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 17, 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 Disclaimer

High level approach. Mathematically rigorous descirption can e found in Kerber (2014).

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

$$C_3H_7OH.$$

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). 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 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. They are two dimensional objects that can be used to represent information about molecules. An example for a molecular graph is shown in Figure 1. 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 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

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

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.

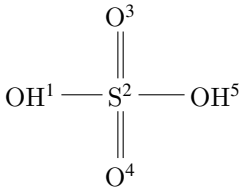

Figure 1: Molecular graph of sulfuric acid.

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

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

Figure 2 shows the adjacency matrix corresponding to the graph in Figure 1. The ordering of the vertices is indicated by superscripts. If we assumed a different ordering of the vertices this would results 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 solve 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 is Figure 3. 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 4.

$$\begin{array}{cccc} O & S & 0H & 1H \\ \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} & \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \end{array}$$

Figure 3: Example feature matrix of the graph in Figrue 1. 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.

$$\begin{array}{ccc} 1 & 2 & 3 \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} & \begin{matrix} (1,2) \\ (2,3) \\ (2,4) \\ (2,5) \end{matrix} \end{array}$$

Figure 4: Example edge feature matrix of the graph in Figure 1. The choses features represent a one-hot encoding of the bond type.

TODO : more sources on moelcular 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.

2.3 Molecular Descriptors

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 differenct 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 desriptors 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 umabiguous 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 inclusidon of structural information (according to the fundamental principle of chemistry that different structures possess differen 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.

2.3.1 Numerical Descriptors

Any attempt to group descriptors into different categories would be quite aribtrary 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 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.

2.3.2 Fingerprint Vectors

TODO: other fingerprints? one popular class?

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.

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

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 oultine 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

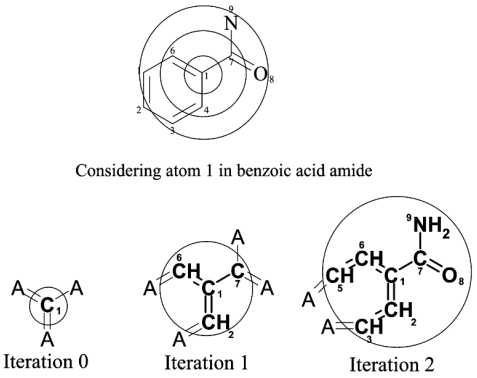

Figure 5: Illustration of the iterative updating in the computation of the ECFPs. In this example the atom type is used as an identifier. In iteration 0 the middle atom' identifier only represents the information about its own type. After the first iteration it has aggregated the information from its 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 5. 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 fingerpint 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 dataset 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 6 and 7 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.

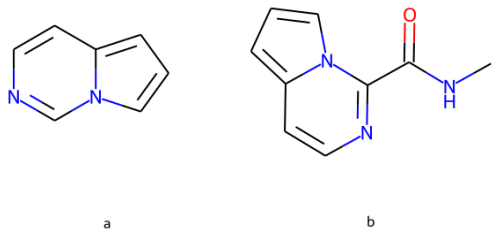

Figure 6: Molecular graphs corresponding to the SMILES strings ‘c1nccc2n1ccc2’ and 1CNC(=O)c1nccc2cccn12’.

ewyerghiehgi eojwojtgowj

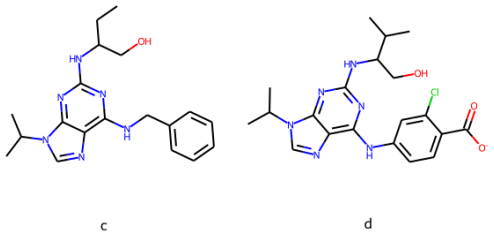

Figure 7: Molecular graphs corresponging to the SMILES strings ‘CCC(CO)Nc1nc(NCc2ccccc2)c2ncn(C(C)C)c2n1’ and ‘CC(C)C(CO)Nc1nc(Nc2ccc(C(=O)[O-])c(Cl)c2)c2ncn(C(C)C)c2n1"

	Morgan Fingerprints			
Molecules	**ECFP2**	**ECFP4**	**ECFP6**	**APFP**
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 6 and Figure 7 respectively

2.4 QSAR AND QSPR MODELS

1962 Hansch

2.5 DESCRIPTOR BASED MODELS

2.5.1 FINGERPRINT

2.5.2 DESCRIPTORS

2.6 APPLICATION IN DRUG DESIGN

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.

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.

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.

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.

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.

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.

Adalbert Kerber. *Mathematical chemistry and chemoinformatics : structure generation, elucidation, and quantitative structure-property relationships [electronic resource]*. Ebook central. Berlin, 2014. ISBN 9783110254075.

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.

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.

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.

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.

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.

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.

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.

LIST OF FIGURES

List of Tables

Appendix

A Similarity values for fingerprints

Used RDKit 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