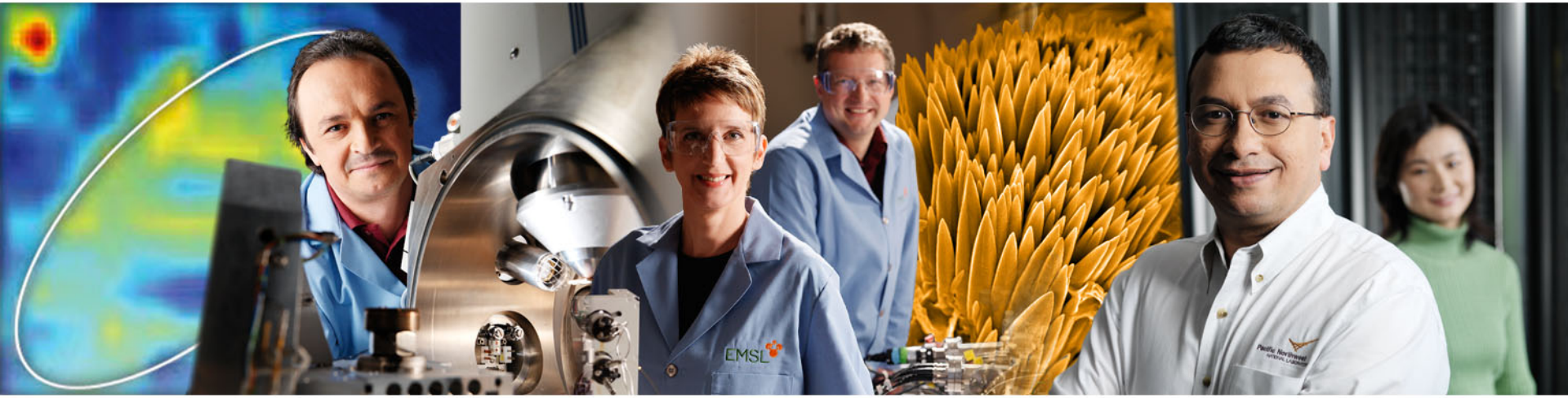
Integrating CML, FoX, Avogadro, NWChem, and EMSLHub to develop a computational chemistry knowledge and discovery base

Wibe A. de Jong, David M. Brown, Andrew Walker, Marcus D. Hanwell



Data is key to scientific discovery

- Multidisciplinary integrated research requires the ability to couple the diverse *semantically rich* data sets from complex experiments and simulations
  - Or, how to enable a researcher to do Google-style chemistry and physics searches

- Semantic Physical Sciences Workshop Series
  - International collaboration centred around the Chemical Markup Language and tools

UNIVERSITY OF CAMBRIDGE

University of BRISTOL

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

CSIRO

Kitware

Science & Technology Facilities Council
Daresbury Laboratory

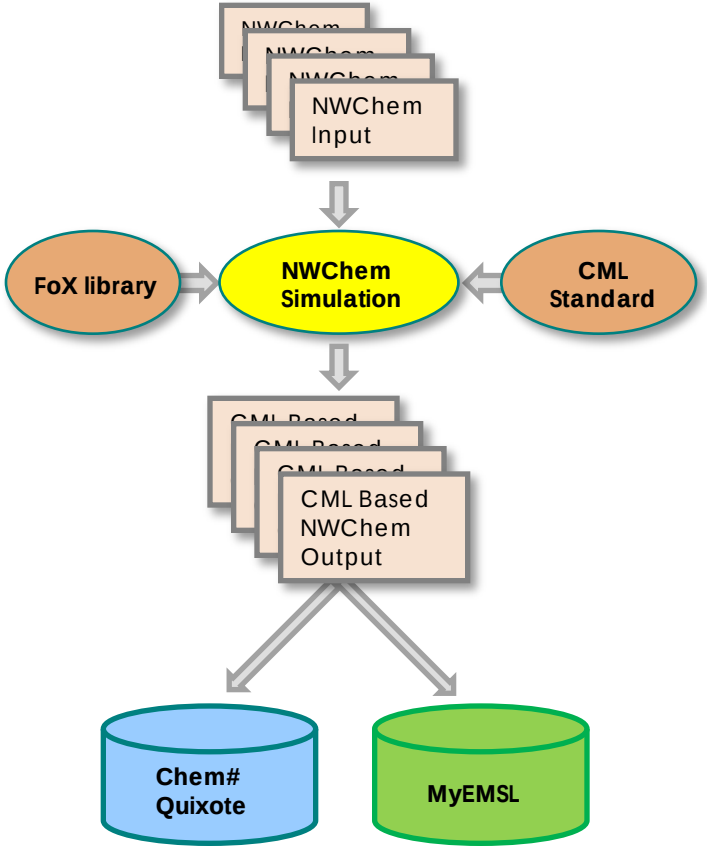
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

U.S. DEPARTMENT OF ENERGY

Generating semantic data with NWChem

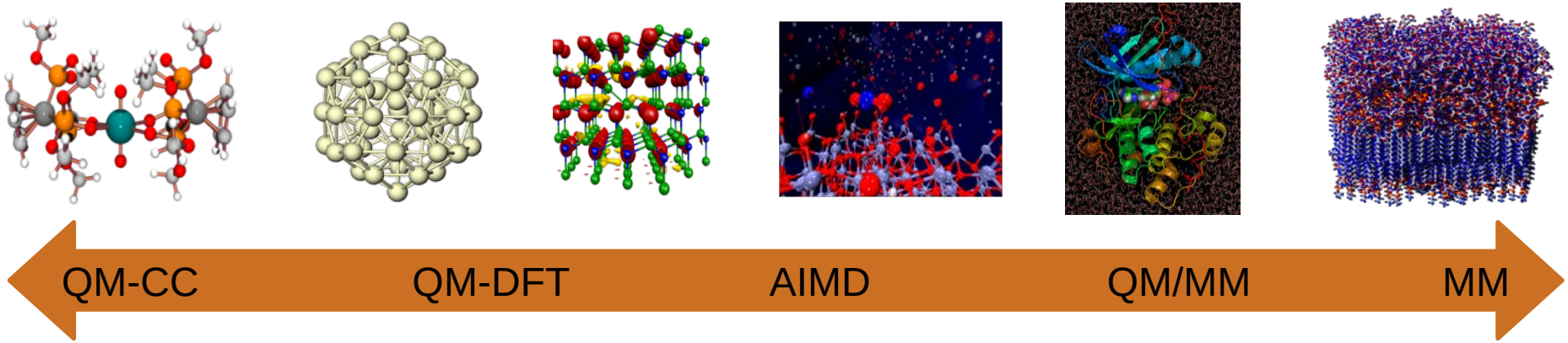
- Approach
  - Cemical Markup Language (CML) generator based on FoX library
  - New dictionary entries and conventions for CML CompChem


```
<?xml version="1.0" encoding="UTF-8"?>
<cml convention="convention:compchem" fileId="test.cml" xmlns="http://www.xml-cml.org/schema"
xmlns:compchem="http://www.xml-cml.org/dictionary/compchem/" xmlns:nwchem="http://www.nwchem-
xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:fpx="http://www.uszla.me.uk/fpx"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:convention="http://www.xml-cml.org/conventi
xmlns:unit="http://www.xml-cml.org/unit/si" xmlns:fmisc="http://www1.gly.bris.ac.uk/~walker/n
  <metadata name="fmisc:UUID" content="de34bf60-8f50-11e1-5315-4b50efbf4e4c"/>
  <module title="NWChem simulation" dictRef="compchem:jobList">
    <module dictRef="nwchem:Input" role="LexicalFile">
      <metadataList>
        <metadata name="fmisc:filename" content="prop_h2o.nw"/>
      </metadataList>
      <scalar dataType="xsd:string">echo</scalar>
      <scalar dataType="xsd:string">start prop_h2o</scalar>
      <scalar dataType="xsd:string">title h2o</scalar>
      <scalar dataType="xsd:string">ecce_print test.cml</scalar>
      <scalar dataType="xsd:string"></scalar>
      <scalar dataType="xsd:string">geometry units au nocenter</scalar>
      <scalar dataType="xsd:string">o       .00000000      .00000000      .11786656</scalar>
      <scalar dataType="xsd:string">h1      .00000000     1.84118838     -.93531364</scalar>
      <scalar dataType="xsd:string">h2      .00000000    -1.84118838     -.93531364</scalar>
      <scalar dataType="xsd:string">end</scalar>
      <scalar dataType="xsd:string"></scalar>
      <scalar dataType="xsd:string">basis</scalar>
      <scalar dataType="xsd:string">  * library cc-pvdz</scalar>
      <scalar dataType="xsd:string">end</scalar>
      <scalar dataType="xsd:string">charge 0</scalar>
      <scalar dataType="xsd:string">task scf </scalar>
    </module>
    <module title="NWChem runtime" dictRef="compchem:environment">
```

Searchable semantic web based chemistry databases

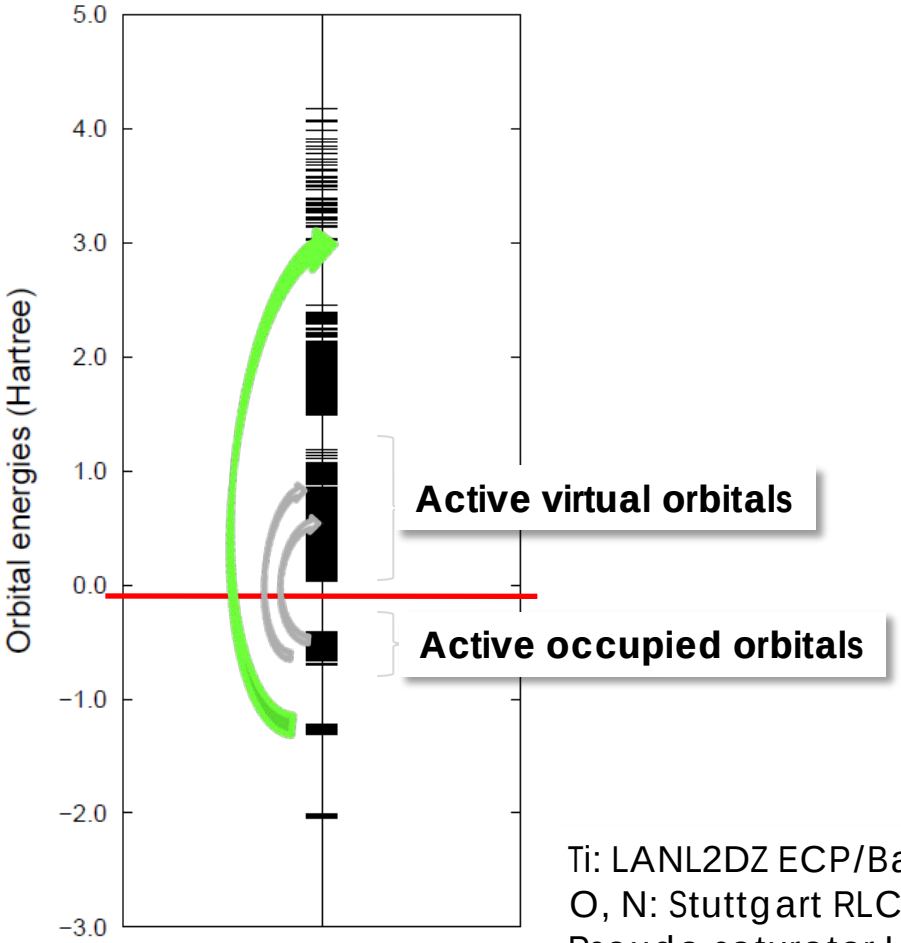
NWChem is Open-Source



- NWChem consortium delivers capabilities and infrastructure for computational chemistry community to build upon
- License is Educational Community License (ECL 2.0)
  - Apache style license

50,000 downloads in 2 years

Active space EOMCCSd for materials



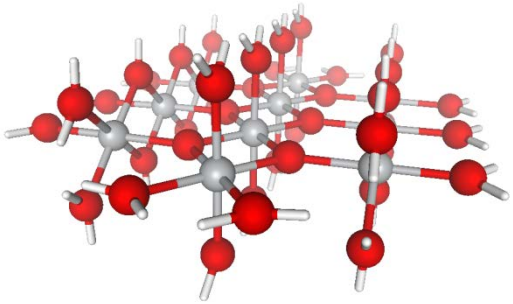
400 correlated electrons

120 active occupied orbitals
360 active virtual orbitals

TiO_2 **EOMCCSd: 3.84 eV**
Delocalized, multiconfigurational character

N-Doped TiO_2 **EOMCCSd: 2.79 eV**
Localized character, excitations dominated from 2p orbitals on N, O → Ti (3d)

110: $Ti_{10}O_{40}$ $(Z_O)_{60}$



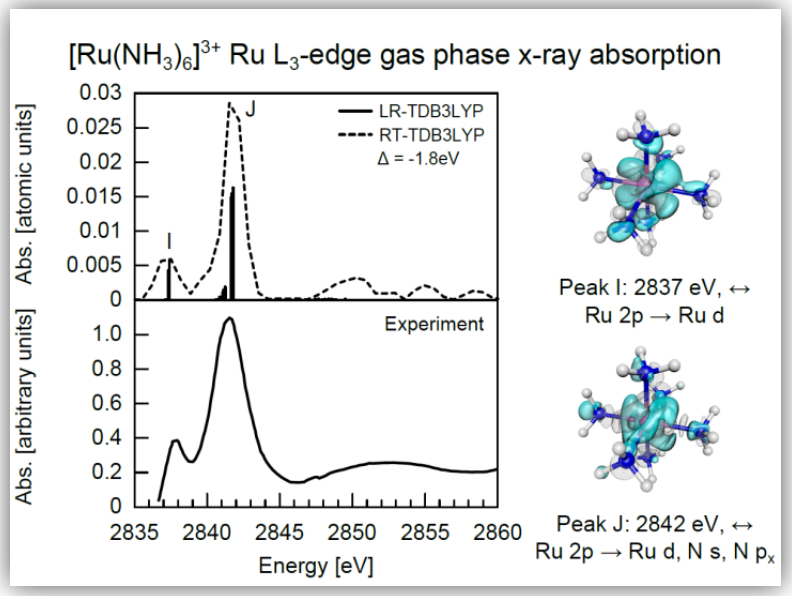
Ti: LANL2DZ ECP/Basis
O, N: Stuttgart RLC ECP/Basis
Pseudo-saturator H: 6-31G
800 basis functions

Govind, Lopata, Rousseau, Andersen, Kowalski, JPC Lett. ***2****, 2696 (2011)*

Understanding X-ray near-edge spectroscopy from first principles

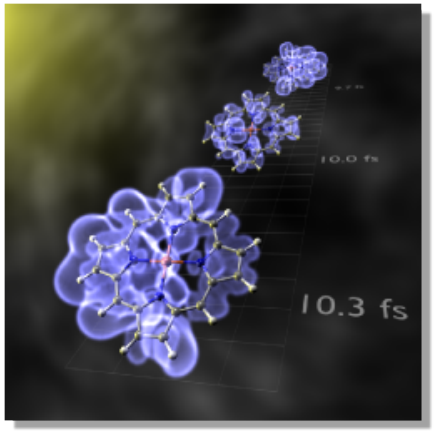
XANES experiments together with x-ray simulations in NWChem provides insight into oxidation states of ruthenium complexes

- Goal is to better understand role of charge transfer in ruthenium complexes in catalytic and solar energy conversion processes
- Calculations elucidated the observed XANES spectra and provided new understanding of the spectral content
- Excited-state calculations required NWChem's one-of-a-kind (real-time) time-dependent density functional theory

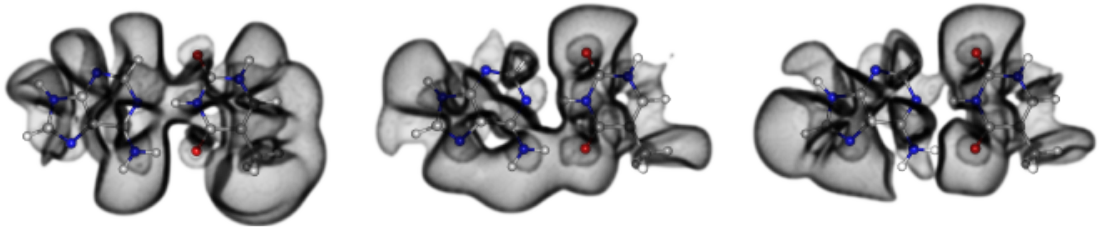


Lopata, Van Kuiken, Khalil, Govind, *J. Chem. Theory Comput.* **8**, 3284 (2012)

Ultrafast Electron Dynamics
Real-time TDDFT

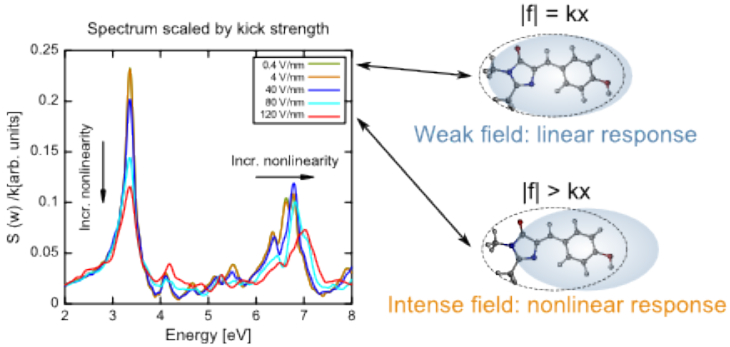


Resonant excitation of zinc porphyrin

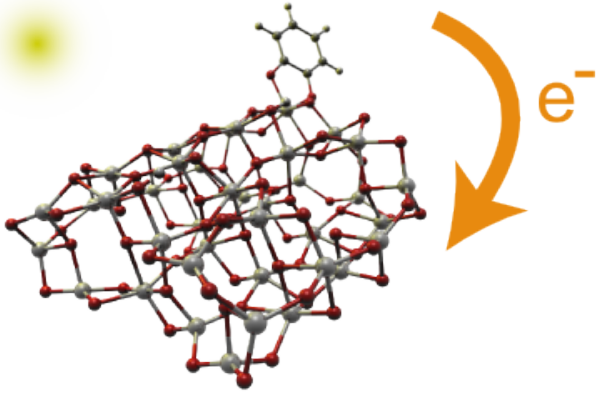


Charge transfer across adenine-thymine base pair using range-separated functional

Ken Lopata



Nonlinear absorption spectrum of green fluorescent protein chromophore

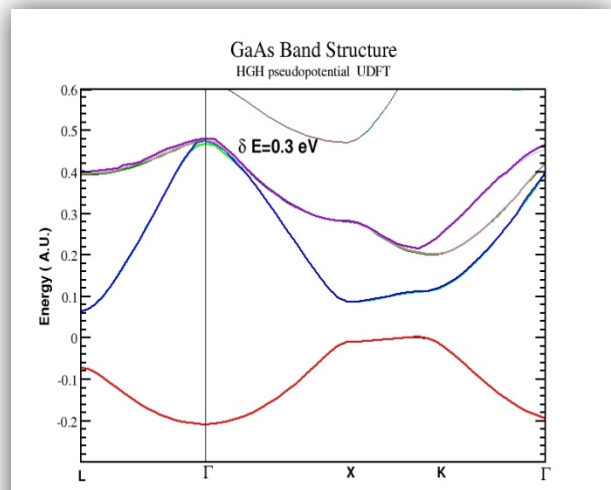


Charge injection from molecule to TiO_2 in dye-sensitized solar cells takes ~10 fs

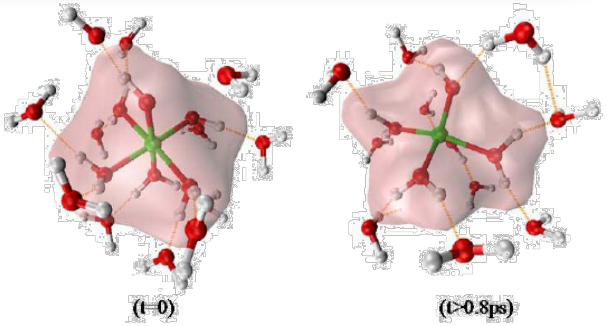
*K. Lopata and N. Govind, JCTC **7**, 1344 (2011)*

Plane wave DFT and dynamics for solution, surfaces, and materials

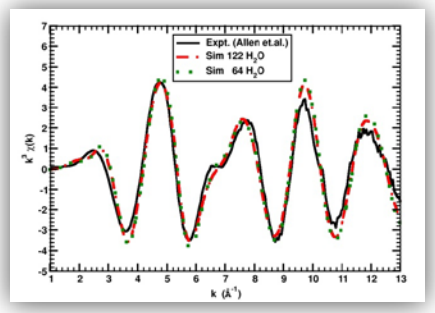
- Plane wave density functional theory
- Extensive dynamics functionality with Car-Parrinello
- AIMD QM/MM molecular dynamics, e.g. SPC/E,CLAYFF solid state MD
- Various exchange-correlation functionals
  - Exact exchange is very efficient
- SIC and OEP for localization
- NMR with spin-orbit ZORA
- Soon full set of PAW libraries (like VASP)

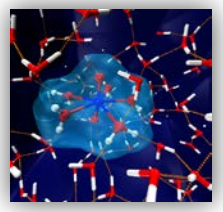


Spin-Orbit splitting in GaAs



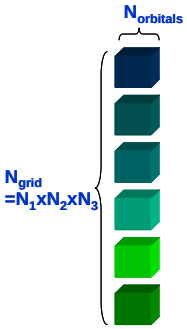
AIMD provides evidence for five-coordinate $Al(H_2O)_4OH^{2+}$
Swaddle et al, ***Science***, **2005**





AIMD accurately models EXAFS of uranyl in water

Strong petaflop scaling of plane waves



NWChem already established a large developers community

EMSL

University of Cambridge

UCSD

Duke University

The University of Wisconsin Madison

Universitat de Girona

Massachusetts Institute of Technology

UCLA

UBC The University of British Columbia

LSU

University at Buffalo The State University of New York

J. Heyrovský Institute of Physical Chemistry Academy of Sciences of the Czech Republic

EMSL

UNT University of North Texas Discover the power of ideas.

Iowa State University

Oak Ridge National Laboratory Managed by UT-Battelle for the Department of Energy

Texas Tech University

Science & Technology Facilities Council Daresbury Laboratory

Argonne National Laboratory

vrije Universiteit amsterdam

The University of Tennessee Knoxville

Universität des Saarlandes

Brookhaven National Laboratory

University of Minnesota Driven to Discover℠

PennState 1855

U.S. Department of Energy

Proudly Operated by Battelle Since 1965

Generating CML data

- Experimental version of NWChem generates semantic data
  - Completed CML generator based on FoX library
  - CML information based on prior ECCE data generator
    - Completely revamped this to align with CML structure
  - Gaussian basis function based quantum methods only for now


```xml
<?xml version="1.0" encoding="UTF-8"?>
<cml convention="convention:compchem" fileId="test.cml" xmlns="http://www.xml-cml.org/schema"
xmlns:compchem="http://www.xml-cml.org/dictionary/compchem/" xmlns:nwchem="http://www.nwchem-
xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:fpx="http://www.uszla.me.uk/fpx"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:convention="http://www.xml-cml.org/conventi
xmlns:unit="http://www.xml-cml.org/unit/si" xmlns:fmisc="http://www1.gly.bris.ac.uk/~walker/r
  <metadata name="fmisc:UUID" content="de34bf60-8f50-11e1-5315-4b50efbf4e4c"/>
  <module title="NWChem simulation" dictRef="compchem:jobList">
    <module dictRef="nwchem:Input" role="LexicalFile">
      <metadataList>
        <metadata name="fmisc:filename" content="prop_h2o.nw"/>
      </metadataList>
      <scalar dataType="xsd:string">echo</scalar>
      <scalar dataType="xsd:string">start prop_h2o</scalar>
      <scalar dataType="xsd:string">title h2o</scalar>
      <scalar dataType="xsd:string">ecce_print test.cml</scalar>
      <scalar dataType="xsd:string"></scalar>
      <scalar dataType="xsd:string">geometry units au nocenter</scalar>
      <scalar dataType="xsd:string">o       .00000000      .00000000      .11786656</scalar>
      <scalar dataType="xsd:string">h1      .00000000     1.84118838     -.93531364</scalar>
      <scalar dataType="xsd:string">h2      .00000000    -1.84118838     -.93531364</scalar>
      <scalar dataType="xsd:string">end</scalar>
      <scalar dataType="xsd:string">basis</scalar>
      <scalar dataType="xsd:string">  * library cc-pvdz</scalar>
      <scalar dataType="xsd:string">end</scalar>
      <scalar dataType="xsd:string">charge 0</scalar>
      <scalar dataType="xsd:string">task scf </scalar>
    </module>
    <module title="NWChem runtime" dictRef="compchem:environment">
```

Expanding the FoX library

- New functionality includes
  - Representation of lexical input files
  - Molecular orbitals
  - Adding reference IDs to atoms in molecule block
  - Minor modifications to Gaussian basis set module

- New release available at

  http://www1.gly.bris.ac.uk/~walker/FoX/

- Note, NWChem not the first to utilize the FoX library
  - SIESTA, GULP, and most recently TURBOMOLE

Writing CML with FoX: Starting a CML file

```
use FoX_wxml
use FoX_wcml
use FoX_common


call cmlBeginFile(xf, filename=‘myCMLfile.cml’, unit=31)
call cmlAddNamespace(xf, prefix='compchem
&            URI='http://www.xml-cml.org/dictionary/compchem/')
call cmlAddNamespace(xf, prefix='nwchem',
&            URI='http://www.nwchem-sw.org/dictionary/nwchem/')
call cmlStartCml(xf, convention='convention:compchem', validate=.true.)
```

Writing CML with FoX: Writing to a CML file

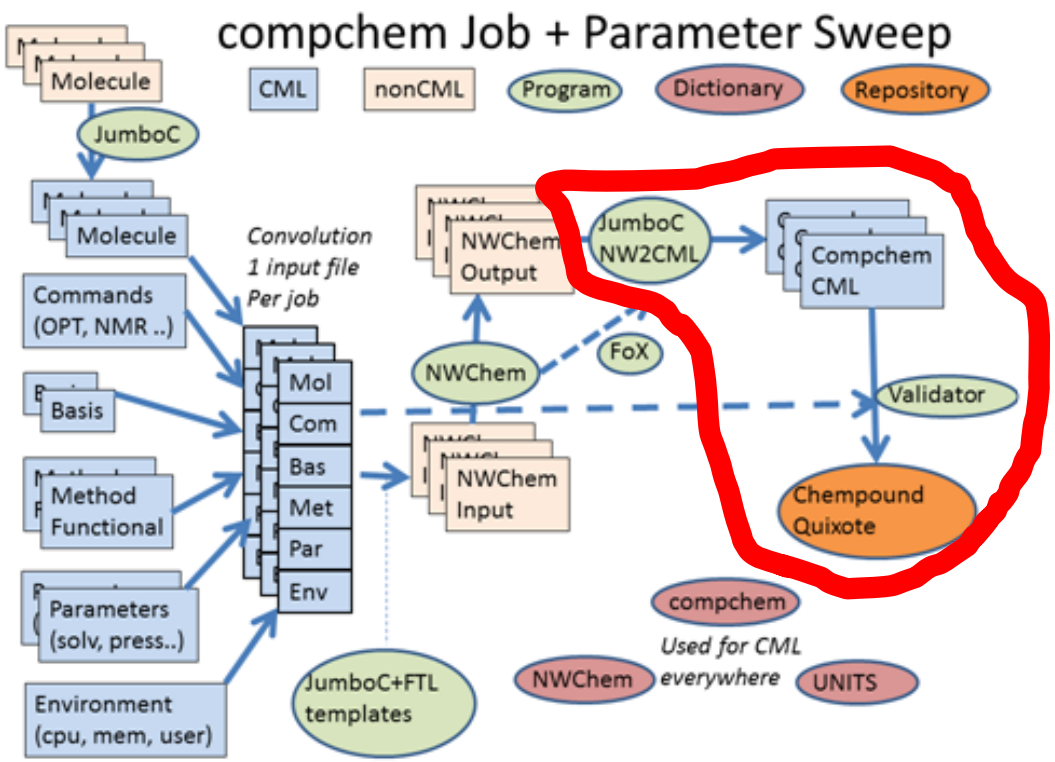
```
   call cmlAddMolecule(xf,natoms=nat, elements=elsym, coords=coord,
&                                atomIds=tags ,style='cartesian', id=trim(name))

   call cmlAddProperty(xf, value='42.0d0', units='nonsi:hartree',
&                                dictRef='nwchem:totalEnergy')

   call cmlAddProperty(xf, units='unit:none', nrows=3, ncols=3,
&                                value=efgArray, dictRef='nwchem:efgtensor')

   call cmlEndCml(xf)
   call cmlFinishFile(xf)
```

Getting old output files into a semantic form



- JumboConverter to convert old NWChem output files into CML

Converters or integration of CML writing in computational chemistry tools

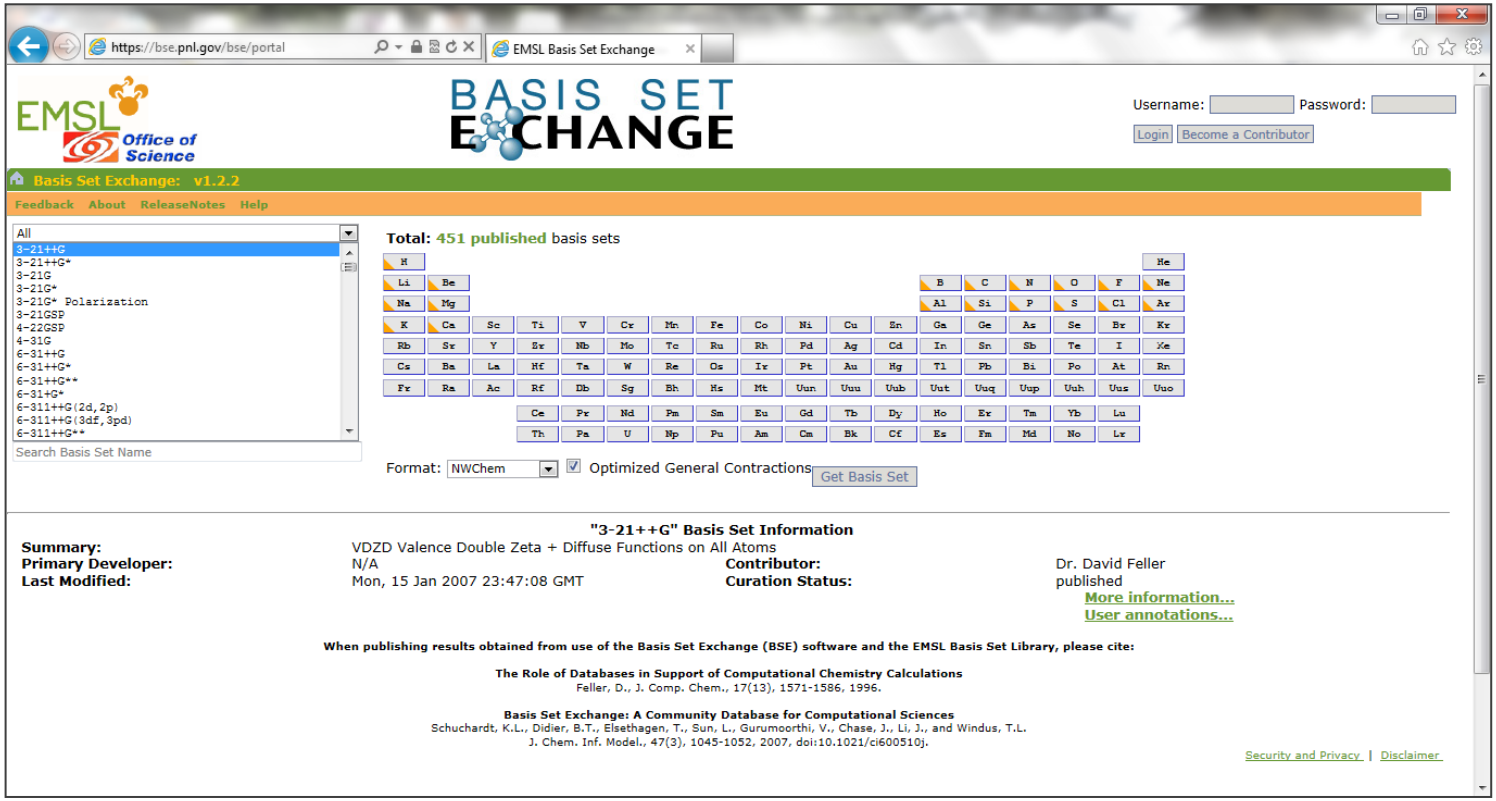
- Output compatibility of converters
  - Change of output, addition of new data breaks converters
  - One converter per software per version needed
- FoX library provides common interfaces
  - For new conventions, dictionaries, concepts
  - Flexibility for developers to adopt new interfaces

Advancing CML language

- Expanding dictionary and adding conventions
  - Molecular orbital representation
  - Various properties utilizing linked property lists
  - Expanding conventions with id and ref connecting linking data
- Keeping large data blocks out of XML/CML
  - Molecular orbital vectors
  - Time stamped trajectories
- Building community advocacy
  - Generate full draft dictionaries and conventions
  - Work towards broader adoption through community workshops

CML dictionary and conventions for Gaussian based molecular orbitals

- Relatively easy as basis sets are well standardized
  - CML CompChem similar to XML in Basis Set Exchange



http://bse.pnl.gov/bse/portal

Drafts for new CML dictionary and conventions: Gaussian based MOs

- Draft for Gaussian based molecular orbitals
  - CML writer for orbitals added to FoX library
  - Complete set of information should allow other codes to read in and reuse the data


```
<list dictRef="molecularOrbitals" id="nmrrun.movecs">
  <array size="10" delimiter="|" dataType="xsd:string" id="aoDescriptions"
dictRef="atomicOrbitalDescriptions">1 H  s        |1 H  s        |1 H  px        |1 H  py        |1 H  pz
|2 H  s        |2 H  s        |2 H  px        |2 H  py        |2 H  pz
  </array>
  <list dictRef="molecularOrbital" id="molecularOrbital1">
    <scalar dataType="xsd:double" dictRef="orbitalEnergy">-4.371860531460e-1</scalar>
    <scalar dataType="xsd:string" dictRef="orbitalSymmetry">a</scalar>
    <scalar dataType="xsd:double" dictRef="orbitalOccupancy">2.000000000000e0</scalar>
    <array size="10" dataType="xsd:double" dictRef="aoVector">3.306455447679e-1
3.187524188824e-1 2.591476635736e-18 9.610825759597e-18 -2.104953907844e-2 3.306455447679e-1
3.187524188824e-1 3.164500292222e-18 7.783890197431e-18 2.104953907844e-2
    </array>
  </list>
</list>
```

CML dictionary and conventions for plane wave molecular orbitals

- General plane wave and Gaussian molecular orbital structure would be pretty similar, however
  - Plane wave pseudopotentials are not standard
    - Pseudopotentials can be proprietary
    - Different definitions are used across software landscape
    - No CML convention has been developed
  - Important differences
    - Data tends to be much larger than Gaussian basis sets
      - Can be reduced if definitions of pseudopotentials and wave functions get standardized
    - Need to develop convention for atomicOrbitalDescriptions

```
<array size="10" delimiter="|" dataType="xsd:string" id="aoDescriptions"
dictRef="atomicOrbitalDescriptions">1 H s |1 H s |1 H px |1 H
py |1 H pz |2 H s |2 H s |2 H px |2 H py |2 H pz
```

CML Conventions and Dictionaries on the web

Firefox | CML Dictionary - Computational Chemi... | +

www.xml-cml.org/dictionary/compchem/

Actions Tools Validation Styles Transformations Views Stats S.O.A. Authoring Help

CML Dictionary

Computational Chemistry - Core Concepts

Namespace

The namespace of this dictionary is: `http://www.xml-cml.org/dictionary/compchem/`

Default Prefix

The default prefix for this dictionary is: `compchem`

Description

Toplevel dictionary for computational chemistry

Concepts in this dictionary are general throughout computational chemistry and are used extensively in the CompChem convention to describe the structure

---

Table of Contents

CML Conventions and Dictionaries on the web

Firefox | CML Dictionary - Computational Chemi...

www.xml-cml.org/dictionary/compchem/#e2Energy

Actions Tools Validation Styles Transformations Views Stats S.O.A. Authoring Help

CML Dictionary

2-electron energy has unit type `unitType:energy`

---

nuclear repulsion energy (ID: nuclearRepulsionEnergy)

Definition

The potential energy arising from Coulombic nuclei-nuclei repulsions.

$$\hat{T}_n = -\sum_i \frac{\hbar^2}{2M_i} \nabla^2_{R_i}$$

Description

The nuclear repulsion energy is the sum of the repulsive Coulombic interaction energies between the positively charged nuclei.

Data Type

nuclear repulsion energy is of data type `xsd:double`

Unit Type

nuclear repulsion energy has unit type `unitType:energy`

---

energy (ID: scfEnergy)

Definition

The Hartree-Fock Self-Consistent Field component of the energy.

Description

This is the SCF energy, where the SCF energy is a compment of another energy, such as the MP2 energy. The scf_energy term should NOT be used for the total energy of an SCF calculation, for this the total energy term should be used, as the exact meaning of the SCF energy will be properly determined by the parameters in the initialization module.

Data Type

energy is of data type `xsd:double`

Unit Type

energy has unit type `unitType:energy`

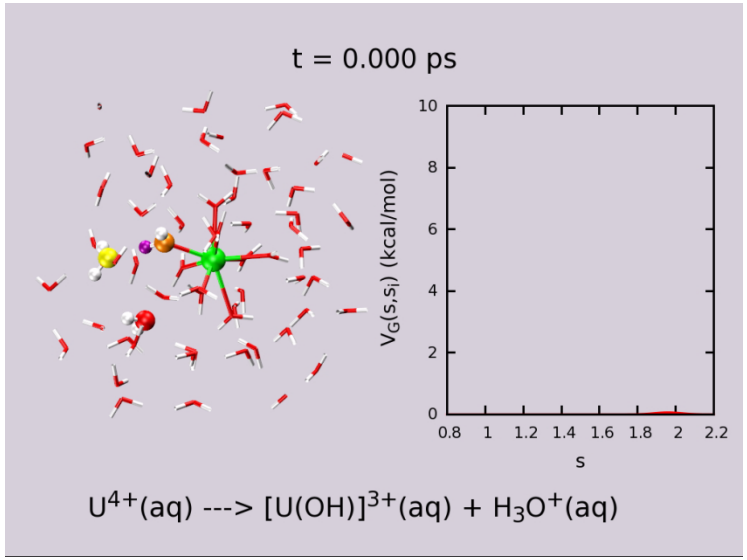
Handling large quantities of data within CML

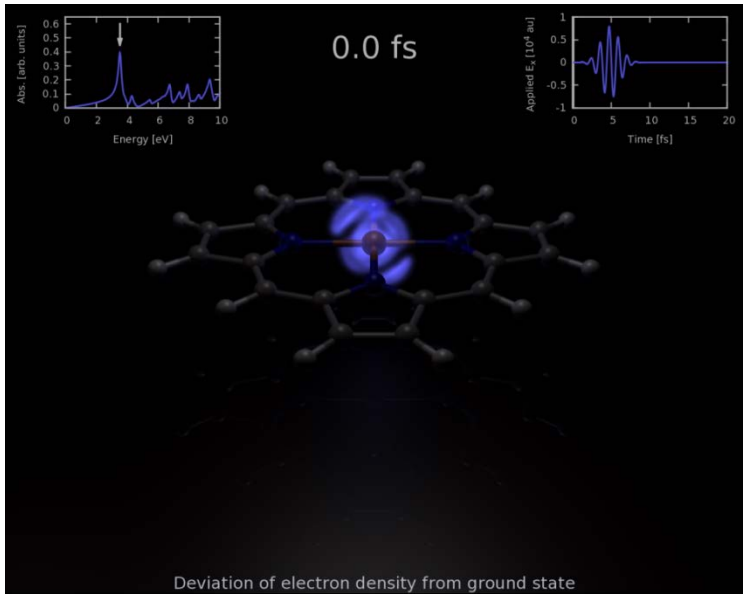
- Both Gaussian and plane wave molecular orbitals can get large
  - Example is a 2500 Gaussian basis function calculation would require 2500^2 doubles (or complex numbers)
  - Even bigger for plane wave orbitals (easily an order of magnitude)
  - Multiple pieces may need to be stored
- A single geometry is small, but a million geometries in a dynamics simulation trajectory becomes large
- Materials genome requires thousands to millions of the data above

Ongoing work with NWChem

- Get "all" NWChem data stored into CML output file
  - Better alignment of CML data writing with flow of NWChem
  - Adding dictionary entries and convention drafts in the process
- Reduce CML data by avoiding replication
  - Using IDs to reference things like basis sets and geometries defined in a module in different modules
- Handling bigger data blocks
  - Molecular orbitals (and others stored in other output files)
    - Normalization, component ordering, basis set, geometry
  - Trajectories
  - Use XDMF and link into CML

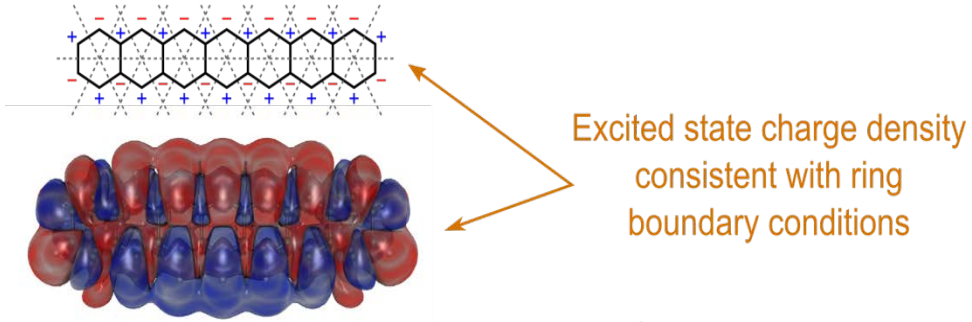
Linking raw data to visualization and interpretation





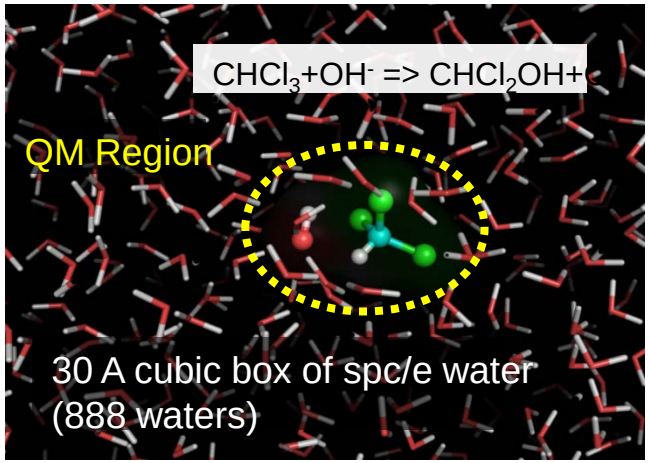
- Real time analysis and visualization of computer simulation for control
  - Creation of free energy surface through dynamics simulation
  - Discovery of rare event processes

- Density information from dynamical simulation can be visualized in various ways

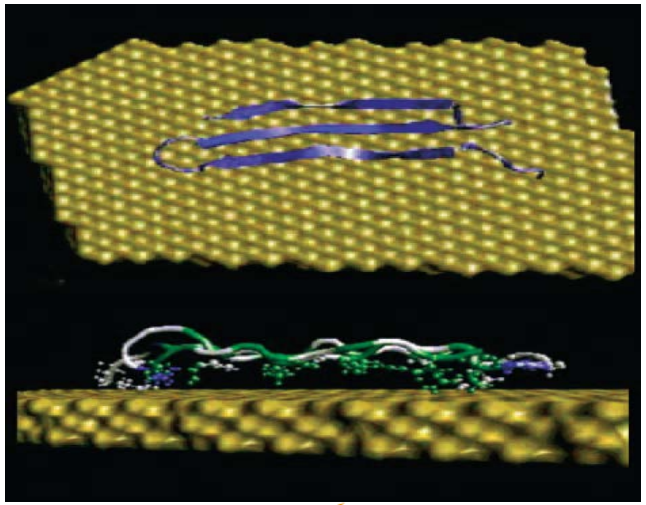


Further out: Representing complex simulations with many components

- One example is QM/MM
  - Linking quantum chemistry methods with molecular mechanics
  - Representations for parts of the system are different, need to describe the interface?!

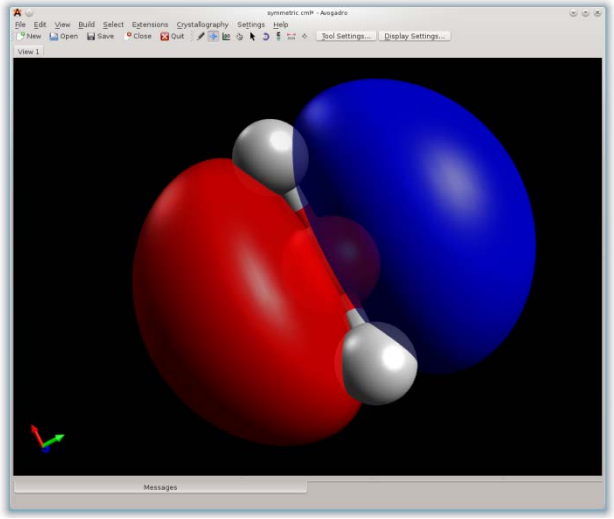


- Mesoscale and nanoscale, linking with continuum models, complex and interacting systems

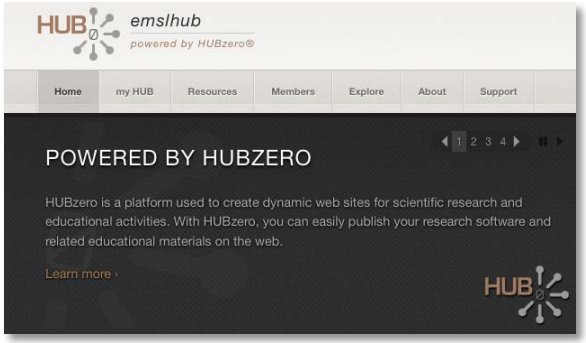


Utilizing NWChem's semantic data

- Open-source Avogadro can extract and visualize NWChem semantic output
  - Reads NWChem's CML and visualizes molecular orbitals and properties

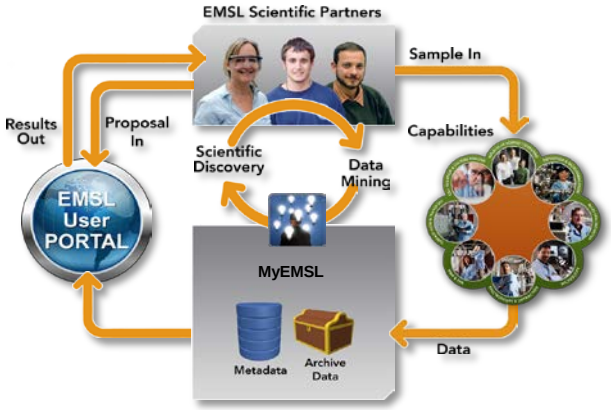


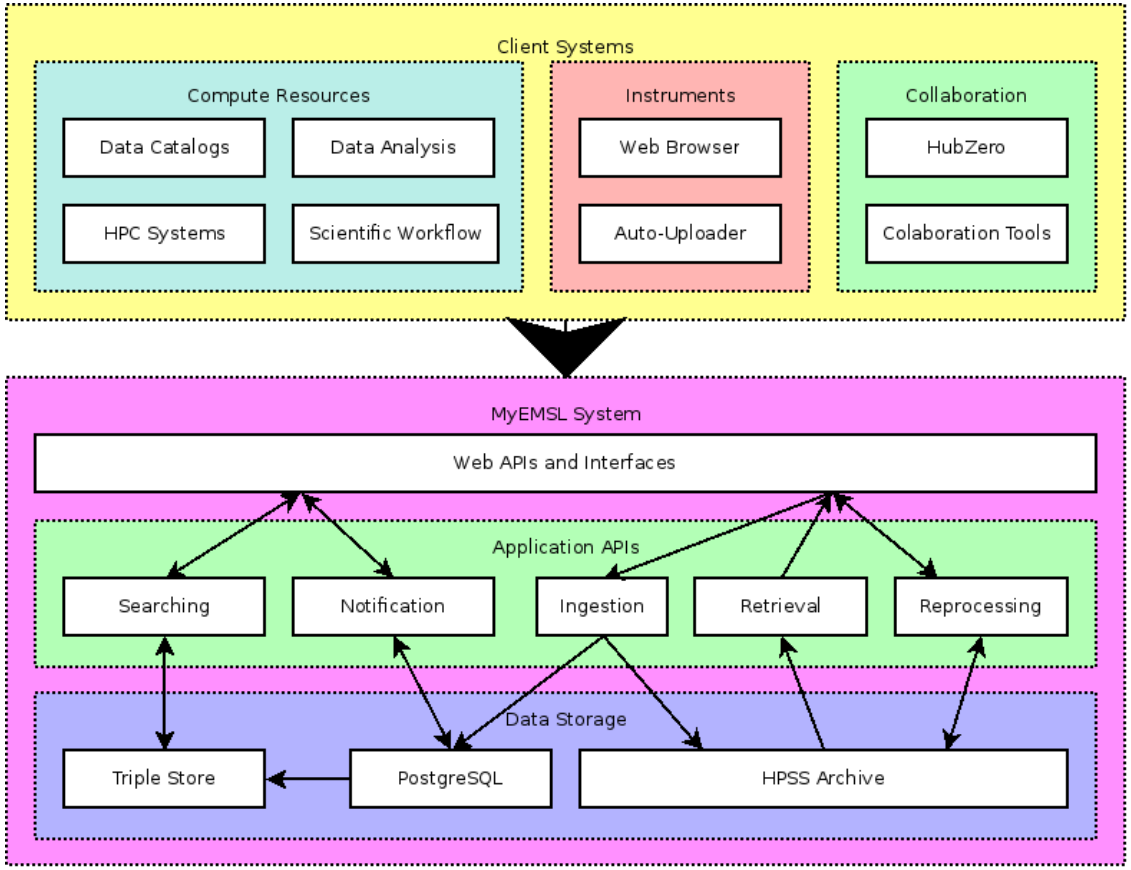
- Demonstrations of integrated access, and visualization of NWChem and NMR data using MyEMSL and EMSLHub



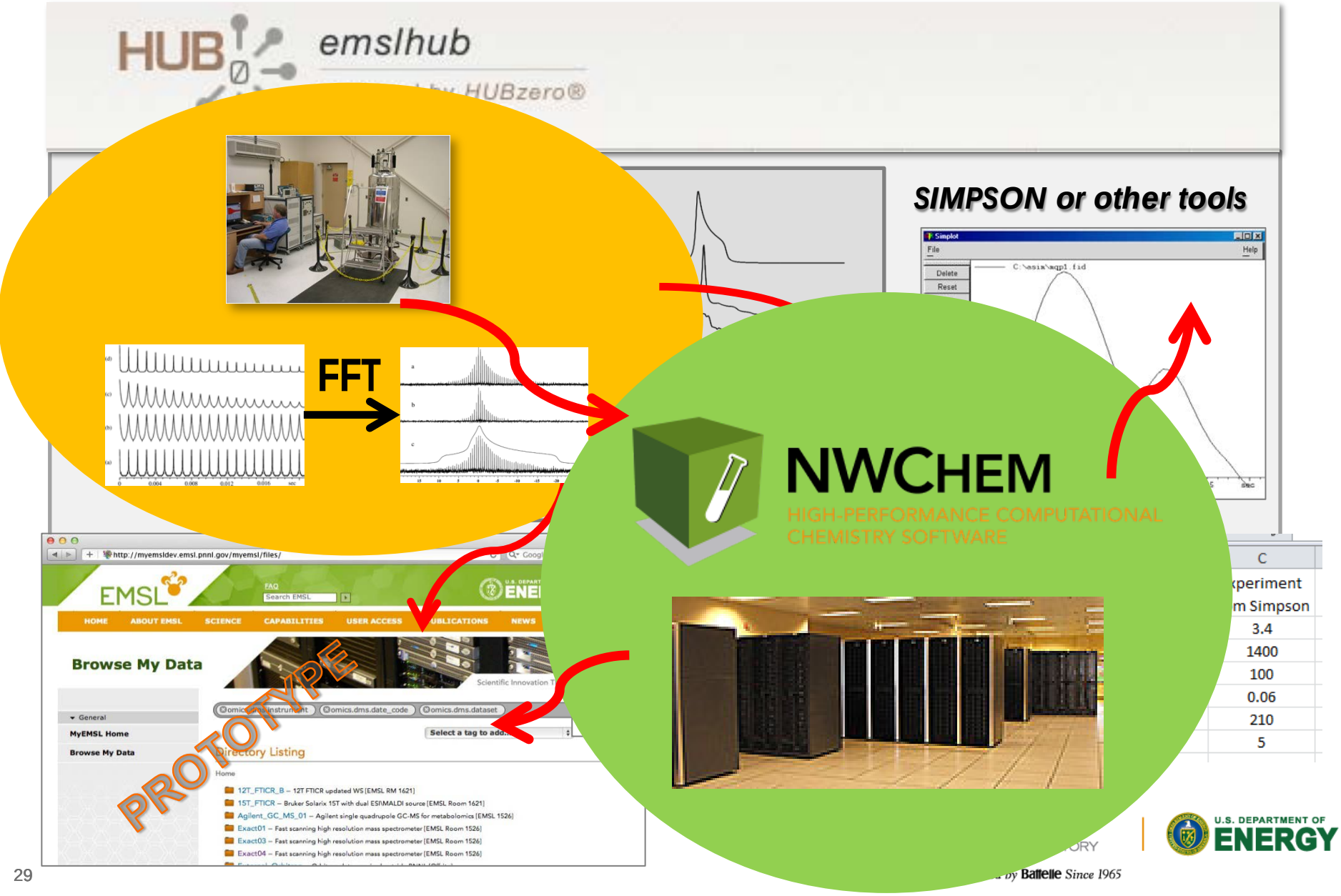
http://avogadro.openmolecules.net/wiki/Main_Page

Storing data: MyEMSL in a nutshell



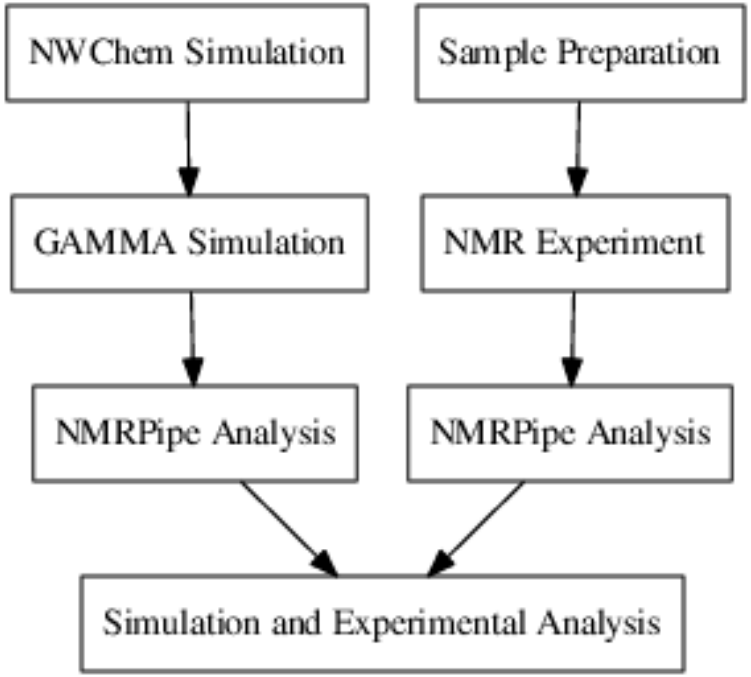


Utilizing NWChem's semantic data



NMR Integration Workflow

- NMR Experiment
  - NMR Capability at EMSL
  - General Workflow
- Simulation
  - NWChem
  - GAMMA
- Analysis and Visualization
  - NMRPipe



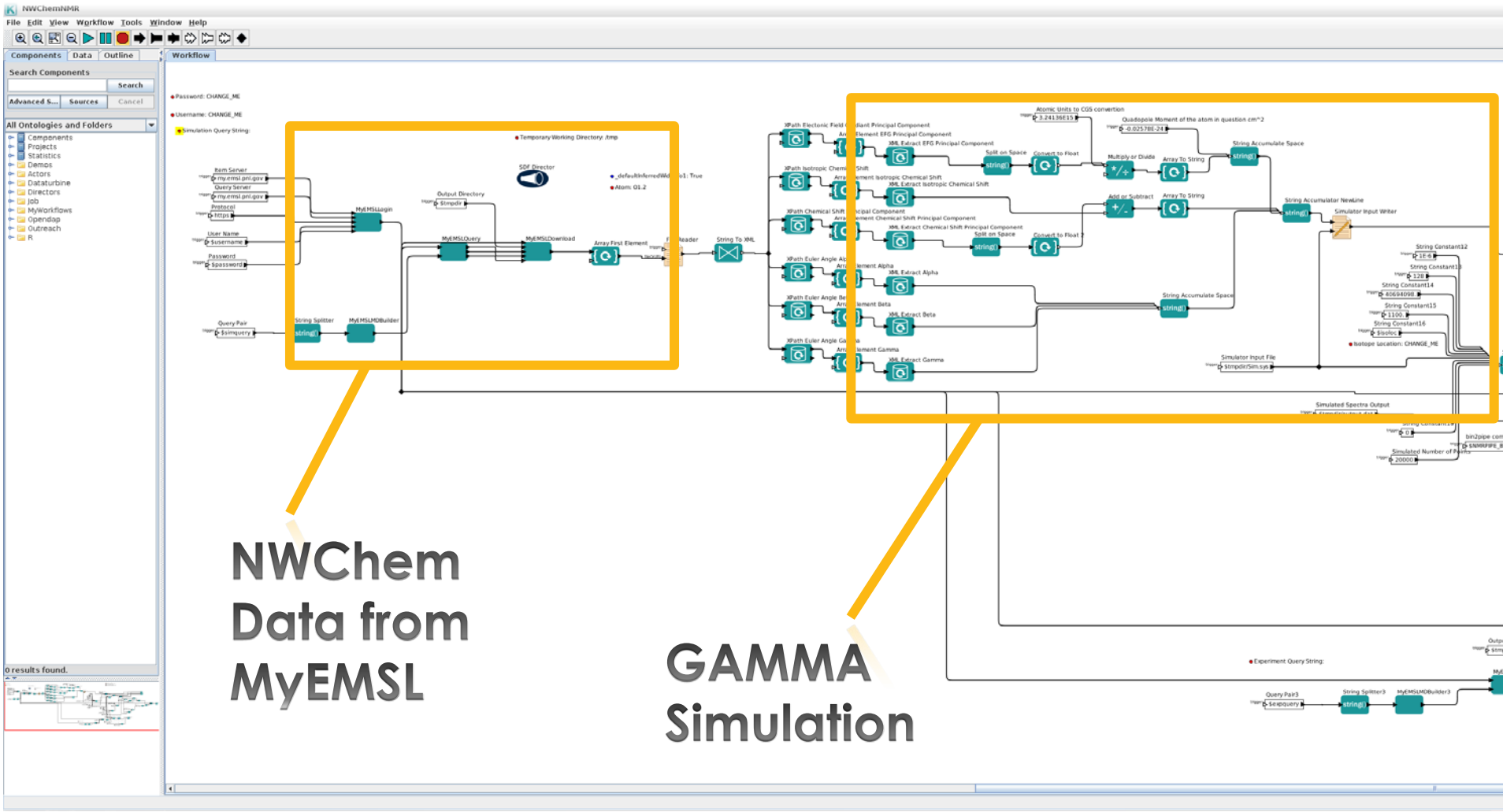
NMR Integration Workflow Inputs

- MyEMSL Query to get CML
- MyEMSL Query to get Experimental Data
- MyEMSL Authentication
- Isotope Information for GAMMA
- Parameters for GAMMA
  - Field Strength, Number of Points
  - NWChem CML Inputs
- Atom to Simulate Spectra
- NMRPipe Command Line Parameters
  - Simulation
  - Experiment
- Upload Metadata

Kepler Scientific Workflow System

- Kepler
  - http://www.kepler-project.org
- Desktop Application
- Directed Acyclic Graph Workflows
- Components for doing Scientific Work
- Open Source Community
- Integrate with MyEMSL and HPC Systems.

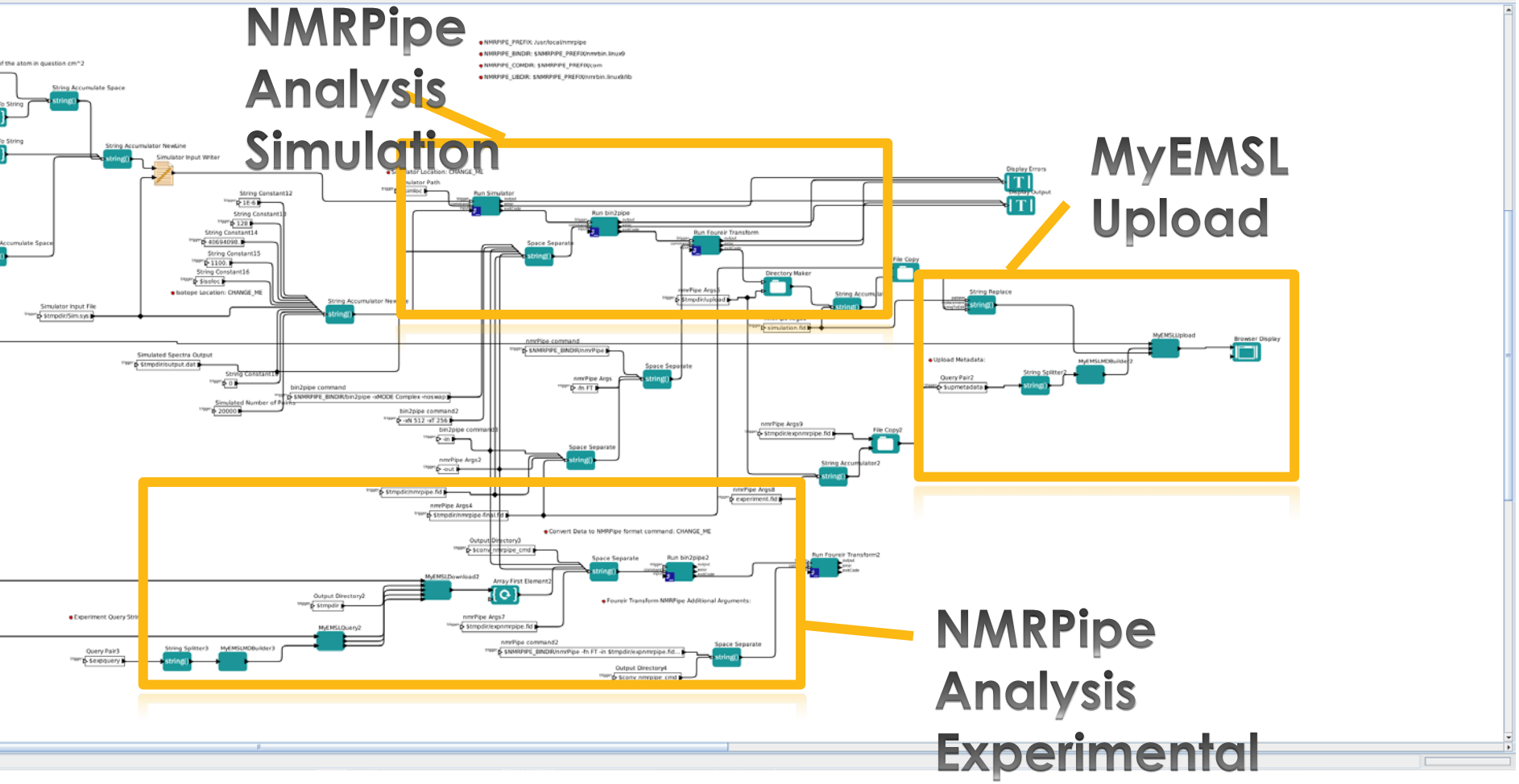
Defining the workflow with Kepler Automation: Simulation to Spectrum



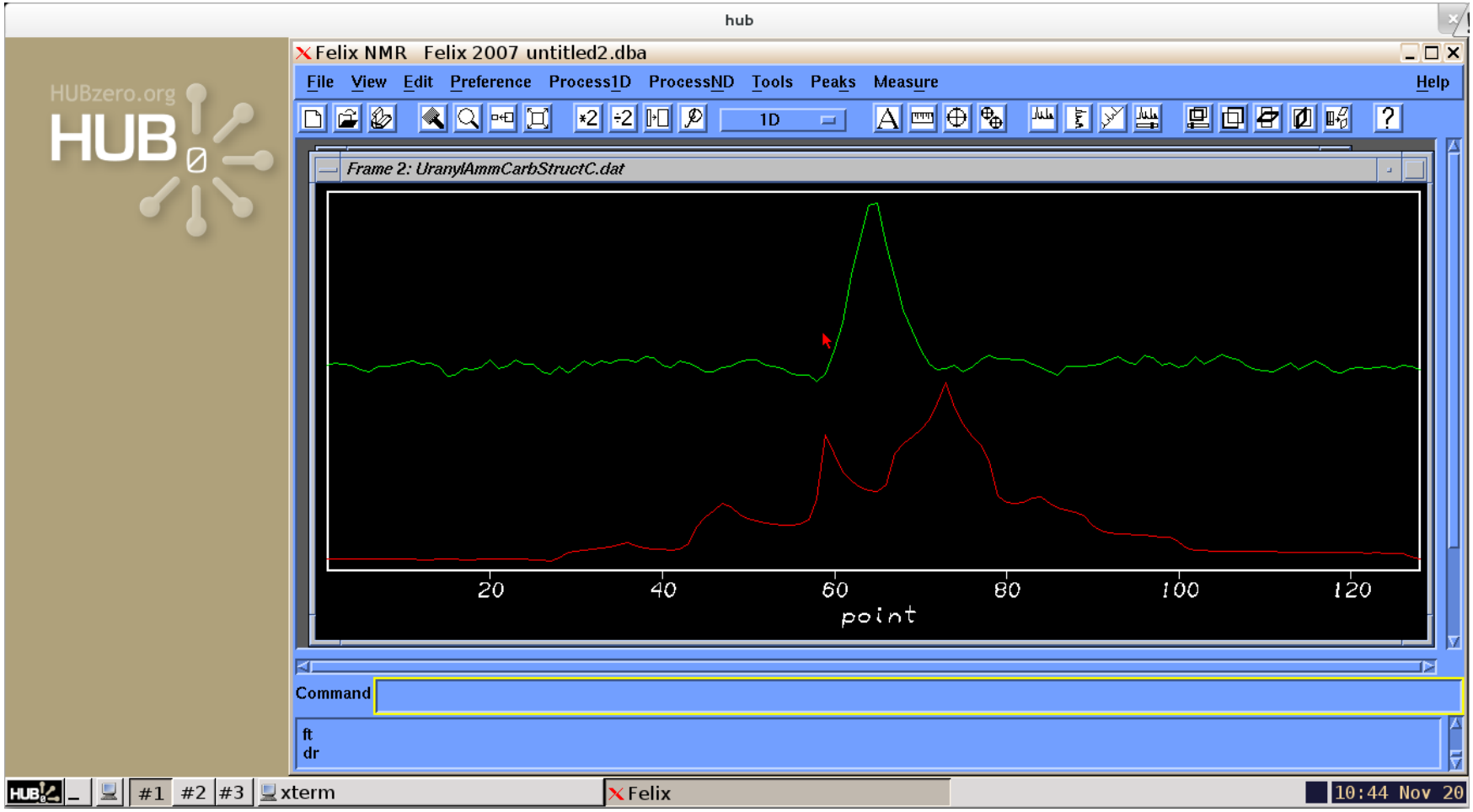
NWChem Data from MyEMSL

GAMMA Simulation

Defining the workflow with Kepler Automation: Spectrum to Visualization



Visualization in EMSLHub



Summary

- NWChem produces CML data
  - For Gaussian basis set modules
  - Utilizing FoX library
- Avogadro reads and visualizes CML data
  - Visualizing molecular orbitals
- FoX library expanded with new functionality
  - Can be utilized by other computational chemistry codes to produce CML data
- Drafts for CML language and conventions defined
  - Molecular orbitals
  - NMR and other properties

EMSL

This research was performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

NWChem development is funded by:

US Department of Energy BER, ASCR, BES offices

PNNL LDRD