Managing large atomic and molecular data sets: HITRAN, ExoMol and CascadesDB

IAEA
International Atomic Energy Agency
Atoms for Peace

Christian Hill
Atomic and Molecular Data Unit
Nuclear Data Section
IAEA

2018 Joint ICTP-IAEA School and Workshop on
Fundamental Methods for Atomic, Molecular and Materials Properties in Plasma Environments

ICTP
The Abdus Salam
International Centre for Theoretical Physics

Summary

1. **Principles** of database design
2. **HITRAN** and **HITRAN*online***: low-temperature, high-resolution spectroscopic database
3. **ExoMol**: high-temperature, high-resolution spectroscopic database
4. **QuantemolDB** and **ALADDIN**: collisional databases for plasma processes
5. **CascadesDB**: collisional cascade molecular dynamics simulation database
6. **Crowdsourcing**

Principles of database design

The FAIR Guiding Principles for scientific data management and stewardship

Findable

Accessible

Interoperable

Reusable

Wilkinson, M. D. et al., *Sci. Data* **3**:160018 doi: 10.1038/sdata.2016.18 (2016).

Principles of database design

Findable data

To be *findable* (meta)data must:

1. be assigned a globally-unique and persistent identifier (a URI such as a DOI)
2. registered in a searchable resource

URI = Uniform Resource Identifier

DOI = Digital Object Identifier

Principles of database design

Accessible data

1. To be *accessible* (meta)data must be retrievable from their identifier using a standardised communications protocol
2. the protocol (e.g. an API) must be open, free and universally implementable
3. the protocol may allow for authentication and authorisation.

Principles of database design

Interoperable data

To be *interoperable* (meta)data must:

1. represented in a formal, shared and broadly-applicable format
2. use vocabularies that follow FAIR principles
3. include qualified references to other (meta)data

Principles of database design

Interoperable data

To be *interoperable* (meta)data must:

1. represented in a formal, shared and broadly-applicable format
2. use vocabularies that follow FAIR principles
3. include qualified references to other (meta)data

Further things to consider

- **Physical Units**
- Phase conventions, reference / fiducial values
- Endianness (for binary data)
- Representation of null / missing / invalid data points

Principles of database design

Reusable data

To be *reusable* (meta)data must be:

1. richly described with accurate and relevant attributes
2. released with a clear data usage licence
3. associated with detailed provenance

Principles of database design

Authentication, Authorization, Accounting

An online database will usually implement a user-management system to:

1. Identify users (usernames, email addresses)
2. Authenticate users (login with password)
3. Account for users’ activity with the database (logs)

Principles of database design

Practical Considerations

An online database must have

1. A stable, highly-available host server(s)
2. Software for managing users (registration, login, logout, password reset)
3. Legal terms and conditions, licence, privacy policy
4. SSL

In addition, it may have:

1. A documented API for automated access by codes, etc.
2. Contact / feedback form
3. An interface for *uploading* data

HITRAN and HITRAN*online*

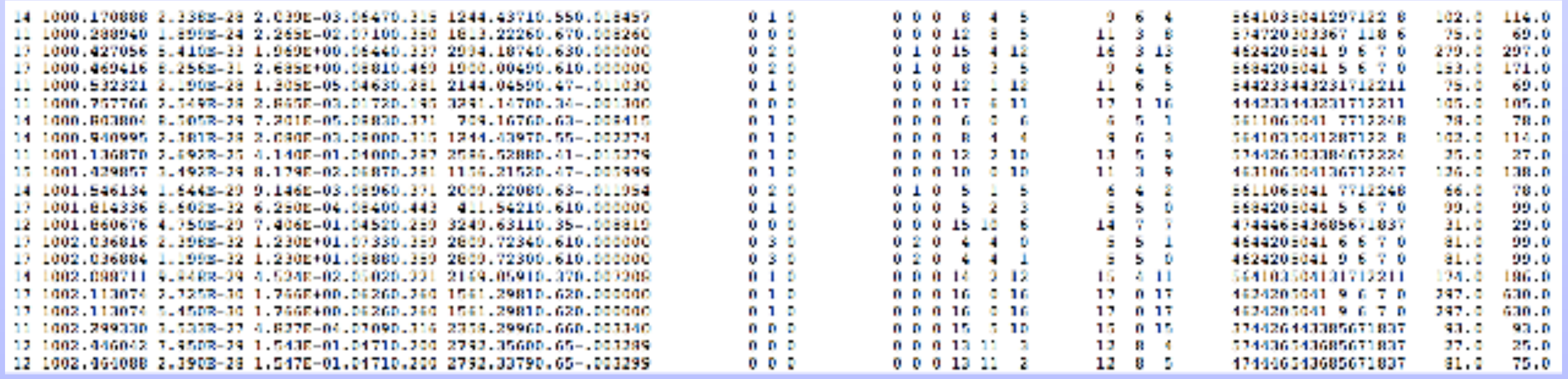
http://hitran.org

- Compilation of spectroscopic parameters for modelling radiative transfer in atmospheres
- Based at the Harvard-Smithsonian Centre for Astrophysics
- Mostly molecules, mostly at “low temperature”
- **5 x 10^6** lines; 600 absorption cross sections
- 365 molecules
- 9000 registered users

HITRAN and HITRAN*online*

1970s

2000s



```
14 1000.170888 2.338E-28 2.039E-03.05470.315 1244.43710.550.018457          0 1 0            0 0 0  8  4  5       9  6  4       5541035041297122 8    102.0   114.0
11 1000.288940 1.899E-24 2.265E-02.07100.350 1813.22260.670.008260          0 0 0            0 0 0 12  3  5      11  3  8       574720303367 118 6     75.0    69.0
17 1000.427056 5.410E-33 1.969E+00.05440.337 2994.18740.630.000000          0 2 0            0 1 0 15  4 12      16  3 13       4524205041 9 6 7 0    279.0   297.0
17 1000.469416 8.256E-31 2.685E+00.08810.459 1900.00490.610.000000          0 2 0            0 1 0  8  3  5       9  4  6       5524205041 5 6 7 0    153.0   171.0
11 1000.532321 2.190E-28 1.305E-05.04630.281 2144.04590.47-.011030          0 1 0            0 0 0 12  1 12      11  6  5       544233443231712211     75.0    69.0
11 1000.757766 2.349E-28 2.865E-03.01720.195 3291.14700.34-.001300          0 0 0            0 0 0 17  6 11      17  1 16       444233443231712211    105.0   105.0
14 1000.803804 8.305E-29 7.201E-05.08830.371  739.16760.63-.008415          0 1 0            0 0 0  6  0  6       6  5  1       5511065041 7712248     78.0    78.0
14 1000.940995 2.381E-28 2.080E-03.08000.315 1244.43970.55-.002274          0 1 0            0 0 0  8  4  4       9  6  3       5541035041287122 8    102.0   114.0
11 1001.136870 2.692E-25 4.140E-01.04000.287 2586.52880.41-.015279          0 1 0            0 0 0 12  2 10      13  5  9       5744263033384672224    25.0    27.0
13 1001.429857 3.492E-29 8.179E-02.06870.281 1136.21520.47-.003999          0 1 0            0 0 0 10  0 10      11  3  9       463106504136712247    126.0   138.0
14 1001.546134 1.644E-29 9.146E-03.08960.371 2009.22080.63-.011954          0 2 0            0 1 0  5  1  5       6  4  2       5511065041 7712248     66.0    78.0
17 1001.814336 8.602E-32 6.250E-04.08400.443  411.54210.610.000000          0 1 0            0 0 0  5  2  3       5  5  0       5584205041 5 6 7 0     99.0    99.0
12 1001.860676 4.750E-29 7.406E-01.04520.259 3249.63110.35-.008819          0 0 0            0 0 0 15 10  6      14  7  7       474446543685671837     31.0    29.0
17 1002.036816 2.398E-32 1.230E+01.07330.359 2809.72340.610.000000          0 3 0            0 2 0  4  4  0       5  5  1       4544205041 6 6 7 0     81.0    99.0
17 1002.036884 1.199E-32 1.230E+01.08880.359 2809.72300.610.000000          0 3 0            0 2 0  4  4  1       5  5  0       4524205041 9 6 7 0     81.0    99.0
14 1002.088711 4.448E-29 4.524E-02.05020.331 2149.05910.370.007208          0 1 0            0 0 0 14  3 12      15  4 11       554103504131712211    174.0   186.0
17 1002.113074 2.725E-30 1.766E+00.06260.360 1561.29810.620.000000          0 1 0            0 0 0 16  0 16      17  0 17       4524205041 9 6 7 0    297.0   630.0
17 1002.113074 5.450E-30 1.766E+00.06260.360 1561.29810.620.000000          0 1 0            0 0 0 16  0 16      17  0 17       4524205041 9 6 7 0    297.0   630.0
11 1002.299330 3.333E-27 4.827E-04.07090.316 2338.29960.660.003340          0 0 0            0 0 0 15  3 10      15  0 15       374426443385671837     93.0    93.0
13 1002.446042 7.350E-29 1.543E-01.04710.300 2792.35600.65-.003289          0 0 0            0 0 0 13 11  3      12  8  4       374436343685671837     27.0    25.0
12 1002.464088 2.390E-28 1.547E-01.04710.300 2792.33790.65-.003295          0 0 0            0 0 0 13 11  2      12  8  5       174446343685671837     81.0    75.0
```

2015+

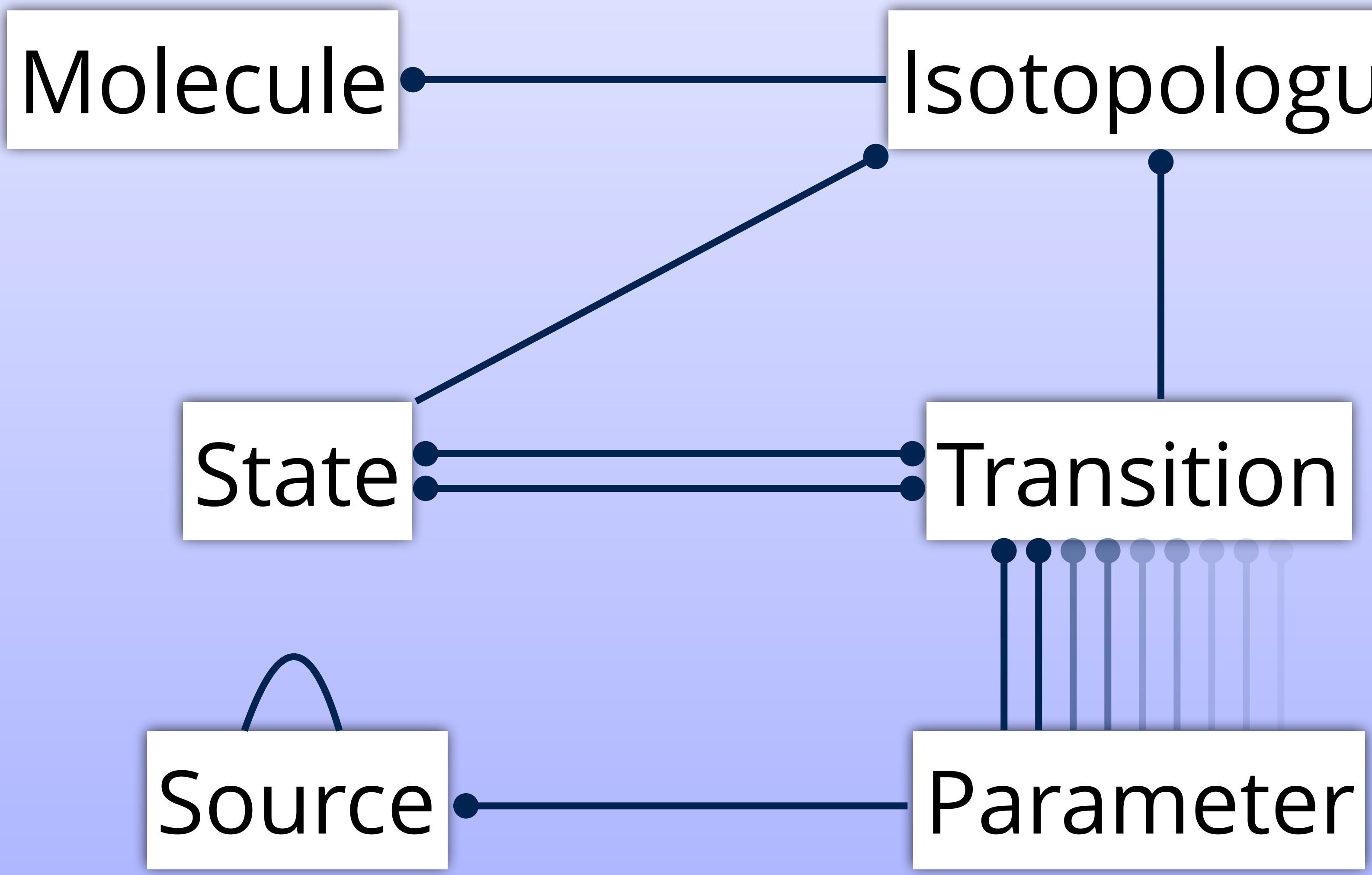
HITRAN and HITRAN*online*

http://hitran.org

- New line shape parameters (beyond Voigt)
- Many broadening species
- Pressure shifts
- Quantum numbers / labels
- Automated bibliography generation
- Uncertainties

HITRAN and HITRAN*online*

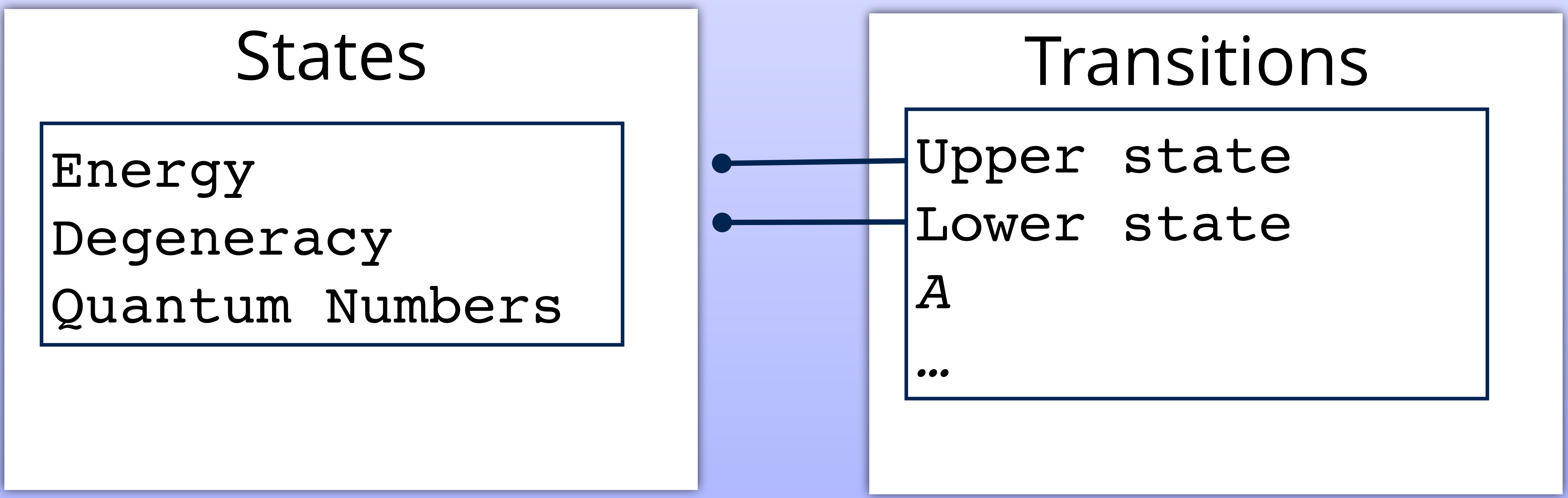
Relational Database Structure (MySQL)



HITRAN and HITRAN*online*

Relational Database Tables

Molecular **states**, linked by (radiative) **transitions**



States

```
Energy
Degeneracy
Quantum Numbers
```

Transitions

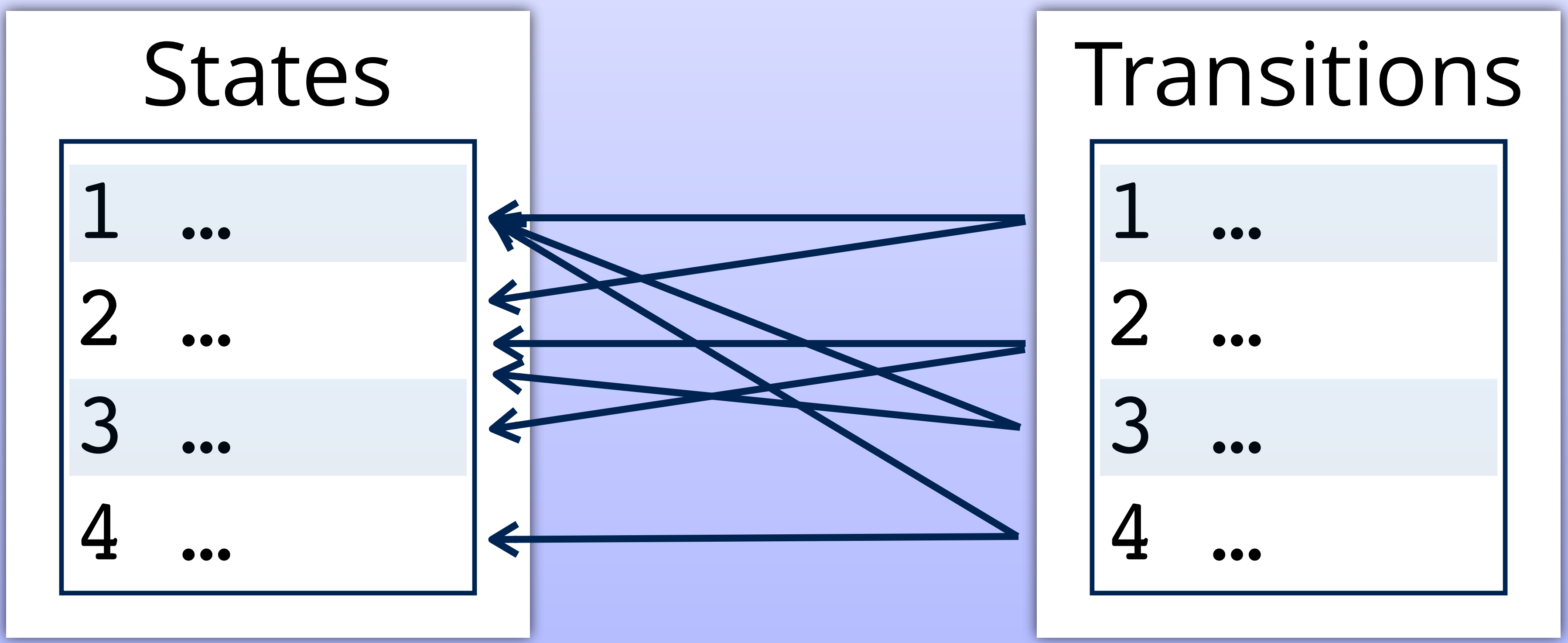
```
Upper state
Lower state
```

A

…

HITRAN and HITRAN*online*

Normalize as far as possible



States

1 …
2 …
3 …
4 …

Transitions

1 …
2 …
3 …
4 …

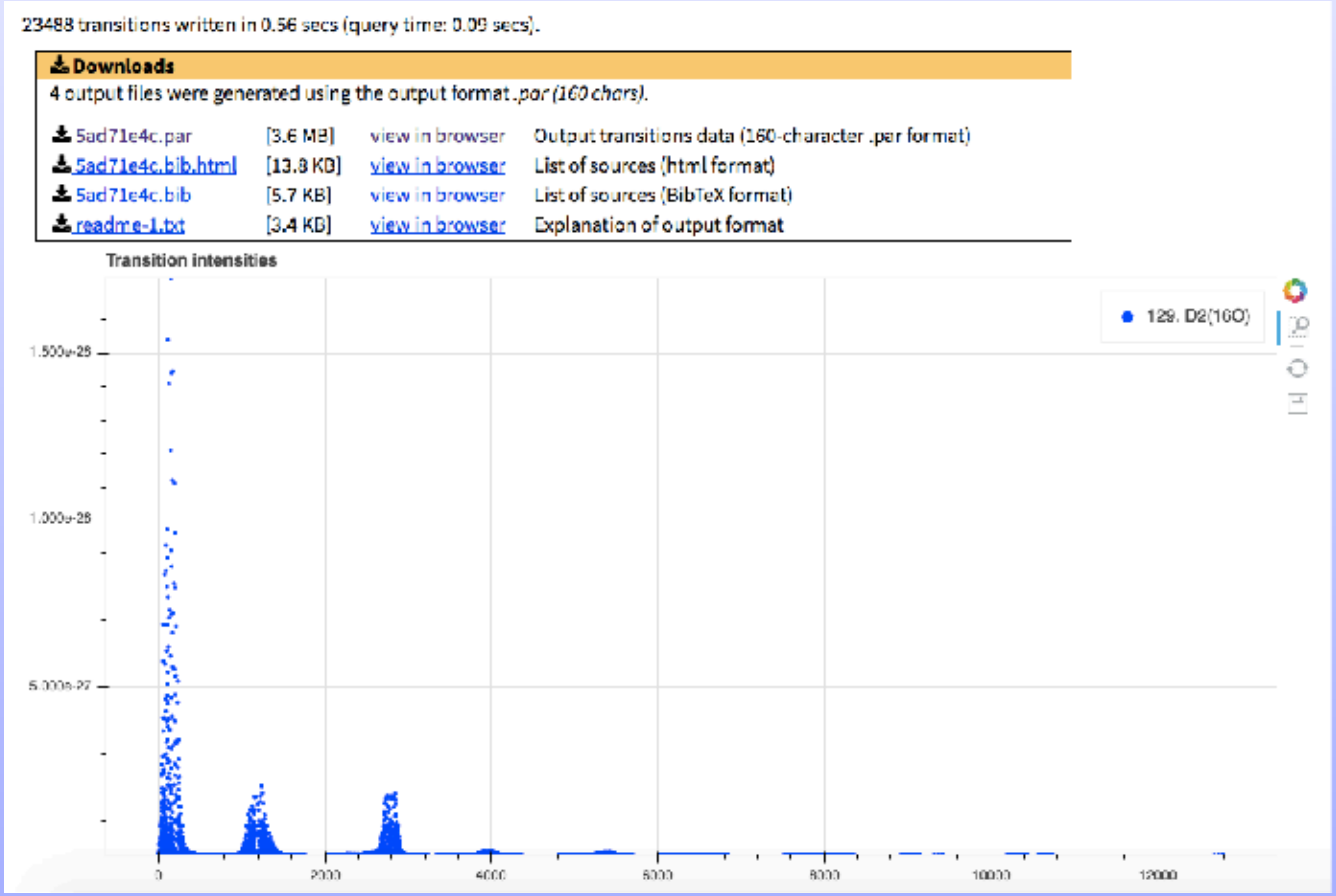
HITRAN and HITRAN*online*

User Interface

- Users register with email address and password
- Email addresses are verified
- Accessible contact form for problems / questions
- User profiles allow **customised output formats**
- Sources (citations) automatically included in output
- Interactive charts for moderate data volumes (<100,000 transitions)

HITRAN and HITRAN*online*

Interactive chart for data visualisation



23488 transitions written in 0.56 secs (query time: 0.09 secs).

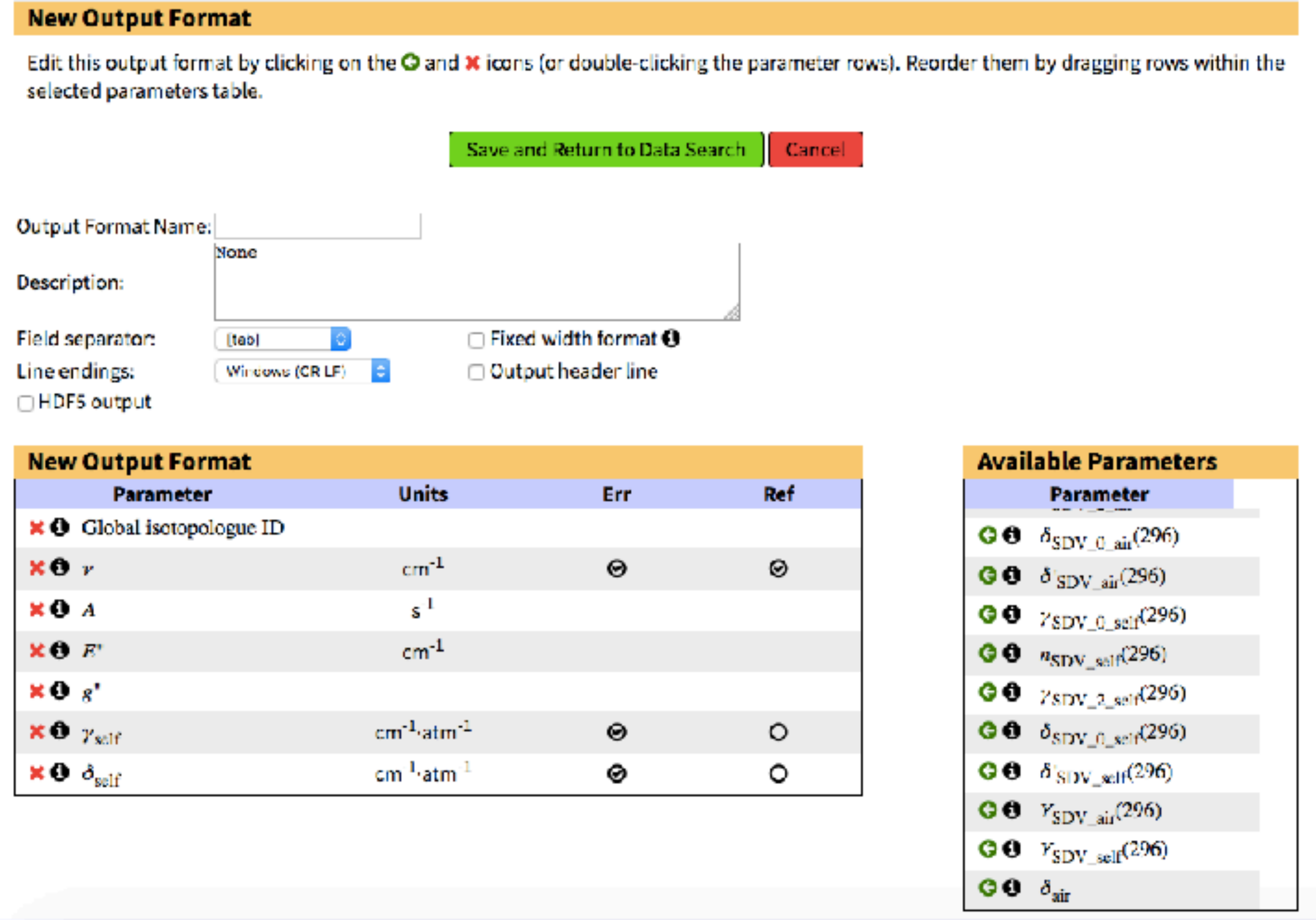
Downloads

4 output files were generated using the output format *.par (160 chars)*.

5ad71e4c.par	[3.6 MB]	view in browser	Output transitions data (160-character .par format)
5ad71e4c.bib.html	[13.8 KB]	view in browser	List of sources (html format)
5ad71e4c.bib	[5.7 KB]	view in browser	List of sources (BibTeX format)
readme-1.txt	[3.4 KB]	view in browser	Explanation of output format

HITRAN and HITRAN*online*

User-defined output formats



New Output Format

Edit this output format by clicking on the ← and ✖ icons (or double-clicking the parameter rows). Reorder them by dragging rows within the selected parameters table.

Output Format Name:

Description: None

Field separator: [tab]

☐ Fixed width format ℹ

Line endings: Windows (CR LF)

☐ Output header line

☐ HDF5 output

New Output Format

Parameter	Units	Err	Ref
✖ ℹ Global isotopologue ID			
✖ ℹ ν	cm^{-1}	✓	✓
✖ ℹ A	s^{-1}		
✖ ℹ E'	cm^{-1}		
✖ ℹ g'			
✖ ℹ γ_{self}	$cm^{-1}\cdot atm^{-1}$	✓	○
✖ ℹ δ_{self}	$cm^{-1}\cdot atm^{-1}$	✓	○

Available Parameters

Parameter
← ℹ $\delta_{SDV\_0\_air}(296)$
← ℹ $\delta'_{SDV\_air}(296)$
← ℹ $\gamma_{SDV\_0\_self}(296)$
← ℹ $n_{SDV\_self}(296)$
← ℹ $\gamma_{SDV\_2\_self}(296)$
← ℹ $\delta_{SDV\_0\_self}(296)$
← ℹ $\delta'_{SDV\_self}(296)$
← ℹ $Y_{SDV\_air}(296)$
← ℹ $Y_{SDV\_self}(296)$
← ℹ δ_{air}

HITRAN and HITRAN*online*

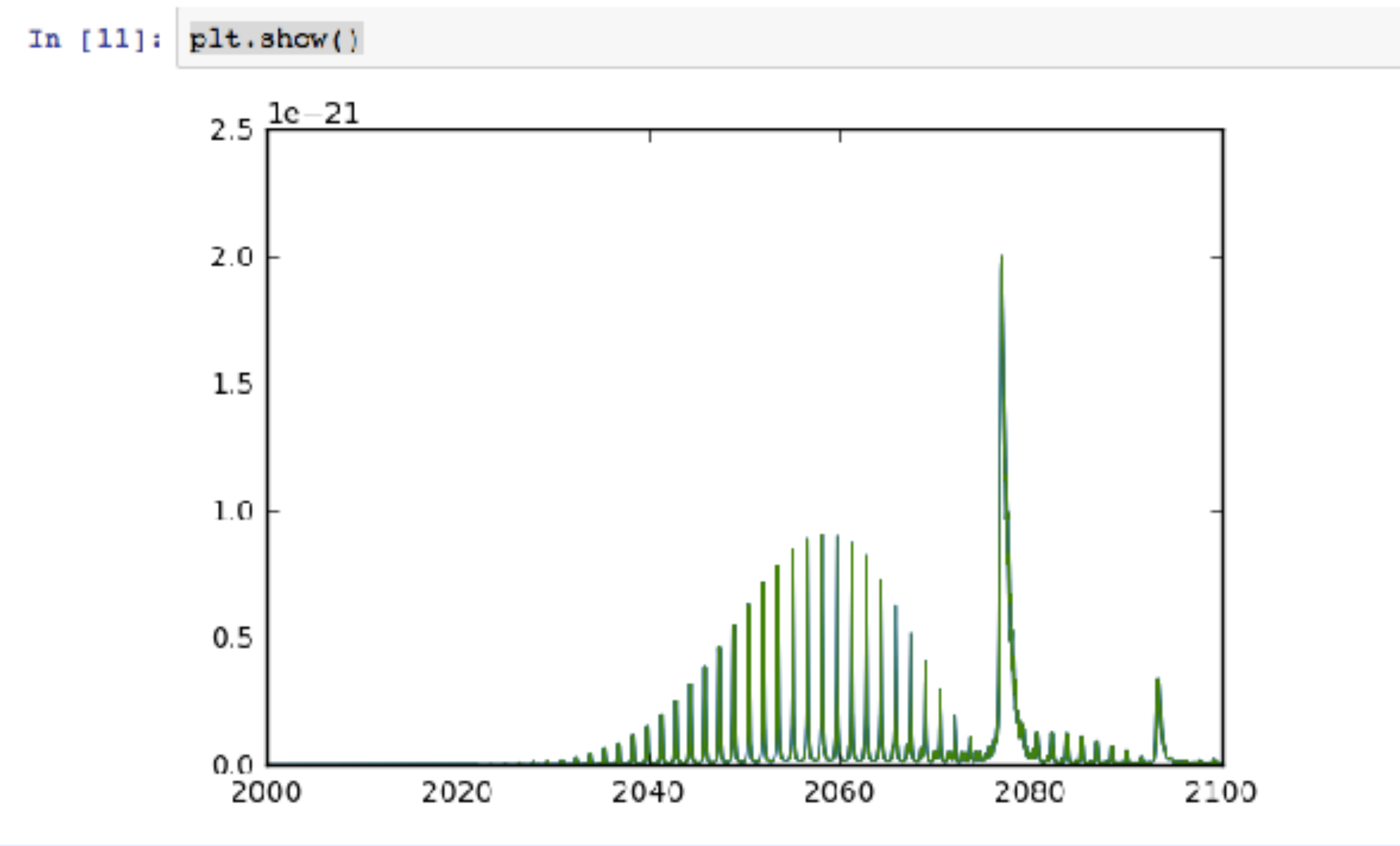
HAPI

- HAPI = HITRAN Application Programming Interface
- Python-based library for accessing HITRANonline programmatically and performing common operations on the data:
  - Non-expert users can utilise advanced line shape formulations
  - Checks for updates of the latest data
  - Allows for flexible, distributable, reusable code

HITRAN and HITRAN*online*

HAPI

```
import hapi
hapi.db_begin('data')
hapi.fetch(‘CO2’, 2, 1, 2000, 2100)
nu, coef = hapi.absorptionCoefficient_Lorentz(SourceTables=‘CO2')
plt.plot(nu, coef)
```



```
In [11]: plt.show()
```

ExoMol

http://exomol.com

- Compilation of molecular spectroscopic parameters for atmospheres of cool stars and exoplanets
- Some data sets can get *extremely large*:
  - 9.8 x 10^9 lines for CH_4 up to 1500 K
  - 1.68 x 10^{10} lines for PH_3 up to 1500 K
  - 2.0 x 10^{10} lines for H_2O_2 up to 1250 K
  - 2.1 x 10^{10} lines for SO_3 up to 800 K
  - 6.27 x 10^{10} lines for SiH_3 up to 1200 K

ExoMol

ExoMol Data Types

- *Ab initio* energy levels
- *Ab initio* transition probabilities (*A* /s^{-1})
- Partition functions
- Heat capacities
- Cooling functions
- Line-by-line pressure broadening parameters

ExoMol

ExoMol Data Storage

Data sets too large for relational databases, so…

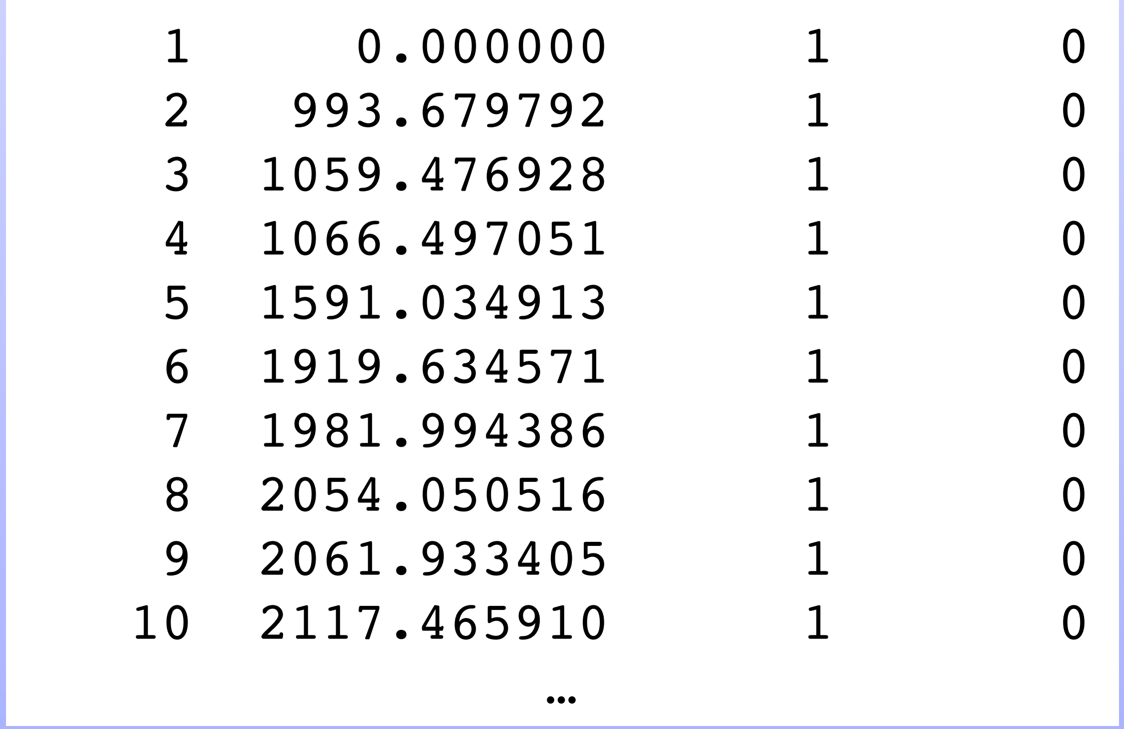
- Store the metadata in a relational database:
  - Molecules and Isotopologues
  - Data Types
  - Sources (citations)
- Store the energies in a single file
- Store the transitions in **compressed archives** over wavenumber intervals, with *references to the energies file* (fully normalized).

ExoMol

ExoMol Data Storage

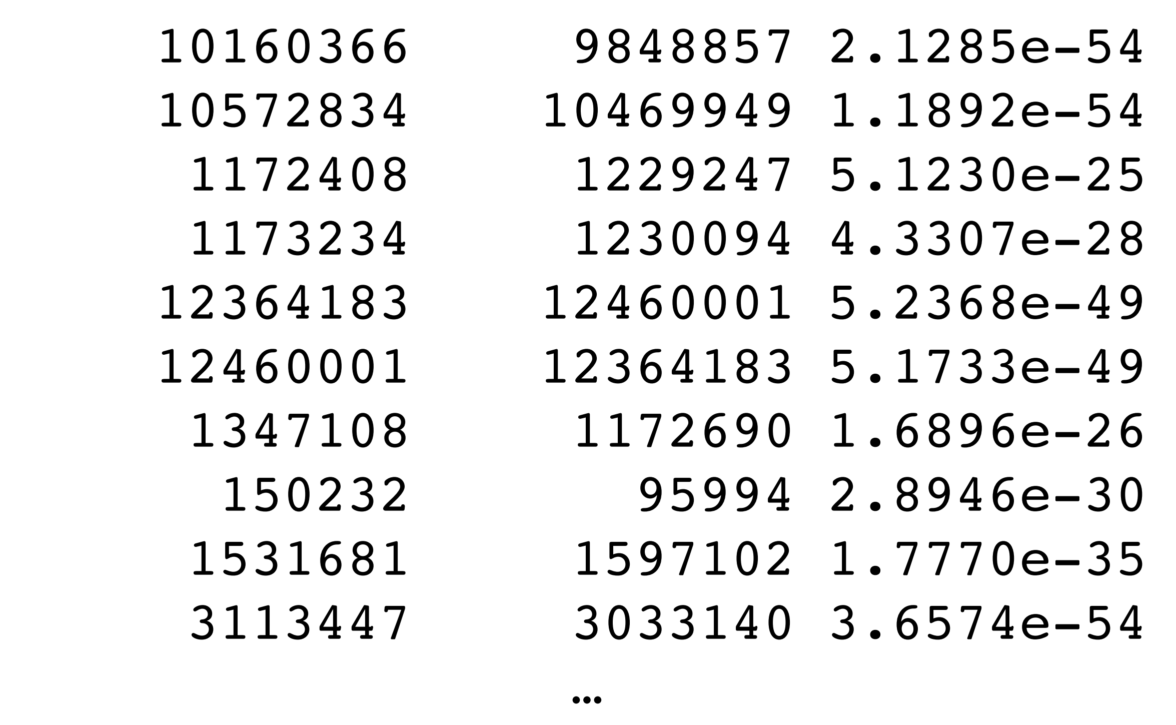
e.g. SO_3: 2 TB → 195 GB (compressed)

states (18530508)



```
    1      0.000000       1         0
    2    993.679792       1         0
    3   1059.476928       1         0
    4   1066.497051       1         0
    5   1591.034913       1         0
    6   1919.634571       1         0
    7   1981.994386       1         0
    8   2054.050516       1         0
    9   2061.933405       1         0
   10   2117.465910       1         0
                 …
```

transitions (21413927818)



```
    10160366      9848857 2.1285e-54
    10572834     10469949 1.1892e-54
     1172408      1229247 5.1230e-25
     1173234      1230094 4.3307e-28
    12364183     12460001 5.2368e-49
    12460001     12364183 5.1733e-49
     1347108      1172690 1.6896e-26
      150232        95994 2.8946e-30
     1531681      1597102 1.7770e-35
     3113447      3033140 3.6574e-54
                 …
```

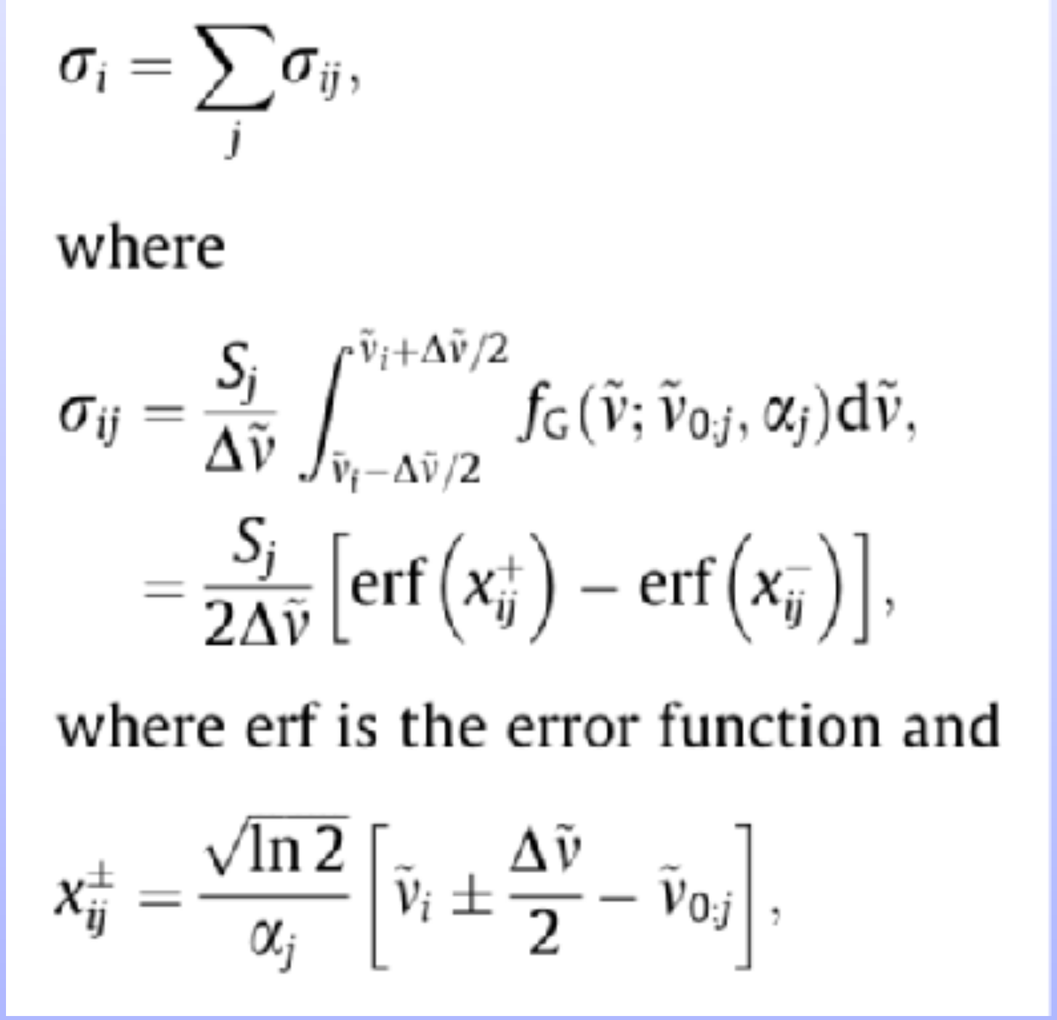
ExoMol

Data Reduction

Not all applications require a full line-by-line treatment.

Full Cross section approach

- Pre-calculate absorption cross sections at high-resolution (wavenumber grid) for a range of *T*.
- Provide a service to interpolate and bin to the requested (*T*, Δ*ν*)



$$\sigma_i = \sum_j \sigma_{ij},$$

where

$$\sigma_{ij} = \frac{S_j}{\Delta\tilde{\nu}} \int_{\tilde{\nu}_i-\Delta\tilde{\nu}/2}^{\tilde{\nu}_i+\Delta\tilde{\nu}/2} f_{\mathrm{G}}(\tilde{\nu};\tilde{\nu}_{0;j},\alpha_j)\,\mathrm{d}\tilde{\nu}, \\ = \frac{S_j}{2\Delta\tilde{\nu}}\left[\mathrm{erf}\left(x_{ij}^{+}\right) - \mathrm{erf}\left(x_{ij}^{-}\right)\right],$$

where erf is the error function and

$$x_{ij}^{\pm} = \frac{\sqrt{\ln 2}}{\alpha_j}\left[\tilde{\nu}_i \pm \frac{\Delta\tilde{\nu}}{2} - \tilde{\nu}_{0;j}\right],$$

ExoMol

Data Reduction

Hybrid approach

- Use cross sections for the very large number of overlapping weak lines
- Retain line-by-line treatment for the strongest lines

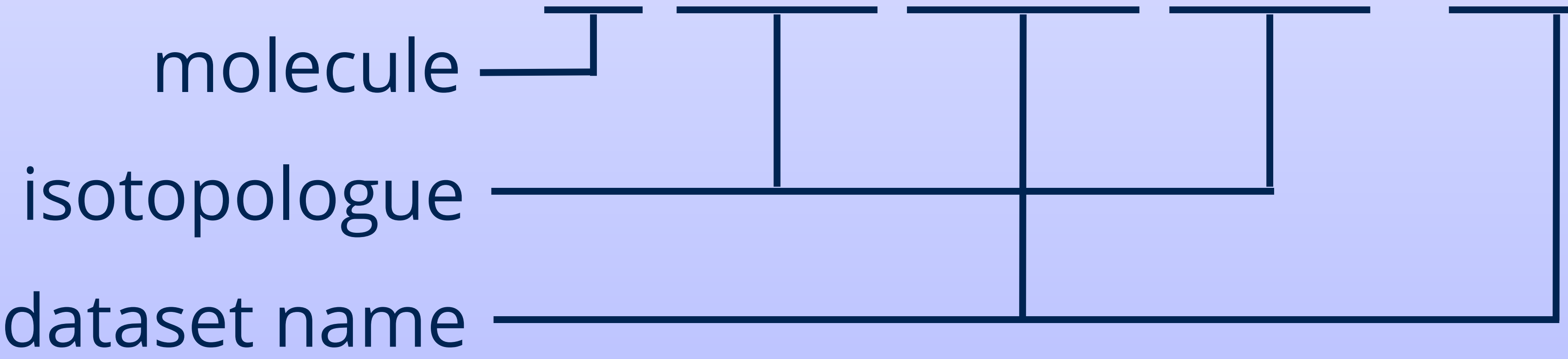
S. N. Yurchenko *et al.*, *A&A* **605**, A95 (2017).

ExoMol

API

Per–dataset “.def” file in predefined and persistent location, e.g.

http://www.exomol.com/db/CH4/12C-1H4/YT10to10/12C-1H4__YT10to10.def



molecule

isotopologue

dataset name

Includes data version / date stamp, data column definitions, dataset file locations.

ExoMol

Data Expansion(!)

- Don’t store redundant data:
  - Don’t store more decimal places than justified by the data accuracy
  - Don’t store arithmetic sequences of (e.g. wavelength or energy grids) explicitly – generate them as needed
- Don’t store the data more than once:
  - Provide scripts to interconvert between commonly used formats (e.g. ExoMol → HITRAN)

QuantemolDB

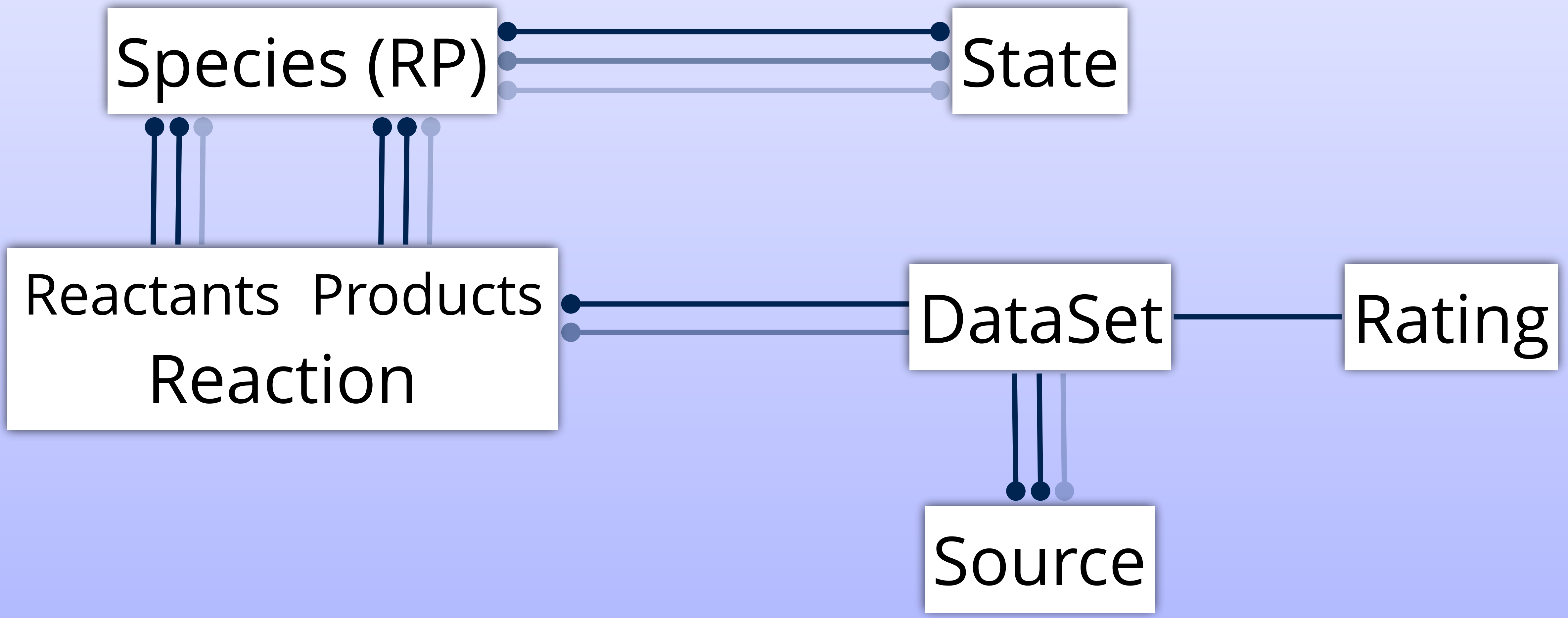
https://quantemoldb.com

Compilation of collision cross sections and rate coefficients for plasma processes:

- 16715 Reactions
- 17491 Data sets
- 1913 Species
- 3113 “Stateful Species”

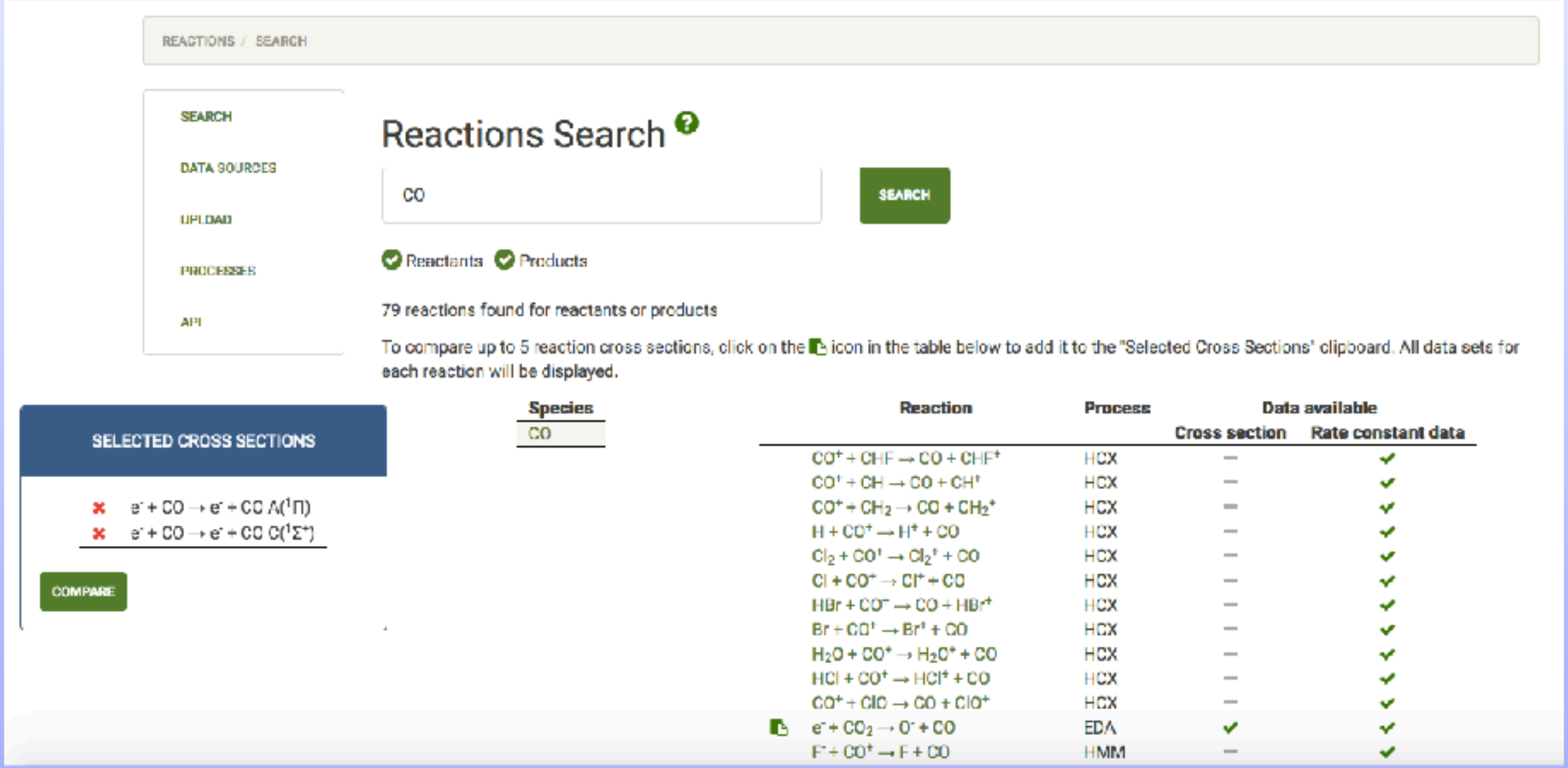
QuantemolDB

Relational Database Structure (MySQL)



QuantemolDB

Cross section search



Reactions Search

CO

79 reactions found for reactants or products

To compare up to 5 reaction cross sections, click on the icon in the table below to add it to the "Selected Cross Sections' clipboard. All data sets for each reaction will be displayed.

SELECTED CROSS SECTIONS

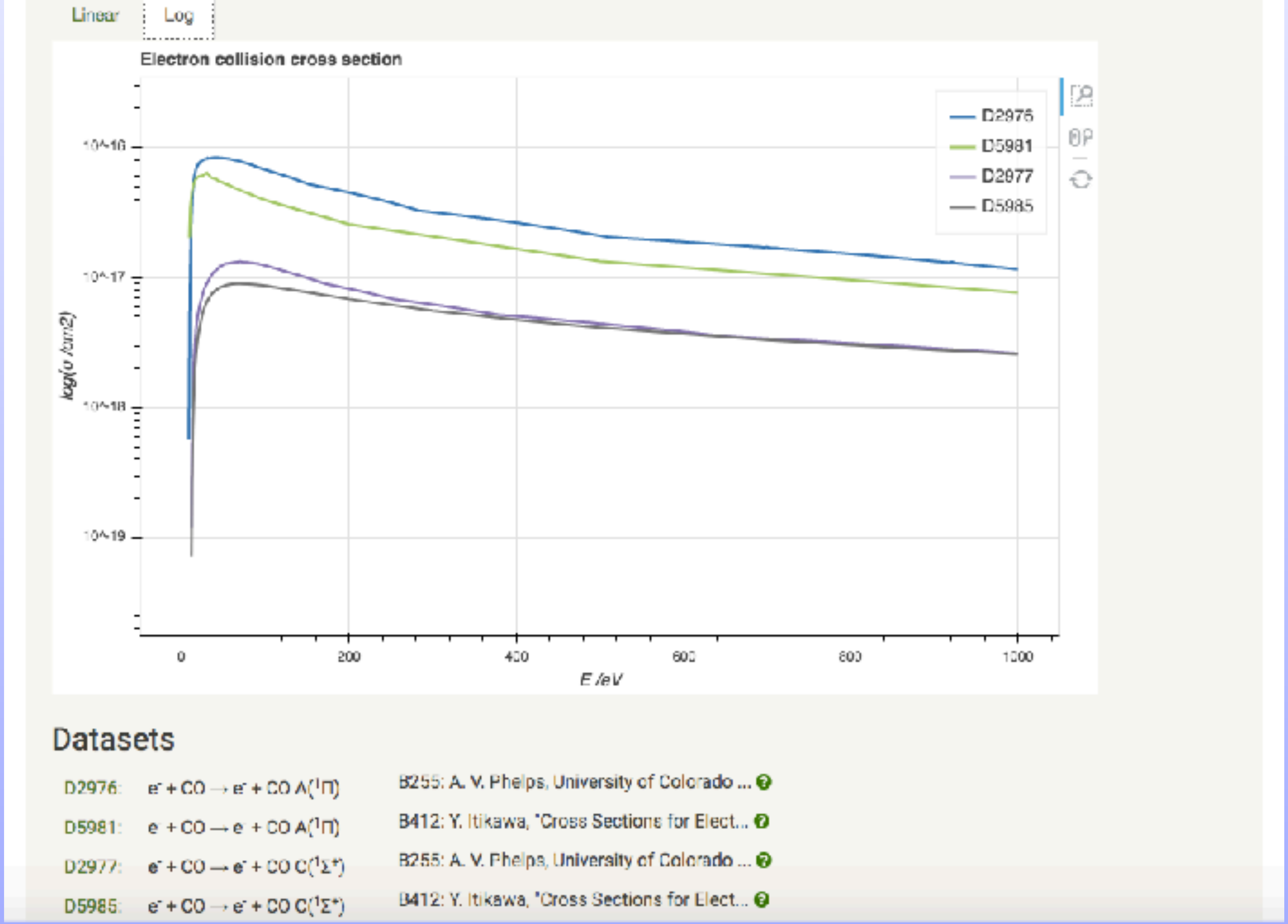
- ✖ $e^- + CO \rightarrow e^- + CO\ A(^1\Pi)$
- ✖ $e^- + CO \rightarrow e^- + CO\ C(^1\Sigma^+)$

Species
CO

	Reaction	Process	Data available	
			Cross section	Rate constant data
	$CO^+ + CHF \rightarrow CO + CHF^+$	HCX	—	✓
	$CO^+ + CH \rightarrow CO + CH^+$	HCX	—	✓
	$CO^+ + CH_2 \rightarrow CO + CH_2^+$	HCX	—	✓
	$H + CO^+ \rightarrow H^+ + CO$	HCX	—	✓
	$Cl_2 + CO^+ \rightarrow Cl_2^+ + CO$	HCX	—	✓
	$Cl + CO^+ \rightarrow Cl^+ + CO$	HCX	—	✓
	$HBr + CO^+ \rightarrow CO + HBr^+$	HCX	—	✓
	$Br + CO^+ \rightarrow Br^+ + CO$	HCX	—	✓
	$H_2O + CO^+ \rightarrow H_2O^+ + CO$	HCX	—	✓
	$HCl + CO^+ \rightarrow HCl^+ + CO$	HCX	—	✓
	$CO^+ + ClO \rightarrow CO + ClO^+$	HCX	—	✓
🗎	$e^- + CO_2 \rightarrow O^- + CO$	EDA	✓	✓
	$F^- + CO^+ \rightarrow F + CO$	HMM	—	✓

QuantemolDB

Interactive cross section comparison



Datasets

D2976:	$e^- + CO \rightarrow e^- + CO\ A(^1\Pi)$	B255: A. V. Phelps, University of Colorado ...
D5981:	$e^- + CO \rightarrow e^- + CO\ A(^1\Pi)$	B412: Y. Itikawa, "Cross Sections for Elect...
D2977:	$e^- + CO \rightarrow e^- + CO\ C(^1\Sigma^+)$	B255: A. V. Phelps, University of Colorado ...
D5985:	$e^- + CO \rightarrow e^- + CO\ C(^1\Sigma^+)$	B412: Y. Itikawa, "Cross Sections for Elect...

QuantemolDB

Quantemol API

- Implemented via GET query in URL
- Authenticate users through API key
- Specify desired output format
- Return zip archive of all matching files …
- … or use a compatible format (COMSOL, HPEM)
- Supports queries for pre-defined “Chemistries” : validated and recommended Data Sets for particular plasma processes

CascadesDB

Molecular Dynamics Simulations of Collisional Cascades

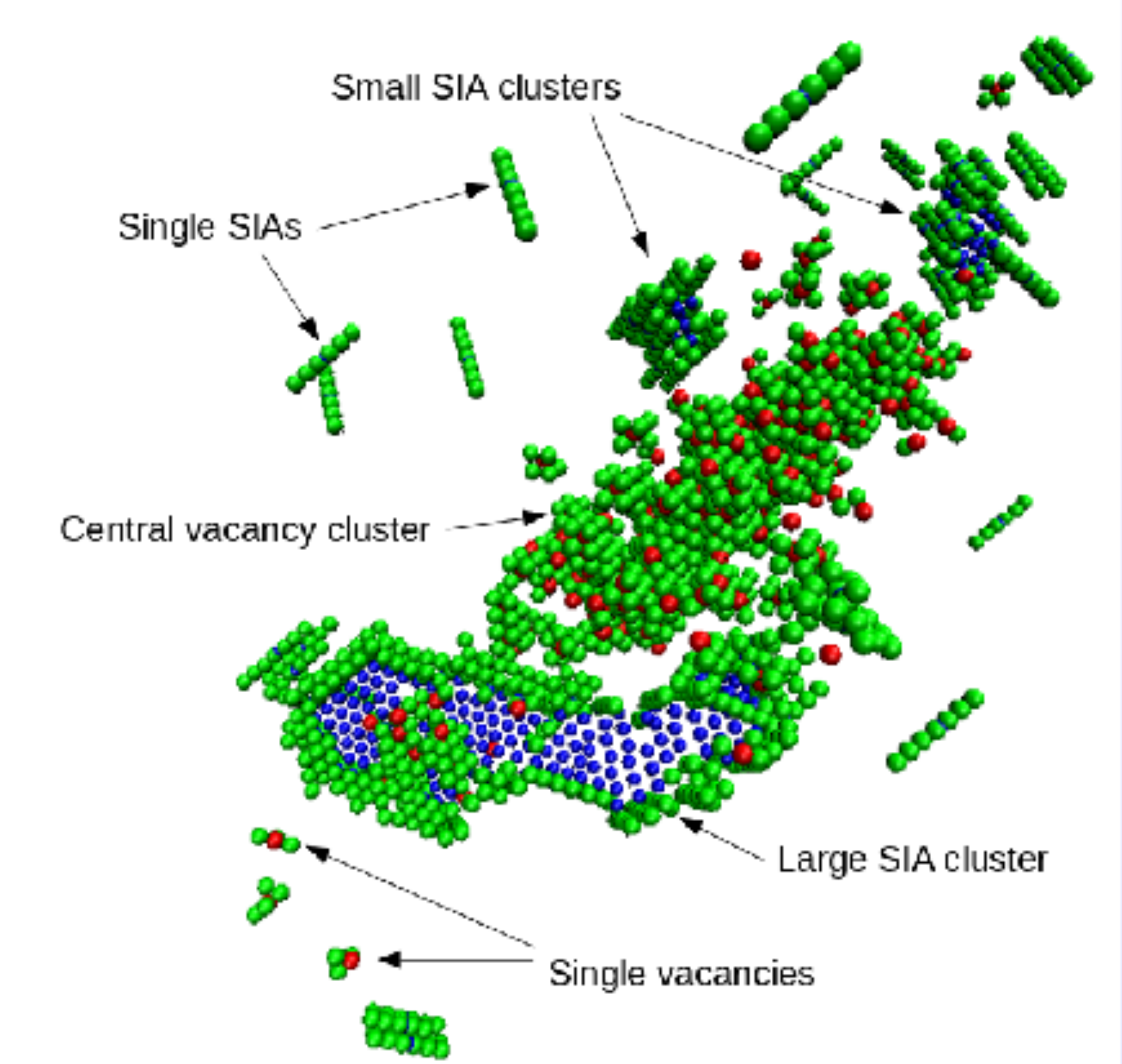


Figure credit: Andrea Sand, U. Helsinki

A repository of simulations of radiation damage in materials of relevance to fusion reactor design

CascadesDB

Data

- Stored as `.xyz` files
- Archived into batches differing only in PKA recoil direction
- (Compressed) archive up to ~10 GB in size

CascadesDB

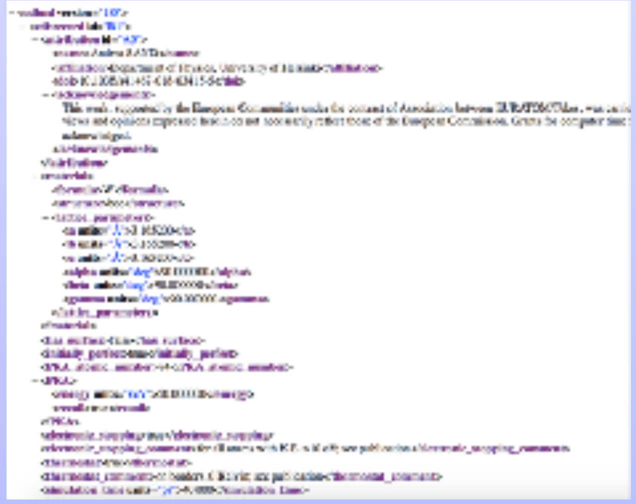
Metadata (searchable)

- Attribution
- Material parameters:
  - Lattice parameters
  - Initial crystal configuration
- Simulation details:
  - Code name and version
  - Temperature
  - Simulation time
  - Interatomic potential used

CascadesDB

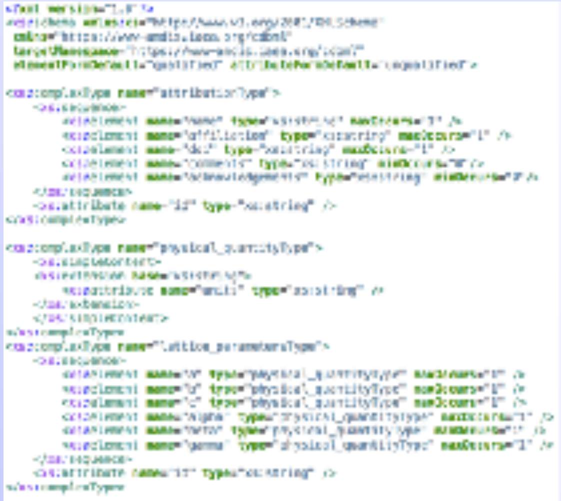
Metadata representations

MySQL:
Searchable



```
- <cdbml version="1.0">
  - <cdbrecord id="R1">
    - <attribution id="A5">
        <name>Andrea SAND</name>
        <affiliation>Department of Physics, University of Helsinki</affiliation>
        <doi>10.1007/s11467-[illegible]</doi>
      - <acknowledgements>
          This work, supported by the European Communities under the contract of Association between EURATOM/Tekes, was carri
          views and opinions expressed herein do not necessarily reflect those of the European Commission. Grants for computer time
          acknowledged.
        </acknowledgements>
      </attribution>
    - <material>
        <formula>W</formula>
        <structure>bcc</structure>
      - <lattice_parameters>
          <a units="Å">3.165200</a>
          <b units="Å">3.165200</b>
          <c units="Å">3.165200</c>
          <alpha units="deg">90.000000</alpha>
          <beta units="deg">90.000000</beta>
          <gamma units="deg">90.000000</gamma>
        </lattice_parameters>
      </material>
      <has_surface>false</has_surface>
      <initially_perfect>true</initially_perfect>
      <PKA_atomic_number>74</PKA_atomic_number>
    - <PKA>
        <energy units="eV">[illegible]</energy>
        <recoil>[illegible]</recoil>
      </PKA>
      <electronic_stopping>true</electronic_stopping>
      <electronic_stopping_comment>[illegible]</electronic_stopping_comment>
      <thermostat>[illegible]</thermostat>
      <thermostat_comment>[illegible]</thermostat_comment>
      <simulation_time units="ps">[illegible]</simulation_time>
```

XML:
Computer readable

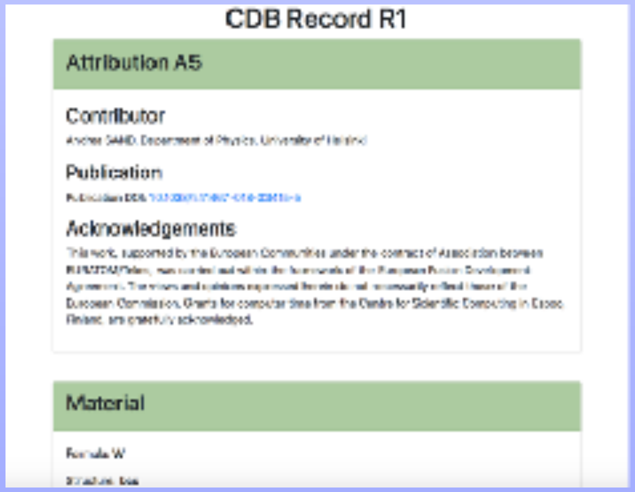


```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
 xmlns="https://www-amdis.iaea.org/cdbml"
 targetNamespace="https://www-amdis.iaea.org/cdbml"
 elementFormDefault="qualified" attributeFormDefault="unqualified">

<xs:complexType name="attributionType">
    <xs:sequence>
        <xs:element name="name" type="xs:string" maxOccurs="1" />
        <xs:element name="affiliation" type="xs:string" maxOccurs="1" />
        <xs:element name="doi" type="xs:string" maxOccurs="1" />
        <xs:element name="comments" type="xs:string" minOccurs="0" />
        <xs:element name="acknowledgements" type="xs:string" minOccurs="0" />
    </xs:sequence>
    <xs:attribute name="id" type="xs:string" />
</xs:complexType>

<xs:complexType name="physical_quantityType">
    <xs:simpleContent>
    <xs:extension base="xs:string">
        <xs:attribute name="units" type="xs:string" />
    </xs:extension>
    </xs:simpleContent>
</xs:complexType>
<xs:complexType name="lattice_parametersType">
    <xs:sequence>
        <xs:element name="a" type="physical_quantityType" maxOccurs="1" />
        <xs:element name="b" type="physical_quantityType" maxOccurs="1" />
        <xs:element name="c" type="physical_quantityType" maxOccurs="1" />
        <xs:element name="alpha" type="physical_quantityType" maxOccurs="1" />
        <xs:element name="beta" type="physical_quantityType" maxOccurs="1" />
        <xs:element name="gamma" type="physical_quantityType" maxOccurs="1" />
    </xs:sequence>
    <xs:attribute name="id" type="xs:string" />
</xs:complexType>
```

XML Schema:
Validation



CDB Record R1

Attribution A5

Contributor

Andrea SAND, Department of Physics, University of Helsinki

Publication

Publication DOI: [illegible]

Acknowledgements

This work, supported by the European Communities under the contract of Association between EURATOM/Tekes, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission. Grants for computer time from the Centre for Scientific Computing in Espoo, Finland, are gratefully acknowledged.

Material

Formula W

Structure bcc

HTML: Human readable

CascadesDB

Metadata links to data

Metadata

[illegible]

URIs

Online resources

- `.xyz` file archive
- Published article
- (Simulation code input file)
- (Interatomic potential)
- (Initial config `.xyz`)

Crowdsourcing

Types of crowdsourcing

- Creation of common goods: e.g. Wikipedia
- Carrying out micro tasks in parallel: e.g. Amazon Mechanical Turk
- Idea competitions / innovation contests
- Creative crowdsourcing: graphic design, architecture
- Crowdsolving
- Crowdfunding
- Crowdsearching
- Collaborative journalism
- Distributed computing

Crowdsourcing

Types of crowdsourcing

- Creation of common goods: e.g. Wikipedia
- Carrying out micro tasks in parallel: e.g. Amazon Mechanical Turk
- **Idea competitions / innovation contests**
- Creative crowdsourcing: graphic design, architecture
- Crowdsolving
- Crowdfunding
- Crowdsearching
- Collaborative journalism
- **Distributed computing**

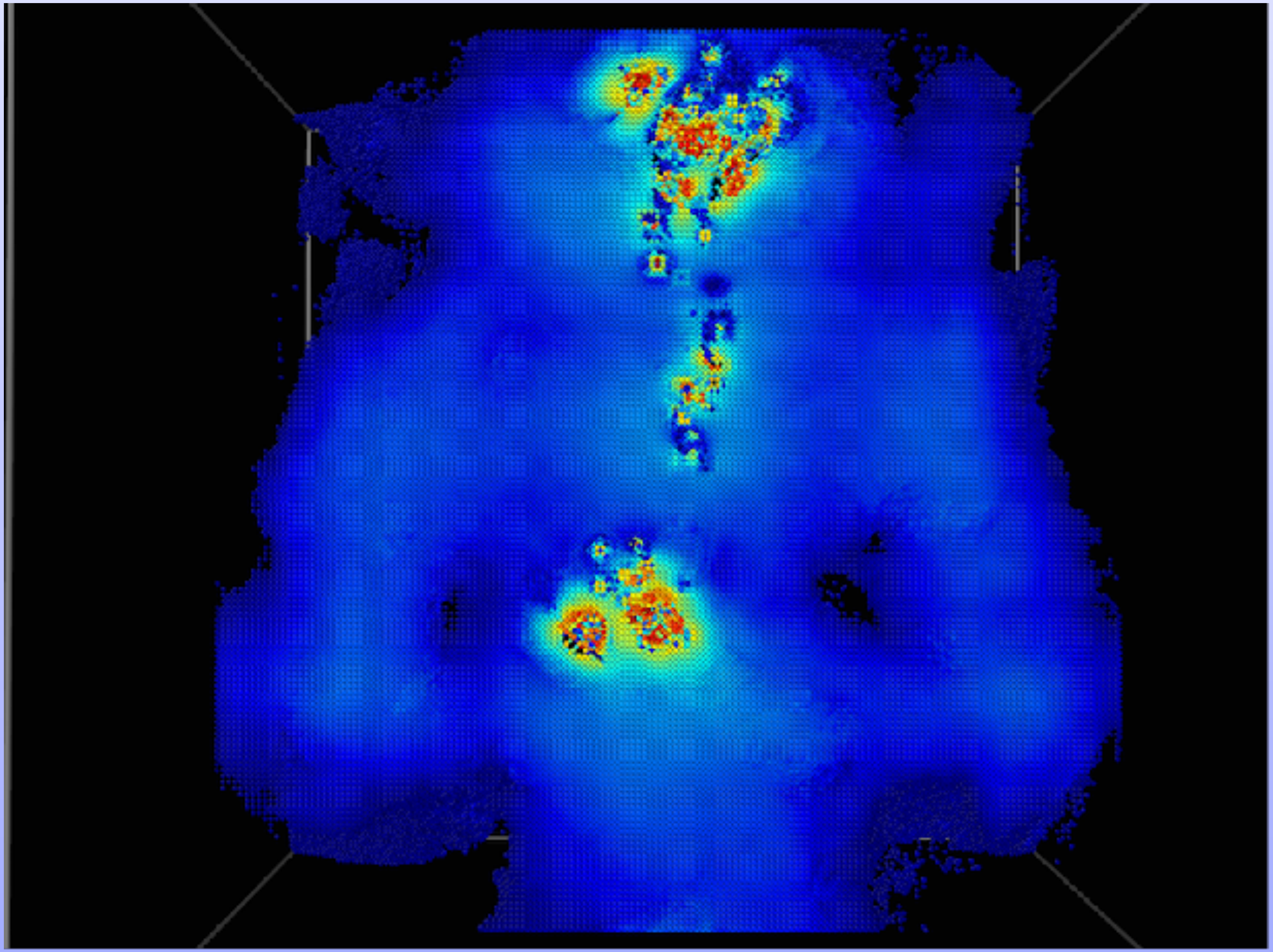
Crowdsourcing

A Brief History of Crowdsourcing

1714: Find a simple and practical method for the precise determination of a ship's longitude at sea (John Harrison: H4 sea watch)
1820: First Montyon prize awarded
1884: First fascicle of the OED (800 volunteers)
1957: Jørn Utzon won the design competition for the Sydney Opera House
1999: SETI@home
2001: Wikipedia
2007: Galaxy Zoo

Crowdsourcing

1. Classifying and visualising radiation damage in fusion-relevant materials



Crowdsourcing

1. Classifying and visualising radiation damage in fusion-relevant materials

- Competition / Challenge with €5,000 prize
- Provided data: ~50 MD simulations (.xyz files) of collisional cascades:
  - Fe and W
  - different PKA energies
  - (different recoil directions)
- Scientific Leads: Andrea Sand (U. Helsinki),
- Sergei Dudarev (CCFE)
- April – June 2018

Crowdsourcing

1. Classifying and visualising radiation damage in fusion-relevant materials

Participants are invited to come up with novel ways to visualise, analyse and explore the provided data. Successful submissions may involve one or more of the following:

- *Novel software for visualizing the material damage represented by the data files in a way that aids its qualitative and quantitative assessment.*
- *New software tools to rapidly and reliably identify, classify and quantify new patterns and structures of particular kinds in the data sets.*
- *Efficient algorithms to depict and summarise the statistical distribution of atom displacements and to analyse the effect of impact energy on this distribution.*

Crowdsourcing

Distributed Computing

- Invite members of the public to download MD simulation software to evolve a virtual crystal after impact damage.
- To be based on the BOINC (Berkeley Open Infrastructure for Network Computing) platform
- Data transferred to and from CascadesDB database

Crowdsourcing

Distributed Computing

Advantages

- Large scale, parallel computing power (cf. 600,000 users for climateprediction.net)
- Uncertainty quantification; interatomic potential validation
- Material discovery

Challenges

- Security
- Bandwidth, storage, scalability
- Maintaining user engagement

Managing large atomic and molecular data sets: HITRAN, ExoMol and CascadesDB

Thank You

Acknowledgments

- J. Tennyson, S. N. Yurchenko (UCL)
- L. S. Rothman, I. E. Gordon (Harvard-Smithsonian CfA)
- A. E. Sand (U. Helsinki) , S. Dudarev (CCFE)
- D. Wallom (U. Oxford)