Searching Chemical Compounds with Fragmentation Codes

Overview

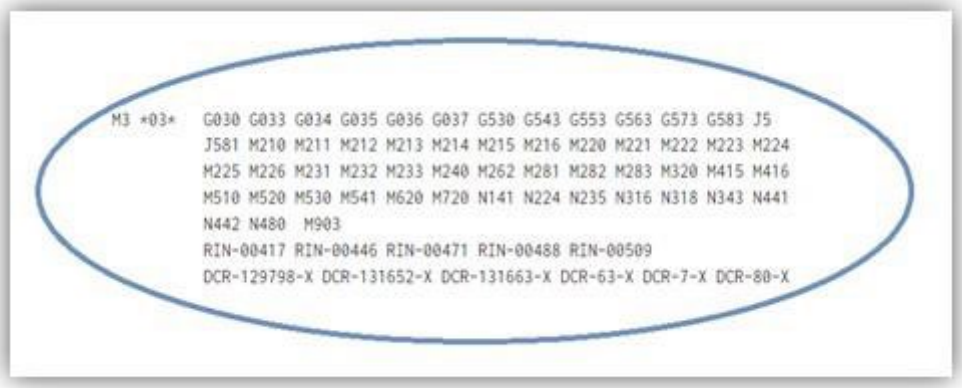
Derwent began fragmentation code indexing in 1963 as an early method for the electronic capture of chemical structure information (the codes themselves correspond to old IBM punch cards). As such, DWPI fragmentation codes are the oldest online resource for searching patent Markush structures in STNext and the only comprehensive source of Markush indexing from 1963-1977 as well as an important resource to capture specific structures. Fragmentation codes provide access to 530.000 DWPI records with unique structure information (chemical compounds without DCR/DWPIM indexing).

The new fragmentation code functionality on STNext is fully integrated in the existing structure editor thereby ensuring comprehensive search capabilities in combination with the existing arsenal of structure search databases. Due to a qualitatively improved code generation process better results in terms of recall and precision can be achieved.

Fragmentation Codes

Fragmentation codes are alphanumeric representations of individual substance fragments:

Complete substances are then indexed as “paragraphs” of fragmentation codes:



Over time, Derwent expanded the scope of indexing to cover wider and wider areas of chemistry:

- 1963 Pharmaceuticals (B)
- 1965 Agricultural Chemicals (C)
- 1970 Code revision + General chemistry (E)
- 1972 Code revision + Ring Index Numbers
- 1981 Code revision + DRNs
- 1987 Markush graphical indexing + DCNs
- 1992 Autogeneration of codes
- 1999 Derwent Chemistry Resource

Function and Features of STNext Fragmentation Code Generator

The new STNext fragmentation code functionality allows a structure to be translated into corresponding alphanumeric fragmentation codes and searched against the WPIX or WPIDS files via a STNext script.

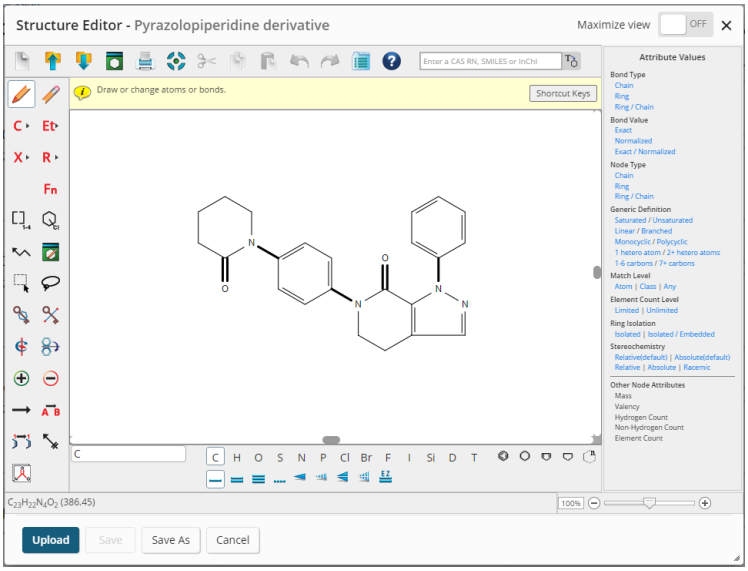
The entry point for generating a search strategy is the STNext structure editor, eliminating the need for a separate editor as well as upfront steps as known from STN Express.

The generated script can be freely edited before submission.

The STNext fragmentation code generator is based on an advanced version of Clarivate's editorial application. As part of a backfile enhancement project, Clarivate added the more precise colored fragmentation codes to older records. This allows the use of a single strategy for Frag Code searching, over all time periods. It is important to note that this single-line code strategy has a potential disadvantage in terms of completeness of the small portion of pre-1981 results. As a corollary, we are currently checking the feasibility of a subsequent implementation of the multi-line time-range strategy as known from STN Express.

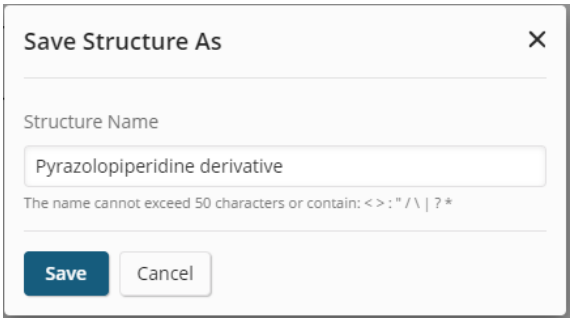
Running a Search

Step 1: Draw a Structure in STNext Structure Editor

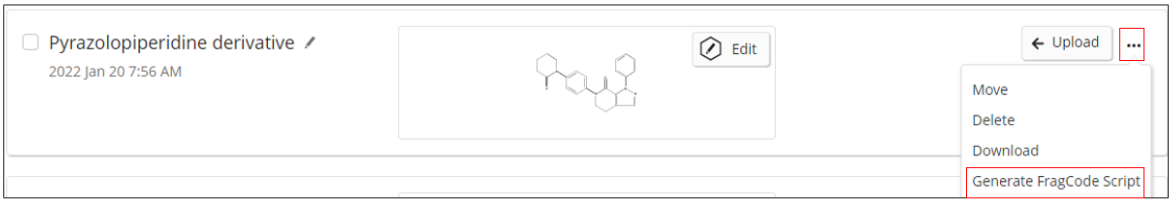


Note: Please see the **FCO Input Format** section of this document for a complete list of supported and unsupported and ignored structure components.

Step 2: Save Structure As



Step 3: Generate Fragmentation Code Script



The system automatically converts a structure drawing into a fragmentation code search and prepares a search strategy in the form of an STN Script. The system also assigns a generic name (e.g., 2022_0005_Structure_FragCode_01) which can be edited to a more meaningful name (50 character max).

Step 4: Edit Fragmentation Code Script (optional)

From the **Edit Script** window, the user may run the search in either the WPIX or WPIDS database or save the script for later use. In addition, the **Edit Script** window can be used to modify the search to allow for substitution by:

- Removing negation codes
- Adding codes with counts (e.g., J012-J014 to consider several COOH substituents)
- Adding codes for derivatives (e.g., hydroxy to ether or ester)

Step 6: Fragmentation Code Search in WPIX

```
=> s (D011(P)D013(P)D830(P)F011(P)F012(P)F433(P)G010(P)G013(P)G100(P)H213(P)J522(P)"L941"(P)"L999"
(P)M511(P)M521(P)M532(P)01188(P)M280(P)M320(P)M540(P)M412)/M0,M2,M3,M4

         12811 D011/M0
        282239 D011/M2
         84855 D011/M3
         46137 D011/M4
         12811 D013/M0
        272302 D013/M2
         83694 D013/M3
         45409 D013/M4
```

. . .

```
        245722 "L8"/M2
         78309 "L8"/M3
           976 "L8"/M4
        439858 M1/M2
        203155 M1/M3
         67743 M1/M4
L3           9 L2(NOTP)(H1 OR H3 OR H4 OR H5 OR H6 OR H7 OR H8 OR H9 OR J0 OR
               J1 OR J2 OR J3 OR J4 OR J9 OR K1 OR K2 OR K3 OR K4 OR K5 OR K6
               OR K7 OR K8 OR K9 OR "L1" OR "L2" OR "L3" OR "L4" OR "L5" OR
               "L6" OR "L7" OR "L8" OR M1)/M2,M3,M4
```

Display of results: In the display the hit codes are highlighted in the CMC section (e.g., d kwic, d cmc)

```
M2 *97*   C000 C100 D010 D011 D012 D013 D014 D019 D020 D021 D022 D023
          D024 D025 D029 D040 D049 D611 D700 D720 D830 D840 D860 D900
          D910 D920 D970 D980 F010 F011 F012 F013 F014 F015 F016 F017
          F019 F020 F021 F029 F423 F433 F710 G001 G002 G003 G010 G011
          G012 G013 G014 G015 G016 G019 G020 G021 G022 G029 G030 G033
          G034 G035 G039 G040 G050 G100 G111 G112 G113 G221 G299 G553
          G563 G599 H100 H102 H103 H121 H122 H123 H141 H142 H2 H211 H212
          H213 H401 H402 H421 H422 H521 H522 H541 H600 H608 H609 H621
          H622 H623 H641 H681 H682 H683 J011 J012 J013 J311 J312 J371 J5
          J521 J522 J523 J592 J599 K130 K140 K620 K640 K830 K850 K899
          K930 L110 L130 L142 L143 L199 L9 L930 L941 L943 L999 M111
          M112 M113 M114 M115 M116 M119 M122 M123 M124 M125 M126 M129 M136
          M141 M143 M210 M211 M212 M213 M214 M215 M216 M231 M232 M233 M240
          M272 M273 M280 M281 M282 M283 M311 M312 M313 M314 M315 M320
          M321 M322 M331 M332 M333 M340 M342 M349 M353 M373 M381 M391
          M412 M431 M511 M512 M513 M521 M522 M523 M530 M531 M532
          M533 M540 M541 M542 M543 M630 M640 M650 M710 M782 M800 P431
          P446 P510 P517 P518 P526 P530 P554 P611 P612 P623.  .  .  P731
          P739 P812 P814 P818 P822 P824 P844 P922 P930 P943 P950  M905
          M904
          RIN: 01160 01168 01172 01174 01175 01186 01188 01190 01262
          01264 01275 01279 07978 11882 46969 47119 47120 48548 50527
          50531 62081
          MCN: 2259-00002-M 2259-00002-N
```

Critical Aspects for your search: Limitations and Best Practices

To ensure that you receive correct results of your fragmentation code search, please see the up-to-date list of **Critical Aspects for Your Search: Limitations and Best Practices.**

More training material including recorded training sessions are available here (Training Center | STN International (stn-international.com)

FCO Input Format – Scope and Limitation

Definitions:

Supported: correct frag code strategy is generated

Not-supported: an error message appears or wrong codes are generated

Ignored: functions/features not considered in frag code strategy

Bond Types

- supported: chain, ring
- ignored: ring/chain

Bond Values

- supported: exact, exact/normalized, normalized

Bond Representation

- supported: single, double, triple
- not supported: unspecified (error message)
- ignored: stereo bonds, E/Z bonds

Generic Nodes

In order to create free sites use the nodes A or Q.

- supported: CHK, CHE, CHY, ARY, HEA, HEF, HET, CYC, A35, AMX, TRM, LAN, ACT, HAL, X
- not supported (wrong codes generated): Ak, Cb, Cy, Hy, Id, ACY, DYE, POL, XX, PEG, PRT, UNK

Structural features not supported (wrong codes generated)

- -OH and –SH substituents directly bonded to aromatic rings (such as e.g. phenols, thiophenols). For those chemotypes the respective Hydrogen(s) must not be present, i.e. use Phe-O for phenols.
- -OH and –SH substituents directly bonded to saturated heterocycles. For those chemotypes the respective Hydrogen(s) must be present. If just –O or –S (without explicitly present Hydrogen) is used wrong codes are generated.

Features and Attributes not supported

- R-groups (if generic nodes are present) (wrong codes generated)
- Variable Point of Attachment (VPA) (error message)
- ring isolation (error message)

Features and Attributes ignored

- Generic definitions: Cycles
    - saturated/unsaturated
    - branched/linear
    - number carbon atoms: less than 7 / 7 or more
    - number of hetero atoms: exactly 1 / 2 or more
- abnormal valency
- abnormal/specific mass
- charges
- match level
- element count level
- Non-Hydrogen Count (NHC), Hydrogen Count (HC)

Atom lists

- Supported: atoms per list (R-group) is limited to 5 atoms
- Not-supported: atom list with more than 5 atoms (error message)