



**Figure S1. Machine learning approach for feature selection and classification. (a)** The misclassification rate as a function of the number of included variables. The light grey lines in the graph represent the validation performance of individual inner segments, while the darker grey lines show the average validation performance of all inner segments. The minimal-optimal ("Min") and all-relevant ("Max") models represent the two extremes of variable selection, with validation performance (measured in misclassifications) minimized within 5% of the minimum. The minimal-optimal model represents the minimum number of variables needed for optimal method performance, such as in biomarker discovery, while the all-relevant model includes all variables with relevant signal-to-noise for the research question, including redundant but non-erroneous variables. The "Mid" model represents a compromise between the "Min" and "Max" models, found at the geometric mean. **(b)** A Comparison of classification error rates for five different classification algorithms: SVM, RF, GLMnet, Featureless, and Random Guessing on Panel A (the overlapping proteins between MUVr and Boruta). The Featureless algorithm assigns all samples to the largest class, while the Random Guessing algorithm randomly assigns samples to classes. These algorithms are included for comparison. The SVM algorithm showed the best performance, with error rates similar to those of the Random Forests and GLMnet algorithms and much better than Featureless and Random Guessing algorithms. **(c)** The performance characteristics of different classification algorithms.