FUNDAÇÃO GETULIO VARGAS

ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp

CURSO DE GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

LEARNING DRUG REPRESENTATIONS FOR SIDE EFFECT FREQUENCY PREDICTION

por
Iago Riveiro Santos Dutra

Rio de Janeiro
2025

FUNDAÇÃO GETULIO VARGAS

ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp

CURSO DE GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

LEARNING DRUG REPRESENTATIONS FOR SIDE EFFECT FREQUENCY PREDICTION

"Declaro ser o único autor do presente projeto de monografia que se refere ao plano de trabalho a ser executado para continuidade da monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador".

Iago Riveiro Santos Dutra

Orientador: Alberto Paccanaro

Rio de Janeiro
2025

IAGO RIVEIRO SANTOS DUTRA

LEARNING DRUG REPRESENTATIONS FOR SIDE EFFECT FREQUENCY PREDICTION

"Projeto de Monografia apresentado à Escola de Matemática Aplicada – FGV/EMAp como requisito parcial para continuidade ao trabalho de monografia."

Aprovado em _____ de _____ de 2025.
Grau atribuído ao Projeto de Monografia: _____ .

Professor Orientador: Alberto Paccanaro
Escola de Matemática Aplicada – FGV/EMAp
Fundação Getulio Vargas

Sumário

1 INTRODUCTION

The production of simple numerical embeddings from complex data has enabled the advance of many branches of machine learning. For example, at the core of the large language models (LLMs) are text tokenizers and transformers, an efficient way of enriching representation vectors with context from the whole sentence. The process of summarizing the input in relatively low-dimensional vectors is crucial for the success of downstream tasks, such as next word prediction, question answering and so on.

Much like natural language processing (NLP) applications, the study of chemicals relies on some kind of process capable of digesting molecular structures, which usually are modelled as graphs (where each atom is a node and each bond is an edge), generating embedding vectors. In this domain, some of the NLP challenges reappear. For instance, molecules have varying amounts of atoms, whereas sentences have different amounts of words. Moreover, atoms, like words, have no “natural” numeric representation. It is up to researchers to determine the best way of extracting useful initial features for the algorithm.

The purpose of this study, thus, is to evaluate different algorithms capable of generating these embeddings and discuss the relevant strengths and limitations of each one. In the concerning literature, some of most used models are Variational Auto-Encoders (VAEs), Transformers, Graph-Convolutional Neural Networks (GCNNs) and Message-Passing Graph Neural Networks (MPGNNs). The baseline strategy for the embedding production is to extract relevant chemical features to create what are called “fingerprints”. These usually encode information like the number of aromatic rings, the number of atoms, the number of double bonds, the number of acidic hydrogen atoms, and the number of hydroxyl groups. The foundational question motivating this project is whether these algorithms are capable of producing more expressive embeddings while maintaining the explainability of the representation. In particular, the proposed task is to predict the frequency of the side effects caused by drugs using minimal chemical information.

2 EXPLORATORY LITERATURE REVIEW

2.1 Distributed Representations and Numerical Embeddings

The origins of what would become modern machine learning can be traced to the early works on brain memory modelling. One of the intriguing features of biological memory is that retrieval is content-based, instead of address-based as is the case with computers. Furthermore, a piece of information can be recovered from partially contradictory or wrong cues. For example, one might recall oneself of that “tall blond man with glasses we met last week” even if the referred person didn’t in fact wear glasses.

In that context, Hinton, McLelland and Rumelhart (1986) discussed a model inspired by cognitive psychology to represent a set of entities using a network of computing elements. Unlike local representations, where each unit conveys a single concept, distributed representations encode entities as patterns of activation across all units. This approach takes advantage of the correlation between entities, as well as enables the description of relations of type inheritance (e.g. gorillas are types of apes) and composition (e.g. men with wooden legs).

The learning process directs individual units into encoding “microfeatures” of an item, with connection strengths meaning plausible “micro-inferences”. This structure allows the emergence of several desirable properties. For instance, new entities can be incorporated by the model without the need for additional processing units, and the failure of some units, simulating the death of some neurons in the brain, doesn’t completely hinder the representation of an entity. Moreover, the learnt representations are robust to a broad range of perturbations, be it the addition of a Gaussian error in the network weights or the removal of some units.

These initial ideas laid the foundations for much of the current machine learning paradigm. A sign of the importance of the topic is the fact that the seminal paper that introduced backpropagation as a means to train neural networks by Rumelhart, Hinton and Williams (1986) is titled “Learning representations by back-propagating errors”. More recently, Alfonso (2025) has explored applying deep-learning approaches to create drug embeddings capable of capturing relevant chemical information. While the employed methods differ greatly from those from the 80’s, the overall purpose remains the same: mapping entities to numerical representations that can be easily processed.

2.2 Representation Learning and Generative Models

Traditionally, representation learning is the task of mapping data to a simple underlying distribution, whereas generative modelling is the task of learning the implicit distribution from which data comes, i.e., mapping "noise" to data. Each problem can be seen as the inverse of the other (TORRALBA; ISOLA; FREEMAN, 2024).

A simple way of ensuring that generated samples are similar to input data is training the model to be able to reconstruct the observed examples. Training an encoder-decoder with the reconstruction loss makes it possible to sample new datapoints from the latent space. A good training schema guides the model towards learning a structured latent space: one in which relevant features of the data are directions. Recent works have explored how to use the geometrical properties of the learned space to guide generation towards specific goals by interpolating embeddings, for example (ZHANG *et al*., 2023).

For numerical, multidimensional data, a canonical statistical method for producing low-dimensional representations is Principal Component Analysis (PCA), which identifies directions of maximal variance in the data. In the context of deep learning, a multilayer perceptron (MLP) encoder can be seen as a nonlinear generalization of PCA. Unlike linear techniques, neural network-based encoders can capture complex, hierarchical structures in the data, making them suitable for domains where the underlying generative factors are entangled and nonlinear.

2.3 Representing Drugs

Nomenclature always has been a major issue in chemistry. Previous to the efforts by Friedrich Kekulé in 1860 to create international standards for the naming of organic compounds, reproducibility of experiments was often hindered by name mixing. Nowadays, the IUPAC (International Union of Pure and Applied Chemistry) organizes guidelines for the unambiguous naming of compounds.

The system designed by IUPAC, however, is unpractical for computational manipulation. Many standards have been proposed to address this question. Among them, the most relevant are InChI (International Chemical Identifier), SMILES (Simplified Molecular Input Line Entry System), SAFE (Sequential Attachment-based Fragment Embedding), SELFIES (SELF-ReferencIng Embedded Strings) (HELLER *et al*., 2015; WEININGER, 1988; NOUTAHI *et al*., 2023; KRENN *et al*., 2020). These are all string

representations that aim to represent the molecular structure unambiguously while enforcing certain properties like order-invariance, chemical validity or representation uniqueness. Although it is a much discussed topic, there's still no agreement on the best choice (LEON *et al*., 2024). Due to its simplicity, SMILES is usually the standard notation.

2.4 The Standard Bioinformatics Approach: Fingerprints

Applications of machine learning for biology and medicine often rely on computing so called fingerprints as inputs for models. These fingerprints often include counts of specific substructures and fragments, as well as previous knowledge about the substance coming from experimental data. Not only are they often task-specific, but relevant information may be overlooked. The attempt to encode as much information as possible may lead to extremely big fingerprints. For example, the Morgan Fingerprint has a default dimension of 2048 in RDKit's implementation. The literature on chemical fingerprints is as vast as it is dispersed. A few representative examples may be found in Thanh-Hoang *et al*. (2020), Xu *et al*. (2017) and Boldini *et al*. (2024).

2.5 Variational Auto-Encoders (VAEs)

Auto-Encoders encompass a broad range of models with the goal of summarizing an observed datapoint into a descriptive measure in a way that preserves information. The quality of the data-representation mapping (encoder) is measured by the capacity of a simultaneously trained decoder to recover the initial input.

Variational Auto-Encoders learn the distribution of data by applying what has become known as the "reparametrization trick". Once a datapoint is encoded, its embedding vector is treated as the parameters for a multivariate normal distribution. The input that will be passed for the decoder to reconstruct the datapoint must come from that distribution, but directly sampling from it would break backpropagation (since "sampling" is non-differentiable). We then proceed to sample from a standard Gaussian and transform it via by translating and scaling using the encoder-mapped parameters for that specific datapoint. Formally, VAEs learn a continuous mapping that can be expressed as a mixture of an infinite number of Gaussians (TORRALBA; ISOLA; FREEMAN, 2024).

VAEs can be applied to tasks such as dimensionality reduction, image compression and, thanks to its variational approach, data generation. Another important improvement over non-variational auto-encoders is the induced continuity of the embedding space. Deterministic auto-encoders may map data to specific points in space surrounded by noise instead of similar data, creating “holes” in the latent space (MAZUMDER *et al*., 2024).

2.6 Transformers

Transformers are attention-based auto-regressive encoder-decoder models. They first appeared in the now famous article titled “Attention is all you need” by Vaswani *et al*. (2017). Proposed as an alternative to Gated Recurrent Units (GRUs) and Long Short-Term Memory Recurrent Neural Networks (LSTMs), the main feature of transformers is their high scalability due to the parallelizability of hte attention mechanism, allowing a faster training. The success was such that they inspired the application of similar tokenization and processing to other domains beyond NLP. For instance, Vision Transformers (ViTs) are usually treated as benchmark models in computer vision (DOSOVITSKIY *et al*., 2021).

Many alternatives have been proposed to regular Transformers, such as Differential Transformers (YE *et al*., 2024) and, more recently, Titans (BEHROUZ; ZHONG; MIRROKNI, 2025). Although these contributions present desirable properties and may potentialize learning, Transformers still are the predominant state-of-the-art architecture. Thus, for the purpose of scope delimitation, in the following comparisons, we shall stick to the original model.

2.7 Graph Neural Networks (GNNs)

The usage of deep learning for graph-related problems is an ongoing theme in the machine learning community. The nature of the tasks can vary greatly depending on the specific problem. There are intra-graph problems, such as edge prediction and node classification, as well as inter-graph problems, such as determining the boiling point of a molecule. For each of these purposes, several models have been proposed. Since our problem of interest is that of extracting relevant features for a set of graphs in a dataset, we will focus on the major models suitable for that. Those are Graph-Convolutional Neural Networks (GCNNs) and Message-Passing Graph Neural Networks (MPGNNs).

First proposed in 2017 (Kipf & Welling), GCNNs take inspiration from Convolutional Neural Networks, which had, by the time of the publication of the paper, become the gold standard for computer vision tasks. The idea is to perform a convolution in the node feature matrix of the graph but masking it using the adjacency matrix so that a node is updated by the convolution applied to itself and its neighbors.

Similarly, MPGNNs act by updating a node's embedding based on information from neighbors. At each layer, nodes send and receive messages which are then aggregated (e.g., via summation, mean, or attention mechanisms) and used to update their own representations. This message-passing framework enables the model to capture both local and global structural information over multiple iterations. MPGNNs are highly flexible and form the foundation for a wide range of graph-based models, including Graph Attention Networks (GATs), GraphSAGE, and Graph Isomorphism Network (GIN) (VELIČKOVIĆ *et al*., 2017; HAMILTON; YING; LESKOVEC, 2017; XU *et al*., 2018).

2.8 Other Generative Models

A great number of novel architectures have been proposed in order to circumvent the shortcomings of the mentioned models. An exhaustive exploration goes beyond the scope of the current work, but a preliminary introduction is useful to provide a broader landscape of the arrea.

2.8.1 Generative Adversarial Networks (GANs)

Goodfellow *et al*. (2014) introduced GANs as a general training framework that dispenses the need for Markov Chains or explicit likelihood inference. The main idea is to train two models: a generator and a discriminator. The role of the first is to create plausible data examples, while the later tries to distinguish real datapoints from those created by the generator. In their words:

> "The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles."

GANs notoriously suffer from unstable training, and much of the related literature concerns techniques for improving convergence (WIATRAK; ALBRECHT; NYSTROM, 2020).

2.8.2 Energy-Based Models (EBMs)

EBMs define a probability distribution over data by associating a scalar energy to each configuration, with lower energy values corresponding to more likely configurations. Instead of explicitly modeling the likelihood, EBMs learn an energy function that assigns low energy to observed data and higher energy to other inputs. Training typically involves contrastive methods such as contrastive divergence or score matching. EBMs are flexible and can capture complex dependencies, but inference and sampling can be computationally expensive, often requiring techniques like Markov Chain Monte Carlo (MCMC) (SONG; KINGMA, 2021).

2.8.3 Normalizing Flows

Normalizing Flows are a class of generative models that transform a simple base distribution (e.g., a standard Gaussian) into a complex target distribution through a sequence of invertible and differentiable mappings. Each transformation is designed to have a tractable Jacobian determinant, allowing for exact likelihood computation via the change of variables formula. This makes flows capable of combining efficient sampling and density estimation, but a major downside is the fact that the dimension cannot be easily changed between layers (KOBYZEV; PRINCE; BRUBAKER, 2020).

2.8.4 Generative Flow Networks (GFlowNets)

GFlowNets are a recent class of models that learn to sample discrete sequences by modeling the process as a sequence of stochastic decisions. The aim is to generate diverse outputs proportional to a given reward function, rather than maximize likelihood. The problem is usually presented using a reinforcement learning paradigm of state-action optimization, and a flow-matching objective is used to ensure that the probability of reaching a given state is proportional to its reward (SILVA *et al.*, 2025).

2.8.5 Diffusion Models

Diffusion models learn to generate data by reversing a gradual noising process. During training, data is corrupted over many steps by the progressive addition of Gaussian noise, forming a Markov chain. The model then learns to reverse this process (“denoise”)

step by step, ultimately generating realistic data samples from pure noise. Unlike GANs, diffusion models are stable to train and produce high-quality, diverse outputs, especially in image and audio generation tasks. Recent advancements, such as Denoising Diffusion Probabilistic Models (DDPMs) and score-based generative models, have established diffusion methods as state-of-the-art in many generative benchmarks (HO; JAIN; ABBEEL, 2020; SONG; MENG; ERMON, 2020). Notoriously Dall-E and Stable Diffusion were the first major commercial applications that achieved significant image generation quaity, and both are based on a diffusion scheme.

3 DEVELOPMENT AND EXPECTED RESULTS

The contribution of this work lies in comparing the representation space that each method learns and evaluate its utility for the specific task of drug side effect frequency prediction. It is expected that this project will make clear the differences in terms of expressivity and predictive power among the studied techniques. Since no ethical issues may arise from the treatment of data that will be performed and only computational experiments are required, the project is feasible with the already available infrastructure.

Although many studies have addressed the comparison of algorithms for embedding generation, none have specifically focused on these group of methods while investigating the corresponding representation spaces. Wang *et al*. (2023) discuss adequate criteria for the evaluation of molecular graph embeddings, but don't investigate the models of interest. O'Boyle & Sayle (2016) compared different fingerprints to create a benchmark for the creation of improved fingerprints for virtual screening, without mentioning deep-learning methods. Boldini *et al*. (2024) used a dataset of natural products to inspect different molecular fingerprints and concluded that the induced spaces vary significantly, which poses the question of finding a possibly universal congruent fingerprint. Finally, Zagidullin *et al*. (2021) analyze similar algorithms, but don't explore the embedding spaces.

4 METHODOLOGY

The tests will be carried out using a public dataset put together by Galeano *et al*. (2020) and the computational infrastructure present on campus. The data consists of the frequency scores for 750 drugs and 994 side effects. The scores are integers going from 0 to 5 where 0 means that no association is known for the drug-side effect pair, 1 means that the occurrence of the referred side effect for that drug is very rare, up to 5, which means that the side effect is very frequent.

The models will be implemented using the original articles as reference, adapting the algorithm to the problem when necessary. The evaluation of the models will take in consideration the best practices of train/validation/test division of the dataset.

The metrics to determine the best model will be the Root Mean Squared Error (RMSE) for the regression task, i.e., trying to recover the exact score, and Area Under the Receiver Operating Characteristic (AUROC) for the classification task, i.e., predicting whether an association exists or not. The training loss for each task will be the RMSE and the Binary Cross-Entropy (BCE), respectively, with weight decay as regularization. Furthermore, since the dataset is highly imbalanced (only about 4,97% of entries are nonzero), negative examples will be weighted down in the loss to avoid the collapse of the predictions' distribution.

A set of chosen example molecules will be used to qualitatively explore the embedding space: the embedding vectors of these molecules will be projected in three dimensions using principal component analysis (PCA) in order to enable the visualization of their distribution. Furthermore, the embeddings will be evaluated in terms of cosine similarity between similar and dissimilar molecules.

5 REFERENCES

ALFONSO, Aldo Javier Galeano. **Learning Object Representations to Predict and Explain Missing Associations**. PHD Thesis (Doutorado em Matemática Aplicada e Ciência de Dados) – Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, 2025.

BEHROUZ, Ali; ZHONG, Peilin; MIRROKNI, Vahab. Titans: Learning to Memorize at Test Time. **arXiv**, 31 Dec. 2024. https://doi.org/10.48550/arxiv.2501.00663.

BOLDINI, Davide; BALLABIO, Davide; CONSONNI, Viviana; TODESCHINI, Roberto; GRISONI, Francesca; SIEBER, Stephan A. Effectiveness of molecular fingerprints for exploring the chemical space of natural products: Journal of Cheminformatics. **Journal of Cheminformatics**, vol. 16, no. 35, p. 1–16, 25 Mar. 2024. https://doi.org/10.1186/s13321-024-00830-3.

DOSOVITSKIY, Alexey; BEYER, Lucas; KOLESNIKOV, Alexander; WEISSENBORN, Dirk; ZHAI, Xiaohua; UNTERTHINER, Thomas; DEHGHANI, Mostafa; MINDERER, Matthias; HEIGOLD, Georg; GELLY, Sylvain; USZKOREIT, Jakob; HOULSBY, Neil. **AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE**. [*S. l.*: *s. n.*], 3 Jun. 2021.

GALEANO, Diego; LI, Shantao; GERSTEIN, Mark; PACCANARO, Alberto. Predicting the frequencies of drug side effects. **Nature Communications**, vol. 11, no. 1, p. 4575, 11 Sep. 2020. https://doi.org/10.1038/s41467-020-18305-y.

GOODFELLOW, Ian J; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. Generative Adversarial Networks. **arXiv**, 10 Jun. 2014. https://doi.org/10.48550/arxiv.1406.2661.

HAMILTON, William; YING, Rex; LESKOVEC, Jure. Inductive Representation Learning on Large Graphs. **arXiv**, 7 Jun. 2017. https://doi.org/10.48550/arxiv.1706.02216.

HELLER, Stephen R; MCNAUGHT, Alan; PLETNEV, Igor; STEIN, Stephen; TCHEKHOVSKOI, Dmitrii. InChI, the IUPAC International Chemical Identifier. **Journal of Cheminformatics**, vol. 7, no. 1, 30 May 2015. https://doi.org/10.1186/s13321-015-0068-4.

HINTON, G. E.; McCLELLAND, J. L.; RUMELHART, D. E. Distributed Representations. *In* : RUMELHART, D. E; McCLELLAND, J. L.; PDP Research Group. **Parallel Distributed Processing**. Bradford Books. ISBN: 9780262680530. 1986.

HO, Jonathan; JAIN, Ajay; ABBEEL, Pieter. Denoising Diffusion Probabilistic Models. **arXiv**, 19 Jun. 2020. https://doi.org/10.48550/arXiv.2006.11239.

KIPF, Thomas N; WELLING, Max. Semi-Supervised Classification with Graph Convolutional Networks. **arXiv**, 1 Jan. 2016. https://doi.org/10.48550/arxiv.1609.02907.

KOBYZEV, Ivan; PRINCE, Simon J. D.; BRUBAKER, Marcus A. Normalizing Flows: An Introduction and Review of Current Methods. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 43, no. 1, p. 3964–3979, 2020. https://doi.org/10.1109/TPAMI.2020.2992934.

KRENN, Mario; HÄSE, Florian; NIGAM, Akshat Kumar; FRIEDERICH, Pascal; ASPURU-GUZIK, Alán. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. **Machine Learning: Science and Technology**, vol. 1, no. 4, p. 045024, 3 Nov. 2020. https://doi.org/10.1088/2632-2153/aba947.

LEON, Miguelangel; PEREZHOHIN, Yuriy; PERES, Fernando; POPOVIČ, Aleš; CASTELLI, Mauro. Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. **Scientific Reports**, vol. 14, no. 1, 23 Oct. 2024. https://doi.org/10.1038/s41598-024-76440-8.

MAZUMDER, Alokendu; BARUAH, Tirthajit; KUMAR, Bhartendu; SHARMA, Rishab; PATTANAIK, Vishwajeet; RATHORE, Punit. Learning Low-Rank Latent Spaces with Simple Deterministic Autoencoder: Theoretical and Empirical Insights. **2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, p. 2839–2848, 3 Jan. 2024. https://doi.org/10.1109/wacv57701.2024.00283.

NOUTAHI, Emmanuel; GABELLINI, Cristian; CRAIG, Michael; JONATHAN; TOSSOU, Prudencio. Gotta be SAFE: A New Framework for Molecular Design. **arXiv**, 1 Jan. 2023. https://doi.org/10.48550/arxiv.2310.10773.

O'BOYLE, Noel M.; SAYLE, Roger A. Comparing structural fingerprints using a literature-based similarity benchmark. **Journal of Cheminformatics**, vol. 8, no. 36, 5 Jul. 2016. https://doi.org/10.1186/s13321-016-0148-0.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, vol. 323, no. 6088, p. 533–536, Oct. 1986. https://doi.org/10.1038/323533a0.

SILVA, Tiago; ALVES, Rodrigo Barreto; DA SILVA, Eliezer de Souza; SOUZA, Amauri H; GARG, Vikas; KASKI, Samuel; MESQUITA, Diego. When do GFlowNets learn the right distribution? **ICLR**, 2025.

SONG, Jiaming; MENG, Chenlin; ERMON, Stefano. Denoising Diffusion Implicit Models. **arXiv**, 6 Oct. 2020. https://doi.org/10.48550/arxiv.2010.02502.

SONG, Yang; KINGMA, Diederik P. How to Train Your Energy-Based Models. 17 Feb. 2021. **arXiv**. https://doi.org/10.48550/arXiv.2101.03288.

THANH-HOANG NGUYEN-VO; NGUYEN, Loc; DO, Nguyet; LE, Phuc H; NGUYEN, Thien-Ngan; NGUYEN, Binh P; LE, Ly. Predicting Drug-Induced Liver Injury Using Convolutional Neural Network

and Molecular Fingerprint-Embedded Features. **ACS omega**, vol. 5, no. 39, p. 25432–25439, 22 Sep. 2020. https://doi.org/10.1021/acsomega.0c03866.

TORRALBA, Antonio; ISOLA, Phillip; FREEMAN, William T. **Foundations of Computer Vision**. [*S. l.*]: MIT Press, 2024.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. Attention Is All You Need. 5 Dec. 2017. **arXiv**. DOI https://doi.org/10.48550/arXiv.1706.03762.

VELIČKOVIĆ, Petar; CUCURULL, Guillem; CASANOVA, Arantxa; ROMERO, Adriana; LIÒ, Pietro; BENGIO, Yoshua. Graph Attention Networks. **arXiv**, 30 Oct. 2017. https://doi.org/10.48550/arxiv.1710.10903.

WANG, Hanchen; KADDOUR, Jean; LIU, Shengchao; TANG, Jian; KUSNER, Matt; LASENBY, Joan; LIU, Qi. Evaluating Self-Supervised Learning for Molecular Graph Embeddings. **arXiv**, 1 Jan. 2022. https://doi.org/10.48550/arxiv.2206.08005.

WEININGER, David. SMILES, a chemical language and information system. **Journal of Chemical Information and Modeling**, vol. 28, no. 1, p. 31–36, 1 Feb. 1988. https://doi.org/10.1021/ci00057a005.

WIATRAK, Maciej; ALBRECHT, Stefano V.; NYSTROM, Andrew. Stabilizing Generative Adversarial Networks: A Survey. 24 Mar. 2020. **arXiv**. https://doi.org/10.48550/arXiv.1910.00927.

XU, Keyulu; HU, Weihua; LESKOVEC, Jure; JEGELKA, Stefanie. How Powerful are Graph Neural Networks? **arXiv**, 1 Oct. 2018. https://doi.org/10.48550/arxiv.1810.00826.

XU, Zheng; WANG, Sheng; ZHU, Fayan; HUANG, Junzhou. Seq2seq Fingerprint. **Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics (ACM-BCB '17)**, 20 Aug. 2017. https://doi.org/10.1145/3107411.3107424.

YE, Tianzhu; DONG, Li; XIA, Yuqing; SUN, Yutao; ZHU, Yi; HUANG, Gao; WEI, Furu. Differential Transformer. **arXiv**, 7 Oct. 2024. https://doi.org/10.48550/arxiv.2410.05258.

ZAGIDULLIN, B; WANG, Z; GUAN, Y; PITKÄNEN, E; TANG, J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. **Briefings in Bioinformatics**, vol. 22, no. 6, 17 Aug. 2021. https://doi.org/10.1093/bib/bbab291.

ZHANG, Yingji; CARVALHO, Danilo S; PRATT-HARTMANN, Ian; FREITAS, André. LlaMaVAE: Guiding Large Language Model Generation via Continuous Latent Sentence Spaces. **arXiv**, 1 Jan. 2023. https://doi.org/10.48550/arxiv.2312.13208.