Improving the reproducibility of cheminformatics workflows with **`chembl-downloader`**

Charles Tapley Hoyt

orcid:0000-0003-4423-4370

RDKit User Group Meeting - September 21st, 2023

Download https://bit.ly/cth-rdkit-ugm-2023,

Work built on top of ChEMBL goes out of date

*Obviously incomplete lists

Databases	ChEMBL		Reference
	Version	Year	
ExCAPE-DB	20	2015	Sun *et al.*, 2017
Deep Confidence	23	2017	Cortés-Ciriano & Bender, 2019
Consensus Dataset	28	2021	Sigkeit *et al.*, 2022
Papyrus	29	2021	Béquignon *et al.*, 2023

Writing

- Blog posts and software documentation (e.g., practicalcheminformatics.blogspot.com/2022/01/the-solubility-forecast-index)
- Peer-reviewed articles (e.g., Nonadditivity Analysis (Kramer, 2019) used ChEMBL 23)

Current Pain Points

Issues:

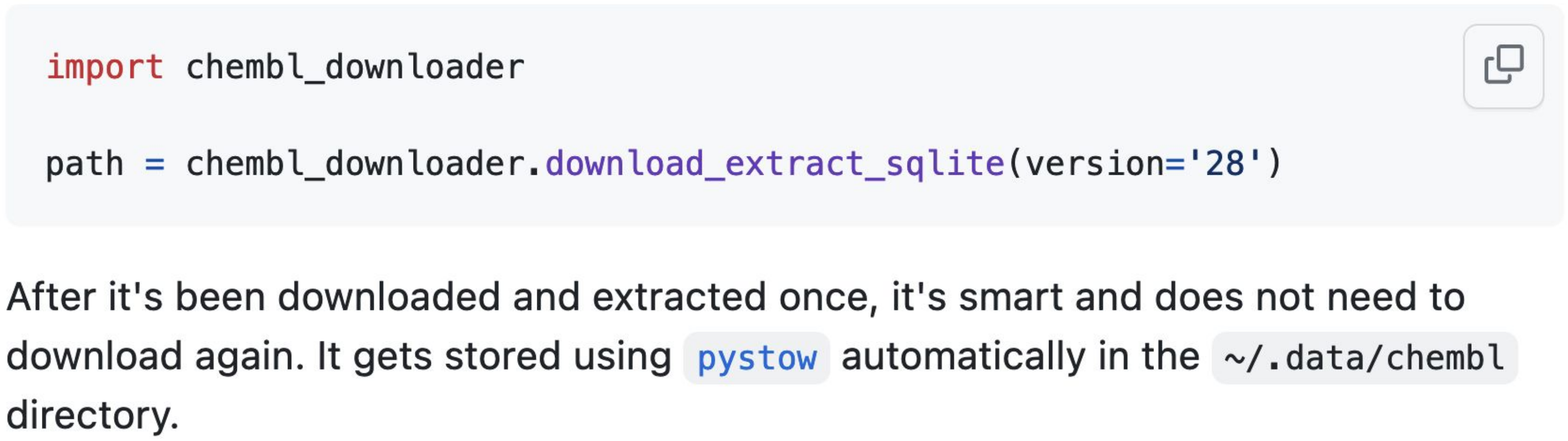
- Manual download and uncompression of data isn't reproducible
- Scripts for processing data often aren't version controlled nor published

Want:

- Automated download and uncompression of data
    - Be able to specify version or just get the latest
- Mid-level utilities for accessing and querying SDF, SQL, and other ChEMBL artifacts
- (optional) High-level tools for common patterns

Solution: `chembl-downloader`

Getting Data

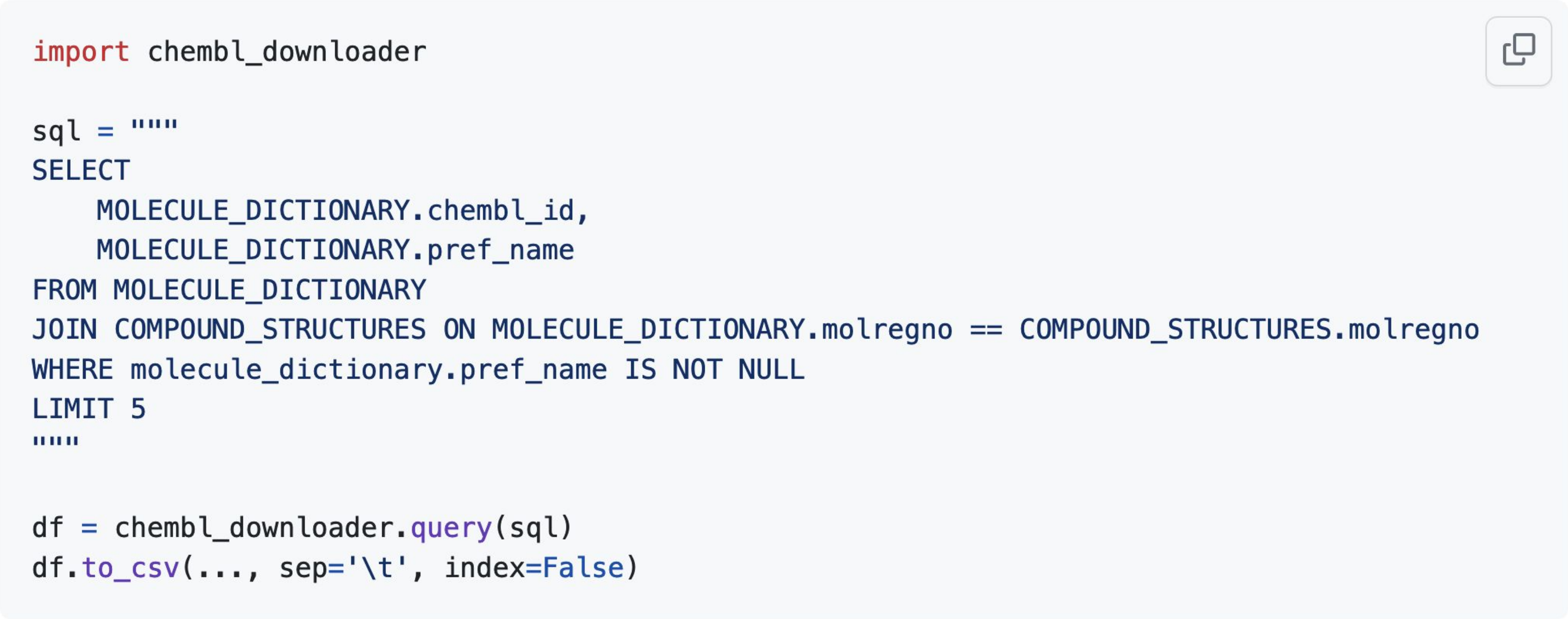


```
import chembl_downloader

path = chembl_downloader.download_extract_sqlite(version='28')
```

After it's been downloaded and extracted once, it's smart and does not need to download again. It gets stored using `pystow` automatically in the `~/.data/chembl` directory.

Querying SQL database

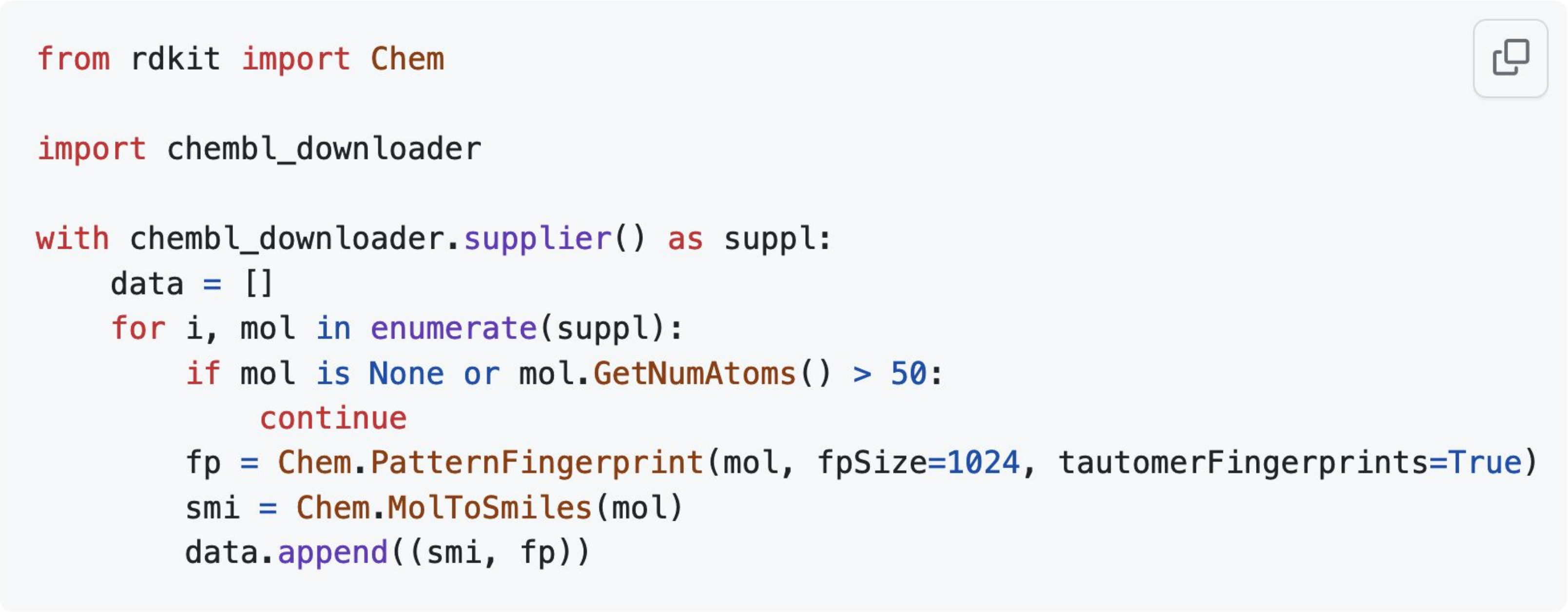


```
import chembl_downloader

sql = """
SELECT
    MOLECULE_DICTIONARY.chembl_id,
    MOLECULE_DICTIONARY.pref_name
FROM MOLECULE_DICTIONARY
JOIN COMPOUND_STRUCTURES ON MOLECULE_DICTIONARY.molregno == COMPOUND_STRUCTURES.molregno
WHERE molecule_dictionary.pref_name IS NOT NULL
LIMIT 5
"""

df = chembl_downloader.query(sql)
df.to_csv(..., sep='\t', index=False)
```

High-level Integrations



```python
from rdkit import Chem

import chembl_downloader

with chembl_downloader.supplier() as suppl:
    data = []
    for i, mol in enumerate(suppl):
        if mol is None or mol.GetNumAtoms() > 50:
            continue
        fp = Chem.PatternFingerprint(mol, fpSize=1024, tautomerFingerprints=True)
        smi = Chem.MolToSmiles(mol)
        data.append((smi, fp))
```

Also for RDKit substructures, pre-build Morgan FPs, chemfp, canned SQL queries, and more

Thanks! Suggestions welcome.

This Presentation: https://bit.ly/cth-rdkit-ugm-2023

Code and Examples: https://github.com/cthoyt/chembl-downloader

Issue Tracker: https://github.com/cthoyt/chembl-downloader/issues

Documentation: https://chembl-downloader.readthedocs.io

Installation: `pip install chembl-downloader`

Users in the wild:
https://github.com/search?q=chembl_downloader+-user%3Acthoyt&type=Code

Funding
Automating Scientific Knowledge Extraction and Modeling (ASKEM) Program - Defense Advanced Research Projects Agency (DARPA) award HR00112220036