Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

R&D Project Proposal

Developing Molecular Representations for Machine Learning

Zuha Karim

Supervised by

Dr. Karl N. Kirschner
Dr. Sebastian Houben

December 2021

1 Introduction

Drug pipeline is a process which involves research and other activities from discovering a molecule to the launch of a drug in the market approved by the authorities. One of the most important translational science activities include drug development and drug discovery as it contributes to human health and growth. The drug pipeline consists of several phases. Each phase is further divided into other phases. Phase 0 consists of discovery & development. Phase 1 consists of preclinical research. Phase 2 consists of clinical development , phase 3 review by authorities, phase 4 consists of post-market monitoring.

This whole process of drug discovery can take approximately 15 years [1] along with unsustainable cost greater than $2 million [2]. It is risky, time consuming and a complicated process, in some cases the cost exceeds to $500 million[7]. The costs of drug development even in the initial stages are very high. The need to lower the cost and time of drug development is increasing. Not only that, third world countries, that have low pharmaceutical investment returns, do not have enough resources for investing in R&D [6].

The ultimate goal of drug discovery is to produce a new medicine in the market which has been tested and proves to cure the illness. This industry plays a vital role in human health.

Companies spend an average of $1 billion and around 13 years in the discovery of a drug. According to a study[Machine learning applications in drug development] from 2016 to 2018, the total development cost per approved medicine nearly doubled (from $1,477 to $2,168 million), nearly doubling in eight years (from 2010 to 2018). Re-purposing pharmaceuticals drugs / compounds is a less dangerous and less expensive process than finding new ones. Reusing drugs can be a solution to shorten the time and lessen the costs of drug development. This process can be optimized by using machine learning and other AI methods. Machine learning and artificial intelligence is speeding up the drug pipeline process[6]. Recent breakthroughs in deep learning, have led to a slew of new drug discovery applications.[2]. Automating the initial process would be really beneficial. In instance, if machine learning could be used to mine current databases for candidate compounds that are likely to have desirable features, drug discovery would be faster and more efficient.[7]

Modern machine learning approaches may now be utilized to model and construct artificial intelligence algorithms that effectively predict how chemical alterations or conformations might affect biological activity of compounds. Many thermochemical and physiochemical properties of compounds have been successfully predicted[8].

The digital encoding employed for each molecule that serves as input for training the deep learning model is referred to as the molecular representation. Molecular representation can be in a variety of ways, each with its own set of advantages and disadvantages in terms of accuracy and computational efficiency.[4]. The aim

to this research is to perform comparative analysis of different machine learning models for generating multiple 3D confirmations based, focusing on their input.

1.1 Problem Statement

Reuse the existing drugs, to reduce the time and cost of drug development is a solution which is gaining popularity.[1] Machine learning models are heavily being used for this repurposing. Finding similarity, between molecules can help predict if both the molecules poses similar properties and can be utilized for treating same illness. To check the similarity, different conformations of each molecule are required.

The aim of this research is to create a neural network model which can produce multiple 3D conformations of the molecules. There has not been much research when it comes to generating multiple conformations in 3D using machine learning model [11]. This research will use 2-3 existing input featurization ways in the created NN model, compare them based on the 3D conformations generated by them. Evaluate the conformations. The outcome of the project is to create multiple conformations candidates of a molecule.

The evaluation metric used in most of the models for generating conformation is the shortest distance, which is based on the RMSD metric.

2 Related Work

2.1 DGSM

Dynamic Graph Score Matching is an approach that predicts the molecular conformation. The previous works only consider local interactions but not the long range ones. This model considers both and performs better. It follows a pattern of an existing work that involves learning the gradient of atomic coordinates. [5]

2.2 ConfGF

This paper presents an approach which is called ConfGF. It is based on Langevin dynamics. The aim of this approach is to predict 3D molecular representations in a single stage as other works involve prediction in more than one stage. It calculates the gradient field of atomic coordinates of molecules. These gradient fields assist in generating the conformations. Upon comparison it with the previous state-of-the-art approaches it outperformed. [14]

2.3 CGCF

This work generates valid conformations using probabilistic framework by a molecular graph. It is a combination of two existing models flow-based and energy-based models. It tries to address an issue of large molecules, the long-distance dependencies of atoms.[16]

2.4 GRAPHDG

This paper describes a probabilistic approach Graph Distance Geometry GRAPHDG, that uses graph representations of molecules to create molecular conformations. Learning methodology is based on Euclidean distance geometry. An exisitng approach CVAE is used for calculating distributions for distance.

2.5 GEOM

The Geometric Ensemble Of Molecules assists in developing models that are capable of predicting 3D molecular representations. It is a dataset that contains 33 million molecular conformers. The dataset consists of some properties like relative energies.[3]

3 Project Plan

3.1 Work Packages

Figure 1:

Work Packages	Tasks/Milestones
Literature Survey	1. Literature search on different featurization ways. 2. Literature search for machine learning models. 3. Literature search on models to produce multiple 3D conformations. 4. Understanding the state-of-the-art approaches for building a new model 5. Documentation of literature search. **M1: Perform in depth literature search**
Generating a model	1. Implementing the knowledge gained from literature search to build featurization vector. 2. Build a learning model using any of the machine learning state of the art approaches. 3. Optimize the model to be able to take more than one way of input. 4. Generate multiple 3D conformations of a molecule. **M2: Generating a model**
Evaluation of model	1. Take a featurization vector and generate conformations. 2. Take another way and then implement the model. 3. Select appropriate metrics for comparison. 4. Perform comparative analysis of the models. **M3: Evaluation of the build machine learning model with other models.**
R&D Project Report	1. Create first draft of the report **M4: Create first draft.** 2. Add improvements if necessary **M5: Create final report**

3.2 Milestones

M1 Perform in depth literature search

M2 Implementing a model to generate 3D conformations.

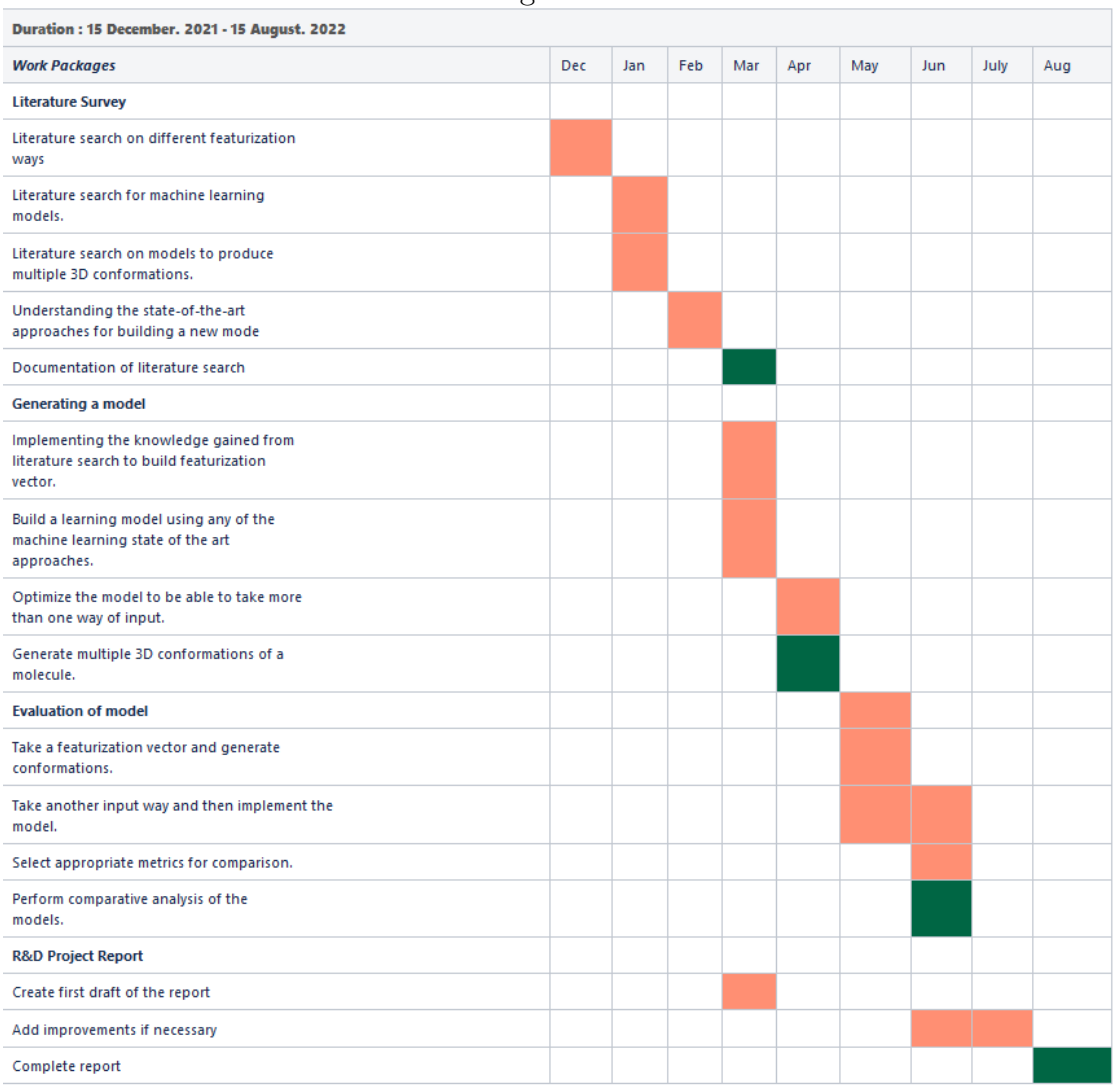
M3 Evaluation of the build machine learning model with other models.

M4 First draft of R&D report

M5 Add improvements and prepare Final R&D report

3.3 Project Schedule

Figure 2:



3.4 Deliverables

Minimum Viable

- In depth literature search on different methods that are using machine learning models to generate 3D multiple conformations.
- Analysis of state of the art.

Expected

- A machine learning model to generate multiple 3D conformations, based on a specific featurization vector.
- Comparative analysis of the generated model with state of the art models.

Desired

- Implement a model that can find similarity between molecules
- Model that is capable of representation learning and explainable AI.

References

[1] Salim Ahmad, Sahar Qazi, and Khalid Raza. Chapter 10 - translational bioinformatics methods for drug discovery and drug repurposing. In Khalid Raza and Nilanjan Dey, editors, *Translational Bioinformatics in Healthcare and Medicine*, volume 13 of *Advances in ubiquitous sensing applications for healthcare*, pages 127–139. Academic Press, 2021. doi: https://doi.org/10.1016/B978-0-323-89824-9.00010-0. URL `https://www.sciencedirect.com/science/article/pii/B9780323898249000100`.

[2] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. 07 2021.

[3] Simon Axelrod and Rafael Gómez-Bombarelli. Geom: Energy-annotated molecular conformations for property prediction and molecular generation. 06 2020.

[4] Jacques Boitreaud, Vincent Mallet, Carlos Oliver, and Jérôme Waldispühl. Optimol: Optimization of binding affinities in chemical space for drug discovery. *Journal of Chemical Information and Modeling*, 60(12):5658–5666, 2020. doi: 10.1021/acs.jcim.0c00833. URL `https://doi.org/10.1021/acs.jcim.0c00833`. PMID: 32986426.

[5] Daiguo Deng, Xiaowei Chen, Ruochi Zhang, Zengrong Lei, Xiaojian Wang, and Fengfeng Zhou. Extracting graph neural network-based features for a better prediction of molecular properties. *Journal of Chemical Information and Modeling*, 61(6):2697–2705, 2021. doi: 10.1021/acs.jcim.0c01489. URL `https://doi.org/10.1021/acs.jcim.0c01489`. PMID: 34009965.

[6] Daniel Elton, Zois Boukouvalas, Mark Fuge, and Peter Chung. Deep learning for molecular design - a review of the state of the art. *Molecular Systems Design Engineering*, 4, 05 2019. doi: 10.1039/C9ME00039A.

[7] Denis Kuzminykh, Daniil Polykovskiy, Artur Kadurin, Alexander Zhebrak, Ivan Baskov, Sergey Nikolenko, Rim Shayakhmetov, and Alex Zhavoronkov. 3d molecular representations based on the wave transform for convolutional neural networks. *Molecular Pharmaceutics*, 15(10):4378–4385, 2018. doi: 10.1021/acs.molpharmaceut.7b01134. URL `https://doi.org/10.1021/acs.molpharmaceut.7b01134`. PMID: 29473756.

[8] Ben Lo, Stefano Rensi, Wen Torng, and Russ Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23, 05 2018. doi: 10.1016/j.drudis.2018.05.010.

[9] Holly Matthews, James Hanison, and Niroshini Nirmalan. "omics"-informed drug and biomarker discovery: Opportunities, challenges and future perspectives. *Proteomes*, 4:28, 09 2016. doi: 10.3390/proteomes4030028.

[10] P. Preziosi. 2.06 - drug development. In John B. Taylor and David J. Triggle, editors, *Comprehensive Medicinal Chemistry II*, pages 173–202. Elsevier, Oxford, 2007. ISBN 978-0-08-045044-5. doi: https://doi.org/10.1016/B0-08-045044-X/00047-X. URL `https://www.sciencedirect.com/science/article/pii/B008045044X00047X`.

[11] Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Learning a continuous representation of 3d molecular structures with deep generative models. *ArXiv*, abs/2010.08687, 2020.

[12] Dan M. Roden. *Principles of Clinical Pharmacology*. McGraw-Hill Education, New York, NY, 2018. URL `accessmedicine.mhmedical.com/content.aspx?aid=1155945053`.

[13] Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Machine learning applications in drug development. *Computational and Structural Biotechnology Journal*, 18:241–252, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2019.12.006. URL `https://www.sciencedirect.com/science/article/pii/S2001037019303988`.

[14] Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. *CoRR*, abs/2105.03902, 2021. URL `https://arxiv.org/abs/2105.03902`.

[15] Gregor Simm and Jose Miguel Hernandez-Lobato. A generative model for molecular distance geometry. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8949–8958. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/simm20a.html`.

[16] Gregor N. C. Simm and José Miguel Hernández-Lobato. A generative model for molecular distance geometry, 2020.

[17] Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, and Jian Tang. Learning neural generative dynamics for molecular conformation generation. *ArXiv*, abs/2102.10240, 2021.