

**Supporting information:**

**"Peptide retention time prediction in hydrophilic interaction liquid chromatography: data collection methods and features of additive and sequence-specific models"**

Oleg V. Krokhin<sup>1,2\*</sup>, Peyman Ezzati<sup>1</sup>, Vic Spicer<sup>1</sup>

<sup>1</sup>Manitoba Centre for Proteomics and Systems Biology, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada

<sup>2</sup>Department of Internal Medicine, University of Manitoba, 799 JBRC, 715 McDermot avenue, Winnipeg, Manitoba R3E 3P4, Canada

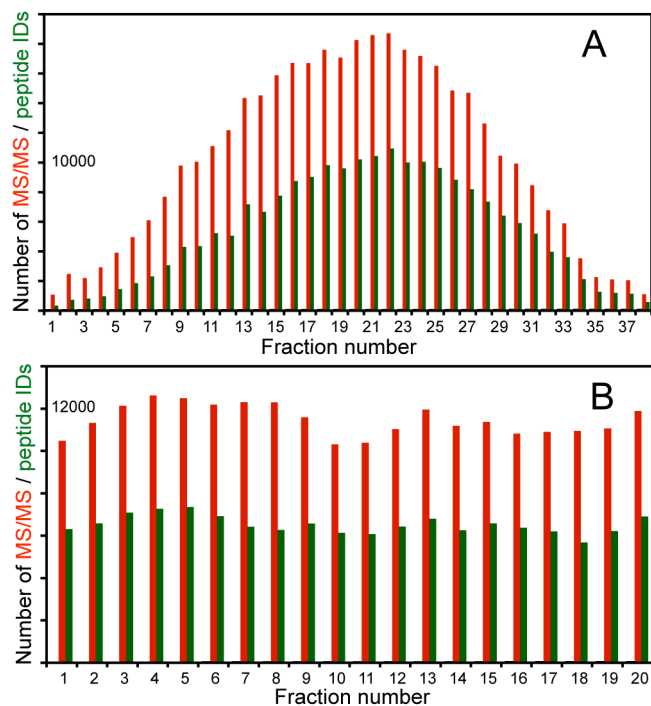
This supporting information contains additional figures (Figure S-1, S-2) and tables (Table S-1 and S-2):

**Figure S-1.** Distribution of number of acquired MS/MS and identified peptides across fractions for 2D LC-MS acquisitions: HILIC-RP and RP-RP (high pH – low pH).

**Figure S-2.** Workflow for retention data filtering and optimization of the SSRCalc HILIC model – expanded version of Figure 2.

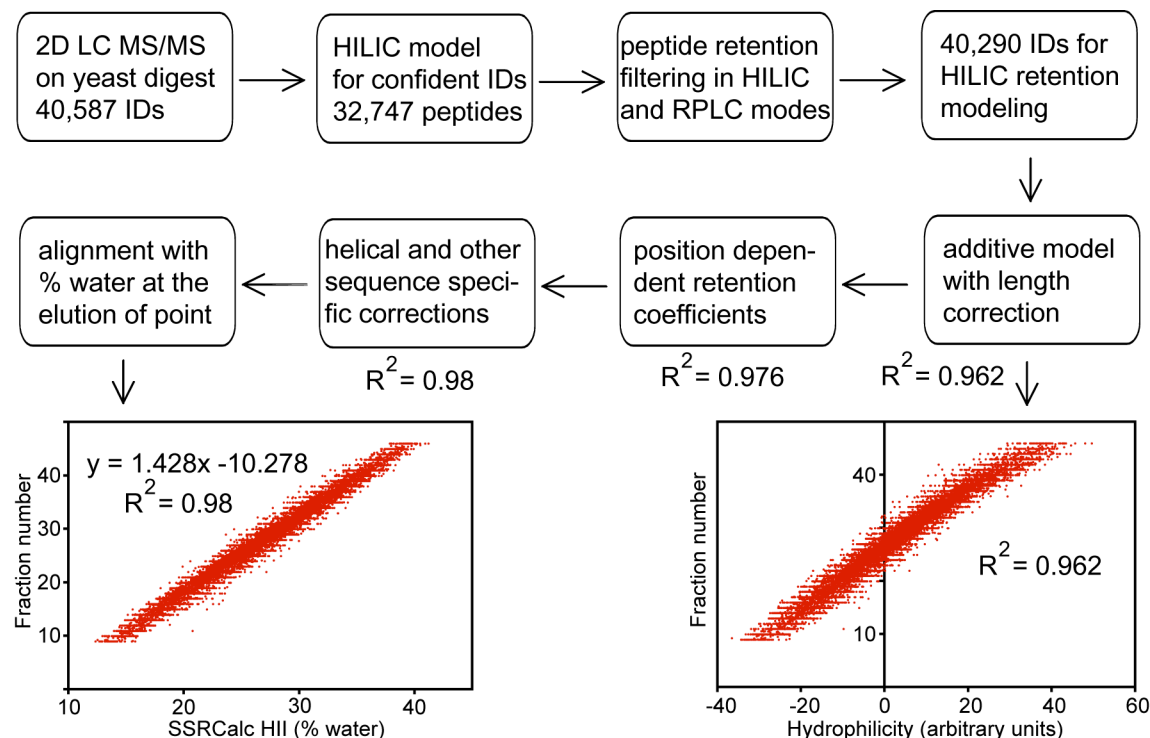
**Table S-1.** Position dependent retention coefficients of individual amino acids in HILIC separation.

**Table S-2.** Typical examples of peptides with largest deviations from HILIC prediction model and their axial helical projections.



**Figure S-1.** Distribution of number of acquired MS/MS (red) and identified peptides (redundant, green) across: A – 38 fractions in HILIC separations from Figure 1 D; B – 20 fractions from RP-RP (high pH – low pH) 2D LC-MS/MS analysis of *S.cerevisiae*.

Figure S-1A shows an expected bell-shaped distribution of number of acquired spectra and identified peptides across 38 HILIC fractions. High pH – low pH (RP-RP) with pairwise fraction concatenation was tailored to maximize utilization of MS/MS time and shows significantly more uniform distribution of the detectable features (Figure S-1B).<sup>26</sup>



**Figure S-2.** Workflow for retention data filtering and optimization of the SSRCalc HILIC model.

The original set of 44,489 identified peptides contained 40,587 non-modified tryptic peptides with expectation values better than -1. A group of 7,840 peptides with low confidence scores ( $-3 < \text{Log}(e) < -1$ ) were excluded for use in the development of our “first pass” approximation of HILIC model. A small portion of these represented false positive identifications, but the majority were short peptides. This first version of HILIC model was used for retention prediction filtering combined with formic acid SSRCalc (second dimension) and database HI values (formic acid) of peptides identified in previous analyses of yeast digests. 0.7% (297 peptides) of all identifications were excluded. The remaining population of 40,290 species was used for model optimization. Tryptic peptides in this dataset were 6-51 residues long (14.1 long on average). Both the amount of the data and the average peptide size were much larger compared to previously published models.

**Table S-1. Position dependent retention coefficients of individual amino acids in HILIC separation on XBridge Amide column (10 mM ammonium formate, pH 4.5).**

Residue	R <sub>C</sub> (internal)	N1*	N2	N3	N4	N5	C5	C4	C3	C2	C1*
K	13.15	14.87	14.74	15.15	14.95	13.66	12.69	13.25	12.4	12.73	12.66
D	12.17	11.84	14.43	12.97	11.35	11.55	12.09	12.17	13.16	13.75	15.34
R	12.15	13	13.15	12.87	12.98	11.82	11.82	11.48	10.06	9.81	9.99
E	12.09	14.33	15.84	12.52	13.01	12.47	13.09	13.09	13.3	14.66	14.5
H	8.98	6.84	7.82	8.85	8.98	8.98	8.98	8.98	8.98	8.98	7.07
N	7.24	3.79	6.85	6.96	6.8	7.16	6.75	7.09	6.34	7.42	8.76
Q	7.21	7.99	8.66	8.29	7.75	7.49	7.49	7.42	7.37	7.78	7.49
C**	5.79	1.99	6.05	6.16	5.72	6.08	6.49	5.92	5.17	5.33	8.29
S	4.57	3.44	4.57	4.9	4.57	4.36	4.16	4.16	4.03	4.57	7.2
P	4	1.38	1.66	1.46	3.21	2.74	3.21	3.03	2.87	2.87	3.41
T	3.25	3.07	3.76	3.33	3.04	2.78	3.04	3.07	2.74	3.4	5.54
G	3.23	3.31	0.68	1.22	2.38	2.23	2.48	1.73	1.64	1.32	5.19
A	1.04	3.66	1.74	1.5	1.58	1.29	1.29	1.17	1.17	1.86	3
V	-3.48	-1.81	-3.45	-3.14	-2.73	-3.27	-2.96	-2.49	-2.99	-3.27	-3.78
Y	-5.08	-5.24	-6.26	-5.24	-4.75	-5.37	-5.03	-4.91	-5.45	-6.37	-5.7
M	-7	-6.33	-7.41	-6.79	-6.79	-6.79	-7.2	-7	-7.25	-7.08	-7.54
I	-7.27	-5.64	-7.6	-7.11	-6.77	-7.11	-6.85	-6.77	-6.77	-7.52	-7.6
L	-10.05	-7.93	-10.97	-10.19	-9.97	-10.38	-10.02	-10.13	-10.22	-10.35	-8.65
F	-11.33	-12.33	-13.5	-12.33	-11.54	-11.87	-11.92	-11.87	-12.33	-13.29	-9.37
W	-11.83	-12.21	-14.58	-13.08	-12.26	-13.08	-13.08	-12.54	-13.08	-13.5	-13.08
Slope and R <sup>2</sup> -value correlation between (R <sub>C</sub> and position dependent coefficients)***											
slope	1.000	0.967	1.130	1.050	1.016	1.017	1.013	1.005	1.007	1.056	1.025
R <sup>2</sup> -value	1.000	0.950	0.982	0.989	0.994	0.996	0.996	0.995	0.99	0.994	0.965

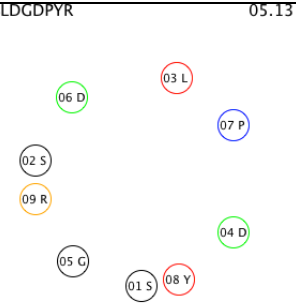
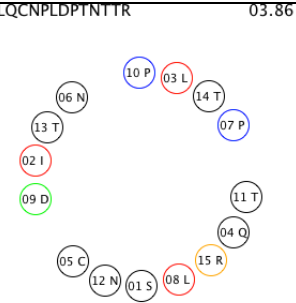
\* - N1 and C1 are N-terminal and C-terminal positions, respectively.

\*\* - Carbamidomethyl-Cys

\*\*\* - R<sub>C</sub> is plotted at X axis, R<sub>N#</sub> and R<sub>C#</sub> at Y axis.

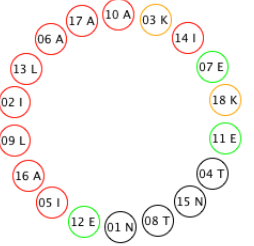
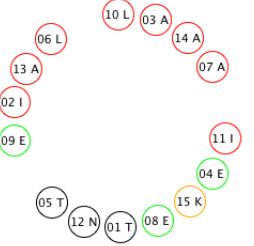
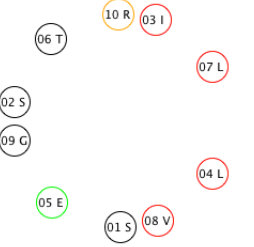
**Table S-2. Typical examples of peptides with largest deviations from HILIC prediction model; possible sequence-specific features that explain this behaviour and their axial helical projections.**

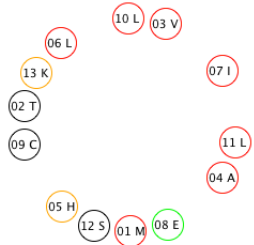
Colour coding in axial projections: red- hydrophobic, green- acidic, yellow- basic, blue- proline, black- all other residues.

Peptide	HILIC prediction error (% water)	RP prediction error (% acetonitrile)*	Agadir helicity	Axial helical projections
Negative prediction errors in HILIC				
Peptides with N-cap motifs (not amphipathic)				
SSLDGDPYR	-6.3	-0.51	0.01	<div> SSLDGDPYR 05.13  </div>
SILQCNPLDPTNTTR	-5.5	3.1	0.01	<div> SILQCNPLDPTNTTR 03.86  </div>

GFSGGPLDPR	-5.1	1.6	0	GFSGGPLDPR 04.81 
SILNPYCVIDPR	-5.1	1.8	0.01	SILNPYCVIDPR 02.50 
VHYDPNGILNPYK	-3.1	0.8	0.01	VHYDPNGILNPYK 08.07 

LVSPSDPTSYMK	-3.0	1.2	0.03	LVSPSDPTSYMK 04.53 
Ala-rich helical peptides				
YATASAIATAVASLVLAR	-5.5	3.8	0.81	YATASAIATAVASLVLAR 03.88 
ANVADILVATAVAAR	-3.6	3.4	0.82	ANVADILVATAVAAR 02.70 

Amphipathic helical peptides (extremely high retention in RPLC)				
NIKTIAETLAEELINAAK	-3.8	8.9	3.41	NIKTIAETLAEELINAAK 09.77 
TIAETLAEELINAAK	-2.9	6.7	2.23	TIAETLAEELINAAK 07.43 
SSILETLVGR	-2.9	3.4	1.04	SSILETLVGR 06.68 

MTVAHLIECLLSK	-1.3	8.0	0.56	<div>MTVAHLIECLLSK06.59</div> <div></div>
---------------	------	-----	------	--

Positive prediction errors in HILIC				
Peptides with multiple Pro, Gly, hydrophobic clusters, multiple positively charged residues				
KFVFNPPKPR	3.8	1.2	0	<div> <div> <div>04.72</div> <div> </div> </div> </div>
KQIAFPQRK	3.8	0.6	0	<div> <div> <div>03.18</div> <div> </div> </div> </div>
GSNFGSSRPPIR	3.8	0.7	0.01	<div> <div> <div>04.49</div> <div> </div> </div> </div>

