Integration of mProphet chromatogram peak identification probability model into Skyline

Brendan MacLean[1], Don M. Marsh[1], Hannes Röst[2], Lucia Espona Pernas[2], Olga Schubert[2], George Rosenberger[2], Ruedi Aebersold[2], Michael J. MacCoss[1]

[1]University of Washington, Department of Genome Sciences, [2]Institute of Molecular Systems Biology, ETH Zurich

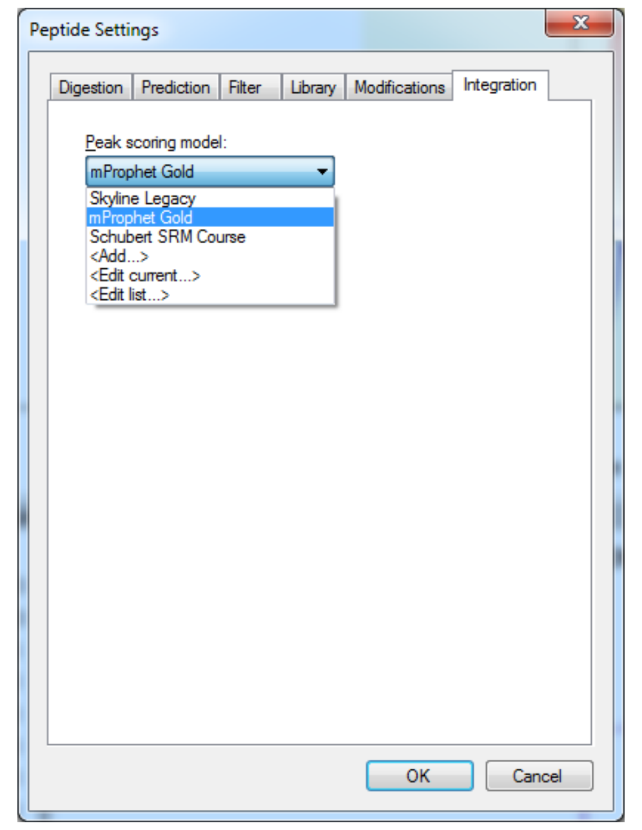
http://proteome.gs.washington.edu/software/skyline

Overview:

The Skyline Targeted Proteomics Environment has distinguished itself as a reliable and useful tool for chromatography-based quantitative proteomics. From its initial focus on selected reaction monitoring (SRM) to its current support for full-scan methods including MS1 filtering, targeted MS/MS and data independent acquisition (DIA – including the approach popularized as SWATH) measurements of peptide abundance have been based on areas under peaks in mass spectrometric chromatograms. Until now, however, peak identification within Skyline has relied on a limited set of features without the ability to derive a statistical confidence metric like a false discovery rate (FDR). To correct this shortcoming, we have integrated the mProphet scoring features, semi-supervised learning (SSL) algorithm and derivation of peak q values into Skyline..

Introduction:

Skyline now supports training an arbitrary number of chromatogram peak scoring models based on acquired data, using the mProphet SSL algorithm. These models consist of a set of coefficients, applied to available peak feature scores to derive an mProphet score, and a mean and standard deviation that define a normal distribution estimate of the null distribution for the mProphet scores. Different models may be trained for different experimental conditions. Once a model has been trained, it may be specified in Skyline and used for any number of experiments under the same conditions, without retraining under SSL. The null distribution parameters allow p value estimates, from which Skyline can estimate q values and FDR using the algorithm described by Storey, et al.



Peak scoring model: mProphet Gold
- Skyline Legacy
- mProphet Gold
- Schubert SRM Course
- <Add...>
- <Edit current...>
- <Edit list...>

Figure 1: Peptide settings form showing a choice of trained mProphet scoring models as well as the legacy Skyline peak scoring model.

Methods:

In order to achieve this, we have implemented the following:

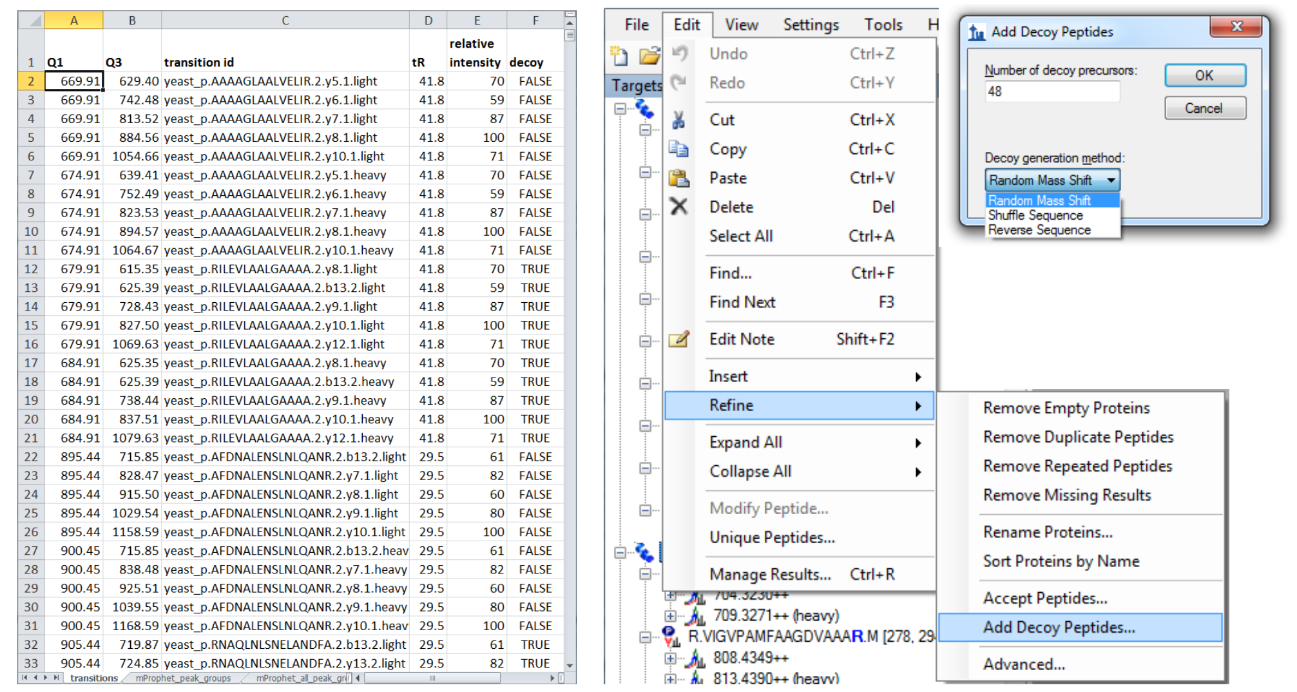
- Decoy peptide and transition generation
- Decoy peptide and transition import
- A calculator architecture allowing two calculator types
  1. Summary – requiring only summary peak attributes
  2. Detailed – requiring chromatogram points
- Calculating and storing detail scores during import
- Calculators for the original 8 mProphet scores
  - ✓ Intensity[1]
  - ✓ Intensity correlation (dot-product) [1]
  - ✓ Coelution[2]
  - ✓ Shape[2]
  - ✓ Reference correlation (dot-product) [1]
  - ✓ Retention time deviation[1]
  - ✓ Reference coelution[2]
  - ✓ Reference shape[2]
- Calculators for several new scores
  - ✓ Coelution count[2]
  - ✓ Reference coelution count[2]
- Export mProphet features for testing with mProphet in R
- The mProphet SSL algorithm converted from R to C#
- Storey-Tibshirani q value calculation algorithm

In all cases, the resulting code was tested against data and results from original publications. Unit tests from OpenSWATH were used to validate some of the scores.

We have created two test Skyline documents:

1. The mProphet gold standard data set
2. A new document with decoy transitions generated in Skyline

Both have had SRM data imported and the new mProphet model editor used to train score coefficients for a composite linear scoring function.



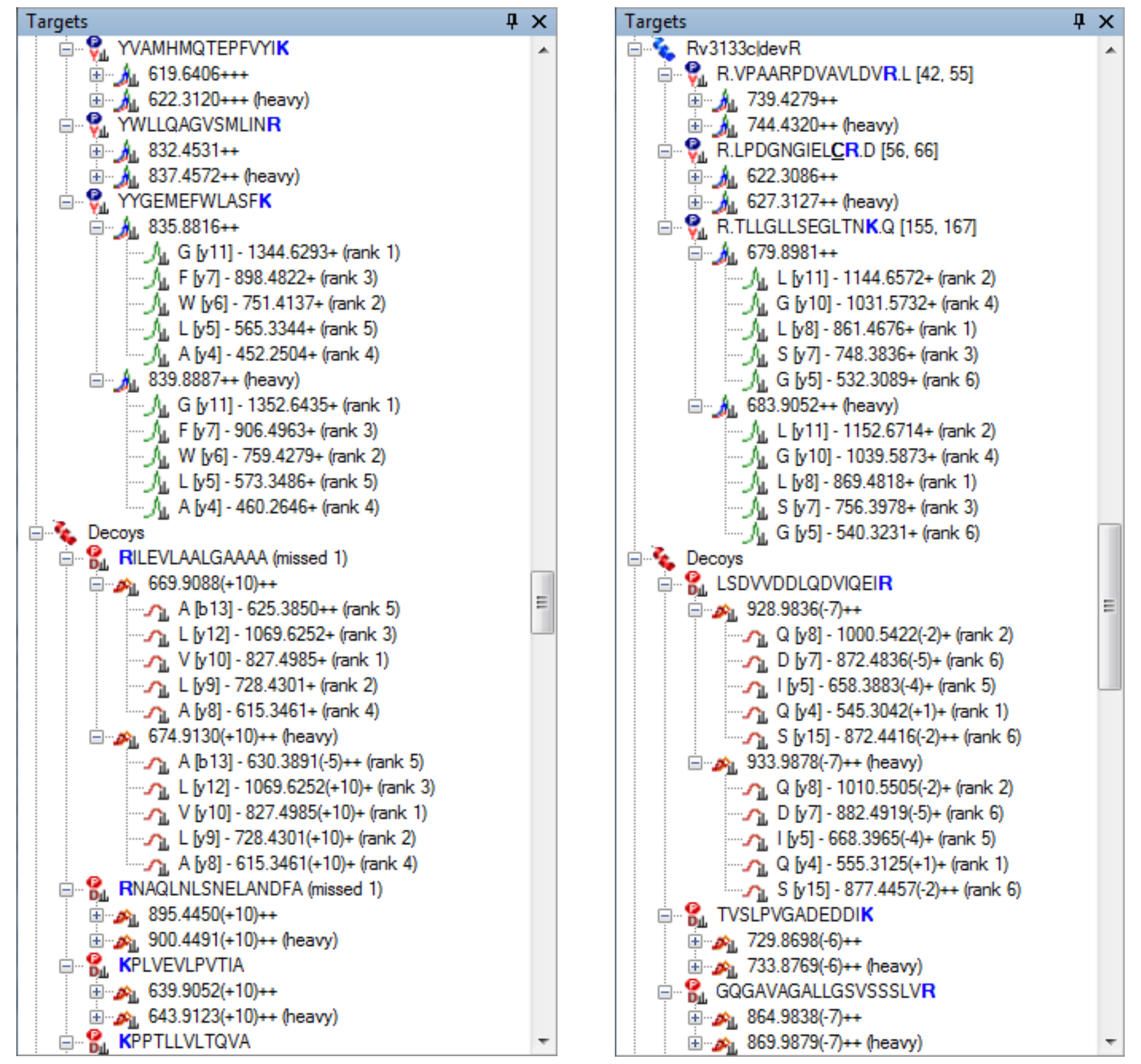
	A	B	C	D	E	F
1	Q1	Q3	transition id	tR	relative intensity	decoy
2	669.91	629.40	yeast_p.AAAAGLAALVELIR.2.y5.1.light	41.8	70	FALSE
3	669.91	742.48	yeast_p.AAAAGLAALVELIR.2.y6.1.light	41.8	59	FALSE
4	669.91	813.52	yeast_p.AAAAGLAALVELIR.2.y7.1.light	41.8	87	FALSE
5	669.91	884.56	yeast_p.AAAAGLAALVELIR.2.y8.1.light	41.8	100	FALSE
6	669.91	1054.66	yeast_p.AAAAGLAALVELIR.2.y10.1.light	41.8	71	FALSE
7	674.91	639.41	yeast_p.AAAAGLAALVELIR.2.y5.1.heavy	41.8	70	FALSE
8	674.91	752.49	yeast_p.AAAAGLAALVELIR.2.y6.1.heavy	41.8	59	FALSE
9	674.91	823.53	yeast_p.AAAAGLAALVELIR.2.y7.1.heavy	41.8	87	FALSE
10	674.91	894.57	yeast_p.AAAAGLAALVELIR.2.y8.1.heavy	41.8	100	FALSE
11	674.91	1064.67	yeast_p.AAAAGLAALVELIR.2.y10.1.heavy	41.8	71	FALSE
12	679.91	615.35	yeast_p.RILEVLAALGAAAA.2.y8.1.light	41.8	70	TRUE
13	679.91	625.39	yeast_p.RILEVLAALGAAAA.2.b13.2.light	41.8	59	TRUE
14	679.91	728.43	yeast_p.RILEVLAALGAAAA.2.y9.1.light	41.8	87	TRUE
15	679.91	827.50	yeast_p.RILEVLAALGAAAA.2.y10.1.light	41.8	100	TRUE
16	679.91	1069.63	yeast_p.RILEVLAALGAAAA.2.y12.1.light	41.8	71	TRUE
17	684.91	625.35	yeast_p.RILEVLAALGAAAA.2.y8.1.heavy	41.8	70	TRUE
18	684.91	625.39	yeast_p.RILEVLAALGAAAA.2.b13.2.heavy	41.8	59	TRUE
19	684.91	738.44	yeast_p.RILEVLAALGAAAA.2.y9.1.heavy	41.8	87	TRUE
20	684.91	837.51	yeast_p.RILEVLAALGAAAA.2.y10.1.heavy	41.8	100	TRUE
21	684.91	1079.63	yeast_p.RILEVLAALGAAAA.2.y12.1.heavy	41.8	71	TRUE
22	895.44	715.85	yeast_p.AFDNALENSLNLQANR.2.b13.2.light	29.5	61	FALSE
23	895.44	828.47	yeast_p.AFDNALENSLNLQANR.2.y7.1.light	29.5	82	FALSE
24	895.44	915.50	yeast_p.AFDNALENSLNLQANR.2.y8.1.light	29.5	60	FALSE
25	895.44	1029.54	yeast_p.AFDNALENSLNLQANR.2.y9.1.light	29.5	80	FALSE
26	895.44	1158.59	yeast_p.AFDNALENSLNLQANR.2.y10.1.light	29.5	100	FALSE
27	900.45	715.85	yeast_p.AFDNALENSLNLQANR.2.b13.2.heav	29.5	61	FALSE
28	900.45	838.48	yeast_p.AFDNALENSLNLQANR.2.y7.1.heavy	29.5	82	FALSE
29	900.45	925.51	yeast_p.AFDNALENSLNLQANR.2.y8.1.heavy	29.5	60	FALSE
30	900.45	1039.55	yeast_p.AFDNALENSLNLQANR.2.y9.1.heavy	29.5	80	FALSE
31	900.45	1168.59	yeast_p.AFDNALENSLNLQANR.2.y10.1.heav	29.5	100	FALSE
32	905.44	719.87	yeast_p.RNAQLNLSNELANDFA.2.b13.2.light	29.5	61	TRUE
33	905.44	724.85	yeast_p.RNAQLNLSNELANDFA.2.y13.2.light	29.5	82	TRUE

Number of decoy precursors: 48

Decoy generation method: Random Mass Shift

Figure 2a & b: Original mProphet gold standard transition list from supplementary information that was imported into Skyline (left) and decoy generation form used to create decoy transitions in new a document.

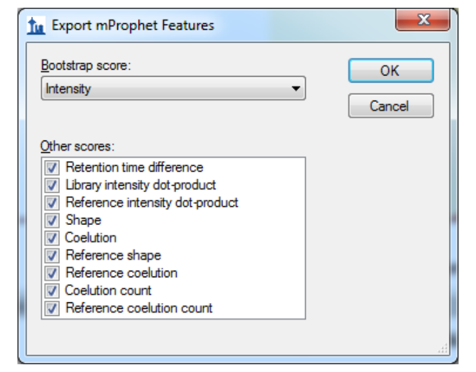
Results:



```
YYGEMEFWLASFK
  835.8816++
    G [y11] - 1344.6293+ (rank 1)
    F [y7] - 898.4822+ (rank 3)
    W [y6] - 751.4137+ (rank 2)
    L [y5] - 565.3344+ (rank 5)
    A [y4] - 452.2504+ (rank 4)
  839.8887++ (heavy)
    G [y11] - 1352.6435+ (rank 1)
    F [y7] - 906.4963+ (rank 3)
    W [y6] - 759.4279+ (rank 2)
    L [y5] - 573.3486+ (rank 5)
    A [y4] - 460.2646+ (rank 4)
Decoys
  RILEVLAALGAAAA (missed 1)
    669.9088(+10)++
      A [b13] - 625.3850++ (rank 5)
      L [y12] - 1069.6252+ (rank 3)
      V [y10] - 827.4985+ (rank 1)
      L [y9] - 728.4301+ (rank 2)
      A [y8] - 615.3461+ (rank 4)
    674.9130(+10)++ (heavy)
      A [b13] - 630.3891(-5)++ (rank 5)
      L [y12] - 1069.6252(+10)+ (rank 3)
      V [y10] - 827.4985(+10)+ (rank 1)
      L [y9] - 728.4301(+10)+ (rank 2)
      A [y8] - 615.3461(+10)+ (rank 4)
  RNAQLNLSNELANDFA (missed 1)
    895.4450(+10)++
    900.4491(+10)++ (heavy)
  KPLVEVLPVTIA
    639.9052(+10)++
    643.9123(+10)++ (heavy)
  KPPTLLVLTQVA
```

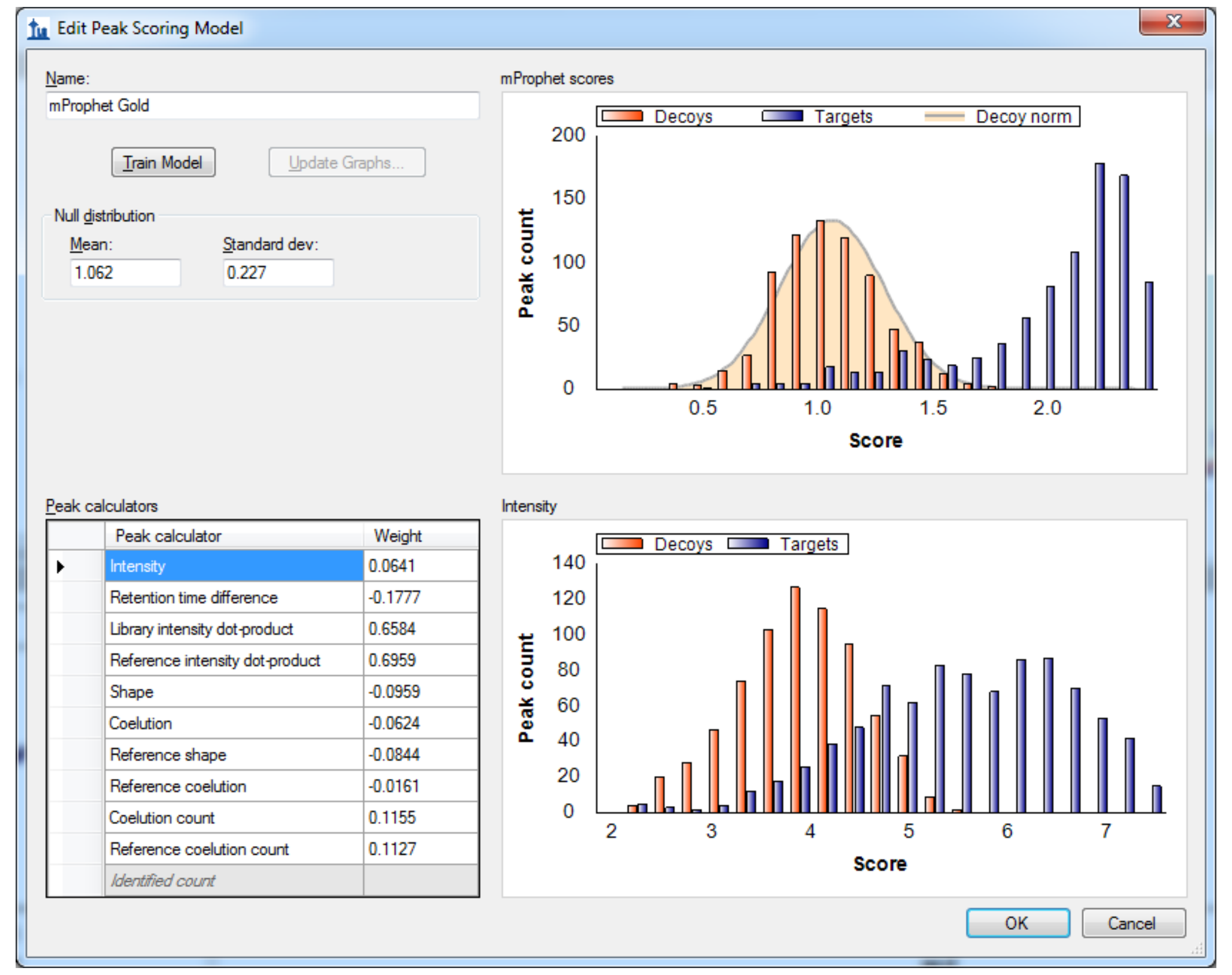
```
    627.3127++ (heavy)
R.TLLGLLSEGLTNK.Q [155, 167]
  679.8981++
    L [y11] - 1144.6572+ (rank 2)
    G [y10] - 1031.5732+ (rank 4)
    L [y8] - 861.4676+ (rank 1)
    S [y7] - 748.3836+ (rank 3)
    G [y5] - 532.3089+ (rank 6)
  683.9052++ (heavy)
    L [y11] - 1152.6714+ (rank 2)
    G [y10] - 1039.5873+ (rank 4)
    L [y8] - 869.4818+ (rank 1)
    S [y7] - 756.3978+ (rank 3)
    G [y5] - 540.3231+ (rank 6)
Decoys
  LSDVVDDLQDVIQEIR
    928.9836(-7)++
      Q [y8] - 1000.5422(-2)+ (rank 2)
      D [y7] - 872.4836(-5)+ (rank 6)
      I [y5] - 658.3883(-4)+ (rank 5)
      Q [y4] - 545.3042(+1)+ (rank 1)
      S [y15] - 872.4416(-2)++ (rank 6)
    933.9878(-7)++ (heavy)
      Q [y8] - 1010.5505(-2)+ (rank 2)
      D [y7] - 882.4919(-5)+ (rank 6)
      I [y5] - 668.3965(-4)+ (rank 5)
      Q [y4] - 555.3125(+1)+ (rank 1)
      S [y15] - 877.4457(-2)++ (rank 6)
  TVSLPVGADEDDIK
    729.8698(-6)++
    733.8769(-6)++ (heavy)
  GQGAVAGALLGSVSSSLVR
    864.9838(-7)++
    869.9879(-7)++ (heavy)
```

Figure 3a & b: Target and decoy peptides, precursors, and transitions from the original mProphet gold method (left), where transitions were imported from the manuscript transition list, and a newly created method, where decoys were generated in Skyline (right).



- ☑ Reference intensity dot-product
- ☑ Shape
- ☑ Coelution
- ☑ Reference shape
- ☑ Reference coelution
- ☑ Coelution count
- ☑ Reference coelution count

Figure 5: Form for exporting peak groups and features to a CSV file for use with mProphet in R.



Name: mProphet Gold

Null distribution
Mean: 1.062
Standard dev: 0.227

Peak calculators

Peak calculator	Weight
Intensity	0.0641
Retention time difference	-0.1777
Library intensity dot-product	0.6584
Reference intensity dot-product	0.6959
Shape	-0.0959
Coelution	-0.0624
Reference shape	-0.0844
Reference coelution	-0.0161
Coelution count	0.1155
Reference coelution count	0.1127
Identified count	

Figure 6: Skyline scoring model training form, showing a fully mProphet model (top) with decoy and target histograms and the estimated decoy normal distribution. The trained linear model coefficients are shown (bottom-left) as well as the single score histograms for the bootstrap Intensity score (bottom-right)

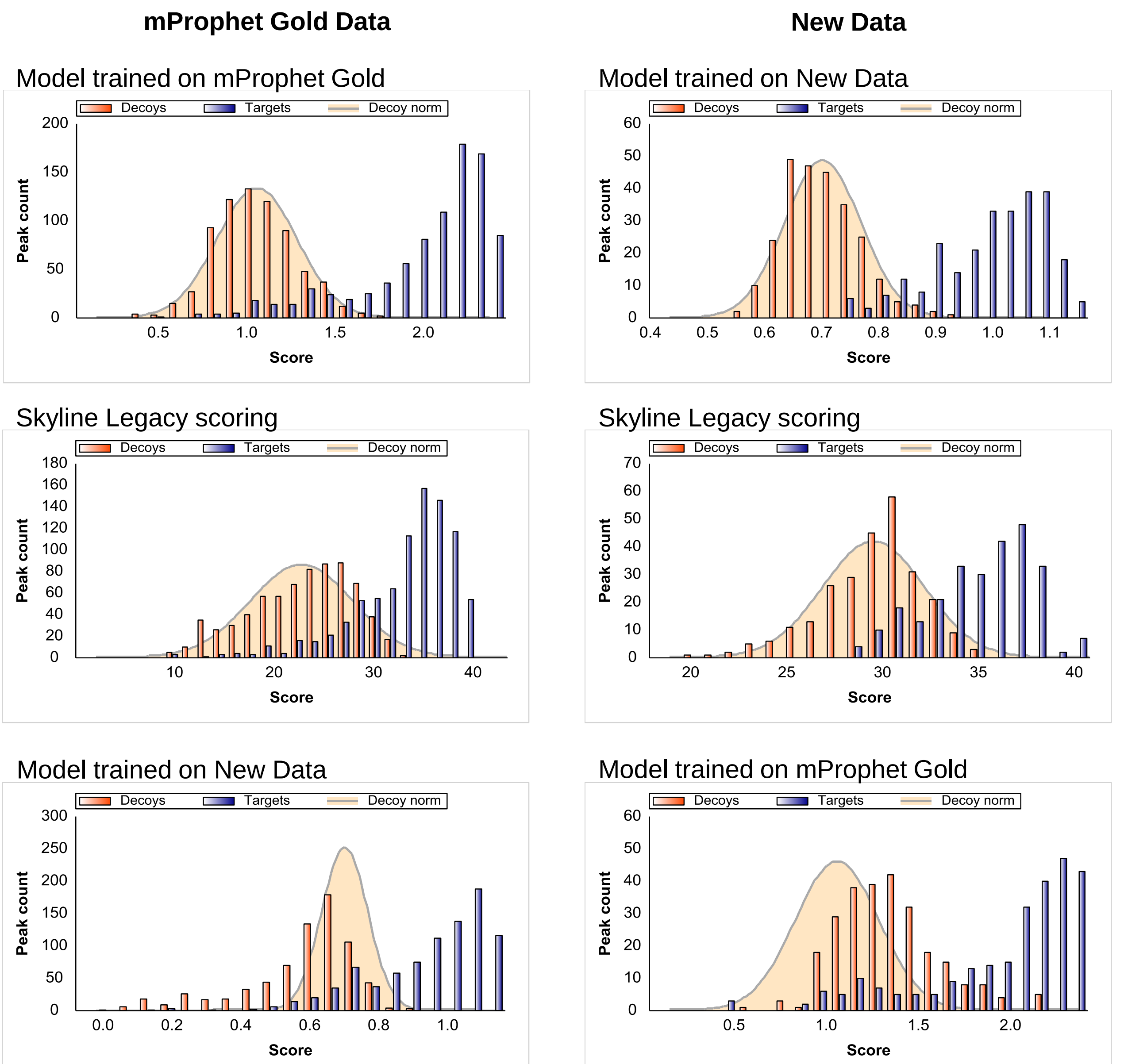


Figure 6: Target-decoy and modeled null distribution graphs copied from the Skyline scoring model training form show improvement in separation of decoy and target peaks from the legacy Skyline peak score, after training with mProphet. Good separation persists, when a model trained on one data set is used on another. Null distribution estimates do not transfer well in this case.

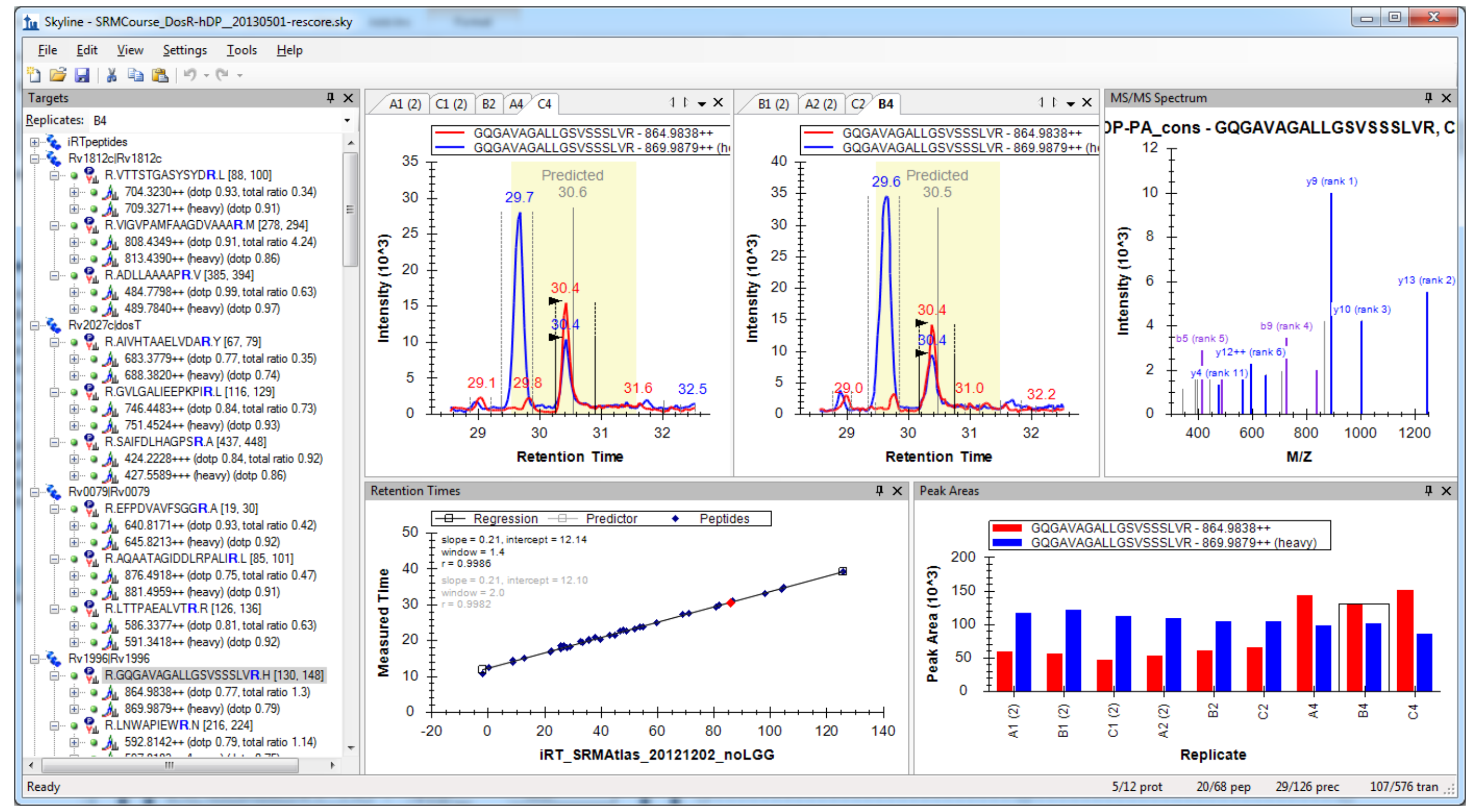


Figure 5: Nine runs imported normally into Skyline for the newly created document. Skyline plots show chromatograms with peaks picked and predicted retention times, matching MS/MS library spectrum, linear regression between measured retention times and iRT calculator scores and light and heavy precursor peak areas by run. The linear regression graph indicates that retention time prediction is unlikely to be a strong indicator of peak validity, since even with decoy peaks the R for the regression is 0.9986, probably because acquisition was scheduled in 4 minute windows around the predicted time.

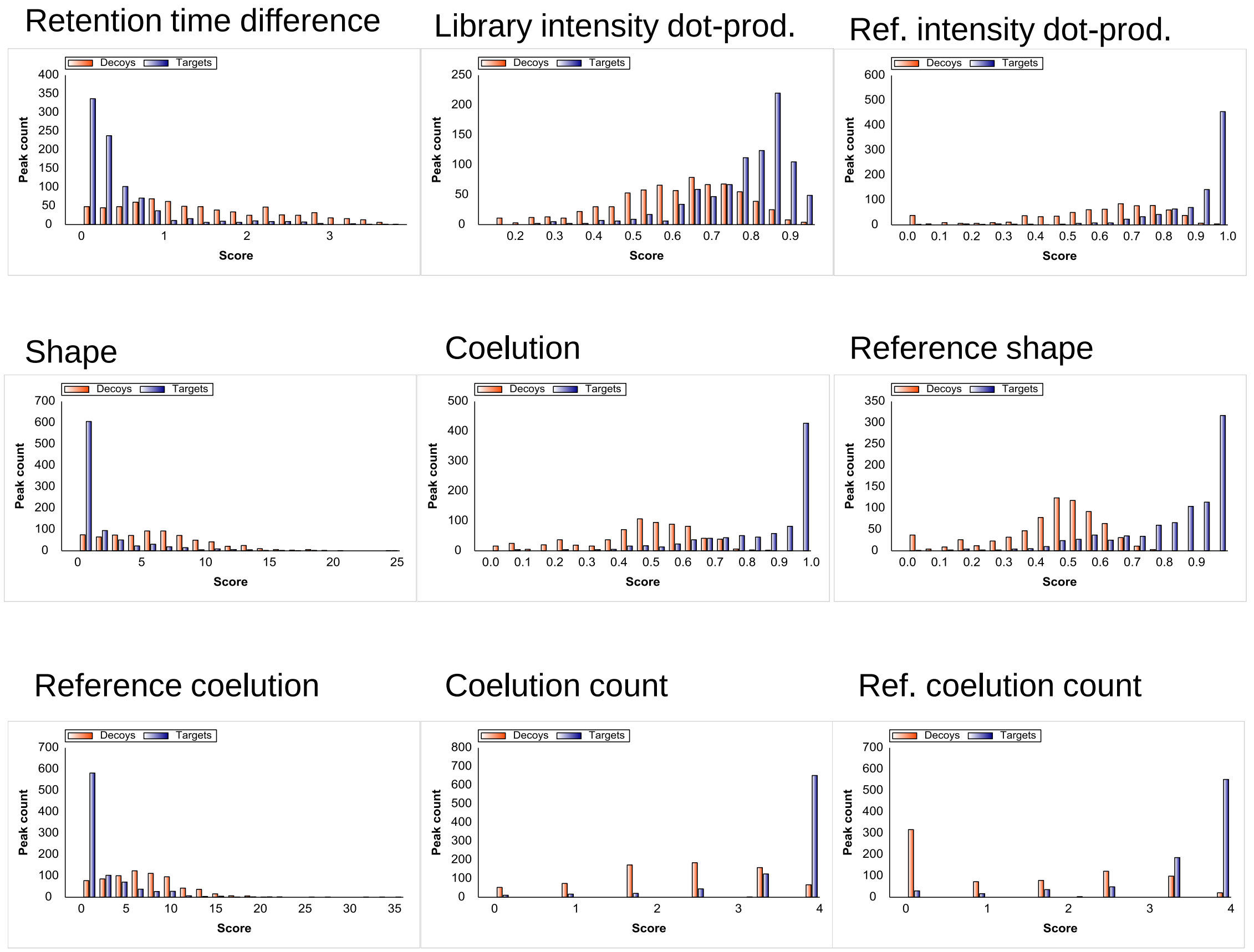


Figure 6: Target-decoy histograms for the remaining 9 scores for the fully trained model shown in Figure 5, copied from the model training form.

Conclusions:

- Good progress has been made on integrating the mProphet scoring model into Skyline in a way that makes important information for assessing model accuracy immediately available.
- Much work remains to be done to understand how much scoring separation will enhance peak picking in Skyline, and whether adding further scores can improve peak picking for non-SRM acquisition.
- Using calculated models to actually perform peak picking and q value estimation will require further implementation work in Skyline.
- Estimating accurate p and q values decoy distributions as a model for the null distribution may be an elusive goal.

References:

(1) Reiter L, Rinner O, et al. Nat. Methods. 2011/05; 8(5):430-5.
(2) Malmstrom L, Malmstrom J, et al. J. Proteome Res. 2012/03; 11(3):1644-53.