Thesis proposal
An alpha version of the draft

Nil Adell Mill
Winter 2020
Institute of Neuroinformatics

Introduction/Background

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 2 and 3 billion USD [Schneider, 2019, Scannell et al., 2012]. 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 *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: **predictive modelling**, meaning modelling the dynamics of the human body—such that any effect relevant to the durg or the disease will be captured by it—and **generative modelling**, namely methods to generate good candidates that are effective at exploring the vast space of possible compounds, an estimated space estimated of 10^{60} compounds[Reymond et al., 2012].

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 [Dargan et al., 2019]. 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 [Chen et al., 2018].

When we deal with this biological and molecular data, it exists a challenge on how to deal with the intrinsic structure of the 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 prior based on the structure of the data. 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 [Wu et al., 2019]. In fact, there have been several models as such being proposed in the literature [Sun et al., 2019].

Another 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. An initial success story on that line is a recently paper [Zhavoronkov et al., 2019] 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]

[I am not completely sure about this paragraph but I leave it here so I don't forget for now] 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 variations when generating a compound or/and evaluating a compunds effect when administered.

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

Aim & Methods

[Should I separate em in two different sections?]

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

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

As mentioned previously, the idea of using GCNNs is not a new one in the literature [Sun et al., 2019]. My project will build upon those ideas presented in the literature, expand them and test their feasibility by implementing them into a wider framework for drug design [Born et al., 2019]. In that context two main areas of application appear. One of them would be to re-desing the drug conditional generator, for instance by reframing the vairational autoencoders, used for molecule generation, to architectures that operate over graphs [Simonovsky and Komodakis, 2018, Li et al., 2018a, Li et al., 2018b]. The second area would be to find better ways to asses the activity of these molecules, and in a wider context, assess their relevance as drug candidates. In the concrete case of the mentioned framework it is done by using a critic network proposed in [Manica et al., 2019]. This could be expanded on a set of different fronts: usign structural data instead of SMILES [Li et al., 2017, Do et al., 2019], by using GCNNs to cover a much wider network of genes [Oskooei et al., 2019, Wang et al., 2019], or by introducing particular scores (rewards) based in the interaction of the compound to certain targets [Yingkai Gao et al., 2018, Zhavoronkov et al., 2019] or the combination of the compund with other drugs [Zitnik et al., 2018] —a common practice in patients with cancer.

All these possible changes on the critic model would apply at different abstraction levels. That opens the door to seek for ways to integrate the representations learnt at those different stages [Ying et al., 2018, Ma and Zhang, 2019, Huang et al., 2019]. On top of that information extracted from here could be then leveraged on the drug generation part of the framework.

References

[Born et al., 2019] Born, J., Manica, M., Oskooei, A., Cadow, J., and Martínez, M. R. (2019). PaccMannˆRL: Designing anticancer drugs from transcriptomic data via reinforcement learning.

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

[Dargan et al., 2019] Dargan, S., Kumar, M., Ayyagari, M. R., and Kumar, G. (2019). A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering*.

[Do et al., 2019] Do, K., Tran, T., and Venkatesh, S. (2019). Graph transformation policy network for chemical reaction prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 750–760.

[Huang et al., 2019] Huang, J., Li, Z., Li, N., Liu, S., and Li, G. (2019). AttPool: Towards Hierarchical Feature Representation in Graph Convolutional Networks via Attention Mechanism. *The IEEE International Conference on Computer Vision (ICCV)*, pages 6480–6489.

[Li et al., 2017] Li, J., Cai, D., and He, X. (2017). Learning Graph-Level Representation for Drug Discovery. Technical report.

[Li et al., 2018a] Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018a). Learning Deep Generative Models of Graphs.

[Li et al., 2018b] Li, Y., Zhang, L., and Liu, Z. (2018b). Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1).

[Ma and Zhang, 2019] Ma, T. and Zhang, A. (2019). Incorporating Biological Knowledge with Factor Graph Neural Network for Interpretable Deep Learning.

[Manica et al., 2019] Manica, M., Oskooei, A., Born, J., Subramanian, V., Saéz-Rodríguez, J., and Rodríguez Martínez, M. (2019). Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. Technical report.

[Oskooei et al., 2019] Oskooei, A., Manica, M., Mathis, R., and Martínez, M. R. (2019). Network-based Biased Tree Ensembles (NetBiTE) for Drug Sensitivity Prediction and Drug Sensitivity Biomarker Identification in Cancer. *Scientific Reports*, 9(1).

[Reymond et al., 2012] Reymond, J. L., Ruddigkeit, L., Blum, L., and van Deursen, R. (2012). The enumeration of chemical space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(5):717–733.

[Scannell et al., 2012] Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. Technical Report 3.

[Schneider, 2019] Schneider, G. (2019). Mind and machine in drug design. *Nature Machine Intelligence*, 1(3):128–130.

[Simonovsky and Komodakis, 2018] Simonovsky, M. and Komodakis, N. (2018). GraphVAE: Towards generation of small graphs using variational autoencoders. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11139 LNCS, pages 412–422.

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

[Wang et al., 2019] Wang, M., Yu, L., Zheng, D., Gan, Q., Gai, Y., Ye, Z., Li, M., Zhou, J., Huang, Q., Ma, C., Huang, Z., Guo, Q., Zhang, H., Lin, H., Zhao, J., Li, J., Smola, A., and Zhang, Z. (2019). Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. Technical report.

[Wu et al., 2019] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A Comprehensive Survey on Graph Neural Networks. Technical report.

[Ying et al., 2018] Ying, R., Morris, C., Hamilton, W. L., You, J., Ren, X., and Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 4800–4810.

[Yingkai Gao et al., 2018] Yingkai Gao, K., Fokoue, A., Luo, H., Iyengar, A., Dey, S., and Zhang, P. (2018). Interpretable drug target prediction using deep neural representation. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2018-July, pages 3371–3377.

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

[Zitnik et al., 2018] Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. In *Bioinformatics*, volume 34, pages i457–i466.