JBC RESEARCH ARTICLE

Machine learning reveals sequence-function relationships in family 7 glycoside hydrolases

Received for publication, November 5, 2020, and in revised form, June 18, 2021 Published, Papers in Press, July 1, 2021,
https://doi.org/10.1016/j.jbc.2021.100931

Japheth E. Gado[1,2], Brent E. Harrison[3], Mats Sandgren[4], Jerry Ståhlberg[4], Gregg T. Beckham[2], and Christina M. Payne[1,*]

From the [1]Department of Chemical and Materials Engineering, University of Kentucky, Lexington, Kentucky, USA; [2]Renewable Resources and Enabling Sciences Center, National Renewable Energy Laboratory, Golden, Colorado, USA; [3]Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA; and [4]Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

Edited by Gerald Hart

Family 7 glycoside hydrolases (GH7) are among the principal enzymes for cellulose degradation in nature and industrially. These enzymes are often bimodular, including a catalytic domain and carbohydrate-binding module (CBM) attached *via* a flexible linker, and exhibit an active site that binds cello-oligomers of up to ten glucosyl moieties. GH7 cellulases consist of two major subtypes: cellobiohydrolases (CBH) and endoglucanases (EG). Despite the critical importance of GH7 enzymes, there remain gaps in our understanding of how GH7 sequence and structure relate to function. Here, we employed machine learning to gain data-driven insights into relationships between sequence, structure, and function across the GH7 family. Machine-learning models, trained only on the number of residues in the active-site loops as features, were able to discriminate GH7 CBHs and EGs with up to 99% accuracy, demonstrating that the lengths of loops A4, B2, B3, and B4 strongly correlate with functional subtype across the GH7 family. Classification rules were derived such that specific residues at 42 different sequence positions each predicted the functional subtype with accuracies surpassing 87%. A random forest model trained on residues at 19 positions in the catalytic domain predicted the presence of a CBM with 89.5% accuracy. Our machine learning results recapitulate, as top-performing features, a substantial number of the sequence positions determined by previous experimental studies to play vital roles in GH7 activity. We surmise that the yet-to-be-explored sequence positions among the top-performing features also contribute to GH7 functional variation and may be exploited to understand and manipulate function.

Cellulose is the most abundant renewable biopolymer on Earth and, thus, holds tremendous potential in transitioning energy production from fossil fuels to a renewable carbon feedstock—a key need to limit anthropogenic climate change. Sugars derived from the deconstruction of cellulose can be converted to biofuels and numerous chemicals *via* myriad biological or catalytic conversion routes. However, efficiently depolymerizing cellulose in a cost-effective manner, such that biofuels can economically compete with fossil fuels, remains a substantial challenge to enabling a lignocellulosic economy (1). In industry, biochemical methods of cellulose deconstruction employing enzymes are promising due to high selectivity, low energy consumption, and low amounts of by-product generation (1–3). As a result, improving the yield of enzymatic hydrolysis of cellulose by enhancing cellulase activity is a major research focus.

In nature, microbial cellulose degradation is primarily achieved *via* a synergistic cocktail of enzymes consisting of processive cellobiohydrolases (CBHs), endoglucanases (EGs), and accessory enzymes such as β-glucosidases and lytic polysaccharide monooxygenases (LPMOs) (2). Organisms can employ these enzymes as free single- or multi-modular constructs or as cellulosomes. Industry tends to employ free enzyme systems, as filamentous fungal hosts are proficient secretors of these types of cellulose-degrading enzymes. EGs act by attacking internal bonds in cellulose, thus creating free chain ends. CBHs attach to free chain ends *via* exo-initiation, or internal regions in the chain *via* endo-initiation, and processively cleave off cellobiose units as they process along the chain. Cellobiose products are consequently hydrolyzed by β-glucosidases to yield glucose (2). Whereas CBHs are known to be processive and to carry out several cellulolytic cuts before detaching from the cellulose substrate, EGs are mostly non-processive or may show little processivity (4–8). Optimum cellulolytic efficiency is achieved by the synergistic action of CBHs and EGs. CBHs, EGs, and β-glucosidases, as well as other glycoside hydrolases (GHs), are currently classified into 168 families in the CAZy database (www.cazy.org) (9).

Family 7 glycoside hydrolases (GH7s) are the powerhouses of cellulose degradation in nature. They traditionally are found mostly in fungi, although sequences have been identified in several nonfungal groups such as Crustacea, Porifera, Alveolata, and Amoeba (10). Because GH7s offer significant cellulolytic potential, they are often the predominant enzymes by mass in the secretomes of many filamentous cellulolytic fungi and constitute the primary components of enzyme cocktails in industrial cellulolytic processes (2, 11, 12).

* For correspondence: Christina M. Payne, christy.payne@uky.edu.

© 2021 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ML reveals GH7 sequence–function relationships

GH7s consist of two main subtypes, CBHs and EGs. Although over 5000 GH7 sequences are known, structural information is presently available for only 21 GH7s (16 CBHs, five EGs) (10, 13–30). GH7 CBH and EG structures share a similar β-jelly roll fold with two antiparallel β-sheets that pack into a curved β-sandwich (14). Loops protrude from the β-sandwich and extend over a tunnel-like active site that spans 40–50 Å across the ends of the catalytic domain (CD). The active site contains at least nine glycosyl subsites for binding cello-oligomers, which are numbered −7 to +2 from the nonreducing end of the cellulose chain (Fig. 1). The cellulose chain is cleaved between the −1 and +1 subsites (2). Despite the overall similarity in fold, structures of GH7 CBHs and EGs are strikingly different in their active-site configuration. Whereas GH7 CBHs exhibit a closed tunnel-like active site, GH7 EGs possess a more open, groove-like active site. These differences arise due to the variation in the residue lengths of the loops that protrude over the active-site groove, labeled A1–A4 and B1–B4 (Fig. 1) (15).

Several structural and mechanistic studies of GH7s have proposed that the differences in functional properties of GH7 CBHs and EGs, such as processivity, endo-initiation, and product inhibition, arise mainly due to the differences in the active-site architecture in the loops (15, 16, 19, 21, 24, 25, 28, 29). Moreover, GH7 CBHs with a more exposed active site tend to exhibit functional characteristics intermediate between typical CBH and EG behavior (6, 19, 31, 32). Besides the differences in the configuration of active-site loops, studies have also indicated that there are key residues in the active site of GH7s that contribute to the variation in GH7 CBH and EG behavior. Several aromatic and charged residues in the active site that interact with the cellulose substrate have been suggested to be crucial for the processive activity of GH7 CBHs (22, 33–36). Furthermore, mutation of these residues notably

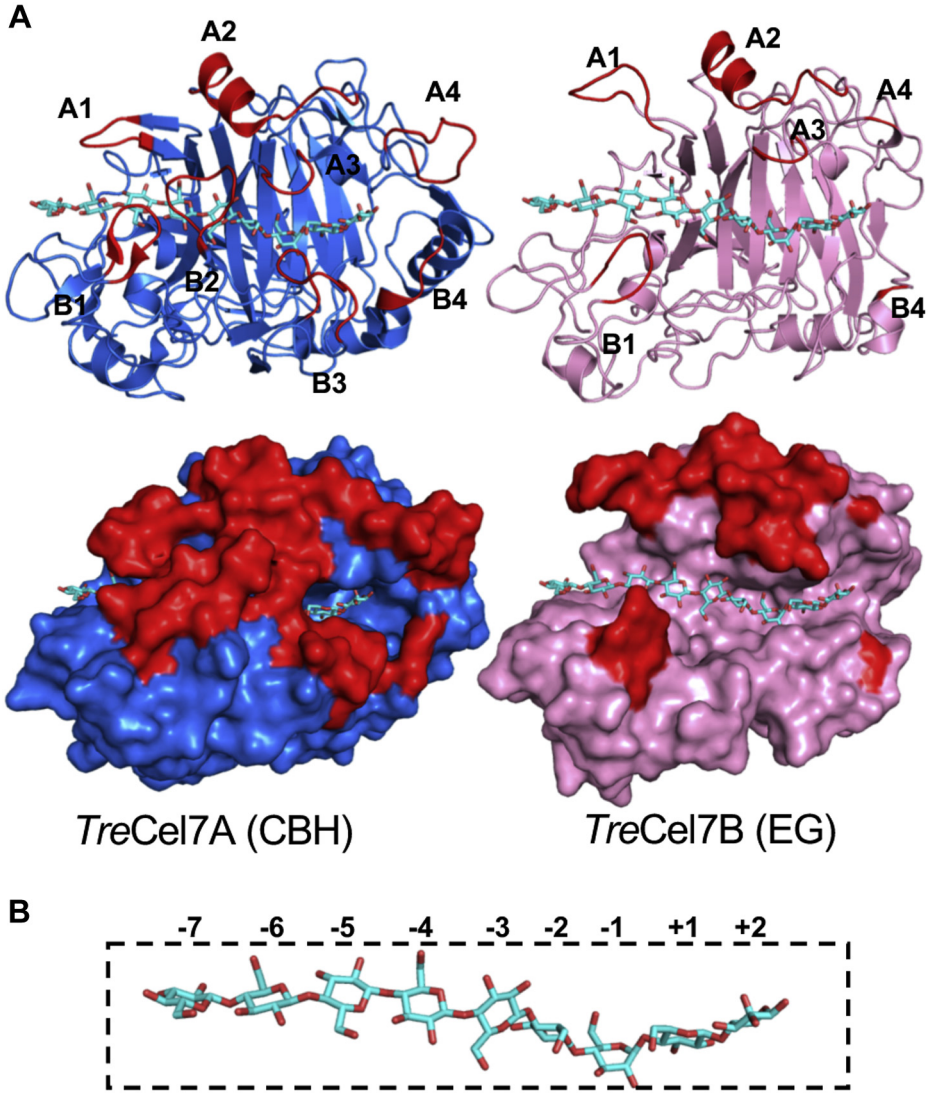


Figure 1. Structures of typical GH7 CBH and EG with a cellononaose ligand in complex. *A*, the CBH (*left*), *Trichoderma reesei* Cel7A (*Tre*Cel7A, PDB code: 4C4C) (23), and the EG (*right*), *Trichoderma reesei* Cel7B (*Tre*Cel7B, PDB code: 1EG1) (26). The eight active-site loops (A1–A4 and B1–B4) are shown in *red*. In the CBH, the active site is tunnel-like, but is more open and groove-like in the EG. *B*, glycosyl binding sites are numbered from the nonreducing end at the active-site tunnel entrance (−7) to the reducing end (+2) where the cellobiose product exits the active site. Bond cleavage occurs between −1 and +1 subsites.

affects the processive activity of GH7 CBHs on crystalline cellulose (37, 38).

Like many other cellulases, GH7s can be bimodular, having a CD attached to a carbohydrate-binding module (CBM) by an intrinsically disordered glycosylated linker peptide (39–43). There are currently 87 families of CBMs in the CAZy database (www.cazy.org/Carbohydrate-Binding-Modules.html) (9), but GH7s mainly utilize family 1 CBMs (2, 44). It is now generally accepted that family 1 CBMs function to increase the affinity of cellulases for crystalline cellulose and, thereby, increase the surface concentration of the enzymes for catalysis. Thus, by facilitating two-dimensional diffusion of the CD on the cellulose surface, the CBM improves catalytic efficiency (39). Furthermore, several studies have revealed that deletion of the CBM-linker domain dramatically reduces CBH activity on crystalline cellulose, especially at low enzyme concentration, but not on soluble substrates (44–49). Takashima *et al.* (50) carried out several mutations in the CBM of a *Humicola grisea* CBH (*Hgr*Cel7A) and observed high positive correlation between the efficiency of the enzyme on crystalline cellulose and the binding affinity of the CBM. Similarly, Srisodsuk *et al.* (48) observed that replacing the CBM of *Trichoderma reesei* Cel7A (*Tre*Cel7A) with the CBM of *Tre*Cel7B, which has a higher cellulose-binding affinity, improved the activity of *Tre*Cel7A on crystalline cellulose. Altogether, these results indicate that CBMs affect GH7 catalytic activity primarily by promoting binding to the cellulose surface.

Despite the tremendous growth in scientific knowledge of GH7s over the last few decades, our understanding of how sequence and structure affect function is far from complete. Although it is known that the exposure of the active site due to truncation in the active-site loops can substantially affect function, little work has been done to elucidate the unique roles that each of the active site loops plays and how the effects of truncation vary with function for the different loops. Recently, Schiano-di-Cola *et al.* (51) studied the effects of deletions in the B2, B3, and B4 loops on the activity and kinetics of *Tre*Cel7A. They found that deletions in the B2 loop, compared with the B3 and B4 loop, most significantly affect CBH behavior of *Tre*Cel7A. Beyond *Tre*Cel7A, there is a need to investigate how variation of active-site loop lengths relates to function across the larger GH7 family.

In this work, we employ machine learning (ML) and bioinformatic analysis to derive relationships between sequence, structure, and function of GH7s using a dataset of 1748 selected protein sequences. The sequences are aligned *via* multiple sequence alignment (MSA) to identify regions of structural similarity and evolutionary importance. Although manual inspection of the MSA may reveal several functional patterns, such as highly conserved positions, many important but nonintuitive relationships are likely to be missed. ML is an especially useful statistical tool when data are abundant and relationships in the data are complex (52). By mapping sequence variation to functional diversity, ML can discover sequence features that are statistically related to function and potentially play critical roles in enzyme activity. In this work, we apply ML to the MSA of GH7 sequences, mapping variation in lengths of the active-site loops to functional subtypes, such that the subtype can be accurately predicted from loop length. We also derive position-specific classification rules to highlight positions that play important roles in CBH/EG function. Lastly, we investigate functional relationships between the CBM and the CD by utilizing ML to predict the presence of CBMs in GH7s using residues in the CD, revealing trans-modular sequence correlation for the first time. It is important to note that, as the current understanding of GH7 function is based on investigation of a few representatives, this present study of 1748 GH7 sequences seeks to identify general sequence–function relationships for the entirety of the GH7 family and the degree to which variation exists. Furthermore, the main objective of this study is neither to employ ML to outperform conventional bioinformatics approaches nor to predict unknown attributes, as is common among ML studies. Rather, our work is distinct in building ML models to consider the importance of sequence features in functional diversity and to subsequently delineate relationships between variation in protein sequence and function across an entire enzyme family. In other words, we apply ML to gain new biological insights into GH7 function rather than in a purely predictive capacity.

Results

Datasets

Three datasets were used in this study. The first dataset contained 1748 full-length GH7 protein sequences retrieved from the National Center for Biotechnology Information (NCBI) nonredundant database. Using a strict keyword search, we queried the NCBI database for the subtype annotation (*i.e.*, CBH or EG) of these 1748 sequences. In total, 427 sequences were clearly annotated as CBH or EG in the database (291 CBHs and 136 EGs), and these 427 sequences comprised the second dataset. For the third dataset, we retrieved 44 GH7 sequences from the manually curated UniProtKB/Swiss-Prot database (53). Accordingly, the subtype annotations of the 44 GH7s (30 CBHs, 14 EGs) are less likely to contain errors than the annotations of the 427 sequences from the NCBI nonredundant database.

Discrimination of GH7 subtypes with hidden Markov models

In the annotation of a protein sequence, several computational prediction methods may be applied. Sequence similarity methods compare an unclassified protein with well-studied proteins and assign the unclassified protein to the same class as the most similar classified proteins (54). Hidden Markov model (HMM) (55, 56), which describes the protein sequence as a probabilistic model, is one of the most sensitive and most accurate methods for discriminating protein functional families with sequence data alone, provided they are built with correct alignments (54). Within a given protein family, HMM can also be applied to discriminate functional subtypes, although the discrimination accuracy varies across different families (57).

We applied HMM to discriminate between GH7 subtypes: CBHs and EGs. The performance of HMM was evaluated by a 5-fold cross-validation technique using the datasets of 427 (NCBI) and 44 (UniProtKB/Swiss-Prot) GH7 sequences. First,

each dataset was aligned and separated into CBH and EG subalignments based on the database annotations. Then, each subalignment was randomly split into 5-folds (Fig. 2*A*). Subtype HMMs (*i.e.*, CBH HMM and EG HMM) were repeatedly built on four out of 5-folds of the CBH and EG subalignment, and the sequences in each left-out fold were used as a test set. To predict the subtype of a sequence, the sequence was aligned separately to both the CBH and EG HMMs, and then the alignment scores were compared. If the CBH HMM alignment score was greater than the EG HMM alignment score, the sequence was predicted to be a CBH; otherwise, it was predicted to be an EG (57). The process was repeated so that all 5-folds were used in training and testing the HMMs.

Figure 2, *B* and *C* show the performance of the HMM method on the UniProtKB/Swiss-Prot dataset (44 sequences) and on the NCBI dataset (427 sequences), respectively. The HMM method achieved perfect accuracy on the UniProtKB/SwissProt dataset. All sequences were correctly predicted, and there was a substantial difference, of at least 120.0, between the CBH alignment score and the EG alignment score. On the NCBI dataset of 427 sequences, which may contain erroneous subtype annotations, the HMM achieved an accuracy of 99.53% and only misclassified two sequences (accession codes: AGY80096.1 and AGY80097.1), which are annotated as EGs. These two sequences may have been erroneously annotated as EGs because they are much more similar to CBHs in overall sequence and loop lengths. Furthermore, the value of the alignment score difference for some sequences in the NCBI dataset is as low as 2.0.

Discrimination of GH7 subtypes with machine learning: relationships between active-site loops and CBH/EG function

In this part of the study, our goal was to use ML to map the variation in amino acid sequence to GH7 CBH and EG activity and to, consequently, determine which aspects of the sequence and structure predominantly affect CBH/EG function. If a particular feature is important for the difference in CBH and EG behavior, we should be able to train ML models on that feature to discriminate GH7 CBHs and EGs with significant accuracy. Otherwise, a feature that has no correlation with activity, but only varies due to phylogenetic diversity, would perform poorly when applied to predict GH7 subtypes with ML.

We used the dataset of 1748 GH7s to test ML algorithms in predicting GH7 subtypes. Since only 427 of the 1748 GH7s are classified as CBH or EG in the databases, we applied the HMM method described previously to derive the functional classes of the unclassified GH7 sequences. Our cross-validation tests showed that the HMM method can correctly classify GH7 subtypes with an accuracy of almost 100% (*i.e.*, consistent with the database annotations). This result is similar to the performance of the HMM method applied to other protein families (57). Moreover, when we trained separate HMMs on the manually annotated dataset of 44 sequences (UniProtKB/Swiss-Prot) and on the "less perfect" dataset of 427 sequences (NCBI) and then applied the HMMs to determine the subtype of the 1748 GH7s, the separate HMMs assigned the same subtype in all but five instances (99.71%). Regardless, misclassification errors of about 1% are not large enough to alter the relationships that we derived from ML on the dataset of 1748 GH7 sequences (58, 59).

In choosing features for the ML models, we capitalized on the observation that crystal structures of GH7 CBHs and EGs differ in their active-site architecture, due to the degree of truncation in the eight active-site loops (Fig. 1). Hence, we used the number of residues in the active-site loops as features for ML to discriminate between GH7 CBHs and EGs. First, a structure-based MSA of all 1748 sequences was carried out (See Materials and Methods for details). For each sequence in the MSA, we counted the number of amino acid residues in the eight active-site loops and derived a vector of the eight loop lengths as features (Fig. 3).

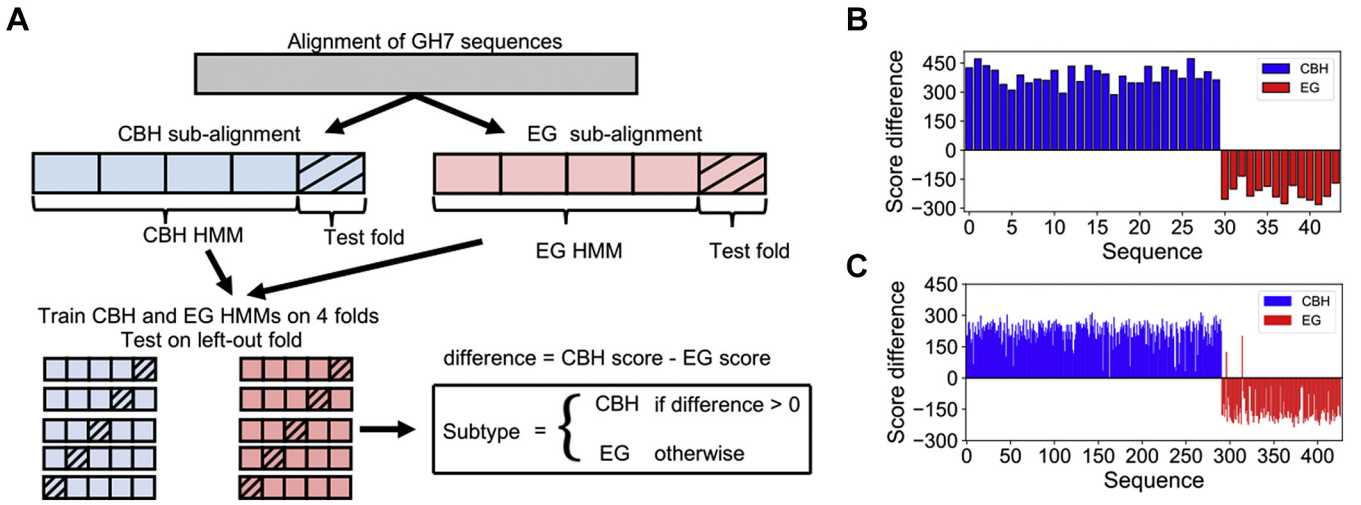
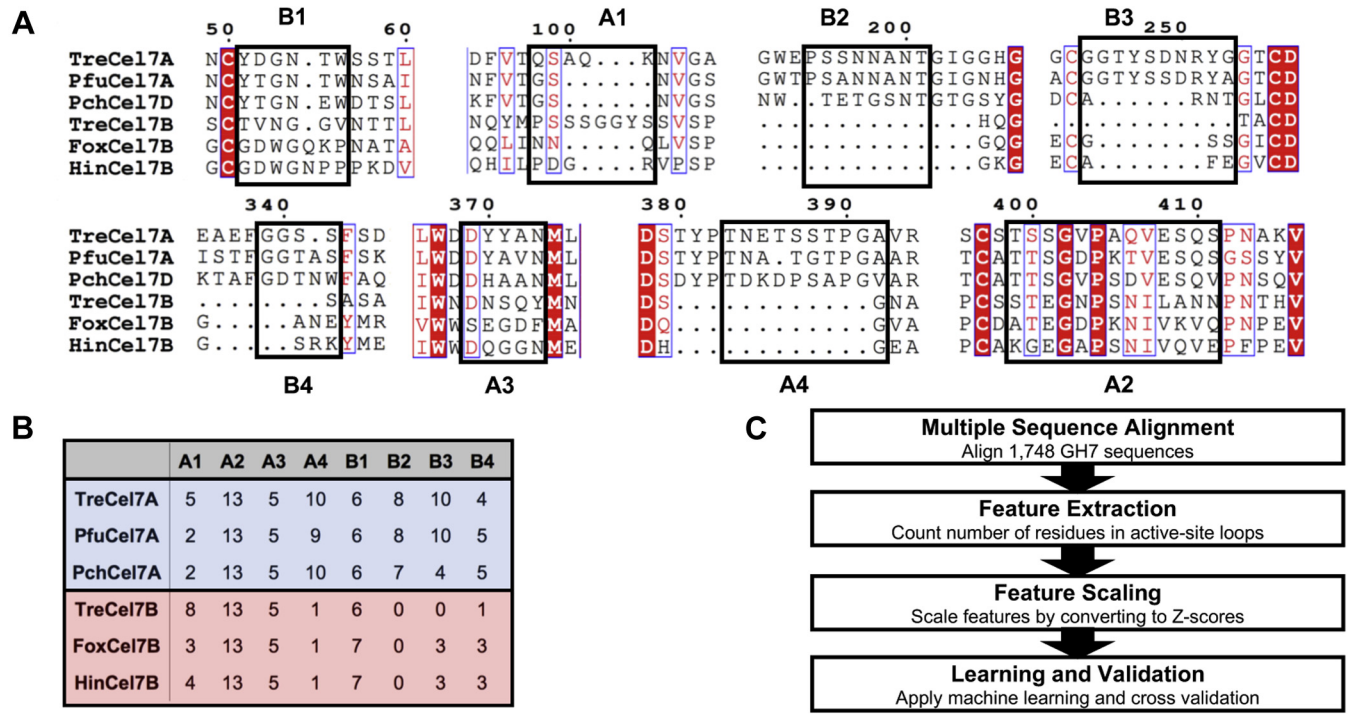


Figure 2. Discrimination of GH7 CBHs and EGs with hidden Markov models (HMM). *A*, 5-fold cross-validation technique for evaluating the performance of HMM. The MSA is split into CBH and EG subalignments and each subalignment into 5-folds. HMMs are repeatedly trained on 4-folds and then tested on the left-out fold. The predicted class (CBH or EG) of a sequence is the class that yields the highest HMM alignment score. *B*, performance of HMM on the dataset of 44 GH7s from the manually curated UniProtKB/SwissProt database. *C*, performance of HMM on the dataset of 427 GH7s from NCBI nonredundant database. Only two EG sequences (GenBank accession codes: AGY80096.1 and AGY80097.1) were misclassified in the NCBI dataset. Note that in *B* and *C*, the assigned sequence numbers (*x*-axes) are arbitrary.



	A1	A2	A3	A4	B1	B2	B3	B4
TreCel7A	5	13	5	10	6	8	10	4
PfuCel7A	2	13	5	9	6	8	10	5
PchCel7A	2	13	5	10	6	7	4	5
TreCel7B	8	13	5	1	6	0	0	1
FoxCel7B	3	13	5	1	7	0	3	3
HinCel7B	4	13	5	1	7	0	3	3

Figure 3. Generating features for discriminating GH7 CBHs and EGs with ML. *A*, segments of a selection of six well-studied GH7s from the structure-based sequence alignment of 1748 sequences showing the active-site loops. The sequences include the CBHs: *Trichoderma reesei* Cel7A (*Tre*Cel7A) (23), *Penicillium funiculosum* Cel7A (*Pfu*Cel7A) (20), and *Phanerochaete chrysosporium* Cel7D (*Pch*Cel7D) (19); and the EGs: *Trichoderma reesei* Cel7B (*Tre*Cel7B) (26), *Fusarium oxysporum* Cel7B (*Fox*Cel7B) (24), and *Humicola insolens* Cel7B (*Hin*Cel7B) (25). *B*, the number of residues in the eight active-site loops as determined from the structure-based alignment. *C*, procedure for generating features for 1748 GH7s. First, the sequences are aligned as in (*A*). Then, a count of the number of residues in each loop is obtained. Residue counts are scaled to Z-scores before ML is applied.

Four ML methods were applied: decision trees, logistic regression, k-nearest neighbors (KNN), and support vector machines (SVM). For each ML method, nine models with different combinations of features were tested. One model involved training the ML algorithms on the lengths of all eight loops, and the remaining eight models involved using each loop length as the sole feature for the training (single-feature models). The performance of the ML models was measured using four metrics: sensitivity (or true positive rate), specificity (or true negative rate), overall accuracy, and Matthew's correlation coefficient (MCC). Here, the sensitivity is the percent of CBHs (the true class) correctly predicted, the specificity is the percent of EGs (the false class) correctly predicted, and the overall accuracy is the percent of both CBHs and EGs correctly predicted. MCC ranges from −1 to +1 and measures the correlation between the predicted and true classifications. An MCC value of +1 indicates perfect prediction, 0 indicates no concordance between predicted and actual classes, and −1 indicates perfect disagreement. MCC has been recommended as the most informative performance metric in evaluating binary classification performance, especially when the dataset is imbalanced since other metrics such as overall accuracy and F1 score can be hugely misleading (60–63). Hence, we use MCC as the primary metric in evaluating the performance of the ML models.

Moreover, we are faced with the problem of an imbalanced dataset: 1306 (75%) of the 1748 sequences in the dataset are CBHs. Ordinarily, imbalanced data will skew the results by causing the ML classifiers to place most of the data in the majority class (CBH). To deal with the imbalance problem, we applied random undersampling (64, 65) to the majority class so that the distribution of CBH and EGs was balanced. We evaluated the performance of the ML models on the redistributed data with 100 repetitions of 5-fold cross-validation, with the dataset undersampled and reshuffled in each repetition (Fig. 4). Repeating the 5-fold cross-validation numerous times is a highly effective way to mitigate the effects of variability in the train-test splits and to ensure that the data space is thoroughly explored despite loss of data in the undersampling step (66).

Our results show that ML is able to accurately discriminate between GH7 CBHs and EGs using only information about the length of the active-site loops (Table 1). However, the performance varied significantly for the different single-feature models (Fig. 5*A*). The models trained on the A2 and A3 loops exhibited the worst performance with MCC values close to 0, indicating that they did not perform better than a random classification. The models trained on A1 and B1 loops showed intermediate performance with MCC values widely varying from −0.08 to 0.79 for the A1 models and from −0.03 to 0.63 for the B1 models. Interestingly, the A4, B2, B3, and B4 models showed very high predictive performance, with MCC values ranging from 0.94 to 0.98 and with much lower variation among the different ML methods. The models trained on these four loops (A4, B2, B3, B4) achieved nearly the same high performance as the models trained on all eight loops (Table 1).

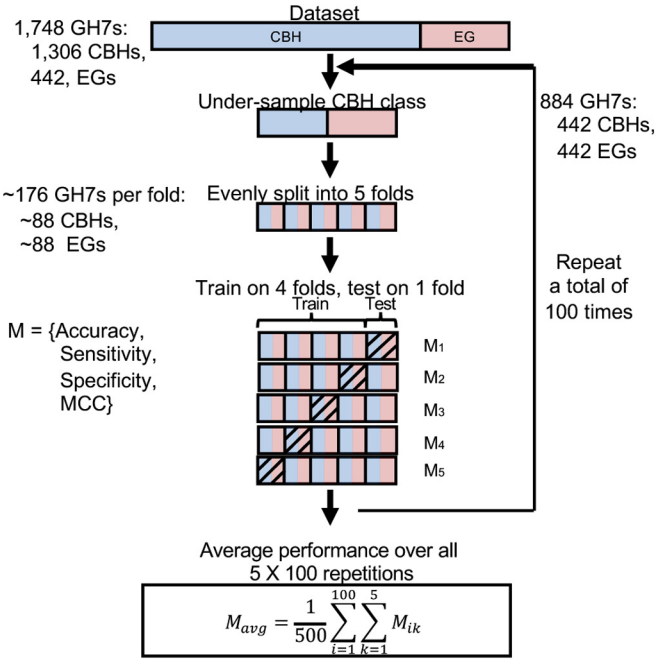


Figure 4. Procedure for evaluating the performance of ML models using 100 repetitions of 5-fold cross-validation with undersampling. The dataset is reshuffled and resampled in each repetition.

Furthermore, we observed that the variation in the lengths of the loops correlates with the discriminative performance of the loops (Fig. 5, *A*–*C*). The loops with very poor discriminatory performance (A2 and A3) show the lowest relative variation in lengths across the 1748 GH7s and nearly identical distributions between CBHs and EGs (Figs. 5*B* and S1). In contrast, loops with intermediate discriminatory performance (A1 and B1) show a greater level of variation in lengths than A2 and A3 loops and noticeably different distributions for CBHs and EGs, although there is a considerable amount of overlap. The loops with near-perfect predictive performance (A4, B2, B3, B4) show the highest variation in lengths.

One major advantage of the tree-based methods over other ML algorithms is the possibility of deriving and visualizing interpretable classification rules (67, 68). In many applications of ML to biological problems, it is desirable to gain knowledge of biological relationships rather than merely applying ML as a predictive tool. Figure 5*D* shows rules derived from the single-node decision-tree classifiers trained on the A4, B2, B3, and B4 loops. A classification accuracy of 96.9% was achieved by the simple rule: if a GH7 has more than four residues in the B2 loop, then it is a CBH, else it is an EG. Overall, the decision trees reveal that GH7 EGs tend to possess three or less residues in the B3 and B4 loops, four or less residues in the B2 loop, and five or less residues in the A4 loop.

Since the lengths of the A4, B2, B3, and B4 loops can independently discriminate between GH7 CBHs and EGs with accuracies greater than 94%, it is expected that there is a substantial degree of correlation between them. We conducted correlation analysis by computing the Pearson's correlation coefficient between the lengths of the eight loops of 1748 GH7s (Fig. 6). As expected, there is significant positive

Table 1
Performance of ML algorithms in discriminating GH7 CBHs and EGs[a]

Features	Decision tree			Logistic regression			K-nearest neighbor			Support vector machine		
	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.
A1	98.6 ± 1.2	45.9 ± 5.0	72.3 ± 3.2	42.0 ± 16.5	52.8 ± 7.3	46.9 ± 6.4	86.5 ± 15.1	88.7 ± 5.4	87.6 ± 5.6	97.0 ± 1.8	85.5 ± 3.4	91.2 ± 2.0
A2	65.9 ± 43.5	37.4 ± 42.2	50.7 ± 3.7	49.3 ± 46.7	50.3 ± 45.3	47.4 ± 2.8	4.6 ± 2.3	97.0 ± 1.9	50.8 ± 3.8	89.2 ± 27.3	18.4 ± 26.0	53.0 ± 3.5
A3	89.0 ± 26.3	16.9 ± 24.8	52.5 ± 3.7	50.8 ± 47.9	49.4 ± 45.4	47.6 ± 3.2	3.0 ± 2.0	97.8 ± 1.6	50.4 ± 3.7	96.7 ± 11.2	11.4 ± 10.9	53.9 ± 3.4
A4	95.7 ± 2.1	99.5 ± 0.7	97.6 ± 1.1	95.8 ± 2.0	99.5 ± 0.6	97.7 ± 1.1	95.8 ± 2.1	99.7 ± 0.5	97.8 ± 1.1	95.6 ± 2.2	99.6 ± 0.6	97.6 ± 1.1
B1	96.8 ± 1.8	44.1 ± 5.5	70.5 ± 3.3	79.3 ± 35.6	34.5 ± 12.2	55.9 ± 13.4	1.3 ± 1.8	98.7 ± 1.6	50.0 ± 3.7	95.1 ± 2.6	72.3 ± 4.4	83.7 ± 2.6
B2	94.6 ± 2.4	99.1 ± 1.2	96.9 ± 1.3	94.7 ± 2.4	98.4 ± 1.2	96.6 ± 1.3	95.3 ± 2.3	97.4 ± 1.8	96.4 ± 1.3	94.8 ± 2.4	98.4 ± 1.6	96.6 ± 1.4
B3	92.4 ± 2.7	99.8 ± 0.5	96.1 ± 1.4	89.9 ± 3.3	99.8 ± 0.5	94.8 ± 1.7	96.3 ± 2.2	98.6 ± 1.1	97.5 ± 1.2	89.7 ± 3.3	99.8 ± 0.4	94.8 ± 1.7
B4	97.9 ± 1.8	98.2 ± 1.3	98.0 ± 1.0	98.2 ± 1.4	98.2 ± 1.2	98.2 ± 0.9	97.6 ± 2.0	98.3 ± 1.3	98.0 ± 1.1	97.8 ± 1.6	98.2 ± 1.3	98.0 ± 1.0
All 8 loops	98.8 ± 1.2	99.1 ± 1.1	98.9 ± 0.8	98.3 ± 1.4	99.2 ± 0.9	98.8 ± 0.8	97.1 ± 2.0	99.4 ± 0.7	98.2 ± 1.1	99.0 ± 1.1	98.9 ± 1.1	98.9 ± 0.7

[a] Each ML model was trained separately with each of the eight loops as a single, independent feature, and then with all eight loops combined (last row). The performance was evaluated by measuring the sensitivity (sens.), specificity (spec.) and accuracy (acc.) in percent. Error represents one standard deviation from the mean.

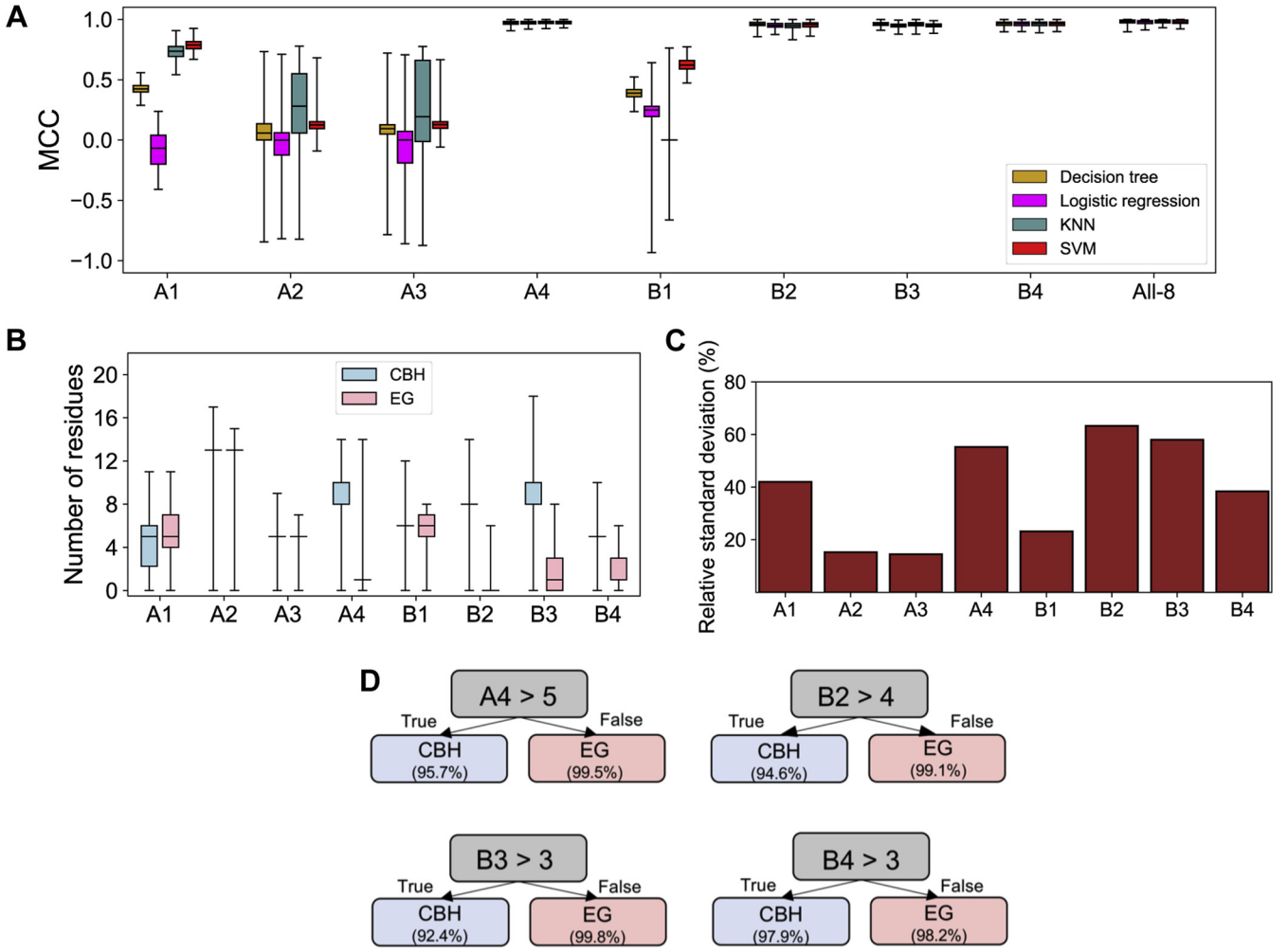


Figure 5. Predictive performance and variation of active-site loops in GH7s. *A*, Matthews' correlation coefficient (MCC) values of four ML algorithms trained separately on the length of each active-site loop and on all eight loops together. The A4, B2, B3, and B4 loops achieve near-perfect performance in discriminating 1748 GH7 CBHs and EGs. Box and whisker plots indicate distribution of MCC values over 100 repetitions of 5-fold cross-validation (*center line*: median, box limits: upper/lower quartiles, *whiskers*: full data range). *B*, distribution of the lengths of active-site loops in 1306 GH7 CBHs and 442 GH7 EGs. Box and whisker plots are as in (*A*). *C*, the relative standard deviation of the length of the eight active-site loops. Generally, variation in the length of a loop correlates with predictive performance of the loop as a ML feature. *D*, rules derived from the single-node decision trees trained on the lengths of the A4, B2, B3, and B4 loops. The accuracy of the rules in discriminating GH7 CBHs and EGs, *i.e.*, the sensitivity and specificity, respectively, are shown in *brackets*.

correlation between the lengths of the A4, B2, B3, and B4 loops ($r \geq +0.76$, $p < 0.0001$). The highest correlations are observed between the A4 and B2 loops (+0.84) and between the A4 and B4 loops (+0.83).

Discrimination of GH7 subtypes with position-specific classification rules: important residues for CBH/EG function

In discriminating GH7 CBHs and EGs with ML, we have used only the lengths of the active-site loops as features without considering the contributions of specific amino acids in the proteins. However, the interactions of specific residues are known to affect GH7 CBH/EG function, and mutagenesis studies have confirmed that certain positions play essential roles in GH7 activity (6, 37, 38, 69). In this section, we investigate the relationships between specific residues in the proteins and the functional subtype.

It is common knowledge that although a protein's function arises from the combined effects of multilevel interactions between all residues in the protein, some residues contribute to function more significantly than others. Consequently, it is likely that in GH7s, if a position is considerably conserved in CBHs such that CBHs tend to utilize a particular amino acid at that position, and EGs tend to not utilize the same amino acid at that position, or vice versa, then that position plays a vital role in the difference in CBH/EG function or structural stability. A typical example is position 40 (*i.e.*, Trp40 in *Tre*-Cel7A). From analysis of the structure-based MSA, we observe that this position is strongly conserved in CBHs with 92.5% exhibiting a Trp at this position, whereas it is notably variable in EGs with only 28.5% exhibiting a Trp at this position (Fig. 7*A*). Considering only this clear difference in the amino acid distribution at position 40, we can infer that Trp40 likely contributes to CBH function. Mutation of Trp40 to Ala has, in fact, been shown to considerably decrease the activity of *Tre*-Cel7A on crystalline cellulose but not on amorphous cellulose (37), indicating that Trp40 is critical for processivity (38). Consequently, we propose that applying a statistical method to mine for positions in GH7s that are conserved but have

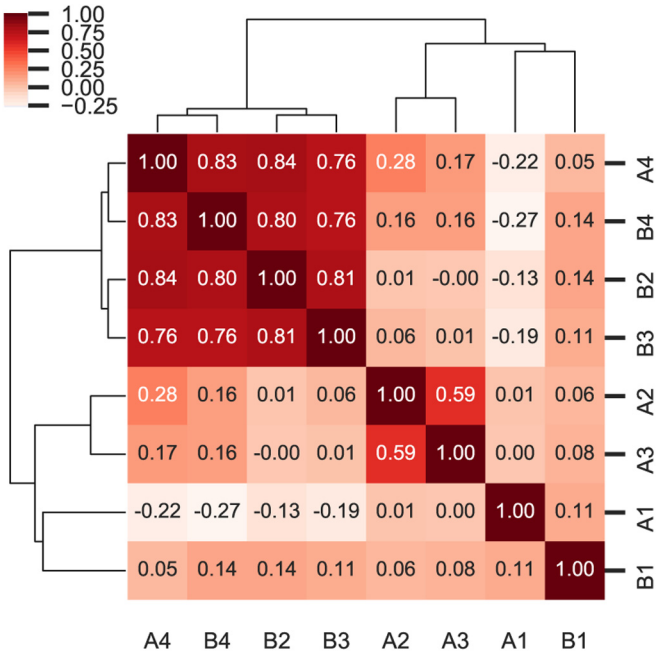


Figure 6. Pearson's correlation coefficient between the lengths of the eight active-site loops in 1748 GH7s. The matrix of correlation coefficients is clustered so that loops with a similar pattern of correlation are grouped together. There is a high degree of positive correlation (*darker red*) between the lengths of the A4, B2, B3, and B4 loops.

remarkably different amino acid distributions between CBHs and EGs can identify positions that play critical roles in CBH/EG function and processivity.

From the amino acid distribution at position 40, we obtain a single-node decision tree with the rule: Trp at position 40 implies CBH, else EG. This simple rule classifies 1748 GH7 CBHs and EGs with an accuracy of 87.2%. Thus, a rational strategy for identifying positions likely associated with CBH/EG function is to derive similar rules for all positions in the MSA and select positions that yield high-performing rules. First, we split the MSA of 1748 GH7 sequences into CBH and EG subalignments and then identified the consensus amino acid and the consensus amino acid type (*i.e.*, aliphatic, aromatic, polar, positive, or negative) for each position in the subalignments. For each position, if X and Z are the consensus amino acids (or type) in the CBH and EG subalignment, respectively, we derived the following classification rules: X=>CBH and Z=>EG, X=>CBH and not X=>EG, and not Z=>CBH and Z=>EG. Applying this strategy to 434 positions in the MSA (*Tre*Cel7A numbering), we derived 1799 classification rules. For each rule, we measured the classification accuracy, sensitivity, specificity, and MCC and tested the statistical significance by conducting Chi-square test of independence. The 1799 rules have fairly normally distributed MCC scores (Fig. 7, *B*), and the top 5% of rules (90 rules) have MCC scores of at least +0.73 and classification accuracies of at least 87% (Tables 2 and S1, Figs. 7 and S2). These 90 rules are derived from 42 positions, which are generally in close proximity to the cellodextrin ligand in the crystal structure. More than half of the top 90 rules are from positions within 5 Å of the cellononaose ligand bound in *Tre*Cel7A structure (PDB code: 4C4C). Moreover, most of the positions are closer to the tunnel entrance where cellulose chains are recruited by the enzyme for processive hydrolysis (Fig. 7, *D* and *E*).

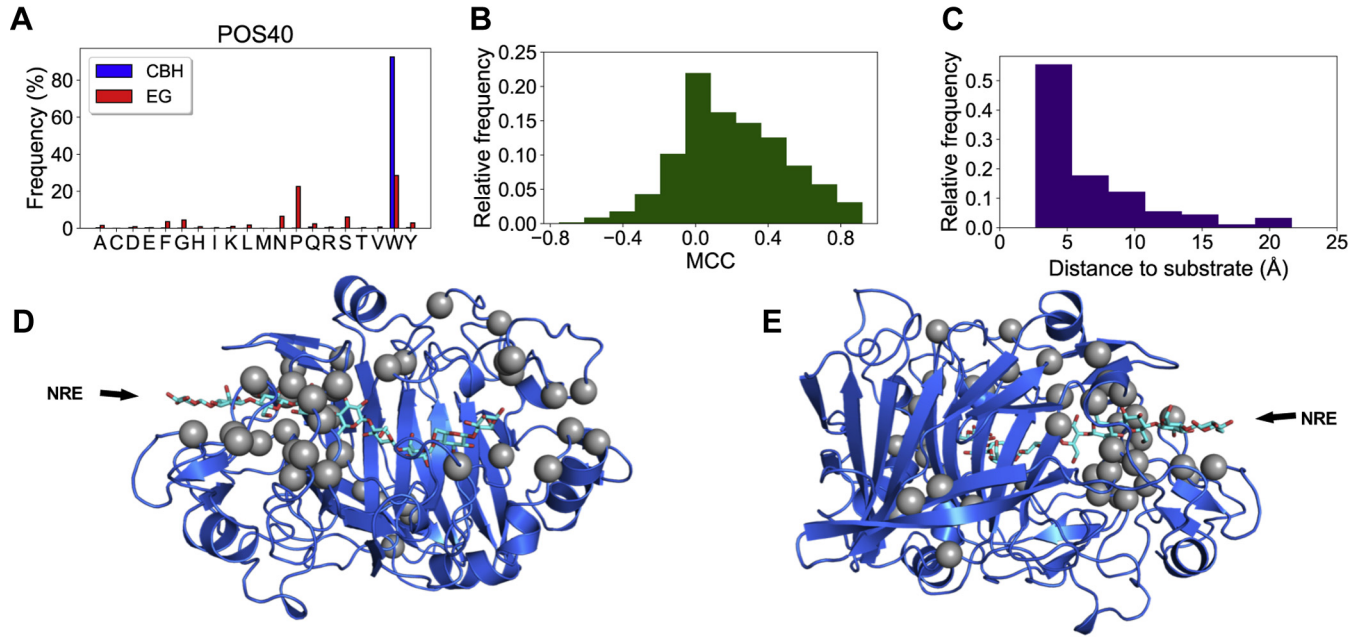


Figure 7. Top-performing position-specific classification rules for discriminating GH7 CBHs and EGs. *A*, amino acid distribution of GH7 CBHs and EGs at position 40 (*Tre*Cel7A numbering). Position 40 is strongly conserved as Trp in GH7 CBHs but not in EGs. *B*, MCC scores of 1799 position-specific classification rules derived from the MSA. The top 90 rules have MCC scores of 0.73 or greater. *C*, Histogram of minimum distance between the cellononaose ligand in *Tre*Cel7A (PDB code: 4C4C) (23) and positions from which the top 90 classification rules are derived. More than half of top 90 rules are derived from positions within 5 Å of the substrate. *D*, alpha carbons of 42 positions from which the top 90 classification rules are derived shown on the structure of *Tre*Cel7A. Most of these positions are near the substrate sites toward to the nonreducing end (NRE). *E*, posterior view of crystal structure.

Table 2
Top-performing position-specific classification rules relating amino acid residues and GH7 subtype (CBH/EG)[a]

*Tre*Cel7A position	Rule	Closest subsite	Distance to closest subsite (Å)	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	Ref.
16	not Thr=>CBH, Thr=>EG	−2	19.0	97.7	82.6	93.9	0.83	
37	Asn=>CBH, not Asn=>EG	−4	3.1	92.8	86.0	91.1	0.77	(79, 82)
38	Trp=>CBH, not Trp=>EG	−4	3.2	93.2	96.2	93.9	0.85	(35, 79, 82)
39	Arg=>CBH, not Arg=>EG	−5	3.6	96.4	79.9	92.2	0.79	(82)
39	Arg=>CBH, His=>EG	−5	3.6	98.2	70.1	91.1	0.76	(82)
49	Asn=>CBH, not Asn=>EG	−7	2.7	90.9	85.3	89.5	0.73	(79, 83)
51	Tyr=>CBH, not Tyr=>EG	−5	3.6	88.4	99.1	91.1	0.80	(82)
53	Gly=>CBH, not Gly=>EG	−5	4.9	90.6	90.7	90.6	0.77	
56	Trp=>CBH, not Trp=>EG	−5	8.9	93.2	99.3	94.7	0.88	
81	Thr=>CBH, not Thr=>EG	−5	4.1	88.1	91.2	88.9	0.74	
82	Tyr=>CBH, not Tyr=>EG	−5	3.8	91.6	86.2	90.2	0.75	(82)
95	Phe=>CBH, not Phe=>EG	−4	7.0	84.0	97.5	87.4	0.74	
97	Thr=>CBH, not Thr=>EG	−5	6.7	89.2	93.4	90.3	0.77	
103	Asn=>CBH, not Asn=>EG	−5	2.7	92.1	87.8	91.0	0.77	(79, 82, 83)
105	Gly=>CBH, not Gly=>EG	−4	4.8	94.8	86.0	92.6	0.80	
105	not Ser=>CBH, Ser=>EG	−4	4.8	99.7	74.9	93.4	0.82	
105	Gly=>CBH, Ser=>EG	−4	4.8	97.2	80.4	93.0	0.81	
106	Ser=>CBH, not Ser=>EG	−2	4.8	89.9	88.7	89.6	0.75	
106	not Pro=>CBH, Pro=>EG	−2	4.8	99.2	86.9	96.1	0.89	
106	Ser=>CBH, Pro=>EG	−2	4.8	94.5	87.8	92.8	0.81	
120	Phe=>CBH, not Phe=>EG	−1	15.7	93.0	83.3	90.6	0.75	
140	Leu=>CBH, not Leu=>EG	−1	8.5	83.2	98.2	87.0	0.73	
146	Phe=>CBH, not Phe=>EG	−1	7.9	91.8	93.9	92.3	0.81	
146	not Leu=>CBH, Leu=>EG	−1	7.9	94.3	79.4	90.6	0.75	
146	Phe=>CBH, Leu=>EG	−1	7.9	93.1	86.7	91.4	0.78	
179	Asp=>CBH, not Asp=>EG	−3	2.6	92.9	99.1	94.5	0.87	(82)
181	Lys=>CBH, not Lys=>EG	−5	2.8	92.0	99.3	93.8	0.86	(79, 82, 83)
192	Trp=>CBH, not Trp=>EG	−4	7.0	93.8	100.0	95.4	0.89	
200	Asn=>CBH, not Asn=>EG	−4	3.5	85.5	99.3	89.0	0.77	(79, 80, 82)
202	Gly=>CBH, not Gly=>EG	−4	6.5	94.0	100.0	95.5	0.89	
204	Gly=>CBH, not Gly=>EG	−4	10.7	95.0	99.5	96.2	0.91	
251	Arg=>CBH, not Arg=>EG	2	3.4	86.9	99.8	90.2	0.79	(6, 35, 36, 82)
262	Asp=>CBH, not Asp=>EG	2	4.1	95.1	97.3	95.7	0.89	(8, 36, 82)
262	not Gly=>CBH, Gly=>EG	2	4.1	98.7	69.0	91.2	0.76	(8, 36, 82)
262	Asp=>CBH, Gly=>EG	2	4.1	96.9	83.1	93.4	0.82	(8, 36, 82)
338	Phe=>CBH, not Phe=>EG	2	7.7	91.8	99.8	93.8	0.86	
340	Asp=>CBH, not Asp=>EG	2	9.1	83.1	99.3	87.2	0.74	
381	Tyr=>CBH, not Tyr=>EG	2	3.5	83.7	99.8	87.8	0.75	(36, 82)
382	Pro=>CBH, not Pro=>EG	2	5.0	93.3	98.4	94.6	0.87	
391	Gly=>CBH, not Gly=>EG	2	6.9	94.6	91.0	93.7	0.84	
394	Arg=>CBH, not Arg=>EG	2	3.1	95.1	96.4	95.4	0.89	(34, 35, 82)
394	Arg=>CBH, Ala=>EG	2	3.1	97.4	72.6	91.1	0.76	(34, 35, 82)
396	not Pro=>CBH, Pro=>EG	2	12.9	85.5	96.6	88.3	0.75	
401	not Glu=>CBH, Glu=>EG	−3	13.5	98.8	72.9	92.2	0.79	
423	not Trp=>CBH, Trp=>EG	−1	18.1	97.4	72.6	91.1	0.76	

[a] All rules discriminate GH7 CBHs and EGs with accuracies of at least 87.0% and MCC scores of at least 0.73. Nearest distance to the nearest glycosyl residues was measured from the *Tre*Cel7A structure (PDB code: 4C4C). Statistical significance was tested by a Chi-square test of independence. All rules are significant at $p < 0.0001$. See Table S1 for rules between amino acid type and GH7 subtype. Positions from which the rules have been derived are shown on the crystal structure of *Tre*Cel7A in Figure 7.

Conserved aromatic residues in the active site of GH7s

GH7s possess several aromatic residues lining the active-site tunnel, which have been suggested to play key roles in cellulolytic bond cleavage and processive action (35). We have conducted bioinformatic analysis of conserved aromatic residues in the active site of GH7s. From the MSA of 1748 GH7s, we selected positions that are located within 6 Å of the cellononaose substrate in the structure of *Tre*Cel7A (PDB code: 4C4C), and that have aromatic residues (Phe, Trp, Tyr, or His) at that position in the consensus sequence of CBHs or EGs (Fig. S3). There are 17 of such aromatic positions in the MSA, and on the protein structure, these positions are distributed across the nine glycosyl subsites.

Furthermore, these 17 positions can be classified into three groups based on the conservation of aromatic amino acids (Table 3). The first group consists of positions that are conserved in both CBHs and EGs such that more than two-thirds of CBHs and EGs utilize aromatic residues at these positions. Positions 145, 171, 216, 228, 367, and 376 (*Tre*Cel7A numbering) fall in the first group. The second group consists of positions that are conserved as aromatic residues (>66%) in CBHs but not in EGs. Positions 38, 40, 51, 82, 252, 370, and 381 fall in the second group. The third group contains positions that are neither conserved (<66%) as aromatic residues in CBHs and EGs although the consensus amino acids are aromatic. Positions 39, 47, 53, and 247 fall in the third group.

When these positions are viewed on the crystal structure (Table 3, Fig. 8), an interesting pattern is observed. Whereas positions that are strongly conserved in both CBHs and EGs (first group) are located near the catalytic center of the active site, positions which are conserved in CBHs but not in EGs flank the catalytic center nearer to the "substrate-binding" sites (−7 to −1) or the "product-binding" sites (+1 to +2).

Predicting the presence of CBMs with machine learning: relationships between the CD and the CBM

The CD of GH7 proteins may be attached to a second domain (the CBM) *via* a flexible linker. The CBM function is

Table 3
Positions within 6 Å of the cellononaose ligand in *Tre*Cel7A (PDB code: 4C4C) with aromatic residues in CBH/EG consensus sequences[a]

*Tre*Cel7A position	*Tre*Cel7A residue	CBH consensus residue	EG consensus residue	Frequency of aromatic residues in CBHs (%)	Frequency of aromatic residues in EGs (%)	Closest subsite	Distance to closest subsite (Å)	Aromatic residues conserved (>66%) in	Ref.
47	S	Y	-	46.8	13.8	−7	3.9	None	
40	W	W	W	93.3	36.0	−6	3.4	CBH	(35, 79, 82, 83)
39	R	R	H	0.0	60.4	−5	3.6	None	(82)
53	G	G	W	0.3	29.4	−5	4.9	None	
51	Y	Y	G	92.6	2.0	−5	3.6	CBH	(82)
82	Y	Y	Y	94.9	31.7	−5	3.8	CBH	(82)
38	W	W	A	94.2	29.4	−4	3.2	CBH	(82)
370	Y	H	E	87.3	1.8	−3	5.3	CBH	
247	Y	Y	-	46.9	0.0	−2	2.7	None	(82)
145	Y	Y	Y	97.9	98.9	−2	2.7	CBH and EG	(82)
367	W	W	W	94.3	98.6	−1	3.0	CBH and EG	(82)
171	Y	Y	Y	98.0	97.3	−1	3.8	CBH and EG	(82)
216	W	W	W	97.9	68.1	−1	5.6	CBH and EG	
228	H	H	H	96.6	97.5	1	2.8	CBH and EG	(82)
376	W	W	W	96.8	99.1	2	3.5	CBH and EG	(35, 36, 82)
252	Y	Y	-	85.0	17.2	2	5.8	CBH	
381	Y	Y	-	92.8	0.2	2	3.5	CBH	(36, 82)

[a] The positions are listed in order of proximity to the glycosyl subsites. Aromatic positions conserved in both CBHs and EGs are near the catalytic center, whereas aromatic positions conserved in only CBHs flank the catalytic center. All conserved positions are shown on the crystal structure of *Tre*Cel7A in Figure 8.

mostly attributed to enhancing the binding of the enzyme to the cellulose substrate and thus, facilitating turnover by increasing enzyme concentration on the cellulose surface (2).

We studied the distribution of family 1 CBMs in our dataset of 1748 GH7s. First, a database of the 1748 sequences was created, and then a BLAST search of *Tre*Cel7A CBM was performed against the database. From a careful manual inspection of the BLAST alignment output, we selected an alignment score of 30 as the threshold so that GH7 sequences that yielded BLAST alignment scores of 30 or greater were determined to possess a family 1 CBM. We compared the distribution of CBMs among GH7 CBHs and EGs in our dataset and determined that 27% of GH7s contain a CBM, with 31% and 15% of GH7 CBHs and EGs exhibiting CBMs, respectively (Table 4). Thus, GH7 CBHs appear to be roughly two times more likely than EGs to contain a CBM. Moreover, a Chi-square test of independence indicated that the relationship between CBM utilization and GH7 subtype (CBH/EG) is significant ($p < 0.001$).

To investigate the relationships between the CD and the CBM, we applied ML to predict the presence of CBMs using the specific amino acid residues in the CD as features. Positions flanking the CD in the MSA were removed, and one-hot encoding was applied to transform the amino acids in the MSA to binary variables (70). Therefore, the MSA was transformed to a matrix such that the rows indicate the sequences and columns denote the amino acid at positions in the MSA (features). Columns are labeled as "residue-position" and can take values of 0 or 1. For example, a value of 1 at columns Q1 and S2 for *Tre*Cel7A indicates that Gln and Ser are present at positions 1 and 2 in the MSA, respectively (Fig. 9, *B*). Subsequently, one-hot encoding resulted in a high-dimensional matrix with 1748 rows and 5933 columns. We implemented the random forest algorithm (71) with 500 trees to predict the presence of a CBM using the 5933 one-hot encoded features. The random forest algorithm is especially suitable for this classification problem because it is capable of robustly dealing with high-dimensional data by performing implicit feature selection in the learning process (72), is more tolerant to noise and overfitting (71, 73), and can be used to evaluate the relative importance of the features (74).

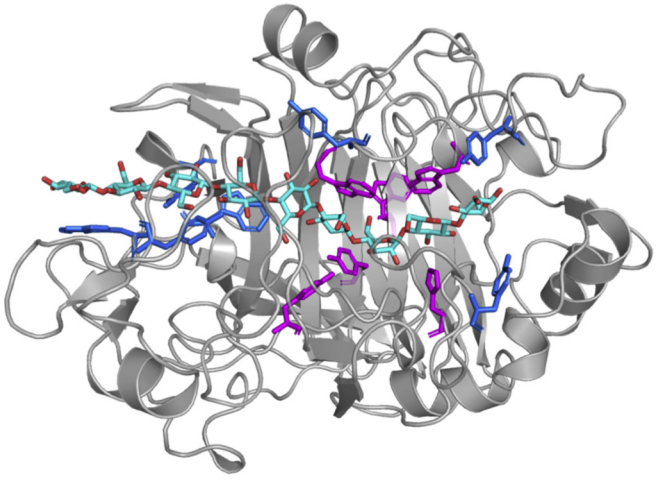


Figure 8. Conserved aromatic residues in the active site of *Tre*Cel7A (PDB code: 4C4C) within 6 Å of the cellononaose ligand. Residues in magenta are conserved (>66% frequency) in both GH7 CBHs and EGs and are found close to the catalytic center between −1 and +1 glycosyl subsites. Residues in blue are conserved in GH7 CBHs but not in EGs and flank the catalytic center.

Table 4
Distribution of CBMs in GH7s showing the relationship between subtype (CBH/EG) and the presence of a CBM[a]

	CBH	EG	Total
Has CBM	407	66	473
No CBM	899	376	1275
Total	1306	442	1748
CBM frequency (%)	31.2	14.9	27.1

[a] GH7 CBHs are roughly two times more likely to possess a CBM than GH7 EGs ($p < 0.0001$, Chi-square test).

ASBMB

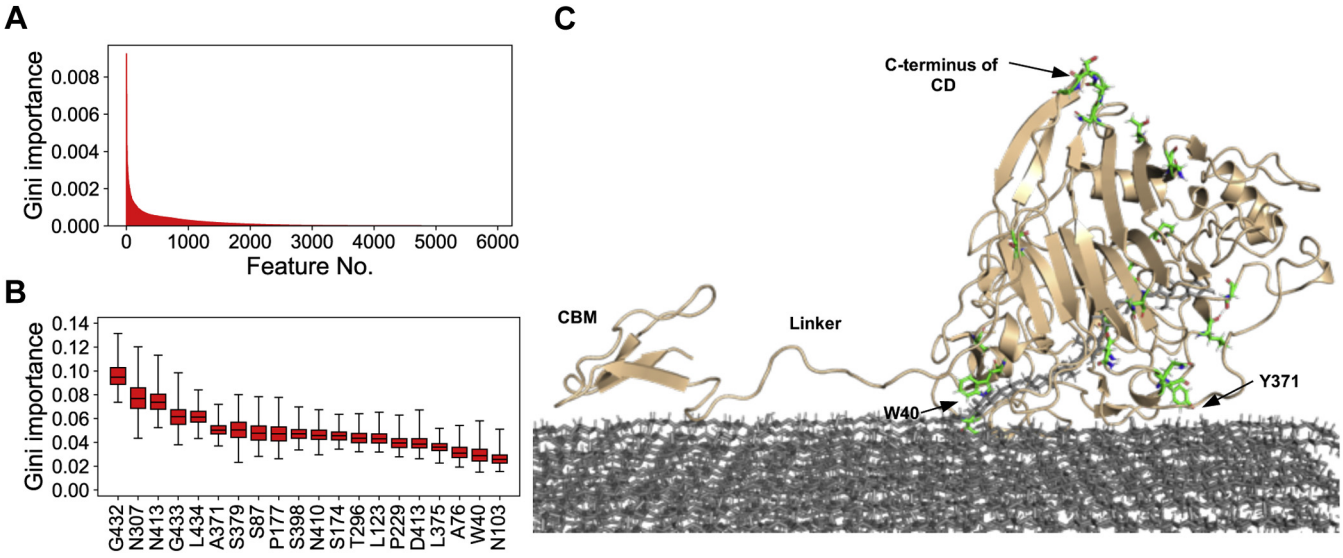


Figure 9. Top-performing features of the random forest classifier in predicting the presence of CBMs in GH7s. *A*, relative importance (Gini) of all 5933 features derived from one-hot encoding of the MSA. Most features provide little information to the model. *B*, relative importance (Gini) of top 20 features in the random forest classifier retrained on only top 20 features. Box and whisker plots indicate the distribution over 100 repetitions of 5-fold cross-validation (*center line*: median, *box limits*: upper/lower quartiles, *whiskers*: full data range). *C*, residues of top 20 features (*green sticks*) shown on the structure of *Tre*Cel7A (*tan cartoon*) on cellulose (*gray sticks*). The structure is derived from a snapshot (t = 0.73 μs) of MD simulations conducted in a previous work (97).

The performance of the random forest classifier was evaluated with 100 repetitions of 5-fold cross-validation with random undersampling, as described previously (Fig. 4). Only 90% of the dataset was used for the cross-validation; 10% of the dataset (174 sequences) was randomly selected and set aside for a separate final test. The random selection of the test dataset was implemented in such a way that a similar distribution (27% CBM, 73% no CBM) was maintained. In the validation routine, an accuracy of 90.8% was achieved by the 500-trees random forest trained on all 5933 features (Table 5). A plot of the relative (Gini) importances (74) of the features shows that most of the 5933 features contribute little or no information to the performance of the random forest classifier (Fig. 9, *A*). We reapplied the random forest algorithm using only the top 50 and the top 20 features with the highest Gini importances. The classifiers trained on only the top 20 and top 50 features showed fairly similar validation performance to the classifier trained on all 5933 features (Table 5). Some residues at the C-terminus of the CD (where the CD connects with the CBM-linker domain) were identified to be among the most important positions in predicting the presence of CBMs (Fig. 9, *B*, Table S2).

Table 6
Distribution of CBMs in GH7s showing the relationship between the presence of the rare disulfide bond (C4-C72 in *Tre*Cel7A) and the presence of a CBM[a]

	Has disulfide bond	Lacks disulfide bond	Total
Has CBM	105	368	473
No CBM	54	1221	1275
Total	159	1589	1748
CBM frequency (%)	66.0	23.2	27.1

[a] GH7s possessing this disulfide bond (mostly CBHs) are roughly three times more likely to possess a CBM than GH7s lacking the disulfide bond ($p < 0.0001$, Chi-square test).

To confirm that the random forest algorithm was not predicting the presence of a CBM mainly by looking at these interdomain connecting residues (S431, G432, S433, G433, T433, L434), we repeated the validation procedure with the top 50 features but excluded features derived from positions near the C-terminus (six features removed, 44 features remaining). The results show that the performance of the new classifier trained on 44 features was only slightly lower, with the accuracy dropping by less than 3%. Moreover, on the separate test set, the classifier trained on the top 20 features achieved an accuracy of 89.7%, confirming that the presence of a CBM can be predicted from a few residues in the catalytic domain with considerable accuracy. In addition, we derived position-specific classification rules with each of the top 50 features, as described previously (*i.e.*, X=>CBM, else, no CBM). As expected, all 50 rules independently performed worse, compared with the random forest classifier trained on all the 50 features (MCC < 0.60, *versus* 0.81, see Table S2). Among these 50 rules, the top six rules are derived from L434, G433, T433, G432 (C-terminus residues) and C4 and C72, which are the Cys residues in *Tre*Cel7A forming a rare disulfide bridge that is virtually absent in GH7 EGs. Overall, disulfide bonds are more frequent in GH7 CBHs than EGs (Table 6 and Fig. 10) (20).

Discussion

In this study, we apply data mining techniques along with the wealth of experimental GH7 data to investigate statistical relationships between sequence and function GH7s and to further identify known and novel sequence features that correlate with function. We are able to accurately discriminate 1748 GH7 CBHs and EGs with ML using only the number of residues in the active-site loops as features. However, whereas

Table 5
Performance (%) of random forest classifiers in predicting presence of CBM[a]

	Validation				Testing
Performance metric	All 5933 features	Top 50 features	44 features (no C-terminus)	Top 20 features	Top 20 features
Accuracy	90.8 ± 2.1	90.9 ± 2.1	88.2 ± 2.5	89.3 ± 2.4	89.7
Sensitivity	93.7 ± 2.8	92.2 ± 2.9	89.6 ± 3.4	90.0 ± 3.2	95.7
Specificity	87.9 ± 3.5	89.7 ± 3.3	86.9 ± 3.7	88.5 ± 3.6	87.4
MCC	0.80 ± 0.05	0.81 ± 0.05	0.76 ± 0.05	0.78 ± 0.05	0.68

[a] Validation and testing are performed on a 90%:10% split of the dataset, respectively. Validation performance is reported as the mean over 100 repetitions of 5-fold cross-validation ± 1 standard deviation.

the ML models trained on the lengths of A4, B2, B3, and B4 loops achieved high predictive performance (>94% accuracy), the models trained on the other loops demonstrated mediocre or poor performance (Table 1, Fig. 5, *A*). These results indicate that the lengths of the A4, B2, B3, and B4 loops are primarily important for the difference in GH7 CBH and EG behavior. Greater exposure of the active site is generally accepted as a hallmark of nonprocessive cellulases (EGs). In addition, the ML results indicate that exposure of the active site in GH7 EGs occurs primarily at the product-binding region (+1 and +2 glycosyl subsites) due to deletions in the A4 and B4 loops, at the region below the catalytic center due to deletions in the B3 loop, and at the region to the lower left of the catalytic center due to deletions in the B2 loop (Figs. 1 and 5*C*).

Earlier works have indicated that GH processivity correlates with ligand-binding affinity, ligand solvation, and the flexibility of catalytic residues (32, 75, 76). In *Tre*Cel7A, binding affinity is stronger at product-binding sites (+1, +2) than at the substrate-binding sites (−7 to −1), and this binding affinity difference has been proposed to be the driving force for the forward processive motion of the cellulase chain (8, 32, 35, 77). Consequently, a logical explanation for why the lengths of the A4 and B4 loops strongly correlate with GH7 CBH/EG function is that deletions in the A4 and B4 loops increase ligand solvation, disrupt protein–substrate hydrogen bonds, and lower binding affinity at the product-binding sites, leading to a decrease in processivity. Similarly, the strong relationship between the lengths of the B2 and B3 loops and GH7 CBH/EG function can be explained by the rationale that deletions in the B2 and B3 loops lead to an increase in solvation and a decrease in protein–ligand interactions in the substrate-binding sites and an increase in solvation and flexibility of catalytic residues. It is interesting that although the A2 and A3 loops also overlay the catalytic center of the active site, their lengths show practically no correlation with GH7 CBH/EG function (Figs. 5*A* and S1), and exposure of the catalytic center in GH7s is achieved primarily by deletions in the B2 and B3 loops instead.

Moreover, the level of variation in lengths of the loops, as measured by the relative standard deviation, positively correlates with the predictive performance of the loops in discriminating GH7 CBHs and EGs (Fig. 5, *A* and *B*). This suggests that variation in the lengths of active-site loops was a major strategy in the evolutionary design of processivity in GH7s such that variation was allowed in the loops that significantly affect processivity and limited in other loops that have little impact on processivity (A2 and A3).

Furthermore, there is a strong positive correlation between the lengths of the A4, B2, B3, and B4 loops (Fig. 6). Hence, in wild-type GH7s, the shortening of any one of these four loops is highly associated with truncation of the other three loops. On our dataset of 1748 sequences, we observed that if the B4 loop of a sequence is shortened, as is typical of GH7 EGs (*i.e.*, possessing three residues or less), the probability that the A4, B2, and B3 loops are all shortened to typical GH7 EG lengths (*i.e.*, five, four, and three residues or less, respectively) is 0.97 (Fig. 5, *D*). In other words, the pronounced concomitant shortening of the A4, B2, B3, and B4 loops observed in all crystal structures of GH7 EGs (26, 29, 78) is remarkably conserved in EGs across the GH7 family. This distinct bimodal distribution (Fig. S1) and strong correlation between loop lengths and GH7 subtype may serve as a valuable

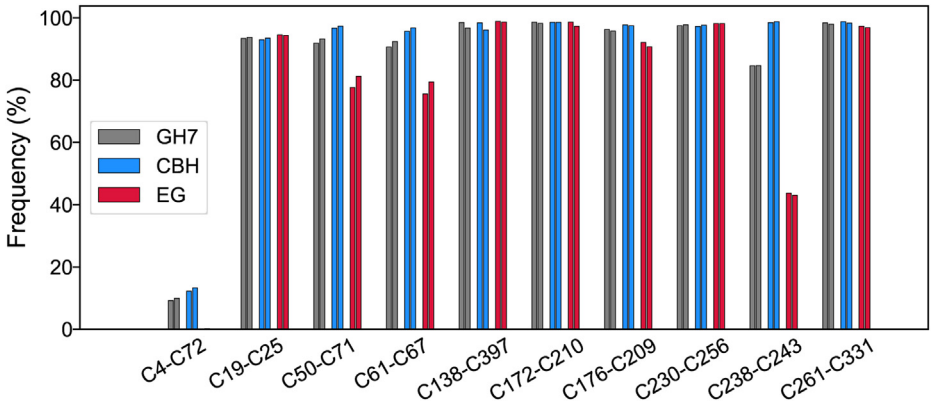


Figure 10. Frequency of Cys at positions forming disulfide bonds in GH7 sequences. Cys positions (*x*-axis) are labeled using *Tre*Cel7A numbering and the frequencies were determined from the structure-based MSA (1748 sequences). GH7 sequences may have up to ten disulfide bonds, nine of which are present in roughly at least 80% of the sequences. A rare disulfide bond, formed by C4 and C72 in *Tre*Cel7A, is present in less than 10% of GH7 sequences and is virtually absent in EGs. Overall, disulfide bonds are more prevalent in GH7 CBHs than EGs.

ASBMB

tool for correct gene annotation. Moreover, the strong conservation of loop lengths also indicates that there are coupled interactions between the A4, B2, B3, and B4 loops (24). This might explain why in the recent work of Schiano-di-Cola *et al.* (51), independent deletions in the B3 and B4 loops did not lead to significant improvements in the activity of *Tre*Cel7A on amorphous cellulose, and deletions in the A4 loop rendered the enzyme inactive. Cooperative synergy of deletions in other loops, as well as point mutations at key positions, may be required to fully exploit the effects of deletions in the B3, B4, and A4 loops.

Beyond the active-site loops, we have derived 90 position-specific classification rules from 42 positions in the MSA such that the specific amino acid, or amino acid type, at these positions can independently predict the GH7 functional subtype, with accuracies ranging from 87% to 97%. The high accuracy of these classification rules implies that there are strong constraints on the specific amino acids, or amino acid types, utilized by GH7 CBHs and EGs at these 42 positions. Such differential constraints likely signify that these positions are imperative to the distinction between GH7 CBH and EG behavior. More than half of these positions are within 5 Å of the cellononaose substrate bound in the *Tre*Cel7A structure, and many of these positions cluster around the B2 loop (Figs. 1 and 9, *D* and *E*). This finding provides a possible explanation for the observation that deletions in the B2 loop led to much greater changes in the CBH behavior of *Tre*Cel7A than deletions in the B3 and B4 loops, relative to *Tre*Cel7B (51). Since more of the important residues that yield high-accuracy position-specific classification rules cluster around the B2 loop than other loops, deletions in the B2 loop likely lead to a disruption of a greater number of interactions necessary for CBH activity than deletions in the B3 and B4 loops.

Many of the 42 positions from which we derived classification rules have been identified and studied in previous works, and mutations at these positions have led to significant increase in catalytic efficiency relative to the wild type (79–81). Trp38, Tyr51, Asn103, Lys181, Asn200, Asp179, Arg251, Asp262, and Arg394 were identified in a docking study as residues that directly interact with and stabilize the cellulose substrate in the active site of *Tre*Cel7A (82). Several of these positions have been further shown to form important stabilizing interactions with the substrate. Arg394 forms hydrogen bonds with the +2 glycosyl residue (34, 35), Arg251 forms a salt bridge with Asp259 and hydrogen bonds with the +1 and +2 glycosyl residues (6, 35, 36), and Asn103 and Lys181 form hydrogen bonds with the -5 glycosyl residue (18, 83). Sørensen *et al.* (80) studied mutants of *Rasamsonia emersonii* Cel7A in which two Asn residues on the B2 loop, Asn194 and Asn197 (Asn197 and Asn200 in *Tre*Cel7A, respectively) were replaced with Ala. They observed that the mutations led to a decrease in substrate affinity and processivity, thus enabling faster enzyme–substrate dissociation and a corresponding increase in activity on crystalline cellulose. In this present work, the Asn200 position yields the following classification rule: Asn implies CBH, and not Asn implies EG, which discriminates GH7 CBHs and EGs with an accuracy of 89%. Similarly, Bu *et al.* (36) conducted computational studies of several *Tre*Cel7A residues including Arg251, Asp262, and Tyr381. These residues were identified to substantially interact with the cellobiose substrate and mutation to Ala resulted in considerably weaker binding of cellobiose in the product-binding site. It was suggested that these mutants would demonstrate improved biomass conversion efficiency due to accelerated expulsion of the cellobiose product. In this present study, these positions (251, 262, and 381) also yield high-accuracy classification rules with accuracies of at least 88%. Additionally, Mitsuzawa *et al.* (79) determined that mutation of Asn63 and Lys203 to Ala in *Talaromyces cellulolyticus* Cel7A (Asn37 and Lys181 in TreCel7A, with classification accuracies of 91% and 94%, respectively) led to a remarkable increase in activity on cellulose.

Some positions farther away from the active site also yielded high-accuracy classification rules. For example, position 401—conserved as Ser in CBHs but as Glu in EGs and more than 13 Å away from the cellodextrin ligand in the *Tre*Cel7A structure—generates a CBH/EG classification rule with an accuracy of 92%. Although residues at positions such as 401 may not directly interact with the cellulose substrate in the active site, they may participate in long-range interactions that affect GH7 CBH and EG behavior. The fact that our approach has accurately returned so many of the known position-specific relationships (Table 2) builds confidence that the novel positions with relationships yet to be determined will impact enzyme function. As such, further studies are warranted to determine the specific roles these correlative positions play in function and structural stability. Altogether, we surmise that the positions that yield high-accuracy classification rules play key roles in GH7 CBH/EG function and, as such, should be carefully considered when engineering the protein at or around these sites.

Bioinformatic analysis of the MSA revealed conserved aromatic positions in the active site that are within 6 Å of the cellulose substrate in *Tre*Cel7A (Table 3, Fig. 8). The results indicate that whereas conserved aromatic residues in the active site of GH7 CBHs span the entire active-site tunnel, conserved aromatic residues in the active site of GH7 EGs are clustered around the catalytic center. Moreover, aromatic positions near the catalytic center are conserved in both GH7 CBHs and EGs. This arrangement of conserved aromatic residues in the active site suggests that while aromatic residues near the catalytic center (Y145, W216, H228, W367, and W376) play major roles in catalytic bond cleavage, conserved aromatic residues that flank the catalytic center (W38, W40, Y51, Y82, Y252, Y370, and Y381) are utilized mainly by CBHs for processive motion. Several experimental and computational studies support this hypothesis (36, 38, 84, 85).

Taylor *et al.* (20) assayed chimeras derived from interchanged subdomains of *Pfu*Cel7A and *Tre*Cel7A. Although the CD of *Pfu*Cel7A exhibited greater efficiency on biomass than the CD of *Tre*Cel7A, interchanging CBM and linker regions did not yield a uniform trend in catalytic efficiency. As a result, it was concluded that there are complex interactions that are not yet well understood between the domains. In this work, we have applied ML to predict, for the first time, the presence of CBMs from amino acid positions in the CD so as to map those nonintuitive relationships between the CBM and CD of GH7s;

such relationships are not readily identifiable by manual inspection of a gene. First, our data indicate that GH7 CBHs are roughly two times more likely to utilize CBMs than GH7 EGs, which is as expected since CBMs likely enable CBHs to stay longer on the cellulose substrate to facilitate consecutive hydrolysis. Furthermore, ML results show that the presence of a CBM in GH7s can be accurately predicted (89.3% accuracy) using only 20 features derived from 19 positions in the catalytic domain (Table 5). This high predictive accuracy largely suggests that there are constraints and key functional relationships between residue positions in the CD and the presence of a CBM in the gene. Interestingly, on the protein structure, these 19 positions are mostly located on loops or at turns all over the protein structure (Fig. 9, *C*). Moreover, the amino acid residues constituting the 20 features are mostly small amino acids (such as Gly, Ser, Thr, Asp, and Asn) that are known to affect the conformational flexibility of proteins (86, 87). Taken together, our ML results, while preliminary, suggest that the presence of CBMs in GH7s correlates with the overall conformational flexibility of the CD and that CBMs may exist, in part, to compensate for highly flexible CDs that are more likely to detach from the cellulose surface. Moreover, the position-specific classification rules we derived from the top 50 random-forest features in predicting CBMs indicate that GH7s possessing a rare disulfide bond (C4-C72 in *Tre*Cel7A) are about three times more likely to possess a CBM than GH7s that lack this disulfide (Tables 6 and S2, Fig. 10). In a previous work, mutation of C4 and C72 in *Tre*Cel7A was shown to increase cellulolytic efficiency and flexibility of the tunnel entrance (20). Since an extra disulfide bridge would generally decrease the flexibility of the CD, the positive correlation of C4 and C72 with the presence of a CBM is contrary to our hypothesis that CBMs compensate for the flexibility of the CD. This paradoxical correlation, thus, warrants further experiments to investigate such relationships.

In conclusion, we have used ML to uncover key positions in GH7 sequences that appear to be related to function and broader statistical relationships between GH7 sequence and functional diversity. Specifically, we identified aspects of GH7 sequence that correlate with activity across the entire family, including the lengths of the A4, B2, B3, and B4 loops and key positions such as 38, 251, 256, and 394. While some of these features have previously been examined for a handful of specific GH7 enzymes, our work here is the first to demonstrate that such features are not just characteristic of a singular GH7 enzyme but are, instead, coevolved with functional distribution across the entire family. Moreover, we also identified other positions that strongly correlate with activity but have yet to be considered in the structural or biochemical literature (such as 16, 146, and 338). By extension, these novel positions are very likely to play similarly critical roles in GH7 activity and can inform future experimental studies. Additionally, we demonstrated that ML applied to a MSA can uncover key positions in an enzyme family that are important for activity. These ML strategies will prove beneficial as a systematic screening technique to identify the positions that strongly correlate with the functional distribution in a given enzyme family. Thus, we anticipate that our findings will inform further propitious studies for the design of more efficient cellulases. While these sequence–function relationships are statistically significant, we stress that they may be influenced by sampling and phylogenetic biases inherent to the dataset. Nonetheless, the strategies we have applied here are extensible to other protein families, particularly where multiple functional classes exist (such as CBH/EG or CBM/no-CBM), and as such, this work provides a solid basis for future statistical investigations to establish sequence–function relationships as well as to identify sequence positions that are promising targets for protein engineering.

Experimental procedures

Sequence datasets

Sequences were retrieved by protein–protein BLAST searches against the NCBI nonredundant database by using *Tre*Cel7A (P62694.1) and *Fox*Cel7B (AAA65586.1) as query sequences. BLAST search was implemented with the NCBI web server (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using default settings. Only sequences with E-values of 1e-20 or better and query cover of 60% or more were retained. The query cover threshold of 60% was applied to exclude the large number of fragment sequences returned by the BLAST search. A total of 2024 sequences were retrieved. A sequence identity threshold of 99% was applied to remove redundant sequences so that only 1748 sequences were left in the dataset. From manual inspection of the BLAST output, 60 of these sequences consisted of multiple domains other than GH7. Other domains were deleted in these sequences leaving only one GH7 domain for each sequence. The UniProtKB/SwissProt dataset of 44 sequences was obtained by a similar BLAST search against the UniProtKB/SwissProt database.

Sequence alignments

Sequence alignments of the UniProtKB/SwissProt dataset (44 sequences) and the annotated NCBI dataset (427 sequences) were conducted with MAFFT version 7 (88) using BioPython (89) with default settings. Due to the greater diversity of the larger dataset (1748 sequences), in order to avoid generating erroneous alignments, a structure-based sequence alignment was implemented for the larger dataset. First, structural alignment of 20 GH7 structures (16 CBHs, four EGs) was conducted with the Promals3D web server (90). The structural alignment was manually edited in UGENE (91) following standard manual adjustment methods (92). Then, an MSA of the 1748 sequences was generated with the MAFFT add-sequences option (88) by adding the sequences to the structural alignment. Sequence alignments were viewed with ESPript (http://espript.ibcp.fr) (93), and sequence logos (Fig. S4) were generated with WebLogo (https://weblogo.berkeley.edu/logo.cgi) (94).

Machine learning and performance evaluation

Profile-hidden Markov models were constructed from the MSAs with a local version of the HMMER software (version

ASBMB

3.1b2) (55, 95). All ML methods were implemented using the Scikit-learn Python package (version 0.20.3) (96). The K-nearest neighbor (KNN) classifier was trained with the "n_estimators" parameter (k) set to an optimal value of 10 (best of 5, 10, and 15). A radial basis function (RBF) kernel was applied in the SVM classifiers, and default settings were used for the logistic regression classifiers. To avoid overfitting with the decision trees, the depth of the trees was limited to the number of features. Hence, single-feature decision trees had a "max-depth" of 1, and the decision tree trained on all eight features had a "max-depth" of 8.

There were severe outliers in the lengths of active-site loops that would have skewed the ML results. For example, from the MSA, a sequence (GenBank accession: CRK24563.1) had 140 residues in the B2 loop. These extremities may have resulted from sequencing or splicing errors. Before the ML procedure, outliers were capped to an arbitrarily selected maximum limit (60) (Fig. S5). All nonbinary features applied in ML were standardized by converting them to Z-scores according to following equation:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

The ML algorithms were applied to discriminate between a positive class (CBH or CBM) and a negative class (EG or no CBM), resulting in four classification outcomes: true-positives (TP), true-negatives (TN), false-positives (FP), and false-negatives (FN). The performance of the ML algorithms was evaluated by computing the sensitivity, specificity, accuracy, and MCC according to the following equations:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (4)$$

$$MCC = \frac{(TN\times TP)-(FP\times FN)}{\sqrt{(TN+FP)\times(TN+FN)\times(TP\times FP)\times(TN+FN)}} \quad (5)$$

Data availability

All datasets and Python scripts used in this study are available at https://doi.org/10.5281/zenodo.4216573.

Supporting information—This article contains supporting information (6, 8, 20, 34–36, 79, 80, 82, 83, 97).

Acknowledgments—This work was also authored in part by the Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308.

Authors contributions—J. E. G., B. E. H., and C. M. P. conceptualization; J. E. G. and B. E. H. data curation; J. E. G., B. E. H., J. S., and G. T. B. formal analysis; J. E. G., G. T. B., and C. M. P. funding acquisition; B. E. H., M. S., J. S., and G. T. B. methodology; J. E. G., G. T. B., and C. M. P. project administration; J. E. G., M. S., J. S., G. T. B., and C. M. P. resources; J. E. G., G. T. B., and C. M. P. software; J. E. G., M. S., J. S., G. T. B., and C. M. P. supervision; J. E. G., B. E. H., and M. S. validation; J. E. G. and B. E. H. visualization; J. E. G. and B. E. H. writing-original draft; J. E. G., B. E. H., M. S., J. S., G. T. B., and C. M. P. writing-review and editing; J. E. G. and B. E. H. investigation.

Funding and addditional information—This work was supported in part by the National Science Foundation (CBET-1552355 to C. M. P. in support of J. E. G.). Funding was provided to G. T. B. by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Bioenergy Technologies Office. This material is also based upon work supported by (while CMP is serving at) the NSF. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: CBH, cellobiohydrolase; CBM, carbohydrate-binding module; CD, catalytic domain; EG, endoglucanase; GH7, family 7 glycoside hydrolase; GH, glycoside hydrolase; HMM, hidden Markov model; KNN, k-nearest neighbor; LPMO, lytic polysaccharide monooxygenase; ML, machine learning; MSA, multiple sequence alignment.

References

1. Himmel, M. E., Ding, S. Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., and Foust, T. D. (2007) Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science* **315**, 804–807
2. Payne, C. M., Knott, B. C., Mayes, H. B., Hansson, H., Himmel, M. E., Sandgren, M., Stahlberg, J., and Beckham, G. T. (2015) Fungal cellulases. *Chem. Rev.* **115**, 1308–1448
3. Lynd, L. R., Weimer, P. J., van Zyl, W. H., and Pretorius, I. S. (2002) Microbial cellulose utilization: Fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506–577. table of contents
4. Zhang, Y. H. P., and Lynd, L. R. (2004) Toward an aggregated understanding of enzymatic hydrolysis of cellulose: Noncomplexed cellulase systems. *Biotech. Bioeng.* **88**, 797–824
5. Bu, L., Nimlos, M. R., Shirts, M. R., Stahlberg, J., Himmel, M. E., Crowley, M. F., and Beckham, G. T. (2012) Product binding varies dramatically between processive and nonprocessive cellulase enzymes. *J. Biol. Chem.* **287**, 24807–24813
6. Von Ossowski, I., Ståhlberg, J., Koivula, A., Piens, K., Becker, D., Boer, H., Harle, R., Harris, M., Divne, C., Mahdi, S., Zhao, Y., Driguez, H., Claeyssens, M., Sinnott, M. L., and Teeri, T. T. (2003) Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D. *J. Mol. Biol.* **333**, 817–829
7. Murphy, L., Cruys-Bagger, N., Damgaard, H. D., Baumann, M. J., Olsen, S. N., Borch, K., Lassen, S. F., Sweeney, M., Tatsumi, H., and Westh, P. (2012) Origin of initial burst in activity for *Trichoderma reesei* endo-glucanases hydrolyzing insoluble cellulose. *J. Biol. Chem.* **287**, 1252–1260

8. Wang, Y., Zhang, S., Song, X., and Yao, L. (2016) Cellulose chain binding free energy drives the processive move of cellulases on the cellulose surface. *Biotechnol. Bioeng.* **113**, 1873–1880
9. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495
10. Hobdey, S. E., Knott, B. C., Momeni, M. H., Taylor, L. E., Borisova, A. S., Podkaminer, K. K., VanderWall, T. A., Himmel, M. E., Decker, S. R., Beckham, G. T., and Stahlberg, J. (2016) Biochemical and structural characterizations of two Dictyostelium cellobiohydrolases from the Amoebozoa kingdom reveal a high level of conservation between distant phylogenetic trees of life. *J. Appl. Environ. Microbiol.* **82**, 3395–3409
11. Vinzant, T., Adney, W., Decker, S., Baker, J., Kinter, M., Sherman, N., Fox, J., and Himmel, M. (2001) Fingerprinting *Trichoderma reesei* hydrolases in a commercial cellulase preparation. *Appl. Biochem. Biotechnol.* **91**, 99–107
12. Martinez, D., Berka, R. M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S. E., Chapman, J., Chertkov, O., Coutinho, P. M., Cullen, D., Danchin, E. G., Grigoriev, I. V., Harris, P., Jackson, M., Kubicek, C. P., *et al.* (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat. Biotechnol.* **26**, 553
13. Moroz, O. V., Maranta, M., Shaghasi, T., Harris, P. V., Wilson, K. S., and Davies, G. J. (2015) The three-dimensional structure of the cellobiohydrolase Cel7A from *Aspergillus fumigatus* at 1.5 Å resolution. *Acta Crystallogr. F Struct. Biol. Commun.* **71**, 114–120
14. Borisova, A. S., Eneyskaya, E. V., Bobrov, K. S., Jana, S., Logachev, A., Polev, D. E., Lapidus, A. L., Ibatullin, F. M., Saleem, U., Sandgren, M., Payne, C. M., Kulminskaya, A. A., and Ståhlberg, J. (2015) Sequencing, biochemical characterization, crystal structure and molecular dynamics of cellobiohydrolase Cel7A from *Geotrichum candidum* 3C. *FEBS J.* **282**, 4515–4537
15. Momeni, M. H., Payne, C. M., Hansson, H., Mikkelsen, N. E., Svedberg, J., Engström, Å., Sandgren, M., Beckham, G. T., and Ståhlberg, J. (2013) Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus *Heterobasidion irregulare*. *J. Biol. Chem.* **288**, 5861–5872
16. Haddad Momeni, M., Goedegebuur, F., Hansson, H., Karkehabadi, S., Askarieh, G., Mitchinson, C., Larenas, E. A., Ståhlberg, J., and Sandgren, M. (2014) Expression, crystal structure and cellulase activity of the thermostable cellobiohydrolase Cel7A from the fungus *Humicola grisea var. thermoidea*. *Acta Crystallogr. Sect. D.: Biol. Cyrstallogr.* **70**, 2356–2366
17. Kern, M., McGeehan, J. E., Streeter, S. D., Martin, R. N., Besser, K., Elias, L., Eborall, W., Malyon, G. P., Payne, C. M., Himmel, M. E., Schnorr, K., Beckham, G. T., Cragg, S. M., Bruce, N. C., and McQueen-Mason, S. J. (2013) Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 10189–10194
18. Parkkinen, T., Koivula, A., Vehmaanperä, J., and Rouvinen, J. (2008) Crystal structures of *Melanocarpus albomyces* cellobiohydrolase Cel7B in complex with cello-oligomers show high flexibility in the substrate binding. *Protein Sci.* **17**, 1383–1394
19. Munoz, I. G., Ubhayasekera, W., Henriksson, H., Szabó, I., Pettersson, G., Johansson, G., Mowbray, S. L., and Ståhlberg, J. (2001) Family 7 cellobiohydrolases from *Phanerochaete chrysosporium*: Crystal structure of the catalytic module of Cel7D (CBH58) at 1.32 Å resolution and homology models of the isozymes. *J. Mol. Biol.* **314**, 1097–1111
20. Taylor, L. E., Knott, B. C., Baker, J. O., Alahuhta, P. M., Hobdey, S. E., Linger, J. G., Lunin, V. V., Amore, A., Subramanian, V., Podkaminer, K., Xu, Q.-S., VanderWall, T. A., Schuster, L. A., Chaudhari, Y. B., Adney, W. S., *et al.* (2018) Engineering enhanced cellobiohydrolase activity. *Nat. Commun.* **9**, 1186
21. Textor, L. C., Colussi, F., Silveira, R. L., Serpa, V., de Mello, B. L., Muniz, J. R. C., Squina, F. M., Pereira, N., Jr., Skaf, M. S., and Polikarpov, I. (2013) Joint X-ray crystallographic and molecular dynamics study of cellobiohydrolase I from *Trichoderma harzianum*: Deciphering the structural features of cellobiohydrolase catalytic activity. *FEBS J.* **280**, 56–69
22. Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J., Teeri, T. T., and Jones, T. A. (1994) The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Science* **265**, 524–528
23. Knott, B. C., Haddad Momeni, M., Crowley, M. F., Mackenzie, L. F., Gotz, A. W., Sandgren, M., Withers, S. G., Stahlberg, J., and Beckham, G. T. (2014) The mechanism of cellulose hydrolysis by a two-step, retaining cellobiohydrolase elucidated by structural and transition path sampling studies. *J. Am. Chem. Soc.* **136**, 321–329
24. Silveira, R. L., and Skaf, M. S. (2018) Concerted motions and large-scale structural fluctuations of *Trichoderma reesei* Cel7A cellobiohydrolase. *Phys. Chem. Chem. Phys.* **20**, 7498–7507
25. Mackenzie, L. F., Sulzenbacher, G., Divne, C., Jones, T. A., Woldike, H. F., Schulein, M., G, W. S., and Davies, G. J. (1998) Crystal structure of the family 7 endoglucanase I (Cel7B) from *Humicola insolens* at 2.2 Å resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate. *Biochem. J.* **335**, 409–416
26. Kleywegt, G. J., Zou, J. Y., Divne, C., Davies, G. J., Sinning, I., Stahlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T. T., and Jones, T. A. (1997) The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 A resolution, and a comparison with related enzymes. *J. Mol. Biol.* **272**, 383–397
27. Kadowaki, M. A. S., Higasi, P., de Godoy, M. O., Prade, R. A., and Polikarpov, I. (2018) Biochemical and structural insights into a thermostable cellobiohydrolase from *Myceliophthora thermophila*. *FEBS J.* **285**, 559–579
28. Borisova, A. S., Eneyskaya, E. V., Jana, S., Badino, S. F., Kari, J., Amore, A., Karlsson, M., Hansson, H., Sandgren, M., Himmel, M. E., Westh, P., Payne, C. M., Kulminskaya, A. A., and Stahlberg, J. (2018) Correlation of structure, function and protein dynamics in GH7 cellobiohydrolases from *Trichoderma atroviride, T. reese*i and *T. Harzianum*. *Biotechnol. Biofuels* **11**, 5
29. Sonoda, M. T., Godoy, A. S., Pellegrini, V. O., Kadowaki, M. A., Nascimento, A. S., and Polikarpov, I. (2019) Structure and dynamics of *Trichoderma harzianum* Cel7B suggest molecular architecture adaptations required for a wide spectrum of activities on plant cell wall polysaccharides. *Biochim. Biophys. Acta Gen. Subj.* **1863**, 1015–1026
30. Schiano-di-Cola, C., Kolaczkowski, B., Sorensen, T. H., Christensen, S. J., Cavaleiro, A. M., Windahl, M. S., Borch, K., Morth, J. P., and Westh, P. (2019) Structural and biochemical characterization of a family 7 highly thermostable endoglucanase from the fungus *Rasamsonia emersonii*. *FEBS J.* **287**, 2577–2596
31. Kurašin, M., and Väljamäe, P. (2011) Processivity of cellobiohydrolases is limited by the substrate. *J. Biol. Chem.* **286**, 169–177
32. Payne, C. M., Jiang, W., Shirts, M. R., Himmel, M. E., Crowley, M. F., and Beckham, G. T. (2013) Glycoside hydrolase processivity is directly related to oligosaccharide binding free energy. *J. Am. Chem. Soc.* **135**, 18831–18839
33. Divne, C., Ståhlberg, J., Teeri, T. T., and Jones, T. A. (1998) High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J. Mol. Biol.* **275**, 309–325
34. Ubhayasekera, W., Muñoz, I. G., Vasella, A., Ståhlberg, J., and Mowbray, S. L. (2005) Structures of *Phanerochaete chrysosporium* Cel7D in complex with product and inhibitors. *FEBS J.* **272**, 1952–1964
35. Knott, B. C., Crowley, M. F., Himmel, M. E., Ståhlberg, J., and Beckham, G. T. (2014) Carbohydrate–protein interactions that drive processive polysaccharide translocation in enzymes revealed from a computational study of cellobiohydrolase processivity. *J. Am. Chem. Soc.* **136**, 8810–8819
36. Bu, L., Beckham, G. T., Shirts, M. R., Nimlos, M. R., Adney, W. S., Himmel, M. E., and Crowley, M. F. (2011) Probing carbohydrate product expulsion from a processive cellulase with multiple absolute binding free energy methods. *J. Biol. Chem.* **286**, 18161–18169
37. Igarashi, K., Koivula, A., Wada, M., Kimura, S., Penttilä, M., and Samejima, M. (2009) High speed atomic force microscopy visualizes processive movement of *Trichoderma reesei* cellobiohydrolase I on crystalline cellulose. *J. Biol. Chem.* **284**, 36186–36190

38. Nakamura, A., Tsukada, T., Auer, S., Furuta, T., Wada, M., Koivula, A., Igarashi, K., and Samejima, M. (2013) The tryptophan residue at the active site tunnel entrance of *Trichoderma reesei* cellobiohydrolase Cel7A is important for initiation of degradation of crystalline cellulose. *J. Biol. Chem.* **288**, 13503–13510
39. Beckham, G. T., Matthews, J. F., Bomble, Y. J., Bu, L., Adney, W. S., Himmel, M. E., Nimlos, M. R., and Crowley, M. F. (2010) Identification of amino acids responsible for processivity in a Family 1 carbohydrate-binding module from a fungal cellulase. *J. Phys. Chem. B.* **114**, 1447–1453
40. Beckham, G. T., Bomble, Y. J., Matthews, J. F., Taylor, C. B., Resch, M. G., Yarbrough, J. M., Decker, S. R., Bu, L., Zhao, X., McCabe, C., and Wohlert, J. (2010) The O-glycosylated linker from the *Trichoderma reesei* Family 7 cellulase is a flexible, disordered protein. *Biophys. J.* **99**, 3773–3781
41. Sammond, D. W., Payne, C. M., Brunecky, R., Himmel, M. E., Crowley, M. F., and Beckham, G. T. (2012) Cellulase linkers are optimized based on domain type and function: Insights from sequence analysis, biophysical measurements, and molecular simulation. *PloS one* **7**, e48615
42. Harrison, M. J., Nouwens, A. S., Jardine, D. R., Zachara, N. E., Gooley, A. A., Nevalainen, H., and Packer, N. H. (1998) Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of *Trichoderma reesei*. *Eur. J. Biochem.* **256**, 119–127
43. Amore, A., Knott, B. C., Supekar, N. T., Shajahan, A., Azadi, P., Zhao, P., Wells, L., Linger, J. G., Hobdey, S. E., Vander Wall, T. A., Shollenberger, T., Yarbrough, J. M., Tan, Z., Crowley, M. F., Himmel, M. E., *et al.* (2017) Distinct roles of N- and O-glycans in cellulase activity and stability. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13667–13672
44. Ståhlberg, J., Johansson, G., and Pettersson, G. (1991) A new model for enzymatic hydrolysis of cellulose based on the two-domain structure of cellobiohydrolase I. *Nat. Biotechnol.* **9**, 286
45. Van Tilbeurgh, H., Tomme, P., Claeyssens, M., Bhikhabhai, R., and Pettersson, G. (1986) Limited proteolysis of the cellobiohydrolase I from *Trichoderma reesei*: Separation of functional domains. *FEBS Lett.* **204**, 223–227
46. Tomme, P., van Tilbeurgh, H., Pettersson, G., Van Damme, J., Vandekerckhove, J., Knowles, J., Teeri, T., and Claeyssens, M. (1988) Studies of the cellulolytic system of *Trichoderma reesei* QM 9414: Analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur. J. Biochem.* **170**, 575–581
47. Reinikainen, T., Ruohonen, L., Nevanen, T., Laaksonen, L., Kraulis, P., Jones, T. A., Knowles, J. K., and Teeri, T. T. (1992) Investigation of the function of mutated cellulose-binding domains of *Trichoderma reesei* cellobiohydrolase I. *Proteins Struct. Funct. Bioinf.* **14**, 475–482
48. Srisodsuk, M., Lehtiö, J., Linder, M., Margolles-Clark, E., Reinikainen, T., and Teeri, T. T. (1997) *Trichoderma reesei* cellobiohydrolase I with an endoglucanase cellulose-binding domain: Action on bacterial microcrystalline cellulose. *J. Biotechnol.* **57**, 49–57
49. Le Costaouëc, T., Pakarinen, A., Várnai, A., Puranen, T., and Viikari, L. (2013) The role of carbohydrate binding module (CBM) at high substrate consistency: Comparison of *Trichoderma reesei* and *Thermoascus aurantiacus* Cel7A (CBHI) and Cel5A (EGII). *Bioresour. Technol.* **143**, 196–203
50. Takashima, S., Ohno, M., Hidaka, M., Nakamura, A., Masaki, H., and Uozumi, T. (2007) Correlation between cellulose binding and activity of cellulose-binding domain mutants of *Humicola grisea* cellobiohydrolase 1. *FEBS Lett.* **581**, 5891–5896
51. Schiano-di-Cola, C., Røjel, N., Jensen, K., Kari, J., Sørensen, T. H., Borch, K., and Westh, P. (2019) Systematic deletions in the cellobiohydrolase (CBH) Cel7A from the fungus *Trichoderma reesei* reveal flexible loops critical for CBH activity. *J. Biol. Chem.* **294**, 1807–1815
52. Alpaydin, E. (2009) *Introduction to Machine Learning*, MIT press, Cambridge, MA
53. Consortium, U. (2009) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142–D148
54. Whisstock, J. C., and Lesk, A. M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**, 307–340
55. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755–763
56. De Fonzo, V., Aluffi-Pentini, F., and Parisi, V. (2007) Hidden Markov models in bioinformatics. *Curr. Bioinform.* **2**, 49–61
57. Hannenhalli, S. S., and Russell, R. B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76
58. Zhu, X., and Wu, X. (2004) Class noise vs. attribute noise: A quantitative study. *Artif. Intell.* **22**, 177–210
59. Pechenizkiy, M., Tsymbal, A., Puuronen, S., and Pechenizkiy, O. (2006) Class noise and supervised learning in medical domains: The effect of feature extraction. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE
60. Chicco, D. (2017) Ten quick tips for machine learning in computational biology. *BioData Min* **10**, 35
61. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16**, 412–424
62. Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biophys. Acta Protein Struct.* **405**, 442–451
63. Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit* **91**, 216–231
64. He, H., and Garcia, E. A. (2008) Learning from imbalanced data. *IEEE T. Knowl. Data En.*, 1263–1284
65. Drummond, C., and Holte, R. C. (2003) C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Datasets II*. National Research Council Canada, Ottawa, Ontario
66. Kim, J.-H. (2009) Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* **53**, 3735–3745
67. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. (2011) An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis. Support Syst.* **51**, 141–154
68. Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215
69. Zhang, Y., Yan, S., and Yao, L. (2013) A mechanistic study of *Trichoderma reesei* Cel7B catalyzed glycosidic bond cleavage. *J. Phys. Chem. B.* **117**, 8714–8722
70. Lin, C.-T., Lin, K.-L., Yang, C.-H., Chung, I.-F., Huang, C.-D., and Yang, Y.-S. (2005) Protein metal binding residue prediction based on neural networks. *Int. J. Neural Syst.* **15**, 71–84
71. Breiman, L. (2001) Random forests. *Mach. Learn.* **45**, 5–32
72. Chen, X., and Ishwaran, H. (2012) Random forests for genomic data analysis. *Genomics* **99**, 323–329
73. Han, P., Zhang, X., Norton, R. S., and Feng, Z.-P. (2009) Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinf* **10**, 8
74. Archer, K. J., and Kimes, R. V. (2008) Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* **52**, 2249–2260
75. Beckham, G. T., Ståhlberg, J., Knott, B. C., Himmel, M. E., Crowley, M. F., Sandgren, M., Sørlie, M., and Payne, C. M. (2014) Towards a molecular-level theory of carbohydrate processivity in glycoside hydrolases. *Curr. Opin. Biotechnol.* **27**, 96–106
76. Payne, C. M., Baban, J., Horn, S. J., Backe, P. H., Arvai, A. S., Dalhus, B., Bjørås, M., Eijsink, V. G., Sørlie, M., Beckham, G. T., and Vaaje-Kolstad, G. (2012) Hallmarks of processivity in glycoside hydrolases from crystallographic and computational studies of the *Serratia marcescens* chitinases. *J. Biol. Chem.* **287**, 36322–36330
77. Colussi, F., Sørensen, T. H., Alasepp, K., Kari, J., Cruys-Bagger, N., Windahl, M. S., Olsen, J. P., Borch, K., and Westh, P. (2015) Probing substrate interactions in the active tunnel of a catalytically deficient cellobiohydrolase (Cel7). *J. Biol. Chem.* **290**, 2444–2454

78. Sulzenbacher, G., Schulein, M., and Davies, G. J. (1997) Structure of the endoglucanase I from *Fusarium oxysporum*: Native, cellobiose, and 3,4-epoxybutyl beta-D-cellobioside-inhibited forms, at 2.3 A resolution. *Biochemistry* **36**, 5902–5911
79. Mitsuzawa, S., Fukuura, M., Shinkawa, S., Kimura, K., and Furuta, T. (2017) Alanine substitution in cellobiohydrolase provides new insights into substrate threading. *Sci. Rep.* **7**, 16320
80. Sørensen, T. H., Windahl, M. S., McBrayer, B., Kari, J., Olsen, J. P., Borch, K., and Westh, P. (2017) Loop variants of the thermophile *Rasamsonia emersonii* Cel7A with improved activity against cellulose. *Biotechnol. Bioeng.* **114**, 53–62
81. Zong, Z., Li, Q., Hong, Z., Fu, H., Cai, W., Chipot, C., Jiang, H., Zhang, D., Chen, S., and Shao, X. (2019) Lysine mutation of the Claw-Arm-like loop accelerates catalysis by cellobiohydrolases. *J. Am. Chem. Soc.* **141**, 14451–14459
82. Mulakala, C., and Reilly, P. J. (2005) *Hypocrea jecorina* (*Trichoderma reesei*) Cel7A as a molecular machine: A docking study. *Proteins Struct. Funct. Bioinf.* **60**, 598–605
83. GhattyVenkataKrishna, P. K., Alekozai, E. M., Beckham, G. T., Schulz, R., Crowley, M. F., Uberbacher, E. C., and Cheng, X. (2013) Initial recognition of a cellodextrin chain in the cellulose-binding tunnel may affect cellobiohydrolase directional specificity. *Biophys. J.* **104**, 904–912
84. Kari, J., Olsen, J., Borch, K., Cruys-Bagger, N., Jensen, K., and Westh, P. (2014) Kinetics of cellobiohydrolase (Cel7A) variants with lowered substrate affinity. *J. Biol. Chem.* **289**, 32459–32468
85. Taylor, C. B., Payne, C. M., Himmel, M. E., Crowley, M. F., McCabe, C., and Beckham, G. T. (2013) Binding site dynamics and aromatic–carbohydrate interactions in processive and non-processive family 7 glycoside hydrolases. *J. Phys. Chem. B.* **117**, 4924–4933
86. Betts, M. J., and Russell, R. B. (2003) Amino acid properties and consequences of substitutions. In *Bioinformatics for Geneticists*, Wiley, West Sussex, UK: 289–314
87. Huang, F., and Nau, W. M. (2003) A conformational flexibility scale for amino acids in peptides. *Angew. Chem.* **42**, 2269–2272
88. Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780
89. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423
90. Pei, J., Kim, B. H., and Grishin, N. V. (2008) PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300
91. Okonechnikov, K., Golosova, O., Fursov, M., and team, U. (2012) Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167
92. Doolittle, R. F. (1996) *Computer Methods for Macromolecular Sequence Analysis*, Academic Press, San Diego, CA
93. Robert, X., and Gouet, P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324
94. Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004) WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190
95. Eddy, S. (2003) *HMMER User's Guide. Biological Sequence Analysis Using Profile Hidden Markov Models*, Howard Hughes Medical Institute, Chevy Chase, MD
96. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Matthieu, P., *et al.* (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830
97. Payne, C. M., Resch, M. G., Chen, L., Crowley, M. F., Himmel, M. E., Taylor, L. E., 2nd, Sandgren, M., Stahlberg, J., Stals, I., Tan, Z., and Beckham, G. T. (2013) Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14646–14651

ASBMB