

---

# FIXED AND LEARNED REPRESENTATIONS IN EARLY STAGE DRUG DISCOVERY

CANDIDATE NUMBER:

In Fulfillment of Assessment for  
'Topics in Computational Biology'

April 20, 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
1.1	Outline of the thesis . . . . .	1
1.2	Disclaimer . . . . .	1
1.3	Molecules . . . . .	1
1.4	Brief history of molecular representations . . . . .	2
2	Fixed Representations . . . . .	3
2.1	Numerical Descriptors . . . . .	3
2.2	Fingerprint Vectors . . . . .	4
3	Learned Representations . . . . .	7
3.1	Molecular Graphs . . . . .	7
3.2	Graph Neural Networks . . . . .	8
3.2.1	Convolutional Neural Networks for Learning Molecular Fingerprints . . . . .	9
3.3	Directed MPP ? Yang et al. (2019a) . . . . .	10
3.3.1	Graph Convolutional Neural Networks . . . . .	10
3.3.2	Graph Attention Networks . . . . .	10
3.3.3	Attentive FP . . . . .	10
3.4	Sequence modeling . . . . .	10

4	Application in Drug Discovery . . . . .	10
4.1	Property Prediction . . . . .	10
4.2	De novo design . . . . .	12
5	Discussion . . . . .	12
6	Conclusion. . . . .	13
	References. . . . .	18
	List of Figures . . . . .	19
	List of Tables . . . . .	20
	Appendix . . . . .	20
A	Similarity values for fingerprints . . . . .	20
B	Kostenrechnung . . . . .	20
C	Ergebnisse . . . . .	20

---

## 1 INTRODUCTION

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* screening 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. Lead compounds were found by isolating natural products from microbiological fermentation, plant extracts and animal sources Gallop et al. (1994). 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 they 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).

On October 5, 1981 a new version of the 'Fortune' magazine was released. 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 a stage of naive euphoria in computational drug design with investments of millions of dollars in hardware and software.

### 1.1 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.

### 1.2 DISCLAIMER

High level approach. Mathematically rigorous description can be found in Kerber (2014).

### 1.3 MOLECULES

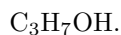
Atoms are the smallest identifiable units of chemical elements which make up all matter in the universe. A fundamental principle of chemistry is that the atoms of different elements can combine to form chemical compounds and a lot of the study in chemistry is centered around understanding what happens 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.

#### 1.4 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



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). Most prominently, the SMILES (Simplified Input Line Entry System) representation (Weininger, 1988) assigns a molecule a string of characters, where atoms are encoded by their atomic symbol and bonds are depicted by one of the following symbols: (-, =, #, \*, ,). Furthermore, branches, rings and charge can be represented by the use of brackets numbers signs (+, -) making SMILES a versatile linear notation that is used to date.

Nowadays, the use of molecular representations has branched. On the one hand, atom based representations like SMILES or InChI are still present in chemical databases in order to uniquely identify a given molecule in a convenient language. These can be used in order to rebuild the molecule based on the representation David et al. (2020). On the other hand, the advent of machine learning methods in all application domains demands a numerical representation of a molecule that represent its properties. Therefore another branch of representations is given descriptors that encode structural or chemical properties. Two established classes of descriptors are described in detail in section 2.

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 FIXED REPRESENTATIONS

Molecular descriptors summarise a class of representations that assign a molecule a fixed vector of numerical values according to some pre-defined properties of that molecule.

According to Todeschini & Consonni (2008) ‘The molecular descriptor is the final results of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment’. This definition highlights the purpose of a descriptor to generate a numerical representation, such as a vector of numbers, from a symbolic representation like a molecular graph. Therefore, descriptors are particularly relevant to applications that require a numerical description of chemical structures like the prediction of chemical or biological properties.

The variety of different descriptors that have been used for QSAR analysis is enormous and depends highly on the considered application. We present a few of the most commonly used descriptors in the following.

The expectations for the usefulness of a descriptor vary a lot depending on the application domain but according to Mauri et al. (2016) these typically include

1. Invariance to node reorderings
2. Invariance to rotations and translations of the molecule
3. Definition by an unambiguous algorithm
4. Well-defined applicability to molecular structures.

These desiderata are supposed to guarantee that the descriptor always gives the same representations for molecules that are considered the same and is generally applicable to all molecules. Beyond that common extra requirements concern the inclusion of structural information (according to the fundamental principle of chemistry that different structures possess different properties), certain discriminative abilities and degeneracy/continuity, i.e. small structural differences result in small but existing differences in the value of the descriptor.

Two classes of descriptors: Numerical descriptors represent the molecule holistically by encoding physical properties. Fingerprints have local approach encoding structural local information among subgroups of atoms.

### 2.1 NUMERICAL DESCRIPTORS

Any attempt to group descriptors into different categories would be quite arbitrary given the sheer amount of different application domains and descriptors. However, Guha & Willighagen (2013) propose an grouping based on the nature of the structural information that they require: Constitutional, topological, geometric and quantum mechanical descriptors.

Constitutional descriptors are the most rudimentary form of descriptors as they do not take into account any spatial information about the molecule but just its basic structural properties. Examples include basic attributes like the molecular weight the number of atoms but also more complex ones such as the sum of atomic van der Waals volumes.

Topological descriptors are based on the connectivity of the atoms in a molecule and encode 2D structural properties using graph invariants of the underlying molecular graphs, i.e. properties that only depend on the abstract mathematical object and not on a particular labeling or ordering of the vertices. Such invariants include the Wiener index Wiener (1947); Nikolić et al. (2001)  $W = \frac{1}{2} \sum_{i,j}^N d_{ij}$ , where  $N$  is the number of non-hydrogen atoms and  $d_{ij}$  is the edge count of the shortest part between atoms  $i$  and  $j$ . A drawback of topological descriptors compared with constitutional descriptors is that they often tend to be less interpretable due to the abstract nature of the underlying graph.

Geometric descriptors receive 3D information about the molecule as their input which may be resourceful to obtain from crystallographic data or molecular optimization Mauri et al. (2016). However, they may also come with more information compared to descriptors that receive lower dimensional inputs. Therefore, they are usually employed in domains when this additional information is critical such as when two conformations are compared (TODO rewrite). An example of a geometric

---

**Algorithm 1:** Morgan Algorithm TODO check with paper

---

**Data:** Molecular graph

**Result:** unique node ordering

Assign each atom the value 1;

**while** *not done* **do**

**for** *atom in atoms* **do**

        | Update value by the sum of the values from the neighbouring atoms;

**end**

**if** *number of different values does not change* **then**

        | break;

**end**

**end**

---

descriptor is given by the 3D Wiener Index which extends the 2D case by weighing the edges by their actual length or the gravitation index Katritzky et al. (1996).

Finally, quantum mechanical descriptors are based on quantum mechanical calculations. An application domain of them are QSAR studies (Reenu & Vikas, 2015; Eroglu & Türkmen, 2007; Senior et al., 2011) to predict toxicity of chemicals for example.

Note that these categories are a non-exhaustive classification of descriptors and many others exist such as auto-correlation descriptors (Broto et al., 1984) (TODO one more?). We conclude that descriptors are a popular method to represent molecules as they are a flexible means to encode the properties that are relevant to the particular application domain. However, this comes also with a downside as the performance of the application may heavily depend on the choice of descriptors and this selection is by no means a trivial task.

## 2.2 FINGERPRINT VECTORS

TODO: other fingerprints? one popular class? Daylight fingerprints encodes linear paths in a molecule within a given length

All descriptors considered so far are derived from performing mathematical computations on the underlying structure and give a holistic representation of the substances considered. Fingerprint Vectors on the other hand are characterised by a more local nature. Specifically, they iteratively aggregate information about substructures of the molecule. Originally, fingerprints were developed for substructure and similarity searching. They depict an way to encode the structure of molecules numerically such that structural similarity reduces to the distance in a high dimensional space. Their simplicity has recently made them a popular means to represent molecules for QSAR machine learning models.

Extended Connectivity Fingerprints (ECFPs) were first introduced by the software Pipeline Pilot in 2000 and then described in detail by Rogers & Hahn (2010). The origin of this representation goes back to Morgan (1965) who introduced the Morgan algorithm on which ECFPs are based. This is why they are also often called Morgan fingerprints (true?). This algorithm assigns numerical values to each atom by an iterative process that does not depend on a specific numbering of the atoms. It is depicted in algorithm 1.

ECFPs adapts this algorithm by stopping the while-loop after a predefined number of steps rather than until completion and storing the intermediate values. We outline each part of the full algorithm in detail in the following paragraphs.

In the first step every atom is assigned an integer identifier that can be chosen arbitrarily as long as it is independent of the node ordering, e.g. the atom’s mass or atomic number. Hydrogen atoms are ignored in this. The ECFP rule explained in Rogers & Hahn (2010) is based on the properties used in the Daylight atomic invariants rule (Weininger et al., 1989) that together are hashed in a 32 bit integer value. A set  $A$  is created containing the initial identifiers of all the atoms. Then, for each atom we add the atom’s own identifier and that of its immediate neighbouring atoms together with their bond order to an array (ordered by the atoms’ identifiers and the order of the attaching bonds). These

values are then hashed to get a single-integer identifier which overrides the initial identifier that the atom was assigned. The updated identifiers are added to the set  $A$  if they are structurally unique as outlined below.

Then, the first step is repeated  $n$  times using the updated identifiers of each atom as the the initial identifiers for the next step. After the completion of the  $n$  steps, numerically equal values are removed from the set  $A$  to arrive at the final ECFP. We clearly see ECFP’s local nature. It manages to generate

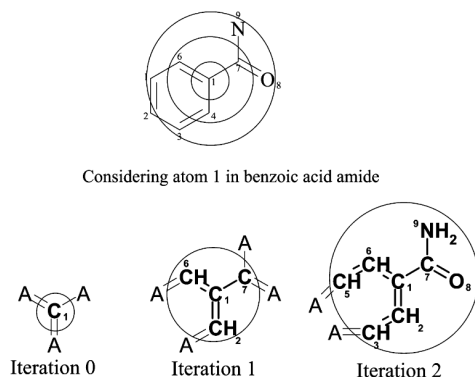


Figure 1: Illustration of the iterative updating in the computation of the ECFPs. In this example the atom type is used as an identifier. In iteration 0 the middle atom’s identifier only represents the information about its own type. After the first iteration it has aggregated the information from its immediate neighbors and after the second iteration the represented substructure has grown even further. Reprinted from Rogers & Hahn (2010).

a global representation by using only local operations thereby implicitly encoding the molecule’s structure. This is opposed to numerical descriptors discussed before which are based on global properties.

Besides numerical equality there is also the notion of structural equality that needs to be taken care of. Consider for example the nitrogen and oxygen atoms at the top and right of the structure shown in Figure 1. Then, after one iteration starting from either of these atoms as the center the exact same substructure is encoded. To avoid this information redundancy, after each iteration it is checked if there are fingerprint features that represent the same bonds generated from the same number of iterations. In this case, these structurally identical values are removed from the set  $A$  that contains the final fingerprints. This also results in fewer feature being generated than in the previous iterations after a couple of steps (maybe delete).

There are two main parameter choices to be made to calculate the fingerprints. On the one hand, the number of iterations  $n$  needs to be specified beforehand which depends on the application domain. Usually  $n = 2$  is used for most applications like similarity or clustering whereas there is a feasible benefit of using a higher  $n$  for activity learning methods Rogers & Hahn (2010). On the other hand, the identifier needs to be chosen which is responsible for the discriminative ability of the fingerprint method. In their paper Rogers & Hahn (2010) describe another fingerprint method FCFP (Functional Class Fingerprints) that is based on the pharmacophore role of the atoms in a molecule. Other types can be used based on different levels of abstraction, e.g. SCFPs (Clark et al., 1989) or LCFPs (Ghose et al., 1998).

There are also other fingerprints that results form small deviations of the algorithm used to compute ECFPs. One example are atom environment fingerprints described extensively in Glen et al. (2006) that use strings in the form of Sybyl atom types as identifiers (Clark et al., 1989) that are iteratively concatenated based on a similar aggregation method as for ECFPs and the final representation is also a circular substructure around each atom. Hashed fingerprints vs keyed fingerprints

As with numerical descriptors fingerprints are also a powerful mean to represent molecules in form of a fixed-size array. But a similar drawback as to numerical descriptors is that the best fingerprints depend strongly on the considered data set which again is non-trivial to find.

In the literature ECFP fingerprints are usually used with  $n = 2$  which is referred to as ECFP4 (4 being the maximum diameter of substructures considered) is considered one of the best performing fingerprints for target prediction Awale & Reymond (2019) and therefore a common baseline for the development of further methods. Another commonly used fingerprint is based on atom pairs (Carhart et al., 1985) which are more suitable for describing large molecules as they are not local as ECFP4 but consider pairs of (non-hydrogen) atoms of arbitrary distance rather than being restricted to radii around atoms.

To investigate potential drawbacks We compare the different Sørensen-Dice similarity values Sorensen (1948); Dice (1945) obtained by using different fingerprints for the molecules illustrated in Figure 2 and 3 respectively. The source code and details about the implementation can be seen in Appendix A. As expected the similarity values decrease when increasing  $n$  and thereby considering larger radii of molecules. Compared to the Atom Pairs scores ECFP2 yields larger similarity scores, especially for molecules c & d, since they are much more complex than a & b. We see that even ECFP6 potentially overestimates the similarity of the two molecules and APFP is presumably more capable of encoding larger molecules.

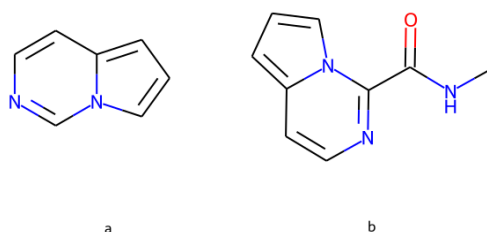


Figure 2: Molecular graphs corresponding to the SMILES strings ‘c1nccc2n1ccc2’ and 1CNC(=O)c1nccc2cccn12’.

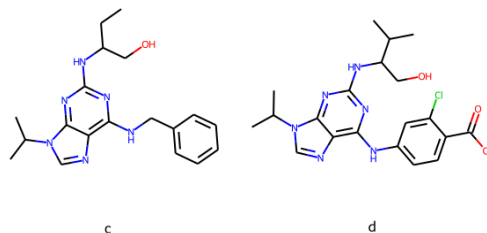


Figure 3: Molecular graphs corresponding to the SMILES strings ‘CCC(CO)Nc1nc(NCc2ccccc2)c2ncn(C(C)C)c2n1’ and ‘CC(C)C(CO)Nc1nc(Nc2ccc(C(=O)[O-])c(Cl)c2)c2ncn(C(C)C)c2n1’

ewyerghiehgi

Molecules	Morgan Fingerprints			APFP
	ECFP2	ECFP4	ECFP6	
a & b	56.25%	46.15%	34.29%	50.88%
c & d	68.66%	58.71%	52.86%	54.47%

Table 1: Dice Similarity values using different fingerprints for molecules in Figure 2 and Figure 3 respectively



### 3 LEARNED REPRESENTATIONS

One of the major drawbacks of using molecular descriptors for drug design is that the performance of the method they are used for is dependent on an a priori selection of features and are therefore biased by expert knowledge Merkwirth & Lengauer (2005). They are designed manually. This means that a huge inductive bias is imposed and the resulting method can only perform as well as the feature selection allows. To remedy this problem an idea is to get away from these representations into a fixed predefined space but rather use machine or deep learning to learn the space itself. A particular method is the use of various kinds of Graph Neural Networks. Instead of just mapping the molecule to this fixed predefined representation Graph Neural Networks operate directly on the Graph and learn the features that are most suitable for the application. In this section we will first introduce Graph Neural Networks in their basic form as well as some extensions and then study the exact workflow in QSAR studies. Continuous degenerated representation due since the representation is encoded by differential mathematical operations and learned via backpropagation.

There are several anticipated benefits of using a learned representation (Shen & Nicolaou, 2019):

1. It does not result in large, sparse representations as fingerprints.
2. It provides a level of interpretability through (Duvenaud et al., 2015) (read)
3. Attention algorithms can be adopted to it (Li et al., 2019; Xiong et al., 2020)
4. It could improve predictive performance on large data sets. (Yang et al., 2019a)

#### 3.1 MOLECULAR GRAPHS

Molecular graphs are the entities that underlie most molecular notations. They are two dimensional objects that can be used to represent information about molecules. An example for a molecular graph is shown in Figure 4. Vertices in the graph correspond to atoms in the molecule and edges represent bonds between them. We also note that the number of edges, i.e. the edge *multiplicity*, may differ. This corresponds to the bond order in the molecule, i.e. the difference between the number of bonds and anti-bonds between two atoms, as introduced by Pauling (1947). However, this graphical representation is not able to encode all information about its underlying molecule such as spatial information. Therefore, these need to be encoded as features of the vertices 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 computers, graphs are represented by a matrix - most commonly by their adjacency matrix  $A$ . The entries of this matrix are given by

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

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.

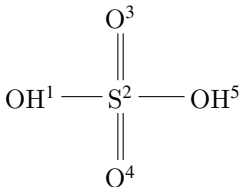


Figure 4: Molecular graph of sulfuric acid.

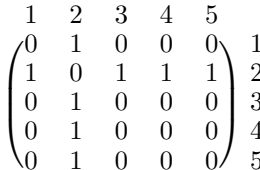

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}$$

Figure 5: Adjacency matrix of the molecular graph representing sulfuric acid given the node ordering.

Figure 5 shows the adjacency matrix corresponding to the graph in Figure 4. The ordering of the vertices is indicated by superscripts. If we assumed a different ordering of the vertices this would result in a permutation of the rows and columns of the adjacency matrix. As we will see, this is a common problem for Graph Neural Network which is attempted to be solved by the introduction of an *inductive bias* devising algorithms that give the same results regardless of a permutation of the matrix.

In order to represent more information about the molecule the adjacency matrix is complemented with two more matrices - a node feature matrix and an edge feature matrix. The node feature matrix has the same number of rows as the adjacency matrix, where row  $i$  corresponds to the feature values for node  $i$ . The number of columns may vary depending on the number of features that are chosen to be encoded. An example feature matrix is shown in Figure 6. Finally, the edge feature matrix contains one row for every edge in the graph, where row  $i$  corresponds to edge  $i$  (TODO edge ordering?) and again the number of columns may vary depending on the number of features, see Figure 7.

$$\begin{array}{cccc} O & S & 0H & 1H \\ \left( \begin{array}{cccc} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array} \right) & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \end{array}$$

Figure 6: Example feature matrix of the graph in Figure 4. The first two columns encode the atom type and the last two columns are a one-hot encoding of the number of implicit hydrogen atoms.

$$\begin{array}{ccc} 1 & 2 & 3 \\ \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array} \right) & \begin{array}{c} (1, 2) \\ (2, 3) \\ (2, 4) \\ (2, 5) \end{array} \end{array}$$

Figure 7: Example edge feature matrix of the graph in Figure 4. The chosen features represent a one-hot encoding of the bond type.

TODO : more sources on molecular graphs While the graphical representation allows for the representation of complex 3D information of molecules, there are some drawbacks of working directly on the graph level. First, not all molecules can be represented as graphs (David et al., 2020) such as those that contain bonds that cannot be explained by valence bond theory. Second, graphs are not a suitable means of depicting molecules whose arrangement of molecules change over time as this would require a reordering of the adjacency matrix every time. Finally, graphs are neither very compact nor easy to process. The adjacency matrix alone has a memory requirement quadratic in the number of atoms in the molecule and depending on the amount of atomic and bond information that is to be encoded the feature matrices might get even bigger. As opposed to this, a linear representation as a single string allows for using substantially less memory while being simultaneously easier to store and process by algorithms. Therefore, graphs are usually used as the basis of more compact representations that we are going to depict in the following subsections.

### 3.2 GRAPH NEURAL NETWORKS

Convolutional Neural Networks (cite) have achieved remarkable success at learning representations of grid-like structures such as images. The idea to generalise these frameworks to less regular structures like graphs motivated the introduction of many Graph Convolutional Neural Networks (GCNNs) like in (Li et al., 2015; Duvenaud et al., 2015; Kearnes et al., 2016; Schütt et al., 2017). An attempt to unify all these approaches in a general framework was made by Gilmer et al. (2017) introducing Message Passing Neural Networks (MPNNs). In the following we will outline how MPNNs work and mention how they restore the previous approaches.

MPNNs manage to represent properties of nodes and edges as well as structural knowledge about the graph. The properties are encoded in the node and edge feature matrices. Structural information is encoded implicitly via a similar aggregation step as for fingerprints in which a node receives knowledge about the neighbouring nodes and updates its own knowledge using that.

An entire forward pass of an MPNN can be divided into two phases: The message passing phase that runs for  $T$  time steps and a consecutive readout phase. Each node stores information about its own features and those of its local environment in a hidden state vector  $\mathbf{h}_v^t \in \mathbb{R}^L$ .  $\mathbf{h}_v^0$  is initialised with the node’s feature vector  $\mathbf{x}_v$ . For each time step during the first phase any node receives ‘messages’

about its neighbours’ hidden states and then updates its own hidden state based on that. Specifically, this can be described as the two equations

$$\mathbf{m}_v^{t+1} = \sum_{w \in N(v)} M_t(\mathbf{h}_v^t, \mathbf{h}_w^t, e_{vw}) \quad (2)$$

$$\mathbf{h}_v^{t+1} = U_t(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}) \quad (3)$$

where  $\mathbf{m}_v^t$  is the ‘message’ node  $v$  receives at time  $t$  which is composed of the sum of the message functions  $M_t$  from its immediate neighbours that can depend on their own hidden state  $\mathbf{h}_w^t$ , the neighbour’s hidden state  $\mathbf{h}_v^t$  and features of the edge connecting them.

After  $T$  time steps, any node  $v$  has now received information about any node  $w$  that are at most  $T$  edges away. This is because after the first step  $w$ ’s neighbors receive information about  $w$ ’s hidden state which is in turn incorporated in their own hidden state. In the next iteration,  $w$ ’s neighbours pass their hidden state, incorporating information about  $w$ ’s hidden state, to their own neighbours. This way, information about  $w$ ’s hidden state is propagated through the graph and after  $T$  iterations,  $v$  receives this information.

The second readout phase now computes a feature vector for the whole graph as given in equation 4

$$\hat{\mathbf{y}} = R(\mathbf{h}_1^T, \dots, \mathbf{h}_{|V|}^T) \quad (4)$$

Different choices for the functions  $M_t$ ,  $U_t$  and  $R$  restore different Graph Neural Networks proposed in the literature. All of them have in common that they are differentiable and learned through backpropagation. Furthermore,  $R$  must be permutation-invariant in order for the MPNN to be insensitive to the node ordering. We illustrate how MPNNs recover two former architectures of Graph

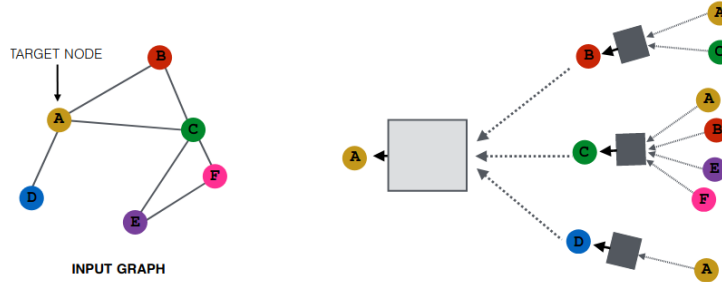


Figure 8: Illustration of the message passing in a MPNN. Reprinted from Hamilton et al. (2018).

Neural Networks. For more detail, we refer to Gilmer et al. (2017).

### 3.2.1 CONVOLUTIONAL NEURAL NETWORKS FOR LEARNING MOLECULAR FINGERPRINTS

This architecture refers to the one proposed by Duvenaud et al. (2015). Here, the message function  $M_t$  is the same across all time steps and given by

$$M(\mathbf{h}_v, \mathbf{h}_w, e_{vw}) = (\mathbf{h}_w, e_{vw}),$$

where  $(\cdot, \cdot)$  denotes concatenation. The update and readout functions are given by

$$U_t(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}) = \sigma(\mathbf{H}_t^{\deg(v)} \mathbf{m}_v^{t+1})$$

which includes learnable parameters as given by the matrices  $\mathbf{H}_t^k$  for all time steps  $t$  and node degrees  $k$ .  $\sigma$  denotes the sigmoid activation function. Finally, the readout function is given by

$$R(\mathbf{h}_1^T, \dots, \mathbf{h}_{|V|}^T) = f\left(\sum_{v,t} \text{softmax}(\mathbf{W}_t \mathbf{h}_v^t)\right)$$

with learnable matrices  $\mathbf{W}_t$  for all time steps  $t$  and a neural network  $f$ .

---

### 3.3 DIRECTED MPP ? YANG ET AL. (2019A)

#### 3.3.1 GRAPH CONVOLUTIONAL NEURAL NETWORKS

These belong to the most popular classes of Graph Neural Networks and was proposed by (Kipf & Welling, 2016). A detailed derivation of the message and update functions can be found on Gilmer et al. (2017). The resulting functions are given by:

$$M_t(\mathbf{h}_v^t, \mathbf{h}_w^t) = \sum_{w \in N(v) \cup \{v\}} (\deg(v) \deg(w))^{-1/2} \mathbf{h}_w^t$$

The update function at time  $t$  is given by:

$$U_t(\mathbf{h}_v^t, \mathbf{m}_v^{t+1}) = \sigma(\mathbf{W}^t \mathbf{m}^{t+1})$$

with a trainable matrix  $\mathbf{W}^t$  and a non-linear activation function  $\sigma$ , e.g. ReLU. This can be thought of as a generalisation of the MPNN framework that includes self-loops in the message passing step.

#### 3.3.2 GRAPH ATTENTION NETWORKS

Graph Attention Networks proposed by Veličković et al. (2018) extend GCNs by replacing the normalisation constants  $\deg(v) \deg(w)$  in the aggregation step by a learned attention score

$$M_t(\mathbf{h}_v^t, \mathbf{h}_w^t) = \sum_{w \in N(v)} \alpha_{vw}^t \mathbf{h}_w^t.$$

The attention score weighs the information from neighbours according to how important they are. Details about how  $\alpha_{vw}^t$  is computed can be retrieved from the paper (Veličković et al., 2018).

#### 3.3.3 ATTENTIVE FP

The most recent state of the art Graph Neural Network architecture was proposed by Xiong et al. (2019). It proceeds similarly to GANs by stacking multiple attention layers together with Gated Recurrent Units (GRUs) to update the nodes' hidden states recursively and allow an atom to focus on its most important neighbors. To generate the final graph embedding, Attentive FP treats the whole molecule as a virtual node that connects to all its atoms. Then, it employs the same architecture as for the individual atoms to learn the molecule's final representation. These GCN, GAN, Attentive FP?

### 3.4 SEQUENCE MODELING

SMILES RNNs, LSTMs? Honda et al. (2019)

## 4 APPLICATION IN DRUG DISCOVERY

Machine learning is has gained a lot of importance to Computer aided drug discovery in recent years (Varnek & Baskin, 2012). Discovery and development of a new drug can take 5000-10000 compounds to screen and 12-15 years to end up with one approved drug requiring costs of more than \$1.3B . Only 2 out of 10 approved and marketed drugs can recover these costs Hecht & Fogel (2009) In this section we study how molecular representations are used in early phases of drug discovery. In particular, we look at the prediction of molecular properties for target prediction or ADME/T and how these representations can be used for de novo drug design.

### 4.1 PROPERTY PREDICTION

One of the primary applications of machine learning in drug discovery is helping researchers to discover relationships of chemical structures and certain desired properties and activities (Lo et al., 2018). For instance, after finding a hit compound in a drug screening campaign researcher would like to understand how its chemical structure can be optimised in order to better (different word not improve) properties like binding affinity, biological responses or physiochemical properties. Fifty

years ago, the only means of solving this problem was through costly and resourceful *in vitro* and *in vivo* screening methods. Nowadays, machine learning methods can be leveraged to model so called quantitative structure-activity/property relationships (QSAR/QSPR) mostly *in silico*.

Early QSAR models, such as the Hansch-analysis (Hansch & Fujita, 1964), were limited by a lack of experimental data and the linearity assumption made for modeling (Lo et al., 2018). The increasing availability of data and more complex AI technologies ... The workflow for molecular property

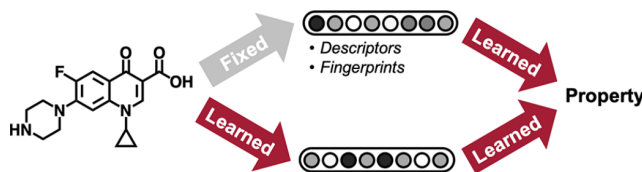


Figure 9: Illustration of the molecular property prediction training workflow. Reprinted from Yang et al. (2019a).

prediction is briefly summarised in Figure 9 and can be divided into two stages where the latter is the same for fixed and learned representations. In the first stage a particular molecule needs to be transformed into a machine-interpretable shape. We have studied two approaches to this, namely fixed and learned representations. For fixed representations expert knowledge is needed in order to select features of the molecules that are relevant to the property of interest. Learned representations such as Graph Neural Networks automate this process by computing a molecular representation automatically based on the property of interest.

After the featurization step, machine learning methods can be employed to learn the relationship of the molecular representations and the desired property. Pretty much any machine learning methods has been applied in this second step Shen & Nicolaou (2019) and popular examples include support vector machines Heikamp & Bajorath (2013); Zernov et al. (2003), extreme gradient boosting Jiang et al. (2020a); Yang et al. (2019b) and random forest Svetnik et al. (2003).

As an example for a full MPP workflow, we choose Stokes et al. (2020) who used a Graph Neural Network to predict the growth of *E. Coli*. Their approach can be divided into three stages. The first stage concerns the training of the model and a classifier according to Figure 9. The molecular representation was built using a directed-message passing neural network Yang et al. (2019a) and can therefore be classified as a learned representation. Similarly to ECFP fingerprints, (D-)MPNNs can struggle to represent global features of molecules, especially if the number of message passing iterations is greater than the longest path in the molecule as discussed in section 3.2. Therefore, the final representation generated by the D-MPNN was augmented with 300 additional molecule-level features. This combined representation was then input in a feed-forward neural network that outputs a number between 0 and 1 as the prediction of the molecule showing growth inhibitory against *E. Coli*. This whole architecture is trained in an end-to-end fashion such that the D-MPNN can generate a representation that is highly attuned to the desired property. The training of this architecture was performed using a set of 2335 molecules that had been classified as hit or non-hit using 80 % growth inhibition against *E. coli* BW25113 Zampieri et al. (2017) as a hit cut-off. On the test data this model achieved an AUC-ROC score of 0.896.

In the second stage, 20 folds of the trained model using different weight initialisations were applied to 6,111 molecules from the Drug Repurposing Hub (Corsello et al., 2017) to predict their probability of growth inhibition against *E. Coli*. The 20 different results were averaged to arrive at the final prediction scores.

Finally, the best scoring 99 molecules were empirically tested for growth inhibition out of which 51 displayed this property. The resulting 51 molecules were ranked according to their clinical phase of investigation, structural similarity to the training data set and their toxicity that was also predicted using a D-MPNN. This resulted in the discovery of the broad-spectrum bactericidal antibiotic halicin with a very low structural similarity to its nearest neighbour antibiotic in the training data emphasising the model's capacity to generalise.

This case study shows the versatility and potential of using Graph Neural Network for property prediction in early drug discovery. They could be employed for both prediction of growth inhibitory

---

effects as well as toxicity and resulted in the finding of a new antibiotic after years of stagnation in this field. Stokes et al. (2020) also reported the prediction scores using Morgan fingerprints and various classifier and the rank of the newly discovered antibiotic halicin was lower in all of them ranging between 773-2644 compared to 69 for the D-MPNN approach. Therefore, it could be argued that halicin would not have been found if molecular fingerprints had been used. However, between there is still some correlation among the top scoring molecules. For instance, both the D-MPNN and Morgan fingerprints predict the same highest ranking molecule and the fourth place for D-MPNN is in second place for Morgan fingerprints. The question that remains to be answered is if this is just a correlation of numerical values and halicin being ranked much higher for learned representations is just a fortunate coincidence or if the predictions of GNNs actually carry more physical relevance.

Despite this breakthrough using the GNN approach, Stokes et al. (2020) still emphasise the importance of a combination of *in silico* and empirical investigations.

ADMET study, other properties?

#### 4.2 DE NOVO DESIGN

### 5 DISCUSSION

Artificial intelligence and machine learning are currently one of the most rapidly evolving research areas and the progress in these fields has direct impacts on a great variety of disciplines. In particular, we have hinted at their potential to revolutionise the entire field of drug discovery coming with significant reductions in time and resources (TODO where? maybe time span to see how little time). Most recently, a variety of Graph Neural Networks has been introduced as a way to automatize the feature selection for molecular property prediction. Instead of relying on expert knowledge to select the most relevant attributes to be used for a computer-interpretable interpretation, which has been shown to heavily impact the performance of the property prediction (Tian et al., 2012), Graph Neural Network manage to learn a continuous vector representation that is highly attuned to the property of concern.

While many studies report that learned representations are superior to fixed representations in term of the property prediction accuracy for a variety of different applications (Wu et al., 2018; Yang et al., 2019a; Korolev et al., 2020), there is still no consensus on this and others report the dominance of descriptor-based approaches and fingerprints (Mayr et al., 2018; Jiang et al., 2020b). This suggests that there are other relevant factors that influence which approach is better. Since there a substantially more parameters involved in learning a representation compared with using a fixed representation a sufficiently large data set is critical to learned approaches. Something else to take into account is the mode of evaluation. As mentioned by Shen & Nicolaou (2019), the evaluation of model performance is critical to molecular property prediction. This is because unlike images there is no standard to generating ground truth labels for the data. These are usually obtained from experiments and experimental procedures can differ and are subject to human errors. Furthermore, baseline models are often not tuned enough to reach peak performance. ( Finally a fundamental assumption of employing and comparing machine different machine learning models is that training and test data are all independently identically distributed. It has been noted that for different molecules this requirement is very hard to verify let alone achieve.) (find source)

In terms of the required computational resources, fixed representations can be computed much quicker than learned approaches

The last aspect to take into consideration is interpretability. Graph Neural Networks like all deep learning algorithms work as a black box. There is no real way to assign any meaning to its final representation in terms of interpretability. For descriptor based models on the other hand the SHAP method (Lundberg & Lee, 2017) allows for a way to interpret the final prediction scores by computing the contribution of each input feature that had been selected. Therefore, it enables an understanding of which features turned out to be the most relevant for a particular property.

I personally think that the future of property prediction is within learned molecular representations. While their lack of interpretability is a considerable drawback, there are two major advantages. First, GNNs are able to achieve state-of-the-art performance and they have already successfully used to impel (word?) areas that were stagnating before their introduction (Stokes et al., 2020). While there

---

are still publications reporting better results for descriptor-based approaches, GNN’s great potential to be adjusted will probably keep improving their results (phrasing). For example, since the message passing approach may struggle to represent global properties of a graph, a global readout (cite) has been proposed helping overcome this. Secondly, GNNs enable their application to property prediction without having to rely on domain experts that need to select appropriate features. This allows for a wider application across disciplines making GNNs a versatile and promising tool for the future.

## 6 CONCLUSION

In this report we have studied fixed and learned molecular representations for molecular property prediction. Two classes of learned representations were introduced, namely descriptor-based approaches and molecular fingerprints. We compared atom-pair descriptors with the most popular Morgan fingerprints to understand how they capture similarities between different molecules. We found that .... After that, we introduced molecular graphs and Graph Neural Networks that operate directly on the graph level as an example for a learned representation. Recent advancements for GNNs were outlined with the general message passing framework and more recent improvements through D-MPNNs, Graph Attention Networks and Attentive FP. These highlight the capability of further improvements for GNNs and henceforth their potential in molecular property prediction. Finally, we compared learned representations with fixed representations in terms of accuracy, computational costs and interpretability. Despite fixed approaches being better in terms of the two latter aspects, we hypothesised Graph Neural Networks to be the future molecular property predictions. On the one hand this was because of their state-of-the-art performance that is probable to be improved further due to their flexibility to be extended (wphrasing) and on the other hand due to their wide applicability given that they do not require expert knowledge to be used.

---

## REFERENCES

- Molecules and molecular compounds. <https://chem.libretexts.org/@go/page/21702>, 2021. Retrieved: April 1, 2021.
- Mahendra Awale and Jean-Louis Reymond. Polypharmacology browser ppb2: Target prediction combining nearest neighbors with machine learning. *Journal of Chemical Information and Modeling*, 59(1):10–17, 2019. doi: 10.1021/acs.jcim.8b00524. URL <https://doi.org/10.1021/acs.jcim.8b00524>.
- P Broto, G Moreau, and C Vandycke. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *European journal of medicinal chemistry*, 19(1):71–78, 1984.
- Alexandr Mikhaylovich Butlerov. Einiges über die chemische structur der körper. *Zeitschrift für Chemie*, 4:549–560, 1861.
- Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985. doi: 10.1021/ci00046a002. URL <https://doi.org/10.1021/ci00046a002>.
- 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.
- Matthew Clark, Richard D. Cramer III, and Nicole Van Opdenbosch. Validation of the general purpose tripos 5.2 force field. *Journal of Computational Chemistry*, 10(8):982–1012, 1989. doi: <https://doi.org/10.1002/jcc.540100804>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540100804>.
- Steven Corsello, Joshua Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi Hirschman, Stephen Johnston, Anita Vrcic, Bang Wong, Mariya Khan, Jacob Asiedu, Rajiv Narayan, Christopher Mader, Aravind Subramanian, and Todd Golub. The drug repurposing hub: A next-generation drug library and information resource. *Nature Medicine*, 23:405–408, 04 2017. doi: 10.1038/nm.4306.
- Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12, 09 2020. doi: 10.1186/s13321-020-00460-5.
- Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.
- 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>.
- David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints, 2015.
- Erol Eroglu and Hasan Türkmen. A dft-based quantum theoretic qsar study of aromatic and heterocyclic sulfonamides as carbonic anhydrase inhibitors against isozyme, ca-ii. *Journal of Molecular Graphics and Modelling*, 26(4):701–708, 2007.
- 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.
- Arup K Ghose, Vellarkad N Viswanadhan, and John J Wendoloski. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of alogp and clogp methods. *The Journal of Physical Chemistry A*, 102(21):3762–3772, 1998.



- 
- Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *CoRR*, abs/1704.01212, 2017. URL <http://arxiv.org/abs/1704.01212>.
- Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, 9(3):199, 2006.
- Rajarshi Guha and Egon Willighagen. A survey of quantitative descriptions of molecular structure. *Current Topics in Medicinal Chemistry*, 12:1946–1956, 01 2013. doi: 10.2174/1568026611212180002.
- William L. Hamilton, Rex Ying, Jure Leskovec, and Rok Soscic. Representation learning on networks. <http://snap.stanford.edu/proj/embeddings-www/>, 2018. Retrieved: April 19, 2021.
- Corwin Hansch and Toshio Fujita.  $\rho$ - $\pi$  analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86, 04 1964. doi: 10.1021/ja01062a035.
- David Hecht and Gary Fogel. Computational intelligence methods for admet prediction. *Frontiers in Drug Design and Discovery*, 4, 01 2009.
- Kathrin Heikamp and Jürgen Bajorath. Support vector machines for drug discovery. *Expert opinion on drug discovery*, 9, 12 2013. doi: 10.1517/17460441.2014.866943.
- 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.
- Shion Honda, Shoi Shi, and Hiroki R. Ueda. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery. *CoRR*, abs/1911.04738, 2019. URL <http://arxiv.org/abs/1911.04738>.
- Dejun Jiang, Tailong Lei, Zhe Wang, Chao Shen, Dong-Sheng Cao, and Tingjun Hou. Admet evaluation in drug discovery. 20. prediction of breast cancer resistance protein inhibition through machine learning. *Journal of Cheminformatics*, 12:16, 03 2020a. doi: 10.1186/s13321-020-00421-y.
- Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Chen Guangyong, Ben Liao, Zhe Wang, Chao Shen, Dong-Sheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. 09 2020b. doi: 10.21203/rs.3.rs-79416/v1.
- Alan R Katritzky, Lan Mu, Victor S Lobanov, and Mati Karelson. Correlation of boiling points with molecular structure. 1. a training set of 298 diverse organics and a test set of 9 simple inorganics. *The Journal of Physical Chemistry*, 100(24):10400–10407, 1996.
- Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8): 595–608, Aug 2016. ISSN 1573-4951. doi: 10.1007/s10822-016-9938-8. URL <http://dx.doi.org/10.1007/s10822-016-9938-8>.
- Adalbert Kerber. *Mathematical chemistry and chemoinformatics : structure generation, elucidation, and quantitative structure-property relationships [electronic resource]*. Ebook central. Berlin, 2014. ISBN 9783110254075.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL <http://arxiv.org/abs/1609.02907>.
- Vadim Korolev, Artem Mitrofanov, Alexandru Korotcov, and Valery Tkachenko. Graph convolutional neural networks as “general-purpose” property predictors: The universality and limits of applicability. *Journal of Chemical Information and Modeling*, 60(1):22–28, 2020. doi: 10.1021/acs.jcim.9b00587. URL <https://doi.org/10.1021/acs.jcim.9b00587>. PMID: 31860296.
- G. Landrum. Rdkit: Open-source cheminformatics. <https://www.rdkit.org/docs/index.html>, 2006.

- 
- 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>.
- Xiuming Li, Xin Yan, Qiong Gu, Huihao Zhou, Di Wu, and Jun Xu. Deepchemstable: Chemical stability prediction with an attention-based graph convolution network. *Journal of Chemical Information and Modeling*, 59(3):1044–1049, 2019. doi: 10.1021/acs.jcim.8b00672. URL <https://doi.org/10.1021/acs.jcim.8b00672>. PMID: 30764613.
- Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.
- Yu-Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546, 2018. ISSN 1359-6446. doi: <https://doi.org/10.1016/j.drudis.2018.05.010>. URL <https://www.sciencedirect.com/science/article/pii/S1359644617304695>.
- Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.
- Andrea Mauri, Viviana Consonni, and Roberto Todeschini. *Molecular Descriptors*, pp. 1–29. Springer Netherlands, Dordrecht, 2016. ISBN 978-94-007-6169-8. doi: 10.1007/978-94-007-6169-8\_51-1. URL [https://doi.org/10.1007/978-94-007-6169-8\\_51-1](https://doi.org/10.1007/978-94-007-6169-8_51-1).
- Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Joerg Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical Science*, 9, 06 2018. doi: 10.1039/C8SC00148K.
- Christian Merkwirth and Thomas Lengauer. Automatic generation of complementary descriptors with molecular graph networks. *Journal of Chemical Information and Modeling*, 45(5):1159–1168, 2005. doi: 10.1021/ci049613b. URL <https://doi.org/10.1021/ci049613b>. PMID: 16180893.
- H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018. URL <https://doi.org/10.1021/c160017a018>.
- Sonja Nikolić, Nenad Trinajstić, and Milan Randić. Wiener index revisited. *Chemical Physics Letters*, 333(3-4):319–321, 2001.
- Linus Pauling. Atomic radii and interatomic distances in metals. *Journal of the American Chemical Society*, 69(3):542–553, 1947. doi: 10.1021/ja01195a024. URL <https://doi.org/10.1021/ja01195a024>.
- Reenu and Vikas. Exploring the role of quantum chemical descriptors in modeling acute toxicity of diverse chemicals to daphnia magna. *Journal of Molecular Graphics and Modelling*, 61: 89–101, 2015. ISSN 1093-3263. doi: <https://doi.org/10.1016/j.jmgm.2015.06.009>. URL <https://www.sciencedirect.com/science/article/pii/S1093326315300176>.
- David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL <https://doi.org/10.1021/ci100050t>. PMID: 20426451.
- Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1), Jan 2017. ISSN 2041-1723. doi: 10.1038/ncomms13890. URL <http://dx.doi.org/10.1038/ncomms13890>.
- Samir A Senior, Magdy D Madbouly, et al. Qstr of the toxicity of some organophosphorus compounds by using the quantum chemical and topological descriptors. *Chemosphere*, 85(1):7–12, 2011.
- Jie Shen and Christos A. Nicolaou. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32-33:29–36, 2019. ISSN 1740-6749. doi: <https://doi.org/10.1016/j.ddtec.2020.05.001>. URL <https://www.sciencedirect.com/science/article/pii/S1740674920300032>. Artificial Intelligence.

- 
- Th A Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34, 1948.
- 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>.
- Vladimir Svetnik, Andy Liaw, Christopher Tong, John Culberson, Robert Sheridan, and Bradley Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43:1947–58, 11 2003. doi: 10.1021/ci034160g.
- Sheng Tian, Junmei Wang, Youyong Li, Xiaojie Xu, and Tingjun Hou. Drug-likeness analysis of traditional chinese medicines: Prediction of drug-likeness using machine learning approaches. *Molecular pharmaceutics*, 9:2875–86, 06 2012. doi: 10.1021/mp300198d.
- Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*, volume 11. John Wiley & Sons, 2008.
- 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.
- Alexandre Varnek and Igor Baskin. Machine learning methods for property prediction in chemoinformatics: Quo vadis? *Journal of Chemical Information and Modeling*, 52(6):1413–1437, 2012. doi: 10.1021/ci200409x. URL <https://doi.org/10.1021/ci200409x>. PMID: 22582859.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.
- 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>.
- David Weininger, Arthur Weininger, and Joseph L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989. doi: 10.1021/ci00062a008. URL <https://doi.org/10.1021/ci00062a008>.
- Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American chemical society*, 69(1):17–20, 1947.
- 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.
- Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Zhong Feisheng, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, H. Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with graph attention mechanism. *Journal of Medicinal Chemistry*, 63, 08 2019. doi: 10.1021/acs.jmedchem.9b00959.
- Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020. doi: 10.1021/acs.jmedchem.9b00959. URL <https://doi.org/10.1021/acs.jmedchem.9b00959>. PMID: 31408336.

- 
- Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019a. doi: 10.1021/acs.jcim.9b00237. URL <https://doi.org/10.1021/acs.jcim.9b00237>. PMID: 31361484.
- Zi-Yi Yang, Zhi-Jiang Yang, Jie Dong, Liang-Liang Wang, Liu-Xia Zhang, Jun-Jie Ding, Xiao-Qin Ding, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. Structural analysis and identification of colloidal aggregators in drug discovery. *Journal of Chemical Information and Modeling*, 59(9):3714–3726, 2019b. doi: 10.1021/acs.jcim.9b00541. URL <https://doi.org/10.1021/acs.jcim.9b00541>. PMID: 31430151.
- Mattia Zampieri, Michael Zimmermann, Manfred Claassen, and Uwe Sauer. Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. *Cell Reports*, 19(6): 1214–1228, 2017. ISSN 2211-1247. doi: <https://doi.org/10.1016/j.celrep.2017.04.002>. URL <https://www.sciencedirect.com/science/article/pii/S2211124717304618>.
- Vladimir V. Zernov, Konstantin V. Balakin, Andrey A. Ivaschenko, Nikolay P. Savchuk, and Igor V. Pletnev. Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of Chemical Information and Computer Sciences*, 43(6):2048–2056, 2003. doi: 10.1021/ci0340916. URL <https://doi.org/10.1021/ci0340916>. PMID: 14632457.

---

## LIST OF FIGURES

1	Illustration of the iterative updating in the computation of the ECFPs. In this example the atom type is used as an identifier. In iteration 0 the middle atom's identifier only represents the information about its own type. After the first iteration it has aggregated the information from its immediate neighbors and after the second iteration the represented substructure has grown even further. Reprinted from Rogers & Hahn (2010).	5
2	Molecular graphs corresponding to the SMILES strings 'c1nccc2n1ccc2' and 1CNC(=O)c1nccc2cccn12'.	6
3	Molecular graphs corresponding to the SMILES strings 'CCC(CO)Nc1nc(NCc2ccccc2)c2ncn(C(C)C)c2n1' and 'CC(C)C(CO)Nc1nc(Nc2ccc(C(=O)[O-])c(Cl)c2)c2ncn(C(C)C)c2n1'.	6
4	Molecular graph of sulfuric acid.	7
5	Adjacency matrix of the molecular graph representing sulfuric acid given the node ordering.	7
6	Example feature matrix of the graph in Figure 4. The first two columns encode the atom type and the last two columns are a one-hot encoding of the number of implicit hydrogen atoms.	8
7	Example edge feature matrix of the graph in Figure 4. The chosen features represent a one-hot encoding of the bond type.	8
8	Illustration of the message passing in a MPNN. Reprinted from Hamilton et al. (2018).	9
9	Illustration of the molecular property prediction training workflow. Reprinted from Yang et al. (2019a).	11

---

## LIST OF TABLES

1	Dice Similarity values using different fingerprints for molecules in Figure 2 and Figure 3 respectively . . . . .	6
---	---	---

## APPENDIX

### A SIMILARITY VALUES FOR FINGERPRINTS

Used RDKit(Landrum, 2006) implementation. Note that this library implements Morgan fingerprints which use the same algorithms as the one proposed in (Rogers & Hahn, 2010) but with a different hashing function

source code

### B KOSTENRECHNUNG

### C ERGEBNISSE