# Thesis proposal
# My Thesis

**Anonymous Author(s)**
Affiliation
Address
email

## Introduction

Drug discovery—the process by which a potential new medicine is identified–is a complex process that encompases the intersection of several fields (such as biology, statistics, chemistry or pharmacology). The entire process is a long and costly endeavour, with a typical time-frame of 10 to 20 years till maket release and an estimated cost between 1 and 2 billion USD. With just a small quantity of the initially identified compounds actually becoming an approved medicine. Many of these dropouts happening at the early stages of the entire pipeline.

It exists, then, a need for better mechanisms for detecting better candidates. One of the most promising directions is to improve the currently used and develop new *in-silico* methods—computational simulations are relatively cheap and quick run that makes them an interesting solution. In-silico simulations then cover two main aspects: modelling the dynamics of the human body, such that any effect relevant to the durg or the disease will be captured by it; and a mathod to generate good candidates that is effective at exploring the vast space of possible compounds.

Among the different computational approaches that have been used in the process of drug discovery deep learning (DL) has shown signs to be a potential game changer. DL has been able to capitalize on the exponential growth of data and the higher availability of computational resources. For examle, DL has had a remarkable success on computer vision (CV) and natural language processing (NLP), and has become the go-to solution for any problem in these two fields. It is, at the same time, penetrating into other fields, drug-discovery being one of them [1].

One of the big challenges is to unify all the aspects of drug-discovery and be able to incorporate all the rellevant biological information when designing possible candidate molecules. A success story on that line is the recently paper published by Zhavoronkov et al. [3] where the authors describe a deep learning method by which they are able to discover inhibitors of discoidin domain receptor 1 (DDR1)—a kinase implicated in fibrosis—in just 21 days.

Those promising results, albeit encouraging, are just the tip of the iceberg. There is still a long way till a model can satisfactorily capture the biological complexity of any arbitrary target and produce promising candidates. On top of that, there is an added dimension as such model should account for the variability from patient to patient and be able to generate a molecule that accomodates for all the genotipic and phenotipic variants, or generate different candidates for each of the genetic populations of interest. [need a ref here]

Even more, in the case of diseases like cancer, an heterogeneous population may appear within a single patient. So the same variant effects arise inside a dynamic ecosystem, where a drug that just targets a subpopulation may lead to an evolutionary pressure complicating further the treatment outlook [reference paper of evolutionary perspective to cancer].

There is then a great need to develop models that can be conditioned based on a large set of biological [conditions?] and meaningfully account for this variatos when generating a compound or/and evaluating a compunds effect when administered.

—-

When it comes to the methodology it exists an orthogonal problem [regarding strcutured data] when we deal with biological data. If we look at the case of deep learning for CV, where we deal with images, a key element of any architecture for it's success was the use of convolutional layers–one will mostly observe convolutional neural networks (CNNs) when analizying the state of the art in CV–which introduce a structural a pripor based on the structure of the input. A similar case can be made for NLP. For that reason, there exists a strong signal to look for models that can leverage the structural equivalent when in molecule or protein data, i.e. leverage graph structures.

[There is already a literature on this and I'm gonna talk about it [2]]

—-

The aim of this thesis will be to explore several ways to tackle this challenge.

—

Deep learning has been applied very successfully in many fields, notably computer vision (CV) and natural language processing (NLP), among those filelds ....

graph convolutional neural networks. We plan to

—

Analyzing graph like structures allows us to

## Aim

The aim of my thesis will be to explore how deep learning applied to the domain of graphs can help capture better biological aspects

## References


[1] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery, 2018.

[2] Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, and Fei Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics*, 2019(0):1–17, 2019.

[3] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R Shayakhmetov, Alexander Zhebrak, Lidiya I Minaeva, Bogdan A Zagribelnyy, Lennart H Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 2019.