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

J. Chem. Inf. Model. **2015**, 55, 1, 95–103

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

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 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

J. Chem. Inf. Model. **2015**, 55, 1, 95–103

FAIR Challenges in Computational Chemistry Data

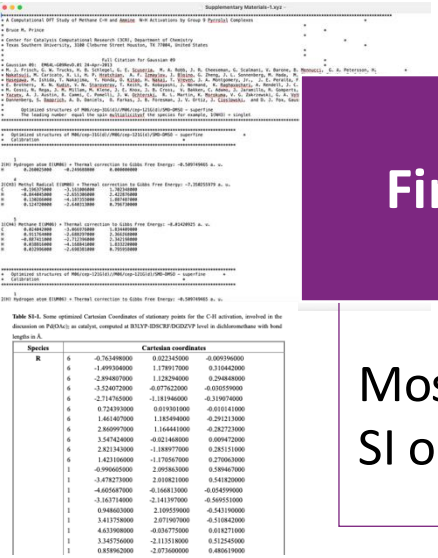


Table S1-1. Some optimized Cartesian Coordinates of stationary points for the C-H activation, involved in the discussion on $Pd(OAc)_2$ as catalyst, computed at B3LYP-IDSCRF/DGDZVP level in dichloromethane with bond lengths in Å.

Species		Cartesian coordinates		
R	6	-0.763498000	0.022345000	-0.009396000
	6	-1.499304000	1.178917000	0.310442000
	6	-2.894807000	1.128294000	0.294868000
	6	-3.524072000	-0.077622000	-0.030559000
	6	-2.714765000	-1.181946000	-0.319074000
	6	0.724393000	0.019301000	-0.010141000
	6	1.461407000	1.185494000	-0.291213000
	6	2.860997000	1.164441000	-0.282723000
	6	3.547424000	-0.021468000	0.009472000
	6	2.821343000	-1.188977000	0.285151000
	6	1.423106000	-1.170567000	0.270063000
	1	-0.996605000	2.095983000	0.589467000
	1	-3.478273000	2.010821000	0.541820000
	1	-4.605687000	-0.166813000	-0.054599000
	1	-3.163714000	-2.141397000	-0.569551000
	1	0.948603000	2.109559000	-0.543190000
	1	3.413758000	2.071907000	-0.510842000
	1	4.633908000	-0.036775000	0.018271000
	1	3.345756000	-2.113518000	0.512545000
	1	0.858962000	-2.073600000	0.480619000

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

DATA AVAILABILITY STATEMENT

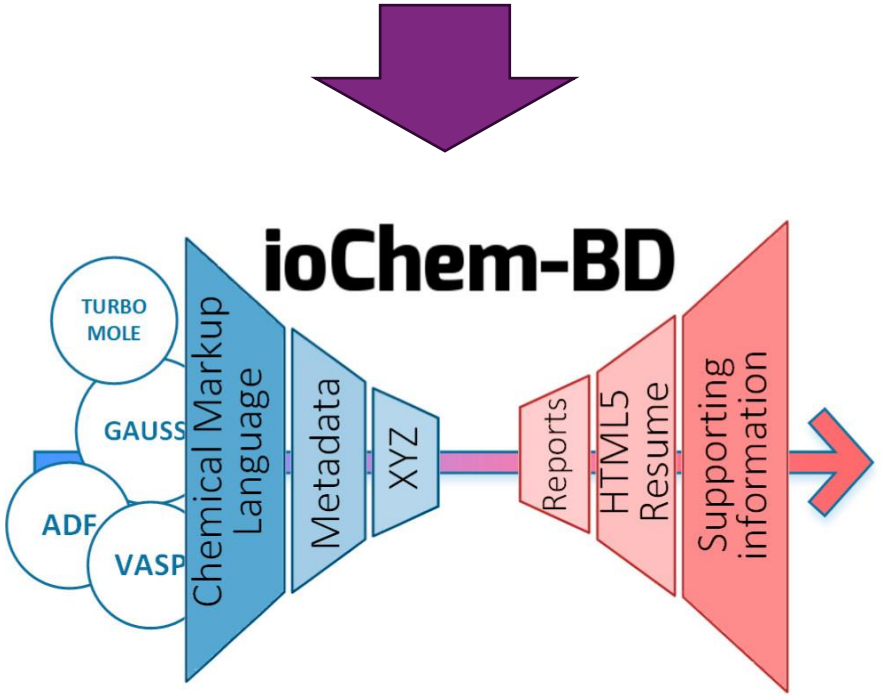
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Research data for this article

Data not available / Data will be made available on request

Motivation

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



ioChem-BD

Search by | Text | Enter search term | Find

Sign up | Login | Download | Docs

Open your research to the world

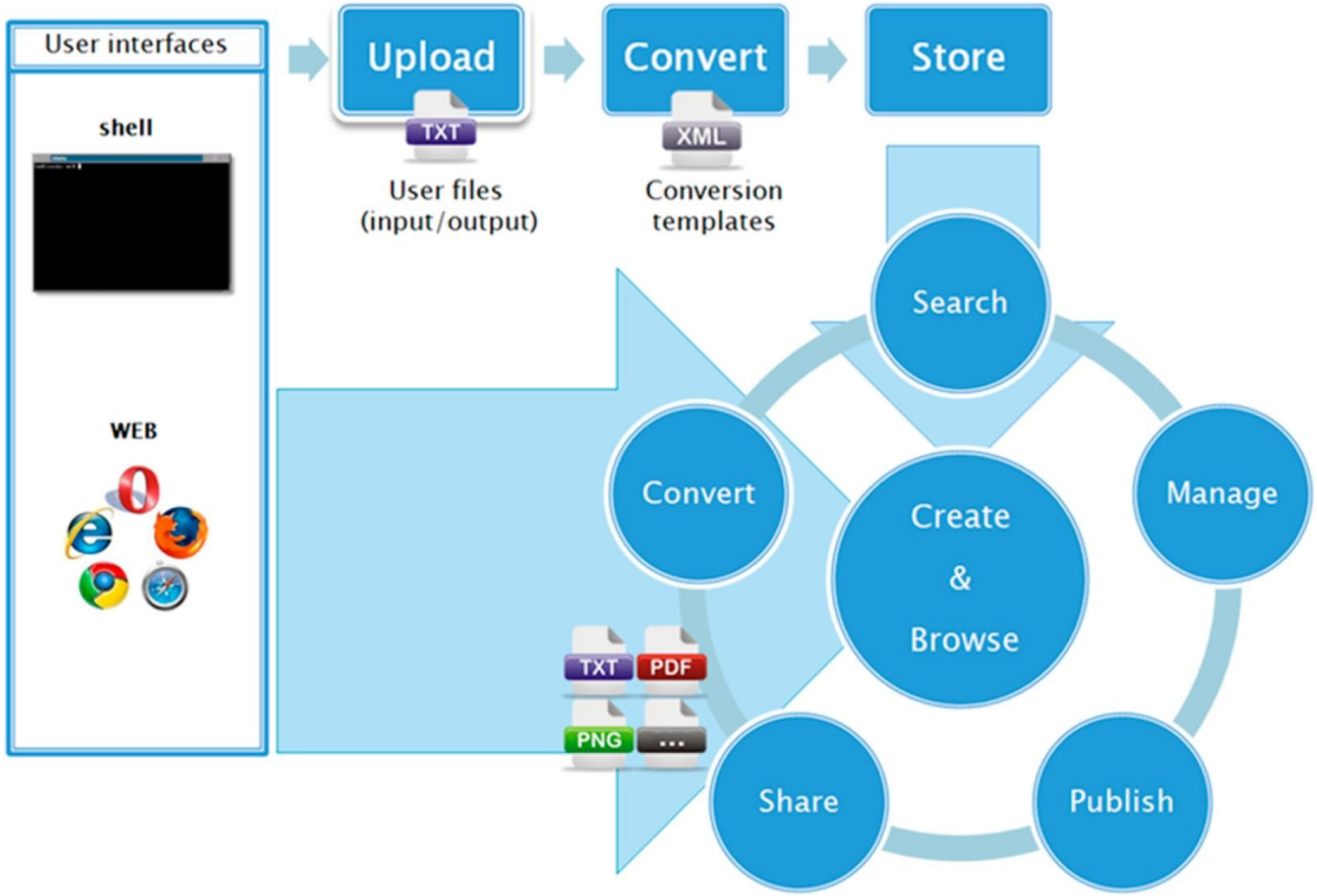
ioChem-BD - The Computational Chemistry Results Repository

1	7	1,006	389,954
Central service	Connected nodes	Collections available	Items indexed

OSCARS
Open Science Clusters' Action for Research & Society

Funded by the European Union

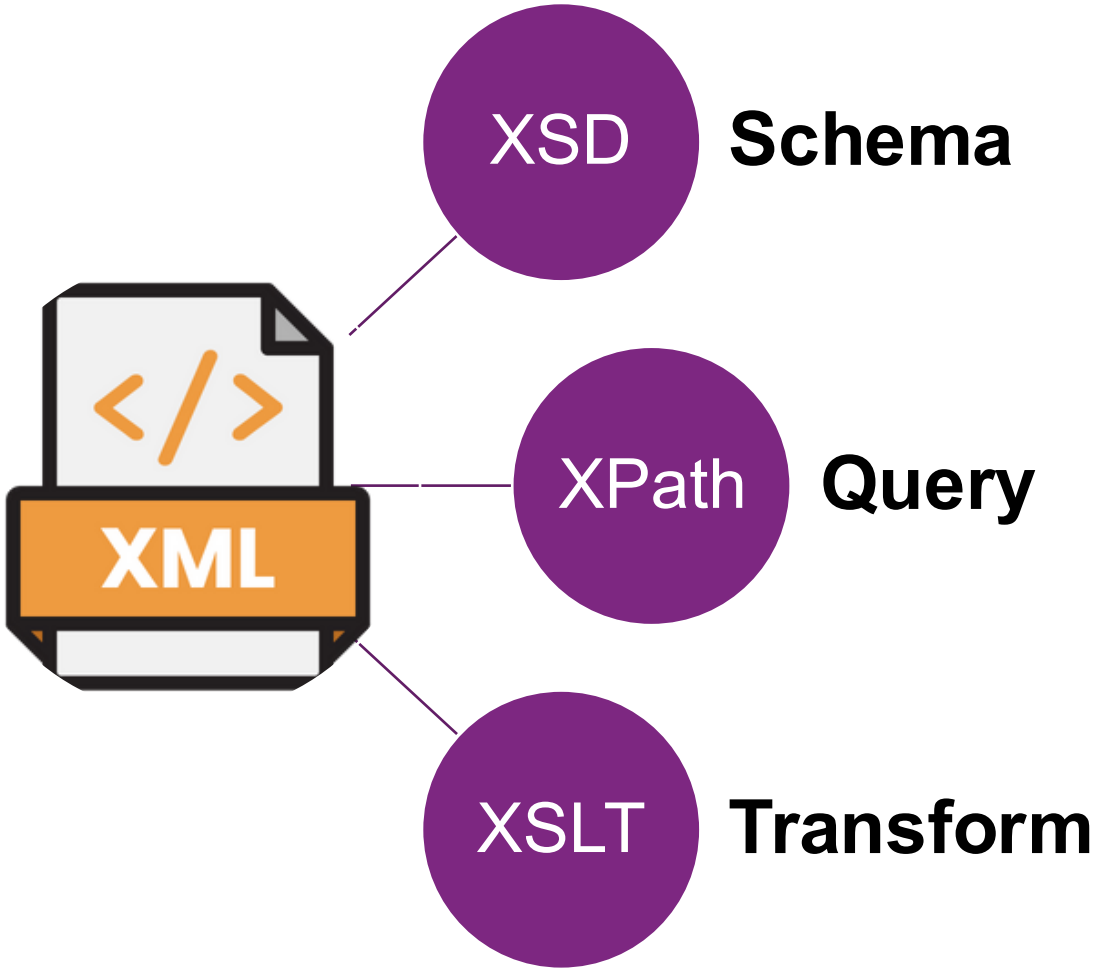
ioChem-BD System Overview



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**

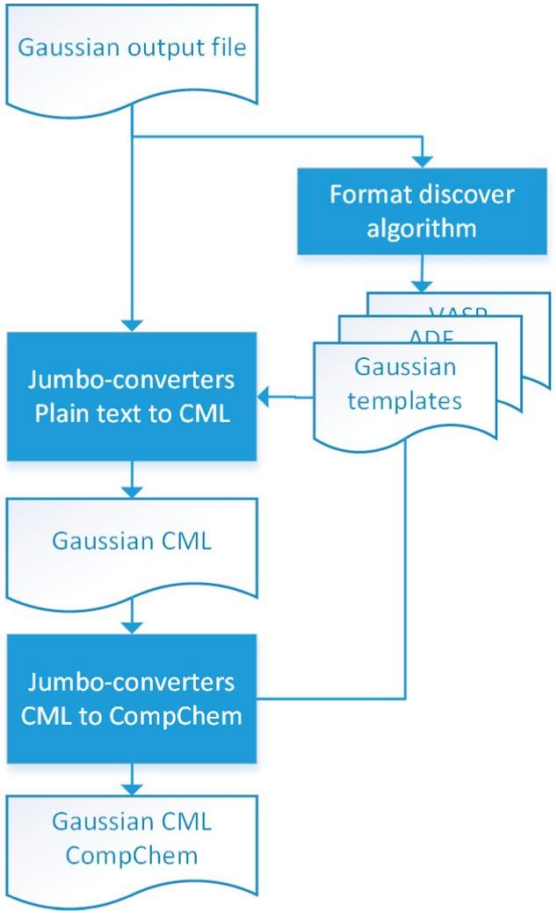
  Quantum chemistry output → Python parser → ORM objects → SQL database

XML

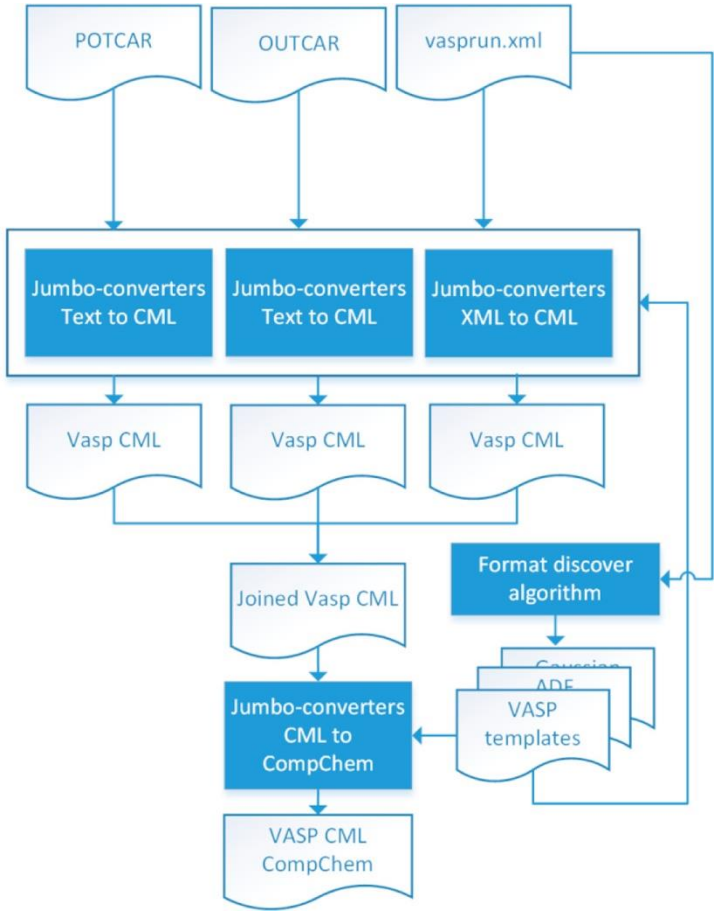
```
<molecule>
  <atomArray>
    <atom id="a1" elementType="C" x3="0.0" y3="0.0" z3="0.0"/>
  </atomArray>
</molecule>
```

```
<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/H2O/h1H2">
      <scalar dataType="xsd:integer" id="auxInfo">AuxInfo=1/0/N:1/rA:3OHH/
rB:s1;s1;/rC:;;;</scalar>
   </formula>
</molecule>
```

Conversion Workflow from Output Files to CML

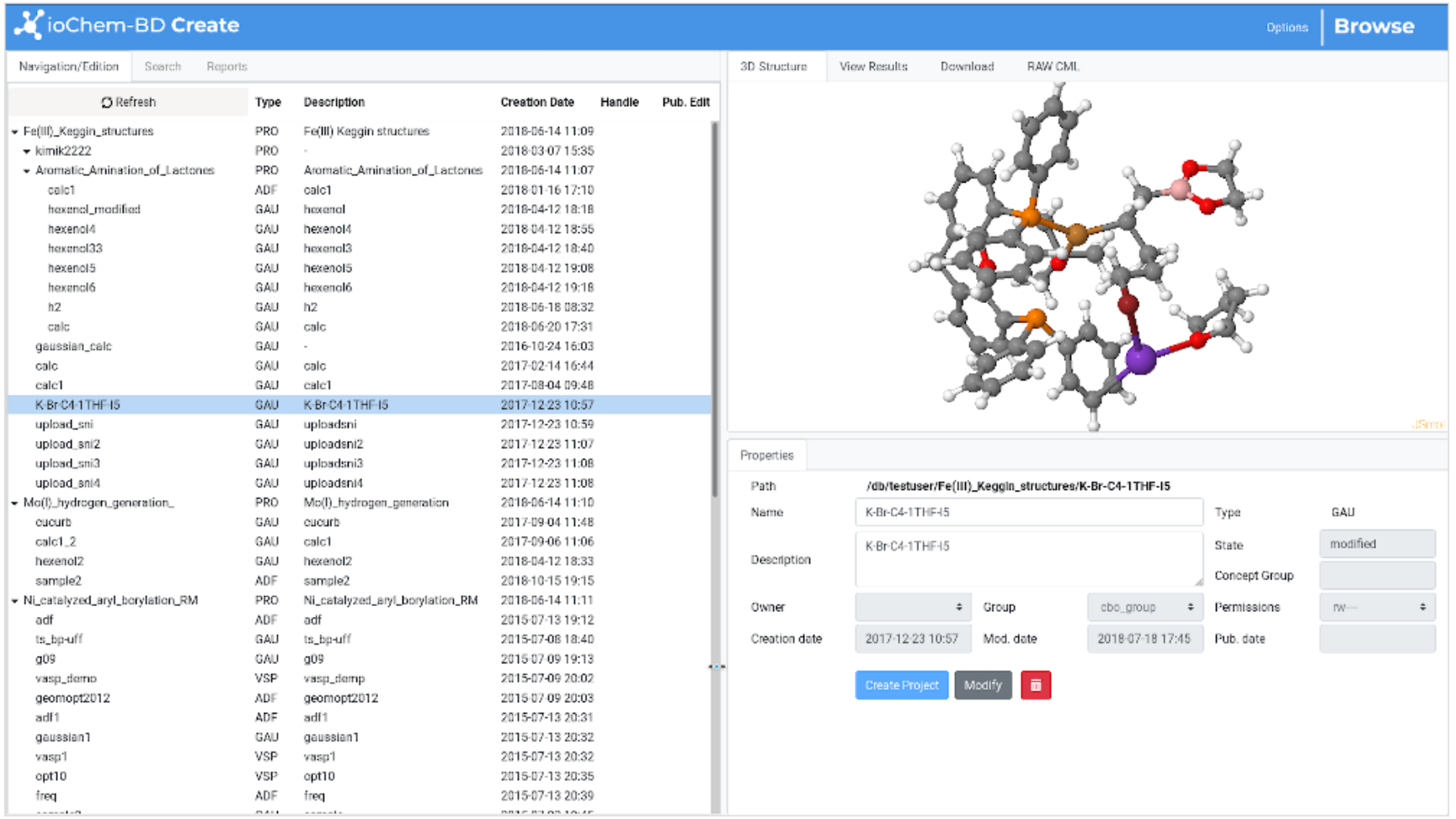


Gaussian and ADF: single output file

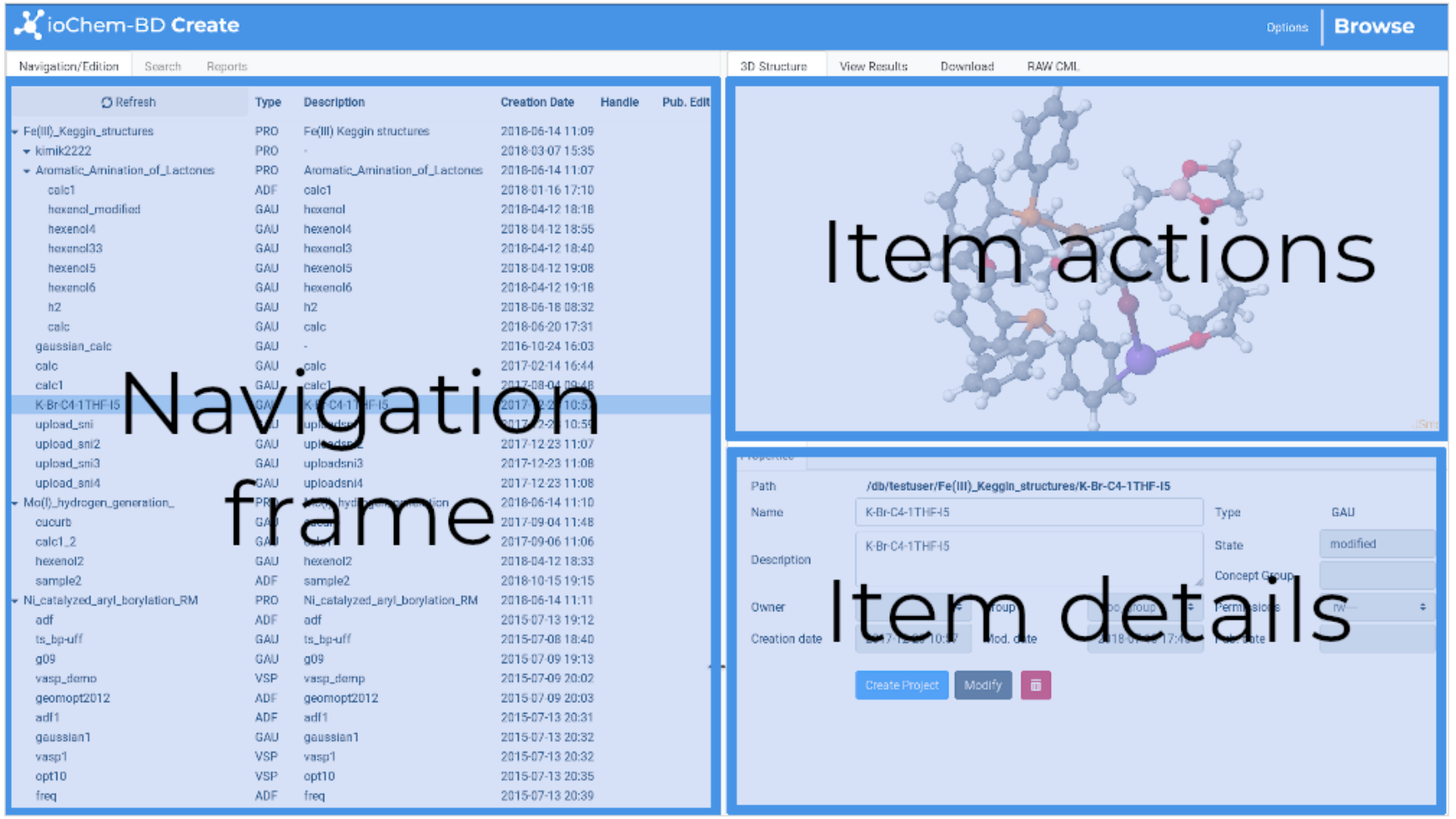


VASP: a group of output files

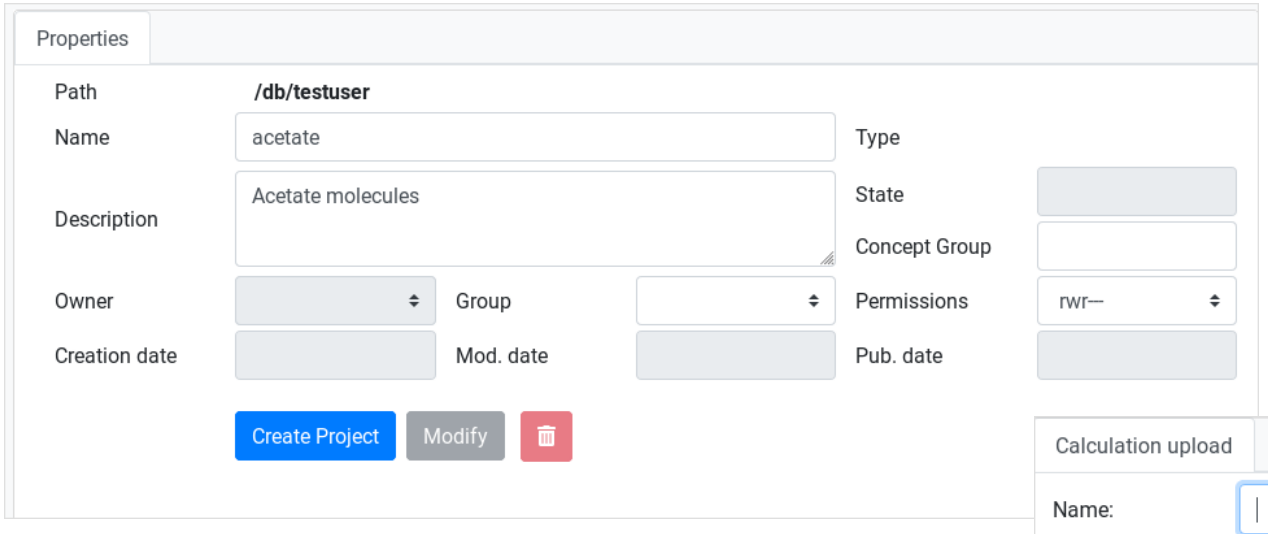
Create Module Overview



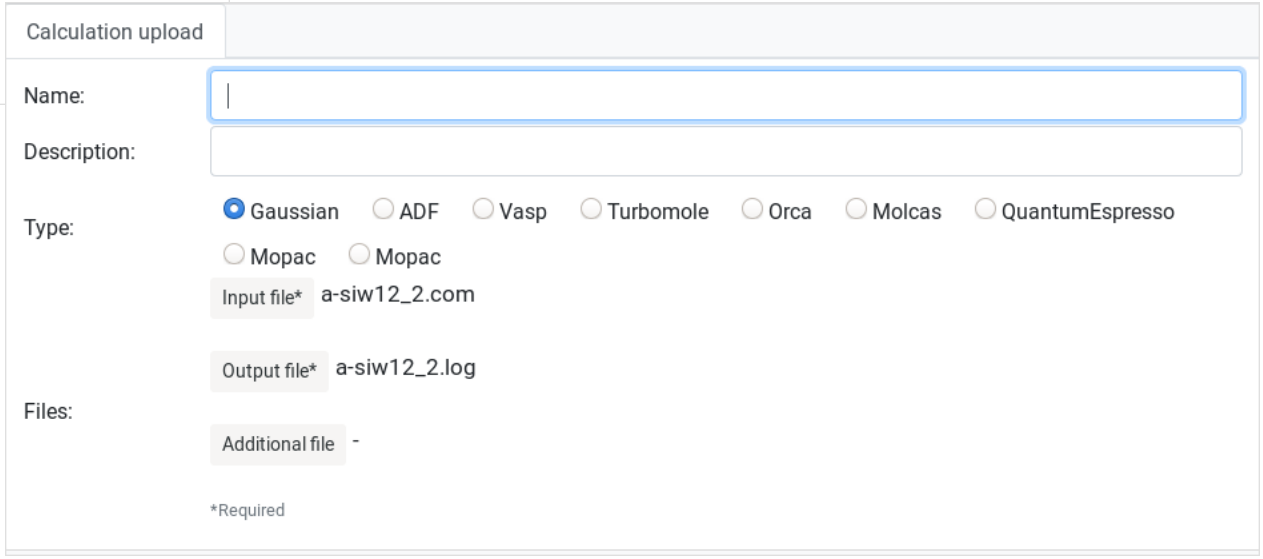
Create Module Overview



Uploading to Create



Creating a project



Upload Calculations

Item Actions frame

3D Structure | View Results | Download | RAW CML

output.cml.html

GENERAL INFO

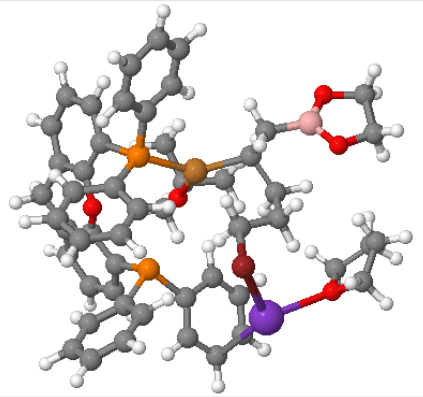
Title:	upload_sni
Program:	Gaussian 09 ES64L-G09RevD.01
Formula:	C52H58BBrCuO5P2K
Calculation type:	Geometry optimization Minimum
Method(s):	RwB97XD - Grimme-D2
Temperature	298.150 K
Pressure	1.00000 atm

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)		

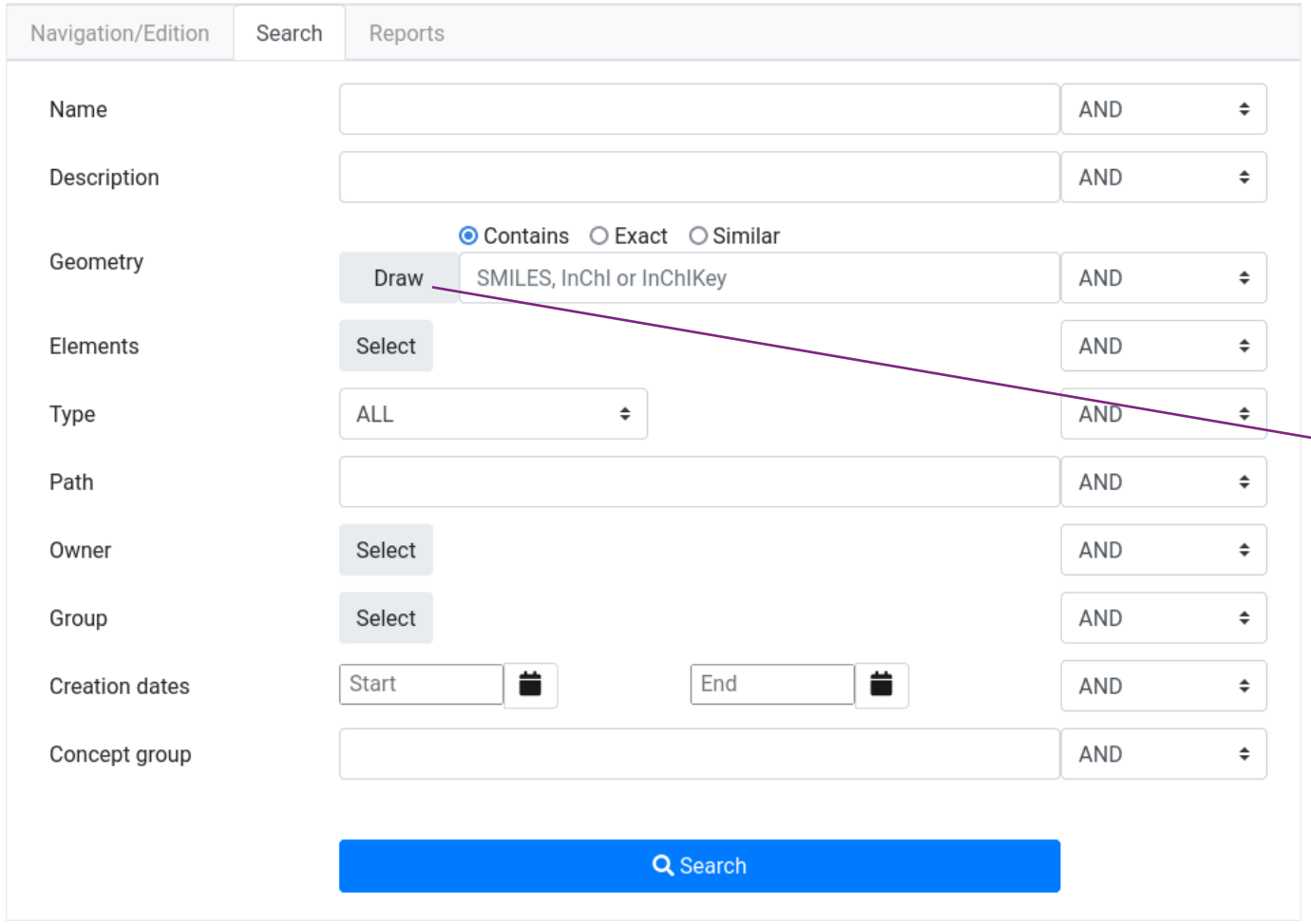
3D Structure | View Results | Download | RAW CML

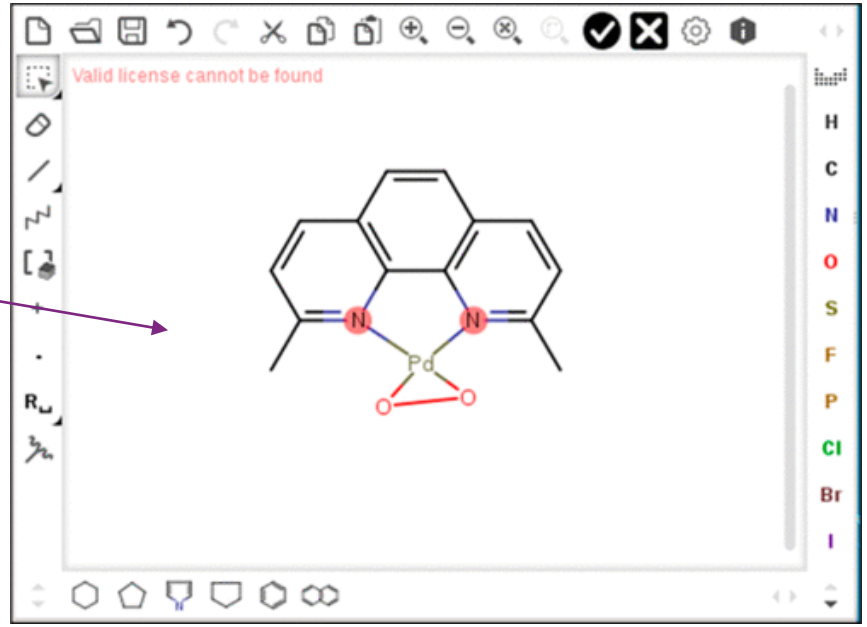


3D Structure | View Results | Download | RAW CML

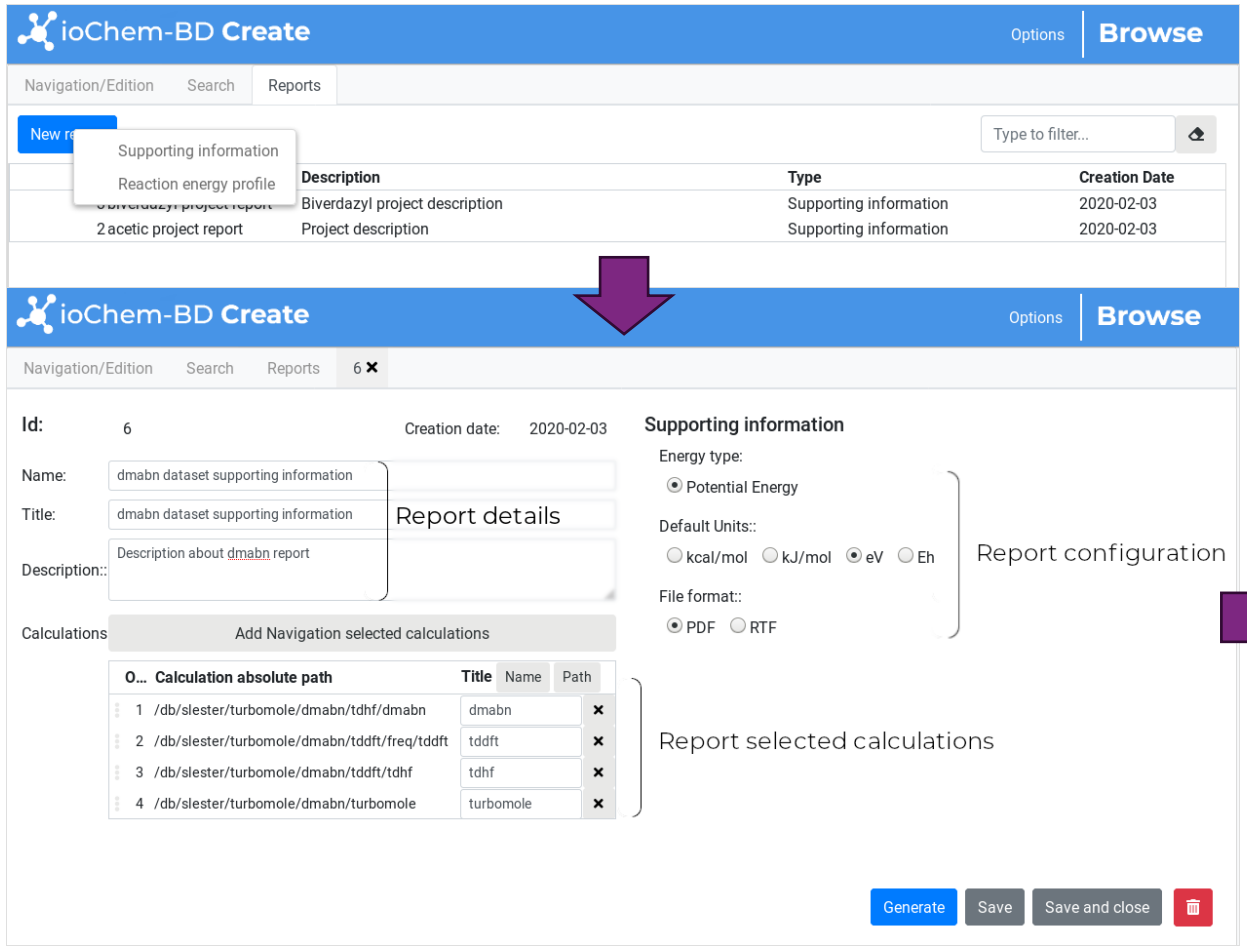
	File name	Mimetype	Size (kB)
Download	output.cml	chemical/x-cml	7,364.14
Download	K-Br-C4-1THF-I5.in	chemical/x-gaussian-input	7.29

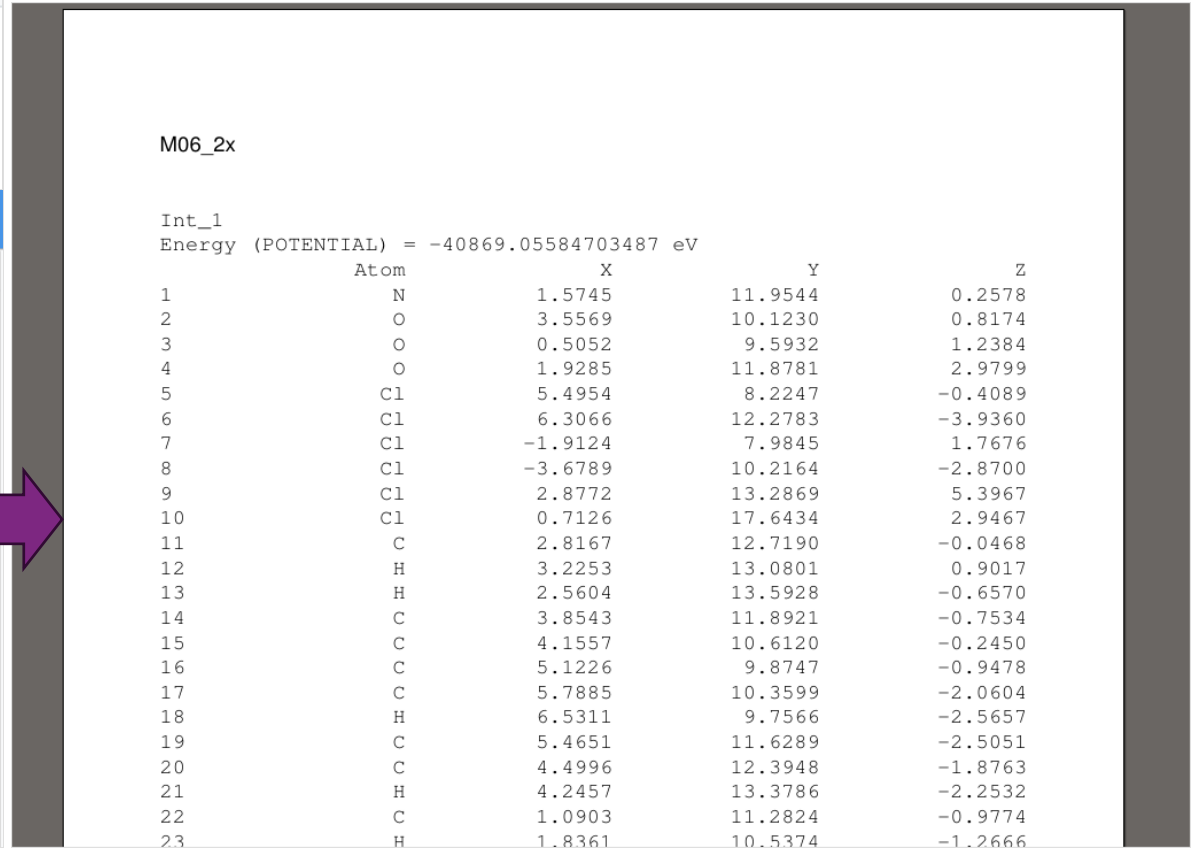
Search Mode


Navigation/Edition
Search
Reports
Name
AND
Description
AND
Geometry
Contains
Exact
Similar
Draw
SMILES, InChI or InChIKey
AND
Elements
Select
AND
Type
ALL
AND
Path
AND
Owner
Select
AND
Group
Select
AND
Creation dates
Start
End
AND
Concept group
AND
Search


Valid license cannot be found
N
N
Pd
O
O
H
C
N
O
S
F
P
Cl
Br
I

Generating Supporting Information





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: 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/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		

MORE VIDEOS

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

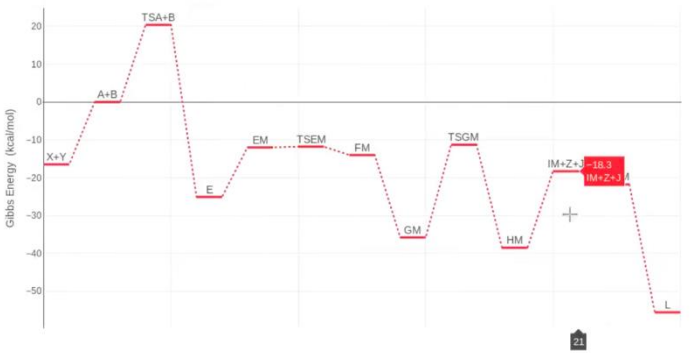
Serie

Name: 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
Variable	Formula

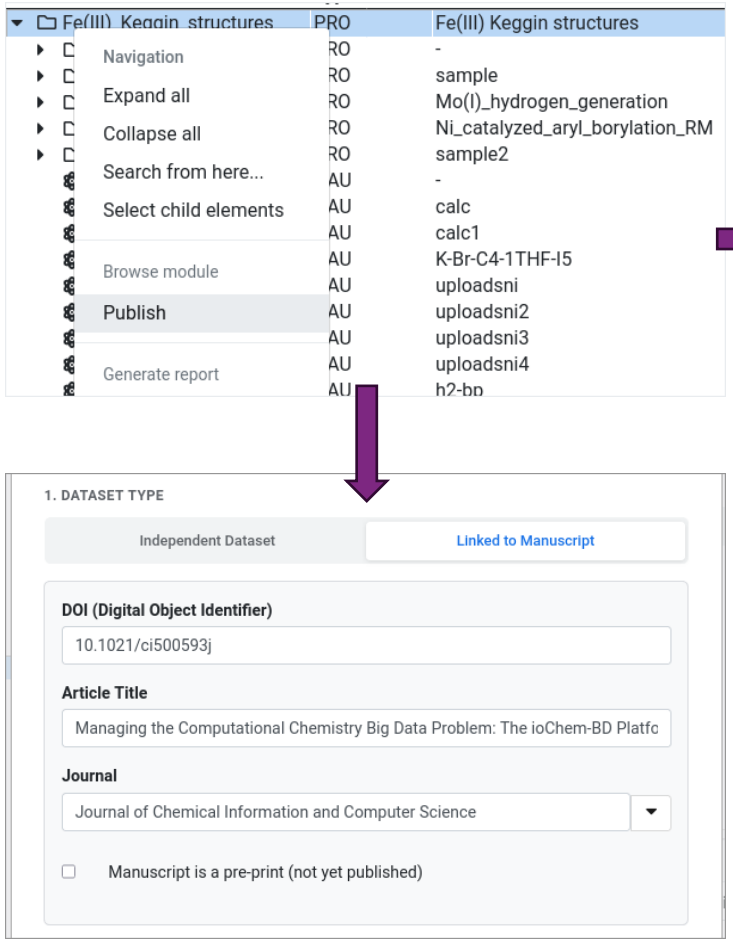
Generate | Save | Save and close

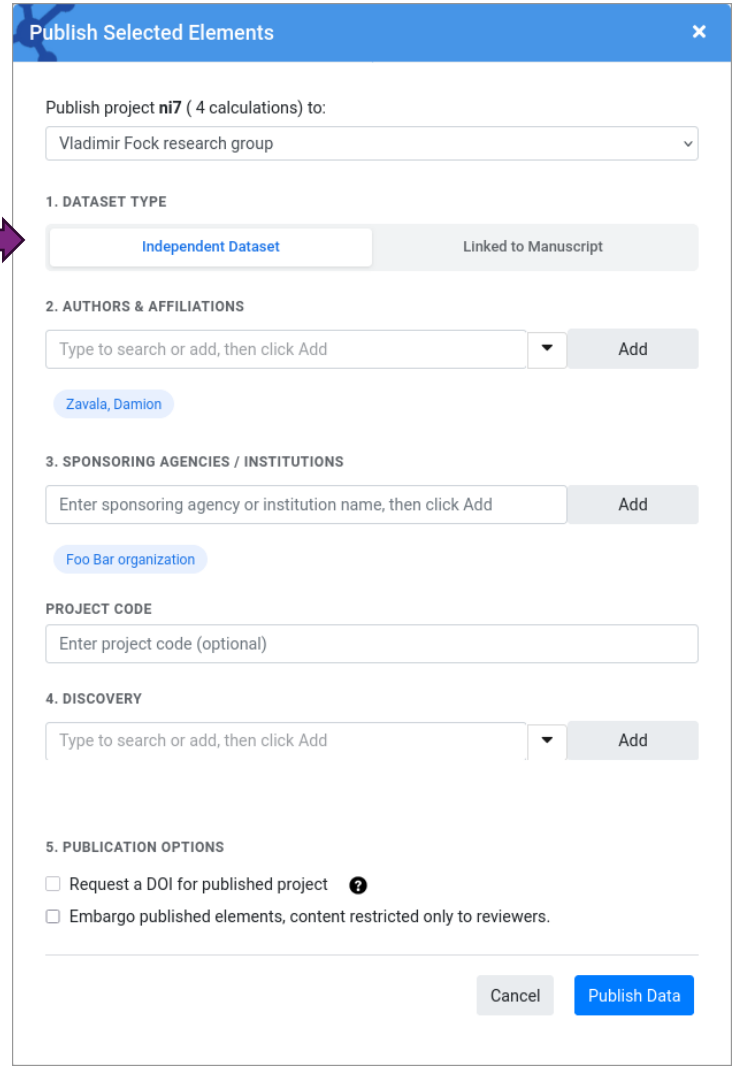


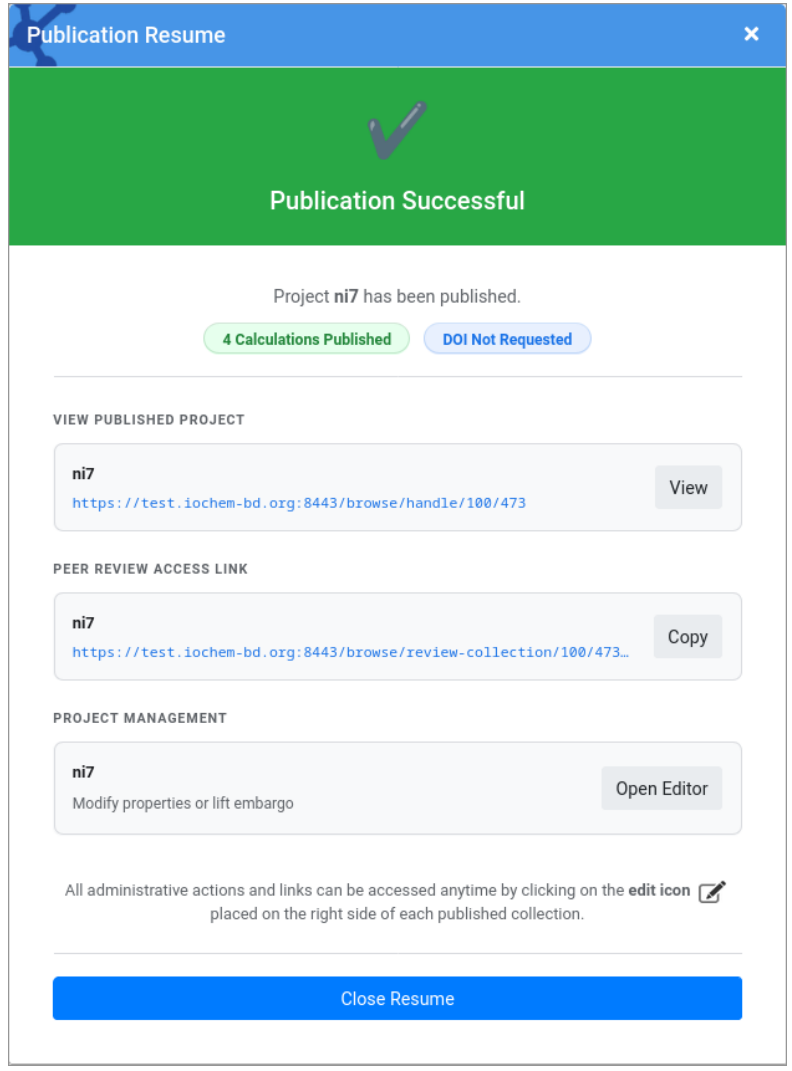
*Any modification done to this chart using Plotly Chart Studio won't be saved at ioChem-BD platform.

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

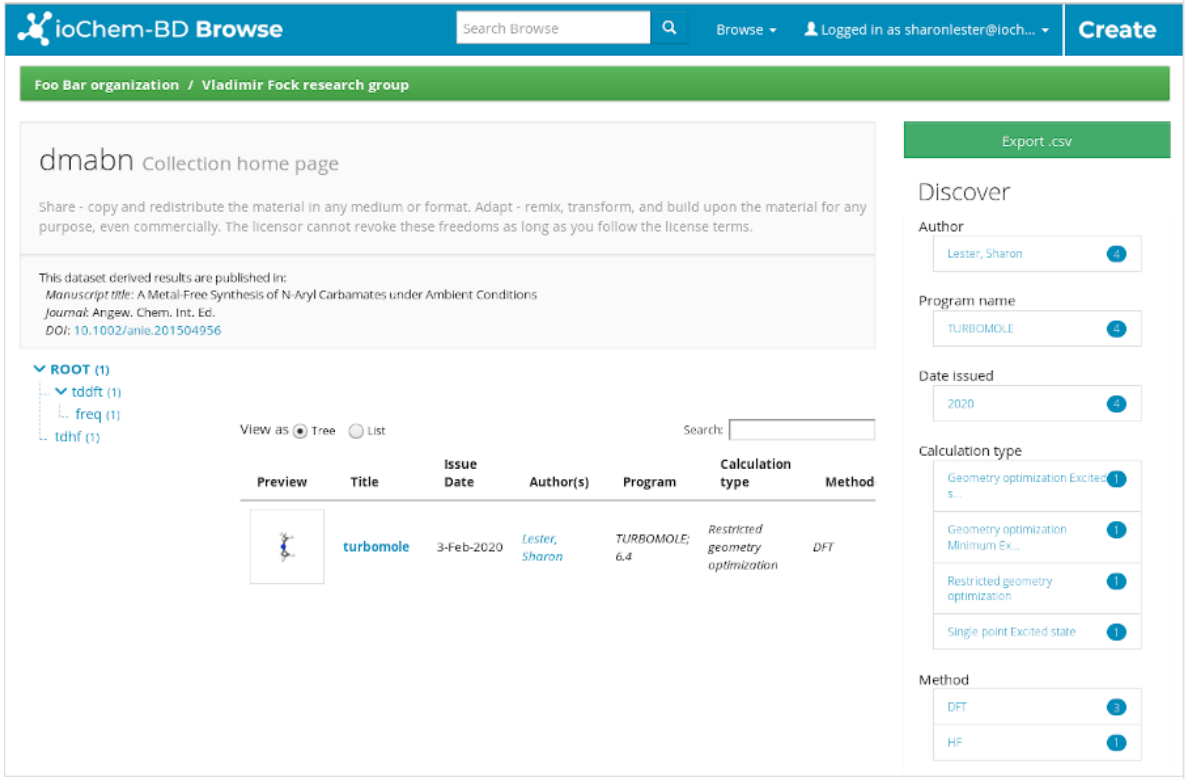
Publish Datasets into Browse

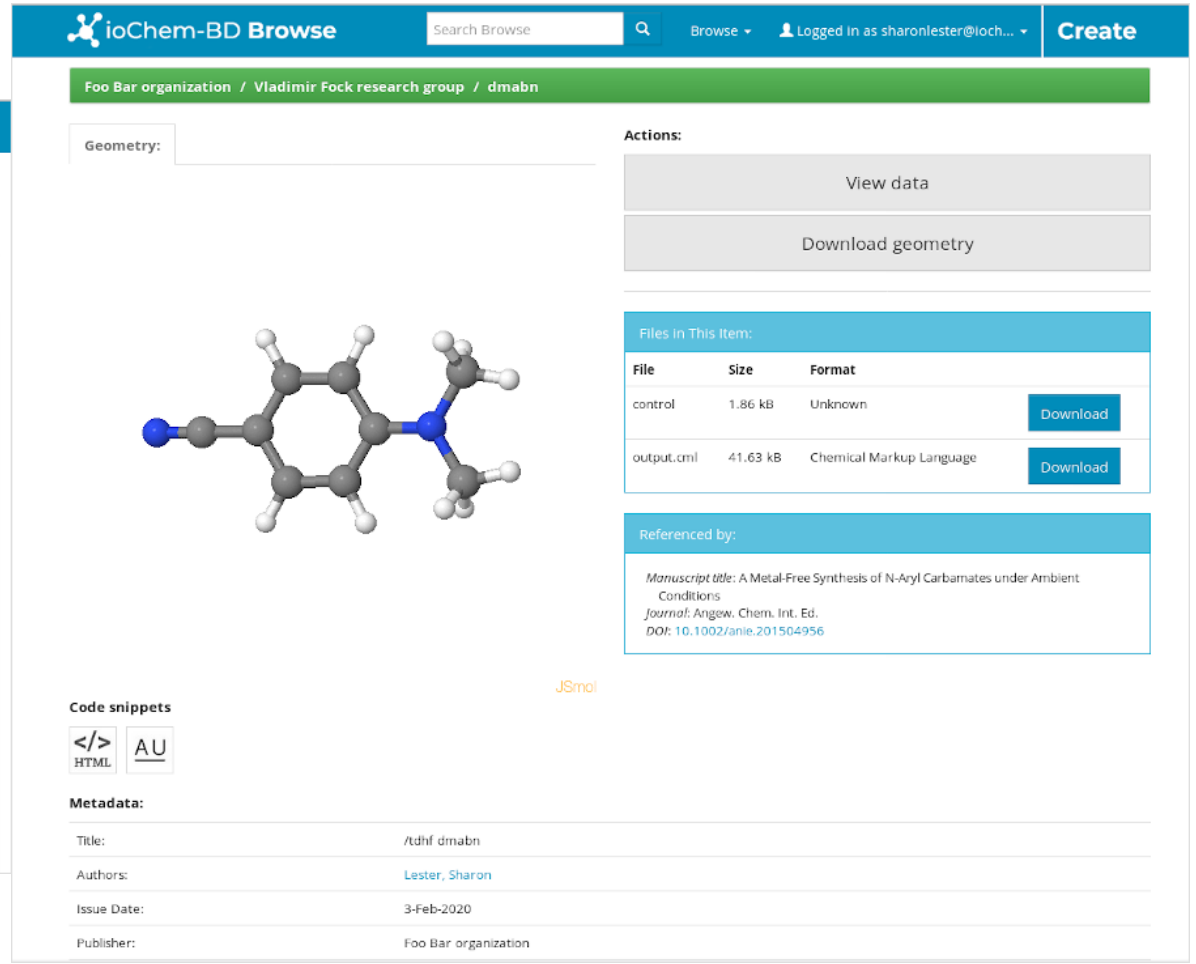






Browse Published Datasets in the Community





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