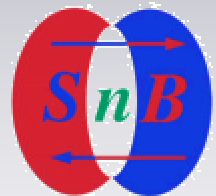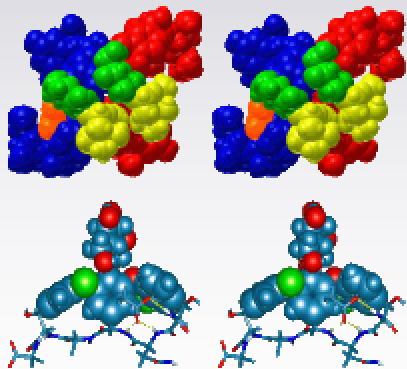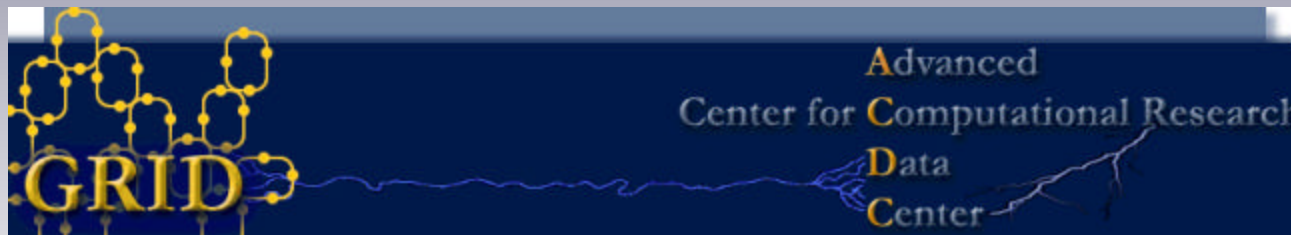# *BnP* on the Grid

Russ Miller[1,2,3], Mark Green[1,2], Charles M. Weeks[3]

[1]Center for Computational Research, SUNY-Buffalo
[2]Computer Science & Engineering SUNY-Buffalo
[3]Hauptman-Woodward Medical Research Institute

University at Buffalo
The State University of New York

Grid Computing

iVDgL

NEESgrid

Data GRID

GLOBAL GF

GriPhyN — Data Intensive Science

European GRID Forum

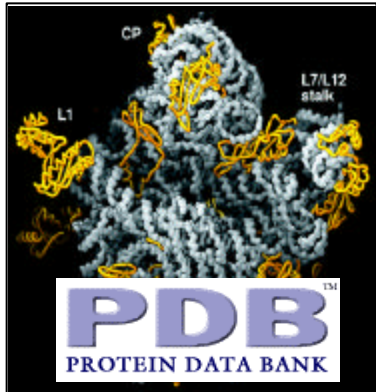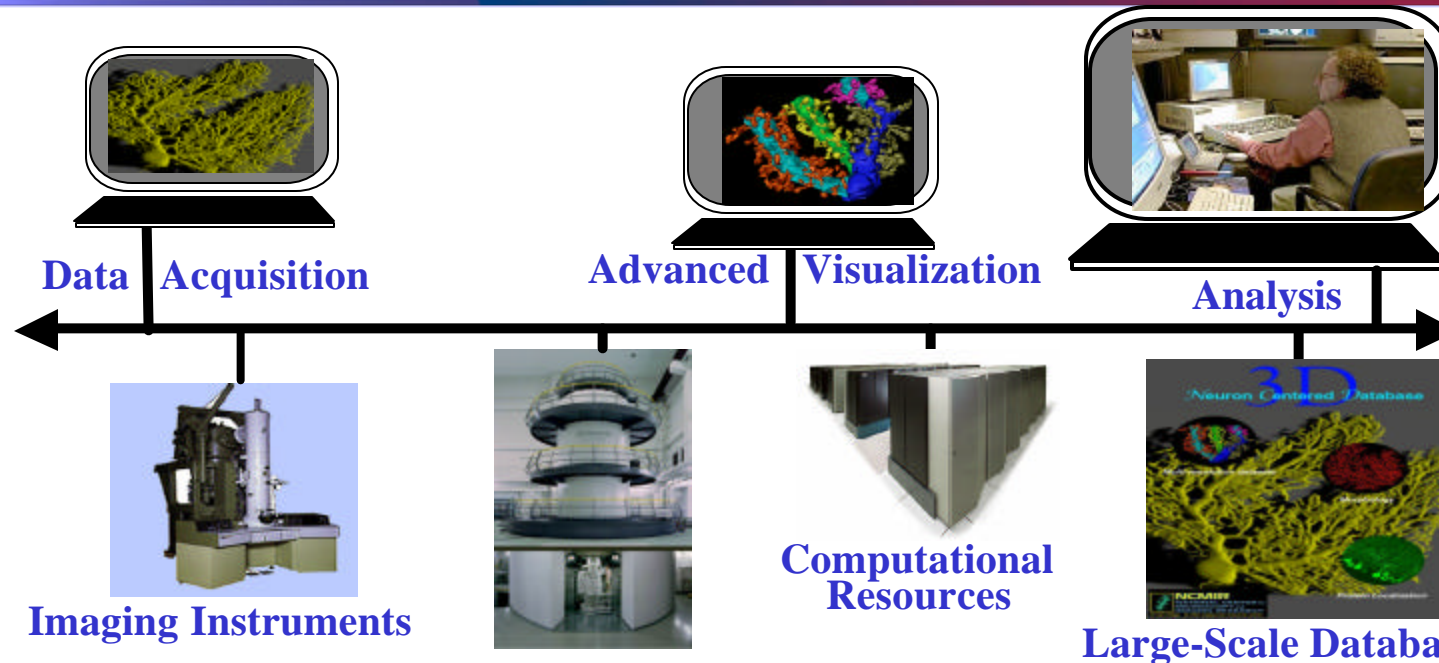SDSC/UCSD · NCSA/UIUC · Caltech · ANL

TERAGRID

NSF PACI

PDB — PROTEIN DATA BANK

United States virtual observatory

DISCOM

SinRG

APGrid

IPG …

APAN

Asia-Pacific Advanced Network

EUROGRID

Grid Computing Overview

Data Acquisition

Advanced Visualization

Analysis

Thanks to Mark Ellisman

Imaging Instruments

Computational Resources

Large-Scale Databases

- **Coordinate Computing Resources, People, Instruments in Dynamic Geographically-Distributed Multi-Institutional Environment**
- **Treat Computing Resources like Commodities**
  - ❑ **Compute cycles, data storage, instruments**
  - ❑ **Human communication environments**
- **No Central Control; No Trust**

Factors Enabling the Grid

- **Internet is Infrastructure**
  - **Increased network bandwidth and advanced services**
- **Advances in Storage Capacity**
  - **Terabyte costs less than $5,000**
- **Internet-Aware Instruments**
- **Increased Availability of Compute Resources**
  - **Clusters, supercomputers, storage, visualization devices**
- **Advances in Application Concepts**
  - **Computational science: simulation and modeling**
  - **Collaborative environments ® large and varied teams**
- **Grids Today**
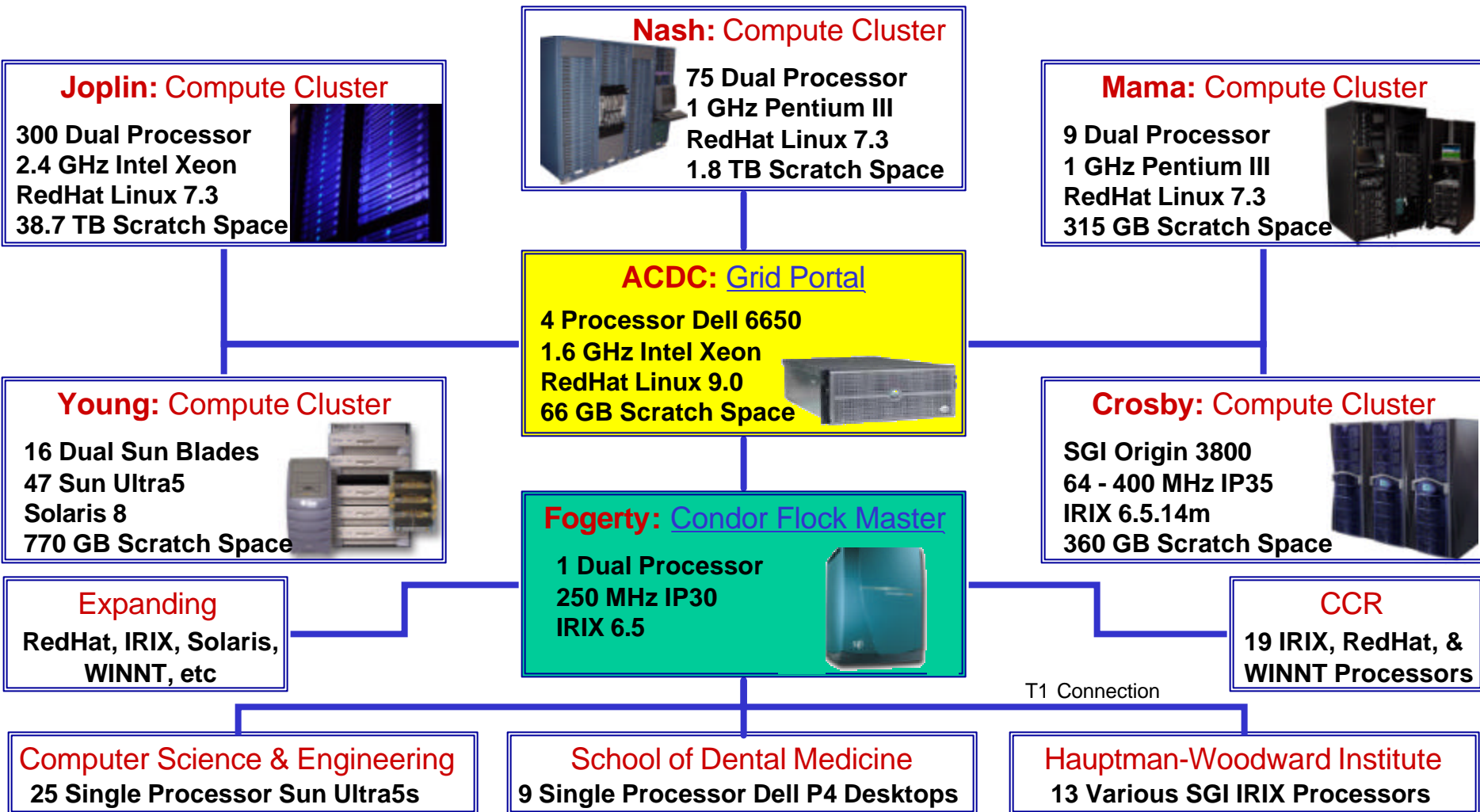 - **Moving towards production; Focus on middleware**

SnB on Grids

- **ACDC-Grid (Western New York)**
  - CIT (UB), CCR (UB), CSE (UB), Dental (UB), HWI
  - Linux, Windows, IRIX, AIX, Solaris
  - Pentium, Itanium, Power, MIPS
- **Grid3+ (International): GriPhyN, PPDG**
  - 29 Sites: ANL, SMU, BNL, BU, CalTech-Grid3, CalTech-PG, FIU, HU, IU, JHU, KNU, OU-HEP, OU-OSCER, PDSF, PSU, Rice, UB, UCSD, UCSD-Prod, UIC, UFL-Grid3, UFL-PG, UMICH, UNM, FNAL, UTA, UWMad, UWMil, Vanderbilt
  - GriPhyN, PPDG, iVDGL, LIGO, ATLAS/CMS/LHC@CERN
  - VOs: iVDGL, LIGO, SDSS, USATLAS, USCMS and BTEV
  - Linux/Pentium, VDT, Globus, ACDC Monitoring, MonaLisa, Ganglia, Condor, PBS, LSF, FBS, PyGlobus, Perl, Pacman
- **IBM NE BioGrid (Northeast USA)**
  - MIT, Harvard, MGH
  - Regatta, Pentium, Linux

Major CCR Resources (12TF & 290TB)

- **Dell Linux Cluster: #22®#25®#38®#95**
  - ❑ **600 P4 Processors (2.4 GHz)**
  - ❑ **600 GB RAM; 40 TB Disk; Myrinet**
- **Dell Linux Cluster: #187®#368®off**
  - ❑ **4036 Processors (PIII 1.2 GHz)**
  - ❑ **2TB RAM; 160TB Disk; 16TB SAN**
- **IBM BladeCenter Cluster: #106**
  - ❑ **532 P4 Processors (2.8 GHz)**
  - ❑ **5TB SAN**
- **SGI Origin3700 (Altix)**
  - ❑ **64 Processors (1.3GHz ITF2)**
  - ❑ **256 GB RAM**
  - ❑ **2.5 TB Disk**
- **SGI Origin3800**
  - ❑ **64 Processors (400 MHz)**
  - ❑ **32 GB RAM; 400 GB Disk**

- **Apex Bioinformatics System**
  - ❑ **Sun V880 (3), Sun 6800**
  - ❑ **Sun 280R (2)**
  - ❑ **Intel PIIIs**
  - ❑ **Sun 3960: 7 TB Disk Storage**
- **HP/Compaq SAN**
  - ❑ **75 TB Disk**
  - ❑ **190 TB Tape**
  - ❑ **64 Alpha Processors (400 MHz)**
  - ❑ **32 GB RAM; 400 GB Disk**

- **IBM RS/6000 SP: 78 Processors**
- **Sun Cluster: 80 Processors**
- **SGI Intel Linux Cluster**
  - ❑ **150 PIII Processors (1 GHz)**
  - ❑ **Myrinet**
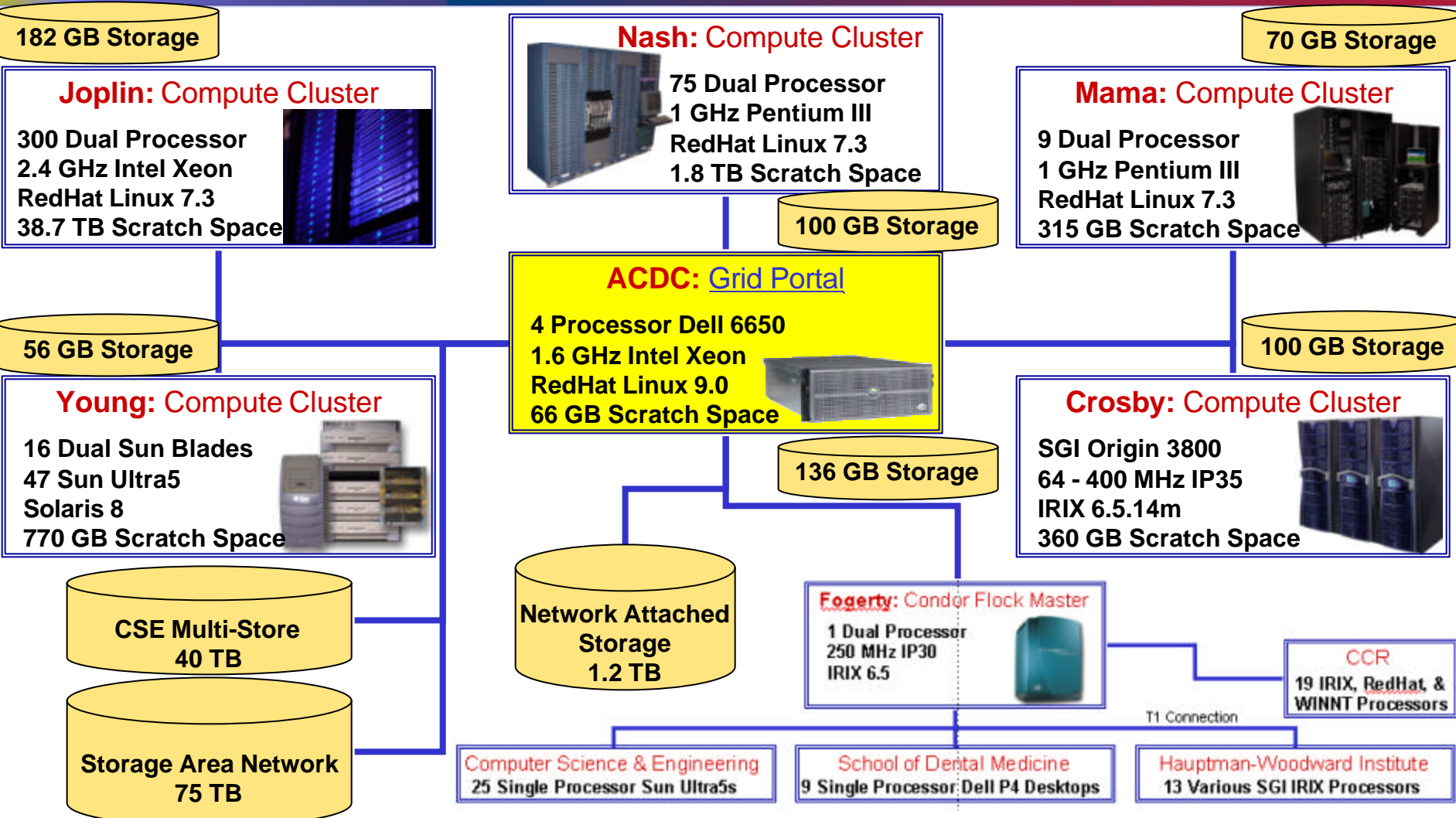
Advanced Computational Data Center
ACDC: Grid Overview

Nash: Compute Cluster

75 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
1.8 TB Scratch Space

Joplin: Compute Cluster

300 Dual Processor
2.4 GHz Intel Xeon
RedHat Linux 7.3
38.7 TB Scratch Space

Mama: Compute Cluster

9 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
315 GB Scratch Space

ACDC: Grid Portal

4 Processor Dell 6650
1.6 GHz Intel Xeon
RedHat Linux 9.0
66 GB Scratch Space

Young: Compute Cluster

16 Dual Sun Blades
47 Sun Ultra5
Solaris 8
770 GB Scratch Space

Crosby: Compute Cluster

SGI Origin 3800
64 - 400 MHz IP35
IRIX 6.5.14m
360 GB Scratch Space

Fogerty: Condor Flock Master

1 Dual Processor
250 MHz IP30
IRIX 6.5

Expanding
RedHat, IRIX, Solaris, WINNT, etc

CCR
19 IRIX, RedHat, & WINNT Processors

T1 Connection

Computer Science & Engineering
25 Single Processor Sun Ultra5s

School of Dental Medicine
9 Single Processor Dell P4 Desktops

Hauptman-Woodward Institute
13 Various SGI IRIX Processors

Note: Network connections are 100 Mbps unless otherwise noted.

University at Buffalo *The State University of New York* **Center for Computational Research**

CCR

ACDC Data Grid Overview
(Grid-Available Data Repositories)

182 GB Storage

Joplin: Compute Cluster

300 Dual Processor
2.4 GHz Intel Xeon
RedHat Linux 7.3
38.7 TB Scratch Space

Nash: Compute Cluster

75 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
1.8 TB Scratch Space

70 GB Storage

Mama: Compute Cluster

9 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
315 GB Scratch Space

100 GB Storage

ACDC: Grid Portal

4 Processor Dell 6650
1.6 GHz Intel Xeon
RedHat Linux 9.0
66 GB Scratch Space

56 GB Storage

Young: Compute Cluster

16 Dual Sun Blades
47 Sun Ultra5
Solaris 8
770 GB Scratch Space

100 GB Storage

Crosby: Compute Cluster

SGI Origin 3800
64 - 400 MHz IP35
IRIX 6.5.14m
360 GB Scratch Space

136 GB Storage

CSE Multi-Store
40 TB

Storage Area Network
75 TB

Network Attached Storage 1.2 TB

Fogerty: Condor Flock Master

1 Dual Processor
250 MHz IP30
IRIX 6.5

CCR
19 IRIX, RedHat, & WINNT Processors

T1 Connection

Computer Science & Engineering
25 Single Processor Sun Ultra5s

School of Dental Medicine
9 Single Processor Dell P4 Desktops

Hauptman-Woodward Institute
13 Various SGI IRIX Processors

Note: Network connections are 100 Mbps unless otherwise noted.

ACDC-Grid Cyber-Infrastructure

- **Predictive Scheduler**
  - ❏ **Define quality of service estimates of job completion, by better estimating job runtimes by profiling users.**
- **Data Grid**
  - ❏ **Automated Data File Migration based on profiling users.**
- **High-performance Grid-enabled Data Repositories**
  - ❏ **Develop automated procedures for dynamic data repository creation and deletion.**
- **Dynamic Resource Allocation**
  - ❏ **Develop automated procedures for dynamic computational resource allocation.**

ACDC-Grid

University at Buffalo *The State University of New York* **Center for Computational Research** **CCR**

ACDC-Grid Administration

Molecular Structure Determination via *Shake-and-Bake*

- *SnB* **Software by UB/HWI**
  - ❑ **"Top Algorithms of the Century"**
- **Worldwide Utilization**
- **Critical Step**
  - ❑ **Rational Drug Design**
  - ❑ **Structural Biology**
  - ❑ **Systems Biology**

- **Vancomycin**
  - ❑ **"Antibiotic of Last Resort"**
- **Current Efforts**
  - ❑ **Grid**
  - ❑ **Collaboratory**
  - ❑ **{Intelligent Learning}**

Source of X-rays   Crystal

File Mode Help

Manual Mode

| About BnP | General Information | Reflections & Invariants | SnB Setup | Run SnB | Evaluate Trials | Protein Phasing |

Title: Trivanco
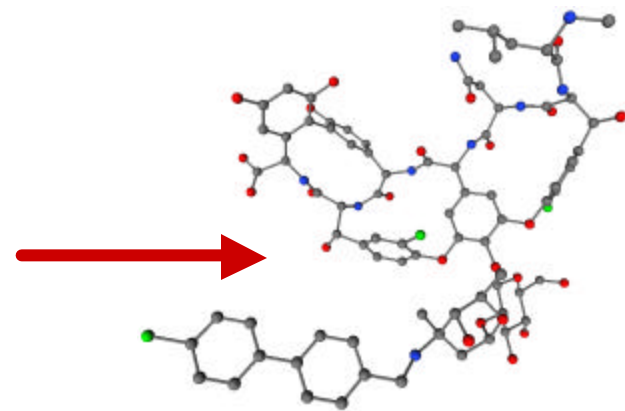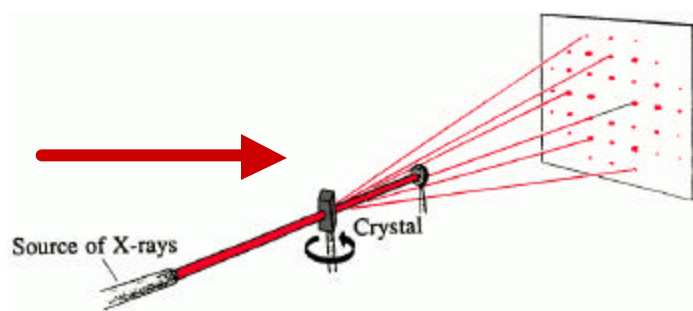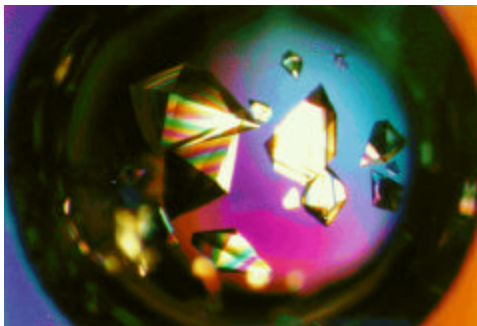
Structure ID (16 char. max.): trivanco   **Space Group:** P1 (1)   ...

Cell Constants and Cell Errors (Cell Errors optional)

| **A:** 21.4 | **+/-** 0.1 | **B:** 24.3 | **+/-** 0.1 | **C:** 24.9 | **+/-** 0.1 |
| **Alpha:** 64.5 | **+/-** 0.2 | **Beta:** 90.2 | **+/-** 0.2 | **Gamma:** 81.0 | **+/-** 0.2 |

Native Asymmetric Unit Contents

Given R residues, then by default use C5R, N1.2R, O1.5R, H8R. **No. Residues (optional):**

Contents: C264H300N36O96Cl8

Examples: C6H12O6 or C6 H12 O6 or C6,H12,O6

Initial Data Sets

[Add] [Delete]

	1
Name (8 char. max.)	vancohkl
File name	trivanco.hkl
File type	F, Sig(F)
Data set type	Native
Wavelength	1.5418
Max. resolution	0.97
Anomalous dispersion	Not measured
Heavy element type	
Nat. element replaced	
No. expected sites	
f'	
f"	

General Information Screen for TriVanco

File   Mode   Help

| About BnP | General Information | Reflections & Invariants | SnB Setup | Run SnB | Evaluate Trials | Protein Phasing |

Normalize reflections and generate invariants for SnB jobs

Output Data Set	Job Type	Input Data Sets Native	Derivative	Normalization Method	Select ?
vancohkl	Basic	vancohkl		Wilson (anisotropic) ▼	☑

Prepare data for SnB runs.

1. Select one or more data sets.
2. If desired, choose a non–default normalization method.
3. Click "Normalize" to compute E values.
4. Click "Generate" to create a structure invariant file.

| Normalize... | Generate... |

Merge and scale reflections for use in PHASES

Output File	File Type	Input Data Sets Native	Derivative	Relative Wilson Scaling Method	Select ?

Prepare data for protein phasing.

1. Select data sets to use.
2. If desired, choose a non–default scaling method.
3. Click "Merge & Scale".

For each selected set, the appropriate merged & scaled file (*.scl) will be created.

| Merge & Scale |

Review Results

Normalize	Generate Invariants	Merge & Scale
vancohkl	vancohkl	

Update List

View Results

Clean Files *

* for Normalize only

Normalize Reflections and Generate Invariants

Setting saved to trivanco.xml

Start | INBOX -... | joplin.cc... | plexus.... | Staff Di... | nash.cc... | vancote... | nash.cc... | GridCo... | trivanc... | Screens... | 10:51 AM

File Mode Help

| About BnP | General Information | Reflections & Invariants | SnB Setup | Run SnB | Evaluate Trials | Protein Phasing |

Invariants

Input invariant file: /nas/gridportal/test/vancoBnP/vancohkl/vancohkl.SnB_inv Browse...

Number of triplet invariants to use: 40400

Trials to Process

Starting phases from: Random Atoms ▼   Random seed (prime): 11909 __

Number of trials: 10000   Input phase file: _____ Browse...

Starting at trial: 1   Input atom file : _____ Browse...

Keep complete (every trial) peak file? ● Yes ○ No

Cycles

Number of Shake-and-Bake cycles: 404   Keep complete (every cycle) trace file? ○ Yes ● No

Terminate trials failing the R-Ratio test? ○ Yes ● No   R-Ratio cutoff: 0.20

Phase Refinement Method

● Parameter Shift ○ Tangent Formula   Number of passes through phase set: 1

Phase shift: 129.0   Number of shifts: 1

Real-Space Constraints

Number of peaks to select: 161   Minimum interpeak distance: 1.0   Fourier grid size: 0.32

Minimum distance between symmetry-related peaks: 3.0   (defines special position excluded volume)

Number of special position peaks to keep: 0

Perform extra cycles with more peaks? ● Yes ○ No   Number of extra cycles: 40   Number of peaks: 404

Twice Baking

Trials for E-Fourier filtering (fourier refinement)? ● None ○ All ○ Best Only

Number of cycles: 40   Number of peaks: 444   Minimum |E|: 0.75

Automatic solution identification criteria for substructures

Rmin Improvement (%): 45.0   Rcryst Improvement (%): 25.0

Define *SnB* Parameters: 10,000 trials; 404 cycles; 40400 triplets

Setting saved to trivanco.xml

Start | INBOX - ... | joplin.cc... | plexus.... | Staff Di... | nash.cc... | vancote... | nash.cc... | GridCo... | trivanc... | Screens... | 10:52 AM

Run *SnB* job on Grid

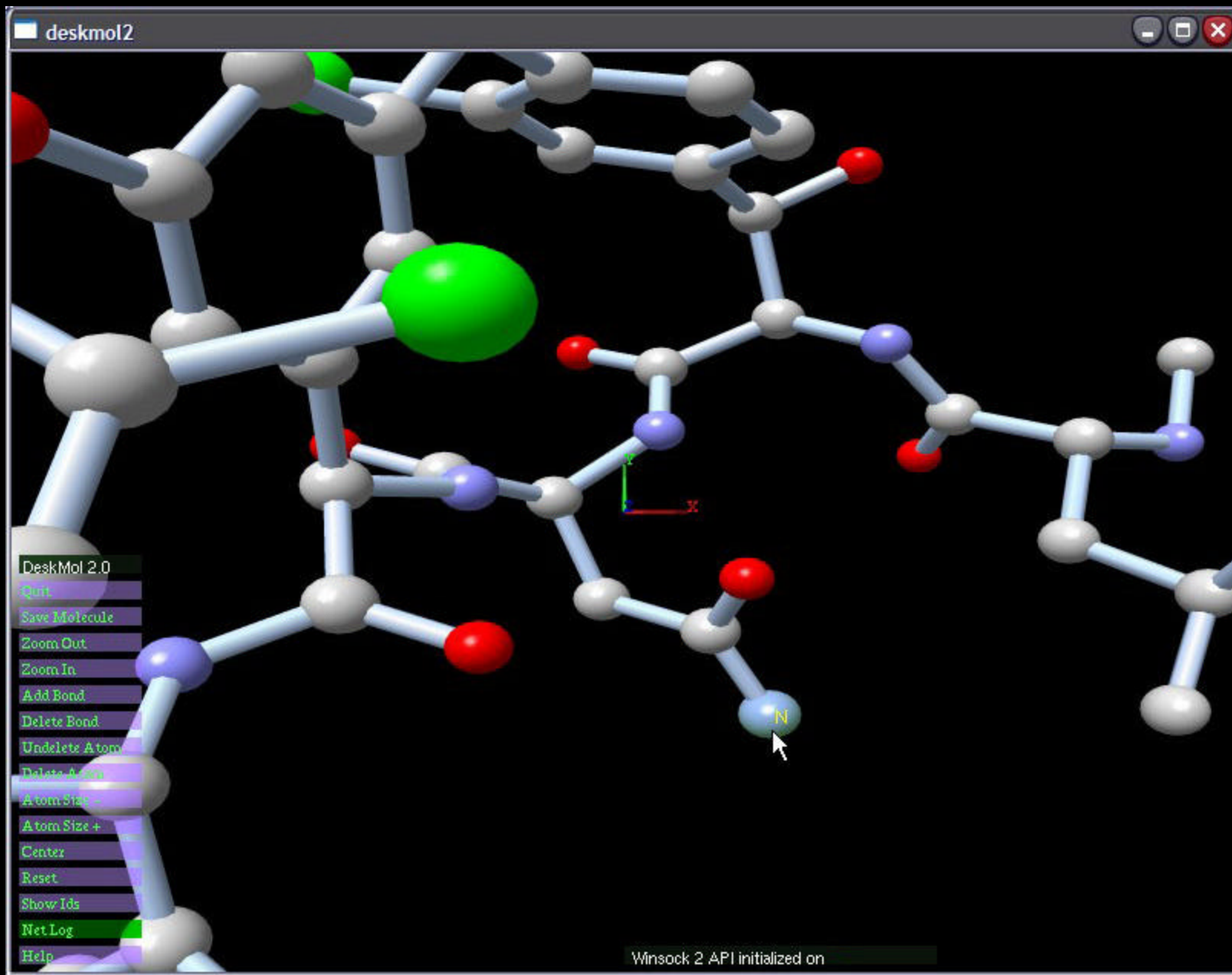Histogram of Final Rmin Values after 200 Trials Completed
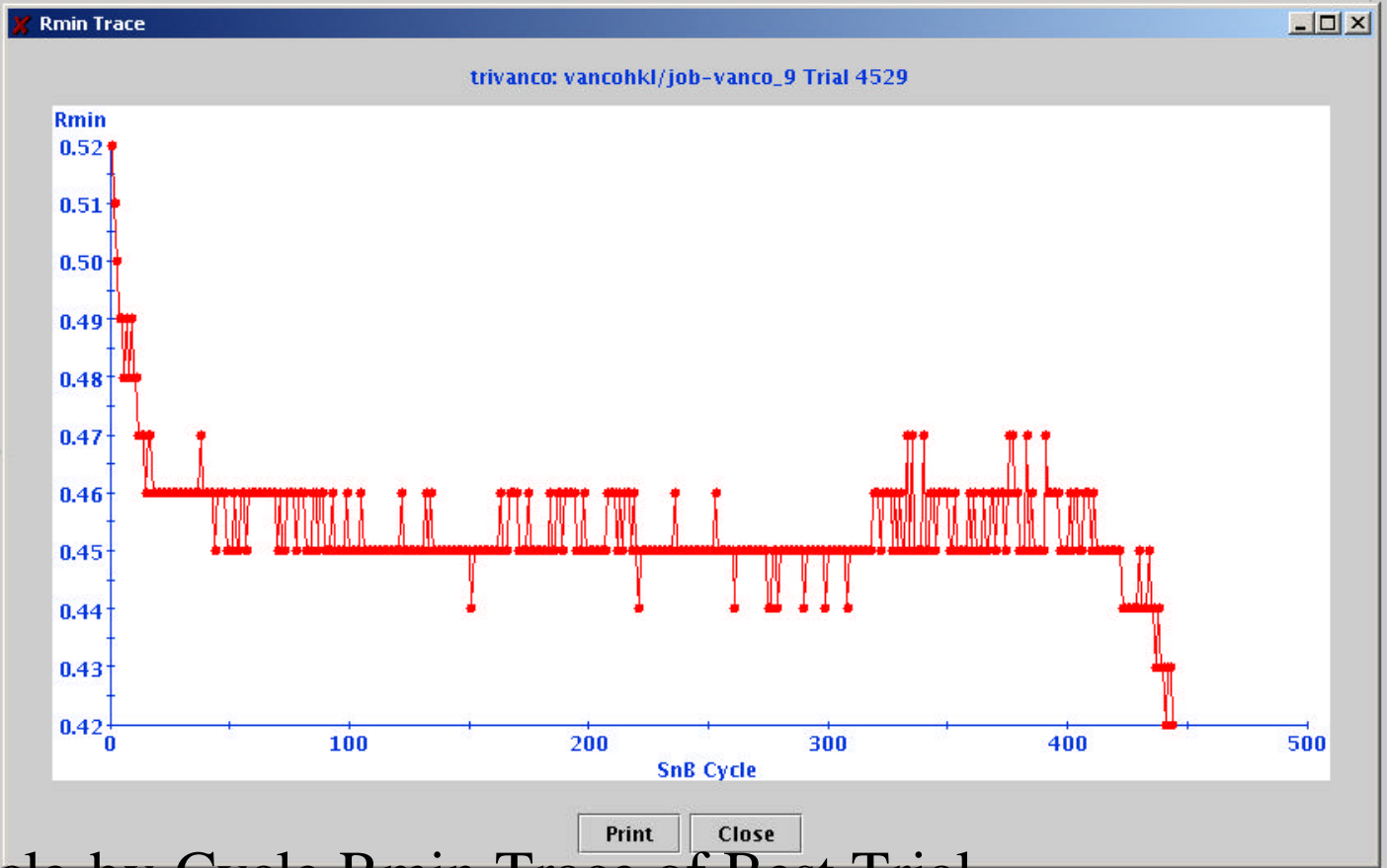
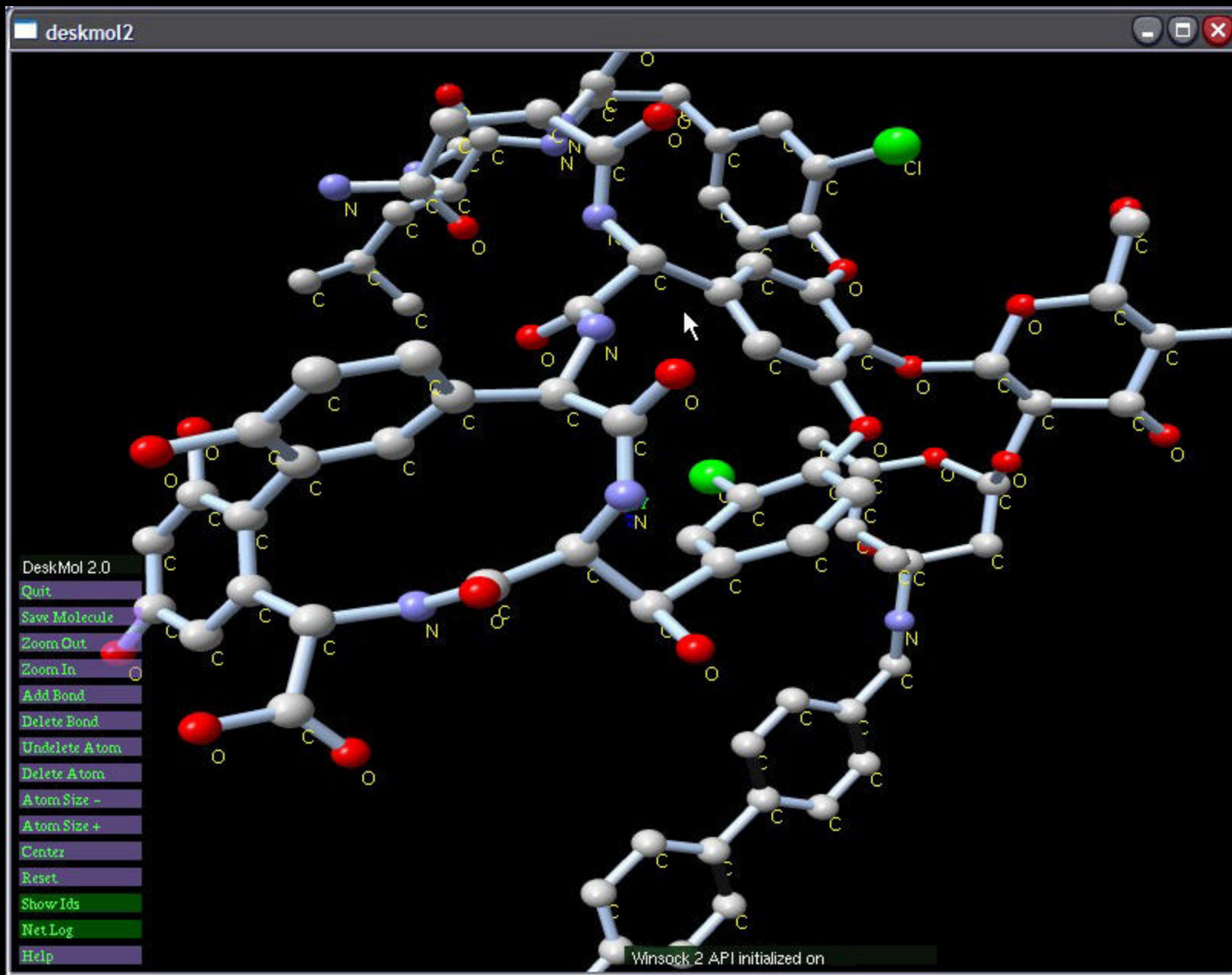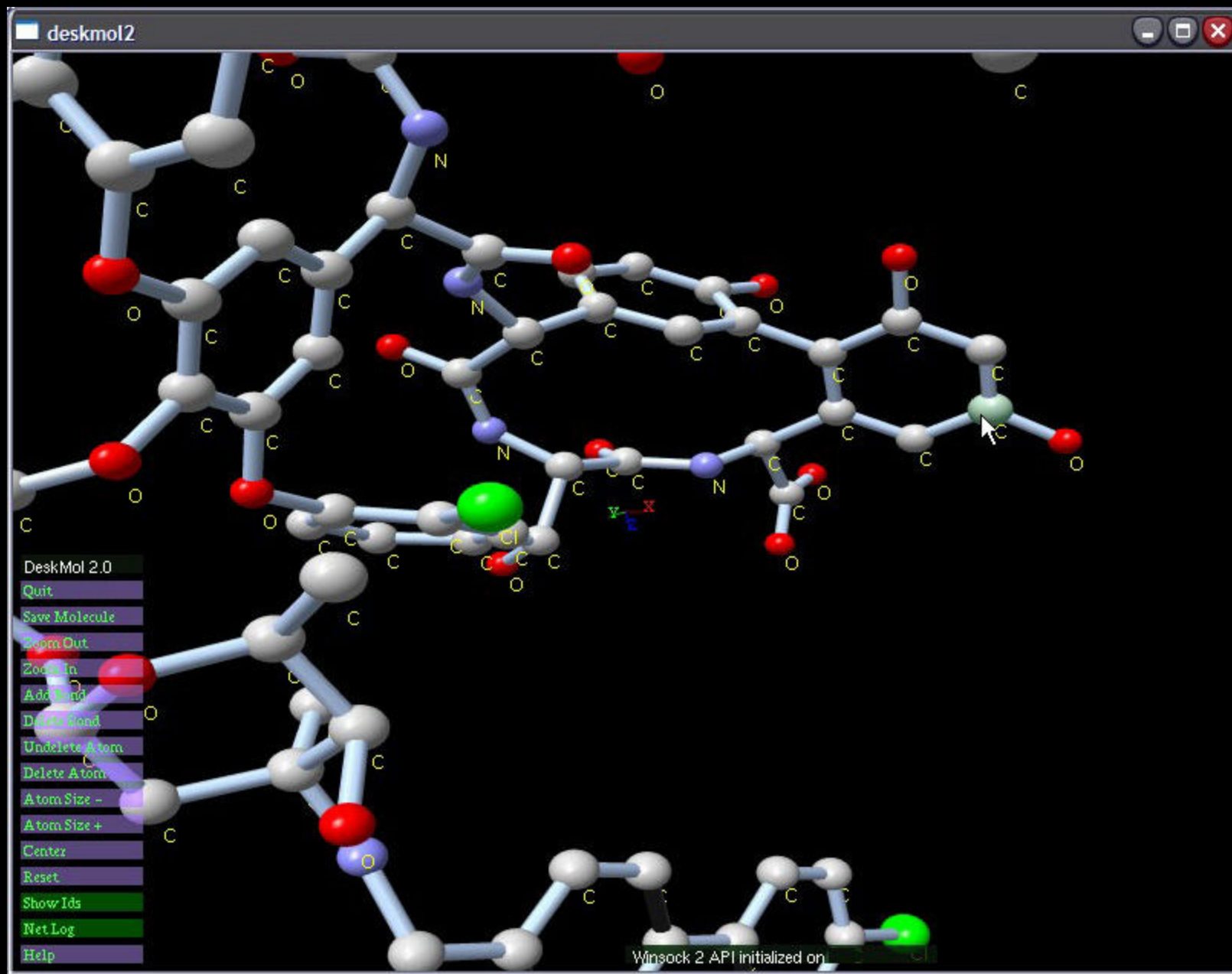Cycle-by-Cycle Rmin Trace of Best Trial

Live Demo

- **Time to Gamble**
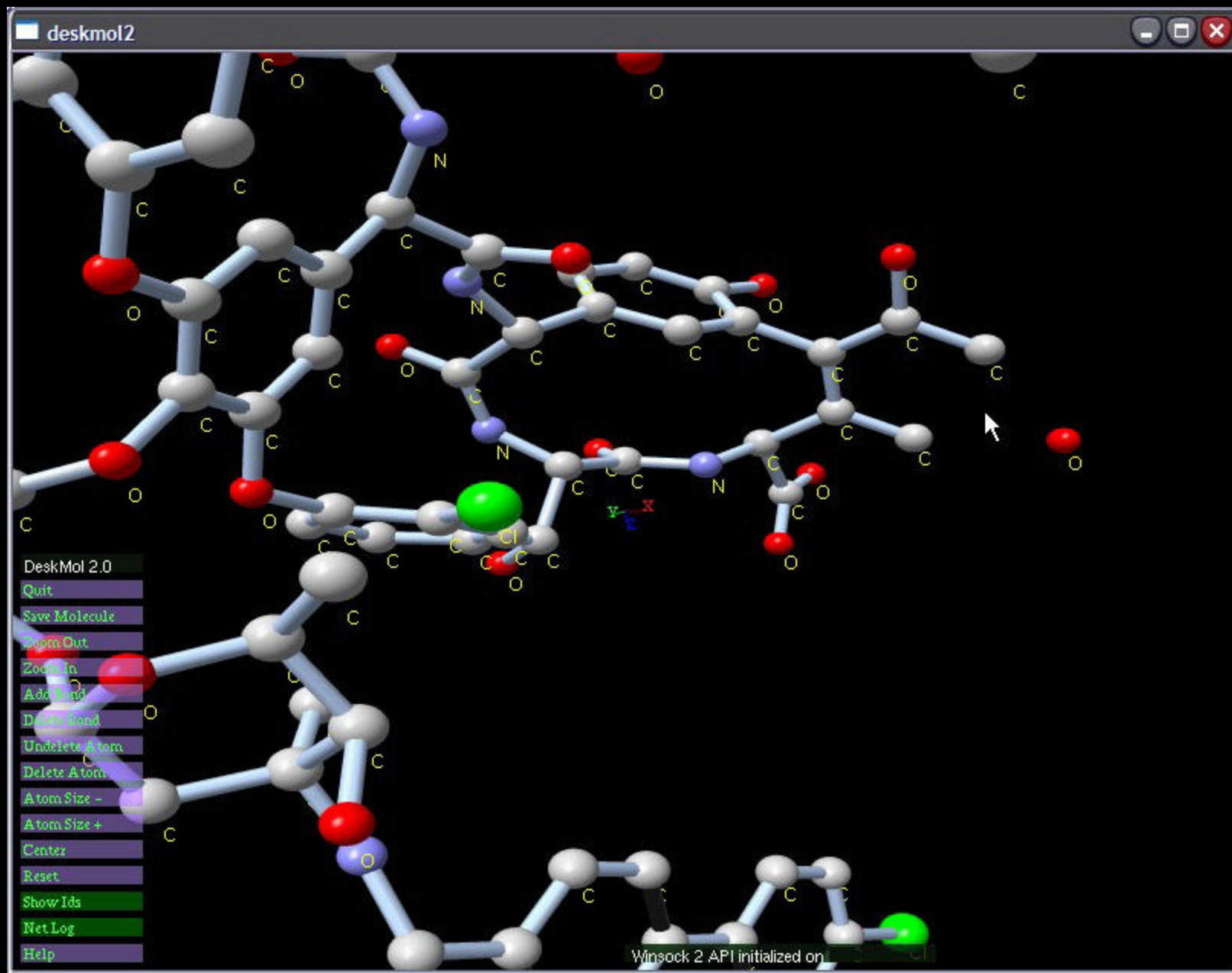
- **Demonstration using ACDC Grid in Buffalo**
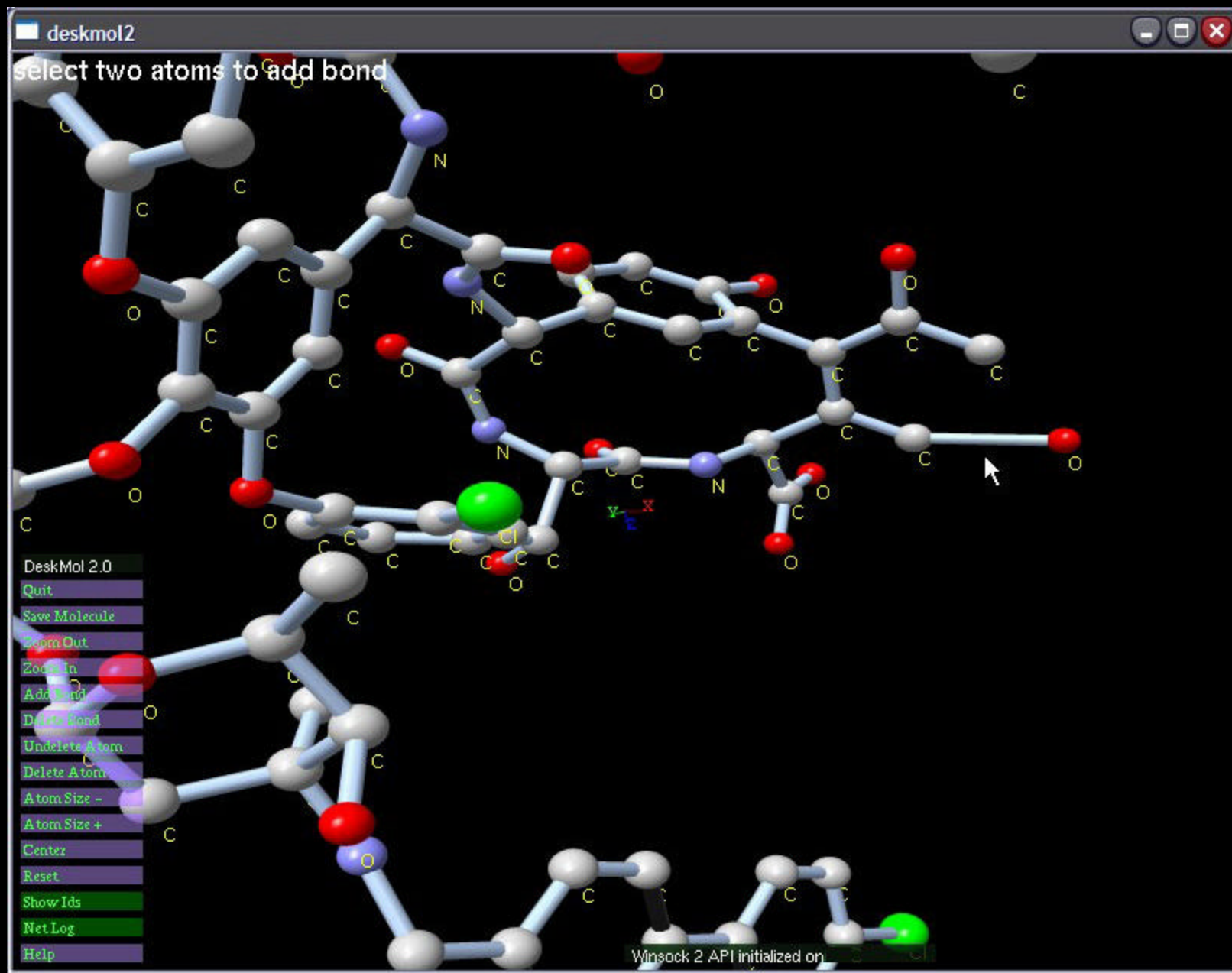
User starts up – default image of structure.

Molecule scaled, rotated, and labeled.
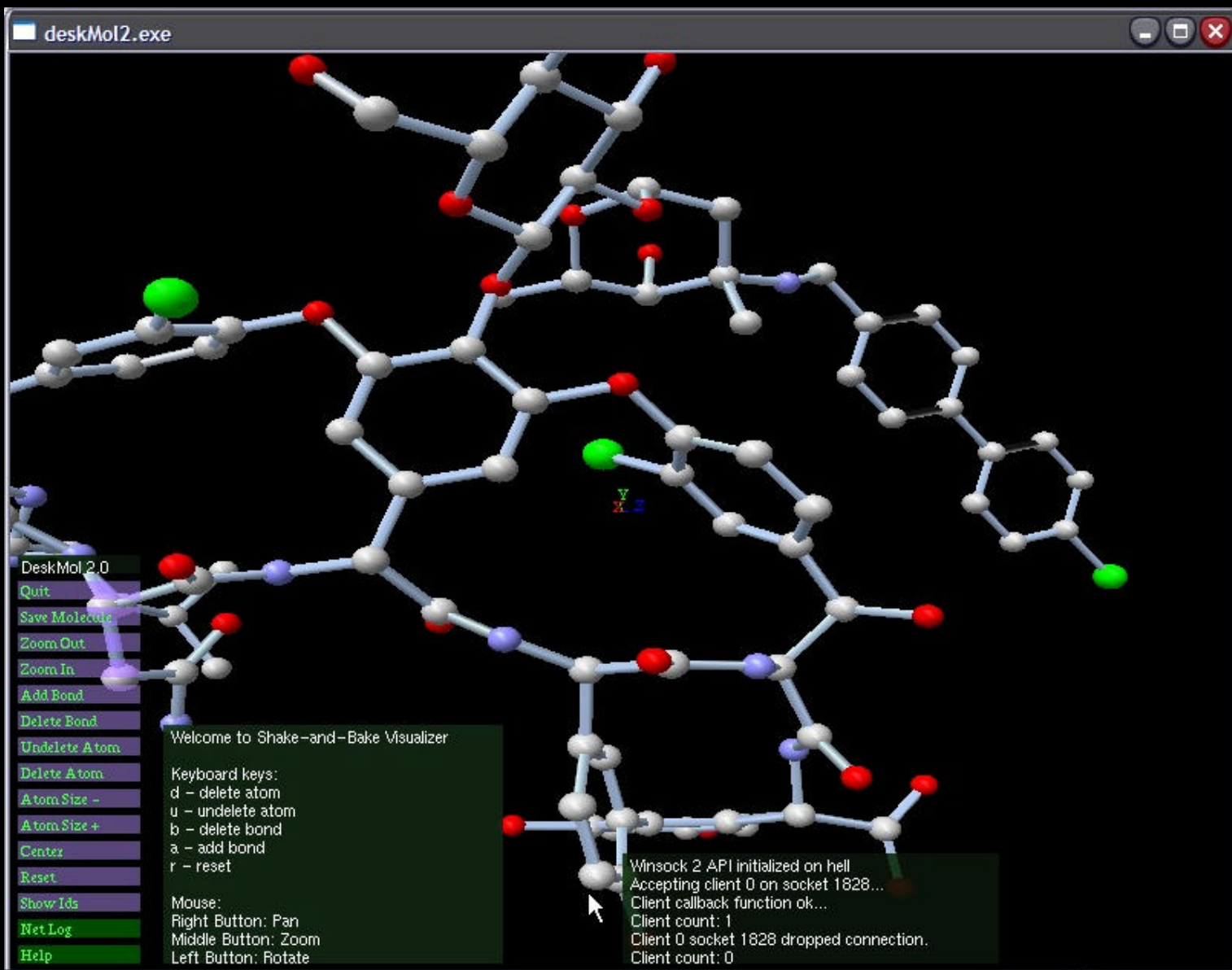
User Mouse to Select Carbon Atoms

Remove Carbon Atoms (and Links)

User Adds Bond Between Atoms

Scale Radius of Atoms

Continue Scaling Atoms
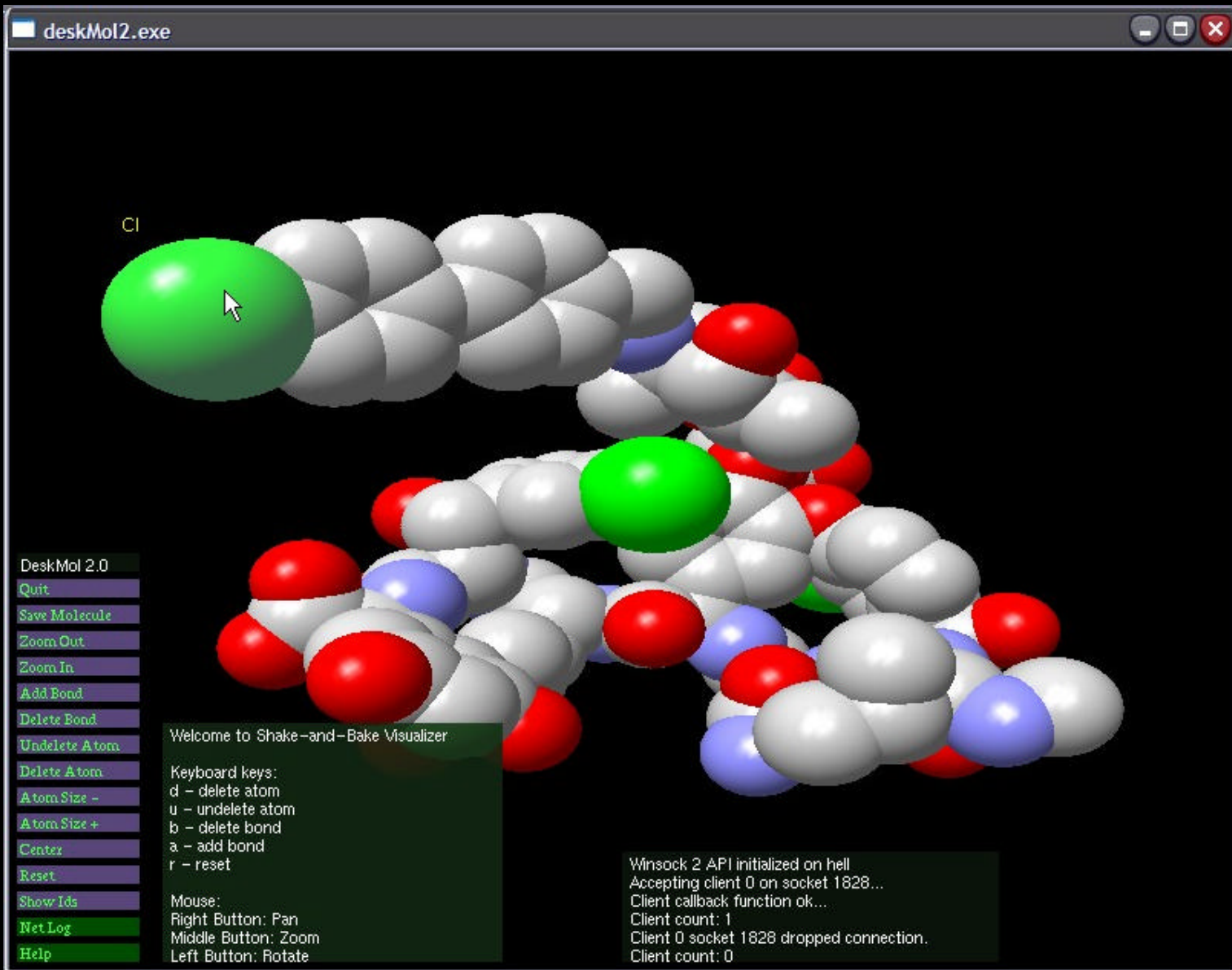
Continue Scaling Atoms

Middleware

- **Grid (Computational and Data)**
  - ❑ **Globus Toolkit 2.2.4 ® direct upgrade WSRF**
  - ❑ **Condor 6.6.0**
  - ❑ **Network Weather Service 2.6**
  - ❑ **Apache2 HTTP Server**
  - ❑ **PHP 4.3.0**
  - ❑ **MySQL 3.23**
  - ❑ **phpMyAdmin 2.5.1**
- **Collaboratory**
  - ❑ **OpenGL (LibDMS, DevIL, GLUT)**
  - ❑ **Windows, IRIX, Mac OS X, Linux**
  - ❑ **CAVE, Desktop**

ACDC-Grid Collaborations

- **High-Performance Networking Infrastructure**
- **WNY Grid Initiative**
- **Grid3+ Collaboration**
- **iVDGL Member**
  - ❑ **Only External Member**
- **Open Science Grid Member**
  - ❑ **Organizational Committee**
  - ❑ **Blueprint Committee**
  - ❑ **Security Working Group**
  - ❑ **Data Working Group**
- **Grid-Based Visualization**
  - ❑ **SGI Collaboration**
- **Grid-Lite: Campus Grid**
  - ❑ **HP Labs Collaboration**
- **Innovative Laboratory Prototype**
  - ❑ **Dell Collaboration**

Acknowledgments

- **Amin Ghadersohi**
- **Naimesh Shah**
- **Stephen Potter**
- **Cathy Ruby**

- **Steve Gallo**
- **Jason Rappleye**
- **Martins Innus**
- **Jon Bednasz**
- **Sam Guercio**
- **Dori Macchioni**

www.ccr.buffalo.edu