THE UNIVERSITY OF OXFORD | THE DEPARTMENT OF STATISTICS

GRAPH NEURAL NETWORKS FOR LEARNING MOLECULAR REPRESENTATIONS

CANDIDATE NUMBER:

In Fulfillment of Assessment for
'Topics in Computational Biology'

April 15, 2021

ABSTRACT

Der Abstract fasst die zentralen Inhalte der Arbeit zusammen. Eine Wertung oder Interpretation erfolgt nicht. Dies hilft, sich einen groben Überblick über Fragestellung, Vorgehen und Ergebnisse zu verschaffen. Bestandteil sollen die Teile a) Hintergrundinformationen, Fragestellung, Zielsetzung, Forschungskontext, b) Methoden, c) Ergebnisse und d) Schlussfolgerungen, Anwendungsmöglichkeiten sein. Der Text ist knapp, vollständig und präzise, zudem objektiv und ohne persönliche Wertung. Achten Sie auf eine einfache und verständliche Sprache. Alle genannten Inhalte müssen auch im Hauptteil aufgegriffen werden. Den Inhalt objektiv und ohne persönliche Wertung wiedergeben. Gehen Sie auf die wichtigsten Konzepte, Resultate oder Folgerungen ein. Verwenden Sie keine Zitate und verzichten Sie auf Abkürzungen. In der Regel sind ca. 200 Wörter ausreichend.

CONTENTS

1 Introduction

1.1 Brief overview of molecular property prediction

Molecules form the smallest identifiable parts of covalent compounds that still retain their chemical properties mol (2021). These covalent compounds can be found in all organisms, since together they form integral parts like proteins or the DNA making an understanding of molecules and their properties key to deciphering the foundations of life. Since molecules are complex physical entities in 3D space consisting of covalent bonds between atoms, identifying their chemical, physical or biological properties is by no means a simple task. *Molecular property prediction* aims to characterise molecules according to their properties. In abstract terms this amounts to finding a nonlinear function from a class of molecules to a set of predefined properties. Classically, *in vitro* screeening and *in vivo* testing were widely used in early stages of drug discovery in order to identify 'druggable' targets that display a desired biological response. However, this process is extremely time and resource inefficient, because ... . More recently, *in silico* methods attempt to embed the molecule into a mathematical representation which can then be used to learn this nonlinear relationship between the embedded molecules and their corresponding properties using statistical and machine learning methods. For instance, J. Stokes et al. achieved a huge breakthrough when the discovered the new antibiotics halicin Stokes et al. (2020) after decades of stagnation in that field. Contrary to previous methods that translate molecules into a fixed predefined mathematical representation, they employed a Graph Neural Network that was able to learn a representation that then served as an input to an Artificial Neural Network to predict the target inhibitory effect against E. coli. Other classes of properties that have been of interest in the past are vast and comprise for example quantum-mechanic, physio-chemical, bio-physical or physiological properties Wu et al. (2018).

Before the advent of computational methods in the process of drug discovery, lead compounds were found by isolating natural products from microbiological fermentation, plant extracts and animal sources Gallop et al. (1994). This involved TODO

On October 5, 1981 a new version of the 'Fortune' magazine was release. Its cover page featured an article titled 'The Next Industrial Revolution: designing drugs by computer at Merck' Van Drie (2007). This marked the begin of stage of naive euphoria in computational drug design with investments of millions of dollars in hardware and software.

1.2 Outline of the thesis

The goal of this thesis is to investigate the role of Graph Neural Networks in the field of Molecular Property Prediction and its application to Drug Discovery being one of its best known representations in biology. In the rest of this section we will give a brief outline of the history of the dominant methods in MPP and elaborate on its role in Drug Discovery. Following that, we will introduce Graph Neural Networks and present the necessary theory behind them in order to understand their advantages and disadvantages. Then, we will compare the performance of GNNs to that of other methods in MPP and assess the benefit of Deep Learning in MPP in general. Finally, a summary of this thesis is given outlining its most important findings.

1.3 History of MPP methods

2 Molecules and their Representation

2.1 Molecules

Atoms are the smallest identifiable units of chemical elements which make up all matter in the universe. A fundamental principle of chemistry is that the atoms of different elements can combine to form chemical compounds and a lot of the study in chemistry is centered around understanding what happens when these compounds are formed. A chemical compound can be defined as a distinct group of atoms that are held together by chemical bonds (cite? Khan modelcules). Similar to the attraction between the positively charged nucleus and the negatively charged electrons that constitutes the structure of atoms, chemical bonds are caused by electrostatic attractions. While there is no clear separation between types of bonding from a physical perspective, it is still convenient to distinguish

between different bonding types from a chemical perspective. The behaviour of the valence electron is the determining factor and this can be responsible for different properties of the resulting substance.

We are primarily concerned with two major types of bonds: Ionic bonds and covalent bonds. In a simplified view, ionic bonding can be classified as the transfer of a valence electron from one atom to the other resulting in the formation of two oppositely charged ions that hence attract each other and are bond together. Covalent bonding on the other hand is the result of electrostatic attraction between one or more electrons to the atomic nuclei of both atoms. This can be regarded as a sharing of the electrons across the two atoms. The structure resulting from covalent bonding is called a *molecule*. These are of particular importance in biology making up the smallest identifiable parts of *covalent compounds* that still retain their chemical properties mol (2021). These covalent compounds can be found in all organisms, since together they form integral parts like proteins or the DNA making an understanding of molecules and their properties key to deciphering the foundations of life.

2.2 Representation of Molecules

2.2.1 Brief history of molecular representations

In 1860 when the first International Chemical Congress was held in Karlsruhe, Germany, Alexander Butlerov predicted that determining the atomic arrangements of molecules would be the future of chemistry (Butlerov, 1861). He was the first person to use the word 'structure' in its modern chemical meaning. This marked the birth of structural chemistry.(Wiswesser, 1968). Since then it took only seven year to develop the main ideas about line-formula conventions in familiar form like

$$\mathrm{C_3H_7OH}.$$

No new practices appeared within 79 years until between 1947 and 1954 structure-delineating notations were introduced such as the Wiswesser line notation (WLN) which became very popular as it was easily interpretable by humans as well as computers. Compared to today's line formulae the WLN was very compact since memory efficiency was a critical factor in computers at that time.

When the advent of technology in the science accelerated in the 1980s, the role of chemical notations began do decline. (Lawlor, 2016) attributes this to two main reasons. On the one hand, computer-manageable connection tables opened up new possibilities to experiment with structures. This meant that rather than working with the chemical formula itself, it was translated in a connection table where algorithms like similarity searches could be run to calculate compute properties, map reactions etc. The second reason is the increasing availability of graphics terminals. Multiple companies like Molecular Design Ltd. or CAS (Dittmar et al., 1983) introduced interactive services that enabled a translation between a graphical representation of compounds and their connections tables. Furthermore, this involved functionalities like searching by structure or substructure diagrams, which allowed chemists to perform the searching by themselves rather than being dependent on their information scientist intermediaries. Thus, a lot of popular representations that are still used today have shifted from prioritising their compactness to being specifically designed for computer applications (Weininger, 1988; Heller et al., 2015; Cereto-Massagué et al., 2015).

Nowadays, the reigning paradigm in molecular representations is given by fingerprint vectors, first introduced in , and descriptors. These methods have been experiencing particular popularity, because these representation can easily be used as the input for machine learning techniques for property prediction. However, this paradigm slowly begins to be challenged by newly emerging deep learning techniques such as Graph Neural Networks. Rather than assigning molecules a fixed representation, these techniques aim to learn a flexible representation depending on the properties of interests of the molecules. This new approach seems equally innovative as crazy (TODO different word) and we will discuss the prospects of this in the following.

2.2.2 Molecular Graphs

Molecular graphs are the entities that underlie most molecular notations. In this part we will introduce the abstract mathematical objects underlying these entities and... A molecular graph is a two dimensional object that can be used to represent information about molecules. Since a graph does not contain any spatial information these need to be encoded by features of nodes and edges.

Formally a graph is defined as a tuple of sets $G = (V, E)$, where V are the vertices of the graph and E are the edges. Any edge $e \in E$ is uniquely identified by a pair of vertices (v_1, v_2), $v_1, v_2 \in V$

that it connects. In a molecular graph the vertices are given by the atoms and edges represent bonds between atoms. Compared to data structures like vectors, graphs are very high dimensional and irregular, simultaneously enabling the representation of more complex information and being harder to process.

In a computer, graphs are represented by a matrix - most commonly by their adjacency matrix A. The entries of this matrix are given by

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } v_i \text{ to } v_j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note that for an undirected graph, like a molecular graph, the adjacency matrix is always symmetric. In order to represent a graph by its adjacency matrix, we need to make a non canonical choice of ordering the nodes. This is inconvenient for molecular graphs since these do not possess any kind of ordering and hence our representation is not well-defined.

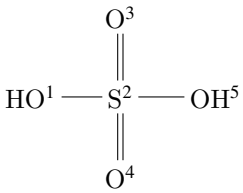

Figure 1: Molecular graph of sulfuric acid.

$$\begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \end{array}$$

Figure 2: Adjacency matrix of the molecular graph representing sulfuric acid given the node ordering

2.2.3 Fingerprint Vectors

2.2.4 Descriptors

2.3 QSAR and QSPR models

2.4 Descriptor based models

2.4.1 Fingerprint

2.4.2 Descriptors

2.5 Application in Drug Design

REFERENCES

Molecules and molecular compounds. https://chem.libretexts.org/@go/page/21702, 2021. Retrieved: April 1, 2021.

Alexandr Mikhaylovich Butlerov. Einiges über die chemische structur der körper. *Zeitschrift für Chemie*, 4:549–560, 1861.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71: 58–63, 2015.

P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, and J. Mockus. The cas online search system. 1. general system design and selection, generation, and use of search screens. *Journal of Chemical Information and Computer Sciences*, 23(3):93–102, 1983. doi: 10.1021/ci00039a002. URL https://doi.org/10.1021/ci00039a002.

M. Gallop, R. W. Barrett, W. Dower, S. Fodor, and E. Gordon. Applications of combinatorial technologies to drug discovery. 1. background and peptide combinatorial libraries. *Journal of medicinal chemistry*, 37 9:1233–51, 1994.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):1–34, 2015.

Bonnie Lawlor. The chemical structure association trust. *Chemistry International*, 38(2):12–15, 2016. doi: doi:10.1515/ci-2016-0206. URL https://doi.org/10.1515/ci-2016-0206.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.01.021. URL https://www.sciencedirect.com/science/article/pii/S0092867420301021.

John Van Drie. Computer-aided drug design: The next 20 years. *Journal of computer-aided molecular design*, 21:591–601, 10 2007. doi: 10.1007/s10822-007-9142-y.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://pubs.acs.org/doi/abs/10.1021/ci00057a005.

William J Wiswesser. 107 years of line-formula notations (1861-1968). *Journal of Chemical Documentation*, 8(3):146–150, 1968.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning, 2018.

List of Figures

List of Tables

Appendix