IDENTIFICATION OF HYDROGEN-BONDING RELATED FUNCTIONAL GROUPS IN MOLECULAR CRYSTALS USING GRAPH THEORY

Aditya Dey

Norwegian University of Life Sciences(NMBU), Norway
Email: aditya.dey@nmbu.no

ABSTRACT

Hydrogen Bonding in compounds is responsible for increasing the boiling and melting temperatures. These hydrogen bonds are formed by functional groups that exist in organic compounds and form inter-molecular or intra-molecular hydrogen bonding. Identification of these hydrogen bonding functional groups allows us to find the solubility of drug-like compounds in aqueous solutions, drug discovery based on the count of the presence of specific functional groups, and various other such applications. In this paper, we propose a method to identify functional groups such as carboxyl, hydroxyl, amines, thiols, aldehydes, and ketones by converting the molecular crystal structure into a network graph using graph theory and further using conditional algorithms to perform functional group detection.

Index Terms— organic compounds, functional groups, graph theory, organic chemistry

1. INTRODUCTION

Organic molecular crystals are made up of carbon and other elements that have an affinity to combine with carbon to make large structures. Even a slight change of one element or group by another in the structure causes a significant change in the physical and chemical properties of the molecule. For example, Methanol(CH_3OH) has a boiling point of $64.7°$C and if the hydroxyl(OH) group is replaced by the thiol(SH) group to form Methanethiol(CH_3SH) the boiling point changes to $5.95°$C which is highly flammable if exposed to a room temperature of $25°$C. Hydroxyl and Thiol are called functional groups defined by IUPAC as "an atom or group of atoms that have similar chemical properties whenever it occurs in different compounds" [1].

These functional groups create a unique bond called a Hydrogen Bond defined as the "electrostatic force of attraction between a Hydrogen(H) atom covalently bound to more electronegative 'donor' atom or group and another electronegative atom bearing a lone pair of electrons-acting as hydrogen bond acceptor" [2]. For example, consider a water molecule(H_2O or $H-O-H$) where oxygen($O^{\delta-}$) has a partial electronegative charge($\delta-$) to which two hydrogens($H^{\delta+}$) are bonded acquiring a partial positive charge($\delta+$).Another oxygen from an adjacent molecule acts as a hydrogen bond acceptor creating a hydrogen bond($O^{\delta-} - -H^{\delta+}$).This increases the bond strength of water molecules due to the inter-molecular bonding among two or more molecules which requires more energy to break these bonds causing its boiling temperature to reach $100°$C. These hydrogen bonds play a larger role in macro-molecules like DNA where the double helical structure is largely due to "hydrogen bonding between base pairs(adenine and thymine)" [3] and polymers such as nylons where they play a significant role in "crystallization of material" [4], reinforcing the material by linking the adjacent chains. According to "Lipinski's or Pfizer's Rule of Five" to evaluate drug-likeness for an orally active drug the criteria suggests no more than 5 hydrogen bond donors and no more than 10 hydrogen bond acceptors [5]. These applications of hydrogen bonds suggest they have a vital role to play and identifying these hydrogen bonding-related functional groups can help us to identify various chemical and physical properties of a molecular crystal.

The functional groups can certainly be identified with a visual inspection of one molecular crystal at a time using tools like Visualization for Electronic and Structural Analysis(VESTA) but if we need to collect information on millions of such crystals as part of data mining for drug testing using machine learning we need to use mathematical methods for feature extraction. Alexandru T.Balaban in his paper "Applications of Graph Theory in Chemistry" [6] suggests the use of graph theory to describe compounds as molecular graphs where points(nodes of the graph) represent the elements and lines(edges of the graph) represent the covalent bond between the elements. We will discuss his work more in Section 2 and will use some of the methods suggested to convert 3-dimensional configurations of crystals into 2-dimensional graphs and identify functional groups that exist that are discussed in detail in Section 3 and 4. Henceforth we compute results based on visual inspection and compare them with output from our algorithms in Section 5 and evaluate the reasons for misidentified and missing functional groups. In Section 6 we discuss further improvements that can be implemented to our methods.

2. RELATED WORKS

In this section, we will discuss the importance of using graph theory as the best-suited method for representing molecular structures and provide examples of its usage over the years.

Graph Theory began as a mathematical branch of topology introduced by Leonhard Euler in his paper "Seven Bridges of Konigsberg" [7] in 1736. Later in 1936, Denes Konig laid the foundation of the mathematical formulation for graph theory which enabled its exploration into chemistry.

In 1985, Alexandru T. Balaban in his paper "Application of Graph Theory in Chemistry" [6] suggests the implementation of graph theory to represent constitutional and valence isomers, quantitative structure-activity relationships(QSAR), search for graph invariants, reaction graphs and synthon graphs. Constitutional isomers for alkanes are represented with the mathematical formula C_nH_{2n+2}. Balaban suggests ignoring all the hydrogen in the alkanes and focusing only on the carbon chains, forming a hydrogen-depleted graph or skeleton graph. Thus for n-butane(C_4H_10), the graph variant will look like Figure 1(1) where each node is a carbon and the linkage is the single bond. Now if we are to plot various other forms of the n-butane graph we will only have one other variant of the graph represented in Figure 1(2) as isobutane. Hence the graph theory suggests that there can only be 2 different graphs that can be formed with 4 nodes which also relates to the fact that butane has only 2 isomeric forms, that is butane and 2-methyl propane. This indicates that graph theory can facilitate in identification of various forms of isomers for alkane and alkyl.

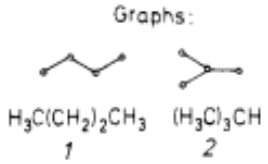


Fig. 1: A hydrogen depleted graph of Butane(1) and IsoButane(2) [6]

QSAR in Balban's article suggested the correlation between physical, chemical, or biochemical properties with the structure of molecules. He hinted towards the molecular structure to be defined into a Topological Indices(TI) and gave an example of its first usage by Harry Wiener's 1947 article "Structural Determination of Paraffin Boiling Points" [8]. According to Wiener, the boiling point is represented as a linear formula $t_{bp} = (98/n_2)W + 5.5P$, where P is the polarity number defined as the number of pairs of carbons atoms separated by three carbon-carbon atoms. The TI in this equation is the parameter "W", the Wiener Index defined as the sum of distances between any two carbon atoms in the molecule. The correlation between the Wiener Index and Boiling Points of lower alkanes is further explored in "Search for Useful Graph Theoretical Invariants of Molecular Structure" [9] where they presented similarity in linear curve of Wiener Index vs boiling points and carbon count vs boiling points in alkanes($C_2 to C_7$).

Today there are more than 200 such TIs available like Randic Index, Hosoya Index, etc. which explore the QSAR and most of them begin with the conversion of molecular structure to a graph to compute the TI.

In modern times with the advent of machine learning, the usage of molecular graph structures for QSAR models grew further. Various methods have been implemented over the years to extract features from the conversion of molecular graphs. "Convolutional Networks on Graphs for Learning Molecular Fingerprints" [10] introduces methods for CNN that can directly operate on graphs. Their pipeline takes inputs as SMILES [11] string encoding and converts them into graphs which are used to create standard circular fingerprints. These fingerprints are unique identification values similar to TI and based on these fingerprints linear model is trained to predict the solubility of the molecules. The circular fingerprint is created by extracting information about the atom and its neighboring substructures from the graph and using a Hashing function to make it unique. Further, they perform Indexing on the hashed output of a molecule that combines all feature vectors into a single fingerprint. Then Canonicalization is performed which sorts the neighboring atoms according to their features and bond features which is the final circular fingerprint made from the conversion of a molecular graph into a modern form of TI.

"Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction" [12] uses molecular graph conversion to fingerprint vectors for training CNN to predict octanol solubility, melting point, and toxicity. Their model begins by representing a molecule as an undirected graph containing nodes(atoms) with Features F_i and edges(bonds) with Features F_{ij}. The features F_{ij} include Bond order, aromaticity, conjugation, ring-like structure, etc, and feature F_i includes atomic identity, estate index, aromaticity, formal charge, etc. These feature vectors are passed through a nonlinear activation function to form a longer fingerprint vector which is finally passed to a neural network for prediction. This demonstrates the conversion of molecules to their respective graph to extract important features.

Similar to the above feature vectors, we determine the functional group as another important feature vector that needs to be extracted to give more uniqueness to a molecule's fingerprint and thus propose a method to determine this feature using various algorithms in Section 4.

3. THEORY

Before we explore the methods, we will briefly discuss the terminologies like graph theory and functional groups to have a general understanding of the underlying concepts and define

in detail the unique chemical bonding properties of functional groups which helps to create the algorithms under Section 4.

3.1. Functional Groups

3.1.1. Hydroxyl Group

As per IUPAC definition, "compounds in which a hydroxy group, –OH, is attached to a saturated carbon atom R3COH" [1] stating hydroxyl group consists of oxygen singly bonded with hydrogen forming a (-OH) group. This group gets attached to a carbon compound with oxygen creating a single bond with carbon creating compounds such as Ethanol ($C_2H_5 - OH$) as shown in Figure 2 where O1 and H1 represent the hydroxyl (-OH) group.

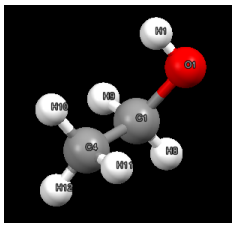


Fig. 2: A hydroxyl group in Ethanol [13]

3.1.2. Carboxyl Group

The Carboxyl group is stated by IUPAC as "oxoacids having the structure RC(=O)OH" [1]. This means the carboxyl group consists of 1-carbon, 2-oxygen where one oxygen creates a double bond with that carbon, and other oxygen creates a single bond with 1-hydrogen. This forms a singular functional group(-COOH), connecting via carbon to organic compounds to create molecules like Acetic Acid ($CH_3 - COOH$) as shown in Figure 3 where C1, O1, O2, and H1 represents the carboxyl(-COOH) group.

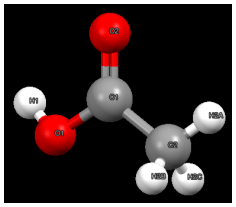


Fig. 3: A carboxyl group in Acetic Acid [13]

3.1.3. Aldehyde Group

As per IUPAC definition, "Compounds RC(=O)H, in which a carbonyl group is bonded to one hydrogen atom and one R group." [1]. It states aldehyde comprises 1-carbon doubly bonded to 1-oxygen, singly bonded to 1-hydrogen. This forms a singular functional group(-CHO) which connects to organic compounds through the aldehyde's carbon to form molecules as Formaldehyde ($H - CHO$)) as shown in Figure.4 where C1, H1, and O1 represent the Aldehyde(-CHO) group.

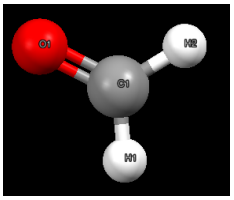


Fig. 4: An aldehyde group in Formaldehyde [13]

3.1.4. Ketone Group

IUPAC states "ketones are compounds in which a carbonyl group is bonded to two carbon atoms R2C=O (neither R may be H)" [1]. This means ketones have mainly 1-carbon and 1-oxygen double bonded to each other. That carbon needs to have singly bonded with 2 other carbons in an organic compound and only then it is considered a Ketone group. Example of ketone is Acetone ($CH_3 - C(O) - CH_3$) as shown in Figure.5 where C1,O1 represents the Ketone group.

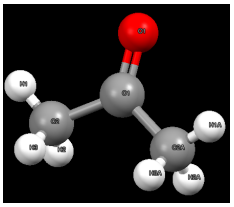


Fig. 5: A ketone group in Acetone [13]

3.1.5. Amine group

As stated by IUPAC, "Compounds formally derived from ammonia by replacing one, two or three hydrogen atoms by hydrocarbyl groups, and having the general structures RNH2 (primary amines), R2NH (secondary amines), R3N (tertiary amines)" [1]. This means Nitrogen can bond with 2-hydrogen and 1-compound to form primary amines, 1-hydrogen, and 2-compounds to form secondary amines, and 3-compounds to form tertiary amines. An example of the tertiary amine is ...

3.1.6. Thiol group

IUPAC defines thiol as "Compounds having the structure RSH where $R \neq H$" [1]. This means sulfur is bonded with 1-hydrogen and the other bond forms with the compound. Example of thiol is Ethanethiol ($C_2H_5 - SH$) as shown in Figure. where S1 and H1 represent the thiol group.

3.1.7. Halogen group

Elements namely Chlorine(Cl), Fluorine(Fl), Bromine(Br) and Iodine(I) belonging to group 17 in the periodic table are halogens. An example of a Halogen group is

3.2. Graph Theory

Graph Theory is a mathematical branch that deals with the study of graphs. According to Williamson, "An undirected graph is made up of vertices also known as nodes that connect using lines called edges" [14]. Mathematically graph is represented as $G = (V, E)$ where, V is a set of nodes and, E is a set of edges ($E \subseteq \{\{x, y\} \mid x, y \in V \text{ and } x \neq y\}$) comprised of pairs of nodes. Figure.6 represents an undirected graph where nodes(V) = $\{1, 2, 3, 4, 5\}$ and edges(E) = $\{(1, 2), (2, 3), (2, 4), (4, 5)\}$. The degree of a node is the number of edges connected to the node therefore $deg(2) = 3$, $deg(1) = 1$, and $deg(4) = 2$.

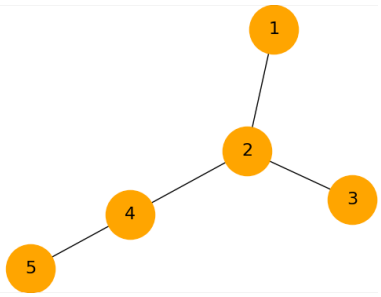


Fig. 6: An undirected graph with 5 nodes

3.3. Crystallographic Information File

Crystallographic Information File(CIF) is a standard text file format for representing crystallographic information published by Hall, Allen, and Brown [15] in 1991 and currently maintained by the International Union of Crystallography(IUCr). It consists of specifications for the molecular view of crystallographic data and allows electronic transmission between individual laboratories, journals, and databases. Within the CIF, our area of interest lies in the three-dimensional coordinate positions and angles of atoms and their symmetry to reconstruct them in VESTA for visualizing or using them to convert from 3D geometry to 2D graphs via programming.

We will extract CIF files from Crystallography Open Database [13] containing records of inorganic, metal-organic, and small organic molecule structural data, and Cambridge Structure Database [16] containing records of published organic and metal-organic small molecule data.

4. PROPOSED METHOD

In this section, we provide a method to convert 3-dimensional coordinates from CIF to a molecular graph. This molecular graph will be used as an input for the algorithms to identify the specific functional groups. The below steps define each action that is performed in the mentioned order for conversion from CIF to a graph.

4.1. Distance Matrix

The distance matrix is a square matrix containing the distances taken pairwise, between each element of the molecule. Since the molecular structure is represented as a three-dimensional coordinate system, the pairwise distance between elements is calculated using equation (1) where (x_i, y_i, z_i) and (x_j, y_j, z_j) represent the x,y,z coordinates of the i^{th} and j^{th} element. The final result is a matrix consisting of distance computation pair-wise between all elements as shown in equation (2).

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{1}$$

$$D_{compound} = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ O_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & O_n \end{array} \\ \begin{bmatrix} D_{C_1C_1} & D_{C_1C_2} & \cdots & D_{C_1O_n} \\ D_{C_2C_1} & D_{C_2C_2} & \cdots & D_{C_2O_n} \\ \vdots & \vdots & \ddots & \vdots \\ D_{O_nC_1} & D_{O_nC_2} & \cdots & D_{O_nO_n} \end{bmatrix} \end{array} \tag{2}$$

4.2. Adjacency Matrix

Adjacency Matrix is a square matrix that indicates if pairs of nodes are adjacent. If we draw a sphere of radius equivalent to the radius of the element then two elements are said to be adjacent if the spheres overlap each other to a certain extent. Mathematically this can be represented by the equation (3) where A_{ij} is equal to 1 if D_{ij} is less than or equal to the sum of radii of i^{th} and j^{th} element else 0, thus forming an adjacency matrix as shown in equation (4).

$$\mathbf{A}_{ij} = \begin{cases} 1 & D_{ij} \leq r_i + r_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$A_{compound} = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ O_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & O_n \end{array} \\ \begin{bmatrix} A_{C_1C_1} & A_{C_1C_2} & \cdots & A_{C_1O_n} \\ A_{C_2C_1} & A_{C_2C_2} & \cdots & A_{C_2O_n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{O_nC_1} & A_{O_nC_2} & \cdots & A_{O_nO_n} \end{bmatrix} \end{array} \tag{4}$$

4.3. Network Graph

The columns or row headers of the adjacency matrix form the nodes of an undirected graph as represented in equation (5) and nodes are said to be connected if $A_{ij} = 1$ hence forming the set of edges as shown in equation (6).

$$\mathbf{V}_{compound} = \{C_1, C_2, H_1, \ldots, O_n\} \tag{5}$$

$$\mathbf{E}_{compound} = \{(C_1, C_2), (C_1, H_1), \ldots, (O_n, H_n)\} \tag{6}$$

4.4. Decomposition

The network graph consists of sub graphs of multiple similar molecules of the compound. In the decomposition process we identify these molecules based on their non-connectivity with each other and segregate them as shown in equation 7 and 8 where V_1 and E_1 represent nodes and edges of a single molecule.

$$V_{compound} = \{V_1, V_2, \ldots, V_n\} \tag{7}$$

$$E_{compound} = \{E_1, E_2, \ldots, E_n\} \tag{8}$$

For the next step, we will only choose the first molecule from the set since they represent similar molecules and denote the nodes and edges of a single molecule as $V_{molecule}$ and $E_{molecule}$.

4.5. Algorithms for Identification

Each functional group will require a different method to identify them based on their chemical bonding properties. They have been subdivided by choosing one major element of the functional group from where we traverse along its adjacent nodes to identify the complete functional group.

In these algorithms we use two functions namely $Degree$ $(node)$ and $Neighbors$ $(E_{molecule}, node)$. $Degree(node)$ function calculates the degree of the specified node. $Neighbors$ $(E_{molecule}, node)$ function returns a list of atoms that are adjacent to the specified node in the molecule.

4.5.1. Halogen Algorithm

Algorithm 1 Identification of Halogens

for $node$ in $V_{molecule}$ **do**
  if $node$ in (F, Cl, Br, I) and $Degree(node) = 1$ **then**
    $adj \leftarrow Neighbors(E_{compound}, node)$
    if $adj = (C)$ **then**
      Halogen group exists.
    end if
  end if
end for

Halogens consist of Fluorine, Chlorine, Bromine, and Iodine and form a single covalent bond with carbon. This bonding property is mathematically represented as Degree(Halogen) = 1. Also as per IUPAC nomenclature, halogens need to bond with carbon only to be called a halogen functional group. If they are bonded to any other element then they form other types of functional groups. Thus using these criteria we propose Algorithm 1 for identifying halogen groups in organic molecules.

First, we iterate through all the nodes of the molecule to check if the halogen element exists and if the degree of halogen is 1. If this condition is met, we extract the adjacent neighbor element and validate if the neighbor is a carbon. If this condition is also met, we declare that the Halogen group exists in the molecule.

4.5.2. Sulphur Algorithm

Algorithm 2 Identification of Thiol

for $node$ in $V_{molecule}$ **do**
  if $node = S$ and $Degree(node) = 2$ **then**
    $adj \leftarrow Neighbors(E_{molecule}, node)$
    if $adj = (C, H)$ **then**
      Thiol group exists.
    end if
  end if
end for

To identify thiol in a functional group sulfur is the key element we use for identifying its existence in the molecule. Sulfur's property is to make two bonds that can be represented mathematically as Degree(Sulphur) = 2 and to be part of the thiol group as per IUPAC, it is mandatory to bond with carbon and hydrogen only. If they bond with any other elements, they form other functional groups like thioethers, etc. Using these key properties we suggest Algorithm 2 to identify thiol functional groups.

As suggested in the algorithm, we iterate over all the nodes of the molecule to check if sulfur exists and if the degree of sulfur is 2. If the condition is met, we check if the neighboring elements are carbon and hydrogen. If this condition is also met, we declare thiol group exists in the molecule.

4.5.3. Nitrogen Algorithm

To identify Amines the key element is Nitrogen. In amines, nitrogen is always bonded with 3 elements having Degree(Nitrogen) = 3 but the elements that bond may differ such that primary amines bond with (2 Hydrogen, 1 Carbon), secondary amines with (1 Hydrogen, 2 Carbons), tertiary amines with (3 Carbons). These chemical bonding configuration allow us to identify amine functional group using the

Algorithm 3 Identification of Amines

for $node$ in $V_{molecule}$ **do**
  if $node = N$ and $Degree(node) = 3$ **then**
    $adj \leftarrow Neighbors(E_{molecule}, node)$
    if adj in $((C, H, H), (C, C, H), (C, C, C))$ **then**
      Amine group exists.
    end if
  end if
end for

Algorithm 3.

According to the algorithm, if during iteration over nodes, nitrogen is encountered and its degree is 3, then we check if neighbors are (carbon, hydrogen, hydrogen) or (carbon, carbon, hydrogen) or (carbon, carbon, carbon). If the criteria match then the Amine group exists.

4.5.4. Oxygen Algorithm

Many of the major functional groups like carboxyl, hydroxyl, aldehyde, ketone, ether, etc. comprise oxygen as a key element. The detection process is more complex since each functional group has a different combination of elements linked to oxygen and also neighbors of the carbon attached to the oxygen need to be checked to identify the complete structure of the functional group. Using Algorithm 4 we try to navigate the complexity of identifying various functional groups based on $Degree(oxygen) = 2$ and $Degree(oxygen) = 1$.

We initialize an empty set called $memorize$ to store oxygen that has been identified already as part of another group to avoid duplicate entries of functional groups. For example, if we are iterating through nodes like previous algorithms, and O_1 is part of a carboxyl group and O_1's neighboring carbon is attached to $O_2(memorize = \{O_1, O_2\})$, then we will memorize both O_1 and O_2 such that when the iteration of nodes reach O_2 we will not check it again.

Next, we perform an iteration over $V_{molecule}$ like previous algorithms and when iteration reaches an oxygen element, we check first if the element does not exist in $memorize$. If the criteria are met then we check if the degree is 1 or 2.

A degree of 1 represents double-bond oxygen which exists in Aldehyde, Ketone, and Carboxyl groups, therefore we validate if the neighboring element of oxygen is carbon. If the criteria are true, then we try to identify all the neighbors of carbon by storing values in $Cadj$.

- If $Cadj = (C, H, O)$ then Aldehyde group exists,
- If $Cadj = (H, H, O)$ then Aldehyde group exists which is a special case of Formaldehyde($H - CHO$).
- If $Cadj = (C, C, O)$ then Ketone group exists,
- If $Cadj = (C, O, O)$ then Carboxyl group exists,
- If $Cadj = (H, O, O)$ then Carboxyl group exists which is a special case of Formic Acid($H - COOH$).

In all the above cases as soon we identify the group we add the oxygen to $memorize$.

A degree of 2 represents a single bond of oxygen connected to two elements that can represent the (OH) part of the carboxyl or the hydroxyl group. We follow a similar process as earlier and store all neighbors of carbon in $Cadj$.

- If $Cadj = (C, O, O)$ then Carboxyl group exists,
- If $Cadj = (H, O, O)$ then Carboxyl group exists which is a special case of Formic Acid($H - COOH$).
- If $Cadj = (C, H)$ then Hydroxyl group exists.

Similar to the earlier process, identified oxygen is stored in $memorize$.

Algorithm 4 Identification of Carboxyl, Hydroxyl, Ketone, Aldehyde

$memorize \leftarrow \{\ \}$
for $node$ in $V_{molecule}$ **do**
  if $node = O$ and $node$ not in $memorize$ **then**
    if $Degree(node) = 1$ **then**
      $adj \leftarrow Neighbors(E_{molecule}, node)$
      if $adj[0] = C$ **then**
        $Cadj \leftarrow Neighbors(E_{molecule}, adj[0])$
        if $Cadj$ in $((C, H, O), (H, H, O))$ **then**
          Aldehyde group exists.
          $memorize.add(carbon\_adj[2])$
        else if $Cadj = (C, C, O)$ **then**
          Ketone group exists.
          $memorize.add(carbon\_adj[2])$
        else if $Cadj$ in $((C, O, O), (H, O, O))$ **then**
          Carboxyl group exists.
          $memorize.add(carbon\_adj[1])$
          $memorize.add(carbon\_adj[2])$
        end if
      end if
    else if $Degree(node) = 2$ **then**
      $adj \leftarrow Neighbors(E_{molecule}, node)$
      if $adj[0] = C$ **then**
        $Cadj \leftarrow Neighbors(E_{molecule}, adj[0])$
        if $Cadj$ in $((C, O, O), (H, O, O))$ **then**
          Carboxyl group exists.
          $memorize.append(carbon\_adj[1])$
          $memorize.append(carbon\_adj[2])$
        end if
      else if $adj = (C, H)$ **then**
        Hydroxyl group exists
        $memorize.append(node)$
      end if
    end if
  end if
end for

Table 1: An Algorithmic Detection of Functional Groups in Organic Compounds

Element Name	Carboxyl	Hydroxyl	Aldehyde	Ketone	Amine	Thiol	Halogen
5-amino-1,3,4-thiadiazole-2-thiol					4		
4-(sulfanylmethyl)-1,3-oxazolidine-2,5-dione	1				1		
Formaldehyde			1				
Acetone				1			
(4-Iodophenyl)propiolic acid	1						1
2-[3-(trifluoromethyl)anilino]benzoic acid	1				1		3
2-azaniumyl-3-(carboxymethanesulfinyl)propanoate	1				1		
2-(2,5-difluorophenyl)pyrrolidin-1-ium 3-carboxy-2-hydroxypropanoate ethanol solvate					1		2
Phenol		1					
4-ethylpiperazin-1-ium 3,5-dinitrobenzoate					1		
(2,3-Dichloro-4-(2-methylene-1-oxobutyl)phenoxy)acetic acid	1			1			2
2,5-bis[(prop-2-en-1-yl)sulfanyl]benzene-1,4-dicarboxylic acid	2						
3-Hydroxy-2-phenylacrylaldehyde		1	1				
3-[2-(2,3-dihydroxy-3-methylpent-4-en-1-yl)-1-methyl-3-methylidene-6-(prop-1-en-2-yl)cyclohexyl]propanoic acid		3					
(1S*,6R*,2'R*)-6-n-Butyl-4-(2'-hydroxypropyl)-3-cyclohexen-1-ol		2					
1-ethyl-3,3,4-trimethyl-1,3-dihydro-2H-spiro[benzo[1,2-b:3,4-b]dipyran-8,2-indol]-2-one	1			1	1		

Table 2: A Visual Inspection of Functional Groups in Organic Compounds

Element Name	Carboxyl	Hydroxyl	Aldehyde	Ketone	Amine	Thiol	Halogen
5-amino-1,3,4-thiadiazole-2-thiol		!			3	1	
4-(sulfanylmethyl)-1,3-oxazolidine-2,5-dione					1	1	
Formaldehyde			1				
Acetone				1			
(4-Iodophenyl)propiolic acid	1						1
2-[3-(trifluoromethyl)anilino]benzoic acid	1				1		3
2-azaniumyl-3-(carboxymethanesulfinyl)propanoate	1				1		
2-(2,5-difluorophenyl)pyrrolidin-1-ium 3-carboxy-2-hydroxypropanoate ethanol solvate	1	2			1		2
Phenol		1					
4-ethylpiperazin-1-ium 3,5-dinitrobenzoate	1				2		
(2,3-Dichloro-4-(2-methylene-1-oxobutyl)phenoxy)acetic acid	1			1			2
2,5-bis[(prop-2-en-1-yl)sulfanyl]benzene-1,4-dicarboxylic acid	2						
3-Hydroxy-2-phenylacrylaldehyde		1	1				
3-[2-(2,3-dihydroxy-3-methylpent-4-en-1-yl)-1-methyl-3-methylidene-6-(prop-1-en-2-yl)cyclohexyl]propanoic acid	1	2					
(1S*,6R*,2'R*)-6-n-Butyl-4-(2'-hydroxypropyl)-3-cyclohexen-1-ol		2					
1'-ethyl-3',3',4-trimethyl-1',3'-dihydro-2H-spiro[benzo[1,2-b:3,4-b']dipyran-8,2'-indol]-2-one				1	1		

5. EXPERIMENTAL RESULTS

Using the algorithms in Section 4 we perform experimentation on a few organic molecular crystals ranging in complexity from small compounds with 1 molecule and 1 functional group to larger compounds with 2-3 molecules and multiple functional groups. We compare the results of algorithmic detection shown in Table 1 with a visual inspection of the molecules using VESTA and Cambridge Structure Database chemical diagrams shown in Table 2. Cells marked with red color in Table 1 represent incorrect detection of the functional group.

From Table 1 we observe there is no incorrect identification of Halogen since the complexity is low due to their

single-bond formation and linkage with carbon only. We have successfully detected the amine group in 5 out of 7 compounds, Aldehyde in 2 out of 2 compounds, and ketone group in 3 out of 3 compounds. This signifies that functional groups that have very distinct chemical bond structures are easier to identify. Hydroxyl identification is successful only for 2 out of 5 compounds. Carboxyl identification is successful for only 5 out of 8 compounds. Thiol group identification failed for all compounds. Hydroxyl and Carboxyl have also been misidentified in 2 compounds.

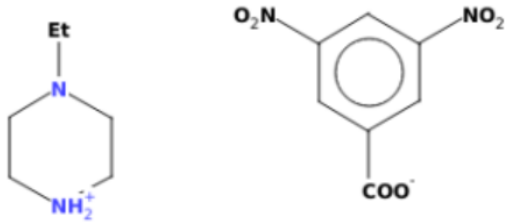
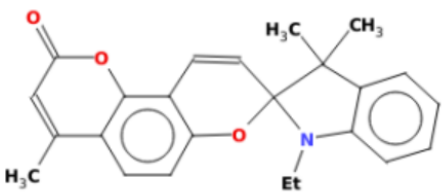


Fig. 7: A chemical diagram of 4-ethylpiperazin-1-ium 3,5-dinitrobenzoate [13]

Figure 7 represents 4-ethyl piperazin-1-ium 3,5-di nitro benzoate consisting of 1 carboxyl, 2 nitrite groups, and 2 amine groups in 2 separate molecules which are part of the compound. In Table 1 for this compound, we can only detect 1 amine group because we assumed that organic molecular crystals only consist of one type of molecule repeated multiple times in the CIF of each compound. Because there are two molecules our algorithm only processes one molecule among them which is the molecule that contains two amines and out of these we only detect tertiary amines and are unable to detect positively charged primary amines. The primary amine has nitrogen connected to two carbon and two hydrogens making 4 bonds and the primary amine's nitrogen is allowed to form 3 bonds only, the extra bond makes the nitrogen positively charged. This behavior of nitrogen to obtain more bonds thus having a degree of 4 has not been accounted for in our algorithm 3 which only checks for amines that have a degree of 3. Hence we observe our algorithm is unsuccessful for compounds that consist of 2 or more different molecules and molecules having cation or anion form of the functional group.

For the compound 1'-ethyl-3',3',4-trimethyl-1',3'-dihydro-2H-spiro[benzo[1,2-b:3,4-b']dipyran-8,2'-indol]-2-one, a carboxyl group is misidentified which does not exist in the compound. This algorithm misidentifies due to carbon being connected to 2 oxygens as shown in Figure 8 and as per the algorithm this was the criteria mentioned and the algorithm did not check for oxygen's hydrogen connection.

These experiments have led to the result that simple conditional algorithms are not sufficient for detecting functional groups in complex compounds. In the next section, we will discuss the improvements that can be applied.



![h]

Fig. 8: A chemical diagram of 1'-ethyl-3',3',4-trimethyl-1',3'-dihydro-2H-spiro[benzo[1,2-b:3,4-b']dipyran-8,2'-indol]-2-one [13]

6. DISCUSSION

As seen in Section 5 we have incorrect results for compounds with 2 or more molecules present. During the Decomposition phase of Section 4, we only choose one molecule from the segregated molecules which is the source of error. If we add a loop in the algorithm to iterate over all molecules of the compound, we will be able to identify the presence of functional groups in the complete molecular crystal.

Secondly, we had incorrect results for molecules whose functional groups have cation and anion forms. This is due to the formation of conjugate bases and acids of the molecules. Oxygen is highly electronegative and tends to form a cation. This will occur in the carboxyl group where it forms a conjugate base called a carboxylate ion(COO^-) by donating one hydrogen connected to an oxygen atom thus leading to one oxygen gaining a formal negative charge of 1 and having $Degree(Oxygen) = 1$. To identify this it will require additional if-then-else conditions in the oxygen algorithm which needs to detect that C_{adj} = (O, O). The conjugate base of primary amine is $(NH_3)^+$ and secondary amine is $(NH2)^+$. In both cases nitrogen forms 4 bonds instead of 3, thus having $Degree(Nitrogen) = 4$. This can be managed by adding an if-then-else condition to the Nitrogen algorithm where $adj = (C, H, H, H)$ and (C, C, H, H). We could not find any molecular crystals having Hydroxyl, Ketone, and Aldehyde in their conjugate form because the double bond of oxygen is stronger and does not have any hydrogen atoms directly linked to donate to form the conjugate base but we had limitations to explore all possibilities in organic chemistry and there may be such molecules which will require the methods to be subjected to change based on the findings. Thus we can surmise that conjugate base-acid forms of many types of functional groups can be identified by the changes in the degree and the combination of surrounding neighboring atoms.

The proposed algorithms are still using very basic if-then-else conditions to match the criteria of a functional group.

Due to time limitations and testing of algorithmic complexities, we could not create various other types of algorithms that can be modified to perform identification but we will suggest a few methods that can benefit us.

Recursive function is defined as a function that calls itself. The proposed algorithm checks the degree of node first and then its adjacent neighbors using nested if-then-else. This can be combined into a recursive function with input as node, degree and an empty list to identify its adjacent neighbors and append set of neighbor of the node like $(H, H, H), (C, C)$, etc. to the list. This way we will iterate over all the nodes and final output will be a list consisting of neighbor's set of each node. Finally we can match the pattern set to identify the type of functional group. This method will allow to combine all algorithms into one single algorithm to form set of neighbors on which pattern matching can be done. This reduces the computational cost of nested if conditions and patterns can be defined easily but the complexity of oxygen related functional group pose a threat where we need to again find ways to identify duplicate set of neighbors that will be created which in our algorithm we were avoiding using $memorize$.

Peter Ertl in "An algorithm to identify functional groups in organic molecules" [17] presents a method to detect by iterative marching through atoms resulting in identification of 3080 unique functional groups. His method begins with identification of all heteroatoms in a molecule including halogens, and further identifies carbon atoms connected by non-aromatic double and triple bond to any heteroatom, acetal carbons connected to two or more nitrogen or sulfurs(O,N and S atoms must have only single bonds) and all atoms in oxirane, aziridine and thiirane rings. Identified atoms are merged based on connected marked atoms to a single functional group. Even though his method does not hints towards usage of any graph theory but his identification of heteroatoms is similar to our method and various other types of atoms suggests more feature extraction from the molecule can help us to identify functional groups in bulk rather than matching set of patterns. His method detects the presence of functional groups but has not provided the solution to classify them which is one of the main criteria for our work.

As discussed in Section 2 earlier, neural networks can play a significant role in performing these identifications. Even though there are no current research papers available for functional group detection using neural networks, the possibility does exist. A CNN model is commonly used for image processing where images are provided in the form of a matrix and features are extracted based on the kernel matrix provided. Similarly, we can also utilize our distance and adjacency matrix to feed to the neural network which can learn relationships in the form of which atoms are always paired together and may create information in the form of an image matrix. Henceforth we can utilize kernel filters that resemble a matrix of functional groups and can attempt to detect using CNN.

There is a need to explore further using neural networks such as CNN and graph neural networks instead of an algorithmic approach as neural networks can learn changes and adapt to the information better than alogrithms where we need to hard code every change that we need.

7. CONCLUSION

Based on the experimentation we have been successful in identifying functional groups in smaller molecular crystals with low complexity where we have only one or two functional groups present. With higher complexity and larger molecules having more than 3-4 functional groups, we observe misclassification and misidentification of carboxyl and hydroxyl and various other groups using our algorithms. Thiol algorithms did not function for any compound even though the degree and neighboring relationship criteria match. Hence we can conclude that the proposed algorithms for halogen, carboxyls, hydroxyls, and amines can work successfully for molecules with lower sizes and low complexity. For molecules with higher complexity and large size, conditional algorithms are not effective and will lead to errors. Feature extraction using neural networks is a better option to explore in the future.

8. ACKNOWLEDGEMENT

I would like to thank my thesis advisor Professor Kristian Berland and co-supervisor Seyedmojtaba Seyedraoufi for providing the necessary guidance in the field of organic chemistry and consistently motivating to steer this paper to be own work.

9. REFERENCES

[1] Alan D McNaught, Andrew Wilkinson, et al., *Compendium of chemical terminology*, vol. 1669, Blackwell Science Oxford, 1997.

[2] Elangannan Arunan, Gautam R. Desiraju, Roger A. Klein, Joanna Sadlej, Steve Scheiner, Ibon Alkorta, David C. Clary, Robert H. Crabtree, Joseph J. Dannenberg, Pavel Hobza, Henrik G. Kjaergaard, Anthony C. Legon, Benedetta Mennucci, and David J. Nesbitt, "Definition of the hydrogen bond (iupac recommendations 2011)," *Pure and Applied Chemistry*, vol. 83, no. 8, pp. 1637–1641, 2011.

[3] M. Spencer, "The stereochemistry of deoxyribonucleic acid. II. Hydrogen-bonded pairs of bases," *Acta Crystallographica*, vol. 12, no. 1, pp. 66–71, Jan 1959.

[4] T.D. Fornes and D.R. Paul, "Crystallization behavior of nylon 6 nanocomposites," *Polymer*, vol. 44, no. 14, pp. 3945–3961, 2003.

[5] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1pii of original article: S0169-409x(96)00423-1. the article was originally published in advanced drug delivery reviews 23 (1997) 3–25.1," *Advanced Drug Delivery Reviews*, vol. 46, no. 1, pp. 3–26, 2001, Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998.

[6] Alexandru T Balaban, "Applications of graph theory in chemistry," *Journal of chemical information and computer sciences*, vol. 25, no. 3, pp. 334–343, 1985.

[7] Leonhard Euler, "Leonhard euler and the koenigsberg bridges," *Scientific American*, vol. 189, no. 1, pp. 66–72, 1953.

[8] Harry Wiener, "Structural determination of paraffin boiling points," *Journal of the American Chemical Society*, vol. 69, no. 1, pp. 17–20, 1947, PMID: 20291038.

[9] Milan Randic, Peter J Hansen, and Peter C Jurs, "Search for useful graph theoretical invariants of molecular structure," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 2, pp. 60–68, 1988.

[10] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams, "Convolutional networks on graphs for learning molecular fingerprints," *Advances in neural information processing systems*, vol. 28, 2015.

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

[12] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen, "Convolutional embedding of attributed molecular graphs for physical property prediction," *Journal of chemical information and modeling*, vol. 57, no. 8, pp. 1757–1772, 2017.

[13] Saulius Gražulis, Daniel Chateigner, Robert T Downs, AFT Yokochi, Miguel Quirós, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail, "Crystallography open database–an open-access collection of crystal structures," *Journal of applied crystallography*, vol. 42, no. 4, pp. 726–729, 2009.

[14] E.A.B.S.G. Williamson, *Lists, Decisions and Graphs*, S. Gill Williamson.

[15] Sydney R Hall, Frank H Allen, and I David Brown, "The crystallographic information file (cif): a new standard archive file for crystallography," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 47, no. 6, pp. 655–685, 1991.

[16] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward, "The cambridge structural database," *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, vol. 72, no. 2, pp. 171–179, 2016.

[17] Peter Ertl, "An algorithm to identify functional groups in organic molecules," *Journal of cheminformatics*, vol. 9, no. 1, pp. 1–7, 2017.