Applications of the ASE database
calculational workflows and data sharing

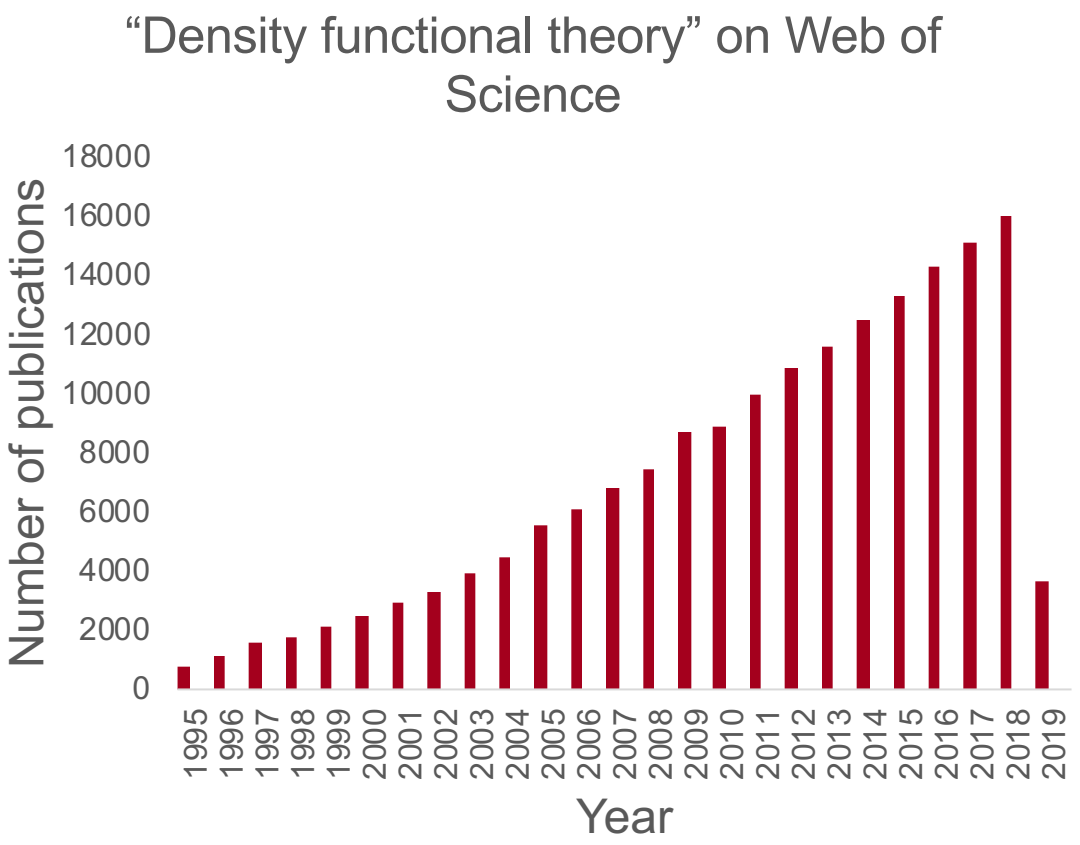
Kirsten Winther
Postdoc, SLAC, Stanford University
Nov 19th 2019

U.S. DEPARTMENT OF ENERGY
Office of Science
SLAC NATIONAL ACCELERATOR LABORATORY

Why do we need databases?

SLAC

- Sharing and handling increasing amount of DFT data.



NOMAD

OQMD

MATERIALSCLOUD

The Materials Project
materialsproject.org

AFLOW
Automatic - FLOW for Materials Discovery

Outline

- Introduction to the ASE database.
- Application of the database
  - Local handling of atomic structures and calculations
  - Server side open databases – catalysis-hub.org
  - Machine learning workflows

- Suggested improvements and future plans.

Overview of the ASE database implementation:

SLAC

J.J. Mortensen

David Kleiven

SQL?

SLAC

- ASE uses Structured Query Language (SQL)
  - Data is organized in well-defined tables.
  - Commands to efficiently select, write, update or delete.

```
sqlite > SELECT * FROM systems WHERE energy < -10 AND fmax < 0.01;
```

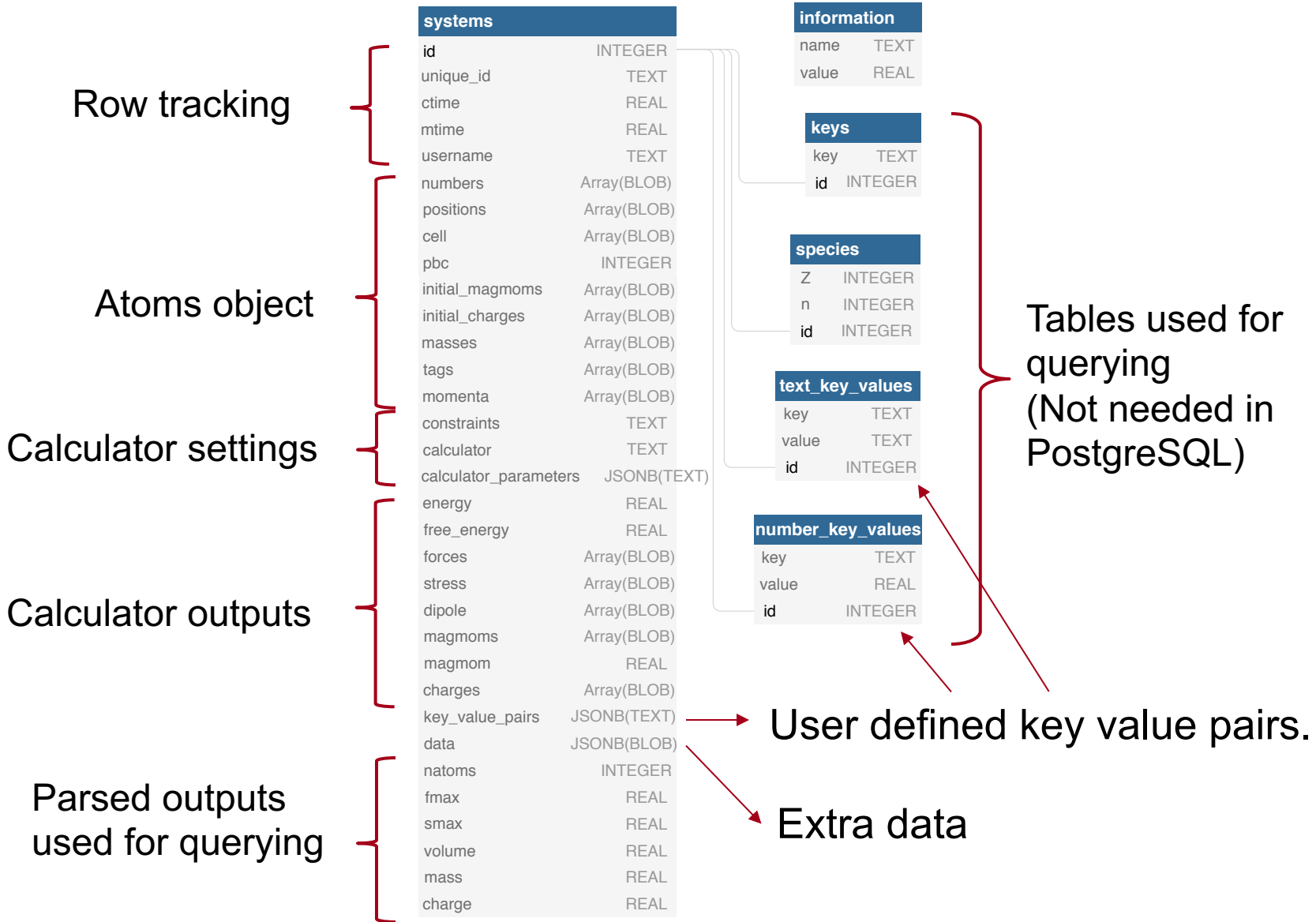
  - Local backends: 

  - Server side backends: 

- NoSQL (key value pair databases) 

  - No fixed database schema/structure -> high flexibility and scalability
  - PostgreSQL also has NoSQL-like features (JSONB data type)

Database structure (schema)



A simple use example

SLAC

```
#!/usr/bin/env python3

from ase.db import connect

db = connect('dbname.db'). # SQLite3 backend
db.write(atoms,
         key_value_pairs={'relaxed':1, adsorbate: 'C'}
         data={'vib': frequency_array})
# Select rows
for row in db.select('Cu', 'O>3', adsorbate='C'):
    print(row.id)
    print(row.energy)
    print(row.forces)
```

More examples at https://wiki.fysik.dtu.dk/ase/tutorials/db/db.html

Command line interface

View and sort calculations:

```
$ ase db AB2.db completed=1,relaxed=1 -s Ef -c
formula,p_name,energy,Ef,fmax,smax
formula|p_name          | energy|    Ef| fmax| smax
Ti2O4  |AB2_2_a_c_102  |-49.776|-8.476|0.008|0.038
Ti2O4  |AB2_2_d_j_84   |-49.766|-8.475|0.150|0.059
Ti4O8  |AB2_4_b_f_64   |-97.931|-8.341|0.013|0.034
Ti4O8  |AB2_4_j_lk_131|-97.738|-8.325|0.020|0.031
Ti4O8  |AB2_4_a_c_205  |-97.198|-8.280|0.007|0.032
Ti2O4  |AB2_2_a_b_224  |-48.386|-8.245|0.000|0.032
Ti4O8  |AB2_2_a_b_224  |-96.771|-8.245|0.002|0.033
```

Run local web-app:

```
$ ase db AB2.db -w
```

View structures in ASE gui:

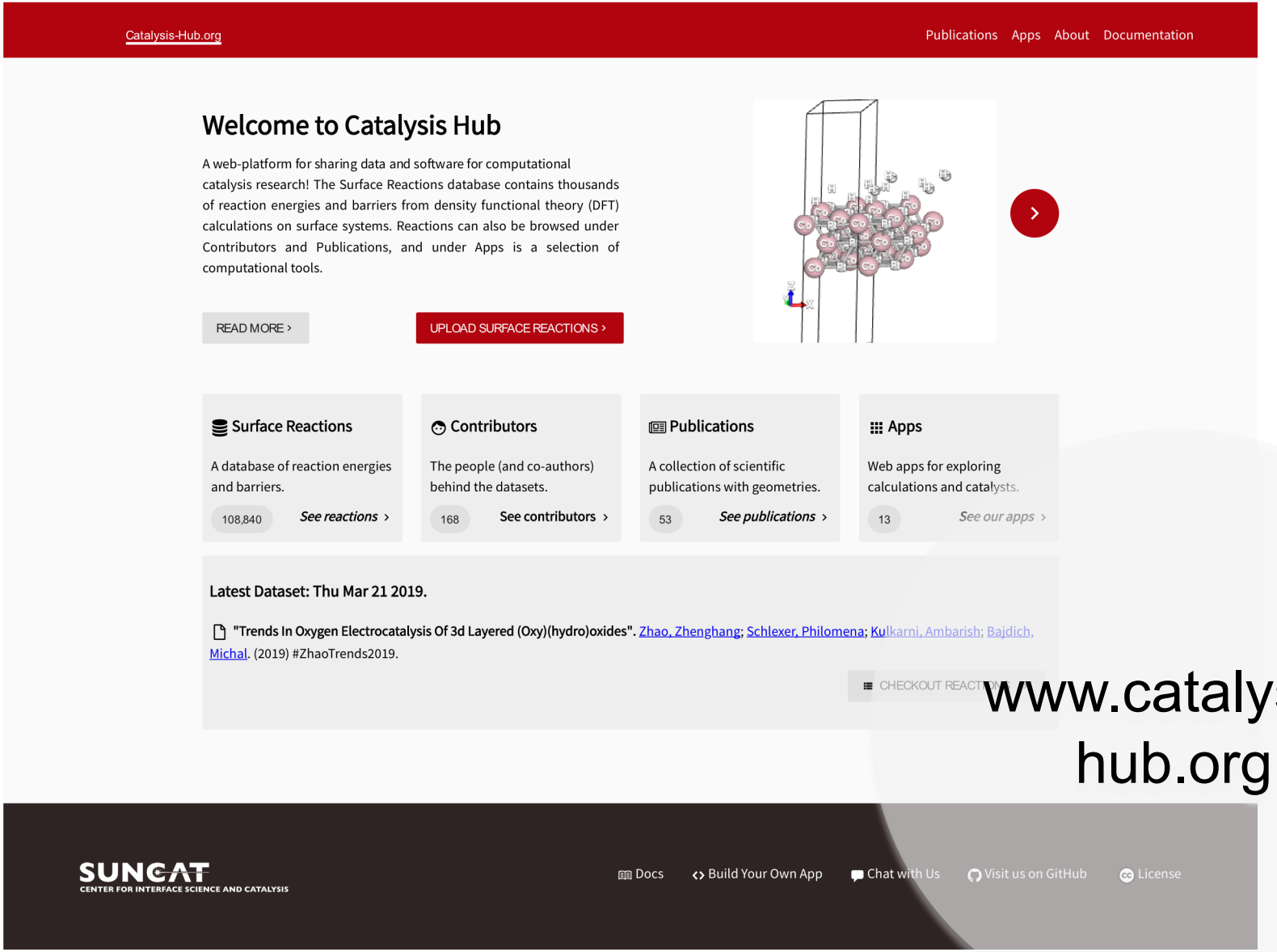
```
$ ase gui name.db@id=30
```

Server side databases using ASE db

- Computational Materials Repository: http://cmr.fysik.dtu.dk

- Katladb: https://nano.ku.dk/english/research/theoretical-electrocatalysis/katladb/

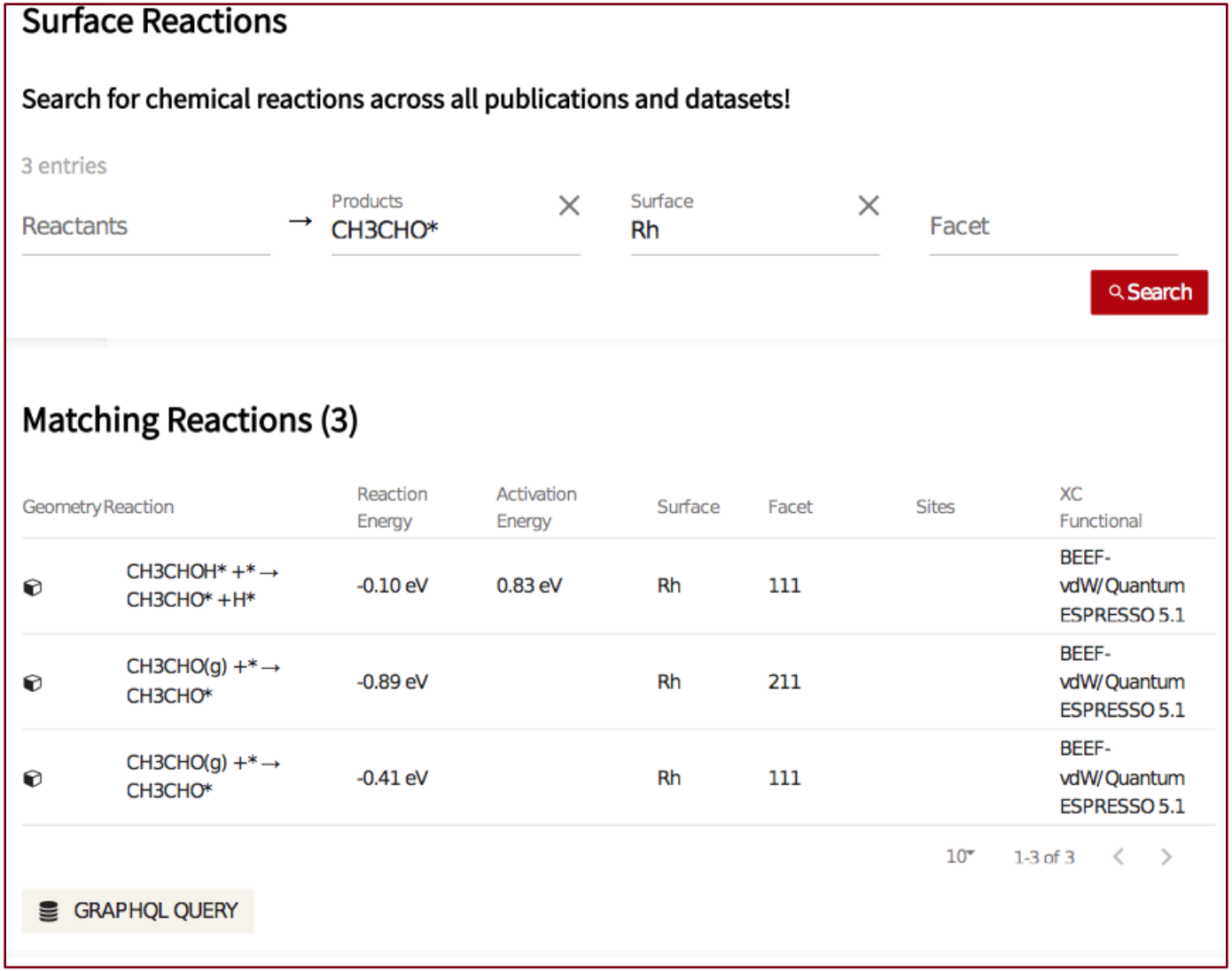
- Catalysis-hub: http://www.catalysis-hub.org



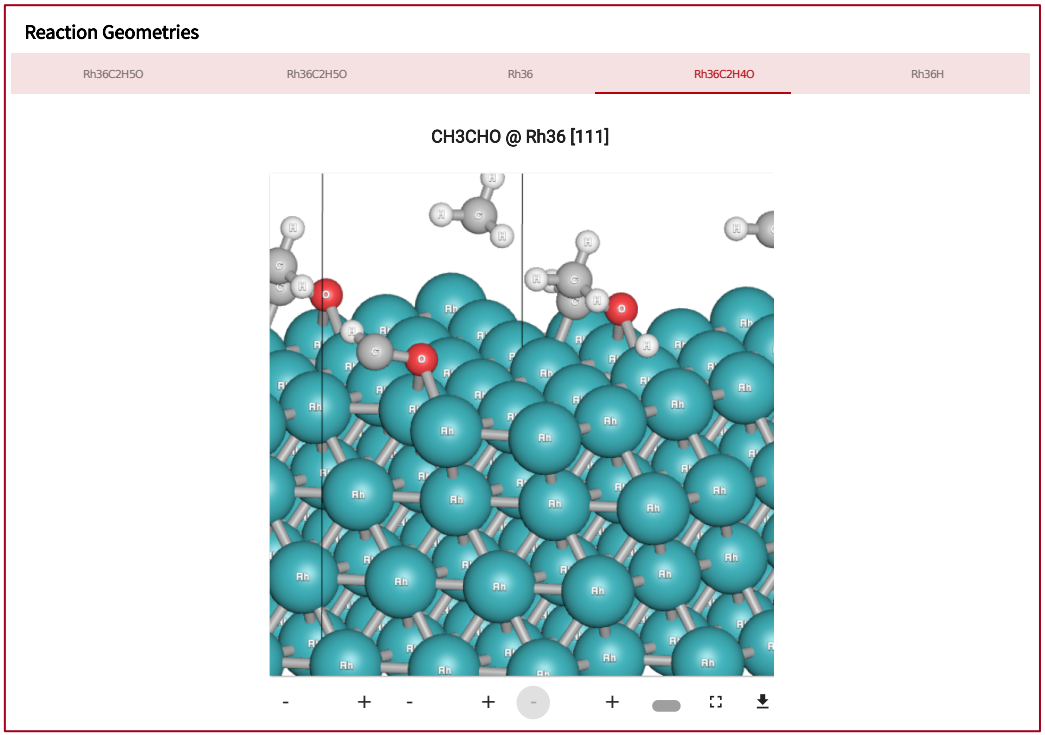
www.catalysis-hub.org

Winther, K.T. , Hoffmann, M. J., Mamun, O., Boes, J.R, Bajdich, M, Bligaard, T. Scientific Data (2019)

Surface Reactions database search



Visualize atomic structures in the browser

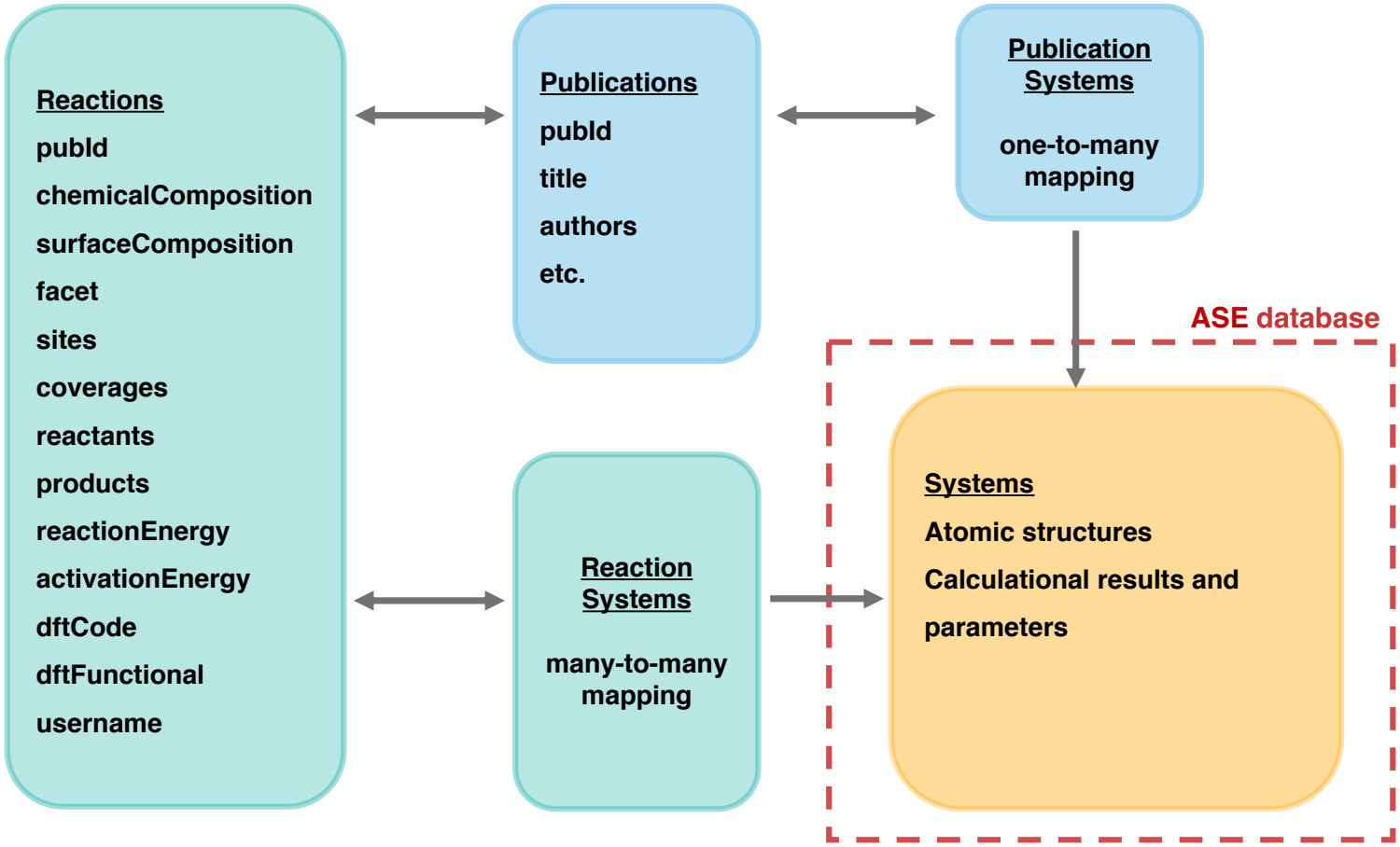


browse through all structures involved

View calculational details

- Formula: Rh36C2H4O
- DFT Total Energy: -115620.80 eV
- DFT Code: Quantum ESPRESSO 5.1
- DFT Functional: BEEF-vdW
- Publication: **"Selectivity Of Synthesis Gas Conversion To C2+ Oxygenates On Fcc (111) Transition-metal Surfaces"**. Schumann, Julia; Medford, Andrew J.; Yoo, Jong Suk; Zhao, Zhi-Jian; Bothra, Pallavi; Cao, Ang; Studt, Felix; Abild-Pedersen, Frank; Nørskov, Jens K.. *ACS Catalysis*. 4, 3447– 3453 (2018)
- Source DOI: 10.1021/acscatal.8b00201
- View all reactions in dataset
- Open GRAPHQL QUERY to view calculational details.

Catalysis-Hub data structure: extra tables

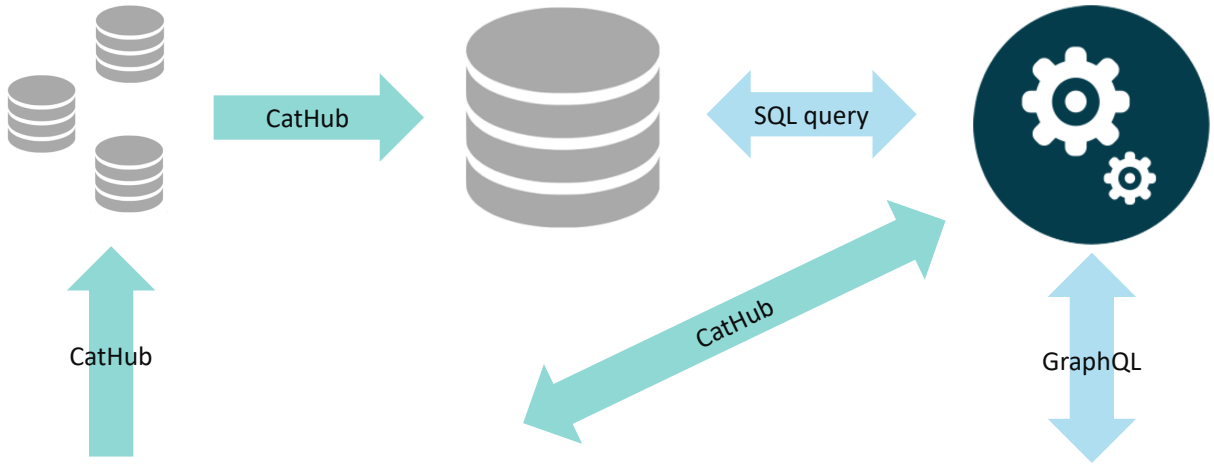


Catalysis-hub infrastructure

Data contributions
(SQLite3)

Database server
(PostgreSQL)

Web API (backend) at
https://api.catalysis-hub.org



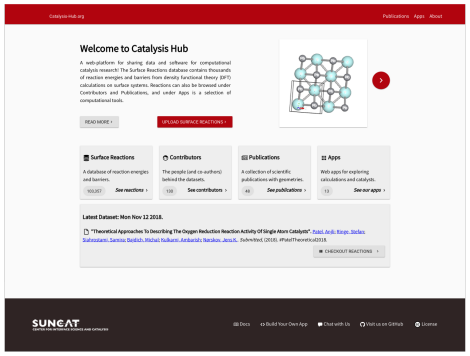

```
~$
~$ cathub reactions -q reactants=CO
Connecting to database at http://api.catalysis-hub.org/graphql

Executing query:

{reactions(first: 10, reactants: "CO") {
 totalCount
  edges {
    node {
      chemicalComposition
      Equation
      reactionEnergy
    }
  }
}}

Getting data from server...
```

Local workstation



Web page at
https://www.catalysis-hub.org

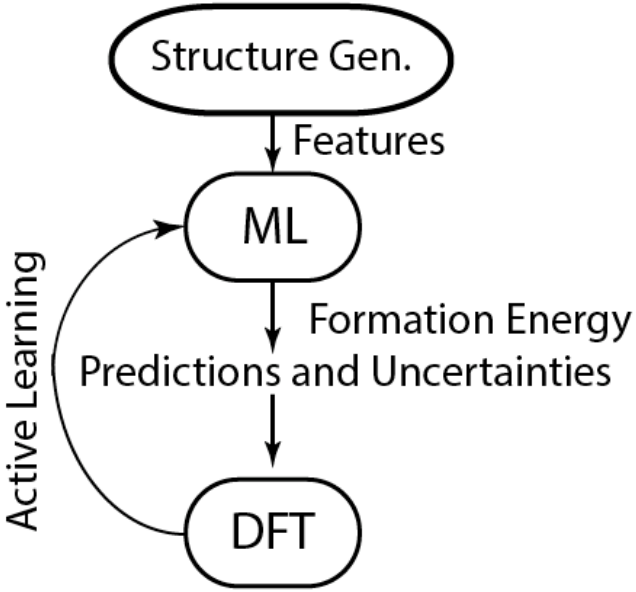
https://github.com/SUNCAT-Center

Outline

- Introduction to the ASE database.
- Application of the database
  - Local handling of atomic structures and calculations
  - Server side open database – catalysis-hub.org
  - Machine learning workflows
- Suggested improvements and future plans.

Motivation: Beyond high-throughput computational screening

- We don’t want to calculate everything with DFT!
- Use machine learning to guide the search which will (hopefully) enable us to search a broader space of materials with less DFT calculations.
- In active learning the ML model is continuously retrained with new calculations

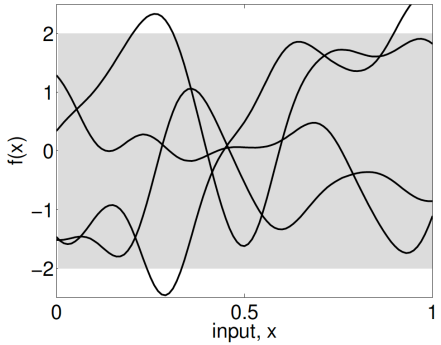


Machine learning for materials discovery

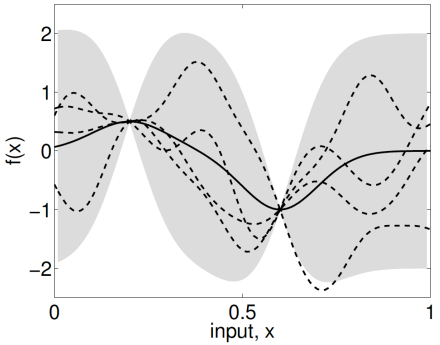
- Supervised regression with Gaussian Processes

$$y:\ \tilde{y} = f(X, \varphi)$$

target input features free parameters

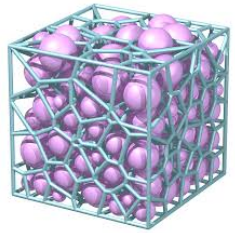


(a), prior



(b), posterior

- Fingerprint: Magpie voronoi tesselation[1]:

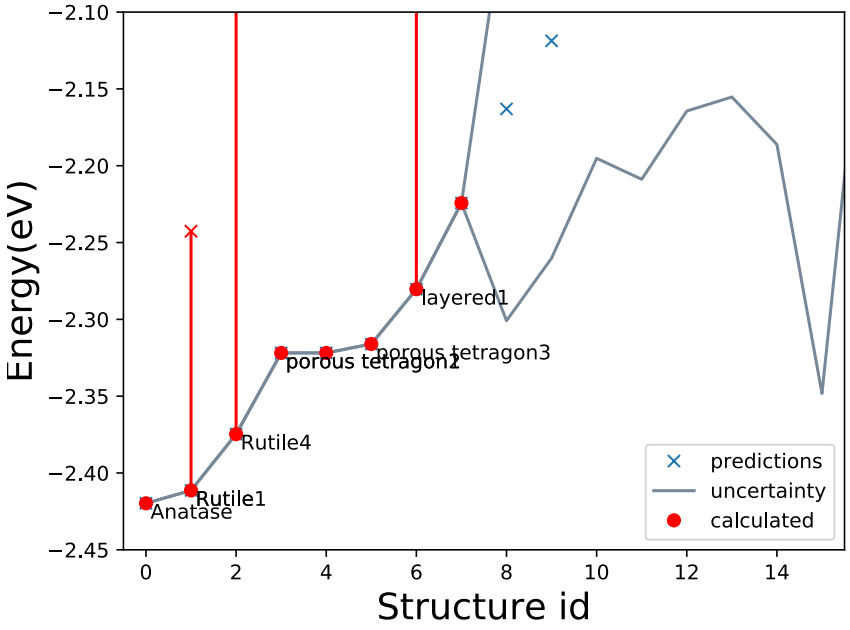


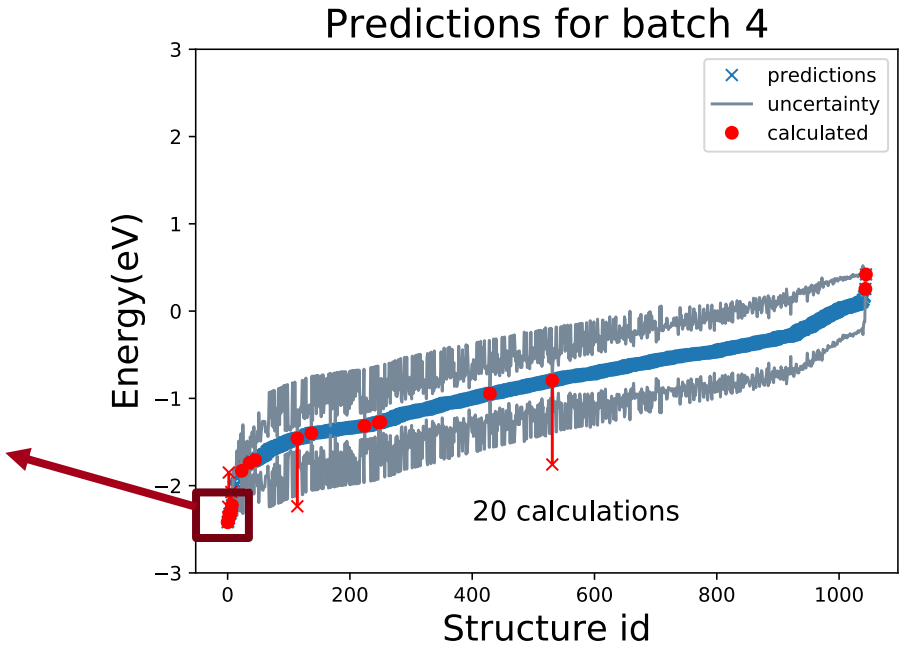
\+ 22 elemental properties (electronegativity, meting temperature, valence electrons, band gap, etc.) → 271 features for each structure

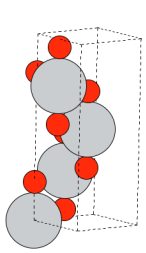
- Acquisition function: LCB = mean(Ef) – std(Ef)

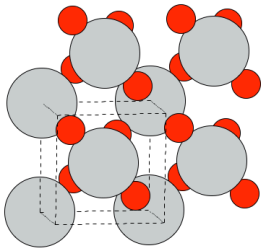
[1] L. Ward, C. Wolverton et al. PRB 96, 024104 (2017)
https://bitbucket.org/wolverton/magpie

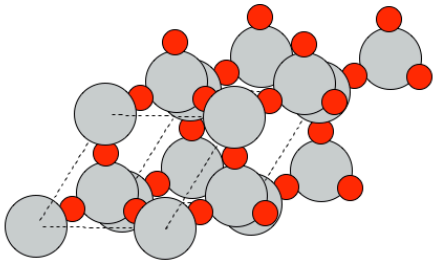
(Re)discovering stable polymorphs of TiO_2

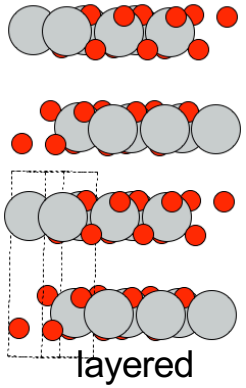





anatase

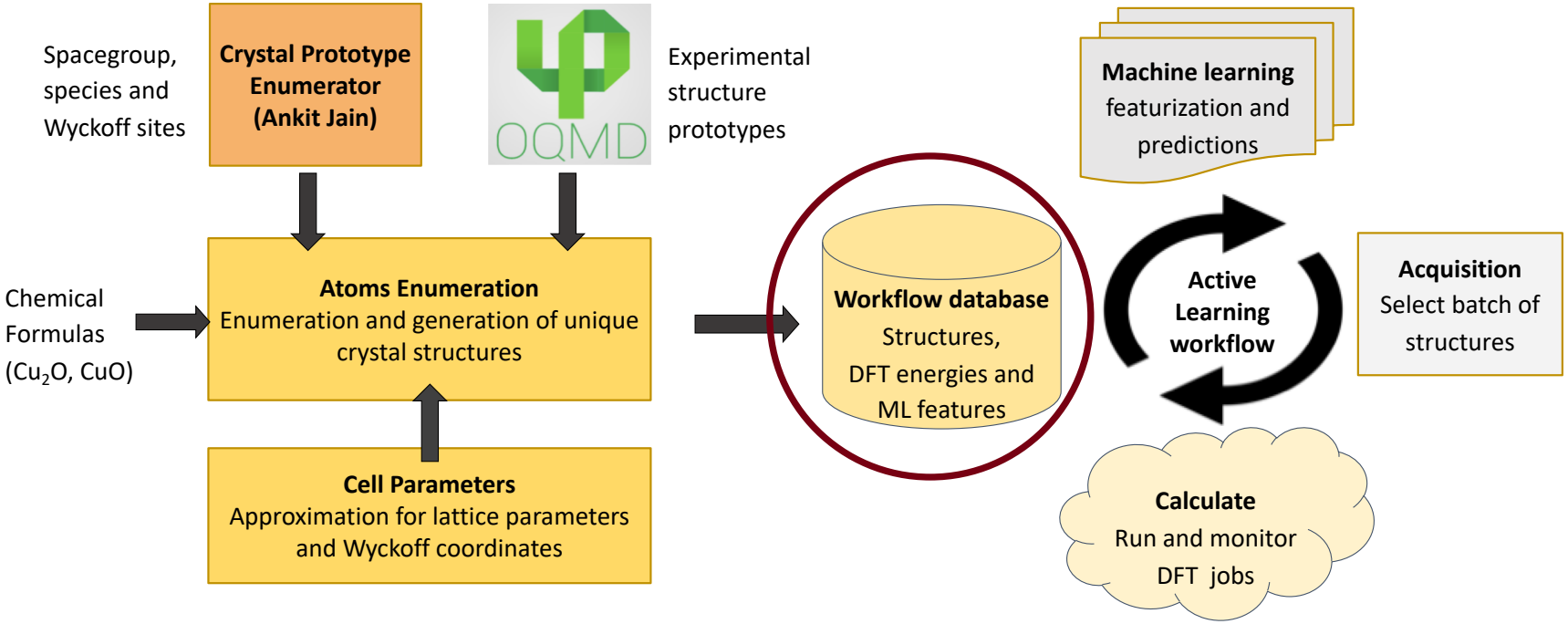

rutile


porous tetrahedral

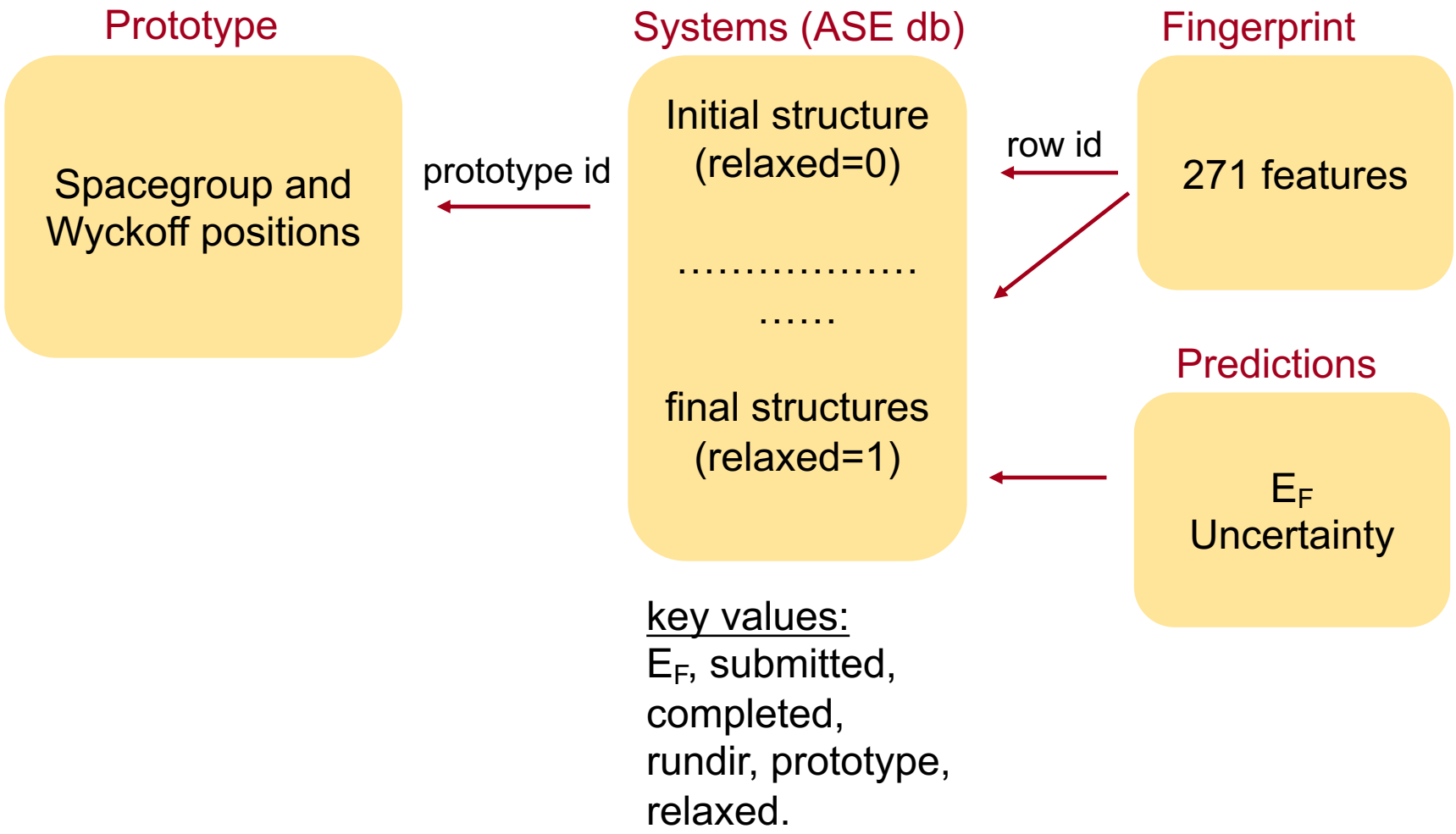

layered

Automated acquisition, job submission and monitoring

SLAC



Workflow database structure



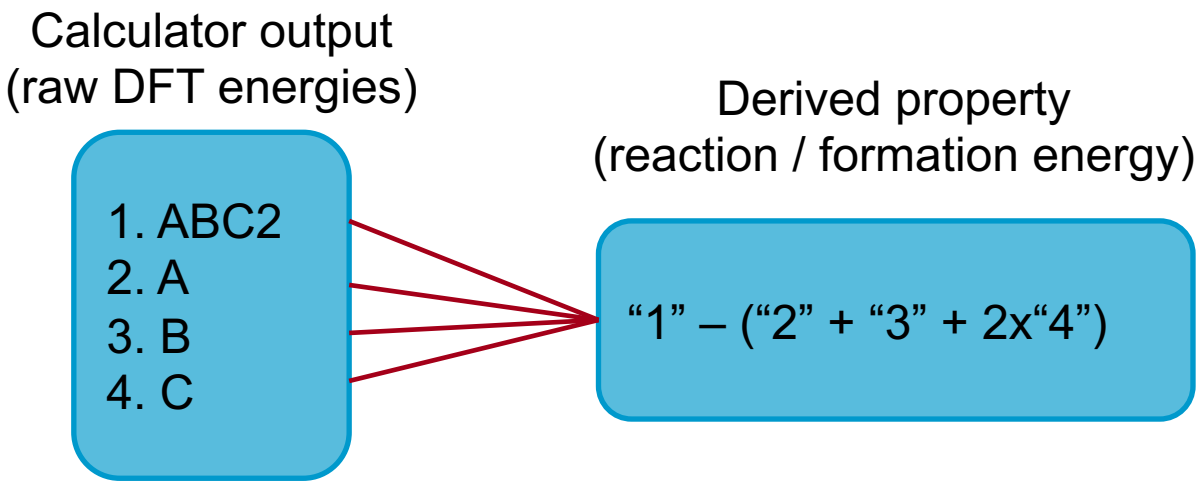
Possible improvements to the database

SLAC

Handle connections between atoms-rows:

- Storing full relaxation pathway / NEB images:

- Connections between row to get derived properties.

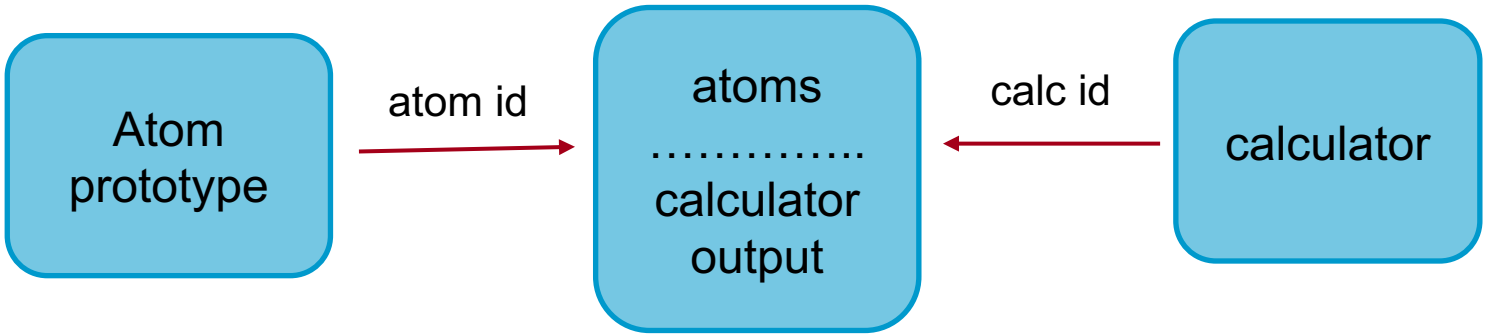


Possible improvements continued …

SLAC

Calculator information for enhanced workflow and data sharing:

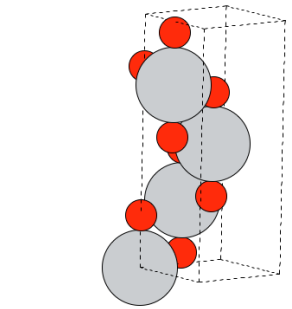
- Fix calculator ios + more tests (reading calculator and parameters from raw input/output files)
- Calculator version + setups/potentials!
- Storing detailed calculator output such as eigenvalues?
- Efficient grouping of entries with same calculator settings in separate 'Calculator' table?



Other possible contributions

- Extension to the Spacegroup module to include Wyckoff positions:

```
crystal(symbols=[‘Ti’, ‘O’],
        wyckoffs=[’a’, ‘e’],
        spacegroup=141,
        cellpar={‘a’: 3.78, ‘b’: 3.78, ‘c’: 9.51,
                 ’alpha’: 90, ‘beta’: 90,
                 ‘gamma’: 90,
                 ‘e1z’: 0.21})
```



anatase

- Structural fingerprints for machine learning?

Ackowledgements

DTU

Thomas Bligaard

Jens J. Mortensen

SUNCAT

Michal Bajdich

Frank Abild-Pedersen

Max Hoffman

Osman Mamun

Jacob R. Boes

Raul Flores

Ankit Jain

Meng Zhao

Chris Paolucci

SUNCAT postdocs

TRI Materials:

Jens Hummelshøj

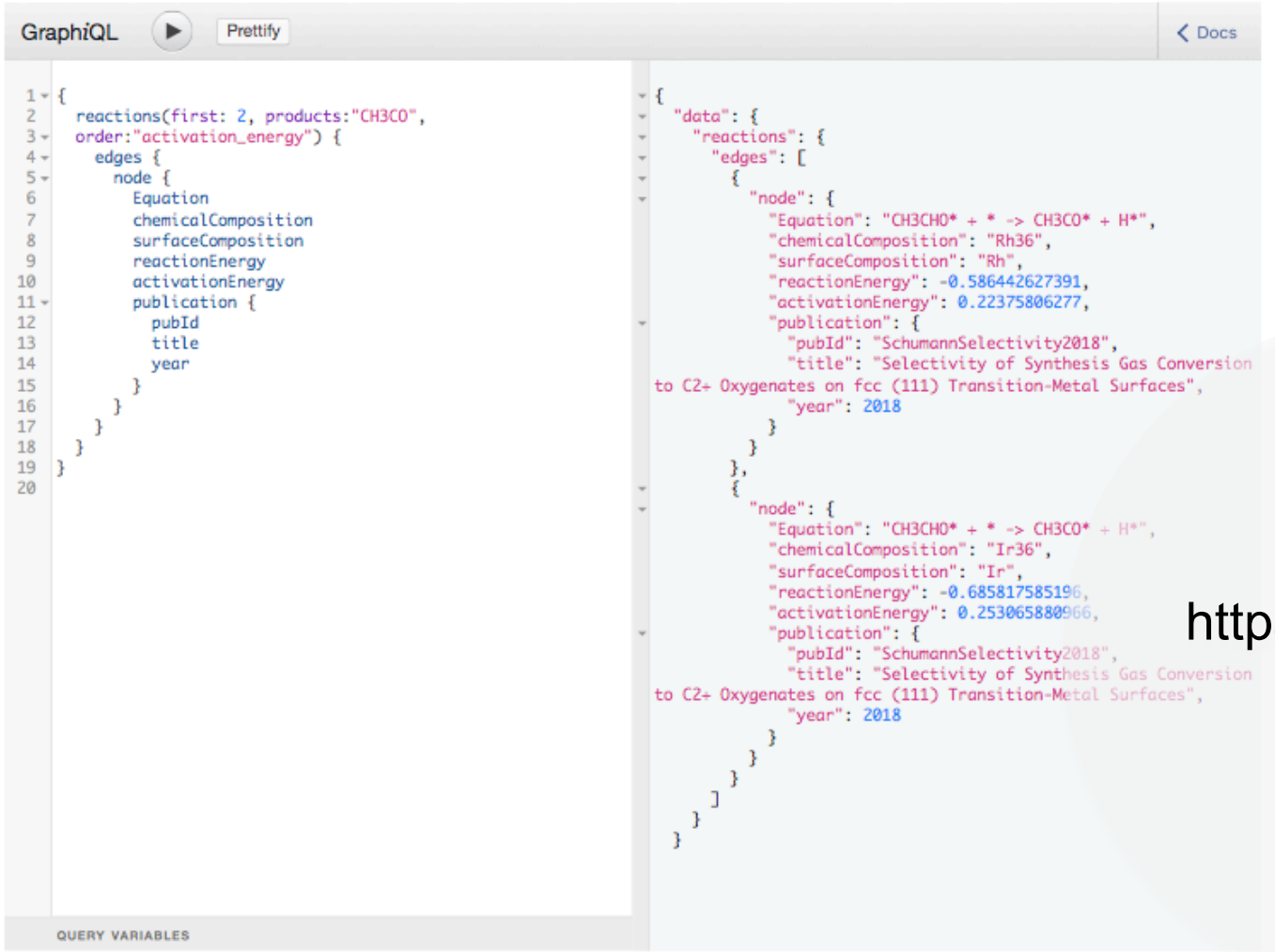
Joseph Montoya

Murat Aykol

Thank you for your attention

… and please share your data!

Direct access to data in web API



https://api.catalysis-hub.org

Materials databases