

The logo for EMEN2, featuring the letters 'EM' stacked above 'EN' in a black box, with a red '2' to the right.

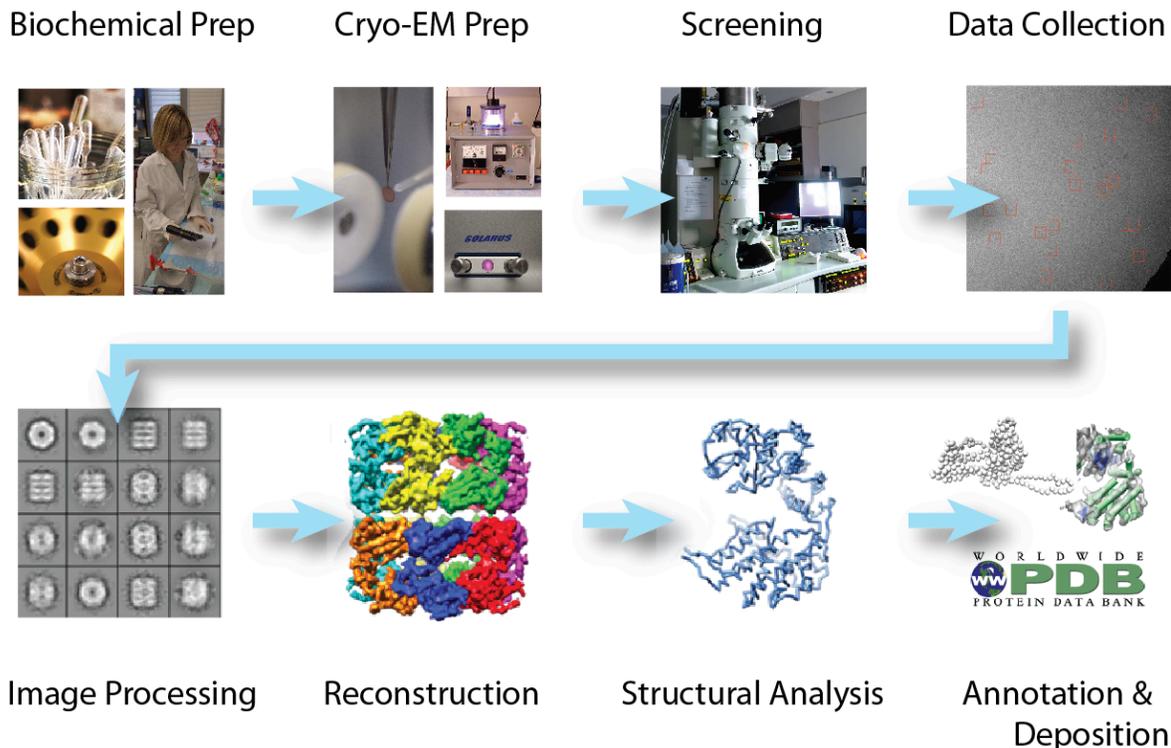
EMEN2 Tutorial

NCMI Workshop on Single Particle Reconstruction, Validation, and Analysis

March 14-17 2011

Ian Rees

Introduction



Cryo-EM data sets are data-intensive and interconnected, with several stages in a typical pipeline: biochemical purification, cryo-EM preparation and freezing, screening and data collection, image processing, reconstruction, modeling and structural analysis, and publication and deposition of results. Well-documented workflows are critical not just for reproducibility, but also for understanding the numerous experimental factors that can influence the quality of the image data and reconstruction. Additionally, workflows themselves are constantly changing as experimental protocols are optimized and techniques improve, and may also vary from group to group. This presents an archiving challenge as any database schema will undergo constant revisions or fail to record data in a searchable and mineable way.

To address these questions we have developed EMEN2, an object-oriented database designed for storage and mining of scientific data in collaborative environments.

EMEN2 builds on our experience with our previous EMEN system (2001-2007) and other database technologies. It is designed to be easy to install and maintain, on simple hardware, without the need for a full-time system administrator. A “Web 2.0” style interface is provided, as well as a programmatic API for client programs and instrument integration.

Architecture

EMEN2 is built around two fundamental concepts: experimental parameters and experimental protocols:

- An experimental parameter consists of a unique name, a description of what the parameter measures or represents, a data type (string, integer, list, etc.), and physical property and default units if applicable.
- An experimental protocol, similar to a 'class' in object-oriented terminology, consists of a plain-text description of the experiment, with embedded parameter names indicating where measurements are taken or decisions are made.

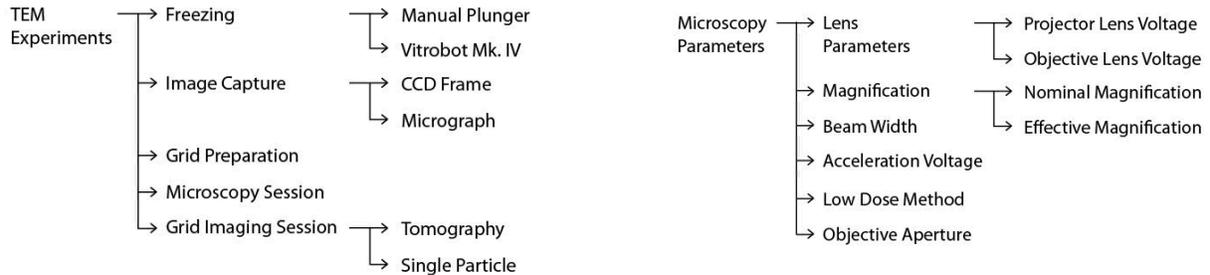
For example, a cryo-EM vitrification protocol might describe a number of steps, with one or more values recorded during each step, such as specimen concentration, blotting time, humidity and temperature, buffer concentrations used, etc. Each of these recorded values is associated with an appropriate parameter type. Parameters defined for one experimental protocol can and should be reused in any other protocol where the parameter has the same contextual meaning.

New protocols are usually derived from existing protocols, and remain associated with the original, allowing tracking of changes in technique. Protocols are also used to represent types of equipment and organizational elements such as projects.

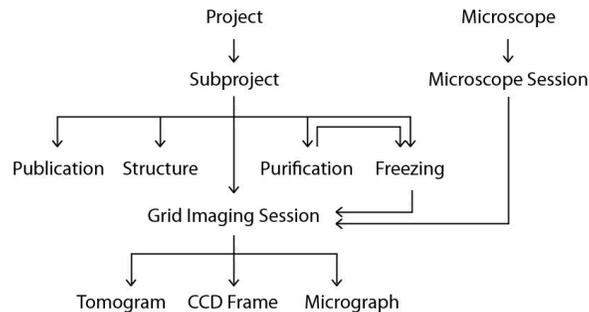
A record is a collection of parameter/value pairs, with the protocol description used as a form to create and view records. While each record is an instance of a protocol, additional "out-of-band" parameters are allowed.

Relationships

A third fundamental concept in EMEN2 is the use of a hierarchy to describe relationships between parameter and protocol definitions, creating an ontology of terms and experiments that provides unique organizational and descriptive power.



For instance, there is a generic protocol for images, `image_capture`, with specific image types as children: `ccd`, `micrograph`, `stack`, etc. These relationships are frequently used to broaden or narrow queries. Each item may have several parents and children.



Records are organized in the same way, allowing a great deal of contextual information to be stored without explicit duplication of data. For example, this capability is used for microscopy sessions: an imaging session has both the project and microscope records as parents, permitting the simple query 'Find all ccd frames in the GroEL project collected on the JEOL2200 microscope.'

Collaborators, Publishing, Deposition



EMEN2 is a multi-user system, with a fine-grained per-record security model. Each record also maintains a full history of all changes, including previous values, the date of the change, the user who made the change, and any comments describing the purpose of the change. Users can subscribe for daily email updates of new and changed records in specified projects, as well as immediate notification of any comments made on a record.

Many journals now support “open science” initiatives, where all raw and intermediate data is made available with a publication. EMEN2 supports marking subsets of a project as “published,” which permits a specified (anonymous, email-required, or registration-and-approval required) level of public access. NCMI public data server: <http://ncmi.bcm.edu/publicdata>

A data harvesting module is also being developed to help prepare submissions to resources such as the PDB / EMDatabank.

Implementation and Availability

EMEN2 is completely open-source, written in pure Python, and is built on top of Berkeley DB, an open-source high-performance key/value embedded database from Oracle. It is fairly simple to extend EMEN2 with custom modules that are accessible via the Web interface or through the API, and we have used this several times to use the software in new situations, including hosting the Pacific Symposium on Biocomputing 2011 Cryo-EM Modeling Challenge.

The system has been in production use at the NCMI for about 3 years, and contains nearly 500,000 records, 16 terabytes of raw data, and more than 700 users.

We are very interested in supporting new users. You may download EMEN2 and access the documentation, including this tutorial, at the EMEN2 Wiki:

<http://blake.bcm.edu/emanwiki/EMEN2>

Note: Python environments

Most EMEN2 commands are started using `python -m emen2.command`. This is a convenient way of running Python modules that requires fewer changes to your shell variables. If you installed EMEN2 into the Python environment included with the EMAN2 binaries for Linux (recommended), you will need to use that interpreter. This will be `~/EMAN2/Python/bin/python` on the workshop computers. For the purpose of this tutorial, it may be simpler to create an alias:

```
$ alias python=~/EMAN2/Python/bin/python
```

On Mac OS X, the EMAN2 binaries use the system's Python environment, so usually just `python` is sufficient. Currently, EMEN2 requires Python 2.6 or higher, available in Mac OS X 10.6 and above. Mac OS X 10.5 uses Python 2.5 and is not supported.

Note: Database environments

EMEN2 requires a directory for the database environment. All database files, log files, temporary files, and raw data will be stored under this directory in the default configuration. You can specify the environment directory as an argument to EMEN2 commands using `-h ~/workshop/emen2/db` or similar, or set the `EMEN2DBHOME` environment variable:

```
$ export EMEN2DBHOME=~/workshop/emen2/db
```

This has already been set on the workshop computers, so the `-h` argument will not be necessary. If you following the tutorial with a laptop, you may want to do this as well.

Load Data

First, we will initialize the EMEN2 database using a data set based on NCMI's published raw data. The data was exported to a JSON-based flat file format, consisting of `records.json`, `users.json`, etc. This data is preloaded on the workshop computers in `~/workshop/emen2/emen2-demo` or can be downloaded from the workshop wiki page (<http://blake.bcm.edu/emanwiki/Ws2011/Emen2>) and untarred.

```
$ cd ~/workshop/emen2/emen2-demo
$ python -m emen2.db.load
[2011-03-15 11:53:54]:LOG_INIT :: Loading config: config.base.json
[2011-03-15 11:53:54]:LOG_INIT:database.py:357 :: Installing default
    DB_CONFIG file: /Users/irees/test_db/DB_CONFIG
[2011-03-15 11:53:54]:LOG_INFO:database.py:327 :: Opening Database
    Environment: /Users/irees/test_db
[2011-03-15 11:54:01]:LOG_INFO:database.py:408 :: Opened database with 0
    records
=== New Database Setup ===
Admin (root) email (default irees@muta.local): ian.rees@bcm.edu
Admin (root) password (default: none):
[2011-03-15 11:54:06]:LOG_INFO:database.py:469 :: Initializing new database;
    root email: ian.rees@bcm.edu
```

A new EMEN2 database environment will be initialized, and you will be prompted for a root account password and email address. This will run for 2-3 minutes while all the schema, users, and data is imported (EMEN2's logging is fairly verbose.)

Start Web Server

Most users interact with EMEN2 using the web interface, which is part of the `emen2.web` module. You can start the server with the command:

```
$ python -m emen2.web.server
```

Access the web server at <http://localhost:8080> and login using “root” and the password you specified earlier.

Note: Internet Explorer is not recommended; please use Firefox, Chrome, or Safari.

There are several elements on the home screen, depending on your account privileges and configuration:

- Profile information and photo
- Pending user accounts
- Welcome message and recent notices from the administrator
- Recently created records that you have permissions to view
- A list of of your projects

EMEN2: Home
localhost:8080

EMEN2
Home Sitemap Query Projects Equipment Params Protocols Users Groups Help Logged in as root Logout

Home

Ian Rees (root) [Edit Profile](#)

Department: Biochemistry
Institution: Baylor College of Medicine

Address: 1 Baylor Plaza
None
Houston TX, 77030 USA

Email: ian@ianrees.net
Phone:
Fax:
Web:



Users (0 pending)

- User Management
- Group Management

Yes	No	Username	Name	Email	Phone
Accept / Reject Users					

Welcome to EMEN2 [Edit](#)

This database stores (or will store) the raw data and experimental records related to our various published structures. We firmly believe that transparency is critical to scientific advancement, and when possible such data should be available to anyone for development or testing purposes. We ask only that you contact us prior to using any of this data in a new publication (wah@bcm.edu or sludtke@bcm.edu).

The web pages you are currently viewing is the main interface for the EMEN2 (Electron Microscopy Electronic Notebook) database that we develop and use internally to store records of all experiments in the NCMI. This server is not our main database, of course, but simply a clone of the portion of the data which has been released to the public. The interface should be fairly simple to learn, but if you'd like a quick tutorial, just go to:

[Simple EMEN2 Tutorial](#)

Currently, data for the following projects is available. We will be adding more over the coming months:

- Demo Microscope
- Subproject 6 Ang GroEL Film Data
- Subproject Data for 4 A structure of GroEL
- Subproject Epsilon 15
- Subproject PSSP7
- Subproject Rice Dwarf Virus

[All Records](#) • [Projects](#) • [My Records](#) • [My Imaging Sessions \(All\)](#) • [My Images \(All\)](#) • [My Lab Notebooks \(All\)](#) • [My Publications \(All\)](#)

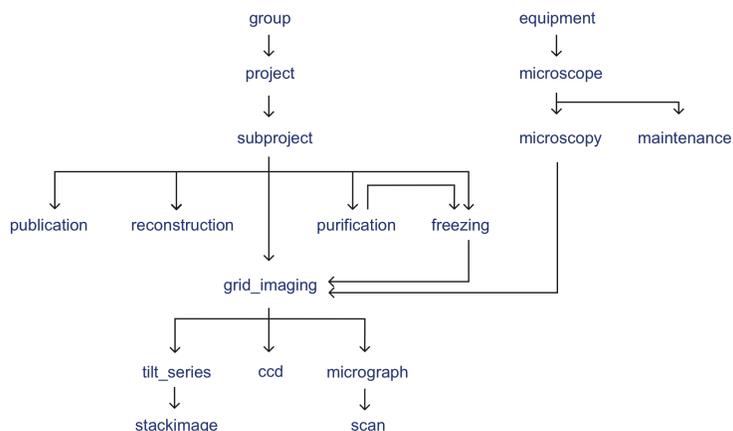
13886 Records Tools Query Rows 1 / 1389 »

rename	thumbnail	Record type	Record ID	Record Creator	Creation time
JEOL 2010F Demo		microscope	13885	Rees, lan	2011/03/15 21:56:30
Demo Microscope		folder	13884	Rees, lan	2011/03/15 21:56:30
Subproject PSSP7		subproject	13883	Rees, lan	2011/03/15 21:56:30
Subproject Epsilon 15		subproject	13882	Rees, lan	2011/03/15 21:56:30
Subproject Rice Dwarf Virus		subproject	13881	Rees, lan	2011/03/15 21:56:30
Boxes		box	13880	Rees, lan	2011/03/15 21:56:30
Boxes		box	13879	Rees, lan	2011/03/15 21:56:30
Boxes		box	13878	Rees, lan	2011/03/15 21:56:30
Boxes		box	13877	Rees, lan	2011/03/15 21:56:30
Boxes		box	13876	Rees, lan	2011/03/15 21:56:30

Projects

Welcome

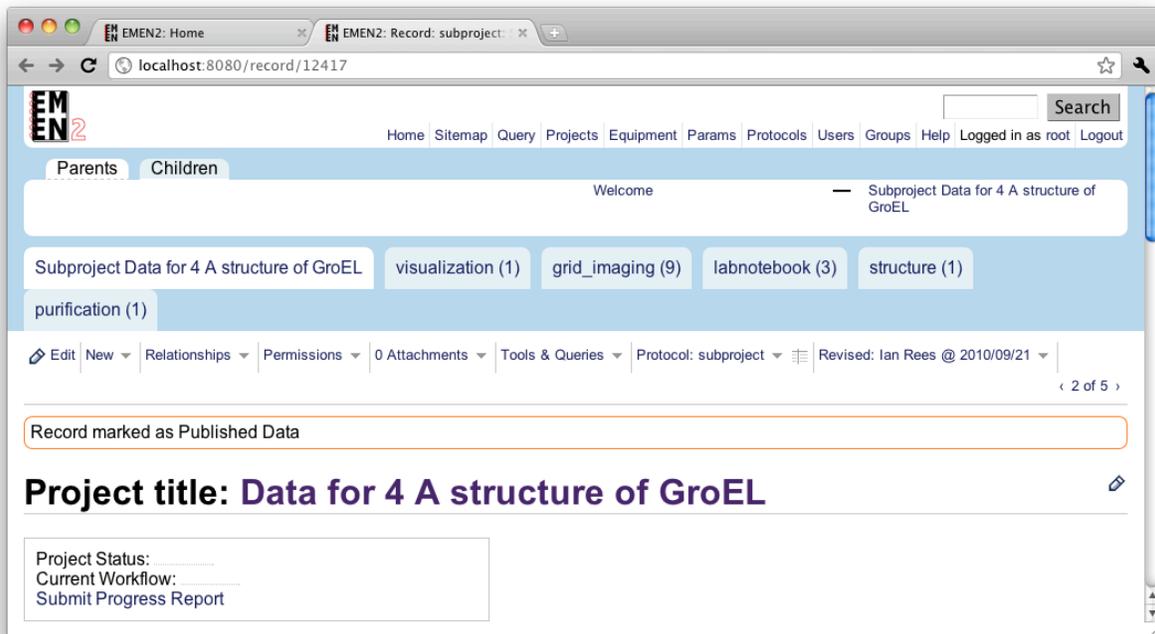
Navigating a project



While records can be arbitrarily linked together with parent/child relationships, EMEN2 will suggest organizational schemes that have worked well at the NCM. Typically, we have a top-level “project” for each collaborator, with several “subprojects” for different parts of the study and containing purifications, freezing and imaging sessions, and reconstruction results.

Records can have multiple parents. For instance, a grid imaging session might be connected to both a subproject and a microscope. This can provide additional context without duplicating data, and is also very useful for queries.

From the home page, click “Data for 4Å structure of GroEL.”



In this public GroEL data set, there 9 imaging sessions, one purification, one structure, etc., child records represented by the tabs. The “Parents” box above that shows the path back to the root node. Switching to the “Children” tab provides another way to navigate by quickly drilling down child relationships.

There are several controls for viewing and editing the record’s details. You can edit parameter values with either the “Edit” button, or by clicking an individual pencil icon. “New” creates a child record, with options for inheriting permissions and parameter values.

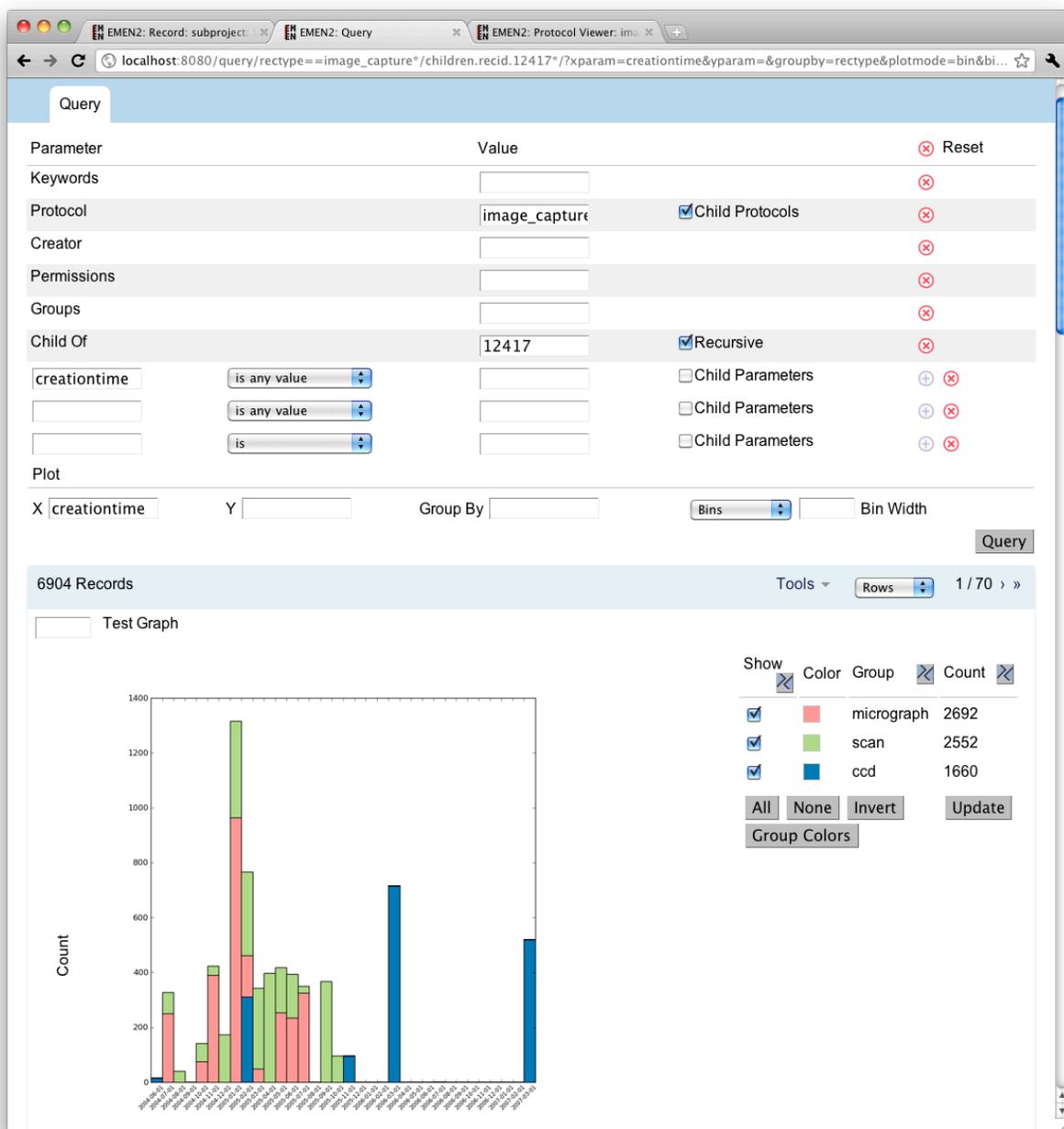
“Relationships” allows you to add and remove parents and children from a given record. To add a parent, open the relationship editor, select “Add Parent” as the current tool, and click the record you would like to add a parent too. A record chooser will appear, allowing you to navigate until you find the new parent. Likewise, use the “Delete” tool to remove a relationship; you will be prompted to confirm the change.

“Permissions” is for viewing and editing access privileges. There are four increasing levels of access: “Read” for basic read-only access, “Comment” may annotate a record using the comments form at the bottom of the page, “Write” permits changes to parameter values, and “Admin,” which is required to change permissions. All changes to parameter values are logged and can be viewed in the “History” tab at the bottom of the page. If you add a Group, members of that group will have a minimum of read-only access. There are a couple built-in and special groups: “authenticated” will give any logged in user read permission, “anonymous” will make the record publicly viewable even without an account, and “published” will mark a record as public data.

The “Tools” menu contains access to commonly used items, as well as some context-sensitive actions, such as the [web.boxer](#) particle picker for micrographs.

Set Magnification	Specimen temp	Purpose	parentvalue title_freezing	# ccd	# micrograph	# scan	Record Creator	Creation time
60.0	4.19	collecting data	HR GroEL 2005	0	150	149	Rees, lan	2011/03/15 21:53:56
60.0	4.18	collecting data	HR GroEL 2005	0	100	99	Rees, lan	2011/03/15 21:53:43
60.0	4.176	collecting data	HR GroEL 2005	0	80	80	Rees, lan	2011/03/15 21:53:43
60.0	4.177	collecting data	HR GroEL 2005	0	69	69	Rees, lan	2011/03/15 21:53:43
60.0	4.18	collecting data	HR GroEL 2005	0	140	141	Rees, lan	2011/03/15 21:53:43
60.0	4.17	collecting data	HR GroEL 2005	0	100	100	Rees, lan	2011/03/15 21:53:30
60.0	4.18	collecting data	HR GroEL 2005	0	198	197	Rees, lan	2011/03/15 21:53:30
60.0	4.18	collecting data	HR GroEL 2005	0	90	90	Rees, lan	2011/03/15 21:53:30
60.0	4.18	collecting data	HR GroEL 2005	0	50	50	Rees, lan	2011/03/15 21:53:30

Child records and query results are usually presented in a table format. The records are sortable (and editable, if you have permissions) using the icons in the header row. There are a number of macros that can be used in these views; here, the total number of three types of images for each session are displayed. Additionally, records meant to store images will have thumbnail previews in the table view.



Any table view can be quickly modified using the “query” drop down. For instance, to find all images in this project, enter `image_capture` for Protocol and check child protocols, which will include additional related Protocols such as `ccd`, `micrograph`, `stack`, etc. Also select the recursive box for the “child of” constraint. When you run the query, there should be ~6900 CCD frames and micrographs returned. To see the breakdown of each image type over time, enter `creationtime` as the X parameter, select “bins” as graph type. “Month” (default), “year”, and “day” are allowed “bin width” values for time-based parameters.

NCMI Database: Record: ccd

ncmidb.bcm.edu/record/56969

NCMI National Center for Macromolecular Crystallography

Home Search Query Projects Equipment Params Protocols Users Groups Help Logged in as ianrees Logout

Parents Children

JEOL 3000SFF — Microscopy: JEOL 3000SFF by Chen, Donghua @ : collecting data — Imaging: on JEOL 3000SFF by Admin @ 2007/03/27: 205 images — CCD ccd_3160

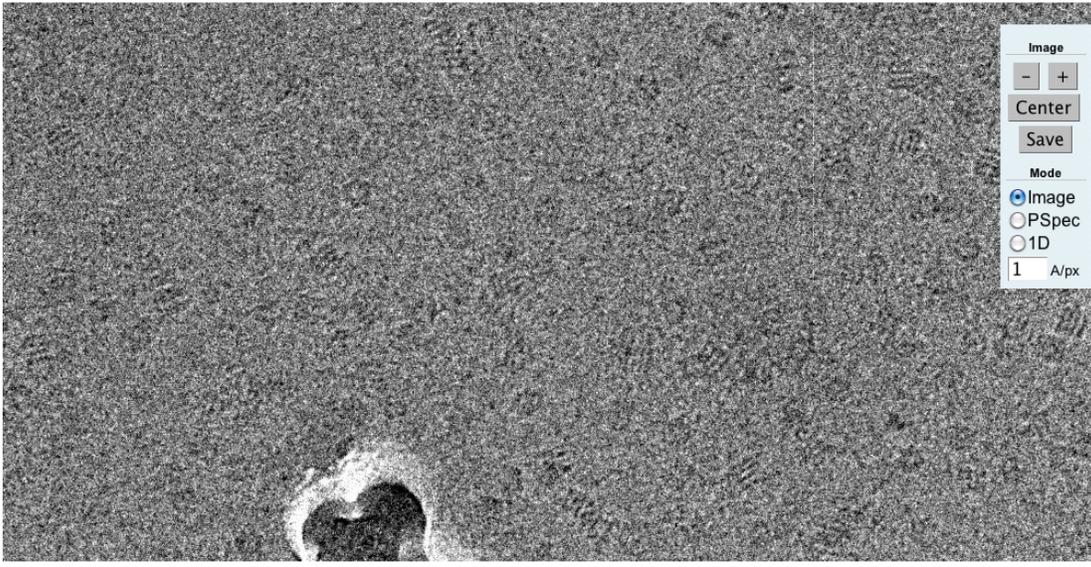
Project GroEL (GroEL) — Subproject High resolution GroEL

Aliquot: GroEL 3/8/2005, 2005/03/08 — Vitrobot: GroEL G8 G9 by Chen, Donghua @ 2005/03/11

CCD ccd_3160

Edit New Relationships Permissions 1 Attachments Tools & Queries Protocol: ccd Revised: Admin @ 2010/09/21 < 205 of 205

Record marked as Published Data



Image

- +

Center

Save

Mode

Image

PSpec

1D

1 A/px

CCD Micrograph

Frame ID : ccd_3160

Exposure #: 1

Set Magnification: 80.0

Set Defocus: 2.0

Dose rate: 18.0

Total dose: 36.0

Exposure time: 2.0

A/pixel: 1.32978723404

Screen current: 7.6

Beam width: _____

Energy filter: 0

Records with images attached usually have a “Google Maps” style micrograph viewer. You can pan by dragging the mouse, and zoom with the - / + buttons on the right-hand side. You can view the image itself, the FFT, or a 1D rotationally averaged power spectrum. “Save” will prompt your browser to save the micrograph.

Note: These image previews are JPEG quality and may not be suitable for publications.

You may notice that this record's parent tree branches. The imaging session is connected to a subproject, a Vitrobot freezing session, and a microscope session. This makes it very simple to find all images captured on a given microscope or from a particular freezing session.

EMDash

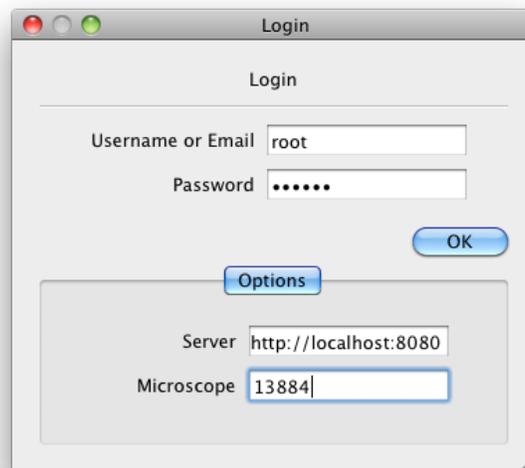
EMDash is a GUI program for uploading data from instruments. It was originally designed to help manage microscope sessions and upload micrographs in the background, but is being expanded to manage other equipment such as Vitrobot, Plasma Cleaner, etc.

First, make sure the EMEN2 web server is running per instructions above. On the workshop PCs, you will need to specify the path to EMDash:

```
$ python ~/emen2-2.0rc4/scripts/emdash.py
```

On laptops, emdash.py should be in your path.

```
$ emdash.py
```



Note: Make sure `http://localhost:8080` is the server.

Note: Use 13884 for “Microscope,” the instrument’s Record ID. This would be in the configuration file if installed on an actual instrument.

After login, you will be presented with a form requesting details about this microscope session. You may enter whatever you like, but choose “CCD” for “Detector.”

After the main window pops up, click the “Grid Imaging” button, then “New Grid Imaging.” There will be a wizard to guide you through the process of creating a new grid. Select the 4Å GroEL project, select “New grid_preparation,” and then select one of the existing Vitrobot sessions. You may fill in the remaining two forms with whatever values you like. Click “Commit” on the last page to complete the wizard.

Select a Project or Subproject

Use a project or subproject from this list

Record Name
Subproject 6 Ans GroEL Film Data
Subproject Data for 4 A structure of GroEL
Subproject Epsilon 15
Subproject PSSP7
Subproject Rice Dwarf Virus

New Grid > Select a Freezing Session

Each frozen grid must be linked against a Freezing Session record. You should only skip this screen if it is not applicable because the grid was not frozen (e.g. magnification calibration.) If you have not created the Freezing Session for this grid, you should exit this wizard and run the 'New Freezing Session' wizard first, to ensure the workflow is complete.

Use a freezing session from this list

Record Name
▼ Purification: GroEL by Jiuli Song @ 2004/05/28: 5 aliquots
▼ Aliquot: 1/11/2005, 2005/01/11
Vitrobot: HR GroEL 2005 by donghua @ 2005/01/14
▼ Aliquot: 11/4/04, 2004/06/02
Manual Freezing (A2, A3, A4)
Vitrobot: A1 by donghua @ 2004/11/04
▼ Aliquot: GroEL 3/8/2005, 2005/03/08
Vitrobot: GroEL G1 G2 by donghua @ 2005/03/11
Vitrobot: GroEL g1 g2 g3 g4 by donghua @ 2005/03/16
Vitrobot: GroEL G3 by donghua @ 2005/03/11
Vitrobot: GroEL G4 G5 by donghua @ 2005/03/11
Vitrobot: GroEL G6 G7 by donghua @ 2005/03/11
Vitrobot: GroEL G8 G9 by donghua @ 2005/03/11
▼ Aliquot: GroEL 6/2/04, 2004/06/02
Vitrobot: GroEL 61, 62, 63, 64 by donghua @ 2004/06/05
▼ Aliquot: GroEL 6/4/04, 2004/06/02
Vitrobot: GroEL 65 66 67 68 69 70 71 72 by donghua @ 2004/06/05

Skip this step: I am ABSOLUTELY SURE that it is not required!

New Grid > New Grid Prep

Please enter the Grid Preparation details for this grid. You may leave the vitrobot-related parameters empty if they are specified in the parent Freezing Session. Please make sure to describe any pre- or post-freezing treatments you applied to the grid.

Grid Type

Grid label:

ID of grid batch:

Grid type:

Grid meshsize:

Grid Preparation

Get a grid

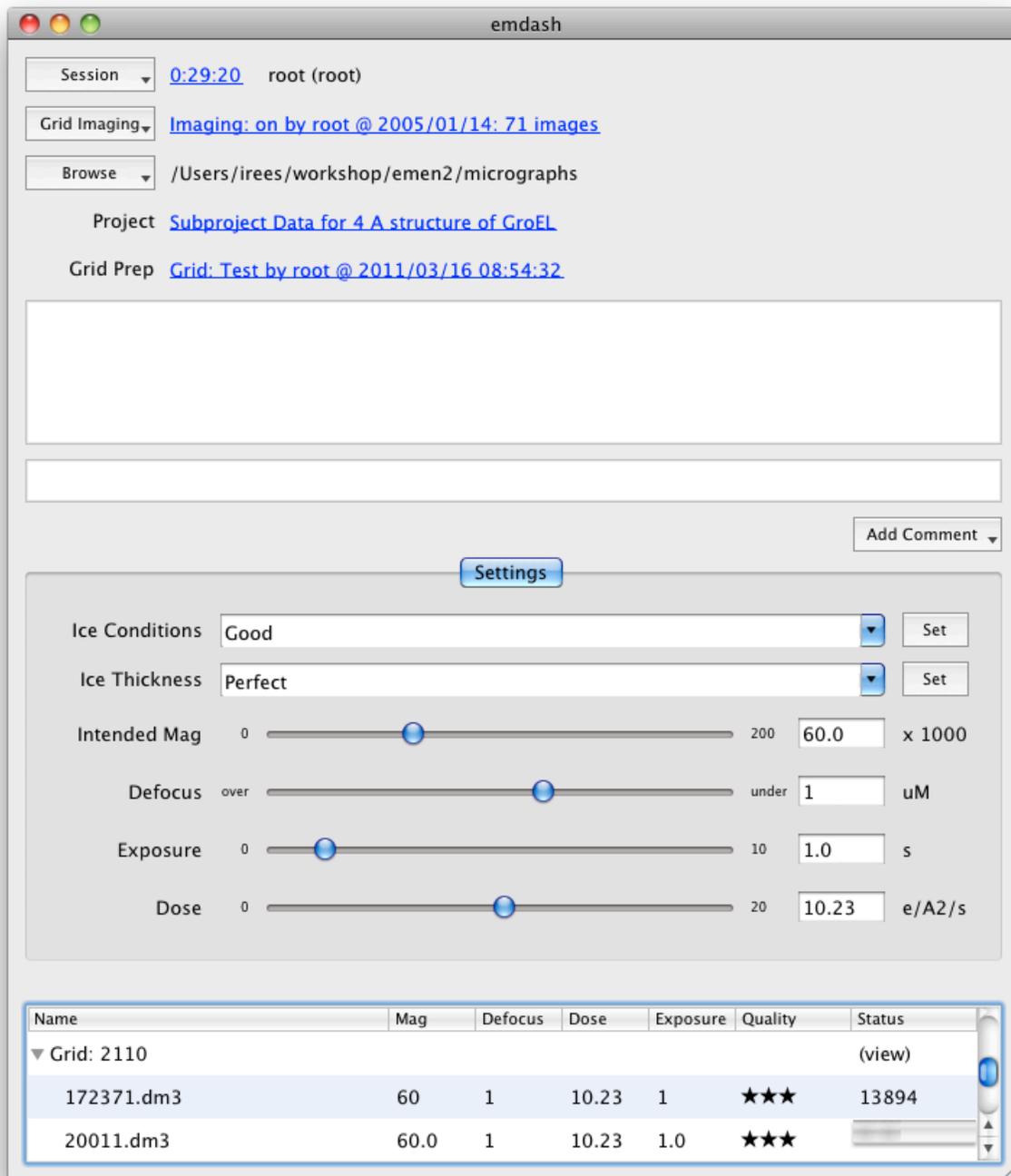
New Grid > Grid Imaging Details

Please describe the purpose of this this Imaging Session; it will help you and your collaborators find your images in the future. The 'default' parameters below will set the initial values for images collected during this session. During the session, you can adjust the sliders/input fields to change the parameters applied to uploaded images. After an image is uploaded (the status column will contain a Record ID) you can click on the values in each column to adjust a value for a single image.

Grid Imaging Session

Purpose:

Cryoholder:



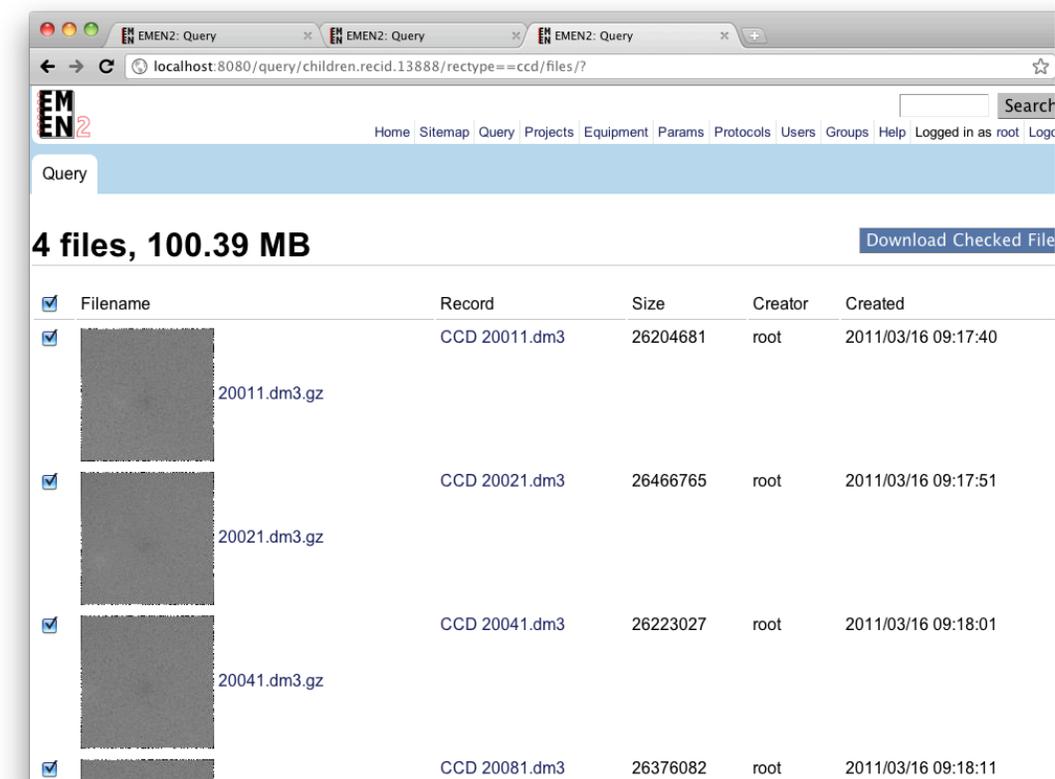
Once your grid imaging session is prepared, you may want to set the various sliders and controls if you are not using data acquisition software that includes this metadata in the micrograph headers (e.g. JADAS.)

EMDash watches a directory for new micrographs, waits a small amount of time, and then uploads the files in the background. Click the “Browse” button and select the ~/workshop/emen2/micrographs directory in the file chooser. The files in the directory

will be found and added to the upload queue, and the directory will be watched for new files. Once a file is uploaded, the "Status" column will show the Record ID for that image in the database, which you can click to view it in your web browser. The other fields may be edited, particularly the "Quality" setting where you may assign a rating of Trash to 5 Stars. We urge everyone at the NCMi to upload all images taken, regardless of quality, and use the rating system to filter out unwanted data.

Download

You can perform batch downloads using either the web interface or the `emen2client.py` command line program.



In the web interface, you can download all the file attachments in a given table through the “Tools” menu and selecting “Download all files.” This will take you to a page where you can select the files you want. The system will provide an estimate of the file size, and create a `.tar.gz` archive of all the files when you click “Download checked files.”

Alternatively, you may use the command line client by specifying the Record ID.

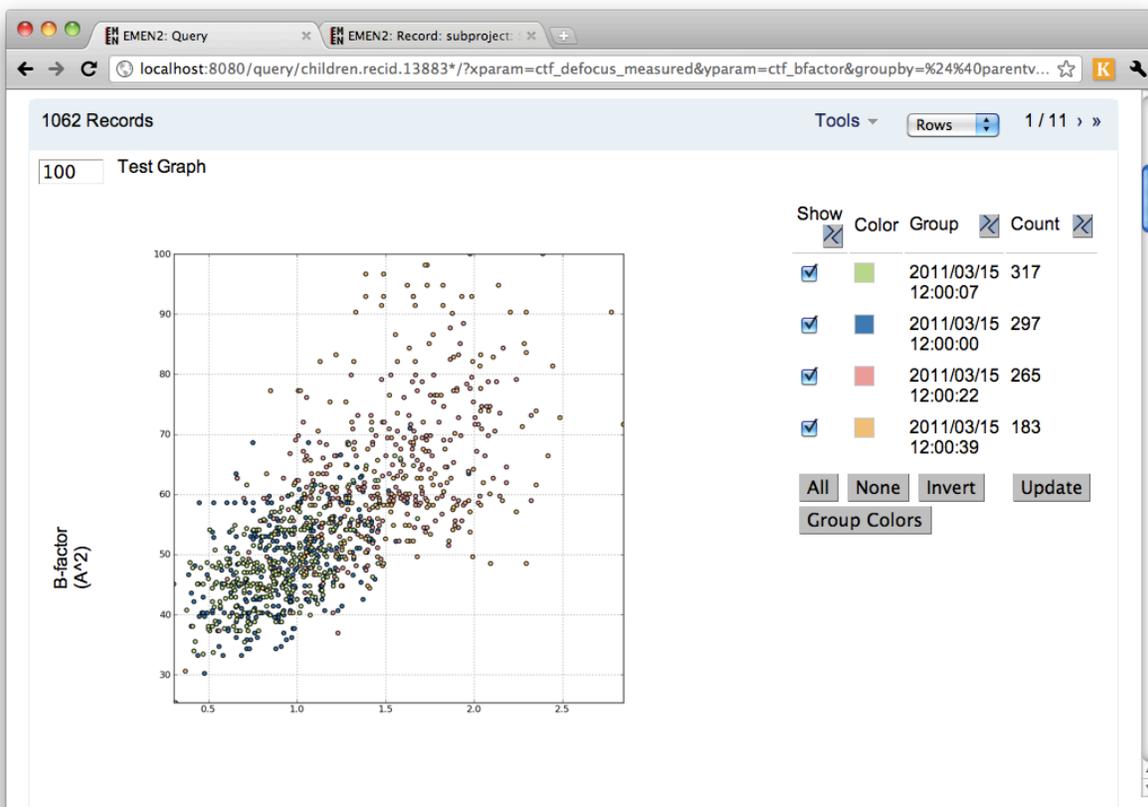
```
$ python ~/emen2-2.0rc4/scripts/emen2client.py download 13888
Username: root
Password:
emen2client version 1.10 is up to date

1 of 1: 13888
  Checking for items to download
  Found 4 items in 5 records
  1 of 4: bdo:2011031600001
    Downloading 20011.dm3.gz -> 20011.dm3...
```

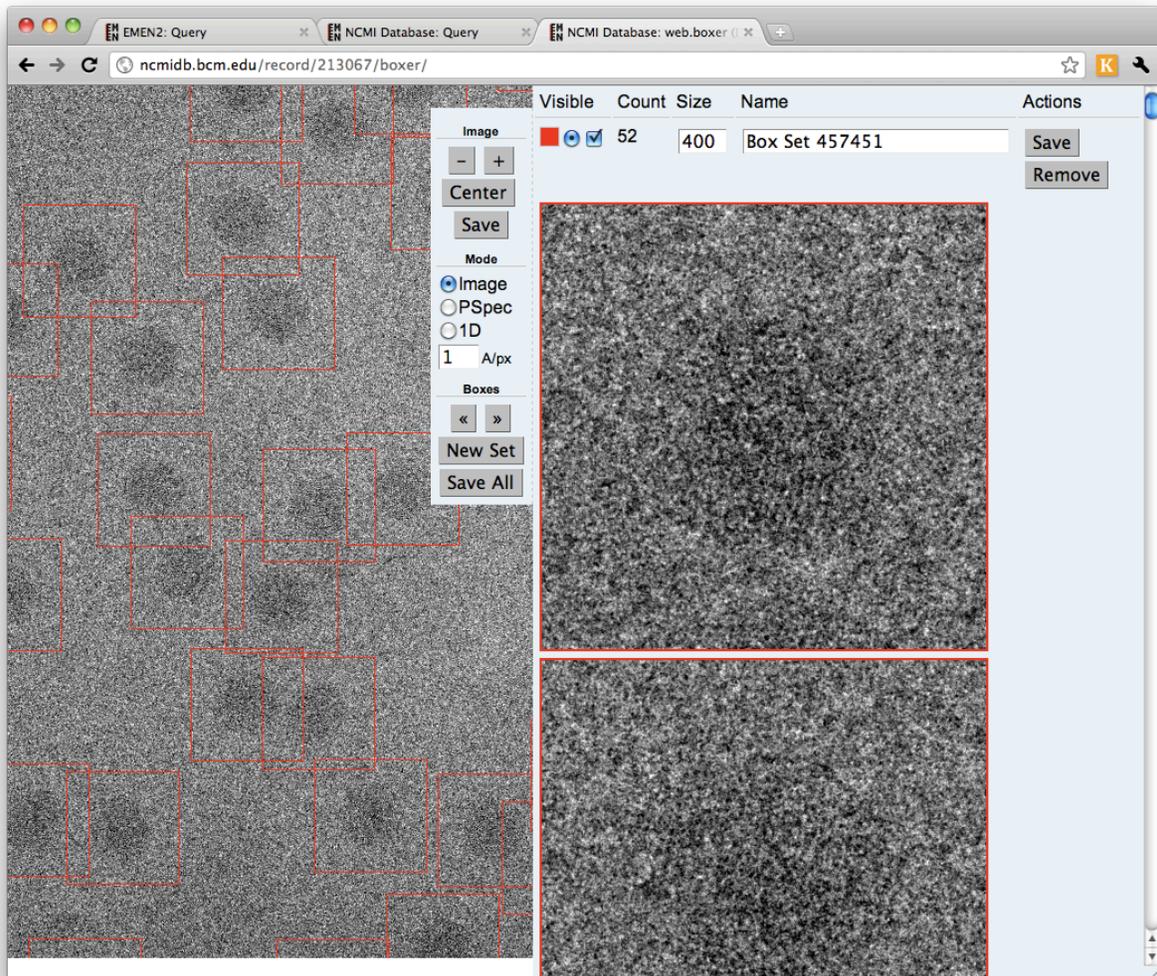
EMAN2 Sync

`emen2client.py` also handles EMAN1 and EMAN2 integration. You can read the documentation for this mode at the EMEN2 wiki: http://blake.grid.bcm.edu/emanwiki/EMEN2/EMAN2_Integration

CTF parameters and particle coordinates are currently supported, and are attached to the original micrograph records in the EMEN2 database.



This plot shows the relationship between defocus and B-factor for the PSSP7 data. To generate this plot, go to <http://localhost:8080/record/13883>, then select "Tools," and "Child images, plot B-factor vs. defocus." You can also group the images by the date of the imaging session by entering `parentvalue(creationtime)` in the "Group By" field.



Particle coordinates that have been uploaded can also be viewed and edited with the [web.boxer](#) particle picker. You can start [web.boxer](#) on one of the images you uploaded with EMDash through the record “Tools” menu.