

# Report I : Learning to Extend Molecular Scaffolds with Structural Motifs

Irem Begüm Gündüz - 7026821

## Contents

Overview of the Approach . . . . .	1
Discussion . . . . .	2
References . . . . .	2

## Overview of the Approach

Recent studies suggest that deep learning-based molecular modelling may improve the speed of in silico drug discovery. We have been able to build molecules one atom at a time and connect them together using generative models. There are many types of deep-learning-based modelling approaches present, but each has its own limitations. Graph-based models that generate molecules sequentially employ perfect validity because they can enforce hard chemical constraints such as valence. Compared to SMILES-based methods, this approach is more efficient and easier to use. Generated molecules are required to fit certain valence constraints. One of the common constraints that are considered in graph-based drug discovery projects is to include a scaffold, which is a predefined subgraph.

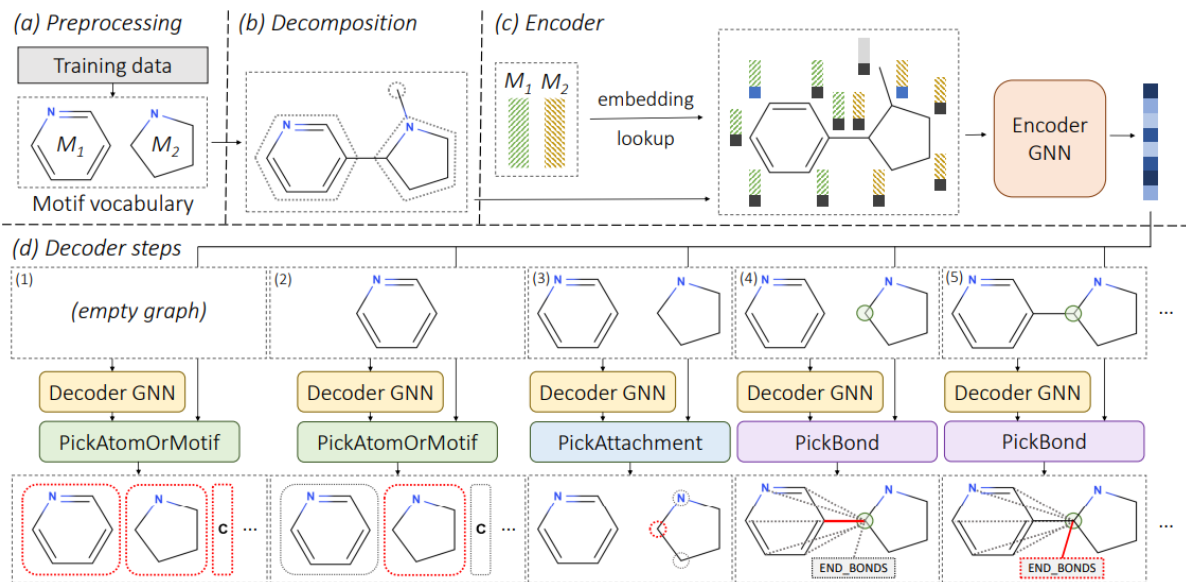


Figure 1: Overview of the methodology.

One of the significant limitations of in-silico drug discovery with deep learning modelling is, that they require a rigid scaffold to be present in order to generate a molecule. However, this existed scaffold may not be always available in rigid form. To overcome this issue, this study proposed a new model *MoLeR*, which is a graph-based model that naturally supports scaffolds as initial seeds for the generation procedure. Maziarz

et al. (2021) *MoLeR* offers a fragment-by-fragment generation of molecules instead of atom-by-atom. Below, the overall framework of *MoLeR* is summarized:

- a) The *MoLeR* method employs common molecular fragments called motifs, which can be used to build molecules fragment-by-fragment. Most partial molecules that are constructed during generation are semantically sensible, meaning they don't contain half-built structures such as partial rings. Thus, the first step of the *MoLeR* method is to discover motifs from data.
- b) The second step includes using the detected motifs to decompose an input molecule into motifs and single atoms.
- c) The deterministic encoder takes features from the bottom layer and embeds them into the top layer, making the motif information available at the atom level.
- d) The last step is the decoder step. Decoder steps are only dependent on the encoder output and the partial graph, so they have to choose one of the valid options.

## Discussion

The *MoLeR* model uses an auto-encoder paradigm to train the model and, graphs to represent molecules. Atoms are represented by vertices, and the connections between them are based on the bonds between them. The *MoLeR* model has a deterministic encoder, which attempts to compress an input molecule into a latent code and, always, returns the maximum likelihood latent code  $z$  (which corresponds to the mean of the predicted Gaussian). The decoder then tries to reconstruct the original molecule using this code.

In order to decompress a short encoding into a graph of any size, the reconstruction process itself is designed to be sequential. As a result, it can complete the graph by adding new atoms or bonds throughout each step. The decoder, in the model makes predictions based on a partial graph and a hidden code at each step. However, what is the hidden code or how it is generated is not stated by the authors.

The model proposed by *MoLeR* is a novel way of generating molecules without relying on past predictions. Besides being able to complete partial or arbitrary scaffolding and, optimizing the scaffolding constraints using motifs, *MoLeR* also outperforms the existing graph-based modelling methods as they are limited by the requirements of scaffolding.

This method brings attention to graph-link prediction. I was wondering if the autoencoders used in *MoLeR* are able to predict the motifs more accurately than the proposed method here Besta et al. (2021). Or replacing motif prediction mechanisms of *MoLeR* with the offered link-prediction framework can increase accuracy or not. How reliable are these predictions, and why not relying on the prediction history improved performance? Finally, is there any change present using *MoLeR* to over-smooth the scaffold constraints? The authors stated even if a molecule does not violate valence constraints, it can still contain unstable or unsynthesizable substructures. With *MoLeR*, they offered an optimization methodology. What are the limitations of smoothing constraints? How can we prevent generating molecules that contain unstable or unsynthesizable substructures? In addition, I would like to see how we can detect if *MoLeR* over-smooths the scaffold constraints. These are the questions that did not specifically point out by the authors or at least, I am not able to answer them completely. The paper needed to state a vocabulary that implies what makes a constraint a good to constrain, when or where should we use optimization, etc.

## References

- Besta, Maciej, Raphael Grob, Cesare Miglioli, Nicola Bernold, Grzegorz Kwasniewski, Gabriel Gjini, Raghavendra Kanakagiri, et al. 2021. "Motif Prediction with Graph Neural Networks," May. <https://doi.org/10.48550/arxiv.2106.00761>.
- Maziarz, Krzysztof, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. 2021. "Learning to Extend Molecular Scaffolds with Structural Motifs," March. <https://doi.org/10.48550/arxiv.2103.03864>.