Table S1. Tunable components of machine learning pipelines: Hyperparameters and corresponding search spaces

ML architecture component		**Hyperparameters**	**Search Space Method**	**Search Space Explored**
Machine Learning Algorithm	PLS (Projection to Latent Structures)	Number of Components	Grid Search	1 to 10
	SVM (Support Vector Machine)	Regularization Parameter (C), Type of Kernel (K), Influence of Single Training Example (Γ), Margin of Tolerance (ε), Use of Shrinking Heuristics (H)	Random Search	C: A number between 0.1 and 5, K: One of 'linear', 'rbf', 'sigmoid', Γ: 'scale', 'auto' or a number between 0.001 and 0.1, ε: A number between 0.0 and 0.2, H: True or False
	CNNR (Convolutional Neural Network for Regression)	Number of Filters in First Convolutional Layer (C1_K), Size of Filters in First Convolutional Layer (C1_S), Number of Filters in Second Convolutional Layer (C2_K), Size of Filters in Second Convolutional Layer (C2_S), Number of Neurons in Dense Layer (Dense), Dropout Rate (Drop)		C1_K: A random integer between 16 and 64, C1_S: 3, 5, or 7, C2_K: A random integer between 32 and 128, C2_S: 3, 5, or 7, Dense: A random integer between 64 and 256, Drop: A random number between 0.0 and 0.5
	RFR (Random Forest Regressor)	Number of Trees (n), Maximum Depth of Tree (maxD), Minimum Number of Samples Required to Split an Internal Node (minsplit), Minimum Number of Samples Required to be at a Leaf Node (minleaf), Number of Features to Consider When Looking for the Best Split (maxF), Method of Selecting Samples for Training Each Tree (bootstrap)		n: A random integer between 50 and 200, maxD: A random integer between 1 and 20, minsplit: A random integer between 2 and 20, minleāf: A random integer between 1 and 20, maxF: One of 'auto', 'sqrt', 'log2', bootstrap: True or False
	XGB (XGBoost)	Number of Gradient Boosted Trees (n), Learning Rate (lr), Maximum Depth of a Tree (maxD), Subsample Ratio of the Training Instances (subs), Subsample Ratio of Columns When Constructing Each Tree (colsub), Minimum Loss Reduction Required to Make a Further Partition (γ), Minimum Sum of Instance Weight Needed in a Child (minC)		n: A random integer between 50 and 200, lr: A random number between 0.1 and 0.6, maxD: 10, subs: A random number between 0.5 and 1.5, colsub: A random number between 0.5 and 1.5, γ: A random number between 0 and 20, minC: A random integer between 1 and 20
Spectral Feature Selection Method	Moving-Window	Window size in spectrum % (S), window overlap in % (O), number of windows (N)	Arbitrated	[S,O,N]: [25%,75%,1], [12.5%,75%,10] and [5%;10%,5]
	Variable Importance to Projection, VIP	Importance in a PLS model using 100% of the spectrum (VIP)		VIP > 0.9, with 0.05 steps
Data pre-processing transformation	Mean Centering and Standard Normalization, MCSN		N/A	
	All data logarithm, Log	Log-shift	Arbitrated	+ 1 for spectroscopy data, + 0.1 for standard analytical data
	Standard analytical data excluvise logarithm, LogOutput			+ 0.1