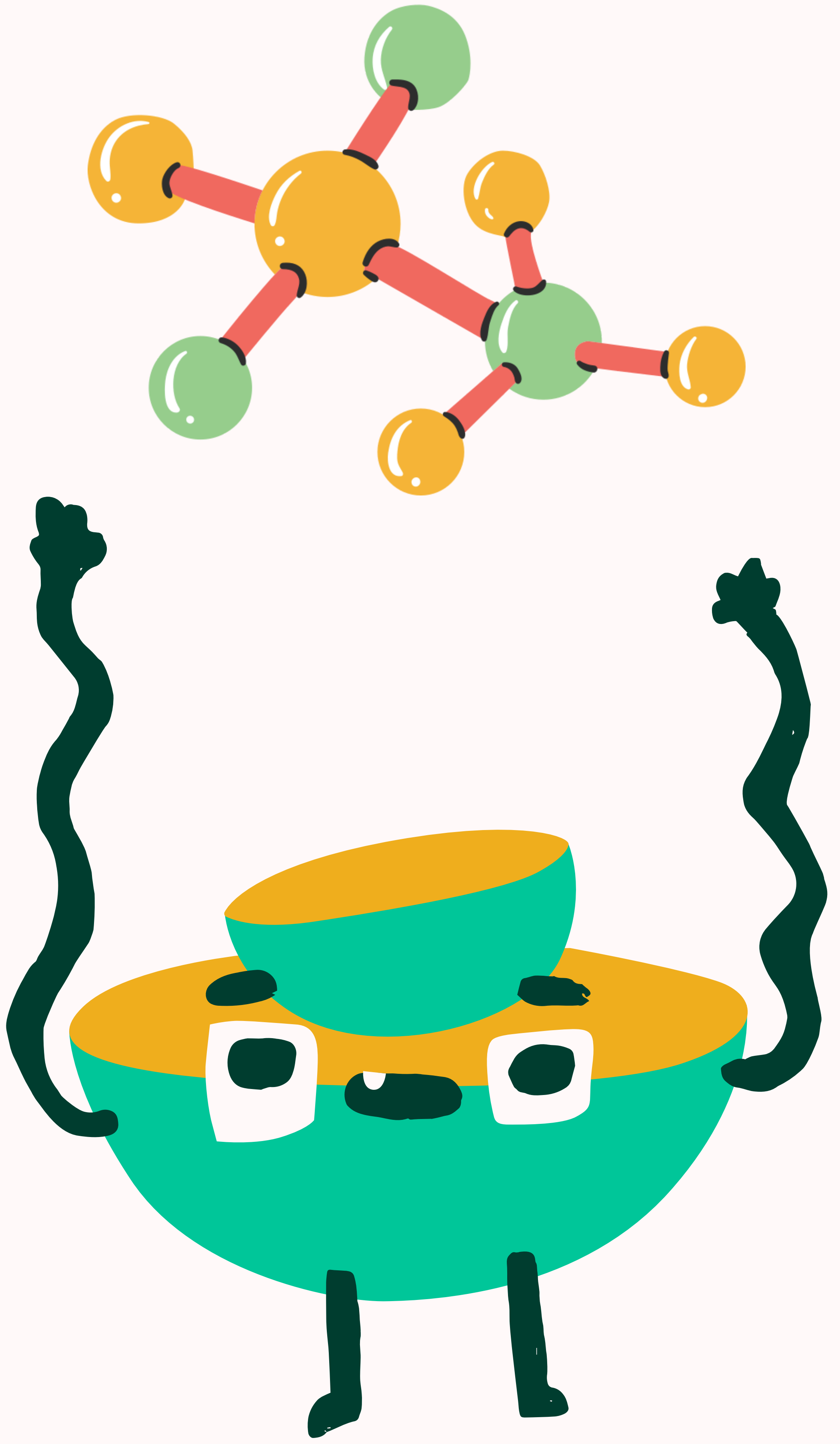
Molecule Design Using Variational Autoencoders



Yassaman Ommi (9613005)

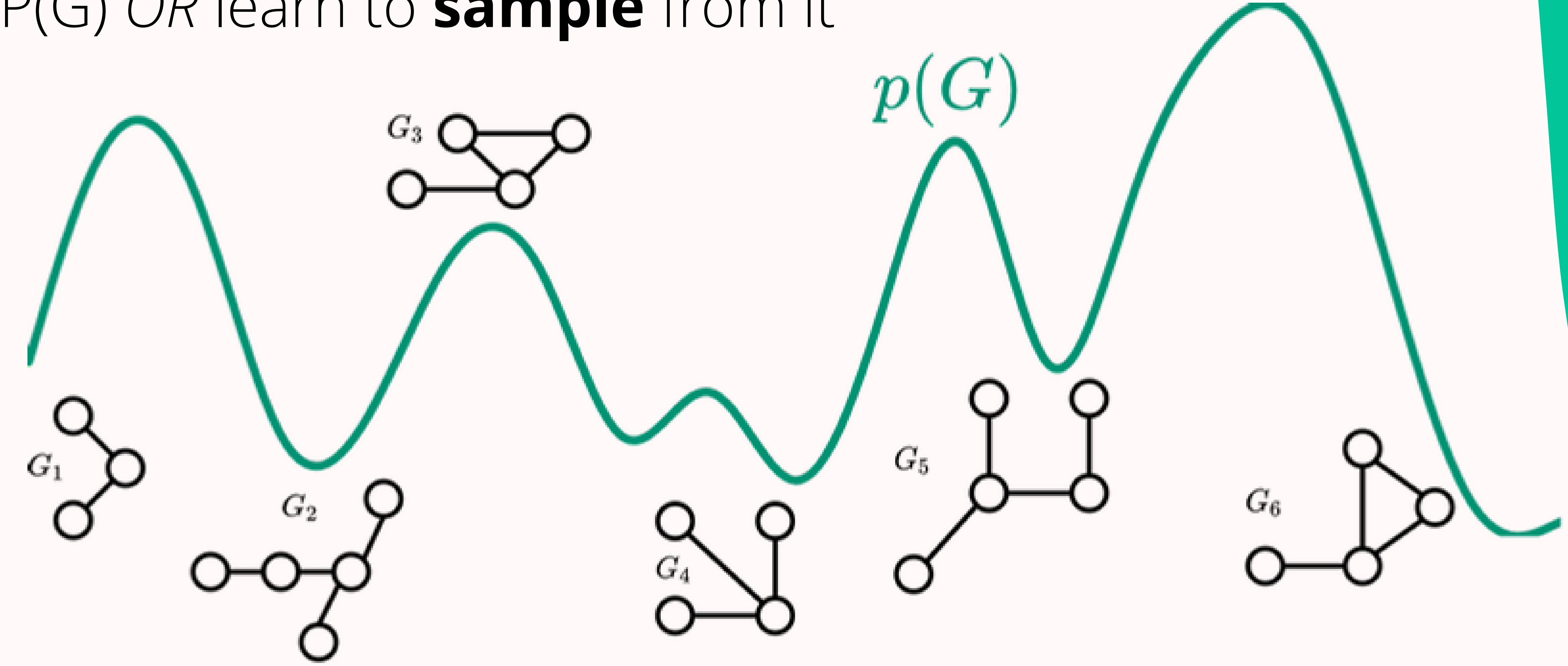
Supervisor: Dr. Amin Gheibi

OUTLINE

- Problem Formulation & Applications
- Related Works
- Proposed Method
- Experiments & Results

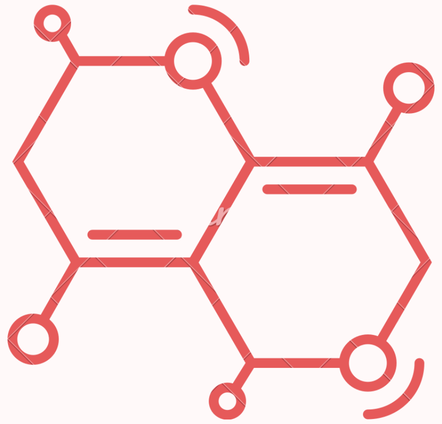
PROBLEM FORMULATION & APPLICATIONS (1/2)

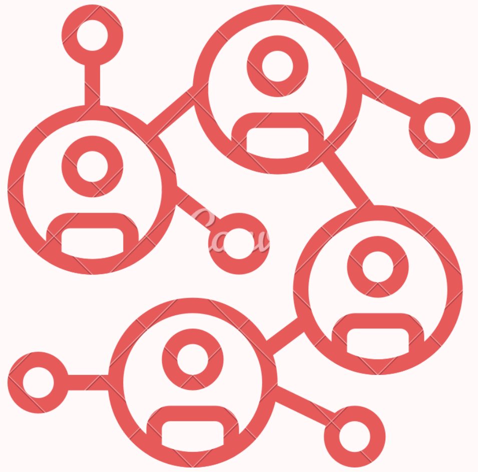
- A rich **history**
- A set of observed graphs **G** with underlying distribution **P(G)**
- Train a model to **estimate** P(G) *OR* learn to **sample** from it



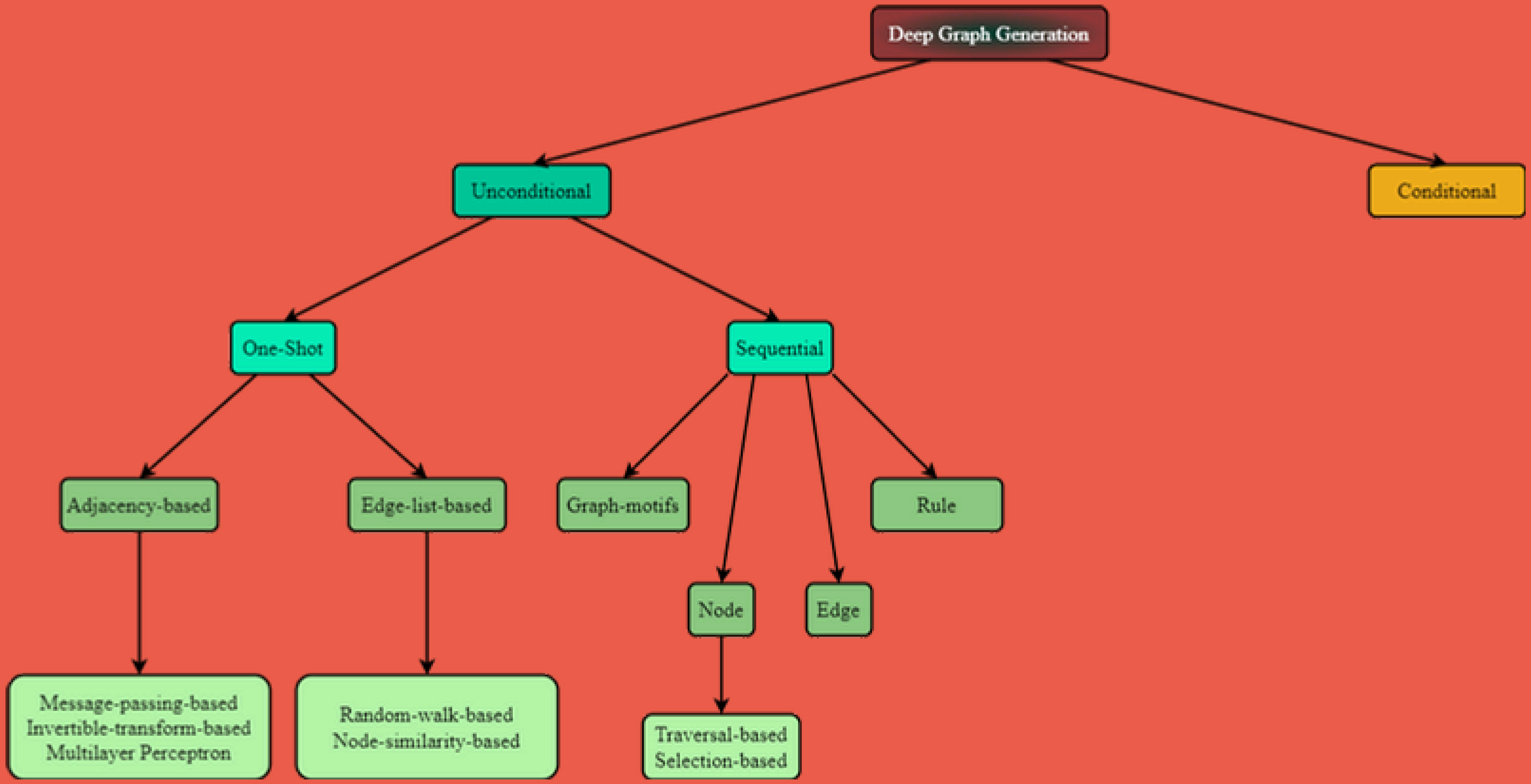
PROBLEM FORMULATION & APPLICATIONS (2/2)

- Molecular graph generation (drug design / material discovery ...)
- Computational social sciences
- Network science (food webs / epidemics ...)
- Semantic parsing
- ...



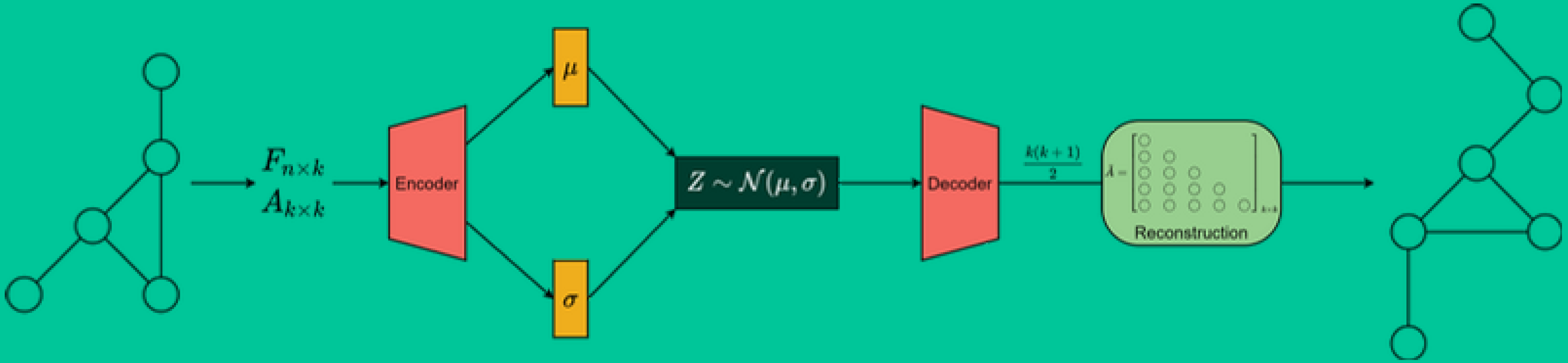


RELATED WORKS



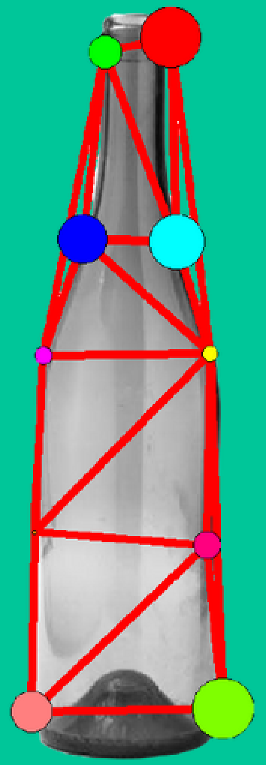
ALAJVAE (1/3)

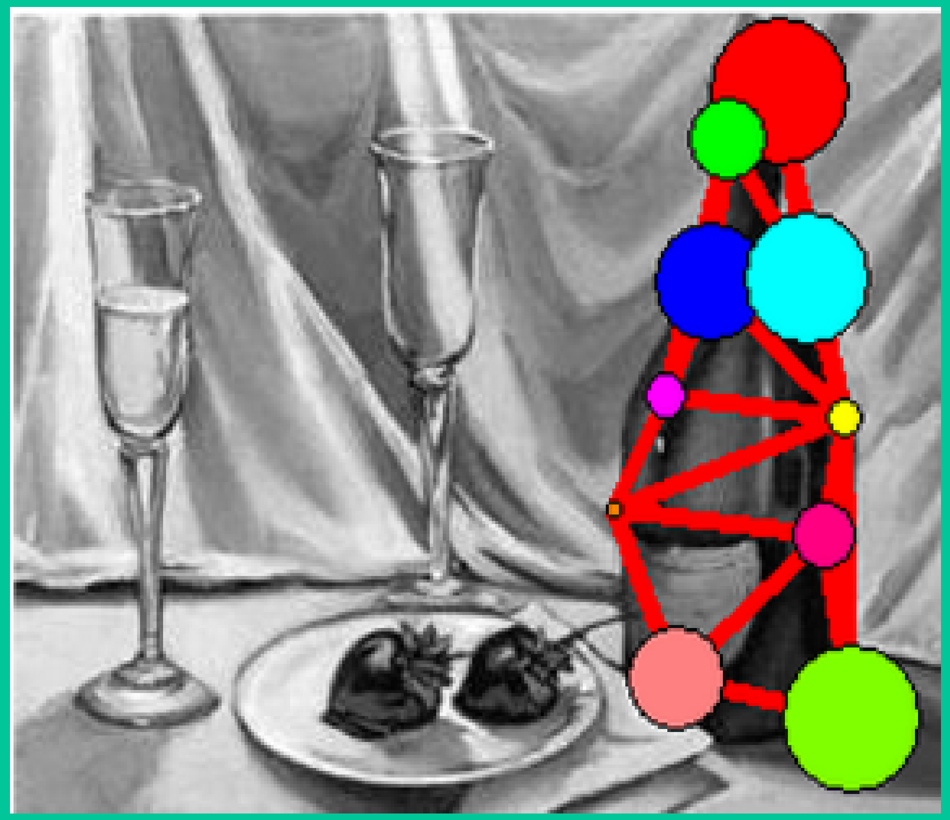
- **VAE**-based generative model
- Input: **Padded** graph's **adjacency** matrix (k nodes) and feature matrix (n features)
- Output: **Probabilistic** adjacency matrix (edge and node existence)

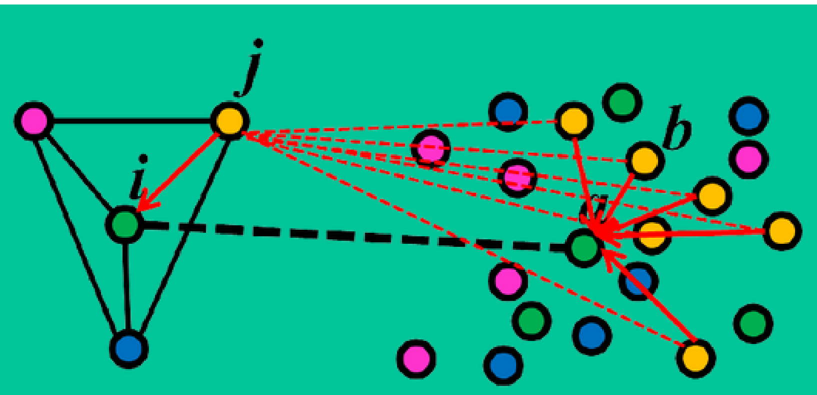


ALAJVAE (2/3)

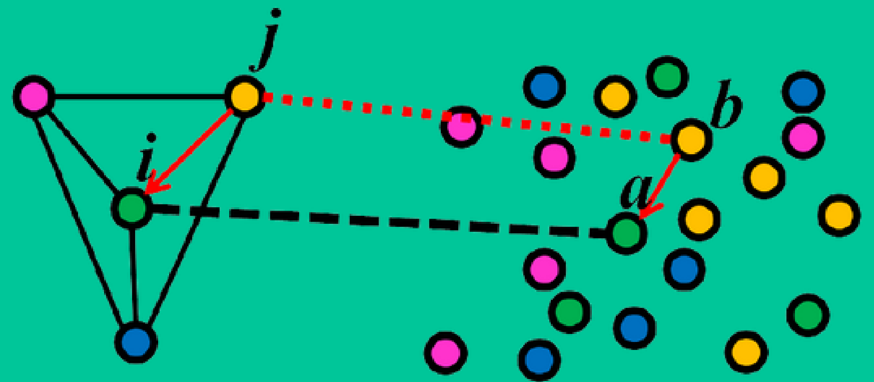
- No specific **node ordering** causes trouble
- Use graph **matching** to align and **compare** output with ground truth
- Find correspondence between a **reference** and a test **scene**
- **Applications** in shape matching, object recognition, and ...



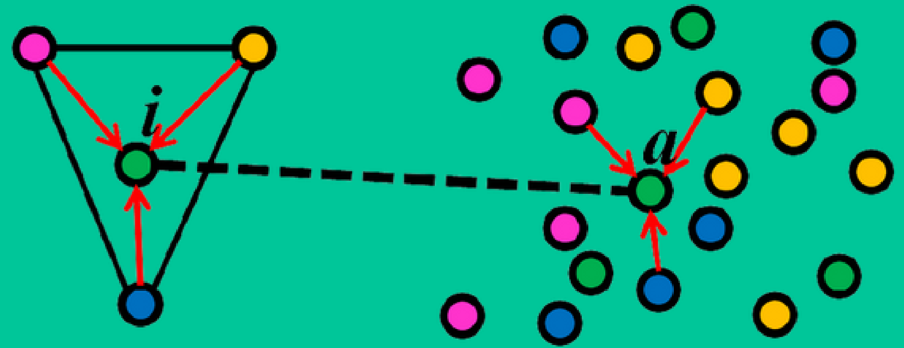




Sum-pooled support from node j to a match (i, a)



Max-pooled support from node j to a match (i, a)



All max-pooled supports to a match (i, a)

ALAJVAE (3/3)

- **Max-pool** matching iterative power method [Cho. 2014]
- Proposed **similarity** function

$$S : (i, j) \times (a, b) \rightarrow \mathbb{R}^{+} \text{ for } i, j \in V \text{ and } a, b \in V'$$

$$\begin{aligned} S((i, j), (a, b)) &= \\ &= \frac{1}{|F_i - \tilde{F}_a| + 1} \cdot \tilde{A}_{a,a}[i = j \wedge a = b] && \longleftarrow \textbf{node} \text{ similarity} \\ &+ A_{i,j} \cdot \tilde{A}_{a,b} \cdot \tilde{A}_{a,a} \cdot \tilde{A}_{b,b}[i \neq j \wedge a \neq b] && \longleftarrow \textbf{edge} \text{ similarity} \end{aligned}$$

EXPERIMENTS AND RESULTS (1/2)

- **General Metrics**

The absolute difference between the generated samples and the dataset is measured

1. *Graph-based statistics*: Node Degree Distribution, Clustering Coefficient, Largest Connected Component, ...
2. *Graph-generation metrics: Uniqueness, Novelty, Validity, ...*

- **Application-based Metrics**

1. *Chemistry-based:* Quantitative Estimate of Drug-likeness (QED), Synthetic Accessibility (SA), Molecular Weight (MW), ...

EXPERIMENTS AND RESULTS (2/2)

- Dataset

Name	Domain	Size	|V|	|E|	|**V**|	|**E**|
ENZYMES	Protein	575	[2, 125]	[2, 149]	3	-

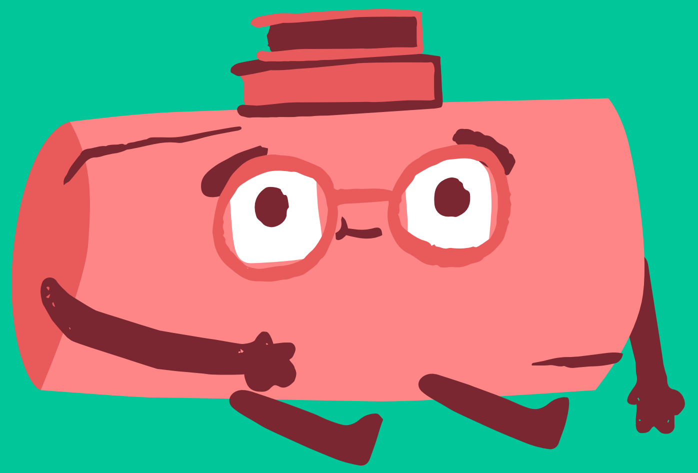
- Results

Dataset	Properties	LCC	Clustering Coef.	Mean Deg.	Gini Coef.	Novelty@100	Uniq@100	Training Time	epochs
ENZYMES	**AlajVAE**	**0.32**	0.206	0.405	**0.0063**	**99%**	98%	*30h*	*30*
	GraphGen	-	0.198	0.243	-	98%	**99%**	3h	4000
	GraphRNN	-	**0.151**	**0.090**	-	**99%**	97%	15h	20900

CONCLUSION & FUTURE DIRECTION

- Proposed **probablistic** method suitable for **small** graphs
- Growth of **GPU** memory requirements
- High **complexity** of the matching algorithm
- Adding atom and bond **types** to nodes and edges respectively
- Reconstructing **features** too

Thank You!



Questions?