

# XtrapolG8 manual

July 19, 2023

This is a manual to use Xtrapol8 v1.2.0 via graphical user interface (GUI).  
Note that not all figures have been updated since v1.1.0.

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Before you start</b>	<b>2</b>
<b>3</b>	<b>Xtrapol8 organization</b>	<b>2</b>
<b>4</b>	<b>Running Xtrapol8 via the GUI</b>	<b>3</b>
<b>5</b>	<b>Configure</b>	<b>5</b>
5.1	Input / Output . . . . .	5
5.1.1	Input Files . . . . .	5
5.1.2	Resolution boundaries . . . . .	6
5.1.3	Output prefix and directory . . . . .	6
5.2	FoFo / Extrapolation . . . . .	7
5.2.1	Xtrapol8 modes . . . . .	7
5.2.2	Occupancies . . . . .	7
5.2.3	Maps and Scaling . . . . .	8
5.2.4	Extrapolated structure factors and map coefficients . . . . .	9
5.2.5	Expert Mode . . . . .	9
5.2.6	Map Explorer . . . . .	9
5.2.7	Negative and Missing reflections . . . . .	11
5.2.8	Scaling resolution . . . . .	11
5.3	Refinement . . . . .	11
<b>6</b>	<b>Output</b>	<b>13</b>
6.1	Output in XtrapolG8 . . . . .	13
6.1.1	Figures in Main tab . . . . .	13
6.1.2	Figures in Occupancies tab . . . . .	19
6.2	Output directory . . . . .	21
6.2.1	Subdirectory per occupancy and Bayesian weighting strategy . . . . .	22
<b>7</b>	<b>Trouble shooting</b>	<b>23</b>
<b>8</b>	<b>Reference</b>	<b>24</b>

## 1 Introduction

Xtrapol8 is software for the calculation of Bayesian-weighted Fourier difference (FoFo) maps, extrapolated structure factors and estimation of the occupancy. It is based on the cctbx toolbox [4] which come automatically with Phenix [5] and uses some CCP4 programs. [9] In order to run Xtrapol8, you will need to have a proper license for Phenix and CCP4 and have both software suites installed. Please use Phenix 1.19 or higher.

Xtrapol8 can be used under Mac osx and Linux operating systems. It has not been tested on Windows but you are free to try.

This manual specifically concerns the usage of Xtrapol8 using the graphical interface.

Please cite us if Xtrapol8 has been useful for your project: De Zitter, E., Coquelle, N., Oeser, P., Barends, T. R. M., Colletier, J.-P., Xtrapol8 enables automatic elucidation of low- occupancy intermediate-states in crystallographic studies, Communications Biology, *accepted*, <https://doi.org/10.1038/s42003-022-03575-7>.

## 2 Before you start

Clone or download the Xtrapol8 repository (<https://github.com/ElkeDeZitter/Xtrapol8.git>) to a place in your PATH or use the full path for running. Take care that all files are stored in the same directory.

Both CCP4 and Phenix should be setup via the command line. A more detailed description on how to setup CCP4 and phenix correctly for Xtrapol8 (Mac):

1. Open a terminal (can be found under *Utilities* or *Other* in *Launchpad*)
2. Add Phenix and the cctbx modules to your PATH:  
Source the file setpaths.sh within the Phenix folder (you can use Finder to find out which Phenix version you have installed and where to find setpaths.sh).  
`source /Applications/phenix-1.20-4459/phenix_env.sh`
3. Add CCP4 to your PATH:  
In the same terminal source ccp4.setup-sh from the ccp4/setup-scripts or ccp4/bin folder (again you can use Finder to find the file).  
`source /Applications/ccp4-8.0/bin/ccp4.setup-sh`
4. Check if Phenix and CCP4 programs can be found using the following commands. They should both return you the complete path.  
`which phenix.refine`  
`which scaleit`

You can add step 2 and 3 to your `~/.profile`, `~/.bashrc`, `~/.zprofile` or `~/.zshrc` file (this depends on your operation system and shell) if you want to avoid doing these steps before you run Xtrapol8. e.g. on Mac:

5. Find out the shell you are using:  
`echo $$SHELL`  
If your shell is `/bin/bash`, then you should add the lines form step 2 and 3 to a file called `~/.profile`. If your shell is `/bin/zsh`, then you should add the lines form step 2 and 3 to a file called `~/.zprofile`.

If you don't do this additional step, then take care to do step 1-4 in the same terminal window and run Xtrapol8 from the same terminal window.

## 3 Xtrapol8 organization

X8\_gui.py is the script to launch the GUI. It will call the other scripts in the folder called gui when launched and launches the Xtrapol8 engine in the background.

Processing consists of 5 steps (Figure 1):

1. Reading and checking input files, data quality assessment.
2. Calculation of (weighted) difference map.
3. Integrate in the difference maps and associate the peaks to the closest amino acid residues.
4. For each occupancy to test:
  - Calculate the requested extrapolated structure factors and map coefficients.
  - Quality assessment of the extrapolated structure factors.

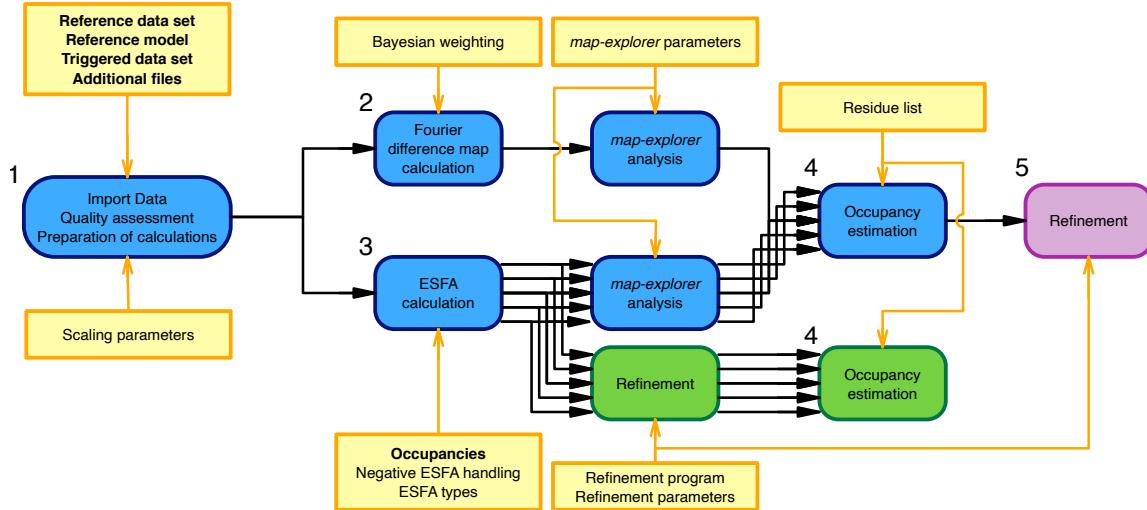


Figure 1: Principal steps taken by Xtrapol8. Main steps are depicted in blue, steps specific to the *slow and curious* mode in green, steps specific to the *fast and furious* mode in purple, input in yellow boxes with obligatory input indicated in bold.

- Real-space and reciprocal-space refinement.
- Estimate the occupancy of the triggered state based on the mFextr-DFc map or based on the structural comparison of the real space refined model with the starting structure.

We refer to the associated publication for more information. [2]

## 4 Running Xtrapol8 via the GUI

X8\_gui.py is the main script and should be called with phenix.python from a terminal. After launching, .  
`phenix.python <where>1/<to>/<find>/X8_gui.py`

To avoid having to run the complete line, you can specify an alias in your `~/.profile`, `~/.zprofile`, `~/.zshrc` or `~/.bashrc` file:

`alias XG8='phenix.python <where>/<to>/<find>/X8_gui.py'`

Then run Xtrapol8 with

`XG8`

instead of `phenix.python <where>/<to>/<find>/X8_gui.py`

In what follows I will replace the `<where>/<to>/<find>/X8_gui.py` just by `X8_gui.py`, but remember that you should always provide the full path or the alias (`XG8`) you defined above.

To get information on all options, the easiest way is to use Xtrapol8 without any argument in the command line (See manual for the command line manual).

`phenix.python Fextr.py`

You can choose to supply the various input files and parameters upon launching Xtrapol8 or to fill in the fields afterwards individually or via an input file. In all cases, each of the files and parameters can be altered afterwards. An example input .phil file with all options and an example with some minimal options can be found on the same location as where you saved the scripts. We refer to the command line manual for more information on the input file.

### Example with filling all the fields after having launched Xtrapol8

- Launch Xtrapol8  
`phenix.python X8_gui.py`

<sup>1</sup>Words and numbers between < and > should be replaced by the actual words and values

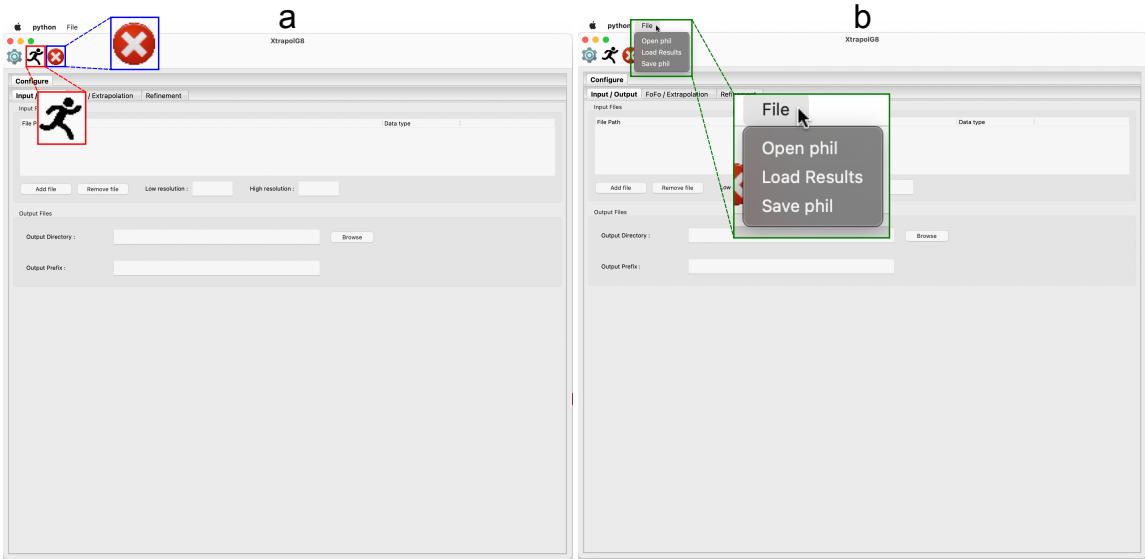


Figure 2: XtrapolG8 window upon launching without any arguments. a, The Run and Abort button are highlighted in red and blue, respectively. b, The File menu allow to open an Xtrapol8 input (.phil) file, load results from a former run and save the Xtrapol8 input file.

2. fill the fields in the Input/Output, FoFo/Extrapolation and Refinement tabs
3. Run Xtrapol8  
*Press the Run button*

#### **Example using an input file upon launching XtrapolG8**

1. Change the Xtrapol8.phil using your favorite editor  
*nano Xtrapol8.phil*
2. Launch XtrapolG8 with the input file as additional argument  
*phenix.python X8\_gui.py Xtrapol8.phil*
3. Optionally change parameters
4. Run Xtrapol8  
*Press the Run button*

#### **Example with opening an input file after having launching XtrapolG8**

1. Change the Xtrapol8.phil using your favorite editor  
*nano Xtrapol8.phil*
2. Launch XtrapolG8  
*phenix.python X8\_gui.py*
3. Load the Xtrapol8.phil file  
*File → Open Phil*
4. Run Xtrapol8  
*Press the Run button*

XtrapolG8 can also be used to generate an input file without running Xtrapol8, which can be useful when one want to launch a long run on a server; or to inspect the results of a former Xtrapol8 run, independent on whether this run was launched via the command line or GUI.

### Example to generate an input file with XtrapolG8

1. Launch XtrapolG8  
*phenix.python X8\_gui.py*
2. Fill the fields in the Input/Output, FoFo/Extrapolation and Refinement tabs.
3. Save the input file.  
*File → Save Phil*

### Example to inspect former results in XtrapolG8

1. Launch XtrapolG8  
*phenix.python X8\_gui.py*
2. Load the directory with the Xtrapol8 results.  
*File → Load Results*

## 5 Configure

### 5.1 Input / Output

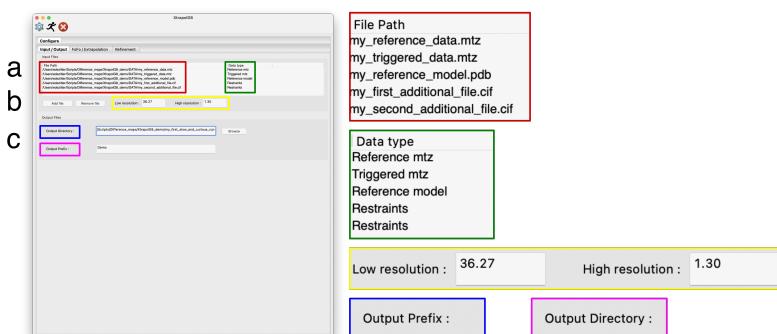


Figure 3: Input / Output panel of XtrapolG8 with Zoom on several fields.

#### 5.1.1 Input Files

Figure 3 a. The file types (Figure 3 green boxes) are automatically guessed but can be changed by right clicking on the type.

##### • Reference mtz

Data for the ground/untriggered/unperturbed state in mtz or mmcif format. For example, in case of a photo-induced process, this is the data when no laser light is applied; in case of a compound-induced process, this is the data before addition or release of the compound.

##### • triggered mtz

Data for the excited/triggered/perturbed state in mtz or mmcif format. For example, this is the data when the crystal(s) have been subjected to laser-light, a compound,... This data usually contains the triggered state with a low occupancy.

Ideally, the unit cell and space group of this data is identical to the reference\_mtz. If this is not the case, the space group and unit cell of the reference\_pdb will be transferred to this data set. But then the results have to be interpreted with great care and Riso and CCiso values will indicate whether or not this was allowed.

##### • Reference model

Model for the ground/untriggered/unperturbed state in pdb or mmcif format. Preferably, this model has been obtained upon refinement with reference\_mtz.

- **Restraints**

Additional library file for non-standard residues, in cif-format. Multiple cif-files can be added upon repetition of the keyword.

If multiple suitable columns are found in the data files, then the following order will be applied for the reference data set:

1. If there are columns containing anomalous and non-anomalous signal, then columns containing the non-anomalous data will get priority over anomalous columns. It should be noted that the Bijvoet pairs will be averaged and merged during processing by Xtrapol8 as the treatment of anomalous data is not yet fully covered.
2. If there are columns containing structure factors, then the structure factors will be directly used. If there are multiple columns with structure factors, then the first encountered columns containing structure factors and sigmas will be used.
3. If there are no columns with structure factors, then intensities will be used and converted to structure factors using the program truncate in CCP4 using French-Wilson scaling.

The following order will be applied in order to extract data from the triggered data set:

1. Knowing the column labels of those columns that were selected from the reference data set, the same columns will be searched in the triggered data set.
2. If the same columns cannot be found, then the same search order as for the reference data set will be applied.

It is thus very important to inspect your input files before using them in Xtrapol8 (e.g. using mtzdmp (CCP4), phenix.mtz.dump (phenix) or ViewHKL (CCP4)), and if necessary generate new input files that only contain your preferred columns (e.g. using cad (CCP4), phenix.reflection\_file\_converter (phenix) or reflection\_file\_editor (phenix GUI)).

### 5.1.2 Resolution boundaries

Figure 3 b.

- **Low resolution**

If no low resolution cutoff is specified, the lowest possible resolution will be used. In practice, this is determined by the common reflections between reference\_mtz and triggered\_mtz. We do not recommend the use of a low resolution cutoff unless you really know what you are doing.

- **High resolution**

If no high resolution cutoff is specified, the highest possible resolution will be used. In practice, this is determined by the common reflections between reference\_mtz and triggered\_mtz.

### 5.1.3 Output prefix and directory

Figure 3 c.

- **Output Prefix**

Name to be used as suffix or prefix in several output files. The name of the triggered\_mtz will be used in case outname is not specified.

- **Output Directory**

Directory in which the output will be stored. If not specified, then the output will appear in a new directory called "Xtrapol8". If the directory already exists, Xtrapol8 will create a new one using the specified outdir name followed by a number.

## 5.2 FoFo / Extrapolation

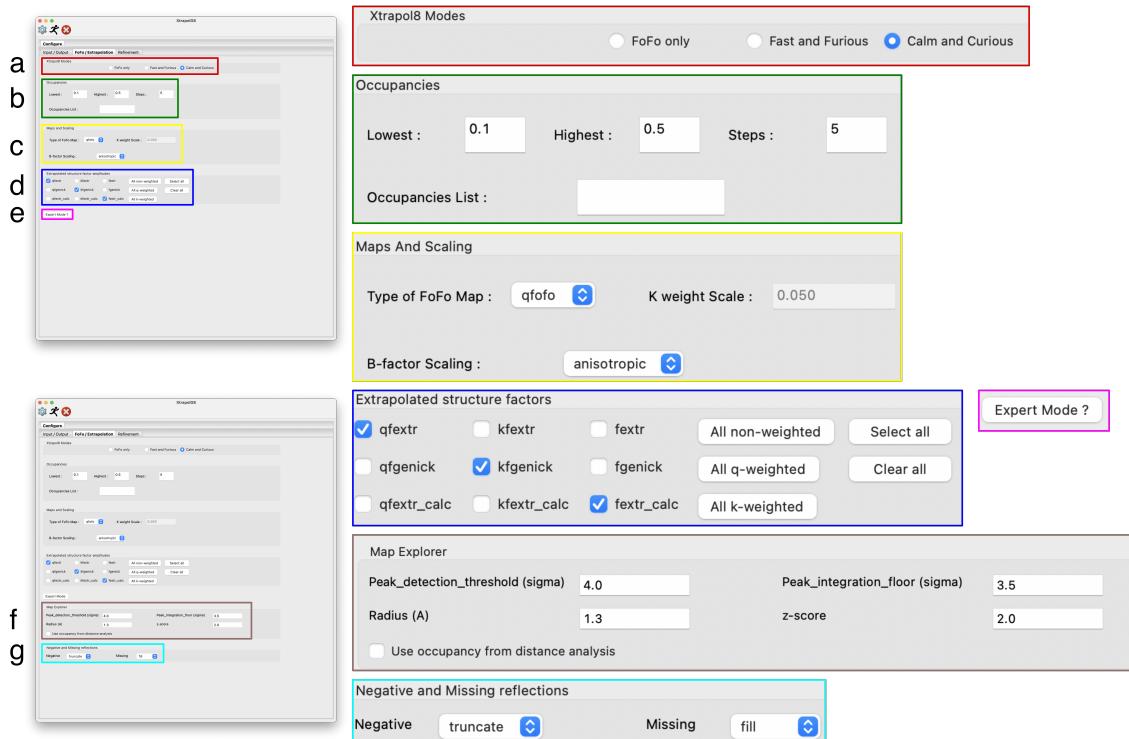


Figure 4: FoFo/ Extrapolation panel of Xtrapol8 in non-expert (top) and expert (bottom) mode with Zoom on several fields.

### 5.2.1 Xtrapol8 modes

Figure 4 a.

Xtrapol8 has two main modes of operation: **Fast and Furious** and **Calm and Curious**. As the name implies, the first mode is a fast mode, ideal for a first run to estimate the parameters or to get a fast result. In *fast and furious* mode, the default parameters will be used for several parameters, thus it can also be seen as an unsupervised mode. The main reason for being much faster than the *calm and curious* mode is that reciprocal and real space refinement will only be run using the extrapolated structure factors and maps of the for the estimated occupancy instead of running the refinements with all extrapolated structure factors, and this only for the qFextr map type. We refer to the command line manual and publication [2] for more information. Fixed parameters are:

- Type of FoFo Map = qfofo
- Extrapolated structure factors = qfextr
- Negative and Missing = truncate and fill

In **FoFo only** mode, Xtrapol8 will only calculate the Fourier difference map.

### 5.2.2 Occupancies

Figure 3 b.

You need to specify a range of occupancies to test. This can be done in two ways:

- **Lowest, Highest and Steps**

Define the lowest and highest occupancies to test and the number of steps between them

- **Occupancy List**

Provide the specific occupancies to test (separated by a space). If a list is provided, it will overwrite the previous method to define the occupancies.

### 5.2.3 Maps and Scaling

Figure 4 c.

Two Bayesian weighting schemes are provided in Xtrapol8 and can be applied to the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors as well as for extrapolated structure factor amplitudes:

- **q-weighting**

Q-weighting uses Bayesian statistics to reduce the contribution of uncertain structure factor values to the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors. More information can be found in Ursby and Bourgeois, 1997. [8]

- **k-weighting**

K-weighting is related Bayesian weighting scheme to reduce the contribution of uncertain structure factor values to the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors introduced by Ren *et al.* [6] In the k-weighting scheme, the user has the option to specify the K weight factor to be more or less stringent about outlier rejection (**kweight\_scale**, default is 0.05).

Using the possibility of q-weighting, k-weighting and no weighing, three possible  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factor amplitudes can be calculated which are used to calculate the Fourier difference map (Table 1):

- **qfofo**

Q-weighted  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors. This type of weighted Fourier difference map is calculated by default.

- **fofo**

Non-weighted  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors.

- **kfofo**

K-weighted  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors.

Table 1: Three types of  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference structure factors

type	Difference structure factors	Map Coefficients of Fourier difference map
qfofo	$q(F_o - F_o) = \frac{q}{\langle q \rangle} \times (F_{\text{trig}} - F_{\text{ref}})$	$mq(F_o - F_o), \phi_{\text{ref}}$
fofo	$F_o - F_o = F_{\text{trig}} - F_{\text{ref}}$	$m(F_o - F_o), \phi_{\text{ref}}$
kfofo	$k(F_o - F_o) = \frac{k}{\langle k \rangle} \times (F_{\text{trig}} - F_{\text{ref}})$	$mk(F_o - F_o), \phi_{\text{ref}}$

Only a single type of FoFo map can be calculated per Xtrapol8 run.

The B-factor scaling option is used to specify the scaling method of the Triggered mtz versus the Reference mtz dataset in scaleit:

- **anisotropic**

Scale the Triggered mtz to the Reference mtz using an anisotropic B-factor scaling scheme.

- **isotropic**

Scale the Triggered mtz to the Reference mtz using an isotropic B-factor scaling scheme.

- **no**

Don't scale the two datasets. This option should only be used in case the two datasets are already scaled, e.g. when crysFEL is used for the data processing and the data are scaled and merged using partialator with custom-split option; or when XSCALE is used and the reference data set is added as reference.

Changing of the resolution range to be used during scaling can only be altered in the command line version of Xtrapol8 for now.

### 5.2.4 Extrapolated structure factors and map coefficients

Figure 4 d.

Three different methods for calculation of extrapolated structure factors can be used. The additional possibility to weight the structure factor difference, gives this rise to nine possibilities (Table 2):

- **qfextr, kfextr and fextr**

As described in Coquelle et al., 2018. [1]

- **qfgenick, kfgenick and fgenick**

As described in Genick et al., 2007. [3]

- **qfextr\_calc, kfextr\_calc and fextr\_calc**

As described in Terwilliger and Berendzen, 1995. [7]

Table 2: Nine types of Extrapolated structure factors and map coefficients

type	Extrapolated structure factors	Map Coefficients	
		$2F_{\text{extrapolated}} - F_{\text{calc}}$ type	$F_{\text{extrapolated}} - F_{\text{calc}}$ type
qfextr	$qF_{\text{extr}} = \alpha \times \frac{q}{\langle q \rangle} \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{obs}}^{\text{ref}}$	$2m qF_{\text{extr}}  - D F_c , \phi_{\text{ref}}$	$m qF_{\text{extr}}  - D F_c , \phi_{\text{ref}}$
kfextr	$kF_{\text{extr}} = \alpha \times \frac{k}{\langle k \rangle} \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{obs}}^{\text{ref}}$	$2m kF_{\text{extr}}  - D F_c , \phi_{\text{ref}}$	$m kF_{\text{extr}}  - D F_c , \phi_{\text{ref}}$
fextr	$F_{\text{extr}} = \alpha \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{obs}}^{\text{ref}}$	$2m F_{\text{extr}}  - D F_c , \phi_{\text{ref}}$	$m F_{\text{extr}}  - D F_c , \phi_{\text{ref}}$
qfgenick	$qF_{\text{genick}} = \alpha \times \frac{q}{\langle q \rangle} \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{obs}}^{\text{ref}}$	$m_{\text{dark}} qF_{\text{genick}} , \phi_{\text{ref}}$	$m_{\text{dark}} qF_{\text{genick}}  - D F_c , \phi_{\text{ref}}$
kfgenick	$kF_{\text{genick}} = \alpha \times \frac{k}{\langle k \rangle} \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{obs}}^{\text{ref}}$	$m_{\text{dark}} kF_{\text{genick}} , \phi_{\text{ref}}$	$m_{\text{dark}} kF_{\text{genick}}  - D F_c , \phi_{\text{ref}}$
fgenick	$F_{\text{genick}} = \alpha \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{obs}}^{\text{ref}}$	$m_{\text{dark}} F_{\text{genick}} , \phi_{\text{ref}}$	$m_{\text{dark}} F_{\text{genick}}  - D F_c , \phi_{\text{ref}}$
qfextr_calc	$qF_{\text{extr\_calc}} = \alpha \times \frac{q}{\langle q \rangle} \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{calc}}^{\text{ref}}$	$2m qF_{\text{extr\_calc}}  - D F_c , \phi_{\text{ref}}$	$m qF_{\text{extr\_calc}}  - DF_c , \phi_{\text{ref}}$
kfextr_calc	$kF_{\text{extr\_calc}} = \alpha \times \frac{k}{\langle k \rangle} \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{calc}}^{\text{ref}}$	$2m kF_{\text{extr\_calc}}  - D F_c , \phi_{\text{ref}}$	$m kF_{\text{extr\_calc}}  - DF_c , \phi_{\text{ref}}$
fextr_calc	$F_{\text{extr\_calc}} = \alpha \times (F_{\text{obs}}^{\text{trig}} - F_{\text{obs}}^{\text{ref}}) + F_{\text{calc}}^{\text{ref}}$	$m F_{\text{extr\_calc}}  - D F_c , \phi_{\text{ref}}$	$m F_{\text{extr\_calc}}  - D F_c , \phi_{\text{ref}}$

$$\alpha = \frac{1}{\text{occupancy}}$$

Multiple maps and types can be selected in one run. The buttons **All non-weighted**, **All q-weighted**, **All k-weighted**, **Select all** and **Clear all** facilitate filtering the desired types.

### 5.2.5 Expert Mode

Figure 4 e.

In **Expert Mode** the fields for altering “Map Explorer”, “Negative and Missing reflections” and additional scaling options become visible.

### 5.2.6 Map Explorer

Figure 4 f.

Xtrapol8 has several build-in methods to estimate the occupancy of the triggered state. All methods that are compatible with the input parameters will be run and results shown, but only one will be used for further steps within an Xtrapol8 run:

#### 1. Difference map method: difference map maximization

The *difference map* method uses the Fourier difference ( $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$ ) and extrapolated difference maps ( $|F_{\text{extrapolated}}| - |F_{\text{calc}}|$ , Table 2 last column) to estimate the occupancy of the triggered state. Practically, this is done by *map-explorer*, which needs the following parameters:

- **peak\_detection\_threshold**

Peak detection threshold, expressed in sigmas (r.m.s.d.). Only peaks with an absolute value equal or above this value will be integrated.

- **peak\_integration\_floor**

Floor value for blob integration, expressed in sigmas (r.m.s.d.) (blob = peak from with more than 2 voxels). Peaks will be integrated from their maximum value towards this lower bound to avoid integration of noise.

- **radius**

Maximal radius for blob annotation to closest amino acid residue. If not defined, the high resolution cutoff (defined or implied by the mtz files) will be used.

- **z\_score**

Z-score cutoff to select only the highest integrated peaks.

The integrated peaks around the selected residues will be used to estimate the occupancy of the triggered state and annotate the most probable extrapolated structure factors. This is done by:

- i Using the `peak_detection_threshold` parameter, we select the peaks in the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  map that are within the `radius` of an atom in the reference model.
- ii Integrate the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  peaks from their maximum value down to the `peak_integration_floor` (if we integrate to zero we'll get one big blob).
- iii Make a histogram of the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  peaks (they should have a normal distribution) and select only the highest using the `z_score` parameter. All map voxels outside of these peaks are masked in the following step.
- iv Using the mask, integrate the same peaks in the  $|F_{\text{extrapolated}}| - |F_{\text{calc}}|$ , sum the positive peaks, sum the negative peaks and sum of the absolute values, divide by the positive/negative/all peaks of the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  map and plot their values (green:positive, red:negative and black: sum of both) and search for the maximum of the sum.

## 2. Difference map method: **difference map Pearson CC**

The second submethod of the *Difference map* method is a standard Pearson correlation coefficient between the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  map and the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  maps. Here, the complete map is used over the complete asymmetric unit, no selection of peaks. This should generate a very similar result to the first plot unless artifacts, such as big peaks in the solvent channel, appear. For now it is not possible to select the occupancy that is given by this plot, except in case of failure to estimate the occupancy based on the maximization of the peaks fails.

## 3. Distance analysis method

In case of running Xtrapol8 in *calm and curious* mode with running the refinement steps, another method is applied to estimate the occupancy. This method is based on a structural comparison of `reference.pdb` with the real space refined models after reciprocal space refinement (`<outname>_occ<occupancy>_<extrapolated structure factor type>-DFc_reciprocal_space_real_space.pdb`, see section 6).

Also in the *distance analysis* method, the residue list arriving from the largest Fourier difference map peaks (`residlist_Zscore<z-score>.txt`) as found by map-explorer will be used. Hence, even though the maps are not used to estimate the occupancy, the map-explorer parameters (**peak\_detection\_threshold**, **peak\_integration\_floor**, **radius** and **z\_score**) should be properly set to select the most useful amino acid residues.

The *difference distance* method can be useful in cases where the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  does not show high and pronounced peaks, but can be very dependent on the refinement parameters used. Therefore, an additional script is provided **refiner.py** in which reciprocal and real space refinement parameters can be altered and refinement can be run again. Currently, **refiner.py** only works using `phenix.refine` and `phenix.real_space_refine`. **Refiner.py** takes an input file which requires the `Xtrapol8_out.phil` file (see section 6) and an input file for reciprocal and real space refinement. Just as for Xtrapol8, it can be launched using `phenix.python`. To get the available parameters:

```
phenix.python refiner.py
```

The occupancy determination using the *difference map* and *distance analysis* methods can be re-run after a complete Xtrapol8 run in a standalone mode in which map-explorer parameters can be altered or custom residue list can be provided (**differencemap\_analysis.py**, **distance\_analysis.py**). The scripts do not use an input file and all arguments have to be passed through by the command line. The standalone modes of the *difference map* and *distance analysis* methods is currently only available through the command line. To get all options:

```
phenix.python differencemap_analysis.py
```

```
phenix.python distance_analysis.py
```

### 5.2.7 Negative and Missing reflections

Figure 4 g.

Extrapolated structure factor amplitudes can become negative. These are not taken into consideration by structure refinement programs and thus lead to a lower apparent completeness. Therefore, several options are provided to treat these negatives. In addition, missing reflections can be estimated (filled) upon map calculations. We refer to the Xtrapol8 publication for more information and suggested strategies to deal with the negative and missing reflections. [2]

- **Negative** The extrapolated structure factor amplitude distribution can be estimated using the truncate method, in which the extrapolated intensities are calculated by squaring the extrapolated structure factors and multiplication by the initial sign. Thereafter the ccp4 program truncate is run to recalculate extrapolated structure factor amplitudes. Otherwise the negative extrapolated structure factor amplitudes can be replaced by their observed or calculated reference counterparts ( $F_{\text{obs}}^{\text{ref}}$  or  $F_{\text{obs}}^{\text{calc}}$ ), rejected, or set to zero. There is also an option to don't treat them ("keep") in which case they are written to the output files as such. Since Genick suggests to reject the negative ESFAs [3], they will be rejected for "fgenick" type of ESFAs (independent of the method selected and other ESFA types).
- **Missing** The options "fill" and "no\_fill" allows the user to specify whether or not missing reflections should be estimated to calculate electron density maps.

Electron density maps with the "keep\_no\_fill" are calculated in any case and can be found in the "maps-keep\_no\_fill" subfolders (see section 6.2.1)

### 5.2.8 Scaling resolution

A specific resolution range for scaling the the reference and triggered data sets can optionally be given:

- **high\_resolution**
- **low\_resolution**

These resolution values will only be applied for scaling and will not be used to the truncate the data.

## 5.3 Refinement

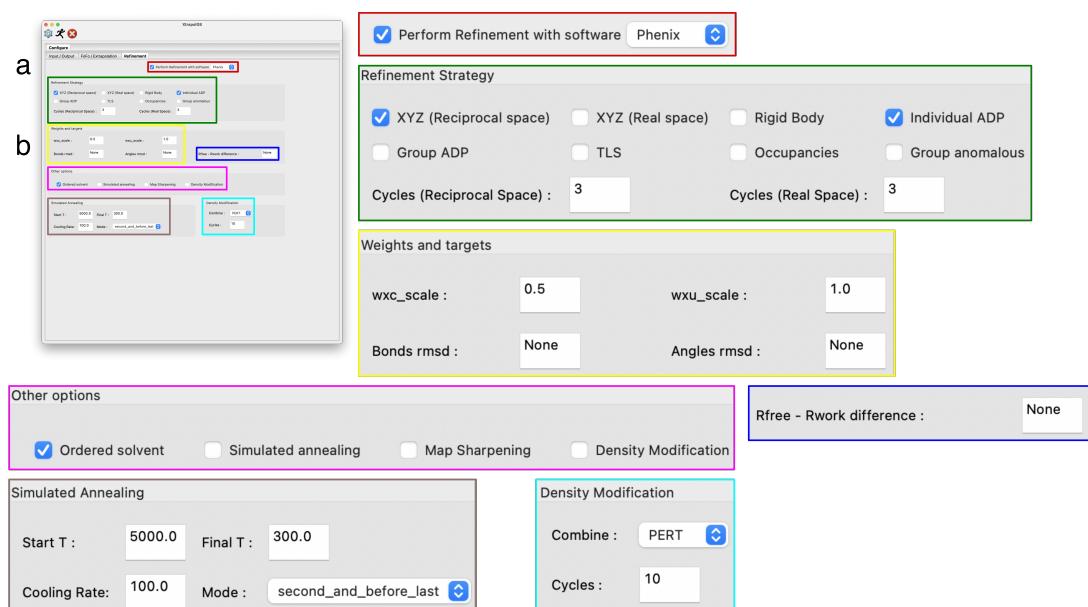


Figure 5: Refinement panel of XtrapolG8 for use with phenix.refine and phenix.real\_space\_refine with Zoom on several fields.

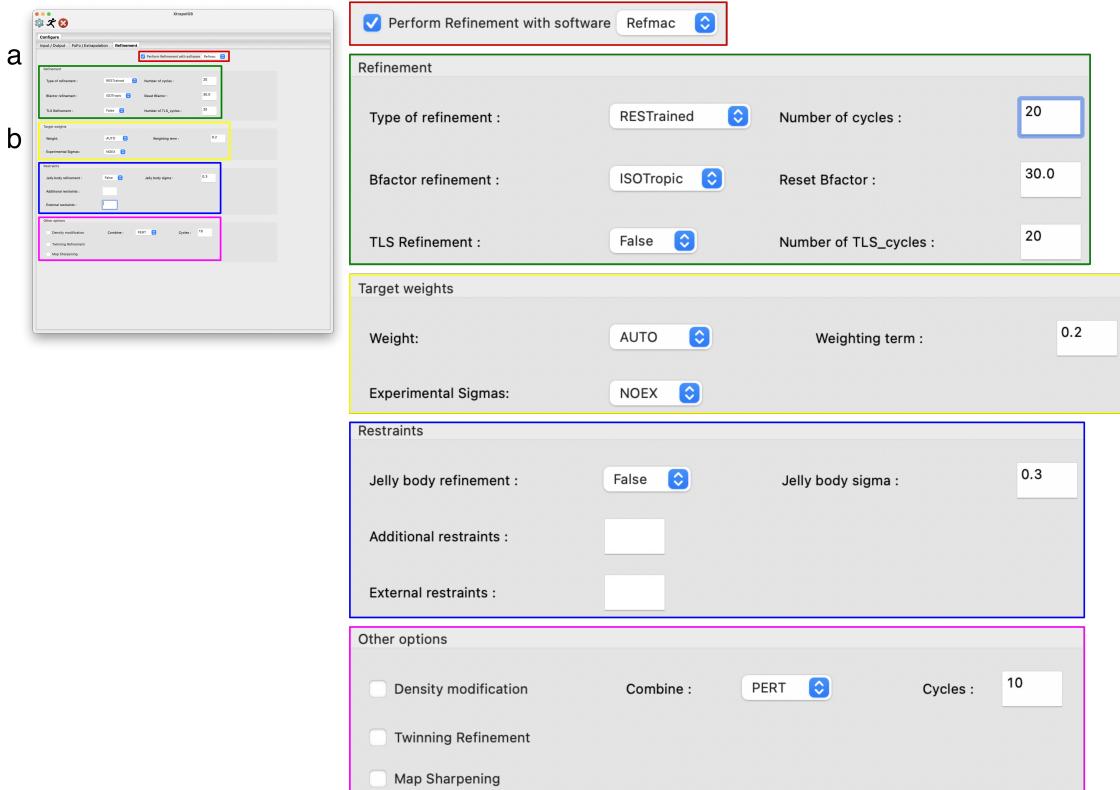


Figure 6: Refinement panel of XtrapolG8 for use with phenix.refine and phenix.real\_space\_refine with Zoom on several fields.

At the point of refinement, 3 different refinements will subsequently be run:

1. Reciprocal reciprocal space refinement using the extrapolated structure factors and reference\_pdb.
2. Real space refinement using the extrapolated map coefficients and reference\_pdb.
3. Real space refinement using the map coefficients and model originating from the reciprocal space refinement (refinement 1).

These reciprocal-space and real-space refinements can be carried out using the couple phenix.refine and phenix.real\_space\_refine (Figure 5) or the couple Refmac5 and Coot (Figure 6) and are selected from the dropdown menu next to **Perform Refinement with software** (Figure 5 a and 6 a). For the refinement parameters (Figure 5 b and 6 b) we refer to phenix and CCP4 documentation. In both cases, if **Density Modification** is ticked, then density modification is performed using the program dm (CCP4), for which we also refer to its documentation.

Unticking the box **Perform Refinement with software** avoids doing any type of refinement. This is useful for any case in which a manual intervention is needed before refinement. This will be the case for ligand binding studies as the ligand is not present in the input model, when the conformational changes are too large to be captured by the automatic refinement or a bond break takes place. The automatic refinements and analyses can be still be executed using the `refiner.py` script form the command line as explained above.

In *fast and furious* mode, refinement will only be performed using the extrapolated structure factors and map coefficients for the estimated occupancy and this only for qFextr extrapolation (see 5.2.1). In *calm and curious* mode, the three refinement steps will be performed for each occupancy estimation and for all of the requested extrapolated structure factor types.

The refinements will be performed using all the data between the high and low resolution cutoff as defined by the Xtrapol8 input parameters. For extrapolated structure factors, a different resolution cutoff might be more appropriate. This cutoff will be different for each type of extrapolated structure factors and occupancy. Even though a high resolution cutoff will be suggested in the generated output, it is the user's responsibility to check the data and decide on the actual application of such a resolution cutoff, and rerun the automatic refinements using `refiner.py` is necessary.

## 6 Output

### 6.1 Output in XtrapolG8

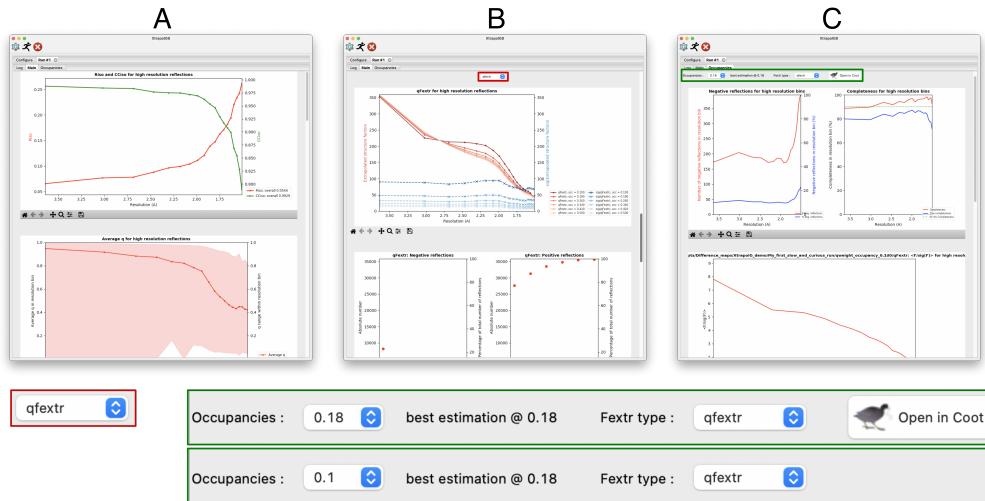


Figure 7: Xtrapol8 output as visualized in XtrapolG8 with Zoom on useful details.

Per Xtrapol8 run, launched from XtrapolG8 or opened via “Load Results”, a tab called “Run <number>” will be generated. Therein three tabs will appear:

- **Log** This tab depicts the Xtrapol8 status and shows the Xtrapol8 log file. When Xtrapol8 is run in “FoFo only” mode, a button to load the model and map is activated in this tab.
- **Main** This tab shows the occupancy independent graphs (Figure 7 A and B). For extrapolation type dependent graphs, the user should select the type for which the graphs should be depicted (Figure 7 B, red box). This dropdown menu is shown below the extrapolation type independent graphs.
- **Occupancies** This tab shows the occupancy dependent graphs. Using the dropdown menu the user can select the2 occupancy and extrapolation type (Figure 7 C, green box). A button to load the models and maps in Coot is provided for each extrapolation type at the estimated occupancy (Figure 7 C, green box).

#### 6.1.1 Figures in Main tab

The figures which are depicted in the “Main” tab are also present in the output directory, and are provided in pdf and png format.

- **Riso\_CCiso.png**

Plot showing Riso and CCiso of reference.mtz and triggered.mtz versus resolution (Figure 8). An overall Riso value below 10 % indicates that the two data sets are strongly isomorphous. However, this is often not the case which can deteriorate the quality of the maps and final results. An Riso value below 0.25 and CCiso above 0.75 for the highest resolution shell are highly recommended.

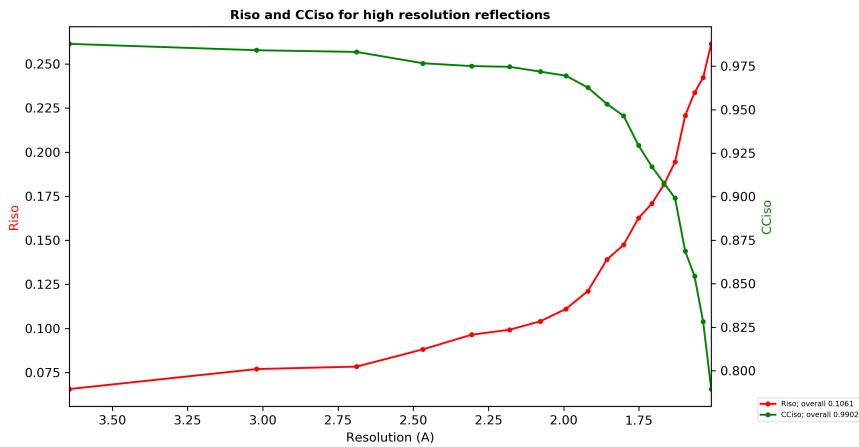


Figure 8: Example of plot showing the average Riso and CCiso values versus resolution. In this example, a high resolution cutoff at 1.9 Å is advised.

- **q\_estimation.png**

Plot showing the average q values and its range versus resolution. (Figure 9)

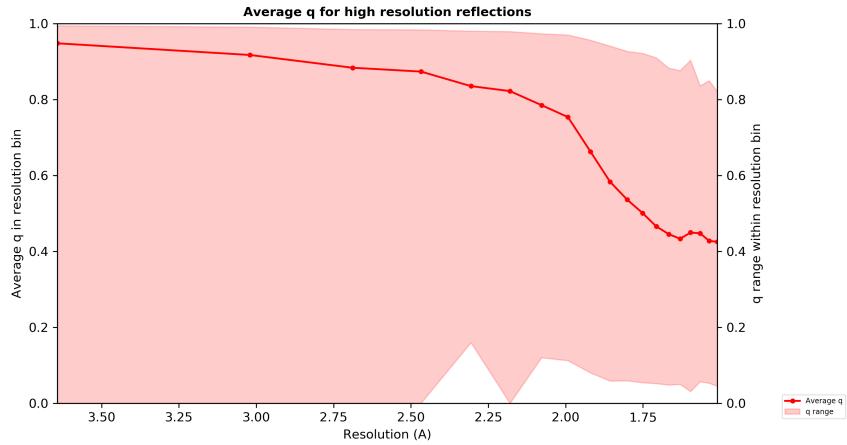


Figure 9: Example of plot showing average q value and its range versus resolution. In this example, the weight of the high resolution reflections is reduced as compared to the low resolution reflections.

- **k\_estimation.png**

Plot showing the average k-values and its range versus resolution. (Figure 10).

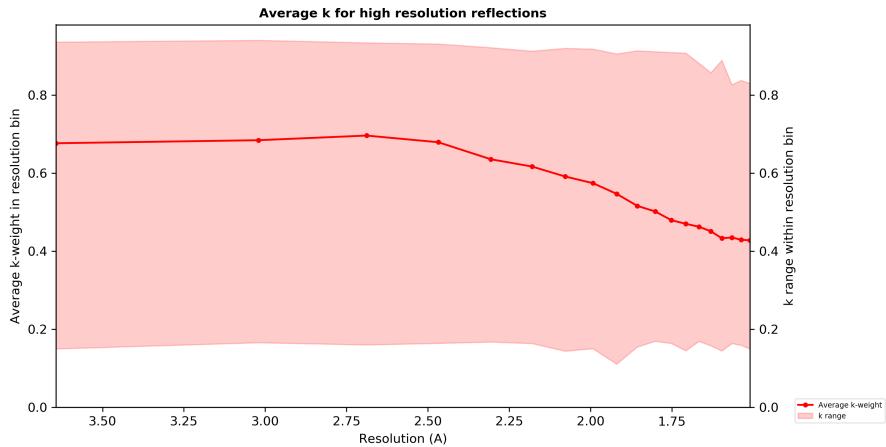


Figure 10: Example of plot showing average k value and its range versus resolution.

- **summed\_difference\_peaks.png**

Plot showing the positive and negative integrated peak area versus amino acid and secondary structure (Figure 11).  $\alpha$ -helices are depicted as pink bars whereas  $\beta$  sheets are depicted by blue triangles; positive and negative peaks in the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference map are depicted as green and red bars, respectively. Ligands and water molecules, if they have associated difference map peaks as determined by map-explorer, are shown on a separate plot at the bottom. For the ease of comparison, the y-axis has the same range in all subplots.

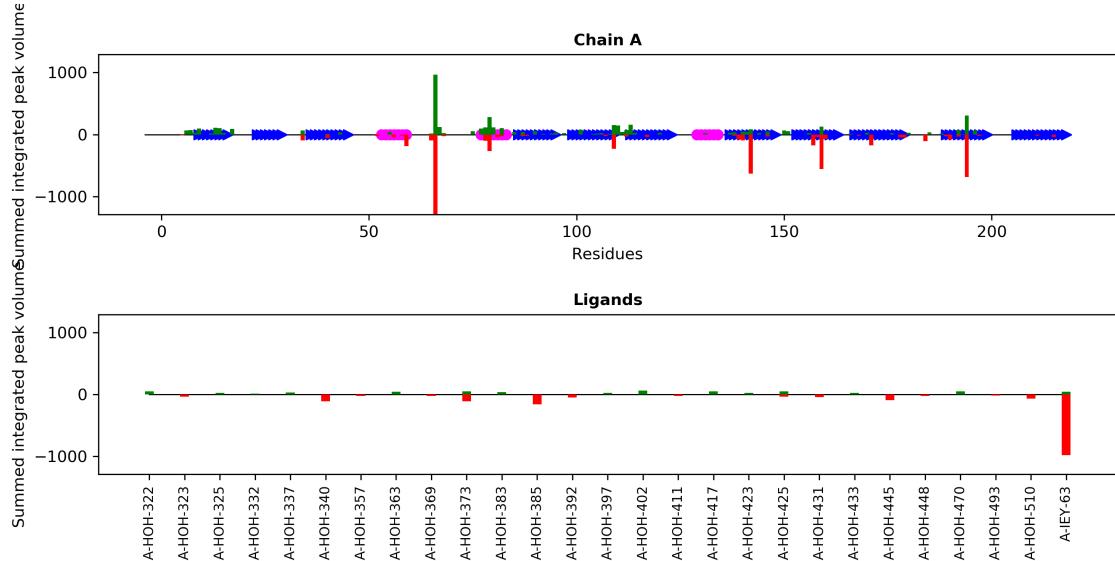


Figure 11: Example of secondary structure plot indicating the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference map peaks. The example model has four chains and several ligands and water molecules to which difference map peaks are annotated.

- **<q/k>FoFo\_sigmas.png**

Plot showing the average  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  values (full lines and spheres for data points, red, left axis) and their error estimation (dashed line with crosses for data points, blue, right axis) versus resolution (Figure 12).

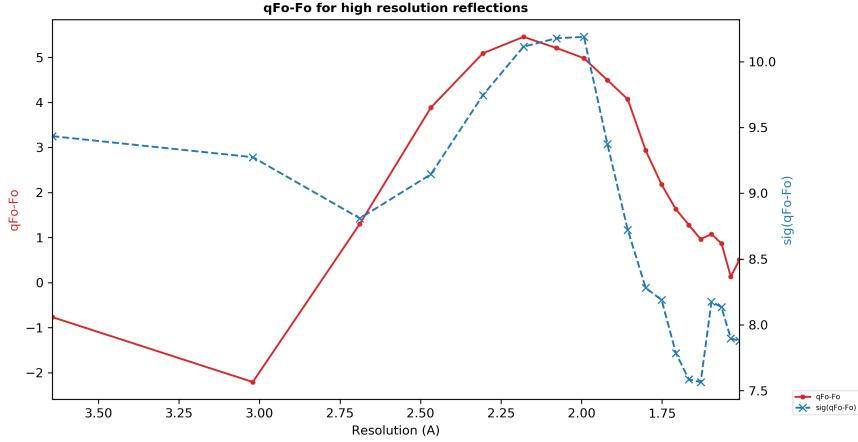


Figure 12: Example of plot showing the average  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  structure factors and estimated errors. In this example, there are more negative differences than positive differences in each high resolution shell. The increase in sigma values at high resolution indicates that the sigma values are no longer correctly estimated

- <extrapolated structure factor type>.sigmas.png

Plot showing the average extrapolated structure factors value (red to salmon) and their error estimation (blue to light blue) versus resolution. This plot is generated for each type of requested extrapolated structure factors (Figure 13).

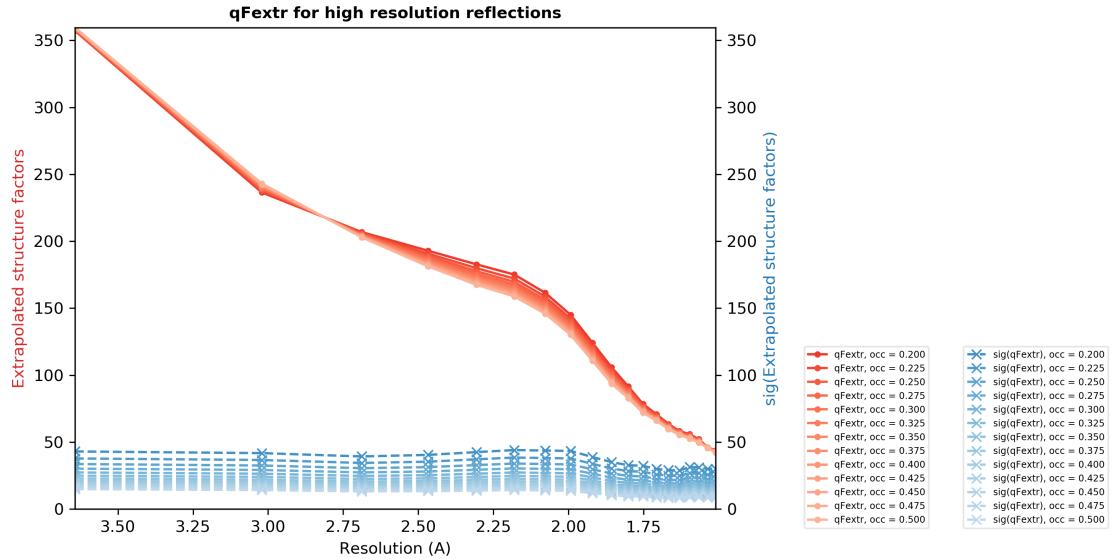


Figure 13: Plot showing the extrapolated structure factors and estimated errors for each type of extrapolated structure factors and occupancy. The bottom plot shows that the data strength decreases fast but at all resolution shells remain higher than the error values.

- Neg\_Pos\_reflections\_<extrapolated structure factor type>.png

Plot showing the absolute number (spheres, left axis) and percentage (right axis) of negative and positive extrapolated structure factors versus occupancy. This plot is generated for each type of requested extrapolated structure factors (Figure 14).

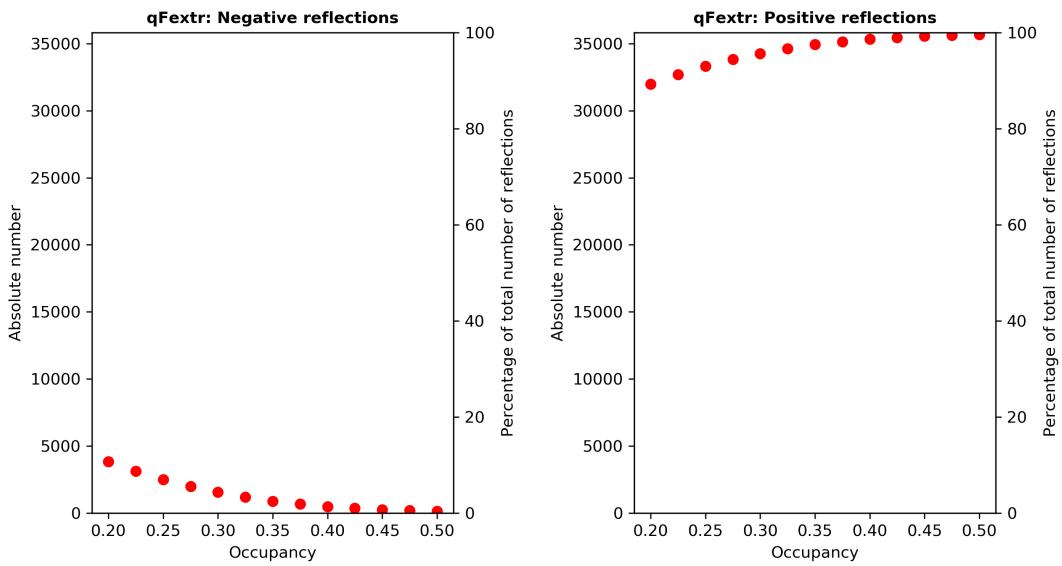


Figure 14: Example of plot showing the number of negative extrapolated structure factors versus the occupancy.

- **alpha\_occupancy\_determination\_<extrapolated structure factor type>.png**

Plot showing the normalized comparison of the extrapolated difference map peaks in order to estimate  $\alpha$  and the occupancy, based the selected residues (residlist\_Zscore<z-score>.txt), all residues and the selected ones with an enhanced signal-to-noise ratio (Figure 15). The  $\alpha/\text{occupancy}$  with a normalized ratio of one, is the most likely  $\alpha/\text{occupancy}$ , but manual inspection should be performed. A different picture is generated for each type of requested extrapolated structure factors.

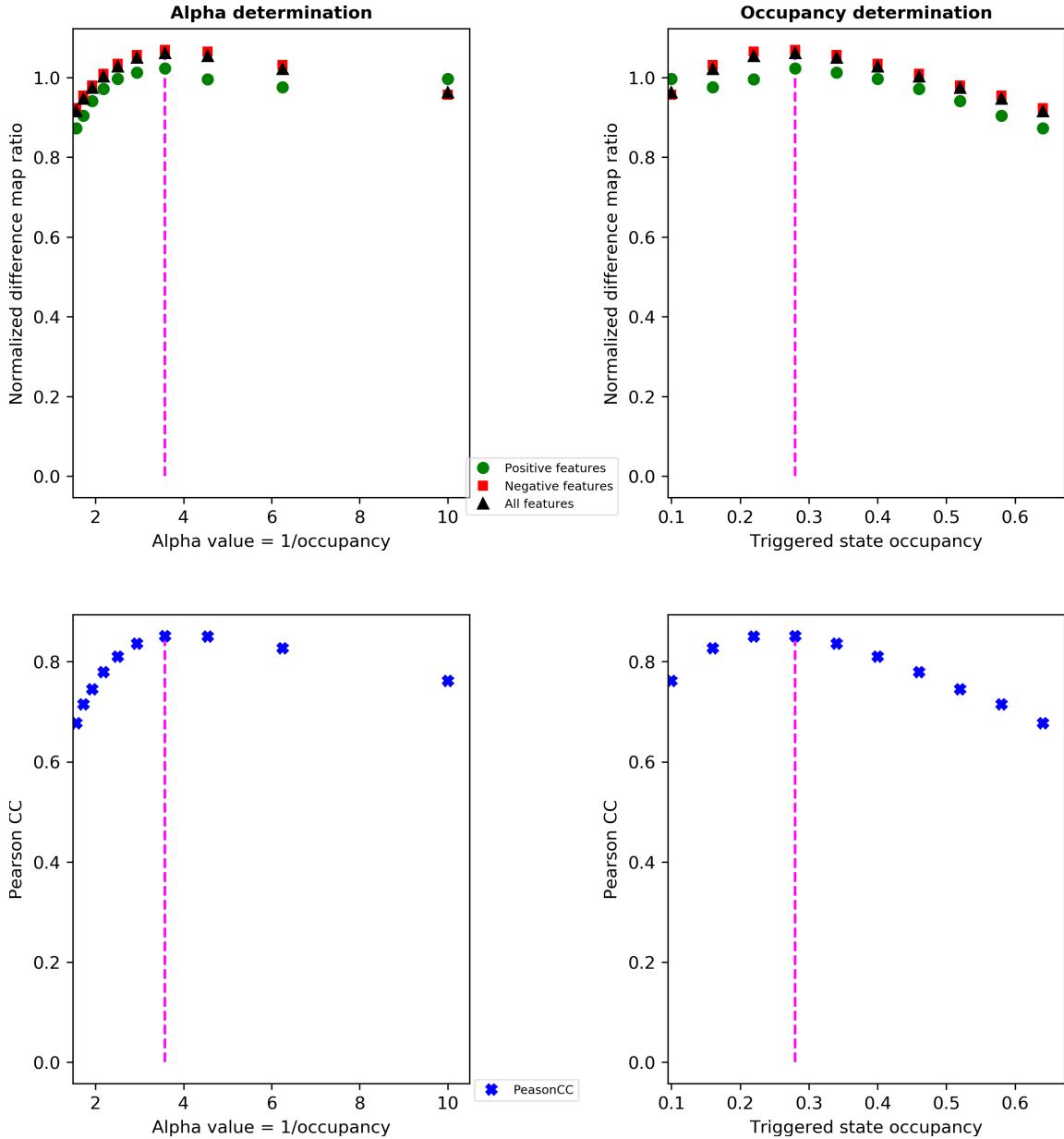


Figure 15: Example of plot showing the results of the *difference map* analysis. The top plots show the  $|F_{\text{extrapolated}}| - |F_{\text{calc}}|$  over  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  ratio of selected peaks versus occupancy and  $\alpha$  value for a single extrapolated structure factor type. The bottom plot shows the Pearson correlation coefficient between the  $|F_{\text{extrapolated}}| - |F_{\text{calc}}|$  and  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  maps versus occupancy and  $\alpha$  value for a single extrapolated structure factor type. The pink dashed line indicates the maximum in each of the plots.

- **Distance\_difference\_plot\_<extrapolated structure factor type>.png**

Plot showing the distance differences between reference\_pdb and the real space refined models (after reciprocal space refinement first) versus  $\alpha$ .

Briefly, all interatomic distances between the atoms of the refined and reference pdb files are calculated for the residues in the list and plotted versus  $\alpha$ . If an sigmoidal fit can be made, they are maintained as significant. If less than 40 distances are significant, they are all plotted, otherwise they are not. The  $\alpha$  is estimated based on the average of the fits of all distances at 99% of reaching the plateau level. Further, the average of all significant distances is plotted and an sigmoidal curve fitted and is used as an alternative  $\alpha$  estimations.

This plot is only generated upon running Xtrapol8 in *calm and curious* mode because it is based on the real space refined models.

A different picture is generated for each type of requested extrapolated structure factors.

#### Distance analysis between reference model and real-space-refined models

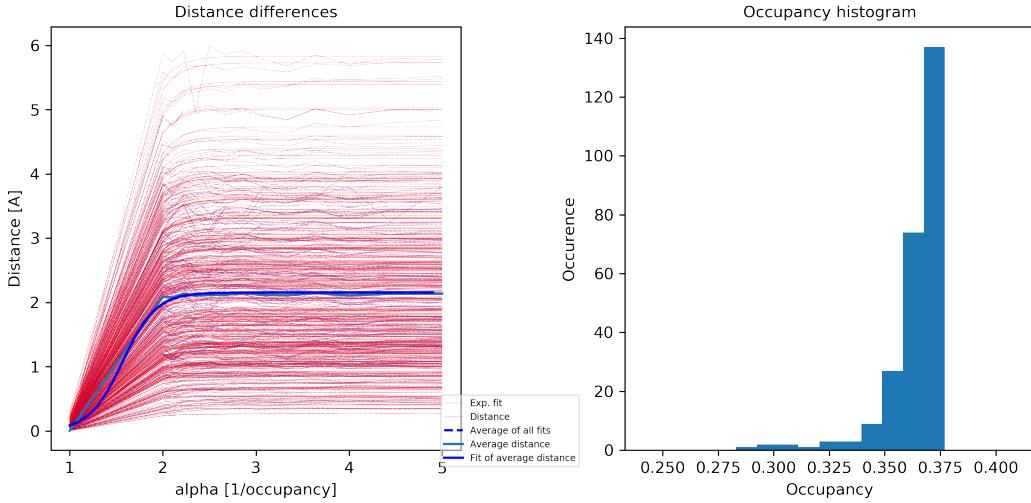


Figure 16: Example of plot showing the distances between atoms in the real-space refined models with those of model\_in versus  $\alpha$  value for a single extrapolated structure factor type (left). The purple lines indicate the experimental distances, their fits are shown by red lines. The average experimental distance is indicated in blue-green, the average fit and fit of the average distances are shown in blue with a dashed and full line, respectively. In this case, the two latter curves superpose. The right curve depicts the distribution of occupancy values as calculated by the individual distances fits.

- <extrapolated structure factor type>\_refinement\_R-factors\_per\_alpha.png

Plot showing the Rwork, Rfree and Rwork-Rfree gap after reciprocal space refinement for each occupancy.

This plot is only generated upon running Xtrapol8 in *calm and curious* mode because it is based on the reciprocal space refinement run for each occupancy.

A different picture is generated for each type of requested extrapolated structure factors.

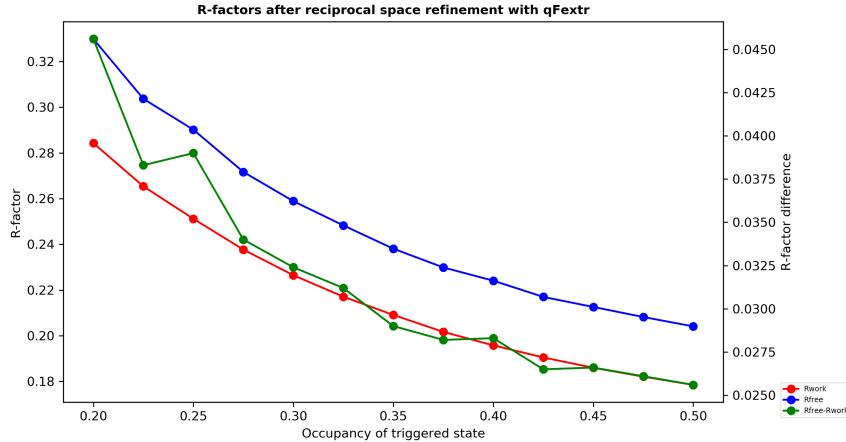


Figure 17: Example of plot showing the refinement R-factors after reciprocal space refinement versus occupancy for a single extrapolated structure factor type.

#### 6.1.2 Figures in Occupancies tab

Three figures are generated for each type of requested extrapolated structure factors and occupancies. The first two figures give an overview of the data quality of the extrapolated structure factors and can help to decide on a resolution cutoff for further refinement and usage.

- <extrapolated structure factor type>\_negative\_reflections.png

Plot showing the absolute number (red, left axis) and percentage (blue, right axis) of negative extrapolated structure factors versus resolution, and plot showing the completeness versus resolution before data was manipulated to eliminating negative extrapolated structure factors (Figure 18). "True completeness" is the completeness of the positive extrapolated structure factors only. A completeness/True completeness above 90% is highly recommended.

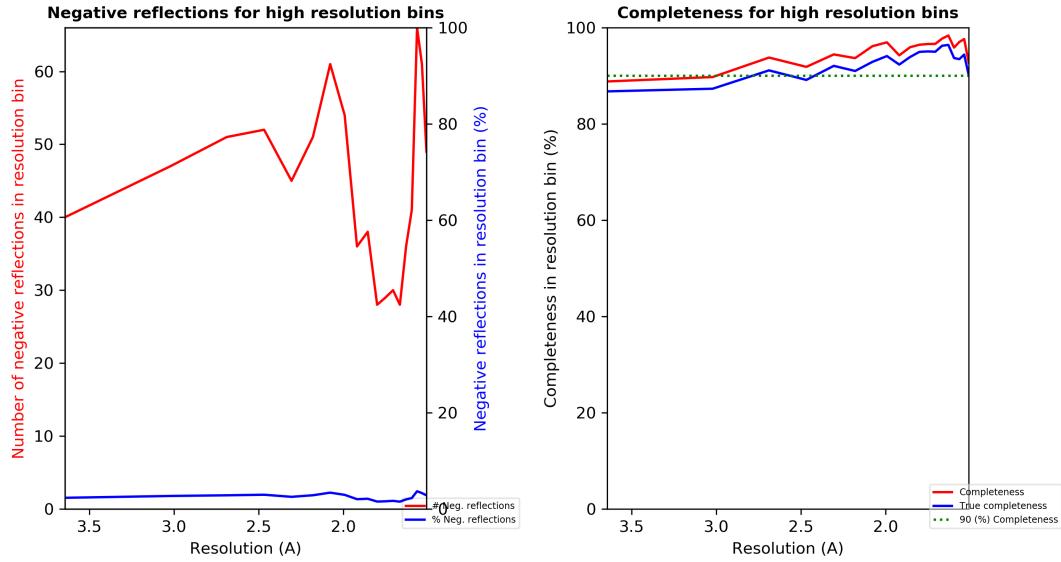


Figure 18: Example of plot showing negative extrapolated structure factors versus resolution on the right, and plot showing the completeness and true completeness on the right.

- <extrapolated structure factor type>\_occupancy<occupancy>-FsigF.png

Plot showing the average extrapolated structure factors over the estimated error versus resolution (Figure 19).  $\langle F/\sigma(F) \rangle$  above 1.22 is highly recommended ( $\sim \langle I/\sigma(I) \rangle > 1.5$ ).

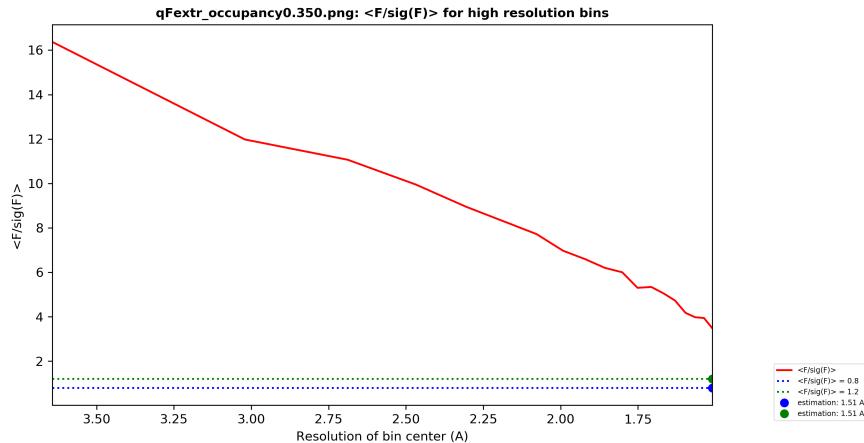


Figure 19: Example of plot showing  $\langle F/\sigma(F) \rangle$  versus resolution. In this example, the resolution cutoff is badly estimated because the signal-to-noise ratio is high in each resolution shell and hence the full resolution range can be used.

- ddm-<reference\_pdb>-<reciprocal\_space\_real\_space\_refined\_model>.png

Plot showing the atomic differences between the reference\_pdb and the real space refined model. This plot is only generated for the models at estimated occupancy.

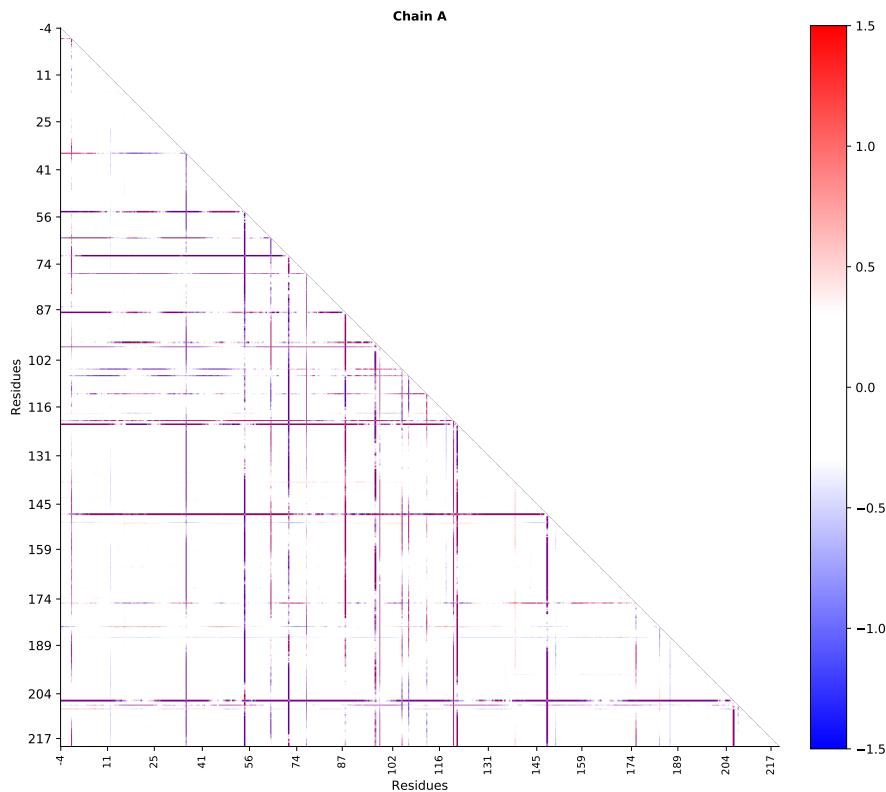


Figure 20: Example of plot showing a distance difference matrix (ddm) indicating for each protein chain the main atomic differences between the input model and the model (reciprocal space + real space refined) at the estimated occupancy.

## 6.2 Output directory

The output directory contains all output files and subdirectories:

- <date-and-time>**\_Xtrapol8.log**  
Log-file with most important output, such as a repetition of the input parameters, data quality statistics in table form, location of output files and occupancy determination.
- **Xtrapol8\_in.phil** and **Xtrapol8\_out.phil**  
For easy rerun and checking of the parameters.  
Whereas the input phil-file should be a copy of the input file and additional command line parameters, the output file takes changes that were made during the program into account.  
The refiner.py script will need this Xtrapol8\_out.phil file as argument in order to correctly find all files.
- <outname>**\_m<weighting>FoFo**  
The  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference map coefficients in three formats: mtz, ccp4 and xplor. The xplor-file is only used for map-explorer analysis and can be removed afterwards to reduce disk space.  
See Table 3 for column names in the mtz-file.
- The output files from map-explorer:
  - **peakintegration.txt**  
Contains the summary of the  $|F_{\text{obs}}^{\text{trig}}| - |F_{\text{obs}}^{\text{ref}}|$  difference map peaks analysis (from map-explorer).
  - **residlist.txt**  
Contains all residues that have at least one associated difference map peak.

- **residlist\_Zscore<z-score>.txt**

Contains all residues that have at least one of the highest associated difference map peak. If the peaks are not normally distributed, no z-score can be calculated and residlist\_Zscore<z-score>.txt will be equal to residlist.txt.

- Figures

See 6.1.1

- A subdirectory per occupancy that is tested:

- Subdirectories called **qweight\_occupancy\_<occupancy>**

Contain all output from q-weighted extrapolated structure factors for a specific occupancy.

- Subdirectories called **kweight\_occupancy\_<occupancy>**

Contain all output from k-weighted extrapolated structure factors for a specific occupancy.

- Subdirectories called **occupancy\_<occupancy>**

Contain all output from non-weighted extrapolated structure factors for a specific occupancy.

See 6.2.1

- **pymol\_movie.py**

Script to open all maps and models in Pymol with models as different frames from highest to lowest occupancy.

Can be opened with Pymol (if installed)

*Pymol pymol\_movie.py*

or from a pymol window:

*run → <path/to/Xtrapol8/output/directory/>pymol\_movie.py*

### 6.2.1 Subdirectory per occupancy and Bayesian weighting strategy

- Each subdirectory from the output directory contains the output specific to a tested occupancy and whether or not q/k-weighting is applied:

- **<outname>\_occ<occupancy><extrapolated structure factor type>.mtz**

The extrapolated structures factors. These can be used as input data for further refinement.

See Table 3 for column names in the mtz-file.

- **<extrapolated structure factor type>\_peakintegration.txt**

Contains the summary of the peaks in the extrapolated difference map.

- **<extrapolated structure factor type>\_residlist.txt**

Contains all residues that have at least one associated difference map peak.

- Extrapolated map coefficients

- \* **<outname>\_occ<occupancy>\_<extrapolated structure factor type>-DFc.mtz**

$2F_{\text{extrapolated}} - F_{\text{calc}}$  and  $F_{\text{extrapolated}} - F_{\text{calc}}$  type map coefficients associated to a certain map type in mtz-format.

See Table 2 for calculation of the map coefficients. See Table 3 for column names in the mtz-file.

- \* **<outname>\_occ<occupancy>\_2<extrapolated structure factor type>-DFc ccp4**

$2F_{\text{extrapolated}} - F_{\text{calc}}$  type map coefficients associated to a certain map type in ccp4-format.

See Table 2 for calculation of the map coefficients.

- \* **<outname>\_occ<occupancy>\_<extrapolated structure factor type>-DFc ccp4**

$F_{\text{extrapolated}} - F_{\text{calc}}$  type map coefficients associated to a certain map type in ccp4-format.

See Table 2 for calculation of the map coefficients.

- \* **<outname>\_occ<occupancy>\_<extrapolated structure factor type>-DFc.map**

$F_{\text{extrapolated}} - F_{\text{calc}}$  type map coefficients associated to a certain map type in xplor-format.

See Table 2 for calculation of the map coefficients.

This file that is used for the map-explorer analysis and can be removed afterwards.

- **maps-keep\_no\_fill**

This folder contains the  $2mF_{\text{extrapolated}} - DF_{\text{calc}}$  and  $mF_{\text{extrapolated}} - DF_{\text{calc}}$  electron density maps using the “keep\_no\_fill” strategy. This means that negatives are not treated and missing reflections are not filled. The figure of merit has been updated (“m” is used, and not “ $m_{\text{ref}}$ ”).

Table 3: mtz column labels

type	Difference structure factors		Map Coefficients	
	amplitudes	sigmas	amplitudes	phases
qfofo	QFDIFF	SIGQFDIFF	QFOFOWT	PHIQFOFOWT
fofo	FDIFF	SIGFDIFF	FOFOWT	PHIFOFOWT
kfofo	KFDIFF	SIGKFDIFF	KFOFOWT	PHIKFOFOWT
type	Extrapolated structure factors		Map Coefficients*	
	amplitudes	sigmas	amplitudes	phases
qfextr	QFEXTR	SIGQFEXTR	2QFEXTRFCWT	PHI2QFEXTRFCWT
kfextr	KFEXTR	SIGKFEXTR	2KFEXTRFCWT	PHI2KFEXTRFCWT
fextr	FEXTR	SIGFEXTR	2FEXTRFCWT	PHI2FEXTRFCWT
qfgenick	QFGENICK	SIGQFGENICK	QFGENICKWT	PHIQFGENICKWT
kfgenick	KFGENICK	SIGKFGENICK	KFGENICKWT	PHIKFGENICKWT
fgenick	FGENICK	SIGFGENICK	FGENICKWT	PHIFGENICKWT
qfextr_calc	QFEXTR_CALC	SIGQFEXTR_CALC	2QFEXTR_CALCFCW	PHI2QFEXTR_CALCFCW
qfextr_calc	KFEXTR_CALC	SIGKFEXTR_CALC	2KFEXTR_CALCFCW	PHI2KFEXTR_CALCFCW
fextr_calc	FEXTR_CALC	SIGFEXTR_CALC	2FEXTR_CALCFCW	PHI2FEXTR_CALCFCW

\* The same logic is applied for creation of the columns for the  $F_{\text{extrapolated}} - F_{\text{calc}}$  difference map.

- Output of refinements

In case of running Xtrapol8 in *fast and furious* mode, the following files will only be present in the folder of the estimated occupancy. In *slow and curious* mode this output is present in all subdirectories.

- `<outname>_occ<occupancy>_<structure factor type>-DFc_reciprocal_space.mtz` or  
`<outname>_occ<occupancy>_<structure factor type>-DFc_refmac.mtz` and  
`<outname>_occ<occupancy>_<structure factor type>-DFc_reciprocal_space.pdb` or  
`<outname>_occ<occupancy>_<structure factor type>-DFc_refmac.pdb`  
Output mtz and pdb file from reciprocal space refinement with phenix.refine or refmac, respectively, using the extrapolated structure factors and reference\_pdb as input.
- `<outname>_occ<occupancy>_<structure factor type>-DFc_real_space.pdb` or  
`<outname>_occ<occupancy>_<structure factor type>-DFc_coot_real_space_refined.pdb`  
Output pdb file from real space refinement with phenix.real\_space\_refine or Coot, respectively, using the extrapolated electron density map and reference\_pdb as input.
- `<outname>_occ<occupancy>_<structure factor type>-DFc_reciprocal_space_real_space.pdb` or  
`<outname>_occ<occupancy>_<structure factor type>-DFc_refmac_coot_real_space_refined.pdb`  
Output pdb file from real space refinement with phenix.real\_space\_refine or Coot after reciprocal space refinement with phenix.refine or refmac, respectively. These pdb-files are used in the distance analysis and ddm-plot, unless the reciprocal space refinement was unsuccessful (in the latter case the pdb files from the simple reciprocal or real-space refinement are used).

- `coot_all_<extrapolated structure factor type>.py`

Script to open all relevant models and maps in Coot for a certain type of extrapolated structure factors.  
`coot -script coot_all_<extrapolated structure factor type>.py`

This script is only present in the folder with the estimated occupancy.

If the Coot executable is found in the PATH and `output.open_coot=True`, then this Coot session is automatically opened at the end of a successful Xtrapol8 run.

## 7 Trouble shooting

- *Not all graphs are being updated during an Xtrapol8 run*  
The GUI still contains some bugs.

→ Try by (re-)selecting the type of extrapolation via the dropdown menus (even if you run with only a single type of extrapolation). If that does not help, try to reload the results using the “Load Results” option from the “File” menu.

- *Not all graphs are shown after an Xtrapol8 run*

The GUI still contains some bugs.

→ Try to reload the results using the “Load Results” option from the “File” menu.

- *Tabs appear to have disappeared*

The GUI still contains some bugs.

→ Try to alter the size of the GUI in order to reload the layout.

- *Error with your input files.*

One of the obligatory input files (reference.mtz, triggered.mtz or reference.pdb) is not defined or not defined correctly.

→ Check your input files.

- Problems with Refmac, statistic calculation, etc.

→ Check if the resolution cutoff is properly chosen.

- *OSError: [Errno 2] No such file or directory*

There is a problem with the output directory.

→ Make sure you have the permission to create the output directory (make sure you have write permission).

- *File "/Applications/phenix-1.19-4080/modules/cctbx\_project/mmtbx/bulk\_solvent/f\_model\_all\_scales.py", line 177, in update\_solvent\_and\_scale\_2. assert approx\_equal(self.r.all(), r.all\_from\_scaler). AssertionError*

This happens upon calculation of F\_model, most probably during the calculation of extrapolated electron density maps. This might be caused by a low data set completeness as a result of rejection of many negative structure factors.

→ Check if there are lot of negative extrapolated structure factors and alter your strategy to handle the negative extrapolated structure factors or increase the lowest occupancy. Note that when the number of negatives is high for a certain set of extrapolated structure factors, truncate will fail and the strategy for that set will be changed to rejecting the negatives, leading to rejection of a large fraction of structure factors and hence failing of the subsequent scaling algorithm.

- *SyntaxError: Non-ASCII character '\xe2' in file /mntdirect/\_programs/x86\_64-linux/phenix/1.20.1-4487/phenix-1.20.1-4487/conda\_base/lib/python2.7/site-packages/scipy/stats/\_continuous\_distns.py on line 3346, but no encoding declared; see http://python.org/dev/peps/pep-0263/ for details*

This can happen on Linux systems and Phenix 1.20. It is caused by a unicode issue in scipy. As Xtrapol8 uses phenix.python and thus the python packages that are shipped with Phenix, the package versions cannot be altered by the user unless by changing Phenix version.

→ Always checkout the latest Phenix nightly build. Otherwise, use Phenix version 1.19 instead of version 1.20.

## 8 Reference

### References

- [1] Nicolas Coquelle, Michel Sliwa, Joyce Woodhouse, Giorgio Schirò, Virgile Adam, Andrew Aquila, Thomas R M Barends, Sébastien Boutet, Martin Byrdin, Sergio Carbajo, Eugenio De la Mora, R Bruce Doak, Mikolaj Feliks, Franck Fieschi, Lutz Foucar, Virginia Guillon, Mario Hilpert, Mark S Hunter, Stefan Jakobs, Jason E Koglin, Gabriela Kovacsova, Thomas J Lane, Bernard Lévy, Mengning Liang, Karol Nass, Jacqueline Ridard, Joseph S Robinson, Christopher M Roome, Cyril Ruckebusch, Matthew Seaberg, Michel Thepaut, Marco Cammarata, Isabelle Demachy, Martin Field, Robert L Shoeman, Dominique Bourgeois, Jacques-Philippe Colletier, Ilme Schlichting, and Martin Weik. Chromophore twisting in the excited state of a photoswitchable fluorescent protein captured by time-resolved serial femtosecond crystallography. *Nat Chem*, 10(1):31–37, 01 2018.

- [2] Elke De Zitter, Nicolas Coquelle, Paula Oeser, Thomas R. M. Barends, and Jacques-Philippe Colletier. Xtrapol8 enables automatic elucidation of low-occupancy intermediate-states in crystallographic studies. *Communications Biology*, 5(1):640, 2022.
- [3] Ulrich K Genick. Structure-factor extrapolation using the scalar approximation: theory, applications and limitations. *Acta Crystallogr D Biol Crystallogr*, 63(Pt 10):1029–41, Oct 2007.
- [4] Ralf W Grosse-Kunstleve, Nicholas K Sauter, Nigel W Moriarty, and Paul D Adams. The computational crystallography toolbox: crystallographic algorithms in a reusable software framework. *Journal of Applied Crystallography*, 35(1):126–136, 2002.
- [5] Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczki, Vincent B Chen, Tristan I Croll, Bradley Hintze, Li Wei Hung, Swati Jain, Airlie J McCoy, Nigel W Moriarty, Robert D Oeffner, Billy K Poon, Michael G Prisant, Randy J Read, Jane S Richardson, David C Richardson, Massimo D Sammito, Oleg V Sobolev, Duncan H Stockwell, Thomas C Terwilliger, Alexandre G Urzhumtsev, Lizbeth L Videau, Christopher J Williams, and Paul D Adams. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallogr D Struct Biol*, 75(Pt 10):861–877, Oct 2019.
- [6] Z Ren, B Perman, V Srager, T Y Teng, C Pradervand, D Bourgeois, F Schotte, T Ursby, R Kort, M Wulff, and K Moffat. A molecular movie at 1.8 Å resolution displays the photocycle of photoactive yellow protein, a eubacterial blue-light receptor, from nanoseconds to seconds. *Biochemistry*, 40(46):13788–801, Nov 2001.
- [7] T C Terwilliger and J Berendzen. Difference refinement: obtaining differences between two related structures. *Acta Crystallogr D Biol Crystallogr*, 51(Pt 5):609–18, Sep 1995.
- [8] T. Ursby and D. Bourgeois. Improved estimation of structure-factor difference amplitudes from poorly accurate data. *Acta Crystallogr. Sect. A*, 53:564–575, September 1997.
- [9] Martyn D Winn, Charles C Ballard, Kevin D Cowtan, Eleanor J Dodson, Paul Emsley, Phil R Evans, Ronan M Keegan, Eugene B Krissinel, Andrew G W Leslie, Airlie McCoy, Stuart J McNicholas, Garib N Murshudov, Navraj S Pannu, Elizabeth A Potterton, Harold R Powell, Randy J Read, Alexei Vagin, and Keith S Wilson. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*, 67(Pt 4):235–42, Apr 2011.