



香港中文大學  
The Chinese University of Hong Kong



# Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform

Huiwen Tan  
29 Jan 2026



# Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform

M. Álvarez-Moreno,<sup>\*,†,‡</sup> C. de Graaf,<sup>‡,||</sup> N. López,<sup>†</sup> F. Maseras,<sup>†,§</sup> J. M. Poblet,<sup>‡</sup> and C. Bo<sup>\*,†,‡</sup>

<sup>†</sup>Institute of Chemical Research of Catalonia, ICIQ, Av. Països Catalans 16, 43007 Tarragona, Catalonia, Spain

<sup>‡</sup>Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, C/Marcel·lí Domingo s/n, 43007 Tarragona, Catalonia, Spain

<sup>§</sup>Department of Chemistry, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain

<sup>||</sup>Catalan Institution for Research and Advanced Studies, ICREA, Passeig Lluis Companys 23, 08010 Barcelona, Catalonia, Spain

# Background

- **Computational chemistry generates massive data**

These data are hardly usable by any person other than the creator himself

- **Current challenges**

Data often not shared, poorly organized, or not machine-readable

Extracting results to generate tables and figures from raw outputs is largely manual and labor-intensive

Limited compliance with FAIR (Findability, Accessibility, Interoperability, Reusability) standards





# FAIR Challenges in Computational Chemistry Data

## Findable

Data and metadata are uniquely identifiable and locatable, making them easy to find with search tools.

Most computational data are stored in SI or local folders

## Interoperable

Data uses standard formats, allowing different systems and tools to understand and exchange it.

Heterogeneous formats, software-dependent raw outputs

## Accessible

Data is retrievable, though access might be restricted (e.g., behind a login); metadata should always be open.

Data is often not publicly released, requiring direct contact with authors

## Reusable

Data is well-documented with enough metadata to allow users to understand its origin, methodology, and any potential limitations.

Missing metadata, provenance, and standardized format

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Research data for this article

Data not available / Data will be made available on request

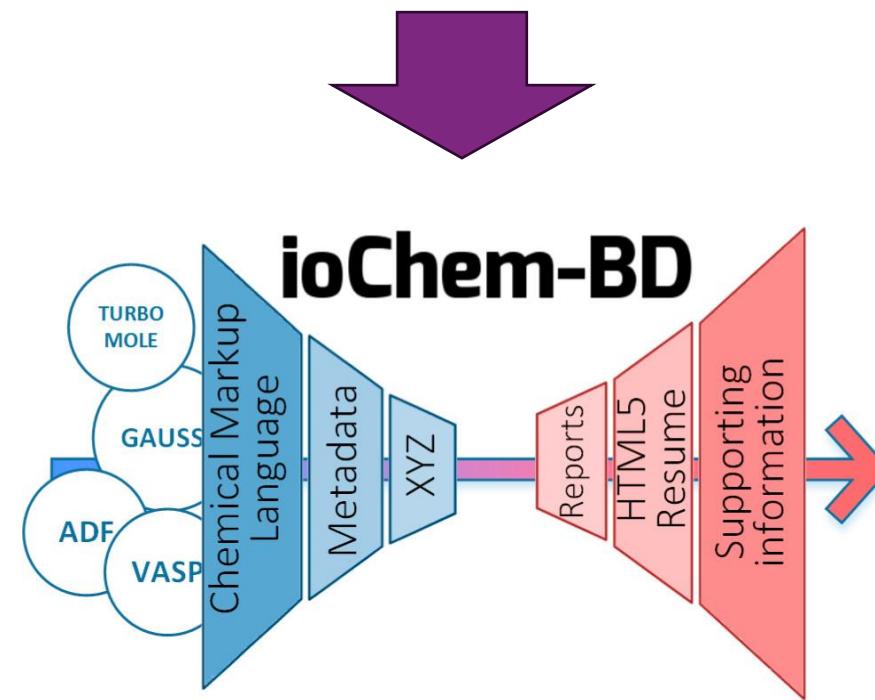


SI or local folders



## Motivation

Computational Chemists need structured, searchable, and reusable computational chemistry databases





Search by

Text

Enter search term



Find

Sign up

Login

Download

Docs ▾

# Open your research to the world

ioChem-BD - The Computational Chemistry Results Repository



1

Central service



7

Connected nodes



1,006

Collections available



389,954

Items indexed

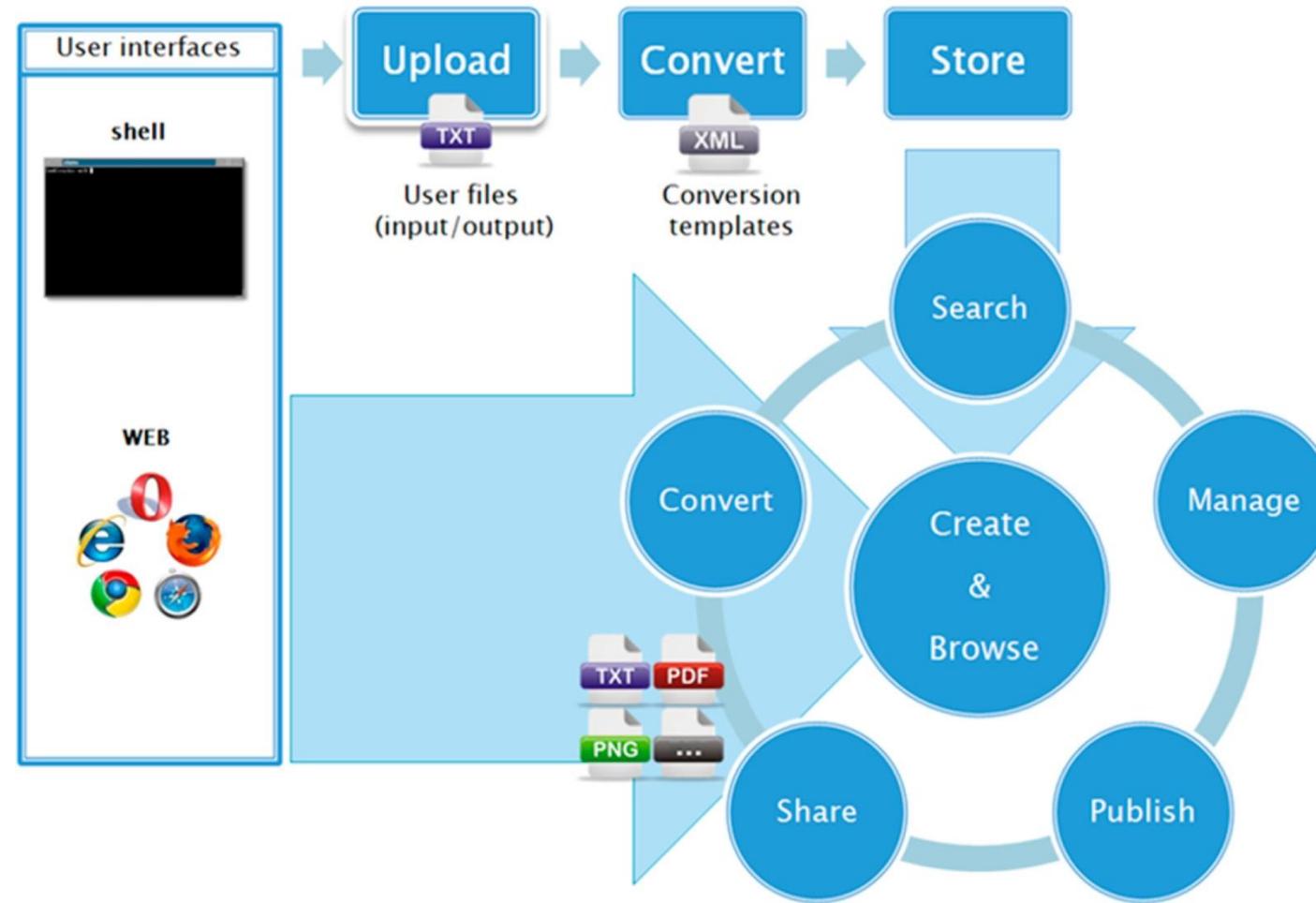


**OSCARS**  
Open Science Clusters' Action  
for Research & Society



Funded by  
the European Union

# ioChem-BD System Overview



# eXtensible Markup Language (XML)

- **What is XML?**

Structured, hierarchical, self-describing data format

Machine-readable and human-readable

- **XML Ecosystem**

**XSD:** schema validation (enforce data structure)

**XPath:** query language for XML trees

**XSLT:** rule-based transformation (XML → HTML/CSV/PDF)

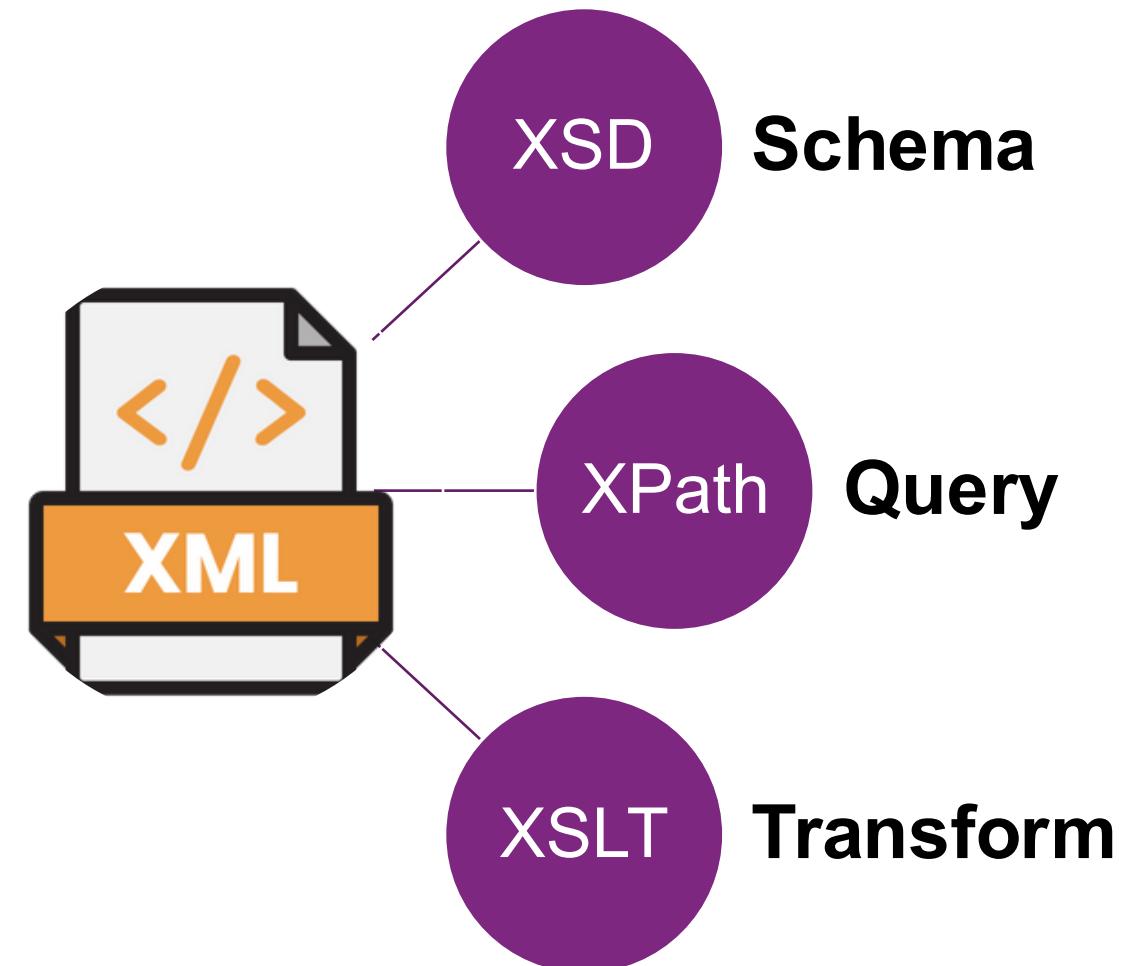
- **Why XML in ioChem-BD?**

Format neutrality and long-term interoperability

Validated and semantically extensible (CML)

XML

```
<calculation>
  <method>B3LYP</method>
  <basis>def2-TZVP</basis>
  <energy units="Hartree">-123.456</energy>
</calculation>
```



# Chemical Markup Language (CML)

- CML extends XML with standardized chemical concepts (molecules, atoms, bonds, computational metadata)
- However, later replaced by simpler, programmatic workflows (JSON, Python, HDF5)
- **CML pipeline**

Quantum chemistry output → CML (XML) → XSLT → Reports / knowledge graphs / databases

- **Modern pipeline**

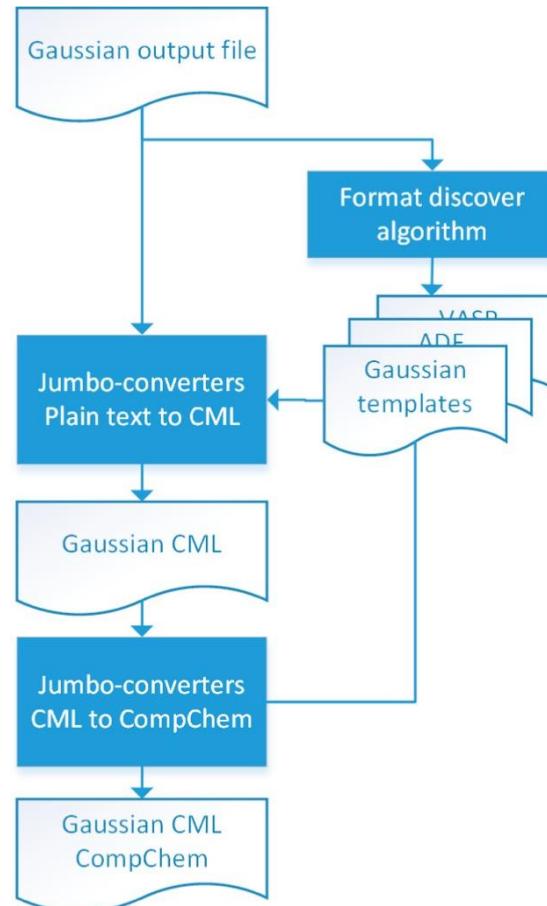
Quantum chemistry output → Python parser → ORM objects → SQL database

XML

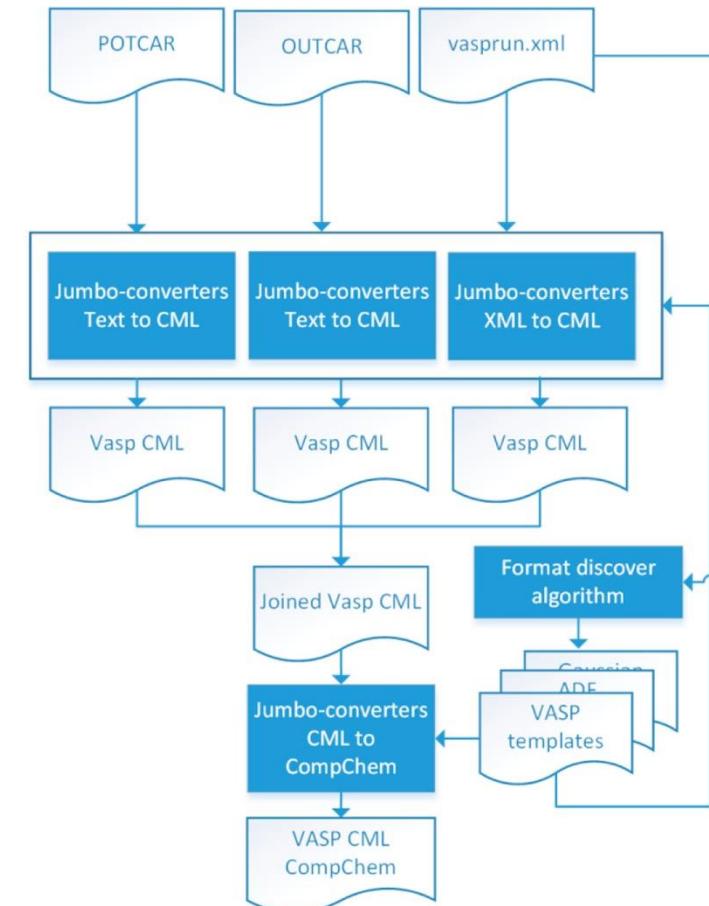
```
<molecule>
  <atomArray>
    <atom id="a1" elementType="C" x3="0.0" y3="0.0" z3="0.0"/>
  </atomArray>
</molecule>
```

```
<molecule cmlx:templateRef="mol"
  formalCharge="0"
  id="zmat"
  spinMultiplicity="1">
  <atomArray>
    <atom elementType="O"
      id="a1"
      x3="-1.07419"
      y3="0.95647"
      z3="0.0000"/>
    <atom elementType="H"
      id="a2"
      x3="-0.26225"
      y3="0.3896"
      z3="0.0000"/>
    <atom elementType="H"
      id="a3"
      x3="-1.82642"
      y3="0.31246"
      z3="0.0000"/>
  </atomArray>
  <bondArray>
    <bond atomRefs2="a1 a3" order="S"/>
    <bond atomRefs2="a1 a2" order="S"/>
  </bondArray>
  <formula concise="H2O"/>
  <property dictRef="cml:molmass">
    <scalar units="unit:dalton">15.9994</scalar>
  </property>
  <list cmlx:templateRef="charge">
    <list>
      <scalar dataType="xsd:integer" dictRef="g:charge">0</scalar>
      <scalar dataType="xsd:integer" dictRef="g:mult">1</scalar>
    </list>
  </list>
  <formula convention="iupac:inchi" inline="InChI=1S/H20/h1H2">
    <scalar dataType="xsd:integer" id="auxInfo">AuxInfo=1/0/N:1/rA:30HH/rB:s1;s1;/rC:;;;</scalar>
  </formula>
</molecule>
```

# Conversion Workflow from Output Files to CML



Gaussian and ADF: single output file



VASP: a group of output files

# Create Module Overview

**ioChem-BD Create**

Navigation/Edition | Search | Reports | Options | **Browse**

3D Structure | View Results | Download | RAW CML

Refresh Type Description Creation Date Handle Pub. Edit

- Fe(III)\_Keggin\_structures PRO Fe(III) Keggin structures 2018-06-14 11:09
- kimik2222 PRO - 2018-03-07 15:35
- Aromatic\_Amination\_of\_Lactones PRO Aromatic\_Amination\_of\_Lactones 2018-06-14 11:07
  - calc1 ADF calc1 2018-01-16 17:10
  - hexenol\_modified GAU hexenol 2018-04-12 18:18
  - hexenol4 GAU hexenol4 2018-04-12 18:55
  - hexenol33 GAU hexenol3 2018-04-12 18:40
  - hexenol5 GAU hexenol5 2018-04-12 19:08
  - hexenol6 GAU hexenol6 2018-04-12 19:18
  - h2 GAU h2 2018-06-18 08:32
  - calc GAU calc 2018-06-20 17:31
  - gaussian\_calc GAU - 2016-10-24 16:03
  - calc1 GAU calc 2017-02-14 16:44
  - calc1 GAU calc1 2017-08-04 09:48
- K\_Br-C4-1THF-i5 GAU K\_Br-C4-1THF-i5 2017-12-23 10:57
- upload\_sm1 GAU uploadsm1 2017-12-23 10:59
- upload\_sm2 GAU uploadsm2 2017-12-23 11:07
- upload\_sm3 GAU uploadsm3 2017-12-23 11:08
- upload\_sm4 GAU uploadsm4 2017-12-23 11:08
- Mo(I)\_hydrogen\_generation\_... PRO Mo(I)\_hydrogen\_generation 2018-06-14 11:10
  - cucurb GAU cucurb 2017-09-04 11:48
  - calc1\_2 GAU calc1 2017-09-06 11:06
  - hexenol2 GAU hexenol2 2018-04-12 18:33
  - sample2 ADF sample2 2018-10-15 19:15
- Ni\_catalyzed\_Aryl\_borylation\_RM PRO Ni\_catalyzed\_Aryl\_borylation\_RM 2018-06-14 11:11
  - adf ADF adf 2015-07-13 19:12
  - ts\_bp-uff GAU ts\_bp-uff 2015-07-08 18:40
  - g09 GAU g09 2015-07-09 19:13
  - vasp\_demo VSP vasp\_demp 2015-07-09 20:02
  - geomopt2012 ADF geomopt2012 2015-07-09 20:03
  - adf1 ADF adf1 2015-07-13 20:31
  - gaussian1 GAU gaussian1 2015-07-13 20:32
  - vasp1 VSP vasp1 2015-07-13 20:32
  - opt10 VSP opt10 2015-07-13 20:35
  - freq ADF freq 2015-07-13 20:39

Properties

Path: /db/testuser/Fe(III)\_Keggin\_structures/K\_Br-C4-1THF-i5

Name	K_Br-C4-1THF-i5	Type	GAU
Description	K_Br-C4-1THF-i5	State	modified
Owner	.cbo_group	Group	cbo_group
Creation date	2017-12-23 10:57	Mod. date	2018-07-18 17:45
Permissions	RW—	Pub. date	

**Actions:** Create Project, Modify,

3D Structure View Results Download RAW CML

Chemical structure visualization: A complex organic molecule with various functional groups, including aromatic rings, hydroxyl groups, and a central metal atom coordinated to multiple ligands.

# Create Module Overview

**Navigation frame**

**Item actions**

**Item details**

The image shows a screenshot of the ioChem-BD Create module interface. The left panel is a navigation frame displaying a hierarchical list of projects and their contents. Projects listed include Fe(III)\_Keggin\_structures, kimik2222, Aromatic\_Amination\_of\_Lactones, Mo(I)\_hydrogen\_generation\_cucurb, and Ni\_catalyzed\_aryl\_borylation\_RM. The right panel shows item actions for a selected item, K-Br-C4-1THF-15, which is a GAU type file. The item details panel shows the item's properties: Path (/db/testuser/Fe(III)\_Keggin\_structures/K-Br-C4-1THF-15), Name (K-Br-C4-1THF-15), Type (GAU), State (modified), Description (K-Br-C4-1THF-15), Owner (testuser), Concept Group (None), Permissions (rw-rw-rw-), Creation date (2017-12-23 10:57), and Mod. date (2018-01-17 01:20). Action buttons for Create Project, Modify, and Delete are also present.



# Uploading to Create

Properties

Path	/db/testuser
Name	acetate
Description	Acetate molecules
Owner	Group
Creation date	Mod. date

Type  
State  
Concept Group  
Permissions  
Pub. date

**Create Project** **Modify** **Delete**

Creating a project

Calculation upload

Name:

Description:

Type:

Gaussian  ADF  Vasp  Turbomole  Orca  Molcas  QuantumEspresso

Mopac  Mopac

Input file\*  a-siw12\_2.com

Output file\*  a-siw12\_2.log

Files:  
 Additional file -

\*Required

Upload Calculations

# Item Actions frame

3D Structure   View Results   Download   RAW CML

output.cml.html

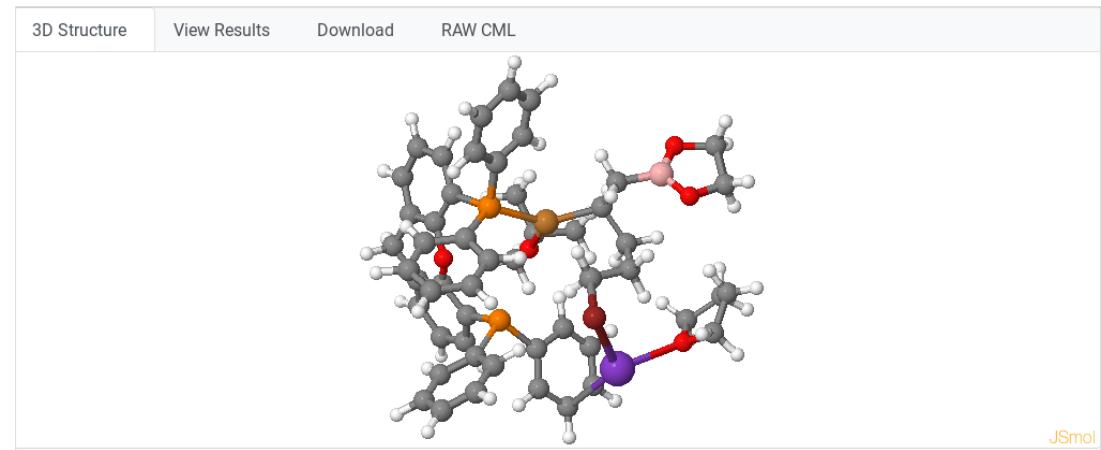
**GENERAL INFO**

Title: upload\_sni  
 Program: Gaussian 09 ES64L-G09RevD.01  
 Formula: C<sub>52</sub>H<sub>58</sub>BBrCuO<sub>5</sub>P<sub>2</sub>K  
 Calculation type: Geometry optimization Minimum  
 Method(s): RwB97XD - Grimme-D2  
 Temperature 298.150 K  
 Pressure 1.00000 atm

**ATOM INFO**

Atomic coordinates [Å] (optimized)

ATOM	x	y	z	TYPE	Core	ECP
1	Cu	0.2319	1.4549	-0.6657 lanl2dz	+	+
2	C	3.2566	-0.0165	0.6511 6-31g(d)		
3	P	-1.4161	-1.9130	0.0031 lanl2dz	+	+
4	P	-1.6259	2.1612	0.6160 lanl2dz	+	+
5	C	-2.3341	1.2641	2.0654 6-31g(d)		
6	C	-3.2503	-0.2076	4.2710 6-31g(d)		
7	C	-1.4320	0.6455	2.9376 6-31g(d)		
8	C	2.7022	1.1463	2.2214 6-31g(d)		



3D Structure	View Results	Download	RAW CML	JSmol
Download	File name	Mimetype	Size (kB)	
Download	output.cml	chemical/x-cml	7,364.14	
Download	K-Br-C4-1THF-I5.in	chemical/x-gaussian-input	7.29	

# Search Mode

Navigation/Edition    Search    Reports

Name

Description

Geometry

Elements

Type

Path

Owner

Group

Creation dates

Concept group

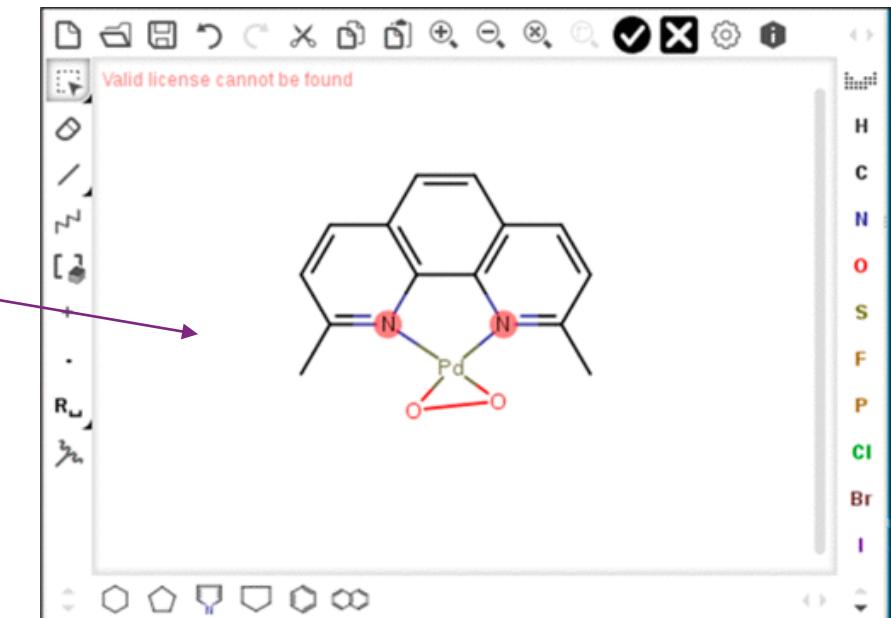
AND

Draw  Contains  Exact  Similar

SMILES, InChI or InChIKey

Search

A screenshot of a search interface. On the left, there's a sidebar with tabs for Navigation/Edition, Search, and Reports. Below these are search filters for Name, Description, Geometry (with options for Contains, Exact, or Similar), Elements (Select), Type (ALL), Path, Owner (Select), Group (Select), Creation dates (Start and End fields), and Concept group. Each filter has an AND dropdown menu to its right. A 'Draw' button is highlighted with a purple arrow pointing to a chemical drawing tool on the right. The 'Contains' radio button is selected. The 'Search' button at the bottom is blue.



# Generating Supporting Information

**ioChem-BD Create**

Navigation/Edition Search Reports Options **Browse**

New report

Supporting information

Description

Type

Creation Date

Reaction energy profile Biverdazyl project description Supporting information 2020-02-03

2 acetic project report Project description Supporting information 2020-02-03

Type to filter...

**ioChem-BD Create**

Navigation/Edition Search Reports 6 × Options **Browse**

**Report details**

**Supporting information**

Energy type:

Potential Energy

Default Units::

kcal/mol  kJ/mol  eV  Eh

File format::

PDF  RTF

**Report configuration**

**Report selected calculations**

**Calculations** Add Navigation selected calculations

O... Calculation absolute path Title Name Path

1 /db/slester/turbomole/dmabn/tdhf/dmabn	dmabn	x
2 /db/slester/turbomole/dmabn/tddft/freq/tddft	tddft	x
3 /db/slester/turbomole/dmabn/tddft/tdhf	tdhf	x
4 /db/slester/turbomole/dmabn/turbomole	turbomole	x

Generate Save Save and close

M06\_2x

Int\_1

Energy (POTENTIAL) = -40869.05584703487 eV

Atom	X	Y	Z
1	N	1.5745	11.9544
2	O	3.5569	10.1230
3	O	0.5052	9.5932
4	O	1.9285	11.8781
5	Cl	5.4954	8.2247
6	Cl	6.3066	12.2783
7	Cl	-1.9124	7.9845
8	Cl	-3.6789	10.2164
9	Cl	2.8772	13.2869
10	Cl	0.7126	17.6434
11	C	2.8167	12.7190
12	H	3.2253	13.0801
13	H	2.5604	13.5928
14	C	3.8543	11.8921
15	C	4.1557	10.6120
16	C	5.1226	9.8747
17	C	5.7885	10.3599
18	H	6.5311	9.7566
19	C	5.4651	11.6289
20	C	4.4996	12.3948
21	H	4.2457	13.3786
22	C	1.0903	11.2824
23	H	1.8361	10.5374



# Generating Reaction Energy Profile

**ioChem-BD Create**

Navigation/Edition Search Reports Favourable Martin wB97xD x Options **Browse**

**Id:** 520 **Creation date:** 2019-07-16

**Name:** Favourable Martin wB97xD

**Title:** Favorable Martin wB97xD without entropy corrections

**Description:** Favorable Martin wB97xD without entropy corrections

**Calculations:** Add Navigation selected calculations

Order	Calculation absolute path	Title	Name	Path
1	/db/amatoe/Favourable_Martin/NICOD	NICOD		x
2	/db/amatoe/Favourable_Martin/COD	COD		x
3	/db/amatoe/Favourable_Martin/Fluorobenzene	Fluorobenzene		x
4	/db/amatoe/Favourable_Martin/Cat	Cat		x
5	/db/amatoe/Favourable_Martin/TS1_bis	TS1_bis		x
6	/db/amatoe/Favourable_Martin/1Int_cis_bis	1Int_cis_bis		x
7	/db/amatoe/Favourable_Martin/phosphine	phosphine		x
8	/db/amatoe/Favourable_Martin/Int1_1_mono	Int1_1_mono		x
9	/db/amatoe/Favourable_Martin/TSiso1_mono	TSiso1_mono		x
10	/db/amatoe/Favourable_Martin/Int1_2_mono	Int1_2_mono		x
11	/db/amatoe/Favourable_Martin/adduct	adduct		x
12	/db/amatoe/Favourable_Martin/Prev_TS2_adduct_mono	Prev_TS2_adduct_mono		x
13	/db/amatoe/Favourable_Martin/TS2_adduct_mono	TS2_adduct_mono		x
14	/db/amatoe/Favourable_Martin/Post-TS2-adduct-mono	Post-TS2-adduct-mono		x
15	/db/amatoe/Favourable_Martin/NaFBOPh	NaFBOPh		x
16	/db/amatoe/Favourable_Martin/Int2_1mono	Int2_1mono		x
17	/db/amatoe/Favourable_Martin/TS3_mono	TS3_mono		x
18	/db/amatoe/Favourable_Martin/BnepBenzene	BnepBenzene		x

**Energy reaction profile**

**Energy type:**  Potential Energy  Gibbs Energy  Zero Point Energy Corrected  Enthalpy

**Default Units:**  kcal/mol  kJ/mol  eV  Eh

**File format:**  CHART

**Serie**

Name:	Step	Variables:
Serie	X+Y	R c3+c4
Steps:	c1+2*c7-(c4+2*c2)	X+Y
0	A+B	R c3+c4+c11
c5-R1	TSA+B	Variable Formula
c6-R1	E	Variable Formula
c8+c7-R1	EM	Variable Formula
c9+c7-R1	TSEM	
c10+c7-R1	FM	
c12+c7-R	GM	
C13+c7-R	TSGM	
c14+c7-R	HM	
c16+c7+c15-R	IM+Z+J	
c17+c7+c15-R	TSIM	
c18+c15+c4-R	L	
Formula	Label	
Formula	Label	

**Gibbs Energy (kcal/mol)**

\*Any modification done to this chart using Plotly Chart Studio won't be saved at ioChem-BD platform.

Series	Step	Delta											
X+Y													
A+B													
TSA+B													
E													
EM													
TSEM													
FM													
GM													
TSGM													
HM													
IM+Z+J													
TSIM													
L													
21													

**MORE VIDEOS**

**Generate** **Save** **Save and close**



# Publish Datasets into Browse

The figure illustrates the process of publishing datasets from a local environment into the ioChem-BD Browse system.

**Left Panel:** A screenshot of a local application interface showing a tree view of datasets under "Fe(III) Keggin structures". The right side shows detailed information for a selected dataset, including navigation, expand/collapse options, and publishing commands like "Publish" and "Generate report". A purple arrow points from this panel to the "Publish Selected Elements" dialog.

**Middle Panel:** The "Publish Selected Elements" dialog. It displays the project name "ni7" and the number of calculations (4). The "Dataset Type" is set to "Independent Dataset". The "Authors & Affiliations" section lists "Vladimir Fock research group" and "Zavala, Damion". The "Sponsoring Agencies / Institutions" section lists "Foo Bar organization". The "Discovery" section is empty. The "Publication Options" section contains two checkboxes: "Request a DOI for published project" (unchecked) and "Embargo published elements, content restricted only to reviewers" (unchecked). At the bottom are "Cancel" and "Publish Data" buttons.

**Right Panel:** The "Publication Resume" dialog, which displays a green header with a checkmark and the message "Publication Successful". It states "Project ni7 has been published." and shows "4 Calculations Published" and "DOI Not Requested". Below this, it lists the published project details: "ni7" and its URL "<https://test.iochem-bd.org:8443/browse/handle/100/473>". It also provides a "View" button. Further down, it lists the "Peer Review Access Link" for "ni7" and its URL "<https://test.iochem-bd.org:8443/browse/review-collection/100/473...>", with a "Copy" button. Finally, it includes a "Project Management" section for "ni7" with a "Modify properties or lift embargo" link and an "Open Editor" button.

**Bottom Left:** The date "29 January 2026" is displayed.



# Browse Published Datasets in the Community

**ioChem-BD Browse**

Search Browse  Search Browse Logged in as sharonlester@loch... Create

Foo Bar organization / Vladimir Fock research group

**dmabn** Collection home page

Share - copy and redistribute the material in any medium or format. Adapt - remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms.

This dataset derived results are published in:  
*Manuscript title: A Metal-Free Synthesis of N-Aryl Carbamates under Ambient Conditions*  
*Journal: Angew. Chem. Int. Ed.*  
*DOI: 10.1002/anie.201504956*

**ROOT (1)**  
 ▾ **tdft (1)**  
 ▾ **freq (1)**  
 ▾ **tdhf (1)**

View as  Tree  List Search:

Preview	Title	Issue Date	Author(s)	Program	Calculation type	Method
	turbomole	3-Feb-2020	Lester, Sharon	TURBOMOLE; 6.4	Restricted geometry optimization	DFT

Export .csv Discover

Author: Lester, Sharon (1)  
 Program name: TURBOMOLE (1)  
 Date issued: 2020 (1)  
 Calculation type: Geometry optimization Excited states (1), Geometry optimization Minimum Energy (1), Restricted geometry optimization (1), Single point Excited state (1)  
 Method: DFT (3), HF (1)

**ioChem-BD Browse** Search Browse Search Browse Logged in as sharonlester@loch... Create

Foo Bar organization / Vladimir Fock research group / dmabn

Geometry:

**Actions:**

View data Download geometry

**Files in This Item:**

File	Size	Format	Action
control	1.86 kB	Unknown	<span>Download</span>
output.cml	41.63 kB	Chemical Markup Language	<span>Download</span>

**Referenced by:**

*Manuscript title: A Metal-Free Synthesis of N-Aryl Carbamates under Ambient Conditions*  
*Journal: Angew. Chem. Int. Ed.*  
*DOI: 10.1002/anie.201504956*

**Code snippets:** </> AU HTML

**Metadata:**

Title:	/tdhf dmabn
Authors:	Lester, Sharon
Issue Date:	3-Feb-2020
Publisher:	Foo Bar organization

JSmol

## Conclusions

- ioChem-BD demonstrates an early large-scale infrastructure for managing computational chemistry data
- XML/CML enabled structured, validated, and transformable chemical data representations
- Highlights the importance of standardized, automated, and reusable computational data workflows

**However,**

- XML/CML is very verbose and complex, so few researchers adopted it
- Parsing and transforming XML data requires significant technical effort
- Difficult to integrate with modern data science and machine learning workflows
- ioChem-BD is not fully open-source, and some versions require paid licenses, which may limit adoption and transparency



# Thank You