DERWENT MARKUSH RESOURCE DATABASE ON STN®

Reference Manual

February 2021

Authors:

Andreas Barth

Thomas Stengel

Content

List of current issues that affect the usability of the database. If possible those issues will be solved and the manual will be updated accordingly.

Issue	Page number
1	10
2	14
3	16
4	29
5	32
6	35
7	36
8	46
9	103

1. Introduction

Patent literature is considered to be an important source of technical knowledge and it has been stated that a large portion of new knowledge is not published elsewhere.[1] A significant part of this literature is classified as Chemistry patents, containing new chemical compounds, new ways of preparation, or new applications of known compounds. In general, the core of a patent claim in Chemistry is based on a chemical structure which could be generalized to broaden the scope of the respective claim. Any serious patent search must be able to find all relevant substances claimed in the patent literature, and this includes both specific structures as well as generalized (Markush) structures.

Figure 1 shows the timeline of patent documents in the database Derwent World Patent Index (DWPI) by category in the period from 1996 to 2015 (20 years). The lines show the total number of DWPI documents (grey), the Chemistry patents (blue), the patents with chemical codes (green), the patents with specific chemical structures (orange), and the patents with Markush structures (red). In order to fit into the diagram the total number of patents has been divided by a factor of 5 and the number of Chemistry patents has been divided by a factor of 2. The trends for chemical patents as well as those containing chemical codes or chemical structures generally follow the trend for the total number of patents. In contrast, the number of Markush patents per year is stable (around 30.000/year).
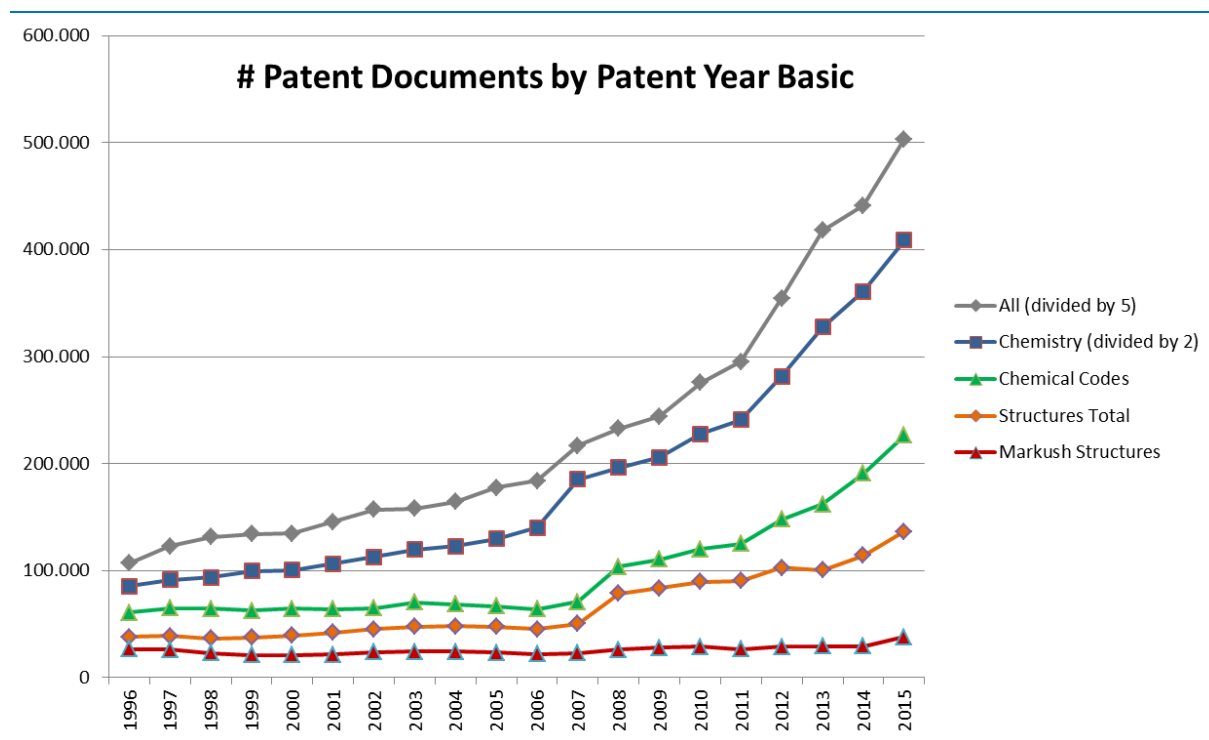
Figure 1. Timeline of patent documents in DWPI by category between 1996 and 2015.

---

[1]   The original source for this statement is from 1977: '*8th USPTO Technology Assessment and Forecast Report. Section II: "The uniqueness of patents as a technological resource"*, USPTO 1977'. In this document it states that 80% of new knowledge found in patents is not published anywhere else. In a recent study Anthony Trippe showed that "95% of the patented substances did not appear in the non-patent literature references" (*Revisiting an Old Standard – 80% of Technical Information is Found Only in Patents,* Anthony Trippe, 2014, http://www.patinformatics.com/).

Definition of a Chemical Patent

"A patent is a document, issued, upon application, by a government office (or a regional office acting for several countries), which describes an invention and creates a legal situation in which the patented invention can normally only be exploited (manufactured, used, sold, imported) with the authorization of the owner of the patent. "Invention" means a solution to a specific problem in the field of technology. An invention may relate to a product or a process. The protection conferred by the patent is limited in time (generally 20 years)" (www.wipo.int). In Chemistry it is possible to protect:

- Chemical compounds,
- Process of preparation, and
- Application or use.

Definition of a Markush Patent

In 1924, Dr. Eugene Markush (1887-1968) applied for a new patent on "*Pyrazolone dye and process of making the same*", where he used the formulation *"where R is a group selected from ..."*, in order to patent a large class of compounds (US Patent 1506316). Markush succeeded in the patent litigation process and since that time it is possible to use generalized chemical structures in patents. Hence, a Markush patent can be defined as a chemical patent which contains generalized chemical structure formulas with one or more structural variations.

Markush Structures

A Markush structure is a generalization of a specific structure (s. Figure 2). It consists of an invariant core structure (Markush master structure), usually shown in diagrammatic form together with a set of Markush variations. In the example shown, 4 types of variations are applied. First, the methyl group has been replaced by a variable group consisting of H, CH_3 or OH (**substitution variation**). Second, the position of the chlorine atom has a variable point of attachment (**position variation**). Third, the methyl bridge connecting the benzene ring and the amino group has become a repetition group (**frequency variation**). Finally, the ethyl group has been generalized to a carbon chain (**homology variation**).
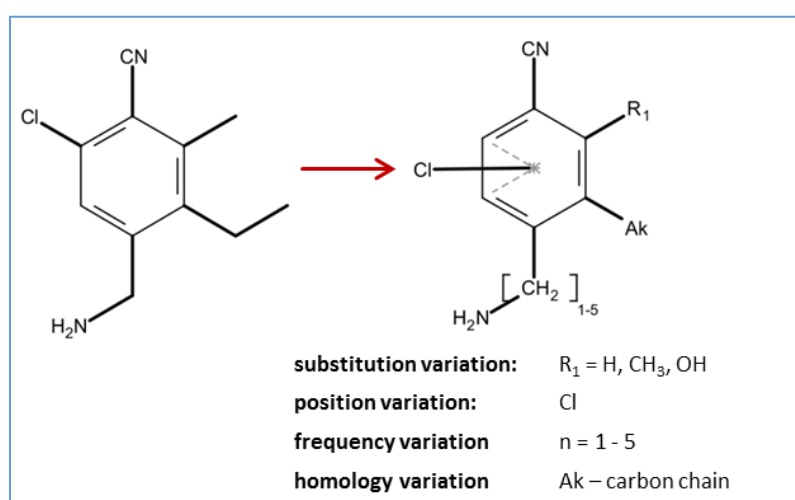
Figure 2. Creation of a Markush structure from a specific structure.

The example structure shown in Figure 2 is actually a rather simple Markush structure. In general, chemical structures in patents may include many nested G-groups, and each G-group may consist of several fragments. In addition, the description of Markush structures may contain further specifications such as node attributes and element counts, conditional logic, or shared variables to form a ring.

1.1. DWPIM on STNext and new STN

Derwent Markush Resource, file DWPIM, is accessible on new STN as well as on STNext. The content and functionality of DWPIM is the same on both platforms, however, there are differences with regard to the workflow. This section summarizes the typical characteristics of STNext which are important for proper use of DWPIM.

The corresponding bibliographic databases on STNext can be WPINDEX, WPIX and WPIDS dependent on the respective contractual situation. In the interest of claritiy those different types are summarized as DWPI in this manual.

1.1.1. Structure Search

For uploading a chemical structure a structure can either be drawn in the structure editor or selected from the folder "MyFiles". The uploaded structure is displayed with all associated attributes and definitions as well as match level information as shown in Figure 3. This information is also provided in the transcript.

1.1.1.1. Overview on attributes online and in transcripts
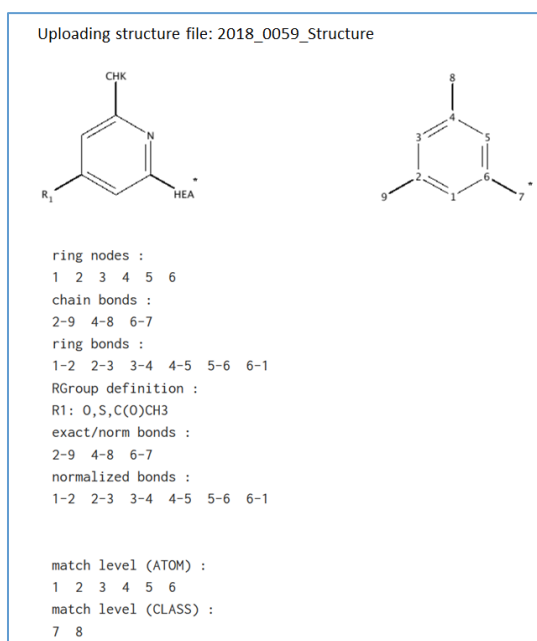
Figure 3: Detailed information associated with the query structure

> **Issue 1:** Ring lock is only indicated with **bold bonds** in the structure drawing but not mentioned textually. Unspecified bonds are shown as punctuated lines in the structure drawing but named incorrectly in the text as „exact/normalized".

1.1.1.2.   Description of search types and modes

There are two different search types, Substructure search (SSS) and Closed substructure search (CSS), three different search scopes, namely sample search, full search and batch search.

In the following some aspects are described in more detail:

Default setting for structure search is the sample search which retrieves every tenth record with a maximum of 50 records. The commands for sample, full substructure and closed substructure search are as follows:

```
=> S L-NUMBER SAM

=> S L-NUMBER FULL SSS

=> S L-NUMBER FULL CSS
```

The search time for SSS and CSS is limited to 60 minutes.

By using the batch mode the search time period is extended to 90 minutes. A batch search can be done as follows:

```
=> BATCH

ENTER QUERY L# FOR BATCH REQUEST OR (END): L-NUMBER

ENTER BATCH REQUEST NAME OR (END): NAME/B

ENTER TYPE OF SEARCH (SSS) OR CSS:.

ENTER SCOPE OF SEARCH (FULL) OR RANGE:RANGE

If Range is selected:  ENTER RANGE OR (ALL):.
```

After completion of batch search the results can be retrieved as follows:

```
=> D SAVE

=> ACT NAME/A
```

The batch search may have the following advantages compared to the online mode :

- Increased search time per structure compared to the online mode may reduce number of iteration incompletes

- Workload is shifted
- Extended search time of one hour may increase number of completed searches

1.1.1.3. Iteration Incompletes

If a hit structure is unequivocally obtained within the system limits the iteration status is "complete". If a search for a hit structure cannot be completed within the system limits the iteration status is "incomplete". The number of incompletes is indicated in the header of the structure search answer set. In these cases the validity of the hits cannot be established unequivocally and has to be checked manually by displaying the full record.

Affected records can be separated by the following commands *(L1 represents the substructure query)*:

```
=> S L1 SSS FULL      -> L2 (answer set of completes and incompletes)

=> S L2/INC           -> L3 (answer set of incompletes only)

=> S L2 NOT L3        -> L4 (answer set with completes only)
```

Please note that there is currently no hit structure support for the "iteration complete" answer set obtained by a search using L2/COM *(COM for completes)*.

1.1.1.4. Information after sample and structure search

There is no full file projection for online and batch search available. Consequently the information content of the number of iterations is low and in the case of the full file search it is even inapplicable.

Information on Iterations:

- Sample search: the number to iterate equals the number of records in the DWPIM file (In Figure 4 it is 2254007 records).
  The number of iteration value indicates how many iterations the system has conducted to yield 50 sample records (In Figure 4 there are 1316 Iterations required to get the maximum of 50 sample hits).
- Full Search: there are always 0 iterations displayed. This value has no meaning and should be ignored.

```
SAMPLE SEARCH INITIATED 16:01:16
SAMPLE SCREEN SEARCH COMPLETED -  2254007 TO ITERATE

  0.1% PROCESSED      1316 ITERATIONS                      50 ANSWERS
SEARCH TIME: 00.00.09

L4          50 SEA SSS SAM L1

=> s l1 sss ful

FULL SEARCH INITIATED 16:01:50
FULL SCREEN SEARCH COMPLETED -       0 TO ITERATE


100.0% PROCESSED          0 ITERATIONS                  84921 ANSWERS
SEARCH TIME: 00.00.17

L5       84921 SEA SSS FUL L1
```

Figure 4: Example for information on "Iterations" after sample and full search

1.1.2.   Display results

A clear and understandable display of Markush chemical groupings can significantly support the efficient and reliable evaluation workflow following a substance search. To address this requirement, STN offers several display options comprising different depth of detail. The following display types are available:

- **Full display**: comprises the complete record with all available information (*i.e.* core fragment G0, all pertaining G-groups with all possible alternatives) including highlighting.
- **Brief display**: reduction to the "hit", besides the core fragment G0 only generic groups that were part of the structure query are displayed including highlighting. The summary view is the default display format which is automatically displayed for each result record after a structure search.
- **Assembled display** (default): display of a "hit" in **one** structure by aligning enumerated structures to the Markush core structure. Thereby only query-relevant fragments are automatically assembled in their correct location in the chemical structure. Especially in the case of very generic Markush records with nested G-groups the answer assembly may provide an easier way to interpret the view of the Markush results.

Which rules are applied for delivering highlighting information?

- If more than one alternative atom or fragment within a G-group matches a query, the system will only highlight one of them.

Which rules are applied for the assembly of the assembled display?

If more than one alternative atom or fragment within a G-group matches a query, the system will automatically select one relevant atom or fragment, respectively.

All display types provide highlighting (in red) of the query fragment in the core fragment (G0) and/or in the G-groups, respectively.

The commands for retrieving the assembled display (asb), brief display (brief) and full display (all) are as follows:

```
=> D L-NUMBER ASB

=> D L-NUMBER BRIEF

=> D L-NUMBER FULL
```

Optionally a range can be defined by:

```
=> D L-NUMBER RANGE DISPLAY TYPE
```

1.1.3. Substance Descriptors (SDM) and Markush Descriptors (MDE)

The search by substance descriptors and Markush descriptor can be done as follows:

```
=> S SUBSTANCE DESCRIPTOR/SDM

e.g. => S M/SDM; M = metals, alloys

=> S L-NUMBER AND MARKUSH DESCRIPTOR/MDE

e.g. => s L1 and S/MDE; S = single specific structure
```

For more information on SDM's see chapter 3.8.4, for MDE's see 3.8.5.

1.1.4. Crossover from DWPIM to DWPI

The crossover from DWPIM results to DWPI is established by performing a search of the respective DWPIM L-number in DWPI.

```
=> FIL DWPIM

=> S L-NUMBER SEARCH TYPE (E.G. L1 SSS FUL)

=> FIL WPIX

=> S L-NUMBER

=> D L-NUMBER
```

The assembled display is the default display in DWPI. The commands for the assembled, brief and full hit structure in DWPI are as follows:

```
=> D L-NUMBER AHITSTR

=> D L-NUMBER BHITSTR

=> D L-NUMBER FHITSTR
```

Issue 2: DWPIM hitstructures are only preserved in WPIX/WPINDEX answer sets within the same session (including log off hold). Otherwise hitstructures are lost and the full Markush record is diplayed. This limitation may also affect results from SDI's and batch searches.

Recommendation: hitstructure information within DWPIM is preserved. Therefore all answer sets should always be also saved in DWPIM.

Records from DWPIM and WPIX/WPINDEX can be mutually allocated by Markush number (AN in DWPIM corresponds to MCN in WPIX/WPINDEX).

1.1.5. Crossover of Compounds from DWPI to DWPIM

Markush compounds from DWPI records can be extracted and displayed in DWPIM by applying the transfer command. It is important to note that this step requires a reassignment of compound suffix MCN to AN.

The workflow is described in the following:

```
FIL DWPIM

TRA L-NUMBER WPIX RECORD [RANGE] MCN /AN
```

Example: Tra L1 1-3 MCN /AN

1.1.6. Application of Roles in DWPI

The DCR and DCN roles as described in Chapter 3.8.2 can be applied by the following commands:

```
=> FIL DWPIM

=> S L-number search type (e.g. S L1 SSS FUL)

=> FIL WPIX

=> S L-NUMBER (T) ROLE/MCN (e.g. S L1 (T) PRD/MCN)

=> D [RANGE] HIT (e.g. D 1-3 HIT)
```

1.1.7. Application of Fragmentation Codes

Besides specific and Markush substance indexing (backfile to 1963) fragmentation codes also encompass non-structural topics such as pharmaceutical and agrochemical activity. Application of Non-structural codes can enhance search accuracy by specifically linking activity concepts to the DWPI chemical indexing for a substance.

Fragmentation Codes can be applied by the following commands:

```
=> FIL DWPIM

=> S L-NUMBER SEARCH TYPE (E.G. S L1 SSS FUL)

=> FIL WPIX

=> S L-NUMBER (P) FRAGMENTATION CODE/M0,M1,M2,M3,M4,M5,M6
(E.G. S L1 (P) P960/M0,M1,M2,M3,M4,M5,M6)
CODE P960 = PERSONALIZED MEDICINE

=> D [RANGE] CMC (OR HIT) (E.G. D 1-3 CMC)
```

1.1.8. Alert options

1.1.8.1. Automatic structure-based alerts (SDI's), including GUI support

```
=> FIL DWPIM

=> S L-number search type (e.g. L1 full sss)

=> SDI L-number
```

The settings can be defined in the following steps (not shown here).
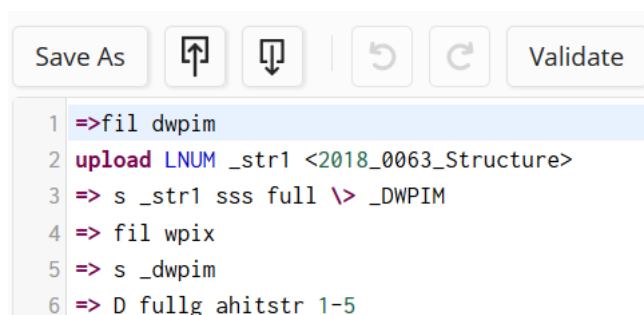
A SDI name has the syntax "SDI name/s".

An existing SDI can be edited by "SDI EDIT" command.

Issue 3: The SORT command is currently not supported for SDI alerts.

1.1.8.2. Use of scripts to upload structures and cross-over to DWPI

By using script language the upload and search for structures and the subsequent cross-over to the bibliographic database DWPI can be automated. An workflow example is shown in Figure 5 . The name of the example structure <2018_0063_Structure> corresponds with the respective entry in the "My Files" folder.


```
  Save As    田    田    |    ↺    ↻    Validate

1  =>fil dwpim
2  upload LNUM _str1 <2018_0063_Structure>
3  => s _str1 sss full \> _DWPIM
4  => fil wpix
5  => s _dwpim
6  => D fullg ahitstr 1-5
```

Figure 5: Example script for structure upload, search and cross-over.

By running the example script the system automatically uploads the defined structure, performs a full substructure search, initiates a cross-over and search in DWPI and finally displays the bibliographic information including the assembled hit-structures for the first five records.

1.1.9. Subset Search

A subset search can be done as structure-, text-, and combined structure-text-search.

Example for combined structure-text-search:

- *L-number from structure-search*, e.g. S STR1 SSS FULL: *L1*

- *L-number from text-search*, e.g. S Y/SDM: *L2*

- S *L1* subset=*L2* full: *L3* gives all records from *L1* which have Y as an associated substance descriptor SDM.

Example for structure-in-structure-search:

- S *structure 1* full: *L1*

- S *structure 2* full: *L2*

- S *L1* subset=*L2* full: displays hit structures from L1 which overlap with L2

- S *L2* subset=*L1* full: displays hit structures from L2 which overlap with L1

1.2. Included Databases – Derwent Content Domain

Chemical information is indexed by Clarivate Analytics (formerly Thomson Reuters, originally Derwent) in 3 databases:

- DWPI (Derwent World Patent Index) covers all patent literature (incl. Chemistry)
- DCR (Derwent Chemistry Resource) contains specific structures, referenced in DWPI
- Derwent Markush Resource (DWPIM) contains Markush structures, referenced in DWPI.

For the DWPIM database, Clarivate Analytics indexes the patent literature from 33 patent issuing authorities worldwide (s. Table 1).

Table 1. Patent issuing authorities which are currently indexed in DWPIM including the starting years of indexing.

Country	Start	Country	Start
Argentina	2015	Malaysia	2010
Austria	1987	Mexico	2015
Australia	1987	Netherlands	1987
Belgium	1987	New Zealand	1987
Brazil	2010	PCT/WO	1978 (Pharma); 1982 (Agro + Gen. Chem.)
Canada	1987	Poland	2011
China	2008	Russia	1993-98 and 2010 to date
European Patents	1978 (Pharma); 1982 (Agro + Gen. Chem.)	Singapore	2014
France	1961-76 (Fr-M)[2]; 1978 (Pharma); 1982 (Agro + Gen. Chem.)	South Africa	1987

[2] FR-M refers to French Medicament series patents. These were issued as a separate series by the French patent office INPI and then discontinued around 1976. In DWPI they have kind code FRM.

Country	Start	Country	Start
Germany	1980 (Pharma); 1983-84 and 1987 to date (Agro + Gen. Chem.)	Spain	2010
Great Britain	1980 (Pharma); 1983-84 and 1987 to date (Agro + Gen. Chem.)	Sweden	1987
Gulf Cooperation Council	2008	Switzerland	1987
India	2000	Thailand	2010
Indonesia	2013	Turkey	2015
Ireland	1987	United States	1978 (Pharma); 1982 (Agro + Gen. Chem.)
Japan	1987	Vietnam	2010
Korea	2008		

For the major patent offices the indexing started in 1978 with Pharma documents, followed in 1982 by Agro and General Chemistry. In addition, the back file data indexed by INPI for French patents has been included (1961-1976).

The Clarivate Analytics Chemistry Data Content comprises more than 3.2 million patent families with specific and/or Markush structures. Currently, there are more than 2.3 million Markush structures available in the DWPIM file, and the DCR file contains more than 3 million specific structures.[3] The records in DWPIM are structure based, *i.e.* one record contains a single Markush structure with all variations. The Markush compound number is linked to the corresponding record in DWPI. The content of DWPIM includes the following substance classes:

- Organic compounds
- Organometallic compounds
- Inorganic compounds (salts, metals and non-metallic elements, transition metal complexes)[4]
- Polymers[5]
- Fullerenes
- Peptides.

The three databases DWPI, DCR, and DWPIM build the Clarivate Analytics content domain where the databases are connected via the structure keys. Possible scenarios are

- Scenario 1: Start with a structure search in DWPIM and DCR, crossover to DWPI and combine the results with Derwent roles and/or any bibliographic and patent information
- Scenario 2: Start with bibliographic and patent information in DWPI, crossover to DWPIM and DCR, and confine the results with a subset structure search.

---

[3] The figures are obtained on October 2020.
[4] Indexing of alloys and intermetallics only for pharmaceutical (B) and agrochemical patents (C).
[5] Only for pharmaceutical (B) and agrochemical (C) patents. Polymers have to relate to the invention (e.g. therapy, purpose of the patent), otherwise they are not indexed.

While in new STN DCR is a separate database in STNext is integrated in the corresponding bibliographic database. As a consequence searching for DCR records on STNext has to be done within DWPI by using the command "S L-number" (either as substructure or closed substructure search). The corresponding bibliographic records can be retrieved by using the command "S L-number/DCR". For Markush searches it is important to note that respective records are indexed with the identifier AN in DWPIM and MCN in DWPI (see also Figure 6).
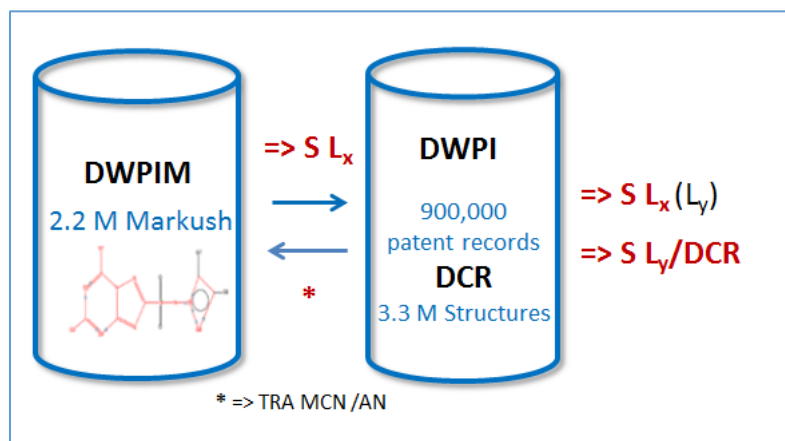
Figure 6: Clarivate Analytics content domain consisting of DWPI, DCR, and DWPIM on STNext (L_x is L-number of uploaded chemical structure)

Prior to the introduction of DCR in 1999/2000 the indexing policy was different. Both specific and generic structures were covered by Markush structures, often as part of the same structure. Where specific compounds were claimed as preferred embodiments of a Markush structure, these specifics were incorporated into the indexed Markush structure by including the appropriate values into the variable groups (e.g. structure 9830-C8901, Figure 7). In cases where the patent claimed just a new single specific compound, this was indexed using the normal Markush indexing process, with a new Markush structure record and Markush Compound Number created (e.g. structure 9213-F8101, Figure 7).
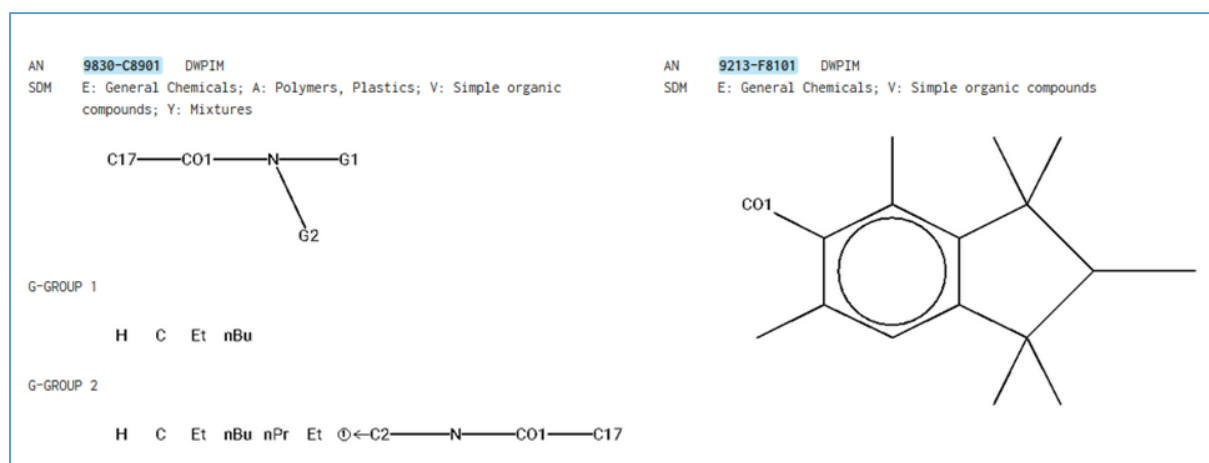
Figure 7. Examples for specific structures in DWPIM.

1.3. Indexing Limitations

Current indexing limits for DWPIM include:

- Connectivity: maximum 8
- Number of G-groups: maximum 50
- Number of fragments in a G-group: maximum 50
- Nesting between G-groups: maximum 4 levels
- In case no specific ratio is given the ratio 1-255 is used whereby 255 is the maximum allowable upper limit allowed for entering ratios of components. The system does not allow for the option of leaving the upper or lower value blank.

Current indexing limits for Markush structures and DCR structures in DWPI:

- For each DWPI document the upper limit to the sum of Markush and DCR structures indexed is 99.

- All DCR structures which cannot be indexed as specific structures due to the existing limit will be covered by a Markush structure containing several single specific structures. Thereby it is ensured that the whole chemistry content of a patent is retrieved after a structure search - either as a specific structure and/or as a Markush structure.

1.4. Working Example for Indexing and Representation of Markush Structures in DWPIM

Markush structures in DWPIM may contain specific atoms, shortcuts, superatoms (representing generically-described groups in the structure), or variable groups. To provide an understanding of the indexing and representation of Markush structures in DWPIM, the relationship between the original patent, the corresponding DWPI record and the DWPIM record is illustrated in the next 4 figures.

- Figure 8 shows an example of a Markush patent (WO 2007031739 A1)
- Figure 9 is an extract of the DWPI document corresponding to the Markush patent WO 2007031739 A1
- Figure 10 is a collection of the Markush record (0348-32901), containing the scaffold and an extract of some G-groups
- Figure 11 shows an example of a Markush structure from record 0348-32901 with preferred definitions according to the corresponding DWPI record.

The structure of formula I (s. Figure 8) together with a set of derivatives and salts has been claimed in patent WO 2007031739 A1 as a new and useful drug, *e.g.* for the treatment of type 2 diabetes. The title of the original patent is *"Heterobicyclic compounds as glucokinase activators"*. All the variations are described in detail for the list of R-groups. In addition, the patent claims other compound classes as new and describes the corresponding preparations. Here, we have restricted ourselves to a simplified extract of the patent full-text. In DWPI we find the complete documentation of the invention (s. Figure 9) with the Derwent title *"New pyrrolopyridine derivatives are glucokinase activators useful in the preparation of medicament for treating non-Insulin dependent diabetes"* (Accession number:

2007-495110). Among other references the DWPI document refers to the corresponding Markush record in the indexing section (MCN 0348-32901).

Heterobicyclic compounds as glucokinase activators
WO 2007031739 A1

Summary

Compounds of Formula (I) wherein R1 to R11, A and X1 to X3 are as described in the specification, and their salts, are activators of glucokinase (GLK) and are thereby useful in the treatment of, for example, type 2 diabetes. Processes for preparing compounds of formula (I) are also described.

Claims:

1. A compound of Formula (I):
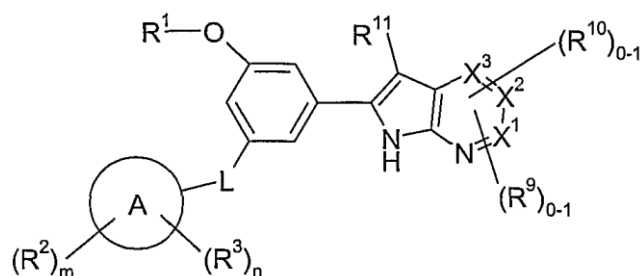
(I) wherein:

Ring A is selected from phenyl and HET-I

X^1, X^2 **and** X^3 are each independently CH or N, with the proviso that only one of X^1, X^2 and X^3 may be N; …

L is a linker selected from -O- and -(l-3C)alkylO- …

R^1 is selected from (l-6C)alkyl, (2-6C)alkenyl, (2-6C)alkynyl, (3-6C)cycloalkyl, …

HET-I and HET-Ia are independently a 4-, 5- or 6-membered, C- or N-linked saturated, partially or fully unsaturated heterocyclyl ring containing 1, 2, 3 or 4 heteroatoms independently selected from O, N and S, …

R^2 is selected from -C(O)NR^{4}R^5, -SO$_2$NR^{4}R^5, -S(O)$_P$R^4 and HET-2;

HET-2 is a A-, 5- or 6-membered, C- or N-linked saturated, partially or fully unsaturated heterocyclyl ring containing 1, 2, 3 or 4 heteroatoms independently selected from O, N and S, …

R^3 is selected from halo, fluoromethyl, difluoromethyl, trifluoromethyl, methyl, (1-4C)alkoxy, carboxy and cyano

R^4 is selected from hydrogen, (l-4C)alkyl …

R^5 is (independently at each occurrence) selected from hydrogen, (l-4C)alkyl and (3-6C)cycloalkyl; …

R^4 **and** R^5 are independently selected from hydrogen and (l-4C)alkyl; …

R^6 is selected from (l-4C)alkyl, -C(O)(I -4C)alkyl, …

R^7 is selected from -OR^5, (l-4C)alkyl, -C(O)(I -4C)alkyl…

HET-3 is an N-linked, 4 to 7 membered, saturated or partially unsaturated heterocyclyl ring, …

when R^8 is a substituent on carbon it is selected from halo, -OR^5, (l-4C)alkyl, (2- 4C)alkenyl, (2-4C)alkynyl, …

R^9 is selected from (l-4C)alkyl, halo, cyano, hydroxy (l-4C)alkyl, …

R^{10} is selected from methoxy, methyl and halo

R^{11} is selected from hydrogen and (1 -4C)alkyl; … ; m is O or 1; n is O, 1 or 2; or a salt thereof.

Figure 8. Example of a Markush patent (WO 2007031739 A1). The text has been extracted from the full-text (R-groups are simplified).

In the section **Extended Abstract** the DWPI document also lists the Preferred Definitions for the Markush structure variations. Derwent uses a different set of generic groups for indexing (G-groups) and the correspondences are given in parenthesis, *e.g.* R11 in the original patent corresponds to G6 in the DWPIM record.

- R^{11} = H (G6)
- A = phenyl (G3)
- L = O or (1-3C) alkylO (G2)
- R^1 = 1-6C alkyl (optionally substituted by OH or 1-4C alkoxy) (G1 + G19)
- R^2 = methylsulfonyl or azetidinylcarbonyl (G4 = G10 + G11 + G24)
- R^3 = fluoro, chloro, CN, OCH3 or carboxy (G5, G48)
- R^9 = halo, CH3 or OCH3 (G8)
- R^{10} = absent (G9)
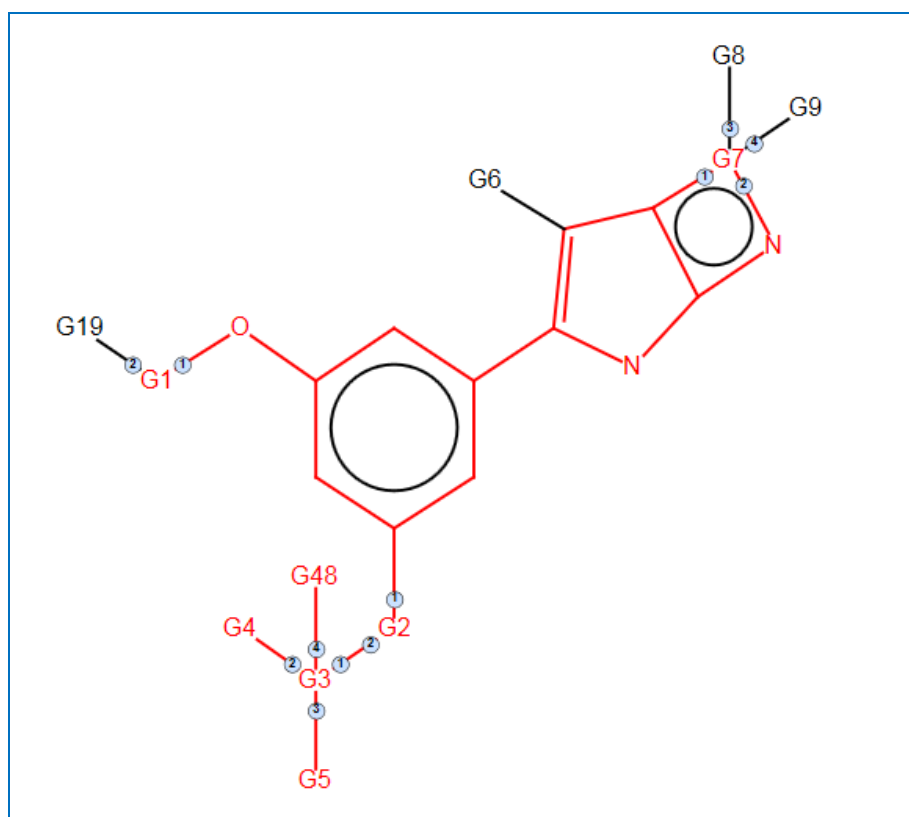- X^3 = CH; X^2, X^1 = N (G7)
- m, n = 0 or 1.

---

1. New pyrrolopyridine derivatives are glucokinase activators useful in the preparation of medicament for treating non-insulin dependent diabetes
Entered STN: 27 Jul 2007
Last update: 22 Dec 2011

Invention
Title Extensions

| Title Terms: | NEW DERIVATIVE ACTIVATE USEFUL PREPARATION MEDICAMENT TREAT NON INSULIN DEPEND DIABETES |

Document Identification

| Accession Number: | 2007-495110 [200748] |
| Document Number: | C2007-182461 |

| Derwent Class: | B02 |
| : | |

Thomson Indexing

Markush Compound Number (MCN)

Index Terms:	[1] MCN-0348-32901-CL MCN-0348-32901-NEW MCN-0348-32901-PRD; DCR-1497594-CL DCR-1497594-NEW DCR-1497594-PRD; DCR-1497595-CL DCR-1497595-NEW DCR-1497595-PRD; DCR-1497596-CL DCR-1497596-NEW DCR-1497596-PRD; DCR-1497597-CL DCR-1497597-NEW DCR-1497597-PRD
File Segment:	CPI
Manual Code(s):	B06-D05 B06-D08 B14-E12 B14-L01A2 B14-S04A

| Field Availability: | DNC, PA, PACO, IN, PN, AI, FDT, PRAI, IPC, IPCI, EPC, NCL, FTERM, CPC, TT, TI, AB, TECH, ABEX, GI, MC, CMC, RIN, DCN, MCN, DCR, IT |

Derwent Abstract

Patent Number

WO 2007031739 A1 [20090317]

| Novelty (NOV): | Pyrrolopyridine derivatives (I) and their salts are new. |

Figure 9. DWPI document describing "new pyrrolopyridine derivatives".

The Markush record in Figure 10 consists of a master structure or scaffold and a set of variations encoded as structure fragments in the G-groups. To provide a better overview of the Markush structure the figure shows only a small selection of the variations. The master structure looks rather similar to formula I in the original patent. However, there are some differences:

- The ring system with the fused 5- and 6-membered heterocycle from the original patent is indexed as a 5-membered heterocycle (pyrrole) fused with an "aromatic" 4-membered ring. The latter contains a group with several fragments with three nodes to describe all the variations allowed for X^1, X^2, and X^3 (example: the fragment of G7). An insertion of these fragments yields a 6-membered ring. This construction allows the indexer to write a more compressed Markush structure.
- R^1 is described by G1 + G19.
- In the Markush structure G3 corresponds to a phenyl ring with different possible substitution positions. Different from the description in the original patent where the phenyl ring (A) carries only two additional substituents (R^2 and R^3) we find three substituents in the Markush structure (G4, G5, and G48). In fact, G5 and G48 are identical. This construction is necessary

to take into account that the repetition factor of R^3 (G5, G48) is n = 0, 1, 2. In the case of 2 it is necessary to have two identical groups with independent variations of the fragments.

Figure 10. Markush document 0348-32901 corresponding to patent WO 2007031739 A1. Encircled in red are the preferred definitions from the DWPI document.

Figure 11 shows an example of a Markush structure from Markush record 0348-32901 with preferred definitions (assembled structure).
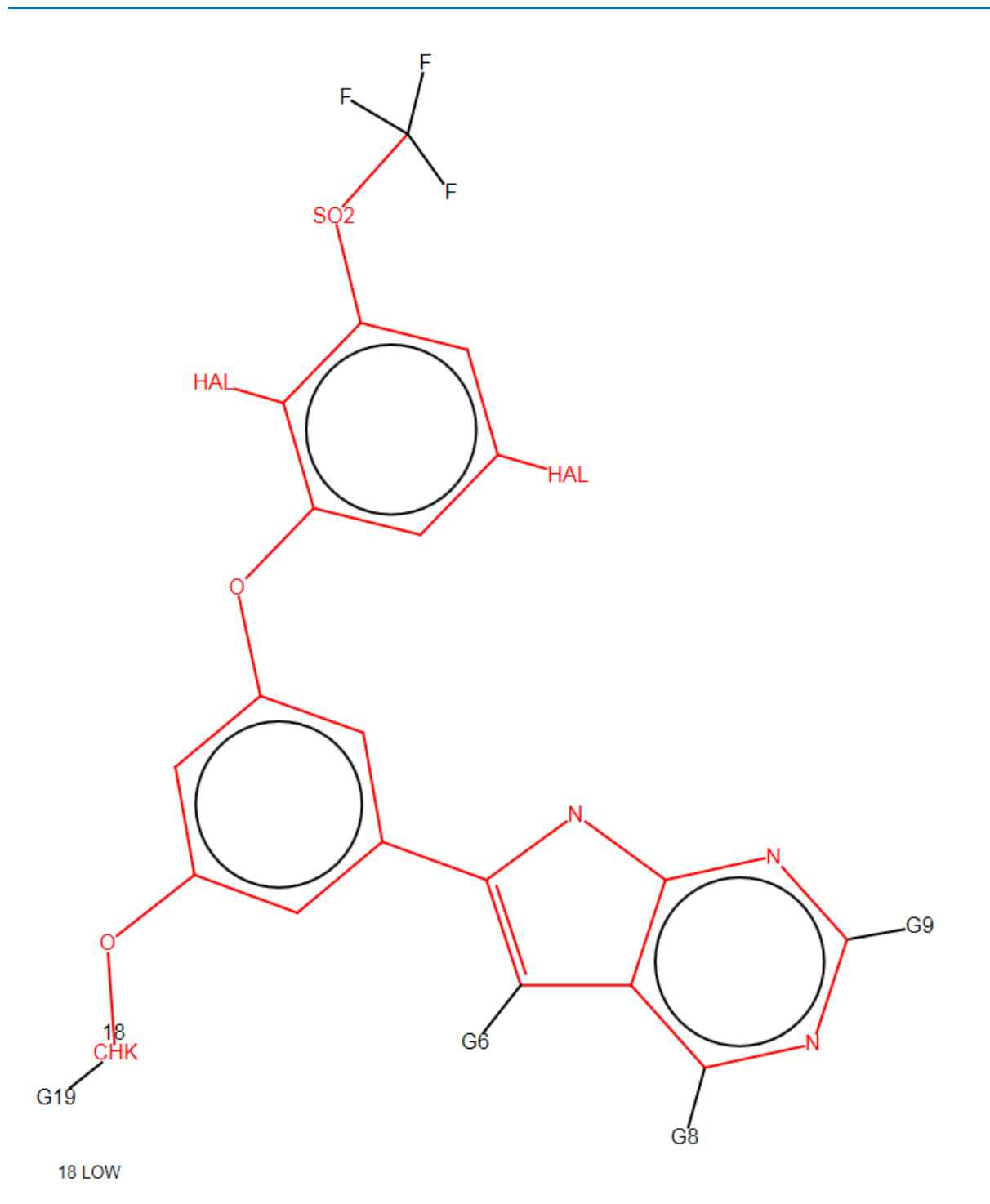

Figure 11. Example from the Markush record 0348-32901 with preferred definitions.

2. Derwent Markush Concept

In most cases chemical structures in patents are presented as generic or Markush structures. Formally, these structures consist of nodes, bonds, attributes, and other descriptions. Nodes could be

- Atoms (specific nodes)
- Shortcuts (common used fragments or functional groups)
- G groups (generic groups)
- Superatoms (generic nodes).

Bonds between the nodes are defined by bond type (chain, ring, ring/chain) and by bond value (single, double, etc.). In the sequel the elements of a Markush structure as indexed by Clarivate Analytics are described in detail. Note: the structure elements used for the formulation of a structure query in STN are sometimes different from those which are used by Clarivate Analytics for the indexing of Markush structures.

2.1. Chemical Nodes

2.1.1. Atoms and the Derwent Element Groups

Atoms are indexed using the common element symbols. However, the grouping of elements is specific to Derwent and is done according to the organization of the Derwent fragment codes (CPI: Chemical Codes). The periodic system with the atomic numbers, the element symbols, and the normal valences is shown in Figure 12. The colors indicate the different Derwent groups.

Group/Period	1	2	3		4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H 1																		2 He 0
2	3 Li 1	4 Be 2												5 B 3	6 C 4	7 N 3	8 O 2	9 F 1	10 Ne 0
3	11 Na 1	12 Mg 2												13 Al 3	14 Si 4	15 P 3	16 S 2	17 Cl 1	18 Ar 0
4	19 K 1	20 Ca 2	21 Sc 3		22 Ti 4	23 V 5	24 Cr 3	25 Mn 2	26 Fe 3	27 Co 2	28 Ni 2	29 Cu 2	30 Zn 2	31 Ga 3	32 Ge 4	33 As 3	34 Se 2	35 Br 1	36 Kr 0
5	37 Rb 1	38 Sr 2	39 Y 3		40 Zr 4	41 Nb 5	42 Mo 6	43 Tc 7	44 Ru 3	45 Rh 3	46 Pd 2	47 Ag 1	48 Cd 2	49 In 3	50 Sn 4	51 Sb 3	52 Te 2	53 I 1	54 Xe 0
6	55 Cs 1	56 Ba 2	57 La 3	★	72 Hf 4	73 Ta 5	74 Cr 3	75 Re 7	76 Os 4	77 Ir 4	78 Pt 2	79 Au 3	80 Hg 2	81 Tl 2	82 Pb 4	83 Bi 3	84 Po 2	85 At 1	86 Rn 0
7	87 Fr 1	88 Ra 2	89 Ac 3	★★	104 Rf	105 Db	106 Sb	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn						

				4	5	6	7	8	9	10	11	12	13	14	15	16	17
Lanthanides	★			58 Ce 3	59 Pr 3	60 Nd 3	61 Pm 3	62 Sm 3	63 Eu 3	64 Gd 3	65 Tb 3	66 Dy 3	67 Ho 3	68 Er 3	69 Tm 3	70 Yb 3	71 Lu 3
Actinides	★★			90 Th 4	91 Pa 5	92 U 6	93 Np 5	94 Pu 4	95 Am 3	96 Cm 3	97 Bk 3	98 Cf 3	99 Es 3	100 Fm 3	101 Md 3	102 No 3	103 Lr 3

Legend:
- Hydrogen
- Alkali and Earth Alkali Metals
- Group 13 - 15 Metals (formerly IIIa to Va)
- Transition Metals
- Lanthanides (without Lanthan)
- Actinides (including Actinium)
- Halides
- Nonmetals
- Noble Gases

Figure 12. Periodic system with the Derwent specific grouping of chemical elements.

28

2.1.2. Shortcuts

Shortcuts are abbreviations for commonly used chemical groups. A list is given in Table 2. For structure searches the shortcuts are automatically replaced by the corresponding structure fragments.

Table 2. List of shortcuts used for indexing by Clarivate Analytics.

Shortcut	Structure Fragment	Shortcut	Structure Fragment
CO_2	COOH	nPr	n-propyl
CO_1	CO	iPr	iso-propyl
SO_2	SO_2	iBu	iso-butyl
SO_3	SO_3H	nBu	n-butyl
PO_3	PO_3H2	sBu	sec-butyl
PO_4	OPO_3H_2	tBu	tert-butyl
CN	CN	Cn	divalent straight carbon chain, n represents the chain length
NO_2	NO_2	Ph	phenyl
Ace	$C(O)CH_3$	oBe	ortho-phenylene
Et	ethyl	mBe	meta-phenylene
		pBe	para-phenylene

Note: The shortcuts oBe, mBe and pBe can also be integrated into a ring system of a query structure. A search for a dihydro azetidine fragment containing oBe would e.g. yield indoles in the answer set (s. Figure 13, left image). Furthermore these shortcuts can also be present in the assembled structures (right image). In some cases the divalent carbon chain Cn can also be integrated into a ring system (right image).
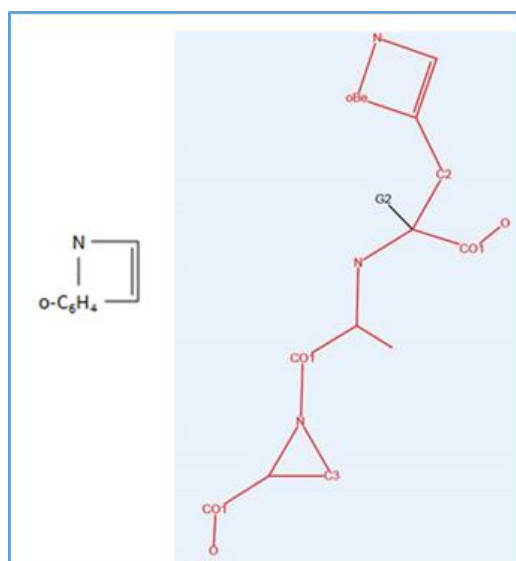

Figure 13. The shortcuts oBe, mBe and pBe can be integrated in ring systems (left image: query structure; right image: assembled display). The divalent carbon chain Cn within a ring system of an assembled structure (C2, right image).

Issue 4: The following shortcuts do not work (result in 0 hits) because their bond value is firmly defined as normalized while those bonds are indexed as exact: COOH, CO2H, COSH,CSSH, CS2H, OPO3H2, PO3H2, PO2H, OSO3H, SO3H, SO2H, SeO3H, S2O2H (Figure 14).
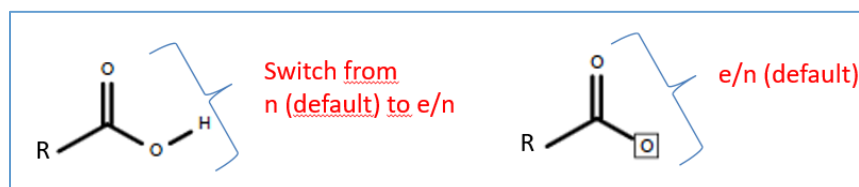
Possible ways how to handle this issue:

Figure 14: Change bond value for certain shortcuts for queries in DWPIM

2.1.3. Superatoms

In Derwent nomenclature generalized chemical elements are called superatoms. These superatoms represent

- Chemically defined open sets, *e.g.* alkyl defines the superatom CHK (Table 3)
- Chemically defined closed sets, *e.g.* halogen defines the superatom HAL (Table 4)
- Set defined by a property, *e.g.* the chromophore or fluorescent group is defined by the superatom DYE (Table 5).

There are 22 superatoms available plus an additional set of 30 amino acid shortcuts. The next tables and figures list the various sets of superatoms together with some examples (excluding amino acid shortcuts for peptides[6]).

Table 3. Superatoms representing organic fragments (open sets).

Acyclic Hydrocarbons	
CHK	Alkyl, Alkylene
CHE	Alkenyl, Alkenylene
CHY	Alkynyl, Alkynyline
Carbocyclic Systems	
ARY	Aryl (carbocyclic system, optionally fused, containing at least one benzene ring or quinoid moiety)
CYC	Cycloaliphatic ring system (optionally fused, no benzene ring)
Heterocyclic Systems	

---

[6] Amino acids (peptide) shortcuts are described in chapter 2.1.4 .

HEA	Heteroaryl (aromatic, monocyclic)
HET	Heterocycle (nonaromatic, monocyclic)
HEF	Heterocycle (fused)

Since some of these superatoms are rather special it is necessary to provide some examples. The definitions for the carbon chains are obvious

- CHK is a carbon chain containing only single bonds, *e.g.* methyl, ethyl, t-butyl
- CHE is a carbon chain containing at least one double bond (no triple bonds), *e.g.* ethenyl, propenyl
- CHY is a carbon chain containing at least one triple bond, *e.g.* ethinyl, propinyl.

ARY (aryl) is specified as a carbocyclic system which contains at least one benzene ring or a quinoid moiety, while CYC (cycloaliphatic ring system) comprises all the other carbocycles. Some examples for ARY and CYC are given in Figure 15. It should be emphasized, that the distinction between ARY and CYC is based on the chemical notation of classical aromaticity – both ring systems can be isolated or fused.
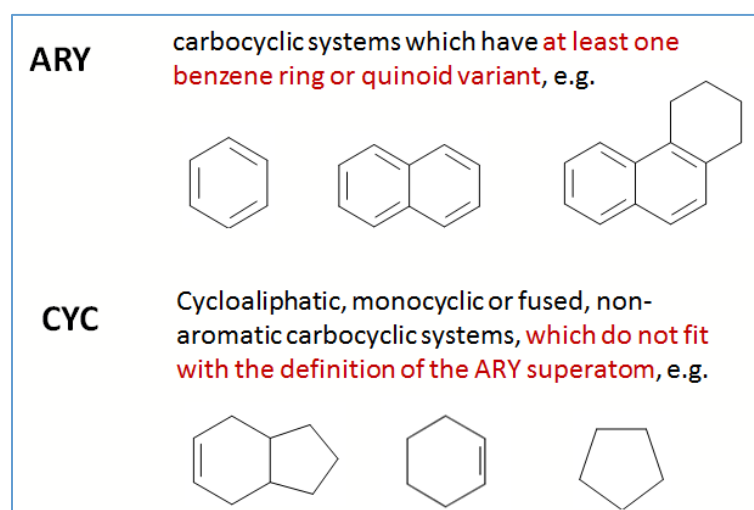

Figure 15. Examples of carbocyclic structures which correspond to the superatoms ARY and CYC.

As chemists know very well aromaticity may be extended to heterocyclic compounds. Derwent has defined heteroaryls as monocyclic five-membered rings with two double bonds or fully unsaturated six-membered rings which have at least one heteroatom in the ring. The corresponding superatom is called HEA. All other isolated monocyclic heterocyclic compounds belong to the superatom HET. Finally, all fused heterocycles belong to the superatom HEF. As a consequence HEF may also contain fused heteroaryls. Obviously, Derwent has applied different segmentations for carbocyclic and heterocyclic compounds. Some examples for heterocyclic compounds are shown in Figure 16.
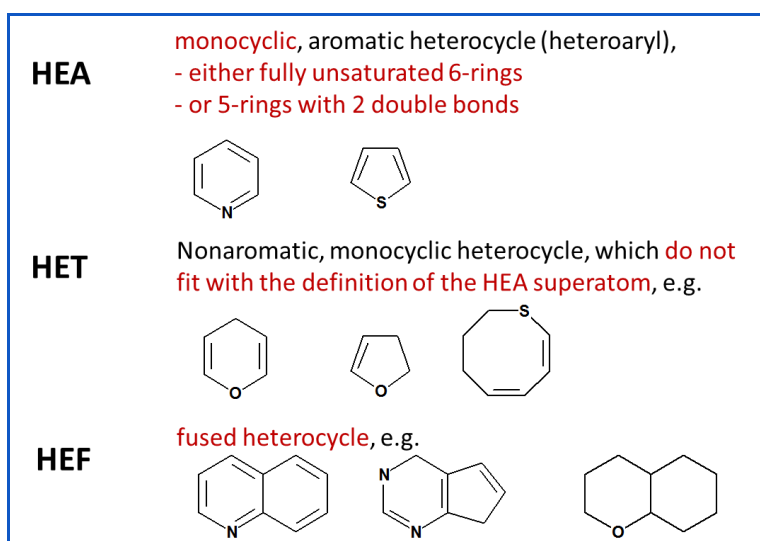
Figure 16. Examples of heterocycles which correspond to the superatoms HEA, HET, and HEF.

Halogens (HAL, in STN queries: X) and metals (MX, in STN queries: M) represent another set of superatoms, a closed set of specific chemical elements (see Table 4). Metals are further divided into five subgroups according to the organization of the Derwent chemical fragment codes (CPI codes): AMX (alkali and alkaline earth metals), A35 (metals from groups 13 – 15, formerly groups III A to V A metals), TRM (transition metals), LAN (lanthanides), and ACT (actinides).

Table 4. Superatoms representing halogens and metals (closed set).

Halogens	
HAL	Halogens
Metals	
MX	Metal (any)
AMX	Alkali and alkaline earth metal
A35	Group 13 - 15 Metals (IIIa-Va)
TRM	Transition Metals
LAN	Lanthanides
ACT	Actinides (including Actinium)

The definitions of halogens and all metal superatoms together with the corresponding section of the CPI code are shown in Table 5 (see also Figure 12). It should be noted that there are some differences with respect to STN standards:

- The halogen group does not contain Astatine (At); this belongs to the superatom ACT.
- Lanthanides do not contain La; instead this is included in the superatom TRM.
- The Alkali and alkaline earth metal group (AMX) does not include the radioactive elements Fr and Ra. These elements are included in ACT.

Table 5. Definition of Derwent metal and halogen superatoms.

Node	Definition	Elements	#	CPI Code
A35	Group III A-V Metal	Al, Ga, In, Tl; Ge, Sn, Pb, Sb, Bi	9	A3
ACT	Actinides (incl. Actinium)	Po; At; Fr; Ra; Ac; Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, ...(>92)	19+	A8
AMX	Alkali and Alkaline Earth Metals	Li, Na, K, Rb, Cs; Be, Mg, Ca, Sr, Ba	10	A1 + A2
LAN	Lanthanides	Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu	14	A7
TRM	Transition Metals	La, Sc, Y; Ti, Zr, Hf; V, Nb, Ta; Cr, Mo, W; Mn, Tc, Re; Fe, Ru, Os; Co, Rh, Ir; Ni, Pd, Pt; Cu, Ag, Au; Zn, Cd, Hg	30	A4 + A5 + A6
HAL	Halogens	F, Cl, Br, I	4	C0
	Other Elements	B; C, Si; N, P, As; O, S, Se, Te	10	B + C

Issue 5: metal superatoms A35, ACT, AMX, LAN and TRM do not work correctly within ring systems and lead to 0 hits. Possible Workaround: Use generic node M (any metal) within rings.

Additional superatoms are defined by specific chemical properties (see Table 6), *e.g.* the superatom ACY represents the functional group acyl. It is only used when the patent explicitly states the term "acyl" and is meant to cover any acid residue produced by removing the OH group from an acid such as alkyl/arlcarbonyl, alkyl/aryl sulphonyl, alkyl/aryl phosoporoyl etc. If a markush definition is e.g. arylcarbonyl the analyst would index the more precise ARY-CO1* etc. The ACY superatom has a default Valency of 1 and never appears as a divalent group.

All other superatoms are used for an unknown open set of chemical nodes or groups. DYE represents a chromophore or fluorescent group, POL is a macromolecule residue, PEG is a polymer end group, and PRT a protecting group.

Table 6. Other superatoms which are defined by a property.

Other Superatoms (defined by property)	
ACY	Acyl
DYE	Chromophore or fluorescent group
POL	Macromolecule residue (defined very broadly: includes polymers, polypeptides, antibodies, enzymes, DNA, polysaccharides, etc.)
PEG	Polymer end group
PRT	Protecting group
XX	Unspecified atom or group, excluding hydrogen
UNK	Unspecified atom or group (unknown)

The superatom UNK (undefined) is the most general superatom in this group and describes any atom or group including hydrogen. Similarly, the superatom XX describes any atom or group excluding hydrogen. In the early 1990's it has been decided by Derwent that there was no need for two such general nodes. Hence, the superatom UNK was no longer used and has been replaced by the superatom XX.

- XX is used when there is definitely a substituent, *e.g.* substituted alkyl would be indexed as CHK-XX, while CHK-O-XX presumably represents an ether group, some type of ester, possibly a peroxy group or a hydroxylamine (but not a hydroxyl alkyl group)
- XX is used when a group is described as substituted but no examples of the substituents are given, *e.g.* groups described as optionally substituted are shown unsubstituted and with XX attached (optionally substituted aryl would be indexed as both ARY and ARY-XX).
- The associated attributes are fixed defined to support maximum retrieval, i.e. match level is always ANY and the node type is always ring/chain.
- Since the early 1990's instead of UNK it is necessary to index two variations: XX and H.

The superatoms from Table 3 (organic fragments) and Table 4 (halogens and metals) form a hierarchy of nodes as shown in Figure 17. It should be noted that the other superatoms (see Table 6) as well as the peptide superatoms are not integrated in this hierarchy.[7]
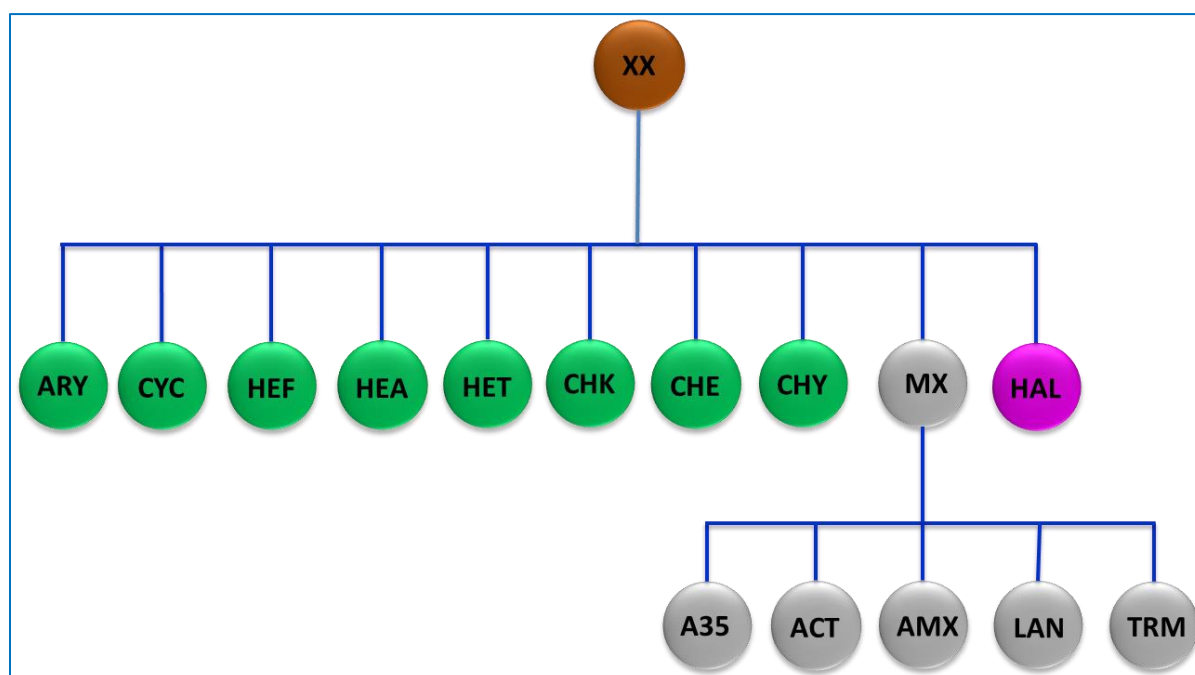

Figure 17. Hierarchy of superatoms.

---

[7] The identical hierarchy is applied at Questel for the MMS database.

2.1.4. Amino Acid Shortcuts

Peptides were indexed with amino acid shortcuts[8] (s. Table 7) until the year 2000. Since then peptides are indexed with their full chemical structure. In case of the amino acid shortcuts directional bonds[9] were used to indicate the C- and N-terminal part of the amino acids.

Table 7. List of 30 amino acid shortcuts. Non-standard amino acids are indicated by '*'.

ABU aminobutyric acid (*)	**GLY** glycine	**ORN** ornithine (*)
ALA alanine	**HCY** homocysteine (*)	**PHE** phenylalanine
ARG arginine	**HIS** histidine	**PRO** proline
ASN asparagine	**HSE** homoserine (*)	**SAR** sarcosine (*)
ASP aspartic acid	**ILE** isoleucine	**SER** serine
ASU aminosuberic acid (*)	**LEU** leucine	**STA** statine (*)
CYS cysteine	**LYS** lysine	**THR** threonine
GLN glutamine	**MET** methionine	**TRP** tryptophan
GLP pyroglumatic acid (*)	**NLE** norleucine (*)	**TYR** tyrosine
GLU glumatic acid	**NVA** norvaline (*)	**VAL** valine

A peptide must be at least three amino acid moieties long to be indexed with shortcuts (s. Figure 18). Amino acid shortcuts are represented as $*- HN - C(-R)^{10} - C(=O) -*$ if in the middle of the chain. An amino acid at the end of the chain automatically includes the OH at the C terminus and at the N terminus an additional implicit H is included. In order to convert an acid into an ester a regular single bond from the C terminus to an Oxygen atom has to be drawn (e.g. ALA-O-CH_3 for the methyl ester of alanine). Non peptide substituents can be attached at the shortcuts via regular single bonds. Modifications of the N-terminal part are achieved by drawing a substituent directly at the respective terminal shortcut. E.g. Et – ALA → SER → THR describes an N-ethyl substituted alanine bonded with C bonded to the amino group of serine, etc. ALA – SER – THR is shown in Figure 18. For any modifications to non-terminal amino acids it is required to draw out the respective amino acid in full and attach the desired substituent(s).

---

[8] Amino acid shortcuts are described by Clarivate Analytics as peptide superatoms.
[9] Directional bonds are indicated by an arrow. They are not yet implemented in DWPIM.
[10] R stands for the different substituents of the amino acids listed in Table 7, e.g. R=CH_3 for ALA.
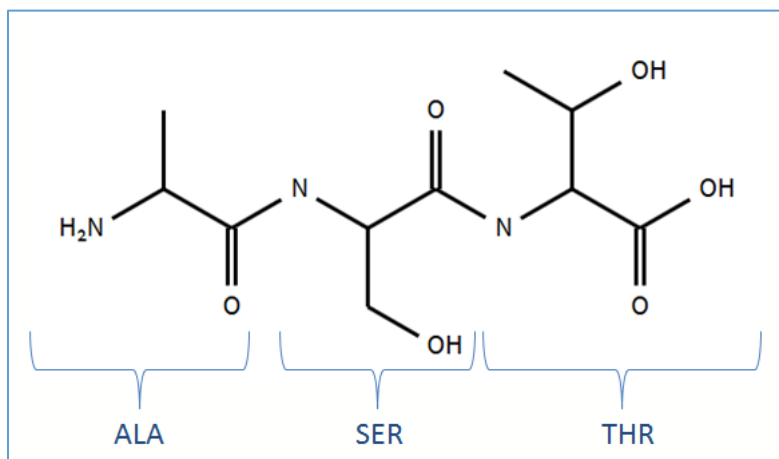
Figure 18. Example of a simple peptide structure consisting of ALA → SER → THR.

Issue 6: The search for peptides using amino acid shortcuts is not supported.

2.2. Variable Groups

A variable group is a group which consists of different organic or inorganic fragments. Variable groups are called R-groups in queries and G-groups in indexed file structures. In the Markush file structures variable groups can be nested, *i.e.* a G-Group may contain another G-group. In file structures nesting is restricted to 4 levels. A typical G-group from DWPIM is shown in Figure 19. The first variation is simply a hydrogen atom H (no substituent), the second is the HEA superatom substituted by another G-group, and the last two examples are pyrrole rings with different positions for the attachment to the father group (in this case the master structure).
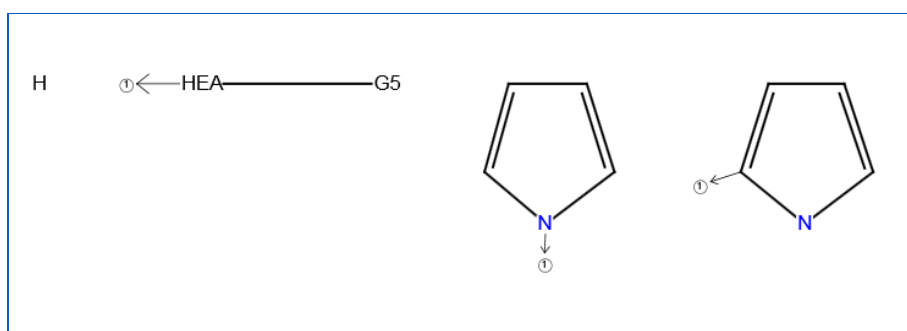

Figure 19. G-group fragments from a Markush record.

2.3. Chemical Bonds

2.3.1. Overview

Nodes are connected by bonds. They are characterized by 2 attributes: bond type, and bond value.

Bond types:

- **Ring/Chain** - system default value that allows you to retrieve structures containing the selected bond in either cyclic (ring) or acyclic (chain) systems in the hit structures.
- **Ring** - allows you to specify that the selected bond must be part of a cyclic (ring) system in the hit structures.
- **Chain** - allows you to specify that the selected bond must be part of an acyclic (chain) system in the hit structures.

Bond values:

- **Exact/Normalized** - search results match answers with either the exact bond value drawn or with normalized bonds. Note: a single bond with bond value exact/normalized is interpreted as single exact or normalized while a double bond with bond value exact/normalized is interpreted as double exact or normalized, respectively.
- **Exact** - search results exactly match the bond value that is drawn.
- **Normalized** - search results match only normalized bonds.
- **Unspecified** – search results match to all bond values.

In Table 8 an overview of all chemical bonds available in the STN structure editor together with the assignable bond types and values is shown.

Table 8. Chemical bonds available in the STN structure editor[11].

Chemical bond	Bond representation	Allowed Bond type	Allowed Bond value
Single	———	ring/chain, ring, chain	exact/normalized, exact, normalized
Double	≡≡≡	ring/chain, ring, chain	exact/normalized, exact, normalized
Triple	☰	ring/chain, ring, chain	exact/normalized, exact, normalized
unspecified (single, double or triple)	- - - - - -	ring/chain, ring, chain	exact/normalized

Issue 7: The stereochemical information of bonds is ignored for the search. However, in the structure display the stereo bonds are drawn.

When performing structure searches the bonding conventions of the system need to be taken into account. In this regard, two important aspects should be considered:

- Normalization in rings (see 2.3.4)
- Tautomerism (see 2.3.5)

---

[11] The stereochemical bond types (upward stereo, downward stereo, double upward stereo, double downward stereo, E/Z double bond) are ignored for search in DWPIM. Therefore these bonds are omitted in the table.

2.3.2. General STN Bond Conventions

In the following the bond conventions used for the STN structure editor, for the STN query and for the indexed data are described (s. Table 9).

- **Structure Editor**: bond values in the structure editor
- **STN-Convention**: adoption of bond values from the structure editor to the STN convention. Note that "exact" from structure editor is translated to single exact or double exact, respectively, depending on the drawn bond type.
- **Indexed bond**: bond value which matches to the value defined by the STN-convention. This shows that *e.g.* single exact can only match to a single bond whereas a single/normalized bond can either match to a single or to a normalized bond in the indexed structure.

Table 9. Overview on bond convention and translation.

Bond representation in STN Structure Editor		STN-Convention	Indexed bond in database
———	**exact**	single exact	single
———	**exact/normalized**	single/normalized	single or normalized
═══	**exact**	double exact	double
═══	**exact/normalized**	double/normalized	double or normalized
≡≡≡	**exact**	triple exact	triple
——— ═══	**normalized**	normalized	normalized

Example Cyclohexane:

- Draw cyclohexane in STN structure editor. Default bond value is **exact/normalized**
- Cyclohexane structure is automatically converted to STN bond convention **single/normalized**
- Query matches to records in database with indexed bonds **single or normalized** (*i.e.* cyclohexane derivatives as well as benzene derivatives are retrieved)

2.3.3. Aromatic Bonds
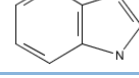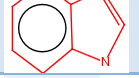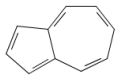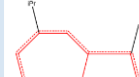
For the definition of aromaticity Hückel's rule applies: a cyclic compound is aromatic if it is planar and if it has $(4n+2)$ π-electrons. In the case of a heterocycle the free lone pair(s) of heteroatom(s) can participate in the aromaticity. In STN only rings up to a size of 10 atoms are considered as aromatic. Aromatic compounds with an even number of atoms are indexed with non-localized (normalized) bonds while aromatic compounds with an uneven number of atoms are indexed with localized single and double bonds. These differences are taken into account in the structure editor which automatically defines normalized or exact/normalized bonds, respectively.

2.3.4. Normalized Bonds in Ring Systems

Normalized bonds are used in ring systems with an even number of atoms containing alternate double and single bonds.

It is important to note that this definition is not fully consistent with the definition of aromatic compounds. As depicted in Table 10, anti-aromatic compounds such as cyclooctatetraene contain normalized bonds while aromatic 5-membered heterocycles such as pyrrole are defined with localized single and double bonds. An exception to the general rule is the cyclopentadienyl anion and the cycloheptatrienyl cation which are indexed with normalized bonds. In these cases the default bond value in the structure editor (exact/normalized) should be changed to normalized in order to find exclusively this type of structures which are indexed with normalized bonds.

Table 10. Bond types in rings.

Example	Search Structure	Default Bond Value in Structure Editor	Indexed Bond in Database	DWPIM Display	MLE CLASS
Benzene		normalized	normalized		ARY
Cyclooctatetraene		normalized	normalized		CYC
Cyclobutadiene		normalized	normalized		CYC
Pyrrole		exact/normalized	single and double bonds		HEA
Indole		normalized; exact/normalized	normalized; exact		HEF
Cycloheptatrienyl cation (tropylium ion)		exact/normalized[12]	normalized		CYC
Azulene		normalized	normalized; exact		CYC
Cyclopentadienyl-anion		exact/normalized[13]	normalized		CYC

2.3.5. Normalized Bonds in Tautomeric Systems

Normalized bonds are used when the chemical structure contains atoms which fall under the "tautomer rules". It is important to note that tautomers as defined in structure searching do not always correspond with the chemists' definition of tautomers. A noteworthy example of deviating tautomer definition is the keto-enol tautomerism of aldehydes and ketones.

In general, tautomeric structures are described by two compounds in an equilibrium state:
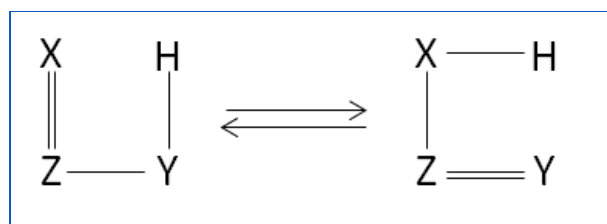
Figure 20. Description of tautomeric structures.

---

[12] Bond value must be changed to normalized in order to find records containing only cycloheptatrienyl cations. No charge to be applied.
[13] Bond value must be changed to normalized in order to find records containing only cyclopentadienyl anions. No charge to be applied.

Where Z can be: B, C, Si, N, P, As, S, Se, Te, F, Cl, Br, I
X and Y can be: O, S, Se, Te, N

These rules apply for the following scenarios:

- The tautomeric bonds can be part of an acyclic chain or a ring system or part of both.
- The end points, X and Y, can be in adjacent rings of a fused ring system. However, a nitrogen atom that is located at the fusion point in such a system cannot take part in tautomerization.
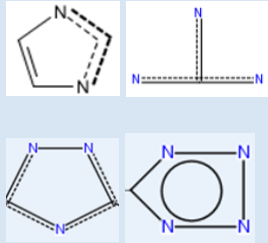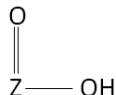
The degree of normalization of bonds depends on the type of the tautomeric system and on where it is positioned within the molecule:

- Type of atoms used for X, Y and Z
- Aliphatic or ring system
- 2n atom ring systems with alternating double bonds
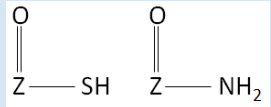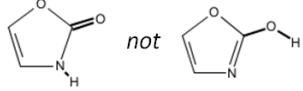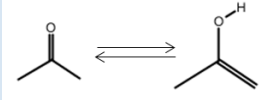
2.3.5.1. Overview of Bond Normalization in Tautomers

Table 11 summarizes the most important cases for bond normalization in tautomers including the corresponding default bond values in the structure editor as well as the bond values of the indexed bonds.

Table 11. Bond normalization in tautomers.

Case	Bond Value in Structure Editor	Indexed Bond in Database	Examples
1) X and Y are both N	exact/ normalized	normalized	Imidazoles, guanidines, 1,2,4 triazoles, tetrazoles
2) X and Y are the same but not N	normalized	single and double bond[14]	Carboxylic acids

---

[14] STN conventions are preserved by automatic transforms

Case	Bond Value in Structure Editor	Indexed Bond in Database	Examples
3) X and Y are different (chains)	Normalized (for thioacids), exact/normalized (for amides)	Single and double bond[15]. The double bond is placed preferentially on the first atom in the sequence: O>S>Se>Te>N	Thioacids, amides
4) X and Y are different (rings with uneven number of atoms)	exact/ normalized	single and double bond. The double bond is placed preferentially on the first atom in the sequence: O > S > Se > Te > N	2-oxazolidinones[16]
5) Z and Y are C (keto-enol tautomerism)	C=C: exact C-OH: exact/ normalized	Single and double bond, drawn in the keto form.	Acetone tautomerism

Practical Consequences:

Case 1.   X and Y are both N: for acyclic systems such as guanidine the STN bonding conventions are preserved. The structure can be searched with exact/normalized bonds (default setting in structure editor) and as a result hits with normalized bonds are retrieved.

Case 2.   X and Y are the same but not N: Carboxylic acids as well as related acid groups cannot be searched using the default setting in the structure editor[17]. The default of those groups is normalized bonds which do not match to localized double and single bonds of the indexed structures. Therefore the bond value has to be adjusted to e or e/n whereby e/n is preferred because it also ensures matching in CAS databases. Another possibility is to draw the respective groups with atom locked heteroatom instead of an Hydrogen. This procedure has the advantage that the default for bond value there is already e/n so that no manual changes are required.

Case 3-5.   The rule for double bonds needs to be considered when drawing the query structure.

2.3.5.2. Special Case of Tautomerism and Aromaticity

Ring systems with 2n atoms containing n alternating double bonds constitute a special case. In these cases normalized bonds take priority over tautomerism conventions. A typical example is the 2-

---

[15] STN conventions are preserved by automatic transform

[16] If "O" is listed in a G-group the representation may differ, *e.g.:*

 G1: O, Hal, CHK
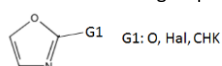
[17] The following groups are affected: -COOH, -CO₂H, -COSH, -CSSH, -CS₂H, -OPO₃H₂, -PO₃H₂, -PO₂H, -OSO₃H, -SO₃H, -SO₂H, -SeO₃H, -SeO₂H.

Pyridinone/2-Pyridinol tautomerism as depicted in Figure 21. The representation with normalized bonds (2-Pyridinol) is preferred over the tautomer with localized single and double bonds. It is important to note that 2-Pyridinol would not be retrieved by searching for the structure of 2-Pyridinone since the exocyclic C-O bonds are not normalized in these structures.
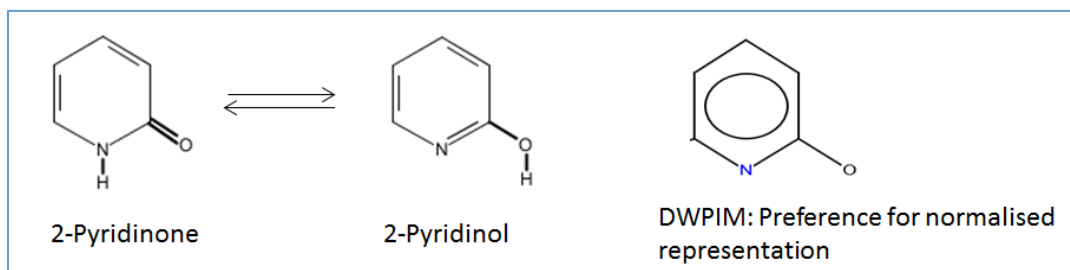
Figure 21. Example for preferred tautomer representation for aromatic rings.

2.3.5.3. Special Case of Keto-Enol Tautomerism

Keto-enol tautomerism applies to compounds of the type:
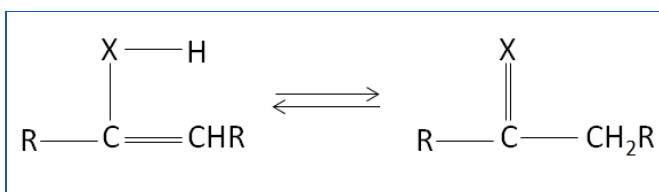
Figure 22. Description of keto-enol tautomerism.

where X can be: O, S, Se, Te.

In case of keto-enol tautomers no bond normalization takes place. An organic compound containing a group capable of keto-enol tautomerism is preferentially indexed and represented in the keto form.
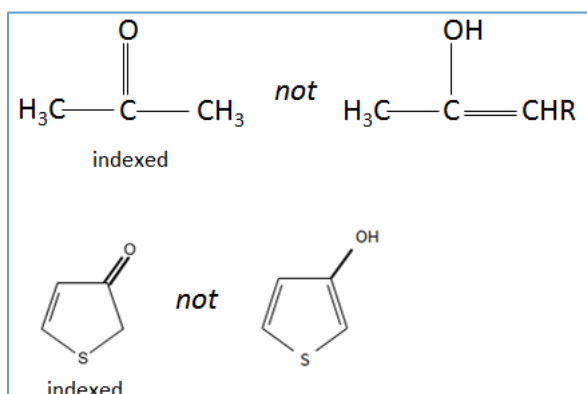
Figure 23. DWPIM indexing of keto-enol tautomers.

2.3.5.4. Special Case of Vinyl Amine - Imine Tautomerization

In contrast to keto-enol tautomerism an imine group of an alicyclic ring is converted to the enamine form provided there is at least one hydrogen atom present on an adjacent carbon for migration.
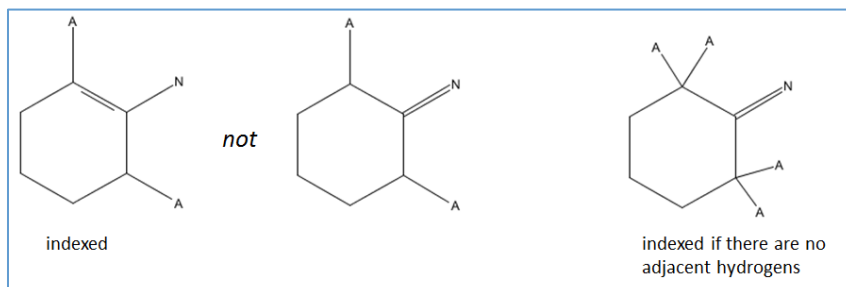
Figure 24. Indexing of imine groups.

For aliphatic systems no such rule is applied for vinyl amine – imine tautomerization. In these cases the structures are indexed as drawn in the patent document.

2.3.5.5. Special Case of Quinoid Systems

In general, bonds in quinoid systems are not normalized (Figure 25).
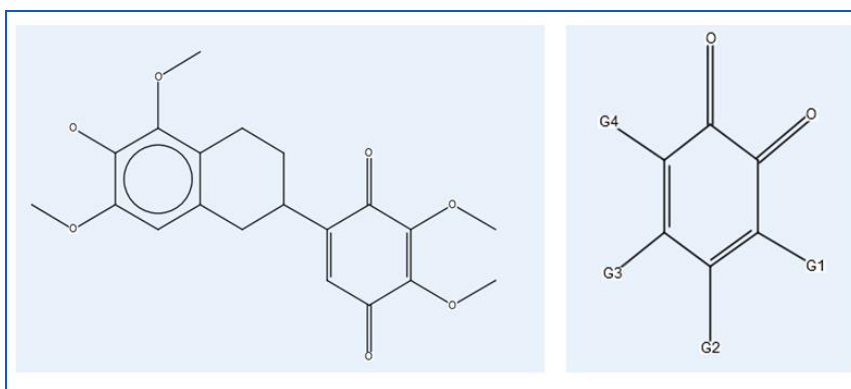
Figure 25. Examples for the representation of quinoid systems (left structure: 1,4 benzoquinone type; right structure: 1,2 benzoquinone type).

An exception is where further fused rings make normalization possible (Figure 26). This is for example the case with anthraquinone where two benzene rings are fused with a para quinone moiety.
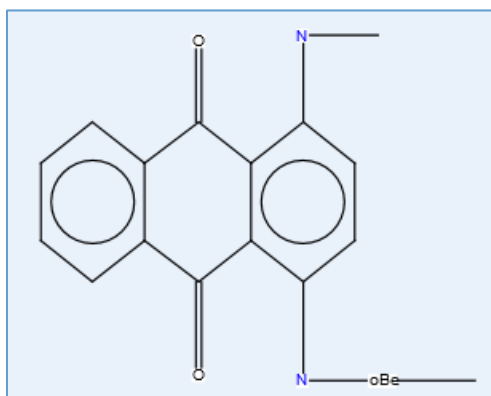
Figure 26. Example for the representation of an anthraquinone.

2.4. Node Attributes

Attributes are additional parameters describing certain properties of chemical nodes. There are four types of attributes:

- Atom attributes, *e.g.* charge
- Superatom attributes, *e.g.* chain length for a carbon chain
- Peptide attributes, *e.g.* configuration
- Polymer attributes, *e.g.* monomer/condensate.

A list of all attributes used for indexing by Clarivate Analytics is given in Table 12.

Table 12. Compilation of attributes of atoms, superatoms, peptides, and polymers.

Type	Attribute	Description	Search
Atom Attributes	abnormal mass		yes
	abnormal valency		yes
	deuterium or tritium count[18]		no[19]
Superatom Attributes	carbon chain	straight/branched[20]; low/mid/high	yes
	ring	monocyclic/fused; saturated/unsaturated	yes
Atom or Superatom Attributes	charge or delocalized charge		yes
	free sites	max. number of free sites allowed at this position[21]	yes
	multiplier	max. degree of substitution which is possible for a given generic term; default value is 1.	no

---

[18] The indexed deuterated and tritiated attributes have been converted to D or T atoms for search purposes on STN (e.g. indexing CHK-O$^{(D=1)}$ is converted to as CHK-OD for STN. See also chapter 2.4

[19] See chapter 0

[20] Branched chains start with C=4

[21] In STN it is not possible to specify directly the number of free sites. It is necessary to specify the number of non-hydrogen substituents. Both concepts are complementary.

Type	Attribute	Description	Search
	numbering	link between node and text (only	no
Peptide Attributes	configuration	D or L	no
	position	position of substituents	no
Polymer Attribute	role of node in the polymer	MC (monomer/condensate); XL (cross linker); EG (end group); DE (modifier/derivative); GM (grafting	no

2.5. Other Structural Features

2.5.1. Variable Points of Attachment

A Variable Point of Attachment (VPA) specifies multiple positions on a ring where an atom can be attached. The attached substituent can be:

- an atom, *e.g.* N
- a group of atoms, *e.g.* SO_2Me
- a superatom, *e.g.* Hy
- a G-Group

On the Markush structure the variable points of attachment are displayed by dotted lines. An example for a representation of VPA's in a ring system is given in Figure 27.
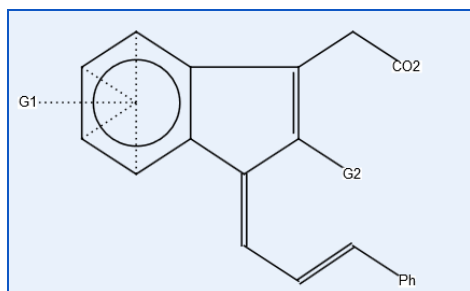
Figure 27. VPA representation for MCN-8337-26301.

Please note that for the structure search the bond type of the bond which connects the VPA group to the main structure is ignored which means it is always set to single bond.

The type of VPA representation (various dotted lines to several positions in the molecule) as shown in Figure 27 was only used for Markush structures in the time period between 1961 and 1998.

2.5.2. Repeating Groups

A repeating group (rpg) within a Markush substructure allows the atoms in the group to repeat a specified number of times. The repeating group is characterized by square brackets and the allowed ranges. It can consist of atoms, superatoms and G-groups (Figure 28).
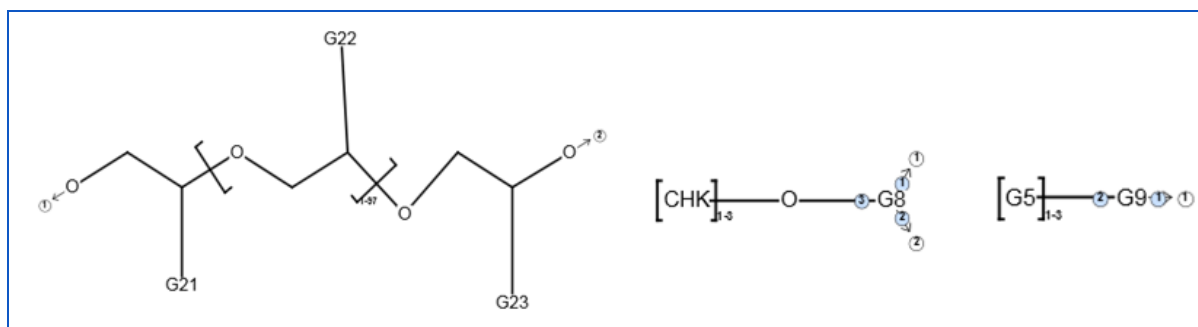
Figure 28. Representation of Markush structures containing repeating groups.

Issue 8: Within rings only one repeating group (rpg) per ring is allowed. In case of two or more repeating groups within the same ring incorrect results are obtained. There is no workaround other than drawing out all possible combinations for every additional repeating group.

Example:
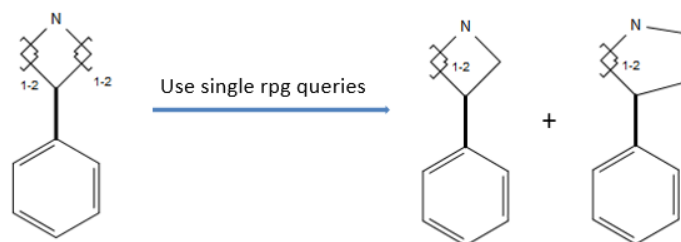
2.5.3. Additional Provisos

Conditional Logic

A Markush structure in a patent may include conditional provisos. For indexing purposes these structures have either to be split into several separate Markush structures or additional G-groups have to be introduced in order to take the various possibilities into account. The procedure is described in the example given in Figure 29. Structure A is a fictitious patent Markush structure which includes conditional logic with regards to the substituents Q and R: If Q is trifluoromethyl then R is F.
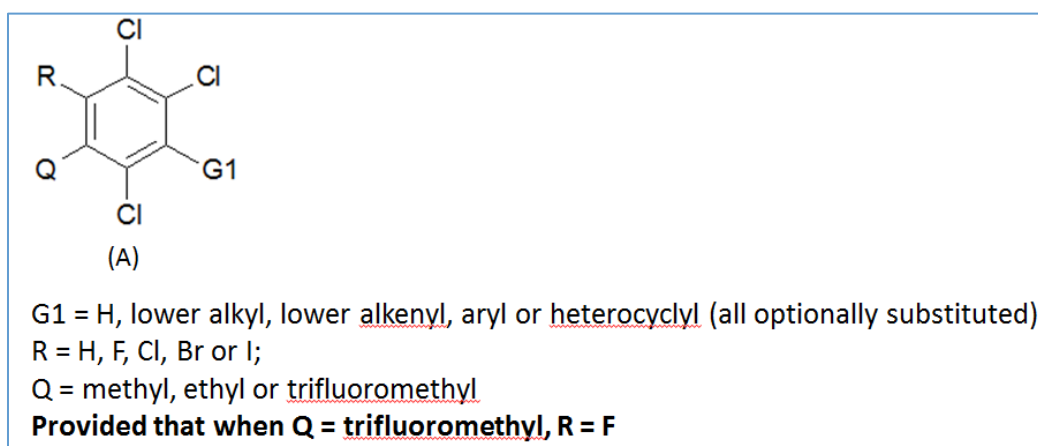
Figure 29. Fictitious patent Markush structure containing conditional provisos (Q, R).

For the representation of the conditional provisos in DWPIM two options exist (Figure 30):

- Option 1: Create two Markush structures - (i) Q = CF$_3$ and R = F, (ii) Q = methyl or ethyl and R = H, F, Cl, Br or I.
- Option 2 (preferred): create two variations of a single Markush structure - take out the portion of the ring containing R, Q and the two ring C atoms as a variable group G2 with (i) one fragment including CF3 and F according to the proviso and (ii) one fragment with the other variations where G3 = R and G4 = Me, Et (Figure 30).
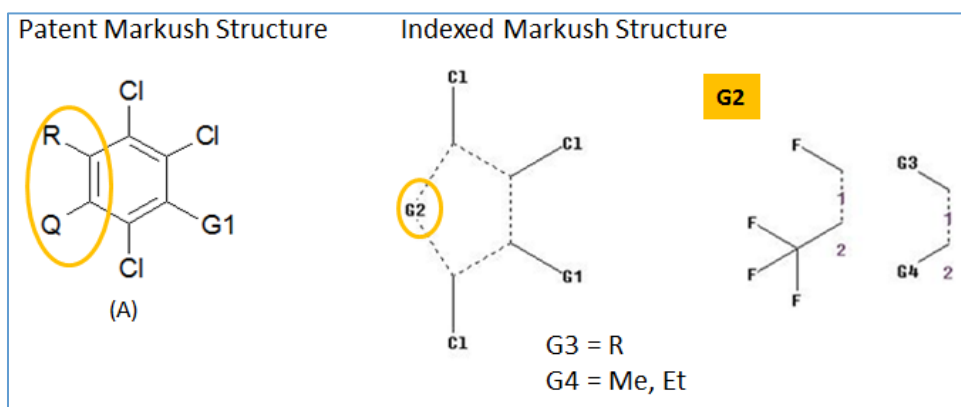
Figure 30. Example for the representation of conditional logic in a Markush structure.

Shared Variables to Form a Ring

Shared variables are sometimes used in Markush structures to cover ring formation. In the fictitious patent example given in Figure 31 the variables R and Q may form a ring which is defined by the structure fragments listed under R + Q.

G1 = H, lower alkyl, lower alkenyl, aryl or heterocyclyl (all optionally substituted);
R = H, F, Cl, Br or I;
Q = methyl, ethyl or trifluoromethyl; or
R+Q = -CH2CH2CH2-, -CH2CH2CH2CH2- or –CH=CH-CH=CH-.
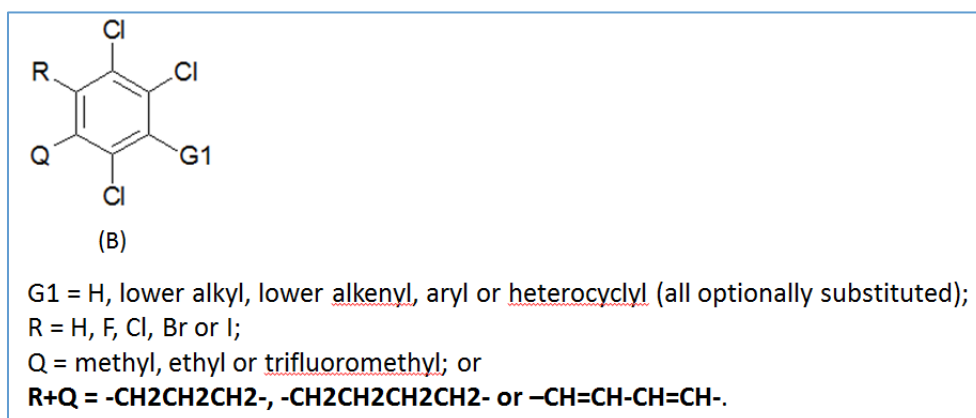
Figure 31. Fictitious patent Markush structure containing shared variables (R, Q).

For the representation of cases like this two options exist:

1. Option 1: Create three Markush structures, one for R and Q separately, one for the cycloalphatic fused rings (*i.e.* –CH2CH2CH2–, –CH2CH2CH2CH2CH2–) and one for the fused benzene (–CH=CH–CH=CH–). It is also possible to combine the latter two in one Markush structure.

- Option 2 (preferred): create several variations of a single Markush structure - take out the portion of the ring containing R, Q and the two ring C atoms as a variable group G2 where R and Q are separately covered by one fragment (defined as G3 and G4) and the three fused rings are covered by three separate fragments (Figure 32).
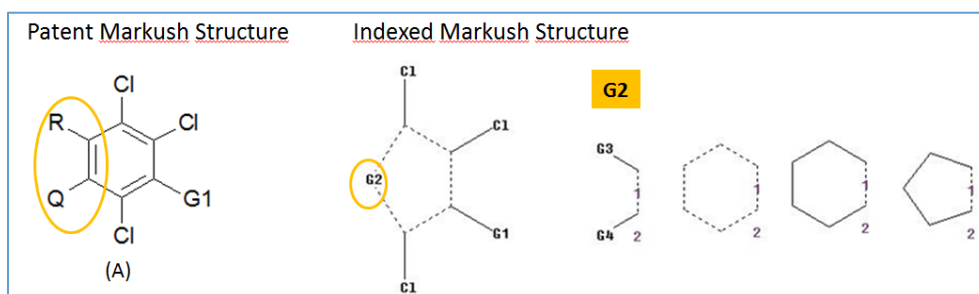
Figure 32. Example for representation of shared variables in one Markush structure.

NULL in G-groups

If rings of different ring sizes are indexed by using one Markush structure it usually comprises the smallest ring size in the given series including a G-group which covers the larger rings. A G-group with NULL logic indicates the absence of a node and therefore leads to direct bond formation (*i.e.* ring contraction) is not used by Clarivate Analytics. In the example given in Figure 33 indexing of piperidine and cyclopentane in a single Markush fragment could be achieved by defining a cyclopentane ring with a G-group containing the fragments –CH2– and –CH2NH–.
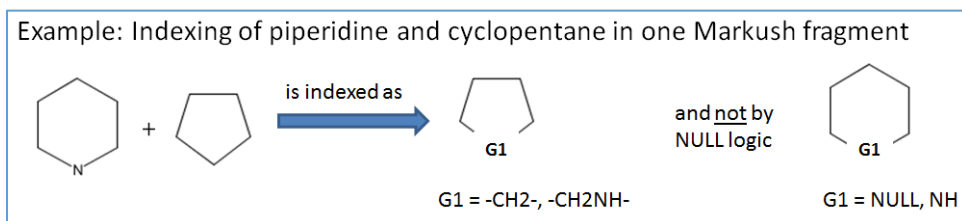
Figure 33. Representation of NULL logic in Markush structures.

3. STN Search Concept

The focus of this chapter lies on the outline of the various possibilities for searching the DWPIM file on STN; in other words the query side is described. In principle, the nodes and bonds in the query must correspond to those in the indexed structures. Some definitions are identical; *e.g.* chemical elements, but others like the generic nodes in the query and the indexed structure could be different.

3.1. STN Structure Building

The STN structure editor can be opened by clicking on the structure icon in the query section of the interface (Figure 34).

Figure 34. Structure editor icon in the STN interface.

In depth information about the general functions of the structure editor can be found in the new STN online helps. For the purpose of this manual only Markush specific topics are described. Some of the concepts that are mentioned in this chapter, *e.g.* "Match Level" and "Element Count Level" will be discussed in more detail in subsequent chapters.

The workflow for assigning Match Level and Element Count Level attribute values to nodes is as follows (Figure 35):

- Right mouse click on a node (specific or generic) in the molecule. The Node Attribute window opens and the default Match Level can be changed.

- For the selection of multiple nodes: Use the selection tool and click on the desired positions in the molecule while holding the Strg key pressed. Blue dots appear at the selected positions. Right mouse click on one of the selected positions opens an interaction window and selections made will apply to all positions.
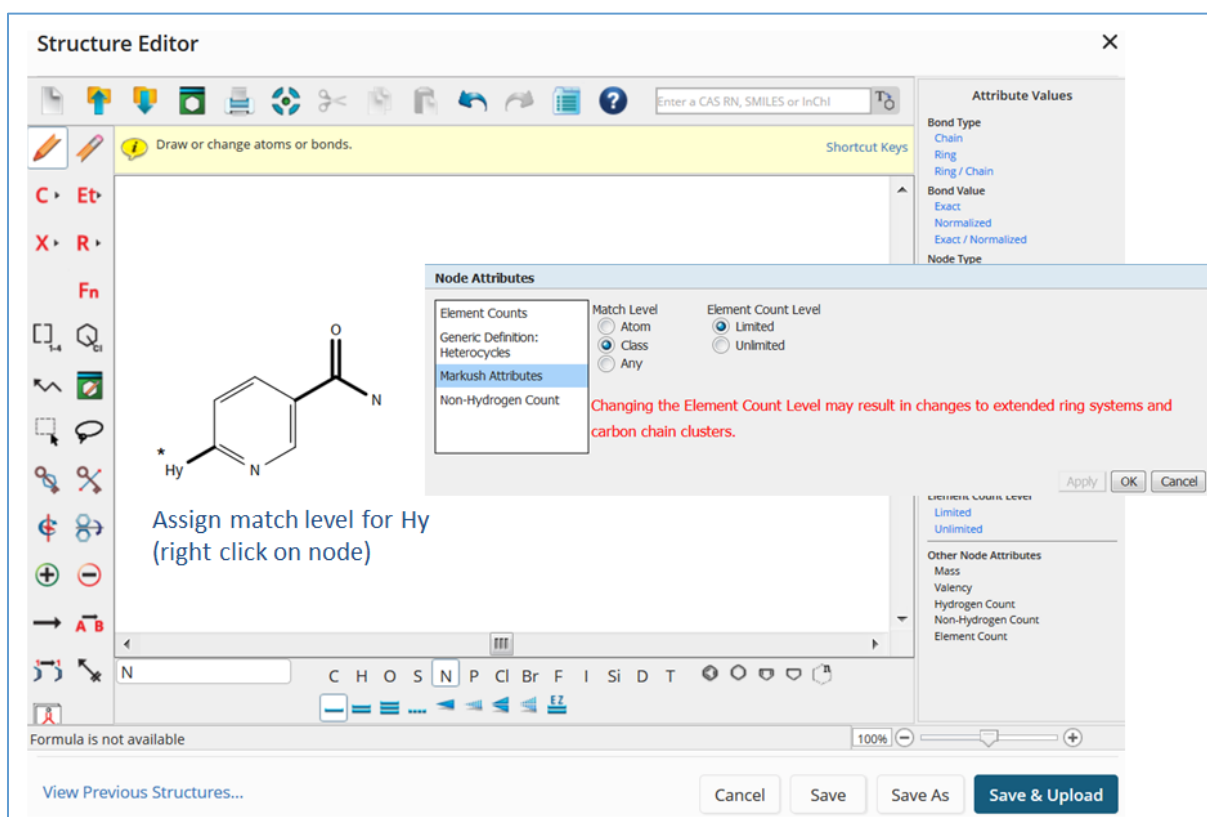
Figure 35. STN structure editor: assigning match level.

In order to change the Element Count Level (*Limited* to *Unlimited*) for generic node(s) it is first necessary to assign an Element Count to the respective node(s). This can be done in the Node Attributes window in the following way (Figure 36):

- Select "Specific" (default is Any)
- Select an element from the list
- Select a count: exact, minimum, maximum or a range between 1 to 255
- Add Element Count
- Make further selections or apply
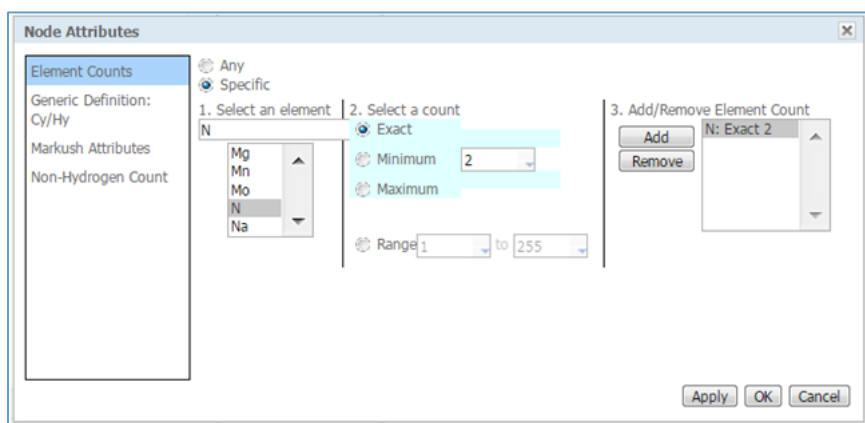
Figure 36. STN structure editor - element counts.

After assignment of an Element Count to a generic node it is now possible to change the Element Count Level from Limited to Unlimited (Figure 37).
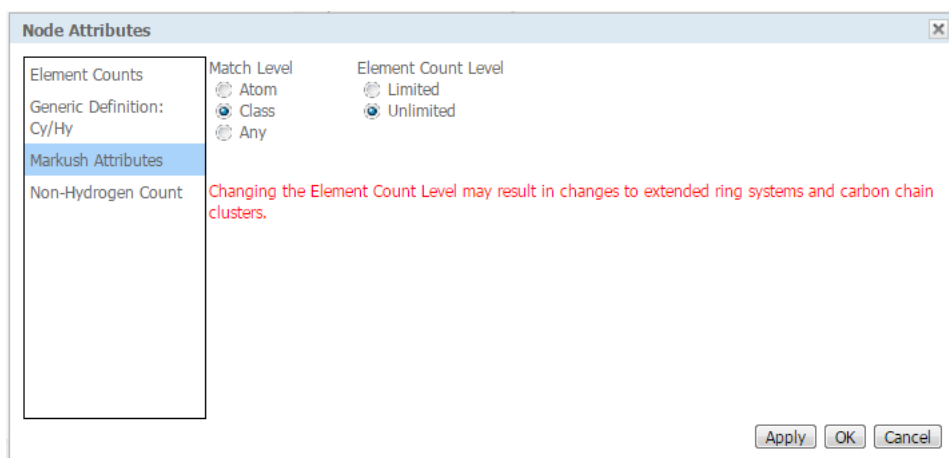
Figure 37. STN Structure Editor – element count level.

While there is no difference in results between ELC *Limited* and *Unlimited* for Match Level ATOM the search results for Match Level CLASS may differ considerably. *Limited* requires an overlap of all the element counts between the query and the structure while *unlimited* ignores the element count(s) of the query. In case of Match Level ANY it is strongly recommended to use ELC Unlimited since otherwise important hits from Match Level CLASS *unlimited* might be lost (see chapter 3.7.2).

Superatoms are integrated in the STN structure editor and are located next to the STN nodes under the icon "X (Variables)" (Figure 38). The superatoms are subdivided in five categories as described in chapter 2.1.3. The superatoms HAL (halogens) and MX (metals) are not separately listed since they match to the corresponding STN nodes X and M, respectively. The superatoms listed under "Miscellaneous" (the individual representatives are shown in Table 6) are only applicable in DWPIM and not in DCR. Superatoms and their associated attributes can be applied to the query structure as described for the STN nodes. A query structure can contain both STN nodes and Superatoms. Please note that the node "Id" is only applicable to CAS databases.
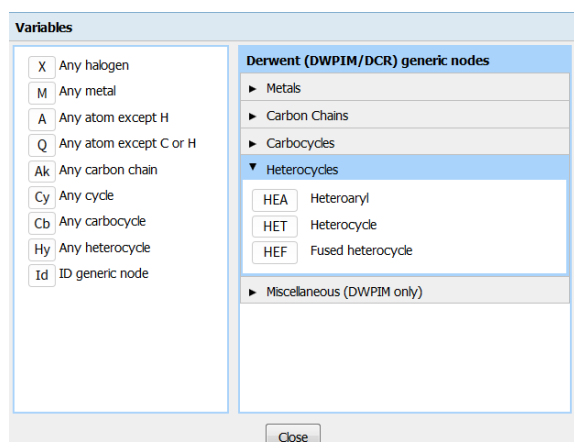
Figure 38. Superatoms in STN structure editor.

3.2.   STN Query Structures

STN Query nodes may consist of

- Atoms (chemical elements)
- STN Shortcuts (different from shortcuts in indexed structures)
- R-groups (variable groups of atoms, STN Shortcuts, STN Generic nodes and user-defined Fragments)
- Attachment points
- Repeating Groups
- STN Generic nodes (different from superatoms in indexed structures).

Atoms

All chemical element symbols are available for searching (s. Figure 12).

STN Shortcuts

STN uses a set of shortcuts for common groups of chemical elements (Figure 39). For searching the shortcuts are replaced by the corresponding group of chemical elements.
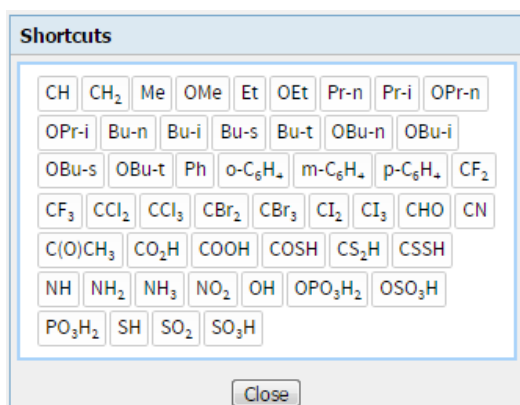
Figure 39. List of STN shortcuts.

R-Groups

R-groups[22] are variable user defined groups which may consist of atoms, STN shortcuts, STN generic nodes, and user-defined fragments. They are used to describe positional variations in Markush queries. Assignment of attributes for an R-group is limited and the scope is defined in the structure editor's preferences.

In order to assign all available attributes user-defined fragments (Fn groups) must be used. These fragments have one or two attachment points and it is possible to assign different attributes to each fragment.[23]

---

[22]   In indexed structures the variable groups are called G-groups.
[23]   When fragments are part of an R-group which is inserted in a ring system it is recommended to adapt the default attribute values to those of the ring system (*e.g.* change "chain" to "ring" or MLE=CLASS as the default for chains to MLE=ATOM as the default for ring systems.

The structure search engine for DWPIM supports the application of nested R-group queries. A simple example of nested R-group query containing two levels where R2 is integrated in R1 is shown in Figure 40.
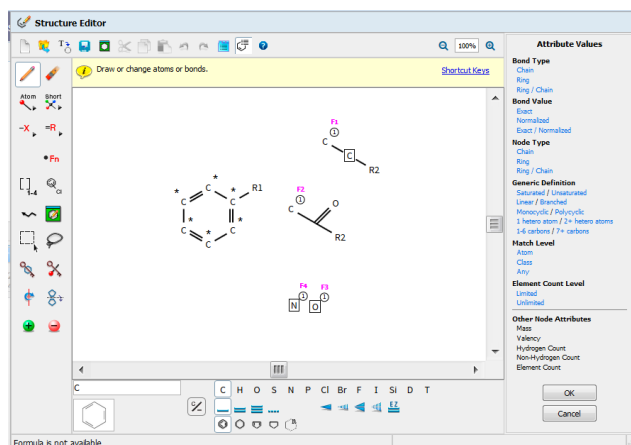
Figure 40. Example of a nested R-Group query in the structure editor. R1=F1,F2 and R2=F3,F4

Even though there is not limit with regard to the nesting level it should be noted that the number of possible combinations defined by respective R-group entries is limited to 1024. Queries which exceed this limit are rejected after submission:

```
=> s l1 sss ful


FULL SEARCH INITIATED 14:23:53
FULL SCREEN SEARCH COMPLETED -  2331830 TO ITERATE


YOUR SEARCH CANNOT BE PROCESSED--PLEASE CHANGE YOUR QUERY STRUCTURE
```

Attachment points and bonds at R-groups

Case 1: If there is an equal number of bonds and attachment points at R-group the upper limit is two. As shown in Figure 41 two the left query structure is accepted while the structure on the right exceeds the maximum number of two connections.
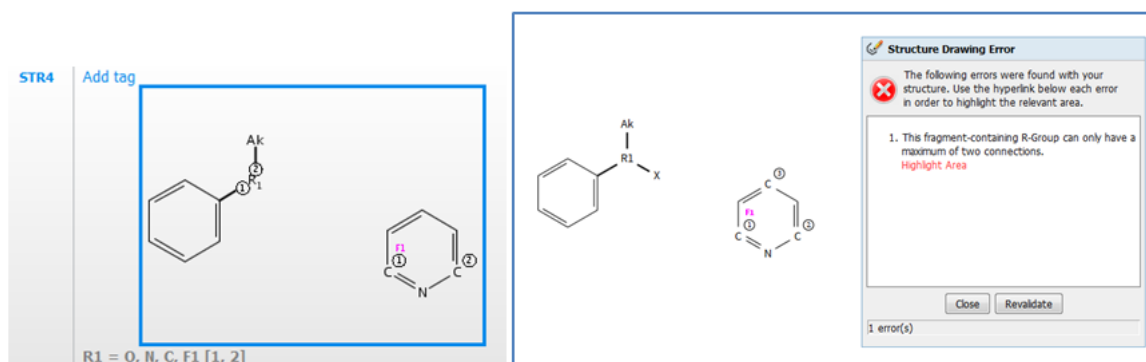
Figure 41. Left side: example of an R-group with two connections (allowed). Right side: Three connections (not allowed)

Case 2: If there is only one bond at the R-group the number of attachment points is limited by a maximum of 20 items allowed per R-group.

R1 = O, N, C, F1 [1, 2]
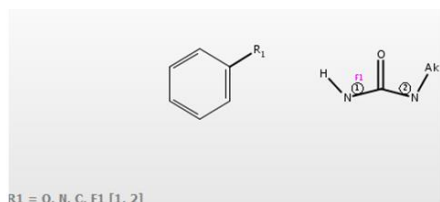
Figure 42. Example of R-group with one connection and two attachment points. A maximum of 17 attachment points would be allowed (limited by the maximum number of items per R-group, which is 20).

Repeating Groups

STN queries may contain a repeating group (frequency variation). Although in most cases this is a rather simple application there are some cases which require a deeper understanding of the search process. In general, a repeating group may contain one or more nodes (specific or generic) and there must be exactly two bonds to nodes outside of the repeating group. The simplest repeating group consists of one node and it can be defined as

$$R_1 \xrightarrow{b_1} [\, A \,]_{a-b} \xrightarrow{b_2} R_2$$

where R_1, R_2 are any possible residues outside the repeating group and A is the node inside the repeating group. Sometimes this is called a link node. R_1, R_2 can be equal or different, the interval a-b is defined as $a \leq f \leq b$, where a ($a \geq 0$) and b are the lower and upper limit and f is the repeating factor running from a to b. b_1, b_2 are external bonds with bond types ring, chain, or ring/chain. This kind of repeating group is most frequently used to specify variable size chains and variable size rings.

A more general definition of a repeating group with two nodes can be formulated as

$$R_1 \xrightarrow{b_1} [\, A_1 \xrightarrow{b_r} A_2 \,]_{a-b} \xrightarrow{b_2} R_2$$
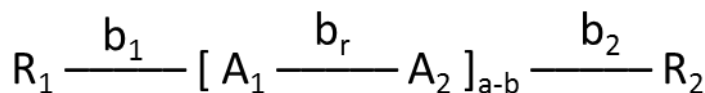
where A_1, A_2 are nodes inside the repeating group and b_r is the user defined bond between nodes A_1 and A_2. All nodes and bonds have a type which may be set to ring, chain, or ring/chain. The part inside the square brackets is called a repeating unit.

Normal applications of repeating groups include variable size chains and variable size rings. Examples are shown in Figure 43. In these cases the bond types of b_1, b_2 and b_r are identical; the bond type is chain for variable size chains and ring for variable size rings. Example 1, 3, and 4 contain a link node and example 2 contains a repeating group with two nodes inside the repeating group.
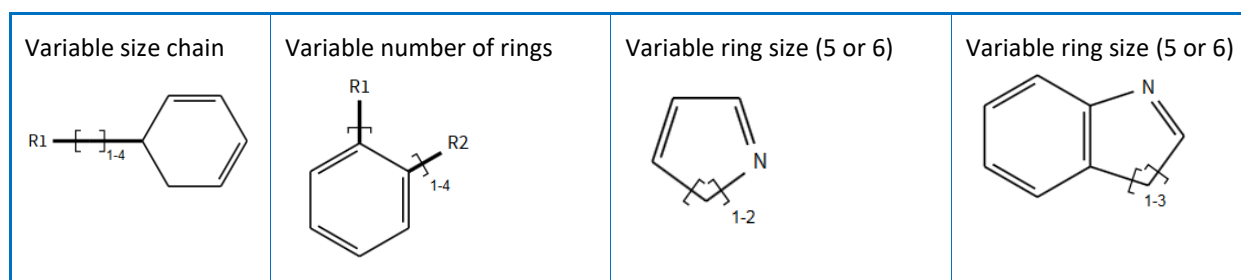
Figure 43. Queries with Frequency variations.

When a repeating group is repeated twice or more the repeating unit is connected with a new bond b_i that cannot be influenced by the user. In principle, the bond type (and value) of b_i are undefined. There are two settings of the structure search software for the bond b_i:[24]

- The <u>bond type</u> of b_i is always set to ring/chain.
- If the <u>bond values</u> of b_1, b_2 are identical this value will be used for the bond value of b_i; otherwise the bond value of b_i will be unspecified.

In order to understand the consequences of these rules it is helpful to look at an example (see Figure 44). In the query all bonds are chain bonds, but the node types of the two carbon atoms are set to ring. This means there could be any type of ring, carbocycles or heterocycles. Although the interval is from 1 to 4 we will only look at the frequency of 2. In this case the query expands to a carbon chain with 4 carbons (all node type ring). The internal bond b_i which connects the two C – C units is undefined and it is set by the software to ring/chain. As a result we have two options: (1) the bond b_i is chain which results in hits with topology type 1 and (2) the bond b_i is ring (all other bond are chain) and this results in hits with topology type 2.
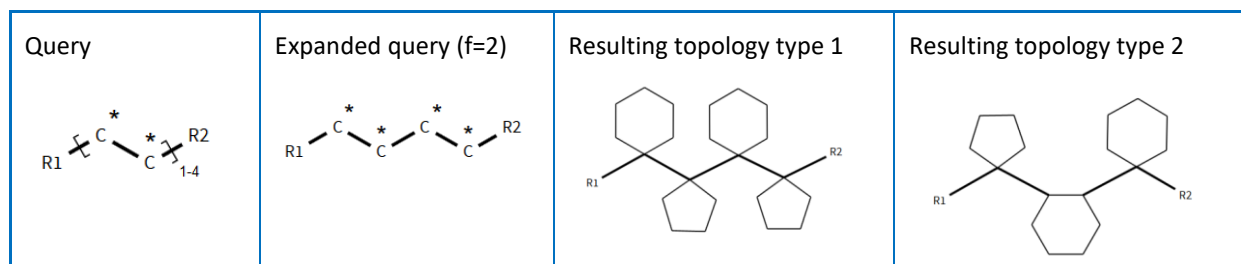
Figure 44. Query with repeating group resulting in two different topologies.

In the case of topology type 1 each carbon atom has its own ring. Our query does not specify the ring and there can be any ring present and all rings may have different ring sizes. Moreover, they can be carbocycles or heterocycles. In other words, our query formulation is similar (not equal) to the use of a generic node Cy. On the other hand, the topology type 2 is rather unusual and it results from the decision to set the bond type of b_i always to ring/chain. This results in rings which connect two repeating units.

An interesting case occurs when the bond types of the external bonds b_1, b_2 are both set to ring while the bond type of b_r is chain, e.g.

$$R_1 \xrightarrow{\quad r \quad} [\, C^R \xrightarrow{\quad c \quad} C^R \,]_{1\text{-}2} \xrightarrow{\quad r \quad} R_2$$
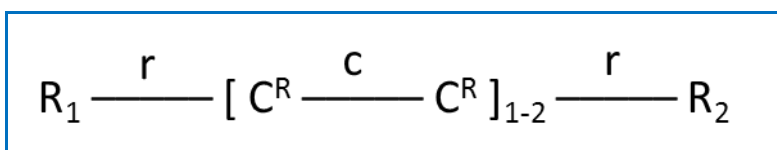
where we use small letters for the bond topologies and large letters for the node topologies. As stated above the bond type of b_i will be set to ring/chain. In this case we obtain two ring bonds between an external node and a node inside the repeating group. Now, this may sound very strange since we have learned that a repeating group can only have exactly two external bonds and now we are saying that we create a ring between R_1 and A_1 and R_2 and A_2. Let us explain how the repeating group is processed. A repeating group in a query structure is only an abbreviation and it creates an independent query for each frequency, i.e. if the interval of the repeating group is 1 - 2 we execute two queries with frequency f = 1 and f = 2. Now the repeating group has disappeared and we end up with two queries which have to be OR'ed in order to obtain the final result of the structure search.

Now, we will expand this query to f = 2 and set both R_1 and R_2 equal to the benzyl group Ph – C (red); the repeating group unit is colored in blue:

$$\text{Ph} \longrightarrow C^R \xrightarrow{\ r\ } C^R \xrightarrow{\ c\ } C^R \xrightarrow{\ r/c\ } C^R \xrightarrow{\ c\ } C^R \xrightarrow{\ r\ } C^R \longrightarrow \text{Ph}$$
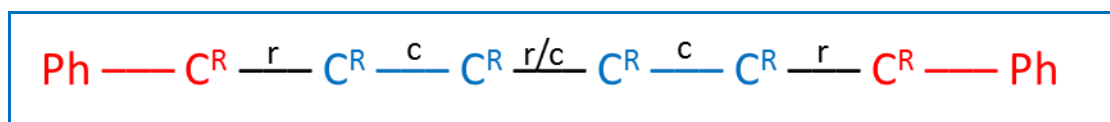
The ring/chain topology of the bond b_i generates two different topologies (type 1 and 2, see Figure 44). Here, we look at type 2, i.e. bond topology ring. Analyzing the above query we expect to find three rings between the two phenol groups (from left to right): C^R —— C^R, C^R —— C^R (between two repeating units), and C^R —— C^R. All other bonds are chain bonds. Finally, it should be noted that in STN it is also possible to assign different bond types to b_1 and b_2, i.e. b_1 to chain and b_2 to ring (or vice versa).

STN Generic Nodes

STN uses a different concept than Derwent for the description of generic nodes (Figure 45). The basis is a group of abstract nodes: Cy (any ring system), Cb (any carbocyclic system), Hy (any heterocyclic system – contains at least 1 non-carbon atom), and Ak (any carbon chain). In addition, variables can be applied as generic nodes: A (any atom except hydrogen), Q (any atom except carbon or hydrogen), M (any metal atom), and X (any halogen atom, incl. At).
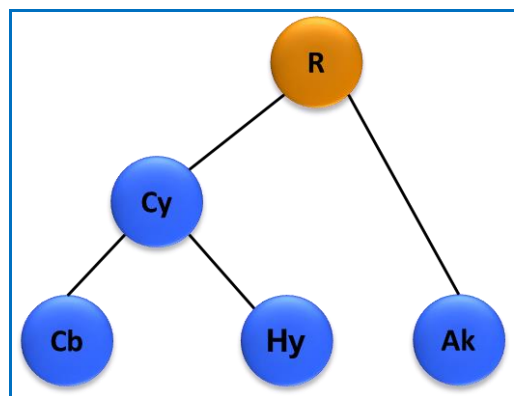
Figure 45. STN query convention for generic nodes.

Comparison of Query Nodes and Answer Nodes

It is important to distinguish between query and answer nodes. In the first release of DWPIM the generic nodes used in the query are different from the generic nodes as they appear in DWPIM (Figure 46). The reason is that the STN query convention and the DWPIM database conventions are different.

STN Query R-groups

- System defined variables: Cy, Cb, Hy, Ak, X, M, Q, A
- User defined R-groups: R1, R2, etc.

Answer G-groups

- MARPAT®: may also contain Cy, Cb, Hy, Ak, X, M
- DWPIM: may also contain ARY, CYC, HEA, HET, HEF, CHK, CHE, CHY, etc.
- G1, G2, etc., defining variables contained in the answer Markush structure

As a result a search using the STN system variables will result in a different variable in DWPIM. For example, a search using the Ak node (carbon chain) yields CHK, CHE, or CHY.
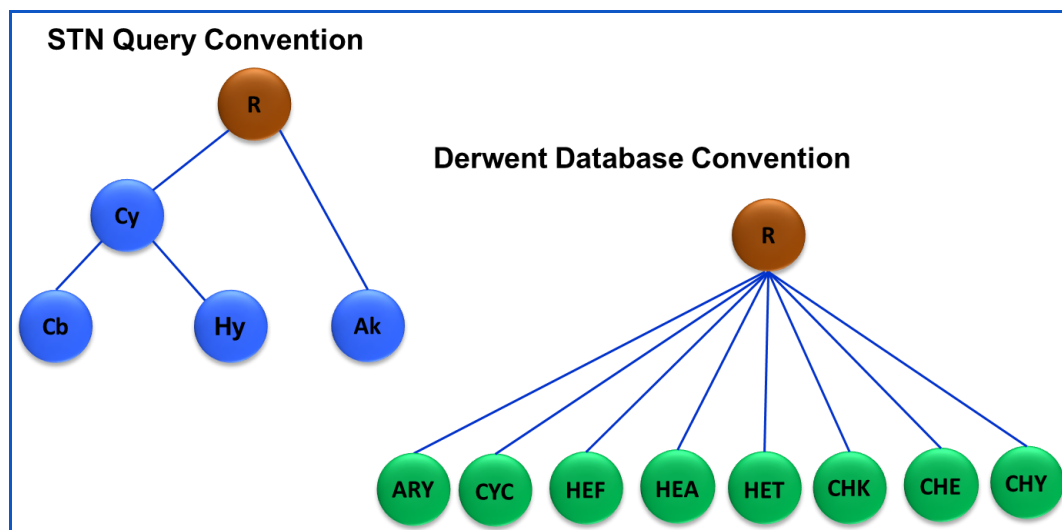
Figure 46. STN query convention for generic nodes and the Derwent database conventions.

3.3. Search Mode and Search Scope

The following **search modes** are applicable in DWPIM:

- Substructure search (SSS): A SSS retrieves substances in which the query appears embedded as a substructure of the molecule or molecule component. Substitution can occur at any open position.

- Closed substructure search (CSS). A CSS retrieves substances in which the query structure appears embedded as drawn. No substitution is allowed on the query except where specifically indicated, such as at forced open positions by use of non-hydrogen count.

The following **search scope** is applicable in DWPIM: Full search. The maximum search duration is set to 60 minutes. If a search cannot be completed within this time period an error message is displayed.

3.4. Spin-Off Generic Nodes

Searches for Markush structures require different types of matches between specific and generic nodes. In order to enable such matches it is necessary to construct additional generic nodes from specific structure fragments. These derived generic nodes are called spin-off nodes and they are denoted with an asterisk in order to distinguish them from original generic nodes. Spin-offs are constructed from structure fragments by applying the following rules:

- The longest carbon chain generates a chain spin-off (MARPAT: Ak*; DWPIM: CHK*, CHE*, or CHY*).
- The largest condensed ring system generates a ring spin-off (MARPAT: Cb*, Hy*; DWPIM: ARY*, CYC*, HEA*, HET*, HEF*).
- Heteroatoms (non-carbon non-metals except halogen) generate Q*.
- Metals and halogens generate the corresponding superatoms (M* and X*), *e.g.* Cl creates X*.

Both the query structure and the Markush file structure are augmented by spin-offs. Illustrations for the generation of spin-offs are shown in Figure 47. Note: in DWPIM the generic node Ak has been changed to the superatom CHK in order to simplify the spin-off.
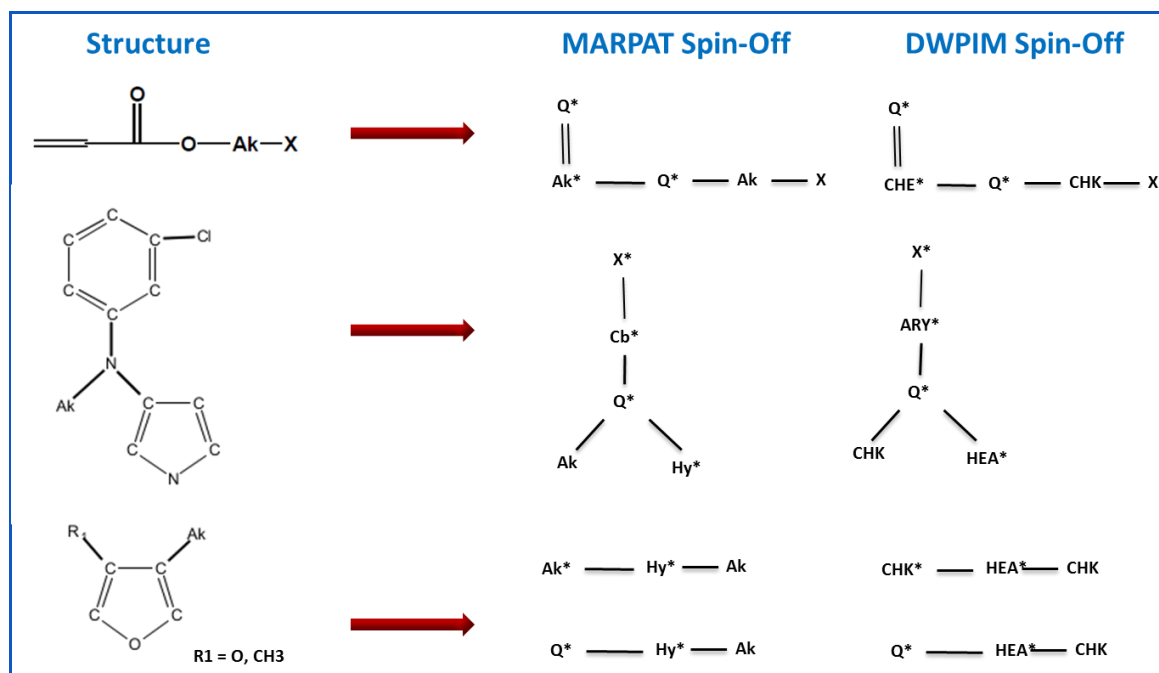
Figure 47. Construction of spin-offs.

When matching is applied in the search process it is very important that certain rules are obeyed

Rule 1:   Specific nodes (chemical elements) from the query match against specific nodes in the Markush structure

Rule 2:   Original generic nodes from the query match with original generic nodes as well as with identical spin-off nodes in the Markush structure

Rule 3:   Spin-off nodes from the query match with identical original generic nodes but NOT with spin-off nodes.

In Figure 48 we have shown how the matching works when spin-off nodes are taken into account. The first example is a substructure search for a pyrimidine derivative and the derived linearized spin-off structure is Q* − HEA* (− X*) − CHK*. This matches against N − HEA (− X) − CHK. The nitrogen from the query matches against the nitrogen in the file structure (rule 1), the spin-off HEA* from the query matches against the original HEA in the file structure (rule 3), the spin-off X* matches against X (rule 3), and CHK* matches against CHK (rule 2). The result is a hit structure. However, the first query structure does not match against the second example since the latter contains a bromine atom instead of a chlorine atom. The spin-off of bromine is X* and it is not allowed to match a spin-off against a spinoff. Hence, this structure is not a hit.

In the second example we start with a rather generic structure O − HEF − Ak (spin-off from O is Q*). This matches against serotonin since the oxygen from the query matches against the oxygen of the hydroxyl group, HEF matches against HEF* (spin-off from indole), and Ak matches against the substituted ethylene.
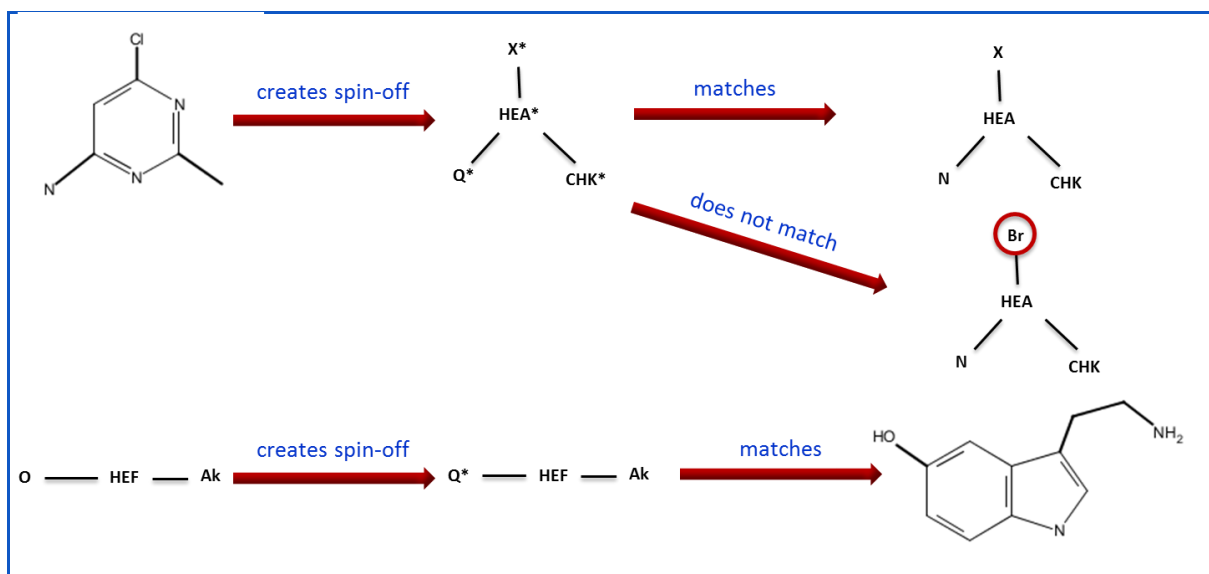
Figure 48. Examples of matches considering spin-off nodes.

3.5. Match Levels

Markush structures consist of both real atoms and generic nodes. Both types of nodes may occur in the query as well as in the file structures. The Markush search is able to compare

- specific elements (real atoms) with specific elements
- specific elements with generic nodes
- generic nodes with specific elements and
- generic nodes with generic nodes.

The degree of matching must be controlled by the user and the corresponding mechanism is called Match Level on STN. In other words: Match levels control the degree of structure query matching between the query structure and the structure in the search file. STN supports 3 match levels: ATOM, CLASS, and ANY, defined as:

- ATOM: retrieves only specific elements and groups of elements
- CLASS: retrieves the results of ATOM plus all generic nodes of the node hierarchy except the generic R-node
- ANY: retrieves the results of CLASS plus the generic R-node

Match levels are assigned as follows:

- Match levels are assigned to each atom and to each generic group/superatom.
- All nodes of a ring system should have the same match level.[25]
- Default Match Level for DWPIM: ATOM for ring nodes and CLASS for chain nodes.

To understand the mechanism how match levels work Figure 49 illustrates the effect on the search of pyridine. A search of pyridine with MLE ATOM yields specific pyridine compounds, with MLE CLASS the results are extended to include the corresponding superatom HEA (heteroaryl), and with MLE ANY results will also include the most generic node R (equal to superatoms XX or UNK).
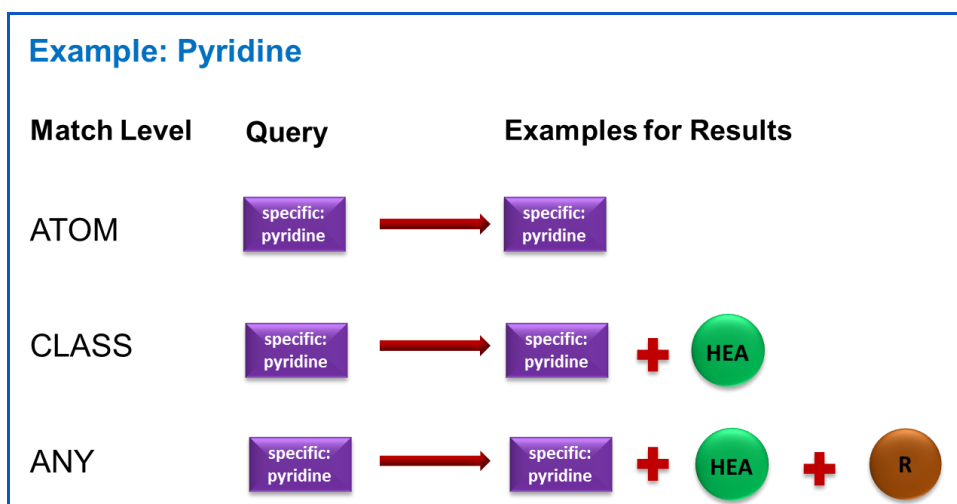
Figure 49. Effect of match levels on specific nodes.

In the case of generic nodes (superatoms) the effect of the match levels is formally identical. However, the set of specific compounds could be much larger. A search for the superatom HEA with MLE ATOM yields all specific heteroaryls, not only pyridine and its derivatives. This is illustrated in Figure 50.

---

[25] Exceptions are for ring systems which contain the superatom XX in the ring system
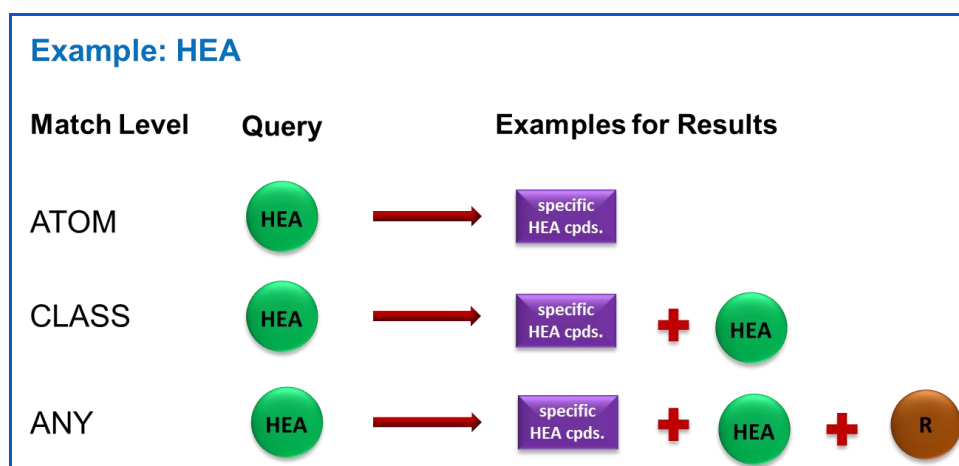
Figure 50. Effect of match levels on superatoms.

3.6. Combined Hierarchy of Generic Nodes and the Effect of Match Levels

An important point is the fact that it is possible to combine the STN query hierarchy and the Derwent hierarchy in a single combined hierarchy of generic nodes (see Figure 51).
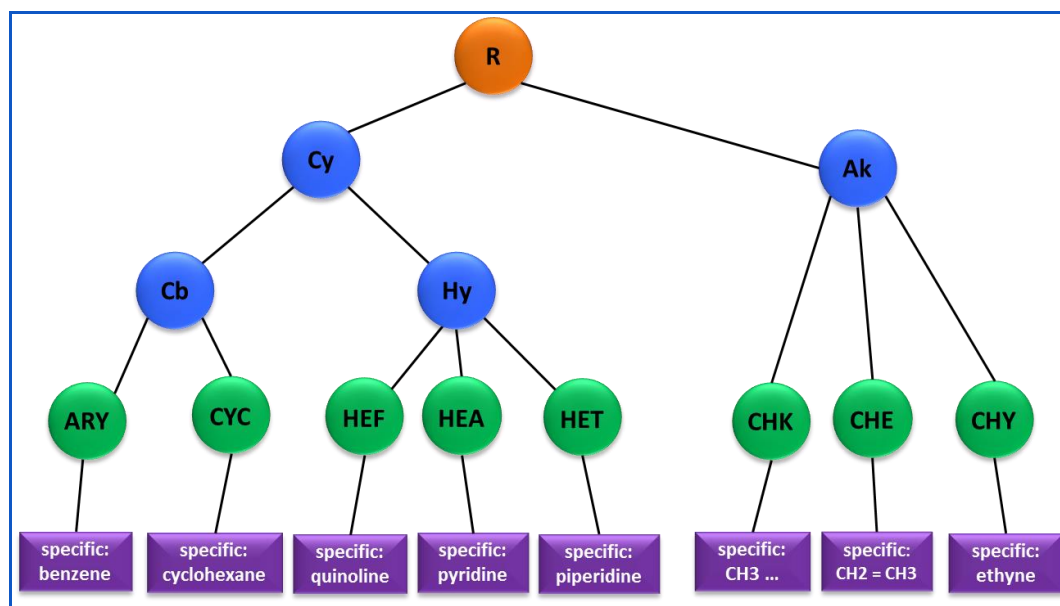
Figure 51. Combined hierarchy of generic nodes.

At first, it is important to note that both hierarchies fit together very well. There are four types of nodes and it is easy to recognize the various relationships.

- **Specific Fragments** (violet): in this case an example is given in each box. These fragments are indexed both by CAS and by Clarivate Analytics (although sometimes with different bonding conventions).
- **Superatoms** (green): they describe a certain level of chemistry (*e.g.* aromaticity). These nodes are only indexed by Clarivate Analytics.

- **STN/CAS Hierarchical Generic Nodes** (blue): they are more general than the Derwent nodes and comprise two or three superatoms. These nodes are only indexed by CAS.
- **R-Node** (brown): a top level node which may contain very general chemical fragments. This is defined differently by CAS and Derwent: in MARPAT, R is only shown in display and cannot be searched directly. In the case of Derwent, the node R is identified with XX or UNK and these nodes are also indexed.

The approach described above opens a number of new possibilities, which will become active with the second release, in which the Derwent nodes will be made searchable in addition to the STN/CAS nodes. As a consequence it will be possible to transfer a query from REGISTRY or MARPAT and refine the query by replacing all or some of the blue nodes with the more specific (Derwent) green nodes. This means that the query can be adapted as far as needed to the Derwent indexing. As a result the blue and green nodes are not an alternative (either … or) for searching but instead they provide different options (… as well as …) that can be chosen as necessary.

Based on the combined hierarchy it is easy to visualize the effect of match levels on searching. Let us assume we search for the specific organic fragment pyridine with different match levels. The first step is a search with match level ATOM (assigned to all nodes of the ring) and the result is an answer set with pyridine derivatives with all kind of substituents (see Figure 52, yellow marking). ATOM yields only specific nodes: chemical elements or fragments consisting solely of chemical elements.
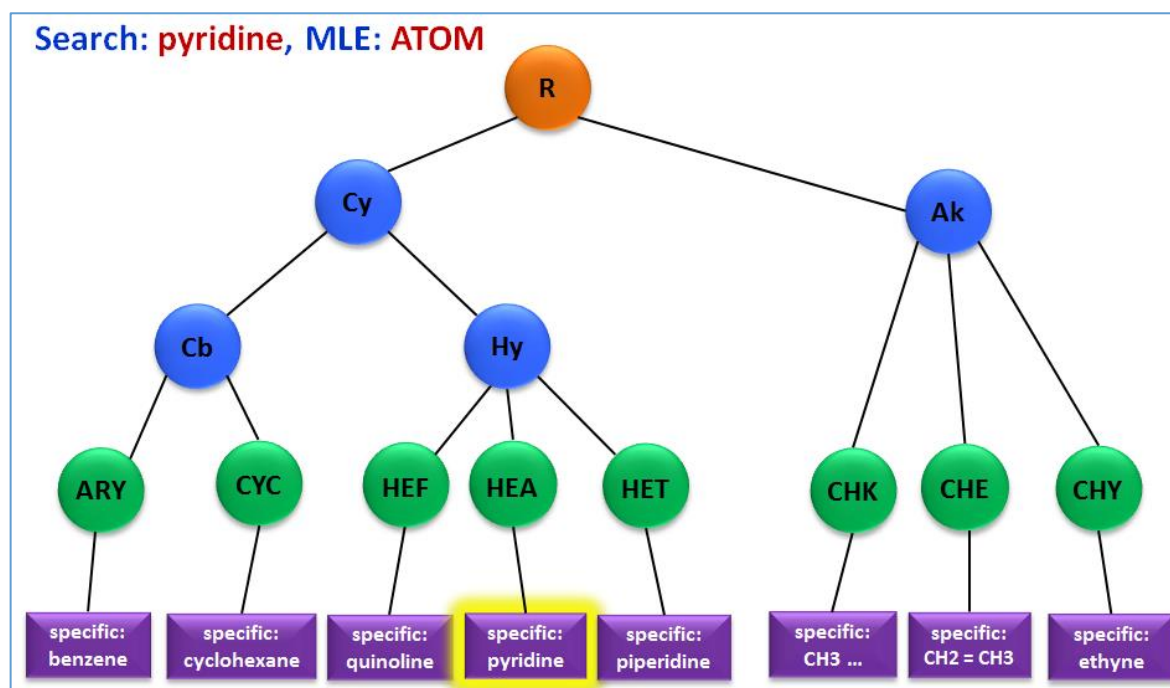
Figure 52. Illustration of the effect of match level ATOM.

In the next step we extend the search and assign match level CLASS to all nodes of the pyridine ring. In addition to the answer set from the search with match level ATOM we also obtain all generic nodes of

the hierarchy, except the R-node. Since Derwent indexes only the green nodes we find only the additional superatom HEA[26] (see Figure 53, yellow marking).[27]
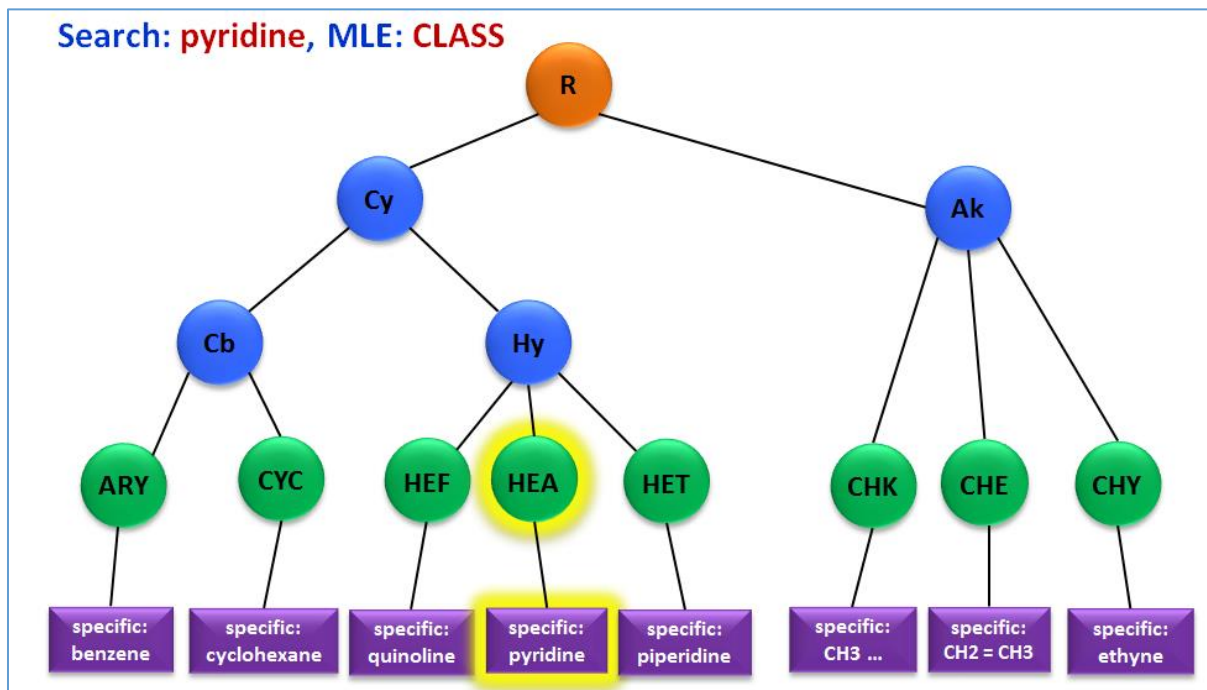
Figure 53. Illustration of the effect of match level CLASS.

Finally, a search with match level ANY results in matching with the nodes of the complete hierarchy. In the case of Derwent we find the additional R-node which is equal to XX or UNK (see Figure 54, yellow marking). As we will see later the extension to match level ANY is the only possibility to obtain some generic ring systems which may be important for the search result. Although the generic nodes Hy and Cb also belong to the pyridine hierarchy they are not retrieved as a hit since they are not indexed by Clarivate Analytics.

---

[26]   Provided that pyridine is isolated (otherwise HEF might also be retrieved)
[27]   Note: in the case of MARPAT one obtains the generic nodes Hy and Cy.
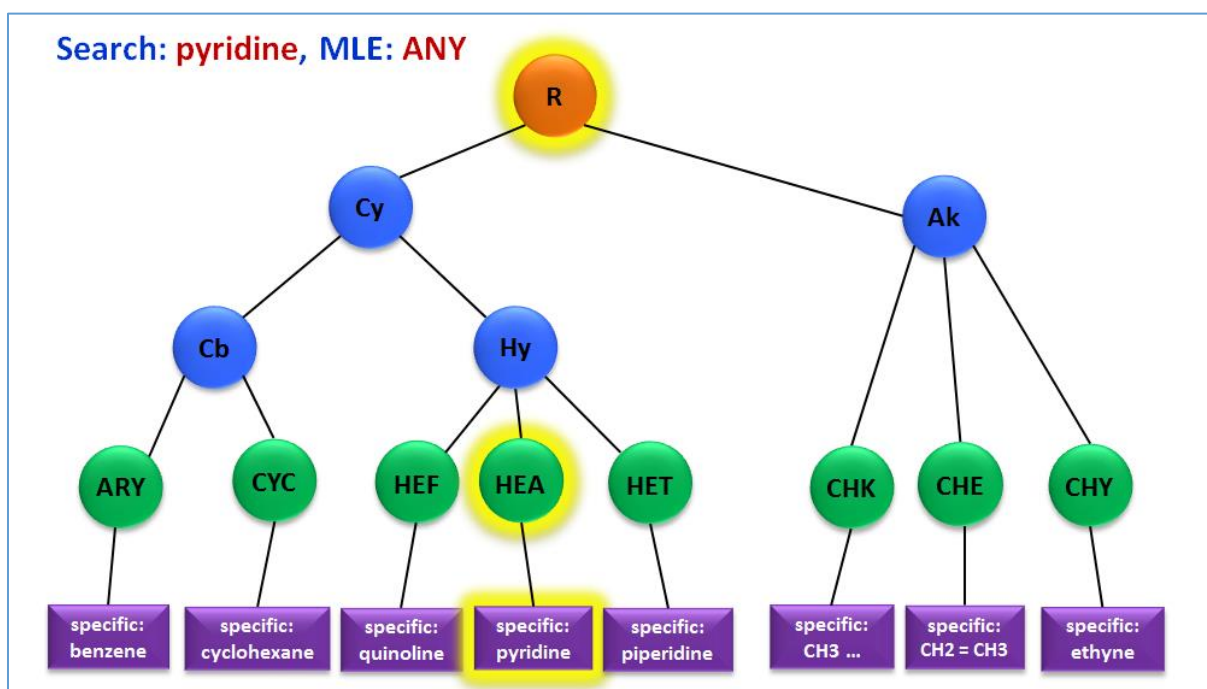
Figure 54. Illustration of the effect of match level ANY.

Looking at the three figures above it is rather easy to deduct the corresponding results for MARPAT. However, it is also very easy to visualize the effect of match levels for other situations, especially for the searches with generic nodes as a starting point. Here we have two cases: start form a Derwent node (green) or start from an STN node (blue). When we search for the superatom HEA with match level ATOM the result includes all specific heteroaryls (pyridine, pyrrole, thiophene, etc.) and not just the pyridine derivatives. The illustration is identical to Figure 52, but the interpretation is different. A search with HEA and match level CLASS yields in addition the HEA superatom (see Figure 53).

Since the first release will only allow the search of the STN nodes (blue) it is important to analyze this situation as well. Let us assume we start from the STN generic node Hy with match level ATOM. This will result in all specific heterocycles (all heteroaryls, all specific mono heterocycles, and all specific poly heterocycles). Match level CLASS will extend this answer set to include also the superatoms HEA, HET, and HEF.

Figure 55 illustrates the combined hierarchy of the element nodes (A-Q Tree), including all the chemical elements and the corresponding generic nodes. Some points are important to note:

- The superatom MX was replaced with the STN generic node M, and the superatom HAL with X despite the tiny difference with the element At.
- The metal subgroups have an additional level in the hierarchy.
- The non-metal atoms do not have an additional superatom.
- Q comprises any element except carbon (and hydrogen). Hence the violet rectangle represents all non-metal elements (except C) and not just N.
- A includes all elements including carbon (but not hydrogen).
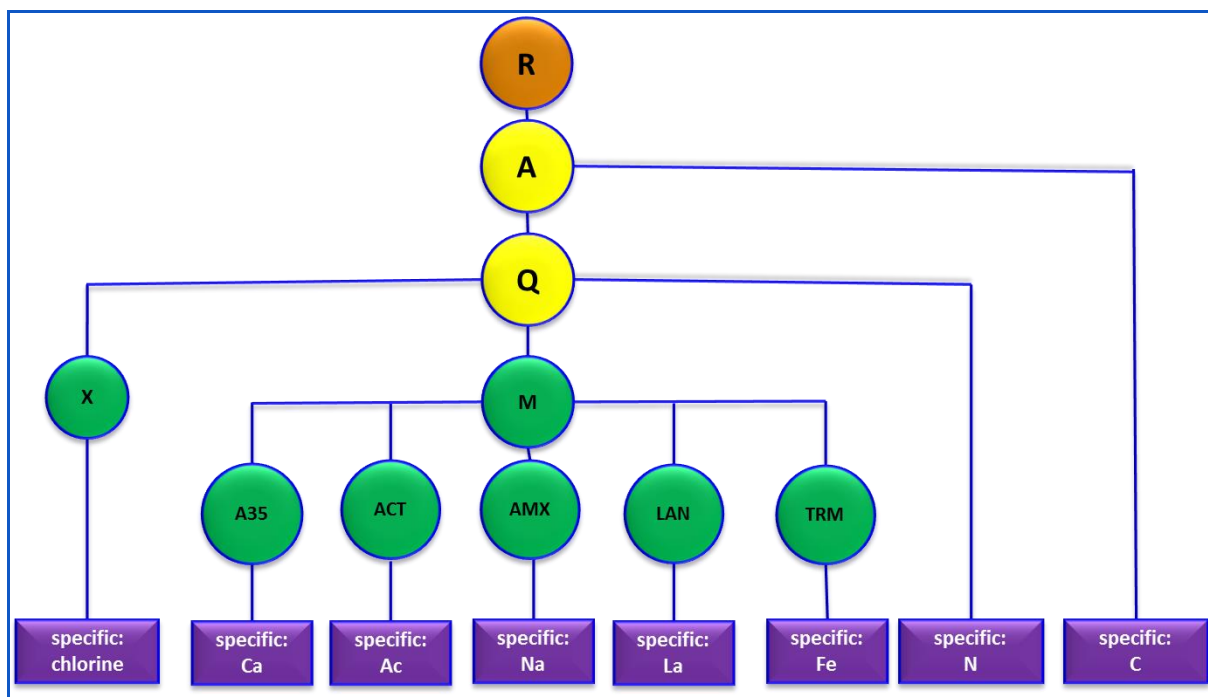
Figure 55. Combined hierarchy of the A-Q Tree.

The A node can match to individual atoms or nodes (not only metal nodes and halogen as shown in Figure 56 but also all other generic nodes and superatoms).

The superatoms ACY, DYE, PEG, POL, and PRT do not fit into the hierarchy of nodes, since they do not contain any set of specific nodes. In other words they are isolated and have no match level. It should be noted that the so-called amino acid shortcuts are not treated as superatoms in our implementation.

3.7.   Query Nodes Attributes

3.7.1.   Mapping of STN Generic Group Attributes to Superatoms

Both specific nodes and generic nodes may be further described by their corresponding attributes (see chapter 2.4, page 44). Although the attributes for STN generic nodes and for superatoms are identical in most cases, their names (and applications) may differ. In Table 13 we provide a mapping between the STN (and CAS) attributes and the corresponding superatoms. Except for the attribute heteroatom count ("exactly 1" or "2 or more") all other attributes are present in both systems.

Table 13. Comparison of STN Generic Group Attributes (GGA) and DWPIM superatom attributes.

STN generic group attribute	STN generic Group	Superatom Attribute	Superatom
Branched, linear	Ak	BRA or STR	CHK, CHE, CHY

STN generic group attribute	STN generic Group	Superatom Attribute	Superatom
0-6 carbons, **7+ carbons**	Ak, Cy, Cb, Hy	LOW, MID HI	CHK, CHE, CHY, ARY[28], CYC[28], HEA[28], HET[28], HEF[28]
Any, Exactly 1, **2 or more**	Cy, Hy	n/a	n/a
Any, monocyclic, **polycyclic**	Cy, Cb, Hy	MON or FU	ARY, CYC; (attribute already included in HEA, HET, HEF)
Any, saturated, **unsaturated**	Ak, Cy, Cb, Hy	SAT or UNS	CYC, HET, HEF; (attribute already included in CHK, CHE, CHY, ARY and HEA)

3.7.2. Mapping of STN Generic Nodes to Superatoms

A detailed overview of all possible matching options between STN generic nodes and superatoms including their associated attributes is given in Table 14. The key properties to be considered are "sat/unsat", "mono/poly" and "heteroatoms". Under "Input" the STN nodes and their associated properties are presented and "Output" encompasses the corresponding superatoms and their attributes. While in the column "Properties" the STN definition is still used (for better illustration) the Derwent terminology as presented in the hit structure displays is shown in column "Superatoms + Properties". Furthermore the "C-count" and "Hetero-Count" (if applicable) of the superatoms are listed in separate columns.

Example:

The STN generic node Cb with attributes unsat and poly (entry 12) can match to the superatoms ARY and CYC. Superatoms and their associated attributes would be displayed in the hit structures as ARY (fu) and CYC (uns, fu), respectively. Since ARY and CYC cover carbocycles only C-counts are considered.

---

[28] The generic group attribute is applicable, however, the corresponding superatom attribute is not shown in the display (compare with table 18)

Table 14. Overview of STN Generic Nodes vs. Superatoms

Rule	Input				Output				
No.	Generic Node	sat / unsat	mono / poly	Heteroatome	Superatoms	Properties	C-Count	Hetero-Count	Superatoms + Properties
1	Ak	N/S	N/A	N/A	CHK, CHE, CHY	N/S	yes	N/A	CHK, CHE, CHY
2	Ak	sat	N/A	N/A	CHK	sat	yes	N/A	CHK
3	Ak	unsat	N/A	N/A	CHE, CHY	unsat	yes	N/A	CHE, CHY
4	Cb	N/S	N/S	N/A	ARY, CYC	N/S	yes	N/A	ARY, CYC
5	Cb	N/S	mono	N/A	ARY, CYC	mono	yes	N/A	ARY (mon), CYC (mon)
6	Cb	N/S	poly	N/A	ARY, CYC	poly	yes	N/A	ARY (fu), CYC (fu)
7	Cb	sat	N/S	N/A	CYC	sat	yes	N/A	CYC (sat)
8	Cb	unsat	N/S	N/A	ARY, CYC	unsat	yes	N/A	ARY, CYC (uns)
9	Cb	sat	mono	N/A	CYC	sat, mono	yes	N/A	CYC (sat, mon)
10	Cb	sat	poly	N/A	CYC	sat, poly	yes	N/A	CYC (sat, fu)
11	Cb	unsat	mono	N/A	ARY, CYC	unsat, mono	yes	N/A	ARY (mon), CYC (uns, mon)
12	Cb	unsat	poly	N/A	ARY, CYC	unsat, poly	yes	N/A	ARY (fu), CYC (uns, fu)
13	Hy, Cy	N/S	N/S	yes	HEA, HET, HEF	N/S	yes	yes	HEA, HET, HEF
14	Hy, Cy	N/S	mono	yes	HEA, HET	mono	yes	yes	HEA, HET
15	Hy, Cy	N/S	poly	yes	HEF	poly	yes	yes	HEF
16	Hy, Cy	sat	N/S	yes	HET, HEF	sat	yes	yes	HET (sat), HEF (sat)
17	Hy, Cy	unsat	N/S	yes	HEA, HET, HEF	unsat	yes	yes	HEA, HET (uns), HEF (uns)
18	Hy, Cy	sat	mono	yes	HET	sat	yes	yes	HET (sat)
19	Hy, Cy	sat	poly	yes	HEF	sat	yes	yes	HEF (sat)
20	Hy, Cy	unsat	mono	yes	HEA, HET	unsat	yes	yes	HEA, HET (uns)
21	Hy, Cy	unsat	poly	yes	HEF	unsat	yes	yes	HEF (uns)
22	Cy	N/S	N/S	N/S	ARY, CYC, HEA, HET, HEF	N/S	yes	N/S	ARY, CYC, HEA, HET, HEF
23	Cy	N/S	mono	N/S	ARY, CYC, HEA, HET	mono	yes	N/S	ARY (mon), CYC (mon), HEA, HET
24	Cy	N/S	poly	N/S	ARY, CYC, HEF	poly	yes	N/S	ARY (fu, CYC (fu), HEF
25	Cy	sat	N/S	N/S	CYC, HET, HEF	sat	yes	N/S	CYC (sat), HET (sat), HEF (sat)
26	Cy	unsat	N/S	N/S	ARY, CYC, HEA, HET, HEF	unsat	yes	N/S	ARY, CYC (uns), HEA, HET (uns), HEF (uns)
27	Cy	sat	mono	N/S	CYC, HET	sat, mono	yes	N/S	CYC (sat, mon), HET (sat)
28	Cy	sat	poly	N/S	CYC, HEF	sat, poly	yes	N/S	CYC (sat, fu), HEF (sat)
29	Cy	unsat	mono	N/S	ARY, CYC, HEA, HET	unsat, mono	yes	N/S	ARY (mon), CYC (uns, mon), HEA (uns), HET (uns)
30	Cy	unsat	poly	N/S	ARY, CYC, HEF	unsat, poly	yes	N/S	ARY (fu), CYC (uns, fu), HEF (uns)

N/A: not applicable
N/S: not specified

3.7.3. Element Counts Limited and Unlimited

Element counts are another specification of structure formulas which can be used to restrict the search for substances. This is especially important together with the application of match levels.

Indexing:

- For specific partial structures element counts are generated from the compound structure. For example pyridine contains N = 1 and C = 5.
- For generic nodes (CHK, …, ARY, …) the element counts have to be specified by the indexer. As an example CHK may contain between 1 - 6 C atoms. This is indexed as CHK (LO). If there are no element counts indexed the superatom contains no element-counts.

Query:

- Element counts are generated from the specific partial structure and may be used for searching.
- The user may assign element counts to generic nodes (CHK, …, ARY, …) with numerical values as exact value, range, minimum, or maximum.
- Default for searching is **Element Count Limited**. This means that the element counts of queries must have an explicit overlap with the element counts of the Markush structures in the answer set.
- The user may also choose **Element Count Unlimited**. In this case the answer set will also contain structures which have no Element Count for the specified element.

Element Count Limited/Unlimited has different effects on the application of match levels (s. examples in Table 15).

- MLE = CLASS: important difference, since ELC Limited requests an explicit overlap of the element between the query and the structure. If ELC Unlimited is specified element counts of the query structure are simply ignored.
- MLE = ANY: important difference with respect to additional hits resulting from CLASS Unlimited, since ANY Limited is based on CLASS limited. In other words ANY Limited collects all the hits from ATOM plus generic nodes from CLASS Limited plus all hits resulting from a match with XX or UNK. ANY Unlimited yields the results from ANY Limited plus the additional hits resulting from CLASS Unlimited.

For DWPIM there is no distinction between ELC Limited and ELC Unlimited for the match level ATOM in case of a specific structure. Figure 56 illustrates the effect of ELC Limited and ELC Unlimited for match level CLASS on a specific structure query (pyridine). Pyridine consists of 5 carbon atoms and 1 nitrogen atom. In the case of MLE = CLASS Limited it is necessary that the element counts for C and N of HEA overlap with those of pyridine. Since HEA comprises only five and six-membered heteroaryls it is obvious that HEA must have C = 5 and N = 1 in order to match with the query. Hence, it must be a 6-membered ring. In the case of Element Count Unlimited all HEA superatoms are retrieved, independent of any element count. In other words CLASS Unlimited simply ignores the element counts in the query.
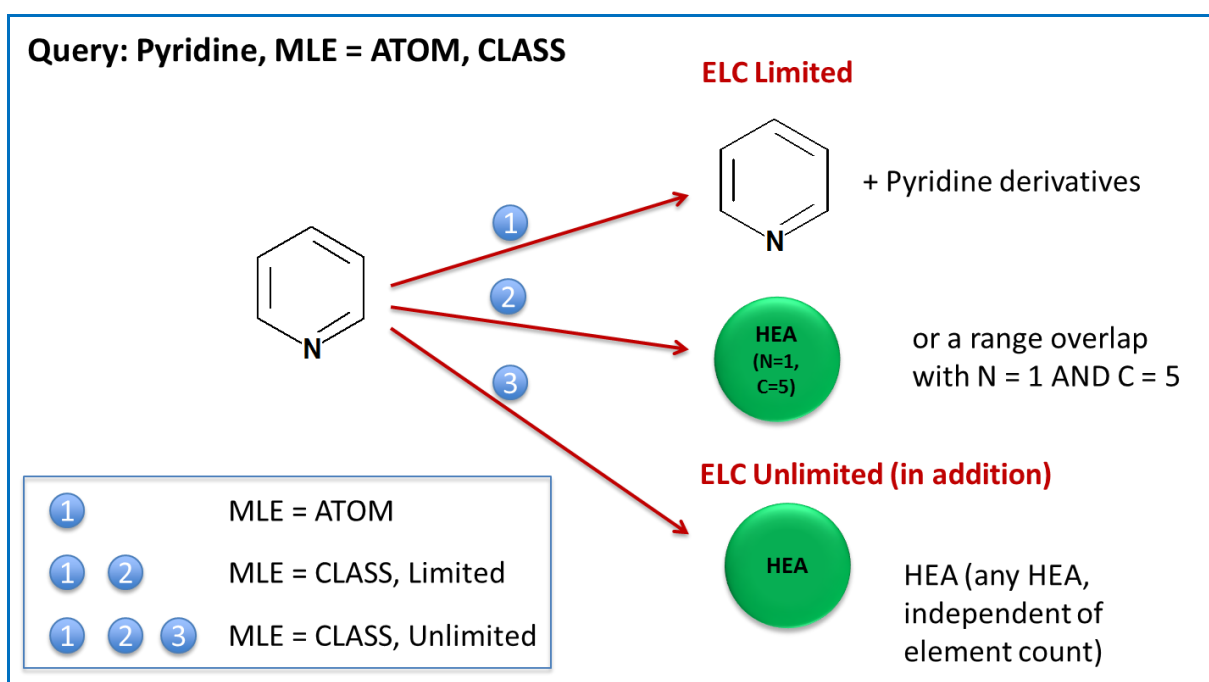
Figure 56. Example for the search of a specific structure query with MLE = CLASS Limited and Unlimited.

A search for HEA (N=1) with ELC Limited and ELC Unlimited is illustrated in Figure 57. For the sake of illustration for the search result for MLE ATOM only the Limited scenario is shown, for a comprehensive overview see Table 15. A search with MLE CLASS Limited requires that the results have 1 or more nitrogen atoms, *e.g.* HEA (N=1) or HEA (N ≥ 1). In order to retrieve all structures with any HEA (independent of any element count) it is necessary to apply MLE = CLASS Unlimited. In other words, for CLASS Unlimited the element counts of the query are simply ignored.
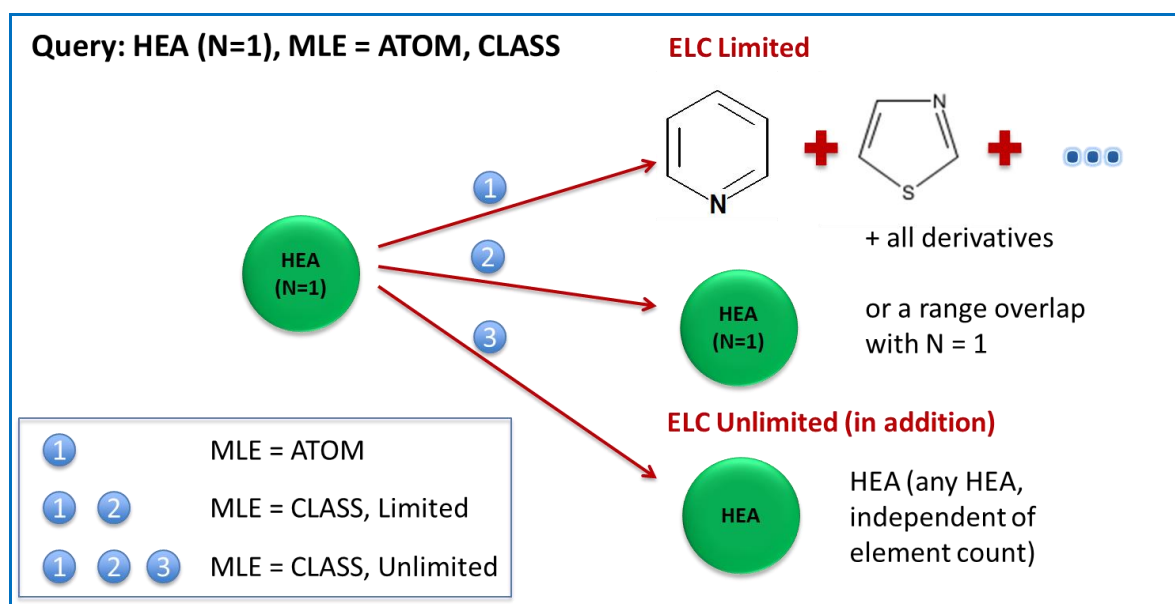
Figure 57. Example for the search of a generic structure query with MLE = CLASS limited and unlimited.

A summary of the search results for pyridine and HEA (N=1, C=5) with MLE = CLASS Limited/Unlimited are collected in Table 15. On each level a new set of hits is included in the answer set (new hits are indicated in red color).

Table 15. Possible results for searching pyridine or HEA (N = 1) with ELC Limited and Unlimited.

Match Level	Element Count	Specific Nodes (Pyridine)	Generic Nodes (HEA, N = 1)
ATOM	Limited	• Pyridine derivatives	• All specific aromatic 5- and 6-rings with N = 1
	Unlimited	• Pyridine derivatives	• All specific aromatic 5- and 6-rings
CLASS	Limited	• Pyridine derivatives • HEA (N=1, C=5) or a range overlap	• All specific aromatic 5- and 6-rings with N = 1 • HEA (N=1) or a range overlap
	Unlimited	• Results from CLASS Limited • HEA (any HEA, independent of any element counts)	• Results from CLASS Limited • HEA (any HEA, independent of any element counts)
ANY	Limited	• Results from CLASS Limited • XX (independent of any attributes or counts)	• Results from CLASS Limited • XX (independent of any attributes or counts)
	Unlimited	• Results from CLASS Unlimited • XX (independent of any attributes or counts)	• Results from CLASS Unlimited • XX (independent of any attributes or counts)

3.7.4. Impact of Superatom Attributes in the Node Hierarchy

Node attributes (chapter 3.7) and element counts (chapter 3.7.2) are associated with the generic node and they are inherited by each lower level in the hierarchy. This is illustrated in the following figures:

Carbon Chains: Figure 58 shows the impact of node attributes and element counts from the generic carbon chain Ak on the superatoms CHK, CHE, and CHY, and further to the specific carbon chains. The parameters are: low/high carbon count, linear/branched, saturated/unsaturated, non-hydrogen attachments (equivalent to free sites), and range of carbon atoms. From this illustration it is clear that the user can distinguish between saturated (CHK) and unsaturated carbon chains (CHE and CHY). However, with the generic node Ak a further distinction between compounds with double and triple bonds is not possible.
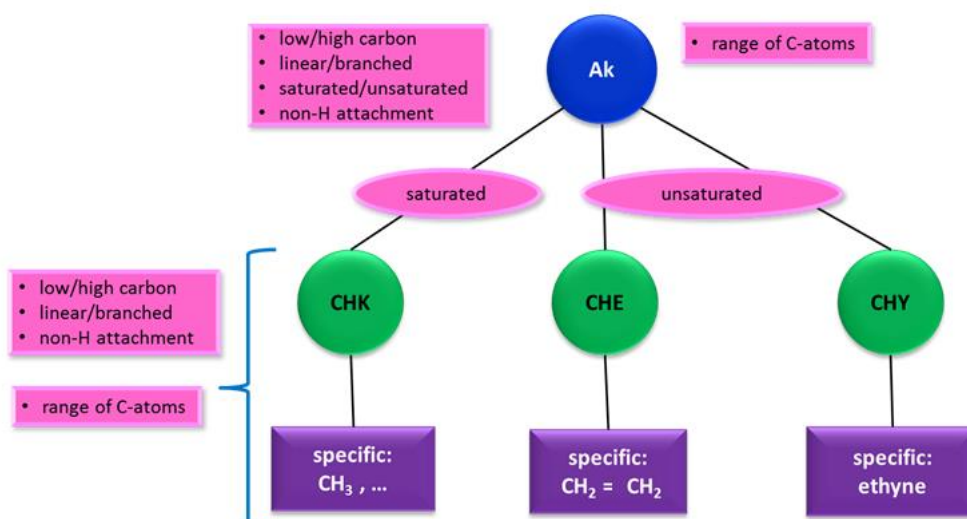
Figure 58. Impact of superatom attributes and element counts for carbon chains.

Carbocycles: Figure 59 indicates the impact of node attributes and element counts from the generic carbocycles Cb on the superatoms ARY and CYC, and further on the specific carbocycles. The parameters are: low/high carbon count, monocyclic/fused, saturated/unsaturated, non-hydrogen attachments (equivalent to free sites), ring isolation, and range of carbon atoms. In the case of ARY and CYC there is no parameter available to distinguish between these superatoms.
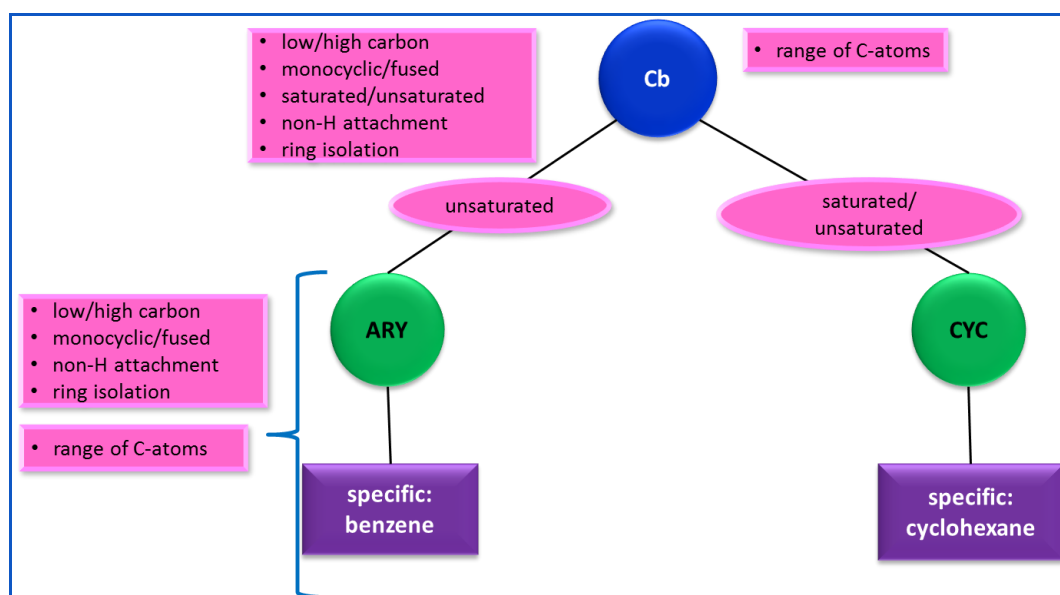
Figure 59. Impact of superatom attributes and element counts for carbocycles.

Heterocycles: Figure 60 demonstrates the impact of node attributes and element counts on the generic heterocycles Hy. In addition to the parameters for carbocycles there are other parameters: number of heteroatoms, and range of elements for each element, including C. It is possible to distinguish between monocyclic (HEA and HET) and fused rings (HEF) but it is not possible to distinguish further between the monocyclic rings.
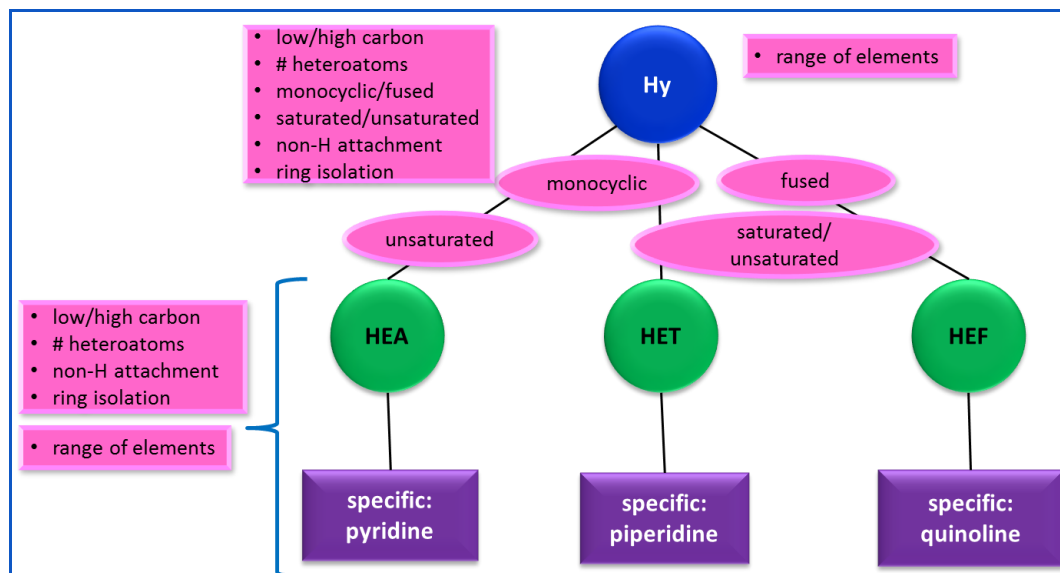
Figure 60. Impact of superatom attributes and element counts for heterocycles.

3.8. Other Search Fields

3.8.1. Markush Compound Number

Markush compound numbers are of the general format

YYWW-CCCSS,

where YY is the year and WW is the Derwent week number for the document. CCC is a three character identifier unique to a document in any given week, and SS is an integer from 01 to 99. One or more of these Markush Compound Numbers may be assigned to any DWPI document containing compounds represented as Markush (generic) structures. Compound Numbers in this format may also be assigned to a single compound if this compound is claimed new.

The source field for Markush in DWPIM is named /AN (Accession Number). A search for specific Markush records can be performed by using the syntax "code number/AN" in the query builder window in STN. Several AN's can be combined with the aid of logical operators (OR, AND, NOT) as depicted in Figure 61.

```
=> que 9540-B3902/AN OR 9523-A5208/AN

L1    QUE 9540-B3902/AN OR 9523-A5208/AN
```

Figure 61. Search for Markush Accession Numbers (AN.M).

The corresponding field code for Markush numbers in DWPI is called /MCN (Figure 62). In a cross-file search in DWPIM/DWPI starting in DWPIM the /AN field is automatically converted to /MCN to ensure compatibility. After the MCN number, DWPI compound number (DCN) roles are indexed in the M-fields in DWPI and the DCR number roles occur in the index terms (IT) section. Markush records can be searched in DWPIM using the syntax "que *accession number*/MCN", e.g. "que 9523-48001-N/MCN" whereby the letter "N" stands for "new compound. Display of "HIT" shows the corresponding entry in the respective M-field as shown in the example 9523-48001-N in Figure 62.

```
L5    ANSWER 1 OF 1  WPIX COPYRIGHT 2019   CLARIVATE ANALYTICS on STN
CMC  UPB   20120924
     M1 *01*   B415 B701 B702 B713 B720 B741 B742 B743 D011 D040 D601 F010 F012
               F014 F020 F021 F111 F211 F423 F431 F521 G001 G002 G003 G010 G013
               G019 G050 G100 G221 H100 H101 H181 H182 H4 H401 H441 H481 H498
               H5 H598 H8 H9 J0 J011 J012 J1 J111 J171 J172 J3 J371 K0 L2 L250
               M210 M211 M212 M213 M214 M215 M216 M271 M280 M281 M311 M312 M313
               M314 M315 M320 M331 M332 M333 M343 M349 M371 M372 M373 M381 M391
               M423 M510 M511 M520 M521 M530 M531 M532 M533 M540 M620 M710 P523
               P528 P616 P633 V812 V902 V911  M903 M904
               MCN: 9523-48001-N
```

Figure 62. Markush Numbers (/MCN) in DWPI.

In the following the complete set of roles which can be used in DWPI after crossover from DWPIM are presented.

3.8.2. DWPI Compound Number and DCR Compound Number Roles

For historical reasons the databases DWPI and DCR contain two partly overlapping sets of roles.

An overview of all available DWPI compound number (DCN) roles is given in Table 16. The indexing of DWPI compound number (DCN) roles started in 1987.

Table 16. Overview of DWPI compound number (DCN) roles.

Derwent Role	Description	Derwent Role	Description
A	Substance Analysed/Detected	Q	Product Defined in Terms of Starting Materials
C	Catalyst	R	Removing/Purifying Agent
D	Detecting Agent	S	Starting Material
E	Excipient	T	Therapeutically Active
K	Known Compound	U	Use of a Single Compound
M	Component of a Mixture	V	Reagent
N	New Compound	X	Substance Removed
P	Known Compound Produced (if the preparation is also claimed role N is assigned as well)	Z	Miscellaneous

An overview of all available DCR number roles is given in Table 17. The indexing of DCR compound number roles started in 1999.

Table 17. Overview of DCR number roles.

Derwent Role	Description	Derwent Role	Description
CL	Claim	RCT	Reactant
EX	Example	RGT	Reagent
DISC	Disclosure	CMP	Component
NEW	New	PUR	Purified
PRD	Produced	REM	Removed
USE	Use	TES	Tested
DET	Detected	ST	Salt

For better illustration and for optimal support of the STN workflow both sets of roles are combined in Table 18.

Table 18. Combined overview of DWPI compound number roles and DCR number roles.

Role	Description	Role	Description
A	Analysed or detected	PUR	Purified
C	Catalyst	Q	Product defined by starting material
CL	Claim	R	Removing/purifying agent
D	Detecting agent	S, RCT	Starting material
DET	Detected	ST	Salt
DIS	Disclosure	T	Therapeutically active
E	Excipient	TES	Tested
EX	Example	U, USE	Use of substance or apparatus

Role	Description	Role	Description
K	Known compound	V, RGT	Reagent
M, CMP	Component of a mixture	X, REM	Substance removed
N, NEW	New compound	Z	Miscellaneous
P, PRD	Produced		

3.8.3. Role Searching Option of DCR and DWPIM Structure Search Results in DWPI

The new role searching option in DWPI supports the refinement of DCR and DWPIM structure search results. The respective role field is "/MCN" and the appropriate proximity operator is T.

Example: The result of a structure search in DWPIM is narrowed down by records encompassing the roles "New compound" and "Produced".

The syntax for this operation is:

=> FIL DWPIM

(Upload structure to create Lx)

=> S Lx SSS FULL *(Ly)*

=> FIL WPIX

=> S Ly(T)(P OR PRD OR N OR NEW)/MCN

=> D FULLG AHITSTR

An complete overview on all available DCR and DCN roles as well as DWPI Registry Number (DRN) roles is also available online by applying the "help roles" command in DWPIM.

3.8.4. Substance Descriptors (SDM)

Derwent Markush Resource Substance descriptors (also known as file segments) are a possibility to categorize the type of structures presented in a particular Markush record. Substance descriptors can be subdivided into the following sections:

- **Sectional substance descriptors** – identify from which chemistry related Chemical Patents Index (CPI) sections of the patent the Markush structure was taken from. These are the descriptors A, B and E. All records must have at least one of them applied.
- **Structural substance descriptors** – identify the type of substance the Markush structure represents (e.g. simple organic, inorganic, complex, peptide etc.). These are the descriptors 7, C, F, L, M, P, V, W and Z. In addition the rarely applied (in most cases only F is applied) polymer related descriptors H, S, J, K, R, N, Q, X, D, G, T, 5 and U.
- **Role related substance descriptor Y** – Used whenever the indexed substance is a component of a mixture. This descriptor is applied whenever the role M is applied to a Markush record in

the indexing. It indicates that at least 2 of the compounds described in the indexing for this patent must be used in the combination to achieve the desired effect of the invention.

- **INPI specific records descriptor 1** – comprises all INPI specific structures from 1961-1998 (those records start with numbers 82 and 83. But please note that there are also generic structures starting with 82 and 83).

By using substance descriptors the retrieved result set can be narrowed in a targeted manner. There are 26 substance descriptors available which cover a wide range of chemical classes (Table 19). A Markush record may be characterized by more than one substance descriptor. A special set of descriptors determining oligopeptides, oligosaccharides, polypeptides and polysaccharides is depicted in

Table 25 in chapter 4.3 (Polymer or Oligomer).

Table 19. Overview of available substance descriptors.

Chemical Class	Code	Description	# Entries in DWPIM[29]
CPI sections	A	polymers, plastics	834,503
	B	pharmaceuticals, agrochemicals	816,142
	E	general chemicals	1,384,477
General	Y	mixtures	1,047,109
	Z	salts (at least one organic component)	537,631
Polypeptides	P	polypeptides	89,526
Miscellaneous compounds	7	inorganics	79,474
	C	coordination compounds, complexes	130,093
	L	oligomers	94,075
	W	extended structures (include fullerenes, zeolites and clays)	164,725
	M	metals, alloys	4,146
	V	simple organic compounds (small molecule)	1,424,240
Polymers	F	any polymers (other than polypeptides)	69,645
	X	cross-linked polymers	120
	D	derived polymers	66
	G	grafted polymers	25
	T	end modified polymers	445
	5	surface modified polymers	18
	U	unmodified polymers	75
Polymer backbone	H	homopolymers	428
	S	simple binary condensates	30
	J	alternating copolymers	27
	K	block copolymers	79
	R	random copolymers	46
	N	natural polymers	8,136
	Q	no backbone	16
INPI specific structures	1	INPI specific structures (1961-1998)	141,081

Substance descriptors can be applied in the query builder. The query syntax for the query builder is: "s L-number AND substance descriptor/SDM" whereas for the term substance descriptor either the code or the corresponding description can be used. For example: "s L1 AND Y/SDM" or "s L1 and mixture/SDM". The Substance descriptors can also be selected by using the Term Explorer in STN.

It should be noted that not all Markush records in DWPIM have substance descriptors associated with them. While the DWPIM database encompasses approximately 2,35 mio records (as of Oct 2020) the number of records with an associated substance descriptor is approximately 2,2 mio.

---

[29] As of 12-10-2020

3.8.5. Markush Descriptors (MDE)

In addition to true Markush structures DWPIM also contains around 180.000 single specific structures. In order to be able to identify those structures the Derwent Markush Descriptors (MDE) have been generated automatically from the Markush structures. There are three different (exclusive) Markush Descriptors (Table 20): (1) structures without any Markush variation = single specific structures (Markush Descriptor: S), (2) structures with at least one Markush variation of the type frequency, position, or substitution (or superatoms X, M), and (3) structures containing at least one true generic node = homology variation.

Table 20. Overview of available Markush Descriptors.

Markush Descriptor	Code	Description	# Entries in DWPIM[30]
Single Specific Structure	S	Structure without any Markush variation	208,295
All Specific Structures	A	Structures with at least one Markush variation of the type frequency, position, or substitution (or superatoms X, M)	491,573
True Generic Structures	G	Structures containing at least one true generic node = homology variation (i.e. CHK, CHE, CHY, ARY, CYC, HEA, HET, HEF, ACY, DYE, POL, PEG, PRT, XX, UNK)	1,653,911

Typical representatives for each Markush Descriptor are shown in Figure 63. The example for S (Single Specific Structure) represents a structure without variations and G-groups. The example for A (All Specific Structures) contains the generic node HAL which represents a closed set of halogen atoms F, Cl, Br and I. The example for G (True Generic Structures) contains the generic nodes CHK and CHE which define open sets of carbon chains, i.e. they cannot be enumerated by a closed set of structures.
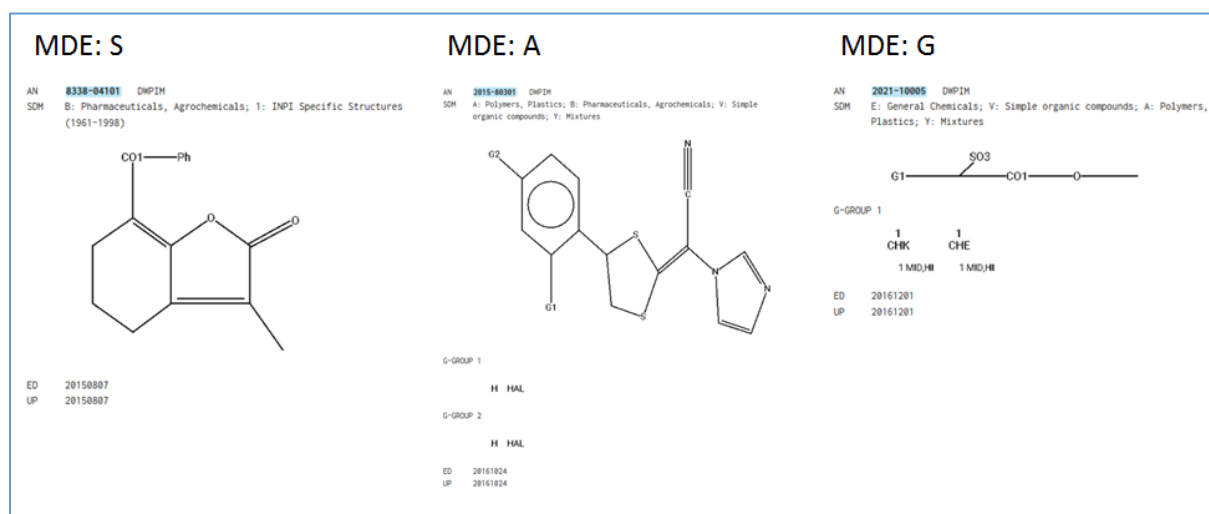

Figure 63. Examples for Markush Descriptors S, A and G.

The Markush Descriptors enable the user to perform a stepwise extension of the search results from specific structures to true Markush structures. Our recommended approach is as follows:

- Start the search in DCR and DWPIM and apply the filter Markush Descriptor (S) to the DWPIM answer set and the resulting answer set contains all single specific structures from Derwent indexing.
- Continue with the addition of all Markush structures containing only specific variations, i.e. frequency, position, or substitution variations using the Markush Descriptor A
- Finally, using the Markush Descriptor G one obtains the Markush structures which have at least one true generic node.

Notes: The classification of the Markush Descriptor type is independent of the location of the structural variation within the Markush record. That means, for example, that the classification of a record as G (true generic structure) is independent of where the indefinite generic node such as HEA is found in the record. It could be located directly in the assembled structure but also in any G-group present.

The major part of single specific structures (S) are INPI specific records from the time period 1961-1998. Consequently those specific structures (106.835) are unique to DWPIM and cannot be found in DCR (starting from 1999). Those 106.835 records can be retrieved by using the Substance Descriptor INPI specific structures (Table 19) following the operation "s S/MDE AND 1/SDM".

3.8.6. Patent Number Kind Code (PNK)

The PNK (Patent Number Kind Code) field can be used as search and display field:

- Example for display: d L1 1-3/PNK
- Example for search: s L1 AND WO2019111107 A1/PNK

3.9. Attributes associated with Superatoms

Information about attributes is displayed in the graphical representation of the core structure (G0) and the G-fragments. Table 21 gives an overview on those attributes.

Table 21. Overview on the superatom attributes in the display.

System	Attribute	DWPIM Display	Attributes apply to
Ring type	Monocyclic	MON	ARY, CYC
	Fused	FU	
Degree of saturation	Fully Saturated	SAT	CYC, HEF[31], HET
	Unsaturated	UNS	
Chain type	Straight	STR	CHK, CHE, CHY
	Branched	BRA	

---

[31] Attribute SAT/UNS is not indexed for the superatom HEF itself. However, if HEF is searched with Match level Atom the attributes are considered for the specific structures.

System	Attribute	DWPIM Display	Attributes apply to
Number of Carbons	Low (1-6 carbon atoms)	LOW	CHK, CHE, CHY
	Mid (7-10 carbon atoms)	MID	
	High (11 or more carbon atoms)	HI	

Figure 64 illustrates the various descriptors associated with generic groups. The numbers above the fragment serve as unique identifiers for the assignment of the attributes to the respective fragment (examples A and C). Green numbers in general indicate the valency (examples B and C) and green numbers in combination with "+" or "-" indicate the charge.
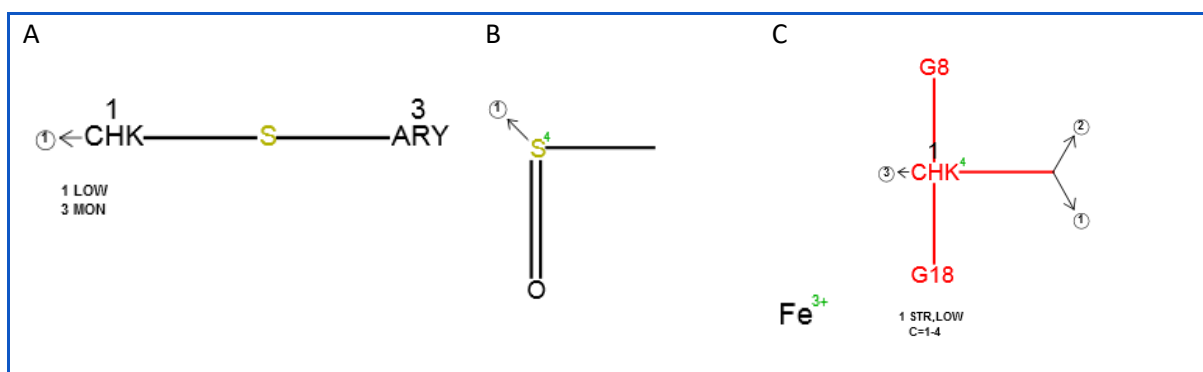
Figure 64. Examples of fragment descriptors.

Ring superatoms can carry attribute information which further specifies the ring system. In Table 22 an overview on chain and ring attributes associated with Markush hit structures is given.

Table 22. Attributes associated with chain and ring superatoms.

Attribute	Description	Appearance
RA	Number of atoms in ring	ARY, CYC, HEA, HET, HEF
NR	Number of rings	ARY, CYC, HEF
X	Any atom	ARY, CYC, HEA, HET, HEF
X=0	No other atom	ARY, CYC, HEA, HET, HEF
E	Number of double bonds (no longer in use, substituted by DBC)	-
DBC	Double bond count	CHE, CHY
Y	Number of triple bonds (no longer in use, substituted by TRC)	-
TRC	Number of triple bonds	CHY
BX	Not applicable (to be ignored)	-
N, O, S, etc	Elements according to the periodic table	HEA, HET, HEF

A typical example of a hit structure including a ring superatom with associated attribute information is shown in Figure 65. The associated attributes to the superatom HEA are listed in the brackets. If the given range of the heteroatoms N, O and S include the value 0 then they are to be considered optional.
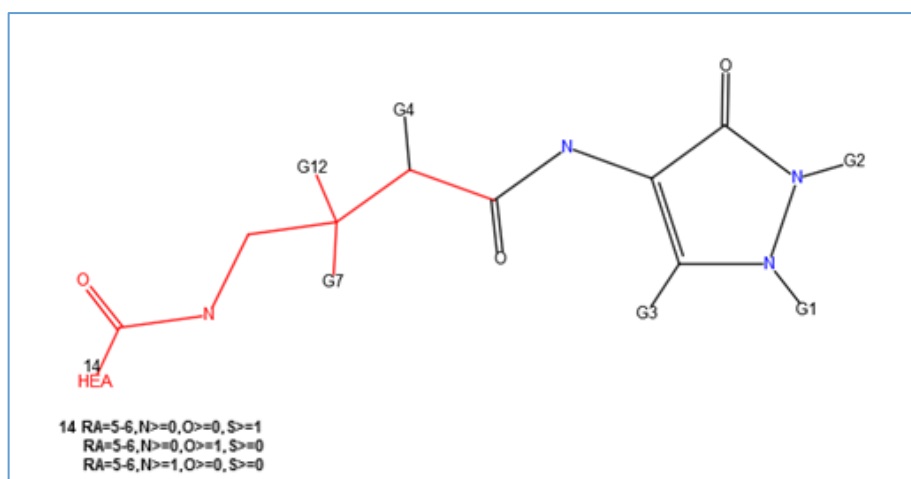
Figure 65. Hit structures containing ring superatom with attribute information.

3.10. Assembled Display

Assembled display: Displays a "hit" in a single structure by aligning enumerated structures to the Markush core structure (see Figure 66). Thereby only query-relevant fragments are automatically assembled in their correct placement in the chemical structure. The assembled view is the default display format which is automatically displayed for each result record after a structure search. Especially in the case of very generic Markush records with nested G-groups the assembled display can be very helpful for fast hit structure evaluation.
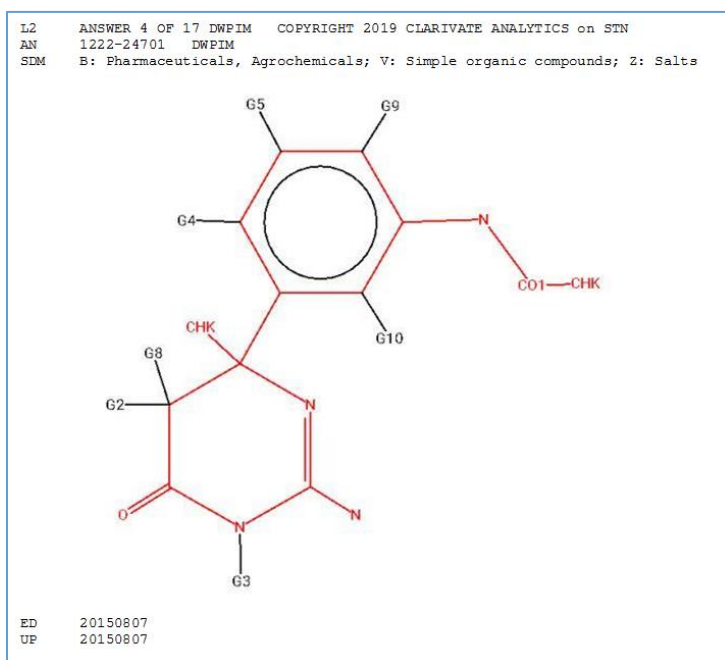
Figure 66. Example of an assembled display.

3.11. Full and Brief Display

Full display: comprises the complete record with all available information (*i.e.* core fragment G0, all pertaining G-groups with all possible alternatives) including highlighting (see Figure 67).
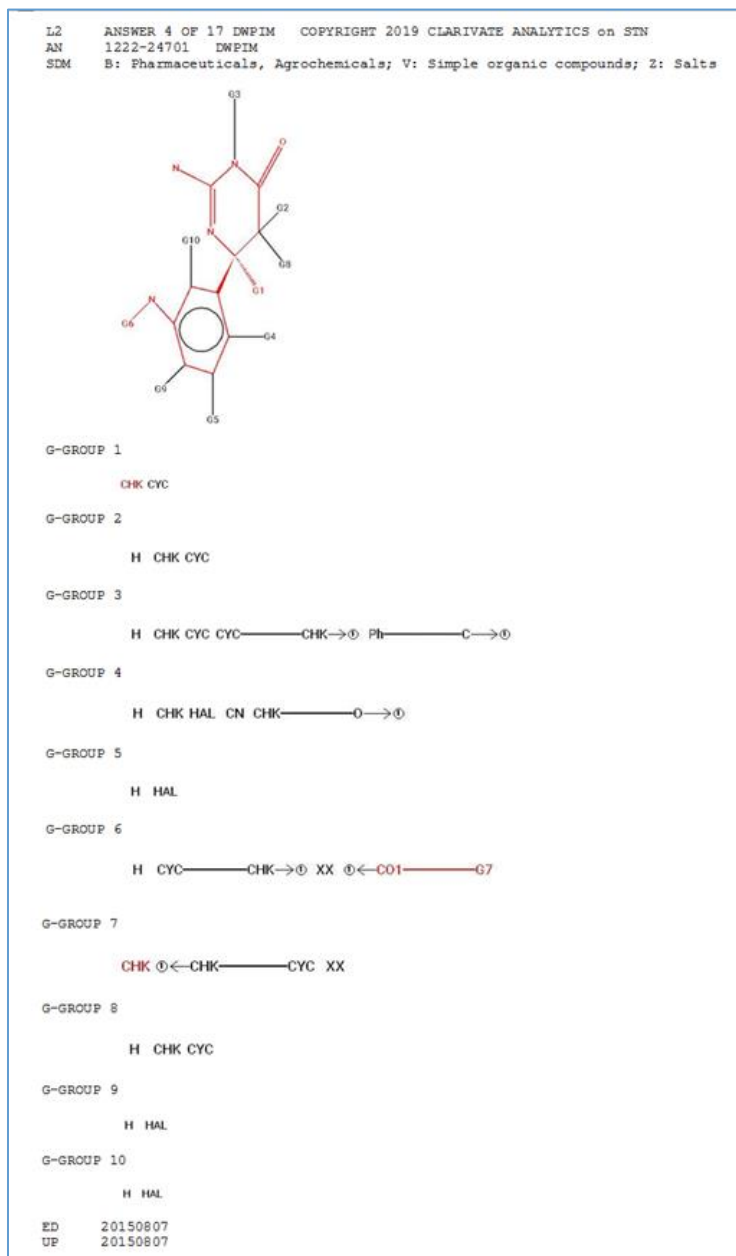
Figure 67. Example of a full display.

Brief display: Reduction to the "hit"; besides the core fragment G0 only generic groups that were part of the structure query are displayed including highlighting (see Figure 68).

```
L2      ANSWER 4 OF 17 DWPIM   COPYRIGHT 2019 CLARIVATE ANALYTICS on STN
AN      1222-24701   DWPIM
SDM     B: Pharmaceuticals, Agrochemicals; V: Simple organic compounds; Z: Salts
```

G-GROUP 1

   CHK CYC

G-GROUP 6

   H   CYC————CHK→① XX ①←CO1————G7

G-GROUP 7

   CHK ①←CHK————CYC XX

```
ED      20150807
UP      20150807
```
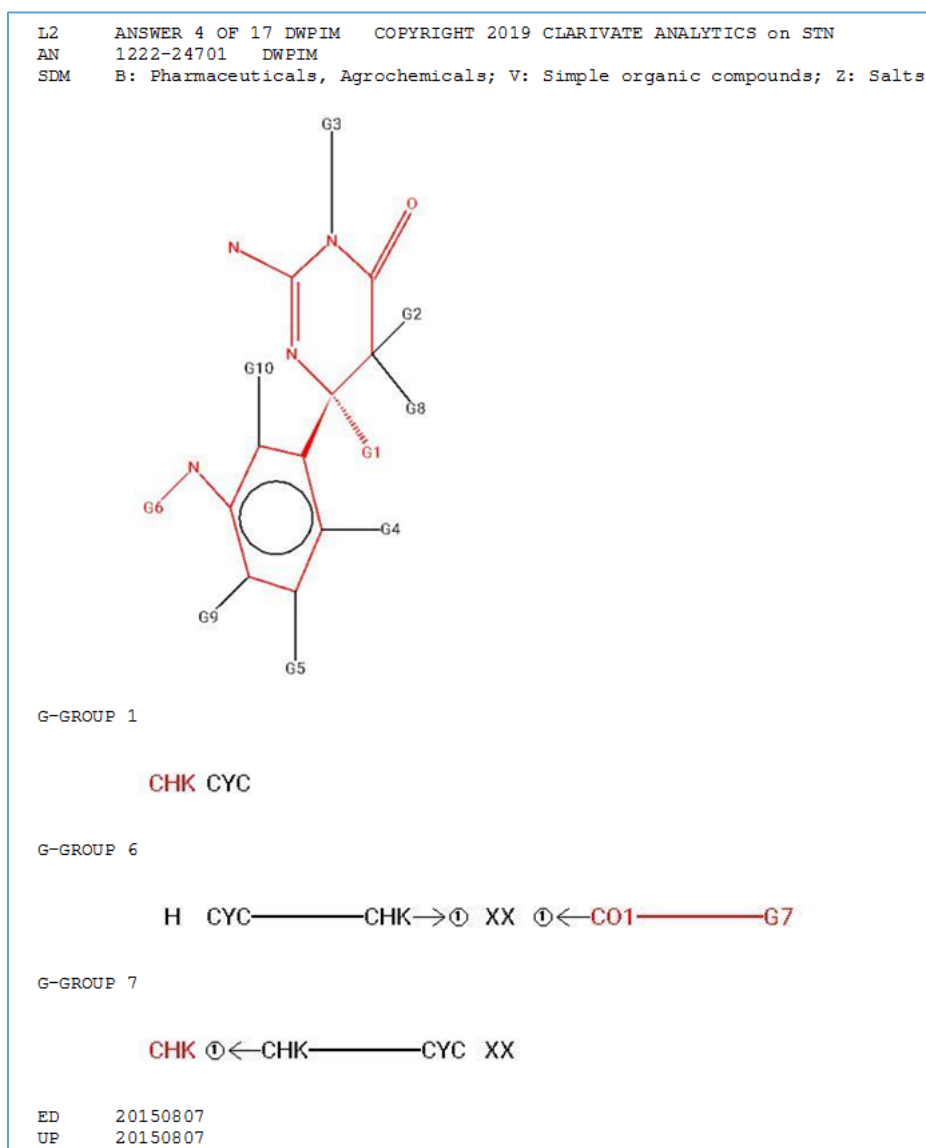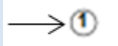
Figure 68. Example of a brief display.

3.12. Manual Assembling of Structures

In order to facilitate the assembly process of a Markush structure appropriate connector symbols are displayed at the respective fragment and atoms. Table 23 provides an overview on connector symbols.

Table 23. Overview assembly descriptors.

Symbol description	Symbol	Function
Blue circles:	①	connects with corresponding fragment carrying a white circle with the same number further down in the hierarchy, *i.e.* higher G-group numbers
White circles:	①	connects with corresponding fragment carrying a blue circle with the same number (usually further up in the hierarchy, *i.e.* lower G-group numbers)

Symbol description	Symbol	Function
Arrow	⟶①	indicates attachment point to the corresponding fragment

In Figure 69 an example is given for the basic assembly principles of a DWPIM Markush structure. The attachment point at G7 is unambiguous as the para position is predefined by pBe fragment. The white circle indicates that there is only one additional bond placed at the pBe fragment.
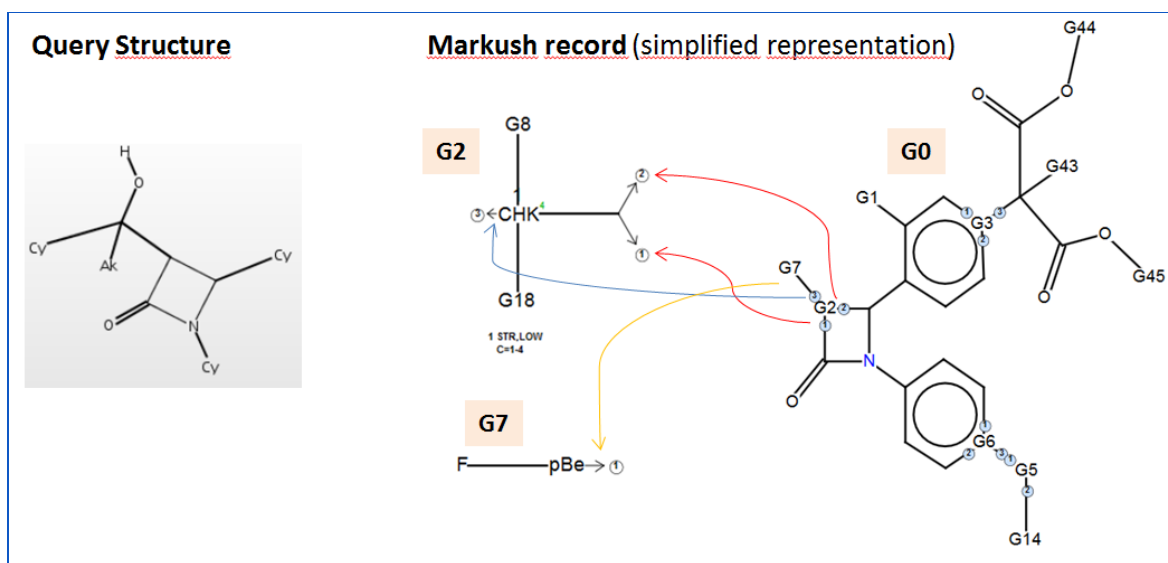
Figure 69. Assembly of Markush structures – basic principle.

In case of superatoms no specific attachment points can be assigned to them. As a consequence pertaining son connectors (white circle symbols) are not explicitly drawn (see G5 in Figure 69).

A G-group in indexed Markush structures can also contain nested G-groups as shown for G14 (contains G15, G46 and G47) in Figure 70. In this case the attachment of G14 at G5 is defined by the white circle symbol at G15 and at the CO1 fragment of G15. The "1" indicates that there is one bond going away from G15 group within G14. No corresponding connector is given at G5 (*i.e.* no blue circle) since the assignment of G14 to G5 is unambiguous (only superatoms listed in G5).
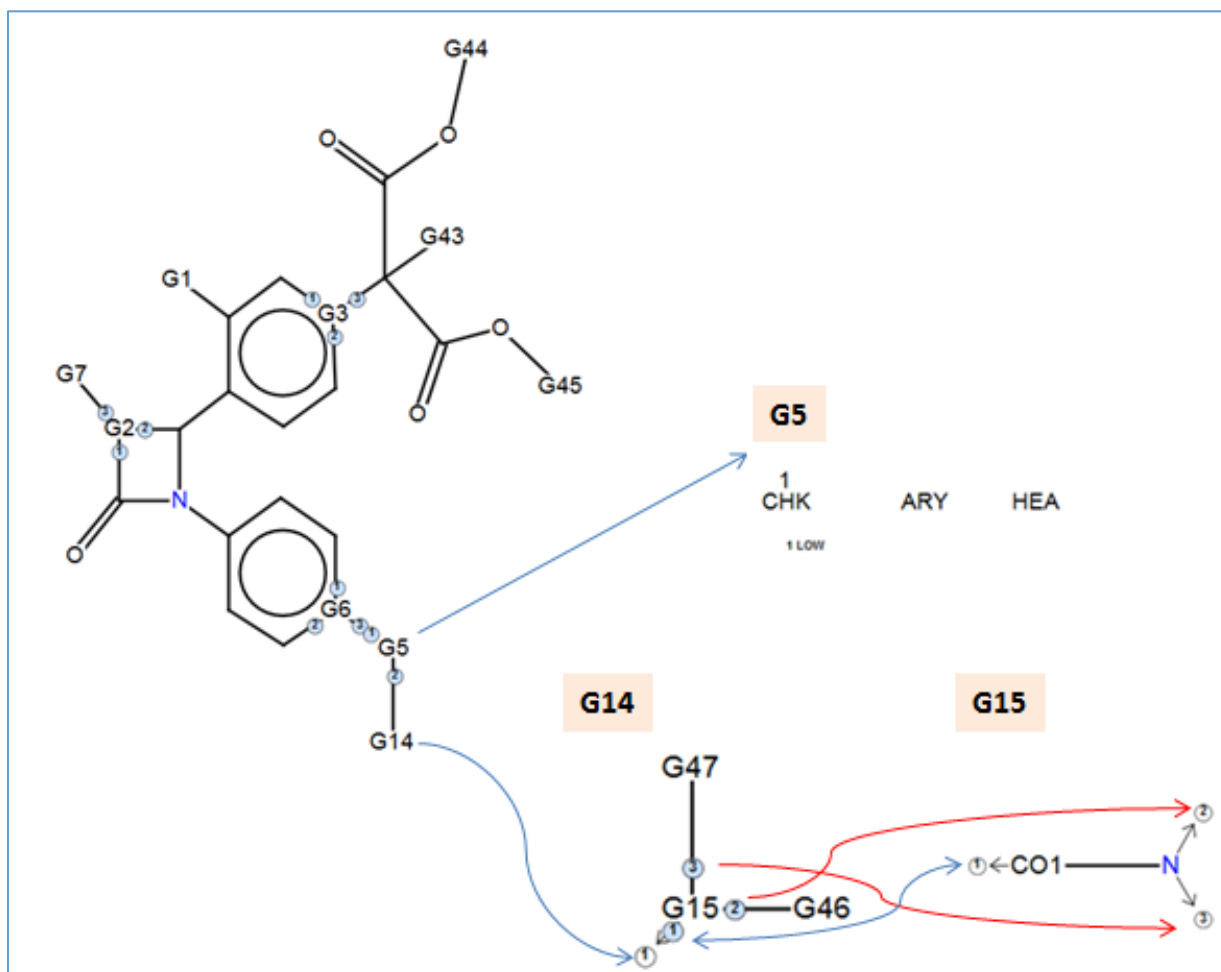
Figure 70. Assembly of Markush structures – nested G-groups.

Rings with different ring sizes are often represented in one compressed Markush ring fragment containing G-group(s). Assembling the structure fragment(s) within those G-group(s) delivers all ring systems covered by the Markush structure (see example in Figure 71).
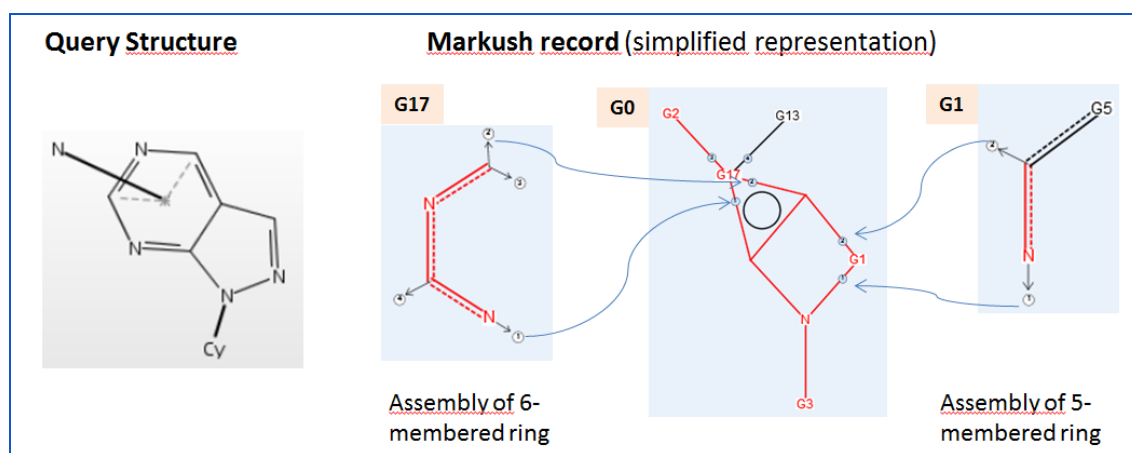
Figure 71. Assembly of Markush structures including ring expansion.

4. Search for Special Compound Classes and Chemical Groups

4.1. Organometallic Compounds

4.1.1. Metal Complexes and Coordination Compounds

For the representation of metal complexes three cases have to be distinguished:

a) The structure of the complex is known: the metal is bonded to the ligand
b) The structure of the complex is not known: multi-fragment representation with the appropriate charge on the metal and no charges on the organic part.
c) Complexes with π-bonded ligands (such as *e.g.* metallocenes): multi-fragment representation with the appropriate charge on the metal and no charges on the organic part (Figure 72).
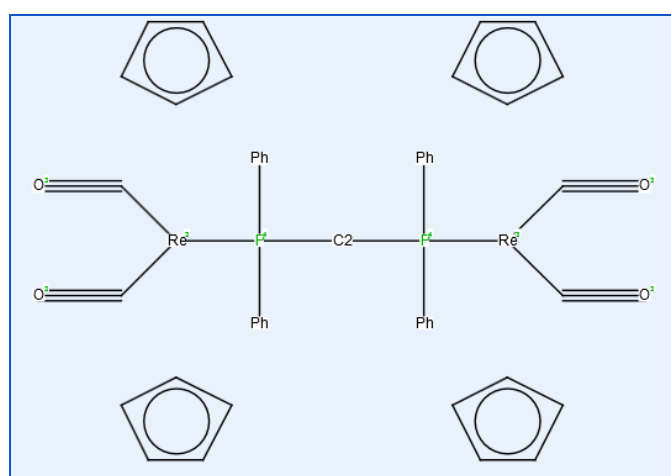
Figure 72. Example for a metal complex containing carbonyl and cyclopentadienyl ligands.

4.1.2. Representation of π-bonding Complexes

Metallocenes have to be searched with multiple fragments. The bonds in the cyclopentadienyl fragments must be defined as normalized bonds. The cyclopentadienyl fragments are drawn as neutral fragments, no negative charge is applied (Figure 73). In DWPIM representation the metal has an abnormal valency of zero as it is disconnected from the cyclopentadienyl rings. However, it is advisable to set the metal valency to "any" in order to take into account possible indexing inconsistencies. Furthermore the application of substance descriptor C (see chapter 3.8.4) can be very useful in connection with searches for metal complexes.
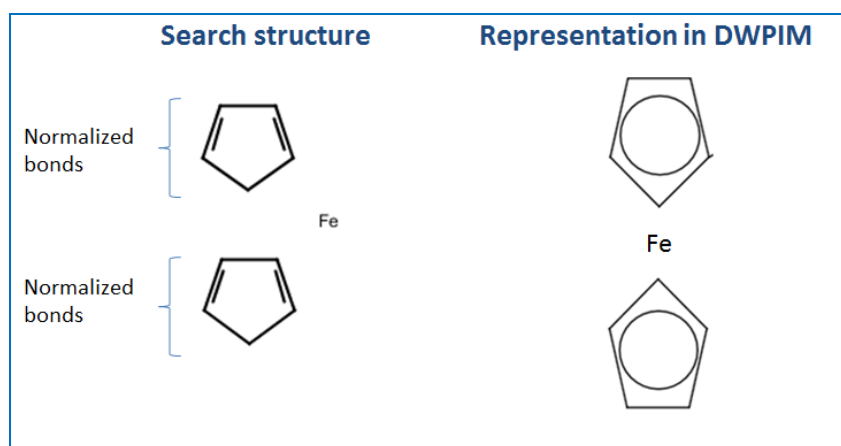
Figure 73. Representation of ferrocene.

4.1.3.   (Metallo)Phtalocyanines and (Metallo)Porphyrines

Phthalocyanines and porphyrines are fully represented with normalized bonds. Normalized bonds are either indicated by ring symbols or partly by dotted bonds as in the case of the corresponding metallo-derivatives (see Figure 74). Phthalocyanines and porphyrines are drawn with the metal bonded to all four N atoms and the valency on the metal is raised accordingly. While oxo ligands present are drawn doubly bonded to the metal, other anionic ligands are drawn disconnected. It is important to note that the indexing rules changed around the year 2000. Before year 2000 the metal was bonded to the ligand and after year 1999 no bonds between metal and ligand are drawn. In case of AMX there have never been bonds between metal and ligand.

9406-E1003/AN.M   9510-D5401/AN.M
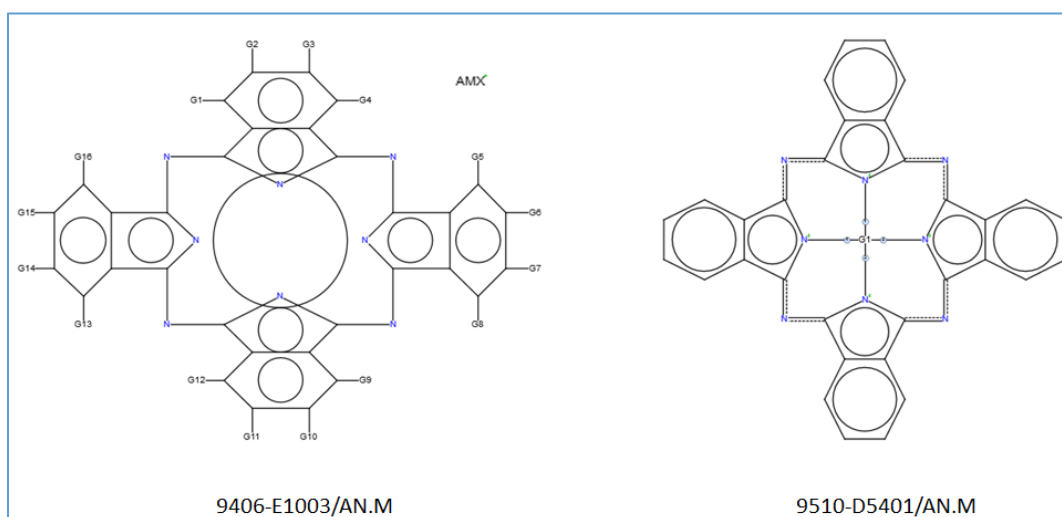
Figure 74. Representation of phthalocyanines and metallophthalocyanines.

4.1.4.   Metal Carbonyls

The carbon monoxide ligand in metal carbonyl complexes can be bound terminally to a single metal atom or bridging to two or more metal atoms. The most important binding modes are terminal, $\mu 2$

and μ3. In case of terminal mode the triply bonded oxygen has abnormal valency of 3 and in case of the μ3 mode the Carbons of the apical carbonyls have abnormal valency of 5.

Table 24. Representation of metal carbonyls.

Binding type	Representation	Abnormal Valency
terminal	M≡O	O: valency 3
μ2		none
μ3		C: valency 5

4.1.5. Acetylacetone Complexes

Acetylacetone and related complexes are indexed with normalized bonds for the carbonyl group and the carbon atoms between them. It is important to note that the bond type for the acetylacetone ligands as well as the node type of the metal must be set to ring/chain or ring in order to retrieve records as shown in Figure 75. It is important to note that the indexing rules changed around the year 2000. Before year 2000 the metal was bonded to the ligand and after year 1999 no bonds between metal and ligand are drawn.

Figure 75. Representation of Cu(acac)$_2$.

4.1.6. TCNQ Complexes or Salts
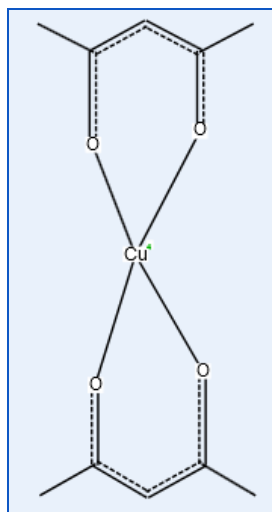
Tetracyanoquinodimethane (TCNQ) complexes are represented as multi-fragment representations with the appropriate charge on the metal and no charges on the organic part.
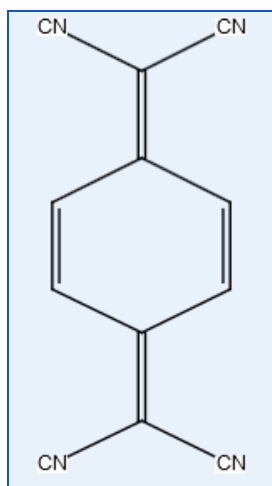
Figure 76. Tetracyanoquinodimethane part in complexes or salts.

4.2. Polymers (other than Peptides)

In principle, both addition and condensation polymers can be indexed in DWPIM. It is important to consider that polymers are for the most part only indexed in DWPIM if the patent is classified in CPI Sections B and / or C. Polymers in Section E (which are not also classified in Section B and / or C) will not be indexed in DWPIM.

- Addition polymers are indexed in terms of the corresponding monomers (each monomer is indexed with the roles M and Q applied)[32]. In case of polymers which contain multiple monomers the monomers are either indexed separately or by using one Markush structure with respective G-group(s). For example ethylene and vinyl chloride could also be indexed as C=C-G1 with G1 = H, Cl. A typical example for an indexed monomer of an addition polymer is given in Figure 77 (structure 1244-13905).

- Condensation polymers will be indexed as the structural repeat unit (2 complete repeats where possible and *M100* type text notes added to identify the repeat unit) whenever this information is provided in the patent. If only a partial structure is given this one will be indexed and the superatom POL will be placed at either end. In cases where no structure is given the starting material is indexed in the same way as for addition polymers. Condensation polymers are indexed with either the substance descriptor F (synthetic) or N (natural), plus any other polymer substance descriptors which apply from Table 19 (chapter 3.8.4). If it is not possible to generate a structure the starting material is indexed instead, however, substance descriptor F is not assigned in these cases. A possible search strategy for these cases is to search for monomer 1 and for monomer 2, execute a *refx* operation to DWPI and apply the Derwent role Q (Product defined in terms of starting materials, also see chapter 3.8.1). A typical example for an indexed condensation polymer based on its structural repeating unit is given in Figure 77 (structure 1250-70601).

---

[32] It should be noted that polymer substance descriptors from Table 19 (chapter 3.8.4) are not indexed for these substances

▪ Macromolecule-Drug Conjugates (e.g. polymer drug conjugates, antibody drug conjugates) can be searched by using the superatom POL whereby POL stands for any macromolecule.

As a consequence addition polymers should be searched for the monomers while condensation polymers should be searched for both repeating units and monomers to ensure complete retrieval.
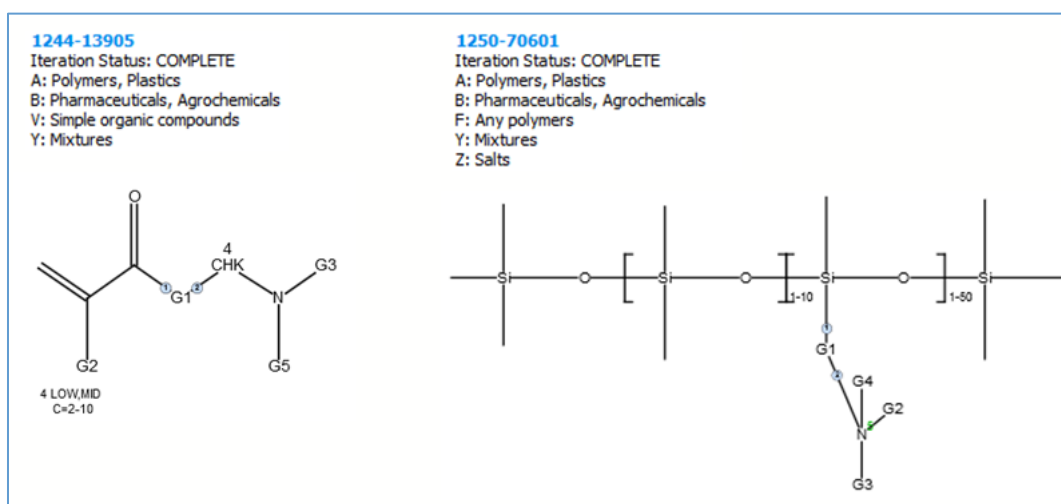
Figure 77. Examples for addition polymers (1244-13905) and condensation polymers (1250-70601). For reasons of simplicity only the core fragment is shown.

In Section E dye patents which contain polymeric dyes the chromophore is indexed up to the point where it bonds to the polymer backbone and this is then replaced by POL.

4.3. Polymer or Oligomer (Peptides, Saccharides)

In

Table 25 the distinction between oligopeptide or oligosaccharide and polypeptide or polysaccharide, respectively, and the distinction towards polymers is given. The BC definition refers to the definition used when indexing pharmaceutical and agrochemical patents (Sections B and / or C). The E definition refers to the definition used when indexing general chemistry patents (Section E). If a patent is classified in section E as well as section B and / or section C the BC definitions are used.

Further polymer descriptors are described in Table 19 in chapter 3.8.4.

Table 25. Definition of polymers and oligomers.

Substance	Substance descriptors	BC definition	E definition
Oligopeptide	VP	3 amino acids	3 amino acids
Polypeptide	P	≥ 4 amino acids	≥ 4 amino acids
Oligosaccharide	L	3-6 sugar units	3-9 sugar units
Polysaccharide	N	≥ 7 sugar units	≥ 10 sugar units[33]
Other oligomer	L	3-8 repeat units	3-9 repeat units
Other polymer	F	≥ 9 repeat units	≥ 10 repeat units[34]

4.4. Deuterated and Tritiated Compounds

Deuterated or tritiated compounds can be directly searched by using the respective D and T symbol of the structure editor (Figure 78).
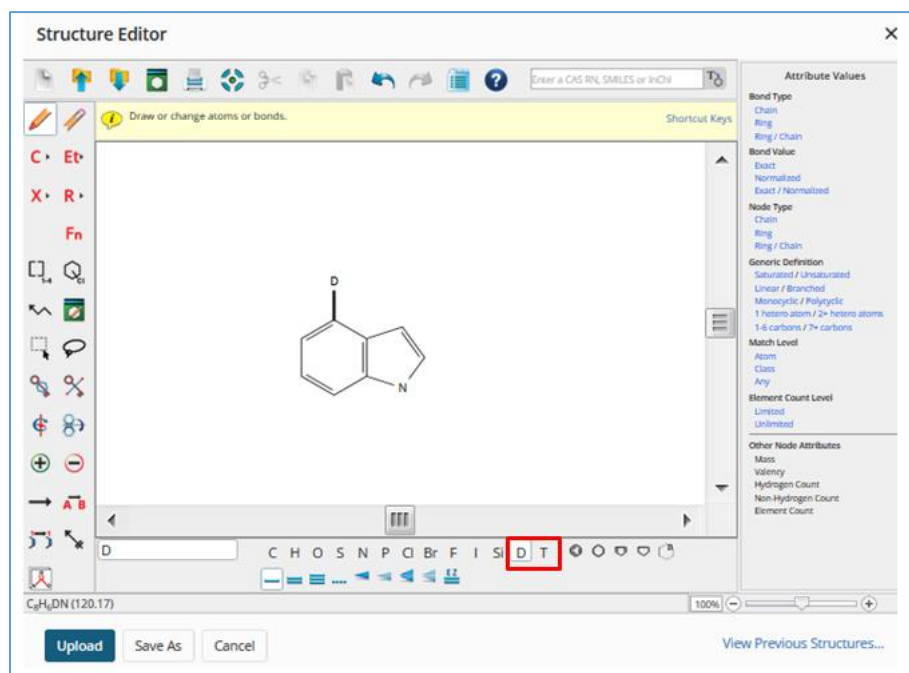
Figure 78. D and T symbols for Deuterium and Tritium.
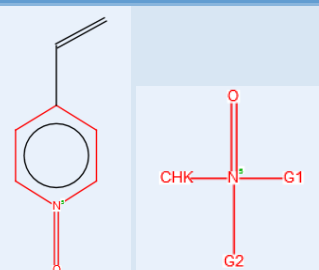
4.5. Amine-N-Oxides

Amine-N-oxides contain the functional group R_3N^+-O^-. The query structure can be defined with or without charge separation. In the display the Nitrogen is preferentially drawn with connectivity of five.

---

[33] Not indexed unless part of a dye molecule
[34] Not indexed unless part of a dye molecule
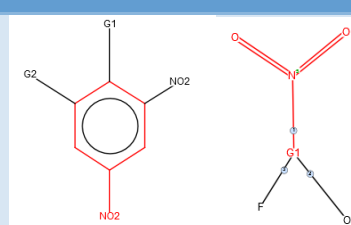
Table 26. Representation of amine-N-oxides.

Search Structure	Preferred Representaion	examples
R1R2R3N=O or R1R2R3N⁺-O⁻	R1R2R3N=O	

4.6. Nitro Group

Nitro groups must be searched with a connectivity of five for the Nitrogen. Therefore two double bonds to the Oxygen atoms have to be drawn. Charge separation is not allowed for the query structure. In the display the Nitrogen is preferentially presented with connectivity of five.
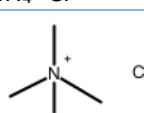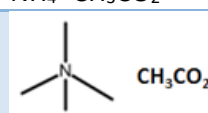
Table 27. Representation of nitro groups.
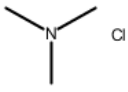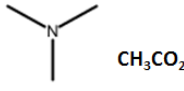
Search Structure	Preferred Representation	Examples
$R1N(=O)_2$	$R1N(=O)_2$ or $R1NO_2$	

4.7. Representation of Salts

For the representation of salts a multi-fragment representation consisting of a separate anionic part and a separate cationic part is used. The type of representation of the anionic and cationic parts, in the charged or uncharged form, is dependent on the organic or inorganic nature of the parts and dependent on the presence or absence of Hydrogen. Metals are never shown covalently bonded to inorganic ions (with exception of the chromate, dichromate, manganate and permanganate ions).

Table 28. Representation of salts derived from inorganic and organic bases.

Base (cation)/Acid (anion)	Inorganic Acid		Organic Acid	
Inorganic base without hydrogen	K^+ Cl^-		K^+ CH_3CO_2	
Inorganic base with hydrogen	NH_4^+ Cl^-		NH_4^+ CH_3CO_2	
organic base without hydrogen				

Base (cation)/Acid (anion)	Inorganic Acid	Organic Acid
organic base with hydrogen		

4.8. Zwitterionic Compounds (Inner Salts)

Zwitterionic compounds containing a positive and a negative electrical charge are indexed in neutral form.
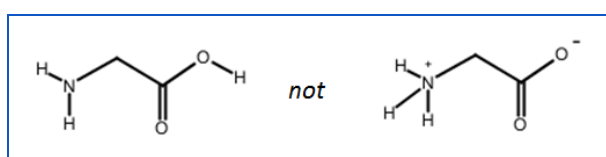
Figure 79. Representation of zwitterionic compounds (e.g. glycine).

5. Special Search Issues

5.1. Effect of Free Sites on Search Results

In STN it is possible to define the number of non-hydrogen attachments. This is the sum of all non-hydrogen substituents of the given node (including those that have been defined explicitly by the user). A complementary view is the concept of free sites: this is the difference between the valency of the node and the number of explicit substituents of the node. In STN, the number of non-hydrogen attachments can be specified by the user while the number of free sites is automatically detected by the software. However, it is possible to switch free sites on or off either locally by using the atom or ring lock tool or globally by using the search mode substructure search (SSS) vs. closed substructure search (CSS).

When specific or generic nodes have free sites the search may result in additional nodes which might come unexpectedly at a first glance. Consider the following example from normal substructure searching:

R−C* where R is a specific rest which has been closed for further substitution, C has one free site and MLE=ATOM. This may yield R−CH3 or R−CH2−CH3, but also R−CH2−CH=CH2. With match level CLASS one may also obtain the compounds R−CHK, R−CHE or R−CHY.

Moreover, the retrieved chains may have more substituents than the number of free sites in the query chain, *e.g.* R−CHE−Cl is another hit structure.

Such an extension of search results is possible for chains, rings, and for single nodes. We can distinguish the following cases:

- Generic chains (1 free site): MLE = ATOM or CLASS
- Specific chains (1 free site): MLE = CLASS
- Generic rings (2 free sites): MLE = ATOM, CLASS, and ANY
- Specific rings (2 free sites): MLE = CLASS

- Single specific node (2 free sites):   MLE = CLASS

The conditions for an extension of the answer set are: chains must have one free site, rings must have two free sites and the ring is not isolated, and the single specific node must have two free sites.

Chain Extensions

Chain extensions may occur for generic chains for match level ATOM and CLASS and for specific chains for match level CLASS. In both cases it is necessary that the chain has one free site.

The complete picture with all possible paths is shown in Figure 80. Green paths are a consequence of normal substructure searches, red paths apply to searches with MLE=CLASS, and blue paths apply to searches of generic nodes with MLE=ATOM. The double arrows in black denote the standard hierarchical relationships between generic and specific nodes. Finally, it should be noted that the extensions may also change the parameters of the original chain, *e.g.* linear/branched, saturated/unsaturated, and the number of carbon atoms.
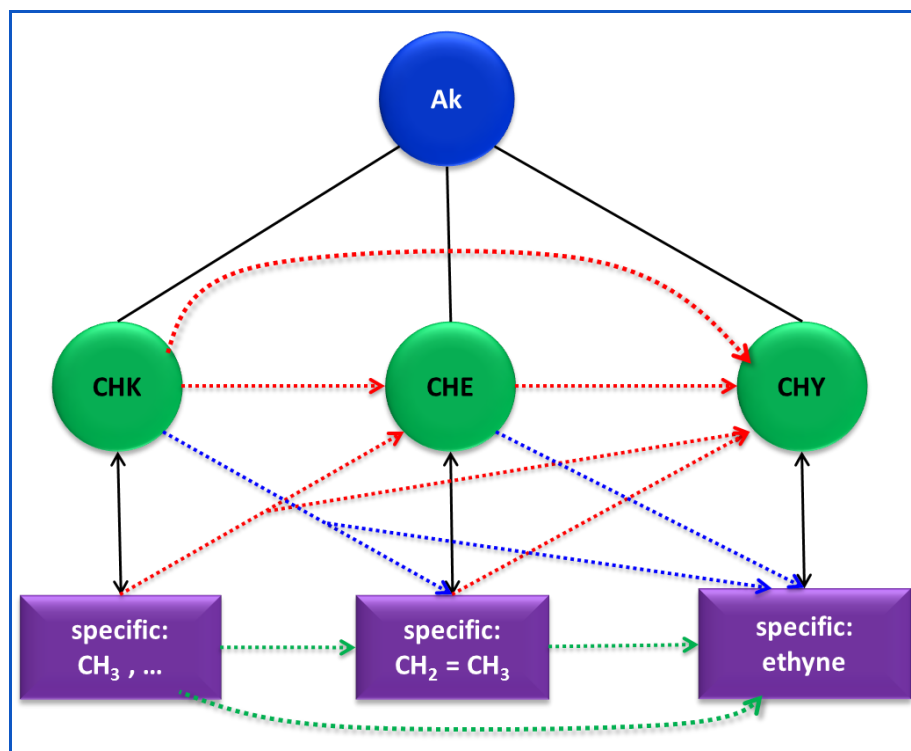
Figure 80. Different paths for searches with free sites on carbon chains.

Ring Expansions

Ring expansions may occur for generic rings for match level ATOM and CLASS[35] and for specific rings for match level CLASS. In both cases it is necessary that the ring has two free sites.

The complete picture with all possible paths is shown in Figure 81. Green paths are a consequence of normal substructure searches, red paths apply to searches with MLE=CLASS and blue paths apply to

---

[35]   Ring extension with match level ANY will be discussed in chapter 5.2.

searches of generic nodes with MLE=ATOM. To compare the situation with MARPAT we have added a grey dotted line: only Cb is extended to Cy.
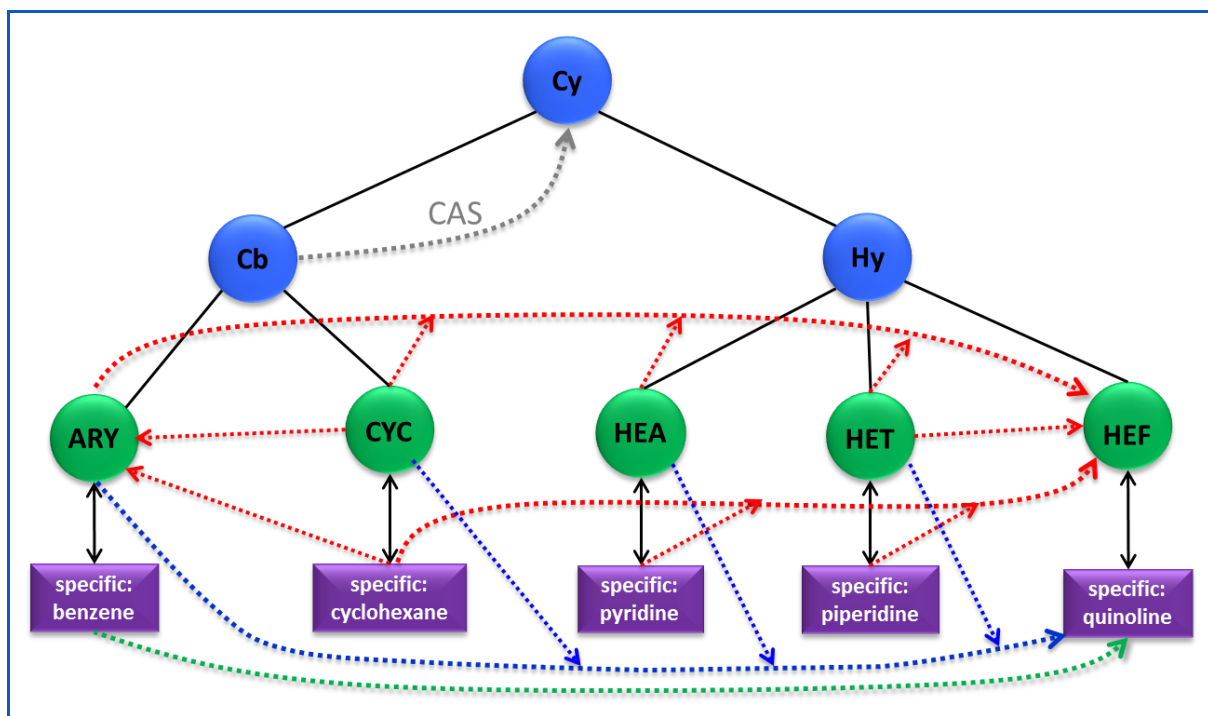
Figure 81. Different paths for searches with free sites on cyclic compounds.

The free sites must reside on the same ring but it is not necessary that they are on adjacent nodes of the ring. When the nodes are not on adjacent nodes one may retrieve also bridged ring systems like norbornane.

Atom Extensions

The hierarchy of chemical elements and the corresponding superatoms is described in Figure 55 with the construction of the A-Q tree. If these elements have two free sites and the node attribute is Ring (or Ring/Chain) it is possible to join the free sites together to form a ring system. Hence, searching for any chemical element or superatom will retrieve additional ring systems. Let us consider an example: a search for Fe (2 free sites, node attribute: Ring) with Match Level CLASS will retrieve:

- Fe, TRM, and MX (or M)
- Specific Fe in mono heterocycles, heteroaryls, and fused heterocycles (each with an element count FE = 1)
- HEA, HET, and HEF (each with an element count Fe ≥ 1)

The same concept applies to all other elements and the corresponding superatoms. Also, for the search of the chemical element C (2 free sites, node attribute: RING) with match level CLASS one may obtain:

- Specific carbocycles and aryls
- Specific heterocyclic compounds (corresponding to the superatoms HEA, HET, and HEF)

- ARY, CYC, HEA, HET, and HEF

5.2. Search for Hybrid Rings Containing the Superatom XX

Ring systems may be described as a specific ring system or as a ring superatom. However, sometimes these descriptions are not sufficient to describe the Markush structure in a patent document. Notations like "two groups which can form a ring" or "N is part of a ring" require additional options for the description of ring systems. Derwent has chosen to use the superatom XX to describe these cases. The possible ring systems are described in Figure 82. As a consequence we obtain a hybrid ring system consisting of at least three nodes. At least one node is a specific node (chemical element) and at least one node is the superatom XX.
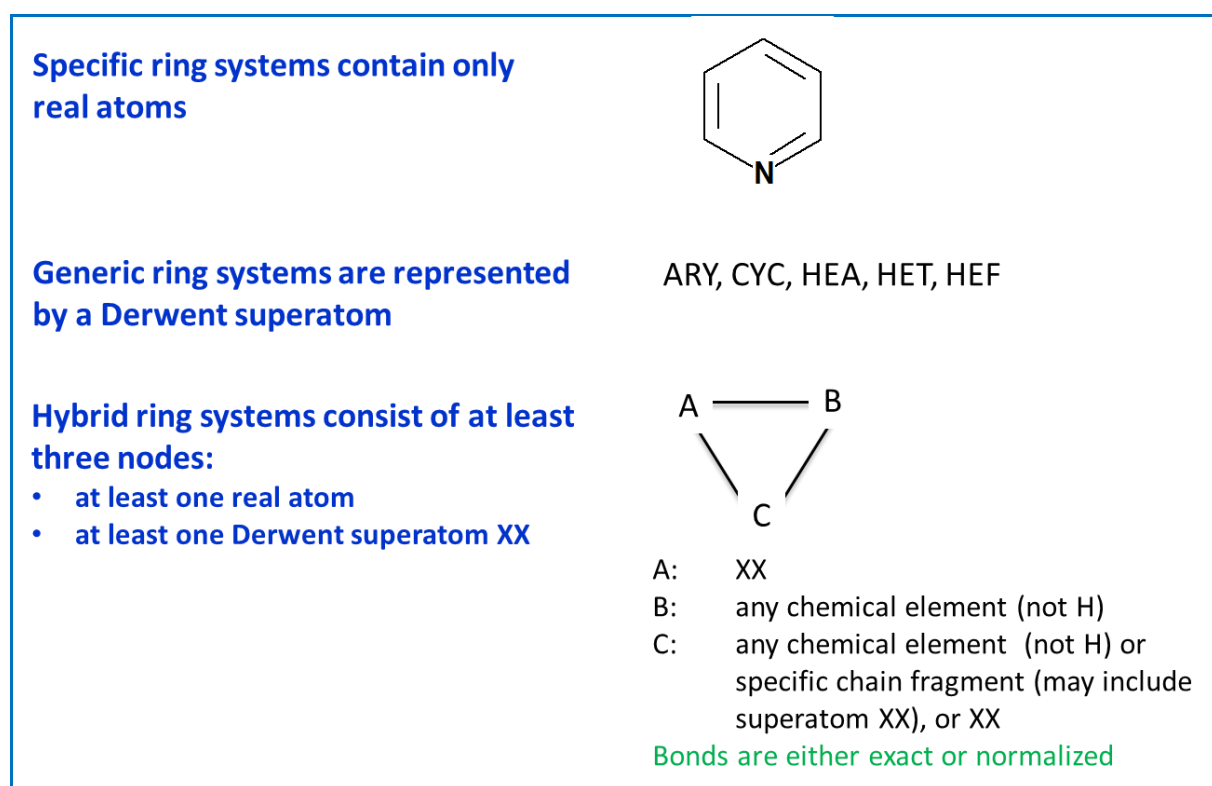
Figure 82. Possible ring systems in DWPIM.

Some examples of hybrid ring systems from real Markush structures are shown in Figure 83. The fragments show rings with C or N with either single or aromatic bonds. In the latter case it must be either a five-membered or a six-membered ring. The last fragment stems probably from the description of a benzene ring with two R-groups forming a ring.
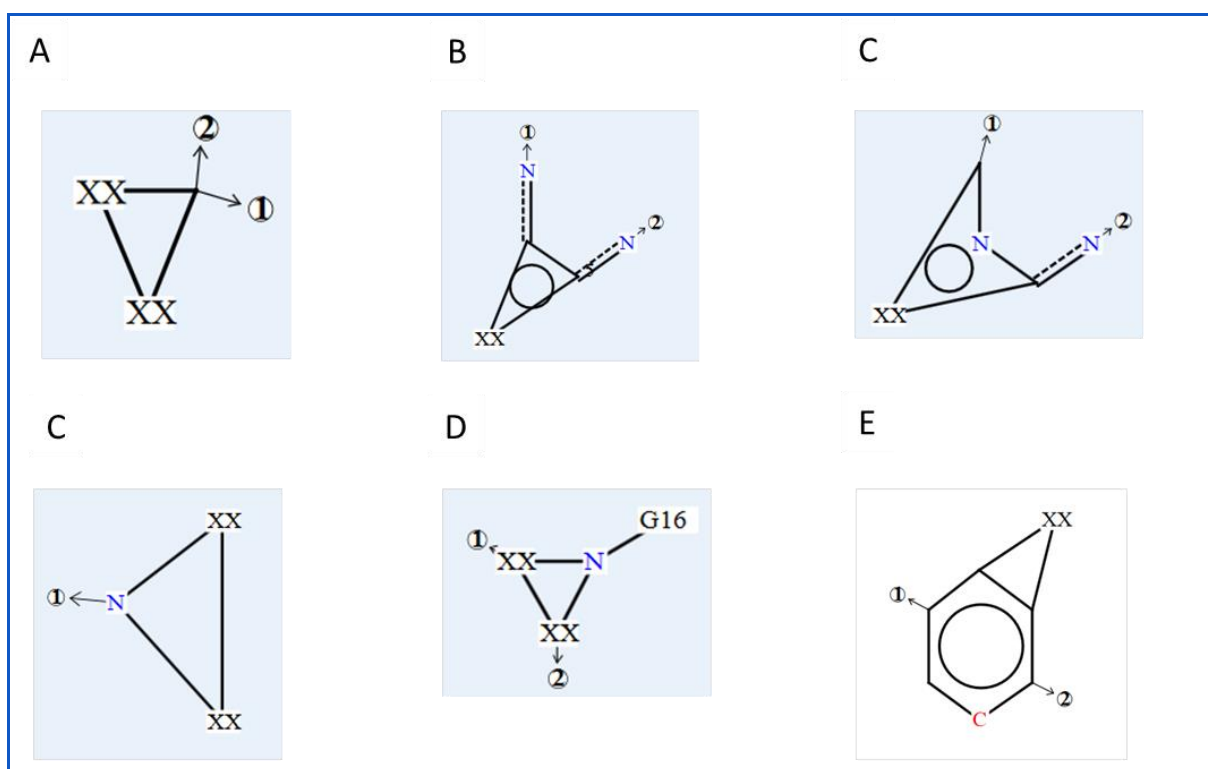
Figure 83. Examples of ring systems which contain the superatom XX.

Normally, all elements of a specific ring system must have the same match level. Examples for these types of searches are illustrated in Figure 84.
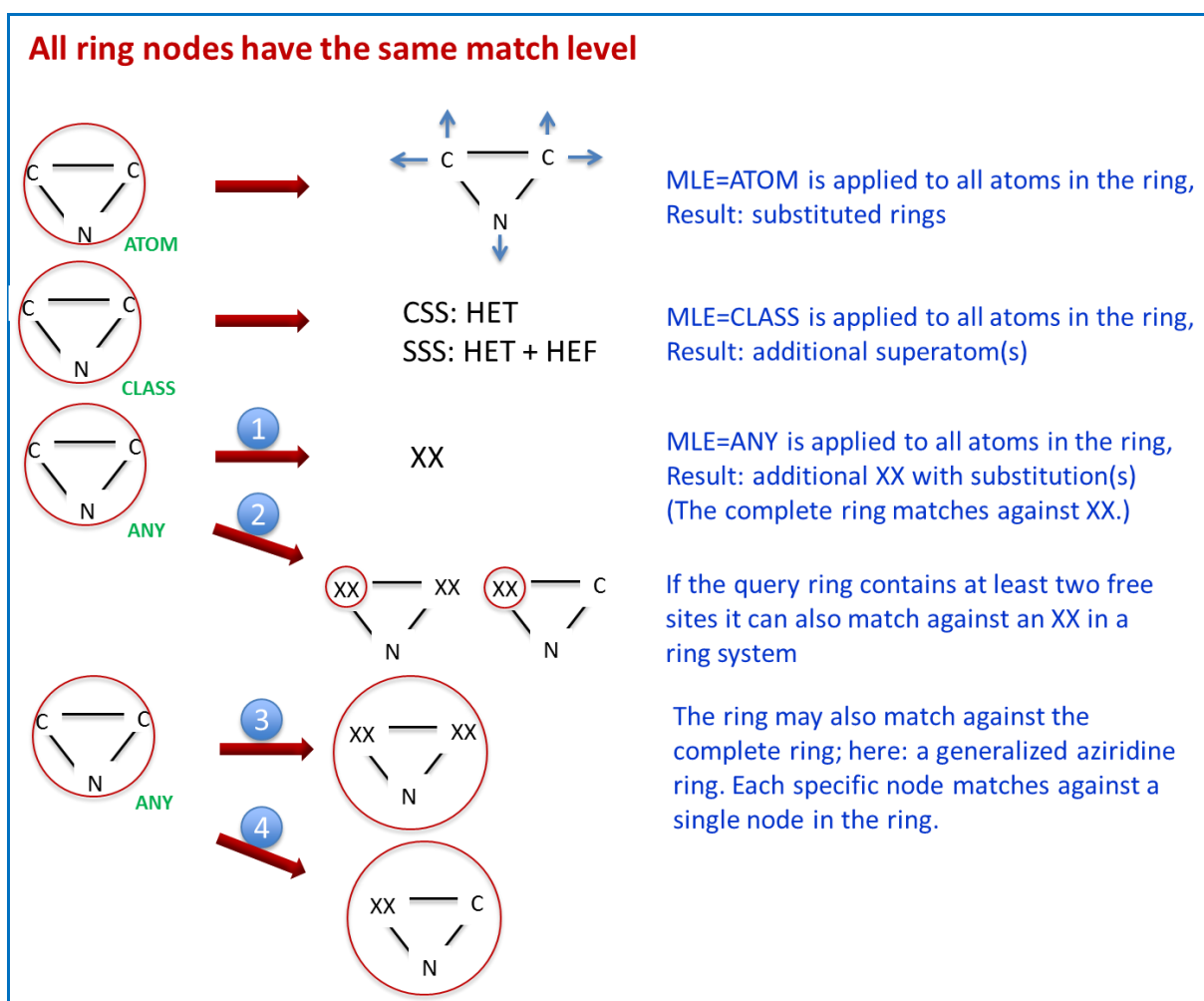
Figure 84. Standard searches for ring systems.

In order to search for hybrid ring systems it is necessary to apply two different match levels to the atoms of a ring system whereby only the combination of ML ATOM and ML ANY is allowed (see Figure 85). Atoms with ML ANY can be substituted by XX. Other combinations of match levels such as ATOM + CLASS, CLASS + ANY und ATOM + CLASS + ANY are not allowed and lead to an alignment of the respective ML's by applying the following rule:

- The most common type of ML is determined. If it is ML ANY, the second most common will be applied for all ring atoms.

- For equal numbers of assigned ML's the lower ML is assigned, e.g. a rig with 3 Atoms ML Atom and 3 Atoms ML Class the overall ML Atom will be assigned.

Ring Contraction does not take place. Ring size of query structure equals ring size of records found. In case of mixed ring systems with specific and generic nodes the specific atom of the query which is set on ML Any will be substituted by XX in the hit records.

Note: Except from XX, metal superatoms and HAL can also be part of specific-generic ring systems (ML Class for the superatom / ML Atom for remaining atoms).
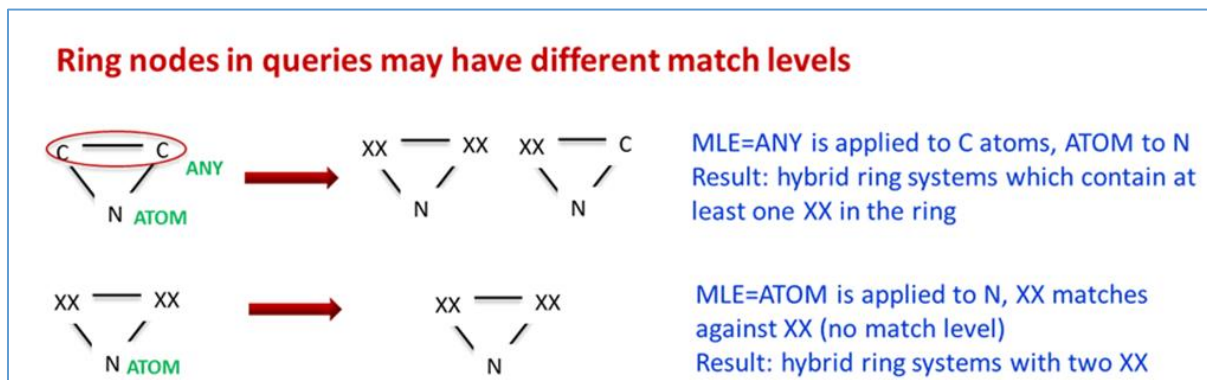
Figure 85. Search for ring systems with different match levels.

5.3. Query structures with Carbon Atoms adjacent to corresponding generic nodes

In case of structures which contain carbonyls adjacent to carbon chains the matching is influenced by the search direction. Therefore usage of query structures containing Carbon atom(s) adjacent to chain generic nodes (Ak, CHK, CHE, CHY) is not recommended. Since there is no overall chain contraction implemented the query may lead to different results dependent on the the search direction determined by the search engine. The search direction of the system for a particular query structure cannot be fully predicted and may be random.

Considering the rules for generation of spin-off nodes (chapter 3.4) the issue is explained in more detail:

1) For real nodes Spin-offs are always generated. Adjacent real nodes as well as generic nodes of the same type (chain or ring) are included.
   E.g. –CO-CH2-CH3:
   Starting from CO or CH3 → CHK (C=3) is generated.
   Starting from CH2 → CHK (C=2) in direction to CO or CH3, respectively.

2) For generic nodes no Spin-offs are generated. Consequently there is no overall ring contraction.
   E.g. –CO-CHK-CH3:
   Starting from CO or CH3 → CHK (C=3) is generated. Chain contraction.
   Starting from CHK → no Spin-off, no chain contraction.

For this reason it is not recommended to use query structures with chain nodes (CHK, CHE, CHY) adjacent to a Carbon.

In the following two examples the implication for the matching process is shown.

Example 1:

Example Query –CO-CH2-CH3

1) Search Direction from left to right

Query: -CO-CH2-CH3   Spin-off: -CO-CHK(C=2)

Target: -CO-CHK

2) Search Direction from right to left

Query: CH3-CH2-CO-   Spin-off: CHK(C=3) O

Target: CHK-CO-

Figure 86. Example 1 for dependency on search direction. Green arrows indicate match, red arrow indicates no match.

Example 2:

- Search direction starting at imidazolinone N
- No hit because there is no match for CO1 moiety

- Search direction starting at NH2
- Hit (matching is complete)
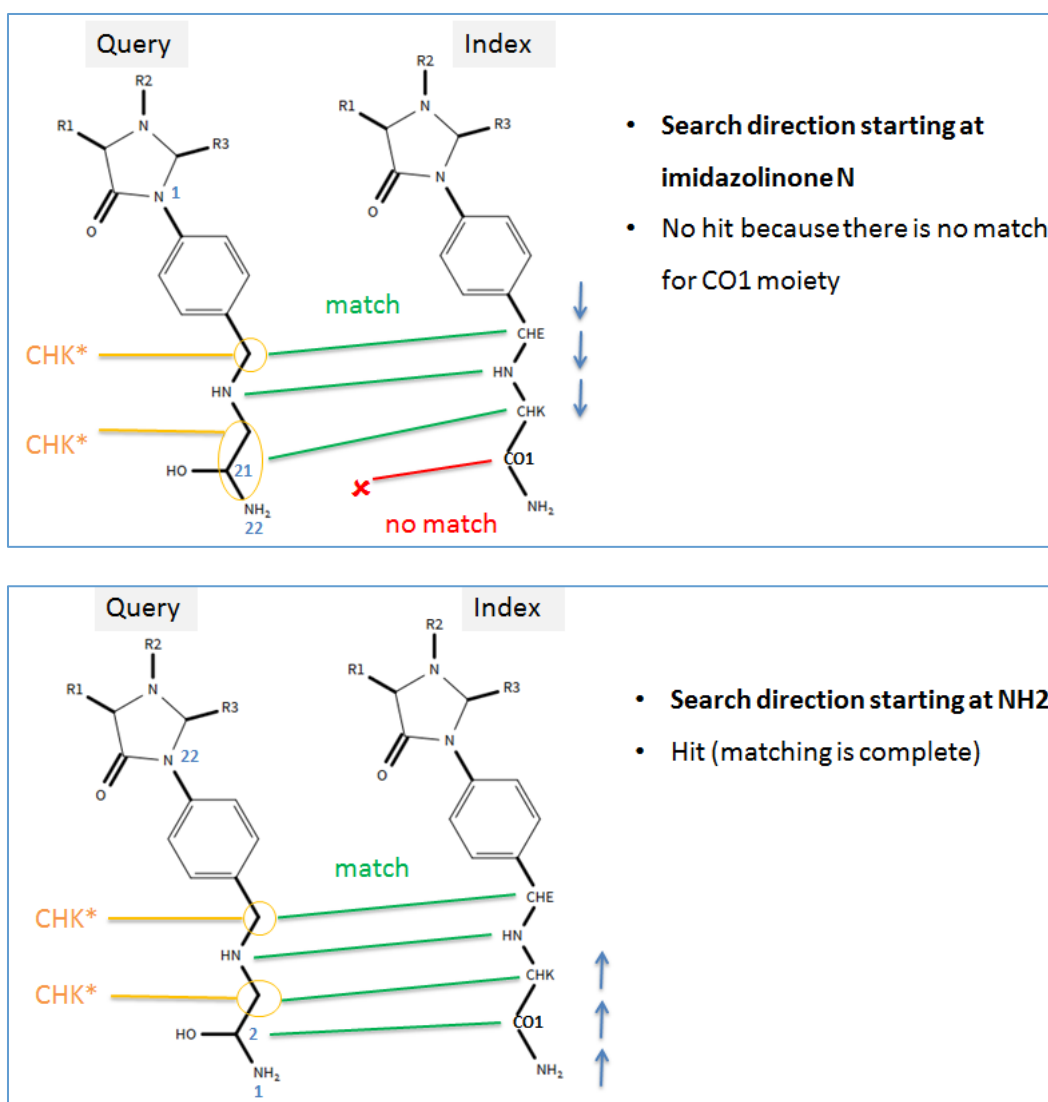
Figure 87. Example 2 for dependency on search direction. Blue arrows indicate search direction.

More details on the fact that the search direction cannot be fully predicted:

There are two rules which are implemented but which are not sufficient to cover every possible case.

> 1) If heteroatoms are present the search engine will start with the highest substituted heteroatom.

> 2) In case of equally substituted hereoatoms the search engine will start with the heteroatom with the highest atomic number.

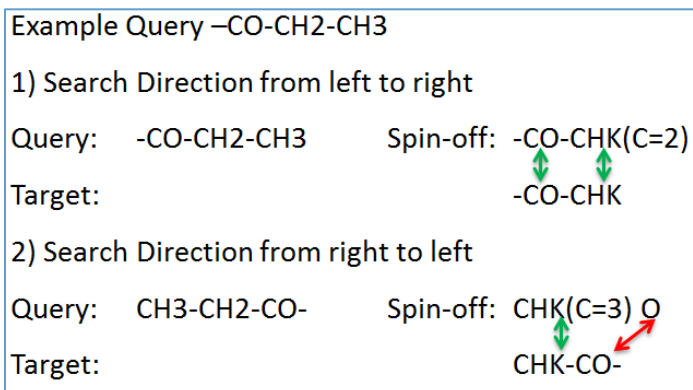If there is no differentiation achieved by applying those rules the numbering assignment for atoms cannot be fully predicted.

5.4.   Markush Generic Terms

Markush fragments in a patent are often abbreviated by a text term, e.g. alkoxy. These text terms are called Markush terms and they must be translated into structure fragments consisting of superatoms, shortcuts, specific elements, and associated free sites. Table 29 lists the Markush terms and the corresponding definitions as they are used by Clarivate Analytics for indexing in DWPIM. As a simple example the term alkoxy is represented by CHK-O★ (where "★" indicates a free site) and an alkoxy-phenyl is represented as CHK-O-Ph. Another example is haloalkyl. This is represented by two terms:

- HAL-CHK★ - an alkyl group with a single halogen attached, plus
- HAL-CHK★-R $_{mu=1,15}$ (R = H,HAL) – an alkyl group with 1 – 16 halogens attached (mu is the abbreviation for multiplier).

Issue 9: It should be noted that sometimes only the first term is indexed for haloalkyls.

The next example is a heteroalkyl which is represented as CHK-XX-CHK★. In this case the heteroatom is not specified as a list of elements but represented by the unspecific superatom XX. As a final example we take oxaalkyl. There are two definitions given

- CHK-O-CHK★ - an alkyl ether, plus
- CHK-O-(CHK-O)$_{n-m}$-CHK★ - an alkyl ether with a repetition group (CHK-O) which may occur between n and m times.

Table 29. Markush patent terms and the corresponding DWPIM indexing terms.

Markush Patent Term	Derwent Markush Resource Index Term
Acyl	ACY plus CHK-CO1★
Alkanoyl	CHK-CO1★
Alkanoyloxy	CHK-CO1-O★
Alkaryl	CHK-ARY★
Alkarylene	★CHK-ARY★ plus CHK-ARY★★
Alkenyl	CHE
Alkenylene	Divalent CHK

Markush Patent Term	Derwent Markush Resource Index Term
Alkoxy	CHK-O★
Alkyl	CHK
Alkylcarbamoyl	CHK-N-CO1★
Alkyloxysulfonyl	CHK-O-SO2★
Alkylene	Divalent CHK
Alkylidene	(CHK-) ★C-(-R) where R is H or CHK, or divalent CHK
Alkyloxycarbonyl	CHK-O-CO1★
Alkynyl	CHY
Alkynylene	Divalent CHY
Amido	CHK-CO1-N★
Aralkyl	ARY-CHK★
Aralkylene	★CHK-ARY★ plus ★★CHK-ARY
Aroyl	ARY-CO1★
Aryl	ARY
Arylene	Divalent ARY
Carbamoyl	N-CO1★
Carboalkoxy	CHK-O-CO1★
Chalcogenide	O, S, Se, Te
Cycloalkyl	CYC
Halo	HAL
Haloalkyl	HAL-CHK★ plus HAL-CHK★-R$_{mu=1,15}$ (R = H,HAL)
Heteroalkyl	CHK-XX-CHK★
Oxaalkyl	CHK-O-CHK★ plus CHK-O-(CHK-O)$_{n-m}$-CHK★
Oxylalkylene	★CHK-O-CHK★ plus ★CHK-O-(CHK-O)$_{n-m}$-CHK★
Phospho	PO3
Sulfamoyl	N-SO2★
Sulfenamido	CHK-S)-N★ or divalent S-N
Sulfinamido	CHK-S(=O)-N★ or divalent S(=O)-N
Sulfinyl	CHK-S★(=O)
Sulfo	SO3
Sulfonamide	CHK-SO2-N★ or divalent SO2-N
Sulfonyl	CHK-SO2★
Sulfonyloxy	CHK-SO2-O★
Thioalkoxy	CHK-S★

6. Summary

Key advantages of DWPIM implementation on STN:

- **Consistent searches can be performed in all structure and Markush databases**

  A single structure query can be used as a basis for searching in all databases. Since there may be different bonding conventions in the databases the user may need to adapt the query accordingly. The same concept for Markush attributes (match levels) applies consistently to all Markush databases on STN.

- **Many parameters enable the user to control recall and precision**

  A large set of parameters is available to control the search precision and recall. Among them are the search modes Closed Structure Search (CSS) or Substructure Search (SSS), element counts, node types and attributes as well as bond types and values.
  Subset and Batch mode offer additional search options.

- **Improved evaluation of Markush structures with hit structure display, highlighting, and assembled structures**

  The results of all structure searches are shown with highlighting of the query. Markush structures can also be viewed in assembled form where the structure includes all the G-group variations which participate in the structure matching.

- **Integrated environment of Derwent files for search and display**

  Efficient crossover capabilities are available for structure (DCR, DWPIM) and text searches (DWPI).

7. Glossary

Term	Definition / Explanation	Page
A35	see **Superatom A35**	31
ACT	see **Superatom_ACT**	31
ACY	see **Superatom_ACY**	32
AMX	see **Superatom_AMX**	31
ARY	see **Superatom_ARY**	30
Aromatic Compounds	Defined by the Hückel rule as planar, fully conjugated, monocyclic systems with (4n+2) π electrons, in DWPIM: mono- or polycyclic systems which contain at least one benzene ring (see **Superatom ARY**)	37
Bonding Conventions	STN and Derwent Markush Resource conventions	37-36
Bond Types	STN structure editor: chain, ring, ring/chain DWPIM: chain, ring	36
Bond Values	STN structure editor: exact/normalized, exact, normalized DWPIM: single, double, triple, normalized (aromatic or tautomeric)	36
CHE	see **Superatom_CHE**	30
Chemical Patent	May protect: chemical compounds, process of preparation, or application or use	8
CHK	see **Superatom_CHK**	30
CHY	see **Superatom_CHY**	30
Conditional Logic	describes dependency between two substituents	46
CYC	see **Superatom_CYC**	30
DCR	Derwent Chemistry Resource, contains specific structures, referenced in DWPI	17
Display Options	Summary, Detailed, Assembled	77
DWPI	Derwent World Patent Index, contains all patent literature (incl. Chemistry)	17
DWPIM	(DWPIM) contains Markush structures, referenced in DWPI	9

Term	Definition / Explanation	Page
DYE	see Superatom DYE	32
Element Counts	number of elements for a fragment or a complete structure (this feature is not yet operational for the beta test)	68
Free Sites	Equivalent to non-hydrogen count (s. chapter 5.1, on the influence on free sites on the results of a Markush structure search)	44
Frequency Variation	see Markush Variations	8
G-Group	Variable group with alternative fragments in the Markush file structure (called R-group in the query), G-groups may be nested	35
Generic Node	Generalization of chemical elements, comprises an open set (*e.g.* halogens) or a closed set of chemical elements (*e.g.* carbon chain or alkyl), Derwent generic nodes are called superatoms (*e.g.* ARY for Aryl compounds) (s.a. Superatom) STN query nodes: X, M, A, Q, Ak, Cy, Cb, Hy (X = halogens, M = metals, A = any element, Q any element except C, Ak = carbon chain, Cy = cyclic compounds, Cb = carbocyclic compounds, Hy = heterocyclic compounds	27
HAL	see Superatom_HAL	31
HEA	see Superatom_HEA	30
HEF	see Superatom_HEF	30
HET	see Superatom_HET	30
Heteroaromatic Compound	(heteroaryls): mono- or polycyclic compounds which contain in addition to a carbon atom at least one other atom (mostly N, O, S; but metal atoms are also allowed) and follow the Hückel rule (see Aromatic Compound); in DWPIM heteroaromatic compounds are restricted to monocyclic five- membered rings with two double bonds and six-membered rings which have at least one heteroatom in the ring (see Superatom_HEA)	30
Homology Variation	see Markush Variations	8
Keto-Enol Tautomerism	Special form of tautomerism between ketones and aldehydes, requires at least one proton in α-position to the carbonyl group; in DWPIM the tautomers are preferentially indexed and represented in the keto form (exceptions are possible)	42
LAN	see Superatom_LAN	31

Term	Definition / Explanation	Page
Markush Substances	Organic, organometallic, inorganic, polymers, peptides, fullerenes & carbon nanotubes, boranes & polypeptides	18
Markush Compound Number	MCN, general format: YYWW-CCCSS YY = Year, WW = Derwent Week, CCC = identifier unique to a document in a given week, SS = integer from 01 to 99	72
Markush Node Attributes	Attributes are additional parameters describing certain properties of chemical nodes. There are different types of attributes: atom, superatom, and peptide attributes.	44
Markush Generic Terms	Markush fragments in a patent are often abbreviated by a text term (Markush term) and they must be translated into structure fragments consisting of superatoms, shortcuts, specific elements, and associated free sites.	95
Markush Master Structure	invariant core structure, usually shown in diagrammatic form, together with a set of variations.	8
Markush Patent	a chemical patent which contains generalized chemical structure formulas	8
Markush Structure	Generalized chemical structure which may consist of a Markush Master Structure and G-groups which contain a set of fragments, nodes in the structure may be chemical elements, shortcuts, superatoms, or G-groups	8
Markush Variations	variable part of the Markush structure: substitution variation (list of possible substituents), position variation (variable point of attachment), frequency variation (number of repetitions), and homology variation (generic nodes).	8
MARPAT®	Produced by CAS. Contains more than 1.099.000 searchable Markush structures from patents covered by CAS from 1988 to the present.	58
Match Level	STN concept to control the degree of structure query matching between the query structure and the structure in the search file (translate option on Questel for MMS): • ATOM: retrieves only specific atoms or groups in the file • CLASS: retrieves specific atoms or groups and corresponding generic groups • ANY: retrieves nodes in the query that match specific atoms, generic nodes, and R nodes Match levels are assigned to each atom and to each generic group/superatom. All nodes of a ring system should have the same match level. Default Match Level for DWPIM: ATOM for ring nodes and CLASS for chain nodes	58

Term	Definition / Explanation	Page
MMS	Merged Markush Services combines the MPHARM database from the French Patent Office and World Patents Index Markush (WPIM) files from Clarivate Analytics on Markush DARC, the file contains both specific structures and Markush structures; Note: on STN the content of MMS has been split into a database with specific structures (DCR) and a database with Markush structures (DWPIM)	
MX	see Superatom_MX	31
Nodes	notation for chemical elements (specific nodes) and generic nodes (superatoms in DWPIM)	27
Node Attributes	parameters of a node, *e.g.* charge or valency for specific nodes or saturated/unsaturated for generic nodes	44
Normalization	Alternating single and double bonds are combined to form normalized bonds (aromatic bonds for aromatic rings)	38
PEG	see Superatom_PEG	32
POL	see Superatom_POL	32
Position Variation	see Markush Variations	8
PRT	see Superatom_PRT	32
Quinoid	Deriving from the class of quinones: A **quinone** is a class of organic compounds that are formally "derived from aromatic compounds by conversion of an even number of –CH= groups into –C(=O)– groups with any necessary rearrangement of double bonds", resulting in "a fully conjugated cyclic dione structure" (IUPAC, *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book") (1997). Online corrected version: (1995) "Quinones"). Typical examples are 1,2-Benzoquinone; 1,4-Benzoquinone; 1,4-Naphtoquinone.	29, 43
Repetition Group	structure fragments which occur several times in the structure; normally the number of repetitions is represented as a range	8
R-Group	Variable group with alternative fragments in the query (called G-group in the Markush file structure), R-groups may be nested	35,53
Shared Variables	Expression used in chemical patents to allow the possibility to form a ring	47
Shortcuts	Abbreviation for a set of frequently used chemical groups (mostly functional groups)	28

Term	Definition / Explanation	Page
Specific Nodes	chemical elements of the periodic system	27
Spin-Off Generic Node	Generic node which is derived from a structure fragment with specific nodes only (appended by an asterisk to distinguish them from original generic nodes), spin-offs are required to enable the software to perform matches between specific nodes and generic nodes	59
Substance Descriptors	code and text to classify the substance, *e.g.* A Polymers & Plastics multiple substance descriptors can be applied to a single substance	75
Substitution Variation	see **Markush Variations**	8
Superatom	Generic node as defined by Clarivate Analytics for the database DWPIM, *e.g.* ARY for aryl	20
Superatom A35	Group III a to V a metal: Al, Ga, In, Tl; Ge, Sn, Pb; Sb, Bi	31
Superatom ACT	Actinides (incl. Actinium): Po; At; Fr; Ra; Ac; Th, Pa, U, Np, Pu, Am, Cm, … (>92)	31
Superatom ACY	Acyl compounds	32
Superatom AMX	Alkali and Alkaline Earth Metals: Li, Na, K, Rb, Cs; Be, Mg, Ca, Sr, Ba	31
Superatom ARY	Aryl compounds, containing at least 1 benzene ring	30
Superatom CHE	Alkenyl & alkenylene compounds	30
Superatom CHK	Alkyl & alkylene compounds	30
Superatom CHY	Alkynyl & alkynylene compounds	30
Superatom CYC	Cycloaliphatic compounds, containing no benzene ring	30
Superatom DYE	Chromophore group	32
Superatom HAL	Halogens: F, Cl, Br. I	31
Superatom HEA	Monocyclic heteroaryl compounds, includes only five- or six-membered rings	30
Superatom HEF	Fused heterocyclic compounds	30
Superatom HET	Monocyclic nonaromatic compounds	30
Superatom LAN	Lanthanides: Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu	31

Term	Definition / Explanation	Page
Superatom MX	Any metal	31
Superatom PEG	Polymer end group	32
Superatom Peptides	Actually amino acid shortcuts: 30 amino acid shortcuts which correspond to standard amino acids commonly found in chemical patents, *e.g.* alanine (ALA)	34
Superatom POL	Polymer	32
Superatom PRT	Protecting group	32
Superatom TRM	Transition Metals: La, Sc, Y; Ti, Zr, Hf; V, Nb, Ta; Cr, Mo, W; Mn, Tc, Re; Fe, Ru, Os; Co, Rh, Ir; Ni, Pd, Pt; Cu, Ag, Au; Zn, Cd, Hg	31
Superatom UNK	Any atom or group including H	33
Superatom XX	Any atom or group excluding H	33
System Limits	For indexing: connectivity: max. 8, Number of G-groups: max. 50, Number of fragments in a G-group: max. 50, Nesting between G-groups: maximum 4 levels.	20
Tautomerization	normalization of bonds in a tautomeric group	39
Clarivate Analytics Content Domain	Also called Derwent Content Domain, consists of the databases DWPI, DCR, and DWPIM; the documents in DWPI are related with the corresponding documents in DCR and DWPIM (and vice versa)	18
Translate Option	Questel concept to control the degree of structure query matching between the query structure and the structure in the search file (match level on STN). • EQ: retrieves the node exactly as specified in the query • NT: retrieves the specified superatoms, superatoms which are lower in the hierarchy, and all corresponding specific atoms • BT: retrieves the specified node and all corresponding superatoms higher in the hierarchy, including the R-node (XX) • ANY: NT + BT	109
TRM	see Superatom_TRM	31
UNK	see Superatom_UNK	33
Variable Group	consists of a set of fragments or Markush variations	34

Term	Definition / Explanation	Page
Variable Point of Attachment (VPA)	specifies multiple positions on a ring where an atom or a group can be attached	8
XX	see Superatom_XX	33

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure