Performance of MolRec at TREC 2011 Overview and Analysis of Results

Noureddin M. Sadawi, Alan P. Sexton, and Volker Sorge

School of Computer Science
University of Birmingham
Email: N.M.Sadawi|A.P.Sexton|V.Sorge@cs.bham.ac.uk
URL: www.cs.bham.ac.uk/~nms|~aps|~vxs

Abstract

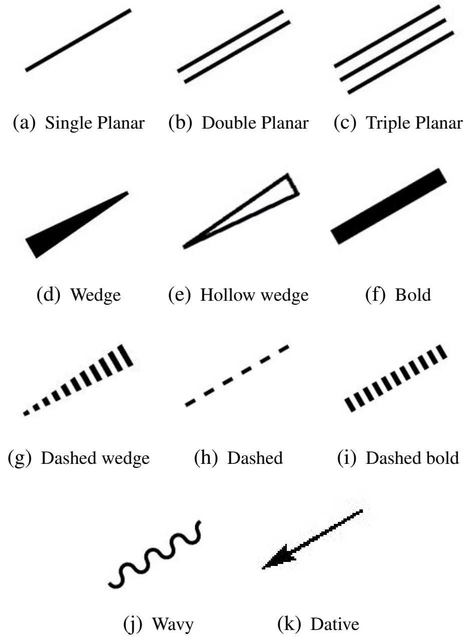
Chemical molecular diagrams are commonly found in documents from the chemical and life science disciplines. We present an overview of the elements of these diagrams and of MolRec, our system for analysing and recognising them. MolRec uses a number of techniques to refine the scanned images and precisely detect line segments and line junctions, structural elements and the atomic formulae that commonly appear in such diagrams. The output of our system is a chemical formula and associated MOL file, a standard representation of molecular structures used in cheminformatics that records precise molecular spatial and connectivity information. When applied to the TREC 2011 test set of 1000 molecular diagrams, MolRec returned in two separate runs 949 and 950 correctly recalled structures, respectively. We discuss these results and present an analysis of MolRec's performance on the test set.

1 Overview of Diagram Elements

Molecular diagrams generally consist of a combination of characters denoting names of atoms (e.g., O) or more complex molecules (e.g., HO) together with graphical elements depicting chemical bonds. The latter can be of a number of different types and their combination determines the overall 3-dimensional structure of the entire molecule.

An overview of the different common types of bonds is given in Figure 1. One or more normal line segments are used to show normal bonds (planar bonds) as in Figures 1(a), 1(b) and 1(c). Parallel line segments close together where each is of approximately the average bond length indicate a double (or triple) bond. Parallel line segments where at least some are shorter than the others indicate a sequence of separate bonds with an omitted carbon atom at the node, called an implicit node, identified by the end of the shorter line segments, as in Figure 2.

To indicate 3-dimensional structure of molecular diagrams, bonds are drawn in different styles to indicate a direction with respect to the drawing surface. These styles determine what chemists term the *stereo-centre* of the bond.


(a) Single Planar (b) Double Planar (c) Triple Planar

(d) Wedge (e) Hollow wedge (f) Bold

(g) Dashed wedge (h) Dashed (i) Dashed bold

(j) Wavy (k) Dative

Figure 1: Common Bond Conventions

The solid wedge, hollow wedge or bold line segment, in Figures 1(d), 1(e) and 1(f) respectively, are used to show bonds coming out of the plane of the drawing surface (towards the viewer). The direction of the solid and hollow wedge bonds is determined by the tip-to-base direction, meaning the stereo-centre is at the narrow end. In the bold bond case both direction and stereo-centre are unspecified and have to be determined somehow, for example by using chemical domain knowledge, for a correct recognition of the diagram.

A dashed wedge, a dashed line segment and a dashed bold line segment, Figures 1(g), 1(h) and 1(i) respectively, are used to depict bonds going behind the plane of the drawing surface (away from the viewer). These also have directions to specify the stereo-centre. For a dashed wedge bond, the stereo-centre is at the shortest dash, so the direction is

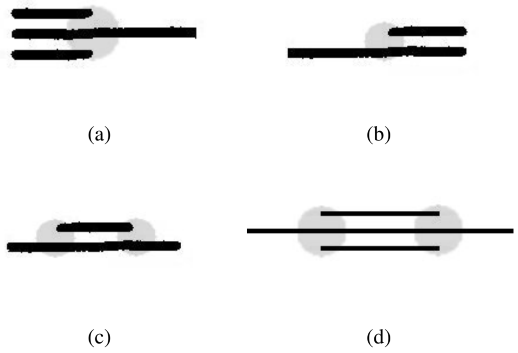


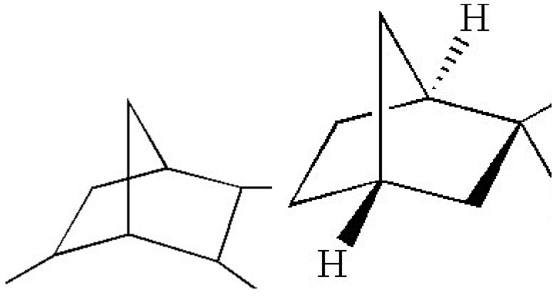
Figure 2: Implicit Nodes (indicated by shaded disks) in Bond Sequences

from the shortest to the longest dash. For a dashed bond and bold bond, the direction and stereo-centre are unspecified and have to be identified.

A wavy bond, as in Figure 1(j), is used to show an unspecified configuration (mixture of up and down).

As Figure 1(k) shows, an arrow is used to illustrate a dative (polar) bond. The direction of the arrow is from source-to-head and it indicates the existence of a negatively charged atom at the head of the arrow.

Further 3-dimensional structure can be depicted with bridge bonds, in case there are multiple different connection paths between different parts of the molecule. These are typically presented in a $2^1/_2$-dimensional perspective drawing form. Such diagrams have one or more foreground bonds drawn crossing one or more background bonds, where foreground and background bonds are not connected where they appear to touch in the diagram. If the background bond is drawn with a gap to make this clear, it is called an open bridge bond, otherwise it is called a closed bridge bond (c.f. Figure 3).



(a) Closed Bridge Bond (b) Open Bridge Bond

Figure 3: Closed and Open Bridge Bonds

Aromatic rings are sometimes drawn with a large circle inside a cycle of bonds, as in Figure 4, instead of separate planar bonds.

Superatoms are names that are embedded in a diagram as if it were an atom (c.f. Figure 5). These names represent whole molecular substructures that can have multiple bonds to the surrounding molecule. Superatom names are typically meaningful to chemists but syntactically in the diagram give little or no clue to their actual content. Hence they can only be supported through some variety of dictionary mechanism.

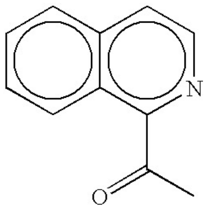


Figure 4: Aromatic Ring

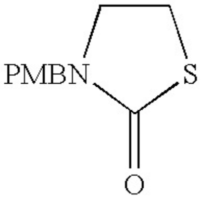


Figure 5: Superatom

2 Implementation

MolRec's recognition procedure consists of a series of steps, of which we present the most important ones in this section.

After initial binarisation of the input image, connected components are labelled and fed into a simple metric space based OCR engine to identify character symbols, which are subsequently combined into character groups. Then we recognise bonds based on a rule set for rewriting basic graphical elements. This forms the basis of a graph structure, which can be translated into the MOL output format, after embedding of superatoms and further resolution of ambiguous stereo bonds.

A full specification of our rule set for bond recognition as well as a more detailed description of the entire recognition procedure can be found in [6].

2.1 Character Grouping

Letters, numerals and some symbols are taken to indicate atoms or superatoms. Any such components identified during the OCR process are grouped to form labels. Grouping is performed horizontally, vertically and diagonally.

Let a and b be elements of $N \cup L \cup S$, where N is the set of digits, L the set of letters and S the set of non-letter, non-digit symbols. If a and b are within a preset distance of each other, a will be grouped horizontally, vertically or diagonally with b as follows:

Horizontal grouping is performed if the spatial relations between a and b is horizontal and one of the following conditions holds:

i) Both a and b are letters

ii) Both a and b are digits

iii) a is a letter and b is a symbol, or vice versa.

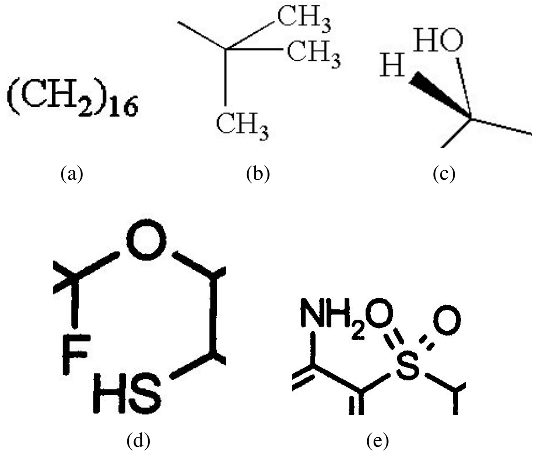
Vertical grouping is performed if the spatial relations between a and b is vertical and:

i) Both a and b are letters

Diagonal grouping is performed if the spatial relations between a and b is diagonal and one of the following conditions holds:

i) a is a letter and b is a digit, or vice versa.

ii) a is a letter and b is a charge sign, or vice versa.



(a) (b) (c)

(d) (e)

Figure 6: Character Groups

Some examples are shown in Fig 6. The top character group of Figure 6(a) shows a letter to the right of a symbol “$(C$”, a letter to the right of another letter “CH”, a digit to the bottom right of a letter “H_2”, a symbol to the top right of a digit “$_2)$”, a digit to the bottom right of a symbol “$)_1$” and a digit to the right of another digit “16”. Also, these rules will not allow grouping of very close characters such as the ones shown in Figure 6(b), 6(c) and 6(d). The example shown in Figure 6(e), which is not entirely clear even to a human reader, is not disambiguated by the previous rules.

Some cases requiring disambiguation include the uppercase letter “O”, lower case “l”, lower case “I”, etc.

2.2 Line Finding and Recognition of Bonds

Any detected character groups are erased from the image and the new image is thinned to unit width thickness. The line segments, which can be free standing or connected as polylines, are traversed and split at junctions where more than 2 line segments join. An average line width is calculated during this stage. This results in a set of polylines that is used as input to the Douglas-Peucker line simplification algorithm [1]. The result of using this algorithm is a set of straight line segments and an average line length.

Parallel line segments indicating double or triple bonds or bond sequences with implicit nodes are all identified by clustering line segments of the same slope that are within a threshold distance of each other. In the bond sequence case, the long line segments are split at the point where the short line segments end, so that a node to hold the implied atom is created.

A sequence of short parallel line segments spaced apart at regular distances are detected by identifying short line segments whose centre points are regularly spaced within a certain tolerance. A sequence of short parallel line segments of monotonically changing length represent a dashed wedge where the stereo-centre is identified as the atom at the shortest line segment of the wedge, while short line segments of similar length represent a dashed bold bond. A more direct approach based on using their slope is not reliable because of the difficulty of accurately finding the slope of such short line segments. Dashed bonds are detected by identifying repeated short line segments of similar length whose centre points are collinear. Again, the stereo-centre in the case of dashed bold bond and dashed bond is unknown. Our method for identifying the stereo-centre is explained in section 3.1.

To extract precise geometric information about bonds in the shape of solid wedges, we use a disc of a radius larger than a measured average line thickness. The measured thickness is obtained dynamically by analysing discovered lines in the image. This disc can fit inside the base of a wedge (triangle) but not in a normal line. We grow the disk until it reaches the largest size possible while still covering only foreground pixels in the original image. Then we walk the position of the disk in any direction that allows it to continue to grow. If this object is indeed a wedge, then when it can grow no more we have found the base of the triangle, thus identifying the stereo-centre for this 3D bond. We can then walk the disk along in the direction of slowest decrease of disk size to find the opposite end of the bond. If it is not a triangle, then the disk size will not change appreciably over the length of the thinned line segment corresponding to this connected component. In this case we recognise the object as a thick solid line segment (c.f. Figure 7), but in this case the stereo-centre in unknown, and we just need to identify the two end points (more on this in section 3.1).

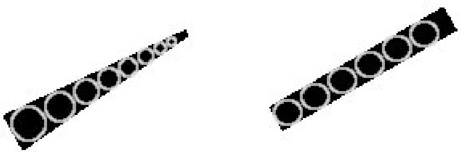


Figure 7: Using Disk to Identify Wedge/Bold Bonds

Wavy bonds, which are drawn as a wave pattern, are reduced by our thinning and line simplification process to a sawtooth pattern polyline of connected short line segments. This is straightforward to identify when following the polyline.

2.3 Graph Construction

At this stage, an initial undirected graph is constructed where each bond is an edge and each junction is a node. This is done by grouping line segment endpoints by distance and by connectivity to the bounding box of character groups in order to construct each node. Each vertex of the graph is labelled with the character group at the corresponding position. Bonds and character groups in the graph are examined for common causes of ambiguity. Disambiguation of lower case “l”, upper case “I”, the digit “1” and a vertical single bond is carried out at this point.

2.4 Superatom Embedding

We mined MOL files in the OSRA dataset [2], and integrated the freely available Marvin abbreviation group collection superatom dictionary [4] to identify complete information about superatoms. Superatoms in MOL file structures identify the superatoms in situ, marking each atom of the whole structure that belongs to the superatom, together with the internal bonds in the superatom and the external bonds between the superatom and the surrounding structure. We then replace the superatom nodes in the graph with an embedding of the structure of the superatom.

This still leaves us with the connection permutation problem, but at least gives us unambiguous internal connection information and identification of the connecting atoms.

2.5 Stereo Bond Resolution

For 3D-bonds with unknown stereochemistry, as in Figure 8, MolRec needs to decide the stereo nature of the bond, i.e. which side of the bond is the stereo-centre, and, in the case of the wavy bond, whether the direction of the bond is towards the background or the foreground of the image.

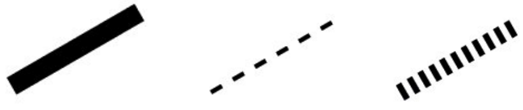

(a) Bold bond (b) Dashed bond (c) Dashed bold bond

Figure 8: Bonds with Unknown Stereochemistry

The syntax of the diagram does not give sufficient information to resolve this correctly so, in the absence of better domain knowledge, we employ a number of heuristics based on observed patterns in diagrams we have analysed. These heuristics use information about the numbers of neighbours on each end of the bond in questions.

Reason	#Images
Incorrect stereochemistry	10
Solid Circles without 3D Hydrogen Bond	5
Image has touching components	6
Image has broken characters	3
Incorrect character grouping	5
Connectivity of superatoms	3
Problematic Bridge Bonds	3
Unhandled bond type	1
Unrecognised syntax	5
Dashed wedge bonds mis-identified	15
Diagram caption confusion	5

Table 1: Reasons for Mis-Recognition of Molecules.

2.6 Output Generation

Finally, the MOL file [8] is generated from the graph. For training and evaluation, we used OpenBabel [5] which provides the ability to compare different MOL files semantically, ignoring unimportant syntactic differences.

3 Analysis of MolRec’s Performance

When run twice on the 1000 images in the TREC11 data set, MolRec achieved a 95% and a 94.9% correct recovery rate, respectively. This corresponds to 50 diagrams mis-recognised in the first run, and 51 in the second. In fact, because most of the diagrams mis-recognised in the first were also mis-recognised in the second, where the internal parameters in MolRec were slightly adjusted, there were a total of 55 different diagrams mis-recognised in one or other of the two runs. Some of these 55 diagrams failed for multiple reasons, so we were able to identify 61 reasons for diagram recognition failures in total.

Table 1 summarises the reasons for failed recognition together with the exact number of mis-recognised images for each reason. In the remainder of this section we discuss each of the cases with some suggestions for future improvements.

3.1 Incorrect Stereochemistry

The stereochemistry of some bonds is not derivable purely from the syntactic properties of the diagram and, in the absence of deeper domain knowledge, our use of heuristics reduces the number of incorrect choices of 3D bond direction but does not eliminate them. When our heuristics guessed the wrong stereochemistry in such cases, a mis-recognition occurs. Adding further domain knowledge should improve recognition rate here.

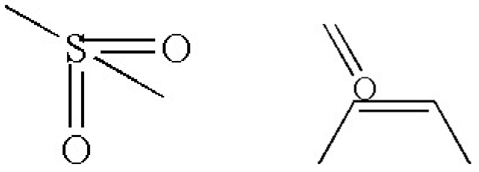
3.2 Solid Circles without 3D Hydrogen Bond

A number of diagrams in the test set use the solid circle notation that indicates a 3D Hydrogen bond, but without that stereo bond information appearing in the solution MOL files. This seems to be a particularity of some of the samples in the test set, an issue that is further discussed in section 4. MolRec's results always generates the stereo bond information in its output MOL files and thus does not match the solution MOL files in these cases.

3.3 Touching Components

MolRec does not currently handle touching components such as in Figure 9. These can include:

- Touching characters, although MolRec handles ligatures such as NH (where the serifs touch).
- Letters touching symbols
- Characters touching bonds, although MolRec handles the common pattern where a diagonal bond is glued to a character from the bottom.
- Some cases of bonds accidentally touching bonds, usually due to ink bleed between close parallel lines making a connection that should not be there.



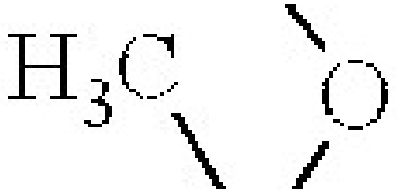
(a) (b)

Figure 9: Example of Touching Components

While a number of solutions have been proposed in the literature to handle touching characters, this problem is notoriously difficult.

3.4 Broken Characters

MolRec does not currently handle broken characters such as in Figure 10. As with touching characters this is a known difficult problem in document analysis.

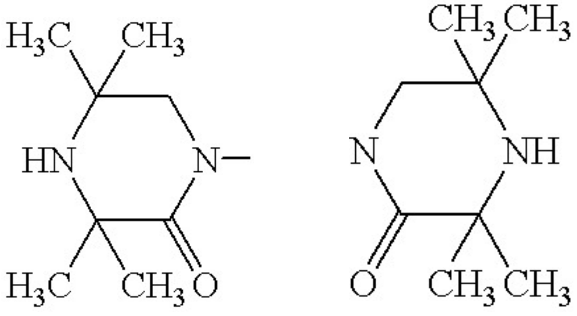


(a) (b)

Figure 10: Example of Broken Characters

3.5 Incorrect Character Grouping

Some characters were incorrectly grouped because they were too close to each other for MolRec to reliably separate, as in Figure 11. This could be improved by adding more knowledge on what molecule groupings are permissible.

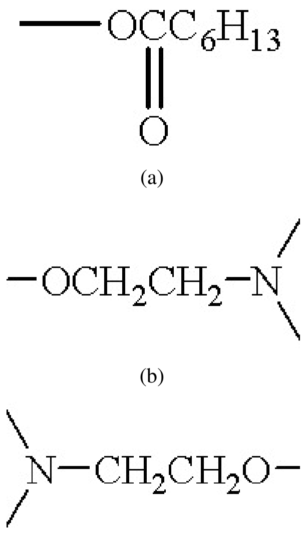


(a) (b)

Figure 11: Incorrect Character Grouping

3.6 Connectivity of Superatoms

If there are two or more bonds between the superatom and the surrounding structure, it is sometimes unclear how to determine the appropriate permutation of connection possibilities to match the actual chemical structures in question.



(a)

(b)

(c)

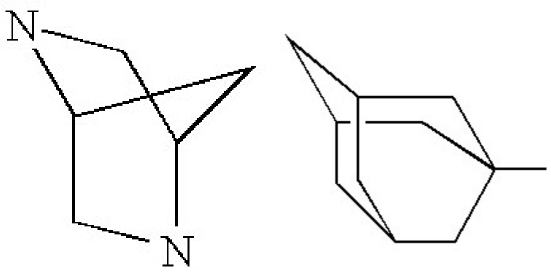
Figure 12: Example of Superatoms

MolRec essentially makes a random guess as to the correct permutation, and therefore gets in wrong in some cases. Again, additional knowledge on the structure of superatoms could reduce this problem.

3.7 Problematic Bridge Bonds

MolRec detects open bridge bonds when a broken straight line with another straight line passing through the gap are found. The broken line is reconnected. Closed bridge bonds are identified by checking the crossing junctions and checking that the lines forming the junction are part of irregularly shaped cycles.

Non-bridge bond cycles are always regularly shaped pentagons, hexagons etc. MolRec detects such irregularities by finding junction angles outside expected ranges and interprets the structure accordingly. However, in some cases in the test set, either the angles involved were outside the thresholds used by MolRec or the perspective was sufficiently extreme that MolRec confused which node should be associated with certain line segment endings. Figure 13 presents two such examples. Here further fine tuning of our recognition approach is needed.



(a) Closed Bridge Bond

(b) Open Bridge Bond

Figure 13: Problematic Closed and Open Bridge Bonds

3.8 Unhandled Bond Type

The dashed dative bond is shown in Figure 14. We have not encountered such a symbol before and are not clear about its intended interpretation. The solution MOL file interprets it as a planar single bond. Since MolRec's bond recognition is rule based, an extension to include currently unhandled bonds should be straightforward.

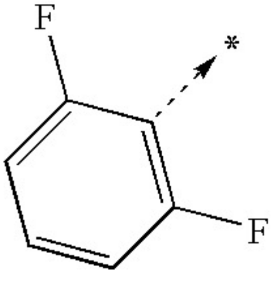
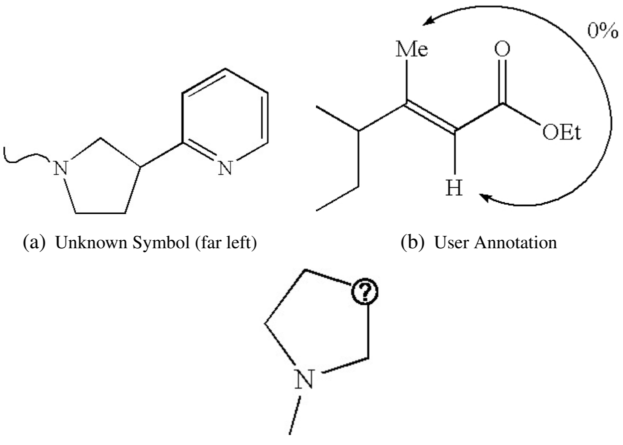


Figure 14: Dashed Dative Bond

3.9 Unrecognised Syntax

We classified a total of five images as possessing a syntax unrecognised by MolRec.

Three images in the dataset included user annotations (c.f. Figure 15). The corresponding solution MOL files do not appear to treat them as part of the structure. In the case of Figure 15(c), the solution MOL file uses a Carbon atom in place of the question mark symbol, i.e. it ignores the symbol.



(a) Unknown Symbol (far left)

(b) User Annotation

(c)

Figure 15: Unrecognised User Annotations

One image, displayed in Figure 16, shows a dashed wedge bond with a wavy line crossing it. We are not familiar with this notation and do not know how to interpret it. An analysis of the corresponding solution MOL file showed the same contents as if the wavy line were not present at all. MolRec interpreted the wavy line as a wavy bond (c.f. Figure 1(j)) and recognised the crossing of the wavy bond with the dashed wedge as 4 bonds connected at the centre: two dashed wedges and two wavy bonds.

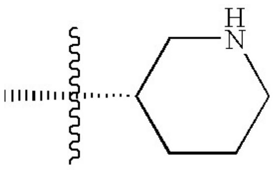


Figure 16: Unrecognised Wavy Line Syntax

Finally, MolRec does not currently recognise structures with frequency variations such as the one in Figure 17, which appears in the test set.

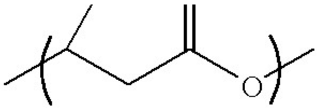


Figure 17: Repetition Structures

While it is unlikely that one can devise a recognition procedure that can take care of arbitrary user annotations, we are hopeful that at least regular Markush structures [3], such as the above frequency variations, could be handled within our rule based approach.

3.10 Dashed Wedge Bonds Mis-Identified

These include incorrect identification of some dashed wedge bonds and of some bridge bonds. For the dashed wedge bonds in question, the short dashes at the narrow end of the bond were considered by MolRec as part of a dashed bond, while the longer dashes were treated as part of a dashed wedge, or dashed bold, bond. This has led to interpreting some dashed wedge bonds as two connected bonds (a dashed bond and a dashed wedge, or dashed bold, bond), which meant an extra non-existent node and bond were added. Further honing MolRec's recognition parameters should take care of this problem in the future.

3.11 Diagram Caption Confusion

On 5 images in the test set a diagram caption appears in the image (c.f. Figure 18). As MolRec is aimed to recognise molecule structures only, it does not do any image segmentation. Consequently it fails to recognise the image because it cannot find a suitable interpretation for the caption as part of the molecule structure.

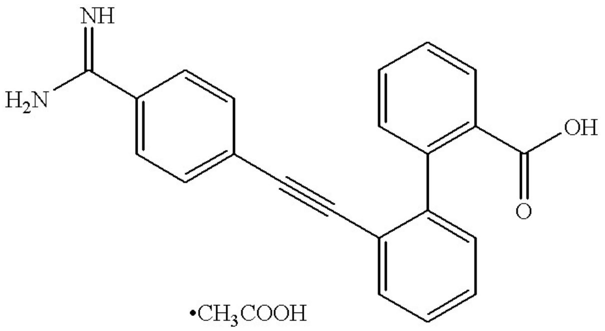


Figure 18: Diagram with Caption

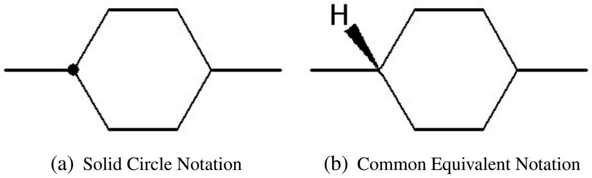
4 Issues with the Test Set

We have not found a definitive graphical syntax specification for molecular diagrams, and it is clear that there exist diagrams which use some graphical elements in different and inconsistent ways from each other. Also there appear to be syntactic notations in these diagrams that do not give sufficient information on their own to uniquely determine the corresponding MOL file up to isomorphism. Finally there are many diagrams to be found in the literature which contain definite errors.

In such cases, it is a difficult choice for a molecular diagram recognition system as to what it should, or could, do. Here we consider the few cases of this nature in the TREC11 test set.

Solid Circles: The test set contains several structures with a solid circle covering a junction. We understand these solid circles to indicate the existence of a hydrogen atom connected to that node via a solid wedge bond (c.f. Figure 19), an interpretation borne out by the corresponding MOL files in the solution set for a number of diagrams with such solid circles in the test set. However, some images in the test set have the solid circle but the provided solution MOL file indicates that the node in question does not have the corresponding solid wedge bond to a hydrogen atom, e.g. US06372153-20020416-C02522.

MolRec recognises and interprets these solid circles in the way we believe correct. However, for the problem cases described above, MolRec obviously produces MOL files that do not match the solution MOL file.



(a) Solid Circle Notation (b) Common Equivalent Notation

Figure 19: These two notations are equivalent

Dative Bonds: The test set contains several structures with dative (polar) bonds (a bond in the form of an arrow as in Figure 20). We understand dative bonds to indicate the existence of a negatively charged atom at the head of the arrow [7]. However, the training set sometimes interprets such arrows as double bonds, e.g. image US20020143030A1-20021003-C00004, and sometimes as a normal single bond, e.g. image US20020143030A1-20021003-C00004. While in the test set's solution MOL files, all dative bonds seem to have been interpreted as normal planar bonds (i.e. as normal single bonds).

MolRec does not yet recognise arrows. Currently it simplifies them into simple line segments, which are then interpreted as normal single bonds. Serendipitously, this treatment agrees with the solution MOL files in TREC11, although one should really consider both MolRec and the solution set to be equally in error in such cases.

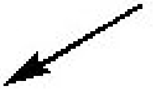


Figure 20: Dative (Polar) Bond

Over Connected Atoms: The test set contains one image, US06334922-20020101-C00005, where 3 carbon atoms had 5 bonds each (the circled atoms in Figure 21). We do not have sufficient chemical domain knowledge to be sure our interpretation is correct but, as we understand it, a carbon atom normally has 4 bonds (or less with omitted hydrogen atoms). When a carbon atom has more than 4 bonds, this means it should be positively charged and a plus sign (+) is used to indicate this charge. However, there were no positive charge signs in the diagram to indicate this, meaning that the diagram is internally inconsistent. The solution MOL file does not indicate the extra positive charge as it should.

MolRec is not currently designed with sufficient domain knowledge to detect this inconsistency, and the MOL file it generates corresponds to the incorrect solution MOL file.

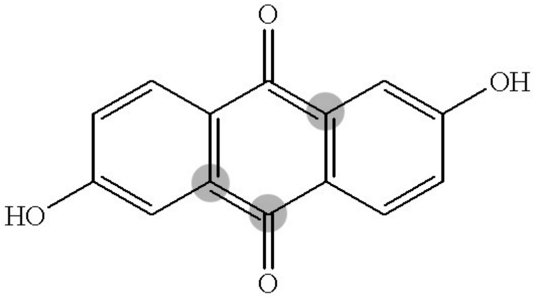


Figure 21: Circled carbon atoms have 5 bonds and no charge

5 Conclusions

Although a MolRec's 95% recognition rate in TREC 2011 is already high, there is still plenty of room for improvement.

Some of the mis-recognition problems we faced are inherently uncorrectable, in the sense that, just like in the real world, some of the test cases either have errors or have incorrect solution MOL files. Such problems must simply be accepted. We believe many of the mis-recognition problems can be solved with some relatively simple enhancements of our system, e.g. the 15 dashed wedge bond misidentifications or the 5 diagram caption confusion cases. For a significant number of the problems we need to incorporate more chemical domain knowledge into our system, e.g. for the 10 incorrect stereochemistry problems or the 3 superatom connectivity problems.

Overall, we are pleased that a number of approaches turned out to be very successful and we recommend them to any who work in this field:

Line finding and simplification: Early experiments using Hough transforms for line finding yielded disappointing results with poor robustness. Instead we start with connected component analysis, filter characters using OCR on each connected component, skeletonise the remaining components, and use the Douglas-Peucker algorithm to simplify the skeletons and remove skeletonisation artifacts, in order to produce clean paths along lines. We can then walk the lines in the original image using the cleaned skeletons to detect and analyse the various types of bonds. This approach has proven to be fast and particularly successful.

Solid triangle and bold line detection: Given our cleaned skeleton paths, we identify and orient solid triangles, and simultaneously detect bold lines, by finding components within which a disc of radius larger than the line width can fit, and then walking the disk along the direction of the component that allows maximal growth in the disk (or minimal shrinkage). This has also proven to be very fast and robust.

Mining superatoms in MOL files: Unavailability of comprehensive superatom dictionaries, and the lack of the level of detail of internal information about the superatoms that are necessary for use in MOL files, led us to mine collections of MOL files for their superatom content. This was a fairly simple process of extracting the superatom definitions from the MOL files and relabelling their contents so that they could be reused in other MOL files. This has significantly increased the number of diagrams we can recognise automatically, although we still face the open problem of connectivity permutations in superatoms which have more than one external connections from different internal atoms.

Breaking joints: Many joints in a molecular diagram touch end-to-end to indicate the presence of an unmentioned Carbon atom. However, often Carbon atoms are explicitly written in to a space separating the bond lines. Rather than have to deal with the combinatorial possibilities of the various ways that bonds might connect, we chose to explicitly break all such connected joints so that we could treat them all in a uniform way. This has significantly simplified our code and the logic for dealing with connections and has yielded unexpected dividends in, to name one example, dealing with implicit nodes such as in Figure 2.

References

[1] D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122, 1973.

[2] I. Filippov and M. Nicklaus. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.*, 49(3):740–743, 2009.

[3] J. Gasteiger. *Handbook of Chemoinformatics: From Data to Knowledge*. Advances in Electrochemical Sciences and Engineering Series. Wiley-VCH, 2003.

[4] The Marvin abbreviation group collection. `http://atchimiebiologie.free.fr/marvin/chemaxon/marvin/templates/default.abbrevgroup`.

[5] Open Babel: The open source chemistry toolbox. `http://www.openbabel.org/`.

[6] N. Sadawi, A. Sexton, and V. Sorge. Chemical structure recognition: A rule based approach. In *19th Document Recognition and Retrieval Conference (DRR 2012)*. SPIE, January 2012. to appear.

[7] Cambridge Soft. Chembiodraw v12.0 user documentation, 2010. `http://www.cambridgesoft.com/software/ChemDraw/`.

[8] Symyx. CTfile formats, 2010. `http://www.symyx.com/downloads/public/ctfile/ctfile.jsp`.