NLP4Chemistry

Introduction and Basic Concepts

Camilo Thorne, Saber Akhondi

May 2024 - COLING/LREC '24

Tutorial Overview

Structure, goals and outline

Structure: This tutorial is broken into 4 blocks of 50 mins. Each block consists of a =< 25 mins "lecture" (talk/slides), followed by a >= 25 mins "lab" (practice with notebooks)

Goal: To cover recent advances in NLP for chemistry, including language models for molecular modeling

Outline:

1. Basic concepts
2. Text mining
3. Language models
4. Exploring the fringe

Authors

Camilo Thorne (Elsevier)

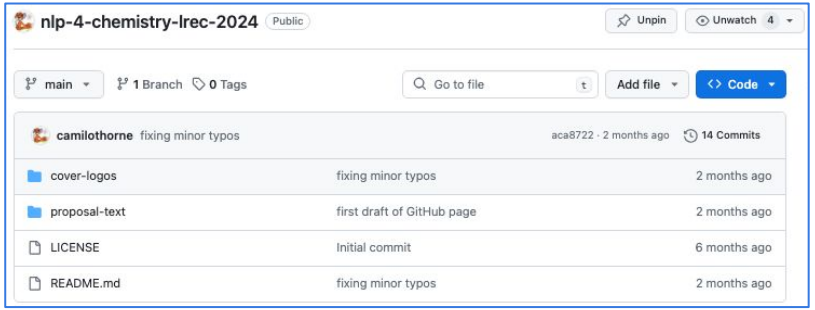
https://tinyurl.com/yc73zahe (LinkedIn)

Saber Akhondi (Elsevier)

https://tinyurl.com/wazcbp68 (LinkedIn)

Materials and communication

- Most materials shared on GitHub (slides, notebooks)
- Model checkpoints and data samples via GoogleDrive
- All materials free for research (non-commercial open source)
- Literature pointers and references shared thru open access URLs
- Slack and/or email for communication



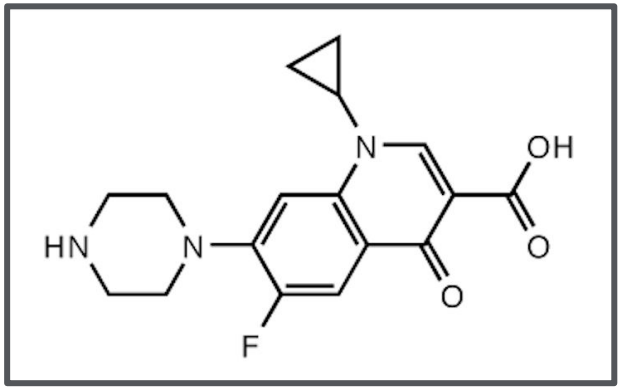
https://github.com/camilothorne/nlp-4-chemistry-lrec-2024
(GitHub)
https://drive.google.com/drive/folders/1LuyMJiL3cfxuYi2KNyBq0FpJ-kpKA9Jg?usp=sharing
(Google Drive)
#nlp-4-chemistry-lrec-coling-2024
(Slack)
c.thorne.1@elsevier.com
(Email)

Basic Concepts

Chemical compound

- A chemical compund is either an atom or a molecule
- A **molecule** is a structured collection of atoms, connected by bonds

- Visually, it can be understood as a graph (not necessarily planar) where:
  - **Vertexes** are **atoms** (of various kinds)
  - **Edges** are **bonds** (of various kinds)



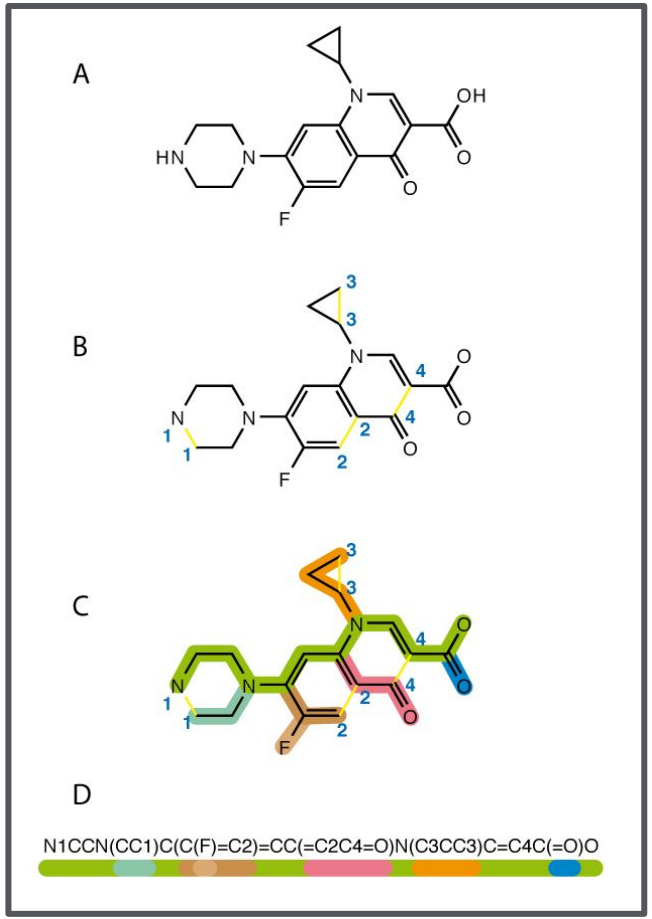
- **Molecular graphs** (MOLs) are the core representation of a molecule in chemistry

Chemical representation formats - SMILES

- **SMILES** stands for:
  "simplified molecular input line entry specification"

- It is a linear molecular representation
- It is obtained as follows:
  1. Compute a spanning-tree of the molecular graph
  2. Choose a random root and
  3. Topologically order the constituent atoms
- The ensuing representation is not-unique

- **Example**: generating SMILES from the molecular graph representation of `ciprofloxacin`



Chemical representation formats – other formats

- **Trivial names**: English names such as `ciprofloxacin`
- **IUPAC names:** names defined by the International Union of Pure and Applied Chemistry) such as `1-Cyclopropyl-6-fluor-4-oxo-7-(piperazin-1-yl)-1,4-dihydrochinolin-3-carbonoxide`
- **Formulas**: expressions referring to the number of constituent atoms, such as $C_{17}H_{18}FN_3O_3$
- **InChI**s: international chemical identifiers is an alternative to SMILES from the InChI Trust, such as `InChI=1S/C17H18FN3O3/c18-13-7-11-14(8-15(13)20-5-3-19-4-6-20)21(10-1-2-10)9-12(16(11)22)17(23)24/h7-10,19H,1-6H2,(H,23,24)`

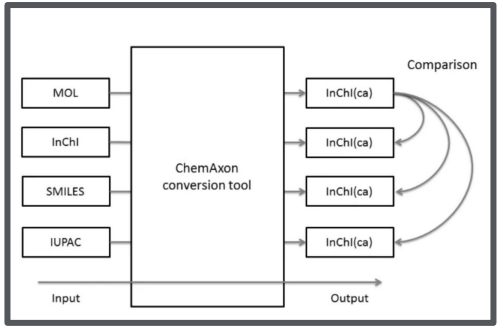
Representations can be ambigous [Akhondi, 2015]

Database	MOL	InChI	SMILES	IUPAC
DrugBank	6506	6391	6504	6489
ChEBI	21367	19076	19725	18798
HMDB	8534	8534	8534	7727
PubChem	5069294	5069293	5069294	4769031
NPC	8024	0	8018	0

Database	MOL	InChI	SMILES	IUPAC
DrugBank	98.9	100	99.1	93.6
ChEBI	90.6	100	96.8	69.8
HMDB	100	99.9	100	38.1
PubChem	100	100	100	92.6
NPC	99.7	-	100	-

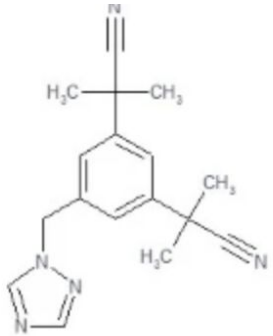
Database	MOL–InChI	MOL–SMILES	MOL–IUPAC
DrugBank	98.2	98.5	90.0
ChEBI	96.5	96.5	75.3
HMDB	89.3	37.2	55.7
PubChem	97.7	97.8	87.2
NPC	-	93.4	-

Representations can be ambigous [Akhondi, 2015]



Normalization across formats doesn't fully eliminate ambiguity!

Anastrozole



SMILES

CC(C)(C#N)c1cc(cc(c1)C(C)(C)C#N)Cn2cncn2
CC(C)(C#N)c1cc(Cn2cncn2)cc(c1)C(C)(C)C#N
CC(C)(C#N)c(cc(cc1C[n]([n]c[n]2)c2)C(C)(C)C#N)c1

IUPAC

2-[3-(1-cyano-1-methyl-ethyl)-5-(1,2,4-triazol-1-ylmethyl)phenyl]-2-methyl-propanenitrile
2,2'-[5-(1H-1,2,4-triazol-1-ylmethyl)benzene-1,3-diyl]bis(2-methylpropanenitrile)
2-[3-(1-cyano-1-methylethyl)-5-(1H-1,2,4-triazol-1-ylmethyl)phenyl]-2-methylpropanenitrile

InChI

InChI=1S/C17H19N5/c1-16(2,9-18)14-5-13(8-22-12-20-11-21-22)6-15(7-14)17(3,4)10-

Representations can be ambigous [Akhondi, 2015]

Database	Compounds	Identifiers	Identifiers/compound
PubChem	4,232,875	15,211,133	3.6
ChemSpider	6,646,902	10,063,709	1.5
ChemSpider-V	654,052	850,601	1.3
HMDB	37,761	308,733	8.2
NPC	14,814	131,290	8.9
TTD	2977	105,407	35.4
ChEBI	15,633	41,956	2.7
ChEMBL	21,398	28,011	1.3
DrugBank	3769	26,780	7.1

Database	Unique identifiers	Ambiguous identifiers	Ambiguity (%)	Compounds/ambiguous identifier
HMDB	173,455	26,430	15.2	6.1
TTD	100,570	4607	4.6	2.1
ChEMBL	26,910	1050	3.9	2.1
NPC	112,717	3455	3.1	2.1
ChemSpider	9,691,277	245,541	2.5	2.5
ChEBI	41,023	827	2.0	2.1
PubChem	14,937,728	201,621	1.3	2.4
ChemSpider-V	842,128	5401	0.6	2.3
DrugBank	26,759	20	0.1	2.1

Computational chemistry tools

- Nowadays **very easy** thanks to Python and...
- RDKit: https://www.rdkit.org/
  - Open source
  - Visualizes molecules
  - Converts between representations
  - Analyzes structural properties (e.g. weight)

RDKit

Open-Source Cheminformatics and Machine Learning

- Alternatives and commercial tools:
  - ChemAxon (C/C++) https://chemaxon.com/
  - OpenBabel (Python/C++) https://openbabel.org/
  - CDK (Java) https://cdk.github.io/

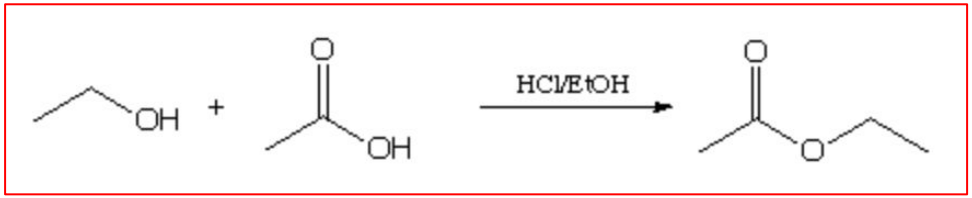
More in the lab!

Reactions and reaction SMILES

- Process transforming one or more reactants into one or more products, involving zero or more reagents
- Reaction SMILES represents reaction equations w. SMILES
- The equation needs to be well-balanced (equal number of atoms between products and reactants)

```
Reaction-SMILES ::= SMILES+ > SMILES* > SMILES+ ;
```

- A dot '.' is used to separate between SMILES units
- '>>' indicates the sense of the eq uation




```
CC(=O)O.OCC > [H+].[Cl-].OCC > CC(=O)OCC
```

The Chemical information Sources

uspto.GOV
The United States Patent and Trademark Office
an agency of the Department of Commerce

WIPO Encouraging Creativity and Innovation
WORLD INTELLECTUAL PROPERTY ORGANIZATION

Google patents

CHEMICAL REVIEWS

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Inorganic Chemistry

Journal of Medicinal Chemistry

PubChem

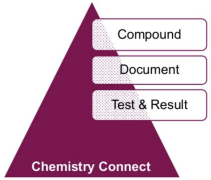
ChemSpider
The free chemical database

Reaxys®

SureChEMBL

CAS®
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

Chemistry Connect



Lab Notes

Patents **Journals** **Databases** **More**

Patents – key source! [Akhondi, 2018]

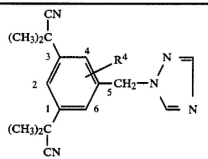
- Multiple authorities
- Different Languages
- Variety of input sources and content:
  - PDF, OCR PDF, Image PDF, XML
- Legal documents
- Obfuscations to hide key inventions
  - Deliberate spelling mistakes
  - Introduced noise
- Complexity of Chemical structures

4,935,437

19

EXAMPLES 49–52

The process described in Example, 1 was repeated, using the appropriate 2- or 4-substituted 2,2-(5-methyl-1,3-phenylene)di(2-methylpropiononitrile) as starting material, to give the following compounds:



Ex	R^4	Position of substitution	Mp.	Footnote
49	NO_2	4	—	1,2
50	Br	4	83–86	3
51	Br	2	128–131	3
52	CN	4	35–37	4

20

EXAMPLE 53

A mixture of 2,2′-(5-chlorodideuteriomethyl-1,3-phenylene)-di(2-trideuteriomethyl-3,3,3-trideuteriopropiononitrile) (0.65 g), dimethylformamide (5 ml) and sodium triazole (0.45 g) was stirred at room temperature for 18 h. The mixture was diluted with water (30 ml) and extracted with ethyl acetate, and the extract was dried and evaporated to dryness under reduced pressure. The residue was purified by flash chromatography, using ethyl acetate as eluant, to give 2,2′-[5-dideuterio-(1H-1,2,4-triazol-1-yl)methyl-1,3-phenylene]-di(2-trideuteriomethyl-3,3,3-trideuteriopropiononitrile), mp 82°–83° after crystallisation from ethyl acetate/cyclohexane.

The starting material from the above process may be prepared as follows:

The process used to prepare methyl 3,5-bis(1-cyano-1-methylethyl)benzoate, described in the later part of Example 8, was repeated, using trideuterioiodomethane instead of iodomethane, to give methyl 3,5-bis[1-cyano-2,2,2-trideuterio-1-(trideuteriomethyl)ethyl]-benzoate, m.p. 83°–84°.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 July 2003 (03.07.2003)

PCT

(10) International Publication Number
WO 03/053438 A1

(51) International Patent Classification[7]: A61K 31/4196, A61P 35/00

(21) International Application Number: PCT/GB02/05554

(22) International Filing Date: 6 December 2002 (06.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0129457.8 10 December 2001 (10.12.2001) GB

(71) Applicant *(for all designated States except MG, US)*: ASTRAZENECA AB [SE/SE]; Sodertalje, S-151 85 (SE).

(71) Applicant *(for MG only)*: ASTRAZENECA UK LIMITED [GB/GB]; 15 Stanhope Gate, London, Greater London W1K 1LN (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: PLOURDE, Paul [US/US]; 1800 Concord Pike, P.O. Box 15437, Wilmington, DE 19850 (US). ROGERS, Anthony [US/US]; 1800 Concord Pike, P.O. Box 15437, Wilmington, DE 19850 (US). VOSE, Brent [GB/GB]; Alderley Park, Macclesfield, Cheshire SK10 4TG (GB). DUKES, Michael [GB/GB]; Alderley Park, Macclesfield, Cheshire SK10 4TG (GB).

(74) Agent: ASTRAZENECA; Global Intellectual Property, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

A1

References

- 1988 - SMILES, a chemical language and information system.
  1. Introduction to methodology and encoding rules - https://people.cs.vt.edu/dbhattacharya/courses/cs6824/SMILES.pdf

- 2009
  - Chemoinformatics—an introduction for computer scientists - https://dl.acm.org/doi/abs/10.1145/1459352.1459353

- 2011 - Chemical Name to Structure: OPSIN, an Open Source Solution
  - https://pubs.acs.org/doi/10.1021/ci100384d

- 2011 - Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications - https://pubmed.ncbi.nlm.nih.gov/27467152/

- 2017 - Computer Representation of Chemical Compounds
  - https://link.springer.com/referenceworkentry/10.1007/978-3-319-27282-5_50

Thank you!