



Chemical Explorer manual

Features

Minesoft's Chemical Explorer is a database of chemicals extracted from the full text and images (US only) of patent documents. It is seamlessly linked to PatBase, a full text patent family database, developed in partnership with RWS.

Minesoft Chemical Explorer allows a user to interrogate the chemical compounds exemplified in US, EP, WO, DE, FR, JP, CN, KR, GB, AU, IL, RU, SU, DD and IN full text patents from the starting point of a chemical structure or a chemical name (including trade and IUPAC names). Features include:

- Full structure drawing capability
- Ability to perform an identity, similarity (based on the Tanimoto threshold) or substructure search
- Search within either the complete patent full text or within the claims only
- Contains over 19 million unique chemical compounds from over 16 million patent documents (Jan 2019)
- Updated daily
- Ability to import chemical structures (Smiles and MOL files)
- Ability to export and save chemical structures (MOL files)
- Links to PubChem, ChemSpider, Wikipedia

Table of Contents

SOURCES	3
COVERAGE	3
SEARCHING	3
DRAWING CHEMICAL STRUCTURES.....	3
FINDING STRUCTURES	7
STRUCTURE SEARCH TYPES.....	8
RETURN STRUCTURES OPTIONS.....	11
TOOLS	12
STRUCTURE RESULTS	12
VIEWING RESULTS IN PATBASE	14
NOTE.....	17

Sources

The PatBase full text collection has been mined using the LeadMine technology from NextMove Software.

Coverage

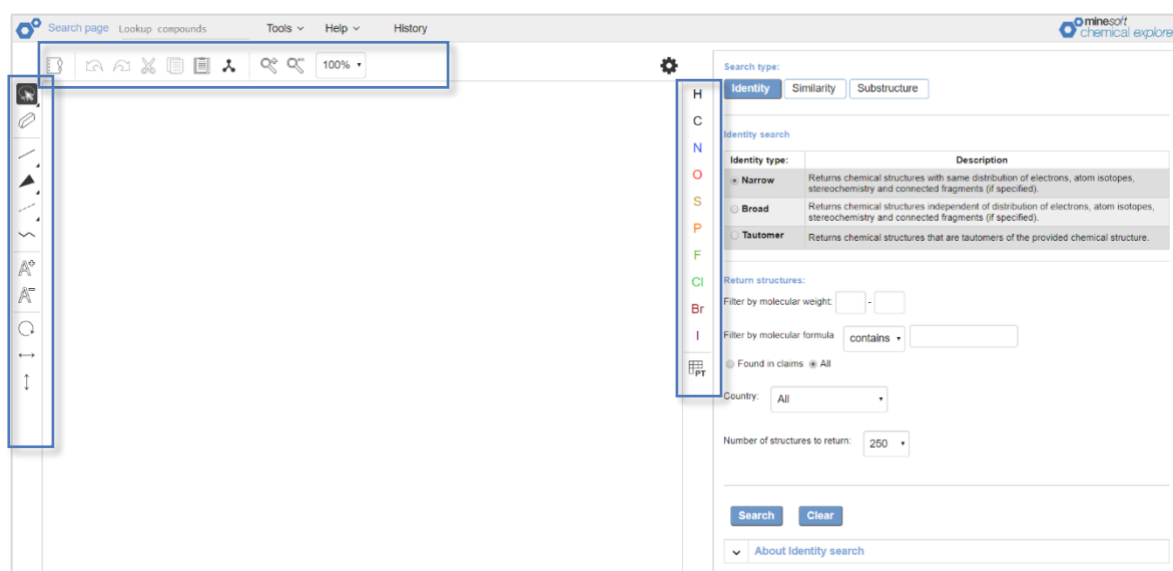
Exemplified chemicals are recognized, extracted and converted into chemical structures from the **full text** of patents from the following countries:

Country	Date
US	1928 to date
US (images)	2001 to date
EP	1978 to date
WO	1978 to date
DE	1920 to date
FR	1960 to date
JP	1998 to date
CN	1985 to date
KR	1994 to date
GB	1917 to date
AU	1990 to date
IL	1971 to date
IN	2006 to date
RU	1992 to date
SU	1924 to dissolution
DD	1957 to dissolution

Searching

Drawing chemical structures

Users can draw chemical structures using the structural formula editor which is surrounded by three toolbars containing the tools you can use in the editor:



Top toolbar



Clear canvas: Clear the entire canvas.


Undo/redo: Undo or redo recent changes.

Cut: Cut the selected part of the structure.

Copy: Copy the selected structure.

Clean: Redraws the sketch with standard bond angles/length

Zoom in/Zoom out: Zoom the sketcher view in and out. This can be modified using the drop-down selection box to the right of these options.

Settings  (far right): Open the Chemical Explorer settings pop-out page to control the default font, size and atom colouring, options on displaying the charge, valency, whether carbon should be displayed explicitly, and how bonds should be displayed within the tool.

Left toolbar



(From top to bottom...)

Selection tools: All these tools can be used to drag the current selection or individual atoms and bonds. You can add/remove atoms and bonds to the selection by clicking them. If you have selected a separate fragment, you can rotate it by dragging an atom in the selection. You can delete the selection using the DEL key. Each tool has different behaviour.



- Lasso select: Select atoms and bonds by drawing a freehand selection area
- Rectangle select: Select atoms and bonds using a rectangular selection area
- Fragment select: Select all atoms/bonds that are connected to the clicked node

Erase: When selected this tool will delete any clicked atom or bond within the sketcher.

Bonds: Allows you to place a single, double or triple bond into the sketcher at the point indicated. Drag and click from one atom to another to form a bond of the selected type between the two atoms.

Up/down & cross bonds: Up, down and cross bonds allow you to place bonds that denote the 3D chemistry of the query molecule, allowing you to select up bonds, down bonds, up/down bonds which will search both variants at the given position or cross bonds, which ignore cis/trans isomerism around the the bond cross bond is placed at.



Any bond: The any bond tool allows you to search a structure where a bond at a given position can be any variation of chemical bond or restricted to a smaller selection of types. (From top to bottom)

- Any bond: The tool will return structures with any bond type at the given location.
- Aromatic bond: The tool will return structures with an aromatic bond at the given location.
- Single/Double: The bond at the given location must be either a single or double bond.
- Single/Aromatic: The bond at the given location must be either a single or aromatic bond.
- Double/Aromatic: The bond at the given location must be either a double or an aromatic bond.

Chain: Create a chain of carbon atoms.

Charge: Increment (+) or decrement (-) the charge of atoms.

Rotate: Click & hold then circle your mouse to rotate the drawn structure within the sketcher.

Horizontal flip: Flip the structure 180 degrees in the horizontal plane.

Vertical flip: Flip the structure 180 degrees in the vertical plane.

Right toolbar

In this toolbar you can select from a number of elements or you can also pick an element from the periodic table using the last button. You can use the element to create new atoms or modify existing atoms.

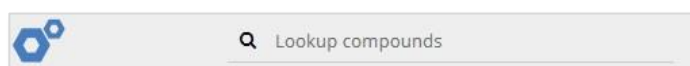


Bottom toolbar

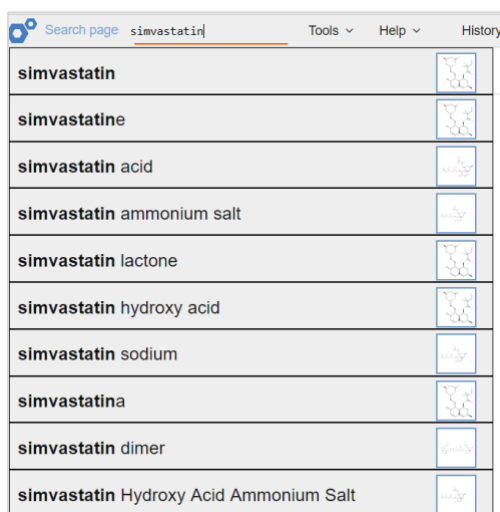
Fragments: Pick one of the fragments (benzene, cyclopropane, etc.) and add Fragments



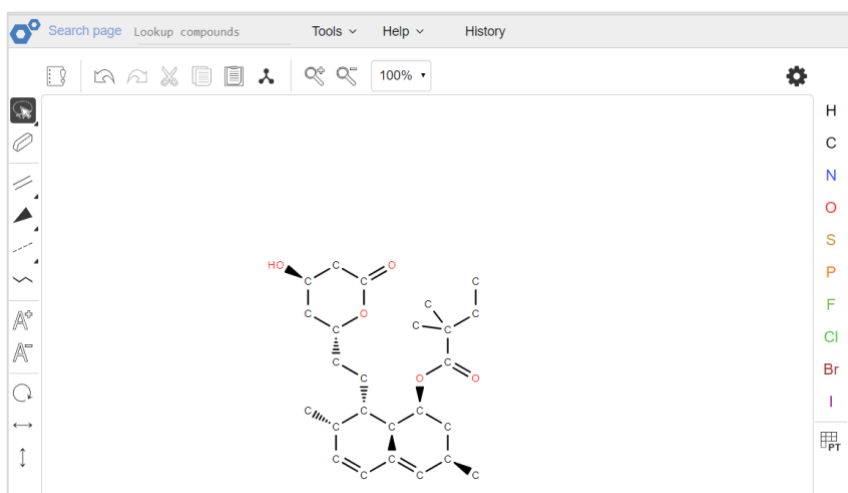
Finding structures



You can lookup compounds or molecules using the search form located on the left side of the menu-bar. Just begin typing what you are looking for and a list of available molecules will appear. For example, typing in simvastatin results in:



When you click on the compound of interest, the chemical structure is populated in the structural editor and can be searched in the database:



Additionally, systematic chemical names can be entered into the box and will be converted to a structure after pressing *Enter*.

Structure Search Types

You can complete different types of structure search:

Identity

Identity search allows you to locate records that are identical to the provided chemical structure with different notions of chemical structure identity. Additional search options allow you to choose the degree of "identity".

Search type:

Identity search

Identity type:	Description
<input checked="" type="radio"/> Narrow	Returns chemical structures with same distribution of electrons, atom isotopes, stereochemistry and connected fragments (if specified).
<input type="radio"/> Broad	Returns chemical structures independent of distribution of electrons, atom isotopes, stereochemistry and connected fragments (if specified).
<input type="radio"/> Tautomer	Returns chemical structures that are tautomers of the provided chemical structure.

Narrow: Most specific search which returns chemical structures with same distribution of electrons, atom isotopes, stereochemistry and connected fragments (if specified in the query).

Broad: Least specific search which returns chemical structures independent of the distribution of electrons, atom isotopes, stereochemistry and connected fragments (if specified in the query).

Tautomer: Returns chemical structures that are tautomers of the provided chemical structure. When selected, additional options are displayed which allow you to select one or more advanced options of:

- Ignoring hydrogen replacement by metal or charges, e.g. COOH → COO
- Where each boundary atom in the tautomeric chain must be one of N, O, P, S, As, Se, Sb or Te
- Where C (not from aromatic ring) is at one end and N,O,P or S at the other end of the tautomeric chain
- Where C from the aromatic ring is at one end and N or O at the other end of the tautomeric chain

Similarity

Similarity search allows you to locate records that are similar to a chemical structure query using pre-specified similarity thresholds.

Similarity is measured using the Tanimoto equation to compare the presence or absence of substructural features in your query to all molecules in Chemical Explorer. These features are encoded as binary "fingerprint". The fingerprint does not consider variation in stereo-chemical or isotopic information.

The threshold at which results are no longer considered similar is set by the Tanimoto threshold. Various predefined thresholds between 100-60% are allowed. Results are always returned in similarity order so

The screenshot shows the 'Similarity' search type selected. The 'Tanimoto threshold %' is set to 95. Under 'Return structures:', there are filters for molecular weight, molecular formula (set to 'contains'), a radio button for 'Found in claims' (which is unselected), a radio button for 'All' (which is selected), a 'Country' dropdown set to 'All', and a 'Number of structures to return' dropdown set to 250.

having a low "Number of structures to return" setting will only exclude results less similar than those that are returned.

Substructure

Substructure search allows one to locate chemical structures that contain a particular connectivity and valence-bond (i.e. bond order) pattern. For example, a substructure search of ethanol (SMILES: OCC) would return, among others, acetic acid (SMILES: OC(=O)C), since ethanol is a substructure of acetic acid.

The screenshot shows a web interface for a substructure search. At the top, there are three buttons for search type: "Identity", "Similarity", and "Substructure" (which is highlighted). Below this is a section titled "Substructure search" containing a table of search options:

Search options:	Description
<input checked="" type="checkbox"/> Resonance	Returns molecules whose resonance forms contain the query molecule.
<input type="checkbox"/> Tautomer	Returns chemical structures that are tautomers of the provided chemical structure.

Below the table, there are several filter options under the heading "Return structures:":

- Filter by molecular weight: [] - []
- Filter by molecular formula: [contains] []
- Found in claims All
- Country: [All]
- Number of structures to return: [250]

There are additional matching options for the substructure searches. These options are provided for further flexibility within a structural query:

- Resonance: Returns chemical structures whose resonance forms contain the query molecule
- Tautomer: Returns chemical structures that are tautomers of the provided query. When selected, additional options are displayed which allows the user to select one or more advanced options of:
 - Ignoring hydrogen replacement of metal bonds and atom charges in tautomeric chains. Where each boundary atom in the tautomeric chain must be one of N, O, P, S, As, Se, Sb or Te

- Where C (not from aromatic ring) is at one end and N,O,P or S at the other end of the tautomeric chain
- Where C from the aromatic ring is at one end and N or O at the other end of the tautomeric chain

Return Structures options

Irrelevant results can be filtered out by adding molecular weight or molecular formula constraints. Molecular formulas take the form of an element followed by the number of times that element appears. If this number is not specified 1 is assumed. Ranges may also be used. For example:

- C₂₀H₄₀ [results must have 20 carbon atoms and 40 hydrogen atoms]
- C₁₀₋₁₅ [results must have 10 to 15 carbon atoms]
- F₀ N [results must have 0 fluorine atoms and 1 nitrogen atom]

If the molecular formula is set to "is", then there must be zero atoms of all atoms not specified.

You can select options to specify the sections of the patent in which to return chemical structure hits, the specific country of interest – the default is all countries, but an individual country e.g. US, EP, WO, CN, JP, KR can be selected – and the number of structures to return, from 50-5000.

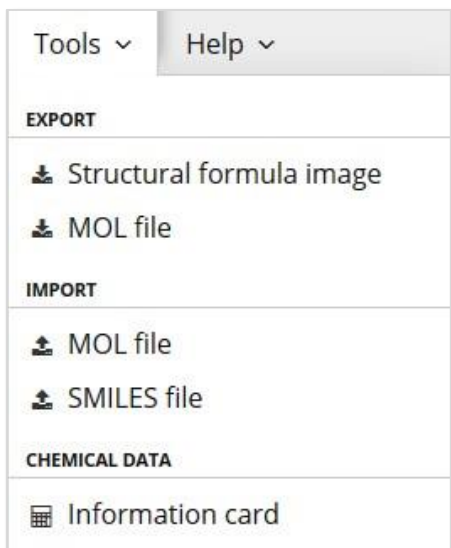
Return structures:

Found in claims All

Country:

Number of structures to return:

Tools



The Tools menu contains several utility functions which are listed below.

Export:

- Structural formula image: Sketcher snapshot (PNG with alpha channel)
- MOL file: MDL Molfile

Import:

- MOL file: MDL Molfile
- SMILES file

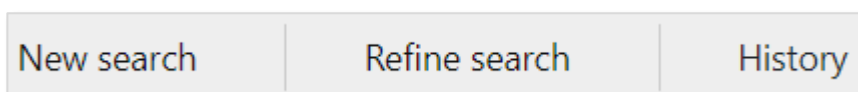
Information card

- This collects and displays information about the current structural formula (loaded from the structural editor).

Structure Results

After completing a chemical structure search the results screen displays the chemical structure hits with details of names, molecular weight, InChIKey, SMILES and formula and allows a user to execute a number of options.

Options





The ability to refine the search or complete a new search or run a previous search from your history.

Search results | New search | Refine search | History | minesoft chemical explorer

Number of structures found: 10

Select all | Clear all | View selected in PatBase | Export to Excel | Search/Filter this page

#	Structure	Details	Publications
1) <input type="checkbox"/>		busulfan IUPAC Name: 4-methylsulfonyloxybutyl methanesulfonate Molecular Weight: 246.302 g/mol InChIKey: CQVZYZSDYWGQREU-UHFFFAOYSA-N SMILES: CS(=O)(=O)OCCCCOS(=O)(=O)C Formula: C ₈ H ₁₄ O ₆ S ₂	121339
		121339 patent publications found View in PatBase Stats Analytics Links: PubChem ChemSpider Wikipedia Save MOL	
2) <input type="checkbox"/>		2-butyn-1,4-diol dimethanesulfonate IUPAC Name: 4-methylsulfonyloxybut-2-ynyl methanesulfonate Molecular Weight: 242.270 g/mol InChIKey: UGFC6BZBAORCFNA-UHFFFAOYSA-N SMILES: CS(=O)(=O)OCC#CCOS(=O)(=O)C Formula: C ₈ H ₁₂ O ₆ S ₂	201
		201 patent publications found View in PatBase Stats Analytics Links: PubChem ChemSpider Wikipedia Save MOL	

▼ Number of structures found: 10

The number of structures found and when expanded (by clicking on the arrow) the details of the search query, like below.

Search: O=S(=O)(C)OCCCCOS(=O)(C)=O
 Search type: Identity (Narrow - Returns chemical structures with same distribution of electrons, atom isotopes, stereochemistry and connected fragments (if specified).)
 Number of structures shown: 1
 Number of patent publications found: 121339
 Country: AU , CN , DD , DE , EP , FR , GB , IN , IL , JP , RU , KR , SU , US , WO

Select all | Clear all | View selected in PatBase | Export to Excel

The ability to select all structures of interest and to view the relevant patent families in PatBase, or to export the selected structures to excel.

The ability to select one or more structures and to view the relevant patent families in PatBase.

57 patent publications found [View in PatBase](#)

The number of instances a specific chemical compound has been identified in the patent and the ability to view these results in PatBase.

Links: [PubChem](#) [ChemSpider](#) [Wikipedia](#) [Save MOL](#)

The ability to link to more information on a chemical of interest in a number of external resources.

Search/Filter this page

The ability to search or filter the results based on a text string.

Viewing results in PatBase

After selecting the chemical structures of interest and clicking on *View in PatBase*, the relevant patent families are searched and identified in PatBase, with the query clearly indicating that the search originates from Chemical Explorer – [Chemical Explorer]: the chemical searched (e.g. Busulfan) and whether the search has been restricted to the claims. Furthermore, if multiple chemicals are selected and searched in PatBase, this is indicated by *Selected Structures* in the PatBase query (e.g. see search strategy 3 below).

3	[Chemical Explorer]: Selected structures (2) (Claims)	2352
2	[Chemical Explorer]: Selected structures (3)	24246
1	[Chemical Explorer]: busulfan	24213

To review the patent families identified in PatBase for the chemical(s) of interest, click on *View* or *Browse*. Within the family table, an additional Minesoft TextMine icon is displayed (🔍) against those publications in which the chemical of interest is found. For example, viewing the above search strategy 2 on the chemical Busulfan identifies a number of publications in which this compound is identified within the full text:

Publication number	Publication date	Application number	Application date	Links
CA2532579 AA	20060116	CA20042532579	20040716	
CA2532579 C	20140218	CA20042532579	20040716	
EP1650826 A1	20060426	EP20040747660	20040716	
EP1650826 A4	20081203	EP20040747660	20040716	
EP1650826 B1	20130501	EP20040747660	20040716	
ES2411659 T3	20130708	ES20040747660T	20040716	
IN225889 B	20090109	IN2006CN00200	20060116	
JP2005008829 A1	20070920	JP20050511857T	20040716	
JP4582458 B1	20101117	JP20050511857T	20040716	
JP4582458 B2	20101117	JP20050511857T	20040716	
KR101201272 B1	20121114	KR20067001080	20040716	
KR20060035767 A	20060426	KR20067001080	20040716	
KR101201273 B1	20121114	KR20117022736	20040716	
KR20110126728 A	20111123	KR20117022736	20040716	
KR101201271 B1	20121114	KR20117022737	20040716	
KR20110126729 A	20111123	KR20117022737	20040716	
KR101290877 B1	20130807	KR20127015293	20040716	
KR20120083525 A	20120725	KR20127015293	20040716	
US2006177742 AA	20060810	US20040564852	20040716	
US8163427 BB	20120424	US20040564852	20040716	
US8722255 BB	20140513	US20120430791	20120327	
WO05008829 A1	20050127	WO2004.JP10194	20040716	

Clicking on the icon next to each publication opens up a new Minesoft TextMine window which displays the full text of the publication, identifies the chemical(s) of interest and number of instances and highlights the chemical(s) of interest appear in the text. For example, we took our results from search line 2, our results will include any documents which mention the chemicals tetramethylene bis(methanesulfonate), Busulfan, or 2-Butynylene bis(methanesulfonate). Where references to these entities are detected, they are highlighted as shown on the next page.

The screenshot displays the Minesoft TextMine interface. On the left, a sidebar shows search filters: Molecule (307/977), Substituent (34/347), Physical (87/195), Generic (27/178), Chemical Role (1/29), Polymer (10/29), and Gene Or Protein (1/4). The main window shows search results for 'tetramethylene bis(methanesulfonate)'. A 'Molecule' window is open, displaying the chemical structure of Tetramethylene bis(methanesulfonate) with its SMILES string: COS(=O)(=O)OCCOCCOCCOCCOS(=O)(=O)C. The interface also shows a list of chemical entities identified in the text, such as Diphenylacetylene, 4-Ethynyltoluene, and Tetramethylene bis(methanesulfonate).

In brackets and next to each of the chemicals identified is the number of instances of this chemical in the text (e.g. Simvastatin occurs 171 times in the full text) and if you mouse over the chemical name, more information appears including the SMILES and variations identified:

SIMVASTATIN (171)

```
CCC(C)(C)C(=O)O[C@H]1C[C@@H](C)C=C2C=C[C@H](C)[C@H](CC[C@@H]3C[C@@H](O)CC(=O)O3)[C@@H]12
```

Instances: 171

Variations:

1. simvastatin
2. Simvastatin
3. Zocor
4. simvastatin

In search result

To move through the instances of the chemicals of interest within the document, you can simply scroll through the document, use the drop-down menu to select the section of interest or alternatively, use the up and down arrows in the right-hand menu.

Title & Abstract
Description
Claims
✓ Full Text

(1/166)

Home, List, Eye, Checkmark

More information can be found on any chemical highlighted within the text by clicking on the highlighted name. A pop-up window appears with the names and structure of the compound and links to external resources.

textmine Discovery Molecule

simvastatin 2

Title
[[1-(S),3-(R),7-(S),8-(S),8-(A)-(R))-8-[2-[(2-(R),4-(R))-4-Hydroxy-6-Oxooxan-2-Yl]Ethyl]-3,7-Dimethyl-1,2,3,7,8,8-(A)-Hexahydronaphthalen-1-Yl] 2,2-Dimethylbutanoate (C₂₅H₃₈O₆)

Abstract
The invention disclosed here...
L'invention ci-decrite concerne...
production de la **simvastatin**...
actuellement de la lovastatin

Description
LovD MUTANTS EXHIBIT


Chemical Structure
CCC(C)(C)C(=O)O[C@H]1C[C@@H](C)C=C2C=C[C@H](C)[C@H](CC[C@@H]3C[C@@H](O)CC(=O)O3)[C@@H]12

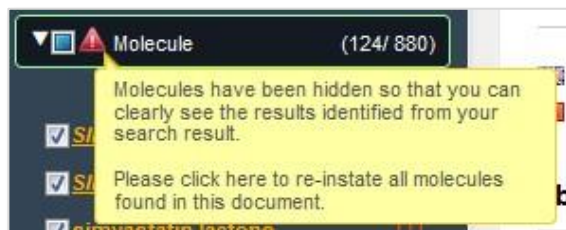
SMILES
CCC(C)(C)C(=O)O[C@H]1C[C@@H](C)C=C2C=C[C@H](C)[C@H](CC[C@@H]3C[C@@H](O)CC(=O)O3)[C@@H]12

InChI
InChI=1S/C25H38O6/c1-6-25(4,5)24(28)30-21-12-15(2)11-17-8-7-16(3)20(23)(17)21)10-9-19-13-18(26)14-22(27)29-19/h7-8,11,15-16,18-21,23,26H,6,9-10,12-14H2,1-5H3/15-,16-,18+,19+,20-,21-,23-/m0/s1

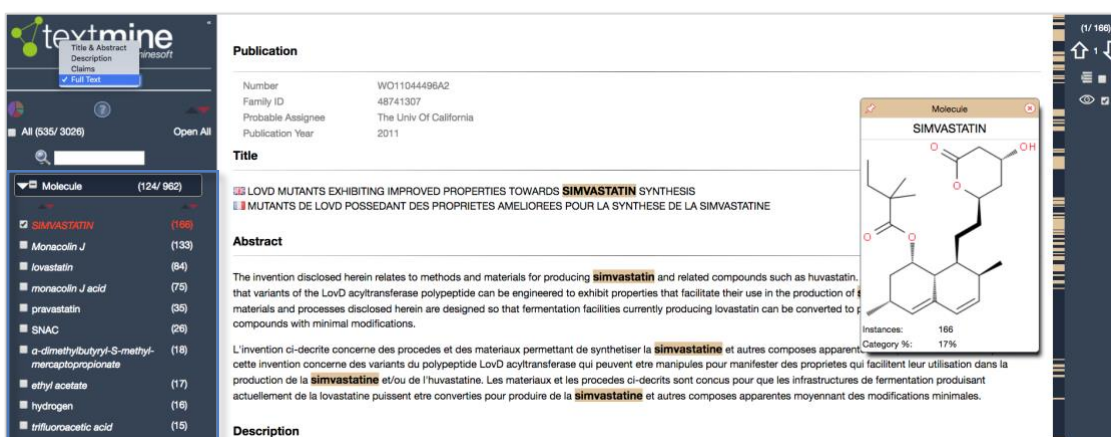
StdInChI
StdInChIKey RYMZZMVNJRUMDD-HGQWONQESA-N

ChemSpider PubChem Close

Furthermore, if you would like to see all the chemicals identified within the text of a publication, click on the filter warning .



You then have the ability to identify and select other chemicals of interest which occur within the full text by using the checkboxes.



The screenshot displays the 'textmine' interface. On the left, a sidebar shows a list of molecules with checkboxes and counts: SIMVASTATIN (166), Monacolin J (133), lovastatin (94), monacolin J acid (75), pravastatin (35), SNAC (26), o-dimethylbutyryl-S-methyl-mercaptopropionate (18), ethyl acetate (17), hydrogen (16), and trifluoroacetic acid (15). The main area shows a patent document for 'WO11044496A2' from 'The Univ Of California' (2011). The title is 'LOVD MUTANTS EXHIBITING IMPROVED PROPERTIES TOWARDS SIMVASTATIN SYNTHESIS'. The abstract discusses methods for producing simvastatin. A chemical structure window for 'SIMVASTATIN' is open on the right, showing the structure and statistics: 166 instances and 17% category.

Please note: Chemical Explorer is not part of a standard PatBase subscription but is a separate product that enables chemical structures to be searched and viewed in the full text.

When reviewing the results of a Chemical Explorer query in PatBase, it may be more efficient to use a split screen or dual screens where both the PatBase bibliographic details and the highlighted full text are visible.

The quality of the patent full text may vary, and this will affect the quality of the chemical data.

No Markush chemical structures are currently being extracted.