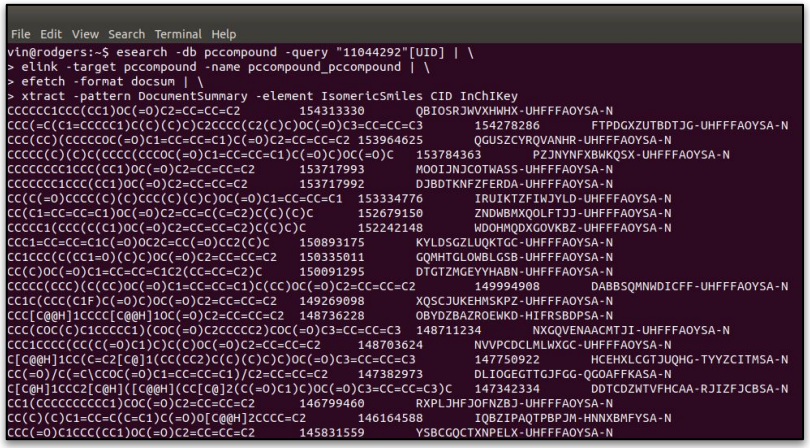
Lowering barriers to teaching programmatic chemical information searching

use-cases demonstrating the NCBI Entrez Direct (EDirect) unix tool



Disclaimer: This work is not affiliated with NCBI/NLM/NIH

ACS Spring 2021 National Meeting
2021-04-16

Vincent F. Scalfani
Science and Engineering Librarian
The University of Alabama
vfscalfani@ua.edu

Chemical Information Searching

What does chemical information searching have to do with (FAIR) data sharing?

- Sharing: machine-readable chemical data sharing, e.g. PubChem, ChemSpider, and Wikidata.
- Search/Retrieval: these databases have robust web services that allow programmatic access.

Hypothesis: If researchers search for chemical information programmatically as part of their regular workflows, they will be more likely to share their own machine-readable chemical data.

If we agree that this hypothesis is reasonable:

Librarians and information educators will need to incorporate programmatic chemical information searching into their regular teaching and reference. In other words, *"The Future of Chemical Information is Now"* [1]

[1] Williams, A. J.; Pence, H. E. *Chemistry International* **2017**, *39* (3).

Current Teaching Tools/Methods

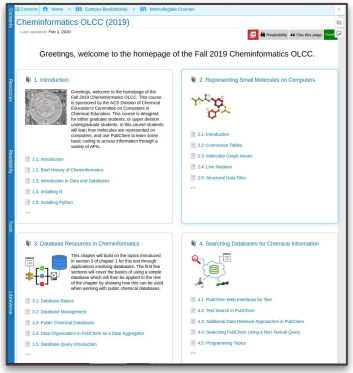


Fig [2]

Cheminformatics Online Chemistry Course (OLCC) [1,2]. 2019 edition includes lessons on accessing PubChem PUG-REST API using **Python/Jupyter Notebooks, R/RStudio, and Mathematica Notebooks.**

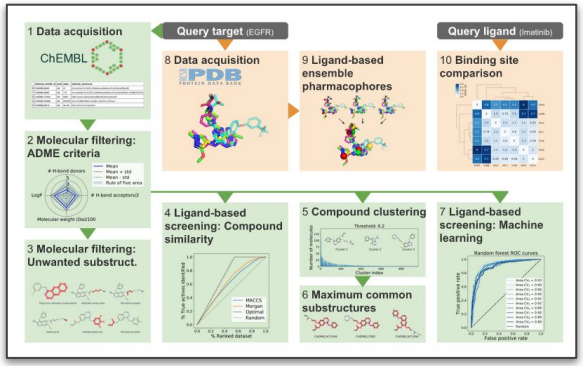


Fig [4]

TeachOpenCADD [3,4], collection of computer-aided drug design tutorials, includes several for programmatic chemical information retrieval from ChEMBL / PDB RESTful API using **Python/Jupyter Notebooks**.

Fig [6]

MATLAB Live Scripts [5,6], collection of notebooks using PubChem PUG-REST API and PubChem Structured Data Query (SDQ) for searching and compiling data.

[1] Kim, S. et al. *J. Chem. Educ.* **2021**, *98* (2), 416–425.; [2] https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics_OLCC_(2019)
[3] Sydow, D.; Morger, A.; Driller, M.; Volkamer, A. *J Cheminform* **2019**, 11 (1), 29. [4] https://github.com/volkamerlab/teachopencadd
[5] Scalfani, V. F. et al. *Chem. Eng. Ed.* **2020**, *54* (4).; [6] https://github.com/vfscalfani/MATLAB-cheminformatics

Limitation of Current Teaching Tools/Methods

Well documented methods and accompanying **digital learning objects** for teaching programmatic chemical information searching/compilation are:

- excellent for independent study and courses
    - i.e. the barrier can be high -- need to introduce API syntax, programming language syntax and concepts, computational notebook setup, and more.
- excellent for users that have prior programming experience.
- more challenging to teach as a chemistry librarian that may only see students in one 60 minute or less workshop.

What to do as a Chemistry Librarian?

It would be great if we could get students (*with no prior programming experience*) started with programmatic chemical information searching/compilation in one workshop.

Need to lower the barrier to get students **started**** with programmatic chemical information searching.

NCBI's Entrez Direct (EDirect) is a tool that I think makes this possible. There is limited syntax to learn and it is approachable for beginners.

***Keyword is **started** here as this is not to say we must teach all programmatic chemical information searching within one workshop.*

What is Entrez Direct (EDirect) [1]?

- Free command-line program from NCBI that allows E-utilities programmatic access to NCBI databases such as PubMed, PubChem, Gene, Taxonomy, etc. directly within a Unix terminal window.
- Can be installed on Unix, Unix-like (e.g., GNU/Linux) distributions, Mac OS, and Windows with Cygwin Unix-emulation.
- Minimal syntax, EDirect constructs the programmatic web URLs for you and includes programs to help you format and process the data into tabular formats.
- ***It's fun to use. Can search for chemical information and compile data all in a terminal window.***

[1] https://www.ncbi.nlm.nih.gov/books/NBK179288/

EDirect Unix Programs [1]

EDirect contains several individual programs. Here is a subset:

1. **einfo -** prints fields and links indexed in each database
2. **esearch** - performs an NCBI Entrez database search based on a specified database and query
3. **efetch** - downloads the esearch query results in a specified format such as XML
4. **xtract** - extracts selected data values from XML
5. **elink** - finds associated records within a specified database
6. **efilter** - limits results (e.g., by date, information type, etc.)

Example syntax:

```
eprogram -argument input
esearch -db pubmed -query "imidazolium AND bacteria"
```

Example workflows with unix pipes:

```
$ esearch | efetch | xtract
$ esearch | elink | efilter | efetch | xtract
$ esearch | elink | efilter | efetch | xtract | sort | uniq
-c
$ esearch | efetch | xtract | openbabel.obabel -arg
```

[1] See the official manual for other EDirect programs and examples: https://www.ncbi.nlm.nih.gov/books/NBK179288/

Available EDirect Guides and Use in the Literature

- NIH/NLM has a 5 part series of using EDirect with PubMed and there are several **PubMed** examples in the EDirect manual [1,2].
- The EDirect manual [2] and EDirect Cookbook [3] has EDirect examples for **Protein, Structure, Gene, Nucleotide, Taxonomy, and Assembly databases.**
- Scanning the ~80 cited references (Google Scholar) to Kans, Jonathan. "Entrez direct: E-utilities on the UNIX command line." *Entrez Programming Utilities Help.* National Center for Biotechnology Information (US), 2020, revealed EDirect literature use cases with **PubMed, Gene, Nucleotide, RefSeq, and GEO.**

I could not find any examples using EDirect with PubChem; that is, for small molecule information.

[1] https://dataguide.nlm.nih.gov/classes.html
[2] https://www.ncbi.nlm.nih.gov/books/NBK179288/
[3] https://github.com/NCBI-Hackathons/EDirectCookbook

Notes on Using EDirect with PubChem

There are limitations when accessing PubChem data with EDirect (uses E-Utilities), but many text or numeric based searches that do not require chemical interpretation are okay [1].

Examples of what you can **not** do with EDirect and PubChem [1]:

1. Substructure or superstructure searches.
2. Access certain tabular data on PubChem Compound pages like bioactivity tables.

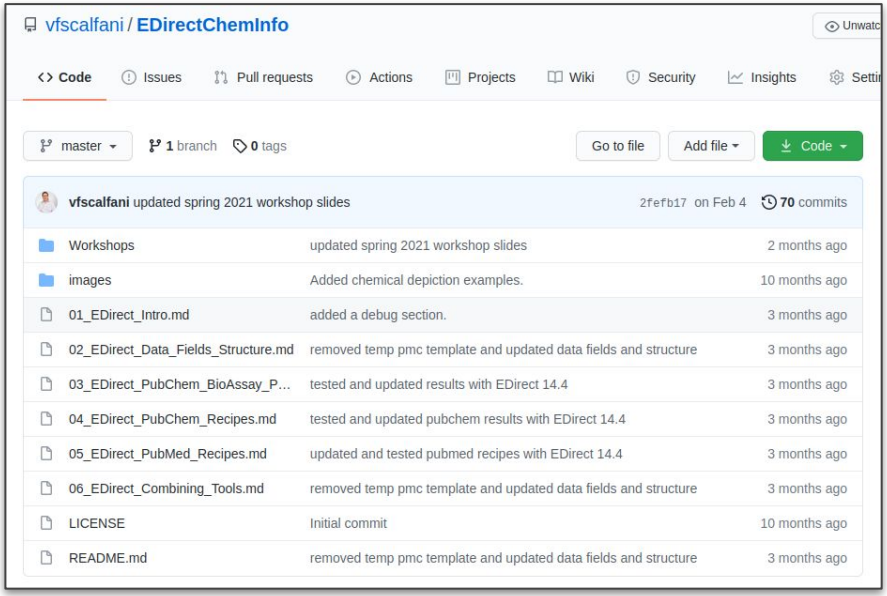
Examples of what you can do:

1. Search compounds by identifier, InChIKey, and specific numeric attributes (e.g., rotatable bond counts) or annotated attributes (e.g., has active bioassay data)
2. Find related information for a compound in other databases like PubChem Substance, PubMed, and PubChem BioAssay.
3. Retrieve SMILES or pre-computed properties like number of chiral atoms.
4. Retrieve pre-computed related compounds such as same connectivity or similar compounds.

[1] Kim, S.; Thiessen, P. A.; Bolton, E. E. Programmatic Retrieval of Small Molecule Information from PubChem Using PUG-REST, Methods in Pharmacology and Toxicology, 2018.

EDirectChemInfo Repository

To demonstrate how to use EDirect for programmatically retrieving small molecule information, I created a GitHub repository with 50+ code “recipes” and workshop materials:



PubChem Compound, PubChem Substance, PubChem BioAssay, and PubMed

structure <> Bioactivity <> literature

https://github.com/vfscalfani/EDirectChemInfo

MIT license

PubChem Compound EDirect Fields and Data [1]

Search Fields, same as online Entrez Database

https://www.ncbi.nlm.nih.gov/pccompound/advanced

AC	ActiveAidCount
ACC	AtomChiralCount
ACDC	AtomChiralDefCount
ACUC	AtomChiralUndefCount
ALL	All Fields
BCC	BondChiralCount
BCDC	BondChiralDefCount
BCUC	BondChiralUndefCount
CDAT	CreateDate
CPLX	Complexity
CSYN	CompleteSynonym
CUC	CovalentUnitCount
DCNT	DepositorCount
DCSY	DepositorCompleteSynonym
DSYN	DepositorSynonym
ELMT	Element
EMAS	ExactMass
FILT	Filter
HAC	HeavyAtomCount
HBAC	HydrogenBondAcceptorCount
HBDC	HydrogenBondDonorCount
IAC	IsotopeAtomCount
IKEY	InChIKey
INCH	InChI
MMAS	MonoisotopicMass
MSHT	MeSHTerm
MW	MolecularWeight
PAID	PharmActionID
PHMA	PharmAction
RBC	RotatableBondCount
SID	SubstanceID
SRCC	SourceCategory
SRC	SourceName
STID	StructureID
SYNO	Synonym
TAC	TotalAidCount
TFC	TotalFormalCharge
TPSA	TPSA
UID	CompoundID
UPAC	IUPACName
XLGP	XLogP

Data available in retrieved docsum XML records

CID
SourceCategoryList
CreateDate
SynonymList
CanonicalSmiles
IsomericSmiles
RotatableBondCount
MolecularFormula
MolecularWeight
MolecularWeightSort
TotalFormalCharge
XLogP
HydrogenBondDonorCount
HydrogenBondAcceptorCount
Complexity
ComplexitySort
HeavyAtomCount
AtomChiralCount
AtomChiralDefCount
AtomChiralUndefCount
BondChiralCount
BondChiralDefCount
BondChiralUndefCount
IsotopeAtomCount
CovalentUnitCount
TPSA
ActiveAidCount
TotalAidCount
InChIKey
ProbeAidCount
InChI

For available related links, see:
https://eutils.ncbi.nlm.nih.gov/entrez/query/static/entrezlinks.html

[1] https://github.com/vfscalfani/EDirectChemInfo

Example 1

search PubChem Compound via InChIKey and retrieve data [1]:

```
$ esearch -db pccompound -query "NJTXJDYZPQNTSM-WMZOPIPTSA-N"[IKEY] | \
> efetch -format docsum | \
> xtract -pattern DocumentSummary -element IsomericSmiles CID MolecularWeight
C[C@]12CCC(=O)C=C1CCC[C@@H]2OC(=O)C3=CC=CC=C3    11044292    284.300
```

[1] https://github.com/vfscalfani/EDirectChemInfo
tested on 2021.01.27, EDirect 14.4, total count was 1.

Example 2

retrieve pre-computed linked similar compounds for a CID in PubChem [1]:

```
$ esearch -db pccompound -query "11044292"[UID] | \
> elink -target pccompound -name pccompound_pccompound | \
> efetch -format docsum | \
> xtract -pattern DocumentSummary -element IsomericSmiles CID InChIKey IUPACName
...
CC(C1=CC=CC=C1)OC(=O)C2=CC=C(C=C2)C(C)(C)C  152679150  ZNDWBMXQOLFTJJ-UHFFFAOYSA-N      1-phenylethyl
4-tert-butylbenzoate
CCCCC1(CCC(C(C1)OC(=O)C2=CC=CC=C2)C(C)C)C   152242148  WDOHMQDXGOVKBZ-UHFFFAOYSA-N
(5-butyl-5-methyl-2-propan-2-ylcyclohexyl) benzoate
CCC1=CC=CC1C(=O)OC2C=CC(=O)CC2(C)C  150893175  KYLDSGZLUQKTGC-UHFFFAOYSA-N
(6,6-dimethyl-4-oxocyclohex-2-en-1-yl) 2-ethylbenzoate
CC1CCC(C(CC1=O)(C)C)OC(=O)C2=CC=CC=C2  150335011  GQMHTGLOWBLGSB-UHFFFAOYSA-N
(2,2,5-trimethyl-4-oxocycloheptyl) benzoate
...
```

[1] https://github.com/vfscalfani/EDirectChemInfo
tested on 2021.01.27, EDirect 14.4, total count was 238.

Example 3

retrieve linked PubMed references for a PubChem CID [1]

```
$ esearch -db pccompound -query 174076[uid] | \
> elink -target pubmed -name pccompound_pubmed | \
> efetch -format xml | \
> xtract -pattern PubmedArticle -element MedlineCitation/PMID -first Author/LastName \
> Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn
22957575   Gabl S     J Chem Phys      2012 137  9     094501
22868451   Zhang Y    Phys Chem Chem Phys   2012 14   35    12157-64
22859056   Malberg    F     Phys Chem Chem Phys   2012 14   35    12079-82
22852554   Zhang Y    J Phys Chem B    2012 116  33    10036-48
22662183   Zhang BB   PLoS ONE   2012 7    5     e37641
...
```

[1] https://github.com/vfscalfani/EDirectChemInfo

tested on 2021.01.26, EDirect 14.4, total count was 102.

Example 4

search PubMed with a text query, then retrieve linked PubChem Compounds [1].

To my knowledge, it's not possible to do this in the new PubMed Database.

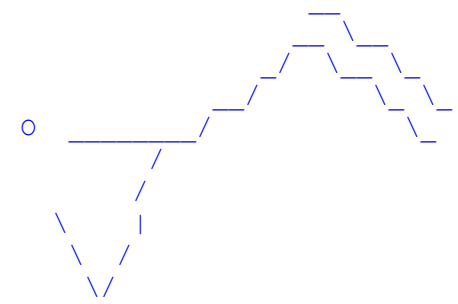
```
$ esearch -db pubmed -query "\"ionic liquids\"[MESH] AND imidazolium" | \
> elink -target pccompound -name pubmed_pccompound | \
> efetch -format docsum | \
> xtract -pattern DocumentSummary -element IsomericSmiles CID InChIKey
C1=CC2=CC(=C(C(=C2C(=O)C(=C1)O)O)O)O135403797   WDGFFVCWBZVLCE-UHFFFAOYSA-N
C1=NC2=C(N1[C@H]3[C@@H]([C@@H]([C@H](O3)CO)O)O)N=C(NC2=O)N  135398635   NYHBQMYGNKIUIF-UUOKFMHZSA-N
CCCCCCCCN1C=C[N+](=C1C2=[N+](C=CN2CCCC)C)C123995430   DRJFJBHYMOHPHX-UHFFFAOYSA-N
CC(=O)OC1=[N+](C=CN1CC=C)C    123614562   XSXMFLUARQMOLS-UHFFFAOYSA-N
C[N+]1=C(N(C=C1)CCCCCCCCCCCCS)OC(=O)OC2=[N+](C=CN2CCCCCCCCCCCCS)C 123431445   DRJOSAVMFCYCSU-UHFFFAOYSA-P
...
```

[1] https://github.com/vfscalfani/EDirectChemInfo
tested on 2021.01.27, EDirect 14.4, total count was 395.

Example 5

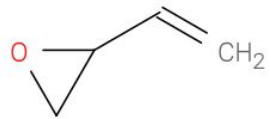
get SMILES from PubChem and depict with Open Babel cheminformatics toolkit [1,2].

```
$ esearch -db pccompound -query "13586"[UID] | \
> efetch -format docsum | \
> xtract -pattern DocumentSummary -element IsomericSmiles | \
> openbabel.obabel -ismi -oascii -xh 10
```



```
1 molecule converted
```

```
$ esearch -db pccompound -query "13586"[UID] | \
> efetch -format docsum | \
> xtract -pattern DocumentSummary -element IsomericSmiles CID | \
> openbabel.obabel -ismi -O 13586.png
1 molecule converted
```

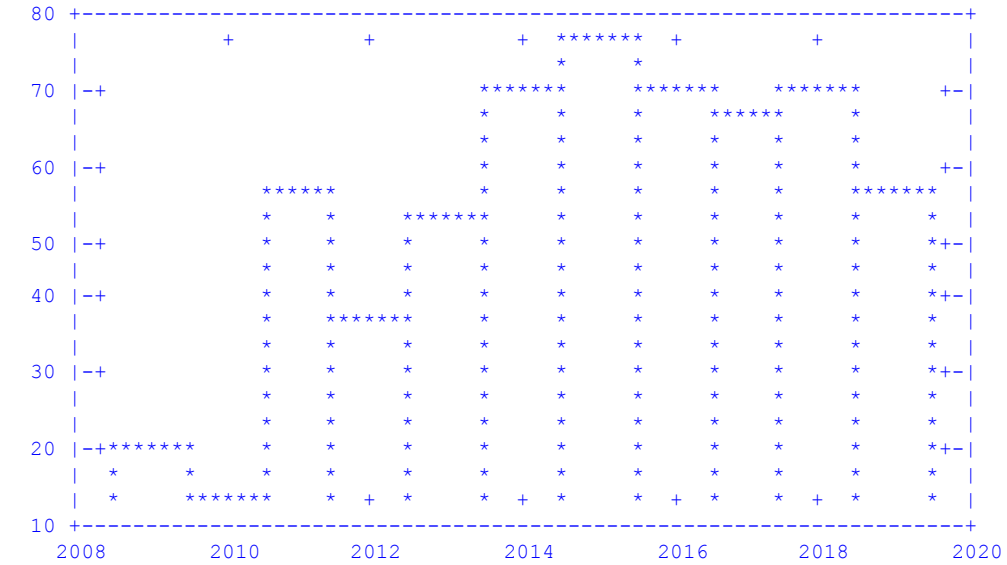


13586

[1] https://github.com/vfscalfani/EDirectChemInfo
[2] O'Boyle, N. M. et al. *Journal of Cheminformatics* **2011**, *3*.
Tested with Open Babel v3.0.0 installed from Snap.

Example 6

```
$ esearch -db pubmed -query "J Cheminform[JOUR]" | \
> efetch -format docsum | \
> xtract -pattern DocumentSummary -element PubDate | \
> cut -d " " -f 1 | \
> sort-uniq-count-rank | \
> sort -k2 | \
> gnuplot -e "set term dumb; plot '-' using 2:1 with boxes notitle"
```



Plot PubMed indexed *Journal of Cheminformatics* articles per year directly in gnuplot [1,2].

Here I plotted as ascii art so it displays in the terminal, but you can of course create more traditional plots outside of terminal window.

[1] https://github.com/vfscalfani/EDirectChemInfo
[2] http://www.gnuplot.info/
Tested with gnuplot-x11 5.2.8.

Initial Observations

Have taught EDirect live a couple times thus far, a few preliminary observations:

1. EDirect allows introduction of a variety of transferable programming concepts (e.g., can later use PubChem PUG-REST)

    a. Construct a query --> retrieve data --> parse/extract data
    b. For-loops for multiple requests

2. Can be combined with Unix utilities, cheminformatics toolkits, and plotting software allowing for discussions about **reproducible** data compilation and analysis workflows.

3. Feels more approachable for beginners, and that I can successfully get users started in a single workshop.

Teaching Limitations

1. The search query/request is more hidden compared to a traditional RESTful query:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/inchikey/NJTXJDYZPQNTSM-WMZOPIPTSA-N/property/IsomericSMILES,MolecularFormula/XML

   The esearch -debug option does show the base e-utilities URL and search terms parameters, but not all together where you could, for example, paste the entire link into a web browser.

```
esearch -db pccompound -query "NJTXJDYZPQNTSM-WMZOPIPTSA-N"[IKEY] -debug

nquire -url https://eutils.ncbi.nlm.nih.gov/entrez/eutils/ esearch.fcgi -retmax 0
-usehistory y -db pccompound -term NJTXJDYZPQNTSM-WMZOPIPTSA-N\[IKEY\]
```

2. Working only with unix utilities can be limiting, but it is straightforward to incorporate shell commands into computational notebook workflows.

Conclusions

- EDirect can be useful for retrieving small molecule and related information. You can do a lot with minimal code.
- Evaluation of EDirect and initial teaching experience suggests it can be a good choice for introducing new users to programmatic chemical information searching and compilation.
- There is a repository available at https://github.com/vfscalfani/EDirectChemInfo which includes 50+ PubChem/PubMed EDirect code “recipes” presented in a tutorial style. There are also workshop slides available, with more coming soon.... Contributions and feedback welcome.
- Future work will include incorporating substructure searches with PubChem PUG-REST into EDirect scripts using unix programs (e.g., with cURL [1])

[1] Kim, S.; Thiessen, P. A.; Bolton, E. E. Programmatic Retrieval of Small Molecule Information from PubChem Using PUG-REST, Methods in Pharmacology and Toxicology, 2018.

Acknowledgments and Notes

Special thanks to:

1. Kans J. Entrez Direct: E-utilities on the Unix Command Line. 2013 Apr 23 [Updated 2021 Apr 15]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK179288/

2. And all other NCBI/NLM/NIH staff for other tutorial content and answering my questions.

I'm not affiliated with NCBI/NLM/NIH, but I'm happy to help you with EDirect PubChem code recipes:

Vincent F. Scalfani
Science and Engineering Librarian
The University of Alabama
vfscalfani@ua.edu

Reminder: before using EDirect, you should definitely read the EDirect Manual (above) and the NCBI Website and Data Usage Policies and Disclaimers: https://www.ncbi.nlm.nih.gov/home/about/policies/