Supplemental Information

Data Preparation.

Cosmic ray artefacts were removed from individuals recording with a median filter. Trend removal was performed using an iterative use of Savitzky-Golay filtering and morphological opening of the signal to avoid removing crucial information from the spectrum. Spectrum were then standardized by spectrum, followed by frequency bin standardization (by feature).

Statistical Inference

Model training and evaluation were performed with sklearn and xboost libraries in Python. Code is available at the following repository: https://github.com/Lmombaerts/raman_gbm_lymphoma. Raw data are available at the following address: TBA.

Logistic Regression

In this study, a L1-regularized Logistic Regression was trained to differentiate between tissue types. The L1 penalty allows for the intrinsic reduction of the entire feature set to a subset of the most relevant features for the discrimination between glioblastoma and lymphoma tumor tissues. This choice of regularization was further motivated by the dimension of the training set: 54 glioblastoma Raman spectrum, 21 lymphoma Raman spectrum and 1613 Raman frequency bins. Due to the relatively small amounts of spectrum and patients (Supplemental Figure 1), the performance of the algorithm was evaluated on 4 folds cross-validation (repeated 5 times) and the regularization parameter was optimized on a [0.1,0.5,1] grid – 0.5 leading to best performance trade-offs. Patient stratification results are displayed on Supplemental Figure 2 and do not suggest patient-specific bias.

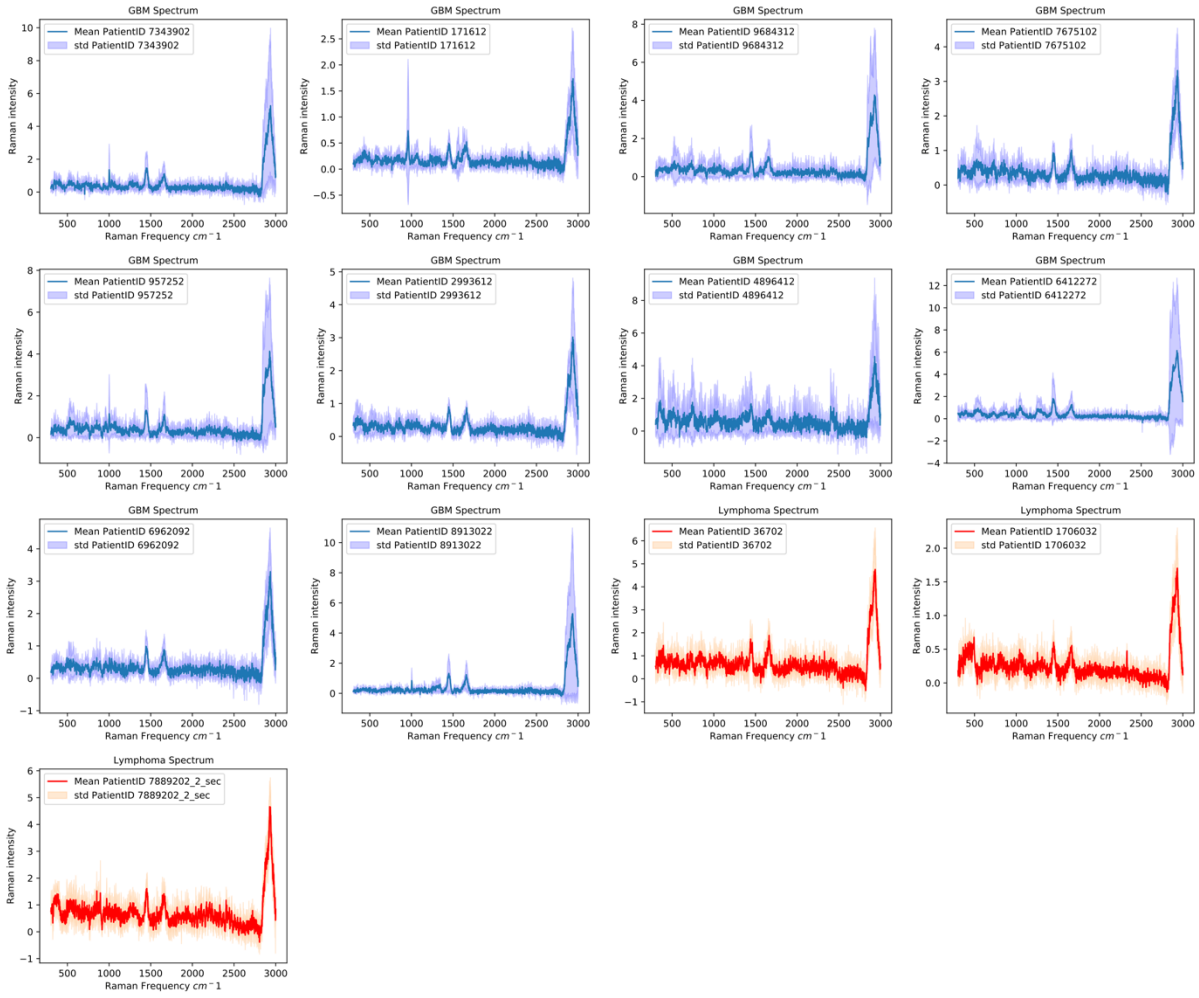
Random Forest

The objective function of the Random Forest has been modified to account for the class imbalance, so that a class-sensitive cost function was optimized to raise the penalty resulting from a misclassification of the minority class (lymphoma spectrum). Feature selection was considered with Recursive Feature Elimination, but did not improve the

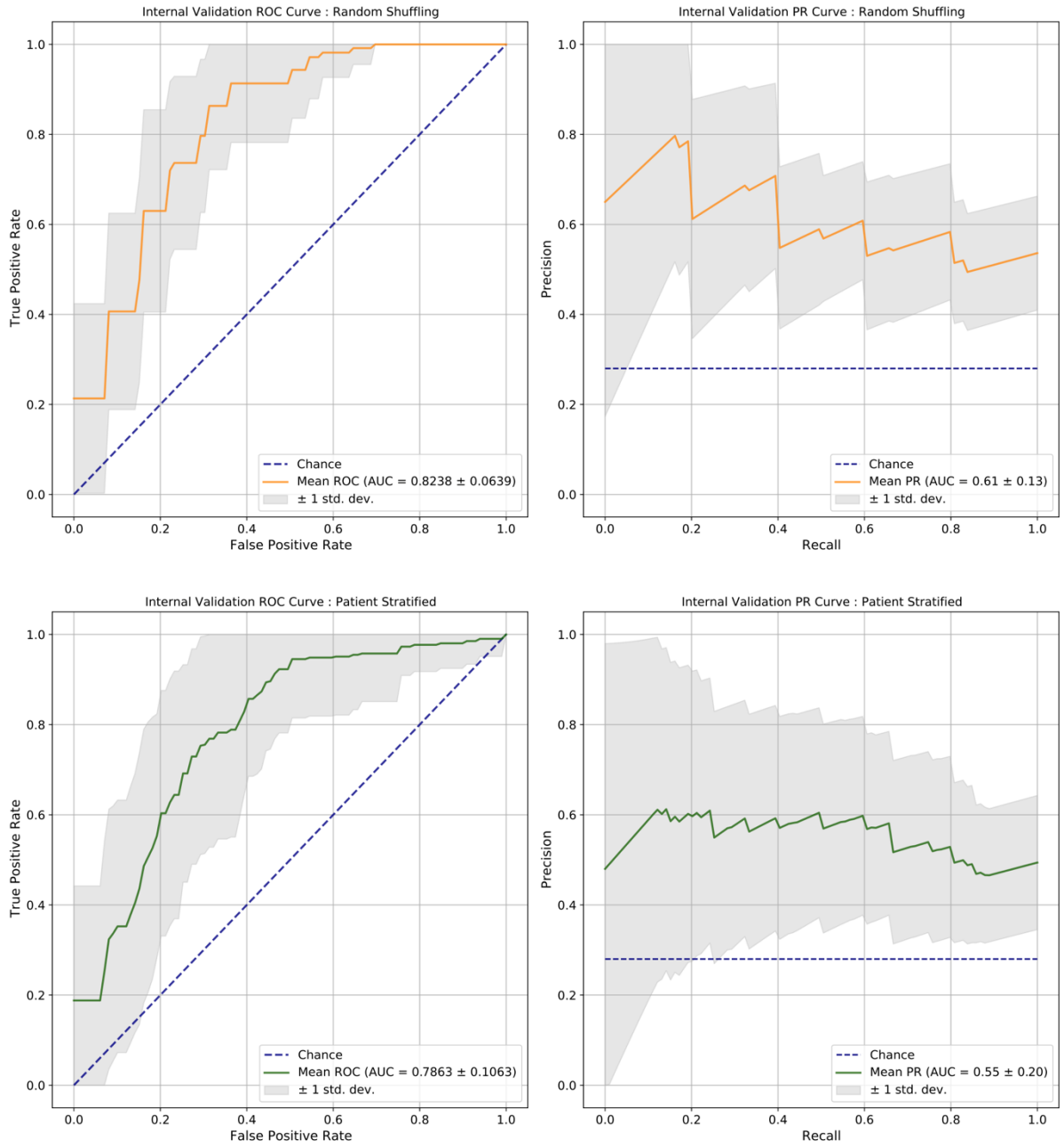
performance of the prediction model. For the SOLAIS classification, the following hyperparameters were optimized using a grid and 4 folds cross-validations (repeated 5 times): number of trees in the forest (or estimators), maximum depth of de decision trees, minimum number of samples to make a split and minimum number of samples to be defined as a leaf. The best classification results obtained were: 115 estimators, maximum depth of 5, 5 minimum samples to make a split and 2 samples minimum as a leaf. For the FFPE classification, the following hyperparameters were optimized using a grid and 3 folds cross-validations (repeated 5 times): number of trees in the forest (or estimators), maximum depth of de decision trees, minimum number of samples to make a split and minimum number of samples to be defined as a leaf. The best classification results obtained were: 100 estimators, maximum depth of 5, 10 minimum samples to make a split and 4 samples minimum as a leaf.

XGBoost

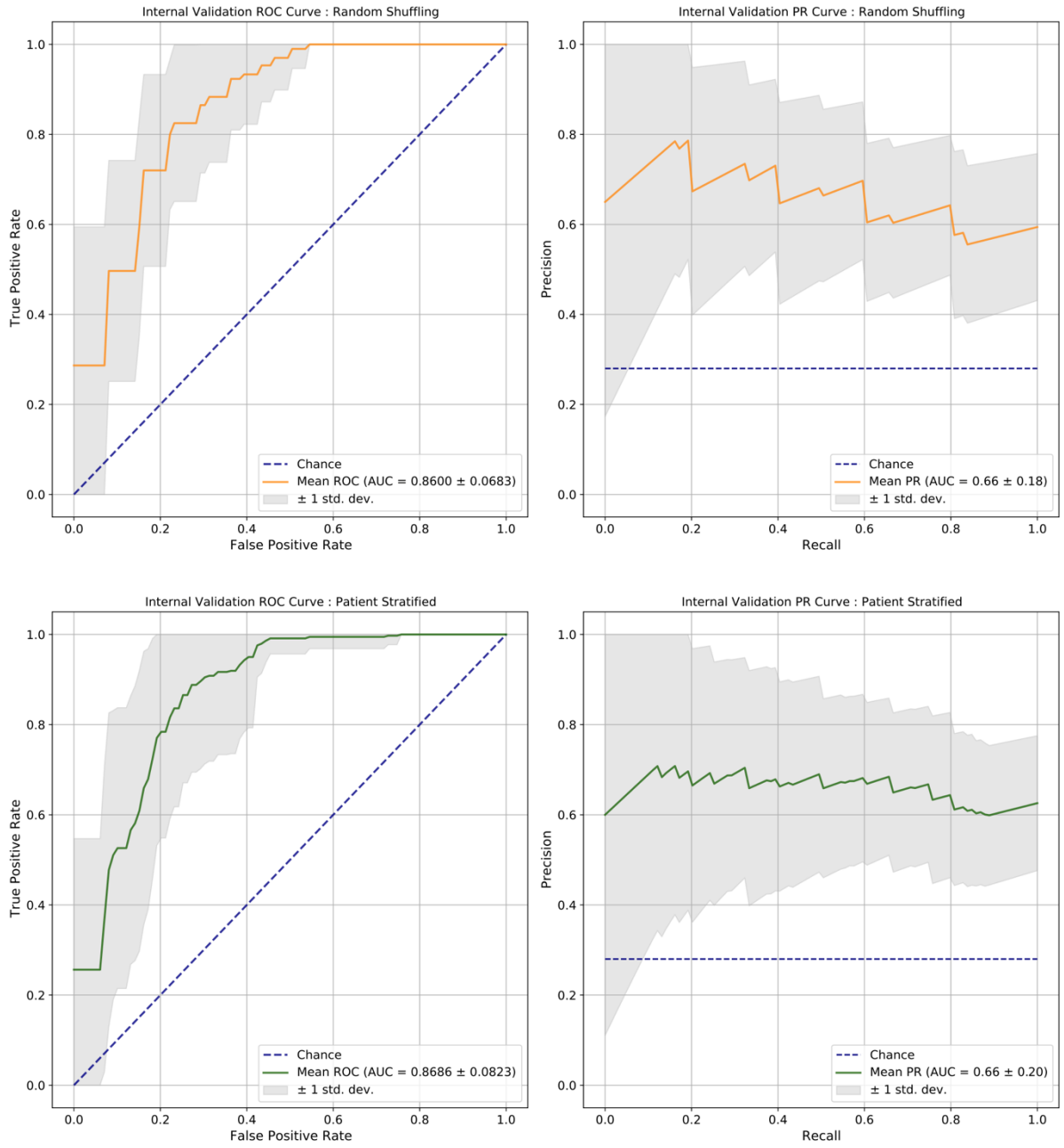
The objective function of the XGBoost algorithm has been modified to account for the class imbalance, so that a class-sensitive cost function was optimized to raise the penalty resulting from a misclassification of the minority class (lymphoma spectrum). The following hyperparameters were optimized on 4 splits cross-validation (repeated 2 times) with a grid search: number of trees (estimators), learning rate, maximum tree depth, gamma regularization parameter, subsample ratio of the training instance, subsampling by tree, subsampling by level and minimum child weight. The best classification results were obtained with 226 estimators, a learning rate of 0.1, maximum tree depth of 3, gamma of 0.1, 0.8 subsample ratio, 0.8 subsampling by tree and by level and 5 minimum child weight.



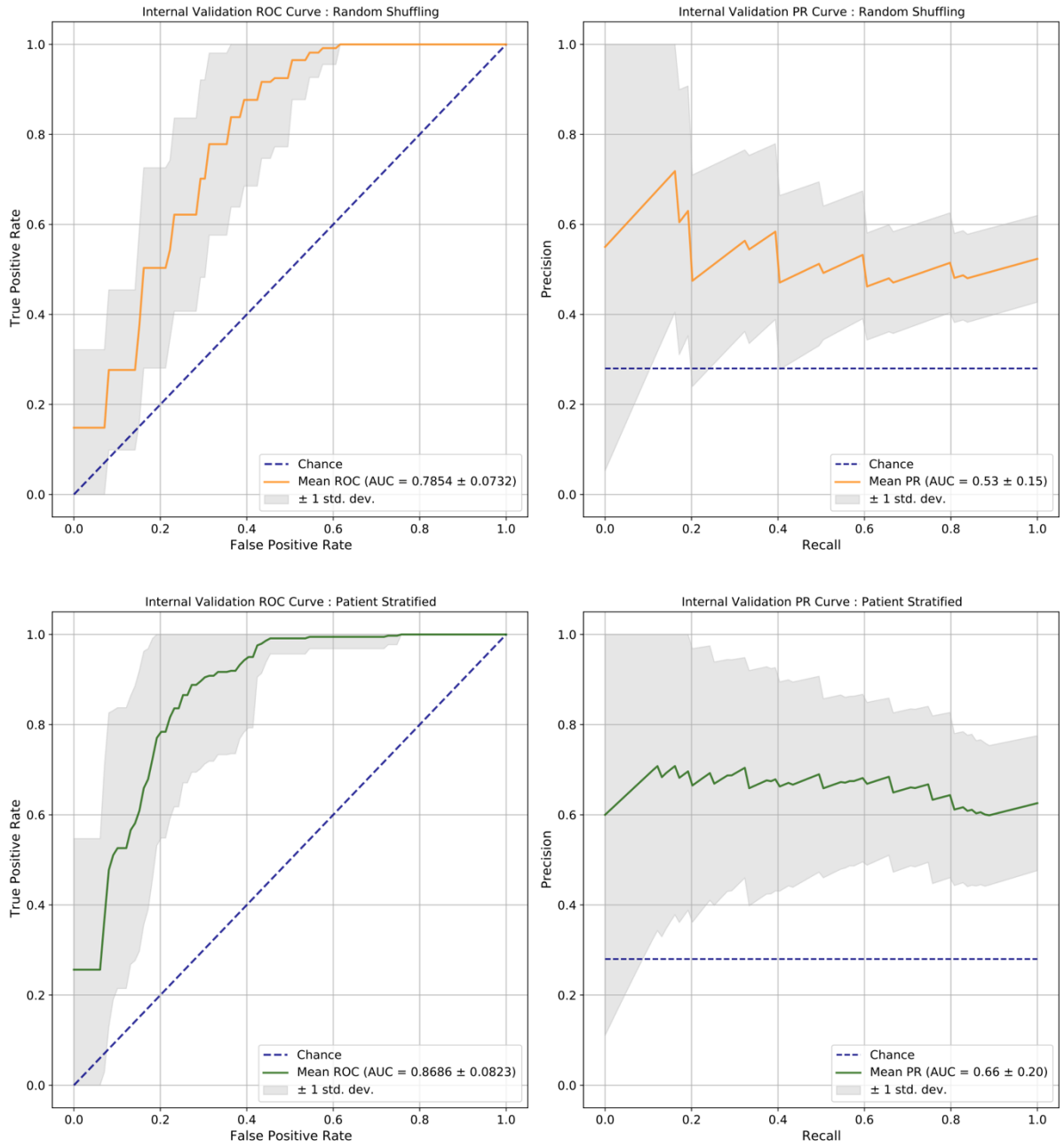
Supplemental Figure 1: Visualization of patient-specific spectrum (SOLAIS).



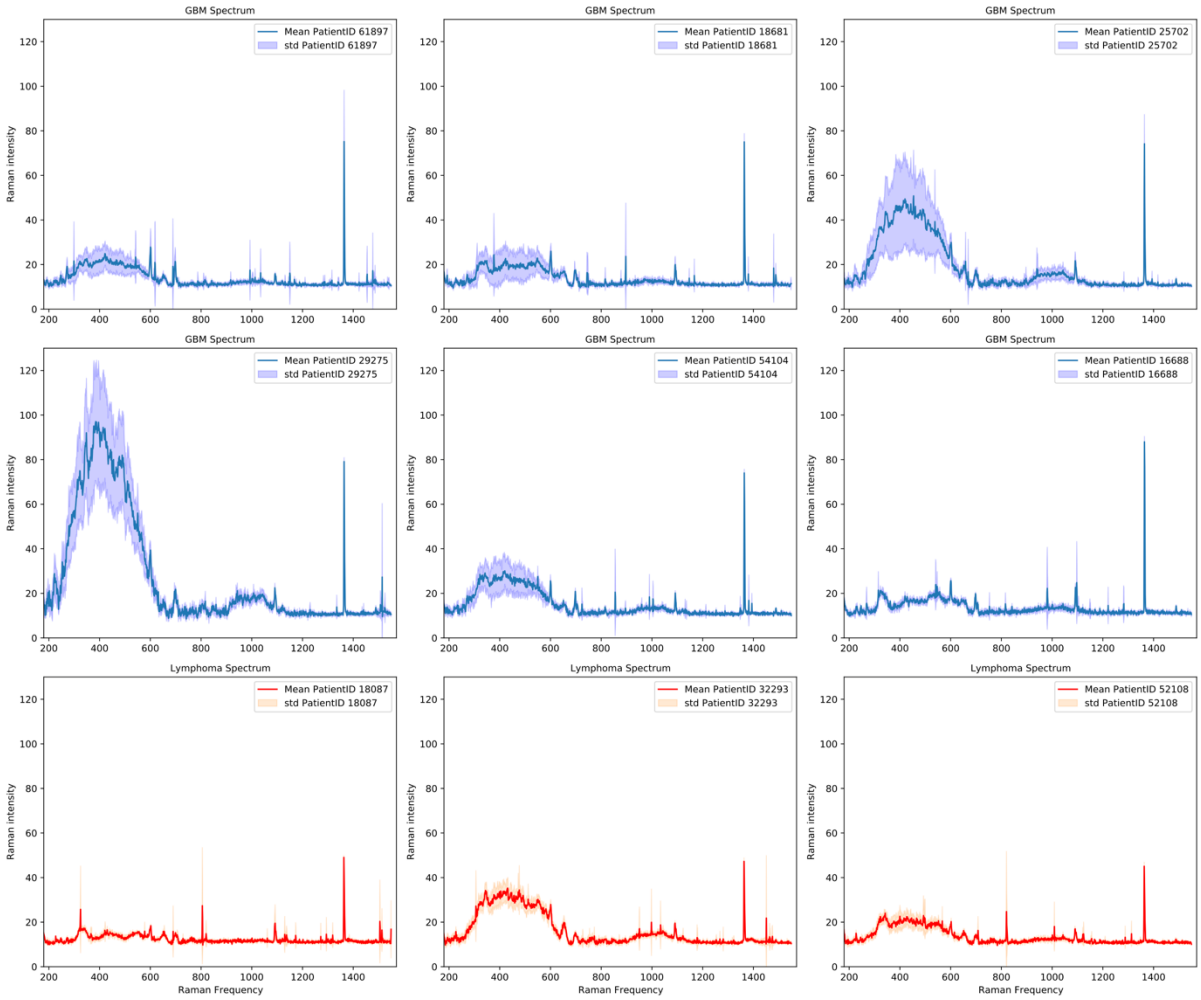
Supplemental Figure 2: Results of Logistic Regression by Random Shuffling (upper) and Patient Stratification (lower) (SOLAIS).



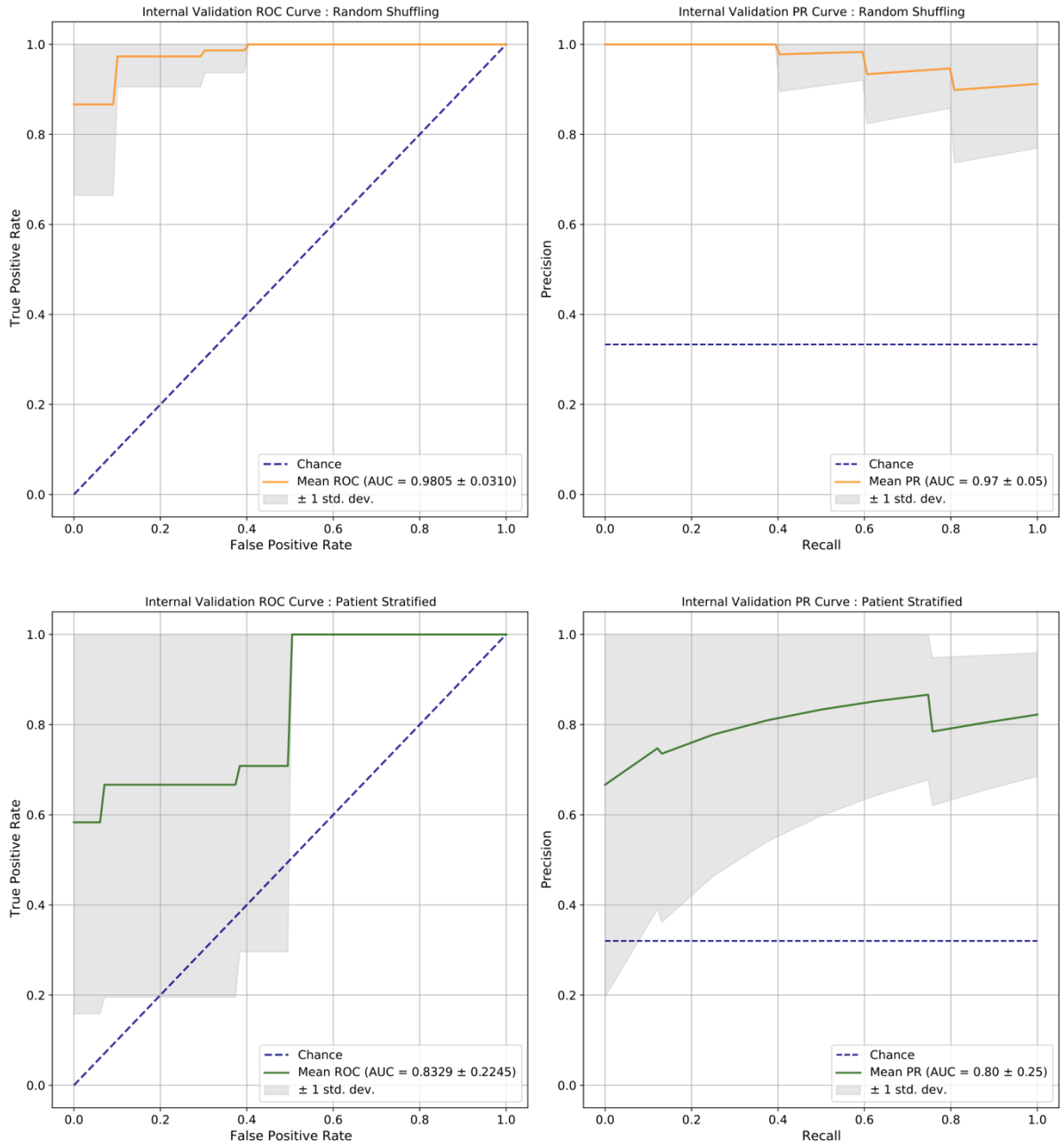
Supplemental Figure 3: Results of Random forest by Random Shuffling (upper) and Patient Stratification (lower) (SOLAIS).



Supplemental Figure 4: Results of XGBoost by Random Shuffling (upper) and Patient Stratification (lower) (SOLAIS).



Supplemental Figure 5: Visualization of patient-specific spectrum (FFPE).



Supplemental Figure 6: Results of Random forest by Random Shuffling and Patient Stratification (FFPE).

tumor sample no°	age	sex	localization	IDH status (wildtype / mutant)	MGMT status (methylated / unmethylated)	ALA positive (+) / negative (-)
1	75	female	right central	wildtype	unmethylated	+
2	64	male	right fronto-dorsal	wildtype	methylated	+
3	72	female	right temporal	wildtype	methylated	+
4	50	female	left temporal	not specified	methylated	+
5	72	male	right opercular	wildtype	methylated	+
6	49	male	left fronto-temporal	wildtype	methylated	-
7	58	male	supratentorial	wildtype	unmethylated	no use of ALA
8	64	female	right occipital	wildtype	methylated	not specified
9	27	male	left frontal	mutant	unmethylated	+
10	67	male	left fronto-basal	wildtype	unmethylated	not specified

Supplemental Table 1. Additional information of the intraoperative measured glioblastoma samples.

tumor sample no°	age	sex	localization	IDH status (wildtype / mutant)	MGMT status (methylated / unmethylated)	ALA positive (+) / negative (-)
1	61	male	right fronto-temporal	wildtype	methylated	not specified
2	42	female	right temporal	mutant	inconclusive	+
3	75	female	right central	wildtype	unmethylated	+
4	56	male	left occipital	wildtype	not specified	not specified
5	64	male	left temporal / parietal	wildtype	methylated	+
6	51	male	right temporal	wildtype	methylated	not specified

Supplemental Table 2. Additional information of the measured glioblastoma samples (FFPE tissue).