Thesis proposal

Nil Adell Mill
January 2020

Institute of Neuroinformatics,
ETH Zurich & Zurich University

Introduction & Background

Drug discovery—the process by which a potential new medicine is identified—is a complex process that encompasses the intersection of several fields (such as biology, statistics, chemistry or pharmacology). The entire process is a long and costly endeavor, with a typical time-frame of 10 to 20 years till market 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—only 1 out of 10000 synthesized molecules gets market approval. 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 modeling the dynamics of the human body—such that any effect relevant to the drug or the disease will be captured by it—and **generative modelling**, i.e. methods to generate good candidates which, in part, means methods that are effective at exploring the vast space of possible compounds, estimated to be on the order of 10^{60}[Reymond et al., 2012]. Several computational approaches have been used over the years, from modeling molecular dynamics simulations to data-driven statistical methods [Hung and Chen, 2014, Kuhn et al., 2016]. Recently 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. DL has had remarkable success in computer vision (CV) [Guo et al., 2016] and natural language processing (NLP) [Young et al., 2018], and has become the go-to solution for any problem in these two fields. For instance, in the case of CV, where we deal with images, a key element of any architecture's success was the use of convolutional layers—one will mostly observe convolutional neural networks (CNNs) when analyzing the state of the art in CV—which introduces a structural a prior based on the structure of the data[Fukushima, 1980, LeCun, 1989, Ulyanov D., Vedaldi A., 2018]. A similar case can be made for NLP. A similar approach, them, can be proposed when we deal with biological and molecular data by leveraging their graph representations. Efforts in generalizing the convolution operator on non-euclidian structures have given rise to graph convolutional neural networks (GCNNs)[Wu et al., 2019]. GCNNs, then, pose an opportunity to drug discovery due to their capacity to deal natively with graph data[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 relevant biological information when designing possible candidate molecules. An initial success story on that line is a recent paper [Zhavoronkov et al., 2019] where the authors describe a deep learning method by which they are able to discover, synthesize, and test in an animal model, inhibitors of discoidin domain receptor 1 (DDR1)—a kinase implicated in fibrosis—in less than two months.

Those promising results, albeit encouraging, are just the tip of the iceberg. There is still a long way until a model can satisfactorily capture the biological complexity of an 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 accommodates for all the genotypic and phenotypic variants. Or otherwise generate different candidates for each of the populations of interest. That is especially important for hypercomplex diseases; for example, in cancer, where a genotypically heterogeneous tumoral

population may appear within a single patient [Boland and Yurgelun, 2017]. 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 [Enriquez-Navas et al., 2015].

There is then a great need to develop models that can be conditioned based on a large set of biological factors and meaningfully account for these variations when generating a compound or/and evaluating a compound's effect when administered. What is, in other words, the need for the wider adoption of precision medicine.

Last of all, in a more holistic view, it is of interest to develop multi-scale models that integrate a system's complexity at all different levels. For instance, a model that can 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[Sun et al., 2019].

Aim & Methods

The aim of this thesis will be two-fold. One the one side, analyze how the explicit use of GCNNs may open new opportunities when dealing with biological and chemical data. On the other side, explore how modeling the biology at different levels (e.g. molecular structure v.s. molecular interaction network) may help create better models. Furthermore, evaluate how these different scales may be integrated.

This precise work will be focused on exploring all these concepts in the context of designing anti-cancer drugs. The work will be done in collaboration with the Computational Systems Biology group at IBM Research (Zurich), currently part of the iPC consortium[1]. As part of such, an end goal of this project is for the results of this project to help in the consortium efforts on paediatric cancer, for instance in contributing to the ongoing research in neuroblastoma, the most common cancer diagnosed on the first year of life [Maris, 2010].

As mentioned previously, the idea of using GCNNs for drug discovery 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 is to re-design the generative model, for instance by reframing the variational 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 is to find better ways to asses the activity of these molecules, and in a wider context, assess their relevance as drug candidates, i.e. improve the predictive model. In the concrete case of the mentioned framework, it is done by using a critic network [Manica et al., 2019]. This can be expanded on a set of different fronts: using structural data instead of SMILES[2] [Li et al., 2017, Do et al., 2019], by using GCNNs over PPI networks in a manner that allows for the use all the information available[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 compound 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.

Data & Baselines

The data from this project will be obtained from multiple sources including, but not limited to, the Genomics of Drug Sensitivity in Cancer (GDSC) databse [Yang et al., 2013], STRING and STICH, [Szklarczyk et al., 2019, Szklarczyk et al., 2016], DrugBank [Wishart, 2006], ChEMBL [Gaulton et al., 2017], PubChem [Kim et al., 2019] and ZINC 15 [Sterling and Irwin, 2015]. For the development of the models PyTorch [Paszke et al., 2019] will be used as a main framework, likely used alongside DeepChem[3].

The novel predictive model will be evaluated against empirically measured properties of drugs in the mentioned data sources. Furthermore model's performance will be compared to that of other models in the literature. For instance, comparing IC50 prediction values for for cell-drug paris (unseen during training) to the empirically

[1] https://ipc-project.eu/
[2] http://opensmiles.org/
[3] https://deepchem.io/

measured ones from the previously mentioned data sources as done in [Oskooei et al., 2019, Joo et al., 2019, Oskooei et al., 2018]. Another prospective set of benchmarks to be used are those proposed by MoleculeNet [Wu et al., 2018].

For evaluation of generative models previous work [Theis et al., 2016] has discussed how different metrics measure different conceptions of performance. In order to deal with different considerations, such as diversity, novelty, molecular activity, or stability, metrics like the Fréchet ChemNet Distance (FCD) [Preuer et al., 2018] have been proposed. In this project, one the metrics will come from the predictive model, which will be used method to evaluate the effectiveness of the generated compounds (i.e. the performance of the generative model). At the same time, another metric for the generative model will be the similarity of the generated compounds to already existing drugs. An approach that has already been used in the literature [Born et al., 2019]. A wider benchmarking framework, GucaMol, was proposed recently [Brown et al., 2019] and could be used in this project. On top of all these metrics, a natural final step for the entire evaluation of the system would be the actual in-vitro synthesis of the generated compounds, as done in [Zhavoronkov et al., 2019], however the complexity of that step makes it unlikely to be carried out during the execution of this project.

References

[Boland and Yurgelun, 2017] Boland, C. R. and Yurgelun, M. B. (2017). Mutational cascades in cancer.

[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.

[Brown et al., 2019] Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. (2019). GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108.

[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.

[Enriquez-Navas et al., 2015] Enriquez-Navas, P. M., Wojtkowiak, J. W., and Gatenby, R. A. (2015). Application of evolutionary principles to cancer therapy.

[Fukushima, 1980] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

[Gaulton et al., 2017] Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954.

[Guo et al., 2016] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48.

[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.

[Hung and Chen, 2014] Hung, C. L. and Chen, C. C. (2014). Computational approaches for drug discovery.

[Joo et al., 2019] Joo, M., Park, A., Kim, K., Son, W.-J., Lee, H. S., Lim, G., Lee, J., Lee, D. H., An, J., Kim, J. H., Ahn, T., and Nam, S. (2019). A Deep Learning Model for Cell Growth Inhibition IC50 Prediction and Its Application for Gastric Cancer Patients. *International Journal of Molecular Sciences*, 20(24):6276.

[Kim et al., 2019] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109.

[Kuhn et al., 2016] Kuhn, M., Yates, P., and Hyde, C. (2016). Statistical Methods for Drug Discovery. pages 53–81.

[LeCun, 1989] LeCun, Y. (1989). Generalization and network design strategies. Technical report.

[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.

[Maris, 2010] Maris, J. M. (2010). Recent advances in neuroblastoma. *New England Journal of Medicine*, 362(23):2202.

[Oskooei et al., 2018] Oskooei, A., Born, J., Manica, M., Subramanian, V., Sáez-Rodríguez, J., and Martínez, M. R. (2018). PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks.

[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).

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.

[Preuer et al., 2018] Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. (2018). Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741.

[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.

[Sterling and Irwin, 2015] Sterling, T. and Irwin, J. J. (2015). ZINC 15 - Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

[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.

[Szklarczyk et al., 2019] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613.

[Szklarczyk et al., 2016] Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1):D380–D384.

[Theis et al., 2016] Theis, L., Van Den Oord, A., and Bethge, M. (2016). A note on the evaluation of generative models. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

[Ulyanov D., Vedaldi A., 2018] Ulyanov D., Vedaldi A., L. V. (2018). Deep Image Prior. Technical report.

[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.

[Wishart, 2006] Wishart, D. S. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(90001):D668–D672.

[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.

[Wu et al., 2018] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530.

[Yang et al., 2013] Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., and Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1).

[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.

[Young et al., 2018] Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing.

[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.