University of Basel

Graph-based Molecule Design with Deep Latent Variable Models

Elias Arnold <elias.arnold@stud.unibas.ch>

Department of Mathematics and Computer Science, University of Basel

30.04.2020

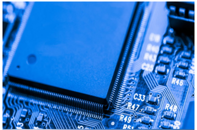
Agenda

1. Motivation
2. Conventional methods
3. Problem
4. Novel approaches
5. Recap of variational autoencoders
6. Constraint Graph Variational Autoencoder (CGVAE)
7. Experiments with CGVAE
8. Conclusion and outlook

Motivation

- Demand for new Molecules steadily increasing
- Number of small organic molecules: $> 10^{60}$ (Number of molecules which can be used to manufacture new drugs)
- Reseach should be efficient (effort, money, ...)

Uses



Semiconductors

Photovoltaics

Drug discovery

Conventional Methods for Molecule Design (Drug discovery)

- **Goal**: Find molecule that binds to a target to (de) activate it
- Two main stages:

1. Find molecules that are capable of binding to the target (Hits)

   High-Throughput-Screening (HTS)

   - Physical tests
   - Search molecule pool
   - Pool size: $\sim 10^3$

   Virtual Screening (VS)

   - No physical molecules needed
   - Computer simulations
   - Various databases available

2. Optimise the molecules found (combine advantages of Hits)

   - Modify hit-compounds (combine advatages)
   - Quantum chemical property estimation
   - Implies solution of Schrödinger equation or Quantum Monte Carlo

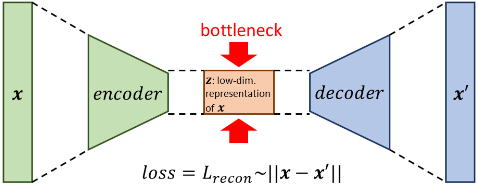
Problem

- Conventional methods scale very poorly
- Conventional methods can not explore all possible compounds
- Random search (*trial and error*) −> **NOT** goal oriented

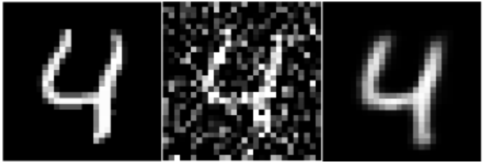
- Recent publications propose statistical methods to tackle this problem
- These approaches also need a dataset for training
- Efficiently optimise the quantum chemical properties of the molecule

The Autoencoder (AE)

- Often, the models used follow the varational autoencoder (VAE) architecture
- To introduce a VAE, we have to understand an (AE) autoencoder first:


bottleneck
$\boldsymbol{x}$
$encoder$
$\boldsymbol{z}$: low-dim. representation of $\boldsymbol{x}$
$decoder$
$\boldsymbol{x}'$
$loss = L_{recon} \sim ||\boldsymbol{x} - \boldsymbol{x}'||$

(a) Illustration of an AE

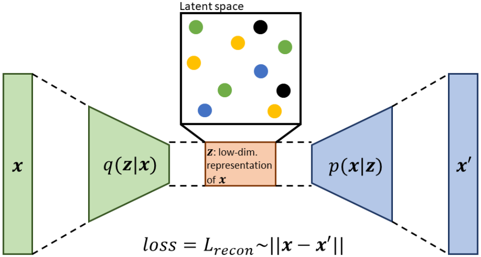


(b) Original | with noise | AE reconstruction

- Vector $\boldsymbol{z}$ is called latent vector

The Autoencoder (AE)

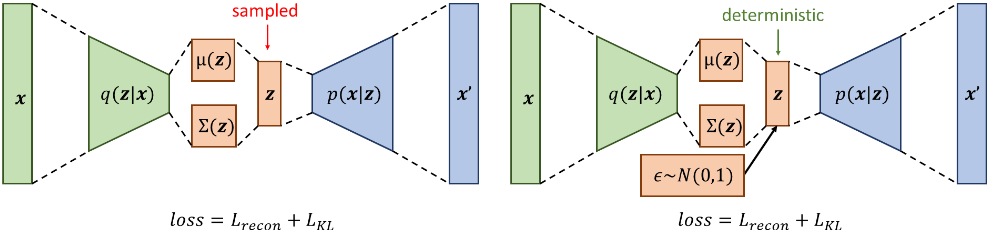
- Assume a dataset of molecules $\boldsymbol{x}$ with a corresponding property y: $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3), \ldots\}$
- AE distributes $\boldsymbol{z}$ all over the latent space, to distinguish molecules
- The colour of the latent vectors $\boldsymbol{z}$ represents the value of y


Latent space
$\boldsymbol{x}$
$q(\boldsymbol{z}|\boldsymbol{x})$
$\boldsymbol{z}$: low-dim. representation of $\boldsymbol{x}$
$p(\boldsymbol{x}|\boldsymbol{z})$
$\boldsymbol{x}'$
$loss = L_{recon} \sim ||\boldsymbol{x} - \boldsymbol{x}'||$

- We want to sample new $\boldsymbol{z}$, to generate unseen $\boldsymbol{x}'$
- **Problem**: Distribution of the latent space not tractable

The Variational Autoencoder (VAE)

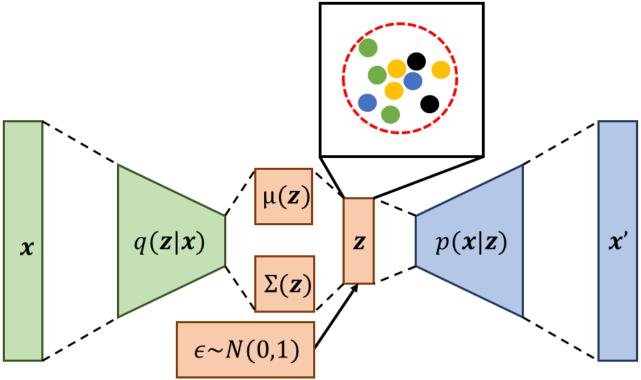
- **Idea**: use KL divergence as a regulariser term to shape latent space according to a tractable distribution (gaussian) –> **Problem**: run backpropagation on a stochastic node
- Use parametrisation trick: $z = \mu + \sigma \odot \epsilon$



(a) How to run backpropagation?

(b) With reparametrisation trick

The Variational Autoencoder (VAE)

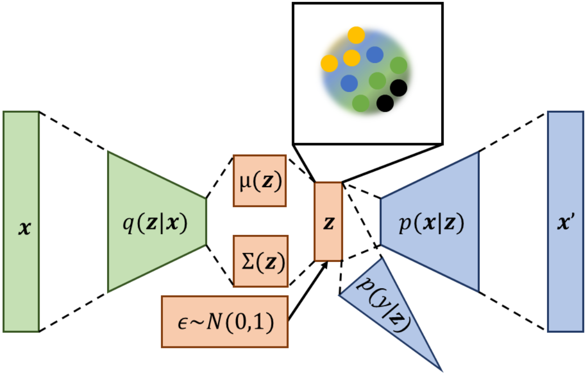


$$loss = L_{recon} + L_{KL}$$

- KL divergence forms latent space according to a gaussian (here, $\mathcal{N}(0, 1)$)
- **Problem**: can we structure the latent space to sample a specific property?

The Variational Autoencoder (VAE) with Property Prediction

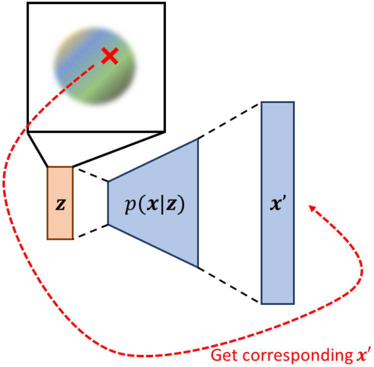
- Structuring of the latent space by second decoder $p(y|\boldsymbol{z})$ with $L_{property}$ loss
- Add $L_{property}$ to the loss function of VAE



$$loss = L_{recon} + L_{KL} + L_{property}$$

VAE to Generate New Data

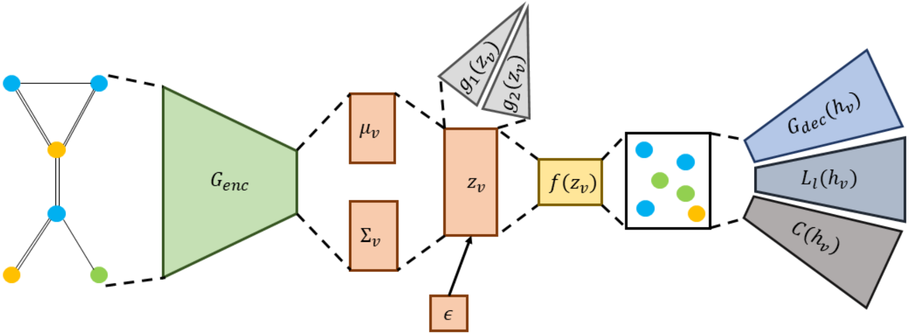
- We can now sample random latent vectors from the latent space
- Moreover, we can sample in regions with interesting property values
- The samples can then be reconstructed with the trained decoder $p(\boldsymbol{x}|\boldsymbol{z})$


$\boldsymbol{z}$
$p(\boldsymbol{x}|\boldsymbol{z})$
$\boldsymbol{x}'$
Get corresponding $\boldsymbol{x}'$

Constrained Graph Variational Autoencoder (CGVAE)

- In this masterthesis, we evaluated CGVAE (Liu et al. 2018)
- **Variational Autoencoder**: CGVAE is based on the VAE architecture
- **Constrained**: CGVAE uses masks to ensure chemical stability of the generated molecules
- **Graph**: CGVAE represents molecules as graphs

Architeture of CGVAE



- CGVAE encodes single nodes (instead of whole molecule)
- Sequential assembly of final graph (in BFS order)
- Encoder and decoder are gated graph neural networks (GGNN) (Li et al. 2015)

Experiments with CGVAE

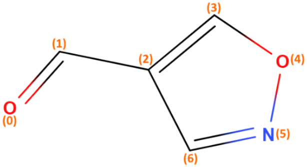
- $L_{CGVAE} = L_{recon} + \lambda_1 L_{KL} + \lambda_2 L_{property}$
- We split our experiments into three parts
- In each part, we include an additional loss term:
  1. Property prediction: $loss = L_{property}$
  2. Reconstruction of molecules: $loss = L_{property} + L_{recon}$
  3. Sampling of new molecules: $loss = L_{property} + L_{recon} + L_{KL}$

Experiments Part 1

- Represent molecules as graphs (like CGVAE)
- **Goal**: predict quantum chemical properties
- This part focuses on: $loss = L_{property}$

Experiments Part 1 - Molecule Descriptor

- Encode molecule as a classical graph
- Graph is a pair of two sets $\mathcal{G} = (V^{9\times 5}, E^{13\times 5})$
- Example: 1,2-oxazole-4-carbaldehyde



```
[[0, 1, 1, -1, -1],
[1, 0, 2, -1, -1],
[2, 1, 3, -1, -1],
[3, 0, 4, -1, -1],
[4, 0, 5, -1, -1],
[5, 1, 6, -1, -1],
[6, 0, 2, -1, -1],
[-1, -1, -1, -1, -1], ...]
```

E **(set of edges)**

Key:
0: Single bond
1: Double bond
2: Triple bond
3: Aromatic bond

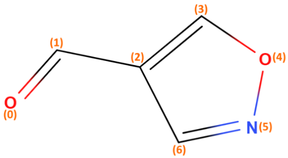
```
[[0, 0, 0, 1, 0],
[0, 1, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 0, 0, 1, 0],
[0, 0, 1, 0, 0],
[0, 1, 0, 0, 0],
[-1, -1, -1, -1, -1], ...]
```

V **(set of nodes)**

Key:
[H, C, N, O, F]

Experiments Part 1 - Molecule Descriptor with Spatial Information

- Graph descriptor has weak descriptive power
- **Idea**: use spatial information (bond lengths/angles) of molecule
- Graph is a tuple of three sets $\mathcal{G}^{47\times5} = (V^{9\times5}, E^{13\times5}, A^{25\times5})$



```
[[0, 1, 1, 1.21, -1],
[1, 0, 2, 1.46, -1],
[2, 1, 3, 1.37, -1],
[3, 0, 4, 1.33, -1],
[4, 0, 5, 1.40, -1],
[5, 1, 6, 1.31, -1],
[6, 0, 2, 1.43, -1],
[-1, -1, -1, -1, -1], ...]
```

E **(set of edges)**

Key:
0: Single bond
1: Double bond
2: Triple bond
3: Aromatic bond
Bond length in angstrom

```
[[0, 1, 2, 2.17, -1],
[1, 2, 3, 2.23, -1],
[1, 2, 6, 2.26, -1],
[2, 3, 4, 1.93, -1],
[3, 4, 5, 1.91, -1],
[4, 5, 6, 1.83, -1],
[5, 6, 2, 1.96, -1],
[-1, -1, -1, -1, -1], ...]
```

A **(set of angles)**

Key:
Bond angle in radian

```
[[0, 0, 0, 1, 0],
[0, 1, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 0, 0, 1, 0],
[0, 0, 1, 0, 0],
[0, 1, 0, 0, 0],
[-1, -1, -1, -1, -1], ...]
```

V **(set of nodes)**

Key:
[H, C, N, O, F]

Experiments Part 1 - Networks

- We used four different networks to predict quantum chemical properties
  - GGNN - used in CGVAE (Li et al. 2015)
  - ARMA - graph learning (Bianchi et al. 2019)
  - RNN - as baseline
  - CNN - as baseline
- Hyperparameters optimised by hand

Experiments Part 1 - Dataset

- Use popular quantum machine 9 (QM9) dataset (Ramakrishnan et al. 2014)
  - 133,885 organic molecules
  - Up to nine heavy (non-hydrogen) atoms
  - 3-d coordinates for each atom
  - 15 chemical properties [hartrees]
- Split data into three disjoint parts
  - Test set: 30'000 molecules
  - Validation set: 20777 molecules (20 % of remaining data)
  - Training set: 83108 molecules (80 % of remaining data)
  - Use split to regress four chemical properties
    - ϵ_{HOMO} - Energy of the highest occupied molecular orbital [kcal/mol]
    - ϵ_{LUMO} - Energy of the lowest unoccupied molecular orbital [kcal/mol]
    - ϵ_{GAP} - Difference of ϵ_{HOMO} and ϵ_{LUMO} [kcal/mol]
    - U_0 - Internal energy at 0 K [kcal/mol]

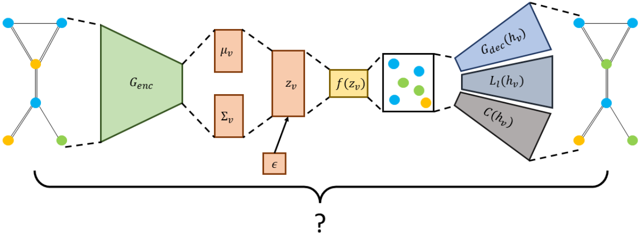
Experiments Part 1 - Results

- Spatial information has big impact on descriptive power
- Good accuracy with baseline models
- Accuracy of simulator: ≈1 kcal/mol

Descriptor	**Architecture**	MAE values in kcal/mol			
		$\epsilon_{\textbf{HOMO}}$	$\epsilon_{\textbf{LUMO}}$	$\epsilon_{\textbf{GAP}}$	$\mathbf{U_0}$
Graph-based with bond lengths and bond angles	CNN	5.60	6.34	7.64	18.89
	RNN	3.52	3.26	5.03	**6.43**
	GGNN	**3.15**	3.29	5.10	7.42
	ARMA	3.32	**2.95**	**4.74**	11.36

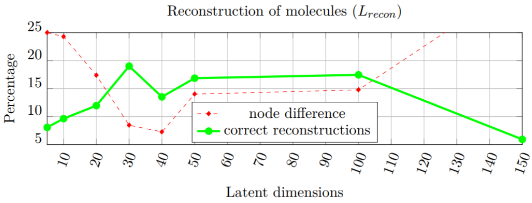
Experiments Part 2

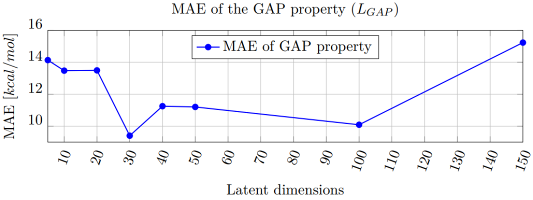
- This part focusses on: $loss = L_{property} + L_{recon}$
- Combine en- and decoder
- **Goal**: find latent size for optimal reconstructions



Experiments Part 2 - Molecule Comparison

- Compare both original and reconstructed molecule in terms of...
  - ... the ϵ_{GAP} property
  - ... the graph structure

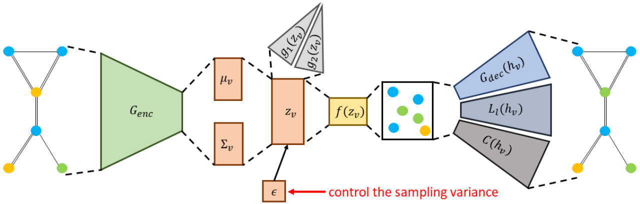




- With 30 latent dimensions, CGVAE makes the most accurate reconstructions

Experiments Part 3

- This part focusses on: $loss = L_{property} + L_{recon} + L_{KL}$
- **Goal**: structuring of the latent space and sampling of new molecules (generation)



control the sampling variance

- CGVAE uses $\epsilon \in \mathcal{N}(0,1)$

Experiments Part 3 - Metrics

- Three prevalent metrics to evaluate generative behaviour

Metric	Description	Reported in CGVAE paper
Novelty	Ratio of generated molecules not present in the training dataset used	94.35 %
Uniqueness	Ratio of duplicates in the set of generated molecules	94.35 %
Validity	Ratio of chemically valid molecules	100.00 %

- Can we reproduce the results reported in the CGVAE paper?

Experiments Part 3 - Our Results

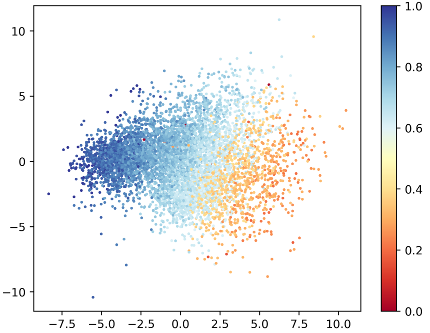
Variance of ϵ	Novelty	Uniqueness	Validity
2.5	84.84 %	89.18 %	100.00 %
2	86.35 %	90.71 %	100.00 %
1.5	88.75 %	96.10 %	100.00 %
1	87.10 %	94.75 %	100.00 %
0.5	79.71 %	96.60 %	100.00 %
0.1	63.78 %	99.14 %	100.00 %
Values in the paper	94.35 %	98.57 %	100.00 %

- Just the validity results were similar as the values reported in the paper
- Opposite trends in novelty and uniqueness of the results
- While novelty **increases** with the variance, uniqueness **decreases**

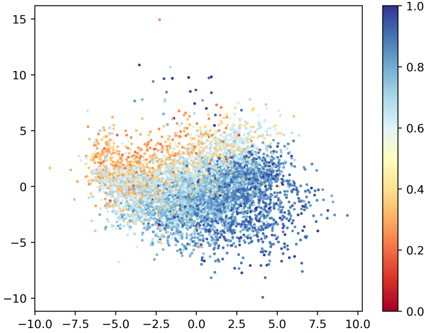
Experiments Part 3 - Visualisation of the Latent Space

- To represent a molecule, we used the mean vector of all nodes
- Then, PCA was used to find a 2-d subspace of the 30-d latent space
- Colours correspond to the ϵ_{GAP} property

$\epsilon \in \mathcal{N}(0, 0.1)$



$\epsilon \in \mathcal{N}(0, 2.5)$



Conclusion

- Spatial information enhances accuracy for property prediction
- Latent space does not have to be high dimensional for good reconstructions
- Trade-off in usually used metrics (Uniqueness and novelty cannot be optimised together)

Outlook

- CGVAE build a molecule sequentially
- This has two donsides:
  - Unlikely to reconstruct exact same molecule
  - Long generation time

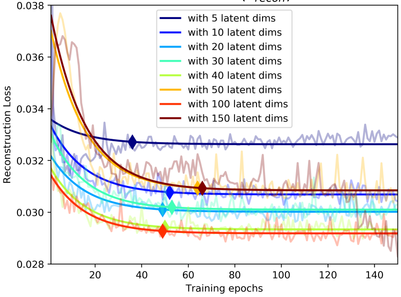
- Future research should investigate one-shot generation of molecules

Questions?

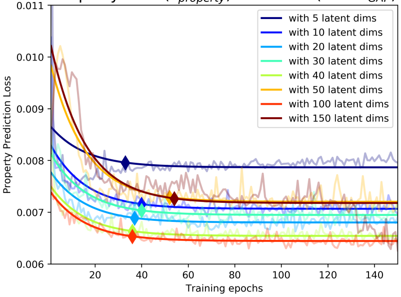
elias.arnold@stud.unibas.ch

Training of CGVAE

- This was our first approach for part 2 of the experiments
- Train models with different sizes of latent space
- Unfortunately, no ranking of models possible due to stochastic nature of training
- However, loss converges after ≤ 70 epochs of training



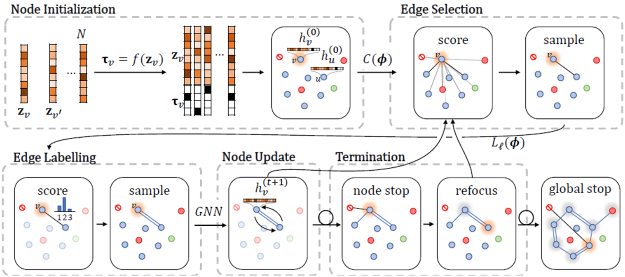
Reconstruction Loss (L_{recon}) of CGVAE



Property Loss ($L_{property}$) of CGVAE (with ε_{GAP})

Generative Procedure of CGVAE

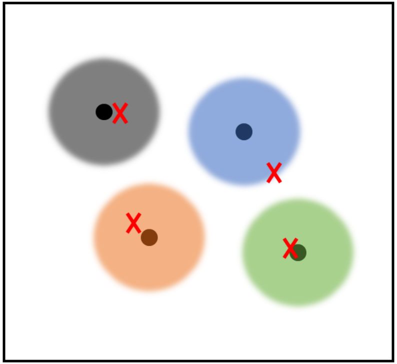
- CGVAE builds a molecule sequentally (node-by-node)
- Therefore, the three networks G_{dec} (GNN), L_l, and C are used
- Initially, the latent space has N unconnected nodes
- N is an upper bound on the atoms of the original molecule



Novelty vs. Uniqueness

- Assume latent vectors represent entire molecules
- Combine latent vectors of CGVAE (e.g. mean vector)

Small variance (increases uniqueness)



Large variance (increases novelty)