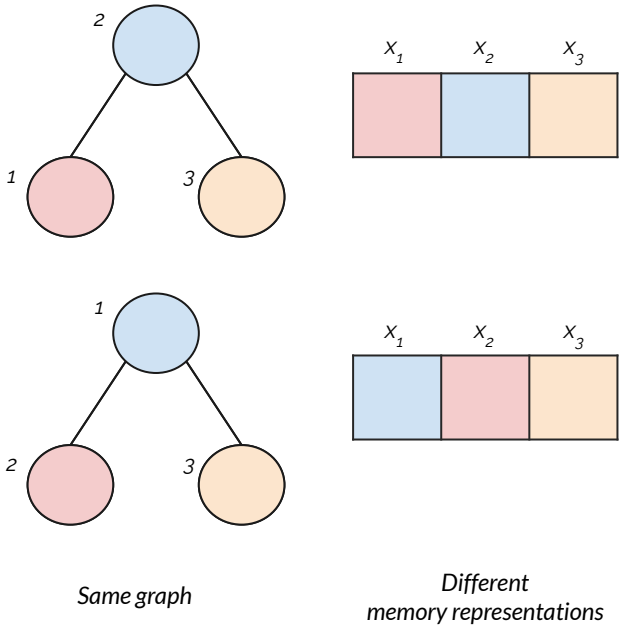
Learning beyond vectors

- If we represent our object as a **graph**, can we perform linear regression on it ?
- Can we learn on objects with mathematical structure ?



Learning on complex objects : The example of graphs

- The representation in a computer is arbitrary : we create structure
  - There is a **permutation symmetry**

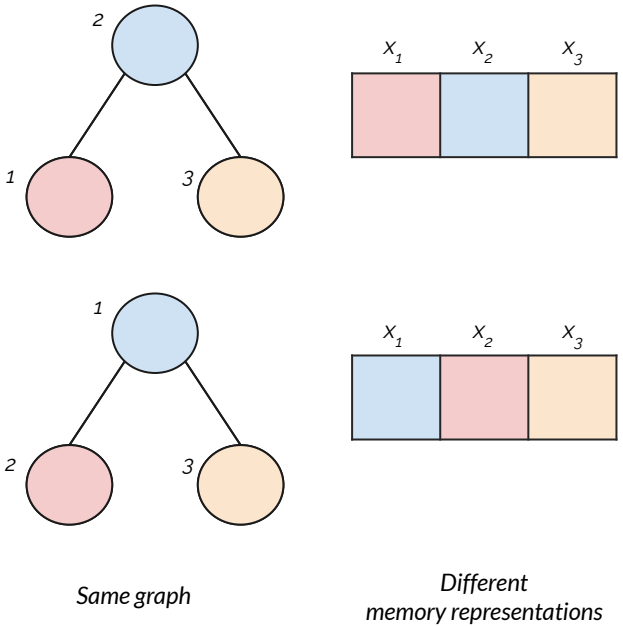


Learning on complex objects : The example of graphs

- The representation in a computer is arbitrary : we create structure
  - There is a **permutation symmetry**

- Order is important for linear regression

$$f : \mathbb{R} \to \mathbb{R}$$
$$x \to f(x) \sim f_\theta(x) = \theta_3 x_3 + \theta_2 x_2 + \theta_1 x_1 + \theta_0$$



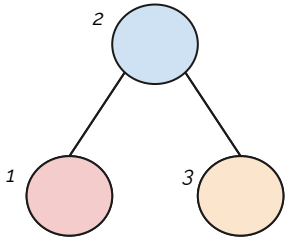
Same graph

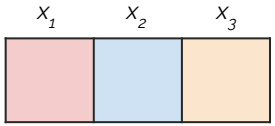
Different memory representations

Learning on complex objects : The example of graphs

- The representation in a computer is arbitrary : we create structure
  - There is a **permutation symmetry**

- The underlying data is structured :
  - There is a **connectivity**



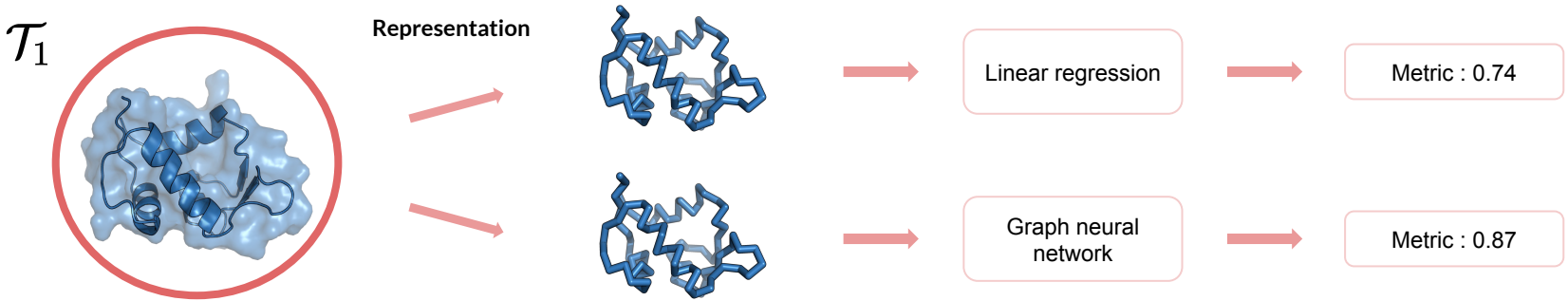


0	1	0
1	0	1
0	1	0

Geometric deep learning aims to **respect these mathematical properties** when dealing with our data !

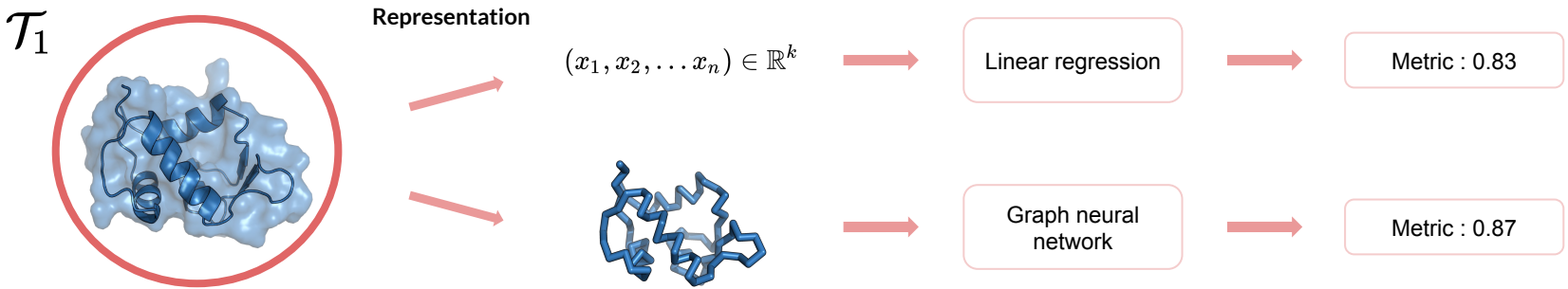
Learning beyond vectors

- Graph neural networks respect those properties and enable learning on graphs
  - They often yield better results !



Learning on biomolecules

- The dual choice of a representation and a learning method underpins a successful learning on biomolecules

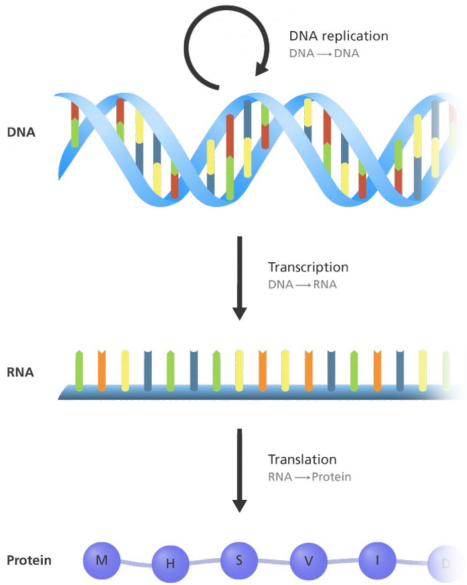


RNA representation as 2.5D graphs

Augmented base pairing networks encode RNA-small molecule binding preferences. Oliver, Mallet et al., NAR, 2020
VeRNAl: A Tool for Mining Fuzzy Network Motifs in RNA. Oliver, Mallet* et al., Bioinformatics, 2022*
RNAglib: A python package for RNA 2.5D graphs. Mallet, Oliver*, Broadbent* et al, Bioinformatics Application Notes, 2022*

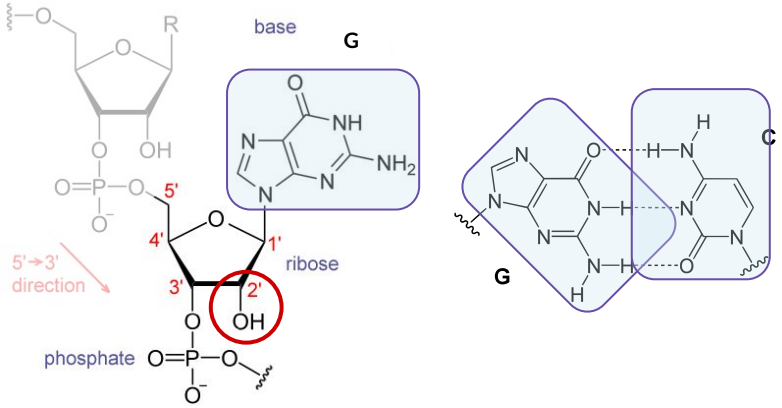
What is RNA ?

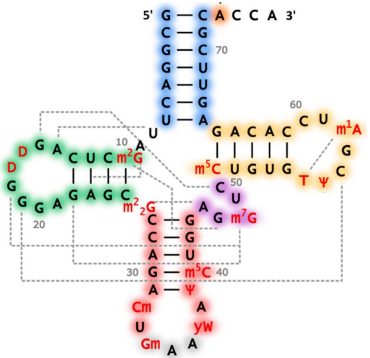
- In between DNA and proteins as a messenger
- Single stranded unlike DNA :
  - allows for complex secondary structures
- Less hydrophobic than protein
  - secondary structure is more prevalent than tertiary structure



RNA 2D graph

- Polymer of nucleotides
- Pairwise interactions form 2D graph
  - Bases are nodes
  - Interactions are edges (in addition to backbone)

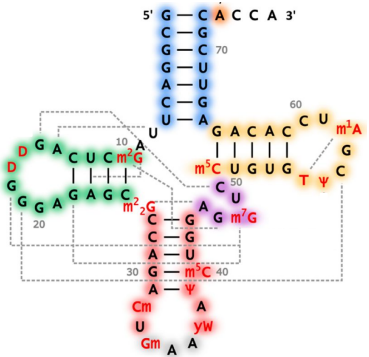




Saccharomyces cerevisiae
cytosolic tRNA^{Phe} (76 nt)

RNA 3D

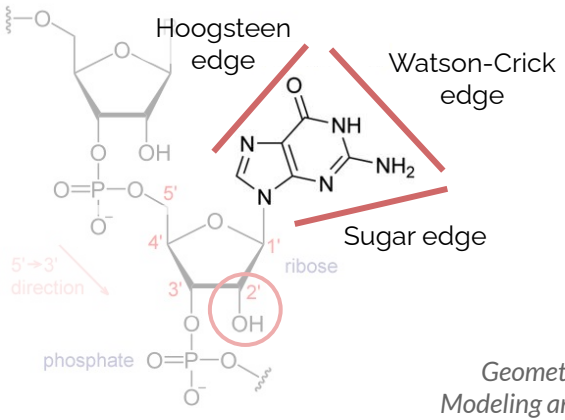
- This 2D structure conditions the 3D structure
- Some information is missing

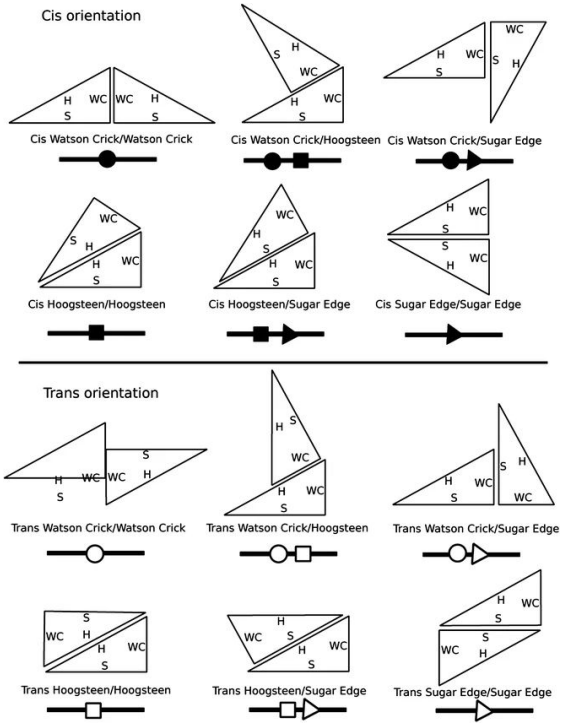


Saccharomyces cerevisiae cytosolic $tRNA^{Phe}$ (76 nt)

2.5D RNA Graphs

- There are other possible interactions !
  - 12 without edge direction
  - 17 with edge direction



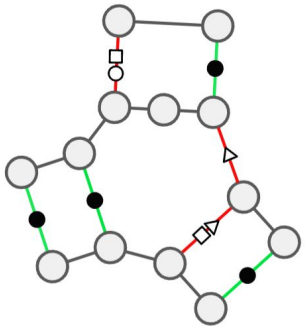


Geometric nomenclature and classification of RNA base pairs, Leontis and Westhof (2001)
Modeling and Predicting RNA Three-Dimensional Structures, Waldispühl and Reinharz (2015)

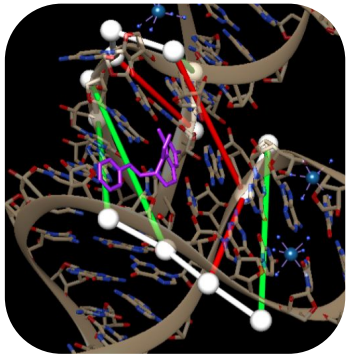
2.5D RNA Graphs

- New interactions => new graph
  - 12 edges types if undirected
  - 17 edges types if directed with a symmetry on certain edges

- These graphs are a finer grained depiction of the 3D



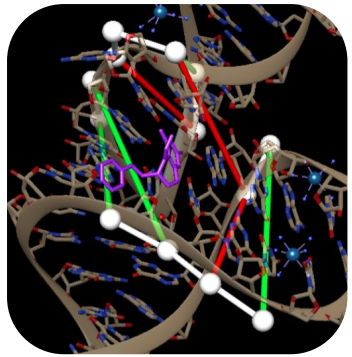
(b) Graph encoding of binding site as an augmented base pairing network (ABPN).



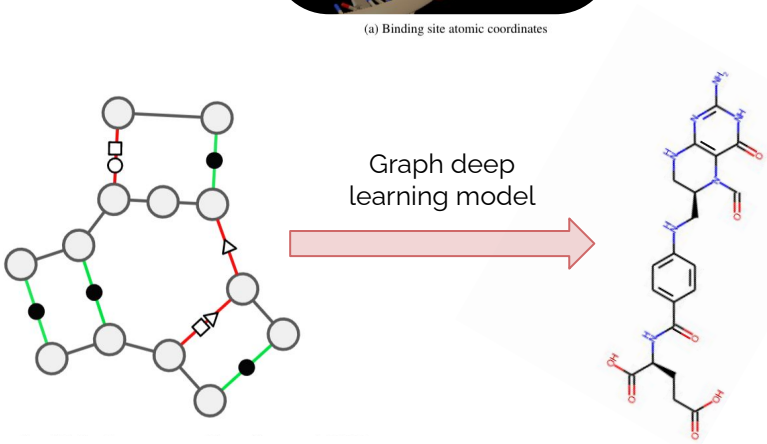
(a) Binding site atomic coordinates

RNAmigos

- Drug discovery task : given a pocket, predict its ligand
- Data : all available RNA-ligand 3D data from the PDB (~800 data points)



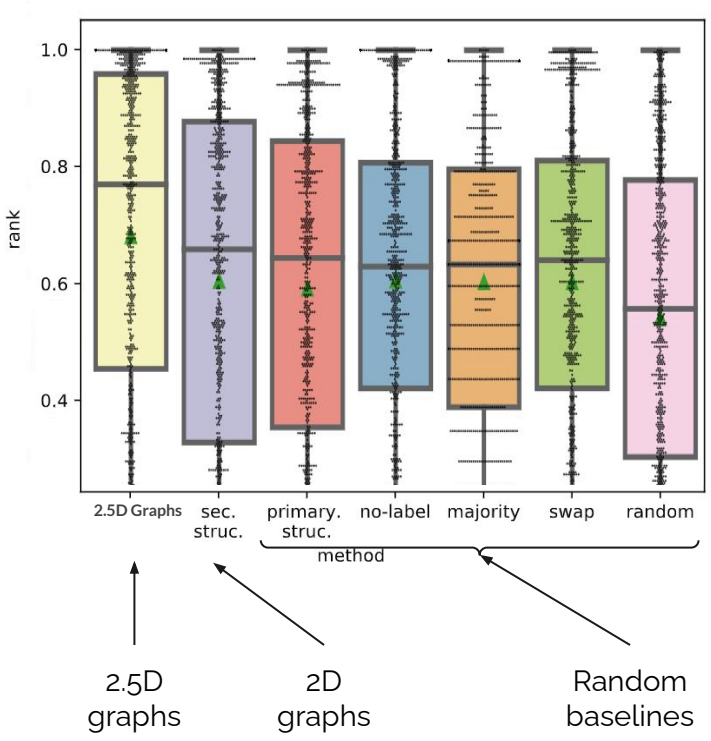
(a) Binding site atomic coordinates



Augmented base pairing networks encode RNA-small molecule binding preferences. Oliver, Mallet et al., NAR, 2020

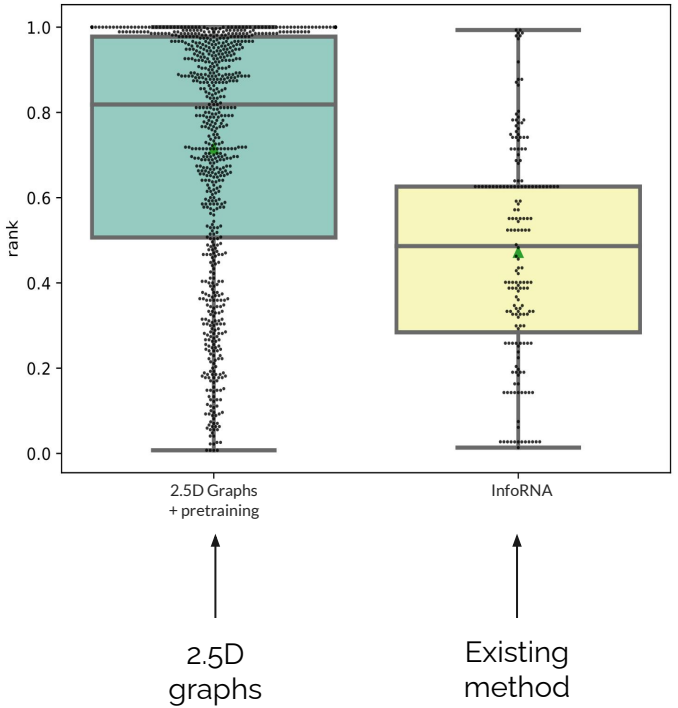
2.5D graphs are relevant

- Using only 2D graphs is comparable to randomized baselines *(p-value = 0.07)*
- Using RNA 2.5D graphs performs significantly better than 2D graphs or the baselines *(p-value = 10^{-11})*



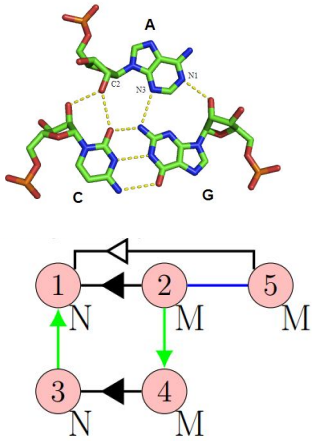
Drug design result

- We have better performance than this tool
- Really the beginning of RNA drug design
- A more in-depth study of the drug design aspect is under construction



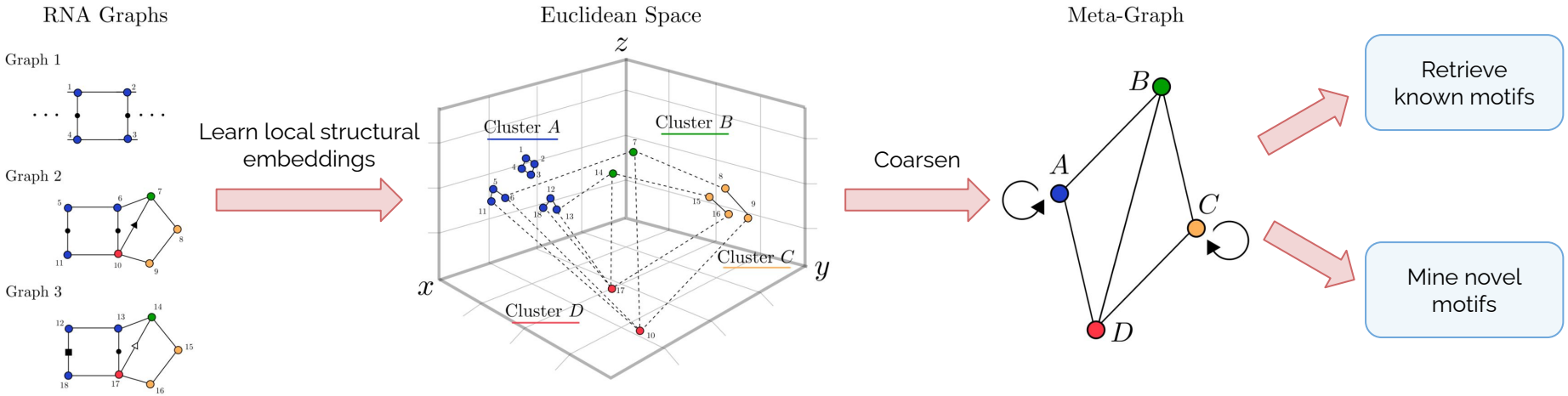
Motifs : recurrent 3D substructures

- Motifs are recurrent 3D structural patterns
  - Roughly subgraphs with similar (or identical) structure that involve non-canonical interactions
- Motifs are functional subunits
  - Enriched at binding sites
  - Useful for structure prediction



The A-minor motif, in 3D and represented as a 2.5D graph

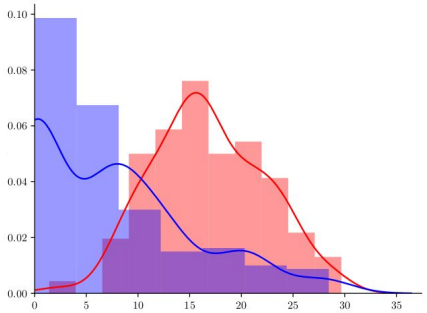
VeRNAl pipeline



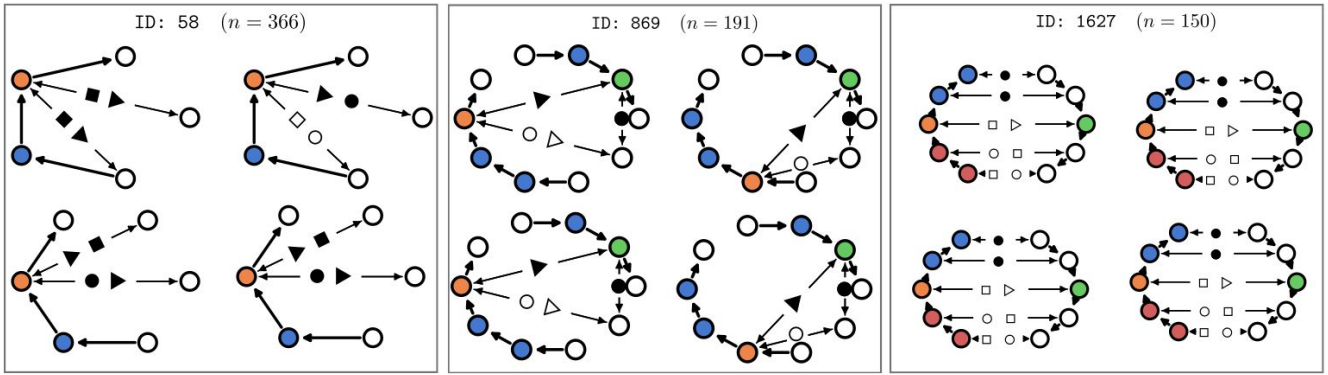
VeRNAl: A Tool for Mining Fuzzy Network Motifs in RNA. Oliver, Mallet* et al., Bioinformatics, 2022*

VeRNAl discovers new motifs

- *VeRNAl* motifs are visually relevant and have low intra-GED
- *VeRNAl* motifs align with existing motifs



Distribution of GED values within (blue) and across (red) motifs

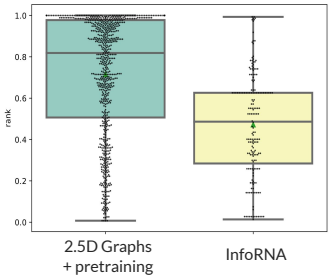


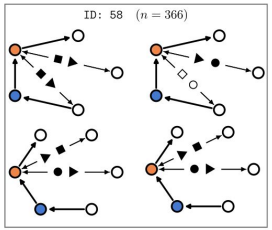
Four instances from three random VeRNAl motifs

Conclusion

2.5D graphs are an efficient representation for learning on RNA

- We successfully used them in drug discovery pipelines
- We successfully used them for motif mining
- We released a pip package (RNAGlib) to promote their use

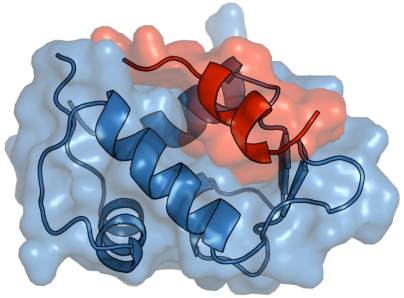




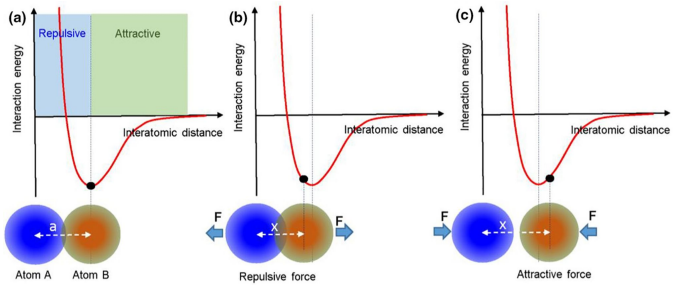
Protein representation as surfaces and beyond

Surface is appealing

- Physics (hence *function*) depend on relative positions
- For interactions, there is **a screening effect** *(2^6 = 64)*
- Going from 3d scaling to 2d scaling



Example structure of MDM2 - p53 complex (PDB code 1ycr)



Atomic potential as a function of distance. This is Leonard Jones with a decrease a D^{-6}

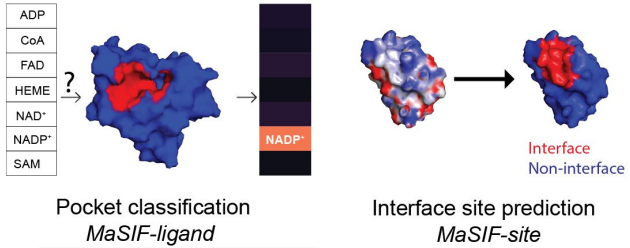
Surface methods are booming

- Plotting, comparison functions are available
- Niche five years ago, now much better results

Method	Accuracy
GWCNN [Ezuz et al. 2017]	90.3%
MeshCNN† [Hanocka et al. 2019]	91.0%
HSN† [Wiersma et al. 2020]	96.1%
MeshWalker† [Lahav and Tal 2020]	97.1%
PD-MeshNet† [Milano et al. 2020]	99.1%
HodgeNet† [Smirnov and Solomon 2021]	94.7%
FC† [Mitchel et al. 2021]	99.2%
DiffusionNet - xyz†	99.4%
DiffusionNet - xyz	99.0%
DiffusionNet - hks†	99.5%
DiffusionNet - hks	**99.7%**

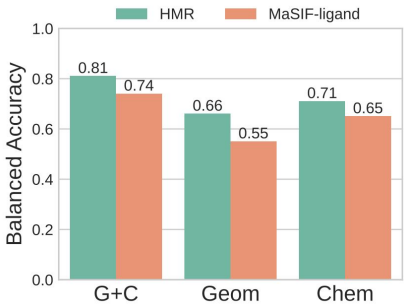
Existing applied methods

- Pioneer work of MaSIF : using GCNN
- Very recent publication applies DiffusionNet to proteins with success



Pocket classification
MaSIF-ligand

Interface site prediction
MaSIF-site

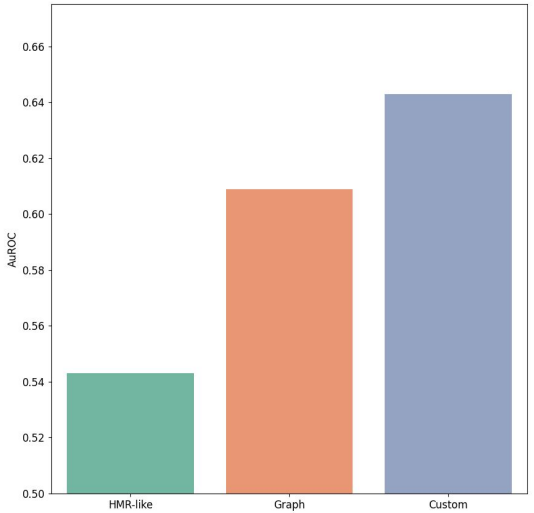


DiffusionNet gives better results than original MaSIF

Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning, Gainza et al., Nature Methods, 2019
Learning Harmonic Molecular Representations On Riemannian Manifold, Wang et al., ICLR, 2023

Work in progress

- Make a stronger assessment of the relevance of surface representation, with benchmarks
- Explore other ways to use the surface

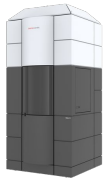


Custom architectures are promising on the benchmark task of mutation stability prediction

ATOM3D : Tasks On Molecules in Three Dimensions, Townshend et al., Neurips, 2021

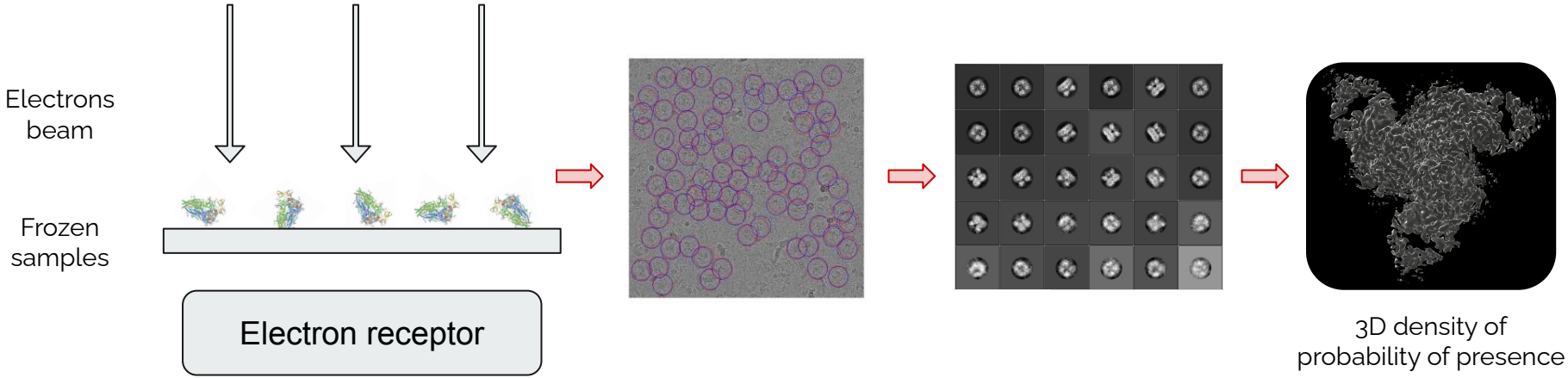
Cryo-EM and antibodies

Structure and Cryo-EM



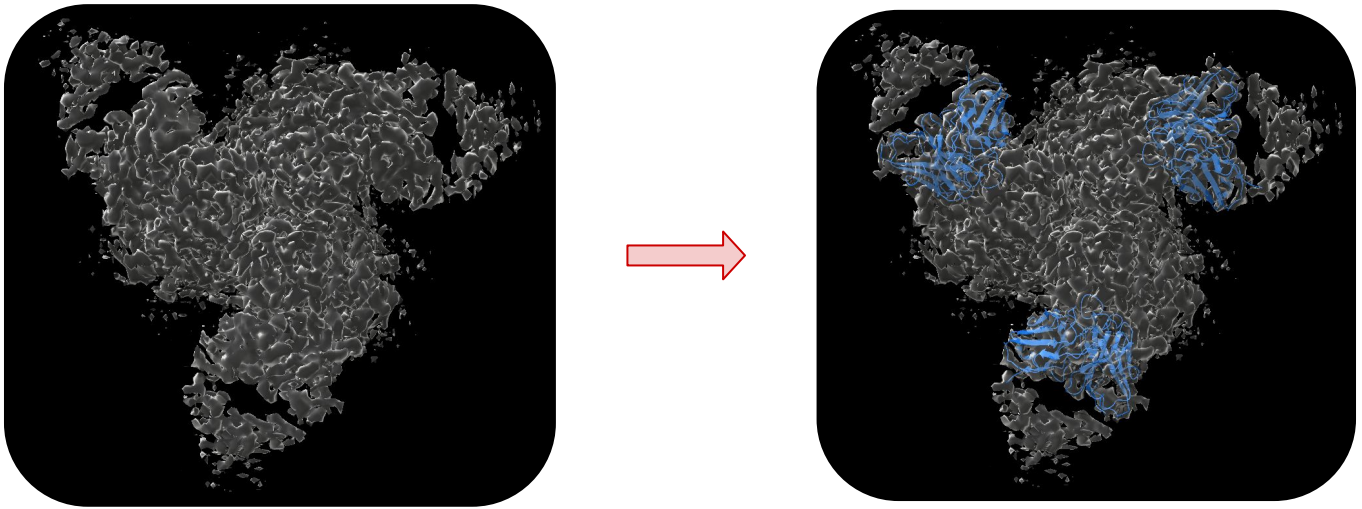
Titan Krios
cryo-EM

- Cryo-EM is a way to get the structure (Nobel prize 2017)



3D density of
probability of presence

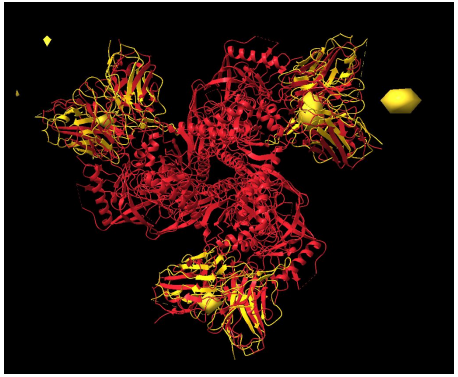
Antibody detection in low resolution cryo-EM maps

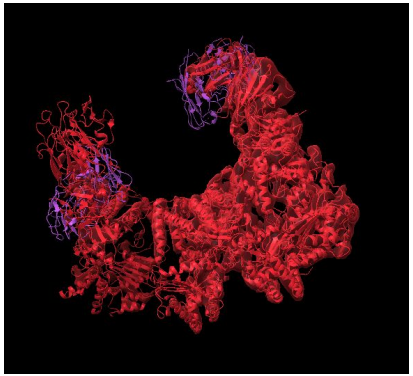


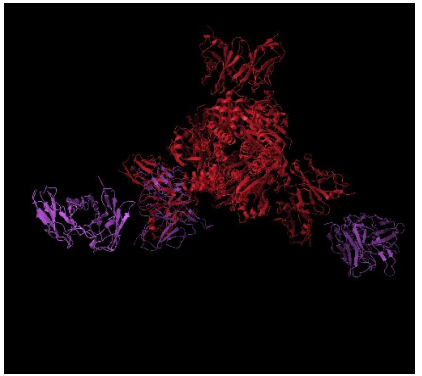
Preliminary results and challenges

- Works well on some examples, ok in some others

- Challenging optimal transport loss (Keops ?)







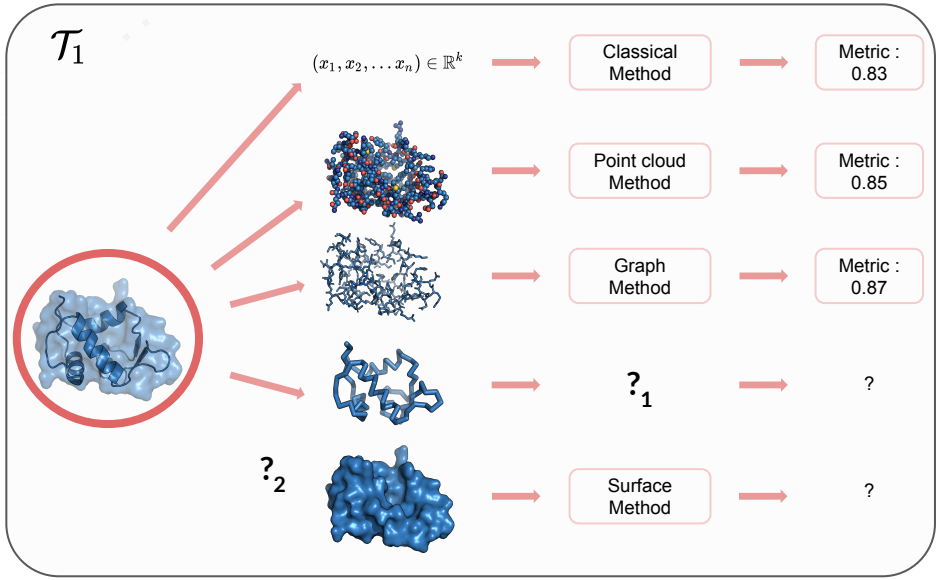
Conclusion

Conclusion

Promising structural biology results using machine learning. This is made possible with coordinated development of :

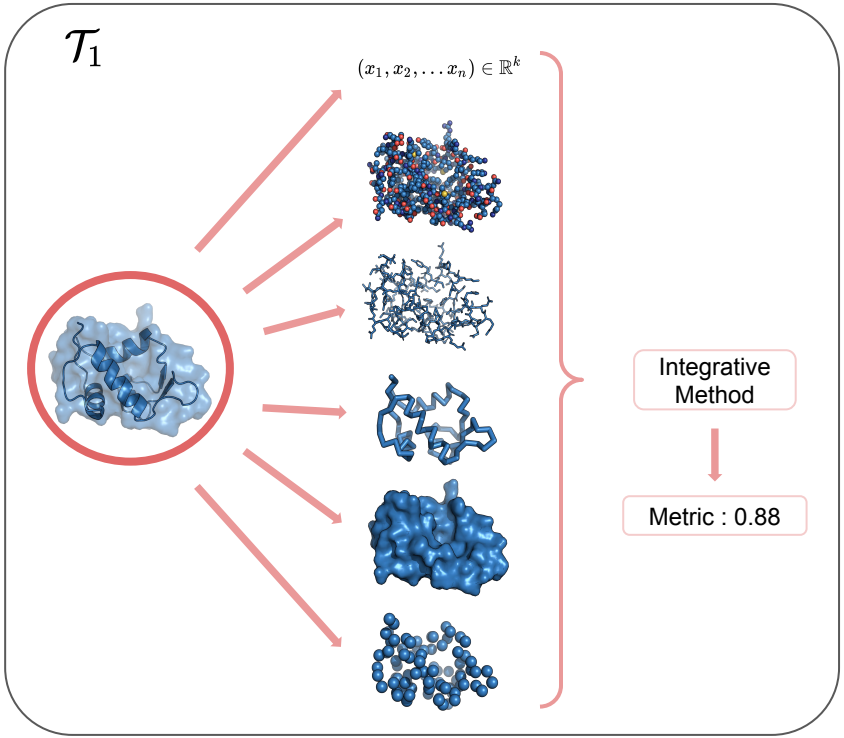
1. Representations of the structure of biomolecules
2. Geometric learning methods that respect the representations properties

…this is still very underdeveloped



Better representations

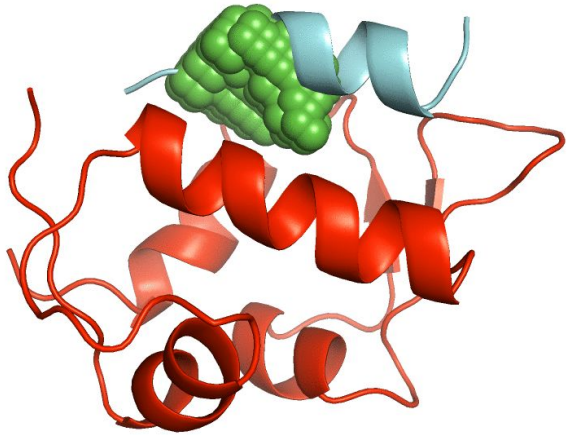
- Integrative approach with several representations at the same time
- Pre-training schemes are promising



Protein Representation Learning By Geometric Structure Pretraining, Zhang et al., ICLR, 2023

Molecules flexibility and dynamics

- Biomolecules are dynamic objects
- Their properties depend on the whole conformational ensemble
- We should use this as a representation

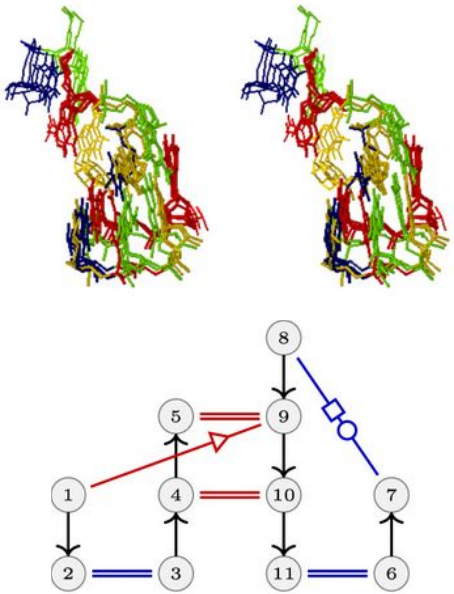


Thanks for your attention !

Questions ?

Existing motif mining tools

- Problem is NP hard : approximated by finding maximum common subgraphs (MCS) on all pairs of graphs[(1)]
  - Very slow
  - Only exact matches
- MCS wastes a lot of time on useless pairs
- MCS misses flexibility, that is important biologically

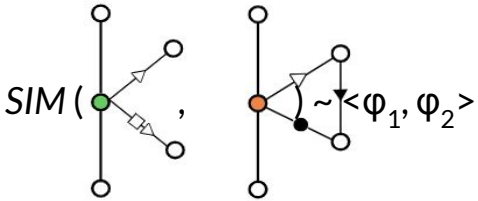


A more complex motif and all its aligned 3D instances

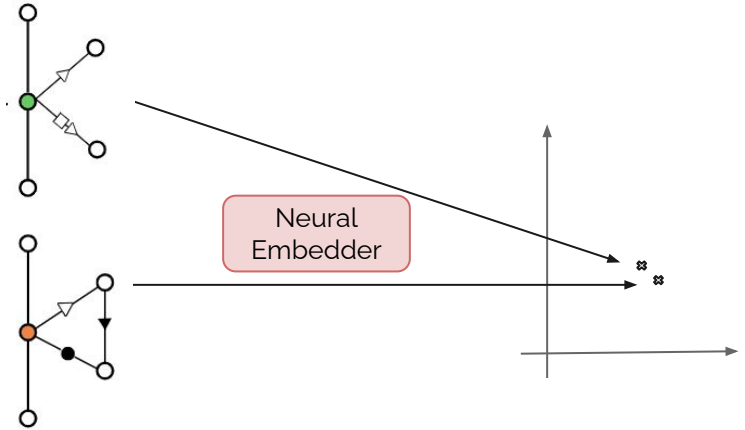
(1) Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families, Reinharz et al. (2021)

Substructure fast comparison

- Approximate a structural comparison *SIM* with dot product of learnt structural embeddings

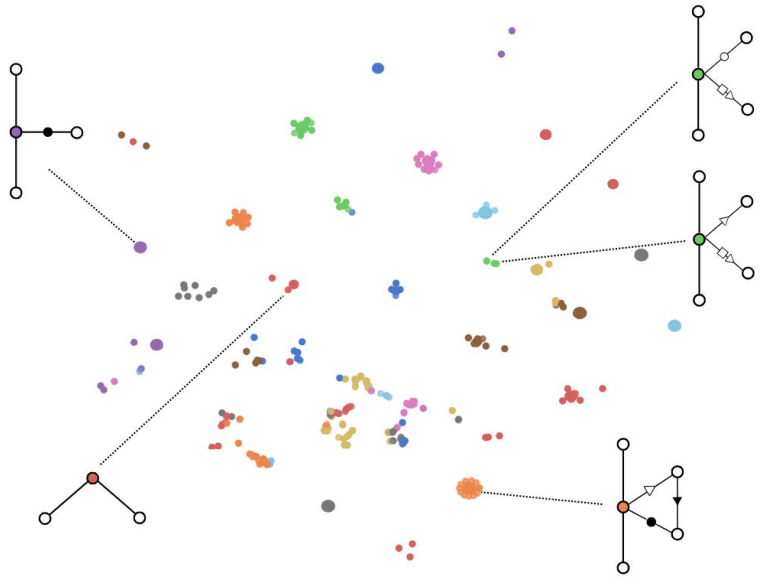


- Here, *SIM* is a custom RNA Graph Edit Distance (GED)



Fuzzy clusters and limitation

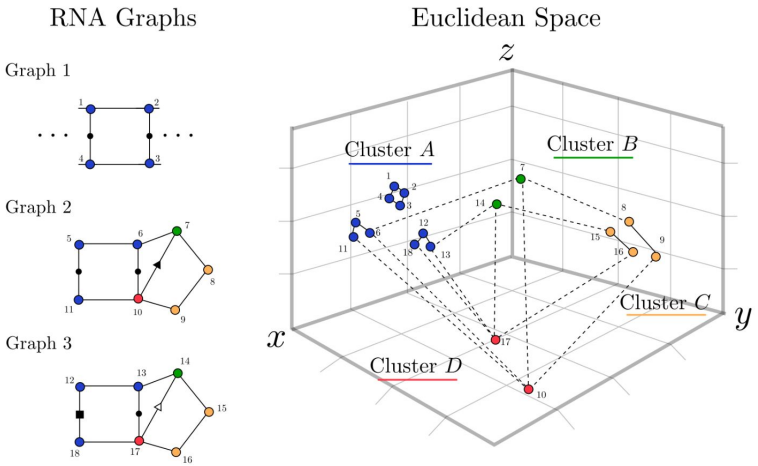
- *Quasi isomorphic subgraphs (**fuzzy**)* are neighbors
- Only rooted subgraphs of fixed size
  => How to go beyond that ?



T-SNE visualisation of the latent representations of RNA bases

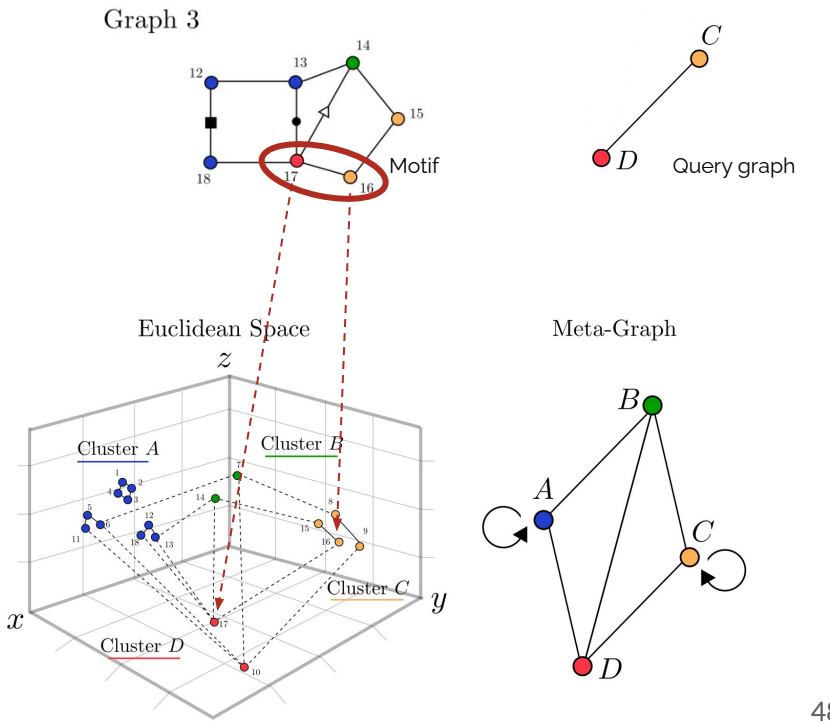
Meta-graph

- Clusters are meta-nodes
- Connections in original graphs are meta-edges
- Close neighbors meta-nodes are *frequently co-occurring adjacent subgraphs*



Motif retrieval - Example

- Example *motif* = {16,17} in Graph 3
- It corresponds to a query with one meta-edge : DC

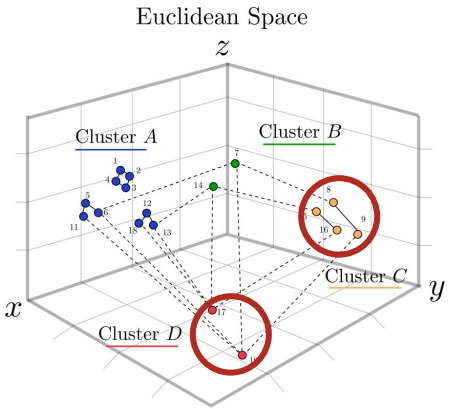


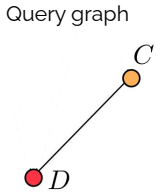
Motif retrieval - Example

1. Start with all nodes in the same clusters (partial hits)

D C

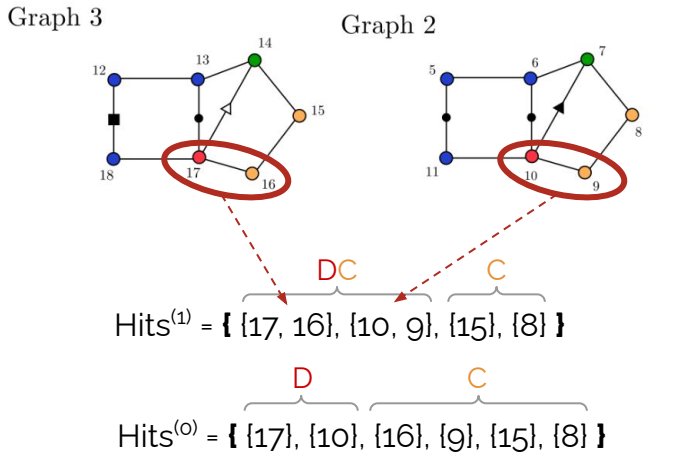
$Hits^{(0)}$ = { {17}, {10}, {16}, {9}, {15}, {8} }

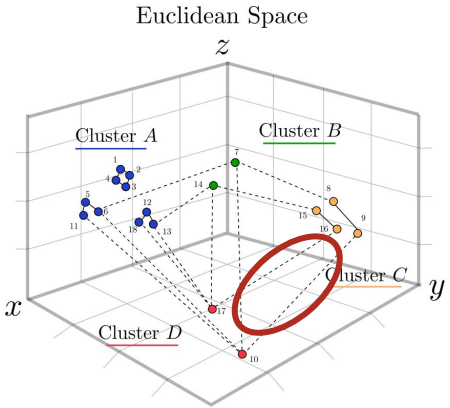


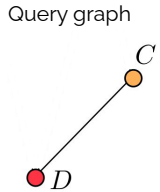


Motif retrieval - Example

1. Start with all nodes in the same clusters (partial hits)
2. Loop over the edges of this query and merge hits that are linked



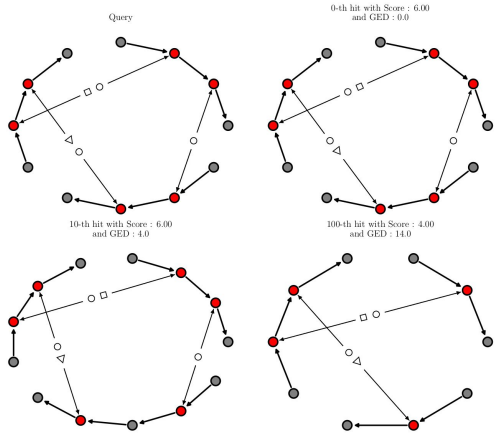




VeRNAl retrieves motifs

- We inspect the hits list for a given query visually and by computing the GED
  - Best hits have low GED
- We compare to three RNA motif mining tools
  - We find most of them
  - We expand them with quasi-isomorphic instances

Rank	1^{st}	10^{th}	100^{th}	1000^{th}	Decoy
Mean GED	3.1 ± 0.3	3.9 ± 0.4	6.2 ± 0.6	9.2 ± 0.8	14.4 ± 0.8



Dataset	Covered	Missed
BGSU Petrov et al. [2013b]	112	14
RNA3DMotif Djelloul [2009]	2	0
CaRNAval Reinharz et al. [2018a]	147	10