Supporting Information

Machine-Learning-Guided Identification of Protein Secondary Structures Using Spectral and Structural Descriptors

Ziqi Wang[1], Kenry[1,2,3,*]

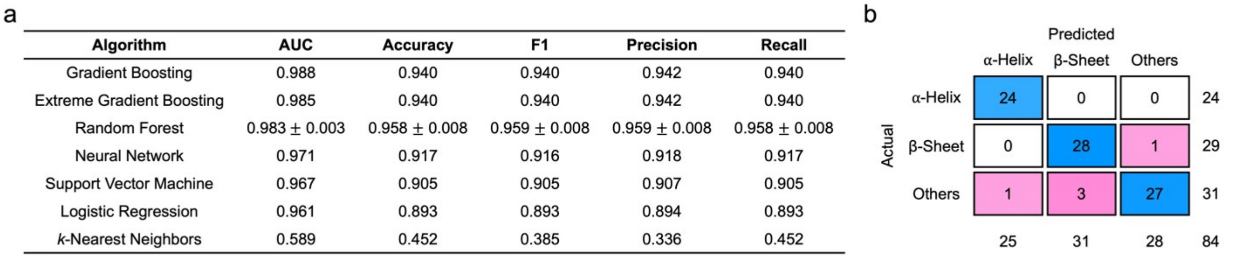
[1]Department of Pharmacology and Toxicology, R. Ken Coit College of Pharmacy, University of Arizona, Tucson, AZ 85721, USA

[2]Clinical and Translational Oncology Program and Skin Cancer Institute, University of Arizona Cancer Center, University of Arizona, Tucson, AZ 85721, USA

[3]BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA

*Email: kenry@arizona.edu

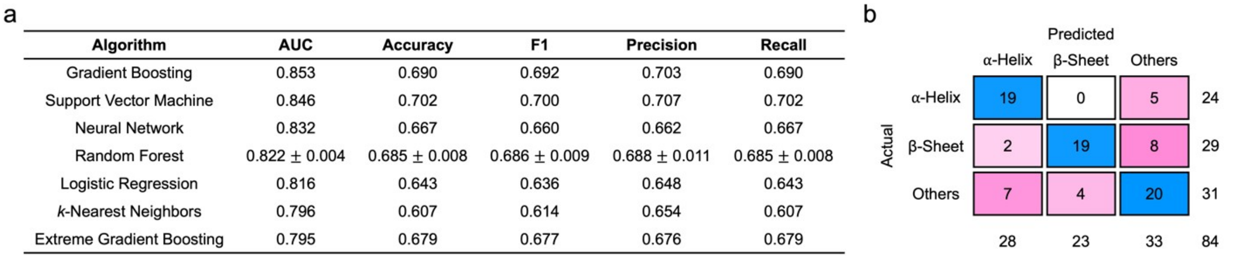
Supporting Figures



Algorithm	AUC	Accuracy	F1	Precision	Recall
Gradient Boosting	0.988	0.940	0.940	0.942	0.940
Extreme Gradient Boosting	0.985	0.940	0.940	0.942	0.940
Random Forest	0.983 ± 0.003	0.958 ± 0.008	0.959 ± 0.008	0.959 ± 0.008	0.958 ± 0.008
Neural Network	0.971	0.917	0.916	0.918	0.917
Support Vector Machine	0.967	0.905	0.905	0.907	0.905
Logistic Regression	0.961	0.893	0.893	0.894	0.893
k-Nearest Neighbors	0.589	0.452	0.385	0.336	0.452

		Predicted			
		α-Helix	β-Sheet	Others	
Actual	α-Helix	24	0	0	24
	β-Sheet	0	28	1	29
	Others	1	3	27	31
		25	31	28	84

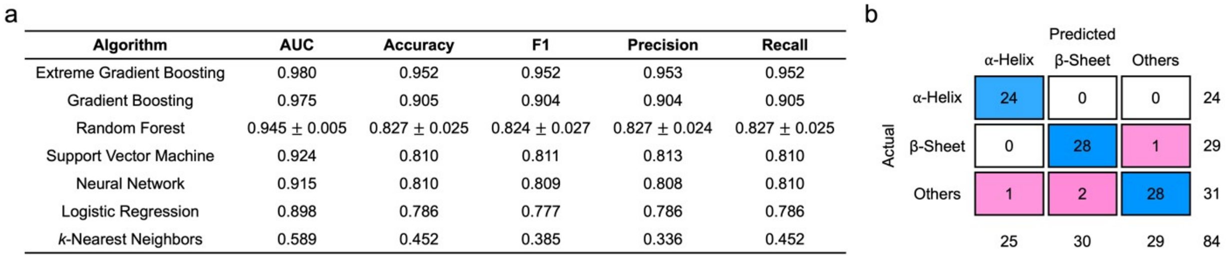
Figure S1. Classification of protein secondary structures based on structural and molecular descriptors during classifier training. (**a**) Comparison of the training performance of all supervised machine learning algorithms leveraging structural and molecular features. *n* = 2 for random forest. (**b**) Confusion matrix of the best performing classifier (i.e., gradient boosting).



Algorithm	AUC	Accuracy	F1	Precision	Recall
Gradient Boosting	0.853	0.690	0.692	0.703	0.690
Support Vector Machine	0.846	0.702	0.700	0.707	0.702
Neural Network	0.832	0.667	0.660	0.662	0.667
Random Forest	0.822 ± 0.004	0.685 ± 0.008	0.686 ± 0.009	0.688 ± 0.011	0.685 ± 0.008
Logistic Regression	0.816	0.643	0.636	0.648	0.643
k-Nearest Neighbors	0.796	0.607	0.614	0.654	0.607
Extreme Gradient Boosting	0.795	0.679	0.677	0.676	0.679

		Predicted			
		α-Helix	β-Sheet	Others	
Actual	α-Helix	19	0	5	24
	β-Sheet	2	19	8	29
	Others	7	4	20	31
		28	23	33	84

Figure S2. Classification of protein secondary structures based on full spectral descriptors during classifier training. (**a**) Comparison of the training performance of all supervised machine learning algorithms leveraging full spectral data. *n* = 2 for random forest. (**b**) Confusion matrix of the best performing classifier (i.e., gradient boosting).



a

Algorithm	AUC	Accuracy	F1	Precision	Recall
Extreme Gradient Boosting	0.980	0.952	0.952	0.953	0.952
Gradient Boosting	0.975	0.905	0.904	0.904	0.905
Random Forest	0.945 ± 0.005	0.827 ± 0.025	0.824 ± 0.027	0.827 ± 0.024	0.827 ± 0.025
Support Vector Machine	0.924	0.810	0.811	0.813	0.810
Neural Network	0.915	0.810	0.809	0.808	0.810
Logistic Regression	0.898	0.786	0.777	0.786	0.786
k-Nearest Neighbors	0.589	0.452	0.385	0.336	0.452

b

		Predicted			
		α-Helix	β-Sheet	Others	
Actual	α-Helix	24	0	0	24
	β-Sheet	0	28	1	29
	Others	1	2	28	31
		25	30	29	84

Figure S3. Classification of protein secondary structures based on full spectral, structural, and molecular descriptors during classifier training. (**a**) Comparison of the training performance of all supervised machine learning algorithms leveraging full spectral, structural, and molecular descriptors. *n* = 2 for random forest. (**b**) Confusion matrix of the best performing classifier (i.e., extreme gradient boosting).

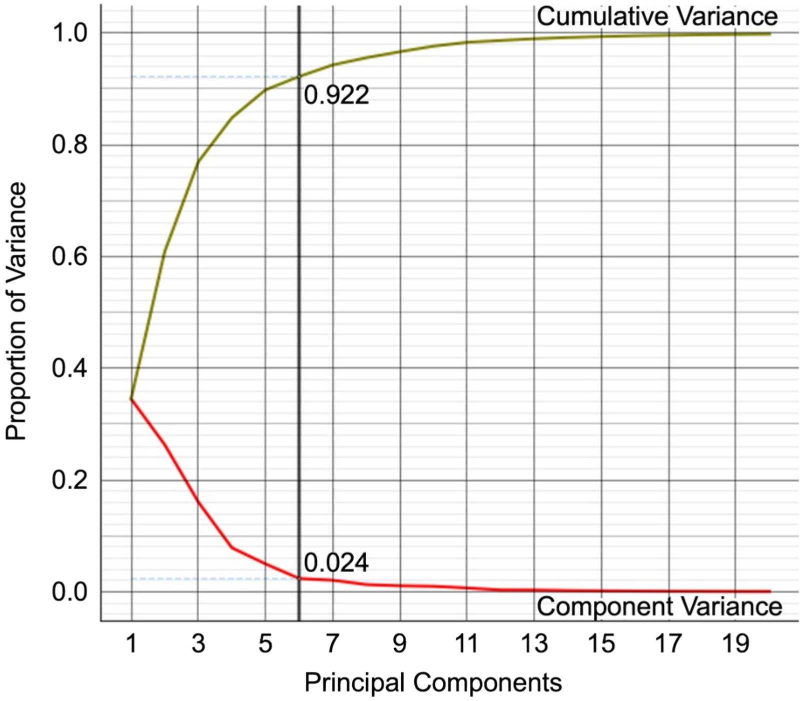
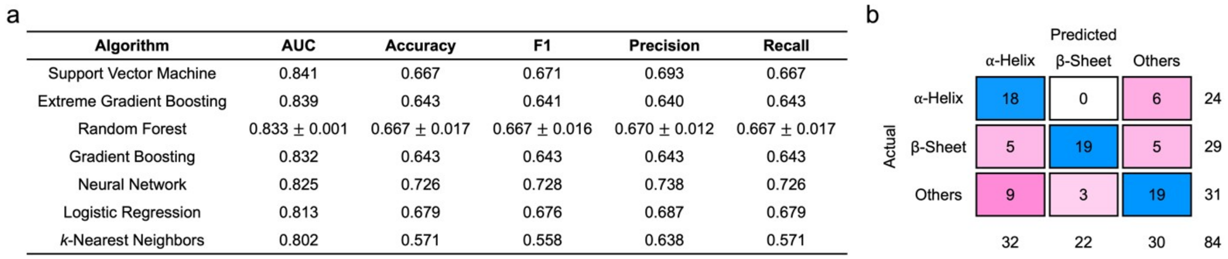


Figure S4. Proportion of spectral data variance as a function of the number of principal components.



a

Algorithm	AUC	Accuracy	F1	Precision	Recall
Support Vector Machine	0.841	0.667	0.671	0.693	0.667
Extreme Gradient Boosting	0.839	0.643	0.641	0.640	0.643
Random Forest	0.833 ± 0.001	0.667 ± 0.017	0.667 ± 0.016	0.670 ± 0.012	0.667 ± 0.017
Gradient Boosting	0.832	0.643	0.643	0.643	0.643
Neural Network	0.825	0.726	0.728	0.738	0.726
Logistic Regression	0.813	0.679	0.676	0.687	0.679
k-Nearest Neighbors	0.802	0.571	0.558	0.638	0.571

b

		Predicted			
		α-Helix	β-Sheet	Others	
Actual	α-Helix	18	0	6	24
	β-Sheet	5	19	5	29
	Others	9	3	19	31
		32	22	30	84

Figure S5. Classification of protein secondary structures based on dimensionally reduced spectral descriptors during classifier training. (**a**) Comparison of the training performance of all supervised machine learning algorithms leveraging dimensionally reduced spectral data. *n* = 2 for random forest. (**b**) Confusion matrix of the best performing classifier (i.e., support vector machine).

Supporting Tables

Table S1. The different classifier hyperparameters and values considered in this study.

Classifier	Hyperparameter	Value
Logistic Regression	Regularization Strength	Lasso, Ridge 1000, 200, 1, 0.02, 0.001
Random Forest	Number of trees	5, 10, 50, 200
Gradient Boosting	Number of trees Learning rate Regularization Limit depth of individual trees	5, 10, 50, 200 0.01, 0.1, 1 0.0001, 0.01, 1, 100 3, 6, 12
Extreme Gradient Boosting	Number of trees Learning rate Regularization Limit depth of individual trees	5, 10, 50, 200 0.01, 0.1, 1 0.0001, 0.01, 1, 100 3, 6, 12
k-Nearest Neighbors	Number of neighbors Metric Weight	1, 3, 5, 10, 20 Euclidean, Manhattan, Chebyshev Uniform, Distance
Support Vector Machine	Cost Regression loss epsilon Kernel Iteration limit	0.1, 1, 10, 100 0.1, 1, 10, 50, 100 Linear, Polynomial, RBF, Sigmoid 10, 100, 10000
Neural Network	Neurons in hidden layers Activation Solver Regularization Maximal number of iterations	10, 100, 500 ReLu, Identity, Logistic, tanh Adam, SGD, L-BFGS-B 0.0001, 0.01, 1, 100 10, 200, 500

Table S2. Optimized classifier hyperparameters during training based on structural and molecular descriptors.

Classifier	Hyperparameter	Value
Logistic Regression	Regularization	Ridge
	Strength	200
Random Forest	Number of trees	50
Gradient Boosting	Number of trees	50
	Learning rate	0.01
	Regularization	0.01
	Limit depth of individual trees	12
Extreme Gradient Boosting	Number of trees	5
	Learning rate	1
	Regularization	0.0001
	Limit depth of individual trees	3
k-Nearest Neighbors	Number of neighbors	20
	Metric	Euclidean
	Weight	Uniform
Support Vector Machine	Cost	100
	Regression loss epsilon	0.1
	Kernel	RBF
	Iteration limit	10000
Neural Network	Neurons in hidden layers	10
	Activation	ReLu
	Solver	Adam
	Regularization	1
	Maximal number of iterations	500

Table S3. Optimized classifier hyperparameters during training based on full spectral descriptors.

Classifier	Hyperparameter	Value
Logistic Regression	Regularization	Ridge
	Strength	1
Random Forest	Number of trees	200
Gradient Boosting	Number of trees	10
	Learning rate	0.1
	Regularization	0.01
	Limit depth of individual trees	6
Extreme Gradient Boosting	Number of trees	5
	Learning rate	1
	Regularization	0.0001
	Limit depth of individual trees	6
k-Nearest Neighbors	Number of neighbors	5
	Metric	Chebyshev
	Weight	Distance
Support Vector Machine	Cost	1
	Regression loss epsilon	10
	Kernel	RBF
	Iteration limit	100
Neural Network	Neurons in hidden layers	100
	Activation	ReLu
	Solver	SGD
	Regularization	1
	Maximal number of iterations	500

Table S4. Optimized classifier hyperparameters during training based on full spectral, structural, and molecular descriptors.

Classifier	Hyperparameter	Value
Logistic Regression	Regularization Strength	Lasso 1
Random Forest	Number of trees	50
Gradient Boosting	Number of trees Learning rate Regularization Limit depth of individual trees	50 0.1 100 3
Extreme Gradient Boosting	Number of trees Learning rate Regularization Limit depth of individual trees	50 1 100 6
k-Nearest Neighbors	Number of neighbors Metric Weight	20 Euclidean Uniform
Support Vector Machine	Cost Regression loss epsilon Kernel Iteration limit	100 50 RBF 100
Neural Network	Neurons in hidden layers Activation Solver Regularization Maximal number of iterations	500 Logistic L-BFGS-B 1 200

Table S5. Optimized classifier hyperparameters during training based on dimensionally reduced spectral descriptors.

Classifier	**Hyperparameter**	**Value**
Logistic Regression	Regularization Strength	Lasso 200
Random Forest	Number of trees	200
Gradient Boosting	Number of trees Learning rate Regularization Limit depth of individual trees	50 0.1 0.01 3
Extreme Gradient Boosting	Number of trees Learning rate Regularization Limit depth of individual trees	200 0.01 0.0001 12
k-Nearest Neighbors	Number of neighbors Metric Weight	20 Chebyshev Distance
Support Vector Machine	Cost Regression loss epsilon Kernel Iteration limit	10 50 Polynomial 10000
Neural Network	Neurons in hidden layers Activation Solver Regularization Maximal number of iterations	10 tanh Adam 0.0001 500