SpaceLight Commandline Documentation 1.0

Sascha Jung & Marcus Gastreich

December 12, 2022

Contents

1 Introduction

All links, references, table of contents lines etc. in this pdf are clickable.

Please note that this package is a commandline package.

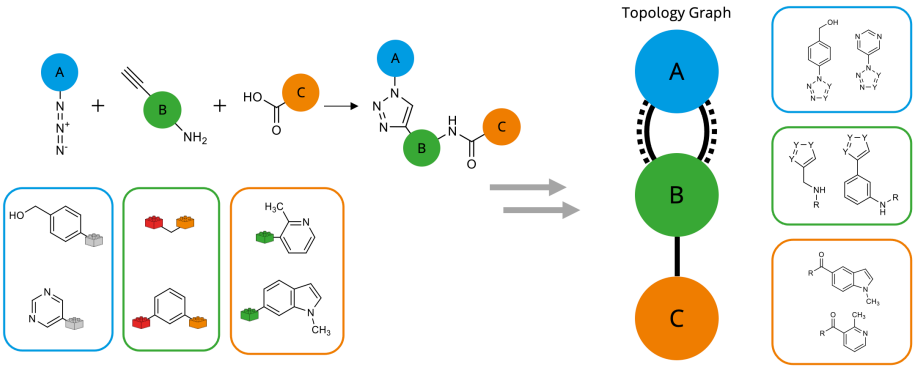
SpaceLight is a unique and novel software to navigate combinatorial chemical spaces ("fragment spaces") of multi-billion size and beyond using classical similarity measures (fingerprints).

SpaceLight is a perfect complement to our fuzzy Feature Trees technology (FTrees). Whereas the FTrees descriptor has a pronounced strength in detecting **distant** neighbors with chemical similarity, SpaceLight will find the **close-by** neighbors in a chemical space. As such, the tools complement each other.

SpaceLight lets you conduct

- very fast, **2D similarity searching** across vast combinatorial Chemical Spaces [3] for your query molecules
- similarity searches using a variety of different fingerprint types and sizes (ECFP and CSFPs)
- similarity searches in enumerated library files (SMILES, SDF or MOL2)

SpaceLight traverses huge Chemical Spaces using Lego-like chemical reaction combinatorics behind the scenes. To conduct quick calculations, we formalize reactions and encode them as pseudo-linking reactions. Per reaction, one so-called "topology graph" is stored and every node of this graph contains formalized, virtual building blocks, or as we sometimes call them, "reaction fate foreseeing" fragments. For more information on space generation, please have a look at our CoLibri package (`https://www.biosolveit.de/products/#CoLibri`).



The search then uses fast combinatorial algorithms and can deliver results in a few seconds only — even for very large, multi-billion sized spaces and beyond.

The implemented 2D similarity descriptors [1] are specially adapted versions of descriptors that you may know from so-called "2D / Tanimoto" similarity calculations from other tools. They need to be specially prepared beforehand for every chemical space. Spaces ready to be used with SpaceLight can be downloaded from our website (`https://www.biosolveit.de/infiniSee/#chemical_spaces`). Alternatively, you can prepare your own spaces with the CoLibri package.

Summarizing, with SpaceLight you can search much bigger spaces than with other methods, and you also require much less time, making it possible to search vast spaces even on modest, standard hardware. Additionally, you can also search traditional "medium sized" enumerated libraries.

2 Jump Start: A Simple Similarity Calculation

Your license is all set? You are familiar with commandline usage? You unpacked the installation archive? You downloaded or prepared a fragment space file (.space or .tfsdb file)? Then here is a typical query call to search a query against a fragment space (either as .space or .tfsdb file):

```
./spacelight -i <path/query.sdf> -s <path/fragmentspace.space>
```

```
./spacelight -i <path/query.sdf> -s <path/fragmentspace.tfsdb>
```

The query can be in sdf, smiles, mol and mol2 format and the file can have multiple entries. For quick tests, the input can just as well be a SMILES string in quotation marks, for example:

```
./spacelight -i "CC(C)C(=O)N" -s <path/fragmentspace.space>
```

To write your results to an output file, use the `-o` option additionally, for example:

```
./spacelight -i "CC(C)C(=O)N" -s <path/fragmentspace.space> -o <path/output.csv>
```

The output file can be either a csv file or a sdf file. It will contain detailed information for every result molecule: result rank, fragment similarity score, whole molecule similarity score, result name, query name, query smiles, space name, reaction name and reagents used to construct the result compound (see section 4.2 for detailed information).

If you do not specify an output file, all output will be printed to the console/terminal.

The fragment connected subgraph fingerprint (fCSFP5) is used as default similarity descriptor. The Tanimoto similarity is used as similarity measure. The similarity values are normalized between 0 (no similarity) and 1 (identical, in the framework of the descriptor).[1] The default descriptor is **not** stereo-aware, but it captures elements, connectivity, valence, and ring membership. As such it will capture "near" neighbor similarities in Chemical Space.

The 100 most similar compounds are written to the output by default. You can adjust this by using the `--max-nof-solutions` option with your desired number:

```
./spacelight -i "CC(C)C(=O)N" -s <path/fragmentspace.space> -o <path/output.csv>
             --max-nof-solutions 30
```

Additionally, you can limit the output to those results which exceed a certain similarity threshold by using the `--min-similarity-threshold` option (value between 0 and 1, see above):

```
./spacelight -i "CC(C)C(=O)N" -s <path/fragmentspace.space> -o <path/output.csv>
             --min-similarity-threshold 0.7
```

All this was too dense? Then let's take it step by step in the paragraphs below.

3 Technical Prerequisites

SpaceLight is a commandline application. Control through a graphical user interface (GUI) will follow later. SpaceLight needs the following to run:

- The **application package** (from `https://biosolveit.de/download`); depending on your operating system, some libraries may have to be installed (get in touch with us if that is the case: `mailto:support@biosolveit.com`; and please mention any errors/warnings that you see in your mail)
- A **shell** (Linux/Unix) or a terminal (macos), or a commandline environment (Windows; e.g.: cmd.exe)
- A valid **license** (from license@biosolveit.com)

The license setup instructions will come with the license that we will send out — or has already been sent out to you.

A "test license" that you can request online and that is sent to you instantaneously can simply be placed next to the executable (SpaceLight.exe, spacelight, or SpaceLight — depending on your operating system). For macos please read on...

macos Specialties On macos, the executable will typically reside inside the *.app package:

```
/Applications/SpaceLight.app/Contents/MacOS/SpaceLight
```

To place the short term test license there, you will have to go into the *.app package using a right mouse click on SpaceLight.app in the Finder, and click on "Show package contents". In there, you will see the Contents/ subfolder, in there the MacOS subfolder, and in there, the SpaceLight executable. If you are about to use the **test license**, place is right there, next to the executable. A longer term license will be handled separately, we will tell you how when we send that very license.

When you call SpaceLight for the first time, go to the Finder, and navigate to the Applications folder. Do a right(!) click on SpaceLight.app, and — if applicable — confirm that you want to open the program. It will flash up once, and you are good to go at the terminal prompt from there on.

To make the first step, call SpaceLight within your shell/terminal/environment.

4 Search and Find Options

4.1 General

An overview of all commmandline options is available by calling SpaceLight with `--help`.

```
./spacelight --help

Program options:
-i [ --input ] arg                  Input query molecule file or single input molecule as smiles.
                                    Supported file types are *.smi, *.smiles, *.mol, *.mol2 and *.sdf.
-s [ --search-files ] arg           Paths to library input molecule files for similarity scoring or to
                                    Topological Fragment Space database files or Fragment Spaces.
                                    Supported file types are *.smi, *.smiles, *.mol, *.mol2, *.sdf,
                                    *.tfsdb, *.space and *zip.
-o [ --output-files ] arg           Output base files (suffixes are required). Supported file types are
                                    *.csv and *.sdf.

Configuration:
-f [ --fingerprint ] arg (=fcsfp5)  Fingerprints for searching: ecfp and csfp variants are supported.
                                    Supported ecfp variants are ecfp0 to ecfp8. Three csfp subtypes are
                                    available: fcsfp, icsfp and tcsfp in variants from 1 to 5. E.g.
                                    ecfp4, fcsfp5, icsfp2 or tcsfp3.
--min-similarity-threshold arg (=0) Similarity threshold below which molecules are discarded [0.0 to 1.0].
--max-nof-results arg (=100)        Number of enumerated results.

General options:
-h [ --help ]                       Print this help message
--license-info                      Print license info
--thread-count arg                  Maximum number of threads used for calculations. The default is to use
                                    all available logical cores.
--version                           Print version info
-v [ --verbosity ] arg (=2)         Set verbosity level
                                         0 [silent]
                                         1 [error]
                                         2 [warning]
                                         3 [workflow]
                                         4 [steps]
```

Please note that the abbreviated, one-letter options are preceded with one dash – whereas the longer, named options are preceded with two dashes: `--`.
If an option needs an argument (arg), you can include or omit the equals sign.

4.2 Program Options

This section describes the arguments you must specify at minimum to successfully run a similarity search. First, you must provide the path to a file containing the query compounds. All well-known data formats are supported (MOL, SDF, SMILES, MOL2). Instead of a file containing the query molecules you can also specify a single SMILES string enclosed by quotation marks. The SMILES string or the molecule file must be passed to the `-i` option. Additionally, you need to specify the path to a topological fragment space database file (tfsdb file) or Fragment Space (.space file) that should be searched. Alternatively, you can also specify a library file (sdf mol2, smiles) to perform an enumerated search instead of a space search. Either the space file or library file have to be specified with the `-s` option. The minimal search prompt then has the following general form:

```
<path to spacelight executable> -i <path to queries> -s <path to fragment space>
```

When you specify the required paths the search prompt might look like the following:

```
./spacelight -i my_queries.sdf -s my_fragspace.space
```

Or, if you specified a SMILES string instead of a file:

```
./spacelight -i "CC(C)C(=O)N" -s my_fragspace.space
```

In the examples above, the output is printed on the console by default, which will look similar to the following example:

```
Query: O(CCCC)C O(CCCC)C
Rank:   1 sim: 0.706 O(CCCOC)C EN300-1717039
Rank:   2 sim: 0.577 O(CCC[NH2+]CCCC)C m_270004cba____8288582____9143638
Rank:   3 sim: 0.577 S(CCCOC)CCCC m_62bba____875776____9108904
Rank:   4 sim: 0.536 O(CCCC[NH2+]CCCC)C m_270004cba____8290272____9143638
Rank:   5 sim: 0.536 S(CCCCOC)CCCC m_62bba____875776____10159086
...
```

For every query molecule, the respective results molecules are printed to the console with information on the rank, similarity score, SMILES representation and name of the molecule. By default, 100 results molecules per query are

printed to the console. If you want to increase or decrease that number, please have a look at section 4.3.

Of course, you can also specify an output file to which the result molecules will be written by using the `-o` option. You can specify either a sdf file or csv file, e.g.

```
./spacelight -i "CC(C)C(=O)N" -s my_fragspace.space -o my_output.csv
```

For every query molecule a separate output file is generated which contains detailed information for every results molecule in addition to its SMILES (csv file) or MOL block (sdf file) representation:

- Rank: This is its ranking among all results for that very query.
- Fragment similarity: Similarity score derived from the weighted score of the fragment fingerprints (relevant only for space searches)
- Whole molecule similarity: Standard similarity score (Tanimoto similarity metric derived from query molecule fingerprint versus result molecule fingerprint)
- Name of the results molecule
- Name of the query molecule
- SMILES of the query molecule
- Name of the space or library
- Reagent1-3: name of building blocks from which the results molecule is constructed (relevant only for space searches)

The parameters in detail Description of program options in detail.

`-i [--input]` Specify a file containing the query molecules. Supported file formats are sdf, mol and mol2 files. You can also provide a text file containing multiple line-separated SMILES (file extension must be .smi or .smiles). Instead of a query file you can also specify a single SMILES string enclosed by quotation marks, e.g. "CNCCC1=CC=CC(=C1)C(O)=O". The `-i` option is required.

`-s [--search-files]` Specify a topological fragment space file (.tfsdb) or a library file (.sdf, .mol2, .smi, .smiles) or a Fragment Space file (.space). The

file is searched for close analogues to the query molecules given with the `-i` option. You can also search multiple spaces or library files at once by using the `-s` option several times with one query, e.g.
`./spacelight -i query.sdf -s space1.space -s space2.tfsdb`.
The `-s` option is required.

`-o [--output-files]` Specify the base name for the output files. Suffixes are required (.csv or .sdf). You can either write a SDF file or a csv file. SpaceLight generates a separate output file for every query molecule contained in the input query molecule file (`-i` option). For example, if you use a query file which contains two molecules you will get two separate numbered output files, both sharing the same base name, e.g.:

```
./spacelight -i query.sdf -s space1.tfsdb -o results.sdf
```

will result in two output files named results_1.sdf and results_2.sdf .

4.3 Configuration

With the parameters described in this section you can adjust the algorithmic parameters of SpaceLight. You can modify the fingerprint used to derive similarity, adjust the minimal similarity threshold below which result molecules are discarded and change the maximum number of results generated for each query molecule. All parameters in this section have default values which are round-bracketed in the following.

`-f [--fingerprint] (=fcsfp5)` With this option, you can adjust the fingerprint type and size used for the determination of similarity between query and result molecules. For more detailed information on the available descriptors see section 4.5. Shortly, you can choose between three different versions of the connected subgraph fingerprint (CSFP), each version with features containing from 1 up to a maximum of 5 heavy atoms (CSFP1-5): fcsfp1, fcsfp2, fcsfp3, fcsfp4, fcsfp5 (default); icsfp1, icsfp2, icsfp3, icsfp4, icsfp5; tcsfp1, tcsfp2, tcsfp3, tcsfp4, tcsfp5. Additionally, you can choose the classical circular extended connectivity fingerprint (ECFP) with a maximum diameter of 8: ecfp0, ecfp2, ecfp4, ecfp6, ecfp8.

`--min-similarity-threshold (=0)` This parameter adjusts the minimum similarity threshold below which the result molecules are discarded. By default, the value is 0, e.g. no result molecules are discarded.

`--max-nof-results (=100)` You can adjust the maximum number of result molecules per query which will be written to the output files. The default value is a maximum of 100 results per query.

4.4 General Options

`-h` / `--help` Displays the commandline help with short descriptions for every argument option. For more information see Section 4.1.

`--license-info` Shows detailed information about the license setup you currently use.

`--thread-count` Specify the maximum number of threads used for your similarity searches. By default, all available logical cores of your computer are used. You may want to reduce the number of threads used if you want to run other computations on your computer at the same time, or if you share the compute resource.

`--version` Displays information on the version of SpaceLight on the commandline. In quoting SpaceLight, please mention this version number.

`-v` / `--verbosity` Set the verbosity level, e.g., the level of console output, with an integer argument. The default value is 2. The following options are available:

- `0` Silent. No messages will be displayed in the console during the similarity search run. Errors will be ignored whenever possible.
- `1` Error. Only error messages will be displayed.
- `2` Warning. The default setting, warnings and error messages will be displayed.
- `3` Workflow. In addition to errors and warnings, information on the different steps of the similarity search are displayed on the commandline.
- `4` Steps. In addition to the 'Workflow' option, the progress of each step is displayed in detail.

4.5 Descriptor Choices

SpaceLight has several similarity descriptors (fingerprints) built in (see page 9). Depending on your use case, one descriptor may be more suited than another.

The acronym CSFP stands for *Connected Subgraph Fingerprints*, they describe all possible chemical features ("substructures") containing up to 5 heavy atoms for a given molecule.[1] In contrast, the "classical" ECFP descriptors (*Extended Connectivity Fingerprints*) describe molecules with a circular collection of features ("sit on one atom, then collect sphere with radius 1 around you, then sphere with radius 2, and so on").[4] CSFP descriptors are particularly optimized for similarity searches in fragment spaces as their set of features minimizes information loss across fragment boundaries. See the table below that has been taken from the original SpaceLight publication; it may serve as an overview of the respective atomic properties stored for the different descriptors along with the respective features.[2]

		ECFP	fCSFP	iCSFP	tCSFP
chemical substructures	circular	x			
	all		x	x	x
	element	x	x	x	x
	connectivity	x	x		x
	connectivity in substructure		x	x	
	valence	x	x	x	
atom properties	valence in substructure		x	x	
	aromaticity				x
	π electrons				
	formal charge	x			
	weight	x			
	ring membership	x	x		

Summarizing, we suggest to use the following descriptors for the respective use cases:

- fCSFP (fragment CSFP): Use this descriptor when the results should be **highly similar** to the query compounds. This is likely the descriptor that you should be able to relate most to when comparing it with the traditional ECFP descriptors. Element, valence state, and connectivity (within the feature and to the surroundings of the feature) are captured for every atom.
- tCSFP (topological CSFP): Compared to the fCSFP, tCSFP does not take into account the atom connectivity within substructures (features) nor the valence states, but on the other hand information on aromaticity is stored. This makes the tCSFP less strict in deriving similarity for a given compound pair compared to the fCSFP. Use this fingerprint in cases you do not find close analogues with fCSFP or ECFP fingerprints.

- iCSFP (independent CSFP): This descriptor is recommended for **substructure retrieval**. The iCSFP describes all structural features of a compound by properties of the atoms that are *independent of the surroundings of its substructures*. Therefore, it is suitable when searching for molecules that contain substructures or very similar substructures of a given query compound in an arbitrary order.
- ECFP: Extended Connectivity Fingerprints are widely used in traditional similarity searching. Use ECFPs when you would like to stay close to or compare to another tool that uses this descriptor type.

If you want to hop "deeper" into chemical space (find "distant neighbors"), then use the "fuzzier" Feature Trees descriptor available in our FTress tool (`https://www.biosolveit.de/download/`) and implemented into our infiniSee platform (`https://www.biosolveit.de/infiniSee`).

5 Further Reading, References

The original ideas behind the SpaceLight method are covered in the original publication by Louis Bellmann at Matthias Rarey's lab (`https://www.zbh.uni-hamburg.de/personen/amd/mrarey.html`, and see below).

Additional information and the tool itself is available at `https://www.biosolveit.de/spacelight-a-spotlight-on-the-analog-hunter-for-chemical-spaces/`.

Complementary tools, especially also the graphical platform infiniSee, can be obtained from the BioSolveIT website (`https://biosolveit.com`).

References

[1] Louis Bellmann, Patrick Penner, and Matthias Rarey. Connected subgraph fingerprints: Representing molecules using exhaustive subgraph enumeration. *J. Chem. Inf. Model.*, 59(11):4625–4635, 2019.

[2] Louis Bellmann, Patrick Penner, and Matthias Rarey. Topological Similarity Search in Large Combinatorial Fragment Spaces. *J. Chem. Inf. Model.*, 61(1):238–251, 2021.

[3] Torsten Hoffmann and Marcus Gastreich. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today*, 24(5):1148–1156, 2019.

[4] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010.

We wish you great success and much joy with SpaceLight!