
Thesis proposal

An alpha version of the draft

Nil Adell Mill
Winter 2020
Institute of Neuroinformatics

1 Introduction/Background

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

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

14 Among the different computational approaches that have been used in the process of drug discovery
15 deep learning (DL) has shown signs to be a potential game changer. DL has been able to capitalize on
16 the exponential growth of data and the higher availability of computational resources. For example, DL
17 has had a remarkable success on computer vision (CV) and natural language processing (NLP), and
18 has become the go-to solution for any problem in these two fields. It is, at the same time, penetrating
19 into other fields, drug-discovery being one of them [Chen et al., 2018].

20 When we deal with this biological and molecular data, it exists a challenge on how to deal with the
21 intrinsic structure of the data. If we look at the case of deep learning for CV, where we deal with
22 images, a key element of any architecture for its success was the use of convolutional layers—one
23 will mostly observe convolutional neural networks (CNNs) when analyzing the state of the art in
24 CV—which introduce a structural prior based on the structure of the data. A similar case can be
25 made for NLP. For that reason, there exists a strong signal to look for models that can leverage the
26 structural equivalent when in molecule or protein data, i.e. leverage graph structures. Sign [Sign?
27 need to rewrite that] of that is the recent advancements in that direction [Sun et al., 2019].

28 Another of the big challenges is to unify all the aspects of drug-discovery and be able to incorporate all
29 the relevant biological information when designing possible candidate molecules. An initial success
30 story on that line is the recently paper published by Zhavoronkov et al. [Zhavoronkov et al., 2019]
31 where the authors describe a deep learning method by which they are able to discover inhibitors of
32 discoidin domain receptor 1 (DDR1)—a kinase implicated in fibrosis—in just 21 days.

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

39 [I am not completely sure about this paragraph but I leave it here so I don't forget for now] Even
40 more, in the case of diseases like cancer, a heterogeneous population may appear within a single

41 patient. So the same variant effects arise inside a dynamic ecosystem, where a drug that just targets
42 a subpopulation may lead to an evolutionary pressure complicating further the treatment outlook
43 [reference paper of evolutionary perspective to cancer].

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

47 In fact it is of interest to develop multi-scale models that capture system complexity at the different
48 levels. For instance, a model that is able to learn protein-compound interactions—commonly known
49 as the docking problem—while at the same time use this information to predict effects of the
50 introduction of the compound on the larger protein-protein interaction (PPI) network.

51 ———

52 **Aim & Methods**

53 [Should I separate em in two different sections?]

54 The aim of this thesis will be two fold. One the one side, analyze how the explicit use of graph
55 convolutional neural networks (CGNNs) may open new oportunities when dealing with biological
56 and checmical data. On the other side, explore how modelling the biology at different levels (e.g.
57 molecular structure v.s. molecular interacton network [okay here I need to develop furhter about PPI,
58 maybe mention NetBite (as Jannis referenced in the mail)]) may help with our understanding [of the
59 biology? of compounds interaction?] and help generate better models. Furthermore, evaluate how
60 these may be integrated toguether.

61 GCNNs:

62 Multi-level:

63 This precise work will be focused around exploring all these concepts in the context of drug design
64 for cancer [...] the work will be done in colaboration with the Computational Systems Biology group
65 at IBM Research (Zurich). [...] The group is currently focused on individualised paediatric cure (iPC),
66 so an end gola of this project if for the end results of this project to help in that effor, for instance in
67 contibuting to the ongoing research in neuroblastoma.

68 **References**

69 [Chen et al., 2018] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The
70 rise of deep learning in drug discovery.

71 [Sun et al., 2019] Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., and Wang, F. (2019).
72 Graph convolutional networks for computational drug development and discovery. *Briefings in*
73 *Bioinformatics*, 2019(0):1–17.

74 [Zhavoronkov et al., 2019] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladin-
75 skiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev,
76 A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy,
77 B. A., Lee, L. H., Soll, R., Madge, D., Xing, L., Guo, T., and Aspuru-Guzik, A. (2019). Deep
78 learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*,
79 37(9):1038–1040.