Optical Recognition of Chemical Structures using classical methods

Dhruv Aggarwal
MIT Manipal
Reg No. 220962336
V Sem (CSE-AIML)

Shrey Singh
MIT Manipal
Reg No. 220962298
V Sem (CSE-AIML)

Jai Shah
MIT Manipal
Reg No. 220962362
V Sem (CSE-AIML)

***Abstract*—Optical chemical structure recognition involves converting an image of a chemical molecule into its standard structural form. The goal of the chemical structure recognition algorithm is to accurately identify the molecular graph, assigning the correct labels to atoms/groups at each node and the correct bond types at each vertex. In this work, we present a new approach aimed at improving current state-of-the-art methods, particularly in handling the added challenges that arise when molecules are hand-drawn. We utilize basic text recognition and corner detection techniques to label the atoms and groups that form the nodes of the molecular graph. Our results demonstrate that our corner detection method outperforms the traditional line vectorization algorithms commonly used in existing systems. To detect bonds between nodes, we apply a Hough transform. A key innovation in our approach is the use of a novel method for classifying bonds based on a variety of feature descriptors derived from sliding-window cross-sections of bonds, which are analyzed using supervised machine learning. In addition to the Hough transform, which is also used for bond classification, we incorporate local maxima detectors on single-pixel slices of bond cross-sections, as well as histogram of oriented gradients (HOG) features from broader bond cross-sections. These are paired with several classifiers, including support vector machines (SVM), logistic regression, decision trees, and neural networks. We compare the performance of these feature descriptors, conducting an evaluation on a dataset of 125 hand-drawn simple molecules, and find that our new bond recognition technique significantly enhances recognition performance compared to the baseline methods.**

I. Introduction

In various fields like biology, chemistry, and medicine, the structural diagram remains the standard way to present organic chemical data. While these diagrams contain all the necessary chemical information about a molecule, they are not suitable for computational analysis. The challenge of optical structure recognition—converting these graphical representations into machine-readable, labeled graph data—remains difficult and prone to inaccuracies in many cases. The widespread availability of such data from scientific patents, journals, textbooks, and other printed materials could significantly advance not only chemistry but also fields like pharmaceuticals, chemical biology, medicine, and many others. A tool for chemical structure recognition would also open up new opportunities for artificial intelligence and data mining, allowing researchers to analyze existing datasets that are currently only available in image formats. Additionally, little research has been conducted on extending optical structure recognition to hand-drawn chemical molecules. Most current approaches focus on refining algorithms designed for computer-generated structures, primarily aiming to improve the recognition of large molecules typically found in scientific patents. However, there has been less attention given to recognizing smaller, hand-drawn molecules (Fig. 1), which present unique challenges from a computer vision perspective. Being able to recognize handwritten molecules, in addition to computer-generated ones, would offer several benefits. For instance, creating molecular structures on a computer can be a time-consuming process, and there is currently no real-time application available for recognizing individual components of hand-drawn structures.

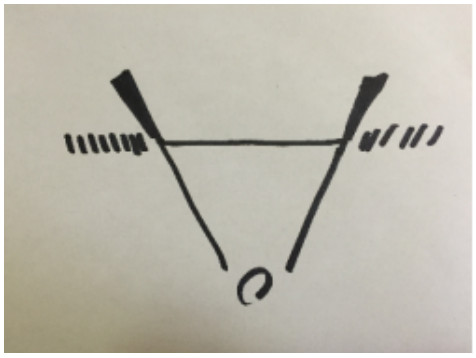


Fig. 1: A molecule from our dataset. Its overall structure is much simpler, but its bonds and labels require different treatment.

II. Review of Previous Work

A. Overview of Previous Research

Previous research in the field of optical structure recognition has primarily concentrated on computer-generated chemical structures. The initial studies in this area date back to the 1990s, with IBM obtaining a patent for the identification of chemical graphics alongside other printed materials on a page. This was accompanied by basic line tracing techniques for recognizing structures. In the same year, researchers from the University of Leeds introduced CLiDE, a similar approach. More recent advancements include ChemReader, developed by the University of Michigan , and the National Cancer

Institute's open-source OSRA. These systems employed more advanced optical character recognition (OCR) and line detection techniques. ChemReader utilized a generalized Hough Transform, while OSRA incorporated the Potrace library. The cutting-edge approach developed by MLOCSR, introduced by Italian researchers Frasconi, Gabbrielli, Lippi, and Marinai, involves a two-step process. The first step is a low-level processing phase that identifies edges, corners, and text, followed by a high-level reasoning engine that leverages Markov logic networks. This engine integrates prior chemical and graphical knowledge to correct any errors from the initial stage. A more recent iteration of CLiDE also incorporates an artificial neural network to classify text labels.

B. Advancements in Current Methods

The primary focus of this paper is on a novel method for the accurate identification of hand-drawn bonds in the low-level processing phase, specifically targeting the correct recognition of atoms and edges without the need for high-level correction using chemical and graphical knowledge. Previous optical structure recognition systems, even state-of-the-art ones, have relied heavily on accurately identifying fine lines, which represent individual elements such as double, triple, and dashed bonds. These methods often fail when applied to imperfect hand-drawn bonds.

Frasconi et al.'s MLOCSR algorithm utilizes the Douglas-Peucker method to approximate the molecular contour with a polygon, selecting the least-vertex polygon that fits within a specified precision. However, we propose that line-detection-based vectorization methods, like the Douglas-Peucker algorithm, may not perform well when bonds are not straight , often leading to an excessive number of vertices. Furthermore, classification algorithms may struggle when dashed lines exhibit irregular patterns or overlap.

Several methods, including the Hough transform, are commonly used for line and segment detection. Applying the Hough transform to hand-drawn images proves to be challenging. Despite efforts to optimize the threshold for the required number of votes to detect lines, the technique still suffers from both false positives and false negatives, particularly due to the inherent characteristics of hand-drawn images.

To address these challenges, we propose using the Hough transform exclusively for bond detection. Once bonds are identified, they are classified based on the features of their horizontal cross-sections. This paper explores a variety of features and classifiers to optimize the accuracy of bond type recognition. These experiments form the majority of this study, with future work planned to enhance text recognition and further investigate the integration of chemical knowledge representations in the higher-level phase.

C. Existing Techniques

Research in OSR began with basic line-tracing algorithms, such as those by IBM and the CLiDE system, which identified basic chemical structures using edge detection and line tracing. Modern methods, like ChemReader and OSRA, incorporate OCR and Hough Transform techniques for improved text recognition and line detection. MLOCSR, a recent approach, combines low-level image processing with high-level reasoning through Markov Logic Networks, addressing limitations in detecting edges and corners.

D. Challenges with Hand-Drawn Structures

Most OSR methods struggle with hand-drawn chemical structures due to irregular line thickness, inconsistent labeling, and the presence of dashed or curved bonds. While the MLOCSR framework utilizes a Douglas-Peucker line-simplification algorithm, it encounters difficulties with dashed bonds and irregular patterns typical in hand-drawn structures. Our approach minimizes reliance on high-level processing, improving low-level detection for hand-drawn molecules by focusing on corner detection and bond classification.

III. Methodology

The proposed methodology includes text recognition, corner detection, bond detection, and classification using a series of preprocessing and machine-learning steps.

A. Overview

The general approach utilized to recognize the chemical structure of molecules involves the following steps:

Detect text labels through scale-invariant template matching Remove text from the image Detect bonds and corners Identify bonds Classify bonds Associate corners with atoms and molecular groups Given the limited availability of hand-drawn chemical data and the small number of labels, we opted for a simpler scale-invariant template-matching technique to detect text in images with reasonable accuracy. Alternative methods, such as Google Tesseract and supervised learning classifiers using histogram of oriented gradients (HOG), can also be explored. However, Tesseract proved difficult to configure because of its general-purpose settings, which rely on multiple language models making numerous assumptions for accurate recognition of structured text. These assumptions, not suited to our specific case, were challenging to eliminate and debug. Although a state-of-the-art OCR engine may be considered in the future, this approach was ultimately unsuccessful. Additionally, HOG classifiers faced challenges, including a high rate of false positives due to a lack of negative training data. While this method could improve with more data, it was not feasible with the limited training set available.

Bond and corner detection were used to identify key points such as intersections of lines representing carbon atoms or endpoints connected to text boxes, signifying atoms or groups. A coarse Gaussian filter was applied to the image before using the Hough transform to locate the lines and points of interest.

Finally, these points were used to locate atoms, groups, and bonds within the molecule. Bond classification was then performed as the final task. Cross-sections of the bonds were analyzed using various feature descriptors, followed by classification using several machine learning classifiers trained on 45 hand-drawn molecules. The success of our method with

such a small training dataset suggests that the approach will improve as more data becomes available. Further details on the bond classification process are presented in the methods section.

B. Data Collection and Preprocessing

The dataset consists of 125 images of nine different simple hand-drawn molecules. These images were drawn using a fine-point black Sharpie marker on standard white printer paper and captured with an iPhone 6 camera at a resolution of 3264 x 2448 pixels, with three color channels (no alpha channel). The images were downsampled to a resolution of 400 x 300 pixels and converted to grayscale using bilinear interpolation. All images were taken under identical lighting conditions, and three different individuals contributed the drawings to ensure that the model would not overfit to a specific drawing style. The images were preprocessed with binarization using a 40% threshold, without any further preprocessing. A subset of 45 images (5 per molecule) was reserved as the training set.

C. Text Recognition

For text recognition, we explored several methods, including scale-invariant template matching, which used five images of each of six templates ("O," "H," "OR," "RO," "N," and "OH"), as well as supervised learning classifiers using four templates ("O," "H," "R," and "N"). Based on the dataset's visual characteristics, we estimated the minimum and maximum scales for the images as 20x20 pixels and 60x60 pixels, respectively. A spatial pyramid sliding window approach was implemented, with window sizes increasing from 20 to 60 pixels in 5-pixel increments. This method was based on a conservative estimate, and we recommend scaling from 0.3

To gather negative training examples, we randomly selected 1,200 image windows from the training set that were confirmed to contain no text. We also collected five examples of each template from the training set. To augment the positive examples, we used 55 images for each template from the open-source Chars74K handwritten dataset. These images were cropped to remove whitespace, and histogram of oriented gradients (HOG) features with 64 bins were extracted from each image. We then compared the performance of a logistic regression classifier, a linear SVM classifier, and a neural network with one hidden layer containing 30 nodes. The results are discussed in section 4.

For the scale-invariant template-matching process, we applied a Gaussian filter with a size equal to half the width of the measured strokes to both the training templates and the image. The spatial pyramid sliding window method was used to match the images, and a tolerance level of 0.77 was selected, as it maximized the F1 score. Non-maximal suppression was applied to remove overlapping bounding boxes. A sample of the successful output from this stage is shown in Fig. 2, while a failed one is also shown in the Fig. 3. The precision and recall graph for the template matching for different tolerance levels is also shown in Fig. 4. The results from this algorithm were then used in subsequent stages of the pipeline. More details on this process are provided in section 4.

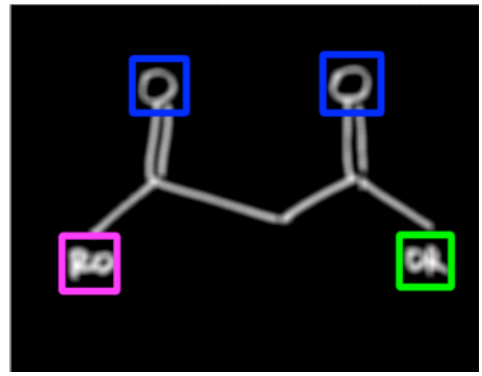


Fig. 2: A result of successful template matching OCR.

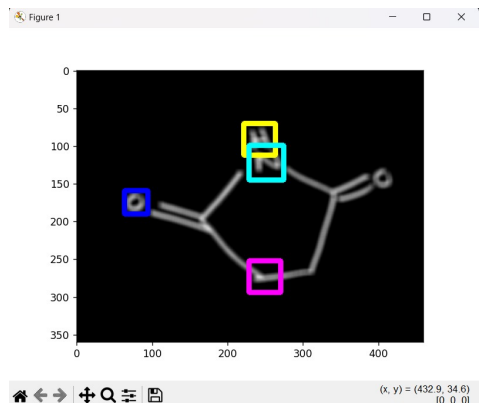


Fig. 3: A result of failed template matching OCR.

D. Corner Detection

We reimplemented the Douglas-Peucker algorithm on our dataset for comparison with MLOCSR, alongside a corner detection algorithm based on a broad Harris corner detector. To clarify, we adopt the terminology from MLOCSR, where a C-point refers to a corner at the intersection of primary bonds representing carbon, a D-point corresponds to the endpoint of a line segment representing a double or triple bond, and a T-point indicates the endpoint of a line segment connected to a text box representing a non-carbon atom or group.

E. Best-Fit Polygon Reimplementation

In line with MLOCSR, we applied the Douglas-Peucker algorithm to detect C-points and T-points and subsequently identified D-points. For each contour, the algorithm iteratively fits n-vertex polygons, increasing n until no point on the contour deviates beyond a specified threshold distance from the polygon. This process returns the vertices of the polygon.

To achieve this, we searched for clusters of all points within polygons fitting the contours of the image after applying a Canny edge detector. We used the threshold of 2 times the edge length, as recommended by MLOCSR.

Next, we applied a basic agglomerative clustering algorithm, setting a maximum distance of 50 pixels between clusters. If a polygon vertex was within 50 pixels of an existing cluster center, it was assigned to that cluster, and the center point of the cluster was updated. Otherwise, a new cluster was initiated. The results of this clustering process, when applied to a molecule image, with blue points representing the final cluster centers. Testing results are detailed in section 4.

F. Harris Corner Detection

The Harris corner detection algorithm is employed to locate key features in the molecule images, specifically the C-points (which correspond to carbon atoms where bonds meet), and T-points (representing endpoints where a bond connects to a non-carbon atom or group). This method is focused on detecting high variations in the image's gradient in two perpendicular directions. The process begins with applying a Gaussian filter to the image. The size of this filter is determined by the estimated stroke width of the drawing, helping to smooth out the image and reduce noise, which makes the corners more detectable.

Next, the Harris corner detector is applied to the filtered image to identify corner points. For accuracy, the algorithm is configured to detect corners that are at least a threshold distance apart, ensuring that the detected points are distinct and not clustered too closely together. When compared to the polygon reconstruction method (used in the earlier step), the Harris corner detection showed superior accuracy, achieving 89% accuracy. In contrast, the polygon method only attained 75% accuracy. This makes Harris corner detection a more reliable method for identifying the C- and T-points in hand-drawn molecular images, where precision in corner detection is crucial for the next stages of the recognition process.

G. Bond Detection

Unlike previous methods in the literature that rely on full line vectorization (which attempts to map the lines representing chemical bonds), this approach specifically focuses on detecting the presence of bonds rather than classifying them directly. Traditional vectorization methods, such as the one used in MLOCSR (a reference method), tend to make errors, especially with hand-drawn molecules. These methods struggle to detect D-points (which indicate the ends of double and triple bonds) with sufficient accuracy, especially when the lines are not perfectly straight or have irregularities.

The method used in this research is simpler but more tailored for detecting bonds in the context of hand-drawn images. The fundamental assumption is that a carbon atom (represented by a node in the image) can form at most four bonds. Therefore, for each node detected in the previous corner detection stage, the algorithm examines the four closest neighboring nodes to check if they are connected by a bond. While molecules could theoretically have more than four bonds at a node, this scenario does not appear in the dataset used, as it focuses on simpler molecules. For more complex molecules, a more generalized approach could involve examining more nodes and using a Markov logic network to filter out incorrect bond detections, though this is omitted for simplicity in this work.

Another heuristic used is that when three nodes are collinear (aligned in a straight line), there cannot be a bond between the two outermost nodes. This heuristic addresses cases where two bonds meet at a 180-degree angle, which is common in structures such as alkynes, and it ensures the algorithm doesn't mistakenly detect a bond between the outer nodes.

Once candidate bonds (potential connections between nodes) are identified, the algorithm refines these candidates by constructing a bounding box around each node-node pair. This bounding box is given a fixed width of 40 pixels, and within this box, the image is divided into windows of a fixed size. The Hough transform is then applied to each window with a very low threshold to detect lines. For a bond to be confirmed, every window in the bounding box must contain a detected line that aligns with the expected direction of the bond. The line orientation is restricted to be within 1 degree of the expected angle for the bond, ensuring that only true bonds are detected.

This method aims to minimize false positives (incorrectly identifying bonds where none exist), which typically occur as erroneous triangular structures that are rarely seen in organic molecules. These false positives are usually removed later by a Markov logic network. However, the method can still produce false negatives (missed bond detections). Most false negatives occur when bonds are missed in ring structures or other predictable configurations. These missed bonds are typically correctable by more sophisticated post-processing techniques, such as the Markov network.

H. Bond Classification

Once bonds are detected, the next step is to classify the type of each bond (e.g., single, double, triple, wedge, or dashed). This classification is necessary to complete the representation of the molecule's structure. The bond classification process begins by extracting cross-sections of each bond from the images. For each bond, a sliding window approach is used, where a window of 10 pixels by 40 pixels slides along the length of the bond. This window captures localized features of the bond's appearance. Typically, each bond will have between 3 to 10 windows depending on its length.

To classify the bonds, Histogram of Oriented Gradients (HOG) features are extracted from each sliding window. These features describe the local structure of the bond by capturing information about the gradient orientations of pixels within the window, which is particularly useful for distinguishing different types of bonds. These features are then used to train various supervised learning classifiers. The classifiers tested in this research include:

Multiclass logistic regression Linear support vector machine (SVM) Decision tree The decision to avoid using a neural network was made due to the relatively small size of the training set, as neural networks are prone to overfitting when data is limited. Each classifier is trained on a set of 62 single

bonds, 33 double bonds, 10 wedge bonds, 10 dashed bonds, and 5 triple bonds, representing different types of bonds found in the dataset.

For each bond in the test set, the same sliding window approach is applied, and the classifier predicts the type of bond based on the HOG features extracted from each window. A voting system is used to aggregate the predictions for each bond. Each window casts a ”vote” for a particular bond type, and the final classification is determined by the majority vote across all windows for that bond.

A sample of the bond classification results is shown in Fig. 5, demonstrating how the sliding windows and classifier predictions lead to the final bond types for the test molecules. The accuracy of the bond classification and overall performance results are detailed in Section 4 of the paper.

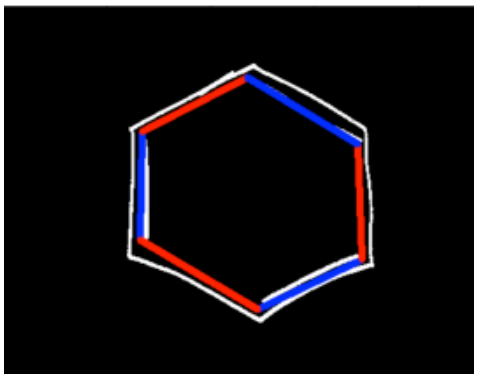


Fig. 4: Final result of bond classification for a benzene ring.

IV. Results and Discussion

A. Scale-Invariant Template Matching

Text recognition was achieved through scale-invariant template matching. The goal was to optimize the tolerance to this method by evaluating precision and recall across the test set. The tolerance value was adjusted to maximize the F1 score, which was found to be optimal. The accuracy was measured at the molecule level, meaning a molecule was counted as correctly recognized only if all its text was correctly identified with no false positives. The results indicated that the method performed well, especially considering the limited training data used for template creation. Further improvement in accuracy could be achieved with a larger training set.

B. Supervised Classifiers

The performance of supervised classifiers, which used HOG (Histogram of Oriented Gradients) features, was assessed using cross-validation. The classifiers, despite performing well on the cross-validation set, struggled to achieve high accuracy on the test set. This was due to the nature of the sliding window detection process, which required perfect matches across numerous windows. The training set was insufficient to provide reliable detection on the test data. Expanding the training set would likely improve the accuracy of the classifiers, but for now, template matching remained the method of choice.

C. Text Recognition Performance

Template matching for text recognition achieved a precision of 0.91 and recall of 0.92 across the dataset, indicating reliable detection of labels such as ”O”, ”H”, ”N”, and ”R”. Despite the dataset’s small size, the template-matching approach proved effective with minimal training data.

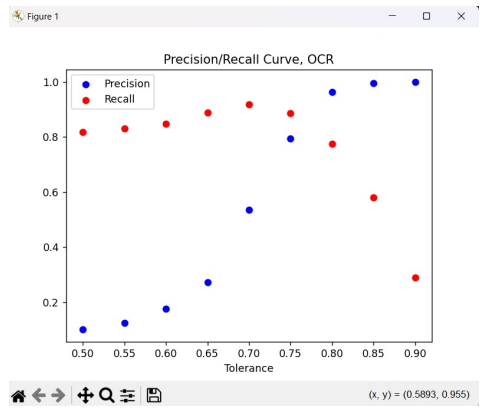


Fig. 5: Precision and recall graph of template matching for different tolerance levels

D. Corner Detection Results

The performance of the Harris corner detector was compared to the MLOCSR (Molecule Level Optical Chemical Structure Recognition) polygon reconstruction method. The Harris method outperformed the polygon method significantly, particularly on molecules with dashed bonds. This was due to the Gaussian smoothing applied to the image before corner detection, which helped blend the dashed bonds into a more recognizable edge. The polygon reconstruction method, which depends on contours of the edges, performed poorly on dashed bonds. The Harris method also showed better results on molecules with curved bonds, such as benzene rings, where it achieved substantially higher accuracy compared to the polygon reconstruction method.

E. Bond Detection and Classification Results

Bond detection, proved to be the weakest performing stage of the pipeline. During this stage, the algorithm was not informed by chemical knowledge regarding the topological structure of molecules. This lack of consideration led to some bond detection errors. For example, the algorithm missed some bonds, which could be rectified in later stages if the rest of the structure was correctly identified. Additionally, issues like false triangular closures occurred when bonds were nearly collinear but not quite, leading to incorrect bond connections. These errors could also be corrected in subsequent pipeline stages.

F. Bond Classification Results

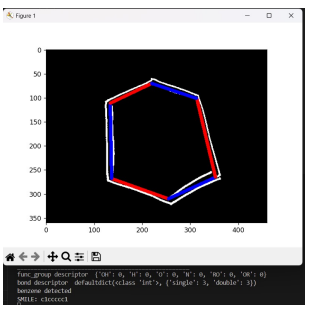
1) Comparison of Classifiers: The bond classification task involved training a set of classifiers and performing cross-validation. Various classifiers were evaluated, including Logistic Regression, Support Vector Machines (SVM), and Decision

Trees. The SVM showed the highest accuracy among the classifiers, and it was chosen for bond classification in the subsequent steps of the pipeline.

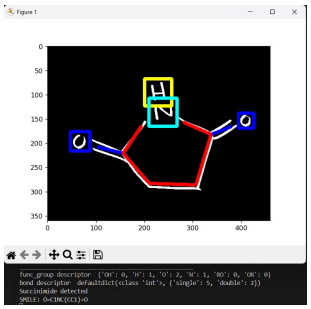
2) Performance on Test Set: The SVM classifier, trained on the cross-validation results, was used to classify bonds on the full set of molecules. The results showed that while the classifier did not confuse one bond type for another frequently, there was a disparity in performance depending on the molecule. Despite having only a small training set, the classifier performed well in correctly classifying bonds, even those with less training data, like triple bonds.

G. Overall System Performance

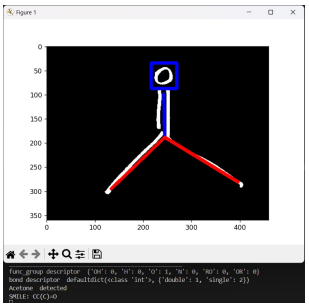
Upon running the entire pipeline on the set of molecules, the results indicated that a significant portion of the molecules were correctly recognized. While the accuracy was not perfect, it outperformed existing optical structure recognition algorithms, which had near-zero accuracy on handwritten data. This was particularly evident when compared to other systems, such as OSRA, which struggled significantly on handwritten datasets. The bond classification method, though trained with minimal data, showed promise and could be improved with additional training data. With more extensive training, nearly 100% accuracy could be achieved in future work. The overall pipeline proved to be effective and was capable of converting recognized molecules into a standard chemical data format.



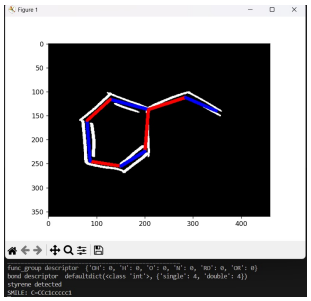
(a) Benzene



(b) Succinimide



(c) Acetone

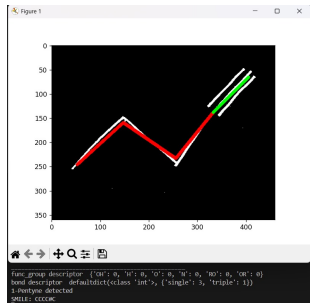


(d) Styrene

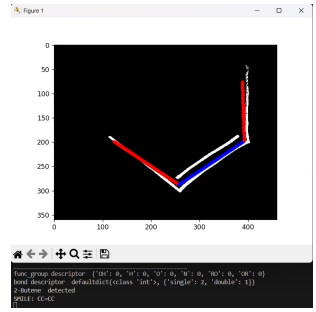
Fig. 6: Final result of molecule detection on different molecules of our dataset

H. Misclassified Results

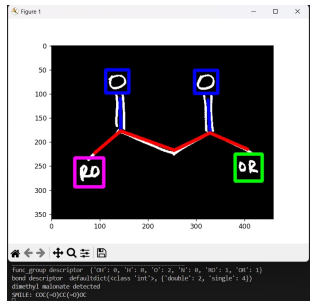
In the current optical chemical structure recognition (OCSR) system, a notable limitation arises in the accurate classification



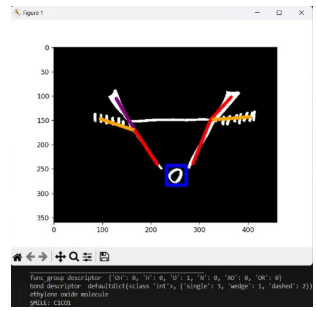
(a) 1-Pentyne



(b) 2-Butene



(c) Dimethyl malonate



(d) Ethylene Oxide

Fig. 7: Final result of molecule detection on different molecules of our dataset

of chemical bonds, particularly double and triple bonds. Due to the similarity in visual appearance under certain conditions, the algorithm occasionally misclassifies bonds, resulting in an incorrect count of single and double bonds within a molecule. This misclassification disrupts the intended molecular structure and leads the system to inaccurately identify molecules or, in some cases, to mark them as unrecognized (Fig. 8).

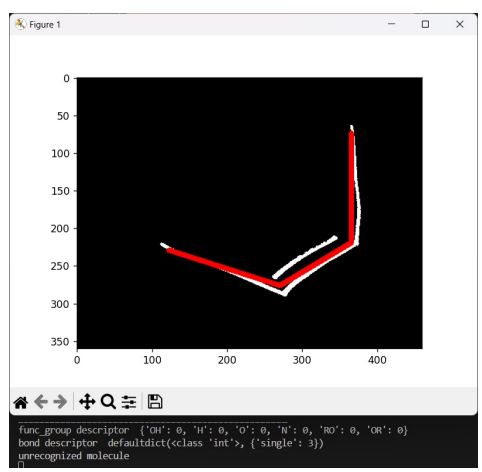


Fig. 8: Misclassified molecule 2-Butene

Another limitation of the current optical chemical structure recognition (OCSR) system occurs in the detection of triple bonds. In certain instances, the algorithm fails to distinguish triple bonds from double bonds, which can lead to misiden-

tification or unrecognizability of the molecule in the dataset (Fig. 9).

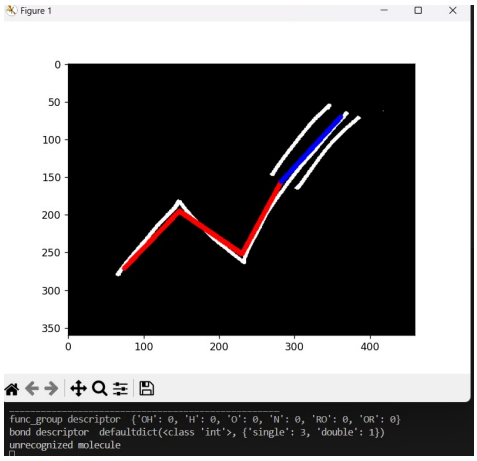


Fig. 9: Misclassified molecule 1-Pentyne

This issue often arises when triple bonds are closely aligned with other bonds or appear slightly obscured, causing the system's bond classification algorithm to interpret them incorrectly as double bonds. Such misclassification disrupts the bond structure required for accurate molecule identification, leading the OCSR system to label the molecule as unrecognized. Improving the system's precision in identifying triple bonds is essential for enhancing its overall recognition accuracy and reliability across complex molecular structures.

TABLE I: OCSR System Results Overview

System	Year	Dataset	Results (%)
OSRA	2009	66 images, C and H structures only	95
chemoCR	2011	1000 structures	65.6
Markov Logic OCSR	2014	937 images-USPOCW	77.6
OCSR	2016	360 structures	26.1
Our model	2024	125 images	83.2

V. Conclusion

Despite the modest accuracy of this approach, our work marks an initial step toward developing reliable recognition systems for hand-drawn chemical structures. The primary reason for the limited accuracy stems from the scarcity of training data, which restricts the model's ability to generalize effectively. Expanding the dataset could notably enhance performance, particularly in text recognition, where advanced OCR techniques could potentially raise accuracy from 77% to near-perfect levels. With more data, it may also be feasible to incorporate convolutional neural networks (CNNs) for bond classification, which would likely yield more accurate results. Currently, our system successfully recognizes 104 out of 125 hand-drawn molecules, achieving an accuracy of 83.2%

This project has primarily targeted the accurate detection of individual atoms and bonds, yet applying higher-level heuristics, like valence checks, could further refine bond identification. For applications such as digital chemistry drawing tools or educational platforms, additional contextual information could be utilized to improve accuracy. For example, incorporating user input characteristics (e.g., stylus speed) might aid in differentiating bond types, and similarity-based algorithms could assist in identifying specific molecules.

Handwritten structure recognition remains challenging due to the inherent variability in hand-drawn inputs. This work addresses these challenges by leveraging sophisticated corner and line detection techniques. A notable innovation of our approach lies in analyzing small segments of bonds to reach a consensus on bond type, rather than assessing entire bonds as in previous methods. While there is room for improvement, these advancements highlight the potential of such techniques for practical applications in the future.

References

[1] Frasconi, P., et al., "Markov Logic Networks for Optical Chemical Structure Recognition," J. Chem. Inf. Model., vol. 54, no. 9, pp. 2380-2390, 2014.

[2] Casey, R., et al., "Optical Recognition of Chemical Graphics," Proc. 2nd Int. Conf. Document Analysis and Recognition, 1993.

[3] Ibison, P., et al., "Chemical Literature Data Extraction: The CLiDE Project," J. Chem. Inf. Comput. Sci., vol. 33, pp. 338-344, 1993.

[4] F. Musazade et al., "Review of techniques and models used in optical chemical structure recognition in images and scanned documents," J. Cheminf., vol. 14, no. 1, p. 1, 2022.

[5] Bradley Emi., "Optical Recognition of Hand-Drawn Chemical Structures," CS231A: Computer Vision (Stanford University), 2016.

[6] Filippov, I. et al., Optical Structure Recognition Software to Recover Chemical Information: OSRA, An Open Source Solution. J. Chem. Inf. Model., 49 (3), pp. 740-743, 2009

[7] Gaulton, A.et al., "Role of open chemical data in aiding drug discovery and design," Future Med. Chem., vol. 2, no. 7, pp. 903-907, 2010.