

# EMAN2 Reconstruction Tutorial

## Using the Workflow interface

### Getting Started

- Install demo data
  - To get started copy the demo data on the NCMI DVD onto your machine. In all there are 6 demo folders (airs, eman, gorgon, modeller, sample\_images, and sparx) that you will need to make a copy of on your local harddrive. To do this e2workflow demo you will need to at least have the eman\_demo folder on your machine. You can put the data anywhere that's convenient, on Windows, the desktop is a good spot.
- Make sure you have the latest available version of EMAN2 installed
  - CD's are circulating around the room with the latest EMAN2 binary for each platform. Please make sure you install one of these new versions, as we have made a number of critical fixes on each platform.
- ➔ EMAN2 documentation is largely provided via a Wiki at: <http://blake.bcm.edu> If you wish to edit the Wiki, create an account for yourself, then send email to [sludtke@bcm.edu](mailto:sludtke@bcm.edu) and we will adjust permissions so you can edit. Previously this was an 'open' wiki, but we had serious spam problems, and now have to individually approve new users. You don't need an account to browse the current contents, of course. During the workshop, with such a large number of users, browsing the wiki may be sluggish.
- ➔ The interface in EMAN2 is fairly standardized and quite similar to EMAN1. While there are alternatives for machines like Macs with the native 1-button mouse, EMAN2 will work much better if you have a 3-button scroll mouse.
  - In most display windows (plots, 2-D images and 3-D volume display), the middle mouse button will open a 'control panel' for the widget with many options to control the display
  - The right mouse button is used for panning in 2-D or 3-D image windows, and can be used to zoom (by dragging), and to reset the zoom (clicking) in plot windows.
  - The left mouse button has various purposes in various contexts.
  - The scroll-wheel will generally act as a zoom. Use the control-panel for more precise control
  - If you have a one button mouse, one of the modifier keys (depending on platform) combined with a mouse click will serve the same role as a middle-click. You may need to try them (alt, command, ctrl, shift) to discover which works on your machine.
  - In the control panels, and other places in the EMAN2 interface you may encounter 'ValSliders'. These are widgets where a slider is attached to a text-box with a number in it. Dragging the slider controls the number, and entering a number will change the slider. In addition, the text-box can be used to control the range of the slider and get more precise control. By typing '<value' or '>value' in the text box you can change the limits of the slider. Note that it is also possible to enter values which are outside the current slider range.

### Introduction

EMAN2 can be used at many different levels ranging from writing code in C++ or Python to a high-level integrated 3-D desktop interface. In this tutorial, we will be using the Workflow interface, which will guide you through the single particle reconstruction process, and keep track of appropriate intermediate data generated during the process in a self-consistent way.

The workflow can be used as part of the integrated desktop (e2desktop.py) or as a standalone program (e2workflow.py). Unfortunately the desktop isn't quite perfected yet, and seems to be very machine dependent. On some machines it performs extremely well, but on other machines it behaves strangely.

The advantage of the desktop is that it will save you a fair bit of desktop space, which is particularly nice on laptops. On the downside, it is fairly 3-D graphics intensive. So, if A) you experience strange behavior or bugs with the desktop, or B) don't have very speedy 3-D graphics, you may want to use the workflow without the desktop.

- Open `e2desktop.py` or `e2workflow.py`
  - On Windows
    - double click the appropriate icon on your desktop
    - in the 'EMAN2 tasks' window, scroll down to the 'directory' task and click on it
    - Browse to the `eman2_demo/raw_data` directory
  - On Mac and Linux
    - open a terminal window
    - `cd` to the `eman2_demo/raw_data` directory
    - type the name of the program
- ➔ If you did this step successfully you will have two windows : 'Running tasks' and the other will be titled 'EMAN2 tasks'. If you hover your mouse over 'Directory', you should see that you are in the `eman2_demo/raw_data` directory.

### Setup Project and Import Data

The workflow interface is an expandable tree. Each level of the tree has a form with associated information or parameters, even the levels which just appear to be containers for the levels below them. The first item in the workflow is **SPR** (Single Particle Reconstruction). This item can be expanded into a list of the individual stages of the single particle reconstruction process, but it can also be selected itself to provide information about the overall reconstruction you are performing in the local directory.

Note that the database used to store the information about your project resides in the local directory, so each project must reside in its own directory. The workflow will make a number of named subdirectories (folders) to store specific types of information.

- Select **SPR**
  - This will open a window where you enter basic information about your project. The demo data for the workshop has the following parameters : 2.0 A/pix, 200 kV, Cs=1.0.
  - For a laptop you may want to use only 1 CPU, even if you have a dual core machine, to reduce heating, but it will make things run slower.
  - The workflow will attempt to detect how much RAM you have but this may or may not work on your platform. This parameter isn't really used at the moment, but you may as well enter the correct value, as it will be used in future.
  - Click 'OK' once you have entered the parameters. If you select a different workflow item without saying 'OK', your values will not be stored.
- Expand **SPR**, causing 6 subtasks to appear: **Micrograph/CCD**, **Particles**, **CTF**, **Refine2D**, **Initial Model**, and **3D Refinement**.
- ➔ In some steps you have a choice between using data already imported into the project, or importing the data directly. For example, if you wish to use particles already boxed out using another (non-EMAN) application, you can skip forward to the **Particles** step. However, the earlier in the process you import your data, the more flexibility you will have in recovering relevant metadata. For example, if you import your CCD frames or micrographs, when particles are selected, EMAN2 will keep track of where each particle came from in each micrograph, offering many possibilities for future analyses that could be performed. This flexibility would be lost if you skip forward in the process.

- Select the **Micrograph/CCD** task. You will see a form appear with an empty data list. This form is just a status display showing you all of the frames which have already been imported. This is NOT where we want to import our data. Close this form with 'Cancel'.
- Expand **Micrograph/CCD** and select **Import Micrograph/CCD**. This will open a similar form to the one you just closed. Click on the browse button. You should see the sample micrographs in the raw\_data directory. Select them all by first clicking on image 1160.mrc, holding shift and then clicking on 1792.mrc. This should highlight all images. Then hit OK. The browser dialog should close and the selected images should appear in the list on the first form.
- In the Import Micrographs form make sure that the 'Edge norm' and 'Thumbnails' check boxes are selected and hit OK.
  - The Edge norm option will adjust the images so the mean value around the edge of the frame is zero and the standard deviation of the pixel values is 1. This normalization helps regularize the data for later processing and minimizes problems with brightness and contrast.
  - Generating thumbnails will save time later when e2boxer is used for boxing.
  - The invert option should be used if necessary to make your particles white on a dark background. Depending on acquisition method and whether your data is in ice or stain, you may or may not need this. For the workshop demo data, you do not need to invert.
  - The X-ray pixel option will apply a filter to remove x-ray pixels from CCD frames. The demo data for the workshop is scanned film data, so this is not necessary.
- Check that the micrographs have been correctly imported into the project using the **Micrograph/CCD** task. This task will now display the images you just imported. If you imported more images later, they would be added to this list. Double-clicking on one of the images in the list will display the image on the screen.
  - You'll note that the filenames that appear in this list aren't simply "1160.mrc", but much longer and complicated looking 'bdb:raw\_data#1160.mrc'. What you have done when you imported your micrographs into the project was to copy the data into an internal embedded database. The files embodying this database are stored in the 'EMAN2DB' directory. The specifications for bdb database access are fairly straightforward:
    - 'bdb:dbname' which refers to the database (think of it as an image file) in the local directory named 'dbname'. 'dbname' can contain individual images, numbered sets of images, named images and other named metadata
    - 'bdb:/path/to#dbname' allows you to specify a database residing in a different directory. Note that '#' is used for the final separation between the path and the database name
    - files in the EMAN2DB directory should NEVER be renamed, individually deleted, moved, etc. It can cause the system to become 'confused' and produce a range of seemingly unrelated errors.
    - If you really insist on deleting files in the BDB directory, make sure you aren't running any EMAN2 programs, remove the files, then also remove /tmp/eman2db-username

## Boxing your data

- Expand the **Particles** entry in the EMAN2 tasks window. You should see : **Interactive boxing, Auto boxing, Generate output, and Particle import**.
- Go to the **Interactive boxing – e2boxer** task and hit OK.
- You should see the e2boxer interface form appear which displays a table. The left-most column lists the image names of the micrographs in the project. The other columns, most probably blank at this point, are titled Boxes in DB, Particles on disk, and Particle dims.
  - Note that you can double click on any of the image names in the left column and the associated image will appear in an interface for viewing (Try it). As a general rule of thumb, anything that has an icon is viewable by double-clicking.

- If you have a decent workstation with a reasonable amount of RAM, you could open all of the micrographs at once, but on a laptop, this probably isn't wise. Choose 2 or 3 images (or just 1 if you have problems with more, implying your machine is too low-end), enter a boxsize of 128, then hit OK.
- Wait a moment while e2boxer loads.

### **e2boxer – the GUI**

- You should see at least 3 windows appear. If you are using e2desktop.py, the windows will be positioned for you in specific locations. If you are running e2workflow.py you will need to arrange the windows yourself. On a laptop with a low-resolution display, this can be a challenge. The windows are:
  - The main controller - looks like a regular window; it has an assortment of buttons and text entry boxes
  - the main image display window - shows the full 2D micrograph currently selected
  - the particle display window - (with nothing in it) that will eventually show the boxed particles (this should be obvious from its icon).
  - micrograph thumbnails - This will only appear if you selected 2 or more micrographs in the previous step. Clicking on one of these will select the current image to be boxed.

### **Getting started with autoboxing**

In the main controller you should see a check box labeled 'Dynapix', we recommend that you turn this on. If this option is selected, e2boxer you will autobox your images in real time as you interactively specify reference boxes (which are black).

Turn Dynapix on then go to the main image display, find a particle and click on it. Try to center the particle in the box fairly well. You can move it after clicking by dragging it in either window. In the particle window you should see your newly selected box appear and, if e2boxer was able to do so successfully, more boxed particles should be displayed that were autoboxed automatically. Continue adding black reference boxes in the main image display until a reasonable proportion of the particles in the image have been selected.

- ➔ If you put a black box on something that caused a lot of bad particles to be selected, hold shift down and left click on it to remove it – the automatic boxes will update automatically to reflect this change (providing Dynapix is on)
- ➔ Bad particles or boxes that contain just noise can be damaging to your reconstruction, as they permit noise/model bias to become stronger. It is much better to miss a few good particles if it permits you to exclude obvious 'bad' particles.
- ➔ One caveat to the above, if the good particles that get excluded are all in one orientation, that would be a bad thing, of course
- ➔ If you just want to start again : Hit the clear button in the main controller – all reference (black) and auto (green) boxes will be removed and you can start again on this image. This functionality is particularly useful when you are just getting to know how e2boxer works.
- ➔ Don't hit the 'classify' button. This is a planned future feature which will make it much easier to eliminate bad particles, but it isn't complete yet.

### **Building up a boxed particle set with e2boxer**

- Now that you are familiar with how the Dynapix option behaves, the idea is to select a set of references that results in the autoboxing of as many good particles as possible. Once you are satisfied the autoboxing results are as good as you are going to make them, you go through the particles in either window and manually delete bad particles or manually add missed good particles (without changing the autoboxing).

- ➔ To exclude large regions of the micrograph: Choose the erase option in the main controller and drag over the bad areas of your micrograph (e.g. ice contamination and particle aggregates). Unerasing will bring back particles that you accidentally removed, as long as you have not autoboxed after you erased. You change the radius of the erased region in the main controller.
  - ➔ To delete bad particles individually: Hold down shift and left click on the bad particles, either in the main image display OR in the particle image display. In the main image display you can also hold down shift and left click and drag the cursor around to delete many particles at one time.
  - ➔ To add manual particles: Choose the manual button in the main controller and then click on the center of the particle you want to add manually. Manual boxes appear white and have the advantage of not changing the autoboxing results.
- Once you have finished boxing the first image you should freeze it – this prevents e2boxer from autoboxing the image automatically in the future if it detects that the underlying autoboxing parameters have changed.

### **Boxing multiple images interactively**

- ➔ Note that this section applies only if you have more than one image loaded into the e2boxer interface.
- Choose the next image in the image thumbnails image. This should load the new image into the main display window, and if Dynapix is on you should also observe that autoboxing has occurred.
- At this point you have several options:
  - you can add more reference boxes and see if the autoboxing results can be improved,
  - you can be satisfied with results and go straight to the cleaning stage
  - you can hit Clear in the main controller and start the dynamic autoboxing procedure from scratch for this (and subsequent) images
- ➔ If Dynapix is on and you have added a reference box and you are noticing that the autoboxing is running slowly it can be because you have specified reference boxes in multiple images. Try pressing clear and starting afresh in the current image. This should only be a problem on lower end machines.
- ➔ There are also more advanced options which give you finer control over the autoboxing process, but these should not be necessary for the demo. Just be aware that there is more there.

### **Finishing e2boxer**

- Once you are done boxing the images you may be tempted to press the ‘generate output’ button, but it is more efficient to do this in the workflow interface instead once you have finished boxing all of the micrographs. When you have finished boxing each set of micrographs, simply click Done in the main controller and return to the e2workflow interface. We will deal with ‘generate output’ later.
- If you have run e2boxer correctly you should be able to click on the Interactive boxing task and observe that the Boxes in DB column now has entries in it. These entries should correspond to the total number of boxes currently stored in the database for the given image.
- At this point you have two options. To get familiar with the workflow you may try both approaches.
  1. You can run automated boxing on the remaining micrographs. Go to the Autoboxing task (e2boxer) and choose images with no boxes stored in the database (column 2 should be blank) and then hit OK. This will tell the workflow to spawn processes on your operating system that complete the autoboxing procedure. You can monitor the progress of these processes in the Running tasks window, which should display the percent of the task which is complete, the name of the program, and the time at which it was initiated. If the task doesn’t appear in the Running tasks window immediately just wait a few moments.

- ➔ On lower end machines try autoboxing only one image at a time. On Windows we are still experiencing some problems. If the process in the Running tasks list stops running (the percentage entry freezes at some value less than 100%) then you should definitely try just doing one image at a time.
  - ➔ Even if you plan to manually clean up the autoboxing results, performing autoboxing in this way can save you the time you would normally spend waiting in the e2boxer interface for autoboxing to occur.
2. Repeat the process using the e2boxer GUI for the other micrographs
- ➔ For the purposes of the workshop before proceeding to the next step, you should have boxed out all of the micrographs you imported into the project.
  - Generate boxed particle output by going to the **Generate output (e2boxer)** task in the **Particles** section of the workflow. You should see a form appear that lists the micrographs you have in the project as well as the number of boxes you have stored for each image in the database.
    - Select all of the images
    - enter a boxsize of 144
    - make sure the normalize.edgemean option is chosen
    - make sure the output image format is “bdb”
    - Hit Ok. Once again this will spawn processes on your operating system that you can monitor in the Running tasks window.
  - ➔ On lower end machines , especially Windows machines, you may experience problems and be forced to run the output generating tasks one at time.
  - Check that your particle images exist and are stored in the database by going to the **Particles** task. This will display a table that tells you precisely which particles exist, how many there are and what dimensions they have.
  - If this table is complete and the dimensions of the particles are correct (144x144) you are ready to go the next stage in the workflow.

## CTF

EMAN2 uses a substantially different CTF model than EMAN1. EMAN2 uses a data-based background curve. For this process to work properly, and to have phase-flipping work optimally, it is important that the box size used for the particles is somewhat larger than the actual particles. This is the reason for expanding the 128x128 boxes used for particle picking to the larger 144x144 box size. The edges of the images are used in determining the background noise spectrum. After determining the background automatic fitting of defocus and B-factor takes place. The uncertainty in B-factor tends to be fairly substantial, largely because it is only a very approximate representation of the true envelope function of the data coming out of modern FEG microscopes.

- ➔ EMAN1 used a complicated 10-parameter CTF model, and required a lot of painstaking manual fitting, with several complicated difficult to describe tasks related to determining a structure factor. Almost all of this work has been completely eliminated in EMAN2, but the models are not compatible. While EMAN2 can understand EMAN1 CTF parameters as well, if you want to work

with EMAN2, you are much better off going back to non-phase-flipped particles and let EMAN2 deal with the CTF entirely internally.

### Determining CTF parameters and generating CTF/Wiener filtered particles

- Expand the **CTF** entry in the workflow (the **CTF** under **SPR**, not the standalone **CTF** entry near the bottom of the workflow). You should see three entries appear underneath the CTF entry called: **Automated fitting** , **Interactive tuning** , and **Generate output**
  - Run automated CTF determination on your particle data by choosing the **Automating fitting (e2ctf)** task.
    - A form should appear with a table that lists the particles that you generated in the previous stage of the workflow. Other columns (which are probably blank at this point) are titled Defocus, B factor, SNR and Sampling .
    - You must decide whether to check the 'Auto high pass' checkbox. If selected, this will modify the SNR curve to eliminate the first sharp peak that appears in virtually all single particle data. This sharp peak can cause some issues leading to incorrect 2-D alignment, and hence classification. However, it is also responsible for much of the contrast which makes particles visible by eye. To see the visual difference, you can run it first with then without this option checked if you like. For purposes of the rest of this tutorial, please check this box.
    - You should also decide on an oversampling factor. For purposes of the demo, 2 is a good value. If you have very far from focus images or small particles, larger values may produce more accurate CTF fitting).
    - To proceed, select all of the images, make sure the microscope voltage, spherical aberration and the angstrom per pixel parameters are correct, and hit OK. This will cause the workflow to spawn a set of jobs to complete automatic CTF determination. You can monitor the progress of the task(s) in the Running tasks window.
  - When the fitting tasks are complete, click on the **CTF** task in the workflow. This should display a form that has been updated with automatically determined defocus, bfactor, ... parameters.
- ➔ Proceed to the next step when you certain that automatic CTF parameters have been generated for all particle data in the project. If necessary, you can rerun automatic fitting on one or more sets.
- Select the **Interactive tuning (e2ctf)** task. This will display a (by now) familiar looking table that lists particle file names and CTF data. Select all or a subset of the images and hit OK. This will launch the e2ctf interactive interface from which you can fine tune the automated fitting results. If you do alter the CTF parameters be sure to hit the Save parms button, which will store the changed parameters in the EMAN2 database. The interactive interface also gives you a variety of tools for assessing the quality of the individual images.
- ➔ Note that unlike in EMAN1, accurate per-micrograph B-factors are *not* critical for a good reconstruction or proper CTF correction (though it never hurts)
- Select the **Generate output (e2ctf)** task and in the subsequently appearing form select all of the images, make sure you select both Phase flip and Wiener, and use the same oversampling factor you used in the previous step, and hit OK. This will spawn output writing processes that you can monitor in the Running tasks window.
  - Finally check that your phase flipped and Wiener filtered data exist and are correctly stored in the EMAN2 database. Do this by choosing the CTF task and carefully inspecting the displayed table. The number of regular, phase flipped and Wiener filtered particles should match for all image entries in the table as should all particle image dimensions. You may also use the browser to look in 'particles' and examine the raw particles, the phase-flipped particles and the Wiener filtered particles.

## Generating reference free class averages (2D refinement)

- Expand the **Refine2D** entry in the EMAN2 tasks window in e2workflow. There is a single entry : **Generate classes – e2refined2d**
- Select the **Generate classes (e2refine2d)** task and hit OK. This will pop up a form asking you to choose from your regular, phase flipped and Wiener filtered particles (you can also specify files, but ignore this option for the purposes of the workshop). Choose the Wiener filtered data and hit ok. (You may experiment with other choices at a later date). This will pop up a form asking you to specify the parameters required to run e2refine2d.py. The default parameters will produce reasonable results. To make it run faster, you may wish to select only a subset of your data, but you may use as much as you like. Select the data and hit OK. This will spawn a process that you can monitor in the Running tasks window.
  - Tool tips display useful information regarding the specific parameters. Just hold your mouse still over one of the parameters for a few seconds to see the tooltips. If you want to obtain more information you can also go to the command line and type e2refine2d.py -h.
  - If you want more information on the particular parameters you can pass to the aligners and/or comparators go to the command line and type 'e2help.py aligners -v' or 'e2help.py cmps -v'. Omitting the '-v' will produce a less detailed listing.
  - You can monitor the progress of the refinement in the 'Running Tasks' window, or by running 'e2history.py' in the appropriate directory.
- Once e2refine2d has finished you will want to view the reference free class averages that were generated. You can do this in a number of ways. The first approach is to choose the Browse option in the EMAN2 tasks window. When the browser pops up you should see a folder called r2d\_00. Note that if you have run e2refine2d several times you will most likely see several folders that start with 'r2d\_' and end with a two figure number (such as r2d\_01,r2d\_02). Navigate into the most recent r2d entry (mostly likely r2d\_00 at this stage). The reference free class averages are called classes\_init, classes\_01, and classes\_02 etc. The highest numbered classes file contains the final results.
  - Unfortunately the class averages produced in EMAN2 seem not to be quite as good as those produced by EMAN1 at this point. Fixing this is on our list of things to do, but the averages produced now are adequate for generating initial models, so it isn't a problem. Expect this program to improve in the weeks following the workshop.
  - Another way to view the results of e2refine2d in the workflow is to select the **Refine2D** task in the EMAN2 tasks window. There is one entry in this form for every time you have executed e2refine2d. This entry lists the most recently generated class averages (viewable by double clicking on the entry). Note that this form can also be used to monitor e2refine2d processes as they are running, by allowing you to view the most recently generated class averages during the run.

## Making an initial model

There is a lot of controversy in the cryoEM community on this point. Some people feel that initial model generation is the most critical step in refinement, and you need to use difficult and time-consuming experimental methods such as random conical tilt or +45 degree tilt methods to get an initial model before you can proceed. We disagree completely with this philosophy. In the vast majority of cases, the simple approach used in EMAN can give a very reliable initial model with no additional experimentation required. However, there are a few caveats here:

- Handedness cannot be determined from single particle data without tilting. If determining handedness directly from your data is important, you will have to do some sort of tilt experiment. Personally I find tomography to be the most appealing approach, since it is becoming very standard in most labs



nowadays, and issues with direction of rotation, etc. likely have already been dealt with. You simply need to collect a tomogram with sufficient particles that the handedness can be observed.

- Heterogeneity is an issue. If you have a particle that is highly heterogeneous, the EMAN initial model strategy is likely to fail to produce a unique answer (since there isn't one). Again, single particle tomography may offer the best solution towards understanding the heterogeneity in your specimen. EMAN1 has a 'multirefine' procedure for refining data with many types of heterogeneity, but this has not yet been reimplemented in EMAN2 (it will be soon).
- Poor angular distribution. If your particles have a strongly preferred orientation, especially if this is combined with a low symmetry, there may not be enough information to produce an unambiguous starting model. However, it is also important to note that in this situation, even if you get a good starting model, refinement will also tend to degrade rather than improving the model. To perform a proper 3-D reconstruction, you must have a reasonable number of particles in orientations either spanning the equator of the unit sphere, or along a line connecting the pole to the equator (corresponding to a complete tomographic series). We will discuss this point more in the workshop.

GroEL is actually a fairly difficult case for the EMAN approach, as it tends to be found predominantly in the side view orientation, with only a few end-on views present, and very few particles in between. This leads to a substantial number of bad starting models. However, as you will see, identifying bad starting models can be quite straightforward, and in most cases the correct starting model will be obvious.

- Expand the Initial Model entry in the EMAN2 tasks window in e2workflow. You should see two entries appear beneath Initial Models called **Make model (e2initialmodel)** and **Import model**.
  - Choose **Make model (e2initialmodel)** and hit OK. This will pop up a form displaying the available reference free class averages. This form will also require you to input some parameters such as the number of iterations and the number of models to try (to generate). Enter d7 for the symmetry, select the class averages you wish to use for generating the initial model and hit OK. This will spawn a process (e2initialmodel) that you can monitor in the Running task window.
  - Once the e2initialmodel process has completed, go to the **Initial Models** task in the EMAN2 tasks window. This will show you a table listing the initial models that were produced by e2initialmodel. You can double click on entries to view the results. Before going to the next stage you should probably decide on your best initial model, this will be used to seed 3D refinement in step 6. In theory, the initial models should be sorted in order of quality. In most cases, you will see that the best scoring model will look very much like GroEL, however, due to the small number of 'top' views of GroEL, there is one false positive (obviously wrong) which can sometimes score better than the correct solution. By browsing through the results you can see some of the other failure mechanisms of this method. You can also use the browser to go into the initial\_models directory and see some other files you can use to evaluate the models.
- ➡ You can use the Initial Models table to view models that are generated as the e2initialmodel process is running. Try waiting until the e2initialmodel process has completed to 20% and then open this table.

This ends Part I of the tutorial