Conditional use of search-engine-determined integration boundaries in Skyline 25.2

Nicholas Shulman[1]; Michael J MacCoss[1]; Brendan MacLean[1]
[1]University of Washington, Seattle, WA

Introduction

Skyline is used for proteomics and small-molecule data analysis, providing tools for chromatogram extraction and peak boundary management. Skyline has always had the option to use peak boundaries specified in spectral libraries such as .blib file from peptide search results or .elib files from EncyclopeDIA.

Users have always been able to disable this behavior in cases where Skyline's peak detection performed better.

The next version of Skyline will introduce options to use peak boundaries from the spectral library if they are available, but also to impute the boundaries in individual replicates if the spectral library has not provided peak boundaries, or if the spectral library peak boundaries are too far away from the boundaries found in other replicates.

Methods

Skyline is open-source software primarily written in C#.

Data analyzed primarily came from the TEI-REX project analyzing irradiated mouse tissue and hair samples.

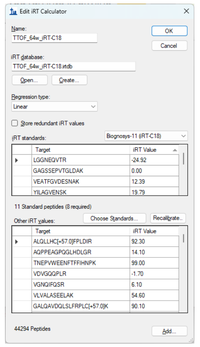
Retention time alignment

Skyline uses information from the runs where peptides were easily detected to choose the correct chromatographic peaks in the runs where the peptide was harder to detect.

Retention time alignment can be used to impute peak boundaries which are more precise and accurate.

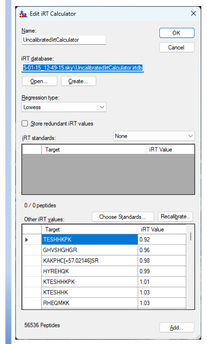
Skyline can perform several types of retention time alignment.

Alignment using iRT standards

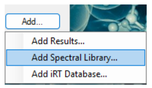


Skyline has supported the use of iRT libraries to predict retention times. Peptides are assigned an iRT score, which is typically a number between 0 and 100 and which is proportional to the expected retention time of the peptide. A small subset of the peptides in the iRT library are designated to be standards. When extracting chromatograms, Skyline first extracts full-gradient chromatograms for the standards and then performs a linear regression between the observed retention time of the standards and their iRT score. Skyline will discard up to 20% of the standard points in order to achieve a regression correlation coefficient (R) of at least 0.99.

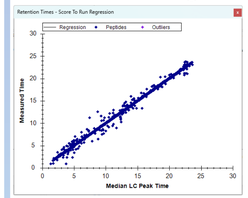
Uncalibrated iRT databases



Skyline 25.2 will allow the creation of iRT libraries which have no standards. These uncalibrated iRT libraries cannot be used for retention time prediction but can be used for alignment between runs.

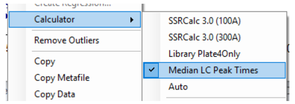


Score to run regression window



The Score to run regression window is typically where one can verify the correlation between a retention time calculator and the observed values in the Skyline document.

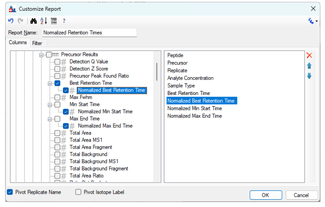
In addition to being able to view iRT calculators, Skyline can also display on the X-axis median retention times from libraries or document chromatograms



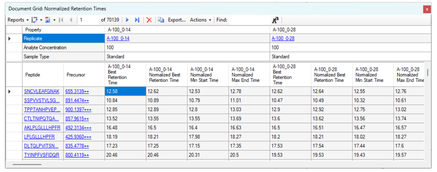
Displaying normalized retention times in the Document Grid

The Document Grid in Skyline is a very flexible way to view, navigate, and export data.

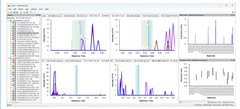
The Report Editor can be used to choose which values to display in grid columns.



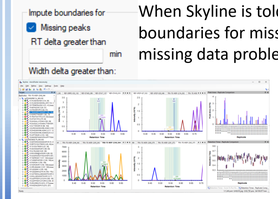
Retention time columns all have a child column which will display the "Normalized" retention time value instead of the raw measured value.



Imputing boundaries for missing DIA-NN results

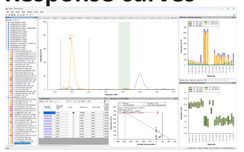


DIA-NN search results often give Skyline no peak boundaries when the peptide could not be confidently identified in a particular sample



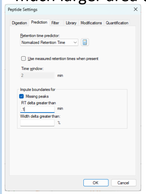
When Skyline is told to impute boundaries for missing peaks, the missing data problem is solved.

Response curves

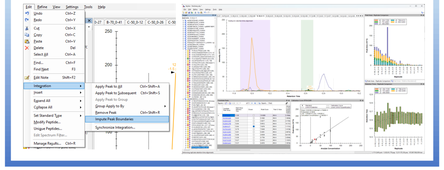


Skyline's peak detection algorithm has traditionally performed poorly in samples that are below the limit of detection.

This is a particular problem in response curve datasets because Skyline always chooses the highest scoring peak. For samples below the limit of detection, the highest scoring peak will usually have a much larger area than the correct peak.

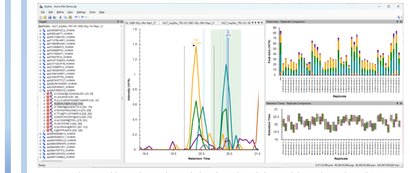


If Skyline is told that to re-impute boundaries whenever the chosen peak is more than X minutes away from the best scoring replicate, the correct response curve shape is achieved.

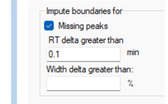


Constraining peaks chosen by EncyclopeDIA

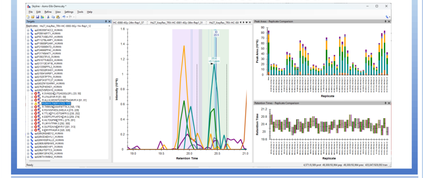
EncyclopeDIA (version 2.12.30) sometimes choose peak boundaries which are clearly wrong, and inconsistent with boundaries chosen in other runs.



Dotted vertical lines show where Skyline has placed the peak boundaries based on information in .elib library file.
Green shaded rectangle show "imputed bounds": where Skyline has mapped the peak boundaries from the highest-scoring replicate



When Skyline is told to move all peak boundaries where the peak is more than 0.1 minutes away from the highest-scoring replicate, the problematic boundaries are fixed.



Imputing boundaries using Skyline's peak scoring model

DIA search engines such as EncyclopeDIA or DIA-NN create libraries which contain explicit peak boundaries. When Skyline is using peak boundaries from libraries like this, Skyline is able to do peak boundary imputation during the same step as extracting chromatograms.

When Skyline is using its own peak scoring algorithm, Skyline does not know which replicate is going to have the highest scoring peak until after all chromatograms have been extracted.

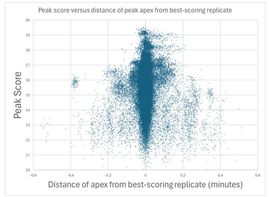
For this reason, peak imputation needs to be initiated by the user in a separate step after chromatogram extraction. This can be done for the currently selected peptide(s) using the **Edit > Integration > Impute Peak Boundaries** menu item, or for the entire document using the **Refine > Reintegrate** menu item.

Conflict of Interest Statement

Five major mass spectrometry vendors (Agilent, Bruker, Shimadzu, ThermoFisher and Waters) provide financial support for the development and maintenance of Skyline. Michael J. MacCoss is a paid consultant for Thermo Fisher Scientific.

Choosing a threshold at which to impute boundaries

Most datasets exhibit the behavior that the highest scoring peaks tend to be chosen at retention times that are consistent across replicates.

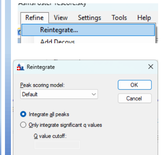


When lower-scoring peaks are chosen, they are relatively more likely to be found far away from other replicates.

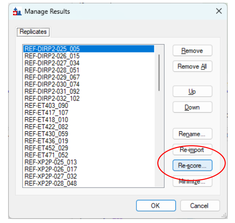
Imputing boundaries in existing Skyline documents

In any Skyline document, first go to the "Prediction" tab in Peptide Settings to choose imputation options.

If Skyline's peak detection algorithm was originally used, then tell Skyline to pick peaks again using the Refine > Reintegrate menu item



If Skyline used peak boundaries from a .blib or .elib file, then use the "Rescore" button in the Edit > Manage Results dialog



Conclusions

These features will be available in Skyline-daily in late summer 2025.

A preview version can be installed from here:
https://proteome.gs.washington.edu/~nicksh/SpecialSkylines/PeakImputation/