Supporting Information

LC-MS data acquisition

Sample preparation was performed as previously described (1). The amount of urine injected into the LC-MS system was normalized to 50 nmol of creatinine. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

Preparation of spiked urine samples

600 µL carbonic anhydrase (CA) solution of 22 mg/mL dissolved in 50 mmol/L NH_4HCO_3 buffer at pH 7.8 were divided into 6 equal aliquots. Ten µL of 100 mM DTT were added to each aliquot and the solution was incubated at 50°C for 30 min followed by addition of 40 µL of 137.5 mM iodoacetamide and incubation at room temperature for another 60 min. Reduced and alkylated CA was digested by adding 40 µL of 0.5 µg/µL trypsin and subsequent incubation at 37°C over night. The reaction was stopped by addition of 10 µL pure formic acid (FA). The excess of DTT and iodoacetamide was removed by solid-phase extraction using a 100 mg Strata C-18 SPE column with the following protocol: the column was conditioned with 2 mL methanol, followed by one washing step with 2 mL water. Each aliquot of digested CA was loaded on the SPE column and the column was subsequently washed with 2 mL of 5% aq. methanol. Peptides were eluted with 1 mL of 80% aq. methanol. The eluate was dried in a vacuum centrifuge and re-dissolved in 200 µL 30% acetonitrile (ACN) and 1% FA. Finally 500 µL of digested CA were mixed with 200 µL of a stock solution of the synthetic peptides resulting in a standard mixture stock solution with a calculated digested CA concentration of 240 µM and the following concentrations (in µM) for the 7 synthetic peptides: VYV, 83; YGGFL, 57; DRVYIHPF, 29; YPFPGPI, 46; YPFPG, 60; GYYPT, 54; and YGGWL, 57.

No compound signal interfering with spiked-in peptides was observed in urine at the lowest spiking level (spiked with a peptide solution at 2000× dilution) and in the non-spiked pooled urine sample.

Reversed-Phase LC−MS

All LC−MS analyses were performed on an 1100 series capillary HPLC system equipped with a cooled autosampler (4°C) and an SL ion trap mass spectrometer (Agilent Technologies, Santa Clara, CA, United States). Samples were desalted on an Atlantis dC18 precolumn (Waters Corporation, Milford, MA, USA, 2.1 × 20 mm, 3 µm particles, 10 nm pores) using 0.1% FA in 5% ACN at a flow rate of 50 µL/min for 16 min. Compounds were back-flushed from the precolumn onto a temperature-controlled (25°C) Atlantis dC18 analytical column (1.0 × 150 mm, 3 µm particles, 30 nm pores) and separated over 90 min at a flow rate of 50 µL/min during which the percentage of solvent B (0.1% FA in ACN) in solvent A (0.1% FA in ultrapure H_2O) was increased from 5.0 to 43.6% (eluent gradient of 0.43%/min). Settings of the electrospray ionization interface and the mass spectrometer were as follows: nebulization gas, 40.0 psi N_2; drying gas, 6.0 L/min N_2; capillary temperature, 325°C; capillary voltage, 3250 V; skimmer voltage, 25 V; capillary exit voltage, 90 V; octapole 1 voltage, 8.5 V; octapole 2 voltage, 4.0 V; octapole RF voltage, 175 V; lens 1 voltage, −5 V; lens 2 voltage, −64.6 V; trap drive, 67; scan speed, 5500 m/z s^{-1}; accumulation time 50 ms (or 30 000 ions); scan range, 100−1500 m/z; a Gaussian smoothing filter (width 0.15 m/z) was applied for each mass spectrum; rolling average was disabled, resulting in a rate of approximately 70 mass spectra per minute. Spectra were saved in profile mode.

Following the gradient, both columns were washed with 85% B for 5 min and equilibrated with 5% B for 10 min prior to the next injection. Different volumes of the standard mixture (CA digest plus peptides) were injected on the pre-column prior to injection of the pooled urine sample to obtain the desired final concentrations. The injection system was cleaned with 70% ACN after each injection and

filled with 0.1% FA in 5% ACN. Mass spectrometry settings were optimized for detection of singly- and doubly-charged ions of DRVYIHPF without provoking upfront fragmentation. Raw data converted to mzXML format are available at http://tinyurl.com/statisticsComparison. After the LC-MS analysis, the raw LC-MS profile data was exported in mzXML format using CompassExport v1.3.6.

Assignment of features in the spiked human urine and porcine CSF datasets that are derived from spiked peptides

A list of features, that are derived from the added peptides (CA digest and 7 synthetic peptides), was assigned based on 4 analyses of samples containing only peptides used for spiking at a 100-fold dilution of the stock solution and analyzing the resulting data with the Threshold Avoiding Proteomics pipeline2. A feature was considered to belong to one of the spiked peptides if it was detected by one of the workflows in at least two separate chromatograms. All features fulfilling these criteria were combined in one set. The resulted list was verified manually by visual inspection of the corresponding EICs in five urine samples spiked at 12.5- (B) and 2000-fold (F) dilution of the standard stock solution resulting in a final list of 151 identified features. This list corresponded to standard peptides, and constituted our reference list to identify features related to the spiked peptides in feature lists obtained from urine samples. A similar identification procedure was applied to the identification of spiked-in features for porcine CSF samples, however in this case using chromatograms obtained from non-spiked samples or samples spiked with 100 fmol of a horse heart Cytochrome C digest. The list of identified standard features is given in Table S2a and b.

Source code

The program source code and preprocessed LC-MS data with descriptions of spiking levels and spiked compounds and indices of spiked-in-peptide-related features available through the source code repository of Netherlands Bioinformatics Centre at https://trac.nbic.nl/biomarkerfeatureselection.

1. Kemperman, R. F., Horvatovich, P. L., Hoekman, B., Reijmers, T. H., Muskiet, F. A., and Bischoff, R. (2007) Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: method development, evaluation, and application to proteinuria. *J Proteome Res* 6, 194-206.
2. Suits, F., Hoekman, B., Rosenling, T., Bischoff, R., and Horvatovich, P. (2011) Threshold-avoiding proteomics pipeline. *Anal Chem* 83, 7786-7794.

id	mz	rt (min)
1	290.5	50.43
2	290.8	45.94
3	298.0	59.28
4	321.6	43.25
5	324.6	42.96
6	338.4	48.91
7	349.4	50.81
8	350.4	48.96
9	352.0	50.40
10	353.2	43.56
11	353.3	41.11
12	355.0	51.35
13	356.4	49.10
14	369.7	50.56
15	371.1	50.67
16	371.8	59.34
17	375.8	49.34
18	380.1	50.83
19	380.1	37.28
20	382.8	59.99
21	384.3	50.58
22	395.5	63.57
23	396.3	51.50
24	399.1	50.85
25	401.7	43.31
26	410.7	51.88
27	411.3	59.67
28	412.4	55.05
29	415.1	50.49
30	419.3	50.68
31	422.5	50.93
32	424.5	43.59
33	428.3	59.44
34	445.6	60.38
35	448.9	43.47
36	449.7	46.27
37	451.0	48.20
38	451.7	56.65
39	453.2	50.61
40	458.4	50.95
41	464.3	59.77
42	470.1	51.42
43	470.6	57.07
44	485.9	44.65
45	487.1	51.41
46	489.2	59.50
48	490.1	48.99
49	493.1	43.38
50	501.2	46.66
51	501.3	54.78
52	505.0	44.61
53	506.2	50.52
54	506.6	56.93
55	509.6	43.68

id	mz	rt (min)
56	511.4	59.55
57	515.7	52.12
58	518.5	41.10
59	520.4	50.41
60	521.0	59.57
61	523.9	50.98
62	526.0	56.77
63	528.0	51.71
64	528.4	61.37
65	529.6	46.85
66	534.0	49.58
67	538.3	44.69
68	541.0	40.81
69	543.5	48.81
70	547.1	52.08
71	550.3	67.37
72	552.4	59.81
73	552.6	48.87
74	553.1	52.57
75	555.1	67.85
76	556.0	55.67
77	557.9	43.47
78	562.8	60.66
79	566.7	50.50
80	569.8	70.77
81	569.9	50.87
82	570.7	56.94
83	575.6	50.92
84	578.2	53.67
85	580.1	49.70
86	582.4	67.84
87	583.3	56.99
88	583.4	70.46
89	586.6	51.13
90	595.0	59.42
91	596.6	56.39
92	598.2	43.43
93	599.8	51.52
95	604.3	51.06
96	606.8	55.28
97	611.3	59.92
98	611.8	56.50
99	615.7	56.34
100	621.0	68.54
101	626.2	50.74
102	638.5	63.88
103	645.0	66.31
104	646.4	60.12
105	648.0	43.48
106	650.3	51.61
107	652.0	59.39
108	652.7	68.29
109	653.2	46.26
110	657.3	71.50

id	mz	rt (min)
111	667.8	60.54
112	671.3	75.29
113	673.8	46.94
114	675.4	78.01
115	682.0	52.69
116	685.5	67.72
117	693.0	63.68
118	694.4	78.70
119	696.4	60.48
120	700.7	52.34
121	706.0	60.59
122	710.8	57.41
123	710.9	76.29
124	716.4	58.70
125	721.8	41.62
126	723.4	43.54
127	723.8	66.68
128	724.8	61.35
129	728.2	69.06
130	733.1	67.76
131	738.0	69.04
132	740.1	79.17
133	742.6	68.47
134	751.7	77.17
135	752.3	67.72
136	754.6	51.65
137	758.9	70.39
138	765.2	67.75
139	770.8	77.09
140	771.6	68.28
141	776.2	67.85
142	778.2	70.74
143	780.6	59.45
144	789.7	63.57
145	801.3	78.75
146	810.5	71.75
147	839.1	71.70
148	874.2	70.14
149	894.0	55.78
150	966.7	66.19
151	984.4	43.71

Table S1a. Retention times and mass to charge ratios of the monoisotopic standard peaks derived from spiked peptides in human urine samples. Ions with different charge states and distinguishable isotopologues are considered as separate features.

id	mz (Da)	Rt (min)
1	533.966	29.60
2	533.637	29.66
3	737.375	30.91
4	736.878	30.92
5	491.590	30.93
6	736.394	30.96
7	735.894	30.96
8	490.931	30.96
9	491.272	31.10
10	391.288	38.25
11	391.749	38.28
12	390.759	38.34
13	390.258	38.35
14	534.311	38.78
15	585.851	38.84
16	391.580	38.85
17	390.913	38.86
18	390.579	38.87
19	390.243	38.87

id	mz (Da)	Rt (min)
20	585.347	38.89
21	584.851	38.90
22	391.241	38.91
23	534.840	38.97
24	451.270	41.11
25	450.948	41.26
26	599.346	42.47
27	598.849	42.55
28	483.807	43.58
29	483.307	43.59
30	482.807	43.59
31	484.285	43.66
32	542.311	45.11
33	541.975	45.11
34	542.644	45.12
35	542.979	45.12
36	749.401	48.17
37	748.402	48.18
38	499.280	48.23

Table S1b. Retention times and mass to charge ratios of the monoisotopic standard peaks derived from spiked peptides for porcine CSF samples. Ions with different charge states and isotopologues are considered as separate features.

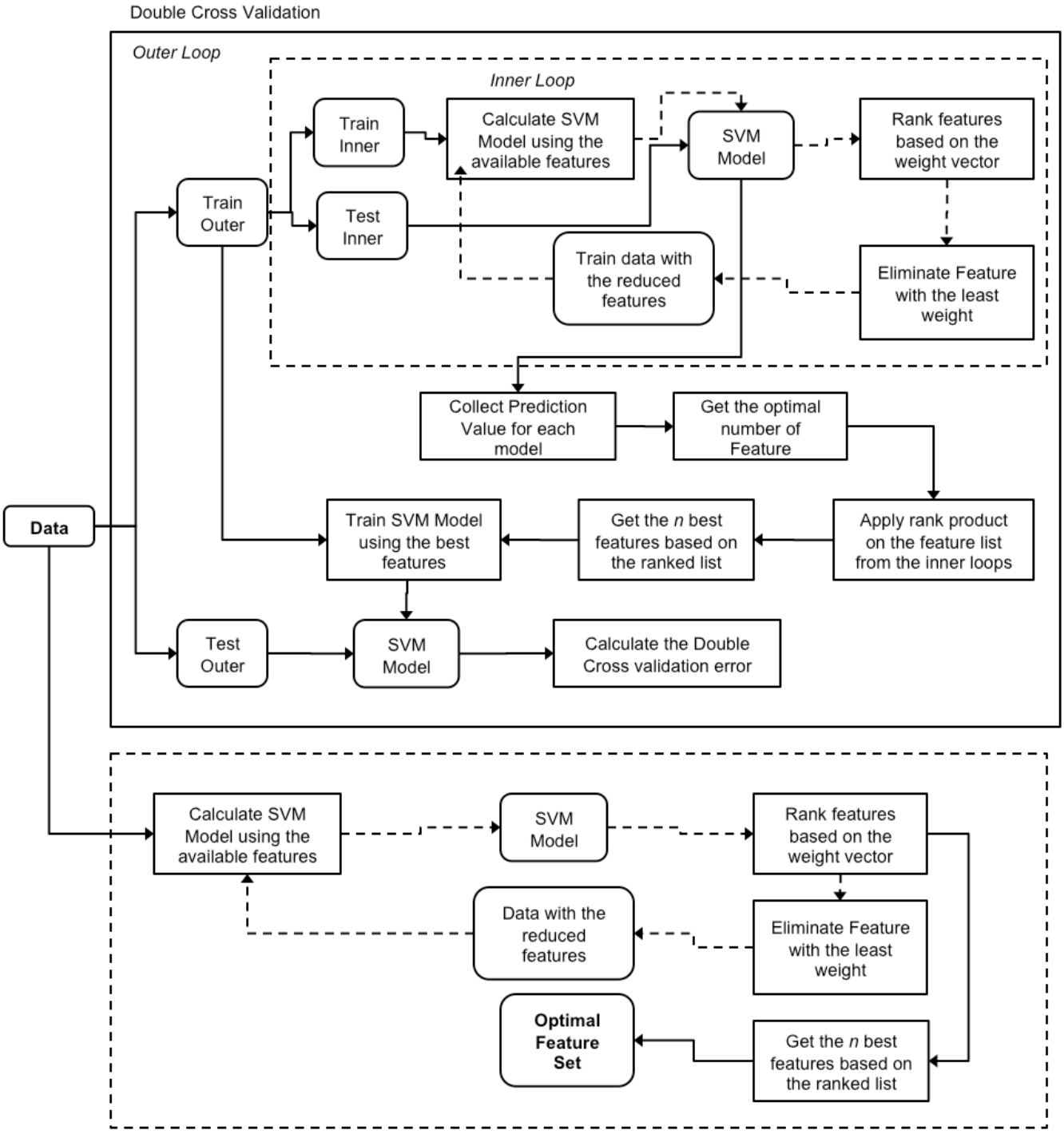

Figure S1. Double cross validation scheme for a Support Vector Machine combined with Recursive Feature Elimination (SVM-FRE). The optimal number of features is obtained in the inner loops. The optimum model is then tested against the test data in the outer loop to obtain the overall classification error.

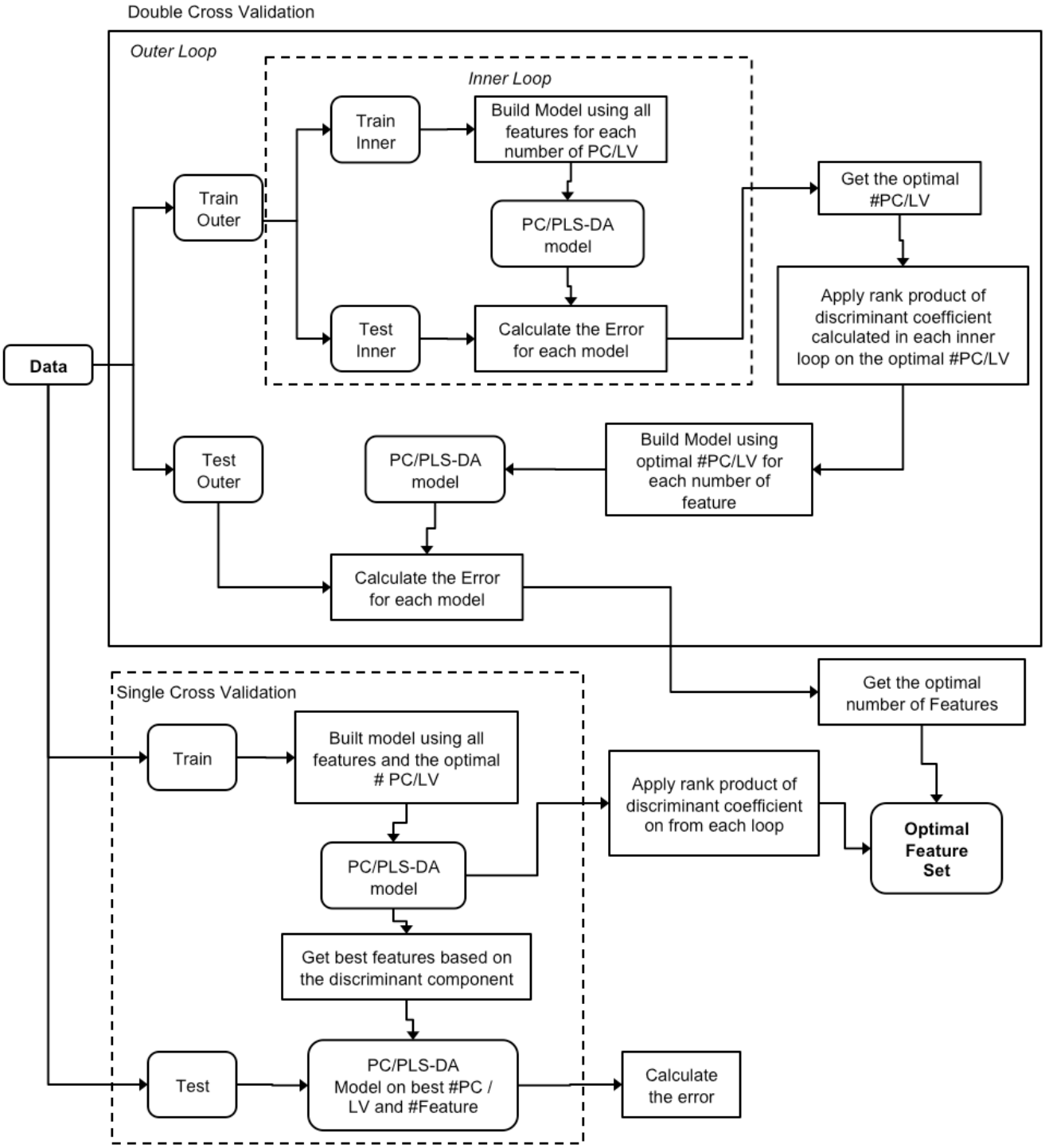

Figure S2. Double cross validation scheme for PCDA and PLSDA. The optimal number of PC/PLS components is obtained in the inner loops. In the outer loop, the optimal number of features is determined by calculating the classification error for each ranked feature set. Once the optimal number of PC/PLS components and the optimal number of features has been obtained, a single cross validation scheme is performed to determine the optimal feature set. The optimum model is then tested against the test data in the outer loop to obtain the classification error.

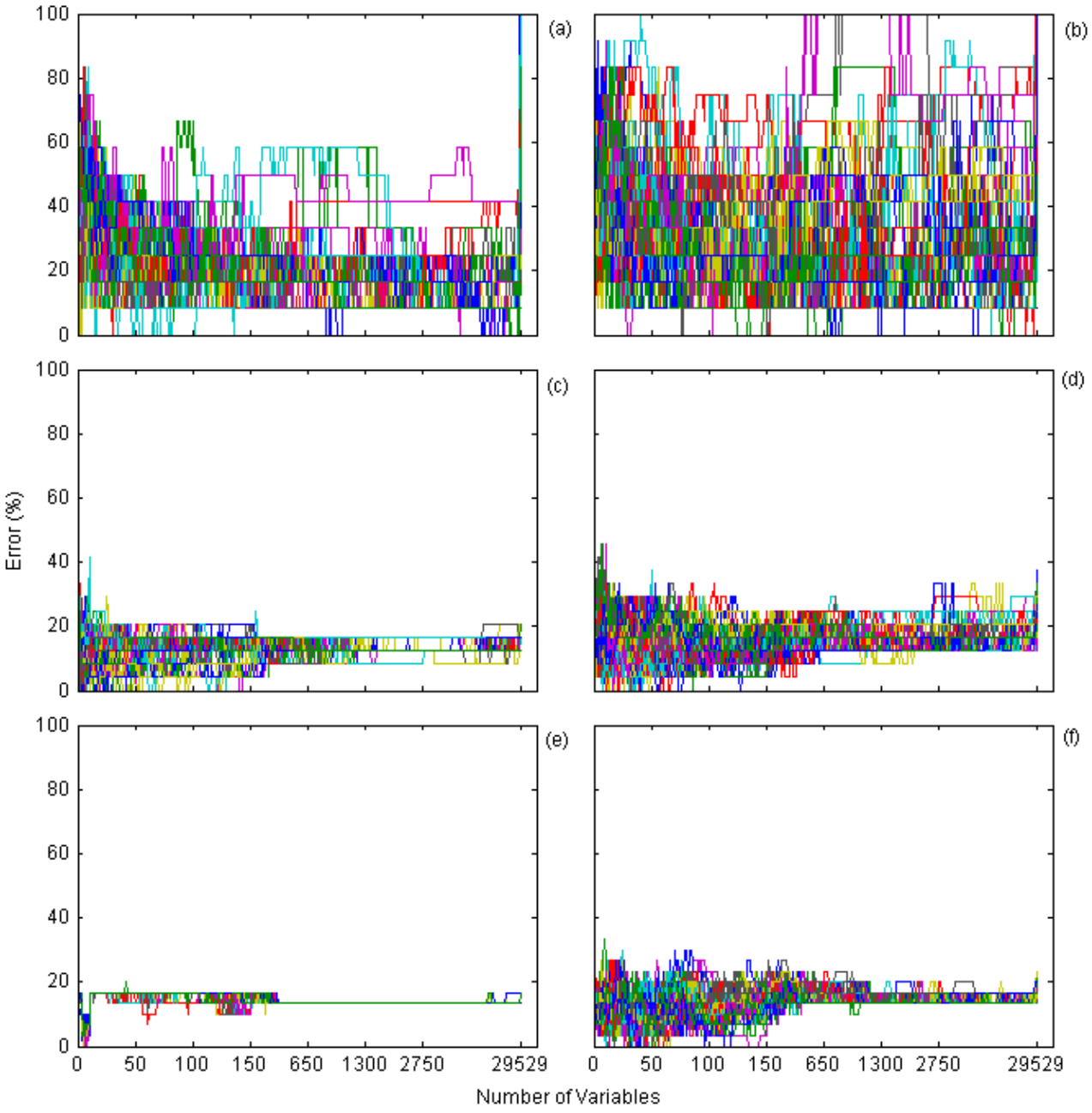

Figure S3. Error plots of the PCDA model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets). The error plots show a decreasing variability with increasing sample size per class.

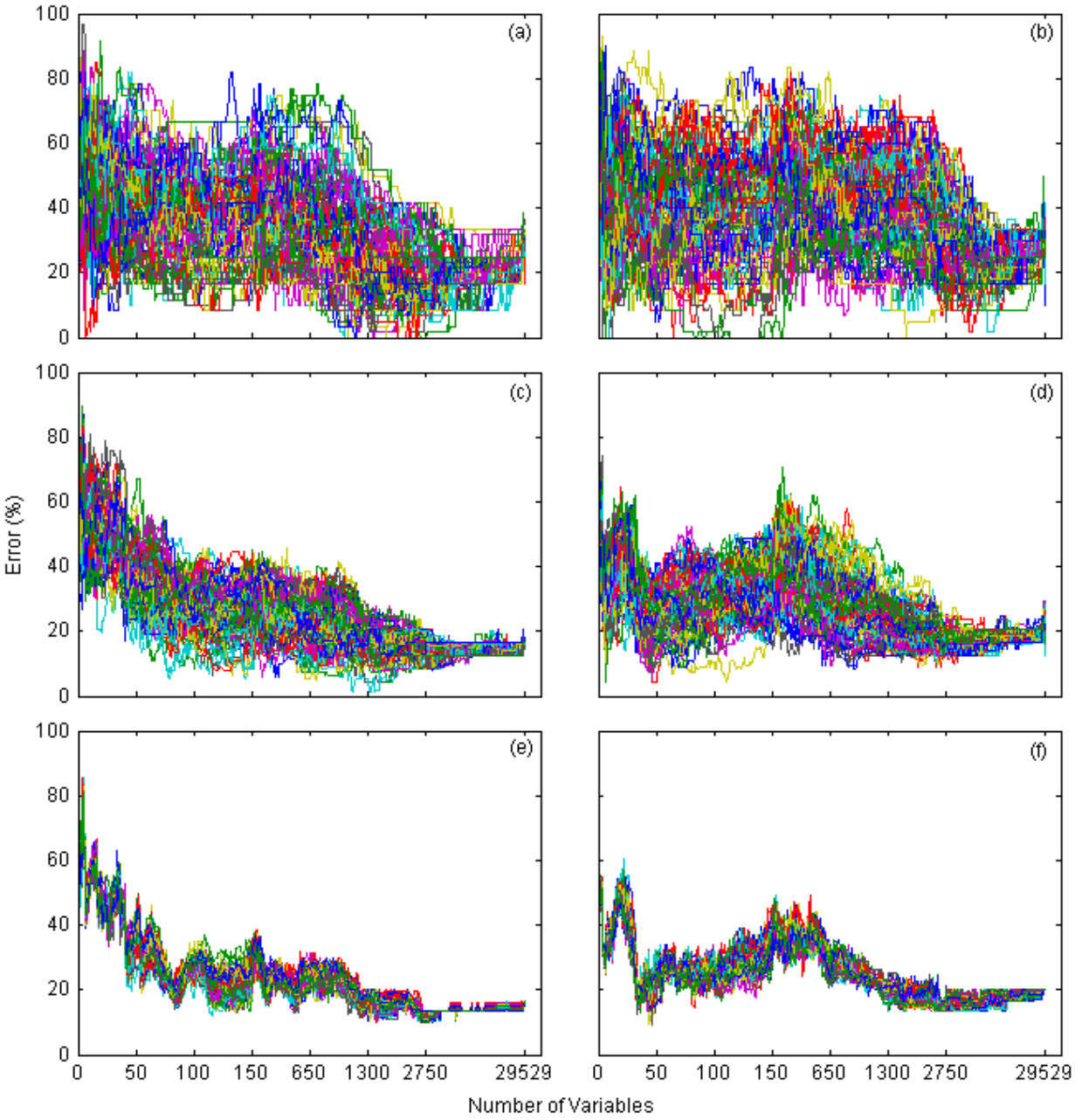

Figure S4. Error Plot of the SVM-RFE model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets).

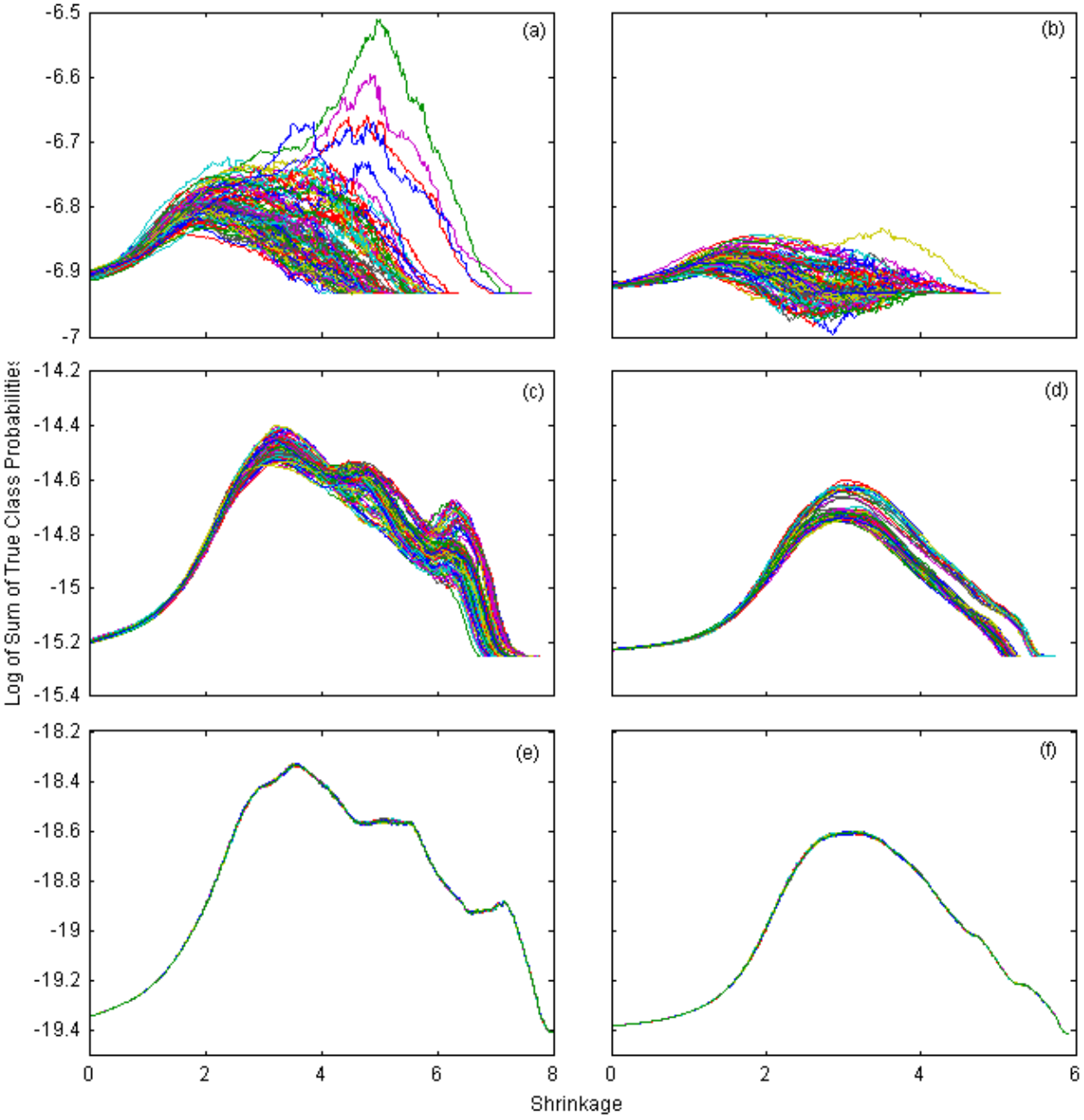

Figure S5. Probability plot of the NSC model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets). The probability plots show a decreasing variability with increasing sample size per class.

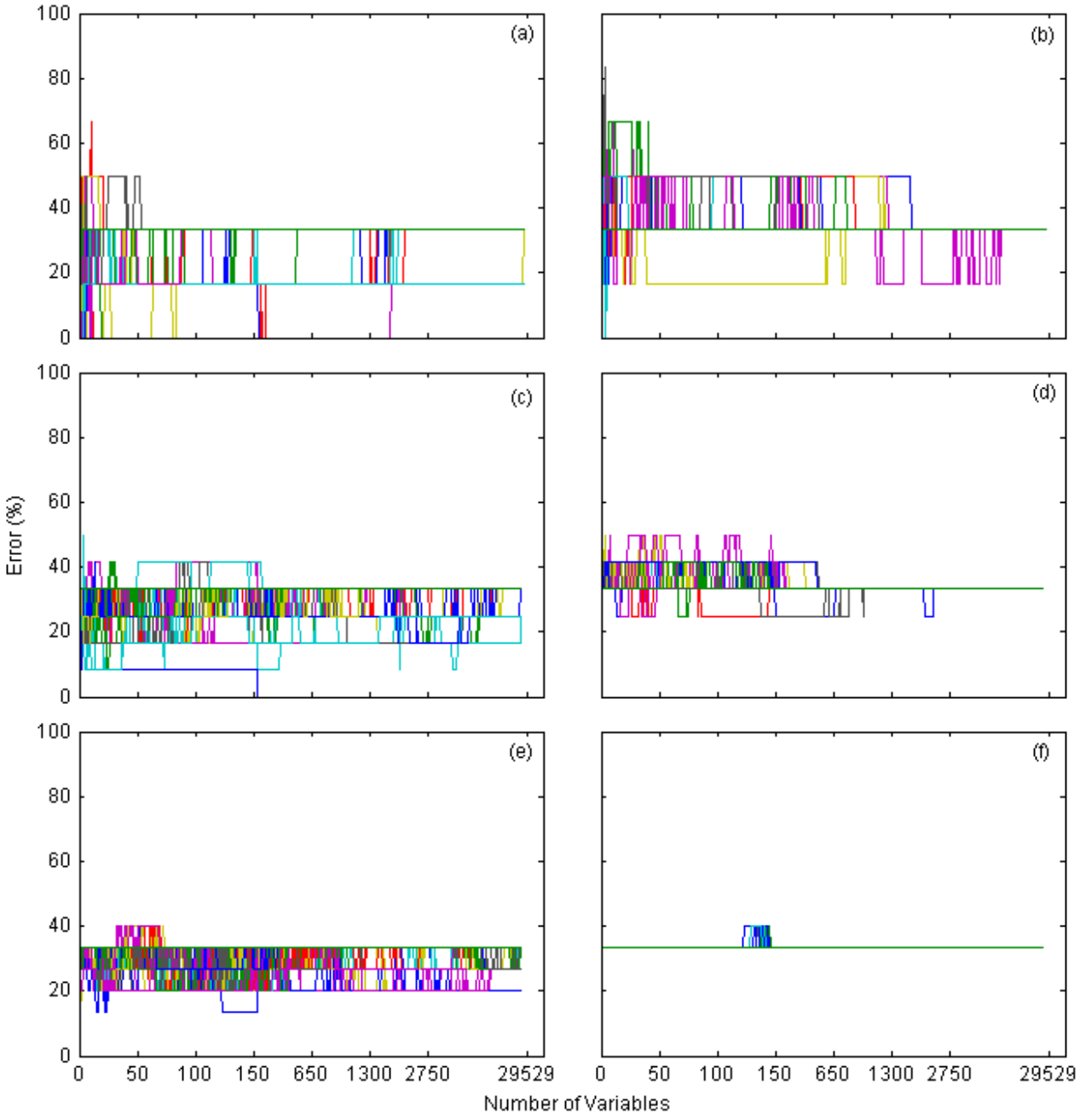

Figure S6. Error plot of the PLSDA model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets).

Methods: *mw*-test				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	0	0	0	0
1b	96	13	366	24
1c	108	108	274	274
2a	0	0	0	0
2b	22	2	34	3
2c	41	41	52	52
Methods: *t*-test				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	1	0	22	0
1b	101	18	348	23
1c	90	90	168	168
2a	1	0	17	0
2b	7	0	17	0
2c	39	39	46	46
Methods: NSC				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	141	1	3352	1
1b	87	47	143	53
1c	59	55	69	65
2a	137	6	10262	6
2b	49	25	82	25
2c	42	36	47	38

Methods: PCDA				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	151	0	28748	0
1b	134	0	857	0
1c	12	1	14	1
2a	151	0	29043	0
2b	89	0	394	0
2c	72	0	142	0
Methods: PLSDA				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	22	2	49	2
1b	77	2	425	2
1c	60	2	221	2
2a	46	2	1298	2
2b	51	2	2044	2
2c	7	2	12	2
Methods: SVM				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	89	0	8428	2
1b	75	0	6356	70
1c	32	1	3236	84
2a	116	0	10756	4
2b	53	0	4331	7
2c	37	0	2976	33

Table S2. Overview of the performance of different methods based on the ratio between unique true positives (Unique TP; selected at least once) and common true positives (Common TP; selected each time). The stability of the delivered feature set can be seen by comparing the number of unique features to the number of common features selected across each of the 100 repetitions (except for the *mww*-test and the *t*-test on data sets 1c and 2c, where repetitions were not possible, since all samples were used). Unique True Positive (Unique-TP) is a spiked-peptide-related feature that is selected at least once in 100 repetitions. Common True Positive (Common TP) a spiked-peptide-related feature that is always selected in each repetition. Unique Positive (Unique-P) is any feature that is included in a selected feature set at least once in 100 repetitions. Common Positive (Common P) is any feature that is always selected in each repetition.