Supplementary Table 3. Evaluation of different machine learning approaches' prediction power on training dataset.

	Apparent error rate	
	10X cross validation on training data	**Human whole cell lysate**
Support Vector Machine	0.28	0.4
Random Forest	0.23	0.33
Neural network	0.23	0.27

Supplementary Table 4. Neural network model training. We started by training the model with all 544 physicochemical features and 20 amino acid counts. Then we remove highly redundant features and used 120 physicochemical features and 20 amino acid counts. Subsequently, backward feature selection was used to reduce the features in half in each step by feature importance. In addition, we also used forward selection to construct feature sets at each step by selecting from unused features that added most weight to the model. Finally, the features showing in both selection methods, were combined to form the 15 final features. In the feature selection step, we applied filtering using sequence coverage and abundance to remove very lowly expressed proteins. Lastly, protein abundance was introduced into the model and the filtering conditions were removed. For model evaluation, we divided the reported probability score into 10 equally spaced bins and computed the ratio of peptides detected [ratio], in terms of *in silico* peptides falling into this category [bin.pep] that are actually detected [detected] in each category. When performance statistics were reported, we used two methods. In the first method [2-Part], we use a cutoff score of 0.5 and treated every *in silico* peptide with a score less than that as not detected. In the second method[3-Part], the divide the peptides into three equal portions we removed the middle portion since this portion mainly contained ambiguous peptides in terms of being detected or not detected. (Var #: Number of features, cov: minimum sequence coverage [%], fmol: minimum absolute amount on column [fmol/column].

		# peptides			PPA score bins										Filter		Type	Sensitivity	Specificity	Precision	Accuracy
	Var#	Total	Dectecd		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	cov	fmol					
All features				bin.pep	487	56	42	44	158	175	295	466	296	56			2-Part	0.9	0.55	0.71	0.74
				detected	59	21	14	15	95	111	199	338	233	44			3-Part	0.83	0.76	0.76	0.79
	564	2075	1129	ratio	0.12	0.38	0.33	0.34	0.6	0.63	0.67	0.73	0.79	0.79	20	7					
Backward feature selection by elimination				bin.pep	296	104	135	175	340	42	48	533	402	0			2-Part	0.84	0.56	0.7	0.71
				detected	17	31	62	68	201	26	27	380	317	0			3-Part	0.75	0.76	0.76	0.76
	140	2075	1129	ratio	0.06	0.3	0.46	0.39	0.59	0.62	0.56	0.71	0.79	0	20	7					
				bin.pep	375	118	89	219	108	60	124	594	388	0			2-Part	0.84	0.65	0.74	0.75
				detected	15	30	43	97	53	39	81	441	330	0			3-Part	0.8	0.8	0.8	0.8
	70	2075	1129	ratio	0.04	0.25	0.48	0.44	0.49	0.65	0.65	0.74	0.85	0	20	7					
				bin.pep	350	130	111	81	101	198	328	472	268	36			2-Part	0.9	0.59	0.72	0.76
				detected	14	34	33	31	55	125	231	353	223	30			3-Part	0.83	0.79	0.79	0.81
	35	2075	1129	ratio	0.04	0.26	0.3	0.38	0.54	0.63	0.7	0.75	0.83	0.83	20	7					
				bin.pep	402	82	80	68	92	117	604	355	260	15			2-Part	0.91	0.57	0.72	0.76
				detected	18	19	33	27	48	82	406	259	225	12			3-Part	0.85	0.75	0.75	0.8
	18	2075	1129	ratio	0.04	0.23	0.41	0.4	0.52	0.7	0.67	0.73	0.87	0.8	20	7					
				bin.pep	329	129	109	110	110	183	288	556	261	0			2-Part	0.9	0.59	0.72	0.76
				detected	19	18	33	46	58	111	194	422	228	0			3-Part	0.83	0.8	0.8	0.81
	9	2075	1129	ratio	0.06	0.14	0.3	0.42	0.53	0.61	0.67	0.76	0.87	0	20	7					
				bin.pep	339	97	92	87	90	140	541	682	7	0			2-Part	0.91	0.55	0.71	0.75
				detected	19	16	24	39	46	75	378	526	6	0			3-Part	0.85	0.73	0.74	0.79
	4	2075	1129	ratio	0.06	0.16	0.26	0.45	0.51	0.54	0.7	0.77	0.86	0	20	7					
Forward feature selection				bin.pep	294	140	115	109	161	172	274	639	170	1			2-Part	0.9	0.58	0.72	0.75
				detected	16	16	34	45	81	111	182	495	148	1			3-Part	0.84	0.79	0.8	0.81
	3	2075	1129	ratio	0.05	0.11	0.3	0.41	0.5	0.65	0.66	0.77	0.87	1	20	7					
				bin.pep	300	116	106	96	111	183	523	536	104	0			2-Part	0.91	0.55	0.71	0.75
				detected	16	16	27	37	62	112	363	405	91	0			3-Part	0.86	0.74	0.76	0.8
	5	2075	1129	ratio	0.05	0.14	0.25	0.39	0.56	0.61	0.69	0.76	0.88	0	20	7					
				bin.pep	361	99	108	118	122	146	271	500	344	6			2-Part	0.91	0.61	0.74	0.77
				detected	10	17	37	43	66	97	186	373	297	3			3-Part	0.84	0.83	0.83	0.84
	10	2075	1129	ratio	0.03	0.17	0.34	0.36	0.54	0.66	0.69	0.75	0.86	0.5	20	7					
				bin.pep	392	94	81	105	113	164	245	547	328	6			2-Part	0.91	0.6	0.73	0.77
				detected	17	18	22	44	60	110	159	418	277	4			3-Part	0.85	0.82	0.82	0.83
	15	2075	1129	ratio	0.04	0.19	0.27	0.42	0.53	0.67	0.65	0.76	0.84	0.67	20	7					
Model with filter				bin.pep	377	115	99	94	134	143	232	483	388	10			2-Part	0.908	0.614	0.737	0.774
				detected	12	19	31	42	68	95	151	375	327	9			3-Part	0.845	0.832	0.829	0.839
	15	2075	1129	ratio	0.03	0.17	0.31	0.45	0.51	0.66	0.65	0.78	0.84	0.9	20	7					
Model with filter and 3 hidden nenrals				bin.pep	235	53	57	56	72	102	153	184	182	83			2-Part	0.885	0.597	0.714	0.750
				detected	17	13	20	22	41	67	105	134	139	68			3-Part	0.819	0.782	0.769	0.799
	15	2075	1129	ratio	0.07	0.25	0.35	0.39	0.57	0.66	0.69	0.73	0.76	0.82	20	7					
Model with filter and 5 hidden nenrals				bin.pep	308	61	37	58	56	63	70	109	131	257			2-Part	0.816	0.633	0.717	0.731
				detected	47	25	17	26	33	38	49	82	85	199			3-Part	0.785	0.761	0.751	0.772
	15	2075	1129	ratio	0.15	0.41	0.46	0.45	0.59	0.60	0.70	0.75	0.65	0.77	20	7					
Full model W/O filter				bin.pep	492	117	128	135	165	239	616	761	189	0			2-Part	0.913	0.535	0.67	0.727
				detected	13	18	35	59	88	136	378	566	152	0			3-Part	0.856	0.74	0.739	0.794
	15	2842	1445	ratio	0.03	0.15	0.27	0.44	0.53	0.57	0.61	0.74	0.8	0	NA	NA					
Full Model with abundance				bin.pep	565	155	137	144	173	227	320	488	626	7			2-Part	0.905	0.618	0.71	0.764
				detected	11	26	39	61	84	119	213	365	522	5			3-Part	0.879	0.822	0.807	0.848
	15 + 1	2842	1445	ratio	0.02	0.17	0.28	0.42	0.49	0.52	0.67	0.75	0.83	0.71	NA	NA					

Supplementary Table 5. Physiochemical features of peptides used in PPA.

Feature	Description	Publication
BHAR880101	Average flexibility indices	Bhaskaran-Ponnuswamy, 1988
DAWD720101	Size	Dawson, 1972
DAYM780101	Amino acid composition	Dayhoff et al., 1978a
JUKT750101	Amino acid distribution	Jukes et al., 1975
KLEP840101	Net charge	Klein et al., 1984
KYTJ820101	Hydropathy index	Kyte-Doolittle, 1982
OLSK800101	Average internal preferences	Olsen, 1980
VENT840101	Bitterness	Venanzi, 1984
ZIMJ680103	Polarity	
ZIMJ680104	Isoelectric point	
ZIMJ680105	RF rank	Zimmerman et al., 1968
FASG760101	Molecular weight	Fasman, 1976
GRAR740102	Polarity	Grantham, 1974
ARGP820101	Hydrophobicity index	Argos et al., 1982
HOPA770101	Hydration number	Hopfinger, 1971

Supplementary Table 6. Area under the ROC for the seven validation datasets of PPA (Figures 2C, 4, Supplementary Figures 3, 5). Datasets that haven't been generated in this study are cited. Results from ESP predictor (ESP, Fusaro *et al* ., 2009) were compared to PPA in three conditions. a) PPA-15 was based only on 15 physicochemical features without any protein abundance information, b) PPA-eSC used the 15 features and external protein coverage from other datasets/publications (referred to as external sequence coverage, eSC), c) PPA-iSC used 15 features and each dataset's own sequence coverage (referred to as internal sequence coverage), d) PPA-CID was trained on a CID-based dataset (Orbitrap Velos; Wisniewski, *et al.,* 2012), PPA-iBAQ used the 15 features and a normalized iBAQ value as protein abundance (referred to as intensity based absolute quantification, iBAQ; Schwanhäusser *et al* ., 2011). n.pos: Number of detected peptides per experiment. n. neg: Number of not detected peptides.

Sample	Platform	ESP	PPA-15	PPA-eSC	PPA-iSC	PPA-CID	PPA-iBAQ	n.pos	n.neg
Single digests	Q-Exactive	0.726	0.791	0.831	0.841			1445	1397
Pooled single digests	Q-Exactive	0.621	0.699	0.706	0.748			3755	8917
Human lysate	Q-Exactive	0.701	0.747	0.791	0.834	0.758	0.846	4392	14924
Human lysate (Wisniewski, *et al.* , 2012)	LTQ-Orbitrap Velos	0.765	0.787	0.815	0.835			9240	10076
Yeast lysate	Q-Exactive	0.78	0.783	0.815	0.861			3529	8581
Yeast lysate	TripleToF 5600	0.689	0.7	0.769	0.815		0.819	7402	4708
Yeast lysate (Hebert *et al.,* 2013)	Orbitrap-Fusion	0.755	0.761	0.805	0.83	0.762		21617	34361

Supplementary Table 7. Statistical analysis of the area under the ROC for the validation datasets (Figures 2C, 4). We applied a modified z-test to compute a p-value (http://www.vassarstats.net/roc_comp.html): a) PPA-15 was based only on 15 physicochemical features, b) PPA-eSC used the 15 features and external protein coverage from other datasets/publications (referred to as external sequence coverage, eSC), c) PPA-iSC used 15 features and each dataset's own sequence coverage (referred to as internal sequence coverage).

Sample	Platform	PPA-15 vs ESP		PPA-eSC vs ESP		PPA-iSC vs ESP		PPA-eSC vs PPA-15		PPA-iSC vs PPA-15		PPA-iSC vs PPA-eSC	
		z value	p-value	z value	p-value	z value	p-value	z value	p-value	z value	p-value	z value	p-value
Single digests	Q-Exactive	5.2	<1E-06	8.694	<1E-06	9.661	<1E-06	3.498	<1E-04	4.469	<1E-03	0.973	0.165
Pooled single digests	Q-Exactive	10.158	<1E-06	11.083	<1E-06	16.894	<1E-06	0.917	0.179	6.655	<1E-06	5.484	<1E-06
Human lysate	Q-Exactive	6.926	<1E-06	14.101	<1E-06	21.385	<1E-06	7.14	<1E-06	14.364	<1E-06	7.2	<1E-06
Human lysate (Wiśniewski, et al, 2012)	LTQ-Orbitrap Velos	4.528	<1E-03	10.646	<1E-06	15.307	<1E-06	6.119	<1E-06	10.784	<1E-06	4.67	<2E-06
Yeast lysate	Q-Exactive	0.426	0.335	5.114	<1E-06	12.609	<1E-06	4.687	<1E-05	12.025	<1E-06	7.336	<1E-06
Yeast lysate	TripleToF 5600	1.647	0.049	12.629	<1E-06	20.783	<1E-06	10.98	<1E-06	19.133	<1E-06	8.175	<1E-06
Yeast lysate (Hebert et al, 2013)	Orbitrap-Fusion	1.943	0.026	16.802	<1E-06	25.842	<1E-06	14.857	<2E-04	23.895	<1E-06	9.034	<1E-06

Supplementary Table 8. Neural network classification feature weights in three neural network models. Model 1: 15 physiochemical features (see **Table S1**). Model 2: PPA based on protein abundance as sequence coverage. Model 3 PPA based on protein abundance as average peak intensity of the three highest intense tryptic peptides (TOP-3, Silva *et al* ., 2006). Model 4 PPA based on protein abundance as iBAQ (Schwanhäusser *et al.* , 2011)

		Model 1		**Model 2**		**Model 3**		**Model 4**	
		To hidden neuron 1	**To hidden neuron 2**	**To hidden neuron 1**	**To hidden neuron 2**	**To hidden neuron 1**	**To hidden neuron 2**	**To hidden neuron 1**	**To hidden neuron 2**
First layer input nodes	Intercept	8.281	1.992	-9.251	-1.485	-1.492	-9.483	-8.265	2.126
	BHAR880101	1.844	-1.481	-1.539	0.854	0.868	-1.522	-1.211	0.366
	DAWD720101	0.392	1.478	0.384	-0.568	-0.589	0.411	0.428	-0.116
	DAYM780101	7.566	0.092	-6.237	-0.409	-0.481	-6.41	-5.285	-0.221
	JUKT750101	-8.684	-0.99	3.896	1.349	1.382	4.038	3.626	0.832
	KLEP840101	-1.267	-0.409	-0.298	0.194	0.164	-0.328	-0.321	0.126
	KYTJ820101	2.001	0.303	-0.609	-0.433	-0.442	-0.622	-0.116	-0.244
	OLSK800101	-3.249	-1.279	-0.025	1.342	1.356	-0.026	-0.889	0.671
	VENT840101	0.86	0.573	-0.514	-0.298	-0.303	-0.527	-0.403	-0.139
	ZIMJ680103	-0.29	0.332	-0.398	-0.272	-0.313	-0.42	-0.174	-0.060
	ZIMJ680104	0.21	0.038	9.001	-0.067	0.223	9.305	6.533	-0.539
	ZIMJ680105	4.761	-1.132	-3.209	0.125	0.111	-3.274	-2.419	0.016
	FASG760101	7.447	-5.158	-10.612	3.283	3.309	-10.743	-8.246	1.443
	GRAR740102	-2.57	8.498	1.637	-5.992	-6.31	1.432	2.611	-2.395
	ARGP820101	0.44	0.663	-0.779	-0.145	-0.164	-0.795	-0.449	0.012
	HOPA770101	-0.35	-2.584	1.687	1.469	1.585	1.75	0.779	0.462
	Abundance	NA	NA	-0.179	-0.431	-0.452	-0.185	0.399	-2.307
Hidden layer nodes	Intercept	-13.854		2.998		2.908		10.883	
	Neuron 1	9.138		-173.345		-8.644		-197.752	
	Neuron 2	6.641		-9.078		-208.71		-17.944	
		Hidden layer neuron to output		**Hidden layer neuron to output**		**Hidden layer neuron to output**		**Hidden layer neuron to output**	