Chemical Space Exploration via Semi-Supervised Learning with Molecular Graphs

Felipe V. Calderan[1], Juarez L. F. Da Silva[2], and Marcos G. Quiles[1]

[1]Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil
[2]São Carlos Institute of Chemistry, University of São Paulo, São Carlos, SP, Brazil

Problem Characterization

- Chemical Space exploration is a recurrent activity in MS
- Generative ML methods to fulfill this purpose are emerging, such as VAEs
- VAEs' latent spaces' dimensions do not follow properties
- Current graph-based VAEs have big limitations and could be much better

Objectives

- Graph-based VAE with navigable latent space, emphasizing properties
- Find a cost function for the VAE that corroborates with the objective
- Include disentanglement procedures to make navigation even better
- Project an elastic model for the iterative ampliation of the latent space
- Test the developed technology using chemical datasets of interest

Methods - GNN and Message Passing

- Graphs encode molecules well and store more information than typical string representations.
- GNN receives graphs as input and encodes the vertices and edges as feature vectors.
- MP algorithm is applied so vertices are informed about their neighbors

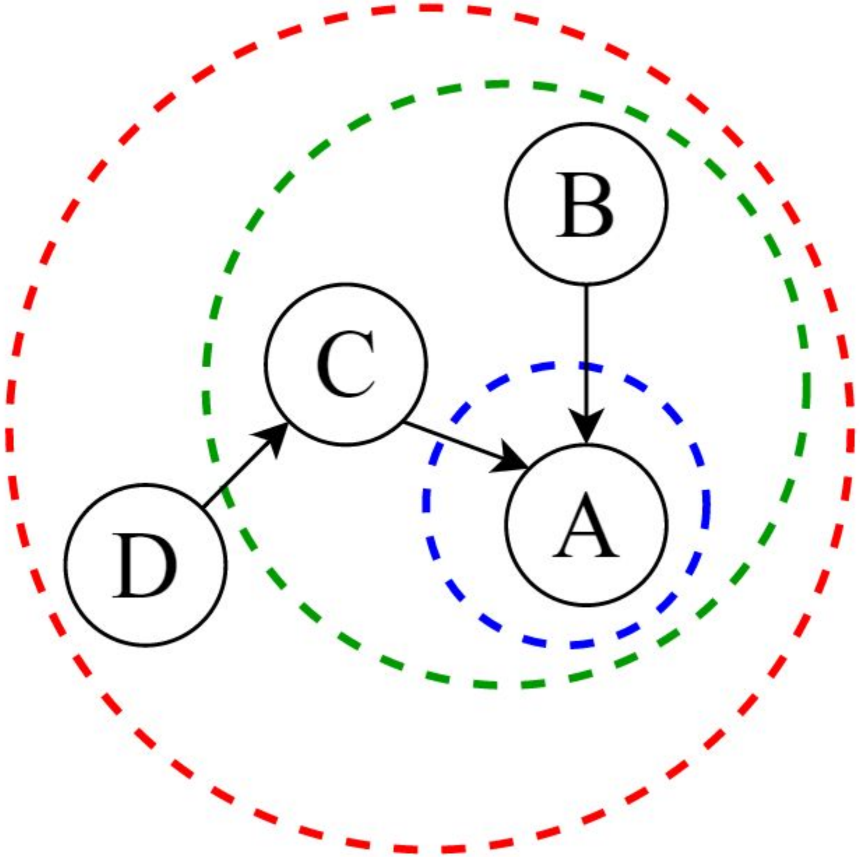

Figure 1: Message Passing representation

Methods - Variational Autoencoder

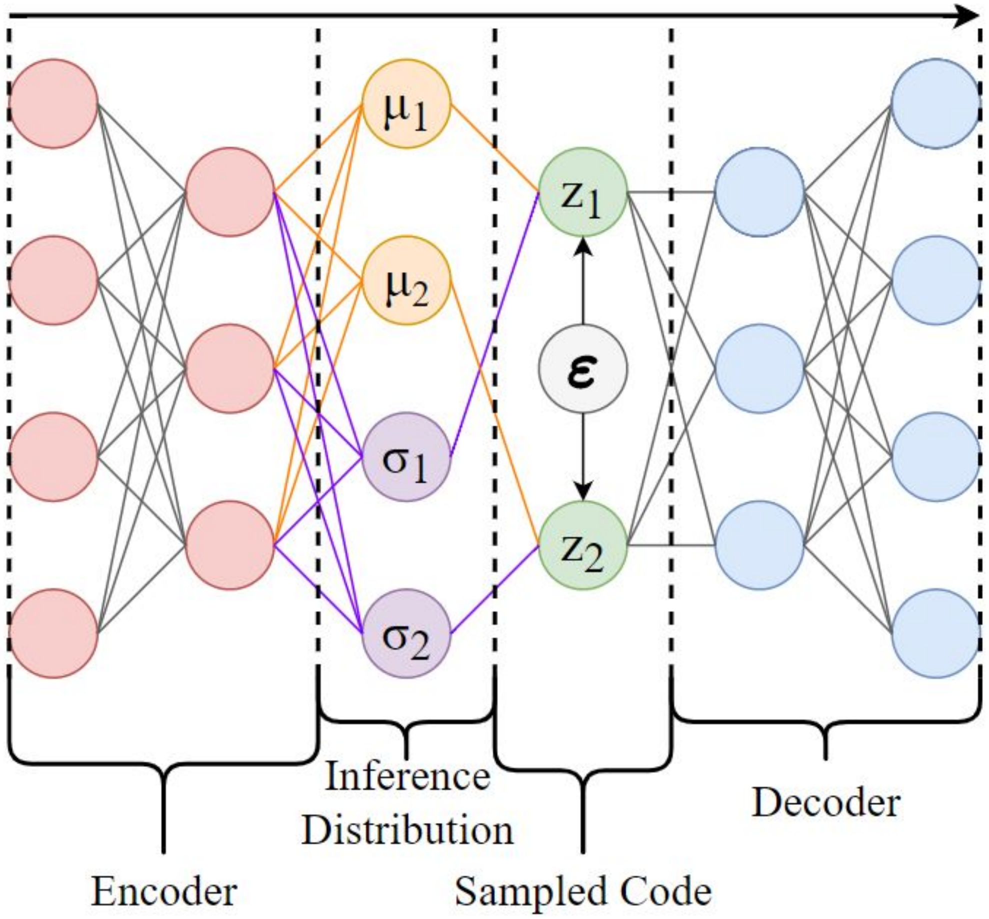

Figure 2: Variational Autoencoder topology

- Autoencoders, by default, work like the identity function
- It is possible to sample data from the latent space of a VAE
- Disentanglement can be applied to a VAE to force different neurons to learn different properties

Methods - Semi-Supervised Autoencoder

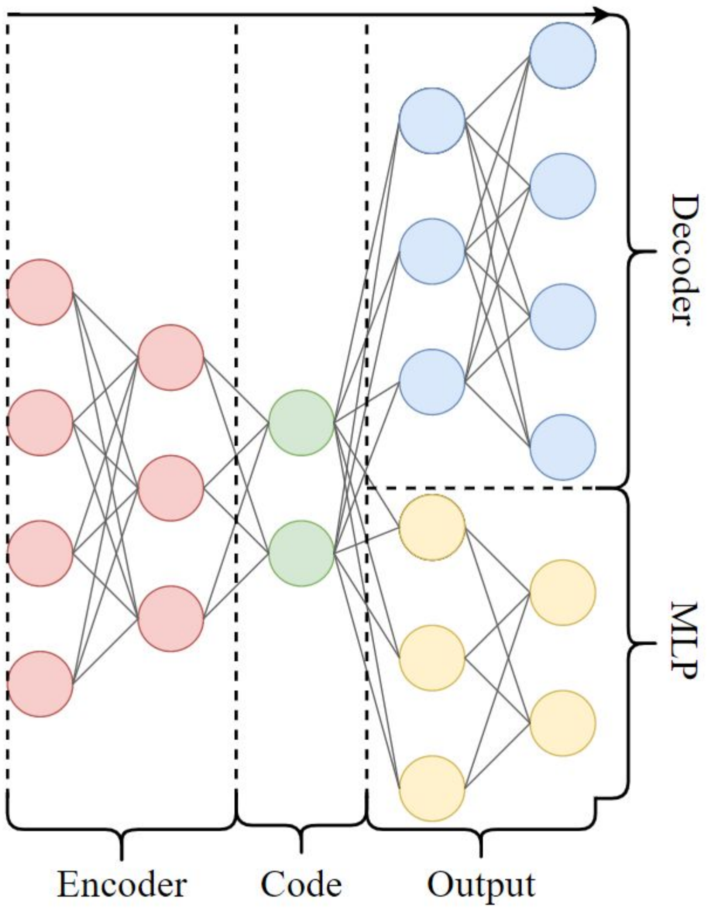

Figure 3: Semi-Supervised Autoencoder topology

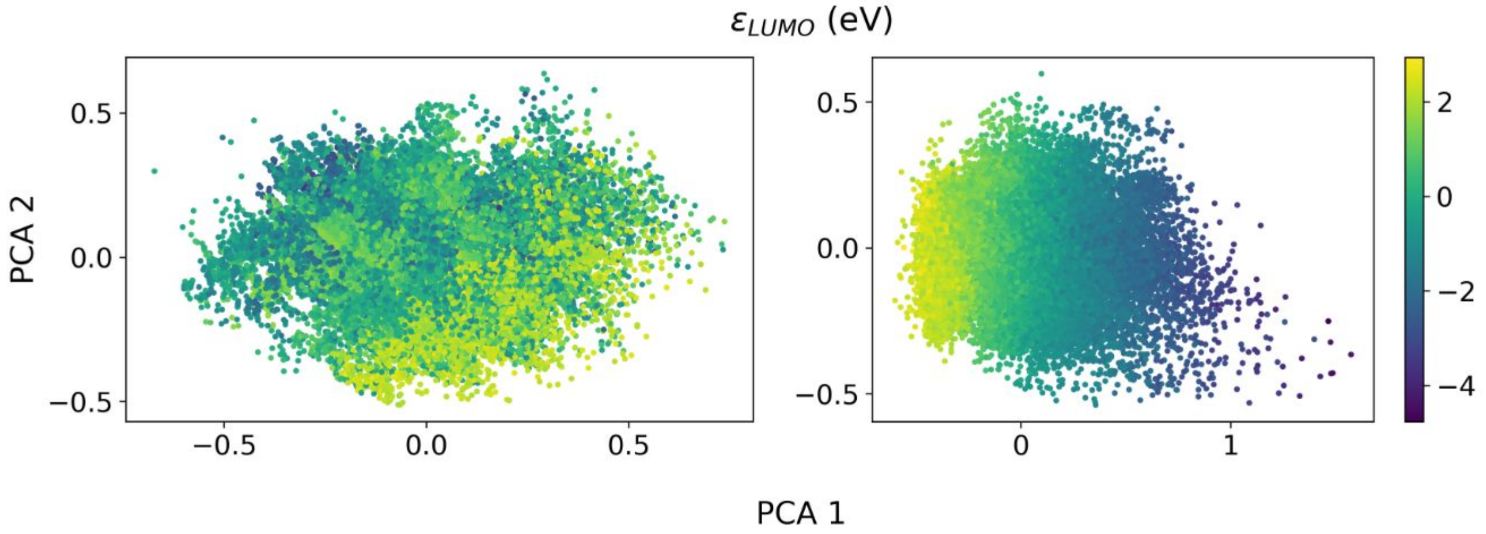

Figure 4: GVAE vs SGVAE (Oliveira, A. F., Da Silva, J. L., Quiles, M. G., 2022)

- It is possible to combine a predictor and a VAE to further organize the latent space

Methods - Grammar AE vs Graph AE

Grammar Autoencoder encoding Alanine represented by SMILES.

Graph Autoencoder encoding Alanine represented by 3D Graph.

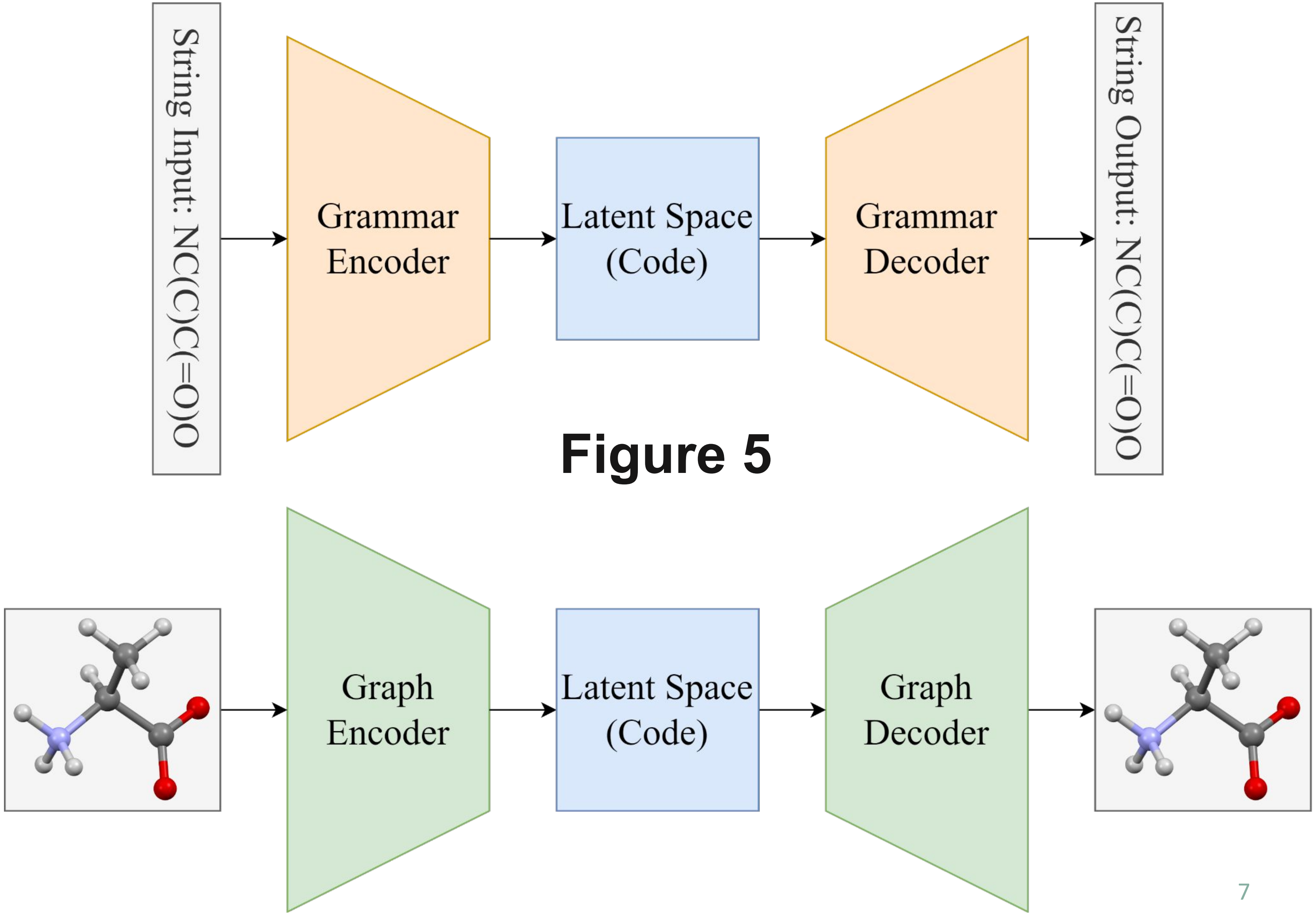

Figure 5

Expected Results

- Semi-supervised learning system capable of finding useful chemical compounds through chemical space exploration
- Reduction on materials screening time (one of the most time-consuming tasks), improving research productivity

Challenges

- Encoding large molecules
- Directing properties
- Configuring hyper-parameters
- Obtaining useful molecules

Acknowledgments

CAPES

FAPESP