GDC
CADD

A stub a day keeps the docstrings at bay

Paolo Tosco

12th RDKit UGM, Mainz
September 21, 2023

NOVARTIS | Reimagining Medicine

Current RDKit Python docstring status

- No hints at all in Visual Studio Code
  - VSCode uses `Pyright` for hints
  - `Pyright` needs `stubs`, and RDKit has none

```
from rdkit import Chem

mol = Chem.MolFromSmiles("CCC")

                (function) GetAtomWithIdx: Any
atom = mol.GetAtomWithIdx(
```

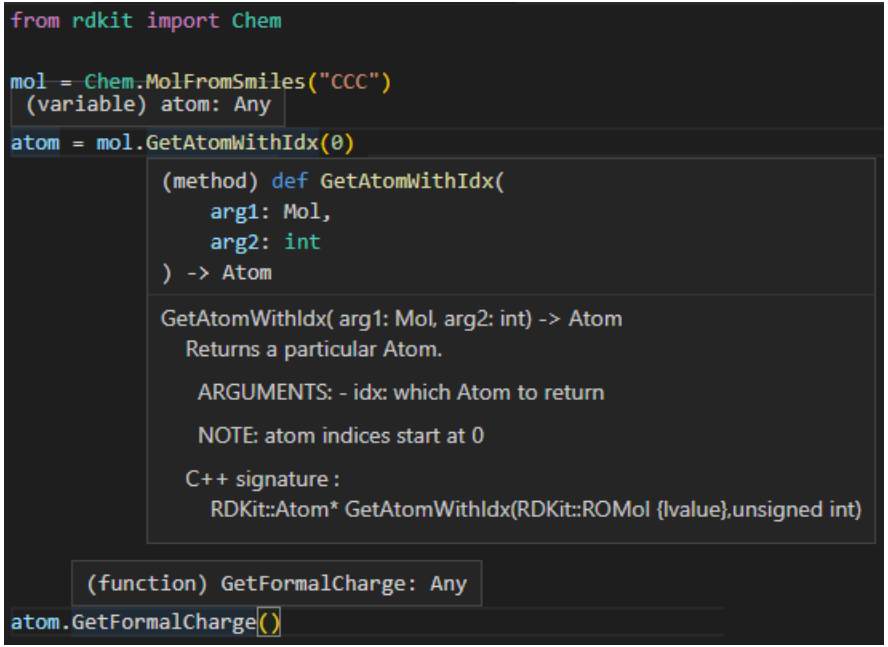
```
[4]: mol = Chem.MolFromSmiles("CCC")
     Last executed at 2023-08-04 12:30:55 in 8ms

[ ]: mol.GetAtomWithIdx(
```

```
Docstring:
GetAtomWithIdx( (Mol)arg1, (int)arg2) -> Atom :
    Returns a particular Atom.

      ARGUMENTS:
        - idx: which Atom to return

      NOTE: atom indices start at 0
```

- Some hints in Jupyter Lab, however
  - `arg1` should rather read `self`
  - `arg2` should rather read `idx`

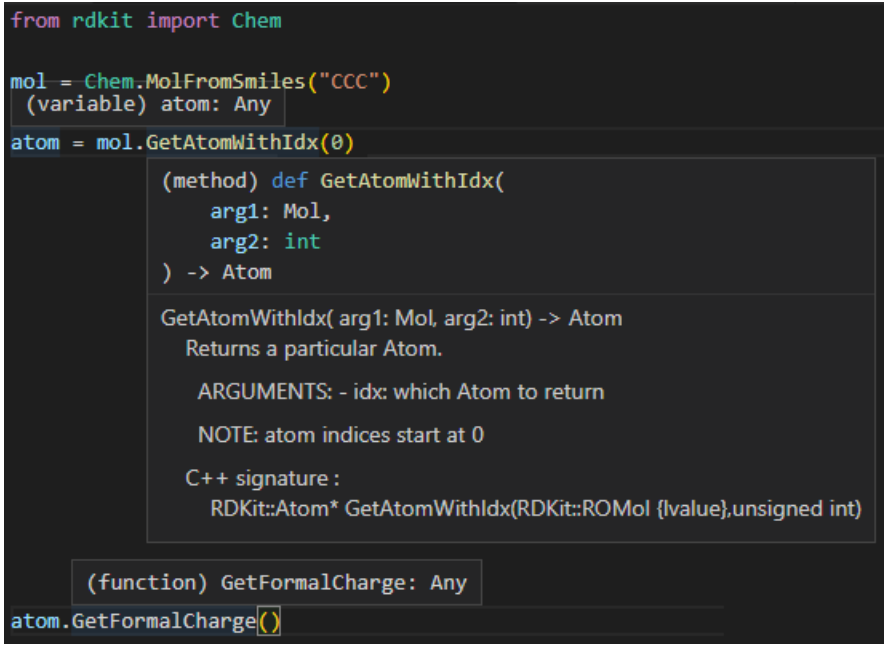
gen_rdkit_stubs.py

- Based on pybind11_stubgen
  - Implements a docstring pre-processing hook to edit `boost::python` docstrings before feeding them to `pybind11_stubgen`
  - Traverses the `rdkit` module to generate all relevant stubs
  - Many non-static methods are incorrectly labelled as `staticmethod` due to lack of `self` parameter in the docstring signature
  - `Any` return types



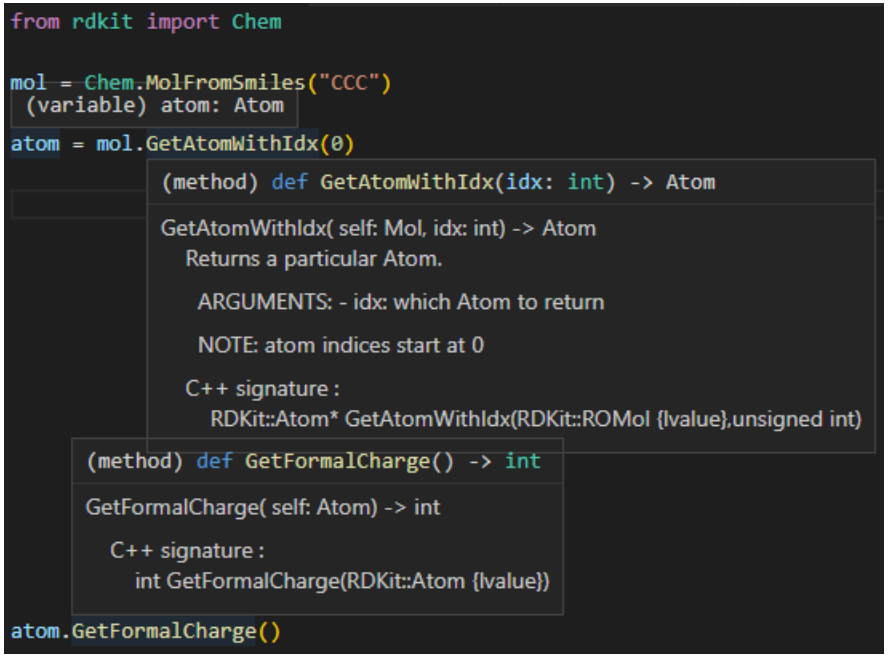
patch_rdkit_docstrings

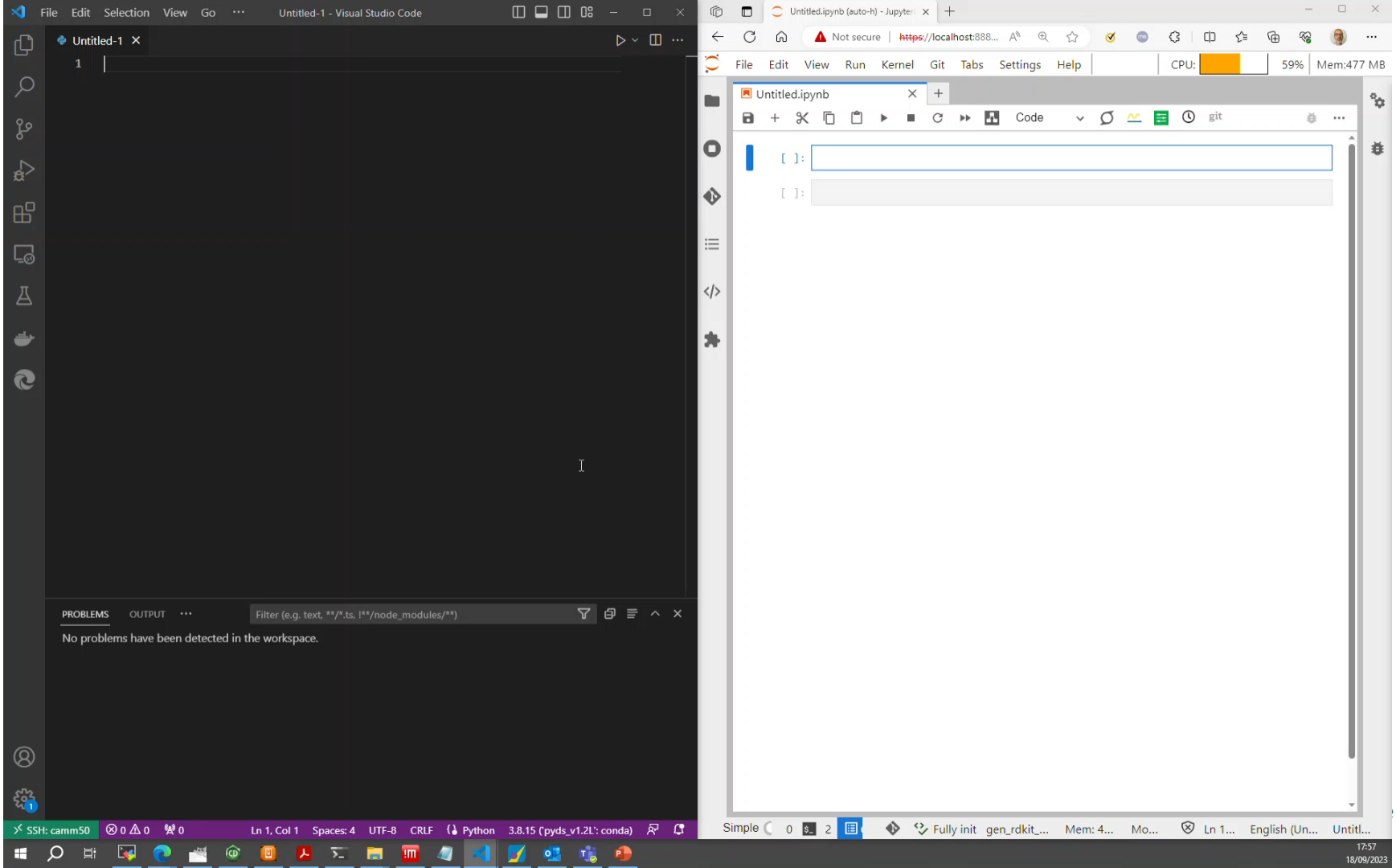
- Based on `clang AST` parsing
  - Runs in parallel through `multiprocessing`
  - Generates AST file for each C++ RDKit Python wrapper
  - Finds docstrings that need missing `self` parameter
  - Finds docstrings with `arg1`, `arg2`, ... parameter names and replaces them with parameter names extracted from C++ function signatures
  - Patches C++ sources



New RDKit D.O.C. strings

- The extent of automated patching is quite large:
  - `$ git diff | grep -c '^+'`
    `1741`
    `$ git diff | grep -c '^-'`
    `1180`
- This is not something that can be done manually, unless one has a lot of time and patience on his/her hands




File Edit Selection View Go
Untitled-1 - Visual Studio Code
Untitled-1
PROBLEMS
OUTPUT
Filter (e.g. text, **/*.ts, !**/node_modules/**)
No problems have been detected in the workspace.
SSH: camm50
Ln 1, Col 1
Spaces: 4
UTF-8
CRLF
Python
3.8.15 ('pyds_v1.2L': conda)
Untitled.ipynb (auto-h) - Jupyter
Not secure
https://localhost:888...
File Edit View Run Kernel Git Tabs Settings Help
CPU:
59%
Mem:477 MB
Untitled.ipynb
Code
git
[ ]:
[ ]:
Simple
Fully init
gen_rdkit_...
Mem: 4...
Mo...
Ln 1...
English (Un...
Untitl...
17:57
18/09/2023

Thank you

NOVARTIS | Reimagining Medicine