Combining Graph Attention and Recurrent Neural Networks in a Variational Autoencoder for Molecular Representation Learning and Drug Design

Alex Müller, Kenneth Atz, Michael Reutlinger, Nicolas Zorn

Small Molecule Research, Pharma Research and Early Development
Roche Innovation Center Basel, Switzerland
F. Hoffmann-La Roche Ltd.

Introduction

To navigate the vast chemical space, computer-assisted drug design approaches necessitate **molecular representations that can be correlated with biological activity**. We are interested in creating **a single representation** that can be **used for multiple downstream tasks** such as quantitative structure–activity relationship (QSAR), virtual screening and de novo molecular design.

Recent advances in molecular representation learning using **chemical language models and geometric deep learning approaches** have shown promising results towards finding chemical foundation model. Here, we combine the best of both worlds and present **G**raph **I**nfused **R**epresentation **A**ssembled **F**rom a multi-**F**aceted variational auto-**E**ncoder (**GIRAFFE**). **GIRAFFE** uses a graph attention neural network[1] as encoder and a RNN with LSTM cells[2] as decoder, constructed as a variational autoencoder (VAE)[3] with property learning.

Methods

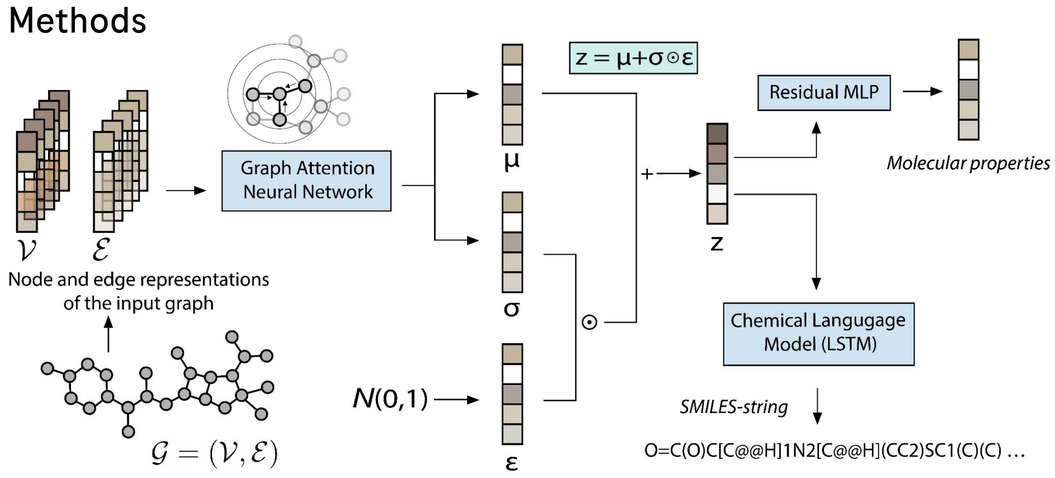

Figure 1: Network architecture of GIRAFFE with GNN encoder, LSTM decoder and property MLP.

Dataset: 10M SMILES strings randomly extracted from PubChem with a maximum length of 128 characters.

Batch size: 256 molecules, randomly sampled from the full 10M in each step.

Graph features: AttentiveFP[1], adapted to 32 atom and 10 bond features.

Reconstruction task: random parent SMILES string in every step.

Property prediction task: 125 calculated RDKit properties (scaled [0, 1]).

Optimizer: Adam[4], with initial learning rate of 0.001 and a step-wise decay of 0.75 every 10 epochs.

Epochs: max. 150, with 1000 steps per epoch

VAE: sampling latent vector z from $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

Loss: $\mathcal{L} = \mathcal{L}_S + \lambda_P \times \mathcal{L}_P + \beta \times \mathcal{L}_{KLD}$

β: cyclical annealing[5] to initially focus on prior.

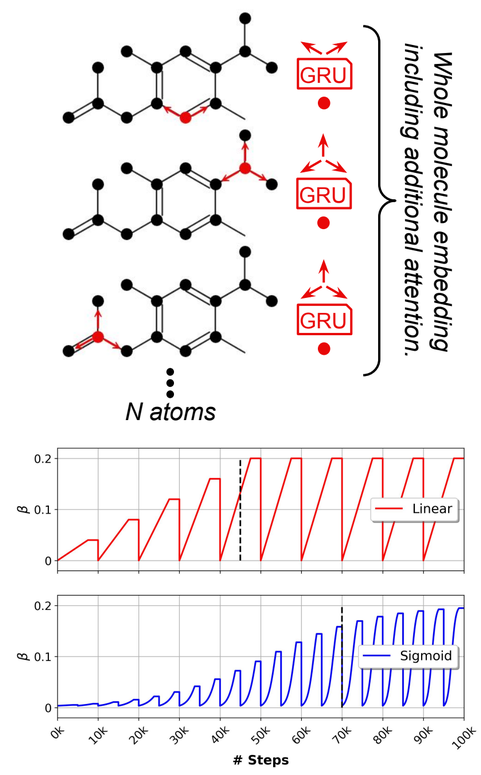

Figure 2: AttentiveFP[1] attention mechanism and best performing cyclical annealing schedules for factor **β**.

Results

Table 1: Regression benchmark[6] RMSE results.

DESCRIPTOR	ESOL	FREESOLV	LIPOP
RDKIT	0.69 ± 0.08	1.67 ± 0.45	0.74 ± 0.04
ECFP4	0.90 ± 0.06	2.88 ± 0.38	0.77 ± 0.03
CDDD	0.57 ± 0.06	1.46 ± 0.43	0.67 ± 0.02
MOLBERT	**0.55 ± 0.07**	1.52 ± 0.66	**0.60 ± 0.01**
FP-BERT	0.67 ± 0.07	**1.07 ± 0.18**	0.67 ± 0.02
NONVAE*	0.57 ± 0.07	**1.07 ± 0.34**	0.61 ± 0.01
GIRAFFE*	**0.55 ± 0.08**	1.11 ± 0.31	0.67 ± 0.03

*our models; others: Ref. 7-9.

Table 2: Classification benchmark[6] AUROC results.

DESCRIPTOR	BACE	BBBP	HIV
RDKIT	0.83 ± 0.00	0.70 ± 0.00	0.71 ± 0.00
ECFP4	**0.85 ± 0.00**	0.68 ± 0.00	0.71 ± 0.00
CDDD	0.83 ± 0.00	**0.76 ± 0.00**	0.75 ± 0.00
MOLBERT	**0.85 ± 0.00**	0.75 ± 0.00	0.75 ± 0.00
FP-BERT	–	0.71 ± 0.01	**0.78 ± 0.01**
NONVAE*	**0.85 ± 0.00**	0.72 ± 0.00	0.71 ± 0.00
GIRAFFE*	**0.85 ± 0.00**	0.71 ± 0.00	0.72 ± 0.00

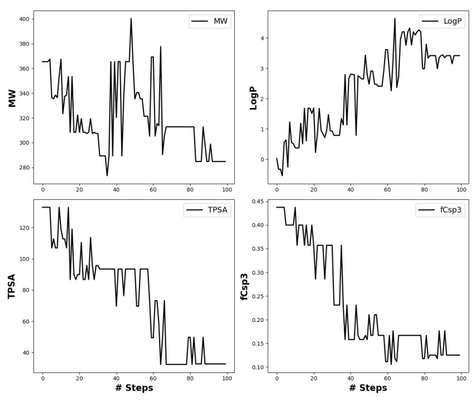

Figure 3: Properties of sampled molecules during linear interpolation in latent space.

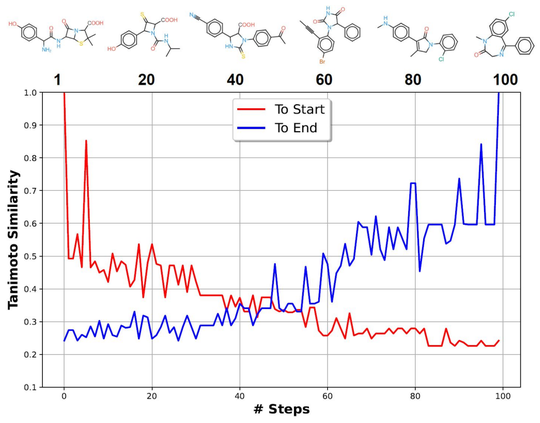

Figure 4: Similarity of sampled molecules during linear interpolation from Amoxicillin to Diazepam.

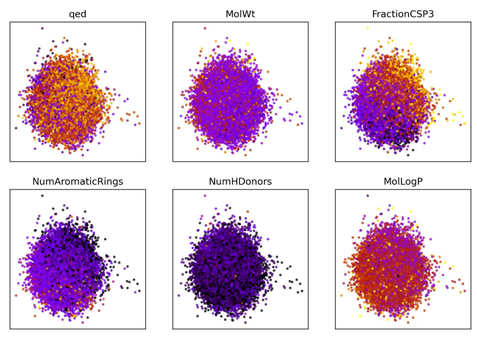

Figure 5: PCA of the GIRAFFE latent space using selected calculated properties relevant for drug discovery projects.

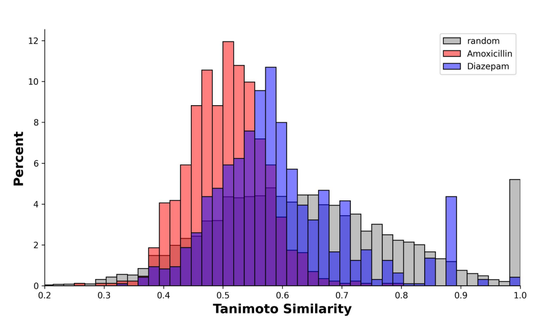

Figure 6: ECFP4 Tanimoto similarity of sampled molecules in GIRAFFE latent space compared to their "parent" compound from the training data.

Conclusions

In the current study, we have achieved the following:

- **Fusion of graph attention neural networks with LSTM and properties in a variational autoencoder framework for molecular representation learning.**
- **Disentanglement of latent space through cyclical annealing of β in VAE loss.**
- **Competitive performance of latent space as descriptor in QSAR benchmarks[6].**
- **High validity and drug-likeness of randomly sampled molecules.**
- **Recreation of chemical space used for training measured as structural similarity and calculated properties.**
- **Robustness for linear latent space interpolation between points of interest.**

Outlook

- Benchmarking the obtained representation on additional datasets.
- Optimization of the presented approach by using additional property endpoints, such as biological assay readouts and physicochemical measurements.
- Employing the GIRAFFE latent space as descriptor to cluster compounds for the organization of compound libraries.
- Incorporating additional priors and constraints to condition the generation of structures on desired scaffolds and properties.
- Getting constructive feedback from the community!

References

1. Xiong, Z *et al. J. Med. Chem.,* 2020, 63, 16, 8749-8760.
2. Hochreiter, S. and Schmidhuber, J., *Neural comput.,* 1997, 9(8), 1735–1780.
3. Kingma, D. P. and Welling, M., *arXiv*, 2013, 1312.6114.
4. Kingma, D. P. and Ba, J.,*arXiv*, 2014, 1412.6980.
5. Fu, H. *et al., NAACL,* 2019.
6. Wu, Z. *et al. Chem. Sci.,* 2018, 9(2):513–530.
7. Winter, R. *et al. Chem. Sci.*, 2019, 10(34):8016–8024.
8. Fabian, B. *et al, arXiv,* 2020, 2011.13230.
9. Wen, N. *et al. J. Cheminform.*, 2022, 14(1): 71.

July 2024