

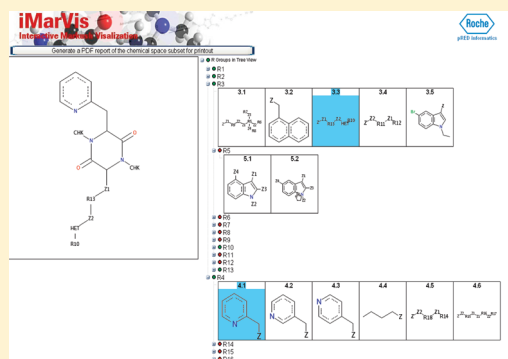
Intuitive Patent Markush Structure Visualization Tool for Medicinal Chemists

Wei Deng, Steven J. Berthel, and W. Venus So*

Roche, 340 Kingsland Street, Nutley, New Jersey 07110, United States

Supporting Information

ABSTRACT: A Markush, or generic structure, is a widely used convention in chemical and pharmaceutical patents. The flexibility and complexity of this format, however, preclude an easy understanding and analysis of chemical space. In this paper, an application package called MarVis (Markush Visualization) is introduced to help chemists visualize Markush structures in chemical patents. MarVis can output a report with the Markush structure showing the query substructure and also an R-group table of all the possible R-groups described in the patent. MarVis also has a unique interactive interface that allows chemists to explore and zoom in the chemical space to find a subset of interest. SMILES, with minimal extensions, was used to facilitate a variety of patent Markush structure studies.



INTRODUCTION

Named after Eugene Markush, a Markush structure is a representation of a compound class in which “functionally equivalent” chemical entities are allowed in one or more parts of the compound class. Currently Markush structures are widely used in chemical and pharmaceutical patents.

In analyzing patent Markush structures, an initial obstacle is that these patent Markush structures are typically archived as digital images and not as machine-readable structures. Two major commercial topological search systems for Markush structures in patents are currently available: MARPAT^{1,2} and MDARC (Markush Documentation and Automated Research of Correlations).³ They contain indexed data that have been manually translated from Markush structures in digital images to data formats that are searchable on each proprietary system. Both systems are widely used in the pharma industry, and both provide substructure search features to identify matching structures embedded in the chemical space described in patents.^{4–6} The MARPAT system was developed based on published work^{7–24} from Lynch’s group. The MDARC system was created based on the study from Dubois’ group,^{25–35} and it was improved by Questel³⁶ to work in a service environment. Later, in the early 1990s, the MDARC system was further improved with the involvement of Lynch’s group.

Reviewing the results from these substructure searches can be time-consuming, partly because the substructure query may be embedded in R groups of the Markush structure, making it difficult to understand how the query substructure matches the Markush structure of the patent hit.³⁷ This point is illustrated in a simple example in Figure 1. The query substructure in Figure 1A is actually embedded in R1 of the Markush structure of one of the patent results (Figure 1B, details in Figure 6). In this simple case,

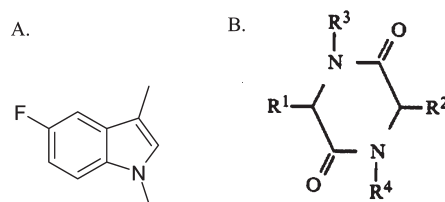


Figure 1. A) An example query substructure used in MMS search. B) Markush core of a hit patent (US4806538). The query substructure and the Markush core have little similarity. A closer investigation revealed that the query substructure matches R1 (details in Figure 6).

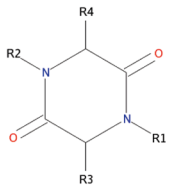
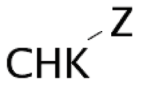
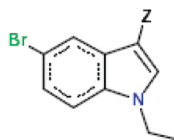
one can still relatively quickly figure out where the query structure is in the patent Markush. However, in most cases, for example when the query substructure matches a nested R group or only matches the enumerated structures (that is, not entirely within an R group or the Markush core), then it can be very time-consuming to understand how the query substructure matches the patent Markush. Therefore, an automated system that summarizes the relevant information from the search and provides a visual representation of how the substructure query matches the compounds described in the patent can be very useful.

An application package, called MarVis (Markush Visualization), has been developed to help analyze patent substructure search results. MarVis can show a graphical table of all possible R-groups described in a patent (the MarVis outcome for the patent in Figure 1B is shown in Figure 6). Following a substructure search, MarVis will expand the Markush structure displaying the query substructure. This feature allows chemists

Received: July 8, 2010

Published: March 07, 2011

Table 1. Example of Extended SMILES Representation of the Markush Structure in Figure 1^a

Structure	SMILES Representation of the structure
Markush core 	<chem>[N]%10%11%12.[C]%10%13=%14.[C]%11%15%16.[R1]%12.[C]%13%17%18.[O]=%14.[C]%15%19=%20.[R3]%16.[N]%17%19%21.[R4]%18.[O]=%20.[R2]%21</chem>
R1 group 	<chem>[Ce]([Z1])</chem>
R3 group** 	<chem>[Z11001]%29%28.[Z11002]%30([Z3]).[R12]%28.[R11]%29%30.[C]=%29%30([Z3]).[C]=%29%31.[C]%30:%32:%33.[N]%31%34%35.[C]:%32%34:%36.[C]:%33:%37.CC%35.[C]:%36:%38.[C]:%37:%39%40.[C]:%38:%39.[Br]%40</chem>

^a A single asterisk (*) denotes the following: CHK = alkyl or alkylene. CHK is represented by rare earth atoms in SMILES (refer to **Superatoms** in the Supporting Information for details). Two asterisks (**) denote the following: Z atoms are attachment points. Z1 and Z2 are attachment points for R11.

to quickly realize how the query substructure matches the Markush structure of a patent hit and hence facilitates further patent analysis. This initial step of understanding how the query substructure matches the patent Markush can take days and weeks as there are usually several dozen or more patent hits. Using MarVis, this step can be completed in minutes. In addition to the R-group table, in the interactive interface of MarVis (iMarVis), the user can recursively select specific R-groups and expand the Markush core to zoom in the patent chemical space in order to obtain a desired subset.

One distinguishing feature of MarVis is its ability to convert the indexed MDARC data into extended SMILES (Simplified Molecular Input Line Entry System).^{38–40} The flexibility of SMILES considerably broadens the scope for patent Markush structure studies. However, because Markush structures can be very complex, novel methods are needed to handle the patent Markush complexity. In particular, two features contribute to this complexity: nested R-groups and multiple connections of an R-group. The methods presented in this study use SMILES to encode the Markush complexity, and if necessary, adding only minimal extensions. Schemes, such as breaking R-groups into multiple fragments, inserting dummy atoms with labels, and temporarily including atomic and bond information in SMILES, were introduced. With these extensions, most of the challenges of handling patent Markush structures were overcome. The resultant extended SMILES can be displayed by general molecular toolkits like the one in Pipeline Pilot.⁴¹ They can also be converted to other format, such as Molfile⁴² for even broader commercial molecular editor acceptance. These conventions

have been developed into a full Markush structure representation system as described in this paper. The underlying novelties of this new system are summarized in Results and Discussion.

METHODS

1. Conversion from MDARC XML to SMILES. Indexed Markush structures are often stored in a proprietary format, such as Questel/Thomson Reuters' binary VMN format^{36,43} used in MDARC. These formats typically require proprietary viewers and cannot be easily converted to other formats. Currently, only a few software providers, such as ChemAxon,⁴⁴ Digital Chemistry,⁴⁵ and Questel,³⁶ offer MDARC patent Markush structure visualization.

For Markush structure representation, a more commonly adopted format is preferred and being sought after.⁴⁶ Several candidates are under consideration, including Chemical Markup Language (CML),⁴⁷ SYBYL Line Notation,^{48,49} and InChI.⁵⁰ For the purpose of this study, SMILES was considered as a promising candidate. Using short ASCII strings for chemical structures,^{38–40} SMILES is a good choice for Markush structure manipulation. It can be easily converted to 2D or 3D structures by molecular toolkits. It is concise and relatively human readable. It is also easy to manipulate by regular expression and string concatenation and therefore ideal for Markush structure enumeration. To preserve these advantages, the method described in this paper tries to add only minimal extensions to SMILES syntax (example in Table 1). These extended SMILES can be recognized and displayed by commercial molecular toolkits such as the one in Pipeline Pilot.⁴¹ In addition, they can also be converted to Molfile format to be

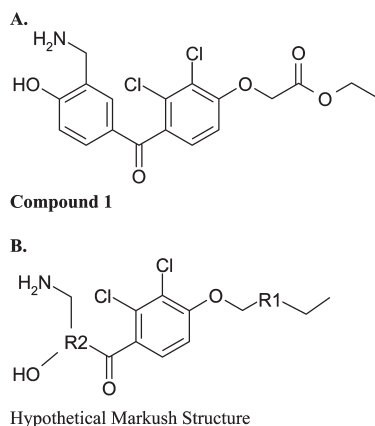


Figure 2. An example **compound 1** (A) and its Markush structure (B) for illustration.

recognized by other molecular viewers like MarvinView.⁵¹ The conversion can be done with Pipeline Pilot. An example PERL conversion script is also included. For details, please refer to **Convert MDARC data to SMILES and Convert extended SMILES to Molfile format** in the Supporting Information.

The MDARC patent Markush structural data are originally indexed by Thomson Reuters Inc. in binary VMN files. Questel Inc. converts them into text XML format. The XML files were purchased from Questel Inc. and used in this study. The contract between Questel and Roche granted Roche access to all patent data in MMS system on a pay-per-use basis. The algorithm of MarVis automatically converts MDARC data in XML format into extended SMILES.

The examples in Figure 2, an actual compound⁵² and its Markush structure, are used in Sections 2–4. Section 2 is focused on the R1 group; Sections 3 and 4 are focused on the R2 group.

2. Ensuring the Correct Orientation in R-Groups with Multiple Connections. A typical patent Markush structure can be very complex.⁵³ A significant challenge is that patent R-groups usually have multiple connections. For the Markush structure in Figure 3A to represent **Compound 1**, R1 should be an ester group. However, R1 can be attached to the Markush core in two possible ways, and only one of them is consistent with **Compound 1**.

In order to solve the issue with multiply connected R-groups, a technique referred to as “breaking an R-group” was developed. As demonstrated in Figure 3B, the R1 group on the core was split into two fragments, each with a new R-group dummy label (1001 and 1002). The two attachment points (Z atoms in Figure 3B) were assigned new labels that were consistent with the new R-group labels. The orientation information is encoded in the MDARC XML files, and the method of using this information is described in the next section. Please also refer to **Markush Representation in SMILES** in the Supporting Information. Notice that no two Z atom labels in one fragment should be identical.

3. Maintaining the Parent-Child Relationship between R-Groups. Another feature contributing to the complexity of the Markush structures found in patents is the nested architecture of R-groups. A patent Markush structure usually has numerous R-groups, many of which are nested. For example, in the MDARC system, a Markush structure can have as many as 50 R-groups and four levels of nesting. In this report, the term “generations” is used for the nesting levels, and the terms “parent”

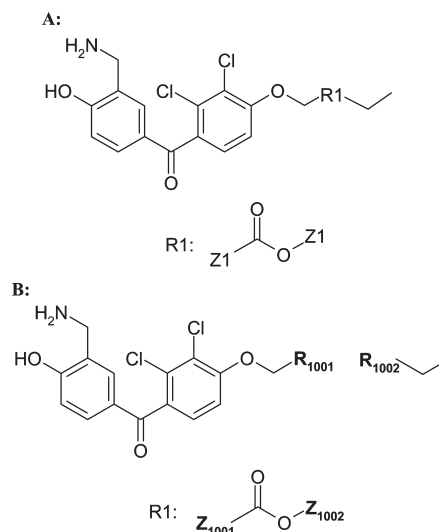


Figure 3. An example to illustrate R-groups with multiple connections. **A)** A hypothetical Markush structure (R2 in Figure 2B is substituted with benzene here) to represent Compound 1 in Figure 2A. There are two ways to attach the ester group (R1) to the core, only one of which is consistent with Compound 1. **B)** Breaking R1 group to R1001 and R1002 and changing the Z labels can ensure correct attachment of R1 group.

and “child” are used for the nested R-groups. Each R-group is connected to its parent R-group (or the Markush core) through attachment points.

When an R-group has multiple connections, the connections may be symmetric or asymmetric. All attachment points are equivalent in an R-group of symmetric connections. However, in the case when the parent and the child R-groups are both asymmetric, matching the R-group’s attachment points and its bonds in the parent R group is critical to ensure correct connectivity (conceptual illustration in Figure 4A). This labeling of an R-group and its parent group in order to correctly connect the attachment points is called the “parent-child labeling”. The labels are available as annotations in the MDARC data. However, including this information in the structural syntax (SMILES in this case) can be challenging.

Figure 4B illustrates how the “breaking R-groups” approach was used to include the correct alignment labeling information from the MDARC data and adapt this parent-child labeling in SMILES syntax. R2 in the core is broken into three fragments. New matching dummy labels (2001, 2002 and 2003) are assigned to the new R atoms and Z atoms. This way, when mapping R2 to the core, only one final structure is obtained having the alignment specified in the MDARC data. In order to incorporate the matching labeling into SMILES, additional information is temporarily inserted into the SMILES strings. This additional information is removed after the correct attachment points are determined. Please refer to **Extension to SMILES to accommodate Parent-Child Labeling for maintaining their relationship** in the Supporting Information.

An extra layer of complexity exists in certain cases where, even with the correct alignment of the parent-child relationships, the child R-group’s attachment points cannot be unambiguously assigned to its grandparent. This observation, referred to as the “grandparent-child alignment (GCA)”, originates from the conversion of VMN to Questel XML files and will be further

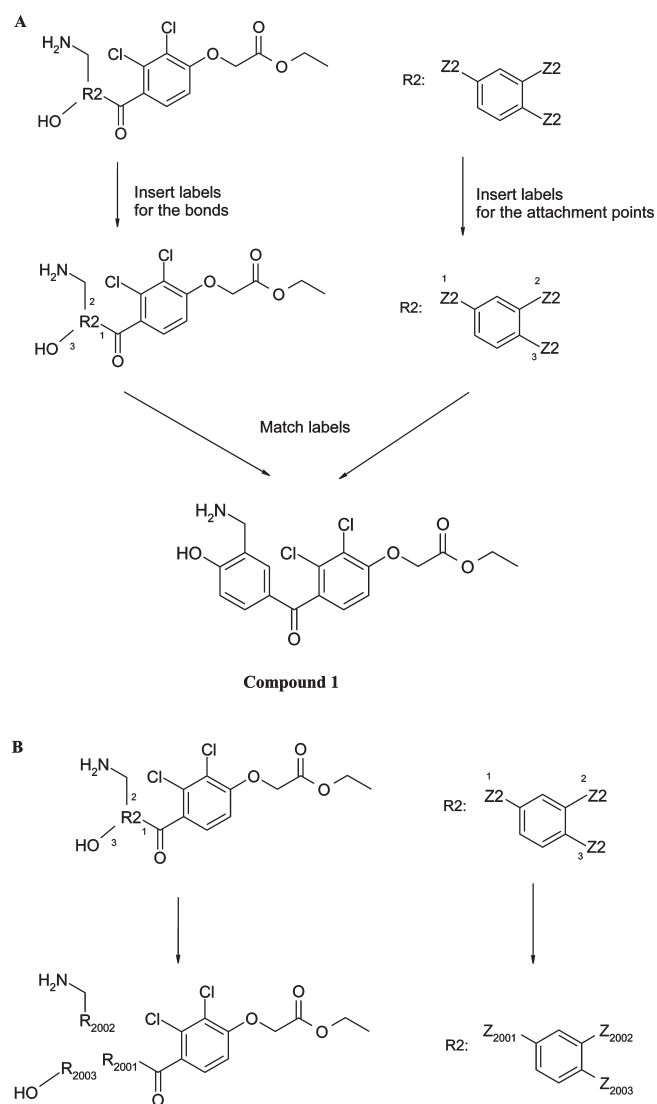


Figure 4. An example to illustrate how to ensure correct connectivity in asymmetric multiple connections situations. The same structure as in Figure 2B is used, except R1 is an ester. A) Both R2 and the three connections on the phenyl ring are asymmetric. Matching labels (e.g., 1, 2, 3) can be used on the R-group attachment points and the bonds of the R group in the core to ensure correct connection. B) In practice, the “Breaking R-groups” approach was applied to achieve the correct connection. R2 is broken into three fragments. New matching labels (2001, 2002, and 2003) are assigned to the new R atoms and the corresponding Z atoms.

illustrated in **Attachment Point Alignment between Grand-parent-child R-groups** in the Supporting Information.

4. Connecting Broken R-Groups for Better Visualization. Although the “breaking R-groups” strategy is an essential component of this method, the broken R-groups formed can be unwieldy to visualize. The ability to clearly depict how R-group attachment points connect to their parents in a simplified manner is indispensable. To address this issue, prior to converting to structures, the broken R-groups are patched together using SMILES. Labels, such as new Z atoms, are inserted into the parent R-group bonds to indicate the corresponding attachment points (Figure 5). Please refer to **Connecting broken R-groups for better visualization** in the Supporting Information.

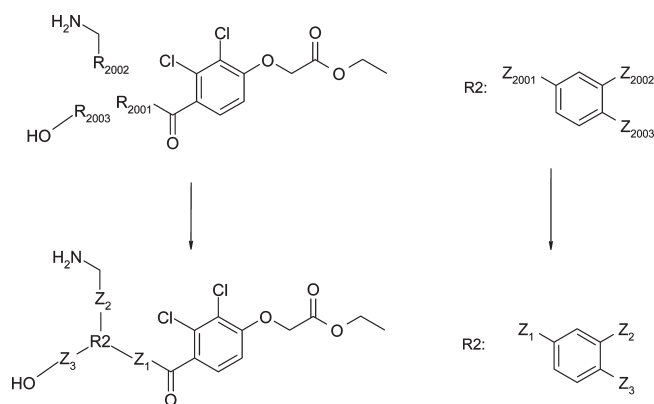


Figure 5. Patch broken R-groups for visualization. Z atom labels are simplified (right panel) and Z atoms are inserted in the corresponding bonds in the parent to indicate the correct way of attachment (bottom left).

5. Representing Superatoms by Rare-Earth Atoms. In indexed patent data, generic groups are represented as short text strings, called “superatoms”. MARPAT and MDARC both use their own superatom annotations. For example, in the MDARC system, CHK is used for alkyl or alkylene groups, and ARY is used for aryl groups.⁵⁴ While in the MARPAT system, “Cb” stands for a carbocyclic group and “Hy” is a heterocycle.⁵⁵ Typically, these superatoms can be difficult to be incorporated into the SMILES syntax.

While superatoms can simplify conversion and search of patent Markush structures, they are difficult to map to specific groups. Matching a Markush structure with superatoms to a chemical structure and vice versa is one of the most challenging problems in patent structure search. The minimum criterion for a new Markush format therefore is to accommodate superatoms in its syntax.

Barnard et al. reported using rare-earth metal atoms to represent superatoms.⁵³ There are several advantages of using rare-earth metals. First, these metal atoms are actual atoms and can be incorporated into the SMILES syntax. In addition, they rarely appear in chemical patents and are therefore unlikely to interfere with existing atoms. Finally, they have high valences and therefore can have multiple bonds. For display purpose, these metal atoms can also be changed back to superatoms using reaction SMIRKS. For these reasons, this approach was adopted in this study. Please refer to **Superatoms** in the Supporting Information.

RESULTS AND DISCUSSION

1. MarVis Report and Web Interface. The goal of this work was to create an automated solution that provides chemists with a much easier and graphical way to review substructure search results and hence allowing them to focus on further patent analysis. A graphical report that contains an R-group table of all the R-group substituents represented by the Markush structures can be automatically generated by MarVis, using the XML file of either the full patent data obtained from Questel Inc.³⁶ or search results from a Questel MMS (Merged Markush Service) substructure search. In the case where Questel MMS substructure search results are available, the R-group matching the query is specified as annotations in the XML file. MarVis adds all matched R-groups to the core, and the expanded core is displayed in the

Patent Number is US4806538

Markush ID is 8743-08701

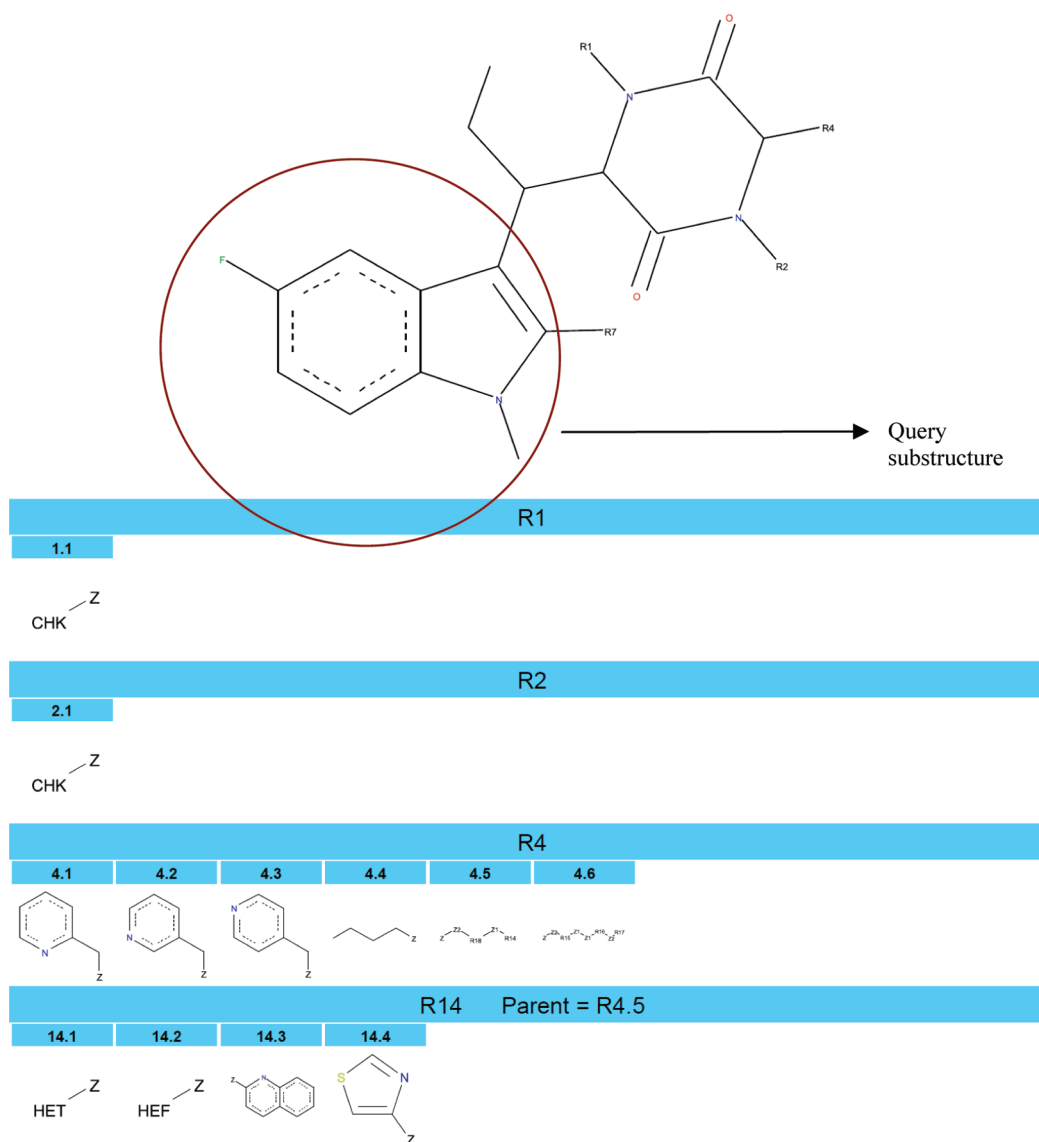


Figure 6. Snapshot of an example MarVis report. The patent search was done using the substructure query in Figure 1A. The search hit's patent number and Markush ID (MDARC internal index) are shown on the top. The Markush structure is expanded with the query substructure, which is circled manually for clarity here. Z atoms are used to indicate the attachment points of parent-child R-groups. It contains the substituents of the parent R-groups (R1, R2, and R4) and substituents for one of child R-groups of R4.5 (R14). Please note that the indexed R group number may be different from that in the patent document.

report, illustrating the proximity of the query substructure relative to the patent Markush structure. This visualization feature can significantly expedite the patent analysis process.

As an illustration to generate a MarVis report, a patent search using the query substructure in Figure 1A was performed, and the search result in XML format was downloaded. The matched R-groups are annotated in the XML file. MarVis converted the XML into extended SMILES format, added the matched R-groups to the core (including merging the R-group substituent to the core according to the parent-child relationships, and changing the final structure for better visualization as described in Methods), and generated a graphical report. The full report on one patent hit is available in the Supporting Information. A snapshot of the report is in Figure 6. The expansion of the

Markush core to show the query substructure is similar to the Markush reduction technique of ChemAxon's JChem.⁵⁶ The R-group table is organized by generations, starting from the top level. Within each parent R group (e.g., R4 in Figure 6), child R groups are listed (e.g., R14). The substituents of an R group are numbered (e.g., R4.1, R4.2), since for nested R groups, specific numbering, like R4.2, allows easy navigation between generations. Each R group has a header indicating the R group number, and the relationship with its parent if applicable (e.g., "R4", "R14 Parent=R4.5").

The MarVis application package was developed on the Pipeline Pilot platform⁴¹ which allows integration with commercial as well as in-house applications. Pipeline Pilot can handle streamline data management, regular expression and graphic user

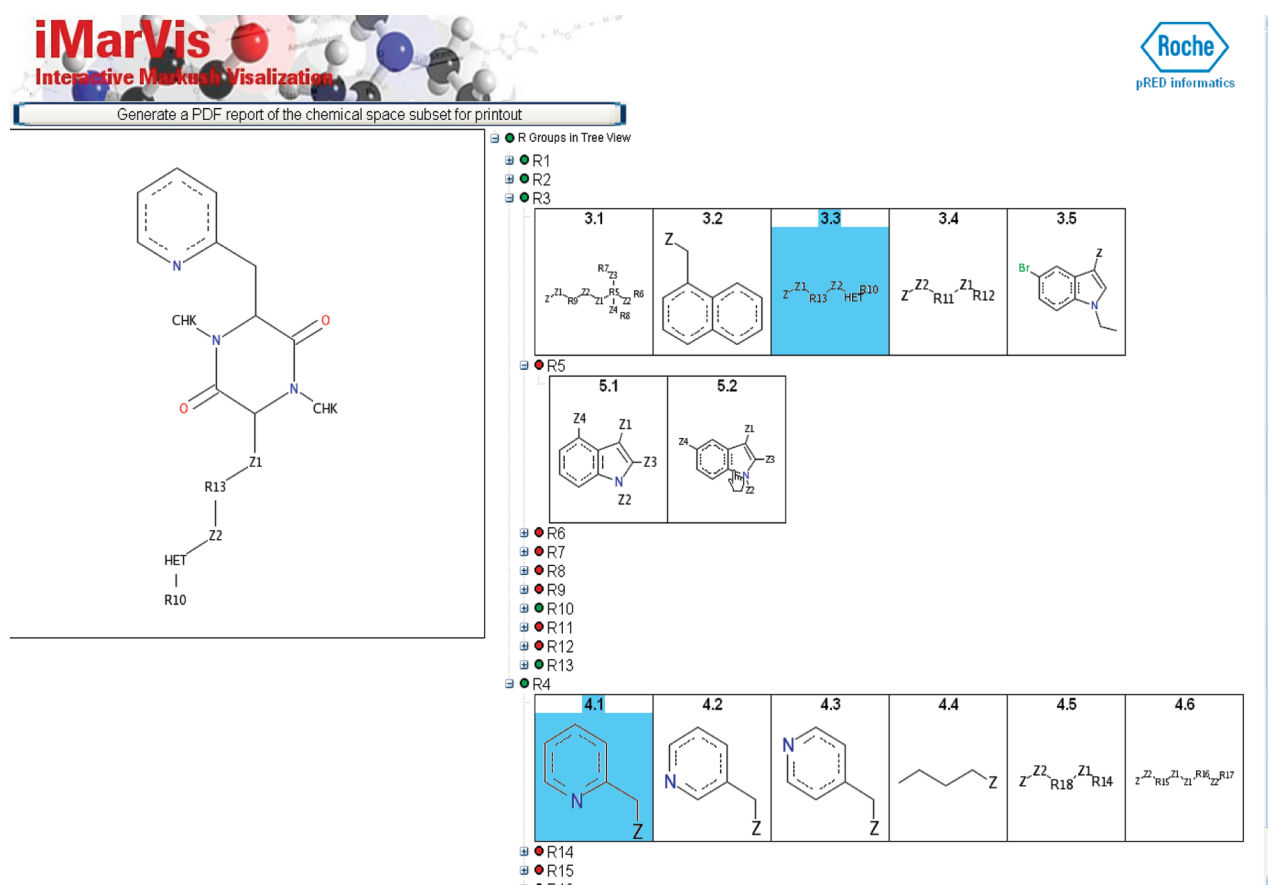


Figure 7. Snapshot of the iMarVis interface. R-groups are listed in tree-view. R-groups can be selected and applied to the Markush core, with the remaining R-groups listed. The colored marks before R group numbers indicate whether the R group is selectable (green = selectable; red = not selectable). A child R group cannot be selected if its parent is not selected.

interface. The MDARC data are first converted to extended SMILES, thereafter structure manipulation and display are based on the SMILES. The typical processing time for generating a graphical report was less than 10 s per MDARC Markush structure. An example of 19 Markush structures was completed in less than two minutes.

An interactive Web interface of MarVis, called iMarVis, has also been developed. Users can select a patent number, view Markush structures, and select a specific fragment for each R-group via an intuitive graphical interface (Figure 7). The layout is similar to a MarVis report with the addition of an R group hierarchy tree. The structure with the selected R-groups is displayed, and the remaining R-groups can be reviewed. The user can then make selections from the remaining R groups, add to the core, and review the outstanding R groups. This can be seen as “zooming in” to the chemical space of a patent to find a suitable subset. The iMarVis selection process is recursive until a subset of interest or an exact structure in the chemical space is identified. As an advanced feature, a child R group cannot be selected (shown as red node) if its parent R group is not selected, and deselect a parent R group will deselect all its children R groups. The user also has the option to print out a MarVis report of the final Markush core and remaining R groups.

2. Enumerating Patent Markush Structures. Based on the algorithm described above, additional Pipeline Pilot components were developed for the analysis of Markush structures in patents. These include top-down and bottom-up enumeration components.

The enumeration algorithm is described in **Markush Structure Enumeration** in the Supporting Information. The components, unlike the original Pipeline Pilot enumeration component, do not have an upper limit of R-group number. They also work on R-groups with multiple connections and produce correctly connected products based on the annotations in XML data. The components take input Markush structures in SMILES format, and manual labeling of the core and R-groups is not necessary.

Significant challenges for enumerating Markush structures are the variety of R-groups and the enormous chemical space claimed. Two R-groups together can be either independent or nested, and different strategies need to be applied. Figure 8 shows an example Markush structure for illustrating the strategies below. In this example, R1, R2, R3, and R4 are on the same level and independent of each other; R5 and R6 are under R1, and R7 is under R2. If the Markush core is considered the “top level”, then the R-groups in the core are on level 1, the R-groups underneath level 1 are on level 2, etc. (Figure 8A). The MDARC indexed patent Markush structures can contain up to 50 R-groups and 4 levels.

Enumerating a nested Markush structure can be achieved by either a “top-down” or “bottom-up” strategy. The top-down method starts from the core and enumerates all R-groups in level 1. The R-groups in successive levels are then enumerated sequentially. The process is complete when all R-groups have been enumerated. Generation-specific partial enumeration is also

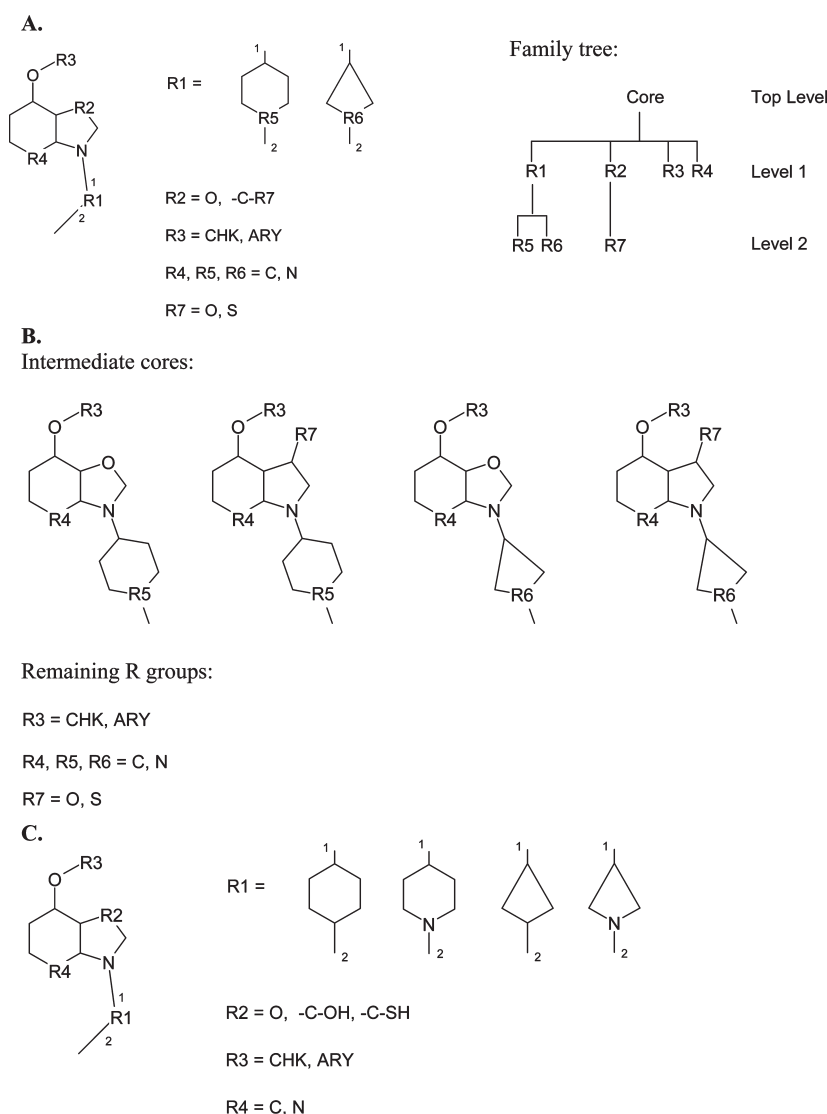


Figure 8. Top-down and bottom-up enumeration. (A) An example Markush structure illustrating the hierarchy. The “family tree” shows the independent R-groups (R1, R2, R3, and R4) and nested R-groups (R5 and R6 under R1 and R7 under R2). (B) Partial top-down enumeration. Intermediate cores after enumerating R1 and R2 in level 1 are shown. (C) Partial bottom-up enumeration. The R groups below level 1 (R5 and R6 under R1 and R7 under R2) are enumerated.

implemented in the Pipeline Pilot component. For instance, a user can enumerate R groups of level 1 and review the intermediate structures (Figure 8B). The advantage of this method is that a family tree is not required prior to enumeration. The disadvantage is that, instead of one Markush core, this process creates many “intermediate cores” which can be difficult to track.

Alternatively, the bottom-up method starts from the lowest level and enumerates the R-groups from the bottom to the top level. Partial enumeration option is also implemented in this component. Given a generation number (e.g., level 1), the component will only enumerate R-groups underneath that generation (see enumerated R1 and R2 substituents in Figure 8C). The enumerated R-groups can then be subjected to sequential analysis. It is useful because when R groups are heavily nested, it would be difficult to recognize the actual functional groups right away. With bottom-up enumeration, all remaining R groups can be enumerated to structures for analysis. The advantage of this method is that only one Markush

core exists during the enumeration process, and it is particularly useful if the final enumeration is computationally expensive.

3. Representing Markush Structure Variations. There are four different types of Markush structure variations (Figure 9). These variations are p-variation (position-variation), s-variation (substituent-variation), f-variation (frequency-variation), and h-variation (homology-variation), and typically two or more coexist in a single Markush structure.¹⁷

In the MDARC system, however, not all of these variations appear. For instance, p-variation is solved by enumerating the attachment point; therefore, each substituent is individually indexed. The “individually” indexed substituent can be converted to extended SMILES. The f-variation, on the other hand, is often indexed as text annotations in MDARC XML data, which is not included in this study.

4. A Novel Markush Structure Representation System. This study was inspired by the SMILES annotation of Markush structures previously reported (Barnard et al.⁵⁷). Although

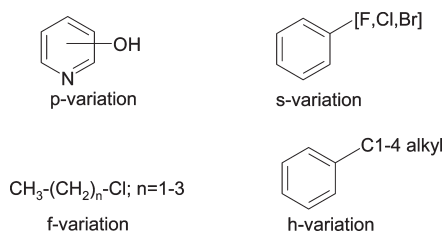


Figure 9. Different Markush structure variations.

handling heavily nested and multiply connected R-groups using SMILES has been implemented in Torus⁵⁸ from Digital Chemistry, the underlying algorithm is not publically available. The method described in this paper expands the basic SMILES annotation and develops it into a full Markush structure representation system. As shown in the Methods, accurate labeling of attachment points is crucial for Markush structure study. One challenge is how to seamlessly incorporate labeling of attachment points into an extended SMILES syntax, so that Markush structures can be displayed without using a proprietary Markush structure viewer. The novel features described in this paper include the following:

First, breaking the R-group to ensure correct alignment of attachment points has not been reported before. After breaking multiply connected R-groups into fragments, the new R-group labels contain both R-group number and attachment point information together. For example, “[R3002]” stands for part of the R3 group with an attachment point number of 2. The new dummy Z atom labels also contain R-group number and attachment point information together. For example, “C[Z3001]” indicates that it is one of the substituents of R3 and should be connected to [R3001]. In this way, an R-group substituent can be accurately aligned to its parent R-group or the core. In addition, the R-group hierarchy can also be recognized from the new labels. For instance, [R4]C-[Z3001] indicates that it is an R3 substituent and it contains the R4 group. Therefore, R4’s parent is R3. The R-group hierarchy is key to the interactive interface and bottom-up enumeration.

Second, the new labels are now part of the structure and not just additional annotations to the structure as in the original MDARC data. The dummy R/Z atoms are now incorporated into the extended SMILES syntax. Since the label is associated with the atoms and the bonds, the structure with labeling can be displayed in general molecular editors (such as the one in Pipeline Pilot), instead of a proprietary Markush structure viewer.

Finally, the extended SMILES in ASCII strings can be easily changed for structure manipulation and clearer display. For instance, for the Markush structure display, the broken R-groups can be patched back together by inserting new dummy atoms and simplifying the labels. Compared to the proprietary data, the extended SMILES is a relatively more “open format”. It can also be converted to other formats (such as Molfile) for broader molecular editor acceptance.

5. Future Prospective. As mentioned earlier, when converting MDARC data into SMILES, most information is retained. However, text annotation, e.g. an alkyl group is “low”, “medium”, or “high” in size, is currently not included. MDARC also has stereochemical information, which is not captured currently but will be considered in a future study.

In addition to MarVis, other applications for patent Markush structure analysis can be developed using the structural data in SMILES. One possible application could visualize the chemical space defined by the patent Markush structure, along with the exemplified

structures for comparison. DecrIPt Inc.⁵⁹ offers a random enumeration approach.⁶⁰ Alternative approaches that can provide a better representation of the chemical space would be valuable.

The advantage of using extended SMILES is that it is recognized by commercial molecular toolkits, like the one in Pipeline Pilot in this study or in Daylight.⁶¹ In addition, it is relatively human readable and easy to store and manipulate. Additional information, like atom labels, parent-child labeling, etc., can be easily added or removed by string manipulation. It can also be easily converted to other format, like the Molfile shown in this study. The extended SMILES used in this study can also help detect ambiguous grandparent-child alignment. On the other hand, the potential pitfalls of using SMILES include coincidence of rare earth metal atoms in patents and potential proprietary issue. In reality, some metals, e.g. cerium, do appear in chemical patents, and they may interfere with the dummy atoms representing superatoms. Due to the number of rare earth metals, only a limited number of superatoms can be represented. Also, even though SMILES has been used widely by commercial software, it is proprietary to Daylight Chemical Information Systems Inc.⁶² To avoid potential issues, other nonproprietary identifier for chemical substances, such as InChI or Open-SMILES,⁶³ may be a good alternative.

The conventions in this study, such as breaking R-groups and renumbering R/Z atom labels, can help other Markush structure studies. For instance, such conventions can be extended to SMIRKS for mapping an exact structure to Markush structure and fragment analysis for drug discovery project teams.

CONCLUSION

A full Markush structure representation system in extended SMILES is reported. Conventions, such as breaking R-groups and renumbering dummy R/Z atom labels, were introduced to ensure accurate alignment of the attachment points in Markush structures. The new labeling system incorporates all the information necessary in a compact syntax and yet readable by existing molecular toolkits.

The new representation system will hopefully contribute to future Markush structure studies. The value was demonstrated by the development of an application package, MarVis, which can help chemists to quickly review substructure search results from Questel MMS. MarVis displays a patent Markush structure with the query substructure expanded, an R-group table that shows the structures of all the described R-group substituents. An interactive interface was also developed, allowing the chemists to explore the chemical space of a Markush structure by selecting different R-group combinations. The new conventions are also under investigation to be extended to SMIRKS for mapping an exact structure to Markush structure and fragment analysis.

ASSOCIATED CONTENT

S Supporting Information. The algorithms of this study and an example MarVis report. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: (973)235-4593. Fax: (973)235-2134. E-mail: venus.so@roche.com. Corresponding author address: Roche, Bldg 76/12, 340 Kingsland Street, Nutley, NJ 07110, USA.

■ ACKNOWLEDGMENT

The authors would like to thank Paul Gillespie, Samuel Megerditchian, Robert Kester, Jefferson Tilley, Sung-Sau So, Kin-Chun Luk, and Rama Kondru at Roche for their helpful discussions. Eric Scott helped to build the interactive Web interface. The authors would also like to thank Questel Inc. and Thomson Reuters Inc. for providing patent data for this study. Financial support was provided by the Roche Postdoc Fellowship Program.

■ REFERENCES

- (1) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability 0.1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–154.
- (2) Ebe, T.; Sanderson, K. A.; Wilson, P. S. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT file. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31–36.
- (3) Benichou, P.; Klimczak, C.; Borne, P. Handling Genericity in Chemical Structures Using the Markush DARC Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 43–53.
- (4) Barnard, J. M. A Comparison of Different Approaches to Markush Structure Handling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64–8.
- (5) Berks, A. H. Current State of the Art of Markush Topological Search Systems. *World Pat. Inf.* **2001**, *23*, 5–13.
- (6) Simmons, E. S. Markush Structure Searching Over the Years. *World Pat. Inf.* **2003**, *25*, 195–202.
- (7) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148–150.
- (8) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.
- (9) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 161–168.
- (10) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Structures in Chemical Patents. 4. An Extended Connection Table Representation for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160–164.
- (11) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57–66.
- (12) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 6. An Interpreter Program for the Generic Structure Description Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66–71.
- (13) Gillet, V. J.; Welford, S. M.; Lynch, M. F.; Willett, P.; Barnard, J. M.; Downs, G. M.; Manson, G.; Thompson, J. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118–126.
- (14) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.
- (15) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 9. An Algorithm to Find the Extended Set of Smallest Rings in Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207–214.
- (16) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 10. Assignment and Logical Bubble-up of Ring Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215–224.
- (17) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 233–253.
- (18) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 12. Principles of Search Operations Involving Parameter Lists: Matching-relations, User-defined Match Levels, and Transition from the Reduced Graph Search to the Refined Search. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 253–260.
- (19) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260–270.
- (20) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 14. Fragment Generation from Generic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 453–462.
- (21) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 15. Generation of Topological Fragment Descriptors from Nontopological Representations of Generic Structure Components. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 369–377.
- (22) Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 16. The Refined Search: an Algorithm for Matching Components of Generic Chemical Structures at the Atom-Bond Level. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1–7.
- (23) Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 17. Evaluation of the Refined Search. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 659–662.
- (24) Lynch, M. F.; Holliday, J. D. The Sheffield Generic Structures Project - a Retrospective Review. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 930–936.
- (25) Dubois, J. E.; Laurent, D. DARC (Documentation and Automation of Correlation Research) System. Population-Correlation Theory, Organization, and Description. *C. R. Séance Acad. Sci., Ser. C* **1968**, *266*, 943–5.
- (26) Dubois, J. E.; Viellard, H. DARC System. VII. Theory of Generation-Description. 1. General Principles. *Bull. Soc. Chim. Fr.* **1968**, 900–4.
- (27) Dubois, J. E.; Viellard, H. DARC System. IX. Theory of Generation-Description. 3. General Description of Structures by DEL. *Bull. Soc. Chim. Fr.* **1968**, 913–19.
- (28) Dubois, J. E.; Viellard, H. DARC System. VIII. Theory of Generation-Description. 2. Establishment of the Uniline Descriptor of a Segment A_i-B_j : the DEL. *Bull. Soc. Chim. Fr.* **1968**, *3*, 905–12.
- (29) Dubois, J. E. Principle of the DARC [Description and Automated Research of Correlation] Topological System. Applications Pointing to Structural Influence on Oxidation of Hydrocarbons. *Entropie* **1969**, *25*, 5–13.
- (30) Dubois, J. E.; Anselmini, J. P.; Chastrette, M.; Hennequin, F. Description and Automation of the Investigations of Correlations (DARC) System. XI. Population-Correlation Theory. 1. Organization of a Chemical Population. *Bull. Soc. Chim. Fr.* **1969**, 2439–48.
- (31) Dubois, J. E.; Aranda, A. DARC [Description and Automation of Correlations Research]: Theory of Population-Correlation. Theory Generation, and Use of Clusters of Points. *C. R. Séance Acad. Sci., Ser. C* **1969**, *269*, 1451–3.
- (32) Dubois, J. E.; Hennequin, F.; Boussu, M. Use of the DARC [Description and Automation of Research of Correlations] Topological System for Documentation: Preparative Methods for Saturated Aliphatic Ketones. *Bull. Soc. Chim. Fr.* **1969**, 3615–23.

- (33) Dubois, J. E.; Laurent, D. Description and Automation of the Investigations of Correlations (DARC) System. XII. Population-Correlation Theory. 2. Description of a Chemical Population. *Bull. Soc. Chim. Fr.* **1969**, 2449–55.
- (34) Cayzergues, P.; Panaye, A.; Dubois, J. E. DARC Code. Description of Complex Stereocenters. *C. R. Séance Acad. Sci., Ser. C* **1980**, 290, 441–4.
- (35) Dubois, J. E.; Sobel, Y.; Mercier, C. Theory of Chemical Graphs. DARC/PELCO Method. Topology-Information Relationships and Regularity Concept. *C. R. Séance Acad. Sci., Ser. II* **1981**, 292, 783–8.
- (36) Questel Inc., 4, rue des Colonnes, 75002 Paris, France.
- (37) Berks A. H. Display for Markush Chemical Structures. US Patent No. 2005/0010603.
- (38) Weininger, D. SMILES, a Chemical Language and Information-System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31–36.
- (39) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique Smiles Notation. *J. Chem. Inf. Comput. Sci.* **1989**, 29, 97–101.
- (40) Weininger, D. SMILES. 3. Depict - Graphical Depiction of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1990**, 30, 237–243.
- (41) Pipeline Pilot, version 7.5; Accelrys Inc.: San Diego, CA, 2009.
- (42) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 244–255.
- (43) Thomson Reuters, 3 Times Square, New York, NY 10036, USA.
- (44) ChemAxon Kft., Máramaros köz 3/a, Budapest, 1037 Hungary.
- (45) Digital Chemistry, 30 Kiveton Lane, Todwick, Sheffield S26 1HL, United Kingdom.
- (46) Barnard, J. Markush Structure Searching. Presented at Chemical Information and Computer Applications Group Meeting. RSC, Burlington House, London, UK, Oct 28, 2009. http://www.rsc.org/images/J_BarnardOct09_tcm18-167682.pdf (accessed Jan 10, 2010).
- (47) Kuhn, S.; Helmus, T.; Lancashire, R. J.; Murray-Rust, P.; Rzepa, H. S.; Steinbeck, C.; Willighagen, E. L. Chemical Markup, XML, and the World Wide Web. 7. CMLspect, an XML Vocabulary for Spectral Data. *J. Chem. Inf. Model* **2007**, 47, 2015–34.
- (48) Tonnelier, C. A. G.; Fox, J.; Judson, P.; Krause, P.; Pappas, N.; Patel, M. Representation of Chemical Structures in Knowledge-Based Systems: The StAR system. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 117–123.
- (49) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. SYBYL Line Notation (SLN): A Single Notation to Represent Chemical Structures, Queries, Reactions, and Virtual Libraries. *J. Chem. Inf. Model* **2008**, 48, 2294–2307.
- (50) Stein, S. E.; Heller, S. R.; Tchekhovski, D. In *An Open Standard for Chemical Structure Representation - The IUPAC Chemical Identifier*; Proceedings of the International Chemical Information Conference, Nimes, France, 2003; Infonortics, 2003, pp 131–143.
- (51) MarvinView, version 5.3.1; ChemAxon: Budapest, Hungary, 2010.
- (52) Lee, C.-M.; Parks, J. A.; Bunnell, P. R.; Plattner, J. J.; Field, M. J.; Giebisch, G. H. [(Aminomethyl)aryloxy]acetic Acid Esters. A New Class of High-Ceiling Diuretics. 4. Substituted 6,7-Dichloro-2,3-dihydrobenzofurans Derived by Ring Annulation. *J. Med. Chem.* **1985**, 28, 589–594.
- (53) Barnard, J. M.; Wright, P. M. Towards In-House Searching of Markush Structures from Patents. *World Pat. Inf.* **2009**, 31, 97–103.
- (54) MMS Mini-Guide UNIX version, Nov 2008, Questel Inc.
- (55) MARPAT User Guide, 2008. CAS. <http://www.cas.org/ASSETS/D5DF848824894DD7AE0CBDC0A5C49122/marpatug.pdf> (accessed Jun 29, 2010).
- (56) ChemAxon. Query Guide - Special search types: Markush structures. http://www.chemaxon.com/jchem/doc/user/query_markush.html#combinatorialMarkushReduction (accessed Jan 10, 2010).
- (57) Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R. D. Use of Markush Structure Analysis Techniques for Descriptor Generation and Clustering of Large Combinatorial Libraries. *J. Mol. Graphics Modell.* **2000**, 18, 452–63.
- (58) Torus, version unknown; Digital Chemistry Ltd.: Sheffield, UK, 2010.
- (59) DecrIPt Inc., 154 Hempstead Street, New London, CT 06320, USA.
- (60) DecrIPt. <http://www.decript.net/company.php> (accessed Jan 24, 2011).
- (61) Daylight. Depict Interactive depiction of SMILES. <http://www.daylight.com/daycgi/depict> (accessed Jan 24, 2011).
- (62) Daylight Chemical Information Systems Inc., 28202 Cabot Road, Suite 300, Laguna Niguel, CA 92677, USA.
- (63) OpenSMILES. <http://www.opensmiles.org> (accessed Sep 23rd, 2010).