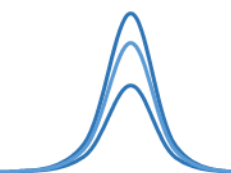
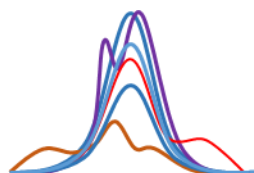


Avant-garde: A Skyline External Tool for automated data-driven DIA data curation

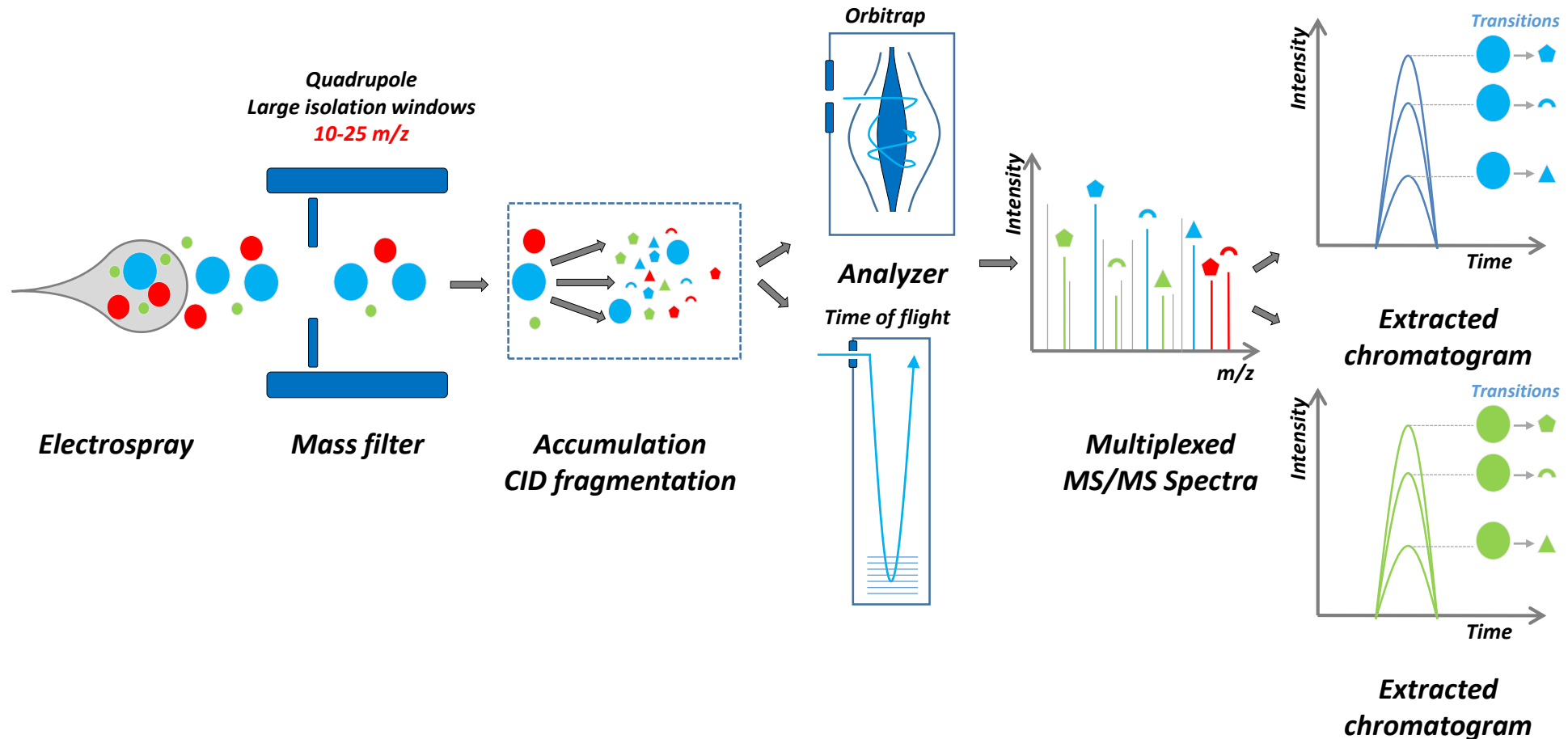
Sebastian Vaca

Proteomics Platform



DIA a powerful but still challenging technique

- DIA is a fundamental technique to decode the complexity of the proteome
- DIA promises to quantify thousands of peptides in complex biological samples

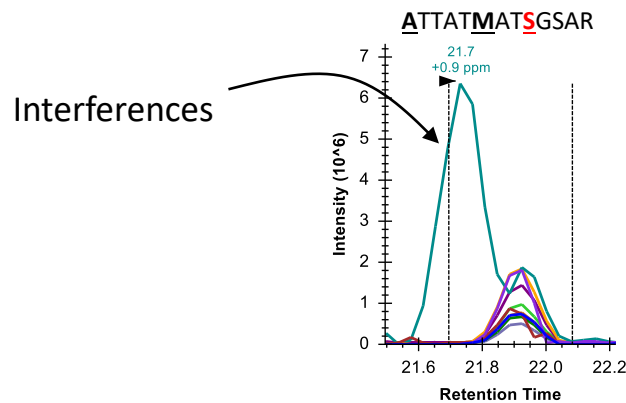


DIA a powerful but still challenging technique

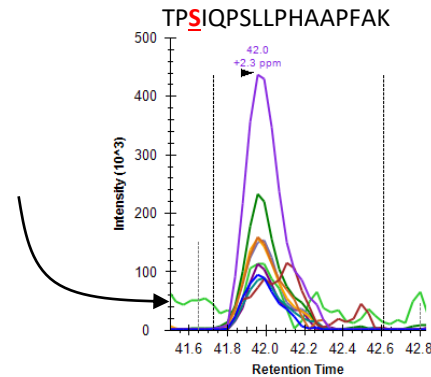
■ Archetype of the ideal DIA signal:

- 1) transitions with same elution peak shape
- 2) relative areas mirroring the relative intensities found in their reference spectrum from a library
- 3) a low mass error
- 4) consistency across all MS runs being compared.

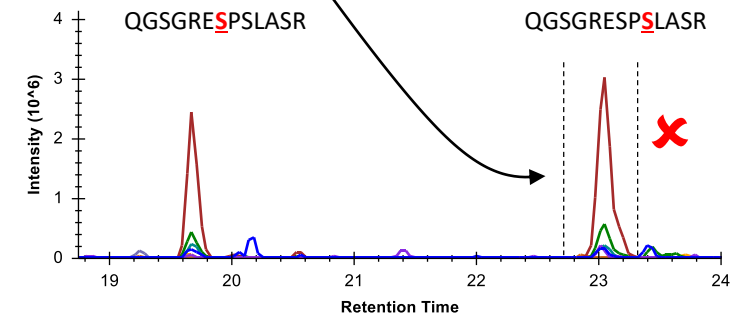
■ In practice, DIA data analysis is not trivial



Noisy signals

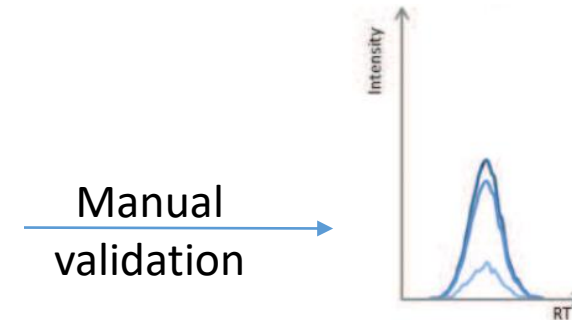


Wrong peak picking



DIA a powerful but still challenging technique

- DIA data analysis often uses statistical validation (target/decoy approach) of peptide identification
- In practice
 - 1) a defined set of transitions is chosen and used to quantify a peptide
 - 2) a score is used to discriminate targets and decoys
 - 3) validation at 1%FDR for protein/peptide identification
- The validation approach might not reflect the quality of the quantitative suitability of a peak in ϵ





Signal processing tool meant to refine the results of DIA/PRM analysis tools:

- Removes transitions subject to interference
- Reduces noise
- Refines peak detection and adjusts peak boundaries
- FDR estimation of analytes for *quantitative suitability*

Outline of the presentation

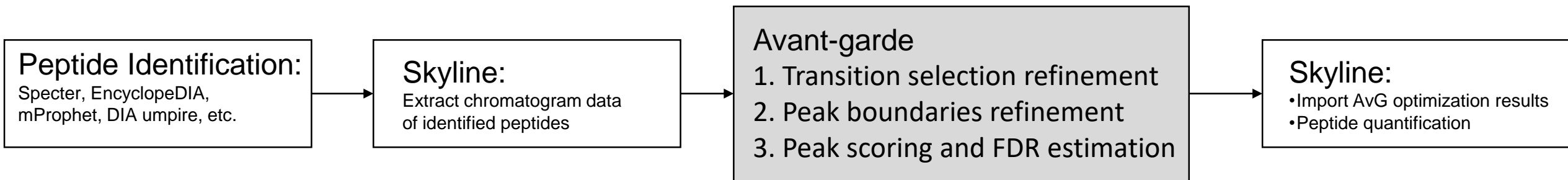


- Principles used by *Avant-garde* to refine DIA/PRM signals
- Example of *avant-garde* on a real example

Avant-garde's workflow



- Avant-garde is a tool designed for automated data curation
- meant to complement common DIA analysis tools





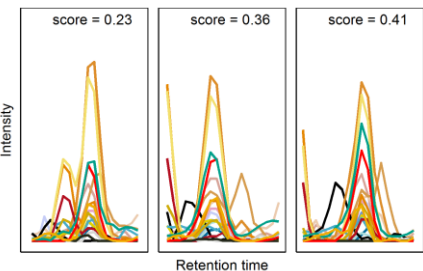
Avant-garde's data-driven and ensemble-driven optimization approach



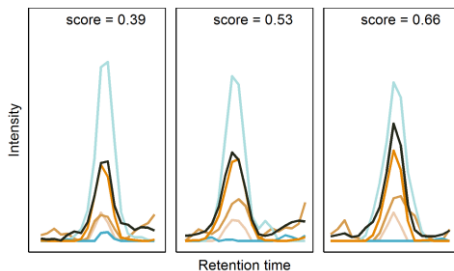
- **Data-driven:** Avant-garde uses both prior knowledge and the DIA data itself to optimize the signals
- **Ensemble-driven:** considers all data from all samples in a given dataset

Avant-garde's genetic algorithm for transition refinement

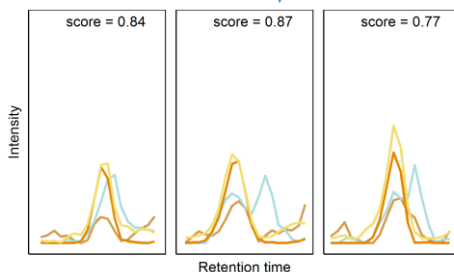
Evolving towards accurate measurements



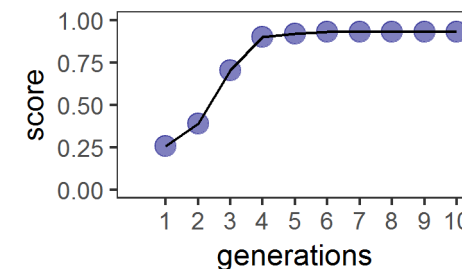
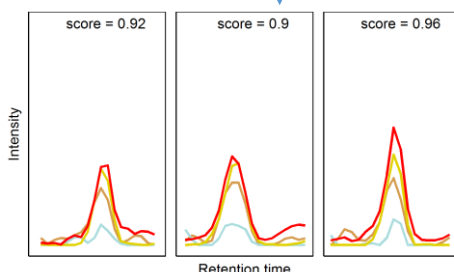
Generation 1



Generation 2



Generation N

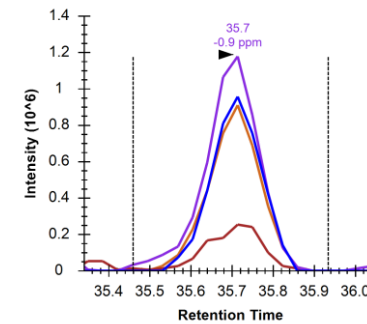
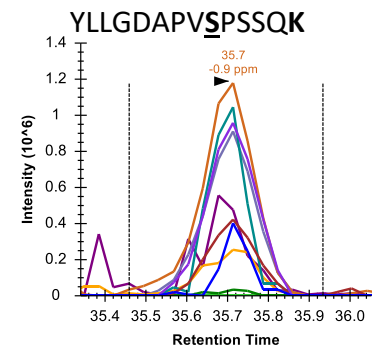
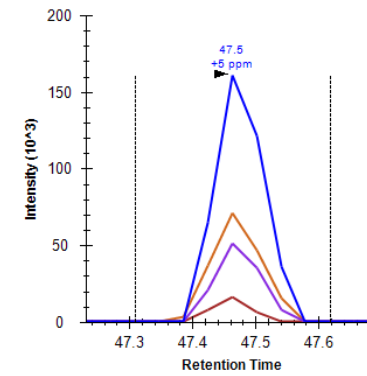
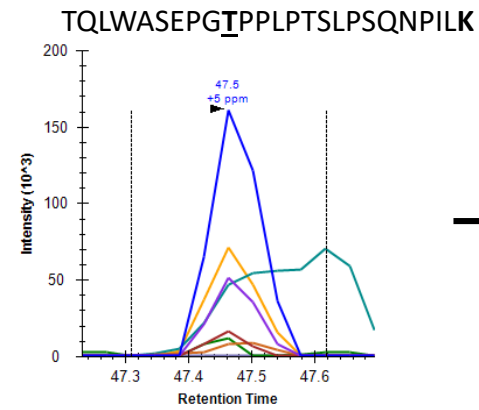
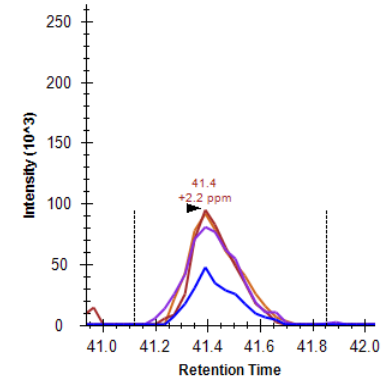
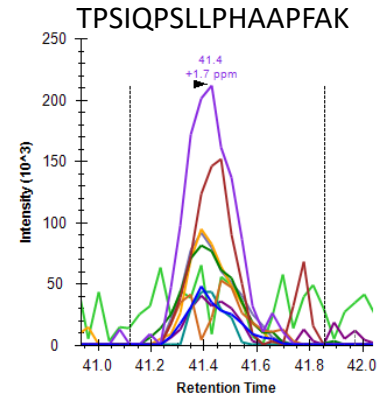


- Clean signals
- Highest intensities
- Similarity between transitions
- Similarity of signals in the entire dataset

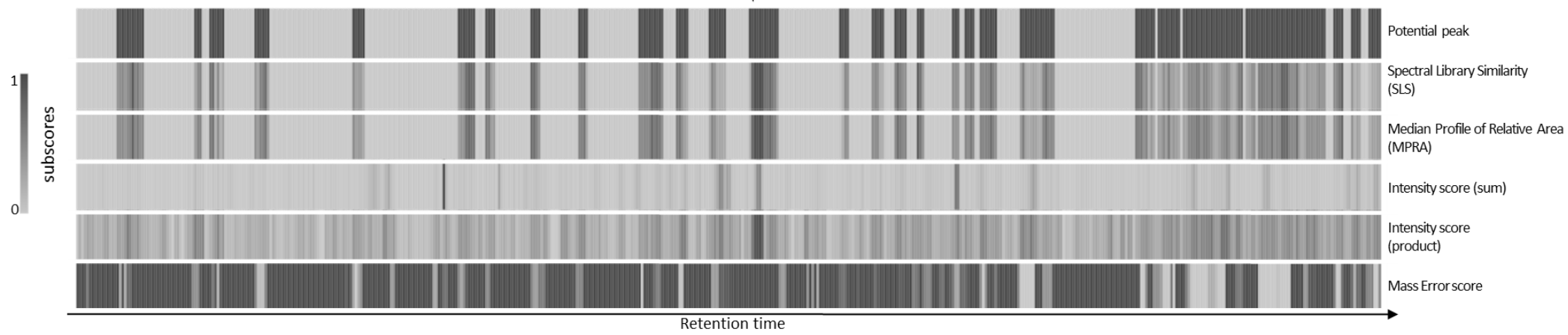
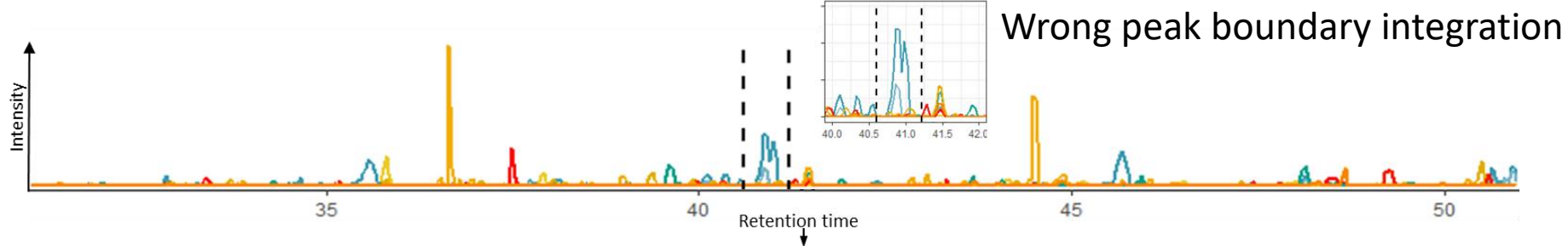
Seeing is believing...

Before

After

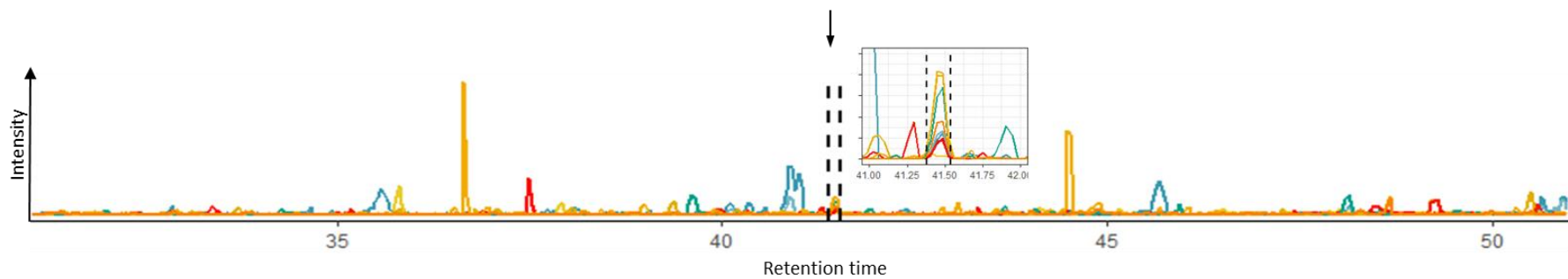
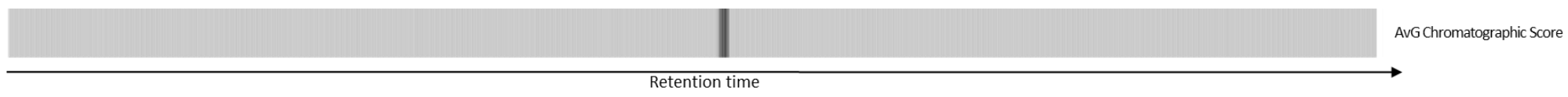


Avant-garde's automated refinement of peak integration boundaries



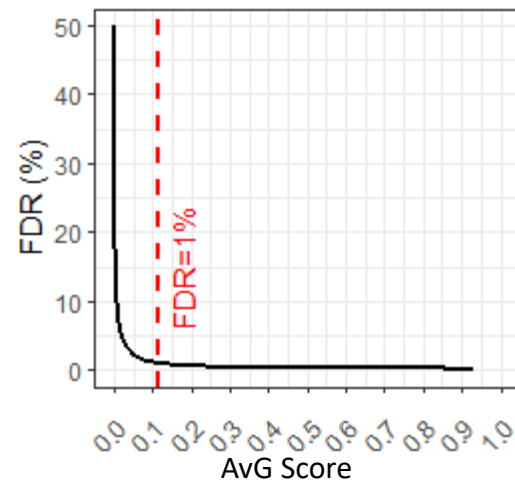
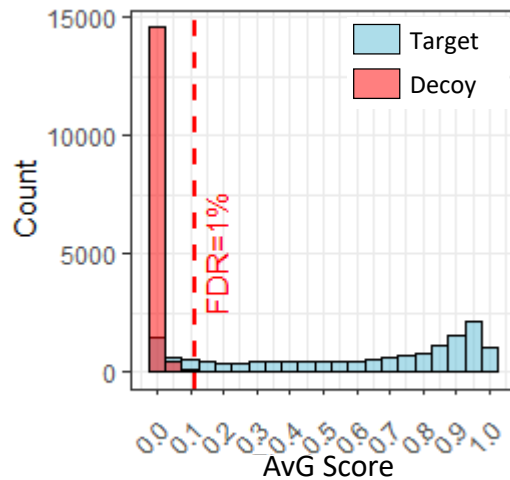
Multiplicative combination

$$\prod scores_i^{\alpha_i}$$

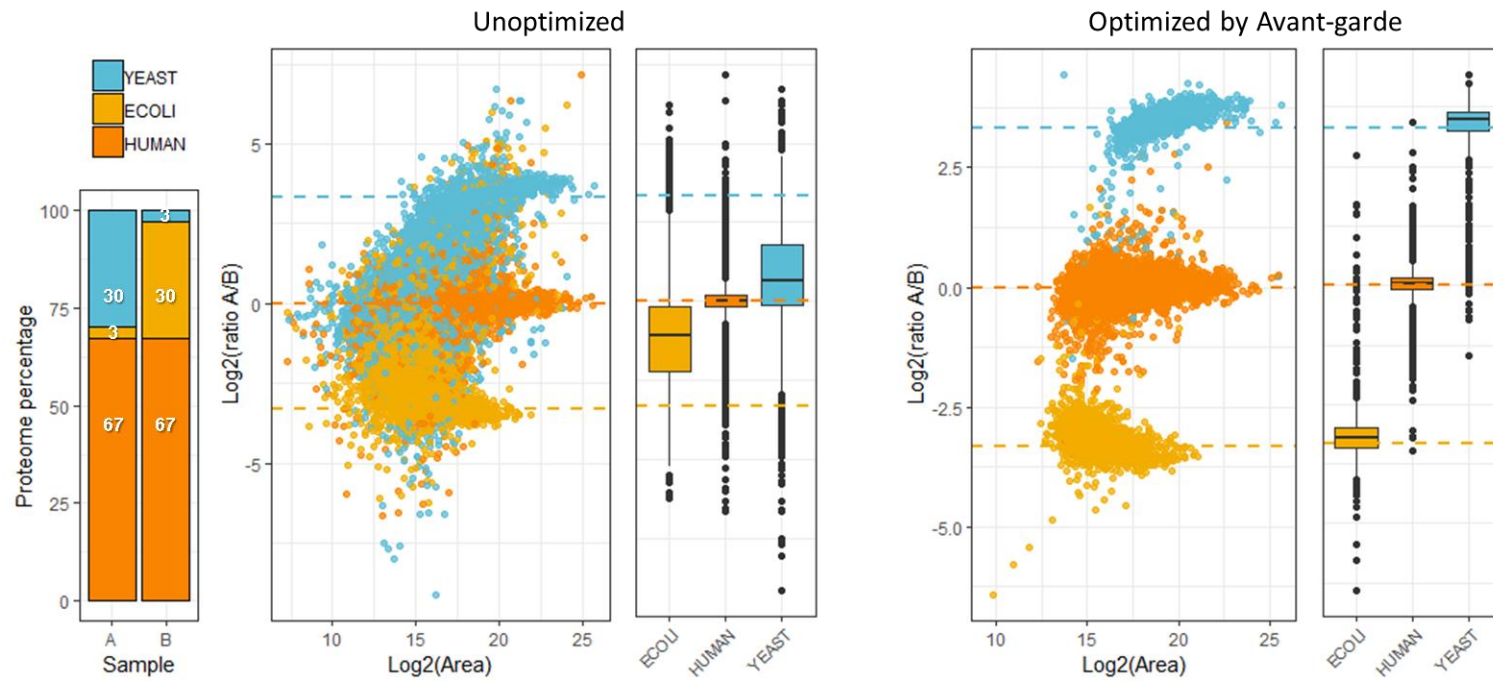


FDR by evaluating the quantitative suitability instead of peak detection

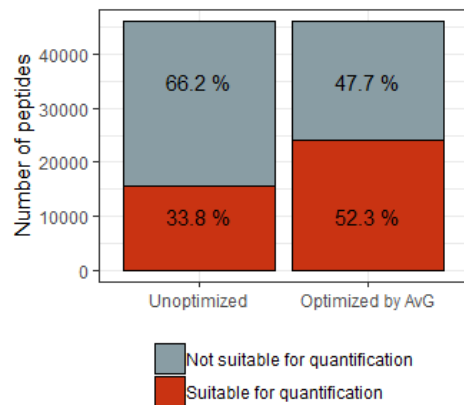
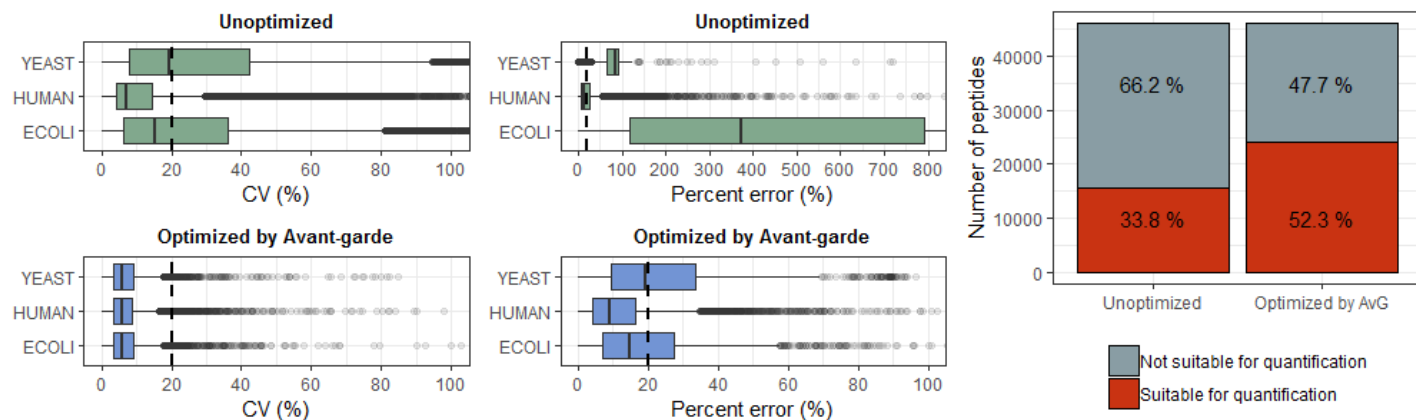
- After curation each peptide is scored again to estimate a dataset-level FDR for **quantitative suitability**, not just detection.
- Avant-garde's ensemble-driven scoring strategy is designed to produce very conservative results by penalizing poor-quality signals.
- **Quantitative suitability** is a metric to evaluate the quality of the signals used to quantify a peptide.



Benchmarking avant-garde against a complex sample



Benchmarking avant-garde against a complex sample



Where can I find avant-garde?



Avant-garde: An automated data-driven DIA data curation tool.

Alvaro Sebastian Vaca Jacome, Ryan Peckner, Nicholas Shulman, Karsten Krug, Katherine C DeRuff, Adam Officer, Brendan MacLean, Michael J MacCoss, Steven A Carr, Jacob D Jaffe

doi: <https://doi.org/10.1101/565523>

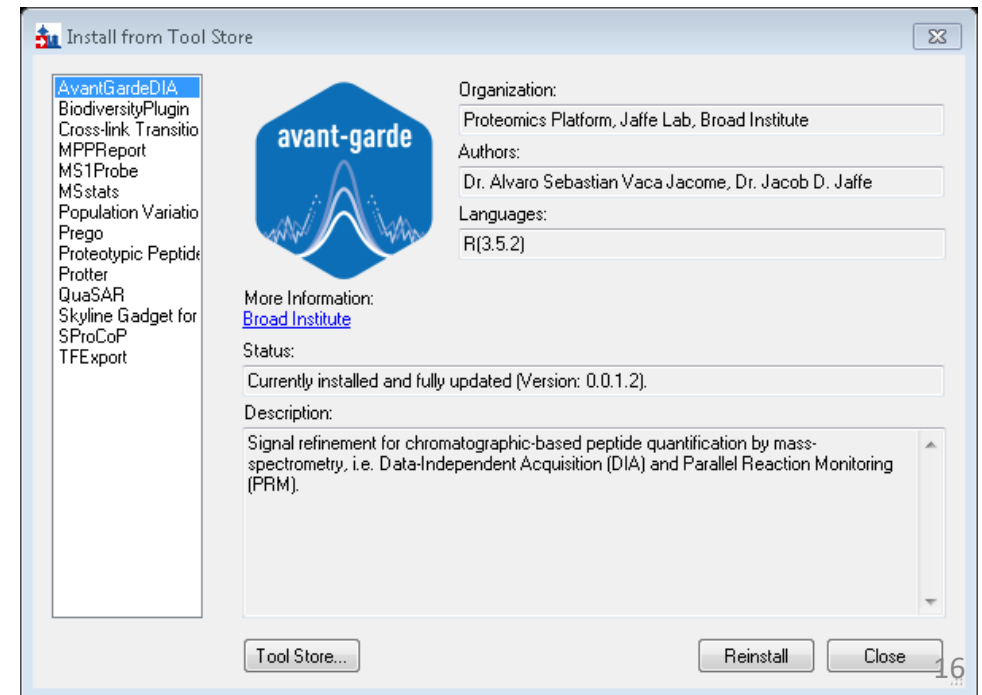
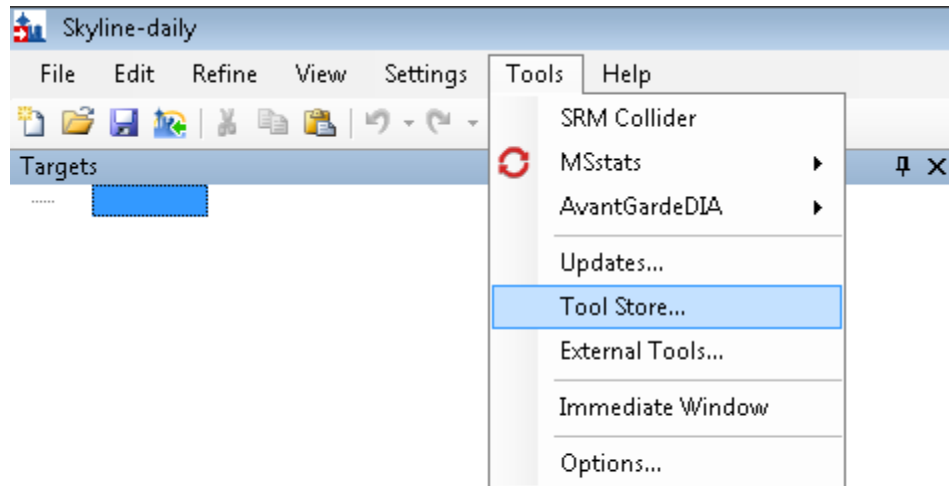


github@SebVaca

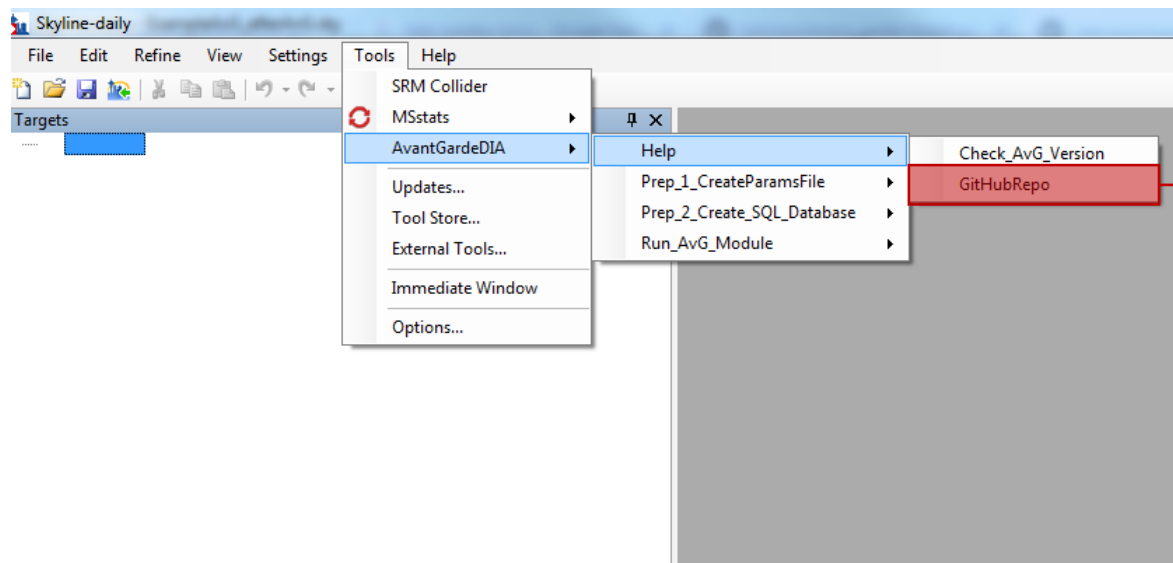


External Tools

Install Avant-garde from the Skyline Tool Store



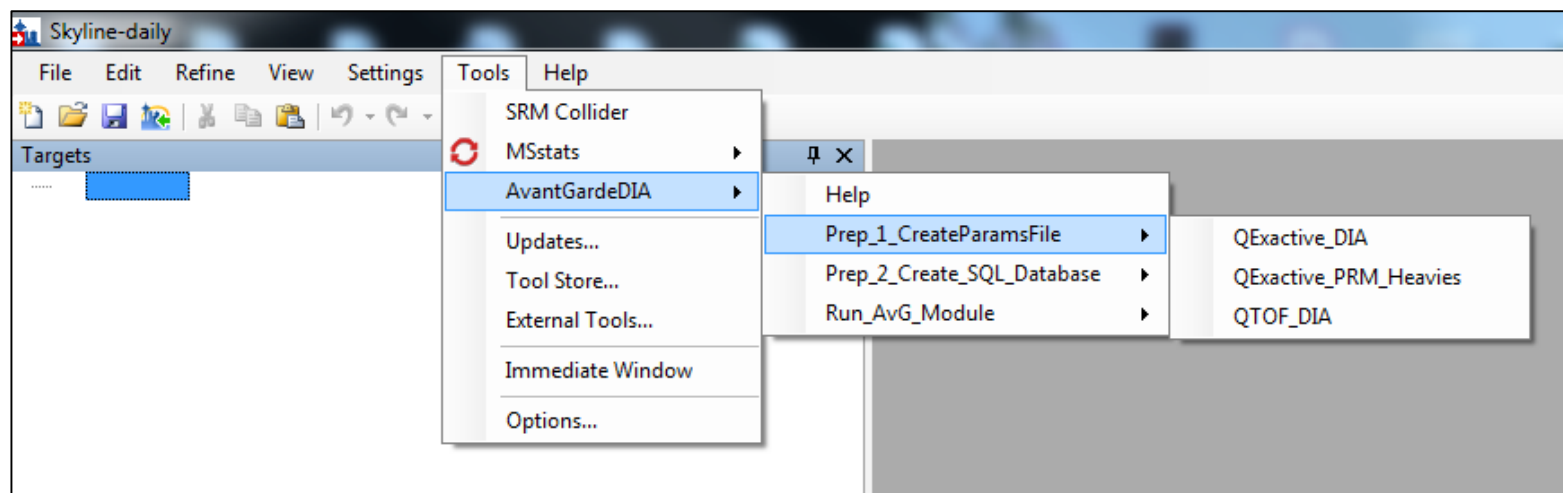
Help! I need somebody. Help! not just anybody...



Opens Avant-garde's GitHub repo on your browser:

- *Tutorial*
- *Latest updates*
- *Support*
- *Demo files*

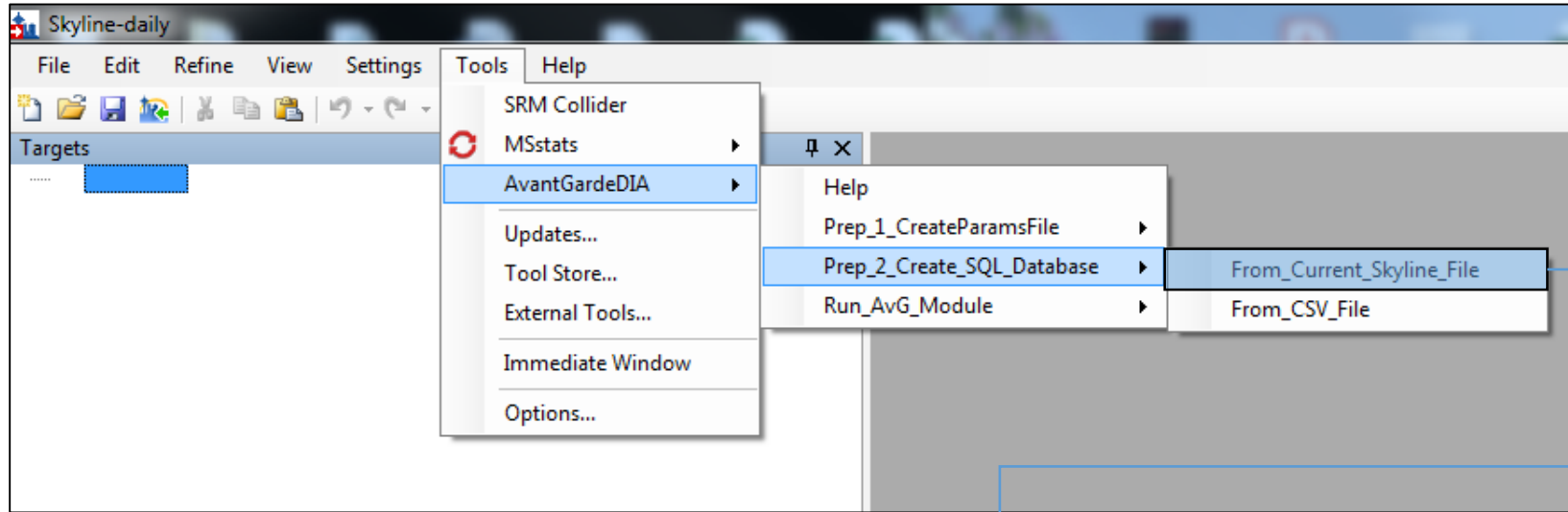
Preparation Step 1: Create parameters file



This step creates:

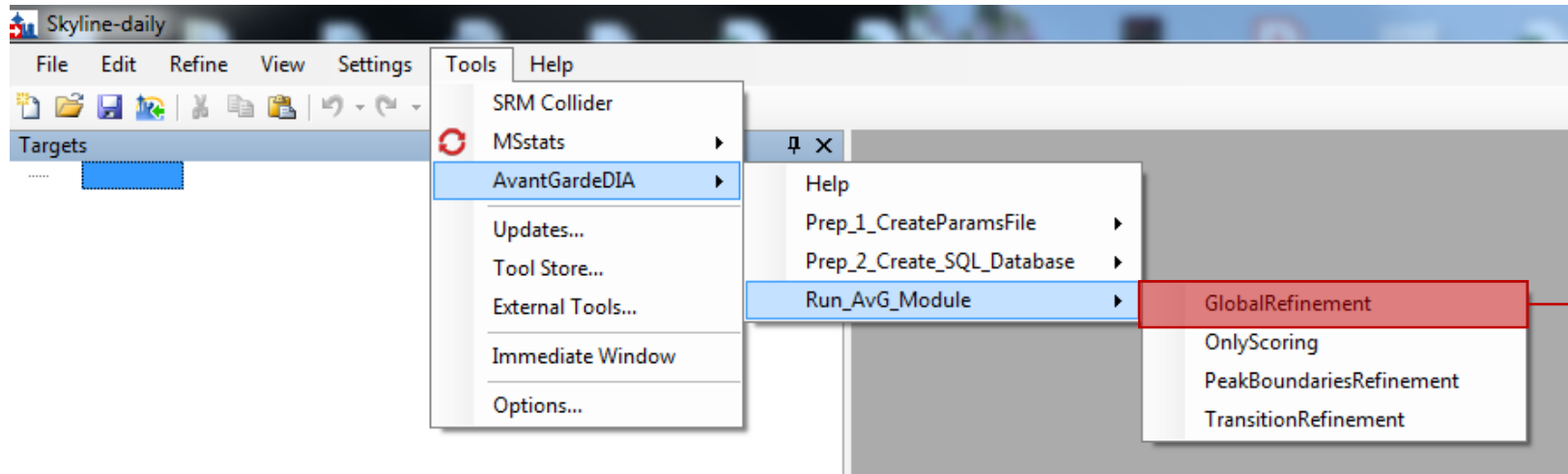
- Parameters file (.R file) with correct format
- Folders for outputted intermediary data and results
- Verifies that the R package is up-to-date with the External tool. If not, the R package is update.

Preparation Step 2



- Exports CSV file into a temporary folder
- Transforms CSV file into a SQLite file

Run *Avant-garde* DIA



Avant-Garde DIA modules:

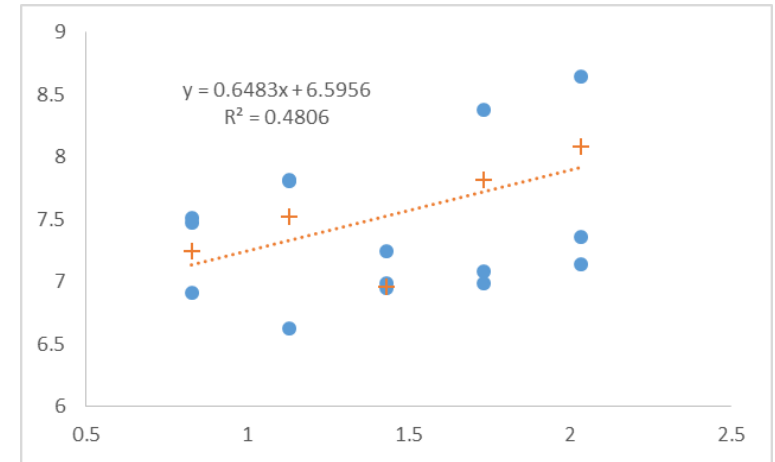
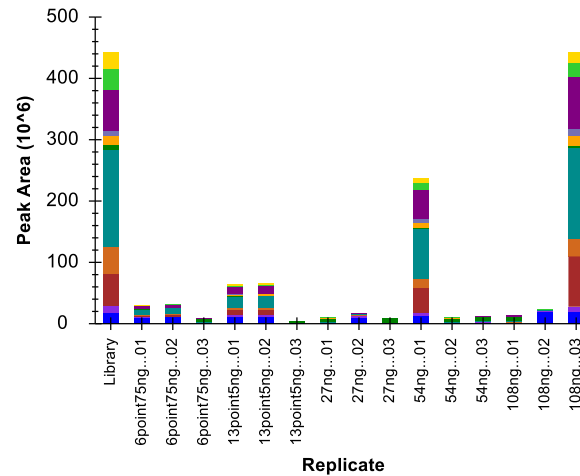
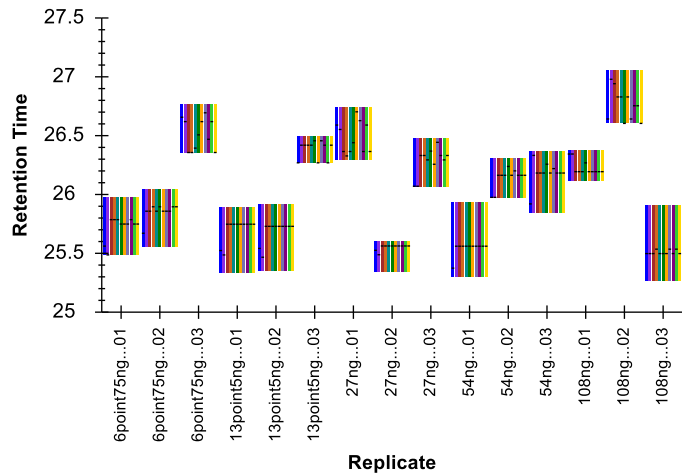
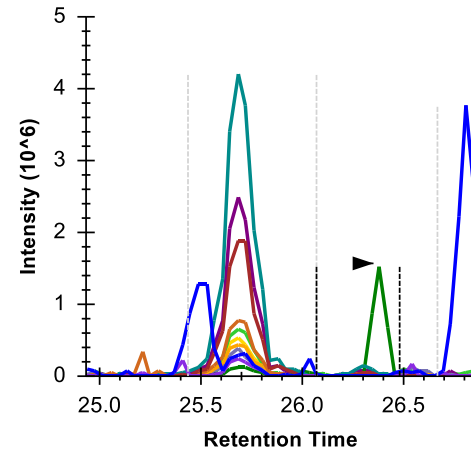
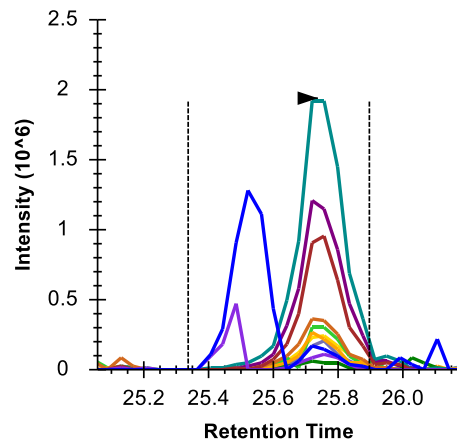
'GlobalRefinement' for

- 1) transition refinement
- 2) peak boundaries refinement
- 3) peak rescoring

A real example: worst case scenario

Demo:
5 point calibration curve
Analyzed in triplicate

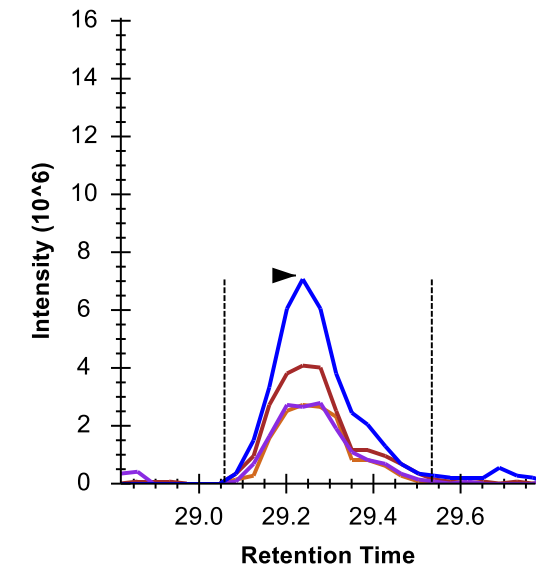
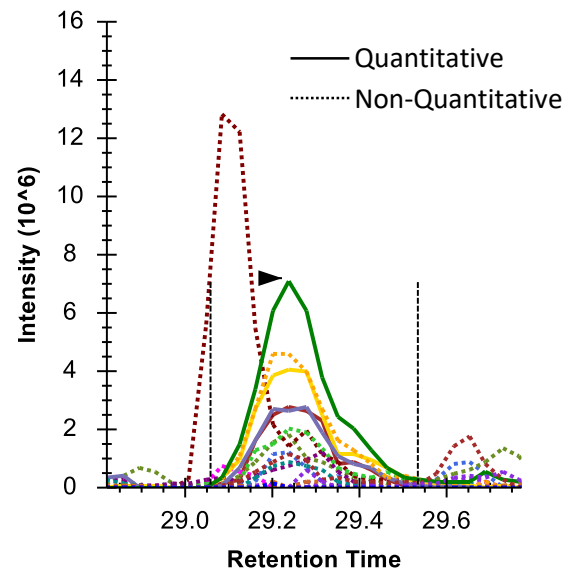
ANASPQKPLDLK



Skyline annotations: Import external information into Skyline

- Avant-garde produces a report compatible with Skyline's annotations
- Adjust peak boundaries
- Select transitions
- It will complete the "quantitative" annotation for each transition in the file.

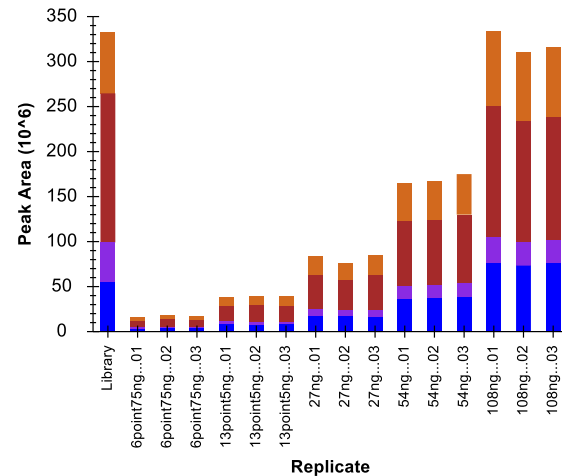
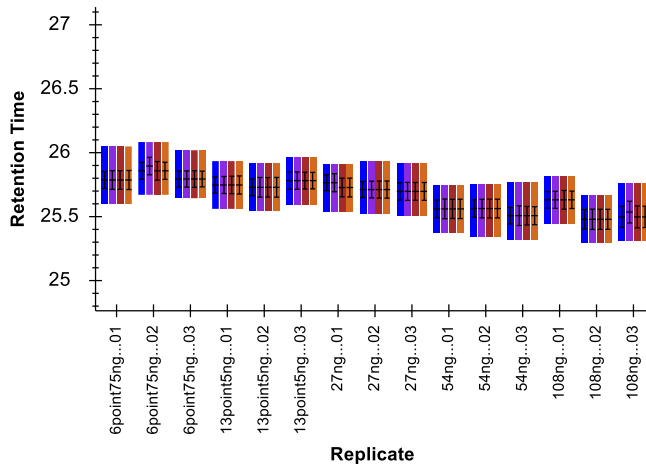
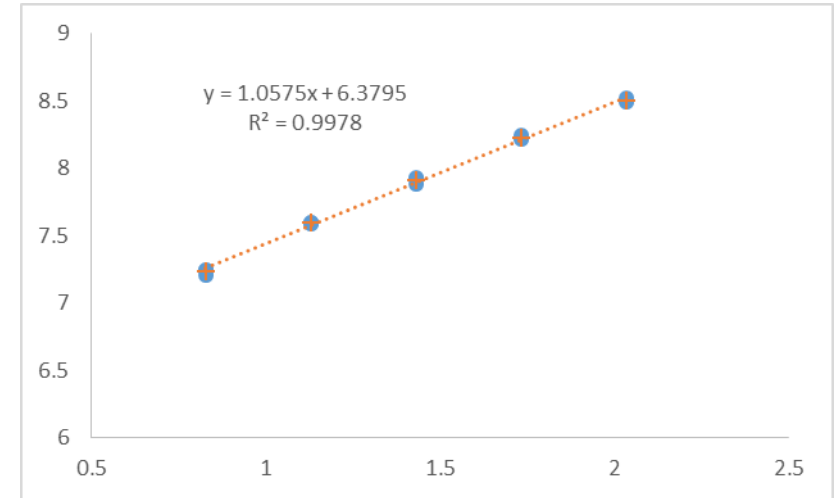
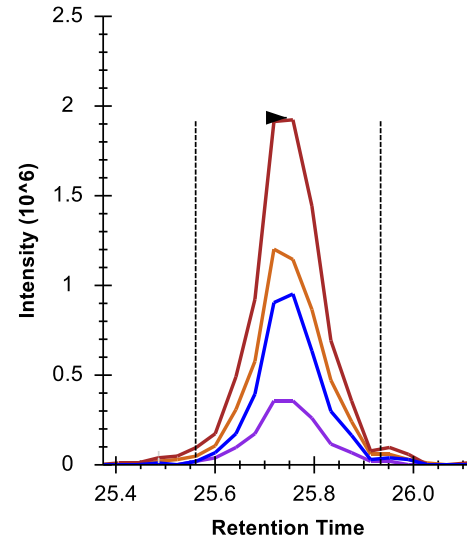
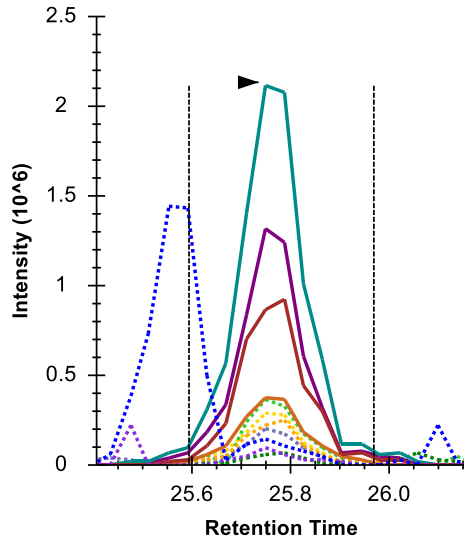
Modified Sequence	Precursor Mz	Quantitative	Product Mz
KPNIFYSGPAS[+80]PARPR	613.308482	<input checked="" type="checkbox"/>	806.423156
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	806.885222
KPNIFYSGPAS[+80]PARPR	613.308482	<input checked="" type="checkbox"/>	757.896774
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	749.863758
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	700.87531
KPNIFYSGPAS[+80]PARPR	613.308482	<input checked="" type="checkbox"/>	644.333278
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	619.787519
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	570.799071
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	489.267407
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	445.751393
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	417.702727
KPNIFYSGPAS[+80]PARPR	613.308482	<input type="checkbox"/>	333.195722
KPNIFYSGPAS[+80]PARPR	613.308482	<input checked="" type="checkbox"/>	298.68499



A real example

Demo:
5 point calibration curve
Analyzed in triplicate

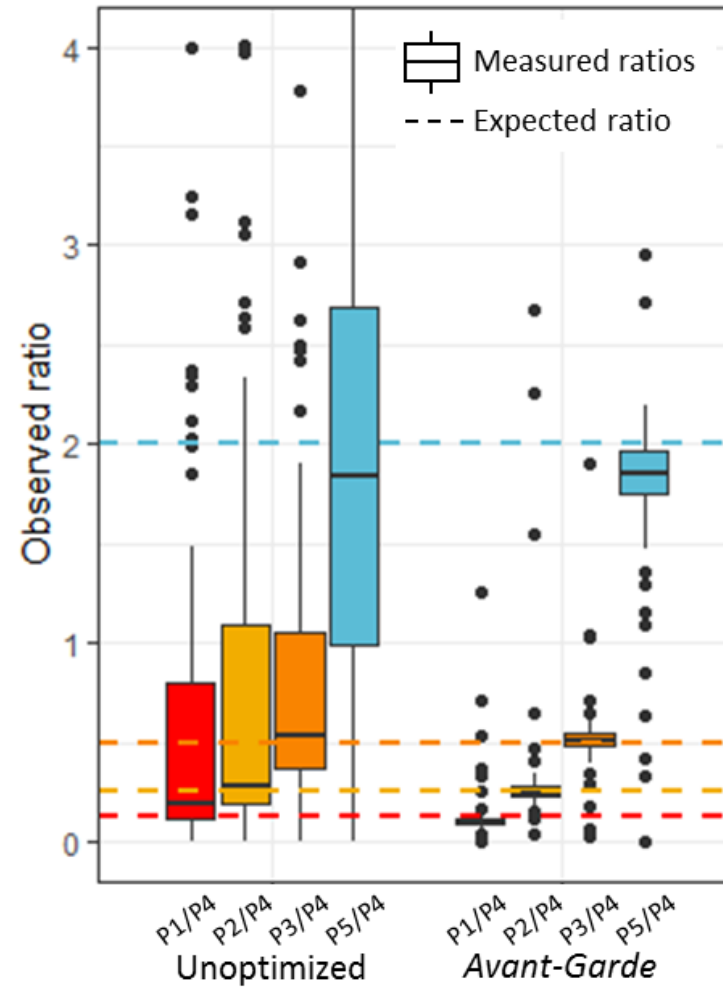
ANASPQKPLDLK



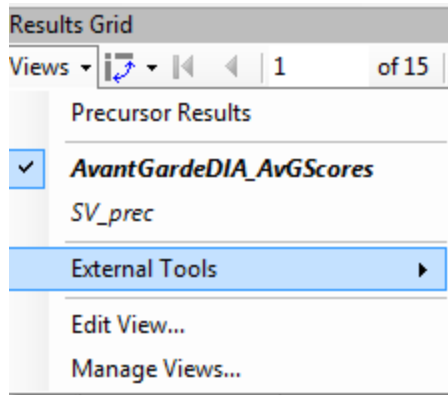
Improved selectivity,
accuracy and reproducibility

A real example

Demo:
5 point calibration curve
Analyzed in triplicate
96 peptides
15 runs



Skyline annotations: Import and view Scores directly in Skyline



RBM17_CL16

SPTGPSNSFLANMGGTVAHK

687.6499+++ (heavy) (dotp 0.92)

Select a precursor on the targeted peptide tree to see the calculated scores.

Precursor Result Locator	Replicate	AvG_Score	AvG_SpectralLibSir	AvG_Similarity_Sco	AvG_MPRA_Score	AvG_MassE	Average Mass Error PPM
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_108ng_Overlap22_01	0.973826588400735	0.999124920628...	0.997627282185...	0.999970305275...	1	5.5
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_108ng_Overlap22_02	0.92286180614668	0.999041486277...	0.992070941463...	0.999331980342...	1	4.9
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_108ng_Overlap22_03	0.9475340615793	0.999021205110...	0.994805563255...	0.999980200567...	1	4.6
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_13point5ng_Overlap22_01	0.635265691601597	0.998903048758...	0.989650638335...	0.999898763349...	0.8693961...	5.7
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_13point5ng_Overlap22_02	0.710025154389462	0.997922415141...	0.974696695527...	0.998547401055...	0.9650670...	10.3
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_13point5ng_Overlap22_03	0.441896240319349	0.998508140442...	0.950892411610...	0.999210440439...	0.8759202...	9.9
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_27ng_Overlap22_01	0.768268029797405	0.998149505228...	0.973509190064...	0.999549029703...	1	8.1
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_27ng_Overlap22_02	0.946890250746204	0.998739686489...	0.994889194866...	0.999560193472...	1	6.1
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_27ng_Overlap22_03	0.920693907755081	0.998979258159...	0.991829284277...	0.999817394096...	1	7.1
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_54ng_Overlap22_01	0.901455990426699	0.998740295351...	0.989754548869...	0.999524134098...	1	6.2
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_54ng_Overlap22_02	0.952769298085154	0.998941664307...	0.995420274480...	0.999979743973...	1	5.6
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_54ng_Overlap22_03	0.854204204480794	0.999180224809...	0.983947789882...	0.999677463215...	1	6
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_6point75ng_Overlap22_01	0.498291014697295	0.998094022166...	0.965195936983...	0.998308098648...	0.8691462...	5.8
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_6point75ng_Overlap22_02	0.205466314558495	0.998638012573...	0.960342348452...	0.997584941695...	0.6210872...	#N/A
PrecursorResult/...	CS20170831_SV_HEK_SpikeP100_6point75ng_Overlap22_03	0.30636326486527	0.997885638078...	0.963523636604...	0.995986285301...	0.7208113...	9.2

Explore and search external data directly within Skyline

Conclusion

- Developed an automated data curation tool to refine DIA (and PRM) results by removing interfered transitions, adjusting integration boundaries and scoring peaks to control the FDR
- Avant-garde's ensemble-driven scoring strategy is designed to produce very conservative results by penalizing poor-quality signals and enables to achieve the archetype of the ideal DIA signal
- Application of Avant-garde improves selectivity, accuracy, and reproducibility of quantitative DIA proteomics data



[github@SebVaca](https://github.com/SebVaca)



Thank you!

Broad Institute – Jaffe Lab and Proteomics Platform



Karen Christianson
Ryan Peckner
Karsten Krug
Kat DeRuff
Shawn Egri
Deborah Dele-Oni
Malvina Papanastasiou
Steve Carr
Jake Jaffe

University of Washington – MacCoss Lab

Nick Shulman
Brendan MacLean
Brian Searle
Vagisha Sharma
Mike MacCoss



github@SebVaca

