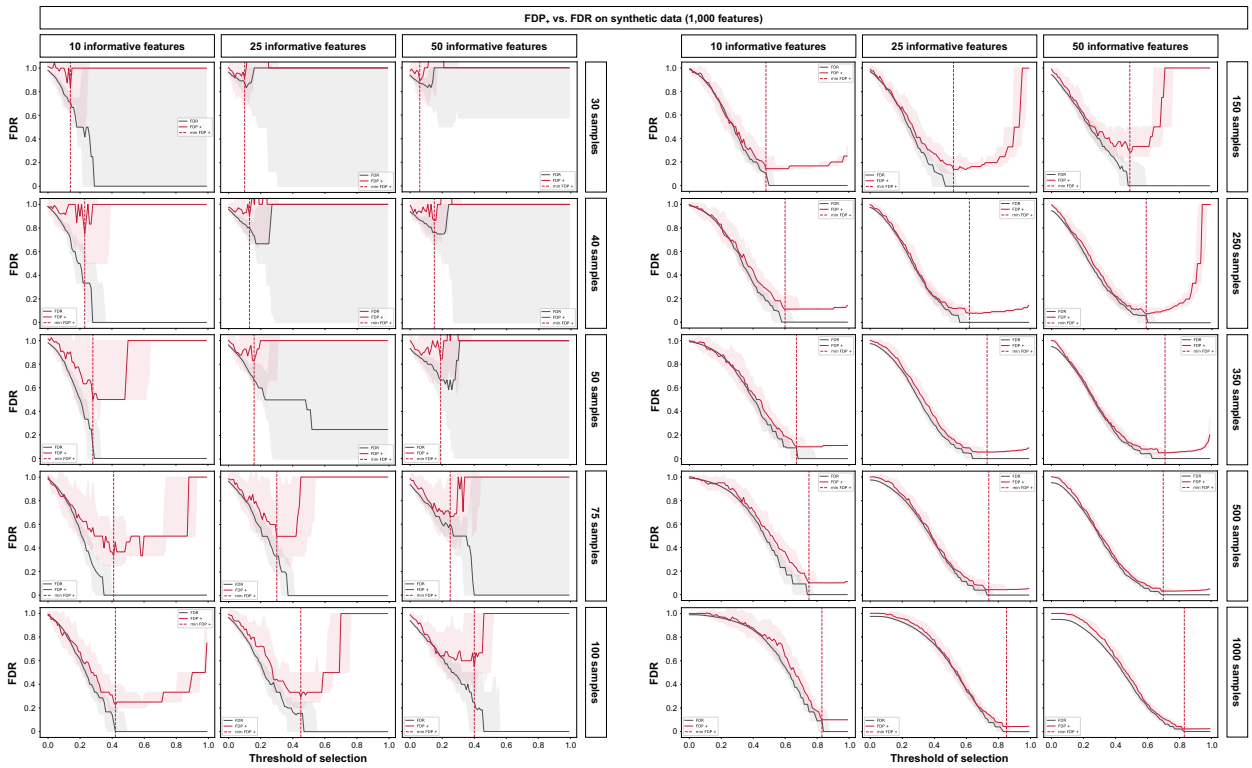
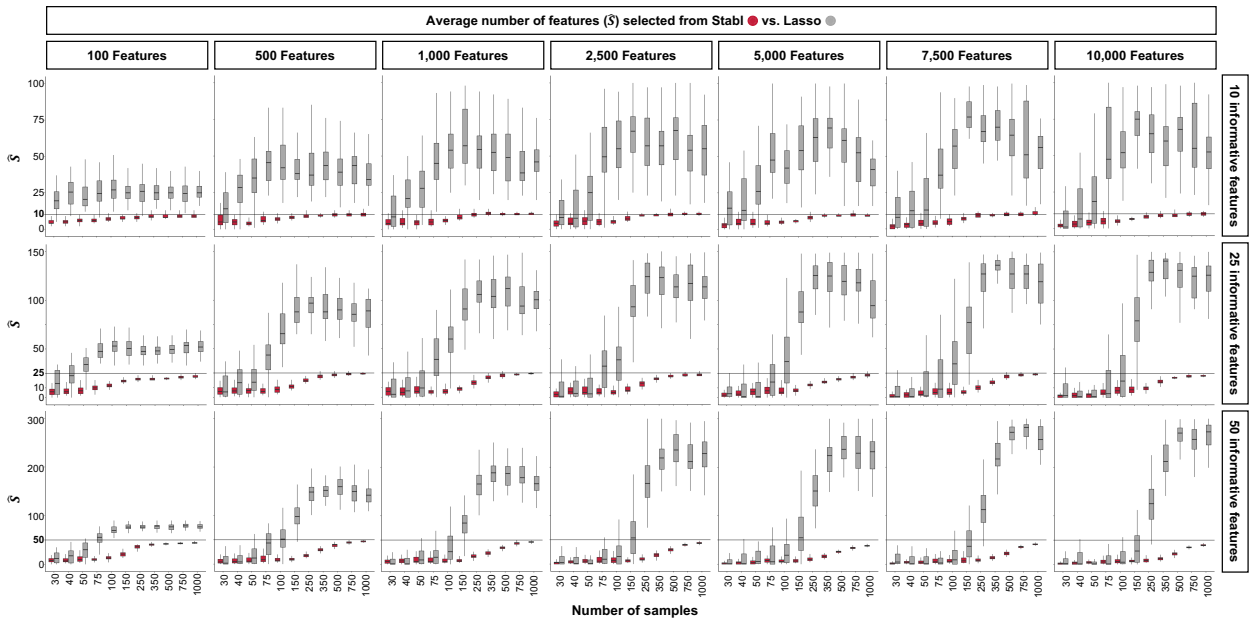
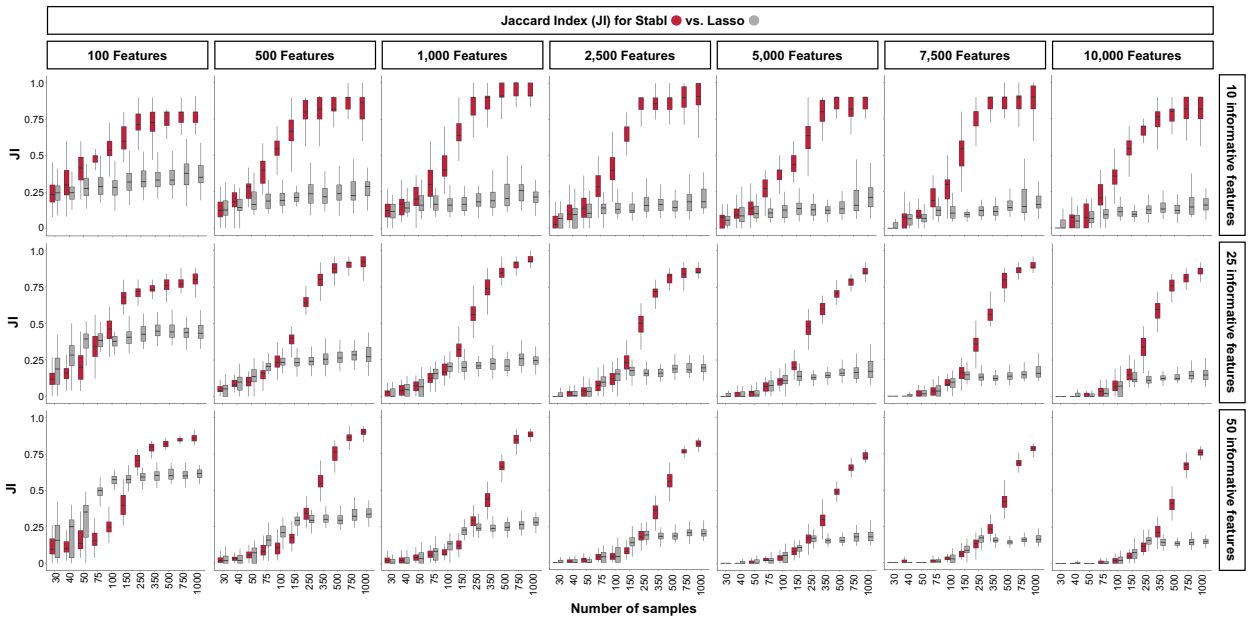
Stabl: sparse and reliable biomarker discovery in predictive modeling of high-dimensional omic data – Extended data



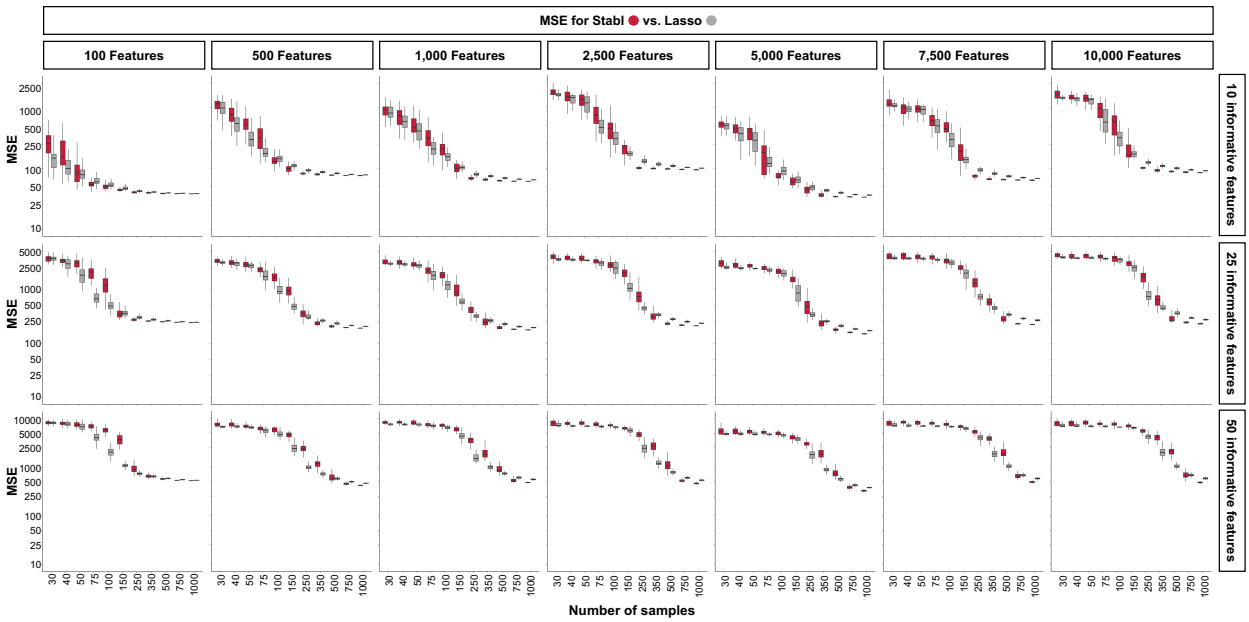
Extended Data Figure S1 | Comparison of FDP$_+$ *vs.* true FDR in synthetic dataset benchmarking. On the generated synthetic dataset, the surrogate for the false discovery proportion (FDP$_+$) and the experimental False Discovery Rate (FDR) were assessed for different dataset sizes ranging from n = 30 to 1,000 samples with 10 (left panels), 25 (middle panels), or 50 (right panels) informative features. The false discovery proportion (red line) and the experimental FDR (black line) are shown as a function of the frequency threshold. The selected reliability threshold (θ, red dotted line) varied across conditions.



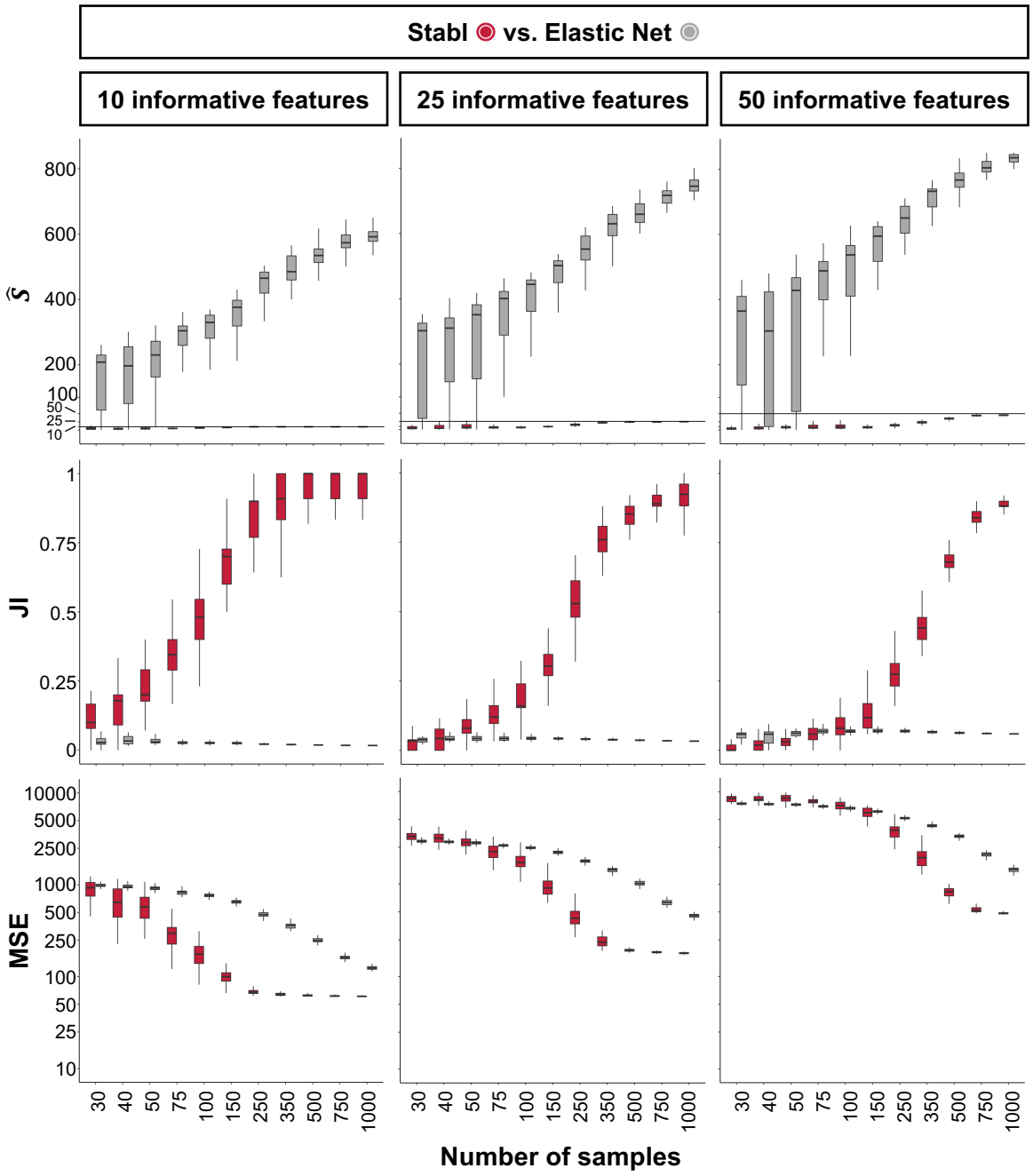
Extended Data Figure S2 | Comparison of Stabl and Lasso sparsity performance on synthetic data. Sparsity (average number of selected features, $\widehat{S}$) of Stabl (red box plots) and Lasso (grey box plots) as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for $p = 100$, 500, 1,000, 2,500, 5,000 and 10,000 total number of features. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



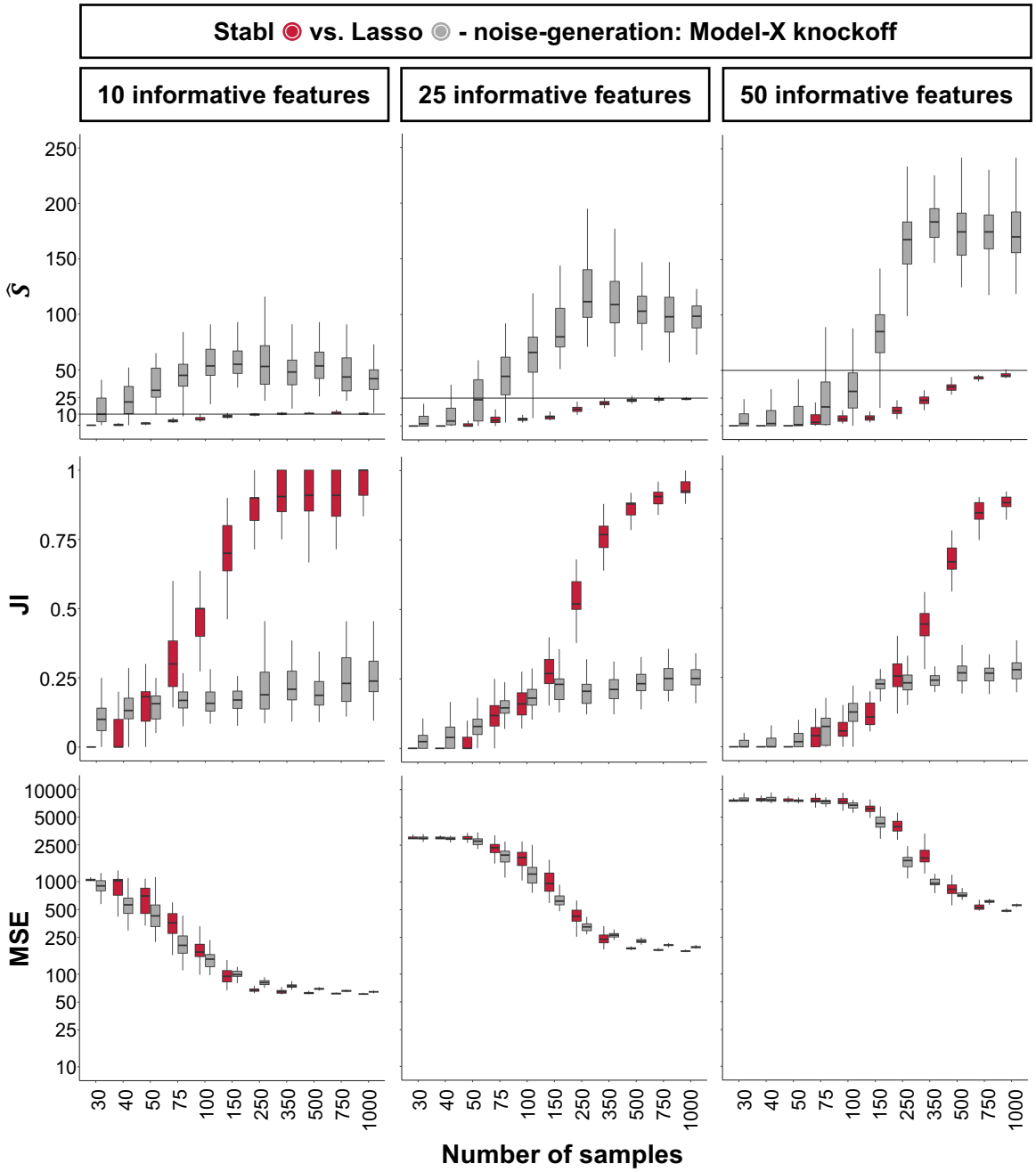
Extended Data Figure S3 | Comparison of Stabl and Lasso reliability performance on synthetic data. Reliability (Jaccard Index, JI, comparing the true set of informative features to the selected features) of Stabl (red box plots) and Lasso (grey box plots) as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for 100, 500, 1,000, 2,500, 5,000 and 10,000 total number of features. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



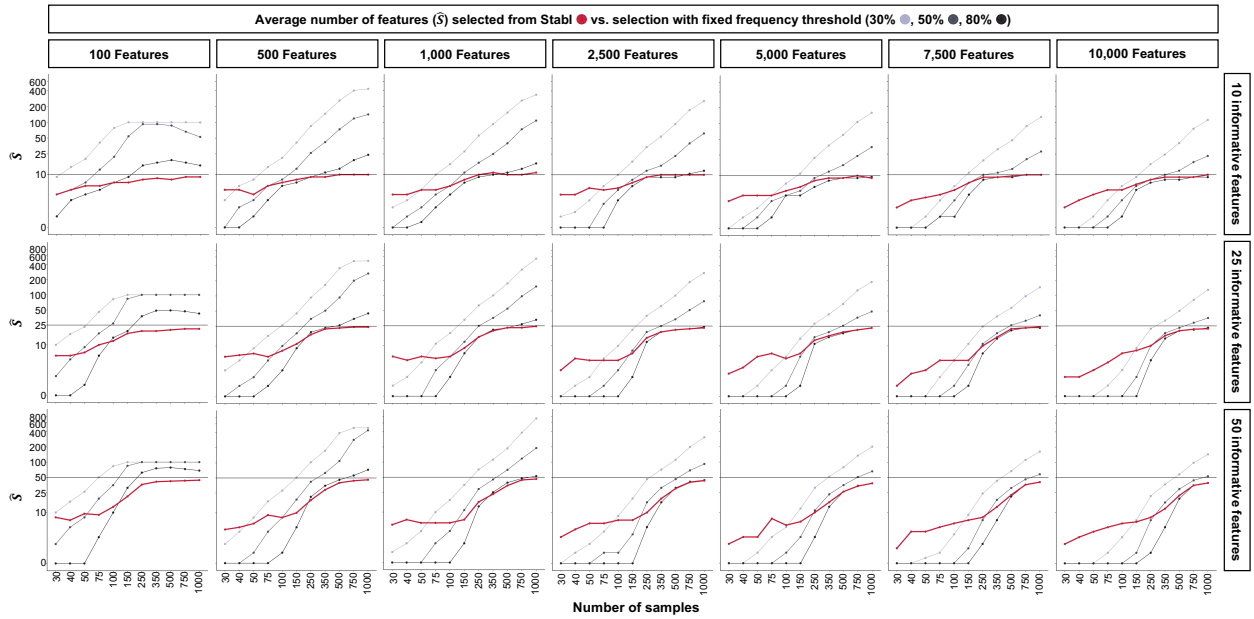
Extended Data Figure S4 | Comparison of Stabl and Lasso predictivity performance on synthetic data. Predictivity (mean squared error, MSE) of Stabl (red box plots) and Lasso (grey box plots) as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for 100, 500, 1,000, 2,500, 5,000 and 10,000 total number of features, evaluated using. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



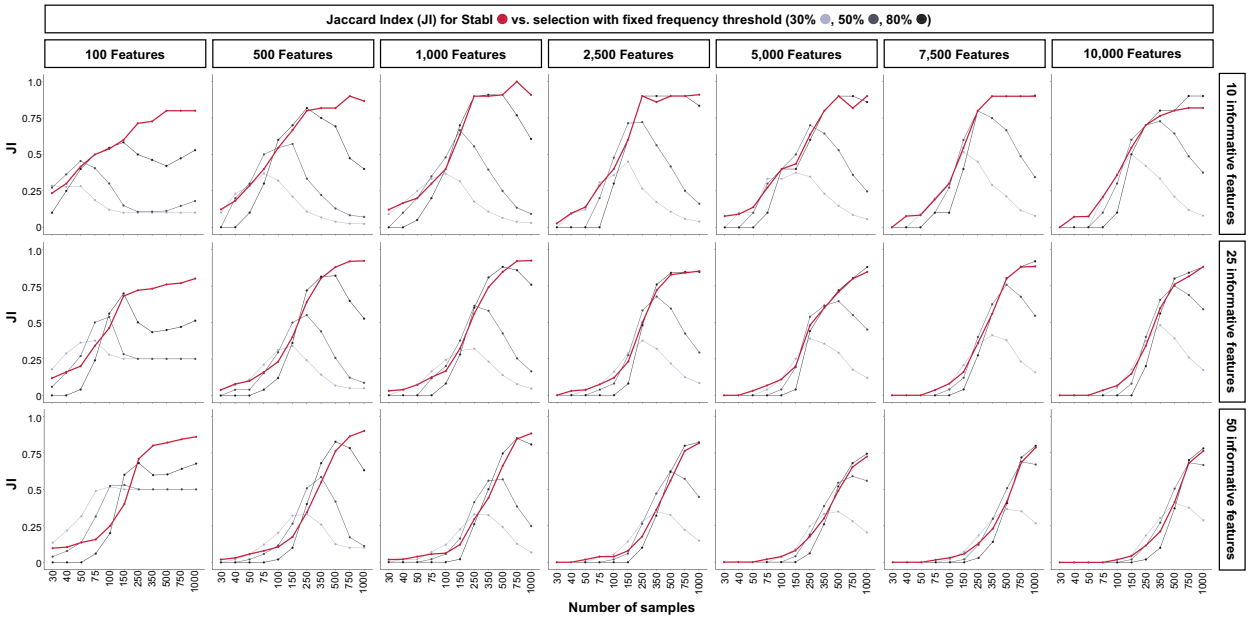
Extended Data Figure S5 | Comparison of Stabl and Elastic Net (EN) sparsity, reliability and predictivity performances on synthetic data. Sparsity (number of selected features, $\hat{S}$, upper panels), reliability (Jaccard index, JI, middle panels), and predictivity performances (mean squared error, MSE, lower panels) of Stabl (red box plots) and EN (grey box plots) as a function of the number of samples (n, x-axis) for 10 (left panels), 25 (middle panels), or 50 (right panels) informative features. A synthetic dataset consisting of 50,000 samples x 1,000 features was generated. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



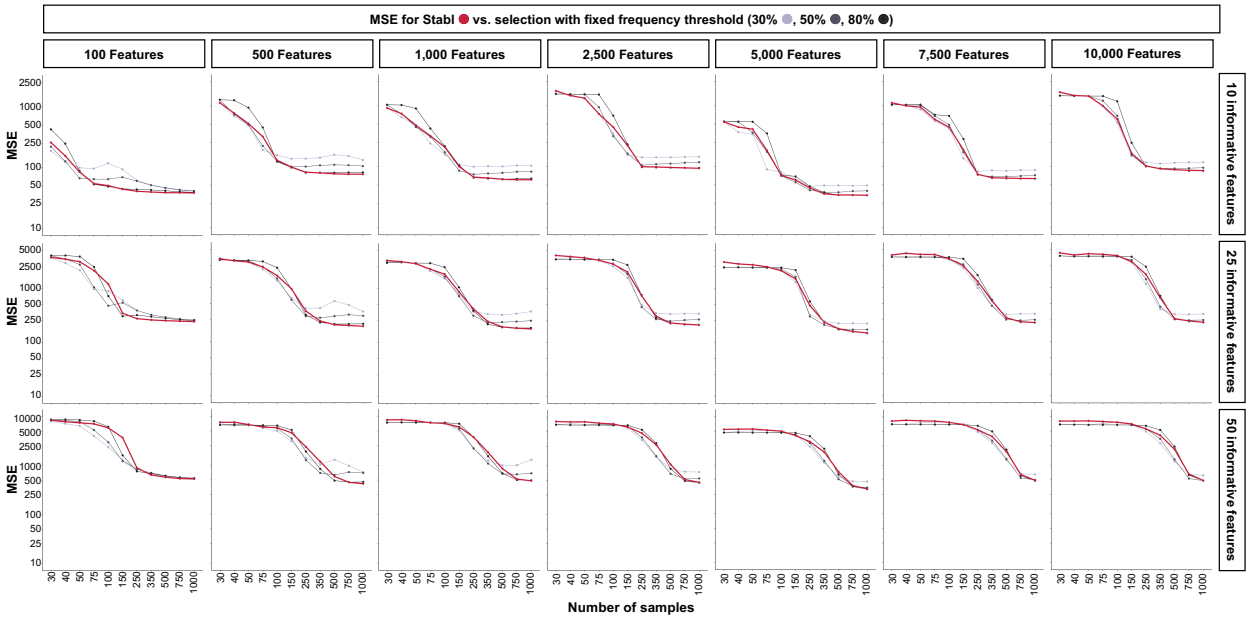
Extended Data Figure S6 | Comparison of Stabl and Lasso sparsity, reliability and predictivity performances on synthetic data using Model-X knockoffs. Sparsity (number of selected features, $\hat{S}$, upper panels), reliability (Jaccard index, JI, middle panels), and predictivity performances (mean squared error, MSE, lower panels) of Stabl (red box plots) and EN (grey box plots) as a function of the number of samples (n, x-axis) for 10 (left panels), 25 (middle panels), or 50 (right panels) informative features. A synthetic dataset consisting of 50,000 samples x 1,000 features was generated. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



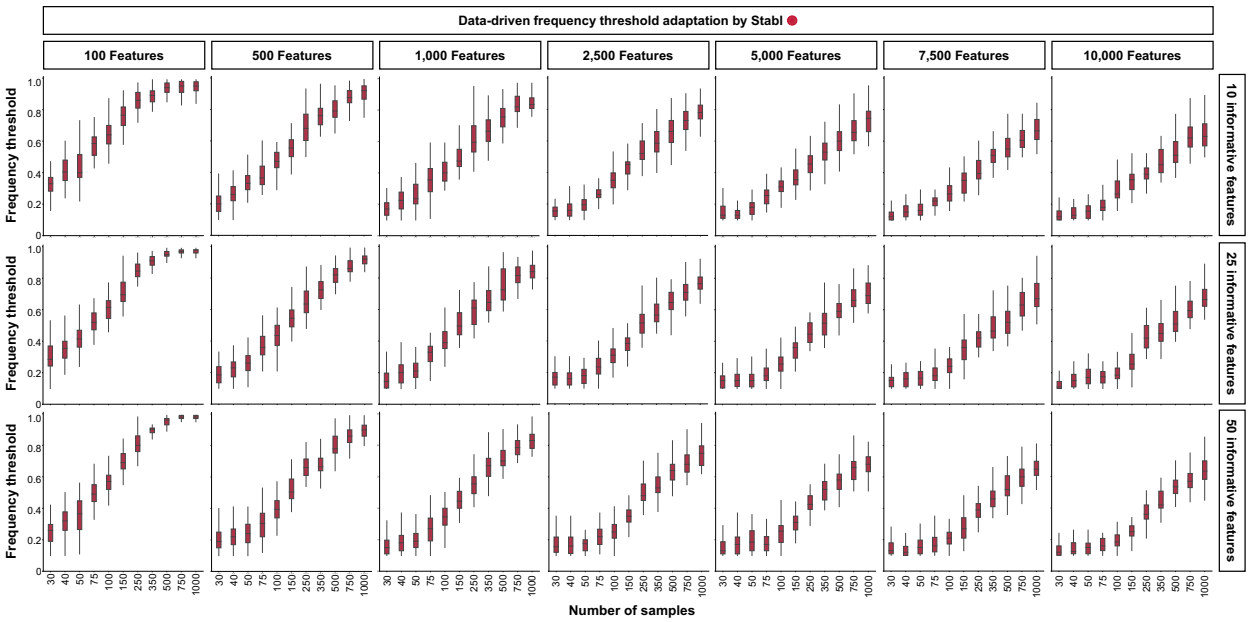
Extended Data Figure S7 | Comparison of Stabl and selection with fixed frequency threshold sparsity performance on synthetic data. Sparsity (average number of selected features, $\hat{S}$) of Stabl (red box plots) and fixed frequency threshold (i.e., SS) of 30% (light grey lines), 50% (Lasso, dark grey lines), or 80% (black lines) as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for 100, 500, 1,000, 2,500, 5,000 and 10,000 total number of features. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



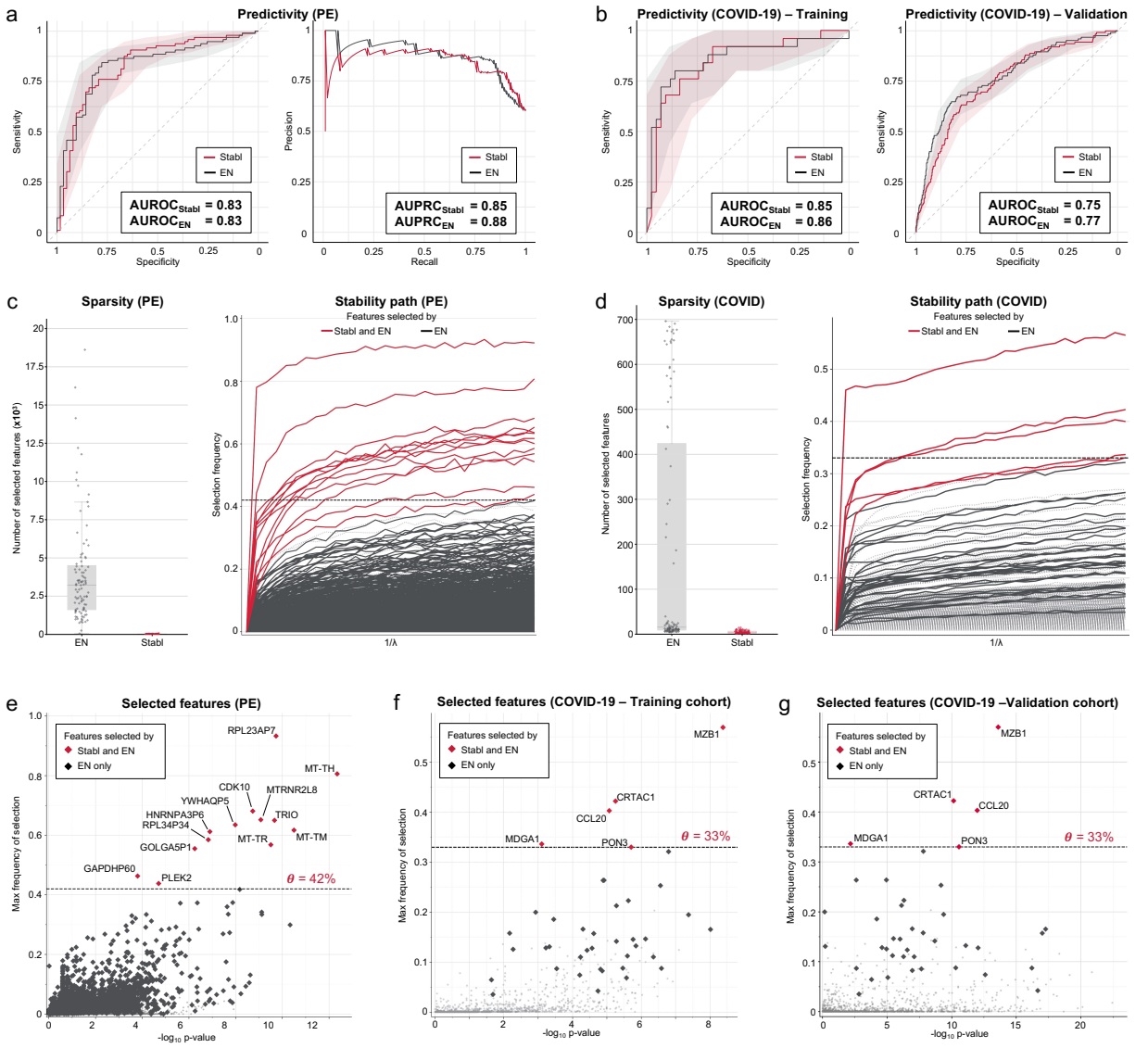
Extended Data Figure S8 | Comparison of Stabl and selection with fixed frequency threshold reliability performance on synthetic data. Reliability (Jaccard Index, JI, comparing the true set of informative features to the selected features) of Stabl (red box plots) and fixed frequency threshold (i.e., SS) of 30% (light grey lines), 50% (Lasso, dark grey lines), or 80% (black lines) as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for 100, 500, 1,000, 2,500, 5,000 and 10,000 total number of features. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



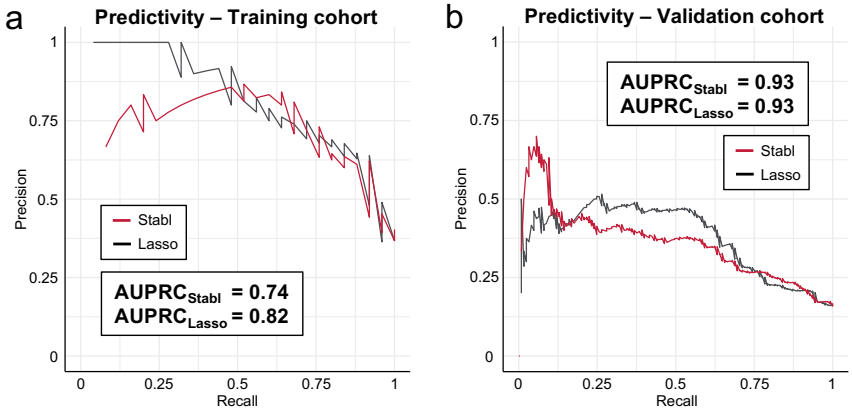
Extended Data Figure S9 | Comparison of Stabl and selection with fixed frequency threshold predictivity performance on synthetic data. Predictivity (mean squared error, MSE) of Stabl (red box plots) and fixed frequency threshold (i.e., SS) of 30% (light grey lines), 50% (Lasso, dark grey lines), or 80% (black lines) as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for 100, 500, 1,000, 2,500, 5,000 and 10,000 total number of features, evaluated using. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



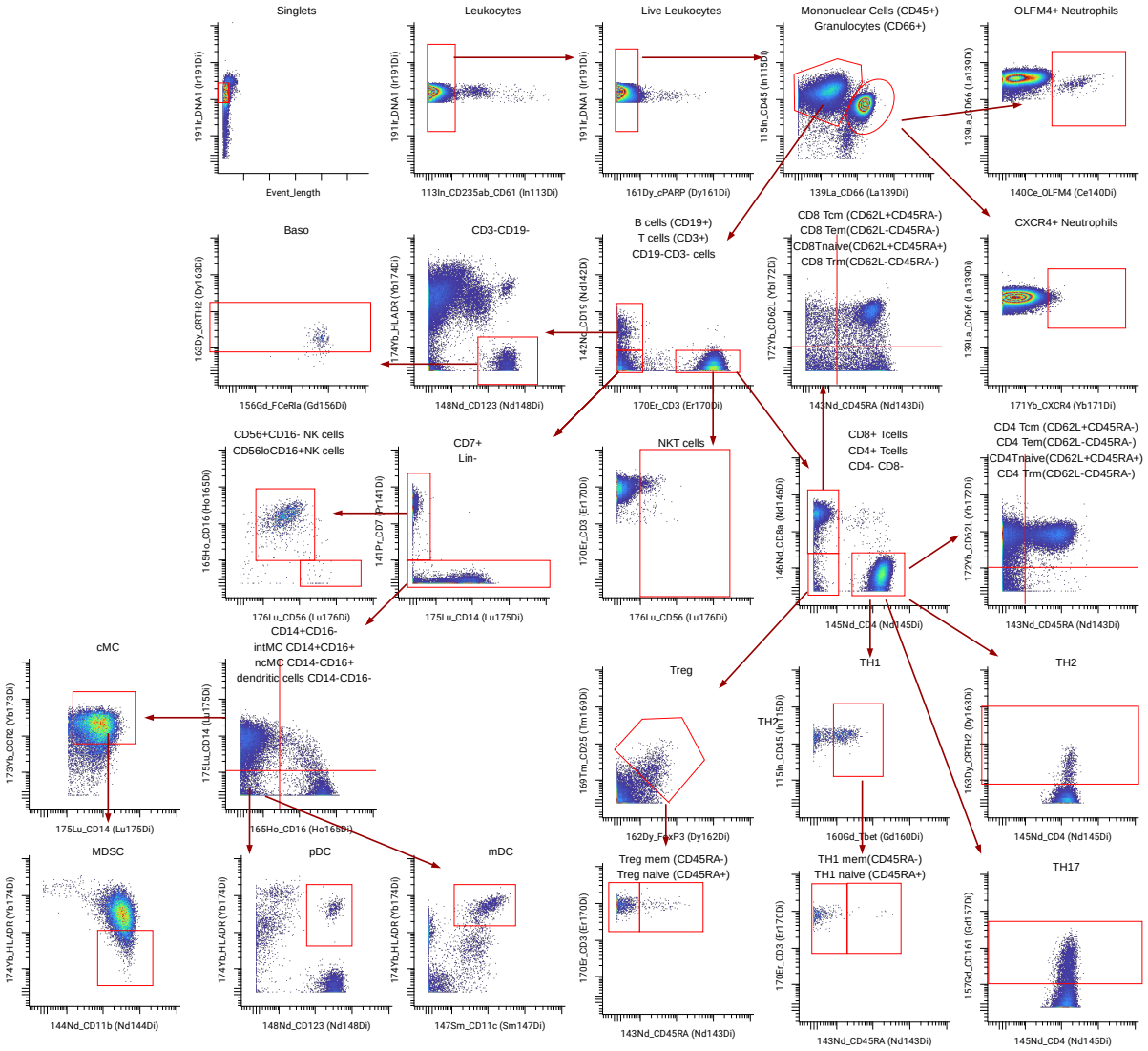
Extended Data Figure S10 | Reliability threshold variation with the number of samples. The reliability threshold chosen by Stabl (θ) is shown as a function of the number of samples (n, x-axis) for 10 (upper panels), 25 (middle panels), or 50 (lower panels) informative features for 100, 500, 1,000, 2,500, 5,000 and 10,000 total number of features. A synthetic dataset consisting of 50,000 samples was generated for each total number of features. 40,000 samples are held out for validation. Out of the remaining 10,000, 50 sets of sizes ranging from 30 to 1,000 are drawn randomly.



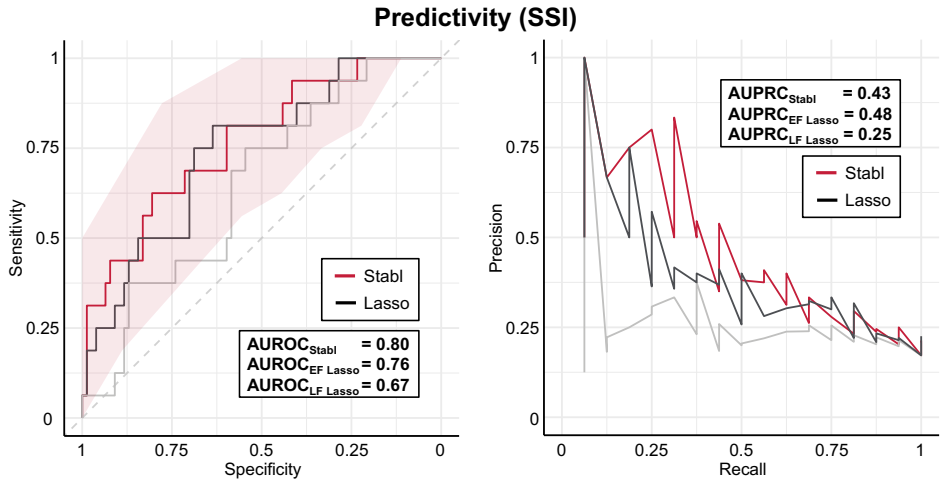
Extended Data Figure S11 | Performance of Stabl compared to EN on transcriptomic (Preeclampsia, PE) and proteomic (COVID-19) datasets. **a,** Predictivity performance of Stabl and EN for the PE datasets. *AUROC$_{Stabl}$ = 0.83 [0.76, 0.90]*, *AUROC$_{EN}$ = 0.83 [0.76, 0.89] (p-value = 0.46, Bootstrap test); AUPRC$_{Stabl}$ = 0.85 [0.77, 0.93]*, *AUPRC$_{EN}$ = 0.88 [0.81, 0.94] (p-value = 0.28).* **b,** AUROC comparing predictive performance of Stabl and EN on training (left panel) and validation (right panel) cohorts for the COVID-19 dataset. Training: *AUROC$_{Stabl}$ = 0.85 [0.74, 0.94]*, *AUROC$_{EN}$ = 0.86 [0.75, 0.95] (p-value = 0.43)*. Validation: *AUROC$_{Stabl}$ = 0.75 [0.71, 0.79]*, *AUROC$_{EN}$ = 0.77 [0.73, 0.82] (p-value = 0.28)*. **c-d,** Left panels. Sparsity performances for the PE (**c**, *median$_{Stabl}$ = 11.0, IQR = [7.8,16.0], median$_{EN}$ = 3220.0, IQR = [1620.8,4494], p-value < 1e-16),* and COVID-19 (**d**, *median$_{Stabl}$ = 7.0, IQR = [4.8,13.0], median$_{EN}$ = 31.5, IQR = [18.8,847.8], p-value < 1e-16)* datasets assessed by the number of features ultimately selected by the algorithm across cross-validation iterations. Right panels. Stability path graphs showing the regularization parameter against the selection frequency. The reliability threshold (θ), is indicated (dotted line) **e-f**, Volcano plots depicting the reliability performances of Stabl and EN for the PE (**e**) and COVID-19 (**f**) datasets. The maximum frequency of selection of each feature is plotted against the -log10 p-value using a univariate Mann-Whitney test. Features selected by Stabl/EN only are colored in red/black respectively. Features selected by Stabl are labeled.



Extended Data Figure S12 | Predictivity of Stabl and Lasso for the training and validation cohort of the COVID-19 dataset. AUPRC comparing the predictive performance of Stabl and Lasso on the (**a**) training *$AUPRC_{Stabl}$ = 0.74 [0.54, 0.92], $AUPRC_{Lasso}$ = 0.82 [0.67, 0.93] (p=0.24)* and (**b**) validation *$AUPRC_{Stabl}$ = 0.93 [0.91, 0.95], $AUPRC_{Lasso}$ = 0.93 [0.90, 0.95] (p-value = 0.38)* cohorts for the COVID-19 dataset.



Extended Data Figure S13 | Gating strategy for mass cytometry analyses (SSI dataset). Live, non-erythroid cell populations were used for analysis.



Extended Data Figure S14 | Predictive performance of Stabl, Early Fusion and Late Fusion Lasso for the SSI dataset. Predictive performance of Stabl, Early Fusion (EF) Lasso, and Late Fusion (LF) Lasso is represented as AUROC (left panel), *$AUROC_{Stabl}$ = 0.796 [0.687, 0.891], $AUROC_{EF\ Lasso}$ = 0.763 [0.627, 0.888] (p=0.38), and $AUROC_{LF\ Lasso}$ = 0.670 [0.536, 0.805] (p=0.13),* and AUPRC (right panel), *$AUPRC_{Stabl}$ = 0.432 [0.243, 0.690], $AUPRC_{EF\ Lasso}$ = 0.475 [0.276, 0.704] (p=0.337),* and *$AUPRC_{LF\ Lasso}$ = 0.249 [0.135, 0.469] (p=0.054)* for the SSI dataset.

Feature	Selected by Lasso	Selected by Stabl	Training cohort Mann-Whitney p-values	Validation cohort Mann-Whitney p-values
CCL20	TRUE	TRUE	8.5E-06	1.1E-12
CRTAC1	TRUE	TRUE	5.6E-06	7.2E-11
DSG4	TRUE	FALSE	2.4E-06	5.0E-07
FOSB	TRUE	FALSE	1.2E-05	1.1E-05
HDGF	TRUE	FALSE	3.8E-05	2.8E-07
LGALS1	TRUE	FALSE	1.6E-07	1.6E-08
MDGA1	TRUE	TRUE	7.9E-04	6.9E-03
MUC13	TRUE	FALSE	3.5E-04	6.4E-05
MZB1	TRUE	TRUE	4.2E-09	2.5E-14
PON3	TRUE	TRUE	2.0E-06	2.7E-11
SPINT1	TRUE	FALSE	5.8E-05	8.1E-08

Extended Data Table S1: Univariate p-values for clinical case study 2: COVID-19.

Dataset: Preeclampsia	Predictivity (ROC AUC)	Predictivity (PR AUC)	Sparsity (number of features)
Stabl	0.829 [0.755, 0.898]	0.850 [0.766, 0.931]	11.000 [7.750, 16.000]
Stability selection (threshold 0.3)	0.832 [0.764, 0.894] (p=0.383)	0.862 [0.778, 0.930] (p=0.281)	10.500 [9.000, 12.250] (p=4.683e-01)
Stability selection (threshold 0.5)	0.777 [0.694, 0.841] (p=0.014)	0.856 [0.783, 0.913] (p=0.424)	2.000 [2.000, 3.000] (p=6.495e-33)
Stability selection (threshold 0.8)	0.464 [0.432, 0.486] (p=0.0)	0.586 [0.507, 0.667] (p=0.0)	0.000 [0.000, 0.250] (p=3.266e-36)
Dataset: COVID-19 training	**Predictivity (ROC AUC)**	**Predictivity (PR AUC)**	**Sparsity (number of features)**
Stabl	0.847 [0.736, 0.935]	0.736 [0.543, 0.918]	7.000 [4.750, 13.000]
Stability selection (threshold 0.3)	0.851 [0.746, 0.939] (p=0.463)	0.739 [0.577, 0.923] (p=0.488)	3.000 [2.000, 4.000] (p=4.864e-18)
Stability selection (threshold 0.5)	0.491 [0.442, 0.532] (p=0.0)	0.364 [0.254, 0.471] (p=0.0)	0.000 [0.000, 1.000] (p=1.026e-35)
Stability selection (threshold 0.8)	0.500 [0.500, 0.500] (p=0.0)	0.368 [0.250, 0.485] (p=0.0)	0.000 [0.000, 0.000] (p=5.108e-39)
Dataset: COVID-19 validation	**Predictivity (ROC AUC)**	**Predictivity (PR AUC)**	
Stabl	0.748 [0.706, 0.789]	0.932 [0.912, 0.952]	
Stability selection (threshold 0.3)	0.778 [0.734, 0.820] (p=0.056)	0.946 [0.930, 0.962] (p=0.011)	
Stability selection (threshold 0.5)	0.722 [0.670, 0.768] (p=0.252)	0.925 [0.901, 0.947] (p=0.289)	
Stability selection (threshold 0.8)	0.500 [0.500, 0.500] (p=0.0)	0.841 [0.814, 0.867] (p=0.0)	

Extended Data Table S2: Predictivity and sparsity comparison for Stabl vs. Stability Selection on single omic datasets.

Features	Coefficients
TRIO	-0.03
PLEK2	-0.99
CDK10	0.40
MT-TM	-0.59
MT-TR	-2.21
MT-TH	3.51
HNRNPA3P6	-0.76
RPL34P34	-0.76
YWHAQP5	0.52
RPL23AP7	-1.51
GAPDHP60	-1.03
GOLGA5P1	-0.82
MTRNR2L8	0.38

Extended Data Table S3: Features selected by Stabl for clinical case study 1: Preeclampsia (PE).

Features	Coefficients
CCL20	2.35
CRTAC1	-0.69
MDGA1	-2.06
MZB1	2.31
PON3	-1.63

Extended Data Table S4: Features selected by Stabl for clinical case study 2: COVID-19.

Feature	Selected by EF Lasso	Selected by Stabl	Pearson R (p-val)	Spearman R (p-val)
AMGO2	TRUE	FALSE	2.7E-01	3.2E-01
Activin.A	TRUE	TRUE	8.0E-23	2.9E-24
Angiogenin	TRUE	TRUE	3.9E-06	5.5E-06
Angiopoietin.2	TRUE	TRUE	4.0E-14	1.2E-14
Angiotensinogen	TRUE	FALSE	4.6E-01	5.5E-01
BLC	TRUE	FALSE	3.0E-01	5.1E-01
CCR5posCCR2posCD8Tem_CREB_IFNa	TRUE	FALSE	3.2E-03	1.1E-03
CD69negCD56loCD16negNK_NFkB_LPS	TRUE	FALSE	4.3E-03	2.2E-03
CD69negCD56loCD16negNK_S6_LPS	TRUE	FALSE	3.8E-02	2.8E-02
CD69posCD56loCD16negNK_CREB_LPS	TRUE	FALSE	2.1E-03	1.7E-02
CD8Tcells_MAPKAPK2_IFNa	TRUE	FALSE	1.0E-05	3.7E-06
FCRL3	TRUE	FALSE	4.0E-02	1.1E-02
GDF2	TRUE	FALSE	4.6E-02	2.2E-01
GFRa.1	TRUE	FALSE	6.5E-04	6.8E-04
Gelsolin	TRUE	FALSE	6.4E-04	1.9E-03
HCE000483	TRUE	FALSE	6.0E-01	5.6E-01
HCE004359	TRUE	FALSE	1.5E-01	2.2E-01
HCG	TRUE	TRUE	1.8E-01	1.1E-01
IL.1.R4	TRUE	TRUE	5.9E-30	8.0E-32
IL.23.R	TRUE	FALSE	1.6E-04	7.9E-03
MIP.1a	TRUE	FALSE	4.8E-01	4.7E-01
MMP.12	TRUE	TRUE	7.4E-12	2.7E-11
MP2K2	TRUE	FALSE	3.3E-01	3.6E-01
NEGR1	TRUE	FALSE	4.6E-04	1.2E-03
NK_STAT1_IFNa	TRUE	FALSE	1.4E-17	1.0E-16
NK_STAT3_IL246	TRUE	FALSE	6.4E-05	2.2E-05
Rb	TRUE	FALSE	3.2E-01	7.4E-02
SHPS1	TRUE	FALSE	6.6E-01	5.0E-01
SLPI	TRUE	TRUE	2.0E-08	1.7E-09
SPINT2	TRUE	TRUE	3.9E-04	5.8E-04
SPTA2	TRUE	FALSE	8.5E-01	3.6E-01
Siglec.6	TRUE	TRUE	1.3E-27	4.4E-26
Thyroxine.Binding.Globulin	TRUE	FALSE	3.5E-03	1.0E-02
Vitronectin	TRUE	FALSE	1.6E-01	8.0E-02
X147.0492_10.2	TRUE	FALSE	5.7E-06	1.7E-07
X151.0615_2.6	TRUE	FALSE	1.6E-07	4.0E-12
X153.0523_0.9	TRUE	FALSE	2.0E-01	4.6E-01
X177.0556_5.4	TRUE	FALSE	6.0E-07	2.3E-06
X187.1442_11.5	TRUE	FALSE	1.6E-01	8.9E-01
X194.046_4.4	TRUE	FALSE	7.8E-03	7.5E-03
X215.0328_0.5	TRUE	FALSE	4.0E-03	3.1E-03
X217.0294_0.7	TRUE	FALSE	3.7E-02	2.6E-02
X218.1501_14.9	TRUE	FALSE	1.7E-07	5.9E-08
X221.1296_4	TRUE	FALSE	1.2E-02	3.2E-02
X258.0748_5.9	TRUE	FALSE	3.3E-03	1.8E-04
X266.1389_6.2	TRUE	FALSE	4.7E-08	1.8E-06
X331.2264_8.1	TRUE	TRUE	5.3E-25	1.5E-29
X335.2226_9.3	TRUE	TRUE	1.7E-04	2.0E-04
X349.2371_7.5	TRUE	TRUE	2.5E-27	5.8E-29
X367.1495_0.8	TRUE	TRUE	2.1E-11	6.6E-10
X371.1895_10.6	TRUE	TRUE	9.0E-11	2.1E-13
X375.2208_3.3	TRUE	FALSE	2.1E-02	1.2E-02
X435.274_13	TRUE	FALSE	6.2E-04	3.1E-04
X441.264_9.5	TRUE	TRUE	2.5E-19	1.9E-22
X441.301_2.6	TRUE	FALSE	1.6E-02	3.2E-02
X443.208_8.4	TRUE	TRUE	1.6E-28	5.7E-34
X459.3107_2.1	TRUE	FALSE	4.5E-03	1.3E-05
X509.2944_0.9	TRUE	FALSE	2.1E-01	1.2E-01
X72.9931_5	TRUE	FALSE	1.3E-02	1.6E-02
X99.0088_7.8	TRUE	FALSE	3.4E-03	5.1E-03
a1.Antitrypsin	TRUE	FALSE	5.0E-04	5.6E-05
calgranulin.B	TRUE	FALSE	2.3E-04	8.9E-04
eIF.5	TRUE	FALSE	4.8E-03	3.2E-03

Extended Data Table S5: Univariate p-values for clinical case study 3: Time to labor.

Features	Coefficients
CD69negCD56loCD16negNK_STAT1_IFNa	4.16
Granulocytes	0.14
Angiopoietin.2	-2.73
Cystatin.C	-1.58
Siglec.6	3.94
Activin.A	4.11
SPINT2	1.79
PAI.1	0.34
Antithrombin.III	-1.23
sTie.2	-3.22
uPA	1.05
IL.1.R4	3.10
SLPI	3.78
MMP.12	-3.21
SLIK5	1.24
VEGF121	0.34
Angiogenin	2.35
HCG	-4.84
PLXB2	2.80
X193.0618_5.3	1.77
X331.2264_8.1	0.22
X331.2264_8.4	-6.65
X335.2226_9.3	-3.03
X349.2371_7.5	6.04
X361.2017_7.1	-0.85
X367.1495_0.8	1.91
X371.1895_10.6	2.13
X441.264_9.5	2.00
X443.208_8.4	8.96
X461.2387_8.7	-1.96

Extended Data Table S6: Features selected by Stabl for clinical case study 3: Time to labor.

Patient characteristic	No Surgical Site Infection, 83% (n=77)	Surgical Site Infection, 17% (n=16)
Age (mean +/- SD)	59.2 +/- 13.9	58.8 +/- 14.2
Male, % (n)	49 (38)	50 (8)
Ethnicity, black, % (n)	1 (1)	0 (0)
Bmi	28.3 +/- 6.5	26.6 +/- 4.6
Surgical indication		
Cancer	58% (45)	25% (4)
Inflammatory bowel disease	6% (5)	12.5% (2)
Other	35% (27)	62.5% (10)
Type of surgery		
Colectomy	64% (49)	56% (9)
Small bowels	3% (2)	6% (1)
Other	34% (26)	37% (6)
Surgical approach		
Minimally invasive	32% (25)	32% (5)
Open surgery	68% (52)	68% (11)
Operative duration, min, mean±SD	213 ±132	234 ±111
ASA classification, median	3	2.5

Extended Data Table S7: Clinical information for clinical case study 4: surgical site infections (SSI).

Features	Coefficients
unstim_CD4Trm_Frequency	4.90
unstim_Tregnaive_Frequency	0.84
IL246_Granulocytes_154Sm_STAT3	-0.02
IL246_Th1naive_153Eu_STAT1	1.89
SAMHD1	2.69
ARL11	3.28
LDLR	3.37
RNASET2	1.65
HSPH1	2.22
WWOX	2.82
MTHFD2	0.67
OGG1	-4.60
ITGAV|ITGB3	-2.78
CNTN3	-0.17
CTNNA3	1.35
IL1B	3.68
CCL3	3.69
FCGR1A	1.88
MAPK8	2.49
CSF2	1.16
PPIB	-0.31
BIRC3	-2.88
IL18	1.92
ASAH1	1.93
ALK	3.04
TNFSF10	1.77
RAP2A	0.37
TOPBP1	0.75
IRF6	0.15

Extended Data Table S8: Features selected by Stabl for clinical case study 4: surgical site infections (SSI).