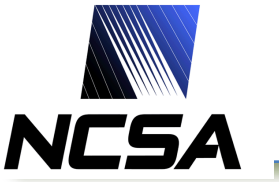
Discovery Environments for Molecular Interactions

Sudhakar Pamidighantam
Science Gateways Group, RT-UITS
Indiana University
pamidisg@iu.edu



Outline

- Discovery Environments
- Some Molecular Science Gateways
- SEAGrid Science Gateway
- Diffraction Workflows using optimized LAMMPS implementation
- Force Field Parametrization workflows
- Data projects in SEAGrid
- Cyberinfrastructures for Medicine

NCSA

INDIANA UNIVERSITY

Motivation

Integrating Services for E-Science and Engineering in Research, Education and Training

Software

- Reasonably Mature and easy to use to address scientists' questions of interest

Community of Users

- Need and capable of using the software

  Some are non traditional computational scientists

  Experimentalists, Non-domain experts

Resources

- Various in capacity and capability
- Distributed and heterogeneous

Discovery Environments

- Integrated HPC Resources
  - Hardware Resources
    - CPU
    - Networks
    - Data Storage Resources
  - Software Resources
    - Application Software
    - Application enabling software
    - Scheduling Software
    - Workflow managers
- Data
  - Hardware specific data
    - System Description (Cores/Accelerators Chip type/OS type etc..)
    - Capacity/Capability
    - System/Queue Level Restrictions
    - System Level Load
    - Queue Level Load
  - Research Data
  - Provenance

- Application Software Specific Data
  - Application Description
  - Capabilities
  - Input Requirements
  - Outputs
  - Pre-processing
  - Post-processing
  - Performance
- Application Enabling Data
  - Soft-environments/ Modules/ Dependent Environments (Compilers, Math Libraries etc...)
  - Encryption
  - IO Specifics
  - Common Environments (Across the Grid)
  - Local Environments (Site Specific)

- Goal of Cyberenvironments is to provide end to end solutions for scientific research
- What is End to End: Integrated (HPC) Resources, Data and Collaboratories to support complete research life cycle
- Problem: The research goals are changing constantly
- Solution: the environments should be adaptable to changes required by research communities well supported with stable and dependable services.

NCSA

INDIANA UNIVERSITY

Outline

- Discovery Environments
- Some Molecular Sciences Gateways
- SEAGrid Science Gateway
- Diffraction Workflows using optimized LAMMPS implementation
- Force Field Parametrization workflows
- Data projects in SEAGrid
- Cyberinfrastructures for Medicine

Science gateways

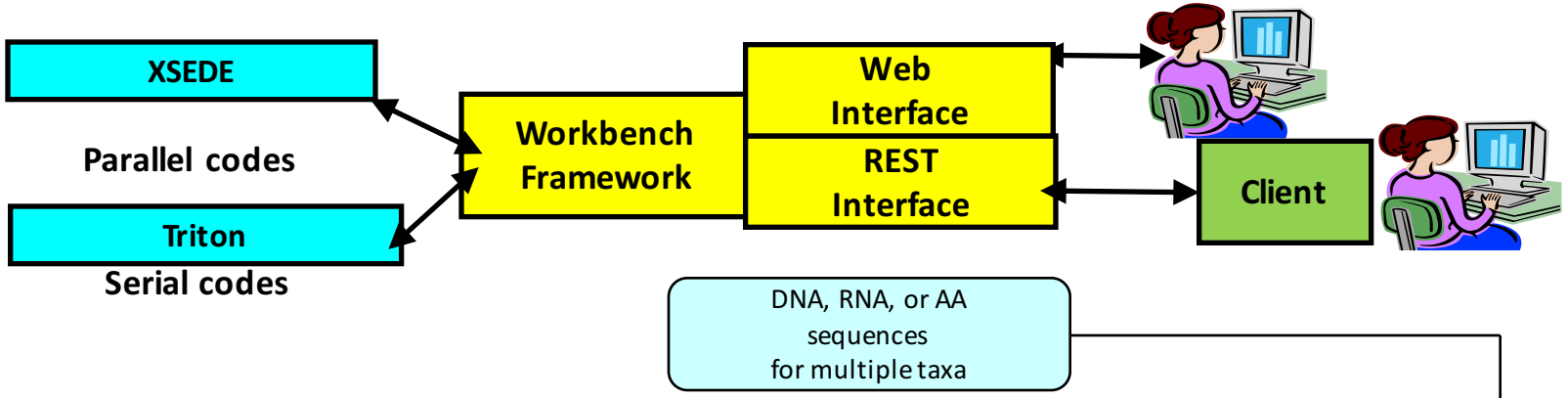
Some Molecular Sciences Oriented Gateways in XSEDE

Gateway	Function/Field
SEAGrid/GridChem	Chemistry
CIPRES	Phylogenetics
Ultrascan	Hydrodynamic with AUC
BioDrugScore	Protein Drug Docking and Drug scores
Rosie	Protein Structures
ParamChem	Forcefield Parametrization
IPlant Collaborative	Plant Genetics Analysis
GAAMP	Automated Atomic Model Parametrization
Robetta	Protein Structure and Interactions
IntegromeDB	Systems Biology
Neuroscience Gateway	Neuroscience
Vlab Gateway	Earth and Planetary Materials

NCSA

INDIANA UNIVERSITY

CIPRES Gateway and Phylogenetic Workflows

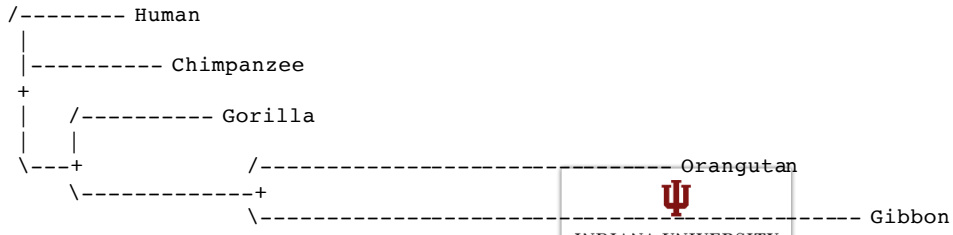


DNA, RNA, or AA sequences for multiple taxa

Multiple sequence alignment is matrix of taxa vs characters

```
                        .   .   .    ......  .   .
Human           AAGCTTCACCGGCGCAGTCATTCTCATAAT...
Chimpanzee      AAGCTTCACCGGCGCAATTATCCTCATAAT...
Gorilla         AAGCTTCACCGGCGCAGTTGTTCTTATAAT...
Orangutan       AAGCTTCACCGGCGCAACCACCCTCATGAT...
Gibbon          AAGCTTTACAGGTGCAACCGTCCTCATAAT...
```

Final output is phylogeny or tree with taxa at its tips



Multiple sequence alignment: ClustalW, MAFFT, …

Aligned sequences

Phylogenetic tree inference: MrBayes, RAxML, …

NCSA

INDIANA UNIVERSITY

Science Gateways

Each code's parameters are specified by the user in a web form…..

Folders

- cactusresponsible
  - Data (1)
  - Tasks (2)
- cactusresponsible
- maryam
- maryam
- will
- mafft carna
- mary
- david
- dppdiv
- Mathieu
- alex mr bayes
- ruth_bone
- b_brito
- martha
- parra
- stine
- guin (frog.girl)
- emily
- lkm
- guan
- beast2
- nclconverter test
- warren
- agnes
- test autoclose sed
- david_bass
- peggy_mauve
- test
- pawan
- dave
- manuel
- thornhill
- janus
- sarahm

Create new task

Task Summary | Select Data | Select Tool | Set Parameters

RAxML-HPC BlackBox: Phylogenetic tree inference using maximum likelihood/rapid bootstrapping on XSEDE. (Alexandros Stamatakis)

Simple Parameters

Maximum Hours to Run (click here for help setting this correctly) * 0.25

Sequence Type * Protein Nucleotide

Outgroup (one or more comma-separated outgroups, see comment for syntax)

Constraint (-g)

Binary Backbone (-r)

Use a mixed/partitioned model? (-q)

Create an input file that excludes the range of positions specifed in this file (-E)

Estimate proportion of invariable sites (GTRGAMMA + I) * yes no

Protein Substitution Matrix * JTT

Use empirical base frequencies? * yes no

Find best tree using maximum likelihood search ☑

Let RAxML halt bootstrapping automatically (HIGHLY recommended) * ☑

Don't use BFGS searching algorithm (--no-bfgs) * ☐

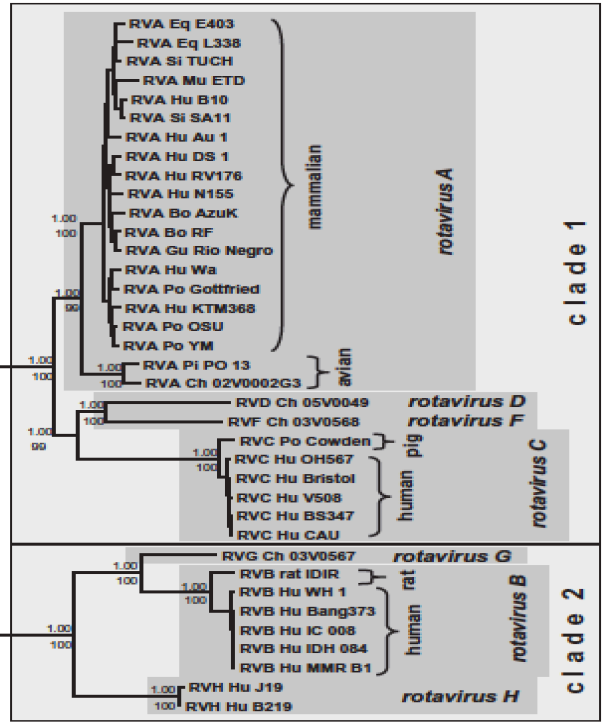
Print branch lengths (-k) ☐

Advanced Parameters

INDIANA UNIVERSITY

Phylogenetics Applications in CIPRES gateway

Code & version	Parallelization	Cores	Computer
MAFFT 7.037	Pthreads	8	Trestles
BEAST 1.7.5	Pthreads	8	Trestles
GARLI 2.0	MPI	≤32	Trestles
MrBayes 3.1.2h	MPI/OpenMPI	10 to 32	Gordon
MrBayes 3.2.1	MPI	8 to 16	Gordon
RAxML 7.6.6	MPI/Pthreads	8, 30, or 60	Trestles
RAxML-Light 1.0.9	bash/Pthreads	≤1,000	Trestles



472,832 TeraGrid/XSEDE jobs submitted by 12,667 unique users.
Average of 297 new XSEDE users registered in each of the last 12 months.
81.3 million core hours of TeraGrid/XSEDE time distributed to scientists.
Used for curriculum delivery by at least 76 instructors.
Supported at least 1570 publications.

NCSA

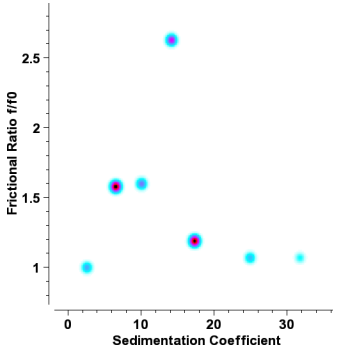
INDIANA UNIVERSITY

Science Gateways

UltraScan3

Analytical Ultracentrifuge Data Analysis Gateway

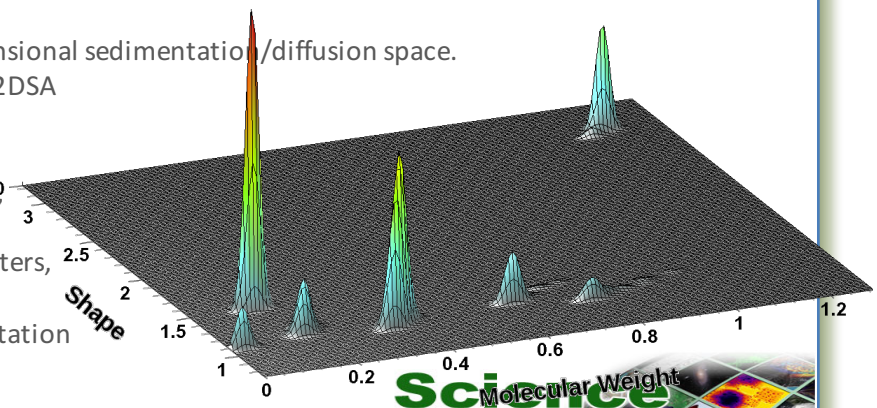
14 Samples at a time, physiological environments, diverse range (10^2-10^8 D) oligomerization states of reversible self- or hetero-associations, ligand binding, slow kinetics and Kd



•How many components? What are their sizes and molecular weights?
•What are their anisotropies?
•What is the partial concentration of each component?
•Conformational analysis-folding/melting properties
•Do the components interact (how fast, strong)?

Data Analysis

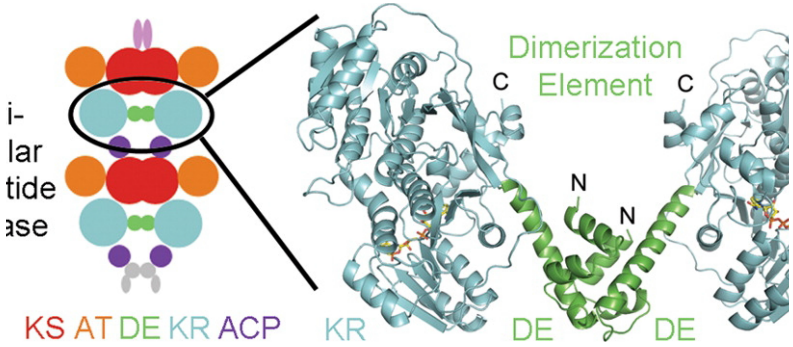
•2-dimensional Spectrum Analysis (2DSA): High-resolution, general and model independent solution for size and anisotropy distributions of non-interacting systems
•Parametrically Constrained Spectrum Analysis (PCSA): Identifies size/anisotropy relationships for polymerizing systems and provides a constrained fit over the 2 dimensional sedimentation/diffusion space.
•Custom Grid Analysis (CG): Takes advantage of prior knowledge to parameterize the 2DSA grid in terms of alternate hydrodynamic variables.
•Genetic Algorithms (GA): Robust non-linear least squares optimization method that provides parsimonious regularization of 2DSA spectra. Also used for fitting of discrete, non-linear models (reversible association, non-ideality, co-sedimenting solvents).
•Monte Carlo Analysis (MC): Used to measure the effect of noise on the fitted parameters, yields parameter distribution statistics
•van Holde – Weischet Method (vHW): Used to generate diffusion-corrected sedimentation profiles which provide finely detailed comparisons between multiple samples.



INDIANA UNIVERSITY

Science Gateways

Polyketide Synthase Dimerization Studies



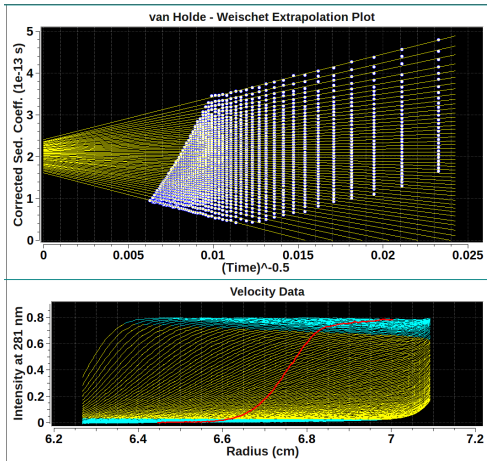
Discrete Model Genetic Algorithm Analysis

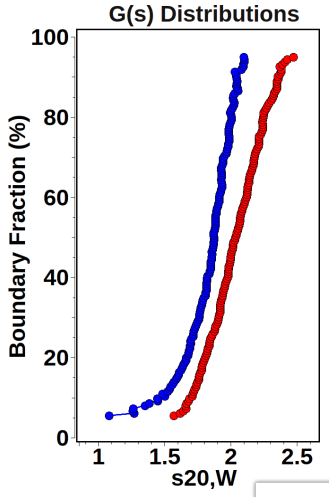
Data Report for Run: *011415_336v1*
Cell 3, Channel B, Wavelength 281, Edited Dataset 1501152211

Data Analysis Settings:

Model:	**monomer-dimer reversible self-association**	
RMSD:	**0.00442867**	
Mol. Weight:	**1.9300e+04**	**(fixed)**
Weight Average s20,W:	**1.6722e-13**	**(fitted)**
Weight Average D20,W:	**9.0958e-07**	**(fitted)**
Total Concentration:	**8.0422e-01**	**(fitted)**
Partial Spec. Vol.:	**0.769186 ml/g**	**(fitted)**
Kd:	**2.6473e-05 Mol**	**(fitted)**
k_off:	**6.6056e-06/sec**	**(fitted)**

Species:	**Molec. Wt.**	**s20,W**	**D20,W**	**f/f0**
Monomer:	**1.9300e+04**	**1.6722e-13**	**9.0958e-07**	**1.3050e+00**
Dimer:	**3.8600e+04**	**2.7038e-13**	**7.3537e-07**	**1.2812e+00**



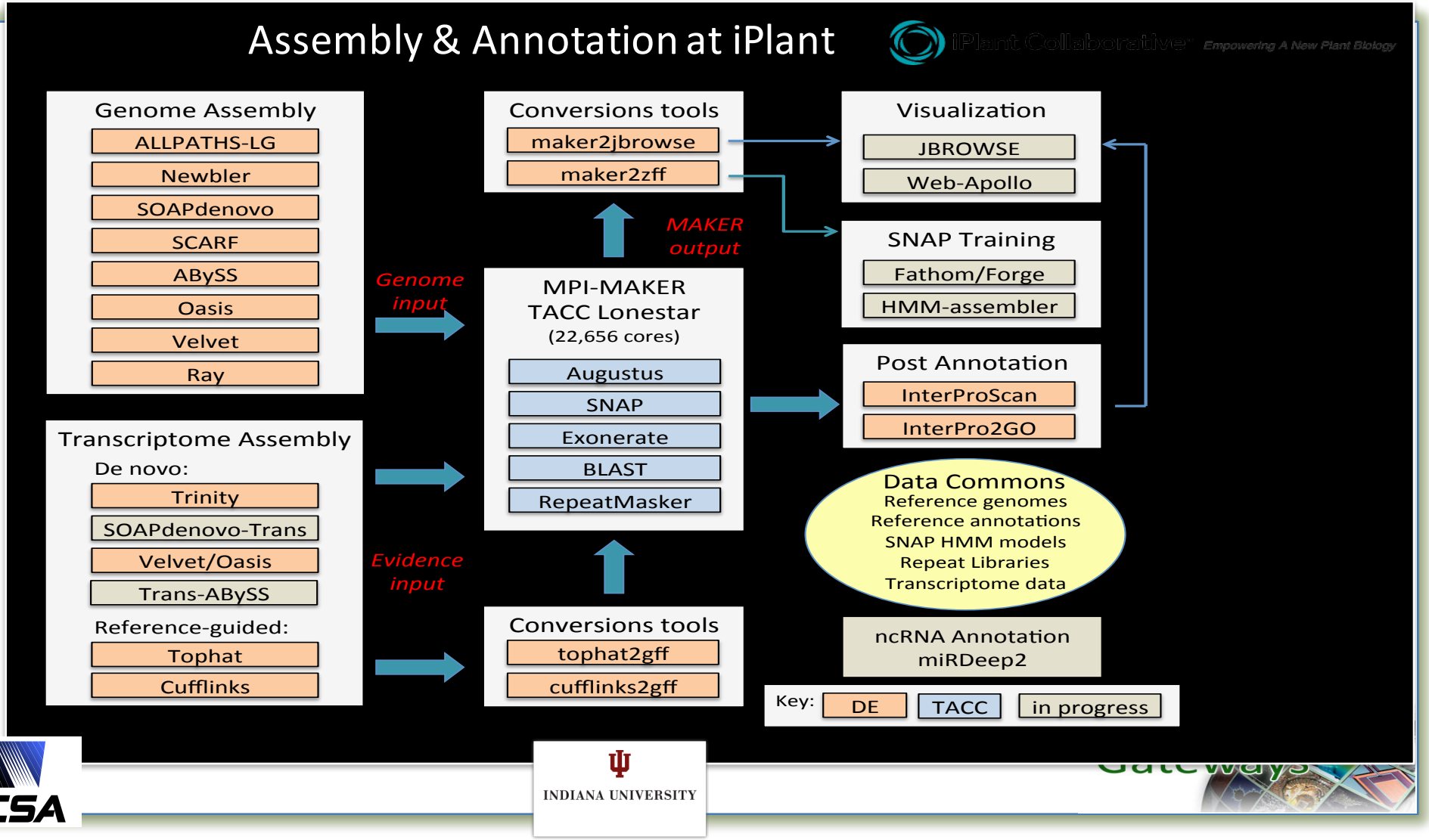


Jianting Zheng, Chris D. Fage, Borries Demeler, David W. Hoffman & Adrian T. Keatinge-Clay.
The missing linker: a dimerization motif located within polyketide synthase modules. ACS Chem. Biol. (2013)

INDIANA UNIVERSITY

Science Gateways

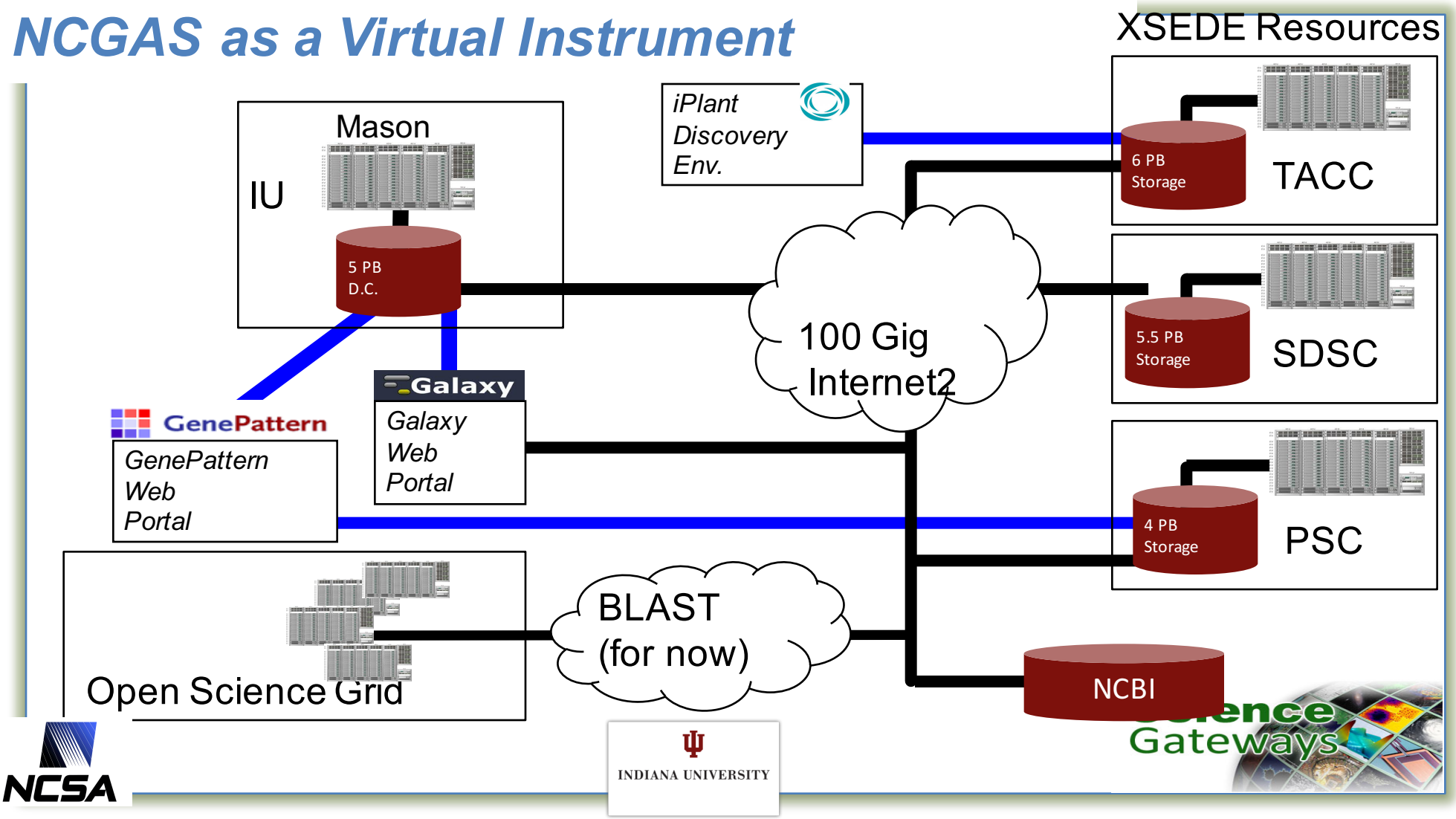
Assembly & Annotation at iPlant



NCSA

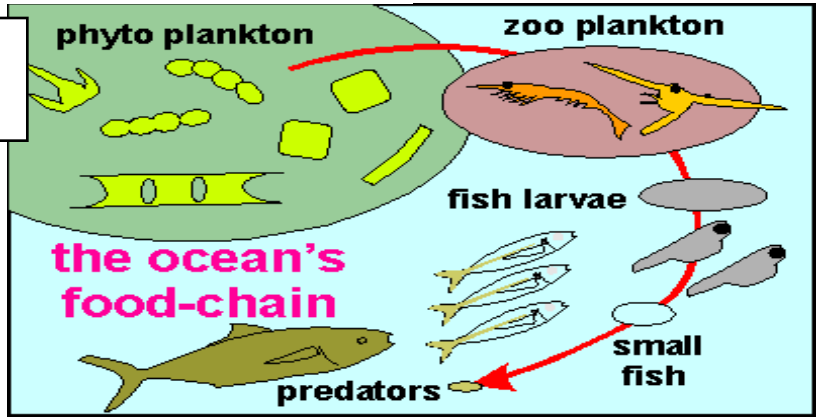
INDIANA UNIVERSITY

NCGAS as a Virtual Instrument



NCGAS Science Stories

Transcriptomes of zooplankton Calanus finmarchicus correlate climate change with decrease in zooplankton and fisheries decline



Study of complete RNA collection of fruit fly uncovers unprecedented complexity in Drosophila melangaster, identifying thousands of new genes, transcripts, and proteins

Outline

- Discovery Environments
- Some Molecular Science Gateways
- SEAGrid Science Gateway
- Diffraction Workflows using optimized LAMMPS implementation
- Force Field Parametrization workflows
- Data projects in SEAGrid
- Cyberinfrastructures for Medicine

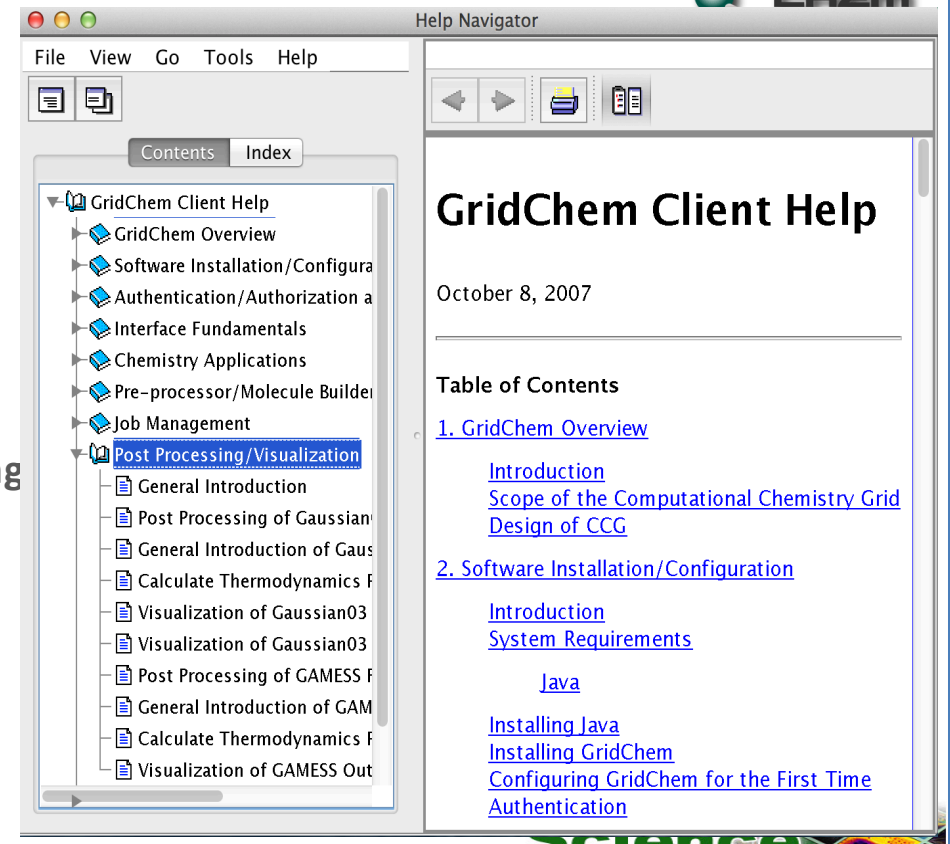
NCSA

INDIANA UNIVERSITY

SEAGrid Virtual Organization

- **Allocation**
  - https://www.gridchem.org/allocations/index.shtml
  - Community and External Registration
  - Reviews, PI Registration and Access Creation
  - **Community User Norms Established**
- **Consulting/User Services**
  - https://www.gridchem.org/consult
  - Ticket tracking, Allocation Management
- **Documentation, Training and Outreach**
  - https://www.gridchem.org/doc_train/index.shtml
  - FAQ Extraction, Tutorials, Dissemination
- **Application deployment, Integration and scripting**
- **Middleware Services and Client**
- **Usage**
  - **569 Users, 15,000 Jobs, 11M XSEDE SUs used last year,**
  - **More than 100 Publications, 10 Dissertations**

Publications: **https://www.gridchem.org/papers/index.shtml**

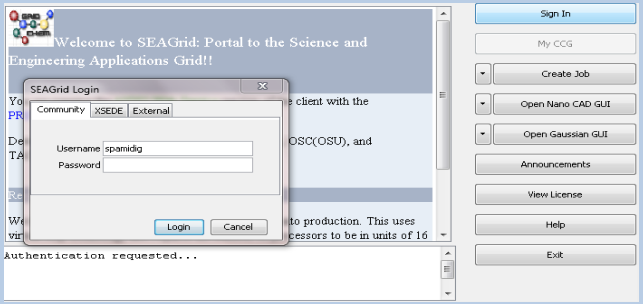


NCSA

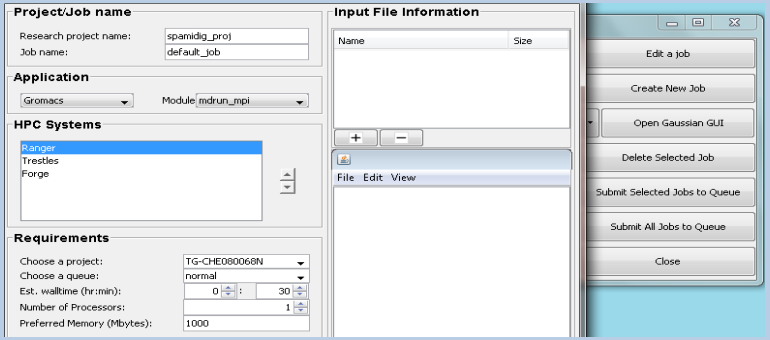
INDIANA UNIVERSITY

Science Gateways

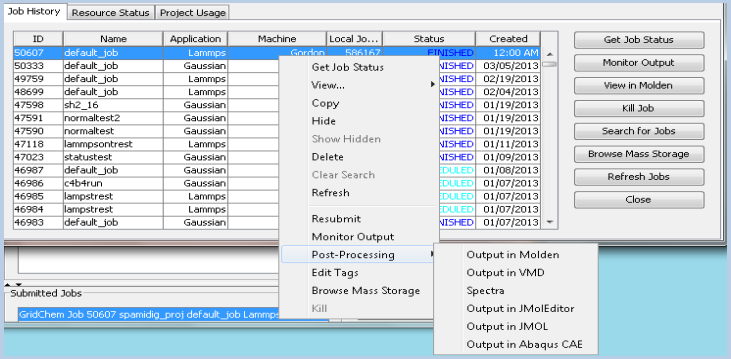
Functions - Client View



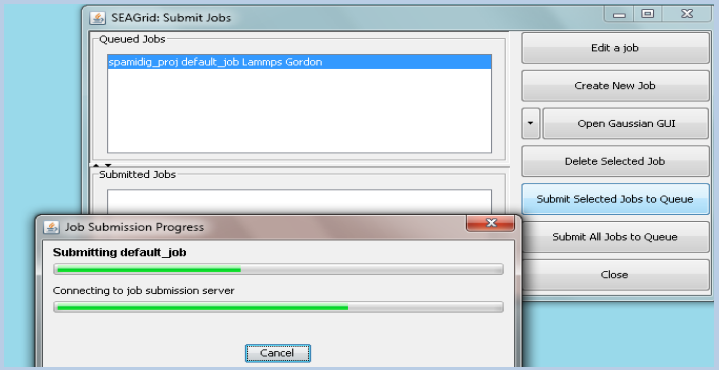
Authentication



Input Selection Job composition



Job Monitoring and Post Processing

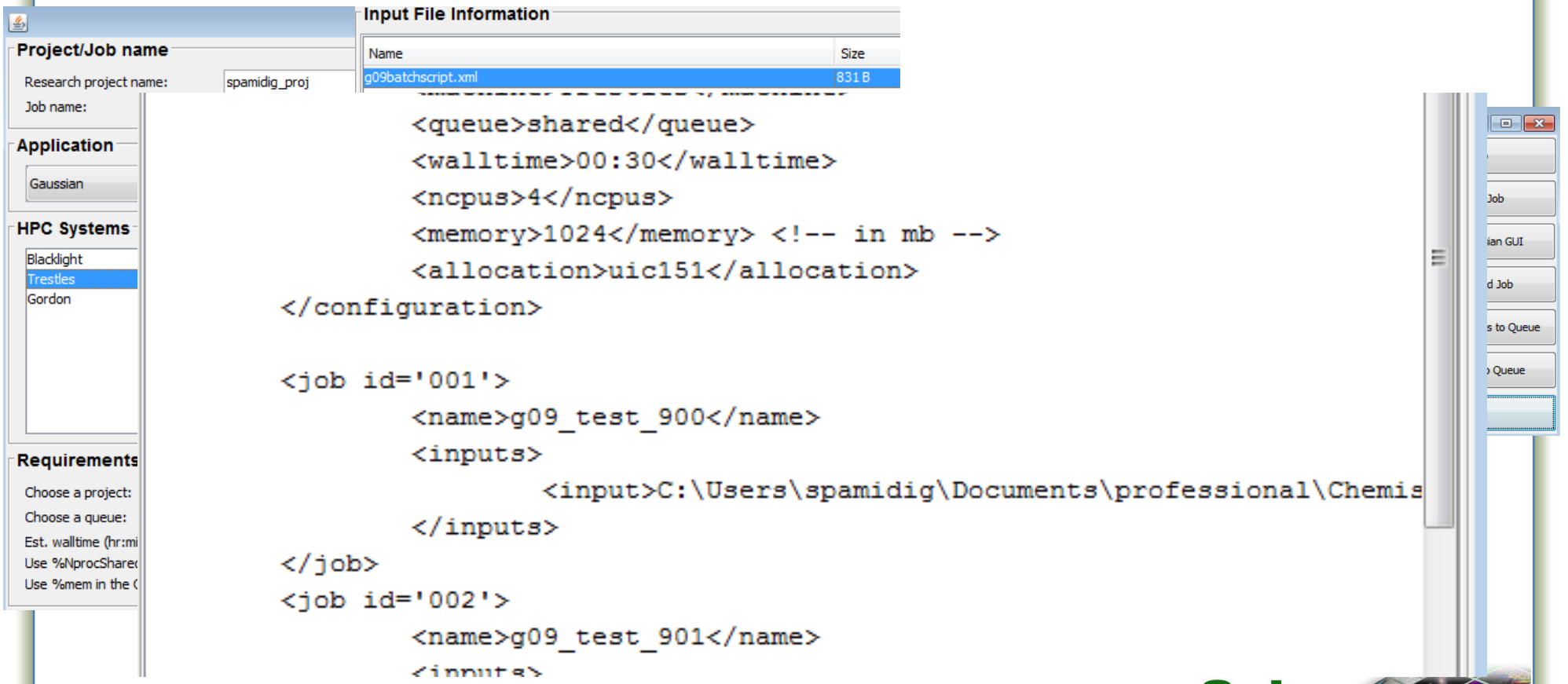


Job Submission

NCSA

INDIANA UNIVERSITY

Script Based high-throughput Job Creation



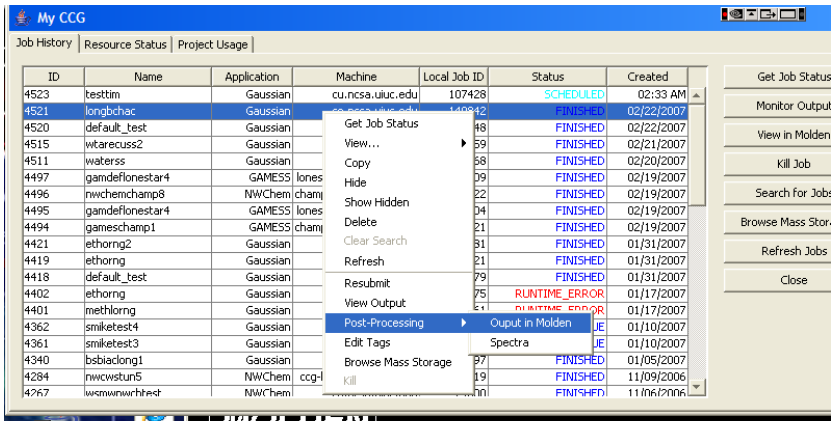
NCSA

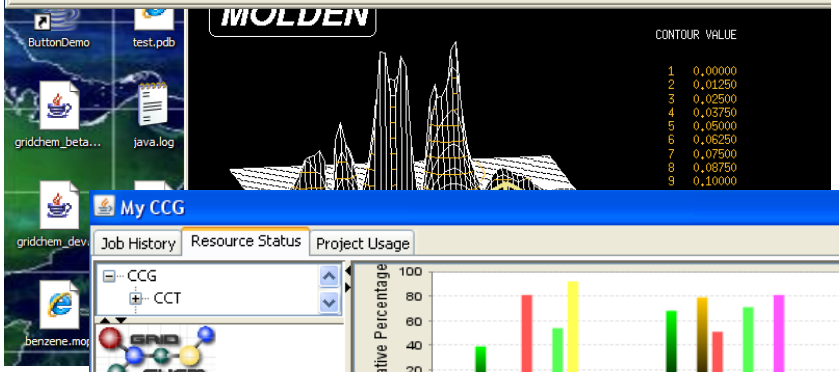
INDIANA UNIVERSITY

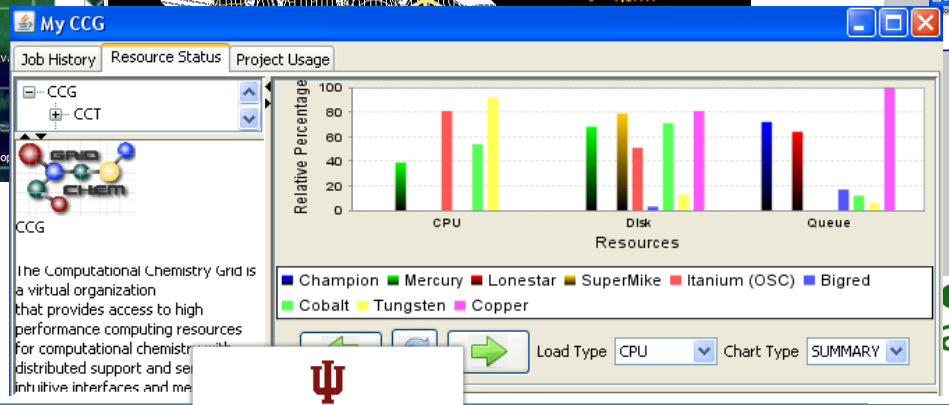
Science Gateways

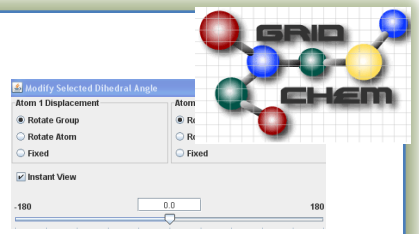
Applications and Postprocessing

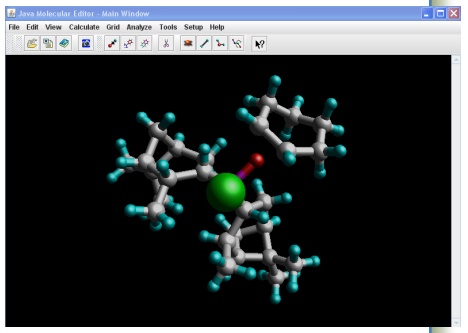
Application	Model
Gaussian	QM/QMMM
GAMESS	QM/QMMM
NWChem	QM/QMMM/MD
Molpro	QM
QChem	QM/QMMM
DDScat	Electron Scattering
Amber	MD
Castep	QM
DMol3	QM
Lammps	MD
Molca	Excited States
Abinit	Periodic Systems
Tinker	MM/Monte Carlo
DFTB+	Semi Empirical QM

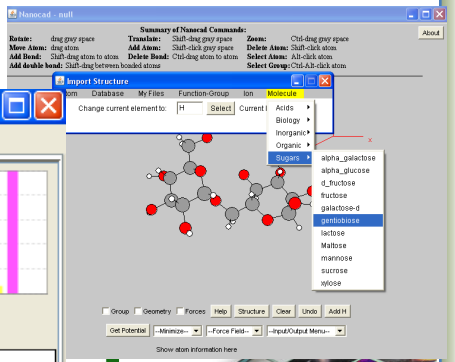








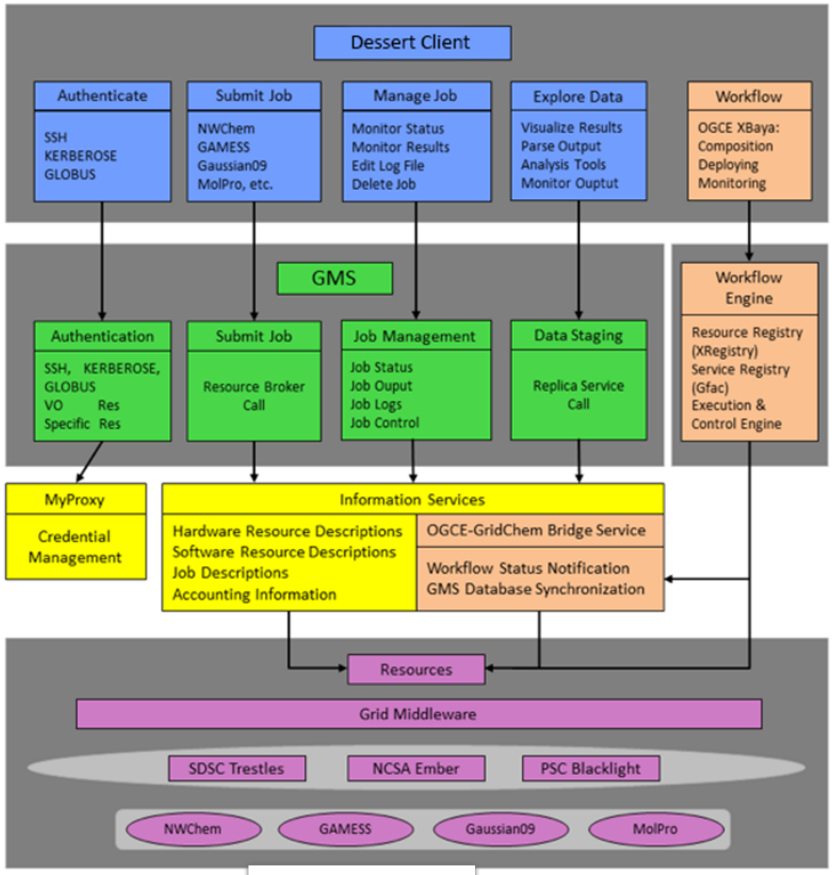




NCSA

INDIANA UNIVERSITY

Schematic Architecture and Grid Middleware Services

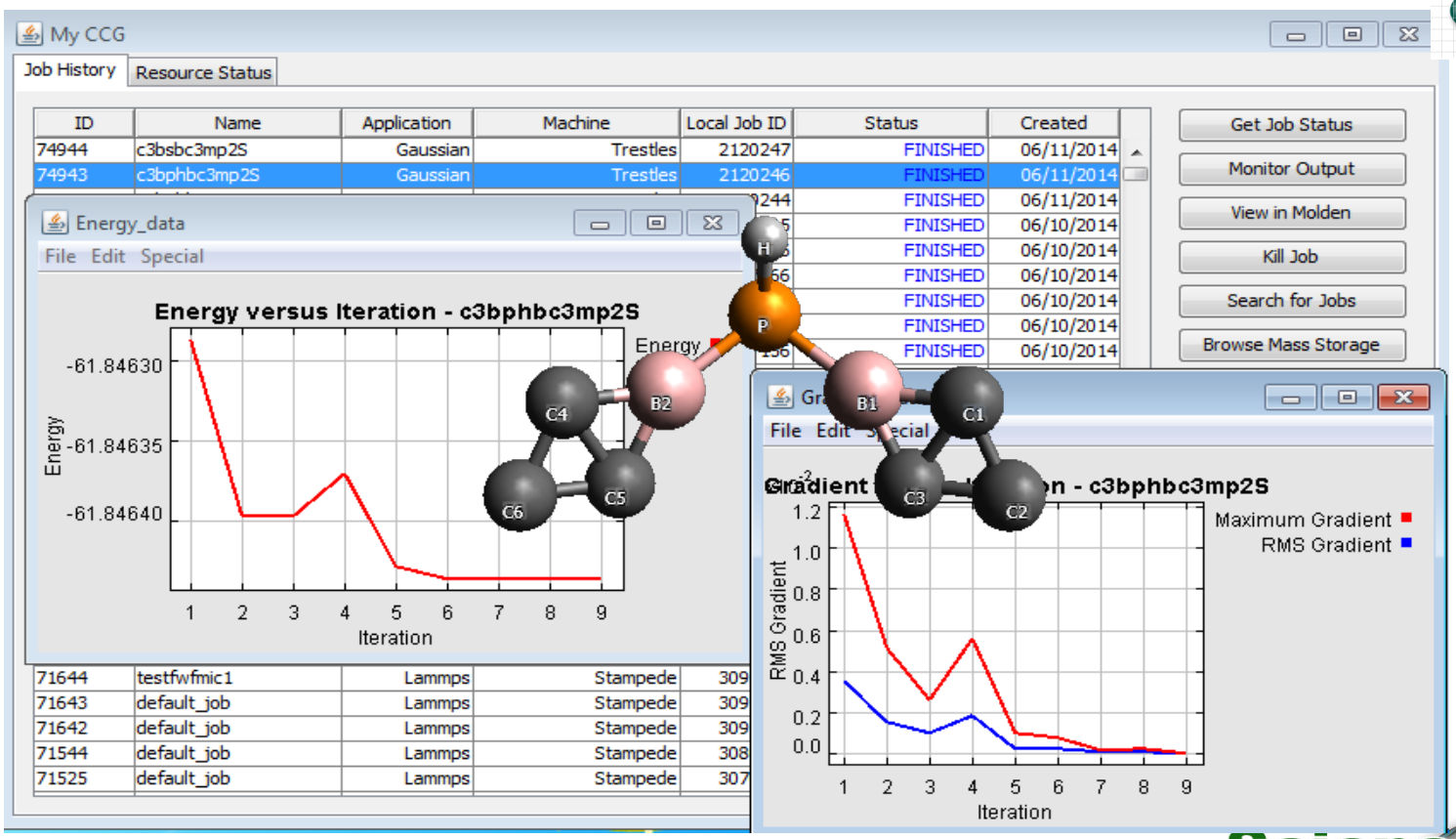


NCSA

INDIANA UNIVERSITY

Science Gateways

Monitoring QM optimization run



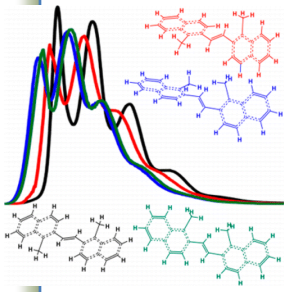
NCSA

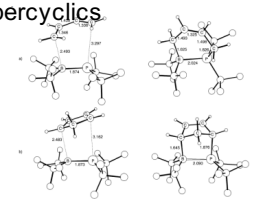
INDIANA UNIVERSITY

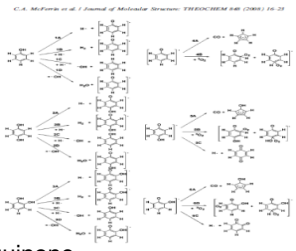
Science Gateways

SEAGrid User Research

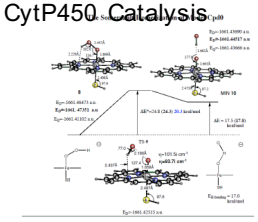
Phosphinoborane percyclics

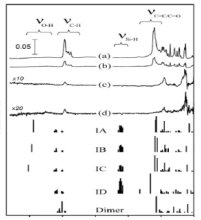




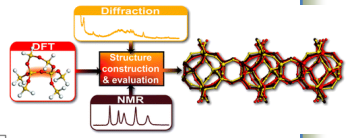


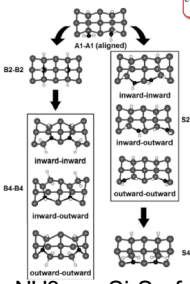
CytP450 Catalysis

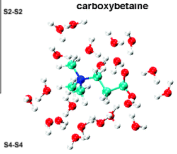


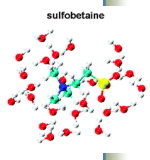


FTIR of Heptanedione on Si



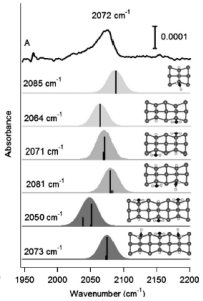




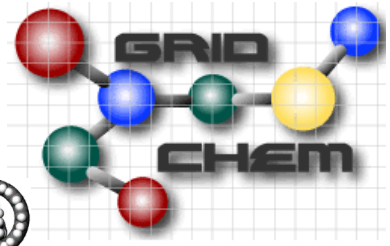


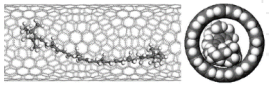
NH3 on Si Surfaces

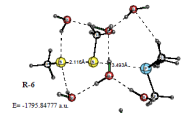
Semiquinone reactions

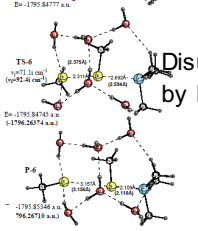


Si Surface IR

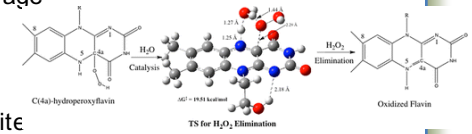




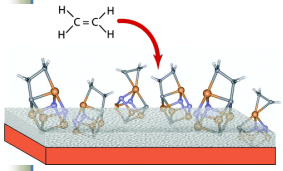


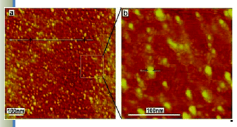


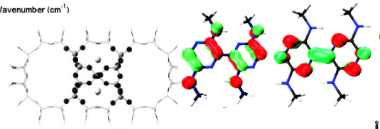
Disulfide clevage by P-



Zeolite Chemistry







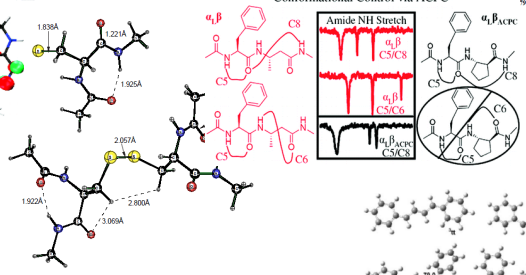
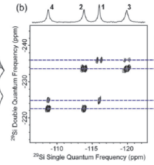
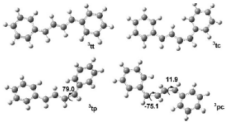
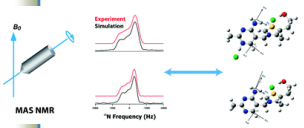


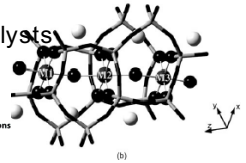
Figure 8. Model protein thiolate and disulfide optimized at the

Conformational Control via ACPC



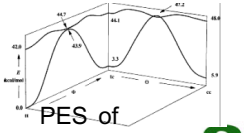




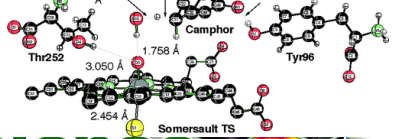


DFT calculations

Thiolate –SS interchange



PES of diphenylbutadienes



INDIANA UNIVERSITY

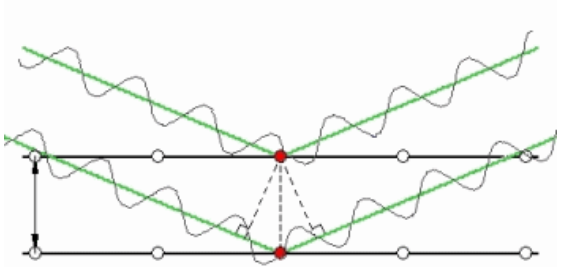
Outline

- Discovery Environments
- Some Molecular Science Gateways
- SEAGrid Science Gateway
- Diffraction Workflows in SEAGrid using optimized LAMMPS implementation – An XSEDE ECSS Project
- Parametrization workflows
- Data projects in SEAGrid
- Cyberinfrastructures for Medicine

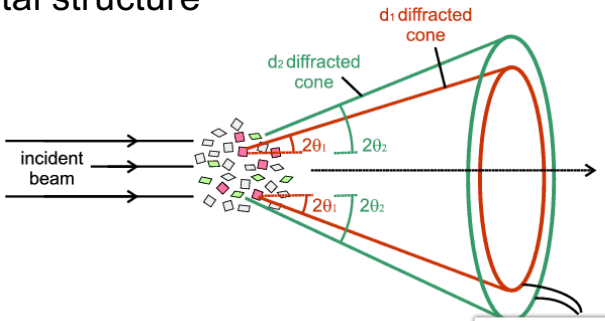
Diffraction Background

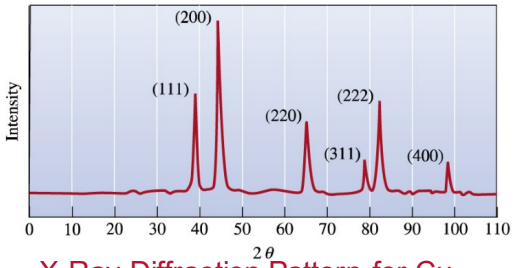
- X-Ray and Electron Diffraction are common experimental techniques to determine the structure of a material

Example: X-Ray Powder Diffraction



- Constructive diffraction will occur at specific orientations between the material lattice and the incoming x-rays
- In powder diffraction, all orientations of the lattice are represented and the resulting diffraction pattern identifies the crystal structure





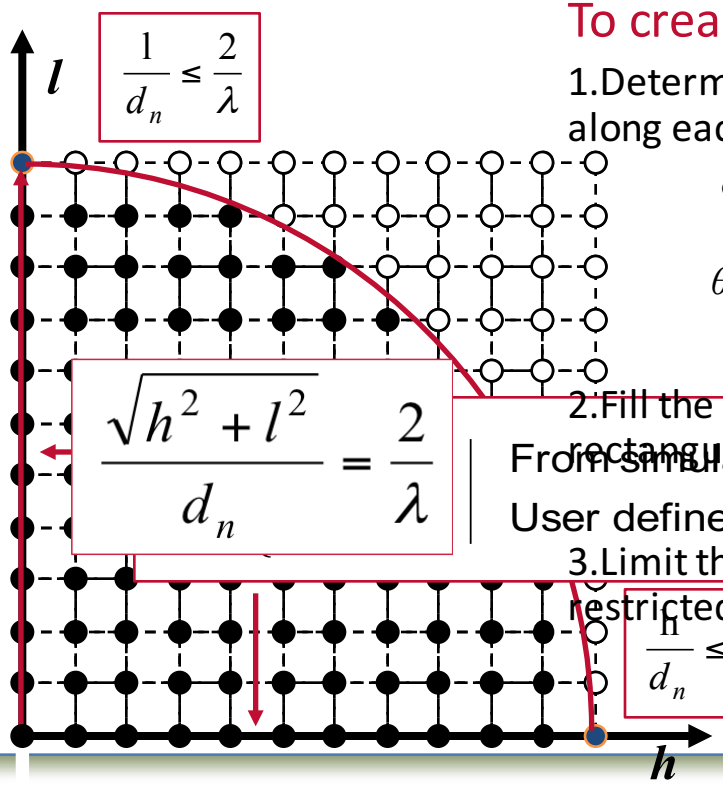
X-Ray Diffraction Pattern for Cu

INDIANA UNIVERSITY

Science Gateways

Step 1: Reciprocal Space Mesh

- The diffraction algorithm computes intensities using a high-resolution 3D mesh of points in reciprocal space



To create the reciprocal space mesh:

1. Determine node spacing, $1/d_n$, and location along each axes
   - Limits introduced by Bragg's Law

$$\theta = \sin^{-1}\left(\frac{\lambda}{2d}\right)$$

Limit: $0 \le \frac{1}{d_n} \le \frac{2}{\lambda}$

2. Fill the volume of reciprocal space with a rectangular mesh
   - From simulation boundaries
   - User defined value
3. Limit the computation within the domain restricted by Bragg's Law

INDIANA UNIVERSITY

Science Gateways

Step 2: Compute Diffraction Intensity

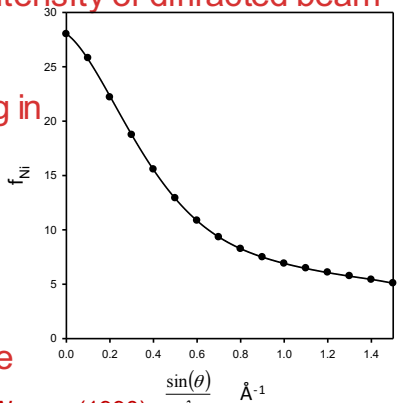
- At each *hkl* node, the diffraction intensity is calculated utilizing the structure factor,

$$I_{hkl} = \mathrm{Lp}_{\mathrm{hkl}} F_{hkl} F_{hkl}^{*}$$ Warren (1990)

$$F_{hkl} = \sum_{n=1}^{\#\,\mathrm{atoms}} f_n \mathrm{e}^{2\pi i\left(hu_n + kv_n + lw_n\right)}$$ = Structure Factor
~ role of crystal structure on intensity of diffracted beam

= Atomic Scattering Factor
~ efficiency of an atom in scattering in a given direction computed via a summation of Gaussian functions

$$\mathrm{Lp}_{\mathrm{hkl}} = \frac{1+\cos^2\left(2\theta_{hkl}\right)}{\cos\left(\theta_{hkl}\right)\sin^2\left(\theta_{hkl}\right)}$$ = Lorenz-Polarization Factor
~ accounts for the variation in the diffracted intensity with angle during x-ray diffraction



Warren (1990)

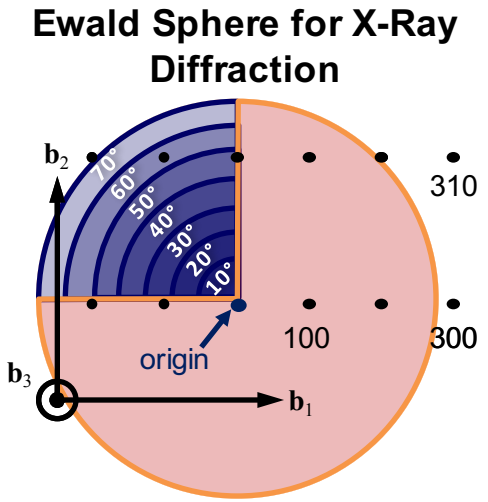
INDIANA UNIVERSITY

Science Gateways

Step 3: Compute Diffraction Patterns

- Diffraction conditions satisfied when a reciprocal lattice point is located on the surface of Ewald sphere
- In x-ray powder diffraction all orientations of the sample are equally probable
- 2θ XRD profiles are produced by binning data by *hkl* nodes with similar diffraction angles

$$\theta_{hkl} = \sin^{-1}\left(\frac{\lambda}{2d_{hkl}}\right)$$



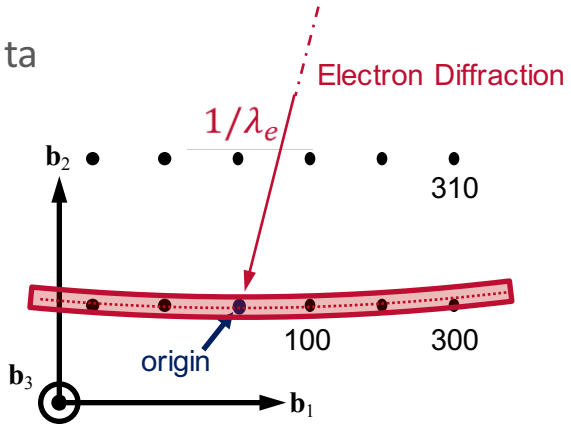
INDIANA UNIVERSITY

Science Gateways

Step 3: Compute Diffraction Patterns

- Diffraction conditions satisfied when a reciprocal lattice point is located on the surface of Ewald sphere
- In electron diffraction, the radius of the Ewald Sphere is much larger
- SAED patterns are produced by filtering the low-intensity data and taking an appropriate slice

Ewald Sphere for Electron Diffraction



MPI/OpenMP/MIC parallelization

LAMMPS: Domain decomposition of atoms parallelized via **MPI**

Each process must have a copy of the reciprocal space mesh

OpenMP parallelization of reciprocal space

- Typically 1 MPI process per node with many OpenMP threads
- MPI Parallelization of atoms across multiple CPU node

Further parallelization of reciprocal space on **MIC coprocessors**

- Implementation based on the OpenMP parallelization

INDIANA UNIVERSITY

Science Gateways

Scaling Study

- OpenMP parallelization with algorithm improvements
  - Machine: TACC Stampede (Dual 8 core Xeon E5-2680 + Xeon Phi SE10P)
  - 256,000 Ni atoms; 9,006,316 reciprocal space nodes; XRD: $10^o < 2\theta < 90^o$; 90% of reciprocal node threads are offloaded to MIC

Parallelization of recip. space

Nodes - MPI/OpenMP/MIC	Speed up	Efficiency (%)	CPU Memory (GB)
1 - 16/0/0	1.69	100%	8.1
2 - 32/0/0	3.21	95%	16.1
4 - 64/0/0	6.16	91%	32.2
8 - 128/0/0	11.71	86%	64.4
16 - 256/0/0	22.17	82%	128.9
1 - 1/16/0	1.78	100%	0.51
2 - 2/16/0	3.56	100%	1.01
4 - 4/16/0	7.12	100%	2.02
8 - 8/16/0	14.16	99%	4.05
16 - 16/16/0	26.94	95%	8.10

With Offloading to MICs

Nodes - MPI/OpenMP/MIC	Speedup	Efficiency (%)	CPU Memory (GB)
1 - 1/16/240	4.61	100%	0.51
2 - 2/16/240	9.19	100%	1.01
4 - 4/16/240	17.43	95%	2.02
8 - 8/16/240	34.65	94%	4.05
16 - 16/16/240	60.93	83%	8.10
1 - 16/0/240	4.37	100%	8.1
2 - 32/0/240	8.38	96%	16.1
4 - 64/0/240	15.94	91%	32.2
8 - 128/0/240	29.51	84%	64.4
16 - 256/0/240	49.32	70%	128.9

Speedup is relative to original code on 16 MPI processes

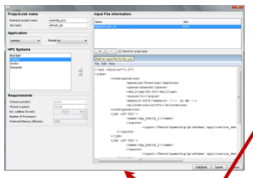
$$\varepsilon = \frac{t_1}{(N_{MPI} N_{Threads}) * t_N} * 100\%$$

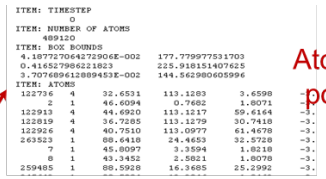
NCSA

Science Gateways

Workflow and Visualization

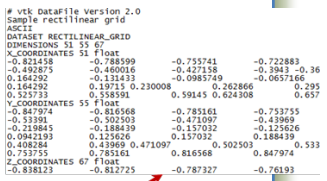
- For large memory calculations a workflow is required to use the appropriate XSEDE resource
  - TACC Stampede: Atomistic simulation of alumina
  - SDSC Gordon: Calculation of diffraction intensities + Visualization
- Workflow implemented through SEAGrid Science gateway
  - Supports a private "DS" LAMMPS build
  - Supports single jobID handle for multiresource job submission
  - Supports the development of a XML script for high throughput job submission
  - Compatible with parallel VisIt executions so that diffraction pattern generation is automated





Atom type, positions (xyz)

Diffraction intensity data (vtk)



SEAGrid Job Submission Lammps_DS MD on TACC Stampede

Stage Required Files to SDSC Gordon and all output to MSS Archive

Generate SDSC Gordon Lammps_DS and Vis Scripts and Launch

Compute Lammps_DS SAED and XRD on SDSC Gordon

Process output and vtk files and setup for VisIt on SDSC Gordon

Parallel VisIt run on SDSC Gordon for all VisIT session files provided

Stage all outputs to MSS Archive – Notify End of Workflow

SEAGrid Job Submission Lammps_DS MD on **TACC Stampede**

Compute Lammps_DS SAED and XRD on **TACC Stampede**

Stage Required Files to **TACC Vis Nodes** all output to MSS Archive

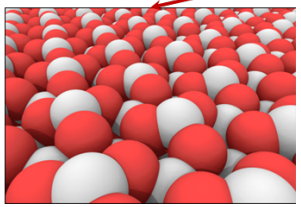
Generate **TACC Vis** Scripts and Launch Dependant Job

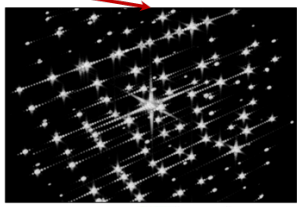
Process output and vtk files and setup for VisIt on **TACC** Vis cluster

Parallel VisIt run on **TACC VIS cluster** for all VisIT session files provided

Stage all outputs to MSS Archive – Notify End of Workflow

Visualization of atoms





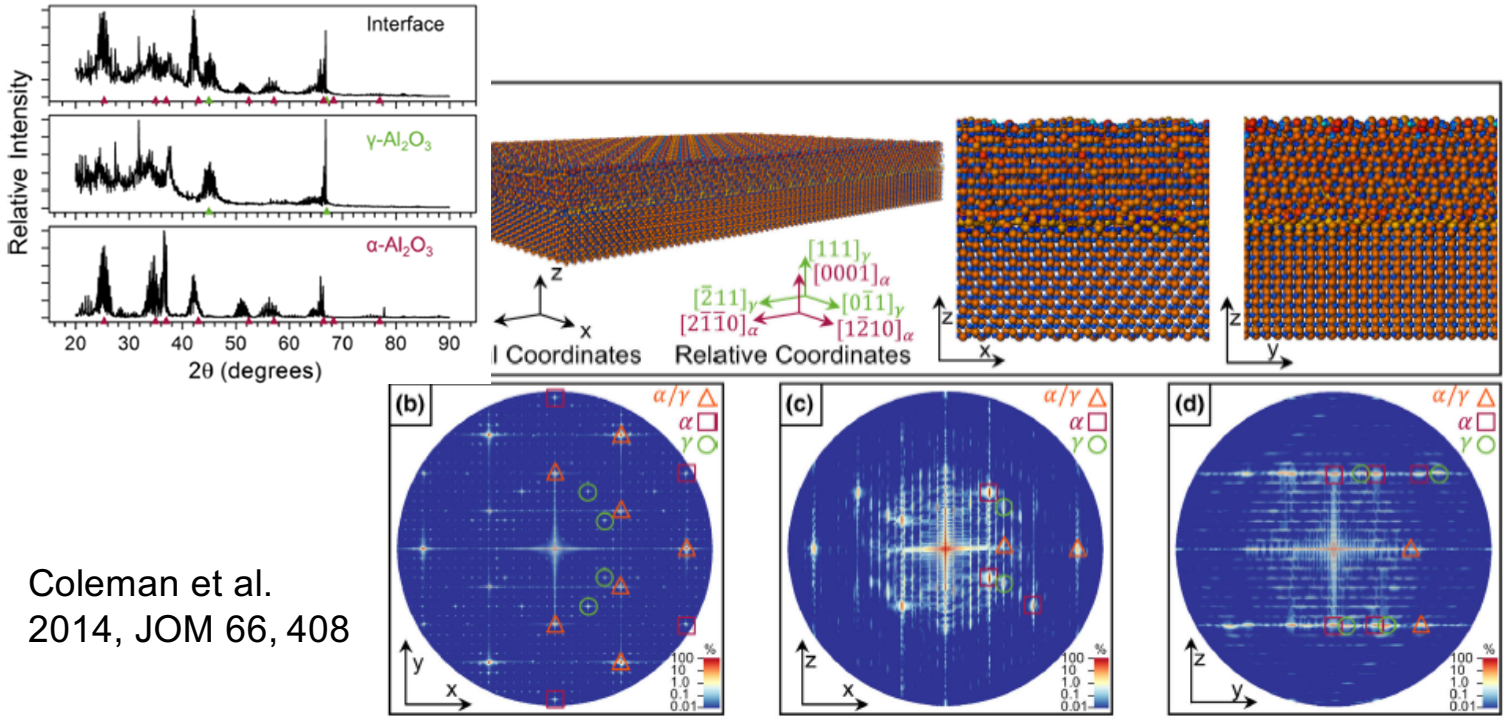
Reciprocal space map

Workflow to run on TACC resources only

NCSA

Science Gateways

Line profiles, Structures and SAED Patterns for heterogonous α-Al2O3 (0001)/γ-Al2O3 (111) interface



Coleman et al.
2014, JOM 66, 408

The colored triangles are positioned at experimentally determined peak locations for bulk α-Al2O3 (red) and γ-Al2O3 (green)33.

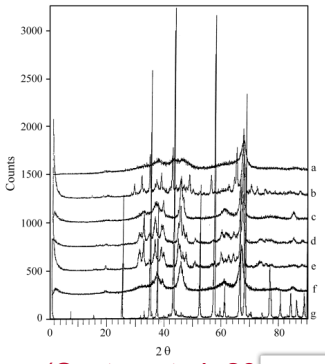
NCSA

Science Gateways

X-ray Diffraction of Bulk Alumina

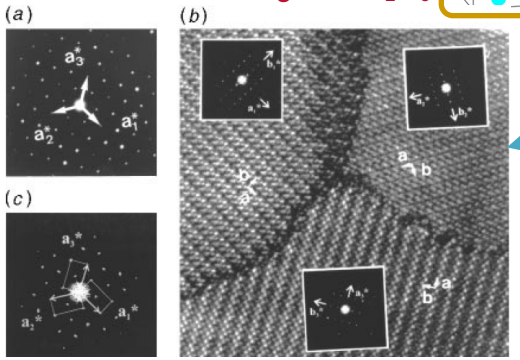
- Characterization methods for non-cubic crystals in atomistic simulations are limited
  - Centrosymmetry analysis shows no clear differentiation between phases
  - Radial distribution function cannot differentiate between transition phases

XRD of Alumina Phases



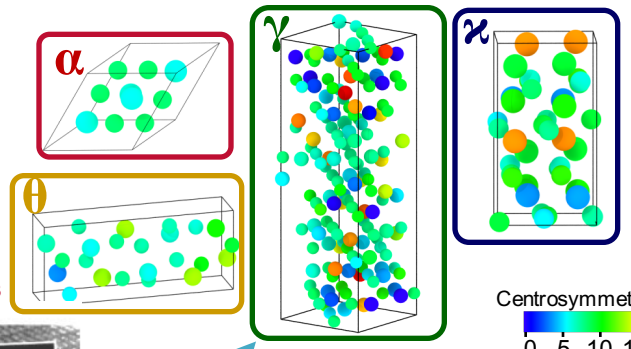
(Santos et al. 20

SAED and HRTEM of 120° twining in κ-Al_2O_3



(Ollivier et al. 1997)

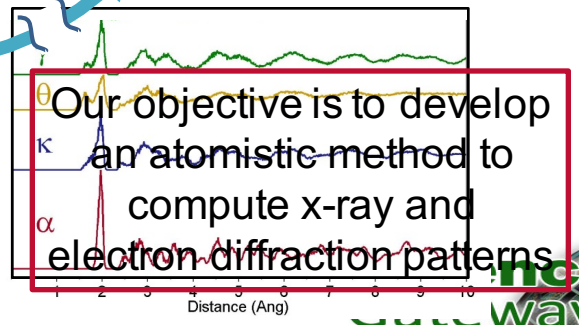
Alumina Phases



Centrosymmetry(Å²)

0 5 10 15 20

Radial Distribution Functions



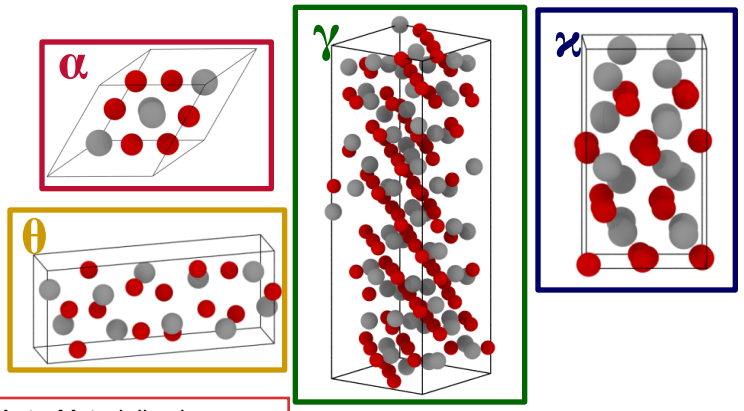
Our objective is to develop an atomistic method to compute x-ray and electron diffraction patterns

NCSA

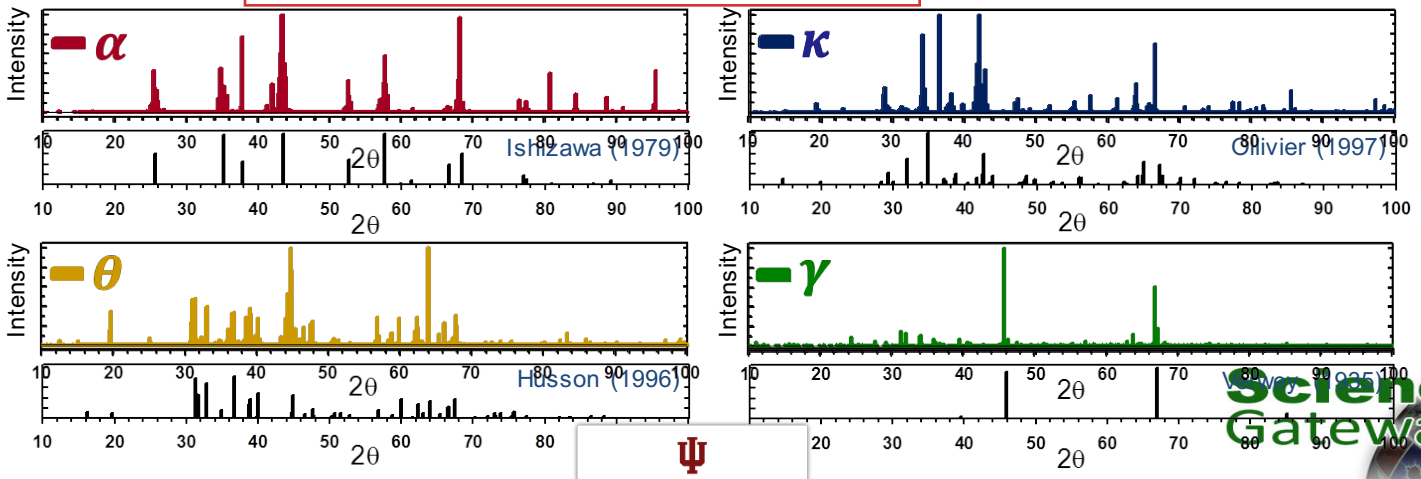
INDIANA UNIVERSITY

Application: Phase Identification in Alumina

- Virtual diffraction can distinguish between alumina phases
  - Al_2O_3 phases modeled with the ReaxFF potential (Sen et al. 2013)



Coleman and Spearot (2014) Acta Materialia, in press.



NCSA

INDIANA UNIVERSITY

Science Gateways

Outline

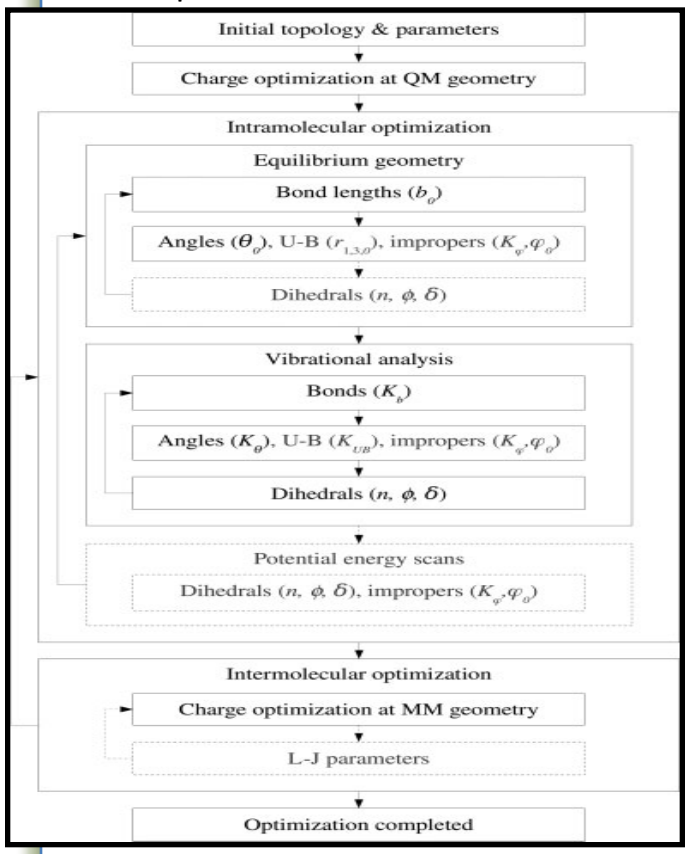
- Discovery Environments
- Some Molecular Science Gateways
- SEAGrid Science Gateway
- Diffraction Workflows using optimized LAMMPS implementation
- Force Field Parametrization workflows
- Data projects in SEAGrid
- Cyberinfrastructures for Medicine

Cyberenvironments for Molecular Force Fields

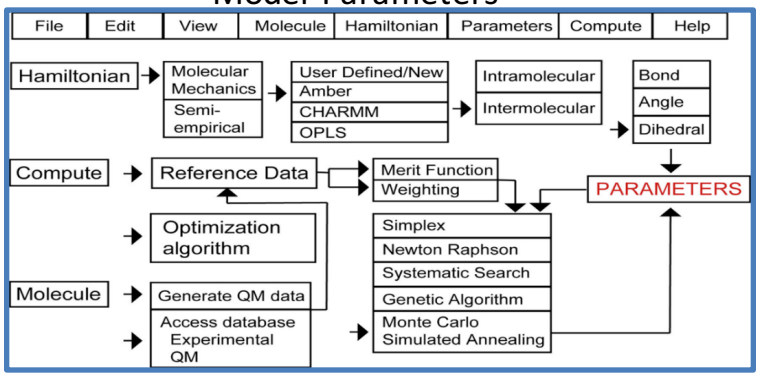
- Molecular ForceFields consist of atomic, bond, angle, dihedral and non bonded interaction parameters to evaluate energetics of model systems
- Extension of currently available models, with the resulting parameters sets to be made available publicly
- Databases of experimental and quantum mechanical reference data to be used in the parameterization process
- Integration of computational resources for data acquisition, automation of QM reference data generation
- Automation Extensible infrastructure for parameterization management for rapid and systematic parameterization of novel Hamiltonians (empirical and semi-empirical)
- Systematic improvement of parameter optimization processes

Parametrization Methodology and Schemes

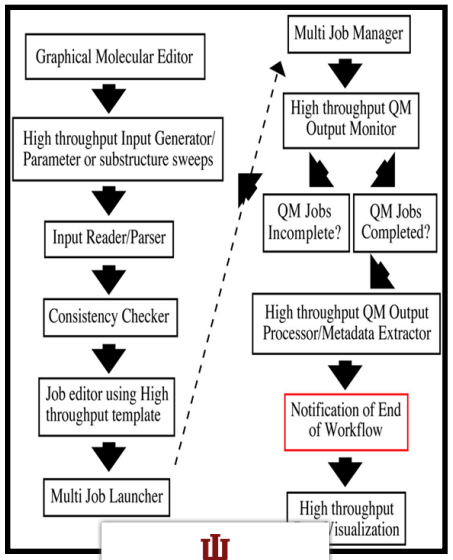
Optimization Scheme



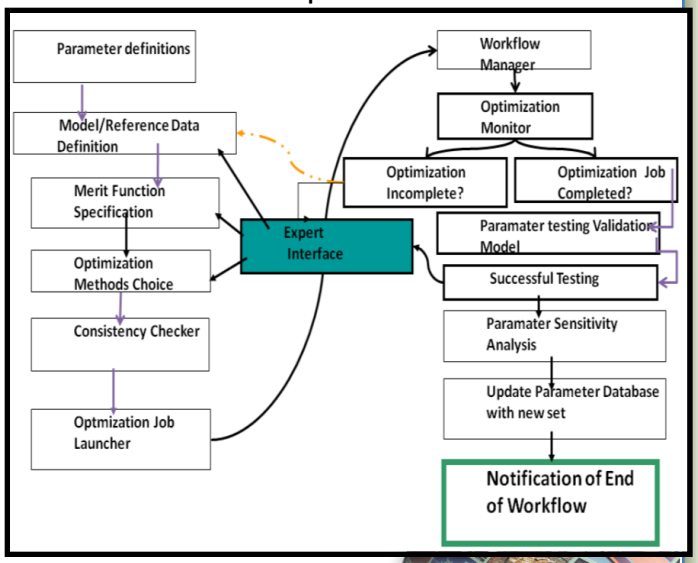
Model Parameters



Reference Data



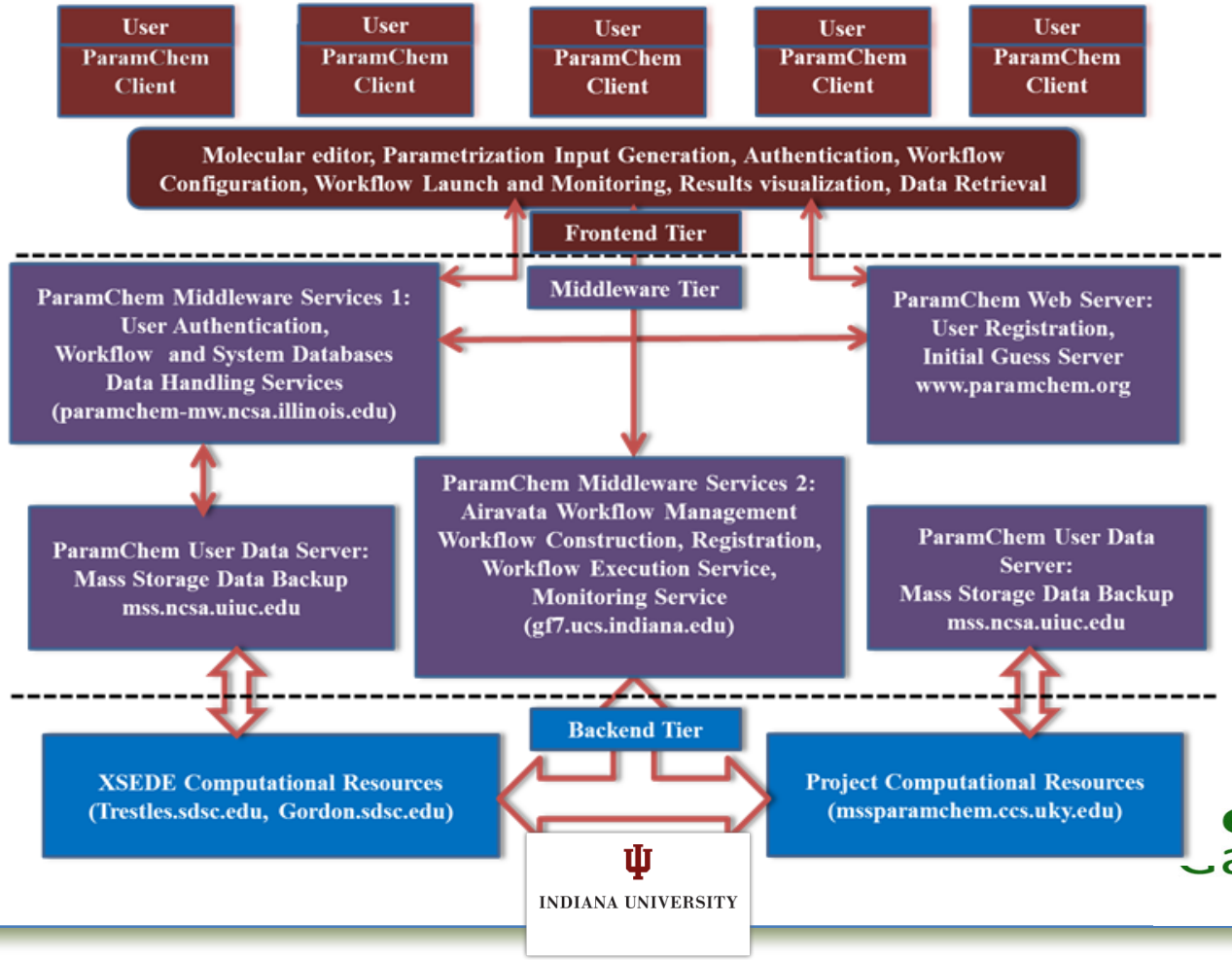
Parameter Optimization



NCSA

INDIANA UNIVERSITY

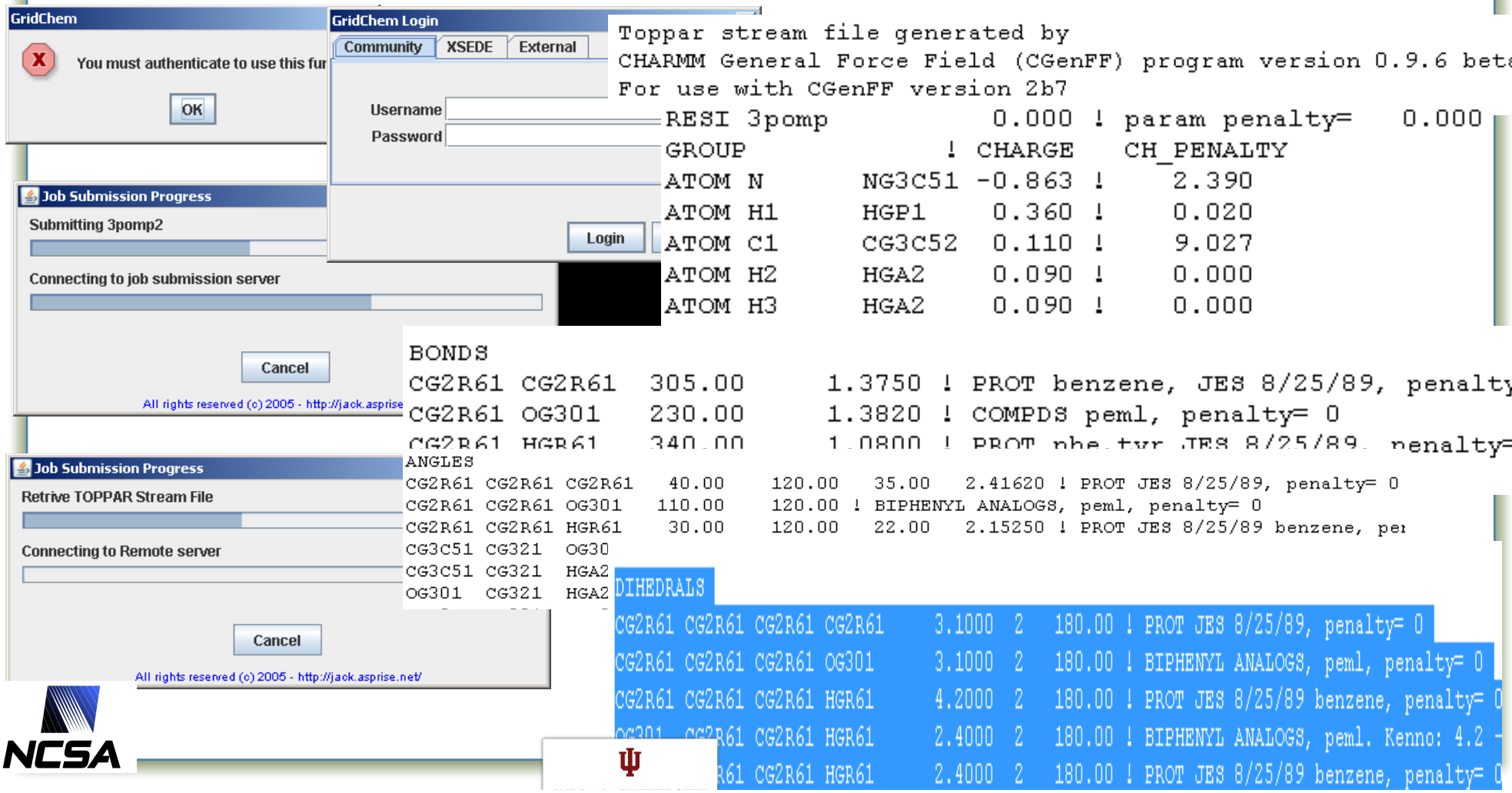
ParamChem architecture



NCSA

cience
ateways

Authentication and Initial Guess generation



NCSA

Ψ

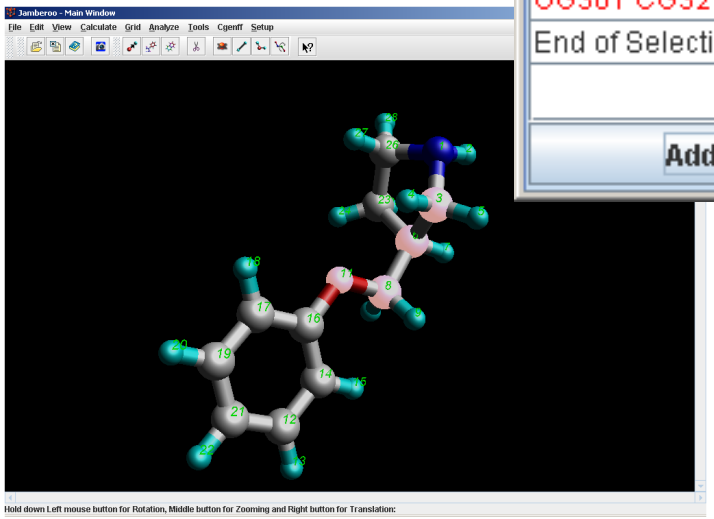
Dihedral selection graphical and data interfaces

Parameters

CHARMM dihedral energy term

$V(\phi) = K\phi\ (\ 1 + (\cos(n(\phi)) + \delta\ )$

- Set $K\phi$ value: [ ]
- Compute initial guess for $K\phi$
- Set n value: [ ]
- Compute initial guess for n
- Set δ value: [ ]
- Compute initial guess for δ



Dihedral Parameters

Parameters around selected bond

All parameters

Dihedral Parameter	Multiplicity
HGA2 CG321 CG3C51 CG3C52	3
OG301 CG321 CG3C51 CG3C52	3
HGA2 CG321 CG3C51 HGA1	3
OG301 CG321 CG3C51 HGA1	3
End of Selection	End of ...

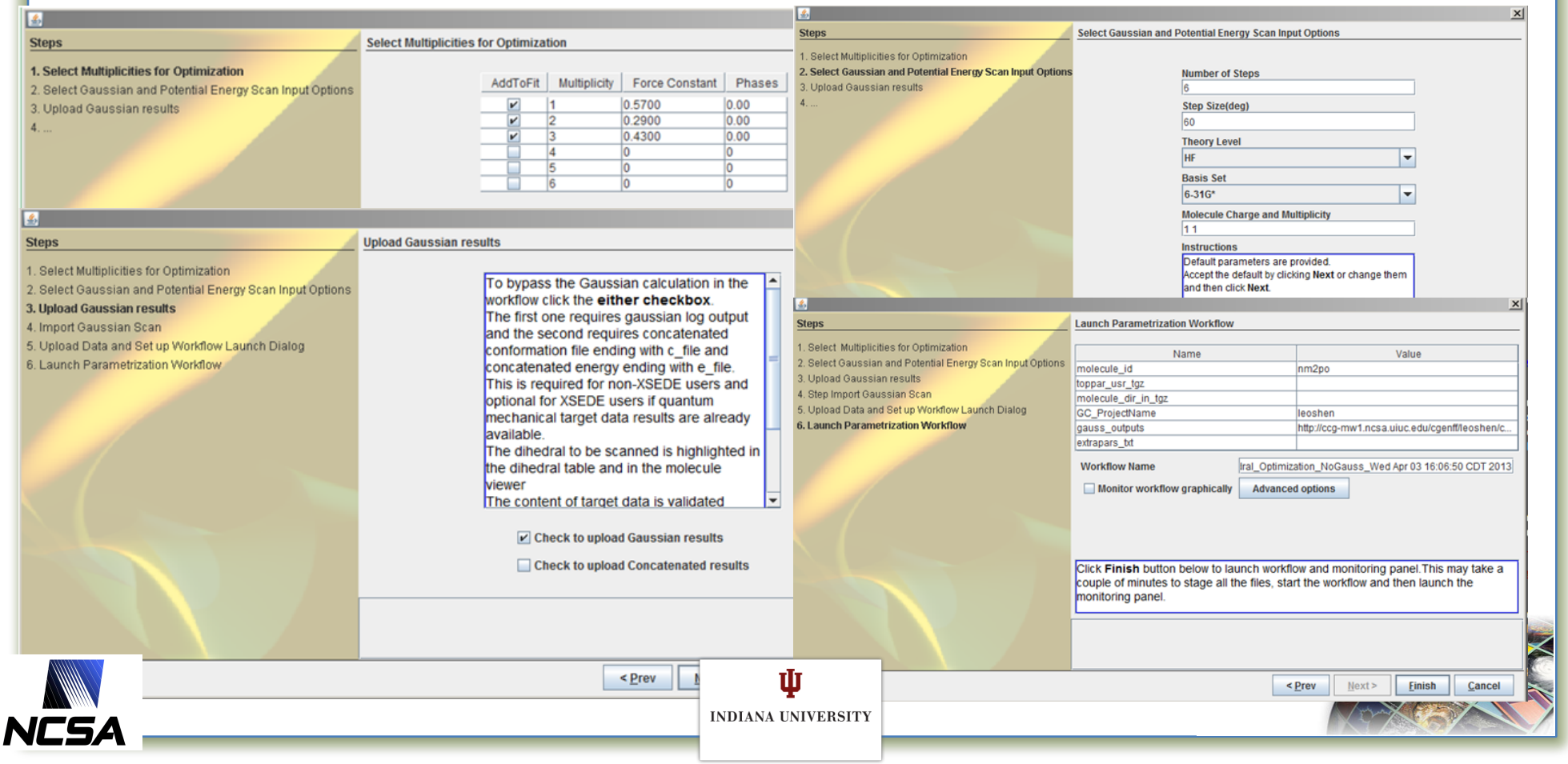
Add Selected to dihedral Table

NCSA

INDIANA UNIVERSITY

Science Gateways

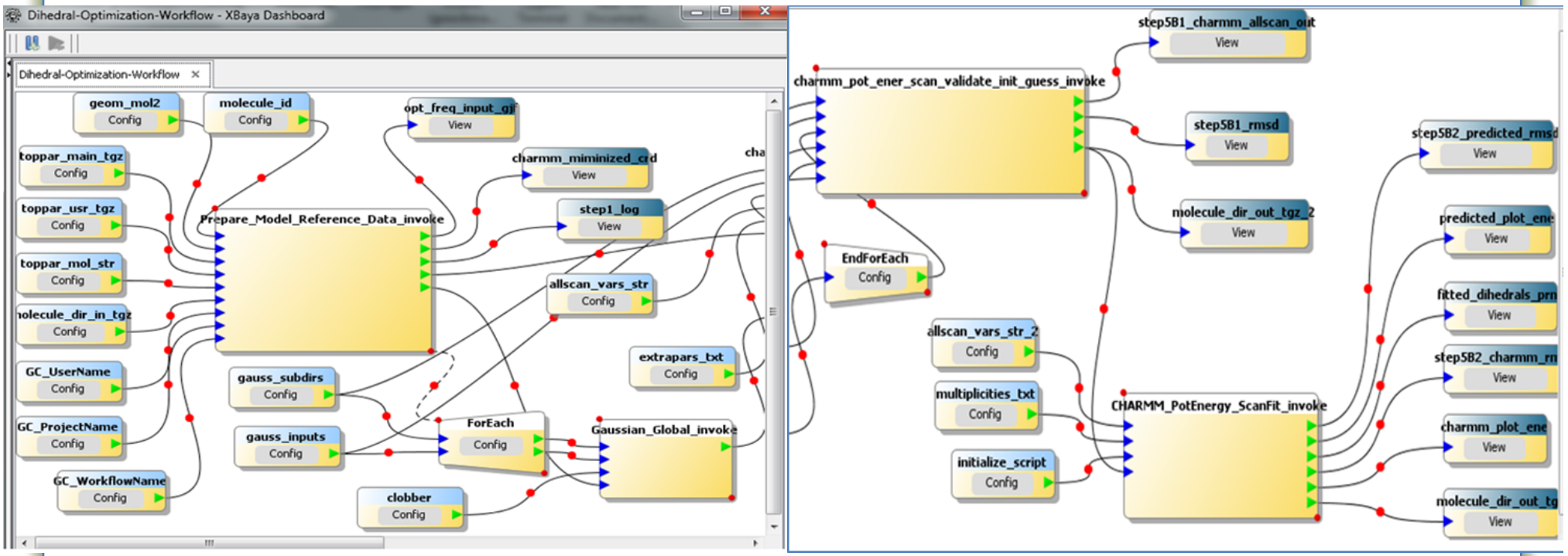
Wizard Guided Dihedral Parametrization Setup



NCSA

INDIANA UNIVERSITY

Apache Airavata Workflow Dihedral Parametrization

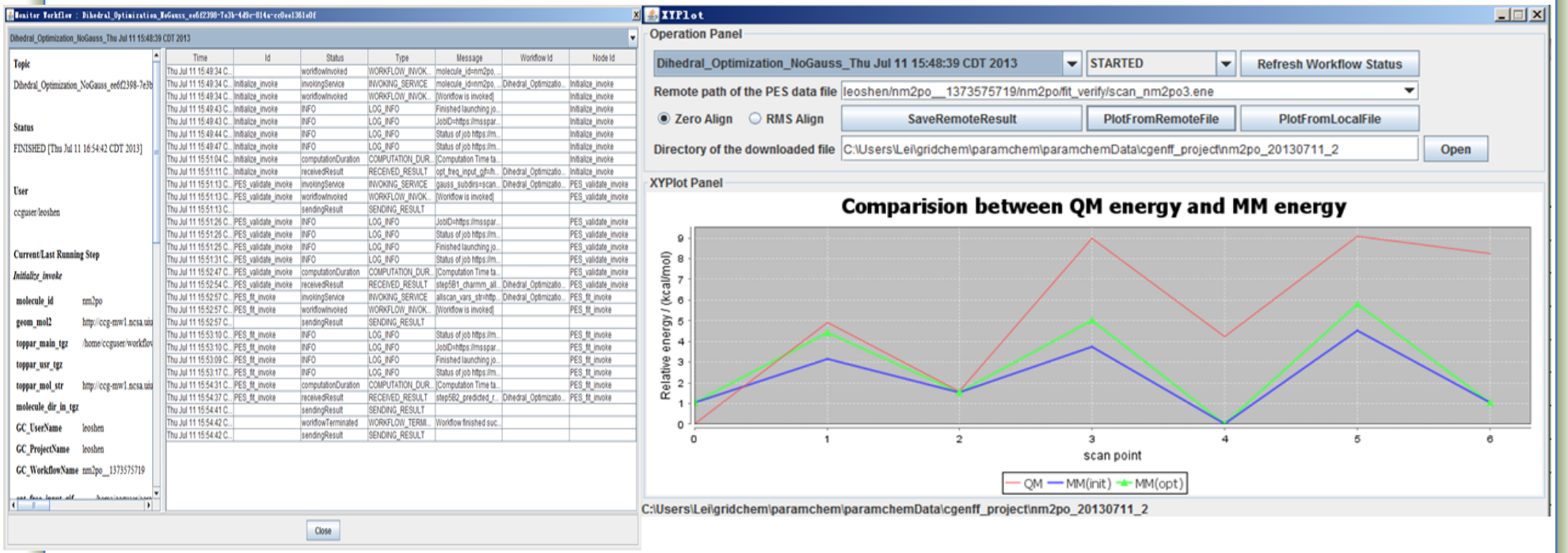


NCSA

INDIANA UNIVERSITY

Science Gateways

Workflow Monitor and Validation of the dihedral parameters



"lsfitpar" program for robust fitting of bonded parameters:
http://doi.org/10.1002/jcc.23897

NCSA

INDIANA UNIVERSITY

Science Gateways

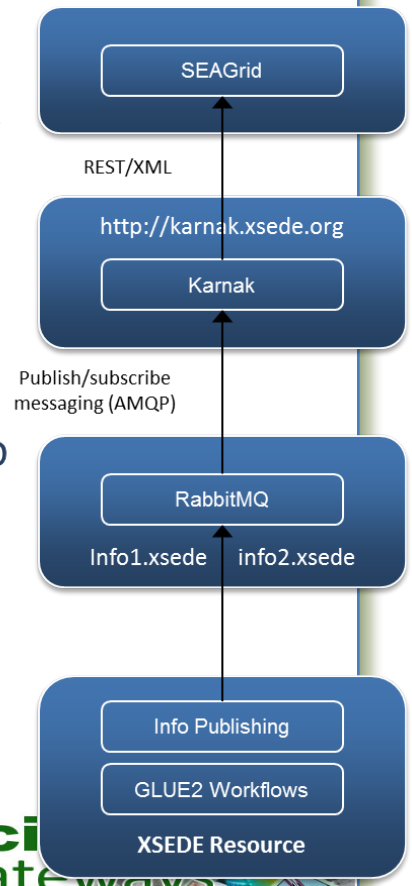
Outline

- **Discovery Environments**
- **Some Molecular Science Gateways**
- **SEAGrid Science Gateway**
- **Diffraction Workflows using optimized LAMMPS implementation**
- **Parametrization workflows**
- **Data analytics projects in SEAGrid**
- **Cyberinfrastructures for Medicine**

Data derived queue predictions

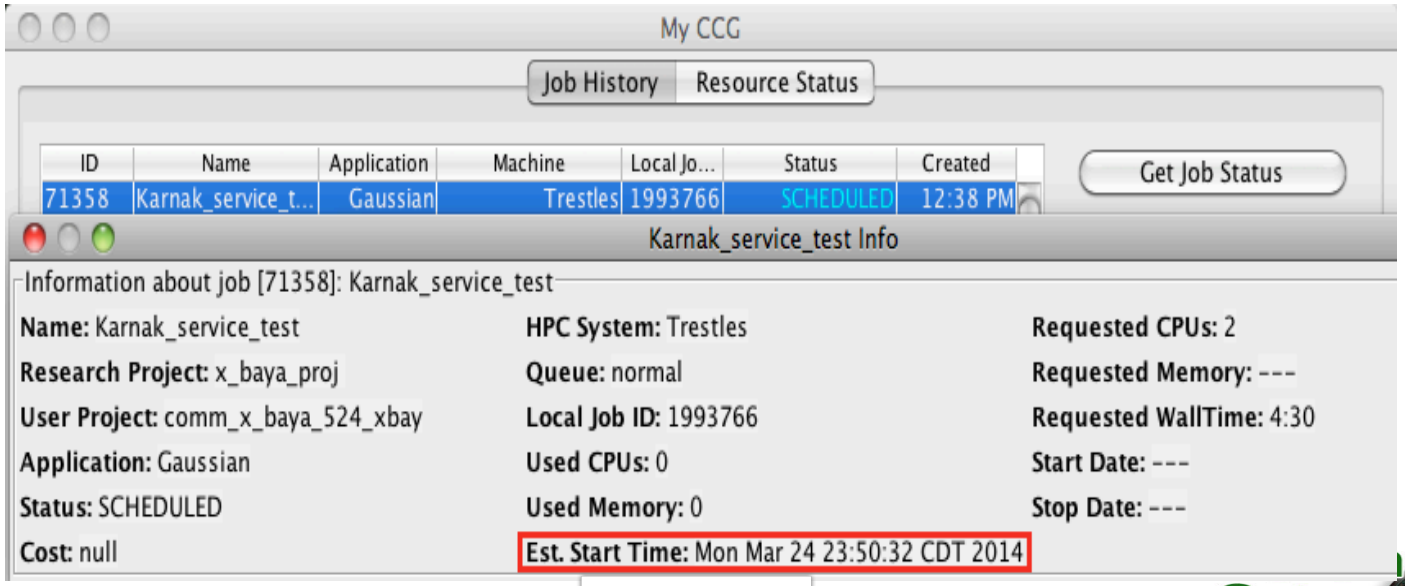
- Help SEAGrid users run jobs more efficiently on XSEDE
  - Provide predictions of when a job would start/complete if submitted
    - Optionally Select system that will complete quickest
- Provide more information to SEAGrid users to improve their experience
  - Provide predictions of when a submitted job will start/complete
    - Information while waiting
- Provide automated estimates for components in a workflow and an aggregated one for the whole workflow

- Modify SEAGrid interface
  - Display predictions in appropriate places
- Modify SEAGrid service
  - Provide predictions to client
- Use Karnak service for predictions
  - Enhanced as needed
- Provide information about systems to Karnak
  - For the specific systems used by SEAGrid
  - Using the new publish/subscribe system XSEDE is deploying



SEAGrid Predictions II

- Enhance job history
- Show estimated start time for waiting jobs
  - Also show estimated completion times



Karnak Instance-Based Learning

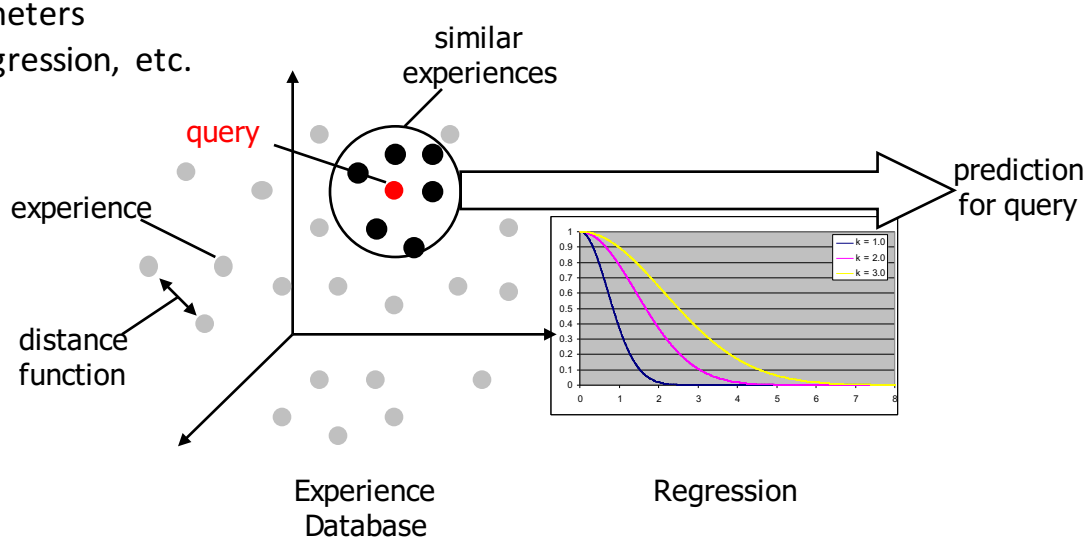
Instance-based learning

Doesn't train a model

Do need to optimize a number of parameters

K-nearest neighbor, locally weighted regression, etc.

Evaluating alternative techniques



- Experience describes an observation
  - Input features describe conditions (e.g. the number of queued jobs)
  - Output feature describes result (e.g. the wait time)
- Store experiences in a database
- Query for similar experiences when making a prediction
  - Defined by a distance function
- Form prediction using output features of similar experiences

Alternative Prediction Techniques

- Instance-based learning has some disadvantages
  - Slow, particularly for large numbers of experiences
    - Slow to optimize (e.g. feature weights, # neighbors, kernel parameters)
    - Relatively slow to make predictions
- Current alternatives
  - Regression tree
    - Construct using training data
    - Traverse the tree using a query, leaf node contains the prediction
  - Boosting
    - Multiple small trees, prediction is the sum (additive approach)
  - Bagging
    - Multiple trees, prediction is the average
- Future alternatives
  - Time series
  - Support vector regression, neural networks

Prediction Performance: Queue Wait Time (Hypothetical Jobs)

System	Mean Wait Time (minutes)	Error (minutes)					
		IBL	Mean	Last from Queue	Tree	Boosting	Bagging
Blacklight	171	**153**	463	209	176	398	160
Gordon	469	**887**	981	474	703	805	667
Trestles	430	**454**	405	426	459	459	433

- How to improve accuracy?
  - Different techniques?
  - Use multiple techniques at once
  - Different representation

NCSA

INDIANA UNIVERSITY

Science Gateways

Gaussian Run Time Predictions

- Most used SEAGrid application
- Extend Karnak to predict
  - Define Gaussian experiences
  - Configure a predictor
  - Create a customized REST interface
- Creating experiences for Gaussian is hard!
  - Extract data from the SEAGrid database & input files
    - Input files are complex
  - Translate that data into Karnak experiences
  - Must identify features that impact run time
  - 11,895 Experiences
  - Currently 42 features
- So far, mean error is 56% of mean run time
- Only use jobs that complete successfully?
- Can we predict certain kinds of runs more reliably than ot
- Treat different sets of features with different techniques

Feature	Type	Notes
System	Category	3 different categories
Processes	Number	
Basis set	Category	~35 different categories
Method	Category	~50 different categories
Time dependent	Boolean	
TD options	Various	4 different features
Intrinsic reaction coordinate	Boolean	
IRC options	Various	9 different features
Optimize	Boolean	
Optimize options	Various	17 different features
Charge	Number	
Multiplicity	Number	
Atoms	Number	
Electrons	Number	
Run time	Number	

Mean Run Time (minutes)	Error (minutes)				
	Mean	Tree	Boosting	Bagging	IBL
350	419	224	212	197	244

NCSA

INDIANA UNIVERSITY

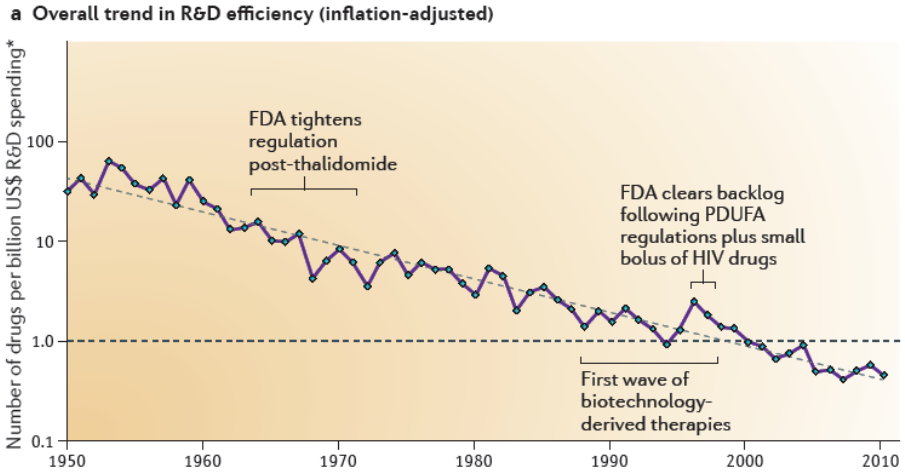
Science Gateways

Outline

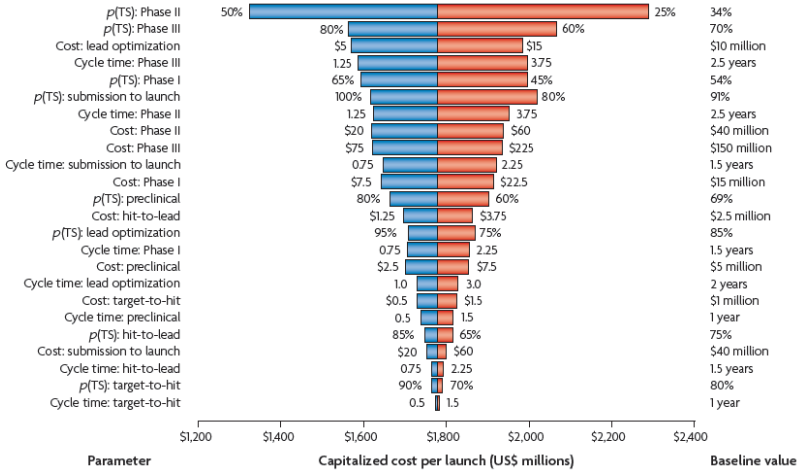
- Discovery Environments
- Some Molecular Science Gateways
- SEAGrid Science Gateway
- Diffraction Workflows using optimized LAMMPS implementation
- Parametrization workflows
- Data projects in SEAGrid
- Cyberinfrastructures for Medicine

  Drug Development and delivery

Unsustainable drug development paradigm



Scanell et al., Nature 191, 2012



Paul et al, Nature 203, 2010

10K molecules – 250 at preclinical – 1 Final New medicine approved
50K molecules in pipeline – Expectation is 10 or so approved medicines?!

NCSA

INDIANA UNIVERSITY

Science Gateways

How do we sustain the well being of world

Education of individuals and populations is going to be the key

Behavioral, preventional and interventional approaches

Ethical/Practical issues in drug development/delivery --- Is it ethical to profit from somebody's suffering/Who pays for the cost

Reduce, Reuse (repurpose), Recycle (if something is still left – do not polute)

Reduce the risk of failure; Reduce the cost of development; Reduce the time to discover and deliver drugs

Reduce unnecessary diagnostics, drug prescriptions, unnecessary dosage and toxicity

Information (data), intelligent analysis

Model every process *in-silico* before taking the process to lab/trial/patient

Commercial and Non-Profit Models

Partnerships with non-profit groups for small populations, unique conditions, experimental, commercially non-viable development – The data obtained from these partnerships should be public to keep the cost low!

Leave drug manufacturing and delivery to for-profit sector for large population where economies of scale can provide viability

Let non-profit organizations to focus on niche medicines

Computer Aided Drug Development and Delivery

Target Selection and Validation –Target Structure

Druggable Genome, Disease Gene/Gene Product mapping (GPCRs, kinases, proteases and peptidases)

PsiPred, Modeler for Target Structure

Binding Site detection and Characterization

CATS ,MOE, Catalyst-HipHop and Hypogen, Phase, LigandScout, DISCO, GASP and GRID for Pharmacophore modeling

POCKET, SURFNET,LIGSITE Q-SITEFINDER,GRID, PocketFinder, PocketPicker, Surface, PockerSurfer, PatchSurfer

Targeted-MD, SWARM-MD, TA-MD, RE-MD.

Docking and Scoring ligands with target

Target Based

MOE-DOCK, GLIDE, and GOLD, DOCK and AutoDock, MOE-Score, GlideScore, GoldScore, ChemScore, and Xscore

LUDI, FLEXX, SEED, SURFLEX, MCDOC, ICM, ROSETTALIGAND

DrugScore,SMOG, BLEEP

Ligand Based methods 3/4/5D QSAR Models

ADRIANA, LigandBuilder, SYNOPSIS, LigandScout and Catalyst

ADME properties and drug optimization

META –Mammalian Xenobiotic Metabolism, GOLD, FlexX, DOCK, AutoDock, and the scoring function C-Score for P450 interactions Catalyst-Hypogen, StarDrop' s Auto-Modeler –HERG blocking/non-Blocking, glucuronidation, sulfation, acetylation, methylation, and glutathione conjugation modeling

QikProp, OSIRIS Property Explorer, MetaSite (Molecular Discovery Ltd, Middlesex UK)

SMARTCyp, SimCYP (Simcyp Ltd, Sheffield UK), NONMEM software package (ICON plc, Dublin, Ireland)

PK/PD modeling

Open Big Data For Drug Discovery

Disease Datasets diseases genotype/phenotype/omim 22,600 entries http://omim.org/statistics/entry
10579 (https://www.ebi.ac.uk/chembl/); http://diseasome.eu/map.html ; http://www.lincsproject.org/

Chemogenomic Datasets http://fitdb.stanford.edu/ ; ChemBL; http://stitch.embl.de/http://tcm.lifescience.ntu.edu.tw/
http://www.dharaonline.org/Forms/Home.aspx

Drug Data 11000 launched ChemBL

Drug Use Data 1000s of Drugs; DailyMeds; http://sideeffects.embl.de/

Biologically Active systems 3.1 Mil Compounds

Biological Effects 1Mil Pharmocological .5Mil PK/PD and .5Mil drug-protein interaction DrugBank.org

Biological Pathways and Networks and Interactions http://www.genome.jp/kegg/pathway.html;
1MIl interactions in 1200 pathways on 572 diseases http://www.enzyme-database.org/stats.php

Publications 17000 journals, 28000 Conferences 200K Press releases

Genetic and MicroArray Data Gene Expression Omnibus ArrayExpress http://www.ebi.ac.uk/arrayexpress/ 25 TB

Biomarkers 10209 and 356K uses EDRN Biomarker Database https://edrn.nci.nih.gov/biomarkers#b_start=0;
https://gobiomdb.com/gobiom/

Clinical Trial Data Clinicaltrials.gov 176,521 studies with locations in all 50 states and in 187 countries

GeneEditing http://www.broadinstitute.org/rnai/public/ (CRISPR/CAS9)

Biomedical Imaging Data BMI Archive https://github.com/NCIP/national-biomedical-image-archive

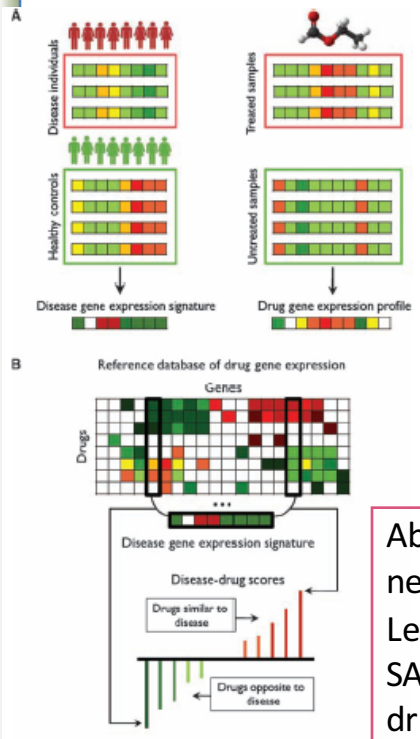
Integrate right data at the right part of the pipeline/workflow

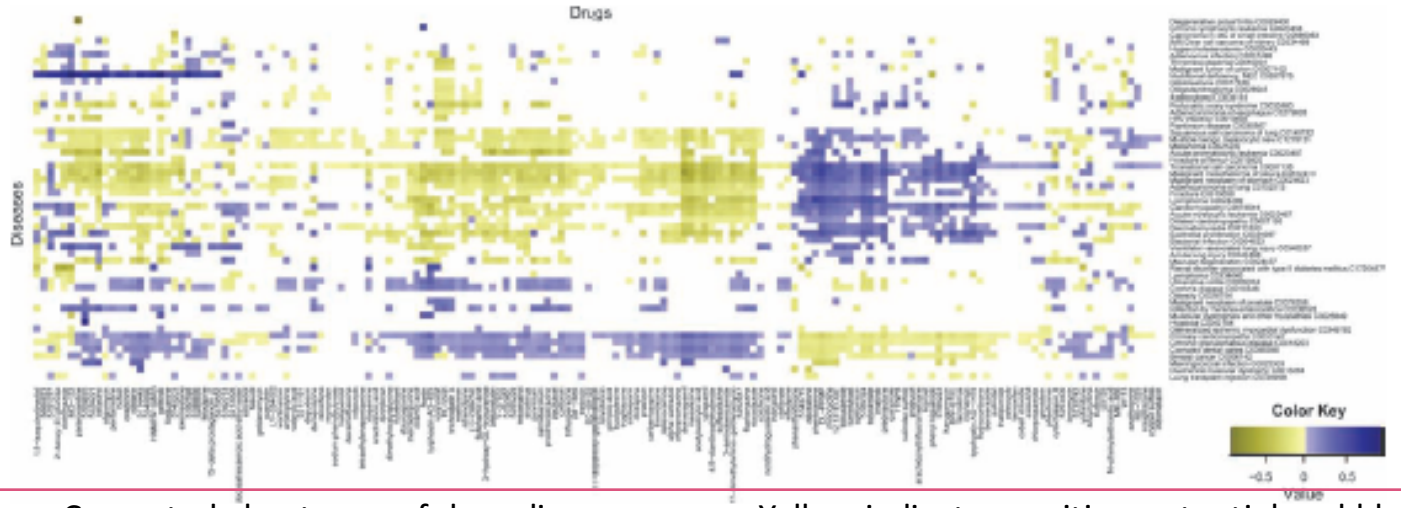
NCSA

INDIANA UNIVERSITY

Science Gateways

Open microarray data and hierarchical clustering based drug repurposing





Above. Computed heat map of drug disease scores. Yellow indicates positive potential and blue negative potential to treat the disease.
Left. A Gene Expresson with and without drug treatment for Disease and Control populations. SAM analysis provides up and down regulations. B. Modified Connectivity Map Method to get a drug-disease score.

Sirota et al. Sci. Transl. Med. 2011, 3, 96ra77. See also Dudley et al 2011, **3** *Sci Transl Med 96ra76*

NCSA

INDIANA UNIVERSITY

Science Gateways

Integrated data and analysis infrastructure for intelligent choices

Understanding the aliments - Psychosomatic origins

Psychological Conditions - CNS Diseases – Physical – Genetic and Non Genetic Origin

•Empirical nature of current treatment (only final outcomes are indication of success) – Vs Mechanism of Action based interventions

•Using (genetically or otherwise) modified bacteria or viruses for treatment

Immunology, Vaccines, Modified cells, Repopulation/rebalancing biome where appropriate

Individual (is in fact a biome) (meta)genomic information analysis may provide ways to reduce the risk of adverse drug interactions or even provide the right drug for right patient.

Other experimental data that could be annotated archived and used in computational drug development

•Time dependent mass spectroscopy (to delineate genetic vs extra genetic chemical signatures)

•Imaging organs-tissues-sections, Hyperspectral IR/Raman, electron cryo-microscopy to atomic structure (XRD,SAED,NS) – Image Analysis and Chemometrics

•structural biology – simulation , Fragment based drug discovery methods

The data integrated computational methods can help improve every aspect of the process

•Diagnostics, Biome management, optimize right clinical trials for reduce failures , provide more successful treatment with reduced toxicity

NCSA

INDIANA UNIVERSITY

Science Gateways

Summary and Outlook

Summary

- Science gateways are productive and widely used showing their importance for community computing
- Scientific workflows can be implemented in such gateways to drive multiple tasks together to provide enhanced coupled computations
- Data collected in Science Gateways can be used to provide improved user experience and potentially automate execution of jobs and workflows

Outlook

- Data mining an management and sharing enhancements in the Science Gateway
- Provide access to external or federated data collections and integrate data into simulations
- Interfacing with SciGaP technologies for workflow orchestration and management
- Explore New opportunities of service in Science, Engineering and Medicine

NCSA

INDIANA UNIVERSITY

Acknowledgements

National Science Foundation
Grants Sudhakar Pamidighantam (SDCI-NMI 1032742 , CRIF: CRF 0823041, SCI-0438312)
#0954505 (Spearot - CAREER), #0963249, #0959124, #0918970 (Computational Resources); TeraGrid/XSEDE

GridChem

Kent Milfeld, Chona Guiang, Rion Dooley, TACC , Michael Sheetz, Vikram Gazula Uky, Suresh Marru, IU, Dodi Heryadi, Joohyun Kim, Yang Liu, Thomas Roney, Ye Fan, NCSA, Stelios Kyriakou, Scott Brozell, Jim Giuliani, OSC.

Difrraction Workflows

Shawn P. Coleman, Douglas E. Spearot, Department of Mechanical Engineering, University of Arkansas; Mark Van Moer, National Center for Supercomputing Applications (NCSA) Yang Wang, Pittsburg Supercomputing Center (PSC); Lars Koesterke Texas Advanced Computing Center (TACC) *Luis Cueva-Parra, Auburn University Montgomery; Paula Romero Bermudez, University of Indianapolis, XSEDE ECSS.*

ParamChem

Alex MacKerell, Kenno Vonnameslaeghe, Univeristy of Maryland, Baltimore, Adrian Roitberg, U. Florida; Ning Shen, UIUC, Remya Puthantodiyil, Narendra Polani, Michael Sheetz, Vikram Gazula U. Kentucky and Suresh Marru, Chaturi Wimalasena, Saminda Wijeratne, Lahiru Gunathilake, Marlon Pierce, Indiana University.

Mark Miller, CIPRES
Boris Demeler, UltraScan

Warren Smith, TACC. XSEDE ECSS.

Thank You!

NCSA

INDIANA UNIVERSITY

Gateways

ॐ सर्वे भवन्तु सुखिनः

Om sarvE bhavantu sukhinah

May every person be happy

Questions?

spamidig@illinois.edu

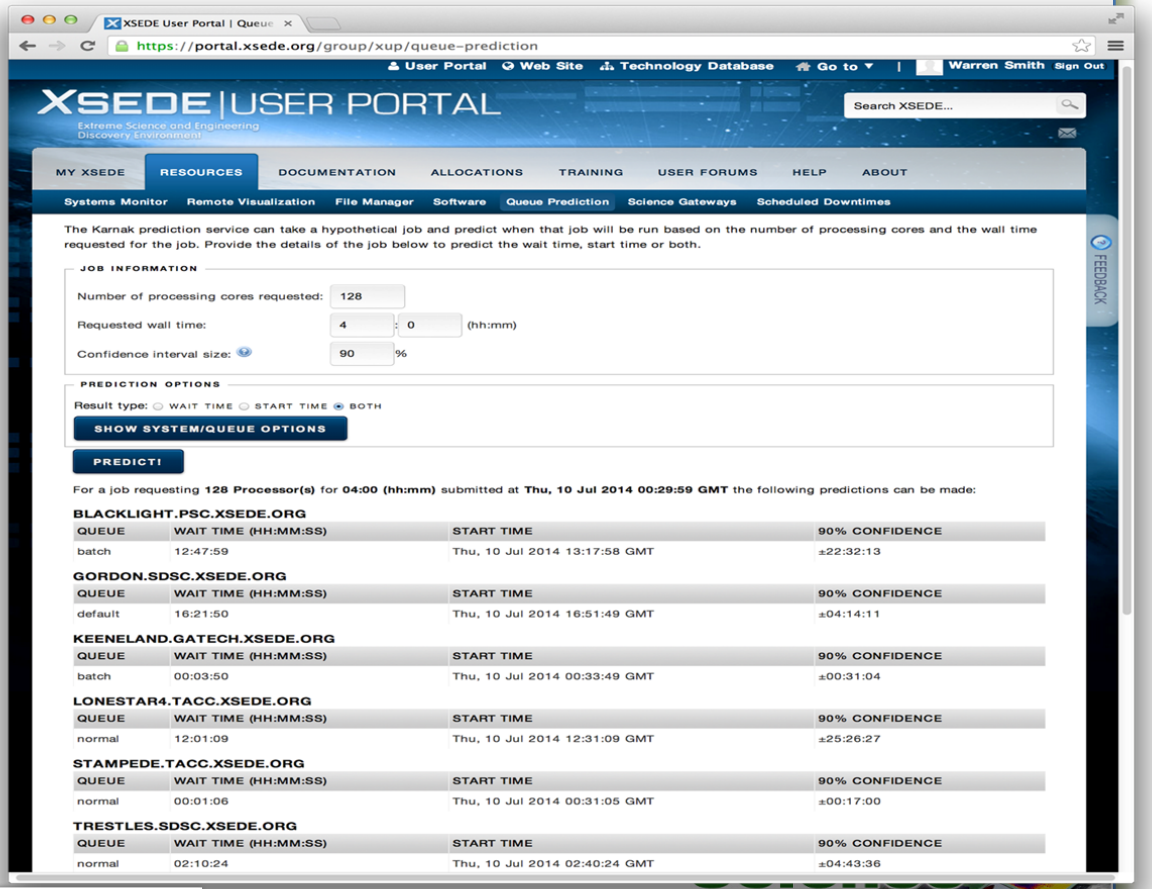
NCSA

INDIANA UNIVERSITY

Science Gateways

Karnak Prediction Service

- Functionality
  - Predict start time of hypothetical jobs
  - Predict start time of queued jobs
  - Provide information about current and recent jobs
- Interfaces
  - REST (XML and plain text), HTML
    - http://karnak.xsede.org
  - Command line programs
  - Java client library
- Available in XSEDE user portal



NCSA

INDIANA UNIVERSITY

Gateways