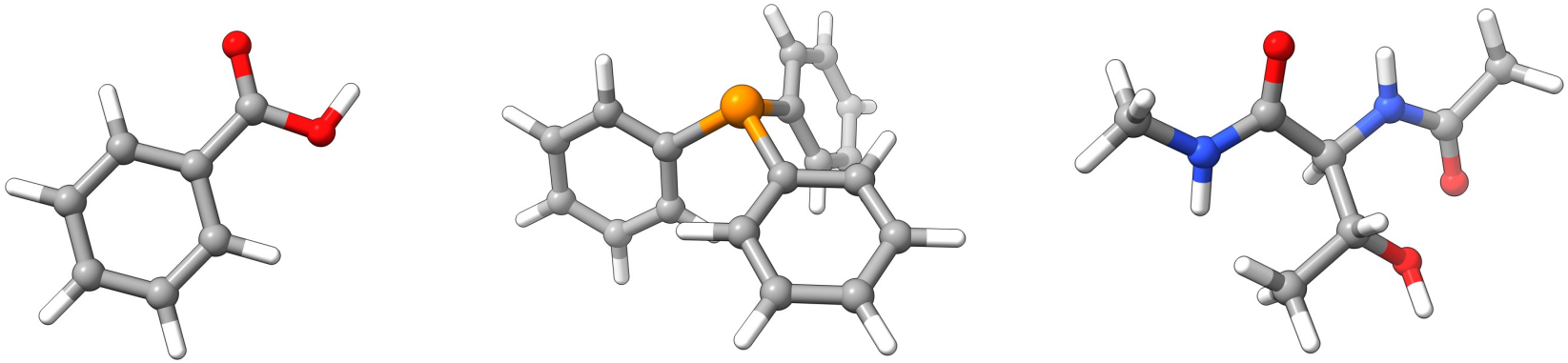
Computational chemistry databases of small organic molecules



Jamie A. Cadge

Lecturer (Assistant Professor) - Department of Chemistry & Department of Chemical Engineering
University of Bath, U.K.

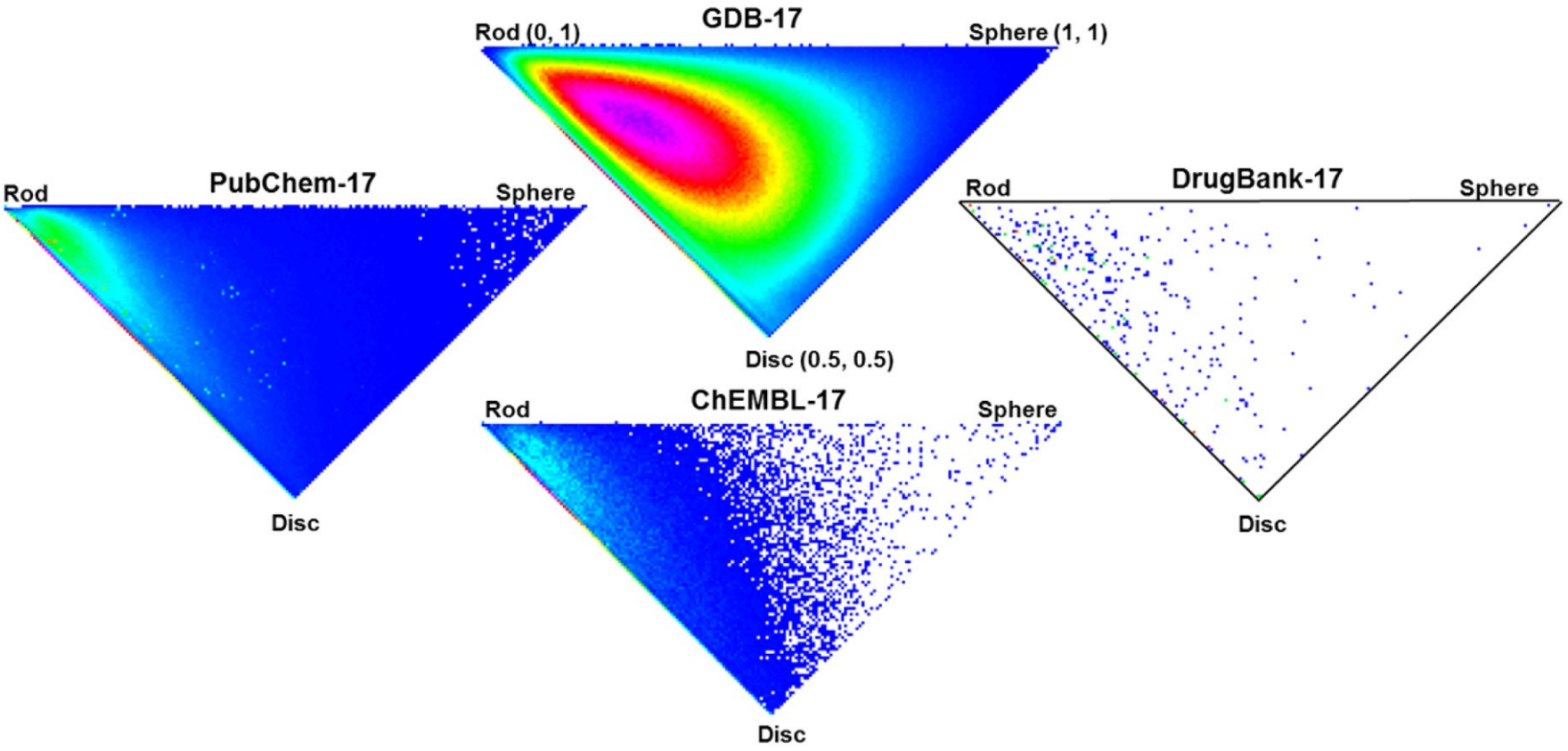
jac340@bath.ac.uk

https://researchportal.bath.ac.uk/en/persons/jamie-cadge

Small molecule chemical space is vast!

The GDB-17 "Chemical Universe Database" contains 166 billion organic small molecules!

J. Chem. Inf. Model. **2012**, *52*, 2864

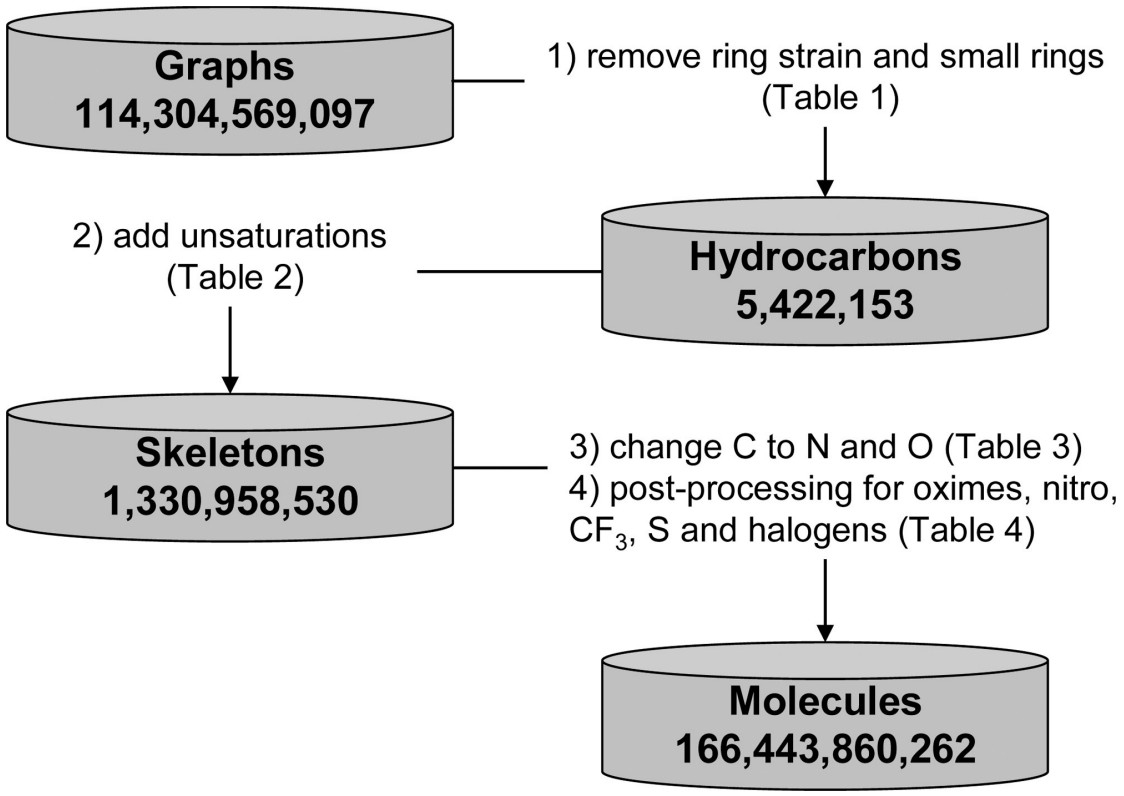


Won't be covering all examples of databases here, but a good comprehensive list can be found in:
Acc.Chem. Res. **2015**, *48*, 722.

Small molecule chemical space is vast!

The GDB-17 "Chemical Universe Database" contains 166 billion organic small molecules!

J. Chem. Inf. Model. **2012**, *52*, 2864



Won't be covering all examples of databases here, but a good comprehensive list can be found in:
Acc.Chem. Res. **2015**, *48*, 722.

What databases are available free of charge?

Examples for structure retrieval (SMILES, Inchi, etc.)

PubChem

https://pubchem.ncbi.nlm.nih.gov/

122,270,562 unique structures
(September 2025)

Many molecular identifiers including biological activity.
Python API for linking the database to your scripts:
https://docs.pubchempy.org/en/latest/

ZINC20

https://zinc20.docking.org/

>750M structures, "purchasable"

Built for virtual screening of compounds, e.g., by docking.
No simple way of interacting with Python but downloadable in a number of formats (SMILES, CSV, etc.)

eMolecules

https://search.emolecules.com/

>9 million building blocks

Large database of organic compounds / building blocks.
Mostly useful for searching for chemical suppliers.
Export options (SDF / CSV / SMILES etc.)

What databases are available free of charge?

Examples of databases which are oriented towards bioactivity / drug discovery applications

https://www.ebi.ac.uk/chembl/

2.5M compounds

Chemical, bioactivity and genomic data. Structures, identifiers and properties can be retrieved in many ways (web interface, direct download of SQL database/SDF, etc.).

https://go.drugbank.com/

>500k compounds

Physicochemical properties, ADME data, etc.
Can be downloaded for free after applying for an Academic Licence

PROC. OF THE 8th EUR. CONF. ON PYTHON IN SCIENCE (EUROSCIPY 2015)

Want Drugs? Use Python.

Michał Nowotka*†, George Papadatos†, Mark Davies†, Nathan Dedman†, Anne Hersey†

https://arxiv.org/pdf/1607.00378

Commercially available compounds

Possible to obtain commercial availability data from databases such as Emolecules or ZINC; however, may not always be reliable.

Some vendors supply their catalogs in downloadable formats:

Enamine

Access to their stock catalog as SDFs free of charge after signing up.

Aldrich Market Select
Small Molecule Screening Library Service

Access to purchasable building blocks and screening compounds.
Need to "Request Data Sharing Agreement".

molport

5.3M in-stock screening compounds; **856k** in-stock building blocks.
From 18 verified suppliers.
Free, but have to sign up for some features.

mcule
·MORE·THAN·MOLECULES·

Similar system to molport.
Free, but have to sign up for some features.

“Synthetically feasible” compounds

https://enamine.net/compound-collections/real-compounds/real-database

Enamine REAL (REadily AccessibLe) library
Largest enumerated dataset of synthetically feasible drug-like molecules (10.1B)
Can download the whole library in SDF format.

https://www.labnetwork.com

Similar to the Enamine REAL library (100M compounds) which can be bought from / made by WuXi.
Can download in SDF / CSV format.

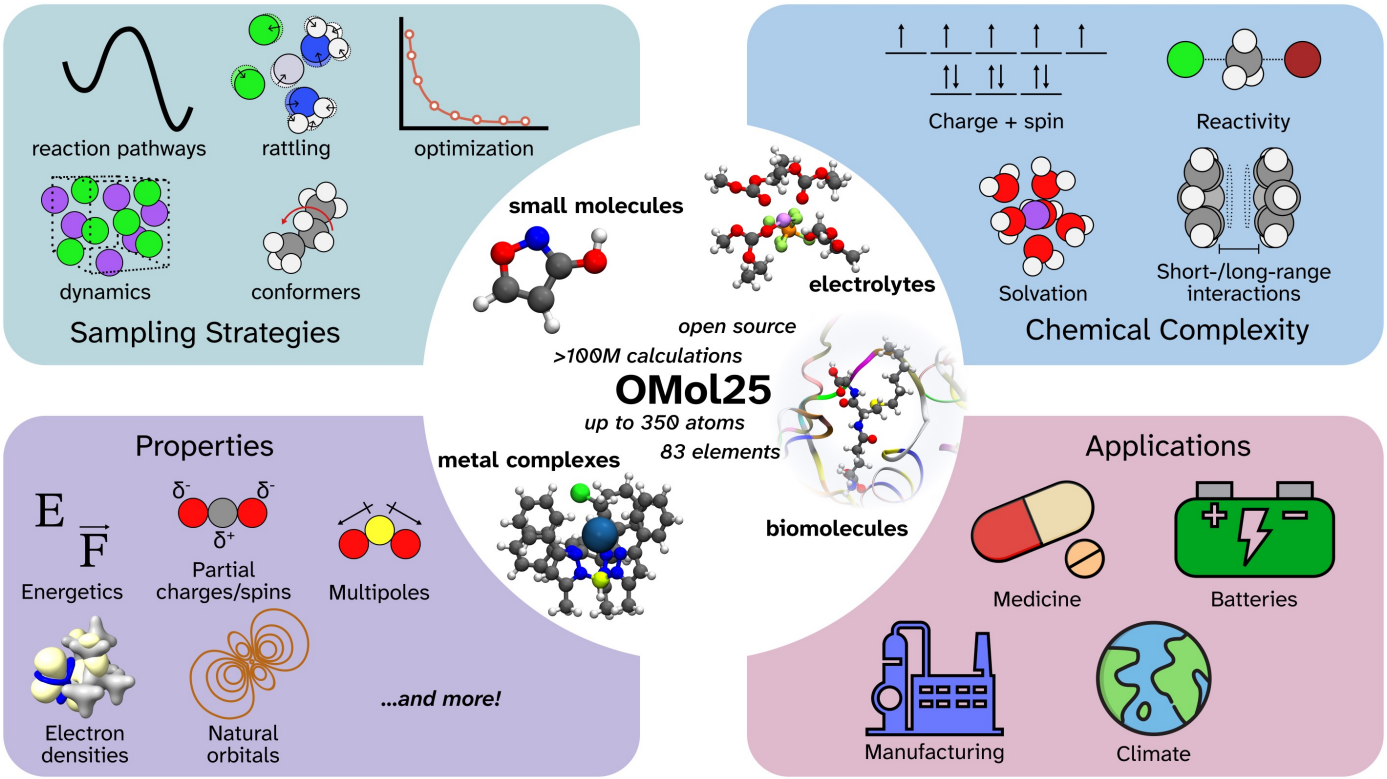
https://ultimate.mcule.com/

Library of synthetically accessible compounds (+100M).
No obvious way of downloading but can search online.

Computational chemistry databases

Databases of compounds which contain features / information from quantum chemical calculations.

Many computational benchmarking sets are available from small molecules to transition metal complexes.



Citation:

https://arxiv.org/abs/2505.08762

OMol25 is the largest high quality molecular DFT dataset spanning biomolecules, metal complexes, electrolytes and community datasets. [**100M DFT single-point calculations** @ ωB97M-V/def2-TZVPD level using ORCA 6].

Designed for the training of molecular chemistry machine learning models (i.e., MLIPs).

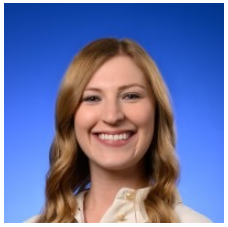
Available for download on Hugging Face: https://huggingface.co/facebook/OMol25

C-CAS / MolSSI Descriptor Libraries Project

Collaboration between the Center for Computer Assisted Synthesis (C-CAS) and the Molecular Sciences Software Institute (MolSSI)

Dr Jessica Nash (MolSSI)

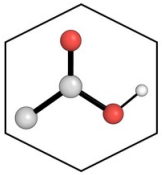
Dr Ellie Peters (Fmr. Sigman Group, now Eli Lilly and Co.)



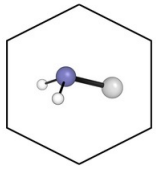
Dr Brittany Haas (Fmr. Sigman Group, now Eli Lilly and Co.)

Dr Melissa Hardy (Fmr. Sigman Group, now GSK)

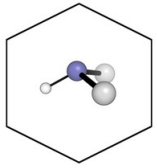
kraken



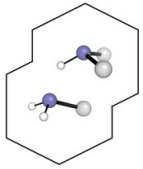
Carboxylic Acids



Primary Alkyl Amines

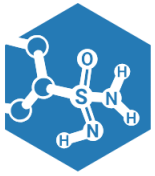


Secondary Alkyl Amines



Combined Alkyl Amines

Anilines



Sulfonimidamides

Cyanoarenes

Sulfonyl Fluorides

How to build a library

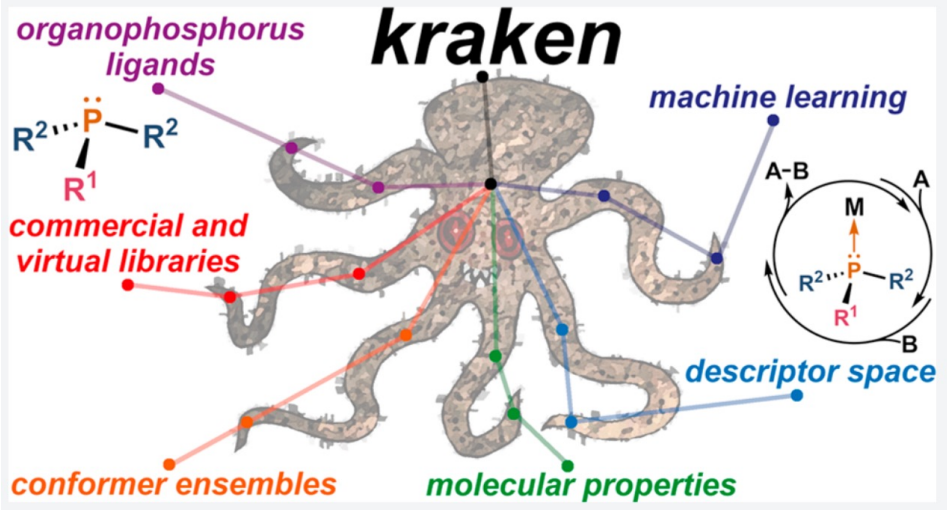
Case study: The *kraken* monophosphine library

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis

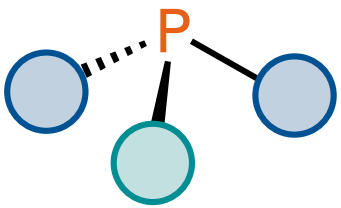
Tobias Gensch,* Gabriel dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, AkshatKumar Nigam, Michael Lindner-D'Addario, Matthew S. Sigman,* and Alán Aspuru-Guzik*

Cite This: *J. Am. Chem. Soc.* 2022, 144, 1205–1217
Read Online



How to build a library

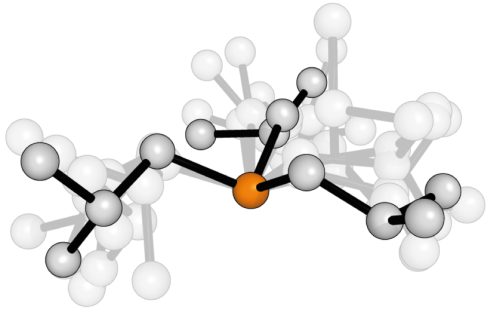
Case study: The *kraken* monophosphine library



Library of >1500 ligands

Commercially available

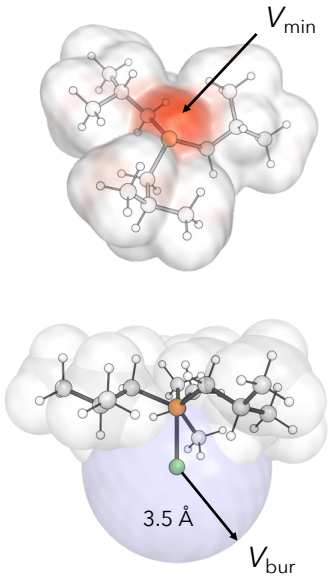
Popular ligands from literature



Conformational ensembles generated

Structures refined using DFT

190 conformationally-weighted descriptors collected



21,417 total conformers – each with **7** DFT calculations

1,381,194 hours total CPU time

How to build a library

Case study: The *kraken* monophosphine library

Extraction directly from a computational output file

```
 NATURAL BOND ORBITALS (Summary):

                                                    Principal Delocalizations
           NBO                 Occupancy    Energy   (geminal,vicinal,remote)
 ================================================================================
  Molecular unit  1  (C70H96P2Pd)
  ------ Lewis ----------------------------------
     1. CR ( 1) P  1             2.00000   -70.46116
     2. CR ( 2) P  1             1.99998   -13.44865
     3. CR ( 3) P  1             1.99999    -4.78513
     4. CR ( 4) P  1             1.99999    -4.78622
     5. CR ( 5) P  1             1.99999    -4.78576
     6. CR ( 1) P  2             2.00000   -70.44278


 Calculating GIAO nuclear magnetic shielding tensors.
  SCF GIAO Magnetic shielding tensor (ppm):
       1  P    Isotropic =   231.6408   Anisotropy =   123.7556
    XX=   243.0521   YX=     7.0705   ZX=   -86.9526
    XY=    25.9896   YY=   212.8392   ZY=     9.1434
    XZ=   -57.7611   YZ=     0.5682   ZZ=   239.0312
```

Use DFT optimized coordinates in Python packages

```
59
pp000067-Pd.log     Energy: -1529698.0941485
Cl        -1.78186       -0.09666        2.55795
P         -1.55034       -0.07122       -0.70104
P          1.58419       -0.07040       -0.65201
C         -0.64491       -0.63171       -2.23980
H         -0.49183       -1.72258       -2.15739
H         -1.25244       -0.44380       -3.14337
C          0.70547        0.09775       -2.30421
H          0.55004        1.17477       -2.49443
H          1.33789       -0.29481       -3.12002
C         -2.37187        1.55002       -1.21052
H         -2.36970        1.49741       -2.31943
C         -3.15542       -1.02515       -0.52023
H         -3.27089       -1.08382        0.57972
C          2.99060        1.15146       -0.67085
```

Molecule stored as an XYZ coordinate file

MORFEUS
molecular features for machine learning

DBSTEP

Considerations to be made

The computational load can be high!

How to run the conformational search? CHEAP!

How many conformers should be taken forward to DFT? EXPENSIVE!

What features to collect? (Min, max, Boltzmann average, LEC, etc.)

What is the best representation of the feature space?

Automated workflows exist:

Auto-QChem (Doyle): *React. Chem. Eng.* **2022**, *7*, 1276

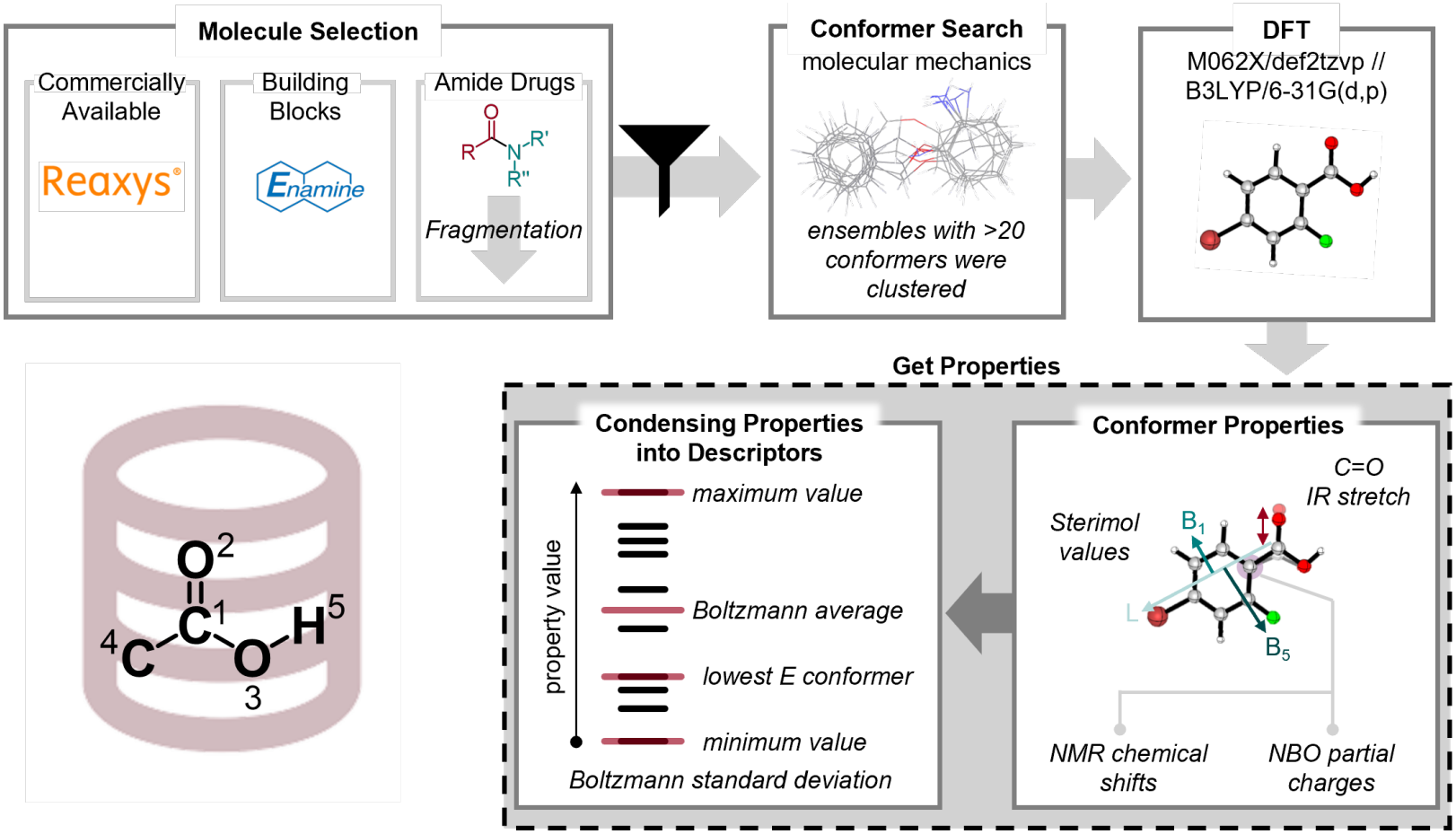
AQME (Paton): *WIREs Comput Mol Sci.* **2023**, *13*, e1663

Molli (Denmark): *J. Chem. Inf. Mod.* **2024**, *64*, 8083

Sigman/Paton/Coley: *Digital Discovery* **2025**, *4*, 222

How to build a library

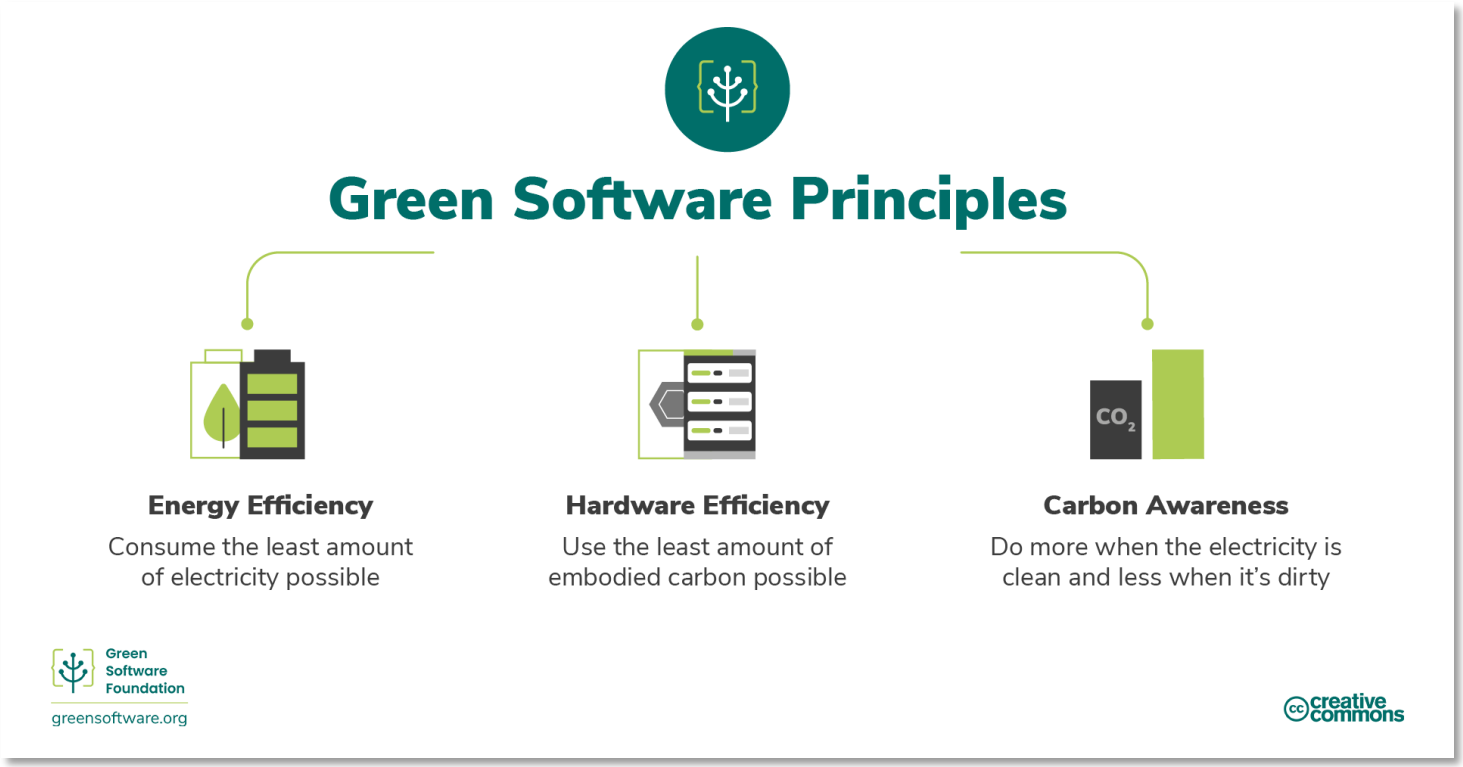
Case study: Carboxylic acid library for predicting amide coupling rates

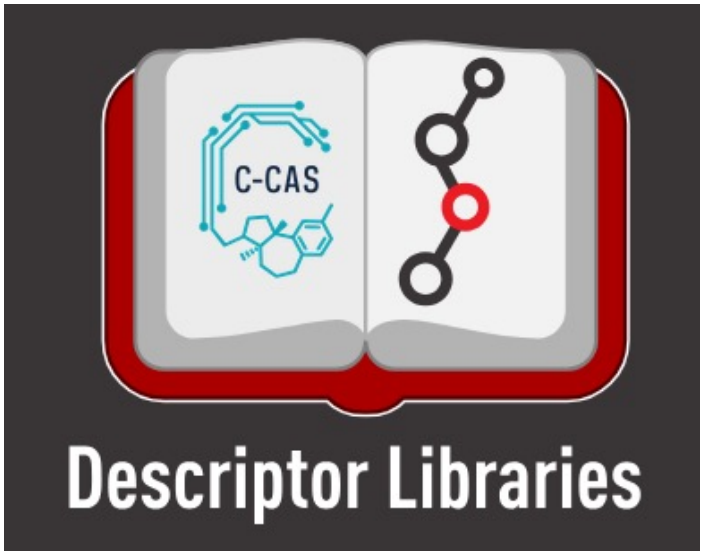


Sustainability and DFT library development

Considerable environmental impact (energy and CO_2 emissions) for running a large set of calculations.

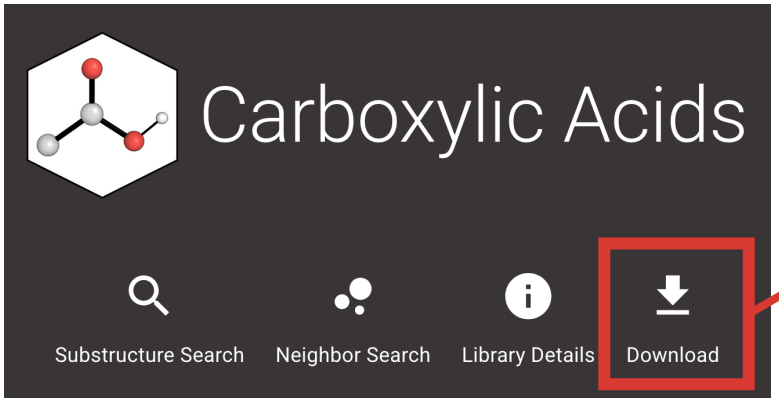
Green Chem. **2024**, *26*, 8669





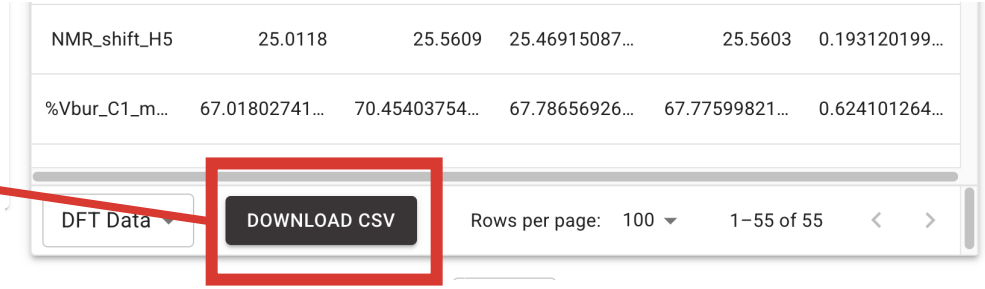
Website Demonstration

Descriptor Libraries project download options



Download full library:
- All reduced features
- Features for each conformer
- SMILES strings
- PCA/UMAP coordinates

Download CSV of reduced features for each compound



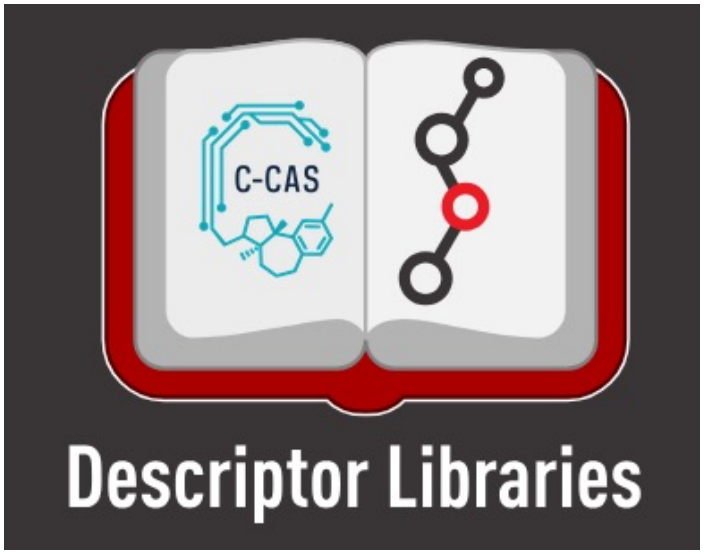
Still beta!

Descriptor Libraries API acids 0.1.0 OAS 3.1

/api/acids/openapi.json

https://descriptor-libraries.molssi.org/api/acids/docs#/current/

Change URL to name of library



API Tutorial

Important: before we begin…

Please make sure that you have downloaded the "requests" package into your Python environment to access the MolSSI API!

```
pip install requests
```

Clone or download the GitHub repository:

```
git clone https://github.com/jcadge/CAMLC2025-MolSSIDatabases.git
```