



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,^{*,†,‡} C. de Graaf,^{‡,||} N. López,[†] F. Maseras,^{†,§} J. M. Poblet,[‡] and C. Bo^{*,†,‡}

[†]Institute of Chemical Research of Catalonia, ICIQ, Av. Països Catalans 16, 43007 Tarragona, Catalonia, Spain

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

[§]Department of Chemistry, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain

^{||}Catalan Institution for Research and Advanced Studies, ICREA, Passeig Lluís 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

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

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

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

Interoperable

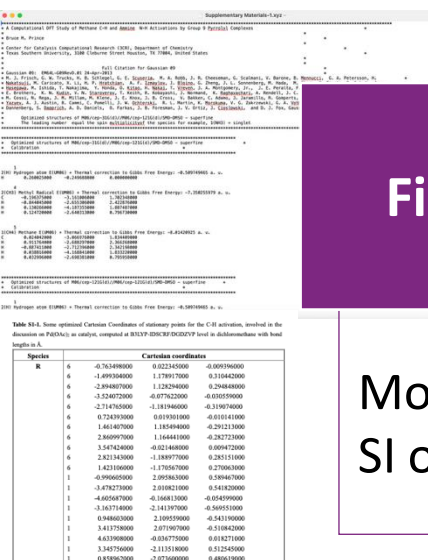
Data uses standard formats, allowing different systems and tools to understand and exchange it.

Heterogeneous formats, software-dependent raw outputs

Reusable

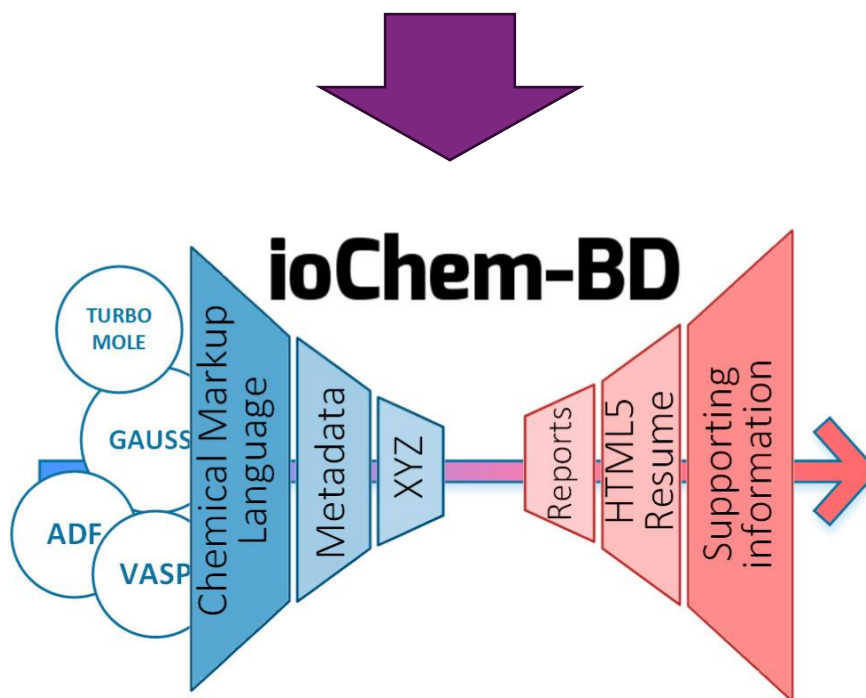
Data is well-documented with enough metadata to allow users to understand its origin, methodology, and any potential limitations.

Missing metadata, provenance, and standardized format



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



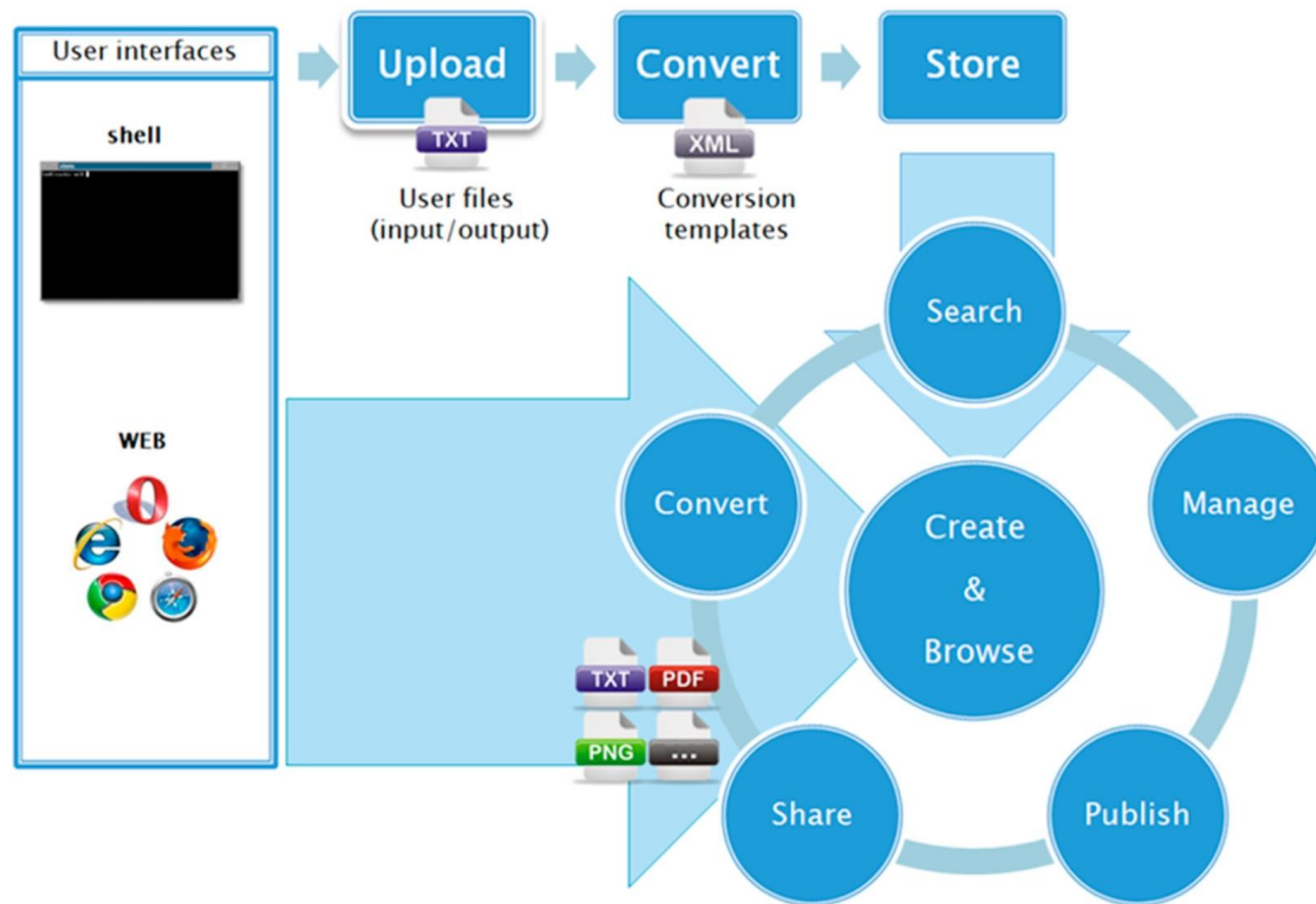
389,954

Items indexed



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

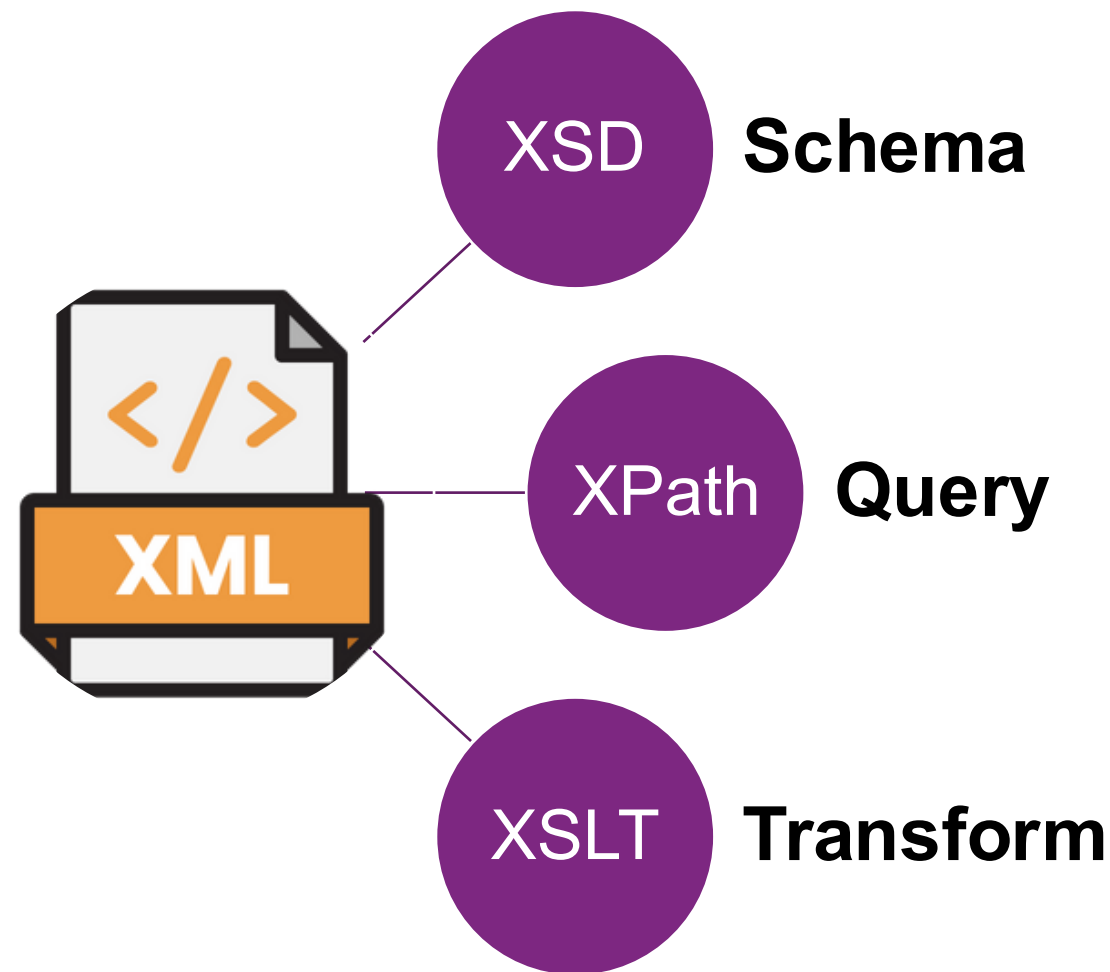
XSLT: rule-based transformation (XML → HTML/CSV/PDF)

- **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)

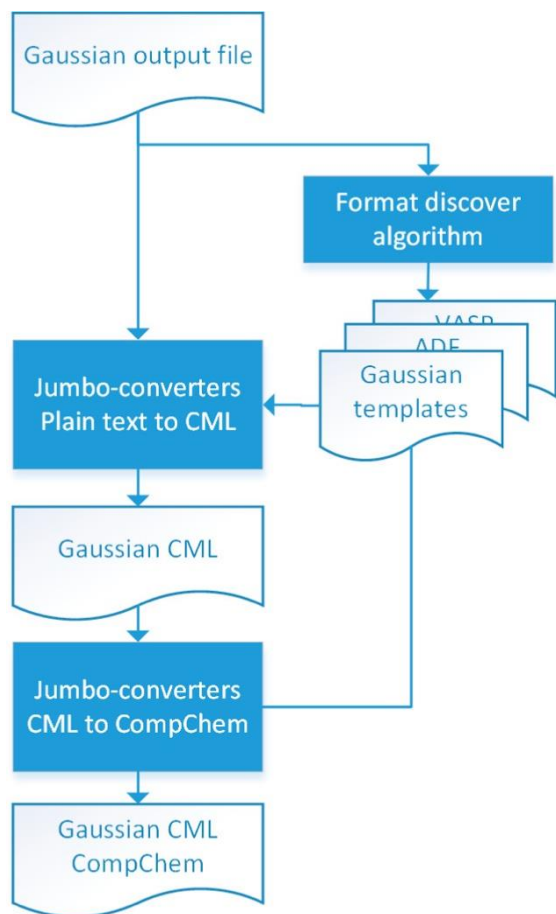
- CML extends XML with standardized chemical concepts (molecules, atoms, bonds, computational metadata)
- However, later replaced by simpler, programmatic workflows (JSON, Python, HDF5)
- **CML pipeline**
Quantum chemistry output → CML (XML) → XSLT → Reports / knowledge graphs / databases
- **Modern pipeline**
Quantum chemistry output → Python parser → ORM objects → SQL database

XML

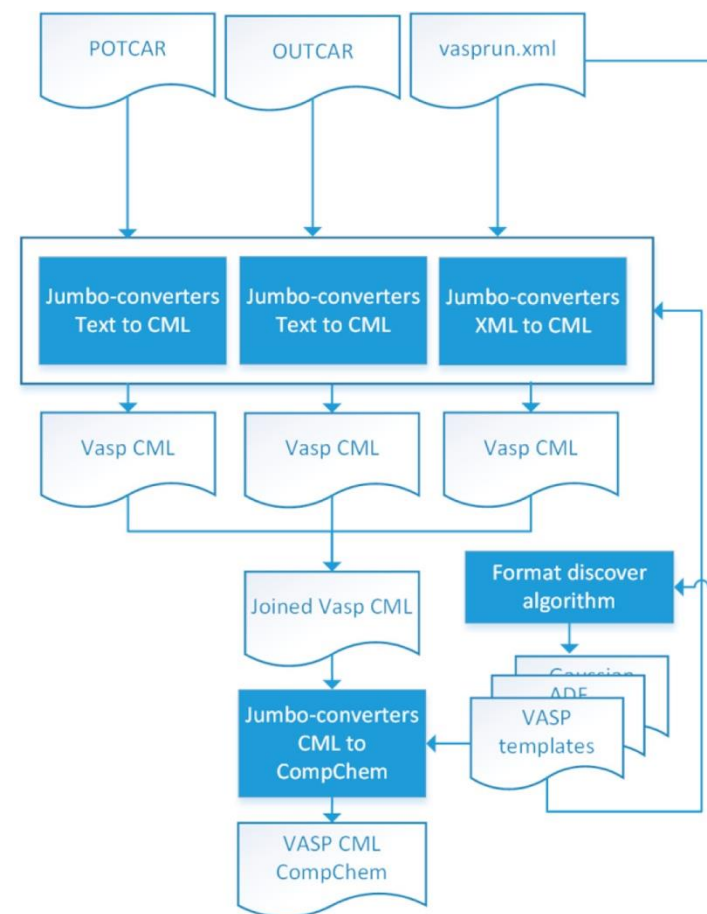
```
<molecule>
  <atomArray>
    <atom id="a1" elementType="C" x3="0.0" y3="0.0" z3="0.0"/>
  </atomArray>
</molecule>
```

```
<molecule cmlx:templateRef="mol"
  formalCharge="0"
  id="zmat"
  spinMultiplicity="1">
  <atomArray>
    <atom elementType="0"
      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/H2O/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

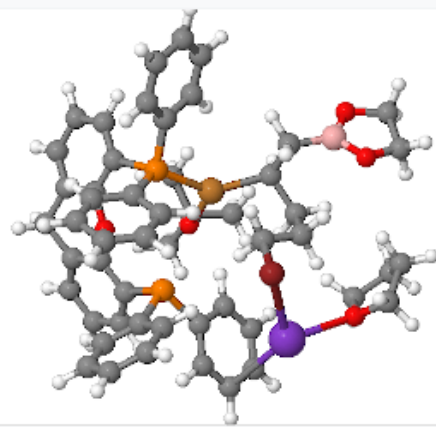
Options | Browse

Navigation/Edition | Search | Reports

Refresh

Type	Description	Creation Date	Handle	Pub. Edit
PRO	Fe(II) Keggin structures	2018-06-14 11:09		
PRO	-	2018-03-07 15:35		
PRO	Aromatic_Amination_of_Lactones	2018-06-14 11:07		
ADF	calc1	2018-01-16 17:10		
GAU	hexenol_modified	2018-04-12 18:18		
GAU	hexenol4	2018-04-12 18:55		
GAU	hexenol33	2018-04-12 18:40		
GAU	hexenol5	2018-04-12 19:08		
GAU	hexenol6	2018-04-12 19:18		
GAU	h2	2018-06-18 08:32		
GAU	calc	2018-06-20 17:31		
GAU	gaussian_calc	2016-10-24 16:03		
GAU	calc	2017-02-14 16:44		
GAU	calc1	2017-08-04 09:48		
GAU	K-Br-C4-1THF-I5	2017-12-23 10:57		
GAU	upload_smi	2017-12-23 10:59		
GAU	upload_smi2	2017-12-23 11:07		
GAU	upload_smi3	2017-12-23 11:08		
GAU	upload_smi4	2017-12-23 11:08		
PRO	Mo(I)_hydrogen_generation	2018-06-14 11:10		
GAU	cucurb	2017-09-04 11:48		
GAU	calc1_2	2017-09-06 11:06		
GAU	hexenol2	2018-04-12 18:33		
ADF	sample2	2018-10-15 19:15		
PRO	Ni_catalyzed_aryl_borylation_RM	2018-06-14 11:11		
ADF	adf	2015-07-13 19:12		
GAU	ts_bp-uff	2015-07-08 18:40		
GAU	g09	2015-07-09 19:13		
VSP	vasp_demo	2015-07-09 20:02		
ADF	geomopt2012	2015-07-09 20:03		
ADF	adf1	2015-07-13 20:31		
GAU	gaussian1	2015-07-13 20:32		
VSP	vasp1	2015-07-13 20:32		
VSP	opt10	2015-07-13 20:35		
ADF	freq	2015-07-13 20:39		
ADF	sample2	2018-10-15 19:15		

3D Structure | View Results | Download | RAW CML



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

Group

cbo_group

Permissions

rw---

Creation date

2017-12-23 10:57

Mod. date

2018-07-18 17:45

Pub. date

Create Project

Modify

Create Module Overview

ioChem-BD Create Options Browse

Navigation/Edition Search Reports 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		
calc	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 10:59		
upload_sni	GAU	uploadsn1	2017-12-23 10:59		
upload_sni2	GAU	uploadsn2	2017-12-23 11:07		
upload_sni3	GAU	uploadsn3	2017-12-23 11:08		
upload_sni4	GAU	uploadsn4	2017-12-23 11:08		
Mo(II)_hydrogen_generation	PRO	Mo(II)_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		
q09	GAU	q09	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		

Item actions

Navigation frame

Item details

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: Concept Group Permissions: rwx


Creation date: 2017-12-23 10:59 Mod. date: 2018-01-16 17:45 File date:

Create Project Modify Delete

Uploading to Create

Properties

Path	/db/testuser		
Name	acetate	Type	
Description	Acetate molecules	State	
		Concept Group	
Owner		Group	
		Permissions	rwr—
Creation date		Mod. date	
		Pub. date	

Create Project Modify 

Creating a project

Upload Calculations

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

Item Actions frame

3D Structure
View Results
Download
RAW CML

output.cml.html

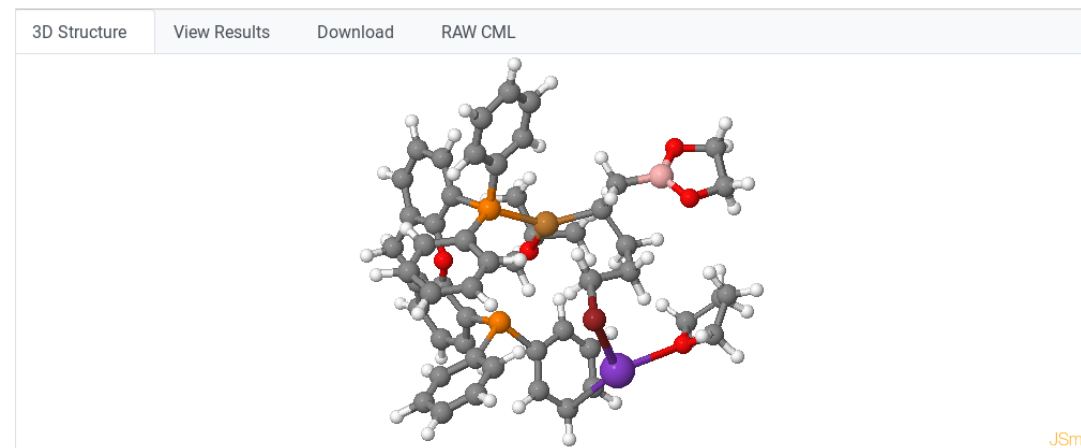
GENERAL INFO

Title: upload_sni
Program: Gaussian 09 ES64L-G09RevD.01
Formula: C52H58BrCuO5P2K
Calculation type: Geometry optimization Minimum
Method(s): Rwb97XD - Grimme-D2
Temperature: 298.150 K
Pressure: 1.00000 atm

ATOM INFO



Atomic coordinates [Å] (optimized)

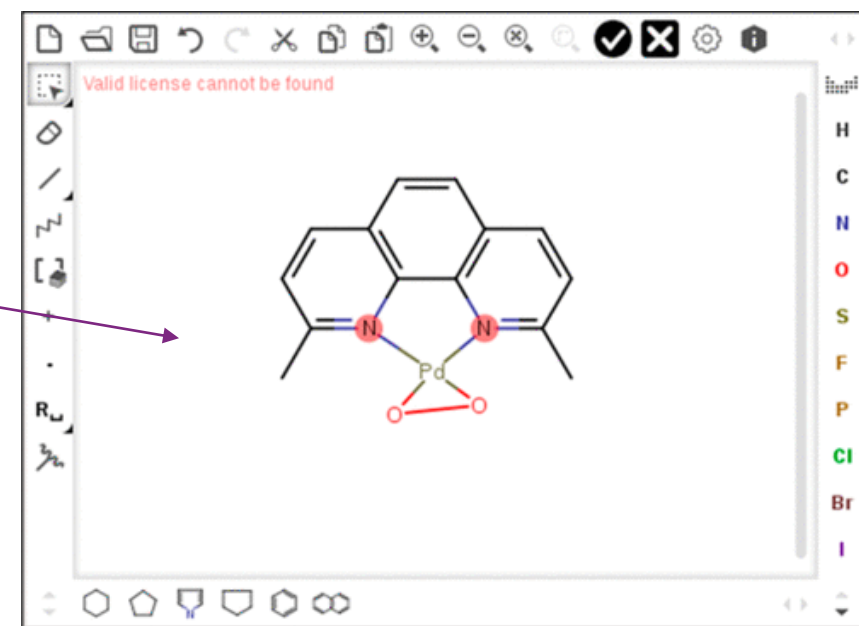
ATOM		x	y	z	TYPE	Core	ECP
1	Cu	0.2319	1.4549	-0.6657	lanl2dz	+	+
2	C	3.2566	-0.0165	0.6511	6-31g(d)		
3	P	-1.4161	-1.9130	0.0031	lanl2dz	+	+
4	P	-1.6259	2.1612	0.6160	lanl2dz	+	+
5	C	-2.3341	1.2641	2.0654	6-31g(d)		
6	C	-3.2503	-0.2076	4.2710	6-31g(d)		
7	C	-1.4320	0.6455	2.9376	6-31g(d)		
8	C	-2.7032	1.1463	2.2214	6-31g(d)		



3D Structure	View Results	Download	RAW CML
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	<input type="text"/>	AND ▾
Description	<input type="text"/>	AND ▾
Geometry	<input checked="" type="radio"/> Contains <input type="radio"/> Exact <input type="radio"/> Similar Draw <input type="text" value="SMILES, InChI or InChIKey"/>	AND ▾
Elements	Select	AND ▾
Type	ALL ▾	AND ▾
Path	<input type="text"/>	AND ▾
Owner	Select	AND ▾
Group	Select	AND ▾
Creation dates	Start <input type="text"/>  End <input type="text"/> 	AND ▾
Concept group	<input type="text"/>	AND ▾
<input type="button" value="Search"/>		



Generating Supporting Information

ioChem-BD Create Options Browse

Navigation/Edition Search Reports

New report

Supporting information
Reaction energy profile
Biverdazyl project report
2 acetic project report

Type to filter...

Description	Type	Creation Date
Biverdazyl project description	Supporting information	2020-02-03
Project description	Supporting information	2020-02-03

ioChem-BD Create Options Browse

Navigation/Edition Search Reports 6 X

Id: 6 **Creation date:** 2020-02-03

Name: dmabn dataset supporting information

Title: dmabn dataset supporting information

Description: Description about dmabn report

Report details

Supporting information

Energy type:
☒ Potential Energy
☐ kcal/mol ☐ kJ/mol ☐ eV ☐ Eh

Default Units::
☐ kcal/mol ☐ kJ/mol ☒ eV ☐ Eh

File format::
☒ PDF ☐ RTF

Report configuration

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

Report selected calculations

Generate Save Save and close

M06_2x

Int_1
 Energy (POTENTIAL) = -40869.05584703487 eV

	Atom	X	Y	Z
1	N	1.5745	11.9544	0.2578
2	O	3.5569	10.1230	0.8174
3	O	0.5052	9.5932	1.2384
4	O	1.9285	11.8781	2.9799
5	Cl	5.4954	8.2247	-0.4089
6	Cl	6.3066	12.2783	-3.9360
7	Cl	-1.9124	7.9845	1.7676
8	Cl	-3.6789	10.2164	-2.8700
9	Cl	2.8772	13.2869	5.3967
10	Cl	0.7126	17.6434	2.9467
11	C	2.8167	12.7190	-0.0468
12	H	3.2253	13.0801	0.9017
13	H	2.5604	13.5928	-0.6570
14	C	3.8543	11.8921	-0.7534
15	C	4.1557	10.6120	-0.2450
16	C	5.1226	9.8747	-0.9478
17	C	5.7885	10.3599	-2.0604
18	H	6.5311	9.7566	-2.5657
19	C	5.4651	11.6289	-2.5051
20	C	4.4996	12.3948	-1.8763
21	H	4.2457	13.3786	-2.2532
22	C	1.0903	11.2824	-0.9774
23	H	1.8361	10.5374	-1.2666

Generating Reaction Energy Profile

ioChem-BD Create

Options | Browse

Navigation/Edition | Search | Reports | Favourable Martin wB97xD ✕

Id: 520

Creation date: 2019-07-16

Name: Favourable Martin wB97xD

Title: Favourable Martin wB97xD without entropy corrections

Description: Favourable Martin wB97xD without entropy corrections

Calculations:

Add Navigation selected calculations

Order	Calculation absolute path	Title	Name	Path	
1	/db/amateo/Favourable_Martin/NICOD	NICOD			✕
2	/db/amateo/Favourable_Martin/COD	COD			✕
3	/db/amateo/Favourable_Martin/Fluorobenzene	Fluorobenzene			✕
4	/db/amateo/Favourable_Martin/Cat	Cat			✕
5	/db/amateo/Favourable_Martin/TS1_bis	TS1_bis			✕
6	/db/amateo/Favourable_Martin/1Int-cis_bis	1Int-cis_bis			✕
7	/db/amateo/Favourable_Martin/phosphine	phosphine			✕
8	/db/amateo/Favourable_Martin/Int1_1_mono	Int1_1_mono			✕
9	/db/amateo/Favourable_Martin/TS-iso1_mono	TS-iso1_mono			✕
10	/db/amateo/Favourable_Martin/Int1_2_mono	Int1_2_mono			✕
11	/db/amateo/Favourable_Martin/adduct	adduct			✕
12	/db/amateo/Favourable_Martin/Prev_TS2_adduct_mono	Prev_TS2_adduct_mono			✕
13	/db/amateo/Favourable_Martin/TS2_adduct_mono	TS2_adduct_mono			✕
14	/db/amateo/Favourable_Martin/Post-TS2-adduct-mono	Post-TS2-adduct-mono			✕
15	/db/amateo/Favourable_Martin/NaFBOPh	NaFBOPh			✕
16	/db/amateo/Favourable_Martin/Int2_1mono	Int2_1mono			✕
17	/db/amateo/Favourable_Martin/TS3_mono	TS3_mono			✕
18	/db/amateo/Favourable_Martin/BnepBenzene	BnepBenzene			✕

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

Series

Name: Serie

Steps:

c1+2*c7-(c4+2*c2)

X+Y

0

A+B

c5-R1

TSA+B

c6-R1

E

c8+c7-R1

EM

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

Variables:

R1

c3+c4

R

c3+c4+c11

Variable

Formula

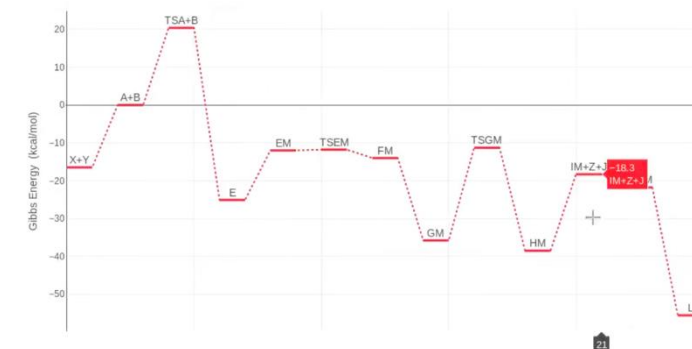
Variable

Formula

Generate

Save

Save and close



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

Series	Step	Step	Step	Step	Step	Step	Step	Step	Step	Step	Step	Step	Step	Delta
	X+Y	A+B	TSA+B	E	EM	TSEM	FM	GM	TSGM	HM	IM+Z+J	TSIM	L	
Serie	-16.5	0	20.4	-25.1	-12	-11.8	-14	-35.8	-11.3	-38.5	-18.3	-21.8	-55.6	36.9

Publish Datasets into Browse

Fe(III) Keggin structures	PRO	Fe(III) Keggin structures
Navigation	RO	-
Expand all	RO	sample
Collapse all	RO	Mo(I)_hydrogen_generation
Search from here...	RO	Ni_catalyzed_aryl_borylation_RM
Select child elements	AU	sample2
Browse module	AU	-
Publish	AU	calc
Generate report	AU	calc1
	AU	K-Br-C4-1THF-I5
	AU	uploadsni
	AU	uploadsni2
	AU	uploadsni3
	AU	uploadsni4
	AU	h2-bo

1. DATASET TYPE

Independent Dataset **Linked to Manuscript**

DOI (Digital Object Identifier)

10.1021/ci500593j

Article Title

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

Journal

Journal of Chemical Information and Computer Science

☐ Manuscript is a pre-print (not yet published)

Publish Selected Elements

Publish project **ni7** (4 calculations) to:

Vladimir Fock research group

1. DATASET TYPE

Independent Dataset **Linked to Manuscript**

2. AUTHORS & AFFILIATIONS

Type to search or add, then click Add

Zavala, Damion

3. SPONSORING AGENCIES / INSTITUTIONS

Enter sponsoring agency or institution name, then click Add

Foo Bar organization

PROJECT CODE

Enter project code (optional)

4. DISCOVERY

Type to search or add, then click Add

5. PUBLICATION OPTIONS

☐ Request a DOI for published project

☐ Embargo published elements, content restricted only to reviewers.

Cancel Publish Data

Publication Resume

Publication Successful

Project **ni7** has been published.

4 Calculations Published DOI Not Requested

VIEW PUBLISHED PROJECT

ni7

<https://test.iochem-bd.org:8443/browse/handle/100/473> View

PEER REVIEW ACCESS LINK

ni7

<https://test.iochem-bd.org:8443/browse/review-collection/100/473...> Copy

PROJECT MANAGEMENT

ni7

Modify properties or lift embargo Open Editor

All administrative actions and links can be accessed anytime by clicking on the edit icon placed on the right side of each published collection.

Close Resume

Browse Published Datasets in the Community

ioChem-BD Browse Search Browse Browse Logged in as sharonlester@ioch... 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)
 ▼ tddft (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

Program name: TURBOMOLE

Date issued: 2020

Calculation type:

- Geometry optimization Excited s...
- Geometry optimization Minimum Ex...
- Restricted geometry optimization
- Single point Excited state

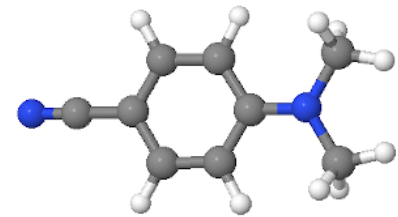
Method:

- DFT
- HF

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

Foo Bar organization / Vladimir Fock research group / dmabn

Geometry:



Actions:

View data

Download geometry

Files in This Item:

File	Size	Format
control	1.86 kB	Unknown
output.cml	41.63 kB	Chemical Markup Language

Referred 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

HTML AU

Metadata:

Title: /tdhf dmabn

Authors: Lester, Sharon

Issue Date: 3-Feb-2020

Publisher: Foo Bar organization



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