UNIVERSITÀ
DI SIENA
1240

Graph neural networks and an application to molecule generation

Franco Scarselli

- DEPARTMENT OF INFORMATION ENGINEERING AND MATHEMATICS
- UNIVERSITY OF SIENA

OUTLINE

- Graph neural networks (GNNs)
- Applications of GNNs
- A method for molecule generation
- Some experimental results

Graph Neural Networks (GNNs)

Graphs

- they allow to represent
  - patterns (nodes with attached vector features)
  - their relationships (edges with attached vector features)

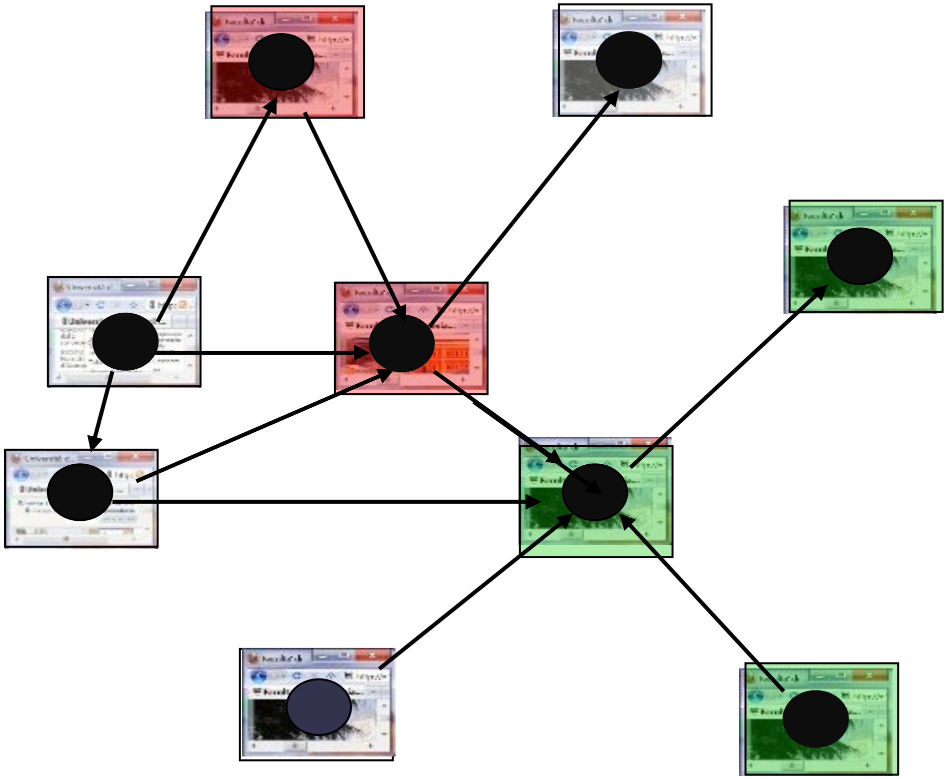
GNNs

- A class of machine learning models for graph processing
- They take in input a graph an return an output at each node/edge

An example: web spam detection

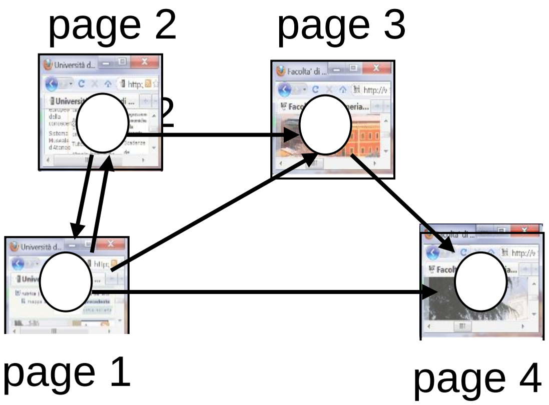
Red=spam
Green=not spam
Black=Unkown

- Goal: learning by examples to detect spam pages using
- Web connectivity
- Page content

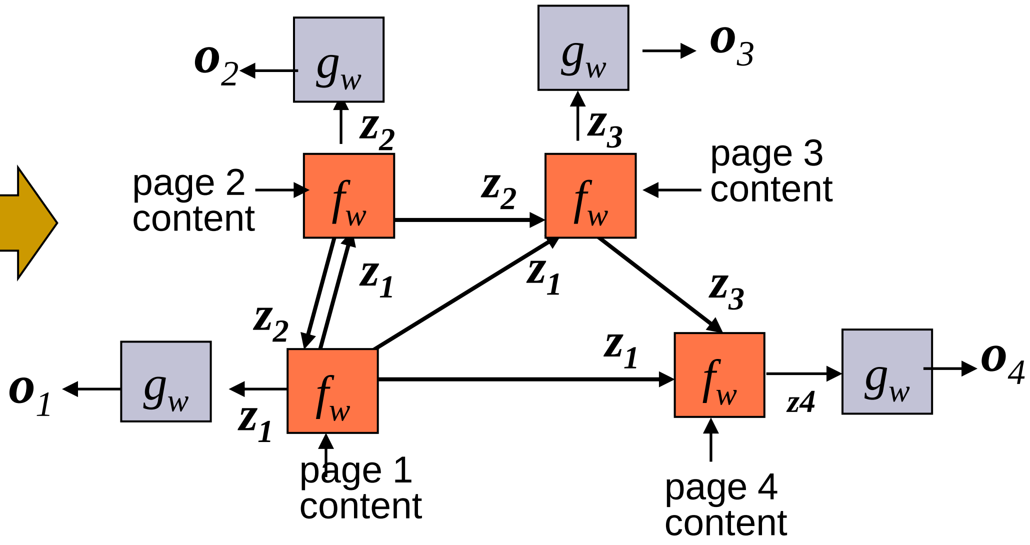


The GNN model (the first version)

- Two f and g neural modules (an instance for each node)
  - f compute a state z_v for each node
  - g compute an output o_v for each node
  - f modules are locally connected so as the graph
- All the modules share the same parameters



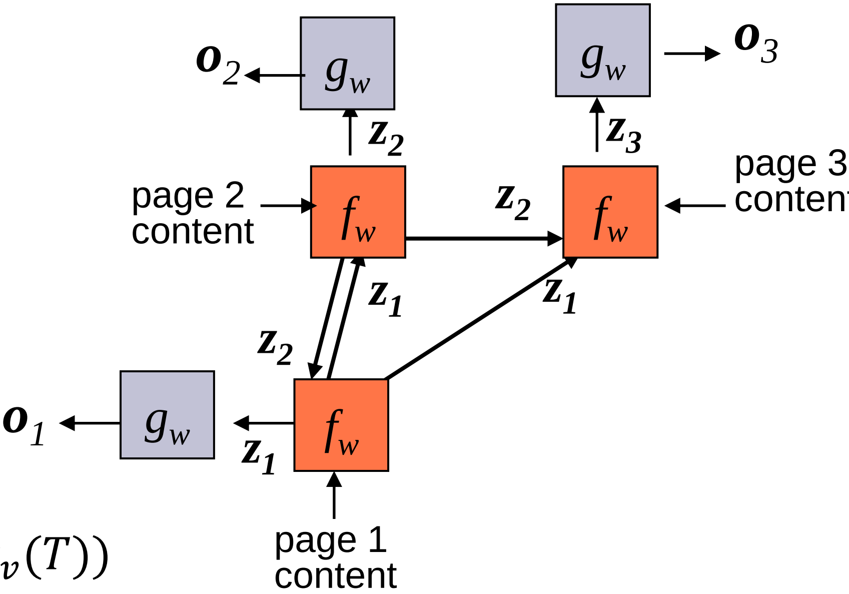
Input graph
i.e., the web



Encoding network

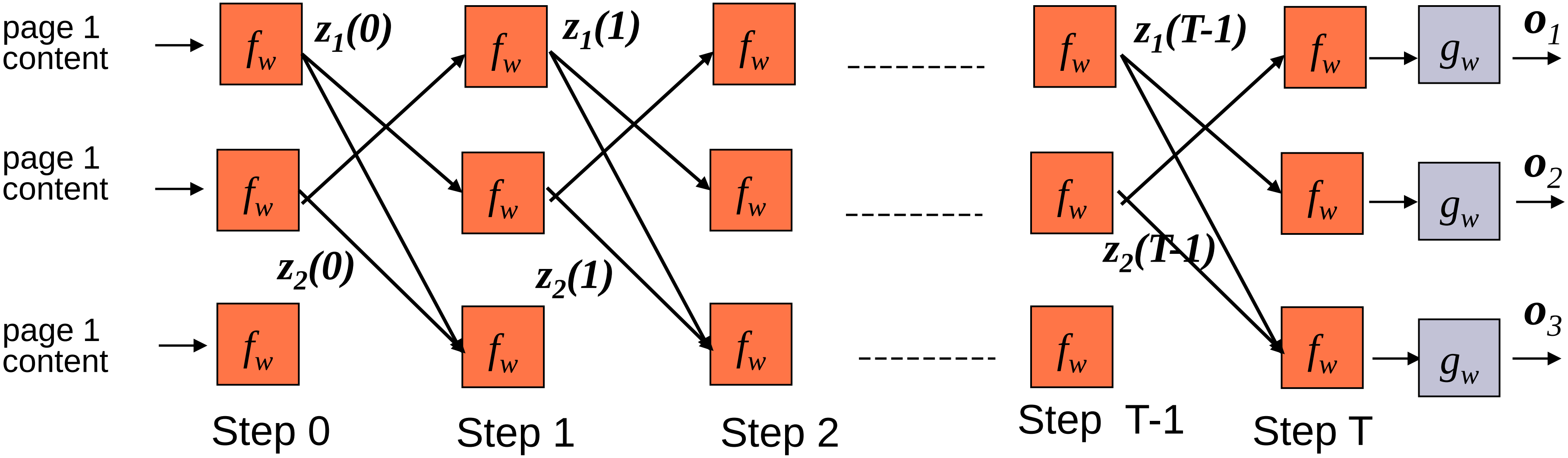
Output computation

- The output is computed by iterating the computation of the state until the state converges



$$z_v(t) = f_w(l_v, l_{ne[v]}, x_{ne[v]}(t-1)) \quad o_v = g_w(l_v, z_v(T))$$

The unfolding network



Modern GNNs

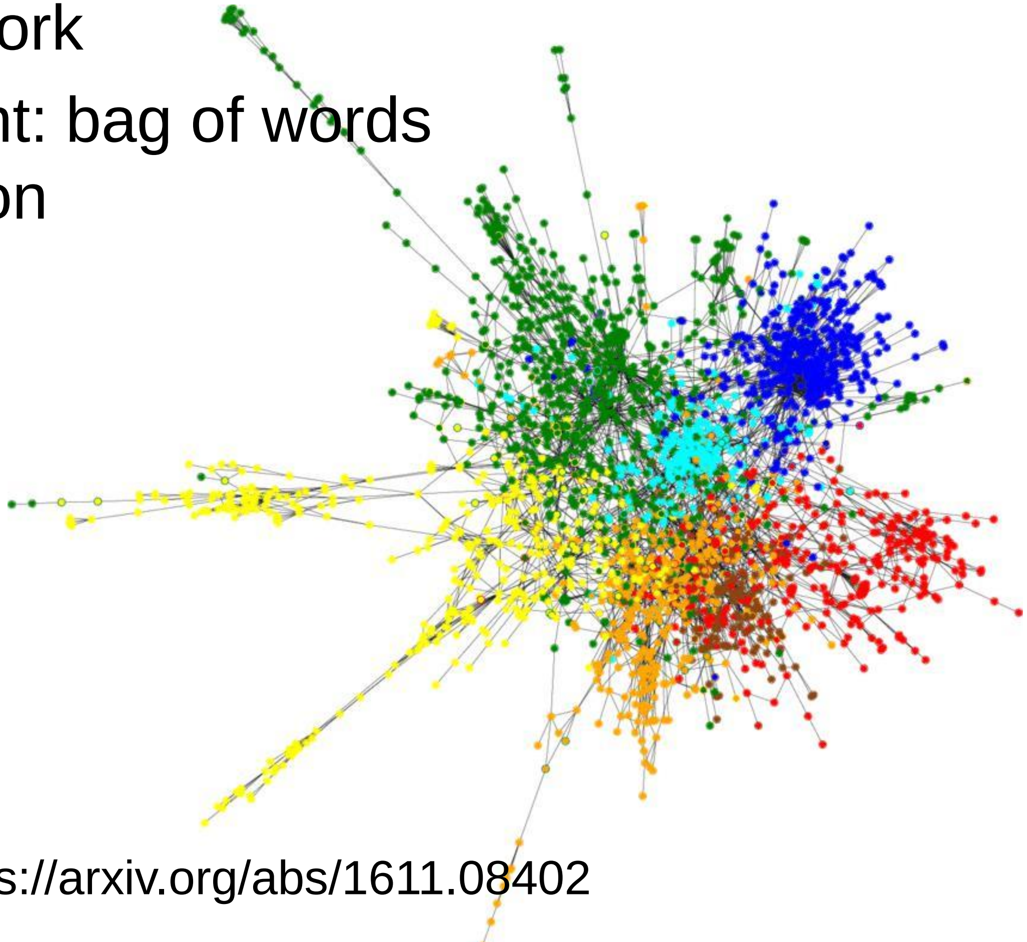
Modern message passage models

$$z_v(t) = COMBINE_w(x_v(t-1), AGGRREGATE_w\,(x_{ne[v]}(t-1))\quad)$$

- Different parameters (and state dimension) at each step
- Special types of modules: pooling, focus of attention, explanation, …
- Different types combination and aggregation functions
- GNNs for dynamic data
- …

Applications: document citation networks

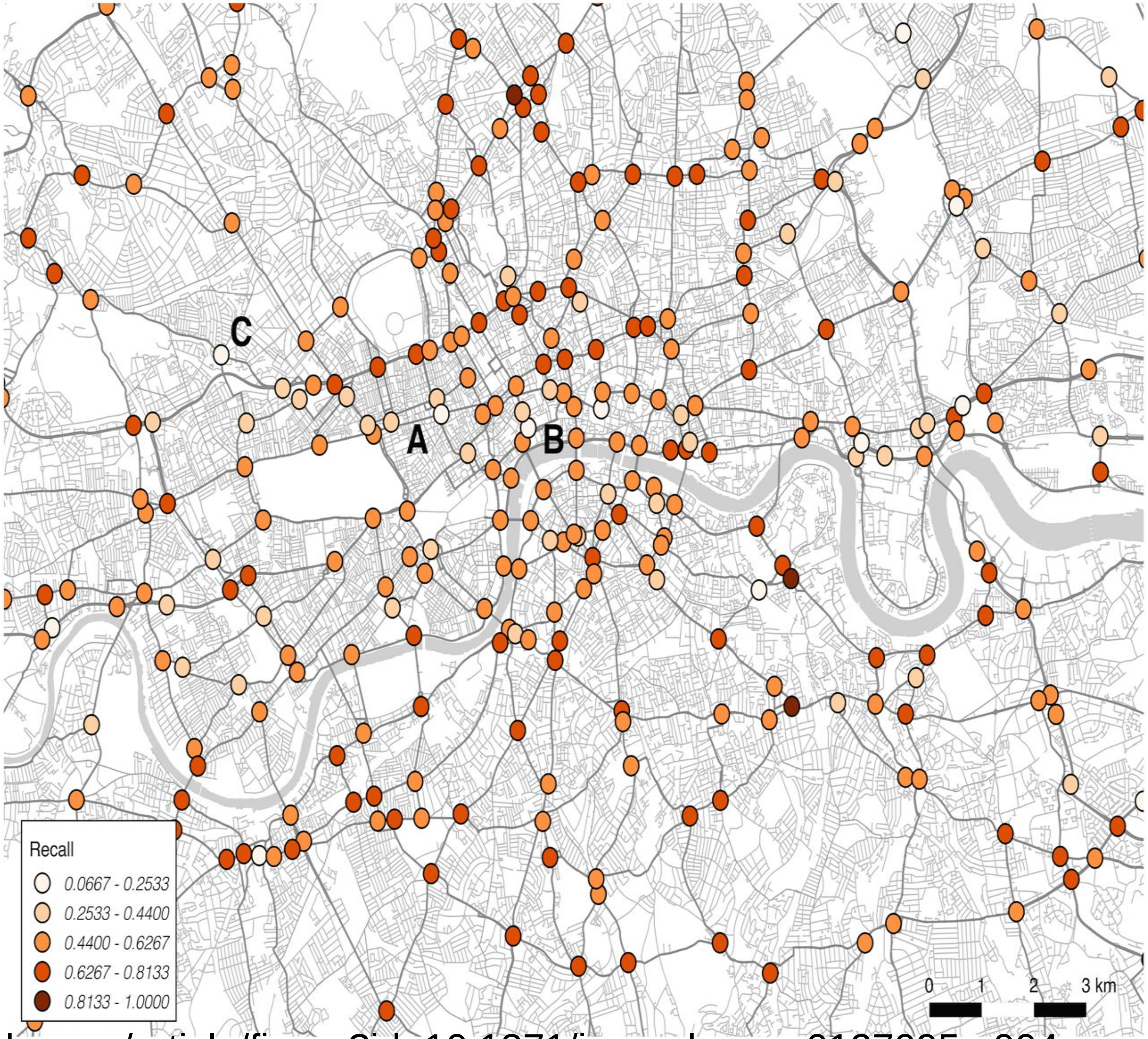
- To classify research papers using
  - citation network
  - paper content: bag of words representation



Source https://arxiv.org/abs/1611.08402

Road networks

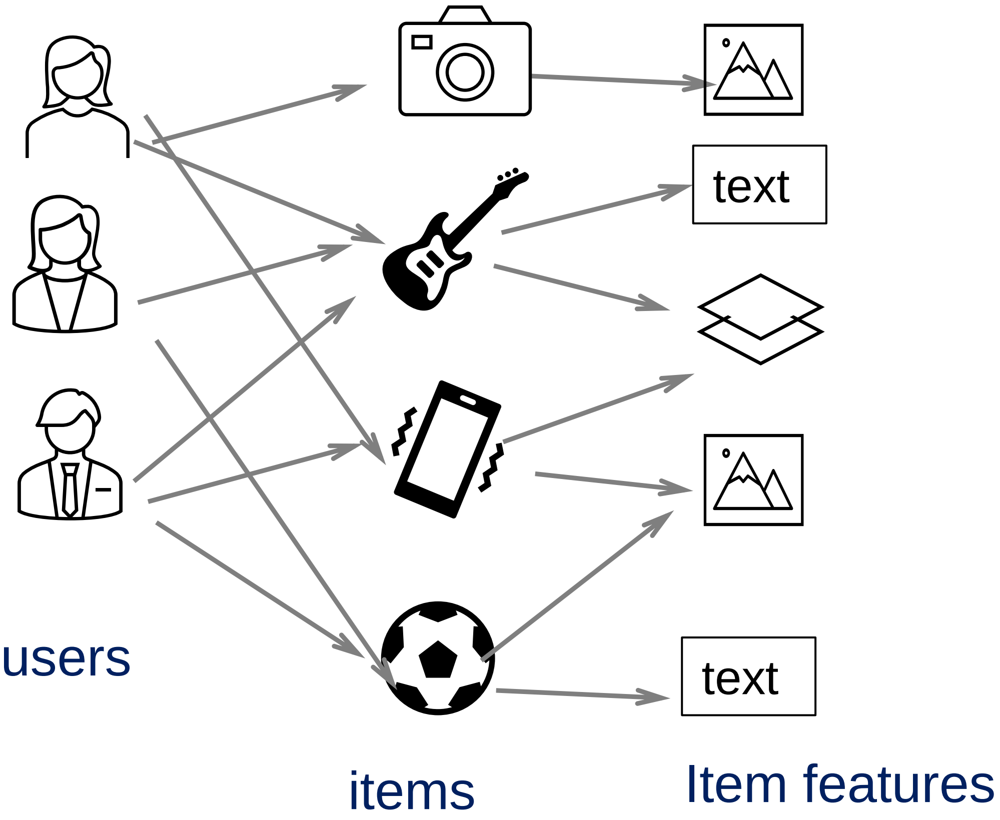
- To predict traffic learning for previous examples of traffic loads



Source: https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0127095.g004

Large scale recommendation system

- To recommend items to users using
  - User-item network
  - Item features, eg images, text, …



Other examples of applications

- Microsoft have experimented GNNs for program understanding
- Deep mind have used GNNs for traffic prediction, protein function prediction, ….
- Facebook has used GNNs to encode wikipedia graph
- Researchers at Large Hadron Collider at CERN will use GNNs to analyse data
- …..

A future general case?

Previous applications regard homogenous graphs for single tasks

A company/organization aggregates into a single data warehouse all its information

- The data look as a single big graph

A lot of nice new problems to play with ….

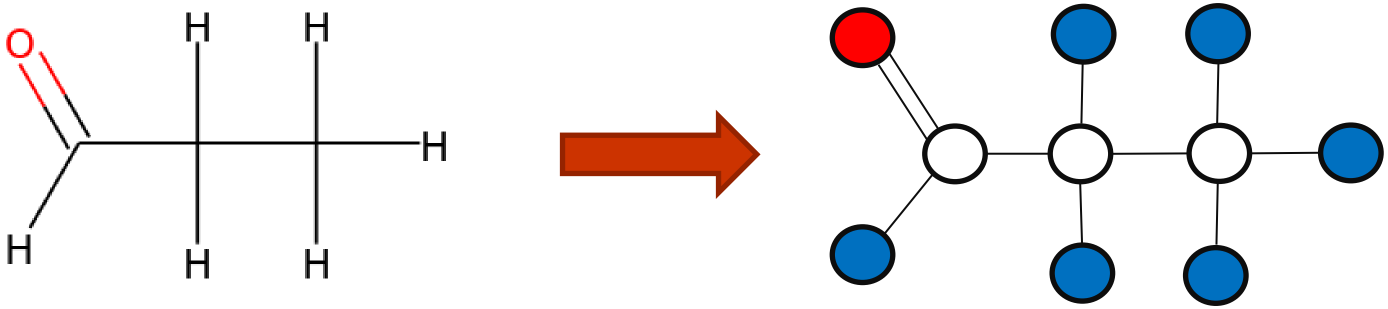
- Heterogenous patterns (nodes/edges)
- Different tasks to be solved (in sequence, contemporaneously)
- Different types of learning involved, e.g. inductive and transductive learning
- …

UNIVERSITÀ DI SIENA 1240

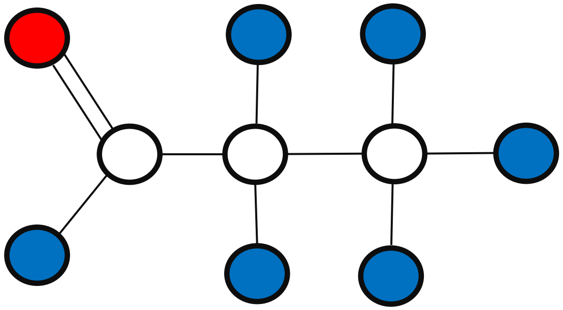
Molecular generative Graph Neural Networks for Drug Discovery (MG^{2}N^2)

with P. Bongini, M Bianchini

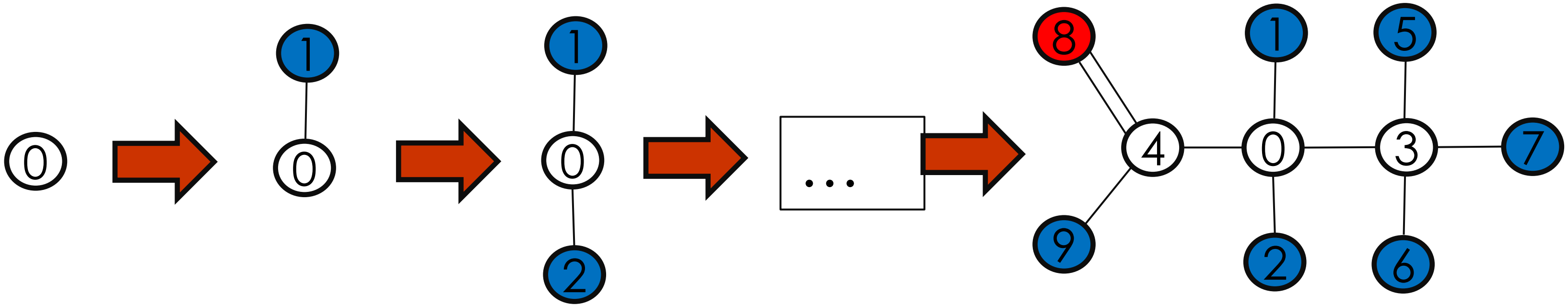
Molecules are usually represented as undirected graphs: atoms correspond to nodes and bonds correspond to edges.



Molecular generative Graph Neural Networks for Drug Discovery (MG^{2}N^2)

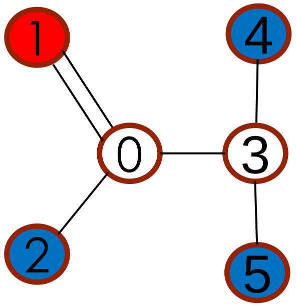


Generation starts from a graph composed of a single atom (sampled from the training set distribution). Each step is a standalone problem.

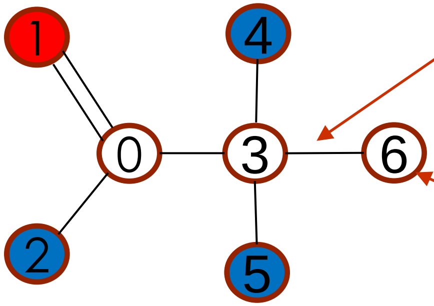


Molecular generative Graph Neural Networks for Drug Discovery (MG^2N^2)

Each step is split into three subproblems



P1: Do I generate another neighbour for node 3? Of which type?

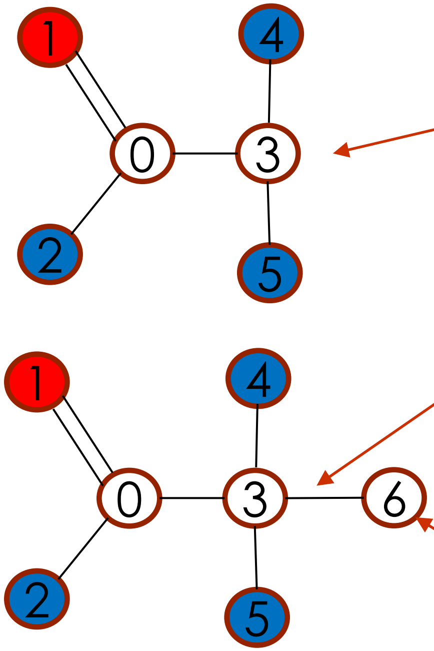


P2: Which is the type of bond (3,6)?

P3: Do I generate any extra bond: (0,6), (1,6), (2,6), (4,6), (5,6)?

GNN Modules

Each step is split into three subproblems

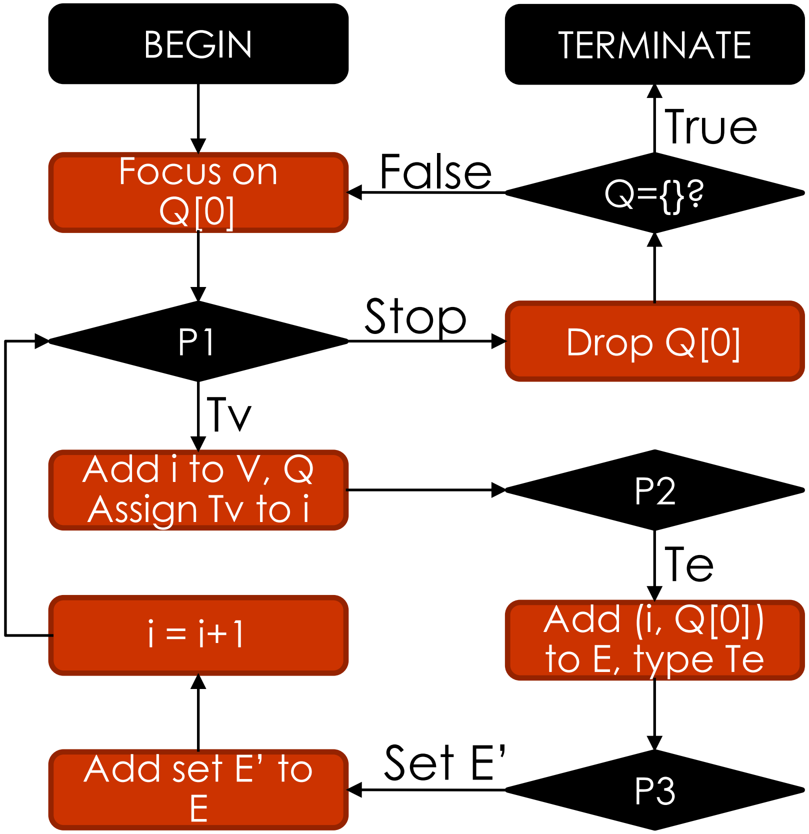


P1: A node-classifier GNN is asked to predict the next neighbour of node 3: (C, H, N, O, F, Stop). If the answer is «Stop», jump to node 4.

P2: Otherwise, an edge-classifier GNN is asked to predict the type of bond (3,6): (I, II, III).

P3: An edge predictor GNN decides if any extra bonds should be generated for the new node (and of which type).

Algorithm Chart

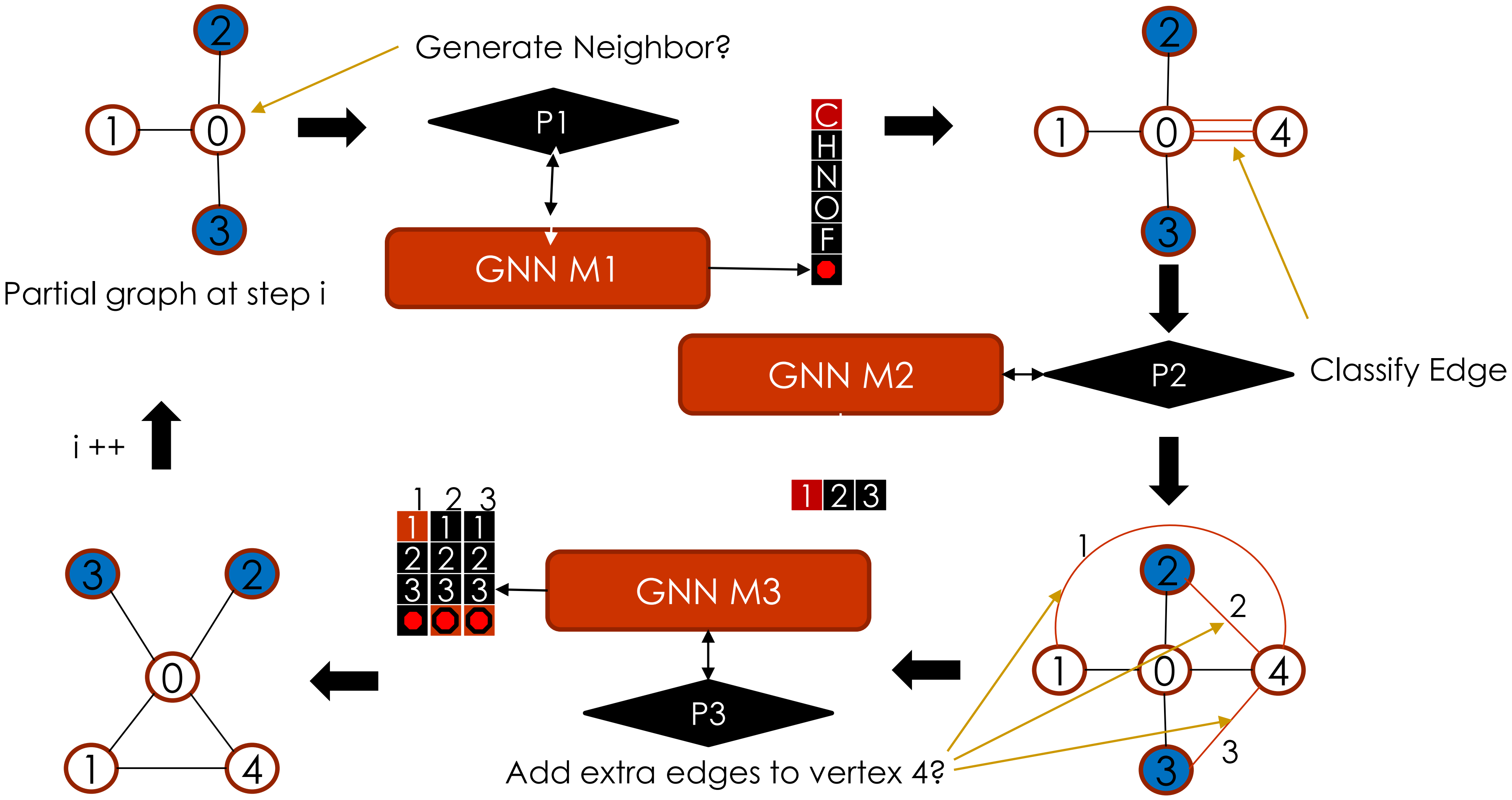


Generation of a graph G:
G=(V,E)
V: Vertex Set
E: Edge Set
Q: Expansion Queue
S: Starting Distribution

P1: Choose if to generate neighbor, and its type.
P2: Choose edge type
P3: Choose which edges to generate, and their types.

Initial conditions: E= {}, V={0}, type of vertex 0 sampled from training set, i=1, Q={0}.

Generation Step Example



Generate Neighbor?

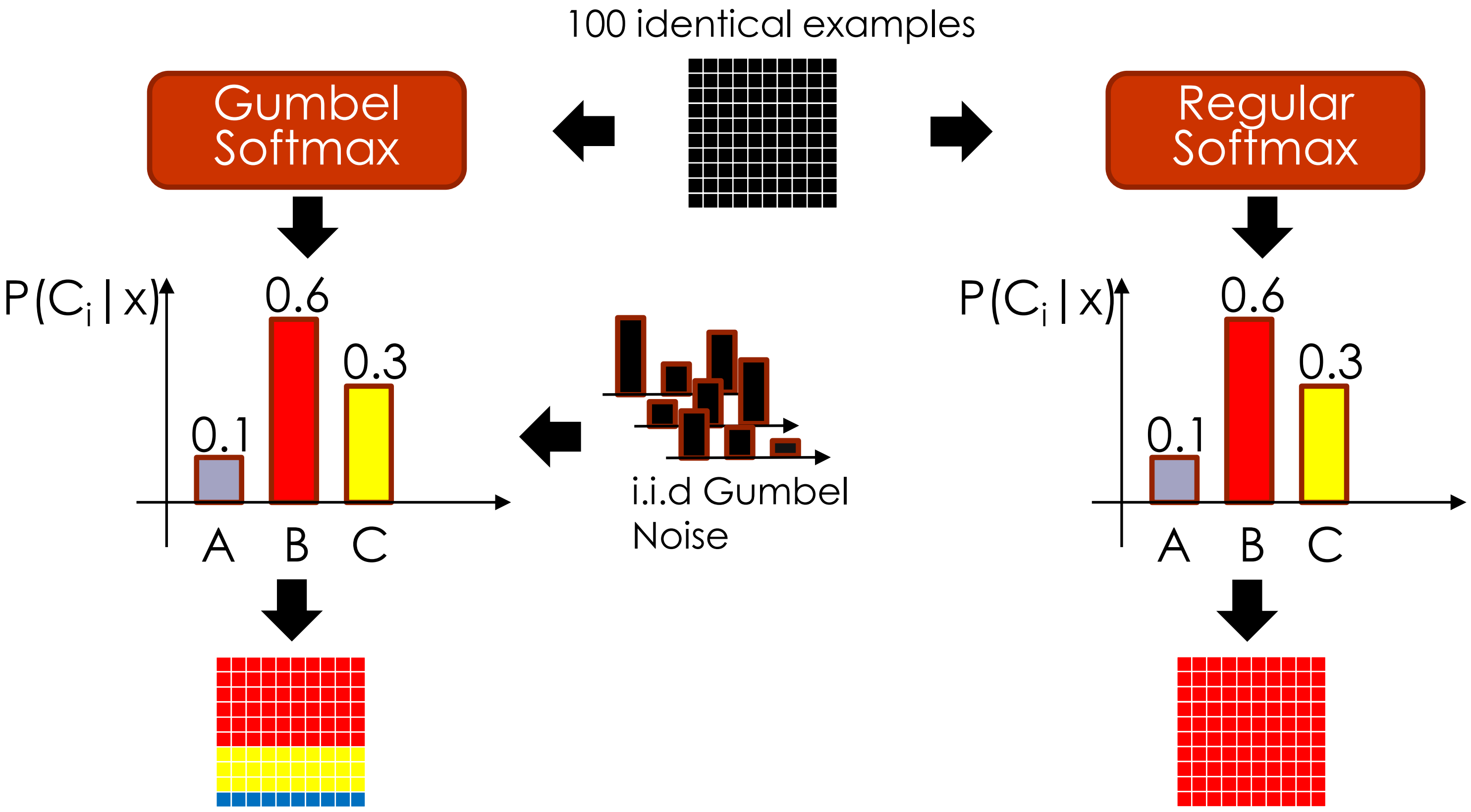
Partial graph at step i

Classify Edge

i ++

Add extra edges to vertex 4?

Gumbel Softmax Classifier



60% of examples are assigned to class B, 30% to class C, 10% to class A.

B has the highest probability, every example is assigned to class B.

About MG^2N^2

- GNN can use the whole graph information to decide
- Generation by iterative methods is "more explainable"
- GNN modules are trained separately
  - Make training/retraining much more flexible
  - It is based on an assumption their independence

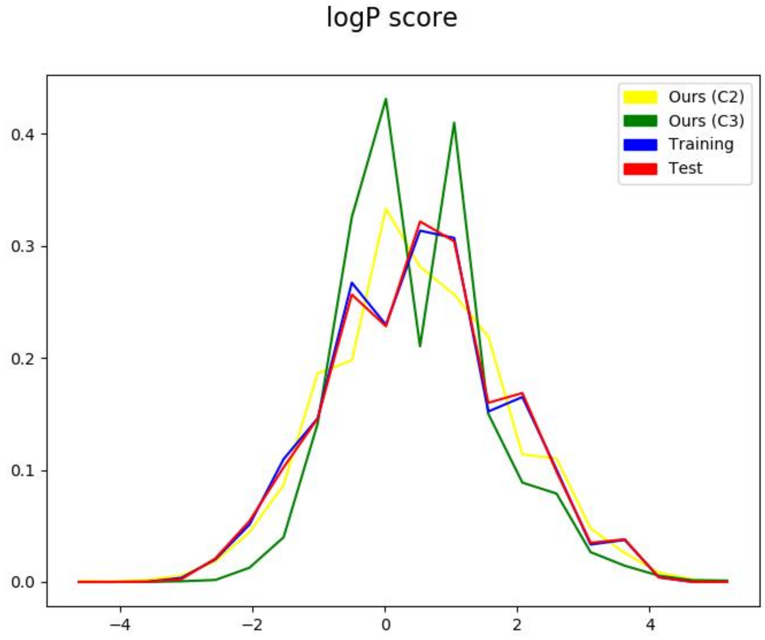
QM9 Dataset

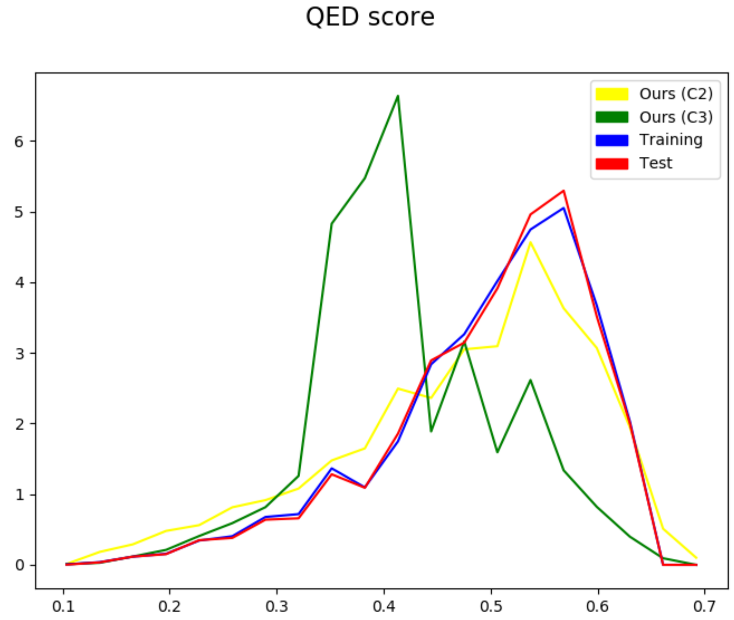
- Dataset of 134k molecules
- 5 atom types (CHNOF)
- 3 bond types (Single, double, triple)
- The objective is to generate new molecules (not found in QM9) which are chemically valid
- New molecules should have similar chemical properties with respect to a held-out test set (proof of generalization)

Experiments

- Each experiment consists in generating 10K molecule graphs
- A held-out test set of 10K graphs will be used to compare their chemical properties
- Chemical Validity, Novelty and Uniqueness are assessed with the RdKit package
- Molecular Mass, logP and QED are measured with RdKit as well

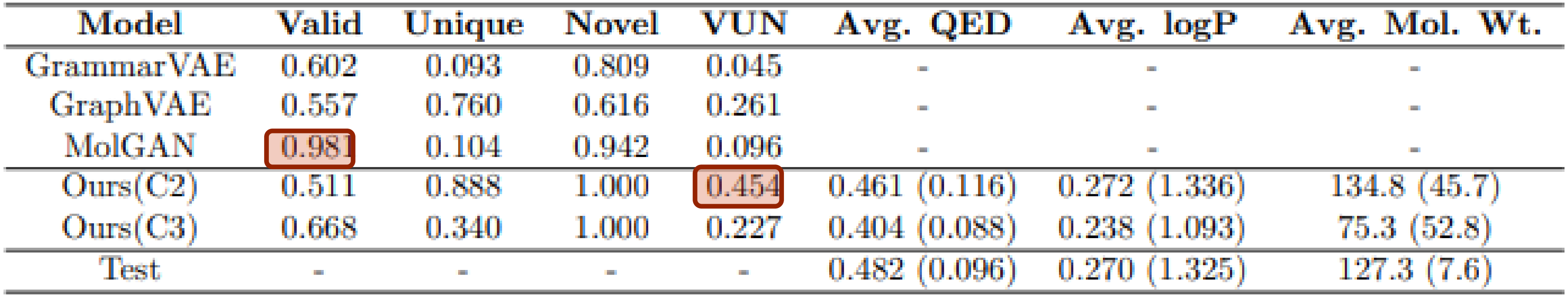
Comparing chemical properties


logP score


QED score

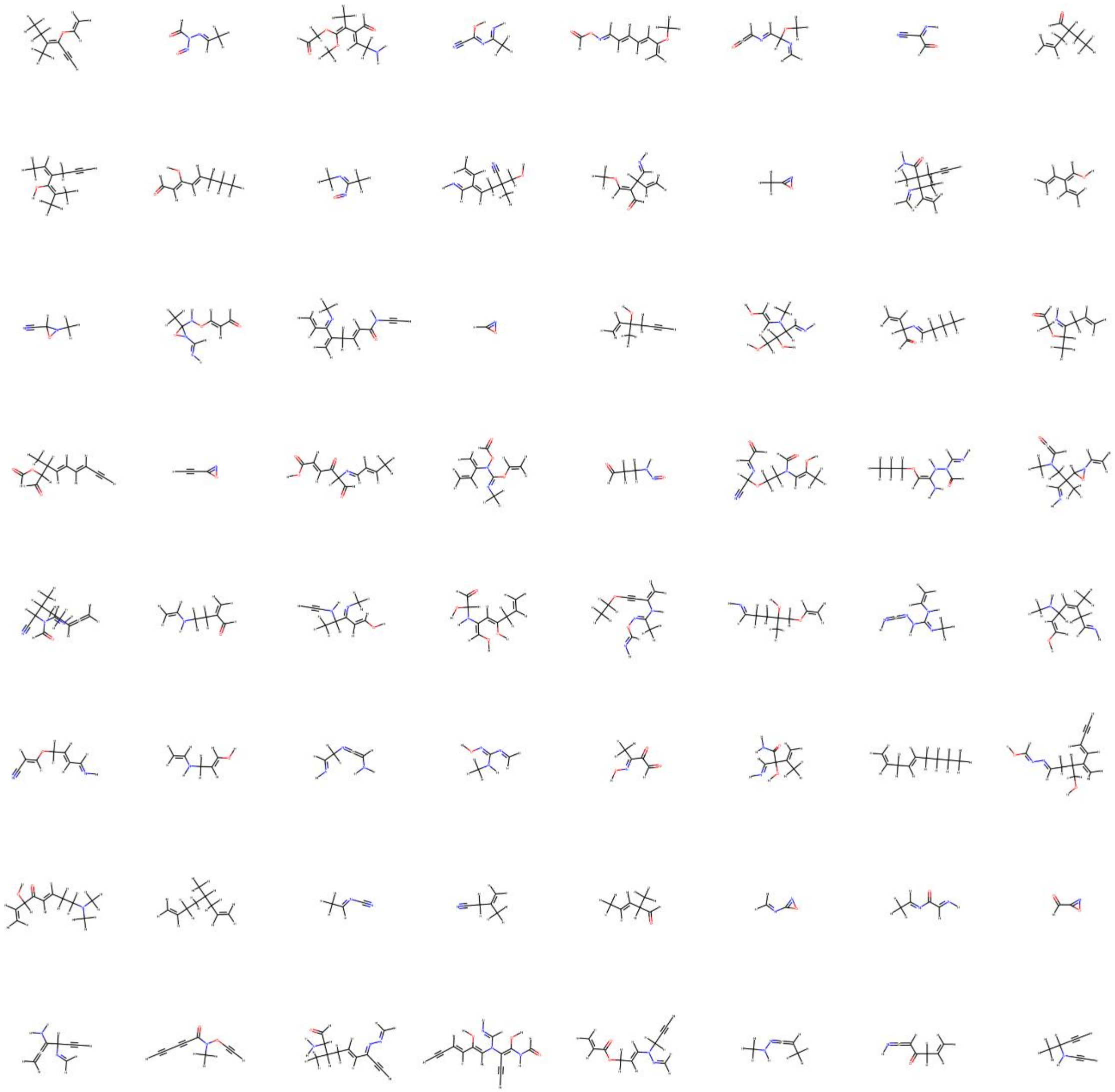
Experimental Results

- A comparison was carried out with other models for the unconditioned generation of molecules.
- GrammarVAE is the best SMILES-based VAE on QM9
- MolGAN (RL-based) is the best sequential model on QM9
- GraphVAE is the best graph-based VAE (and state of the art) for unconditioned generation on QM9



Model	Valid	Unique	Novel	VUN	Avg. QED	Avg. logP	Avg. Mol. Wt.
GrammarVAE	0.602	0.093	0.809	0.045	-	-	-
GraphVAE	0.557	0.760	0.616	0.261	-	-	-
MolGAN	0.981	0.104	0.942	0.096	-	-	-
Ours(C2)	0.511	0.888	1.000	0.454	0.461 (0.116)	0.272 (1.336)	134.8 (45.7)
Ours(C3)	0.668	0.340	1.000	0.227	0.404 (0.088)	0.238 (1.093)	75.3 (52.8)
Test	-	-	-	-	0.482 (0.096)	0.270 (1.325)	127.3 (7.6)

Generated Molecules



Thank you for your attention!

Reference

- Molecular generative Graph Neural Networks for Drug Discovery, P Bongini, M Bianchini, F Scarselli Neurocomputing