Supplement to "Empirical multi-dimensional space for scoring peptide spectrum matches in shotgun proteomics"

Mark V. Ivanov[1,2], Lev I. Levitsky[1,2], Anna A. Lobas[1,2], Tanja Panic[3], Ünige A. Laskay[4], Goran Mitulovic[3], Rainer Schmid[3], Marina L. Pridatchenko[1], Yury O. Tsybin[4], and Mikhail V. Gorshkov[1,2*]

[1]Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow 119334, Russia

[2]Moscow Institute of Physics and Technology (State University), Dolgoprudny 141700, Moscow region, Russia

[3]Medical University of Vienna, Vienna 1090, Austria

[4]Biomolecular Mass Spectrometry Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland

*Correspondence should be addressed to Dr. Mikhail V. Gorshkov, Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow 119334, Russia. E-mail: gorshkov@chph.ras.ru, mike.gorshkov@gmail.com

Tel/FAX: +7(499)137-8257

Supplementary Table S1. Number of "false" identifications (Self-Boosted Percolator) by protein length. Note that, contrary to the below results, peptides from proteins longer than 2000 residues constitute only ~10% of all peptides in the database. As stated in the manuscript, the additional PSMs reported by Self-Boosted Percolator tend to come from the longest proteins in the database.

Replicate	Number of "false" identifications from large proteins (>= 2000 a.a. residues)	Other "false" identifications
UPS2_1	46	1
UPS2_2	49	5
UPS2_3	4	3

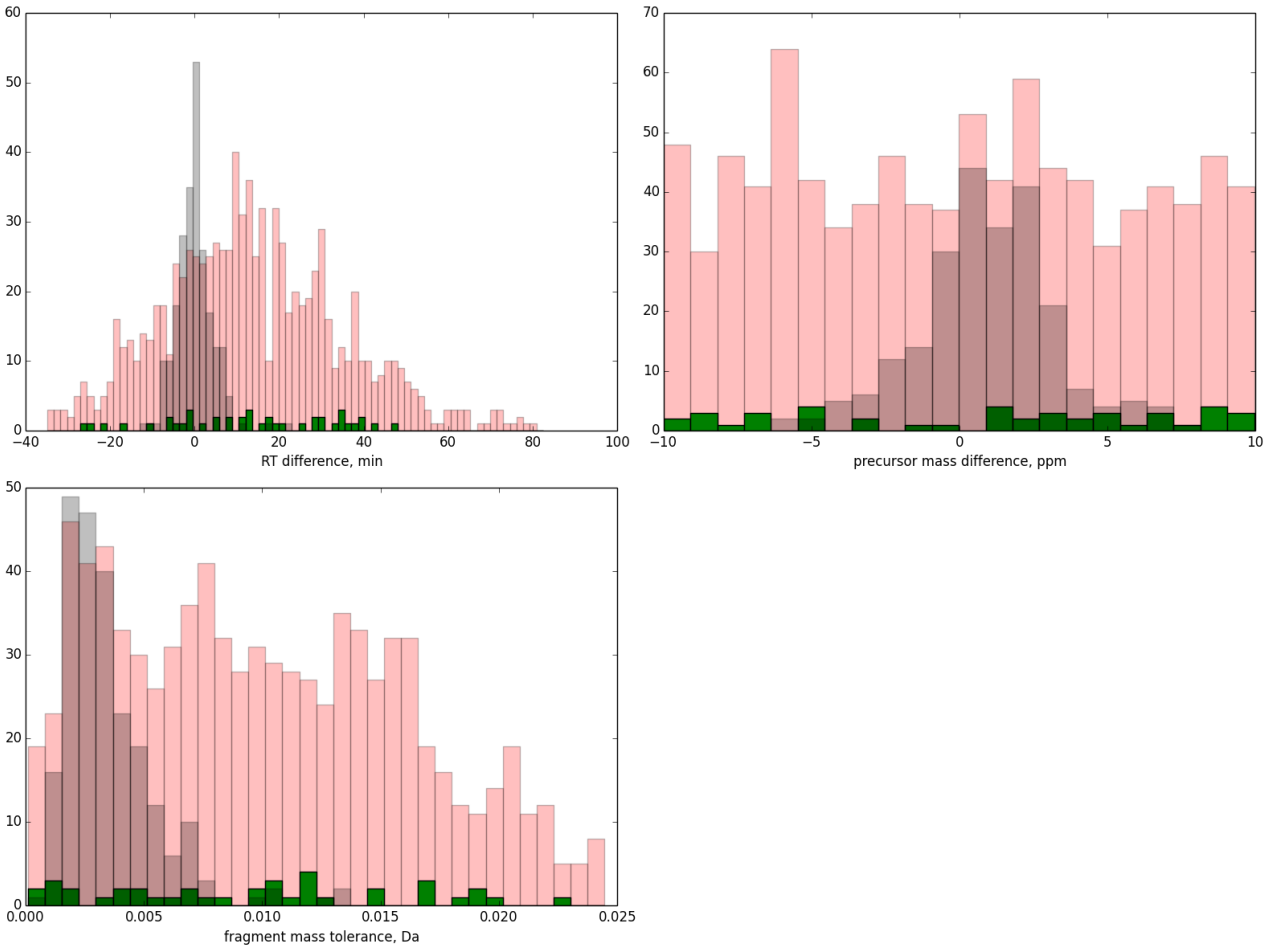
All 9 PSMs from the right column of Table 1 (i.e. "false" PSMs not from long proteins) come from a single protein for all three replicates. This protein (sp|P02042|HBD_HUMAN, Hemoglobin subunit delta) is similar to one of the proteins in UPS (sp|P68871|HBB_HUMAN, Hemoglobin subunit beta).

"False" PSMs reported by MP score and PeptideProphet are distributed as expected for false PSMs, i.e. proteins with length of >2000 residues account for ~10% of these PSMs (4 of 42 for MP score, 5 of 44 for PeptideProphet).

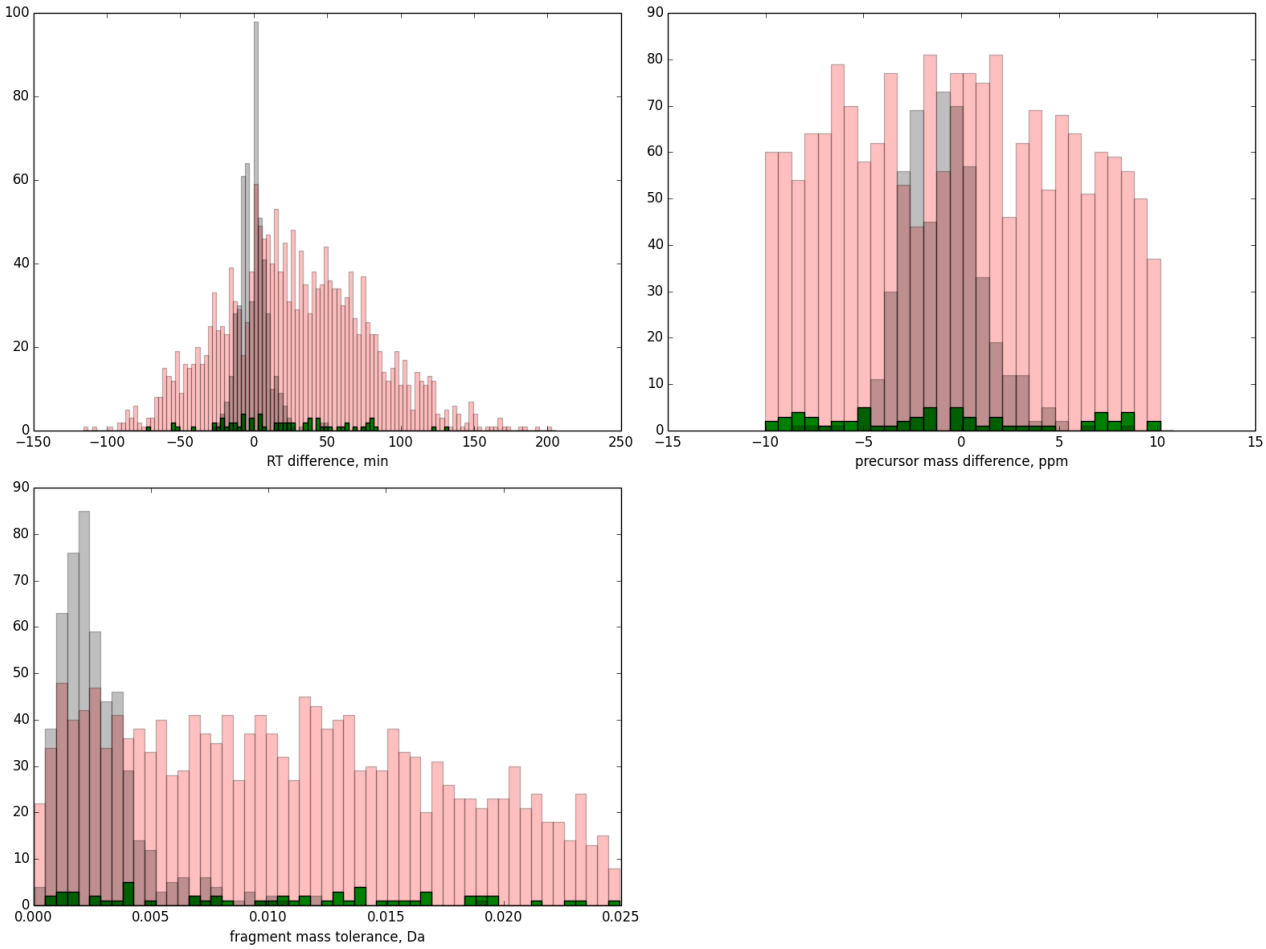
Supplementary Table S2. Protein identifications reported by Self-Boosted Percolator (3 replicates combined). About 90% of PSMs come from titin, the longest protein in the database. Note: the average number of theoretical tryptic peptides for a protein in the database used is around 80.

PSMs	Theoretical Peptides	Protein	Description
90	7070	sp|Q8WZ42|TITIN_HUMAN	Titin
3	1969	sp|Q8NF91|SYNE1_HUMAN	Nesprin-1
1	1531	sp|Q8WXH0|SYNE2_HUMAN	Nesprin-2
1	347	sp|Q9Y566|SHAN1_HUMAN	SH3 and multiple ankyrin repeat domains protein 1
1	1079	sp|Q8IZT6|ASPM_HUMAN	Abnormal spindle-like microcephaly-associated protein
1	1406	sp|Q5VST9|OBSCN_HUMAN	Obscurin
1	492	sp|P50851|LRBA_HUMAN	Lipopolysaccharide-responsive and beige-like anchor protein
1	1653	sp|Q8WXI7|MUC16_HUMAN	Mucin-16

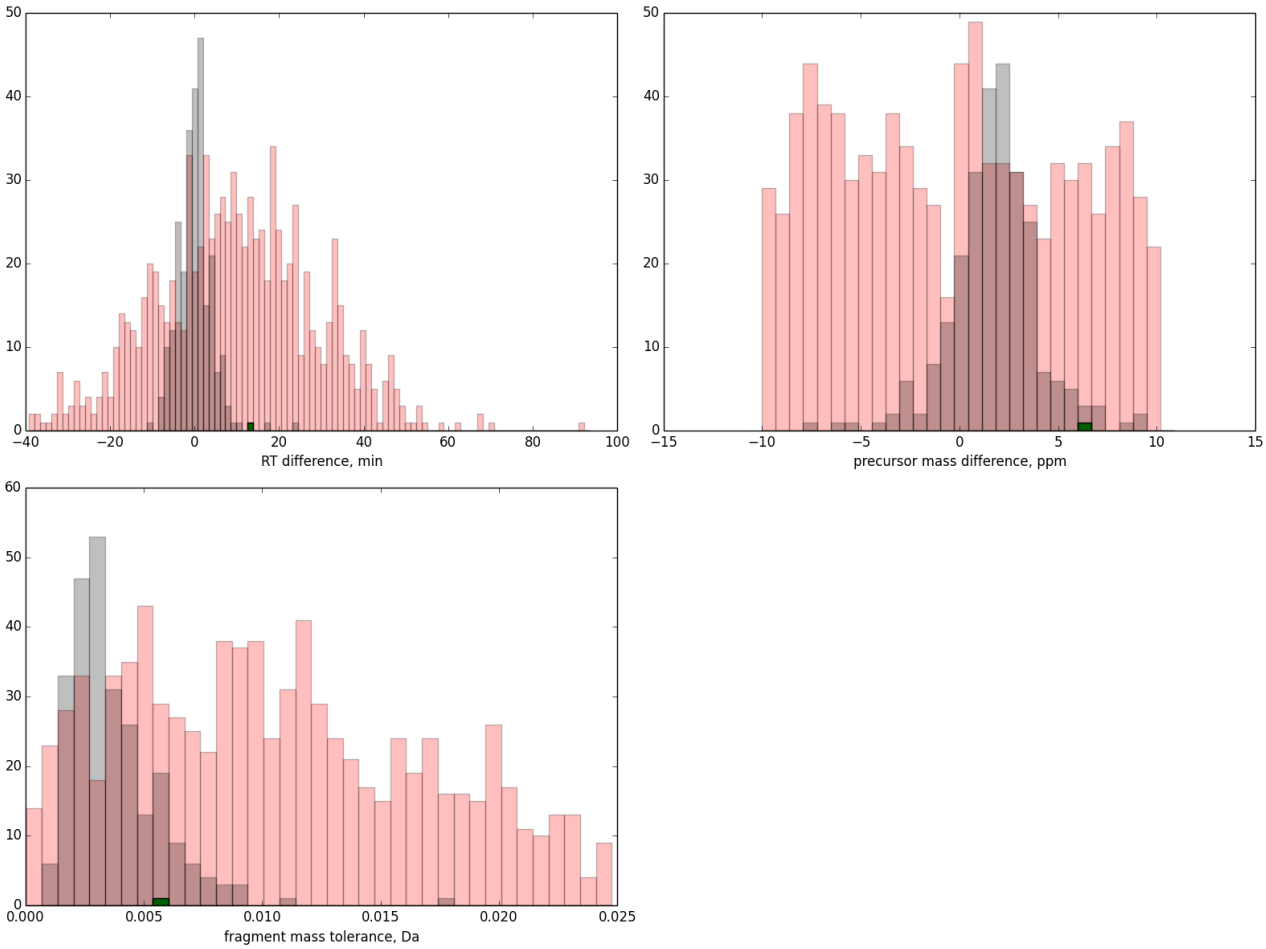
In the figures below we plot the distributions of several descriptors for titin matches for three replicates. As shown in the figures, titin PSMs do not follow the distributions of correct PSMs, but rather those of decoy matches. This observation strongly suggests that the titin PSMs recovered by Self-Boosted Percolator are mostly false and do not indicate superior sensitivity of Self-Boosted Percolator.



Supplementary Figure S1. Distribution of descriptors for UPS2_1 replicate. Bright green: titin PSMs. Bleak red: decoy PSMs. Bleak dark: top 1% FDR PSMs.



Supplementary Figure S2. Distribution of descriptors for UPS2_2 replicate. Bright green: titin PSMs. Bleak red: decoy PSMs. Bleak dark: top 1% FDR PSMs.



Supplementary Figure S3. Distribution of descriptors for UPS2_3 replicate (very few titin matches). Bright green: titin PSMs. Bleak red: decoy PSMs. Bleak dark: top 1% FDR PSMs.