# PeakFilter stages information

This document details each *PeakFilter* stage: when it is executed, if it is an optional stage or under what conditions it is available, and a brief description of what it does. All the matches are done within a tolerance range. The text in green refers to the name of the parameter corresponding to the previous description.

## Step 1: *QC Sample Calculations and Reporting*

**Optional?** It is executed if quality control (QC) samples are present in the input data.
**Description:** For each frame, the mean and the relative standard deviation (RSD) are calculated. The method uses that information to calculate the ratio of RSDs below the lower threshold (`QCRSD`) to RSDs below the upper threshold (`QCRSD`): the higher the ratio the lower the technical variation between the QC samples is overall.

## Step 2: *Solvent removal*

**Optional?** Switchable when solvent (blank) samples are present in the input data.
**Description:** First, the solvent sample replicates for each frame undergo outlier correction. Next, the mean intensity of the remaining solvent replicates for each frame is calculated. Frames where every replicate of every sample have an intensity lower than the corresponding mean intensity times a fold difference (`solventMinFoldDiff`) will be deleted. Remaining sample replicates will have the solvent mean intensity removed from their intensity (with the minimum intensity being zero).

## Step 3: *Low intensity removal*

**Optional?** No.
**Description:** Any replicate intensities less than the given threshold (`intenSignifCutOff`) are considered too insignificant to be processed and are set to zero.

## Step 4: *Mass clustering*

**Optional?** No.
**Description:** Apply hierarchical clustering twice: first to get groups of at least 50 frames, secondly to group frames with very similar m/z together (to be treated as isobaric). The latter groups are the mass clusters.

## Step 5: *Feature set clustering*

**Optional?** No.
**Description:** For non XCMS inputs, frames within each mass cluster are sorted by retention time. Contiguous frames in the same mass cluster that are separated by a small enough retention time (`maxRTDiffAdjFrame`), are grouped and regarded as the same feature set. For XCMS, each frame is already considered a feature.

## Step 6: *Feature peak analysis*

**Optional?** No, but it is only applied to datasets pre-processed with other tool than XCMS.
**Description:** Each feature set is examined individually to check for the existence of sharp, narrow peaks. Wider, flatter features are discarded as contamination.

## Step 7: *In-source ion fragmentation removal*

**Optional?** Yes.
**Description:** Detect and remove in-source ion fragments based on the information provided in the given CSV file (`negIonFragsCSVPath` or `posIonFragsCSVPath`). There are two categories: *remove* and *subtract*. Every frame that matches a *remove* mass is deleted. Every frame which m/z is above the threshold (*mzCutOff* column in the CSV file) that has at least one match in target mass, i.e. m/z plus *subtract* mass (portraying the lost ion fragment), and retention time is removed.

## Step 8: *Mass contaminant removal*

**Optional?** Yes.
**Description:** Remove frames that match the list of contaminant masses found in the given CSV file (`negContaminantsCSVPath` or `posContaminantsCSVPath`). This information comes from known plasticisers or other common contaminants. Retention time is disregarded in the matching process.

## Step 9: *Adduct ion removal*

**Optional?** Yes.
**Description:** Check the existence of adducts based on the list of adduct ion pairings (`negAdductsPairs` or `posAdductsPairs`) and the given CSV file with the list of adducts and their mass differences (`negAdductsCSVPath` or `posAdductsCSVPath`). For every pair of frames matching on m/z and retention time based on this information, the intensity of the frame with the lowest intensity is set to zero (by default).

## Step 10: *Stack removal*

**Optional?** Yes.
**Description:** Lipid and contaminant stacks are removed based on the list of mass multiples for both types of stacks provided in the CSV file (`stacksCSVPath`). Every member of a lipid stack is removed but the parent. Every member of a contaminant stack is removed.

## Step 11: *Replicate retention time correction*

**Optional?** Applied only if the input data has been pre-processed by other tool than XCMS.
**Description:** The retention time for some features in some replicates can be misaligned after the pre-processing and the peak finding process. Move intensities to nearby frames when they are in a sparsely populated frame and it would be better placed in the adjacent densely populated one (within the same feature cluster).

## Step 12: *Outlier correction*

**Optional?** No.
**Description:** Check variation between technical replicates' intensity within the same sample to detect outliers. If they are below the lower threshold (`intensityRSD`) they are removed to reduce the variation. If this solution doesn't solve the problem, all replicates' intensities within the same sample are set to zero, tagging the frame as unreliable.

## Step 13: *Sample mean calculation*

**Optional?** No.

**Description:** For each frame, calculate the mean of every positive sample technical replicate's intensity.

## Step 14: *Mean retention time correction*

**Optional?** Yes.
**Description:** Move mean intensities to nearby frames when they are in a sparsely populated frame and it would be better placed in the adjacent densely populated one (within the same feature cluster).

## Step 15: *Mass reassignment*

**Optional?** No.
**Description:** Replace the m/z of each frame within the same feature cluster or mass cluster (`featMassAssignment`) by the m/z with the highest sample mean intensity.

## Step 16: *Broad retention time contaminant removal*

**Optional?** No.
**Description:** Remove spreads of similar intensity peaks that are detected throughout the retention time range of a mass cluster (contaminants), leaving outlying higher intensity peaks as genuine lipid-like features.

## Step 17: *Isotope removal*

**Optional?** Yes.
**Description:** Remove isotopes of parent analytes: m/z that differ to another m/z by the difference between one $^{13}C$ and one $^{12}C$, and match in retention time.

## Step 18: *Salt cluster removal*

**Optional?** Yes.
**Description:** Remove features identified as salt clusters based on mass defect information provided in the given CSV file (`negMassDefectCSVPath` or `posMassDefectCSVPath`). It will only search in frames up to the retention time threshold specified (`rtCutOff`).

## Step 19: *False Discovery Rate*

**Optional?** Yes.
**Description:** Calculate the False Discovery Rate based on the number of frames identified in the computationally-generated database (COMP_DB) of LIPID MAPS compared to the number of hits in a decoy database, created as a copy of COMP_DB where every mass has been incremented by 0.5 Da. This value provides an estimation of the number of analytes that might be wrongly identified by LipidFinder.