Chasing natural products: the COllection of Open Natural Products COCONUT

Maria Sorokina, Christoph Steinbeck

Friedrich-Schiller University Jena, Germany

https://cheminf.uni-jena.de

ChemBioSys

Cheminformatics and Computational Metabolomics
Friedrich-Schiller-University, Jena, Germany

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

About me

Cheminformatics and Computational Metabolomics
Friedrich-Schiller-University, Jena, Germany

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Steinbeck Lab: https://cheminf.uni-jena.de

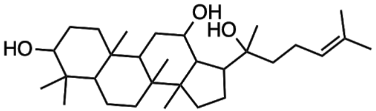
Chem- and bioinformatician

Senior postdoctoral researcher at the Friedrich-Schiller University, in Jena, Germany:

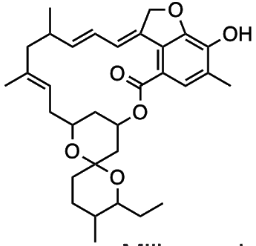
- Natural Products cheminformatics (databases)
- Research Data Management for the ChemBioSys CRC
- Omics for marine diatoms

Natural Products

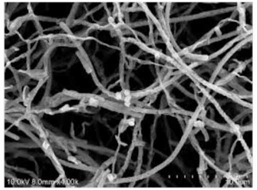
* made by nature
* bioactive
* complex data

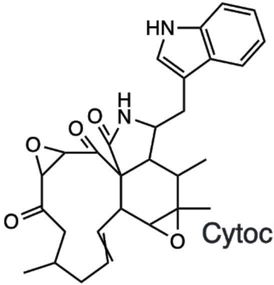


Protopanaxadiol

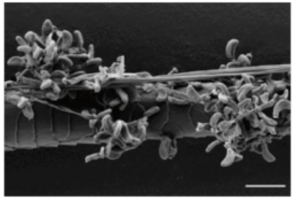


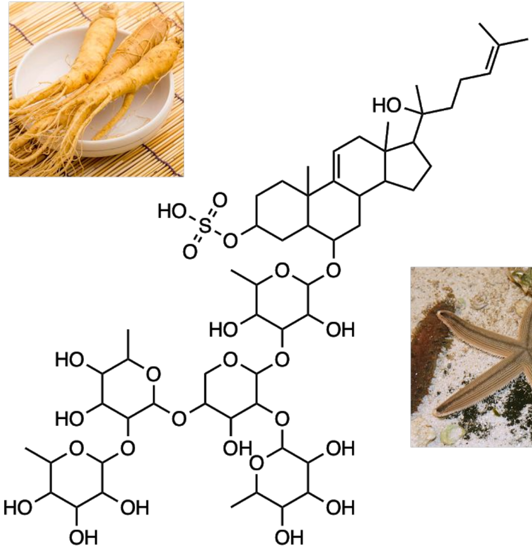
Milbemycin A31



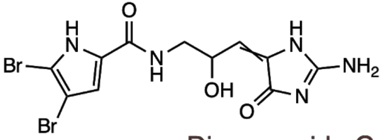


Cytochalasin Z

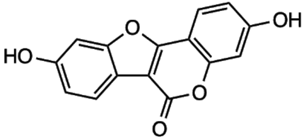




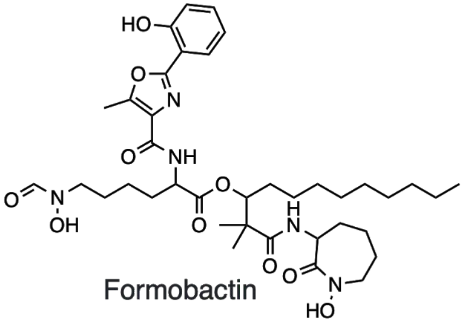
Solasteroside A



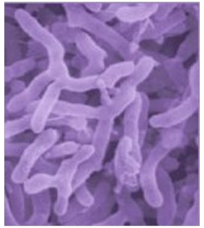
Dispacamide C



Cumostrol



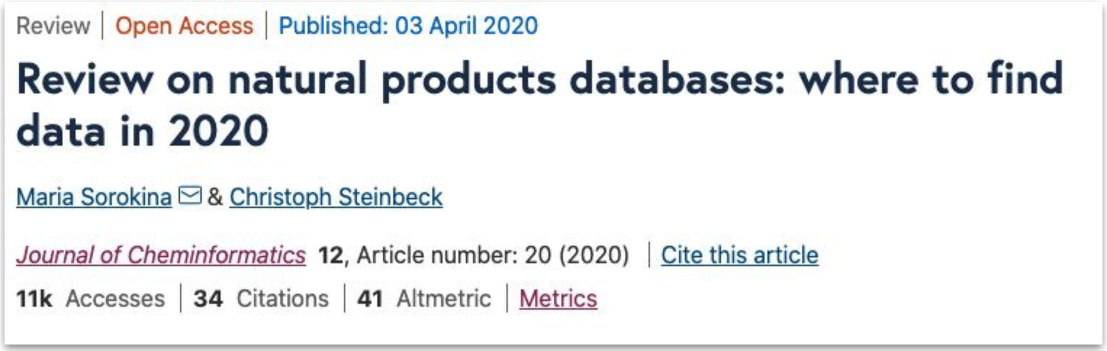
Formobactin



Natural products research: a field (re)gaining in popularity

- Between 2000 and 2020 123 NP databases/datasets were mentioned in the literature
- 90 are open, 50 are downloadable
- Extremely heterogeneous data

https://npreview.naturalproducts.net/



Review | Open Access | Published: 03 April 2020

Review on natural products databases: where to find data in 2020

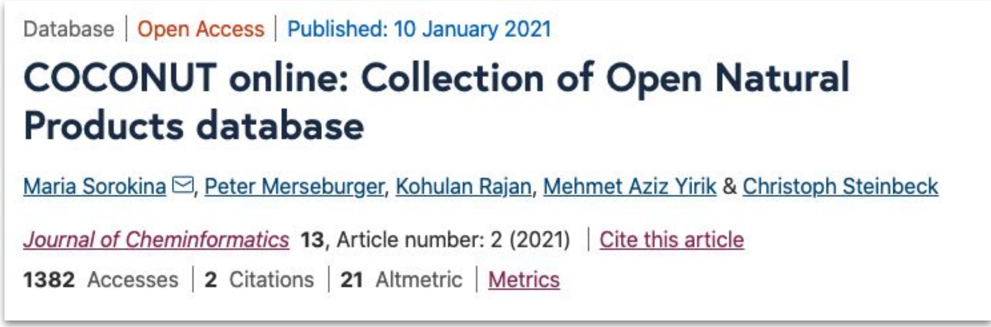
Maria Sorokina ✉ & Christoph Steinbeck

Journal of Cheminformatics **12**, Article number: 20 (2020) | Cite this article

11k Accesses | **34** Citations | **41** Altmetric | Metrics

… so we decided to build yet another NP database

- Which gathers in the same place NP data from 53 (now 55) public databases
- Chemical structure-centred
- Following the **FAIR principles**
- Current version contains 406 ,744 unique “flat” molecules



Database | Open Access | Published: 10 January 2021

COCONUT online: Collection of Open Natural Products database

Maria Sorokina, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik & Christoph Steinbeck

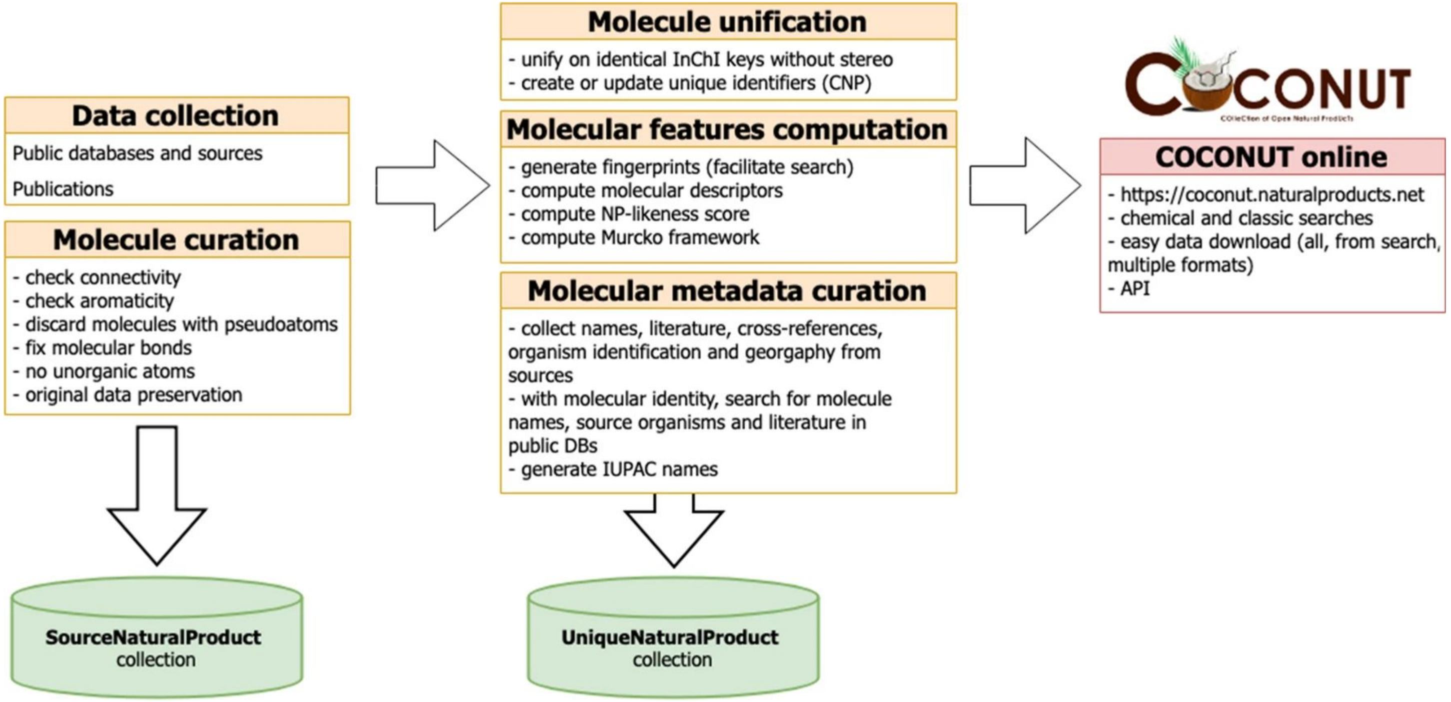
Journal of Cheminformatics **13**, Article number: 2 (2021) | Cite this article

1382 Accesses | **2** Citations | **21** Altmetric | Metrics

https://coconut.naturalproducts.net/

COCONUT data model

mongoDB®



Data collection

Public databases and sources

Publications

Molecule curation

- check connectivity
- check aromaticity
- discard molecules with pseudoatoms
- fix molecular bonds
- no unorganic atoms
- original data preservation

SourceNaturalProduct collection

Molecule unification

- unify on identical InChI keys without stereo
- create or update unique identifiers (CNP)

Molecular features computation

- generate fingerprints (facilitate search)
- compute molecular descriptors
- compute NP-likeness score
- compute Murcko framework

Molecular metadata curation

- collect names, literature, cross-references, organism identification and georgaphy from sources
- with molecular identity, search for molecule names, source organisms and literature in public DBs
- generate IUPAC names

UniqueNaturalProduct collection

COCONUT

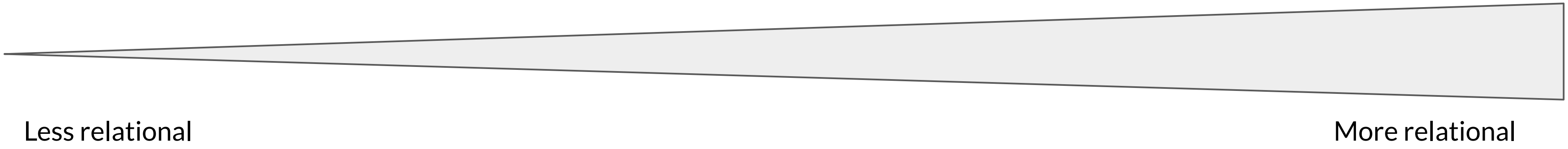
COCONUT online

- https://coconut.naturalproducts.net
- chemical and classic searches
- easy data download (all, from search, multiple formats)
- API

Overview of modern database management systems

SQL: Structured Query Language

noSQL: “not only SQL” rather than “not SQL”



Key-value DBs
Redis,
Voldemort,
Dynamo

Column-oriented DBs
(db-dependent, e.g. CQL)
- CassandraDB, Google’s Big Table

Document DBs
(db-dependent, e.g. mongo query language)
- MongoDB, CouchDB

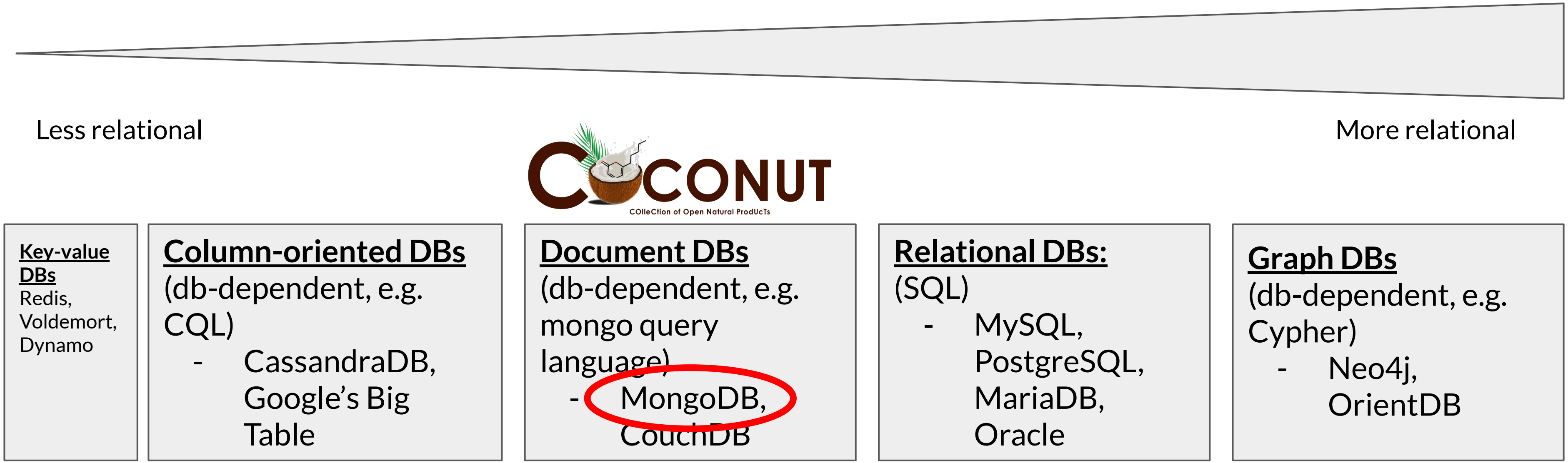
Relational DBs:
(SQL)
- MySQL, PostgreSQL, MariaDB, Oracle

Graph DBs
(db-dependent, e.g. Cypher)
- Neo4j, OrientDB

Overview of modern database management systems

SQL: Structured Query Language

noSQL: “not only SQL” rather than “not SQL”



Less relational

More relational

COCONUT

COllECtion of Open Natural ProdUcTs

Key-value DBs
Redis,
Voldemort,
Dynamo

Column-oriented DBs
(db-dependent, e.g. CQL)
- CassandraDB, Google’s Big Table

Document DBs
(db-dependent, e.g. mongo query language)
- MongoDB, CouchDB

Relational DBs:
(SQL)
- MySQL, PostgreSQL, MariaDB, Oracle

Graph DBs
(db-dependent, e.g. Cypher)
- Neo4j, OrientDB

COCONUT data model - search chemistry

Delegating the search to MongoDB to speed up

- ➔ Structure search: SMILES/InChi identity

COCONUT data model - search chemistry

Delegating the search to MongoDB to speed up

- ➔ Structure search: SMILES/InChi identity
- ➔ Substructure search: query PubChem fingerprints “ON bits” search ($bitsAllSet)

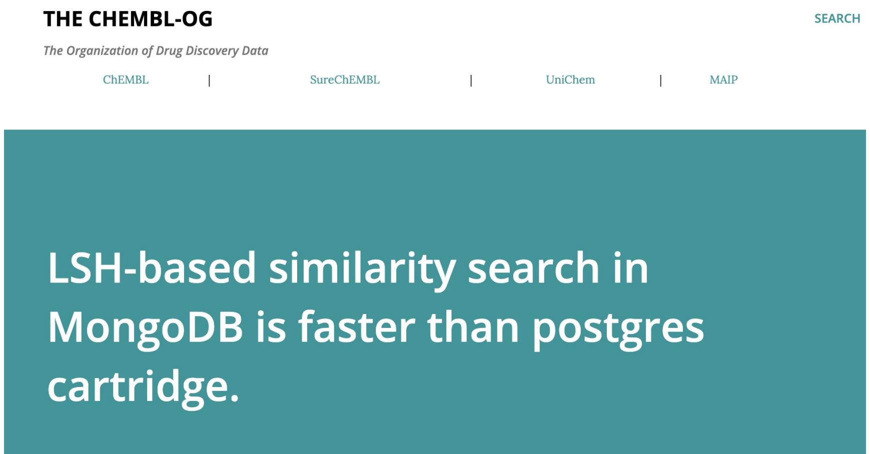
```
{ "_id" : CNP000XX1, "PubChemFP" : "00110110" }
{ "_id" : CNP000XX2, "PubChemFP" : "10110100" }
{ "_id" : CNP000XX3, "PubChemFP" : "01110111" }

db.uniqueNaturalProduct.find( { PubChemFP: { $bitsAllSet: [ 1, 5 ] } } )
> CNP000XX3
```

COCONUT data model - search chemistry

Delegating the search to MongoDB to speed up

- ➔ Structure search: SMILES/InChi identity
- ➔ Substructure search: query PubChem fingerprints “ON bits” search ($bitsAllSet)
- ➔ Similarity search: PubChem fingerprints + inverted indexes + Tanimoto on MongoDB server side



THE CHEMBL-OG

The Organization of Drug Discovery Data

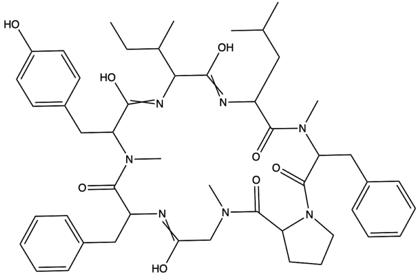
LSH-based similarity search in MongoDB is faster than postgres cartridge.

http://chembl.blogspot.com/2015/08/lsh-based-similarity-search-in-mongodb.html

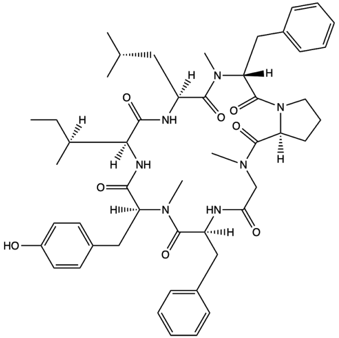
COCONUT data model - the stereochemistry issue

Cordyheptapeptide A

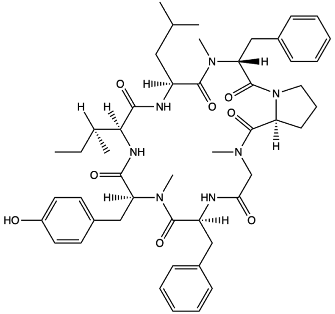
+ no stereo in UNPD



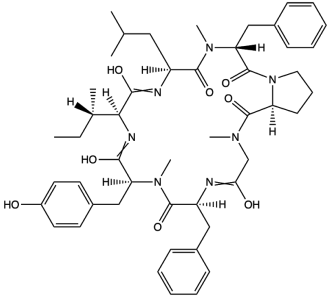
CNP0267851



Supernatural2



NP Atlas & ChEBI



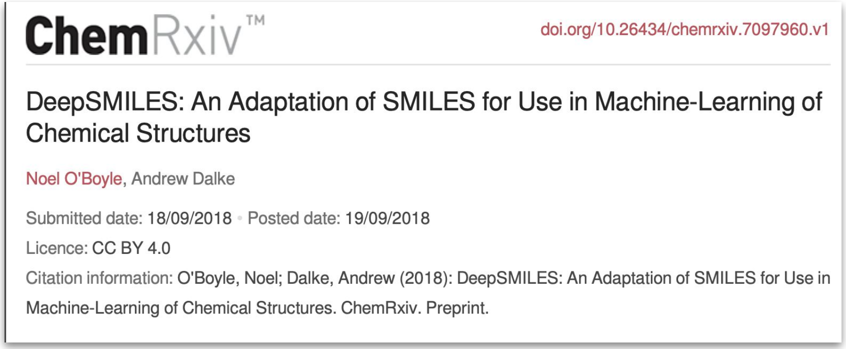
NPASS

COCONUT data model - representations

Classic SMILES, InChI, InChI keys, names and synonyms

Murcko frameworks

Deep SMILES (more suitable for deep & machine learning)



ChemRxiv™ doi.org/10.26434/chemrxiv.7097960.v1

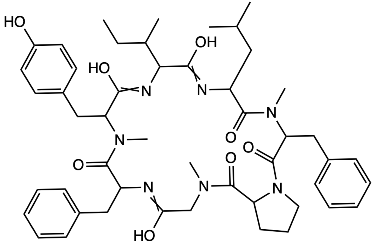
DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures

Noel O'Boyle, Andrew Dalke

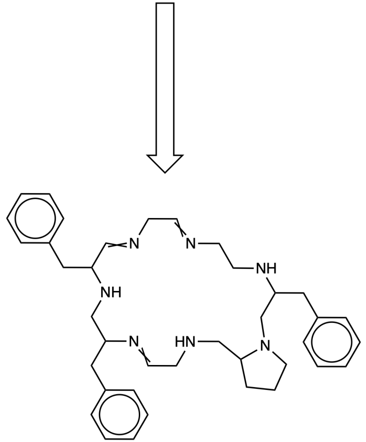
Submitted date: 18/09/2018 • Posted date: 19/09/2018

Licence: CC BY 4.0

Citation information: O'Boyle, Noel; Dalke, Andrew (2018): DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. ChemRxiv. Preprint.



Cordyheptapeptide A (CNP0267851)



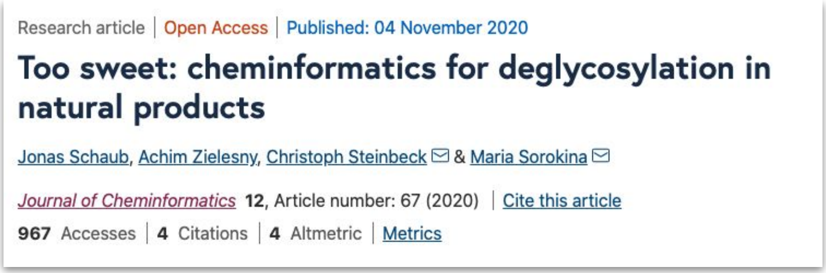
Murcko framework of Cordyheptapeptide A

COCONUT data model - glycosidic moieties

Glycosidic moieties are generally considered as redundant, monotonous substructures that prevent efficient NP structure study

BUT! They actually can greatly quantitatively and qualitatively influence the bioactivity

➔ The glycosylation status of NPs therefore described in COCONUT



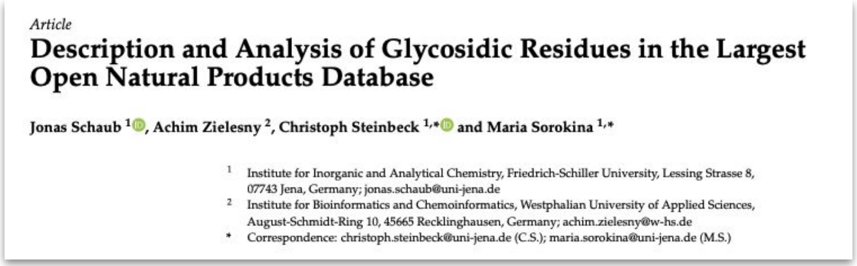
Research article | Open Access | Published: 04 November 2020

Too sweet: cheminformatics for deglycosylation in natural products

Jonas Schaub, Achim Zielesny, Christoph Steinbeck & Maria Sorokina

Journal of Cheminformatics **12**, Article number: 67 (2020) | Cite this article

967 Accesses | **4** Citations | **4** Altmetric | Metrics



Article

Description and Analysis of Glycosidic Residues in the Largest Open Natural Products Database

Jonas Schaub [1], Achim Zielesny [2], Christoph Steinbeck [1,*] and Maria Sorokina [1,*]

[1] Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Lessing Strasse 8, 07743 Jena, Germany; jonas.schaub@uni-jena.de

[2] Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, August-Schmidt-Ring 10, 45665 Recklinghausen, Germany; achim.zielesny@w-hs.de

[*] Correspondence: christoph.steinbeck@uni-jena.de (C.S.); maria.sorokina@uni-jena.de (M.S.)

COCONUT data model - physicochemical properties

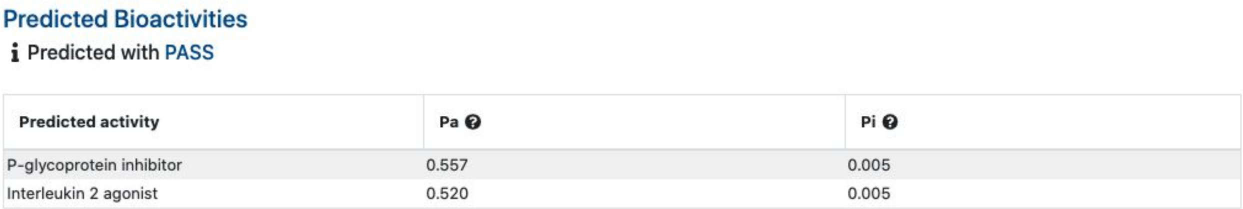
>30 molecular descriptors were calculated for each NP

AlogP, Lipinski Rule of 5 failures, circular fragments, apol, bpol, FMF, fsp3, Kappa Shape Index, Petitjean number, Zagreb index….

https://cdk.github.io/

COCONUT data model - PASS predictions

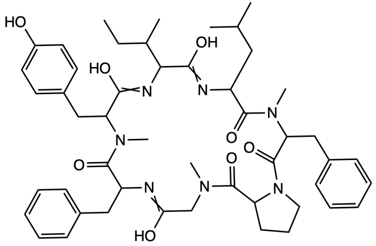
Search by predicted PASS bioactivity is also enabled



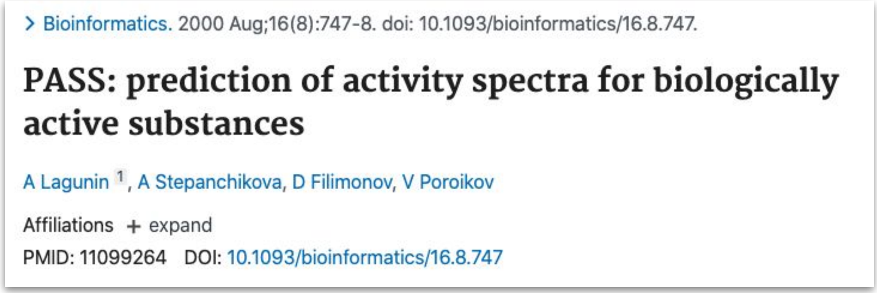
Predicted Bioactivities

ℹ Predicted with PASS

Predicted activity	Pa ❓	Pi ❓
P-glycoprotein inhibitor	0.557	0.005
Interleukin 2 agonist	0.520	0.005



Cordyheptapeptide A (CNP0267851)



› Bioinformatics. 2000 Aug;16(8):747-8. doi: 10.1093/bioinformatics/16.8.747.

PASS: prediction of activity spectra for biologically active substances

A Lagunin [1], A Stepanchikova, D Filimonov, V Poroikov

Affiliations + expand

PMID: 11099264 DOI: 10.1093/bioinformatics/16.8.747

COCONUT data model - annotations

- Taxonomic provenance (~15 %)
- Geographic provenance of the producer organism (~10%)
- Chemical ontology: ClassyFire (NPclassifier classifications will be added)
- Cross-references (can be challenging due to URL organization in the target DB)

Current and future developments

- LOTUS (with J-L. Wolfender & P-M. Allard, Univ. Geneva): lotus.naturalproducts.net/
  - ➔ improvement of COCONUT annotations

- ML-based taxonomic annotations prediction
- Implement user-driven NP submission (this summer)
- **Elaboration of minimal information standards for NP declaration**
- Predicted C13 NMR shifts representations (with J-M. Nuzillard, Univ. Reims)
- Predicted MS spectra representations (with P-M. Allard, Univ. Geneva)
- **And especially: stabilize the server**

Acknowledgements

Cheminformatics and Computational Metabolomics
Friedrich-Schiller-University, Jena, Germany

Chris Steinbeck and the wonderful Caffeine group (cheminf.uni-jena.de)

ChemBioSys CRC

ChemBioSys

COLLABORATIVE RESEARCH CENTER 1127
CHEMICAL MEDIATORS IN COMPLEX BIOSYSTEMS

My projects: https://naturalproducts.net/