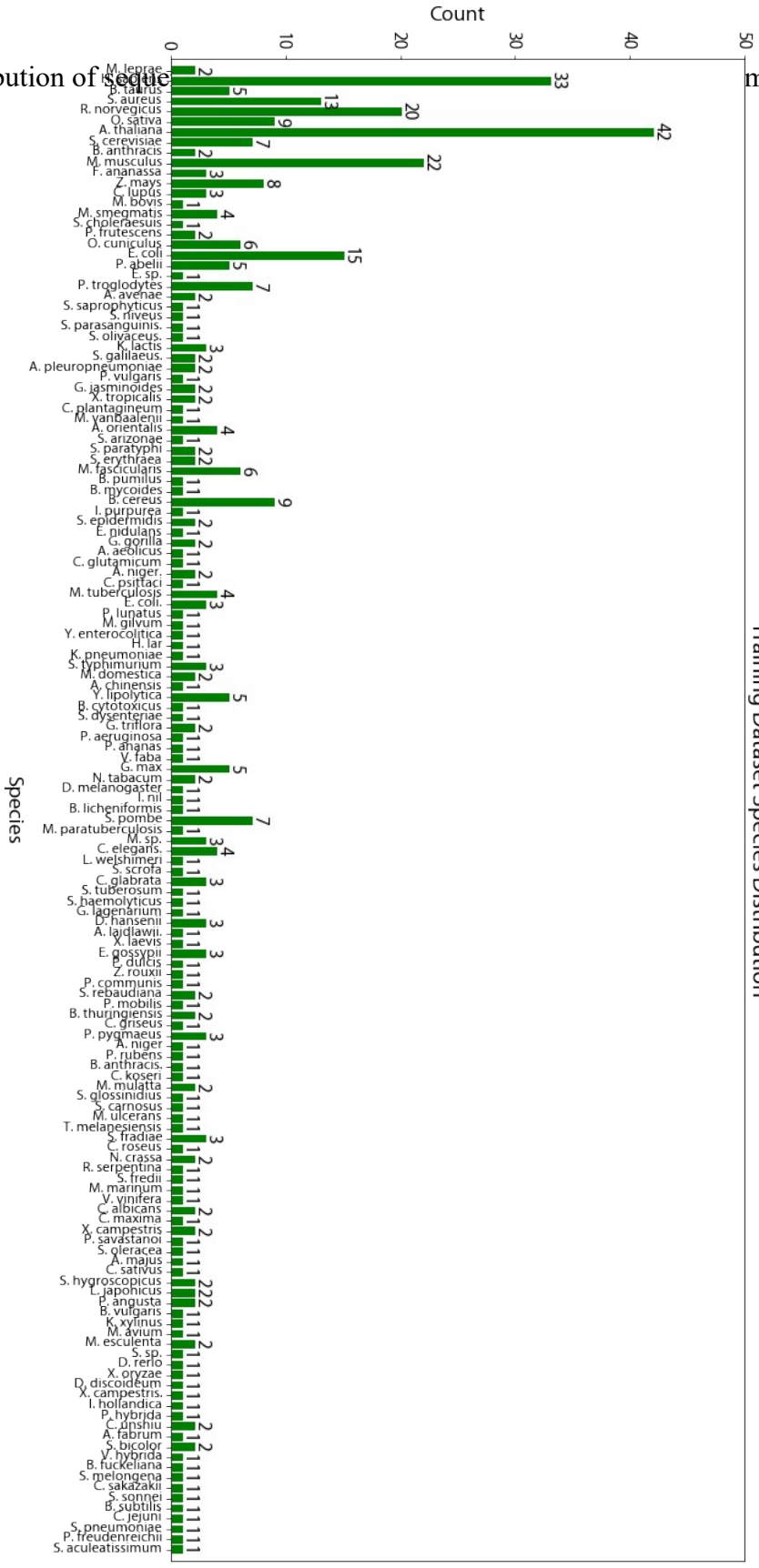
Supplementary Information for

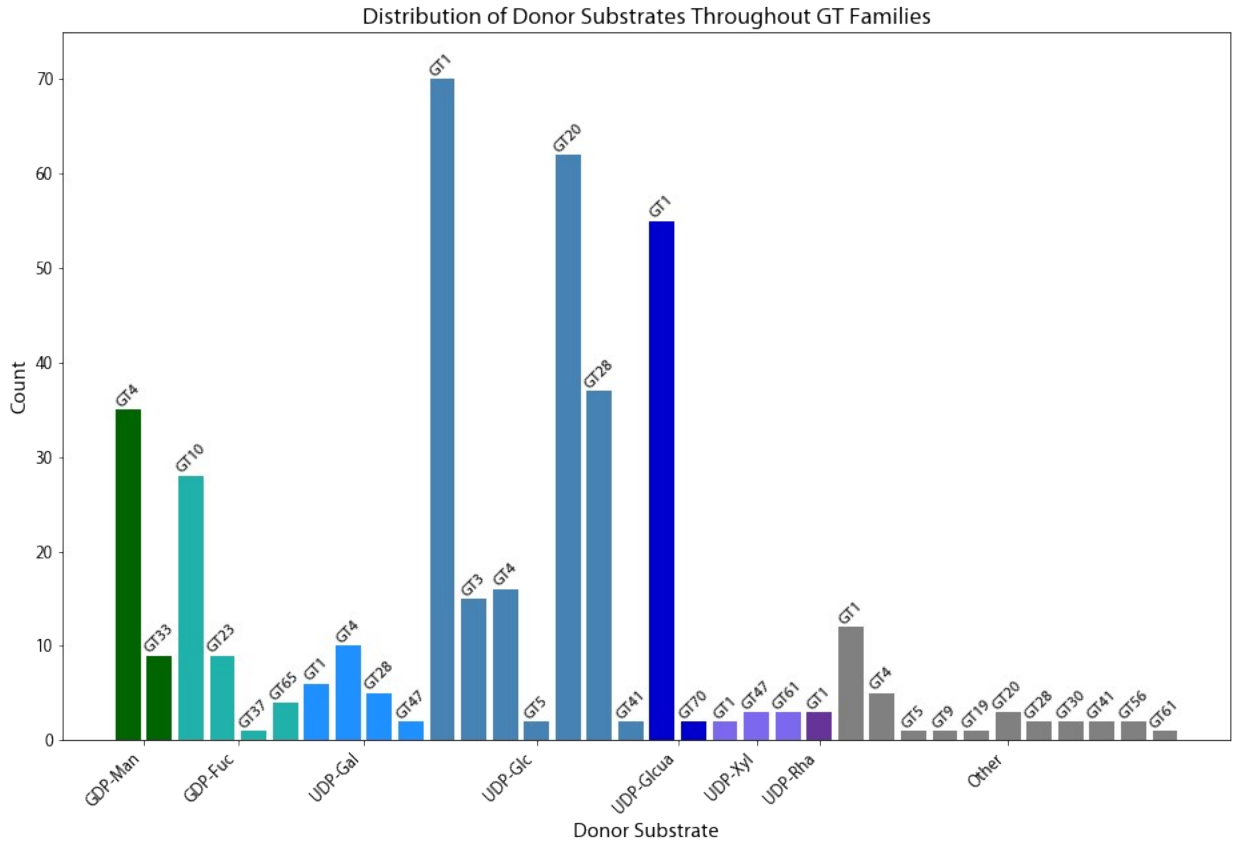
Decoding substrate specificity determining factors in glycosyltransferase-B enzymes – Insights from machine learning models

Samantha G. Hennen[1], Yannick J. Bomble[1], Breanna R. Urbanowicz[3,4], Vivek S. Bharadwaj[2]*

[1]Biosciences Center, National Renewable Energy Laboratory, Golden, CO, [2] Renewable Resources and Enabling Sciences Center, National Renewable Energy Laboratory, Golden, CO,[3] Complex Carbohydrate Research Center, University of Georgia, Athens, GA, [4] Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA.



SI Figure 1. Distribution of sequ mal species.



SI Figure 2. Number of sequences (and their family identity) in the training and test datasets with known activity for each nucleotide sugar donor substrate. Each bar denotes the number of sequences from a family with the color denoting its nucleotide-sugar activity.

SI Table 1: The number of sequences per family in the training dataset is shown below.

Sequences	Family	Donor Substrates
148	GT1	UDP-Glucua, UDP-Glu, UDP-Xyl, UDP-Gal, UDP-Rha, Other
66	GT4	UDP-Glu, GDP-Man, UDP-Gal, Other
65	GT20	UDP-Glu, Other
44	GT28	UDP-Glu, UDP-Gal, Other
28	GT10	GDP-Fuc
15	GT3	UDP-Glu
9	GT33	GDP-Man
9	GT23	GDP-Fuc
5	GT47	UDP-Xyl, UDP-Gal
4	GT41	UDP-Glu, Other
4	GT61	UDP-Xyl, Other
4	GT65	GDP-Fuc
3	GT5	UDP-Glu, Other
2	GT30	Other
2	GT56	Other
2	GT70	UDP-Glucua
1	GT9	Other
1	GT19	Other
1	GT37	GDP-Fuc

SI Table 2: All model hyperparameters used in the training grid search.

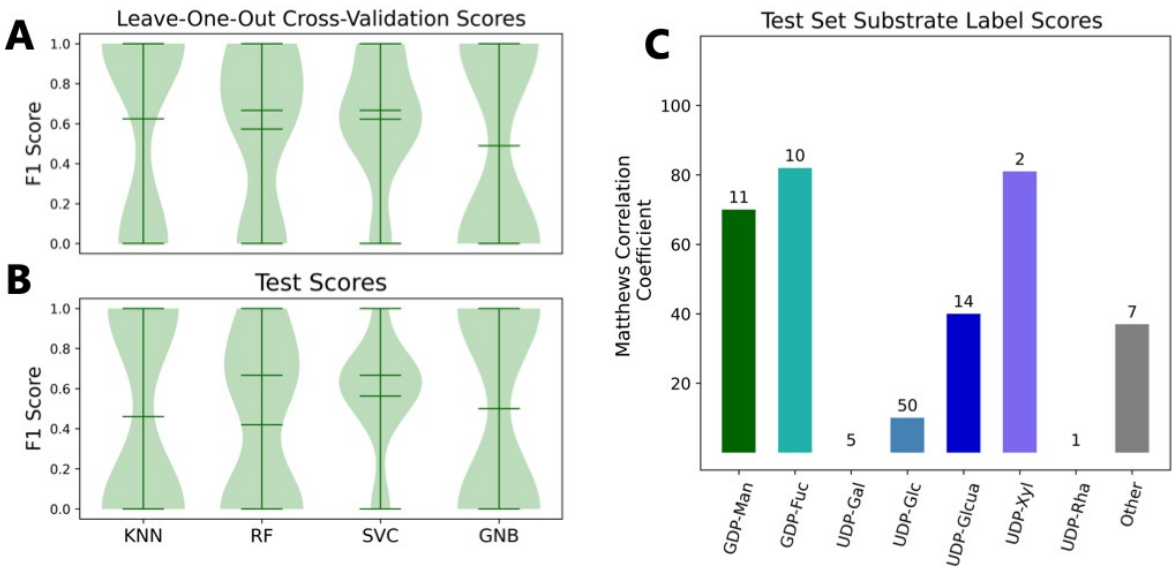
Model Type	**Hyperparameter Ranges**			
Gaussian Naïve Bayes	*Variance Smoothing*			
	100 points in logspace (0,-9)			
K-Nearest Neighbors	*Number of Neighbors*	*Power Parameter*	*Weights*	
	1-10	1, 2	Uniform, Distance	
Random Forest	*Number of Trees*	*Maximum Features per Split*	*Criterion*	*Class Weight*
	20-200, in multiples of 20	0.1 ,0.3, 0.5	Gini, Entropy	Balanced, Balanced Subsample
Support Vector	*Regularization Parameter*	*Maximum Iterations*	*Kernel*	*Gamma*
	0.1, 1, 10	10, 50, 100	Radial Basis Function, Linear, Polynomial	1, 0.1, 0.01

SI Table 3. The mean cross-validation and test set score for each best performing model is shown. The standard deviations are high to only one substrate predicted for each sample, thus making the predictions binary.

Model Type	Cross-Validation F1 Score	Test F1 Score
KNN	94.2% ± 23.4%	85.0% ± 35.7%
RF	89.8% ± 30.2%	44.0% ± 49.6%
SVC	91.5% ± 27.9%	59.0% ± 49.2%
GNB	84.5% ± 32.8%	42.7% ± 48.6%

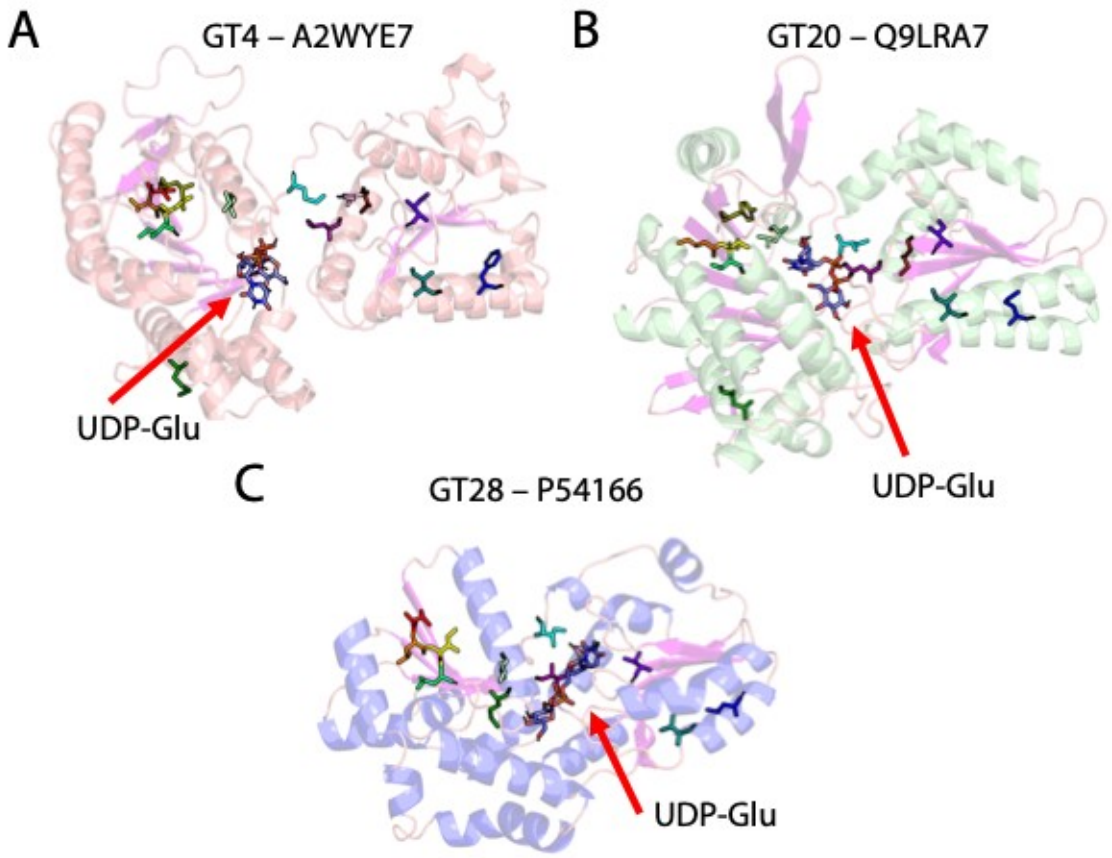
SI Table 4: Optimal hyperparameters and feature lengths of all models found in the training grid search.

Model Type	**Hyperparameter Ranges**				
Gaussian Naïve Bayes	*Variance Smoothing*				*Feature Length*
	1.519911082952933e-09				100
K-Nearest Neighbors	*Number of Neighbors*	*Power Parameter*	*Weights*		*Feature Length*
	1	1	Uniform		550
Random Forest	*Number of Trees*	*Maximum Features per Split*	*Criterion*	*Class Weight*	*Feature Length*
	100	0.1	Entropy	Balanced	500
Support Vector	*Regularization Parameter*	*Maximum Iterations*	*Kernel*	*Gamma*	*Feature Length*
	0.1	100	Linear	1	950



SI Figure 3. F1 cross-validation (A) and test scores (B) from all models trained on only the family number are shown. (C) The Matthews Correlation Coefficient scores of the best performing KNN model for each test set substrate are also shown. As expected, the cross-validation scores of all family-based models are lower than their counterpart models generated with additional features. The best CV set score of 63% ± 48.4% for the KNN model (test score 46% ± 49.8%), is lower than the best test set score of 94.1% ± 23.4%) for the KNN model (Figure 3B) built with the complete feature set (Figure 3) (with test score 85% ± 35.7%)).

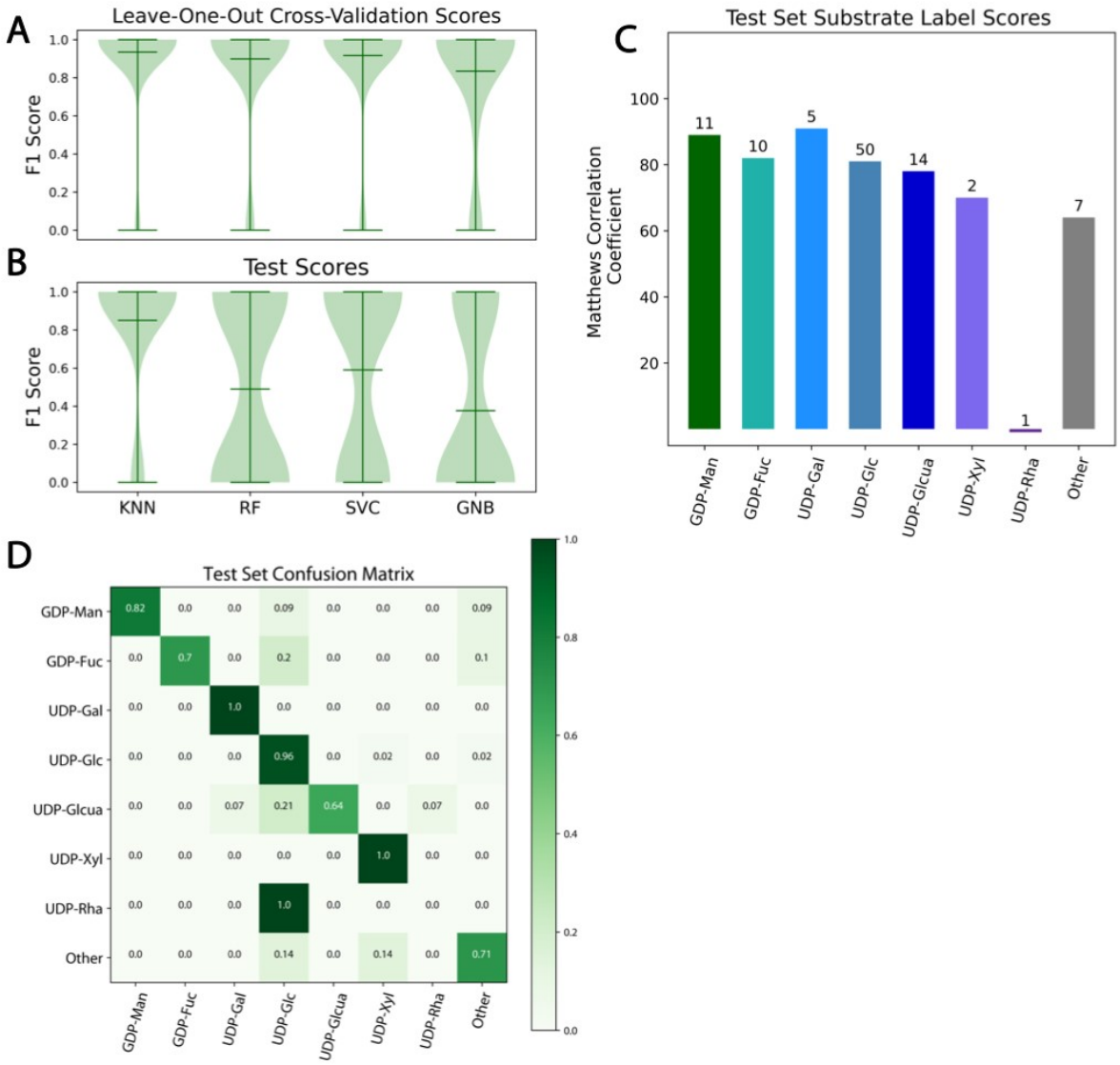
The individual substrate MCC scores are also lower than in the more complex model, showing similar or lower scores on all substrates except UDP-β-L-xylose. The high performance on this single substrate is notable, as it has higher accuracy than in the more complex model. Nonetheless, the poor performance on the additional substrates shows this family-based model's inability to predict nucleotide sugar donor substrates.



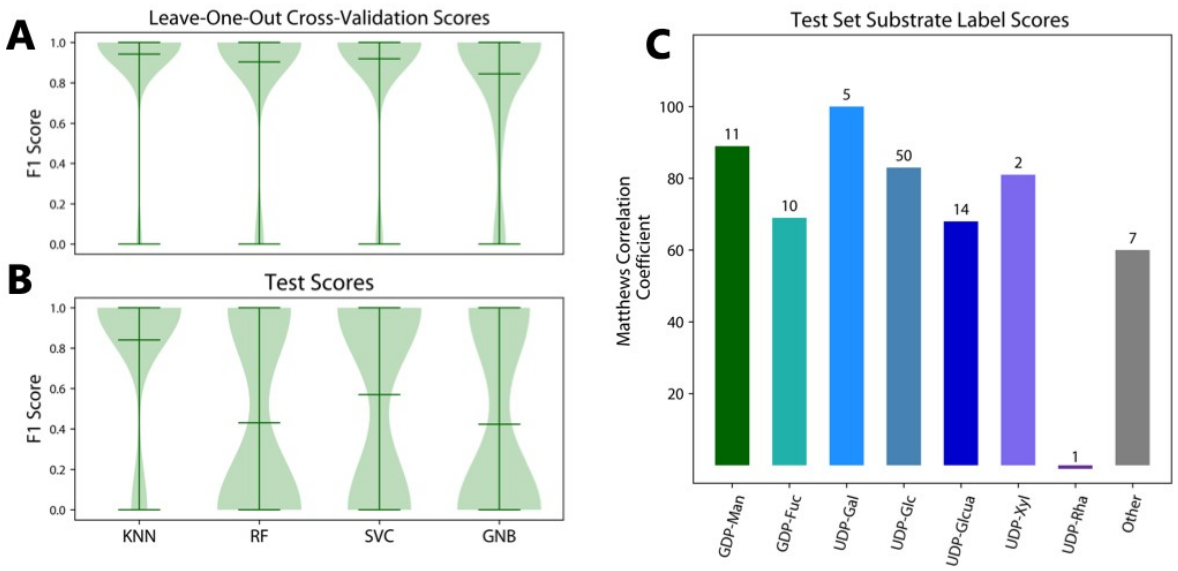
SI Figure 4. (A-C) UDP-α-D-glucose was docked to the truncated representative structures A2WYE7, Q9LRA7, and P54166 from families GT4, GT20, and GT28, respectively. A top pose for each structure is shown, along with residues found to be highly conserved (by residue type) within these families.

SI Table 5: The family distribution of uncharacterized sequences from distinct plant genera datasets.

	Genera			
Family	***Populus***	***Spirodela***	***Eucalyptus***	***Chlamydomonas***
GT1	91	73	3	324
GT4	0	2	4	0
GT5	1	0	3	0
GT10	0	0	1	0
GT28	13	2	7	7
GT37	28	13	1	3
GT41	12	3	1	2
GT47	147	49	139	35
GT61	2	0	0	0
GT92	14	4	3	4



SI Figure 5. F1 cross-validation (A) and test scores (B) from all models trained without solvent accessible surface area and secondary structure values are shown. (C-D) The Matthews Correlation Coefficient scores and confusion matrix of the best performing KNN model for each test set substrate are also shown.



SI Figure 6. Additional models were trained removing the 70% AF2 confidence score minimum to be assigned secondary structure and SASA values. F1 cross-validation (A) and test scores (B) are shown. (C) The Matthews Correlation Coefficient scores of the best performing KNN model for each test set substrate are also shown. These scores are very similar to the KNN model with the restriction.