PH.D. THESIS SUMMARY

Magdalena Wiercioch, MSc

Thesis supervisor: Jacek Tabor, Ph.D Prof., Jagiellonian University

Development of universal data representations with application in chemistry

In recent years, deep learning models have shown their great potential in the field of representation learning. Unfortunately, unlike the established deep learning-based methodologies that have achieved human-level accuracy in various application domains such as computer vision and speech recognition, the development of molecular modeling is still at an early stage. This seems to be mainly caused by the inductive biases of molecules that are completely different from those of image, and the lack of sufficiently large and reliable chemical data.

Therefore, *the main research problem addressed in this Thesis involves learning representation that aims at improving drug discovery and development process*. As a result, three different deep learning - based architectures are offered to create a meaningful embedding given a molecular structure. In order to assess the impact of the models, a variety of tests on classification and regression tasks are performed.

To specify, in this Thesis, the author proposes solutions for three different tasks:

- *The classification task*. Here, the author applies the developed model to detect bioactive chemical compounds.
- *The classification, regression and interpretability task*. Here, the author applies the developed model to predict molecular toxicity.
- *Deep representation learning of graphs and sequences*. Here, the author applies the developed model to evaluate whether the candidate drug and a target protein are interacting.

The Thesis consists of an introduction, four chapters, a conclusion and an appendix.

Chapter I - Introduction

Over the years, scientists have noticed that the choice and quality of the data representation, or features in the data used to train a machine learning model directly affect the final performance of used approach. It is also not surprising that algorithm's usefulness depends on the task. However, one can always indicate sets of features considered as representative that are treated as reflection of what the data is like. Then, these features could be used as input for various tasks such as classification or prediction. Therefore, working on learning representation, in some cases can be beneficial, for example, when data featurization is employed, especially dealing with small datasets.

In general, the concept of representation learning means learning a parameter-function map from the raw input data domain to a feature vector or tensor. The goal is to detect and extract abstract, or higher conceptual level ideas in order to boost the performance of a system over the unseen data. What is more, the dimensionality of the input domain is usually high since the objects such as videos, images, or text are taken into consideration. However, the encoded representation is associated with a low-dimensional manifold. In this regard, although there are many dimension reduction techniques that offer the ability to make high dimensional data space simpler, such methods often do not capture a mapping that is relevant for new data samples. Interestingly, representation learning is developed for doing this job.

In conventional machine learning, one begins with a specific challenge for which there is the training data available. Then, the data is pre-processed, transformed, fed into the machine learning pipeline, and a solution is returned. Here, the learning part includes only a making decision based on the approximation of the data unknown mapping. In turn, one of the driving factors of the success of deep learning lies in its ability to learn compact and expressive representations directly from the observed data. Furthermore, the availability of programmable highly-parallel hardware, especially graphics processing units (GPUs) caused that hand-crafted features have been replaced by feature learning mechanism. In consequence, the development of architecture-engineering has had a tremendous influence in the field of representation learning. In addition, in the last few years a number of novel deep learning architectures and building blocks have been published reporting superior performance.

First of all, there is no single definition of what it means to learn a representation. Undoubtedly, an intuition is that a good representation makes the learning task easier. A few years ago, Bengio, Courville and Vincent [1] focused on a few essential aspects of good representations. According to their investigation, a list of prior factors can be introduced. Examples include local smoothness of input, spatial and temporal coherence in a sequence of inputs is observed, or a hierarchical organization of multiple explanatory factors. In addition, factors are related to each other through simple, usually linear dependencies, and factors that are shared with other tasks, also share the statistical power across tasks.

The significance of representations of molecules have attracted a great deal of interest in the decades of drug discovery research [2]. A molecule is commonly-seen as a group of atoms held together by bonds. Unfortunately, this representation is itself insufficient for understanding chemical space and solving various problems such as properties prediction [3]. Therefore, given the role and applications of molecular design, several new approaches have to be explored. As a result, in the Thesis, a molecular representation $\mathcal{R}$ is a mapping from drug-like molecules $\mathcal{M}$ to some set X.

In spite of the notable advantages of deep learning, challenges in applying deep learning to the cheminformatics domain still remain. For instance, data remains an open challenge. Firstly, the enormous space of valid chemical compounds is estimated to be around 10^{60} [4]. Given the vastness of drug-like chemical space, the efficient and automated methods for development for various applications are needed. Furthermore, the training data is limited for the current challenges in drug discovery. Another aspect that is relevant and should be noted is data bias and data imbalance problems. In addition, in contrast to computer vision or natural language processing fields, the acquisition of the labels to the specific problem is significantly harder to many orders of magnitude. It is caused by the fact that the labels can only be obtained through lab experiments.

In the view of the above, this Thesis, is dedicated to elaborate upon representation learning, whereby the focus is on three big topics: classification, regression, and deep representation learning of graphs and sequences. Besides, model interpretability is discussed in an ongoing fashion. The findings are underpinned through several experiments with a focus on the selected serious challenges that exist in chemistry such as bioactivity prediction, toxicity prediction and drug-target interaction prediction.

Chapter II - Background: cheminformatics and learning representations for molecular data

This chapter provides the foundation on which the study is built, the background information related to existing molecular models and challenges faced by cheminformatics. The author explains that drug design is not straightforward and is still in its infancy. Indeed, the process of drug discovery suffers from the huge computation cost and time-consuming procedures, which limits its application in pharmaceutical industry [5]. To give an illustrative example, existing drug discovery pipelines take 5-10 years with a cost counted in billions of dollars. Therefore, cheminformatics, especially deep neural network-based techniques can be a game changer in various areas of CADD (Computer-Aided Drug Design).

The driving force behind this lies in the fact that deep learning-based approaches enable to learn compact and expressive representations directly from the observed data. Generally, the current works along the line of deep learning for molecules can be categorized into two main groups

according to the input data type of chemical compound, string-based methodologies and graph-based methodologies. Specifically, SMILES (simplified molecular-input line-entry system) is a sequence notation encoding a molecule into a character string that follows a specified grammar [6]. However, a chemical compound can also be naturally seen as a graph with nodes corresponding to atoms and edges corresponding to bonds, and one may learn on a molecular structure. Viewing molecule structure as graph data, leads to graph neural networks-based (GNNs) [7, 8] architectures.

Even though a graph-based approach has already led to the development of the state-of-the-art improvements, more computational methods are required to handle chemical structure that could support more effective DNN-based drug discovery. For instance, the scarcity of labeled data brings serious challenges for deep learning in molecular representation. It is caused by the errors and the fact that lab experiments are costly [9, 10]. In consequence, training datasets used in cheminformatics problems are often limited in size. In turn, this results in overfitting and finally the learned representations lack of generalizability [11]. For this reason, to address the above issue, one has to design a more powerful models that exhibit scalability and accuracy to express a great variety of molecules. In addition, another problem that needs to be discussed is the limited structural information incorporated into existing deep models. Although treating a chemical compound as a set of atoms and bonds is reasonable, one should take into consideration the fact that it also consists of various molecular dependencies that cannot be missed. In particular, structural dependencies between nodes and edges, and interactions must be identified.

Therefore, in the Thesis three different deep learning - based architectures are offered to create a meaningful embedding given a molecular structure. They are introduced in Chapter III, Chapter IV and Chapter V. In order to assess the impact of the models, several experiments on classification and regression tasks are presented.

Chapter III - Learning Hybrid Representation for Classification

An important step in the drug discovery pipeline is to predict molecular bioactivity. This is caused by the fact that the discovery of a new drug involves testing small molecules for their ability to bind to the target receptor [12]. Since the task is to separate the active chemical compounds from the inactives, the classification task is usually suggested. As a result, in the last decades, many computational methodologies have been proposed and widely developed to expedite the process of identification of active molecules. In fact, a series of approaches exploring quantitative structure-activity relationships (QSAR) have been developed [13]. Most of them focus on similarity searching - based methods [14]. In addition, typically, the development of a reliable computational methodology needs a high-quality descriptors [15].

Albeit powerful, the traditional machine learning methods often lead to the insufficient out-

comes. Recently, there has been deep learning architectures successfully applied on molecular data, too. Nevertheless, in practice, dealing with neural networks poses several unique challenges. First, the available models usually need a large and high-quality data. Second, the increasing depth and width of deep architectures has also an influence on growth in computation. Third, as the model is not completely aware of the structural information related to the molecule, it cannot infer any significant molecular dependencies. In Chapter III, the above-mentioned gap is bridged by providing a workflow procedure, named **H**ybrid **D**eep **N**eural **N**etwork (HybNN).

Method

Results

The author uses fourteen binary classification datasets in the experiments to test the performance of HybNN. The datasets are derived from the PubChem database [16]. Each dataset includes binary labels on its bioactivity property toward the targets.

The well-designed experiments demonstrate the effectiveness of HybNN from various aspects: 1) predictive performance on validation sets; 2) predictive performance on test sets; 3) analysis how fast a learning machine improves its behaviour; 4) the impact of chemical diversity; 5) the verification of transferability; 6) the influence of varied number of SGRU layers; 7) the impact of the single blocks. The chapter shows extensive comparisons with various approaches, including traditional machine learning methods and deep learning-based models. The outcomes reveal that HybNN can outperform baselines in predicting bioactivity on all datasets.

Chapter IV - Learning A Fragment-Oriented Representation for Supervised Learning Problems

Undoubtedly, a challenge in machine learning and deep learning is sample size. Overall, the problem is as follows: training a learning model needs enough data to prevent overfitting and extract valuable insight to learn representations. As a consequence, when one has not sufficient data for a learning task, it is hard to use the model and make the model interpretable. Therefore, Chapter IV tackles the challenge of representation learning when the dataset is quite small and biased. In the second contribution, the author casts this problem into a problem involving molecular data where the task is to predict molecular properties.

At the same time, many studies indicate that poor toxicity remain major limiting aspect of drug discovery [17]. One strategy that has been widely employed is *in vivo* methodology. However, time-consuming wet-lab experiments or simulations result in a limited number of chemical compounds with validated properties [18]. In addition, it happens that they do not necessarily scale between

animal models and humans. To address these issues, there has recently been a shift towards *in vitro* and to machine learning based *in silico* techniques.

To foster further development and to better understand the underlying mechanisms of action of various toxic chemicals, the author proposes **S**ubgraph **E**ncoded **N**eural **N**etwork (SENN) that investigates the role of atoms connections in the molecular graph and global molecular features.

Method

The entire architecture of **S**ubgraph **E**ncoded **N**eural **N**etwork (SENN) could be split into seven parts in a high-level discussion. The initial input to the SENN is a graph $\mathcal{G}$ that represents a molecule. Thus, at the beginning, a set of atom attributes is assigned to each vertex and each edge is also associated with its weight. The weight equals the multiplicity of the bond it refers to. Then, the graph $\mathcal{G}$ is preprocessed to obtain the subgraphs, fed into a graph convolutional neural network and the embedding $out_{\mathcal{G}}$ is returned. This representation is concatenated with a vector of properties out_{att}. In this case, the properties are connected with the selected attributes of the chemical compound. They include the features extracted by ChemoPy [19] such as a molecular weight or a number of rotatable bonds. The combined representations form a feature vector that is the input to a few linear layers with a dropout, and a final task layer.

However, as it was mentioned, before the final task is performed, the graph embedding operation and GCN are concerned. Therefore, firstly, for a given graph $\mathcal{G}$, all distinct k-path subgraphs are extracted. Then, a random unit-norm vector is associated with each subgraph. And from that time, the embeddings assigned to vertices are updated by GCN's layers. Specifically, each vector is replaced with the average over all vectors in its neighbourhood. Next, a linear transformation is applied. As a result, the computed vertex embeddings are averaged and a d_0-dimensional graph representation is obtained.

Results

Multiple toxicity-related datasets are adopted for regression and classification tasks. For a classification task, the dataset was taken from the Tox21 Data Challenge [20] in both SDF and SMILES formats. The data consists of approximately 12 000 compounds and includes twelve different subchallenges/tasks. In turn, for a regression task, four toxicity - related data sets were employed [21].

In the experiments, SENN is compared with five state-of-the art models, including both traditional machine learning approach and deep learning methodologies. In general, SENN shows satisfactory outcomes. For instance, on the AR toxicity dataset, SENN obtains significantly better AUC-ROC score (0.802 ± 0.006) than that by the well-known prediction approaches.

Chapter V - Deep Representation Learning of Graphs and Sequences

Among computational approaches to drug development, scientific community has already made tremendous progress in the identification of drug-target interactions (DTI) [22]. In the pharmaceutical sciences, a drug target is a chemical compound that is capable of binding to drugs and producing effects in cells. Proteins are considered as the obvious molecular targets [23].

DTI is significant, especially for finding effective and safe treatments. It is also worth to mention that majority of the existing DTI works have formulated the DTI prediction task as a binary classification. Of course, in the literature, there is a great variety of in silico proposals about DTI prediction. Nevertheless, the existing approaches have found a few relevant drawbacks. Firstly, these methodologies usually need large number of known binding data. In consequence, the prediction results are not satisfactory when one works with a small amount of known data. Secondly, the performance is much worse if the three-dimensional structures of the target protein are not available.

To address the aforementioned shortcomings, a novel DTI prediction methodology, called **T**riplet **E**ncoded **N**eural **N**etwork (TENN) is introduced. TENN aims to identify the drug-target interactions by exploiting the existing topological structure of drug molecules, along with modeling spatio-sequential information.

Method

Triplet **E**ncoded **N**eural **N**etwork (TENN) contains three components. First of all, the heterogeneous network is constructed by integrating a variety of drug and protein related information sources in a form of three components. In the second step, the high-dimensional features of drugs and proteins are combined and reduced by adopting a set of linear layers with dropout, and the low-dimensional representation is obtained. Finally, the association between each pair of drugs and proteins is predicted.

Results

In order to evaluate the proposed methodology, the author used the data from the BindingDB [24] database that contains experimentally determined binding affinities on the interactions of target proteins with small, drug-like molecules. Also, DTIs from DrugBank [25] were employed.

The core advantage of TENN is the ability to handle the low dimensional feature vectors and predict the probability of interaction between each pair of drugs and proteins. The computational experiments reveal that if one extracts the global information of protein sequences and drug compounds, it leads to not only improvement in the efficiency of DTI, but enables to detect more complex interactions. Moreover, the outcomes indicate that TENN is better than the other four

state-of-the-art approaches. Although the research on TENN focuses on the application to the problems in chemistry, the proposed methodology is universal and could be employed to model various interactions in the world.

Chapter VI - Afterword

The conclusion from this Ph.D. Thesis is that the proposed methodologies bring progress. However, important open challenges remain. First of all, the author suggests that future work on explanatory techniques would be necessary for proper usage of the presented approaches in practice. Secondly, the introduced algorithms can further be improved. A possible future direction could be the use of the few-shot learning concept.

Appendix

This chapter first presents a thorough introduction to the most relevant classes of deep learning models to build a ground for the Thesis. In this context, the author starts with discussing state-of-the-art feed forward architectures and temporal neural models. Then, the milestones of supervised learning are briefly mentioned.

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] W. J. Wiswesser, "107 years of line-formula notations (1861-1968)," *Journal of Chemical Documentation*, vol. 8, no. 3, pp. 146–150, 1968.

[3] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. molecular frameworks," *Journal of medicinal chemistry*, vol. 39, no. 15, pp. 2887–2893, 1996.

[4] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: a molecular modeling perspective," *Medicinal research reviews*, vol. 16, no. 1, pp. 3–50, 1996.

[5] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, *et al.*, "Rethinking drug design in the artificial intelligence era," *Nature Reviews Drug Discovery*, vol. 19, no. 5, pp. 353–364, 2020.

[6] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[7] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2, pp. 729–734, IEEE, 2005.

[8] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[9] L. David, J. Arús-Pous, J. Karlsson, O. Engkvist, E. J. Bjerrum, T. Kogej, J. M. Kriegl, B. Beck, and H. Chen, "Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research," *Frontiers in pharmacology*, vol. 10, 2019.

[10] E. B. Lenselink, N. Ten Dijke, B. Bongers, G. Papadatos, H. W. Van Vlijmen, W. Kowalczyk, A. P. IJzerman, and G. J. Van Westen, "Beyond the hype: deep neural networks outperform established methods using a chembl bioactivity benchmark set," *Journal of cheminformatics*, vol. 9, no. 1, pp. 1–14, 2017.

[11] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265*, 2019.

[12] P. Buchwald and N. Bodor, "Computer-aided drug design: the role of quantitative structure–property, structure–activity and structure–metabolism relationships (qspr, qsar, qsmr)," *Drugs Future*, vol. 27, no. 6, pp. 577–588, 2002.

[13] J. C. Dearden, "The history and development of quantitative structure-activity relationships (qsars)," in *Oncology: breakthroughs in research and practice*, pp. 67–117, IGI Global, 2017.

[14] F. R. Burden and D. A. Winkler, "New qsar methods applied to structure- activity mapping and combinatorial chemistry," *Journal of chemical information and computer sciences*, vol. 39, no. 2, pp. 236–242, 1999.

[15] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, *et al.*, "Qsar modeling: where have you been? where are you going to?," *Journal of medicinal chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[16] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, and J. Zhang, "Pubchem bioassay: 2017 update," *Nucleic acids research*, vol. 45, no. D1, pp. D955–D963, 2017.

[17] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal, "Clinical development success rates for investigational drugs," *Nature biotechnology*, vol. 32, no. 1, pp. 40–51, 2014.

[18] G. J. Harry, M. Billingsley, A. Bruinink, I. L. Campbell, W. Classen, D. C. Dorman, C. Galli, D. Ray, R. A. Smith, and H. A. Tilson, "In vitro techniques for the assessment of neurotoxicity.," *Environmental health perspectives*, vol. 106, no. suppl 1, pp. 131–158, 1998.

[19] D.-S. Cao, Q.-S. Xu, Q.-N. Hu, and Y.-Z. Liang, "Chemopy: freely available python package for computational biology and chemoinformatics," *Bioinformatics*, vol. 29, no. 8, pp. 1092–1094, 2013.

[20] N. C. for Advancing Translational Sciences, "Tox21 data challenge 2014," 2014.

[21] K. Wu and G.-W. Wei, "Quantitative toxicity prediction using topology based multitask deep neural networks," *Journal of chemical information and modeling*, vol. 58, no. 2, pp. 520–531, 2018.

[22] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, vol. 50, pp. 71–91, 2019.

[23] Y. Feng, Q. Wang, and T. Wang, "Drug target protein-protein interaction networks: a systematic perspective," *BioMed research international*, vol. 2017, 2017.

[24] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D198–D201, 2007.

[25] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.