Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 2214

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

Samantha G. Hennen,[a] Yannick J. Bomble,[a] Breanna R. Urbanowicz[cd] and Vivek S. Bharadwaj *[b]

Substrate specificity is an essential characteristic of any enzyme's function and an understanding of the factors that determine this specificity is crucial for enzyme engineering. Unlike the structure of an enzyme which is directly impacted by its sequence, substrate specificity as an enzyme attribute involves a rather indirect relationship with sequence as it also depends on structural aspects that dictate substrate accessibility and active site dynamics. In this study, we explore the performance of classifier-based machine learning models trained on curated sequence and structural data for a class of glycosyltransferases (GTs), namely GT-Bs, to understand their substrate specificity determining factors. GTs enable the transfer of sugar moieties to other biomolecules such as oligosaccharides or proteins and are found in all kingdoms of life. In plants, GTs participate in the biosynthesis of plant cell wall biopolymers (*e.g.*: hemicelluloses and pectins) and are an integral part of the enzymatic machinery that enables the storage of carbon and energy as plant biomass. To elucidate the substrate specificity of uncharacterized GT-Bs, we constructed multi-label machine learning models (Support Vector Classifier, *K*-Nearest Neighbors, Gaussian Naïve-Bayes, Random Forest) that incorporate both sequence and structural features. These models achieve good predictive accuracies on test datasets. However, despite our use of structural information, we highlight that there is further scope for improvement in training these models to draw interpretable relationships between sequence, structure and substrate specificity determining motifs in GT-Bs.

Received 21st October 2024
Accepted 2nd July 2025

DOI: 10.1039/d4dd00338a

rsc.li/digitaldiscovery

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

1. Introduction

Plants employ amongst nature's most efficient carbon capture mechanisms, storing much of the world's carbon, and are therefore a valuable resource for conversion to fuels and products. The polysaccharides present in the cell walls that make up the bulk of plant biomass are synthesized, constructed, and modified by a conglomerate of enzymes. A thorough understanding of these enzymes, their substrate specificity and catalytic functions is of utmost importance for our fundamental knowledge of how the carbon fixed *via* photosynthesis is converted and stored in plant biomass, and for facilitating the technology-development to design tailored biopolymers for materials.[1–4] Among the various classes of enzymes involved in plant cell wall biosynthesis, of particular interest are glycosyltransferases (GTs) that catalyze the formation of glycosidic bonds by transferring sugar moieties from sugar-nucleotide donors to oligosaccharide acceptors.[5,6] These enzymes are responsible for the formation of complex glycopolymers that constitute a large portion of the cell wall governing its architecture.[5]

GTs are ubiquitous in both plant and animal species and have thus far been classified into 117 families in the CAZy database on the basis of their sequence similarity.[5,7] GTs are known to adopt one of three major structural folds: GT-A, -B, and -C and comprise 21, 27, and 10 families identified in the CAZy database respectively.[7] Unfortunately, the structural folds of several families have not yet been officially classified by the CAZy database,[7] due to the lack of experimental evidence. While individual members of some families have been proposed to adopt certain folds, such as the fucosyltransferase AtFUT1 from the GT37 family,[8] the structural, functional and mechanistic details of the majority of GTs are still not well understood. The GT-B fold was first identified as a distinct folding superfamily in 2001, and is characterized by a catalytic site localized between two Rossmann-like subdomains.[9] GT-Bs are of particular interest for their important role in the synthesis of non-

[a]*Biosciences Center, National Renewable Energy Laboratory, Golden, CO, USA*
[b]*Renewable Resources and Enabling Sciences Center, National Renewable Energy Laboratory, Golden, CO, USA. E-mail: vivek.bharadwaj@nrel.gov*
[c]*Complex Carbohydrate Research Center, University of Georgia, Athens, GA, USA*
[d]*Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00338a

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

cellulosic plant polysaccharides, a major portion of plant biomass.[5] GT-B enzymes in the GT37 and GT47 families, for example, are key enzymes involved in xyloglucan synthesis, the major hemicellulose in the primary cell walls of dicots.[8,10–12] Despite their important role in plants and the presence of a diverse range of monosaccharides in plants, detailed knowledge of the substrate specificities of many GT-B proteins is yet to be fully understood.

Current approaches used to investigate the molecular mechanisms of biocatalysts often rely on the use of experimentally determined protein crystal structures to develop functional hypotheses that form the basis for subsequent biochemical and mechanistic investigations. While these structures offer highly detailed information, they can be difficult and tedious to generate even for a single candidate.[13–17] This approach becomes intractable for a comprehensive exploration of the impact of mutations or natural variants on substrate specificities in an enzyme family or class. Fortunately, thanks to advances in genomics, there is currently a wealth of available sequence data, allowing for machine learning (ML) approaches that can find patterns throughout large numbers of sequences and relate it to substrate specificity.[18–20] Additionally, there are now over 200 million predicted protein structures available in the AlphaFold2 (AF2) database, allowing for structural information for all protein coding genes from a multitude of species to be easily considered alongside sequence data.[21,22]

Machine learning approaches that leverage this abundance of data are being increasingly used to elucidate properties and activities of enzymes and have been employed in recent studies to predict substrate specificity for some GTs. Yang *et al.* generated an activity assay for GT-1 enzymes, which is a family known to adopt a GT-B fold form, and used decision trees trained on this data to predict donor and acceptor substrate specificity.[23] Taujale *et al.* evaluated the use of tree-based models to predict the donor specificities of GT-A sequences, the most abundant and well-characterized of the three GT folds.[24] There has been no significant efforts focused on predicting the specificities across diverse GT families that adopt a GT-B fold. Predicting substrate specificity is a challenging task. Amongst the multitude of factors that determine substrate specificity, some are directly related to sequence *e.g.*: structure, while others are consequent attributes of sequence *e.g.*: active site dynamics, and many more are totally sequence-independent *e.g.*: substrate chemical environments.[25] Furthermore, many GT families display polyspecificity, with GT47 being a key example wherein members utilize a variety of donors and acceptors, making precise functional predictions even more difficult and unreliable. It is especially challenging to connect sequence to specificity for GT-Bs, as this fold family has little inter-family sequence similarity, and multiple families lack experimentally characterized structures.[24,26]

In this study, we curated sequence-activity data on GT-Bs, built multi-label classifier ML models, compared the performance of four types of models (Random Forest, Support Vector Machines, and K-Nearest Neighbors) on their ability to predict their donor binding specificity, and attempted to connect critical residue features identified from the ML models to the enzyme structure to identify substrate specificity determining motifs in GT-Bs. Our approach began with curating GT-B sequences for training and testing our models and consisted of sequences obtained from the CAZy and Uniprot databases that have been annotated for activity on one of seven diphosphate sugar substrates (GDP-Mannose, GDP-Fucose, UDP-Galactose, UDP-Glucose, UDP-Glucuronic acid, UDP-Xylose, UDP-Rhamnose). To handle samples with other known donor substrates that are not represented by the seven substrates, we employed an eighth substrate class, "Other". This was followed by the description of these GT-B sequences in terms of sequence features such as residue-based polarity, hydrophobicity, and charge, as well as structural features, including solvent accessible surface area and secondary structure (obtained from AF2 predicted structures) to build our ML models. Additionally, due to the large diversity within GT-B families in sequence and structure, only the residues in the Rossmann-like subdomains and catalytic site in the AF2 structural models were aligned and featurized. This was done to generate more meaningful multiple sequence alignments (MSAs), as well as focus the model on positions more likely to dictate substrate specificity. Furthermore, unlike previous studies, these models have been built to predict multiple substrates for each enzyme, as this is an important consideration for GT families that are known to utilize multiple donor substrates, such as GT1, GT4, GT31, and GT47. The trained models have proved to be accurate with the KNN model achieving cross-validation and test scores of 94% and 85%, respectively. We then performed a conservation analysis on the set of residues whose features contribute most to decision-making in the model and translate it to the enzyme structure and results of docking calculations to gauge their importance in substrate-binding.

2. Methods

2.1. Dataset collection

The training dataset comprised GT-B enzymes from 145 species and subspecies gathered from the UniProt database (ESI Fig. 1†). To confirm GT-B identity, only families previously shown to adopt a GT-B fold, as indicated in the Carbohydrate Active Enzymes (CAZy) database, were included in the dataset. Further, only UniProt-reviewed proteins were considered to ensure high quality data.[7] UniProt catalytic activity information was mined to determine donor substrate specificity, with proteins utilizing a donor substrate known to participate in plant carbohydrate biosynthesis selected for the dataset. Five GT47, four GT37, and five GT61 enzymes were also added, as they are known to adopt a GT-B fold, with their substrates were identified from previous literature and the CAZy database.[7,8,10,11,13,27–30] Proteins were labelled with all UniProt-identified substrates. Sequences with their donor substrates listed as a generic NDP-α-D-glucose donor substrate were excluded. Sequences with a donor substrate appearing less than three times in the dataset were also excluded, as at least three examples are needed for representation in the training, cross-validation, and test sets. This curation resulted in a dataset with seven unique donor substrates (GDP-α-D-mannose (GDP-

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 1 The KNN donor prediction F1 scores on the test set subsets are shown, where proteins sequences above the full sequence identity similarity cutoffs to any training set sequences were excluded

Max identity (%)	Sequences	Unique labels	Test F1 score	Average substrate MCC score
75	100	8	85.0% ± 35.7%	71.3%
70	88	8	82.9% ± 37.6%	69.4%
65	81	7	81.4% ± 38.8%	64.1%
60	76	7	80.2% ± 39.8%	63.7%
55	69	7	78.2% ± 41.3%	62.2%
50	57	7	75.4% ± 43.1%	57.6%

Man), GDP-L-β-fucose (GDP-Fuc), UDP-α-D-galactose (UDP-Gal), UDP-α-D-glucose (UDP-Glc), UDP-α-D-glucuronic acid (UDP-Glcua), UDP-α-D-xylose (UDP-Xyl), and UDP-β-L-rhamnose (UDP-Rha)). To account for the models' ability to predict unseen classes *i.e.* substrates outside the seven unique donors, we included an eighth class designated as "Other" in the test and training datasets.[31] This involved adding 39 sequences with a known substrate not included in the seven donors to assess performance on samples with an unrepresented donor.

As the structural diversity of these enzymes is likely to cause inaccurate sequence alignments, only sections of the sequence corresponding to the characteristic Rossmann-like domains[9] and catalytic regions were used in training. Each residue's secondary structure assignment was made using AF2 structures for each sequence and PyMOL.[32] To confirm validity of the AF2 models, residues were assessed for confidence, which found an average 96.4% of residues for each truncated training and testing set structure had a pLDDT score of at least 70%. While a Rossmann-like fold is generally considered to have six to seven β-strands in each sheet, many structures in the dataset contained less, therefore sheets of at least four strands were considered sufficient for our curation. Fourteen proteins from the dataset were excluded from this analysis for lack of clear Rossmann domains. This curation eventually resulted in 513 GT-B proteins (dataset hosted at **https://github.com/vbharadwcomosci/GTB_substrate_prediction/tree/main/Data**) to be further split into training and test sets. The percent shared identity was calculated for all sequence pairs, with the test set sequences chosen if they contained less than 75% identity to any training sequences. This resulted in final training and test datasets of 413 and 100 sequences, respectively. The training dataset contained 19 GT-B families with seven distinct donors and an eighth "Other" class for unrepresented substrates (ESI Fig. 2† and Table 1). To further assess model robustness, additional test subsets were generated to exclude proteins of identity score cutoffs of 70%, 65%, 60%, 55%, and 50% with any training sequence.

2.2. Featurization

The reduced sequences were aligned with Clustal Omega to create a multiple sequence alignment (MSA).[33] Highly gapped residue positions in the aligned sequences were removed from the MSA to select only the most relevant features, and those that might have a relationship to structure. The MSA used in previous ML work on GT-A fold enzymes was curated similarly, with positions of over 15% gaps instead removed.[14] This cutoff was increased to 50% in this work, as the lower cutoff would result in few remaining positions in the highly dissimilar GT-B fold enzymes, resulting in 803 removed MSA positions. The MSA was then converted to a dataset of feature vectors representing each sequence, featurized for residue property values. Each residue within the MSA was featurized with AAIndex assigned values for hydrophobicity, residue volume, accessible surface area (ASA), polarity, and charge.[34] Solvent accessible surface area (SASA) and categorical secondary structure values were also assigned for each residue from AF2 structures by BioPython and PyMOL, respectively.[32,35] ASA and SASA differ as SASA accounts for the position of the amino acid residue in the structure, while ASA is dependent only on residue identity. AF2 residues with pLDDT scores below 70% are considered low-confidence predictions by AF2, and as such were assigned values of 0 for solvent accessible surface area and secondary structure to avoid inputting potentially incorrect data. Feature values were normalized between 0 to 1 for each feature type to prevent overemphasis on features with larger values. SASA values were normalized relative to each structure. As most models only allow non-null feature values, gapped positions received values of 0. Finally, to remove redundant features, 431 features with high correlation (>90%) in the training dataset were removed prior to cross-validation splitting. This refinement was performed as the unfiltered feature vectors could result in inefficient, overfit models. This curation resulted in 1739 features to be further refined during model training.

To maintain identical feature length between training and test sets, each test sequence was individually aligned to the unfiltered training MSA. As the training MSA is a fixed length, the test set sequences always resulted in an alignment within range of the training set length. The test sequences and structures were then featurized for the same features used in the training set.

2.3. Model training

Four classifier ML model types (Random Forest (RF), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), and Support Vector (SV)) from the Scikit-Learn Python package were trained for classification with the 413 sequences in the training set.[36] Due to known GT substrate promiscuity, these models were

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

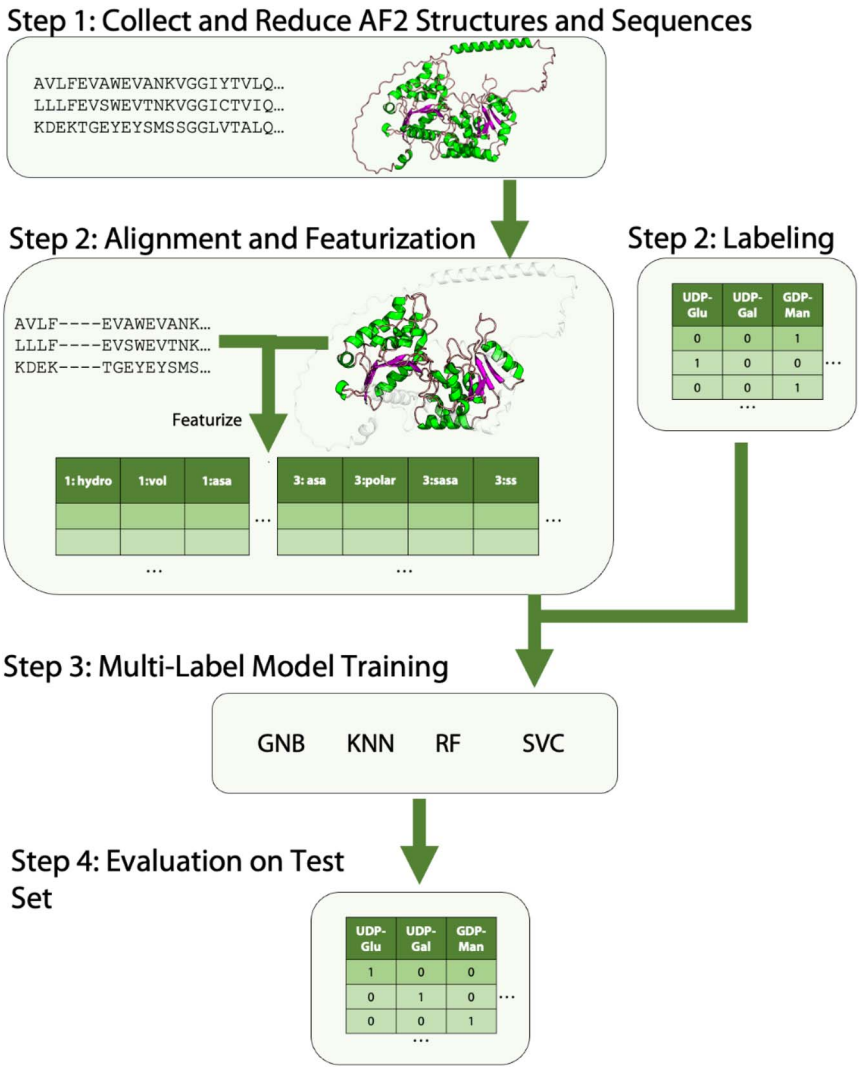

Fig. 1 An overview of model featurization and training for GT-B fold glycosyltransferases. In step 1, the AF2 structures were collected and reduced to the Rossmann-like domains. For step 2, a multiple sequence alignment was constructed and sequences were featurized based on amino-acid properties such as hydrophobicity, volume, accessible surface area, polarity, charge, and structure-based amino acid properties – solvent accessible surface area and secondary structure. Highly gapped MSA positions and highly correlated positions in the featurized MSA were removed. The GTs were also labelled with the substrates to which they are confirmed to bind for prediction. Next, several model types were trained on feature subsets in step 3, and their hyperparameters were tuned to identify optimal values. Finally, these models were assessed on the test set in step 4.

built to allow for multi-label prediction.[23] An overview of the model featurization and training protocol is depicted in Fig. 1.

Hyperparameters, such as the number of trees in an RF model, control a model's complexity and can drastically alter performance, thus it is necessary to evaluate several combinations. Furthermore, different feature lengths should be evaluated to determine the minimum features needed for high accuracy to maximize efficiency and reduce overfitting. Therefore, hyperparameter tuning and feature selection was performed and optimized with leave-one-out cross-validation. A grid search was used to compare every combination of hyperparameters. Test set sequences were kept separate from those used for training and cross-validation for later performance assessment and were not used in the feature selection and hyperparameter tuning grid search. F1 scores were averaged for each multi-label sample and used as the evaluation metric. To select the feature subset with optimal model performance, Scikit-Learn's chi2 SelectKBest was used to select various feature subsets for model training.[37] For each feature length between 50 and 1000, in multiples of 50, the model was trained with every combination of hyperparameters and assessed for performance. All hyperparameter search ranges can be found in ESI Table 2.†

2.4. Family based model

To ensure that the model was learning more significant relationships than simple family identifications, it was also trained with family number as its only feature and its cross-validation

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

and test set performance compared to the model of higher complexity. The training and test sets were kept identical. All models (SVC, RF, KNN and GNB) were trained with this single-family feature and their hyperparameters tuned with an identical protocol as the more complex model.

2.5. Identifying and comparing conserved regions and binding sites

Residue positions used by the optimized model were assessed for consensus to elucidate relationships between conserved residues and substrate specificity. GT-B fold enzymes that bind UDP-Glc were considered here as it is a commonly observed donor substrate with known activity and is utilized by enzymes in several different families. As these families have dissimilar sequences (and structures), this evaluation was intended to explore the presence of structural components that might dictate substrate specificity. Residue positions whose features contributed to the optimized model were assessed for consensus of residue type (hydrophobic, polar, positive *versus* negatively charged) due to the minimal conservation between GT-B fold families. To elucidate the high consensus residues and their potential relationship to ligand binding, docking simulations were performed with representative structures from multiple families. While there are machine learning-based docking softwares available, such as DiffDock, their performance on sugar nucleotide molecules remains untested.[38] Instead, a physics-based model, AutoDock Vina[39,40] was used for simulation and blind-docking simulations were run on GT4, GT20 and GT28 candidate structures with UDP-Glc as the ligand. The GT structures were shortened to the Rossmann regions used in the ML models. 100 potential ligand poses were produced with the exhaustiveness parameter set to 320. High consensus residues were mapped onto the docked enzyme–substrate complexes to assess their role in binding.

2.6. Application to other genera

The trained and tuned models were used to predict substrate specificities of uncharacterized GT sequences from four dissimilar chlorophyte genera exclusive to the training dataset – *Populus*, *Spirodela*, *Chlamydomonas*, and *Eucalyptus*. While the training and testing sets where restricted to only Uniprot-reviewed sequences that are also classified as GT-Bs by CAZy, this curation resulted in a minimal number of uncharacterized sequences. for these genera. Therefore, we included Uniprot-unreviewed sequences and additional UniProt-classified GT-B fold enzymes in addition to the CAZy-classified GT-B sequences. Each protein sequence and structure were pared down to the Rossmann-like domains and the catalytic region. Any sequences without available AF2 structures, with AF2 structures of low confidence (residues < 70% confidence) or lacking clear Rossmann-like domains were excluded. The final sets for *Populus*, *Spirodela*, *Chlamydomonas*, and *Eucalyptus* included 308, 146, 162, and 375 enzymes respectively. After reducing the sequences and structures to the Rossmann-like domains, alignment to the training MSA and featurization, our optimized models were used to predict their potential donor substrate specificity.

3. Results

3.1. The characteristic Rossmann-like fold anchors plant GT-B structures

The GT-B fold was first described for the bacterial T4 β-glucosyltransferase,[41] and has since been found in many GT families including GT28, GT35 *etc.*[42] The characteristic aspects of the GT-B fold consist of two Rossmann fold subdomains and a loop connecting them that plausibly acts as a hinge to mediate catalysis and specificity.[42] The Rossmann fold itself is known to be one of the most ancient, prevalent and functionally diverse protein folds that involve nucleoside-based cofactors[43,44] While experimental structures for GT-Bs from non-plant systems such as bacterial and animal kingdoms have been available for some time,[6,45] structural characterization of plant GT-Bs have been less forthcoming. Recently, the Rossmann-like fold was established as one of the hallmark features for the plant fucosyltransferase from *A. thaliana* classified as a GT37.[15] With particular interest in plant-based GT-Bs and in anticipation of the challenges presented by the inherent structural diversity of GT-Bs, we analyzed the AF2 structures of all the sequences being considered in this study. We decided to parse the sequences to look for sections that correspond the characteristic Rossmann-like subdomains. Our analysis verified that almost all the sequences had characteristic Rossmann-like subdomains, with a few exceptions either too short or containing too few β-strands. This is illustrated in Fig. 2 where these domains in candidate structures from 19 distinct GT-B families are highlighted. All of them have a N-and C-terminal Rossmann-like subdomains. The non-Rossmann-like fold domains of the structures are hidden for the sake of clarity. The GT-B binding and catalytic active site are likely situated at the interface of the two Rossmann-like subdomains.

3.2. Trained models yield high accuracy on cross-validation and test-sets

The ability of the four multi-label ML models to accurately predict the correct donor substrate specificity of the training set sequences was evaluated using cross-validation and test scores. Varying the feature lengths resulted in models with very similar cross-validation scores. As such, the model feature length and corresponding hyperparameters used for further evaluation were chosen by considering optimal test scores (with <1% difference in cross-validation score from the model with best cross-validation score). All models achieved cross-validation F1 scores of at least 84%. The models showed more variance in their test scores, with scores of 43–59% for the GNB, SVC, and RF models, but up to 85% for the KNN model. The KNN model with 550 features had the best balance for cross-validation (94.2% ± 23.4%) and test scores (85.0% ± 35.7%) (Fig. 3 and ESI Table 3†). The standard deviations for the KNN model are high due to all predictions being binary. To confirm that these 550 selected features for the test set are within the applicability domain of the training set, an applicability domain analysis with the standardization approach was performed on the features from both data sets. None of the test set samples were classified as outliers with this method and the full results can be

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

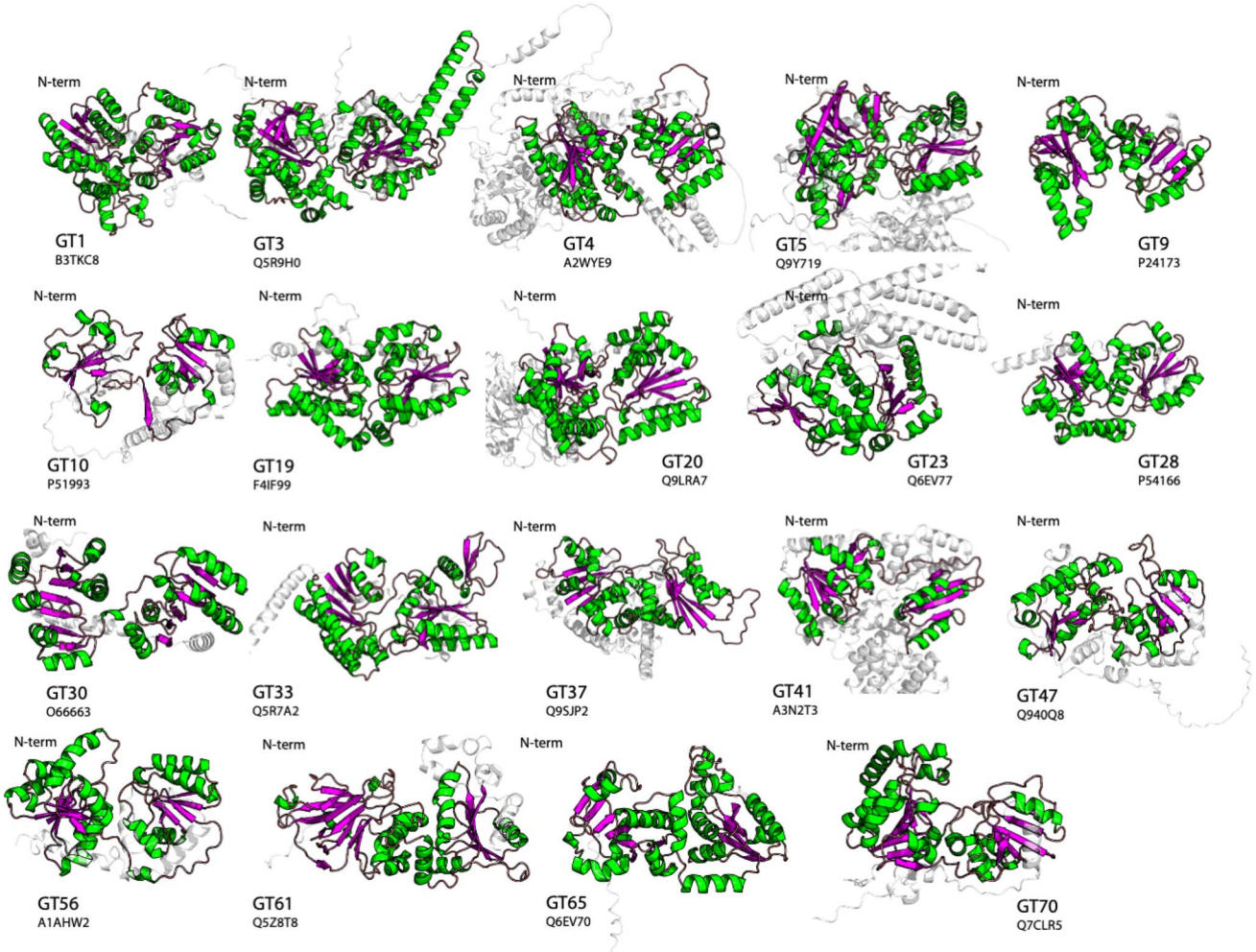

Fig. 2 Representative structures for 19 plant GT-B families represented in the dataset (rendered using PyMOL), with β-sheets shown in magenta and α-helices in green. While there is significant diversity in the structure and size of these proteins, all structures contain the characteristic Rossman-like subdomains. The structures are all aligned such that the N-terminal subdomain is on the left while the C-terminal is on the right. Sections of each structure that are not part of the Rossman-like subdomains are depicted transparently for clarity.

found in the ESI† (AD_Test_Set_Output.csv, AD_Train_Set_Output.csv).[46] The hyperparameters and feature lengths for the chosen models, with optimal cross-validation and test set scores, can be found in ESI Table 4.† As the dataset's substrate representation is imbalanced, an additional metric – the Matthew's Correlation Coefficient (MCC), was evaluated for each of the test set substrates to assess the model's performance on all labels (Fig. 3C). Finally, a confusion matrix was also generated for the test set predictions (Fig. 3D). The model achieved accuracy over 70% for seven of the eight substrates. The remaining substrate, UDP-Rha, appeared only three times in the training set, likely leading to its poorer performance. Substrates with more abundant samples in the training set, even only eight for UDP-Xyl, had much higher accuracy.

3.3. Clustering reveals unique structural motifs within GT-B families

To visualize any trends in model classifications, a dimensionality reduction was conducted for the feature vectors from the KNN model using the t-distributed Stochastic Neighbor Embedding (t-SNE) method.[47] In Fig. 3E, each dot on the plot represents a substrate classification for a test sequence by the model, with shaded regions indicating the GT family to which the sequence belongs. There is significant clustering of GT sequences binding the same substrate, which might be central to the KNN model's high accuracy. While much of the clustering is likely due to family identity, there are some distinctive clustering features. One of these is the fact that some sequences from the same family cluster in different locations based on substrate specificity as exemplified by GT28 sequences organizing into three clusters, two labelled with UDP-Glc and one with UDP-Gal. Notably, the UDP-Gal cluster is located nearer to the GT4 sequences that also have UDP-Gal specificity. The other interesting observation is that the GT1 sequences are a very dispersed cluster, likely due to great donor and acceptor diversity observed for these enzymes.[48] An analysis of the acceptor substrates, curated in a similar fashion as the donor substrates (see Section 2.1), found GT1 enzymes that use UDP-Glc as

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

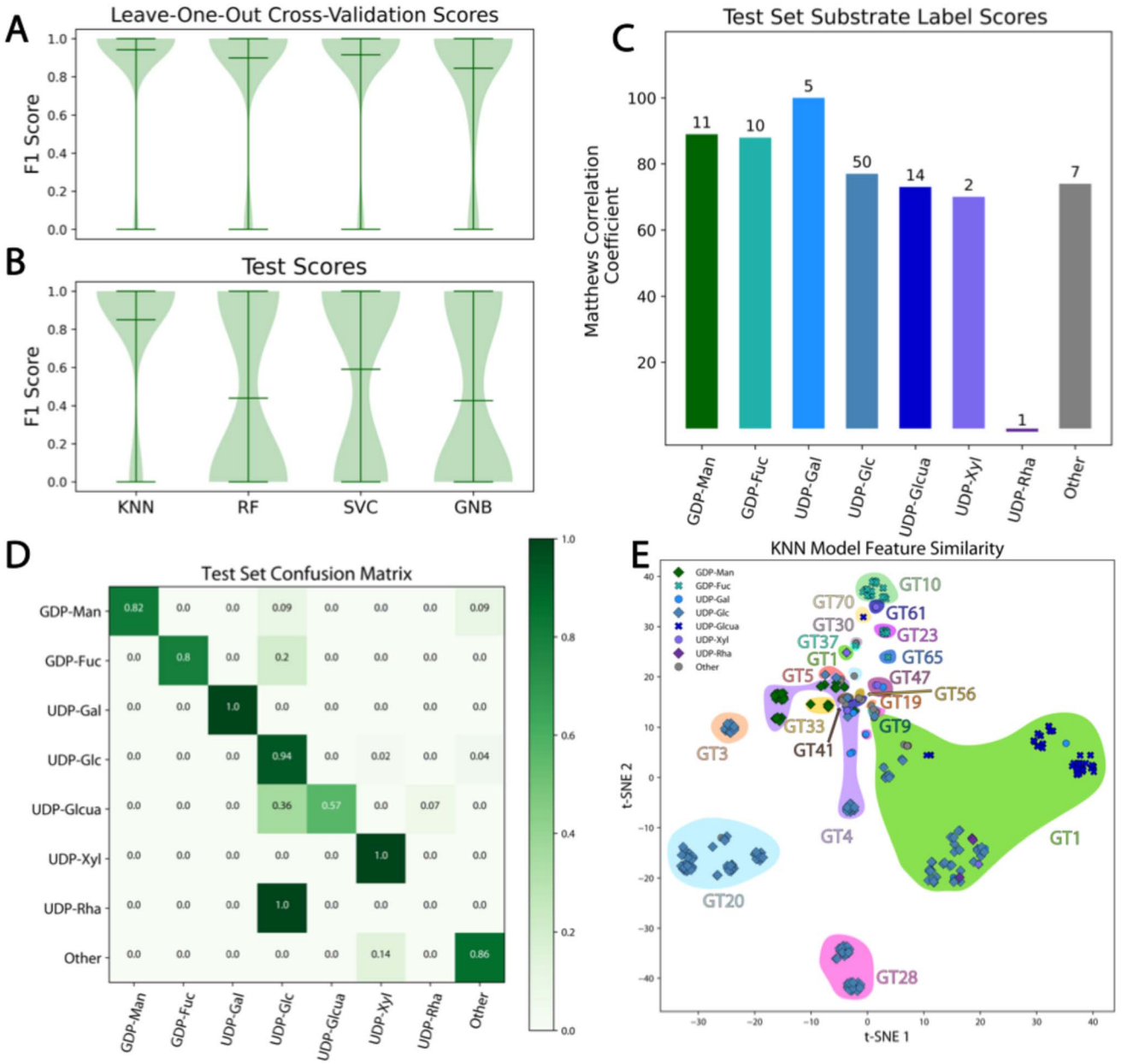

Fig. 3 (A) F1 cross-validation scores and (B) test scores indicate that KNN model performs the best with the GNB model performing the worst. (C) The Matthews correlation coefficient scores reveal substrate-level accuracies for the best performing KNN model. The number of test set samples for each substrate is indicated. (D) A confusion matrix for the test set indicates substrates that are mis-classified by the model. (E) Dimensionality reduction of the 550 selected features from the KNN model with t-distributed Stochastic Neighbor Embedding (t-SNE) reveals unique family and substrate-based clustering.

a donor bind dozens of distinct acceptors, including several different benzoxazinoids and flavonols.

Notably, the t-SNE analysis shows several UDP-Glc sequences as clustering into distinct groups within their respective families, from GT1, GT20 and GT28. This clustering was further inspected by comparing structures from each cluster with significant structural differences found between the clusters (Fig. 4). Two of the GT1 clusters (Fig. 4A(2) and A(3)) are in close proximity in the t-SNE analysis (Fig. 4A(1)), and representative structures from each cluster show high similarity with minor differences in α helix locations. Conversely, a representative structure from a more distant cluster shows much more significant structural difference (Fig. 4A(3)). A similar assessment for the GT20 clusters shows similar structural differences, with two of the clusters containing additional β strands from the third cluster (Fig. 4B(2)–B(4)). The assessment for the GT28 clusters shows more subtle differences, most significantly in shorter β strands in one of the clusters (Fig. 4C(2)–C(3)). This analysis indicates that the 550 feature t-SNE analysis is able to encode structural differences within GT families.

To ensure that the accuracy of results is not driven by high sequence similarity between the testing and training datasets,

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

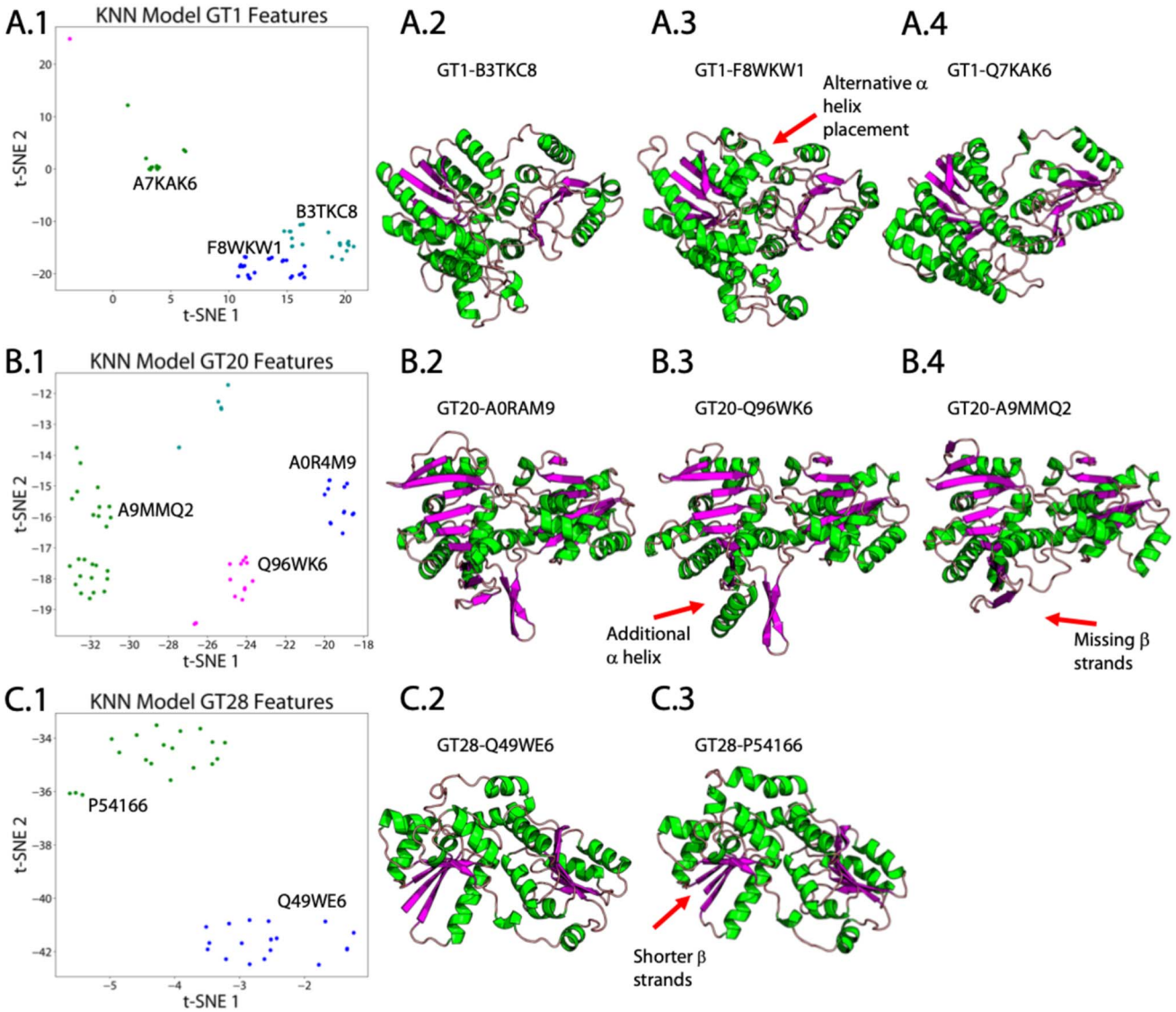

Fig. 4 (A(1), B(1) and C(1)) t-SNE plots derived from Fig. 3E focusing on specific regions is shown only for UDP-Glc binding sequences from families GT1, GT20, and GT28, respectively. (A(2)–A(4), B(2)–B(4), C(2)–C(3)) Structures from each cluster are shown with selected distinguishing structural features indicated.

we created testing data subsets that excluded proteins that share high sequence identity with the training dataset sequences and evaluated the KNN model on these subsets. Unsurprisingly, the model accuracy declines on subsets with lower shared identity. MCC scores were again calculated for each substrate in the given set and the average shown in Table 1. The overall F1 score declines at lower shared identity values but remains ~75% at only 50% shared identity suggesting that the model accuracies not due to high-sequence similarities in the training set. The average MCC score similarly declines at lower identity cutoffs, but maintains high performance on the GDP-Man, GDP-Fuc, UDP-Glc, UDP-Xyl and "Other" substrates (samples with UDP-Gal as their donor are not present at lower identity cutoffs.)

Furthermore, to ensure that our 550-feature KNN model's substrate classifications are not merely family-based, we also built single-feature models and trained them with just the family identifier as its feature (ESI Fig. 3†). This analysis verified the importance of including the complete feature set in model training, with the best test set scores of 85% for the complete feature model and only 46% for the family-based model.

3.4. Relating conserved residues to structure

To assess whether the features used by the ML model could elucidate common structural features between dissimilar families, residues involved in features used by the KNN model were evaluated for consensus. Enzymes active on UDP-Glc were chosen for this evaluation, as these enzymes comprise 204 of the 413 total sequences and represent seven distinct families. GT-B fold families GT1, GT3, GT4, GT5, GT20, GT28, and GT41 participate in UDP-Glc binding and are represented by 70, 15, 16, 2, 62, 37 and 2 sequences respectively. Conservation

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 2 The residues corresponding to the MSA positions with property consensus over 90% within sequences from GT families 4, 20 and 28, and is restricted to enzymes that utilize UDP-Glc as a donor. NC indicates that no conservation was observed in that particular structure, while other structures from the same family do have conservation at that MSA position

	MSA position													
	23	24	25	26	63	89	518	663	759	766	800	1022	1026	1028
A2WYE9 (GT4)	I196	V197	L198	I199	V222	L243	M391	M458	I517	F524	I539	L609	E613	A615
Q9LRA7 (GT20)	NC	I62	I63	V64	L94	Y106	L249	I306	I349	M356	I375	L445	NC	I451
P54166 (GT28)	V7	L8	I9	L10	V19	NC	M134	I179	V218	L225	V236	I284	E288	NC

amongst residues contributing to the 550 substrate-specificity determining features of the KNN model was assessed with a 90% cutoff within each family as well as amongst all families. While there were only two conserved residues between all seven families, higher conservation could be found between subsets of families. We chose to focus conservation analysis on GT4, 20 and 28, as they have abundant UDP-Glc specific samples, with 14 residues conserved in each family over 90% (Table 2). Fig. 5A depicts a Venn diagram that shows the number of conserved residues within each family, between any two families and amongst all three families. For example, within 16 GT4 family enzymes that bind UDP-Glc, there are 68 positions conserved in

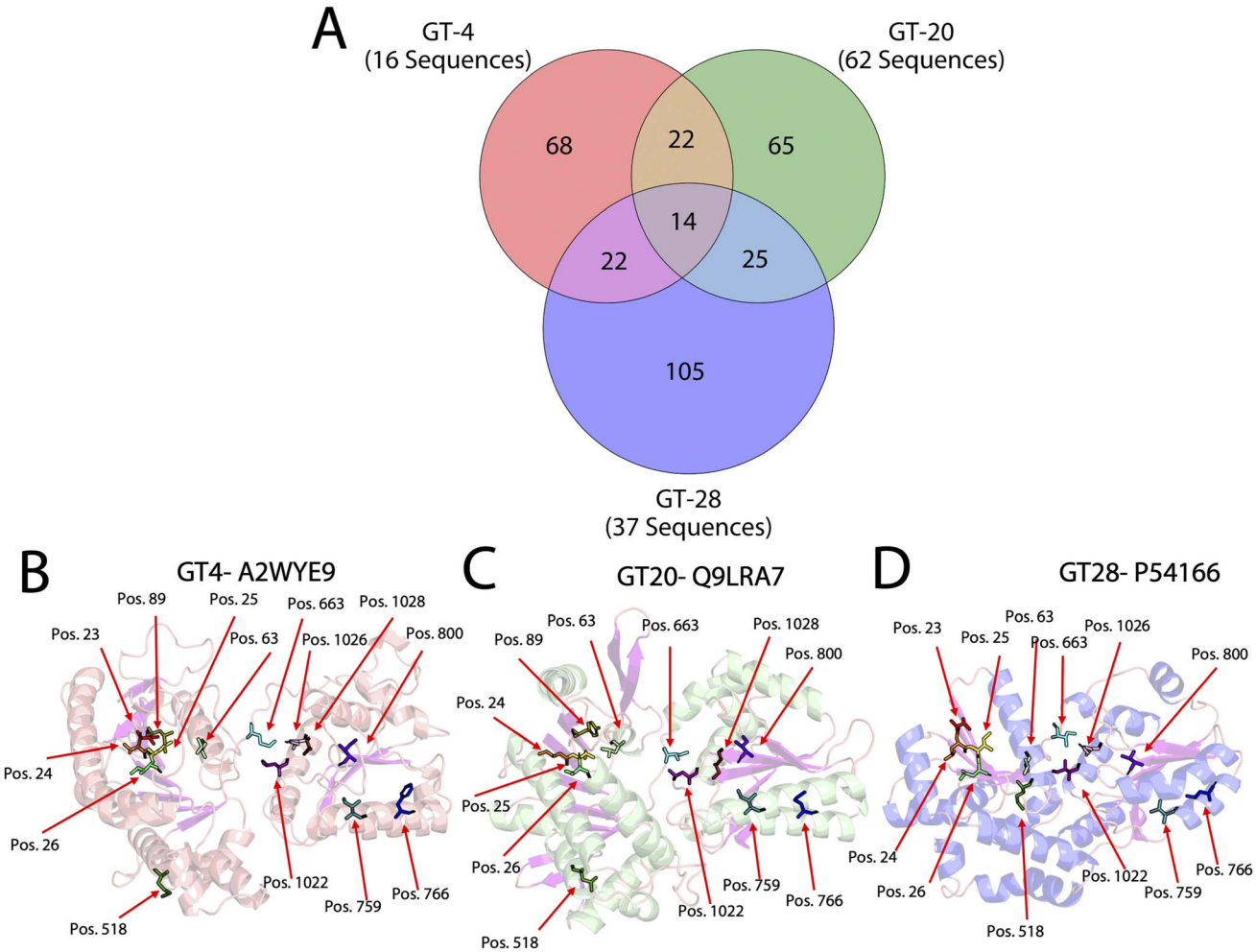

Fig. 5 (A) Venn diagram showing the distribution of residues (MSA positions) that are part of the 550 features used by the KNN model and are involved in detecting substrate specificities for GT4, GT20 and GT28 families constituting 16, 62 and 37 sequences respectively. The overlapped regions indicate the portion of residues (MSA positions) that are commonly conserved amongst any two and all three families for the same property. (B–D) The α-helices are color-coded based on the Venn diagram with the GT4 structures depicted in salmon, GT20 in pale green and GT28 structures in blue. (B–D) Structural regions used in model training are shown with the 14 conserved MSA positions common to all three families depicted in licorice representations. 14, 12, and 12 of these positions are conserved in the example GT4, GT20 and GT28 structures respectively.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

the 550 features used by the KNN model. A similar analysis of 62 GT20 sequences that bind UDP-Glc reveals 65 conserved positions that belong to the KNN feature set. Similarly, amongst the 37 GT28 sequences, there are 105 conserved MSA positions.

Since the GT4, GT20 and GT28 families share UDP-Glc as a common nucleotide sugar donor substrate and exhibit structural similarity in their Rossmann domains, it is not far-fetched to imagine that the 14 residue positions within the KNN features with consensus above 90% in each family might reveal common structural motifs responsible for the selectivity for this substrate (Fig. 5A). Residues corresponding to these positions are listed in Table 2 for a representative structure from each family. Unsurprisingly, these residues exhibit similar placement in all three structures (Fig. 5B–D).

The direct involvement of these conserved residues in substrate binding was investigated by docking UDP-Glc to each of these enzymes using Autodock Vina. As expected, all docked poses (shown in ESI Fig. 4†) for the three candidate enzymes were observed to be at the interface between the N-terminal and C-terminal Rossmann-like subdomains, which contain the presumed active site. Only one position in the GT20 structure and two positions in the GT28 structure are observed to be at the putative substrate binding site. Therefore, a significant relationship between the conserved residues and their role in substrate binding could not be established.

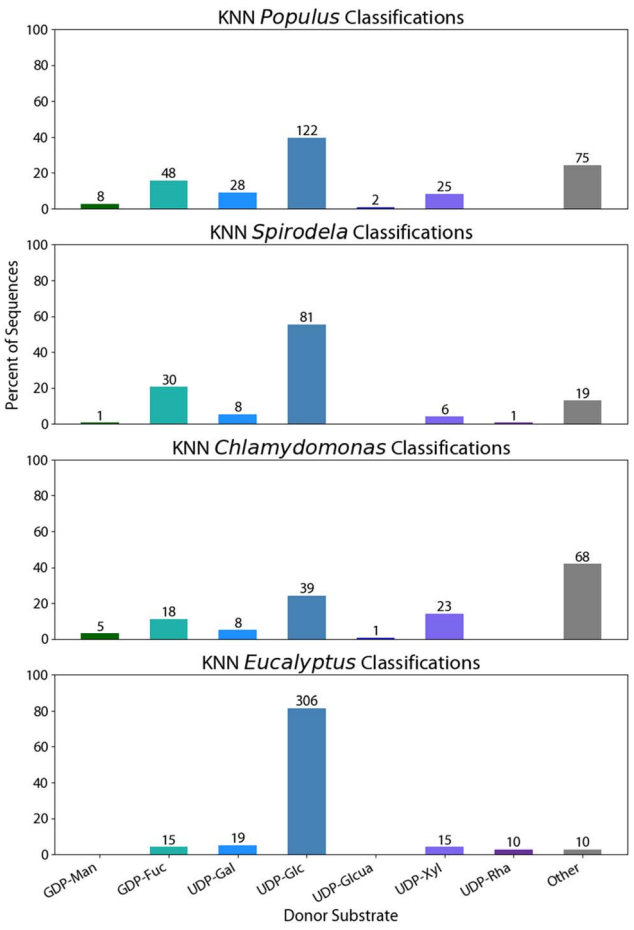

Fig. 6 Application of the optimized KNN model to uncharacterized GT-B fold enzymes in diverse plants. Donor substrate classifications for the uncharacterized proteins from *Populus*, *Spirodela*, *Chlamydomona* and *Eucalyptus* are shown as the fraction of total sequences in the species dataset.

3.5. Model extension to uncharacterized sequences from other plant genera

Finally, the optimized KNN model was applied to uncharacterized GT-B fold enzyme sequences from distinct plant genera to predict their substrate specificities, including *Populus*, *Spirodela*, *Chlamydomonas*, and *Eucalyptus* (Fig. 6). The datasets contain 308, 146, 162, and 375 enzyme sequences, respectively. These genera were chosen in part due to their distinct physiology, a consequence of different carbohydrate composition profiles of their cell walls. The species-specific distribution of families for this dataset is listed in ESI Table 5.† While the KNN model can generate predictions for multiple substrates, no enzymes in any of the sets were predicted to be substrate promiscuous. The predictions for each genera reflect similar distributions of substrates, with a notable difference in the number of "Other" substrates predicted. An abundance of sequences in the *Populus* and *Chlamydomonas* sets receive the "Other" classification, while almost all sequences in the *Eucalyptus* set are classified as specific for one of the seven represented donors. The *Eucalyptus* set also has a much higher percentage of samples classified as specific to UDP-Glc than the other sets.

4. Discussion

Understanding the substrate scope and specificity of glycosyl transferases is crucial for our understanding of natural biosynthetic mechanisms of carbohydrate polymer synthesis. In this work, we develop and evaluate the efficacy of ML classifier models for the prediction of nucleotide donor substrates of uncharacterized GT-B fold enzymes. A major challenge we encountered was the curation of existing experimental data – namely the annotation of sequences and their activity for relevant nucleotide-sugar donor substrates. We collated an extensive dataset of GT-B sequences with experimentally characterized activities on seven donor nucleotide sugars. This data was used to train four different classifier models, amongst which the KNN classifier demonstrated the best performance for predictions on the training set, while also being generalizable to the test set. The feature vectors used by the KNN model formed the basis for further dimensionality reduction and residue conservation analyses. The dimensionality reduction analysis revealed both substrate-based and family-based clustering of sequences. We also noticed multiple sub-clusters for sequences from the same family (GT1, GT20 and GT28) and with activity on the same substrate (UDP-Glc). Structural analysis of sequences within these sub-clusters revealed subtle differences and indicated unique structural motifs for these subclusters. Furthermore, we assessed the KNN model's feature-set for its ability to relate protein sequence to conserved structural regions and binding sites. While some limited

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

structural consensus was revealed, our analysis did not reveal a strong relationship between the suggested binding sites and regions of high consensus.

Previous work has successfully used similar substrate activity classifiers, individually or as ensembles, for a variety of enzyme classes like bacterial nitrilases, thiolases and the GT1 family which adopts the GT-B fold.[23,49,50] A general enzyme prediction model has also been developed using neural networks. However, its accuracy notably declines on substrates not well represented in its training set, such as the donor substrates common to GT-Bs, demonstrating the need for more specific enzyme class models.[51] Thus, RF, SVC and KNN models were applied to our dataset of 413 samples, with an additional challenge presented by these samples being derived from highly dissimilar families adopting the GT-B fold. While enzymes from these families share a distinct structural Rossmann fold motif, they exhibit high sequence dissimilarity. Therefore, we incorporated structural data as well, through both reducing sequences to only the Rossmann domains and featurizing AF2 structures for solvent secondary surface area and secondary structure. Notably, an additional finding of this work is the inclusion of SASA and secondary structure features made little difference in test accuracy (ESI Fig. 5†). Liu *et al.* made a similar discovery in their work on residue pK_a prediction, where they also found SASA data to make little difference in accuracy.[52] A potential reason for this lack of improvement may be that the AF2 structural data is a function of sequence and is already incorporated into the feature data. Additional features instead derived from molecular dynamics simulations and docking simulations, such as RMSF and binding affinity data, may provide necessary physics-based information for substrate classification.[50,53]

A phylogenetic analysis reveals one of the inherent complexities involved in resolving substrate-specificity determining factors in GT-Bs. Fig. 7A shows the phylogenetic tree as generated from the full-length sequence alignment of all 413 training sequences using the NJ method.[54] There is a clear family-based ordering of the sequences in this phylogenetic tree. However, an overlay of the substrate specificities on the same tree (Fig. 7B) reveals that there are families that act on multiple substrates (*e.g.*: GT1), and there are multiple families that act on the same substrate (*e.g.* UDP-Glc being acted on by GT1s, GT4s and GT20s). This clearly is consistent with experimental data that substrate-specificity is not family-based among GTs. Other possible reasons for lack of a relationship between conserved residues and binding sites include allosteric effects from these regions, or simply that common features among these proteins are not necessarily related to binding. Alternative

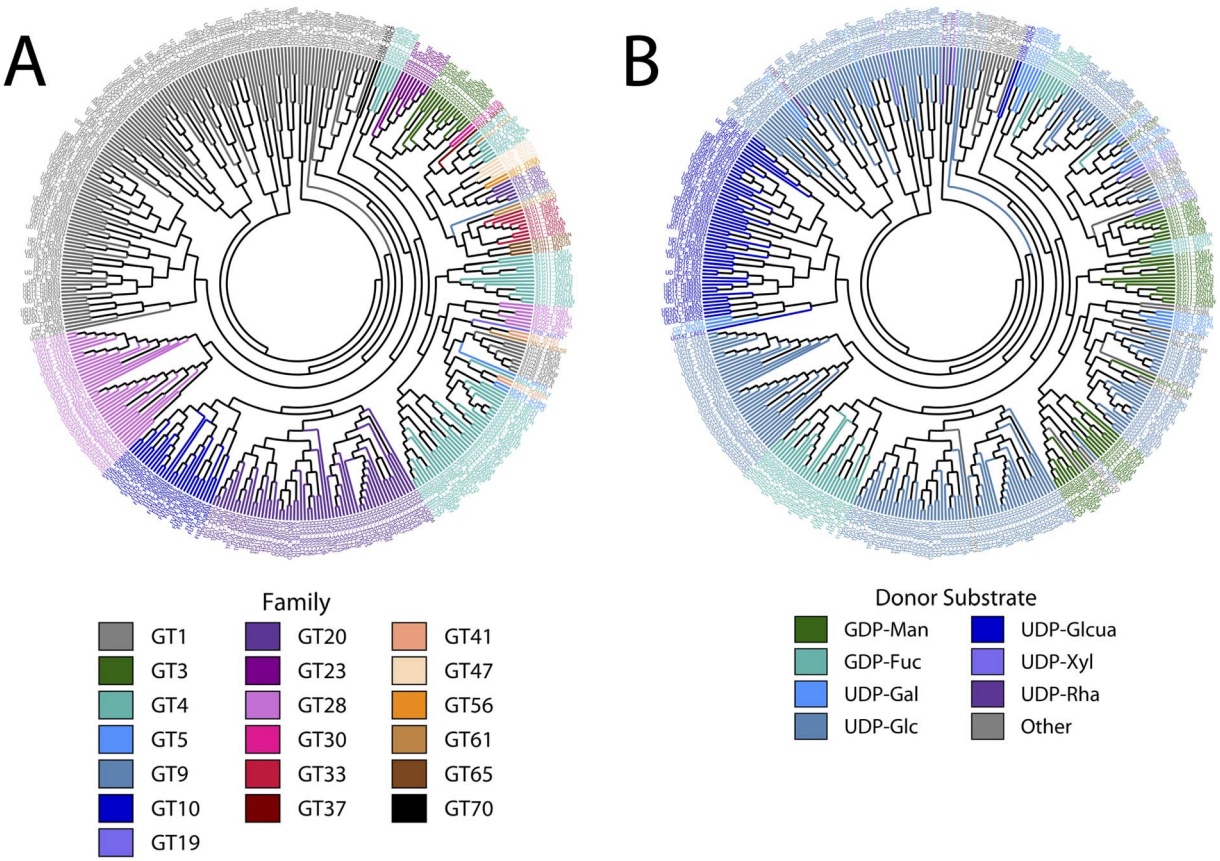

Fig. 7 A phylogenetic tree constructed from the 413 full-length training sequences. (A) shows this tree overlayed with a single donor substrate for each sequence. (B) shows this tree overlayed with the GT family identification for each sequence.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

ML model frameworks which more directly incorporate structural information, such as graph neural networks, may have more success in correlating the identified important residues with the substrate binding site.[55]

A further limitation of this work concerns the training data quality. Many of the training set sequences may in fact have activity on several additional donor substrates, but this has not been experimentally studied and therefore cannot be included in the substrate labels. The enzymes in this dataset had only one substrate listed on the database, but many may very well be less selective polyspecific enzymes. Despite this limitation, most enzymes are listed with the substrate used for their primary function(s), and so the KNN model would remain accurate in predicting the substrates for which the enzyme has the most activity.

While the models trained in this work focus exclusively on donor substrate prediction, extending these models for acceptor prediction is an important consideration for future work. However, this portends some considerable challenges including, due to inconsistent acceptor substrate labelling on databases and the ability of GTs to act on acceptor substrates with varying architectures.[56]

5. Conclusion

GT-B substrate specificity has been a challenge to characterize for diverse families due to a lack of shared sequence similarity, limited curated substrate specificity data and a paucity of structural information. This work presents the first in-depth effort on using a data-driven approach to elucidate the substrate-specificity dictating features in GT-Bs. Our breakthrough approach involved curating the first database of GT-B sequences with experimentally characterized donor activities. We then established that the Rossmann fold domain anchors all GT-B structures and used it as the basis to build effective multiple sequence alignments for this highly diverse fold of enzymes. To account for the fact that many GT-Bs are polyspecific (*i.e.* have activities on different donor substrates), our ML models were built as multi-label classifiers. The models were trained on both features from the protein sequence as well as AF2 predicted structures. While our study demonstrates that (1) current classifier ML models may be adapted to include structural data on these enzymes and (2) can predict substrate specificities for GT-Bs with reasonable accuracies but (3) interpretability does not allow for direct elucidation of structural features and (4) they are still severely limited by the paucity of curated biochemical and structural data on these enzymes. In the future, we envisage the development of other ML approaches *e.g.* large-language models and foundational models for these predictive tasks that might be more adept at learning from a larger protein sequence and structural space, and which may be adapted specifically to the nature of available characterized experimental and physics-based modelling data on these enzymes.

Data availability

The code and datasets used for the GT-B sequence substrate specificity predictor tool (GTBPredict) have been archived and made available to the public at the following link: **https://github.com/vbharadwcomosci/GTB_substrate_prediction**. The code and the associated datasets have also been hosted on Zenodo with the following DOI: **https://doi.org/10.5281/zenodo.15786055**.

Author contributions

VB conceptualized the project, assisted in model development and data-analysis and visualization. SH conceptualized the project, curated the data and developed the ML models, performed data-analysis and visualization. VB, SH, YB and BU wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. This work was 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. This research was supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research, Genomic Science Program grant no. DE-SC0023223. We are also grateful for the access and use of NREL's Computational Sciences' resources (Eagle) supported by the DOE Office of EERE under contract no. DE-AC36-08GO28308. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

1 M. C. McCann and N. C. Carpita, Biomass recalcitrance: a multi-scale, multi-factor, and conversion-specific property, *J. Exp. Bot.*, 2015, **66**(14), 4109–4118, DOI: **10.1093/jxb/erv267**.

2 P. J. Smith, H.-T. Wang, W. S. York, M. J. Peña and B. R. Urbanowicz, Designer biomass for next-generation biorefineries: leveraging recent insights into xylan structure and biosynthesis, *Biotechnol. Biofuels*, 2017, **10**(1), 286, DOI: **10.1186/s13068-017-0973-z**.

3 P. J. Smith, M. E. Ortiz-Soto, C. Roth, W. J. Barnes, J. Seibel, B. R. Urbanowicz and F. Pfrengle, Enzymatic Synthesis of Artificial Polysaccharides, *ACS Sustain. Chem. Eng.*, 2020, **8**(32), 11853–11871, DOI: **10.1021/acssuschemeng.0c03622**.

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

4 H.-T. Wang, V. S. Bharadwaj, J.-Y. Yang, T. M. Curry, K. W. Moremen, Y. J. Bomble and B. R. Urbanowicz, Rational enzyme design for controlled functionalization of acetylated xylan for cell-free polymer biosynthesis, *Carbohydr. Polym.*, 2021, **273**, 118564, DOI: **10.1016/j.carbpol.2021.118564**.

5 C. Breton, L. Šnajdrová, C. Jeanneau, J. Koča and A. Imberty, Structures and mechanisms of glycosyltransferases, *Glycobiology*, 2005, **16**(2), 29R–37R, DOI: **10.1093/glycob/cwj016**.

6 L. L. Lairson, B. Henrissat, G. J. Davies and S. G. Withers, Glycosyltransferases: Structures, Functions, and Mechanisms, *Annu. Rev. Biochem.*, 2008, **77**(1), 521–555, DOI: **10.1146/annurev.biochem.76.061005.092322**.

7 E. Drula, M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat and N. Terrapon, The carbohydrate-active enzyme database: functions and literature, *Nucleic Acids Res.*, 2021, **50**(D1), D571–D577, DOI: **10.1093/nar/gkab1045**.

8 J. Rocha, F. Cicéron, D. de Sanctis, M. Lelimousin, V. Chazalet, O. Lerouxel and C. Breton, Structure of Arabidopsis thaliana FUT1 Reveals a Variant of the GT-B Class Fold and Provides Insight into Xyloglucan Fucosylation, *Plant Cell*, 2016, **28**(10), 2352–2364, DOI: **10.1105/tpc.16.00519**.

9 I. Hanukoglu, Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites, *Biochem. Mol. Biol. Educ.*, 2015, **43**(3), 206–209, DOI: **10.1002/bmb.20849**.

10 M. Madson, C. Dunand, X. Li, R. Verma, G. F. Vanzin, J. Caplan, D. A. Shoue, N. C. Carpita and W.-D. Reiter, The MUR3 Gene of Arabidopsis Encodes a Xyloglucan Galactosyltransferase That Is Evolutionarily Related to Animal Exostosins, *Plant Cell*, 2003, **15**(7), 1662–1670, DOI: **10.1105/tpc.009837**.

11 S.-J. Kim, B. Chandrasekar, A. C. Rea, L. Danhof, S. Zemelis-Durfee, N. Thrower, Z. S. Shepard, M. Pauly, F. Brandizzi and K. Keegstra, The synthesis of xyloglucan, an abundant plant cell wall polysaccharide, requires CSLC function, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**(33), 20316–20324, DOI: **10.1073/pnas.2007245117**.

12 L. Zhang, P. K. Prabhakar, V. S. Bharadwaj, Y. J. Bomble, M. J. Peña and B. R. Urbanowicz, Glycosyltransferase family 47 (GT47) proteins in plants and animals, *Essays Biochem.*, 2023, **67**(3), 639–652, DOI: **10.1042/EBC20220152**.

13 K. R. Acharya and M. D. Lloyd, The advantages and limitations of protein crystal structures, *Trends Pharmacol. Sci.*, 2005, **26**(1), 10–14, DOI: **10.1016/j.tips.2004.10.011**.

14 A. McPherson and J. A. Gavira, Introduction to protein crystallization, *Acta Crystallogr., Sect. F:Struct. Biol. Commun.*, 2014, **70**(1), 2–20, DOI: **10.1107/S2053230X13033141**.

15 B. R. Urbanowicz, V. S. Bharadwaj, M. Alahuhta, M. J. Peña, V. V. Lunin, Y. J. Bomble, S. Wang, J.-Y. Yang, S. T. Tuomivaara, M. E. Himmel, *et al.*, Structural, mutagenic and in silico studies of xyloglucan fucosylation in Arabidopsis thaliana suggest a water-mediated mechanism, *Plant J.*, 2017, **91**(6), 931–949, DOI: **10.1111/tpj.13628**.

16 V. V. Lunin, H.-T. Wang, V. S. Bharadwaj, M. Alahuhta, M. J. Peña, J.-Y. Yang, S. A. Archer-Hartmann, P. Azadi, M. E. Himmel, K. W. Moremen, *et al.*, Molecular Mechanism of Polysaccharide Acetylation by the Arabidopsis Xylan O-acetyltransferase XOAT1, *Plant Cell*, 2020, **32**(7), 2367–2382, DOI: **10.1105/tpc.20.00028**.

17 P. K. Prabhakar, J. H. Pereira, R. Taujale, W. Shao, V. S. Bharadwaj, D. Chapla, J.-Y. Yang, Y. J. Bomble, K. W. Moremen, N. Kannan, *et al.*, Structural and biochemical insight into a modular β-1,4-galactan synthase in plants, *Nat. Plants*, 2023, **9**(3), 486–500, DOI: **10.1038/s41477-023-01358-4**.

18 C. Davis, K. Kota, V. Baldhandapani, W. Gong, S. Abubucker, E. Becker, J. Martin, K. M. Wylie, R. Khetani, M. E. Hudson, *et al.*, mBLAST: Keeping up with the sequencing explosion for (meta)genome analysis, *J. Data Min. Genomics Proteomics*, 2015, **4**(3), DOI: **10.4172/2153-0602.1000135**.

19 Y. Xu, D. Verma, R. P. Sheridan, A. Liaw, J. Ma, N. M. Marshall, J. McIntosh, E. C. Sherer, V. Svetnik and J. M. Johnston, Deep Dive into Machine Learning Models for Protein Engineering, *J. Chem. Inf. Model.*, 2020, **60**(6), 2773–2790, DOI: **10.1021/acs.jcim.0c00073**.

20 S. Goldman, R. Das, K. K. Yang and C. W. Coley, Machine learning modeling of family wide enzyme-substrate specificity screens, *PLoS Comput. Biol.*, 2022, **18**(2), e1009853, DOI: **10.1371/journal.pcbi.1009853**.

21 M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, *Nucleic Acids Res.*, 2021, **50**(D1), D439–D444, DOI: **10.1093/nar/gkab1061**.

22 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, Highly accurate protein structure prediction with AlphaFold, *Nature*, 2021, **596**(7873), 583–589, DOI: **10.1038/s41586-021-03819-2**.

23 M. Yang, C. Fehl, K. V. Lees, E.-K. Lim, W. A. Offen, G. J. Davies, D. J. Bowles, M. G. Davidson, S. J. Roberts and B. G. Davis, Functional and informatics analysis enables glycosyltransferase activity prediction, *Nat. Chem. Biol.*, 2018, **14**(12), 1109–1117, DOI: **10.1038/s41589-018-0154-9**.

24 R. Taujale, Z. Zhou, W. Yeung, K. W. Moremen, S. Li and N. Kannan, Mapping the glycosyltransferase fold landscape using interpretable deep learning, *Nat. Commun.*, 2021, **12**(1), 5656, DOI: **10.1038/s41467-021-25975-9**.

25 B. Ma and R. Nussinov, Enzyme dynamics point to stepwise conformational selection in catalysis, *Curr. Opin. Chem. Biol.*, 2010, **14**(5), 652–659, DOI: **10.1016/j.cbpa.2010.08.012**.

26 T. U. Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Res.*, 2020, **49**(D1), D480–D489, DOI: **10.1093/nar/gkaa1100**.

27 D. M. Brown, Z. Zhang, E. Stephens, P. Dupree and S. R. Turner, Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in Arabidopsis, *Plant J.*, 2009, **57**(4), 732–746, DOI: **10.1111/j.1365-313X.2008.03729.x**.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

28 J. K. Jensen, A. Schultink, K. Keegstra, C. G. Wilkerson and M. Pauly, RNA-Seq Analysis of Developing Nasturtium Seeds (Tropaeolum majus): Identification and Characterization of an Additional Galactosyltransferase Involved in Xyloglucan Biosynthesis, *Mol. Plant*, 2012, **5**(5), 984–992, DOI: 10.1093/mp/sss032.

29 J. K. g. Jensen, S. O. Sørensen, J. Harholt, N. Geshi, Y. Sakuragi, I. Møller, J. Zandleven, A. J. Bernal, N. B. Jensen, C. Sørensen, *et al.*, Identification of a Xylogalacturonan Xylosyltransferase Involved in Pectin Biosynthesis in Arabidopsis, *Plant Cell*, 2008, **20**(5), 1289–1302, DOI: 10.1105/tpc.107.050906.

30 Y. Wu, M. Williams, S. Bernard, A. Driouich, A. M. Showalter and A. Faik, Functional Identification of Two Nonredundant Arabidopsis α(1,2)Fucosyltransferases Specific to Arabinogalactan Proteins, *J. Biol. Chem.*, 2010, **285**(18), 13638–13645, DOI: 10.1074/jbc.M110.102715.

31 L. Shu, H. Xu and B. Liu, Unseen Class Discovery In Open-World Classification, *arXiv*, 2018, preprint, arXiv.1801.05609, DOI: 10.48550/arXiv.1801.05609.

32 *The PyMOL Molecular Graphics System, Version 2.6*, 2015, Schrödinger, LLC.

33 F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Mol. Syst. Biol.*, 2011, **7**(1), 539, DOI: 10.1038/msb.2011.75.

34 S. Kawashima, H. Ogata and M. Kanehisa, AAindex: Amino Acid Index Database, *Nucleic Acids Res.*, 1999, **27**(1), 368–369, DOI: 10.1093/nar/27.1.368.

35 P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, 2009, **25**(11), 1422–1423, DOI: 10.1093/bioinformatics/btp163.

36 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**(null), 2825–2830.

37 N. Elssied, A. P. D. O. Ibrahim and A. Hamza Osman, A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification, *Res. J. Appl. Sci., Eng. Technol.*, 2014, **7**, 625–638, DOI: 10.19026/rjaset.7.299.

38 G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, Diffdock: Diffusion steps, twists, and turns for molecular docking, *arXiv*, 2022, preprint, arXiv:2210.01776, DOI: 10.1002/arXiv:2210.01776.

39 O. Trott and A. J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.*, 2010, **31**(2), 455–461, DOI: 10.1002/jcc.21334.

40 J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings, *J. Chem. Inf. Model.*, 2021, **61**(8), 3891–3898, DOI: 10.1021/acs.jcim.1c00203.

41 A. Vrielink, W. Rüger, H. Driessen and P. S. Freemont, Crystal structure of the DNA modifying enzyme beta-glucosyltransferase in the presence and absence of the substrate uridine diphosphoglucose, *EMBO J.*, 1994, **13**(15), 3413–3422.

42 Y. Bourne and B. Henrissat, Glycoside hydrolases and glycosyltransferases: families and functional modules, *Curr. Opin. Struct. Biol.*, 2001, **11**(5), 593–600, DOI: 10.1016/S0959-440X(00)00253-0.

43 P. Laurino, Á. Tóth-Petróczy, R. Meana-Pañeda, W. Lin, D. G. Truhlar and D. S. Tawfik, An Ancient Fingerprint Indicates the Common Ancestry of Rossmann-Fold Enzymes Utilizing Different Ribose-Based Cofactors, *PLoS Biol.*, 2016, **14**(3), e1002396, DOI: 10.1371/journal.pbio.1002396.

44 K. E. Medvedev, L. N. Kinch, R. Dustin Schaeffer, J. Pei and N. V. Grishin, A Fifth of the Protein World: Rossmann-like Proteins as an Evolutionarily Successful Structural unit, *J. Mol. Biol.*, 2021, **433**(4), 166788, DOI: 10.1016/j.jmb.2020.166788.

45 S. Grizot, M. Salem, V. Vongsouthi, L. Durand, F. Moreau, H. Dohi, S. Vincent, S. Escaich and A. Ducruix, Structure of the Escherichia coli Heptosyltransferase WaaC: Binary Complexes with ADP AND ADP-2-deoxy-2-fluoro Heptose, *J. Mol. Biol.*, 2006, **363**(2), 383–394, DOI: 10.1016/j.jmb.2006.07.057.

46 K. Roy, S. Kar and P. Ambure, On a simple approach for determining applicability domain of QSAR models, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 22–29, DOI: 10.1016/j.chemolab.2015.04.013.

47 L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**(11), 2579–2605.

48 K. Yonekura-Sakakibara and K. Hanada, An evolutionary view of functional diversity in family 1 glycosyltransferases, *Plant J.*, 2011, **66**(1), 182–193, DOI: 10.1111/j.1365-313X.2011.04493.x.

49 D. Banerjee, M. A. Jindra, A. J. Linot, B. F. Pfleger and C. D. Maranas, EnZymClass: Substrate specificity prediction tool of plant acyl-ACP thioesterases based on ensemble learning, *Curr. Res. Biotechnol.*, 2022, **4**, 1–9, DOI: 10.1016/j.crbiot.2021.12.002.

50 Z. Mou, J. Eakes, C. J. Cooper, C. M. Foster, R. F. Standaert, M. Podar, M. J. Doktycz and J. M. Parks, Machine learning-based prediction of enzyme substrate scope: Application to bacterial nitrilases, *Proteins: Struct., Funct., Bioinf.*, 2021, **89**(3), 336–347, DOI: 10.1002/prot.26019.

51 A. Kroll, S. Ranjan, M. K. M. Engqvist and M. J. Lercher, A general model to predict small molecule substrates of enzymes based on machine and deep learning, *Nat. Commun.*, 2023, **14**(1), 2787, DOI: 10.1038/s41467-023-38347-2.

52 S. Liu, Q. Yang, L. Zhang and S. Luo, Accurate Protein pKa Prediction with Physical Organic Chemistry Guided 3D Protein Representation, *J. Chem. Inf. Model.*, 2024, **64**(11), 4410–4418, DOI: 10.1021/acs.jcim.4c00354.

53 N. A. E. Venanzi, A. Basciu, A. V. Vargiu, A. Kiparissides, P. A. Dalby and D. Dikicioglu, Machine Learning

Integrating Protein Structure, Sequence, and Dynamics to Predict the Enzyme Activity of Bovine Enterokinase Variants, *J. Chem. Inf. Model.*, 2024, **64**(7), 2681–2694, DOI: **10.1021/acs.jcim.3c00999**.

54 N. Saitou and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 1987, **4**(4), 406–425, DOI: **10.1093/oxfordjournals.molbev.a040454**.

55 V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, *et al.*, Structure-based protein function prediction using graph convolutional networks, *Nat. Commun.*, 2021, **12**(1), 3168, DOI: **10.1038/s41467-021-23303-9**.

56 A. Biswas and M. Thattai, Promiscuity and specificity of eukaryotic glycosyltransferases, *Biochem. Soc. Trans.*, 2020, **48**(3), 891–900, DOI: **10.1042/BST20190651**.

Open Access Article. Published on 04 July 2025. Downloaded on 11/22/2025 6:40:15 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

© 2025 The Author(s). Published by the Royal Society of Chemistry