NCI CADD Group cheminformatics resources: From web tools to billion-molecule databases

Marc C. Nicklaus

Computer-Aided Drug Design (CADD) Group

Chemical Biology Laboratory
Center for Cancer Research
National Cancer Institute
National Institutes of Health
Frederick, Maryland 21702
Mission of the NCI CADD Group

The CADD Group uses computational methods to

- help develop novel therapeutics against cancer and viral diseases
- investigate issues that impede the drug development process
- move chemoinformatics forward
- make resources publicly available
First public Web GUI for large small-molecule database with advanced capabilities such as full substructure search.

First alpha on server in Erlangen October 1997
**Enhanced NCI Database Browser 2.2**

Database status: 250,250 open structures ready for searching. Bug reports, comments or questions?

<table>
<thead>
<tr>
<th>Query Type</th>
<th>Negate</th>
<th>Query Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSC Number(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAS Number(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formula...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecular Weight Range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact Structure...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Start Search | Reset

Correct query flexibly:
- Max. number of hits and search time: 100 hits, 90 seconds
- Output Format: HTML Table with Samples, preferably 3D
- Output Sort: NSC Number

Web service for NCI/DTP’s Open NCI Database

- 250,250 structure records
- Complex search and output options available
- ~60 million data points: mostly PASS\(^{(1)}\) predictions
- Major update scheduled

---

https://cactus.nci.nih.gov/  Small-molecule centric!

General Information

The NCI/CADD group is a research unit within the Chemical Biology Laboratory at the National Cancer Institute. Read more about the CADD Group's Chemoinformatics Tools and User Services. Or, visit our blog.

Online Services

To search and display chemical structures here, you will need Java/JavaScript to be enabled on your browser.

Accessibility (Section 508): We want to hear from users with disabilities, especially visually impaired users, where they have experienced particular difficulties in using our site. If you are a visually impaired user, please e-mail M.C. Nicklaus and team with your experiences and suggestions.

Chemical Activity Predictor
This service is the first one of our new Apps. It provides the prediction of a (growing) number of small molecule properties calculated by QSAR models created with the GUSAR software. Beta version. --- Due to ever stricter security mandates, this service unfortunately had to be turned off. As an alternative, you can try out the NIH/NCATS Predictor service, which contains many more models than the Chemical Activity Predictor.

Chemical Identifier Resolver
This service works as a resolver for different chemical structure identifiers and allows the conversion of a given structure identifier into another representation or structure identifier. It can be used via a web form or a simple URL API.

Enhanced NCI Database Browser Release 2.2
A web service to the open NCI database compounds (>250,000 structures) with different kinds of output features and links to other services for continued processing (Version 2.2 is a 2013 technical overhaul).

Tautomerizer
Allows you to test a set of 86 different tautomeric transforms with your own molecules. In addition to the standard rules of the chemoinformatics toolkit CACTVS, there are 60+ additional rules compiled in the context of the IUPAC project of Redesign of the Handling of Tautomerism in InChI V2.
GIF Creator

GIF/PNG-Creator for 2D Plots of Chemical Structures

Structure Source
- **Smiles String**: C2CCC(C1C(CCCCC)C)C

Parameters
- **Image Format**: GIF
- **Image Width**: 165 (Pixels)
- **Atom Color**: Black
- **Hydrogen Color**: Default
- **Bond Color**: Black
- **Border Width**: 2 (Pixels)
- **Stereochemistry**: Wedges
- **Plot Alignment**: None
- **Request New 2D Coordinates**: No
- **Header**: Black
- **Footer**: Black
- **Comment**: None
- **Embedded Structure Code**: None

Options:
- **Editor**:
- **Interface**: On
- **Image Height**: 165 (Pixels)
- **Atom Symbol**: Expanded Symbol
- **Carbon Symbols**: None
- **Hydrogen Symbols**: None
- **Background**: White
- **Horizontal Bonds**: Enabled with border of 2 (Pixels)
- **Coordinate Style**: Normal
- **Image Map Style**: None

Buttons:
- **Reset**
- **Submit**
### Online SMILES Translator and Structure File Generator

#### Input Format

<table>
<thead>
<tr>
<th>Input Format</th>
<th>Unique SMILES Output Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>C12C3C4C1C5C4C3C25</td>
<td>- Display on screen</td>
</tr>
<tr>
<td></td>
<td>- SMILES TXT file</td>
</tr>
<tr>
<td></td>
<td>- SDF</td>
</tr>
<tr>
<td></td>
<td>- PDB</td>
</tr>
<tr>
<td></td>
<td>- MOL (only single structure generated)</td>
</tr>
</tbody>
</table>

Start Structure Editor

Please choose this field if you want to submit your own SMILES strings or create a SMILES string using the Structure Editor. A submitted file has precedence, so delete any entry below if you want to submit a new SMILES string.

Choose File

Please choose this field if you want to translate your own files. The service will automatically recognize SD files (single and multiple structure), text files with multiple SMILES fields, MOL files and PDB files (and in fact any other format CACTVS recognizes).

No file chosen

If the input file contains a single structure, the output will also be single structure. Multiple structure input formats will generate multiple structure output for those formats that support this. Otherwise, only the first structure will be used. SD files will contain a UNIQUE_SMILES field for unique SMILES and an USER_SUPPLIED_SMILES field for the user-supplied SMILES (if available).

---

Read about our new web services at /blog

Bug reports, comments or questions?
Last Update: 2017-10-16
Optical Structure Recognition Application (OSRA)

OSRA
https://cactus.nci.nih.gov/osra/

Converts graphical representations of chemical structures in journal articles, patent documents, textbooks, trade magazines etc., into SMILES.

Available as both a downloadable executable and via a Web interface (shown here).

Newest versions of OSRA available at https://sourceforge.net/projects/osra/
Chemical Structure Lookup Service (CSLS)

https://cactus.nci.nih.gov/lookup

“Phone book for chemicals”

Where to find a certain chemical structure?

link(s) to the original database(s)
Chemical Identifier Resolver (CIR)

https://cactus.nci.nih.gov/chemical/structure

- "Resolves" structure identifiers or representations, i.e. converts one structure identifier/representation into another
- Has 450k accesses/day
Chemical Identifier Resolver (CIR)

programmatic URL API:
http://cactus.nci.nih.gov/chemical/structure/”identifier”/”representation”

...but also perfectly usable by humans:
http://cactus.nci.nih.gov/chemical/structure/aspirin/stdinchikey

returns...  InChIKey=BSYNRYMUTXBXSQ-UHFFFAOYSA-N

if a request is not successful: HTTP404 status message
CIR Workflow

**http request**

1. **identifier**
   - identifier is a full structure representation (e.g. SMILES, InChI)

2. **detection of the identifier type**
   - identifier is a hashed structure representation (e.g. InChIKey), or chemical name etc.

3. **database lookup**
   - in CADD Group’s Chemical Structure DataBase (CSDB)

4. **structure**

5. **calculation of the requested structure representation**
   - e.g. InChI, GIF image

6. **http response**
   - e.g. CAS number, chemical name
Tautomerism: Not Just an Academic Question

Tautomeric pairs (conflicts) – via NCI/CADD identifiers\(^1\)


Aldrich Market Select (AMS) database:
5,755,574 molecules (2012-09 version)

31,156 conflicts \rightarrow 62,872 molecules

<table>
<thead>
<tr>
<th>n-tuples</th>
<th>Conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>30,619</td>
</tr>
<tr>
<td>3</td>
<td>514</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Examples (prices per 1 g):

$300$ \hspace{1cm} $188$

Same original supplier!

$313$ \hspace{1cm} $350$

$251$ \hspace{1cm} $300$

Several 100s of cases experimentally confirmed

Guasch, L. \textit{et al.}

Tautomerism is Widespread

Tautomerism is not just interesting, important, and potentially costly - it is widespread:

- Analysis of a combined set of ~401 million molecules:
- Tautomerism possible for an average of 71% structures
- “Tautomerism-critical” cases frequent: chance of missing a match across databases due to tautomerism: nearly 1 in 10

Tautomer Structures Extracted from Experimental Literature: “Tautomer Database”

https://cactus.nci.nih.gov/download/tautomer/

Release 3 - November 2019

2,819 Tautomeric Tuples Comprising 5,977 Structures

Structurally different tuples: 1,776 (comprising 3,884 different structures) since some tuples are differentiated from each other only by experimental conditions such as solvent, spectroscopy method, etc.

See https://doi.org/10.1021/acs.jcim.9b01156 and https://doi.org/10.26434/chemrxiv.10790369.v1 for literature about this database.

# Tautomer Enumeration Tool

https://cactus.nci.nih.gov/tautomerizer/

## Tautomerizer - Predict tautomers based on 80+ rules

**Introduction | Form | Individual Rule Pages | Rules Sources | Help**

### Enter the structure in SMILES format

1. **Input Structure SMILES:**
   - **Submit**

2. **Single step or Multi step:**
   - Single step ☐ Multi step

3. **Activate rules:**
   - ☐ Activate all rules
   - ☐ Activate standard rules
   - ☐ Activate only new rules
   - ☐ Enter your own rule as SMIRKS:
   - ☐ Activate custom rule set via following checkboxes:

   **Select rules:**
   - ☐ PT_02_00 - 1,5 (thio)keto/(thio)enol
     - [O,S,Se,Te;X1:1]—[Cz1H0:2][C:5]—[C:6][C=4z0,NX3:3][#1:4]>[#1:4][O,S,Se,Te;X2:1][Cz1:2]=—[C:5][C:6]=—[Cz0,N:3]
   - Select example: C1=CC(C=C(C1=O)O)=O
   - Run Example

   - ☐ PT_03_00 - simple (aliphatic) imine
     - [#1,a,O:5][NX2:1]=—[Cz[1-2]:2][C=4R(0-2):3][#1:4]>[#1,a,O:5][NX3:1][#1:4][Cz:2]=—[C:3]
   - Select example: [C]1(CC[C][CC1]=—[N]
   - Run Example

   - ☐ PT_04_00 - special imine
     - [Cz0R0X3:1][([C:5]=—[C:2][Nz0:3][#1:4]>[#1:4][Cz0R0X4:1][([C:5]=—[c:2]=—[nz0:3]
   - Select example: C(CC1=NC=C[NH1])(C)C
   - Run Example
**Recent Tool: ReactionCode**

https://cactus.nci.nih.gov/reactioncode/

**ReactionCode: a format for reaction searching, analysis, classification, transform, and encoding/decoding**

**Description**

ReactionCode is a new versatile format for searching, analysis, classification, transform, and encoding/decoding of reactions. ReactionCode is a multi-layer machine readable code, which has been created in the line of ClassCode and BINCODE format. In contrast to these formats, it can fully encode and decode a reaction without any information loss. This open source format is canonical and designed to be flexible, upgradeable and versatile in order to be applied in a broad range of applications. ReactionCode is particularly useful for reaction similarity searching and classification, but is also intended for machine learning applications and as a new transform reaction language.

Further information can be found in the publication: ReactionCode: format for reaction searching, analysis, classification, transform, and encoding/decoding
https://doi.org/10.1186/s13321-020-00476-x
Downloadable Data Sets

Downloads

**PubChem Structure + Assay Download Page**
Download SD files of structures from PubChem with assay data included as properties, suitable for building QSAR or other types of models.

**NCI Database Download Page**
Download the "raw" data in bulk format that were used in building the Enhanced NCI Database Browser.

**FDA SPL Download Page**
Download a Mapping File to, or (older) SD file versions of, Structured Product Labeling (SPL) index files of substances indexed by FDA.

**HIV-1 Integrase Inhibitor Download Page**
Download structures and annotations of HIV-1 integrase inhibitors collected from literature.

**SAVI Products Download Page**
Download products and other associated data of our Synthetically Accessible Virtual Inventory (SAVI) project: Computational generation of a very large database of reliably and inexpensively synthetizable novel compounds with desirable properties for drug development.

**Tautomerism Database Download Page**
Download a spreadsheet with 5,977 structures extracted from experimental literature containing 2,819 cases of tautomeric tuples, annotated with experimental conditions, structure identifiers, bibliographic references, and preliminary analysis of the tautomerism involved in each case. (Release 3)

**Multi-Species Acute Toxicity Database Download Page**
Download a spreadsheet with toxicity measurements for 80,081 unique compounds (compounds that also went into the RTECS® database)

**iRL-Based Screening Sample Download Page**
Download an NCI/CADD Group curated set of over 140 million compounds based on iResearch™ Library (iRL) commercial screening samples (http://www.chemnavigator.com/irli.asp).

**Claimed Small-Molecule Structures in NCI/NIH Patents**
Download a database of about 12,700 structures extracted from patents filed by NCI/NIH, granted through September 2019.
SAVI (Synthetically Accessible Virtual Inventory)

Transforms

LHASA transforms written in CHMTRN/PATRAN

Building Blocks

Enamine Yurii Moroz

Chemoinformatics engine

CACTVS Wolf D. Ihlenfeldt

Product Generation

CADD Group
SAVI – Publicly Downloadable Dataset

SAVI, *in silico* generation of billions of easily synthesizable compounds through expert-system type rules

Hitesh Patel, Wolf-Dietrich Ihlenfeldt, Philip N. Judson, Yuri S. Moroz, Yuri Pevzner, Megan L. Peach, Victorien Delannée, Nadya I. Tarasova & Marc C. Nicklaus

<table>
<thead>
<tr>
<th>Class</th>
<th>SAVI products</th>
<th>Unique</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plus</td>
<td>1,094,782,440</td>
<td>976,051,945</td>
<td>62.61%</td>
</tr>
<tr>
<td>Neg0</td>
<td>609,262</td>
<td>579,532</td>
<td>0.03%</td>
</tr>
<tr>
<td>Neg10</td>
<td>54,775,204</td>
<td>48,036,148</td>
<td>3.13%</td>
</tr>
<tr>
<td>Neg20</td>
<td>82,180,372</td>
<td>80,366,188</td>
<td>4.7%</td>
</tr>
<tr>
<td>Neg30</td>
<td>516,116,725</td>
<td>457,508,945</td>
<td>29.52%</td>
</tr>
<tr>
<td>Total</td>
<td>1,748,464,003</td>
<td>1,526,316,392</td>
<td></td>
</tr>
</tbody>
</table>

Downloadable data set:
https://cactus.nci.nih.gov/download/savi_download/
Summary

- [https://cactus.nci.nih.gov](https://cactus.nci.nih.gov)
- Online for 24 years
- Several services are being renovated
- Interactive services: 8
- Toolkits and software: 4
- Downloadable datasets: 9
- Meeting presentations for CADD Group sponsored workshops: 4
- Total number of downloadable compounds: ~2 billion
- User accesses 2016-2021: ~1 billion; >450,000/day
Acknowledgements

CADD Group Members
• Victorien Delannée

Former CADD Group Members
• Devendra Dhaked
• Hitesh Patel
• Megan Peach
• Yuri Pevzner
• Laura Guasch
• Markus Sitzmann
• Igor Filippov
• Guangyu Sun
• Johannes Voigt
• Bruno Bienfait
• Frank Oellien

Contributors
• Wolf-Dietrich Ihlenfeldt
• Dan Zaharevitz
• Vladimir Poroikov
• Dmitrii Filimonov
• Peter Ertl
• Scott Hutton
• Bret Daniel
• Philip Judson
• Martin Ott
• Nadya Tarasova
• Yurii Moroz