Journal of Dairy Science: link

Paper: Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods

M. Frizzarin,[1,2] I.C. Gormley,[1] D.P. Berry,[2] T.B. Murphy,[1] A. Casa,[1] A. Lynch,[1] and S. McParland[2]

[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland.

[2]Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy P61 P302, Co. Cork, Ireland

Supplemental material

Supplemental Table S1. Prediction performance[1] in the calibration (n=362) and cross-validation (n=120) data sets of alternative prediction models used to predict rennet coagulation time (RCT) expressed in minutes.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	5.76 (0.31)	0.59 (0.02)	6.550 (0.895)	0.125 (0.557)	0.49 (0.06)	1.81 (0.23)	0.88 (0.04)
PCR	6.88 (0.37)	0.42 (0.03)	7.209 (1.106)	0.025 (0.236)	0.37 (0.08)	1.66 (0.26)	0.91 (0.06)
PPR	3.29 (0.33)	0.86 (0.03)	8.823 (0.867)	-0.122 (0.860)	0.32 (0.10)	1.34 (0.18)	0.52 (0.04)
RR	5.61 (0.23)	0.61 (0.01)	6.543 (0.980)	0.068 (0.548)	0.48 (0.06)	1.81 (0.23)	0.90 (0.04)
LASSO	5.45 (0.39)	0.64 (0.03)	6.649 (0.989)	0.179 (0.668)	0.47 (0.04)	1.81 (0.23)	0.87 (0.04)
EN	6.84 (0.30)	0.44 (0.03)	7.163 (0.875)	0.123 (0.170)	0.37 (0.06)	1.81 (0.23)	1.00 (0.06)
Average	5.70 (0.28)	0.60 (0.01)	6.465 (0.947)	0.124 (0.461)	0.49 (0.06)	1.84 (0.27)	0.98 (0.05)
SSR	7.48 (0.39)	0.32 (0.03)	7.662 (1.057)	-0.017 (0.371)	0.28 (0.04)	1.55 (0.21)	1.04 (0.08)
RF	3.05 (0.15)	0.93 (0.00)	7.582 (1.010)	-0.202 (0.495)	0.30 (0.03)	1.56 (0.18)	0.96 (0.07)
Boosting	6.76 (0.30)	0.46 (0.01)	7.486 (1.031)	-0.061 (0.421)	0.31 (0.02)	1.58 (0.19)	1.08 (0.07)
NN	5.05 (0.22)	0.69 (0.01)	6.397 (0.692)	0.092 (0.541)	0.50 (0.03)	1.85 (0.21)	0.90 (0.04)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S2. Prediction performance[1] in the calibration (n=329) and cross-validation (n=110) data sets of alternative prediction models used to predict curd firming time (k20) expressed in minutes.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	2.58 (0.08)	0.43 (0.02)	2.815 (0.244)	-0.028 (0.238)	0.34 (0.07)	1.68 (0.21)	0.86 (0.06)
PCR	2.58 (0.09)	0.43 (0.02)	2.811 (0.230)	-0.018 (0.227)	0.34 (0.06)	1.68 (0.20)	0.87 (0.06)
PPR	0.83 (0.10)	0.94 (0.02)	4.400 (0.217)	-0.569 (0.465)	0.18 (0.05)	1.07 (0.13)	0.32 (0.03)
RR	2.44 (0.20)	0.49 (0.06)	2.801 (0.305)	-0.018 (0.229)	0.34 (0.06)	1.68 (0.21)	0.86 (0.06)
LASSO	2.33 (0.21)	0.54 (0.07)	2.835 (0.256)	-0.036 (0.236)	0.33 (0.04)	1.68 (0.21)	0.83 (0.06)
EN	2.75 (0.05)	0.35 (0.02)	2.938 (0.141)	0.016 (0.343)	0.29 (0.09)	1.68 (0.21)	0.88 (0.07)
Average	2.46 (0.14)	0.49 (0.04)	2.781 (0.256)	-0.016 (0.261)	0.35 (0.06)	1.71 (0.21)	0.92 (0.06)
SSR	2.71 (0.08)	0.37 (0.02)	2.787 (0.282)	-0.018 (0.219)	0.34 (0.06)	1.70 (0.19)	1.04 (0.07)
RF	1.16 (0.04)	0.92 (0.01)	2.916 (0.212)	-0.108 (0.202)	0.29 (0.06)	1.62 (0.19)	0.84 (0.06)
Boosting	2.43 (0.07)	0.51 (0.02)	2.800 (0.272)	-0.006 (0.171)	0.34 (0.08)	1.69 (0.21)	0.98 (0.07)
NN	2.24 (0.20)	0.57 (0.05)	2.770 (0.280)	0.008 (0.239)	0.36 (0.06)	1.71 (0.18)	0.84 (0.05)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S3. Prediction performance[1] in the calibration (n=301) and cross-validation (n=100) data sets of alternative prediction models used to predict curd firmness at 30 min (a30) expressed in millimeters.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	11.30 (0.35)	0.48 (0.02)	12.722 (0.291)	0.006 (1.190)	0.36 (0.03)	1.77 (0.22)	0.83 (0.06)
PCR	11.97 (0.16)	0.41 (0.01)	12.776 (0.239)	0.233 (1.239)	0.35 (0.06)	1.77 (0.27)	0.90 (0.06)
PPR	5.19 (1.32)	0.88 (0.06)	19.095 (1.393)	-0.860 (1.963)	0.13 (0.03)	1.18 (0.17)	0.32 (0.04)
RR	11.23 (0.03)	0.49 (0.02)	12.495 (0.084)	0.203 (1.140)	0.37 (0.05)	1.77 (0.22)	0.92 (0.06)
LASSO	11.57 (0.46)	0.45 (0.04)	12.747 (0.198)	0.313 (1.003)	0.34 (0.05)	1.77 (0.22)	0.92 (0.06)
EN	12.01 (0.22)	0.41 (0.02)	12.716 (0.273)	0.188 (1.080)	0.35 (0.07)	1.77 (0.22)	0.98 (0.07)
Average	11.39 (0.19)	0.47 (0.01)	12.501 (0.090)	0.177 (1.060)	0.37 (0.05)	1.80 (0.25)	0.94 (0.06)
SSR	12.45 (0.09)	0.37 (0.03)	12.800 (0.192)	0.124 (0.819)	0.34 (0.07)	1.76 (0.26)	1.04 (0.07)
RF	5.37 (0.07)	0.92 (0.00)	13.353 (0.368)	0.101 (1.054)	0.29 (0.08)	1.69 (0.28)	0.87 (0.07)
Boosting	11.40 (0.14)	0.48 (0.02)	12.998 (0.413)	0.033 (0.640)	0.32 (0.09)	1.74 (0.28)	0.98 (0.07)
NN	10.81 (0.24)	0.52 (0.01)	12.659 (0.265)	0.204 (1.088)	0.36 (0.03)	1.78 (0.22)	0.88 (0.06)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S4. Prediction performance[1] in the calibration (n=359) and cross-validation (n=119) data sets of alternative prediction models used to predict curd firmness at 60 min (a60) expressed in millimeters.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	9.80 (0.33)	0.30 (0.03)	10.306 (0.898)	0.068 (0.460)	0.25 (0.10)	1.40 (0.21)	0.89 (0.07)
PCR	10.12 (0.32)	0.25 (0.04)	10.477 (1.142)	0.032 (0.534)	0.21 (0.08)	1.38 (0.19)	0.91 (0.08)
PPR	4.17 (0.24)	0.87 (0.00)	14.990 (1.302)	-0.957 (1.707)	0.09 (0.03)	0.96 (0.14)	0.25 (0.04)
RR	9.50 (0.50)	0.35 (0.03)	10.273 (1.035)	-0.063 (0.433)	0.24 (0.09)	1.40 (0.21)	0.91 (0.08)
LASSO	9.87 (0.42)	0.29 (0.02)	10.316 (0.983)	0.045 (0.577)	0.24 (0.10)	1.40 (0.21)	0.96 (0.08)
EN	10.00 (0.29)	0.27 (0.04)	10.344 (0.953)	0.046 (0.608)	0.24 (0.10)	1.40 (0.21)	0.95 (0.08)
Average	9.74 (0.38)	0.31 (0.02)	10.245 (0.972)	0.024 (0.441)	0.25 (0.10)	1.41 (0.20)	0.96 (0.08)
SSR	10.14 (0.32)	0.26 (0.04)	10.380 (1.232)	-0.003 (0.662)	0.24 (0.09)	1.40 (0.19)	1.09 (0.10)
RF	4.36 (0.22)	0.91 (0.01)	10.784 (1.119)	-0.322 (0.694)	0.17 (0.07)	1.34 (0.15)	0.77 (0.08)
Boosting	9.44 (0.36)	0.38 (0.01)	10.489 (1.228)	-0.103 (0.811)	0.21 (0.08)	1.38 (0.19)	1.02 (0.09)
NN	8.87 (0.60)	0.43 (0.07)	10.322 (1.167)	-0.177 (0.605)	0.24 (0.06)	1.41 (0.22)	0.81 (0.07)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S5. Prediction performance[1] in the calibration (n=415) and cross-validation (n=138) data sets of alternative prediction models used to predict casein micelle size (CMS) expressed in millimeters.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	22.64 (0.43)	0.24 (0.01)	25.647 (1.017)	0.145 (1.835)	0.09 (0.02)	1.31 (0.19)	0.56 (0.08)
PCR	24.81 (0.49)	0.09 (0.01)	25.826 (1.706)	-0.129 (1.677)	0.04 (0.02)	1.31 (0.23)	0.60 (0.14)
PPR	13.66 (1.96)	0.72 (0.09)	35.200 (1.613)	-1.785 (2.310)	0.03 (0.01)	0.95 (0.12)	0.15 (0.04)
RR	22.85 (0.34)	0.24 (0.03)	25.339 (1.328)	-0.108 (1.786)	0.08 (0.03)	1.31 (0.19)	0.68 (0.11)
LASSO	23.18 (0.56)	0.22 (0.01)	25.286 (1.294)	-0.233 (1.769)	0.08 (0.03)	1.31 (0.19)	0.72 (0.11)
EN	26.75 (0.41)	0.04 (0.01)	28.344 (1.094)	0.613 (2.360)	0.01 (0.01)	1.31 (0.19)	0.19 (0.08)
Average	23.19 (0.41)	0.21 (0.01)	25.413 (1.013)	0.104 (1.775)	0.08 (0.03)	1.32 (0.20)	0.68 (0.11)
SSR	25.88 (0.58)	0.02 (0.00)	25.971 (1.867)	-0.078 (0.840)	0.01 (0.00)	1.30 (0.24)	1.62 (0.91)
RF	10.48 (0.22)	0.95 (0.01)	26.215 (1.526)	-0.942 (1.745)	0.03 (0.02)	1.28 (0.21)	0.45 (0.13)
Boosting	23.70 (0.53)	0.25 (0.02)	25.691 (1.830)	-0.189 (1.023)	0.03 (0.01)	1.31 (0.24)	0.84 (0.21)
NN	25.41 (0.32)	0.05 (0.03)	25.818 (1.820)	-0.303 (1.289)	0.02 (0.01)	1.31 (0.24)	0.81 (0.26)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S6. Prediction performance[1] in the calibration (n=451) and cross-validation (n=150) data sets of alternative prediction models used to predict pH.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	0.06 (0.001)	0.71 (0.01)	0.061 (0.002)	-0.0006 (0.003)	0.65 (0.04)	2.34 (0.38)	0.96 (0.03)
PCR	0.07 (0.004)	0.52 (0.01)	0.075 (0.002)	-0.0002 (0.005)	0.48 (0.03)	1.90 (0.24)	0.96 (0.04)
PPR	0.04 (0.003)	0.86 (0.02)	0.076 (0.005)	-0.0031 (0.005)	0.53 (0.07)	1.89 (0.27)	0.75 (0.03)
RR	0.07 (0.001)	0.56 (0.01)	0.073 (0.002)	-0.0003 (0.002)	0.53 (0.02)	2.34 (0.38)	1.16 (0.05)
LASSO	0.07 (0.001)	0.53 (0.01)	0.075 (0.002)	0.0001 (0.004)	0.50 (0.03)	2.34 (0.38)	1.16 (0.05)
EN	Convergence	issue	-	-	-	-	-
Average[3]	0.07 (0.001)	0.60 (0.01)	0.070 (0.001)	-0.0003 (0.002)	0.56 (0.03)	2.34 (0.38)	1.09 (0.03)
SSR	0.07 (0.001)	0.57 (0.01)	0.073 (0.002)	-0.0002 (0.003)	0.53 (0.03)	1.97 (0.22)	1.18 (0.05)
RF	0.03 (0.001)	0.93 (0.00)	0.085 (0.005)	-0.0002 (0.003)	0.34 (0.03)	1.68 (0.14)	1.00 (0.06)
Boosting	0.08 (0.001)	0.42 (0.01)	0.088 (0.004)	0.0003 (0.004)	0.29 (0.03)	1.62 (0.15)	1.10 (0.07)
NN	0.05 (0.002)	0.77 (0.02)	0.062 (0.002)	0.0006 (0.003)	0.65 (0.03)	2.32 (0.36)	0.95 (0.03)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average = model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

[3]In the calculation of the results from the model averaging approach, elastic net was not considered due to a convergence issue.

Supplemental Table S7. Prediction performance[1] in the calibration (n=323) and cross-validation (n=108) data sets of alternative prediction models used to predict heat stability expressed in minutes.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	4.36 (0.09)	0.63 (0.02)	5.668 (0.392)	0.161 (0.457)	0.42 (0.07)	1.49 (0.33)	0.82 (0.05)
PCR	5.16 (0.21)	0.49 (0.02)	5.650 (0.771)	0.183 (0.252)	0.40 (0.03)	1.50 (0.25)	0.90 (0.05)
PPR	1.47 (0.09)	0.96 (0.01)	8.845 (0.738)	-0.929 (1.199)	0.20 (0.07)	0.98 (0.30)	0.35 (0.03)
RR	4.46 (0.09)	0.62 (0.04)	5.540 (0.551)	0.186 (0.322)	0.43 (0.05)	1.49 (0.33)	0.91 (0.05)
LASSO	4.47 (0.22)	0.61 (0.08)	5.717 (0.620)	0.191 (0.436)	0.40 (0.05)	1.49 (0.33)	0.86 (0.05)
EN	5.10 (0.25)	0.50 (0.02)	5.647 (0.669)	0.181 (0.416)	0.40 (0.03)	1.49 (0.33)	0.95 (0.06)
Average	4.48 (0.05)	0.62 (0.04)	5.517 (0.587)	0.179 (0.388)	0.43 (0.05)	1.53 (0.28)	0.93 (0.05)
SSR	5.79 (0.26)	0.37 (0.01)	6.028 (0.903)	0.058 (0.554)	0.31 (0.01)	1.40 (0.19)	1.15 (0.08)
RF	2.63 (0.08)	0.94 (0.00)	6.545 (0.845)	-0.255 (0.457)	0.18 (0.02)	1.28 (0.19)	0.90 (0.09)
Boosting	5.84 (0.23)	0.41 (0.02)	6.616 (0.827)	-0.006 (0.586)	0.17 (0.03)	1.27 (0.20)	1.12 (0.13)
NN	3.83 (0.30)	0.72 (0.03)	5.464 (0.390)	0.286 (0.305)	0.45 (0.04)	1.55 (0.32)	0.84 (0.05)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S8. Prediction performance[1] in the calibration (n=335) and cross-validation (n=112) data sets of alternative prediction models used to predict alpha s1 casein expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	1.58 (0.04)	0.56 (0.02)	1.790 (0.138)	0.009 (0.132)	0.45 (0.07)	1.92 (0.24)	0.89 (0.05)
PCR	1.67 (0.06)	0.51 (0.02)	1.801 (0.131)	0.002 (0.118)	0.44 (0.06)	1.91 (0.22)	0.92 (0.05)
PPR	0.81 (0.14)	0.88 (0.05)	2.715 (0.141)	-0.207 (0.147)	0.20 (0.06)	1.27 (0.21)	0.39 (0.04)
RR	1.57 (0.10)	0.57 (0.04)	1.755 (0.140)	-0.022 (0.106)	0.47 (0.05)	1.92 (0.24)	0.94 (0.05)
LASSO	1.61 (0.15)	0.54 (0.08)	1.789 (0.132)	-0.039 (0.074)	0.44 (0.07)	1.92 (0.24)	0.95 (0.05)
EN	1.73 (0.04)	0.48 (0.02)	1.762 (0.143)	-0.004 (0.102)	0.46 (0.04)	1.92 (0.24)	1.01 (0.05)
Average	1.60 (0.08)	0.55 (0.04)	1.745 (0.136)	-0.014 (0.097)	0.47 (0.05)	1.97 (0.22)	0.98 (0.05)
SSR	1.73 (0.04)	0.47 (0.02)	1.769 (0.144)	-0.011 (0.104)	0.46 (0.04)	1.94 (0.21)	1.03 (0.05)
RF	0.74 (0.03)	0.93 (0.00)	1.785 (0.163)	-0.030 (0.045)	0.44 (0.07)	1.93 (0.25)	0.96 (0.05)
Boosting	1.59 (0.04)	0.57 (0.02)	1.779 (0.167)	-0.027 (0.068)	0.45 (0.05)	1.94 (0.23)	1.07 (0.06)
NN	1.44 (0.11)	0.64 (0.04)	1.816 (0.180)	0.006 (0.158)	0.44 (0.06)	1.90 (0.24)	0.85 (0.05)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S9. Prediction performance[1] in the calibration (n=345) and cross-validation (n=115) data sets of alternative prediction models used to predict alpha s2 casein expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	0.69 (0.04)	0.46 (0.01)	0.738 (0.111)	-0.006 (0.045)	0.38 (0.02)	1.66 (0.15)	0.91 (0.05)
PCR	0.71 (0.04)	0.42 (0.01)	0.740 (0.113)	-0.004 (0.046)	0.37 (0.02)	1.66 (0.15)	0.94 (0.06)
PPR	0.35 (0.04)	0.86 (0.03)	1.134 (0.032)	-0.079 (0.114)	0.15 (0.05)	1.10 (0.29)	0.33 (0.04)
RR	0.69 (0.03)	0.45 (0.01)	0.743 (0.110)	-0.005 (0.055)	0.37 (0.04)	1.66 (0.15)	0.93 (0.06)
LASSO	0.68 (0.03)	0.47 (0.02)	0.741 (0.099)	-0.004 (0.055)	0.37 (0.04)	1.66 (0.15)	0.93 (0.06)
EN	0.74 (0.04)	0.37 (0.01)	0.757 (0.121)	0.0002 (0.064)	0.34 (0.04)	1.66 (0.15)	0.96 (0.06)
Average	0.69 (0.03)	0.45 (0.02)	0.734 (0.113)	-0.004 (0.053)	0.38 (0.04)	1.67 (0.15)	0.97 (0.06)
SSR	0.75 (0.04)	0.36 (0.01)	0.756 (0.113)	-0.0003 (0.050)	0.34 (0.05)	1.63 (0.17)	1.07 (0.07)
RF	0.32 (0.02)	0.93 (0.00)	0.783 (0.091)	-0.022 (0.046)	0.30 (0.08)	1.57 (0.19)	0.93 (0.07)
Boosting	0.70 (0.03)	0.46 (0.02)	0.781 (0.108)	-0.003 (0.046)	0.30 (0.05)	1.57 (0.16)	1.05 (0.08)
NN	0.68 (0.03)	0.48 (0.02)	0.750 (0.103)	-0.001 (0.055)	0.36 (0.02)	1.64 (0.17)	0.90 (0.06)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S10. Prediction performance[1] in the calibration (n=337) and cross-validation (n=112) data sets of alternative prediction models used to predict beta casein expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	1.69 (0.05)	0.39 (0.02)	1.800 (0.151)	0.012 (0.085)	0.31 (0.06)	1.66 (0.17)	0.91 (0.06)
PCR	1.61 (0.04)	0.45 (0.02)	1.791 (0.108)	0.036 (0.059)	0.33 (0.06)	1.67 (0.15)	0.86 (0.06)
PPR	0.73 (0.06)	0.89 (0.02)	2.422 (0.086)	-0.101 (0.196)	0.23 (0.08)	1.24 (0.21)	0.41 (0.04)
RR	1.55 (0.04)	0.49 (0.03)	1.759 (0.128)	0.013 (0.076)	0.35 (0.05)	1.66 (0.17)	0.91 (0.06)
LASSO	1.57 (0.06)	0.47 (0.06)	1.802 (0.147)	0.023 (0.063)	0.32 (0.04)	1.66 (0.17)	0.88 (0.06)
EN	1.72 (0.05)	0.37 (0.02)	1.803 (0.168)	-0.0004 (0.060)	0.30 (0.05)	1.66 (0.17)	0.96 (0.07)
Average	1.61 (0.03)	0.45 (0.03)	1.767 (0.150)	0.012 (0.058)	0.33 (0.04)	1.69 (0.16)	0.96 (0.06)
SSR	1.74 (0.06)	0.36 (0.01)	1.774 (0.184)	0.010 (0.065)	0.33 (0.04)	1.69 (0.19)	1.06 (0.07)
RF	0.74 (0.03)	0.92 (0.00)	1.788 (0.136)	-0.010 (0.076)	0.33 (0.08)	1.68 (0.21)	0.92 (0.06)
Boosting	1.60 (0.04)	0.47 (0.02)	1.812 (0.173)	0.014 (0.046)	0.30 (0.06)	1.66 (0.20)	1.03 (0.07)
NN	1.38 (0.06)	0.59 (0.07)	1.793 (0.10)	0.015 (0.047)	0.33 (0.07)	1.67 (0.16)	0.84 (0.06)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S11. Prediction performance[1] in the calibration (n=340) and cross-validation (n=113) data sets of alternative prediction models used to predict kappa casein expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	1.06 (0.01)	0.45 (0.03)	1.097 (0.024)	-0.003 (0.079)	0.42 (0.08)	1.72 (0.31)	0.96 (0.05)
PCR	1.06 (0.02)	0.45 (0.03)	1.110 (0.025)	-0.008 (0.071)	0.41 (0.09)	1.70 (0.31)	0.95 (0.05)
PPR	0.54 (0.12)	0.85 (0.08)	1.616 (0.045)	-0.043 (0.201)	0.19 (0.05)	1.16 (0.19)	0.39 (0.04)
RR	1.06 (0.01)	0.46 (0.03)	1.095 (0.027)	-0.002 (0.074)	0.42 (0.09)	1.72 (0.31)	0.99 (0.05)
LASSO	1.05 (0.01)	0.46 (0.03)	1.099 (0.027)	0.004 (0.077)	0.42 (0.08)	1.72 (0.31)	0.99 (0.06)
EN	1.06 (0.01)	0.46 (0.03)	1.101 (0.030)	0.002 (0.083)	0.41 (0.09)	1.72 (0.31)	0.96 (0.05)
Average	1.05 (0.01)	0.46 (0.03)	1.095 (0.027)	0.0003 (0.078)	0.42 (0.09)	1.72 (0.31)	0.98 (0.05)
SSR	1.07 (0.01)	0.44 (0.03)	1.100 (0.029)	0.0004 (0.070)	0.42 (0.09)	1.71 (0.30)	1.04 (0.06)
RF	0.47 (0.02)	0.92 (0.01)	1.157 (0.024)	0.009 (0.093)	0.36 (0.10)	1.64 (0.34)	0.89 (0.06)
Boosting	0.99 (0.01)	0.54 (0.03)	1.134 (0.015)	-0.004 (0.067)	0.38 (0.10)	1.66 (0.31)	1.07 (0.06)
NN	1.01 (0.02)	0.51 (0.04)	1.095 (0.027)	-0.003 (0.068)	0.42 (0.09)	1.72 (0.32)	0.95 (0.05)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S12. Prediction performance[1] in the calibration (n=343) and cross-validation (n=114) data sets of alternative prediction models used to predict alpha lactalbumin expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	0.23 (0.01)	0.43 (0.01)	0.256 (0.014)	0.001 (0.010)	0.28 (0.05)	1.54 (0.25)	0.82 (0.06)
PCR	0.24 (0.01)	0.33 (0.02)	0.266 (0.021)	-0.0003 (0.019)	0.22 (0.02)	1.48 (0.21)	0.82 (0.07)
PPR	0.11 (0.00)	0.86 (0.01)	0.373 (0.025)	-0.024 (0.038)	0.13 (0.09)	1.08 (0.27)	0.29 (0.04)
RR	0.24 (0.01)	0.38 (0.01)	0.258 (0.021)	0.002 (0.015)	0.26 (0.02)	1.54 (0.25)	0.92 (0.07)
LASSO	0.24 (0.01)	0.38 (0.02)	0.256 (0.019)	-0.0003 (0.014)	0.27 (0.03)	1.54 (0.25)	0.95 (0.07)
EN	0.25 (0.01)	0.28 (0.03)	0.272 (0.026)	0.003 (0.018)	0.19 (0.02)	1.54 (0.25)	0.78 (0.07)
Average	0.23 (0.01)	0.40 (0.02)	0.255 (0.020)	0.001 (0.013)	0.27 (0.02)	1.54 (0.22)	0.95 (0.07)
SSR	0.25 (0.01)	0.29 (0.02)	0.265 (0.024)	0.001 (0.017)	0.22 (0.02)	1.48 (0.19)	1.15 (0.10)
RF	0.11 (0.01)	0.94 (0.01)	0.281 (0.029)	-0.008 (0.017)	0.14 (0.07)	1.40 (0.17)	0.81 (0.10)
Boosting	0.25 (0.01)	0.38 (0.01)	0.285 (0.023)	0.003 (0.018)	0.10 (0.03)	1.38 (0.19)	1.03 (0.15)
NN	0.20 (0.02)	0.54 (0.05)	0.256 (0.018)	0.0001 (0.011)	0.18 (0.05)	1.54 (0.23)	0.80 (0.06)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S13. Prediction performance[1] in the calibration (n=347) and cross-validation (n=115) data sets of alternative prediction models used to predict beta lactoglobulin A expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	1.01 (0.04)	0.27 (0.01)	1.057 (0.100)	-0.0004 (0.044)	0.19 (0.06)	1.55 (0.16)	0.86 (0.09)
PCR	1.03 (0.03)	0.22 (0.03)	1.066 (0.118)	-0.004 (0.043)	0.17 (0.04)	1.54 (0.14)	0.86 (0.09)
PPR	0.46 (0.12)	0.84 (0.08)	1.485 (0.093)	-0.064 (0.093)	0.07 (0.01)	1.10 (0.14)	0.24 (0.04)
RR	0.99 (0.03)	0.28 (0.03)	1.050 (0.113)	0.004 (0.037)	0.19 (0.04)	1.55 (0.16)	0.92 (0.09)
LASSO	0.99 (0.02)	0.28 (0.06)	1.062 (0.113)	0.002 (0.049)	0.18 (0.04)	1.55 (0.16)	0.87 (0.09)
EN	1.02 (0.04)	0.23 (0.02)	1.256 (0.117)	0.011 (0.055)	0.18 (0.04)	1.55 (0.16)	0.91 (0.09)
Average	1.00 (0.03)	0.26 (0.03)	1.052 (0.112)	0.004 (0.046)	0.19 (0.05)	1.56 (0.14)	0.92 (0.09)
SSR	1.06 (0.04)	0.18 (0.02)	1.074 (0.133)	0.003 (0.046)	0.15 (0.05)	1.53 (0.13)	1.08 (0.12)
RF	0.44 (0.01)	0.94 (0.00)	1.125 (0.105)	-0.022 (0.036)	0.11 (0.07)	1.45 (0.11)	0.65 (0.09)
Boosting	0.97 (0.03)	0.35 (0.01)	1.086 (0.114)	-0.007 (0.033)	0.15 (0.07)	1.51 (0.12)	0.95 (0.11)
NN	0.93 (0.04)	0.37 (0.08)	1.063 (0.099)	0.002 (0.025)	0.20 (0.05)	1.54 (0.15)	0.80 (0.08)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S14. Prediction performance[1] in the calibration (n=347) and cross-validation (n=116) data sets of alternative prediction models used to predict beta lactoglobulin B expressed in g/L.

	Calibration		Validation				
Method[2]	RMSE (SD)	R^2 (SD)	RMSEV (SD)	Bias (SD)	R^2 (SD)	RPIQ (SD)	Slope (SE)
PLSR	1.19 (0.06)	0.50 (0.02)	1.461 (0.114)	-0.030 (0.094)	0.28 (0.06)	1.74 (0.18)	0.75 (0.06)
PCR	1.49 (0.05)	0.21 (0.02)	1.577 (0.120)	-0.027 (0.120)	0.14 (0.06)	1.61 (0.14)	0.79 (0.09)
PPR	0.63 (0.11)	0.86 (0.04)	1.947 (0.110)	-0.084 (0.148)	0.14 (0.03)	1.31 (0.17)	0.35 (0.04)
RR	1.31 (0.04)	0.41 (0.02)	1.478 (0.110)	-0.018 (0.095)	0.24 (0.07)	1.74 (0.18)	0.90 (0.08)
LASSO	1.21 (0.04)	0.47 (0.01)	1.453 (0.108)	-0.025 (0.098)	0.27 (0.06)	1.74 (0.18)	0.84 (0.07)
EN	1.49 (0.04)	0.22 (0.03)	1.577 (0.136)	-0.017 (0.115)	0.14 (0.08)	1.74 (0.18)	0.91 (0.11)
Average	1.25 (0.05)	0.47 (0.01)	1.443 (0.117)	-0.022 (0.097)	0.27 (0.08)	1.76 (0.16)	0.97 (0.08)
SSR	1.63 (0.07)	0.08 (0.02)	1.649 (0.188)	-0.014 (0.135)	0.06 (0.04)	1.54 (0.08)	1.06 (0.26)
RF	0.66 (0.02)	0.94 (0.00)	1.699 (0.171)	-0.014 (0.183)	0.03 (0.01)	1.50 (0.09)	0.42 (0.12)
Boosting	1.50 (0.06)	0.30 (0.01)	1.653 (0.179)	-0.007 (0.148)	0.04 (0.01)	1.54 (0.08)	0.69 (0.16)
NN	1.10 (0.04)	0.58 (0.02)	1.455 (0.102)	-0.019 (0.114)	0.28 (0.05)	1.75 (0.17)	0.78 (0.06)

[1]RMSE = root mean square error; RMSEV = root mean square error in validation data set; R^2 = Coefficient of determination; RPIQ = ratio of performance to interquartile distance; SD = standard deviation; SE = standard error.

[2]PLSR= Partial Least Square Regression; RR= Ridge Regression; EN= Elastic Net; Average= model averaging approach; PCR= Principal Component Regression; PPR= Projection Pursuit Regression; SSR= Spike and Slab Regression; RF= Random Forest; NN= Neural Network.

Supplemental Table S15. Average number of wavelengths selected by LASSO, EN, and SSR for each trait.

Trait	LASSO	EN	SSR
RCT	106.50	121.75	13.00
k20	65.50	88.00	12.25
a30	28.00	102.00	11.25
a60	22.00	56.25	10.50
CMS	48.00	319.75	0.00
pH	18.00	-	6.50
Heat stability	87.00	165.50	20.50
Alpha s1 casein	50.25	75.50	13.25
Alpha s2 casein	39.00	75.25	11.25
Beta casein	40.75	107.25	13.00
Kappa casein	13.50	78.00	13.25
Alpha lactalbumin	27.50	187.25	8.00
Beta lactoglobulin A	30.00	90.75	6.50
Beta lactoglobulin B	94.50	171.50	4.25

Supplemental Table S16. Prediction performance[1] for rennet coagulation time (RCT), curd firming time (k20), curd firmness at 30 and 60 min (a30, a60), casein micelle size (CMS), pH, and heat stability classification using different approaches in the calibration and cross-validation data.

Trait	Method[2]	Calibration				Validation			
		AUC (SD)	Sensitivity[3] (SD)	Specificity[3] (SD)	Accuracy (SD)	AUC (SD)	Sensitivity[3] (SD)	Specificity[3] (SD)	Accuracy (SD)
RCT	PLS-DA	0.81 (0.02)	0.85 (0.02)	0.78 (0.03)	0.81 (0.02)	0.75 (0.03)	0.80 (0.04)	0.69 (0.07)	0.75 (0.03)
	RF	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	0.70 (0.01)	0.72 (0.05)	0.68 (0.07)	0.70 (0.01)
	Boosting	0.93 (0.01)	0.93 (0.01)	0.92 (0.01)	0.93 (0.01)	0.70 (0.02)	0.70 (0.09)	0.69 (0.06)	0.70 (0.02)
	SVM	0.83 (0.01)	0.84 (0.02)	0.83 (0.01)	0.83 (0.01)	0.75 (0.06)	0.77 (0.08)	0.73 (0.06)	0.75 (0.06)
k20	PLS-DA	0.80 (0.03)	0.80 (0.04)	0.79 (0.02)	0.80 (0.03)	0.70 (0.01)	0.70 (0.03)	0.69 (0.03)	0.70 (0.01)
	RF	0.99 (0.00)	0.99 (0.00)	1.00 (0.00)	0.99 (0.00)	0.71 (0.04)	0.69 (0.07)	0.72 (0.03)	0.71 (0.04)
	Boosting	0.94 (0.01)	0.92 (0.04)	0.97 (0.03)	0.95 (0.00)	0.70 (0.06)	0.69 (0.12)	0.71 (0.07)	0.70 (0.05)
	SVM	0.83 (0.02)	0.84 (0.02)	0.83 (0.03)	0.83 (0.02)	0.73 (0.03)	0.75 (0.05)	0.72 (0.02)	0.73 (0.02)
a30	PLS-DA	0.73 (0.03)	0.71 (0.02)	0.76 (0.03)	0.73 (0.03)	0.72 (0.04)	0.71 (0.07)	0.74 (0.09)	0.72 (0.04)
	RF	0.99 (0.00)	1.00 (0.00)	0.99 (0.01)	0.99 (0.00)	0.71 (0.03)	0.68 (0.07)	0.73 (0.09)	0.71 (0.03)
	Boosting	0.94 (0.01)	0.95 (0.01)	0.93 (0.01)	0.94 (0.01)	0.70 (0.04)	0.70 (0.05)	0.70 (0.09)	0.70 (0.04)
	SVM	0.80 (0.01)	0.79 (0.01)	0.81 (0.03)	0.80 (0.01)	0.73 (0.03)	0.73 (0.09)	0.73 (0.09)	0.73 (0.03)
a60	PLS-DA	0.71 (0.02)	0.64 (0.03)	0.78 (0.02)	0.71 (0.02)	0.69 (0.07)	0.63 (0.06)	0.76 (0.08)	0.69 (0.07)
	RF	0.99 (0.00)	1.00 (0.00)	0.99 (0.01)	0.99 (0.00)	0.66 (0.07)	0.66 (0.07)	0.67 (0.09)	0.66 (0.07)
	Boosting	0.93 (0.01)	0.95 (0.02)	0.90 (0.03)	0.93 (0.01)	0.68 (0.07)	0.69 (0.07)	0.68 (0.09)	0.68 (0.07)
	SVM	0.74 (0.02)	0.65 (0.03)	0.82 (0.01)	0.74 (0.02)	0.68 (0.05)	0.58 (0.07)	0.78 (0.03)	0.68 (0.05)

CMS	PLS-DA	0.65 (0.05)	0.63 (0.08)	0.67 (0.04)	0.65 (0.05)	0.58 (0.03)	0.59 (0.07)	0.58 (0.06)	0.58 (0.03)
	RF	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.54 (0.04)	0.59 (0.04)	0.49 (0.10)	0.54 (0.04)
	Boosting	0.93 (0.01)	0.92 (0.04)	0.95 (0.03)	0.93 (0.01)	0.58 (0.02)	0.59 (0.11)	0.57 (0.08)	0.58 (0.02)
	SVM	0.72 (0.02)	0.72 (0.02)	0.72 (0.02)	0.72 (0.02)	0.62 (0.03)	0.63 (0.09)	0.62 (0.05)	0.62 (0.03)
pH	PLS-DA	0.86 (0.01)	0.87 (0.00)	0.86 (0.02)	0.86 (0.01)	0.80 (0.03)	0.81 (0.03)	0.78 (0.03)	0.80 (0.03)
	RF	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.65 (0.02)	0.63 (0.06)	0.68 (0.03)	0.65 (0.02)
	Boosting	0.91 (0.00)	0.93 (0.01)	0.90 (0.01)	0.91 (0.00)	0.69 (0.02)	0.68 (0.08)	0.70 (0.05)	0.69 (0.02)
	SVM	0.87 (0.01)	0.87 (0.02)	0.87 (0.00)	0.87 (0.01)	0.80 (0.02)	0.80 (0.02)	0.80 (0.01)	0.80 (0.02)
Heat stability	PLS-DA	0.80 (0.01)	0.72 (0.02)	0.88 (0.01)	0.80 (0.01)	0.74 (0.04)	0.66 (0.10)	0.82 (0.05)	0.74 (0.04)
	RF	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.63 (0.03)	0.62 (0.05)	0.64 (0.02)	0.63 (0.03)
	Boosting	0.96 (0.00)	0.98 (0.00)	0.94 (0.01)	0.96 (0.00)	0.64 (0.04)	0.66 (0.03)	0.63 (0.08)	0.64 (0.04)
	SVM	0.82 (0.02)	0.71 (0.03)	0.92 (0.00)	0.82 (0.02)	0.74 (0.05)	0.62 (0.12)	0.86 (0.04)	0.74 (0.05)

[1]AUC= Area Under the receiver operating characteristic Curve.

[2]PLS-DA= Partial Least Square Discriminant Analysis; DT= Decision tree; RF= Random Forest; SVM= Support Vector Machine; SD = standard deviation.

[3]Sensitivity= samples correctly classified as “high”; Specificity= samples correctly classified as “low”.

Supplemental Table S17. Prediction performance for alpha s1 casein, alpha s2 casein, beta casein, and kappa casein classification using different approaches in the calibration and cross-validation data sets.

	Alpha s1 casein		Alpha s2 casein		Beta casein		Kappa Casein	
	Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
Method[1]	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)
PLS-DA	0.55 (0.05)	0.41 (0.09)	0.52 (0.06)	0.40 (0.04)	0.54 (0.04)	0.43 (0.02)	0.51 (0.04)	0.42 (0.03)
RF	0.99 (0.00)	0.48 (0.02)	1.00 (0.00)	0.38 (0.03)	0.99 (0.00)	0.45 (0.03)	0.99 (0.00)	0.45 (0.02)
Boosting	0.99 (0.01)	0.47 (0.03)	0.99 (0.00)	0.36 (0.02)	0.99 (0.00)	0.45 (0.06)	0.99 (0.00)	0.43 (0.02)
SVM	0.67 (0.02)	0.44 (0.03)	0.62 (0.01)	0.39 (0.04)	0.65 (0.01)	0.46 (0.04)	0.64 (0.01)	0.41 (0.04)

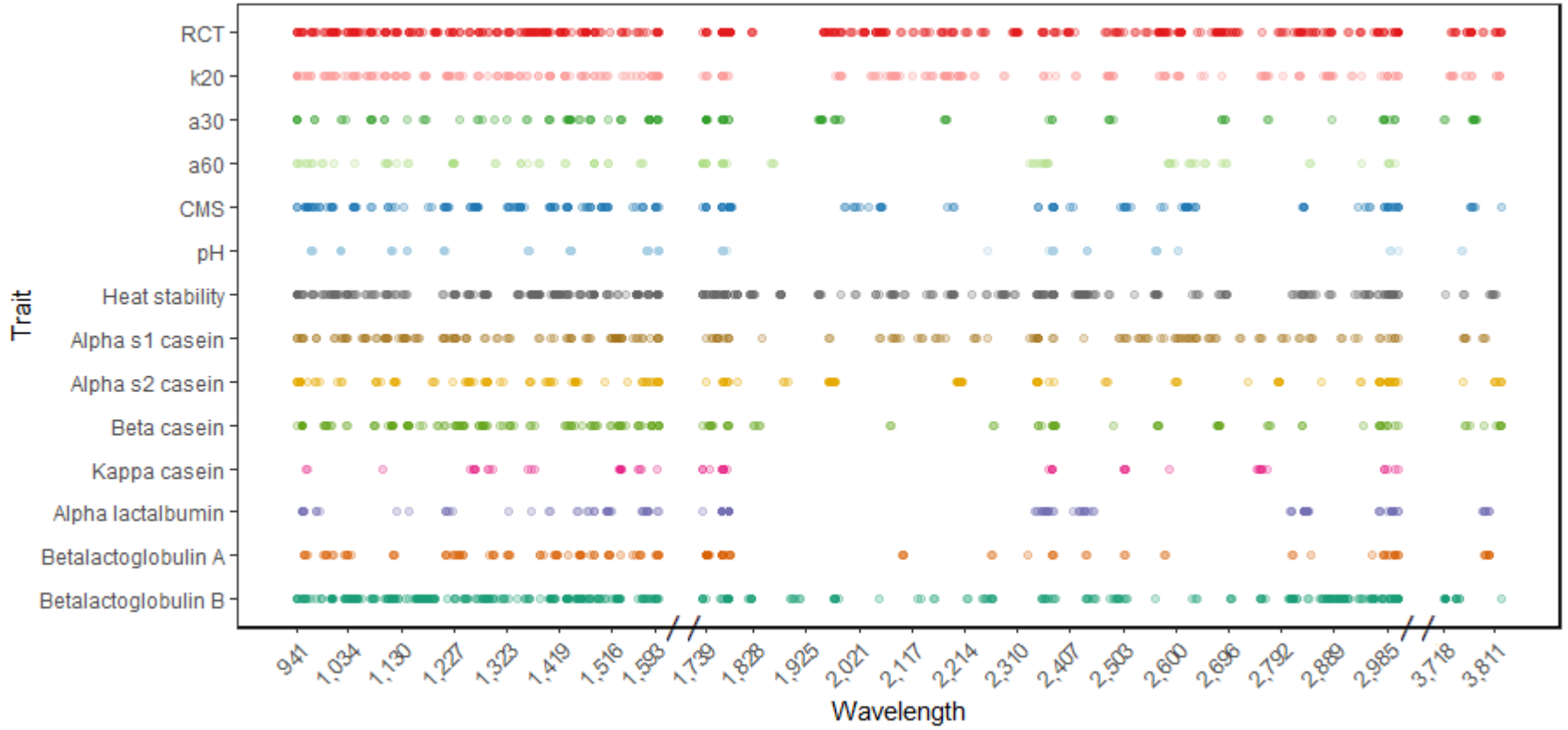
[1]PLS-DA= Partial Least Square Discriminant Analysis; DT= Decision tree; RF= Random Forest; SVM= Support Vector Machine.
[2]SD = standard deviation.

Supplemental Table S18. Prediction performance for alpha lactalbumin, beta lactoglobulin A, and beta lactoglobulin B classification using different approaches in the calibration and cross-validation data sets.

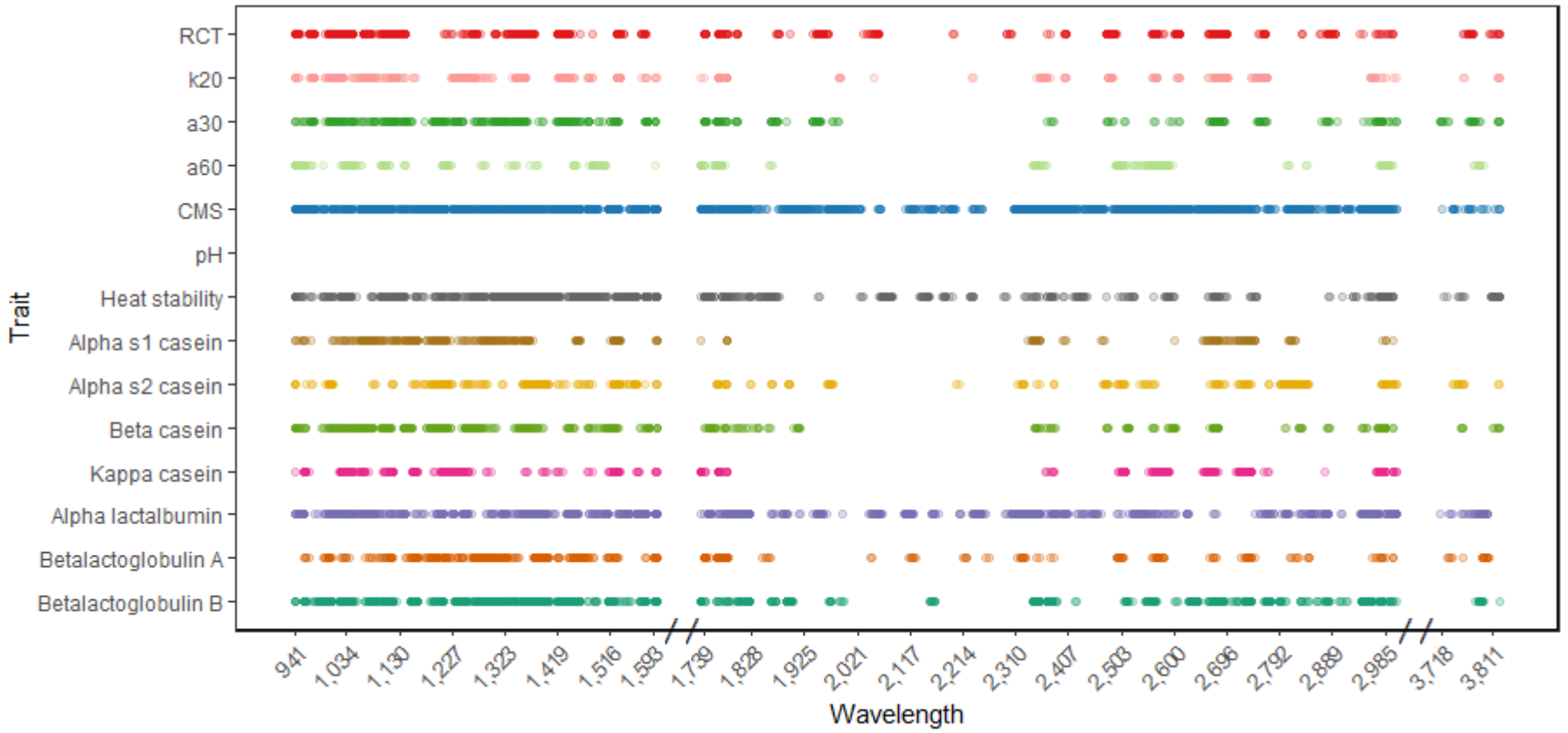
	Alpha lactalbumin		Beta lactoglobulin A		Beta lactoglobulin B	
	Calibration	Validation	Calibration	Validation	Calibration	Validation
Method[1]	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)	Accuracy (SD)
PLS-DA	0.52 (0.01)	0.40 (0.03)	0.42 (0.02)	0.42 (0.03)	0.55 (0.03)	0.41 (0.04)
RF	1.00 (0.00)	0.34 (0.05)	0.99 (0.00)	0.31 (0.03)	0.99 (0.00)	0.29 (0.01)
Boosting	0.99 (0.01)	0.36 (0.06)	0.99 (0.00)	0.33 (0.01)	0.98 (0.00)	0.33 (0.03)
SVM	0.59 (0.01)	0.43 (0.03)	0.55 (0.01)	0.37 (0.03)	0.59 (0.03)	0.40 (0.04)

[1]PLS-DA= Partial Least Square Discriminant Analysis; DT= Decision tree; RF= Random Forest; SVM= Support Vector Machine.
[2]SD = standard deviation.

Supplemental Figure S1. Wavelengths selected by LASSO. The opacity of the dot represents the number of times a specific wavelength has been selected across folds, ranging from missing (0 folds) to full color (4 folds).



Supplemental Figure S2. Wavelengths selected by elastic net. The opacity of the dot represents the number of times a specific wavelength has been selected across folds, ranging from missing (0 folds) to full color (4 folds). pH is not presented in the Figure, as EN had convergence issues during pH prediction.



Supplemental Figure S3. Wavelengths selected by spike and slab regression. The opacity of the dot represents the number of times a specific wavelength has been selected across folds, ranging from missing (0 folds) to full color (4 folds).