nature chemical biology

ARTICLES

https://doi.org/10.1038/s41589-018-0154-9

Functional and informatics analysis enables glycosyltransferase activity prediction

Min Yang[1,6,7], Charlie Fehl[1,7], Karen V. Lees[2], Eng-Kiat Lim[3], Wendy A. Offen[4], Gideon J. Davies[4], Dianna J. Bowles[3], Matthew G. Davidson[5], Stephen J. Roberts[2] and Benjamin G. Davis[1]*

The elucidation and prediction of how changes in a protein result in altered activities and selectivities remain a major challenge in chemistry. Two hurdles have prevented accurate family-wide models: obtaining (i) diverse datasets and (ii) suitable parameter frameworks that encapsulate activities in large sets. Here, we show that a relatively small but broad activity dataset is sufficient to train algorithms for functional prediction over the entire glycosyltransferase superfamily 1 (GT1) of the plant *Arabidopsis thaliana*. Whereas sequence analysis alone failed for GT1 substrate utilization patterns, our chemical–bioinformatic model, GT-Predict, succeeded by coupling physicochemical features with isozyme-recognition patterns over the family. GT-Predict identified GT1 biocatalysts for novel substrates and enabled functional annotation of uncharacterized GT1s. Finally, analyses of GT-Predict decision pathways revealed structural modulators of substrate recognition, thus providing information on mechanisms. This multifaceted approach to enzyme prediction may guide the streamlined utilization (and design) of biocatalysts and the discovery of other family-wide protein functions.

Subtle evolutionary divergence within a protein family enables an enormous breadth of functional activities to occur within a versatile core scaffold[1,2]. The reutilization of common scaffolds in the design of de novo protein functions is also a current major goal. Several large architecturally related protein families are known, among which the group-transfer-enzyme proteins are of particular interest, because several use multiple modular domains upon which relevant functional groups are evolutionarily selected[1]. Multiple group-transfer-enzyme superfamilies, including certain acetyltransferases and glycosyltransferases (GTs), share a conserved β-sheet/α-helical core upon which they exploit variable domains to generate selectivity toward (in some cases thousands of) substrates[3,4]. Some have binding sites that are readily understood by virtue of their narrow substrate range (for example, the lysine acetyltransferases that necessarily bind acetyl CoA and lysine) and hence are tractable to accurate substrate prediction[5]. In contrast, GTs represent the other extreme, in that their activities in vitro unite highly variable substrates, and phylogenetic analyses have provided only limited insight into the evolution of substrate recognition and specificity[6,7]. This lack of insight is despite the high scaffold conservation among GTs[8], which has been exploited in only select examples[9], therefore suggesting that subtle mutations in the background of these scaffolds have profound effects on chemical function. Thus, there remains a general difficulty in understanding the basis for active site plasticity within many enzyme families[10], and GTs in particular represent a striking example of this limitation to understanding, which is exacerbated by a dearth of solved three-dimensional structures[11]. This example is made all the more pertinent by the existence of an excellent database for GTs in the carbohydrate–active enzymes database (CAZy);[4] indeed, the curators of CAZy have highlighted functional prediction as an important future goal[4].

As a primary hurdle, there is currently no general informatics strategy to accurately assess the functional effects of changes between key features of otherwise similar isoforms of biocatalysts in a manner equivalent, for example, to strategies to model and predict subtle stereoelectronic effects in homogeneous small-molecule-catalyst performance[12]. Notably, de novo protein-design methods, although powerfully enabling the creation of rigid structural scaffolds for housing putative function, still fail regarding the finer details associated with the positioning of key catalytic residues[13]. Therefore, bridging this gap between the prediction and structure of precise active site features might yield valuable additional insight into the discovery of desired protein functional activities.

Here, we show that functional profiling (Fig. 1) using broad, unbiased sampling methods of a full GT family present in a single species (the 107-member GT1 family of the plant *A. thaliana*) enables construction of chemical–bioinformatic models that encapsulate family-wide recognition patterns for both electrophilic sugar-donor and nucleophilic acceptor substrates. We observed extreme scattering in activity patterns, as scored by phylogenetic linkage analysis alone, thus confirming that sequence-based assessments cannot explain substrate recognition. However, by incorporating relevant physicochemical parameters such as size, hydrophobicity, and nucleophilicity, predictive algorithms can be trained to annotate function with high accuracy for these promiscuous dual-substrate enzymes.

Results

Strategy for functional profiling of an enzyme superfamily. To date, informatics or computational strategies for predicting GT1 enzyme activity have made only limited progress, as further exacerbated by the limited number of solved three-dimensional structures[11]. High-confidence phylogenetic trees for a complete GT1 family were previously reported by some of us[6], wherein a limited set of substrates was tested for common activity. Little correlation was found between primary sequence alignment and enzymatic

[1]Chemistry Research Laboratory, Oxford University, Oxford, UK. [2]Department of Engineering Science, University of Oxford, Oxford, UK. [3]Center for Novel Agricultural Products, Department of Biology, University of York, York, UK. [4]York Structural Biology Laboratory, Department of Chemistry, University of York, York, UK. [5]Centre for Sustainable Chemical Technologies, Department of Chemistry, University of Bath, Bath, UK. [6]Present address: UCL School of Pharmacy, London, UK. [7]These authors contributed equally: Min Yang, Charlie Fehl. *e-mail: ben.davis@chem.ox.ac.uk

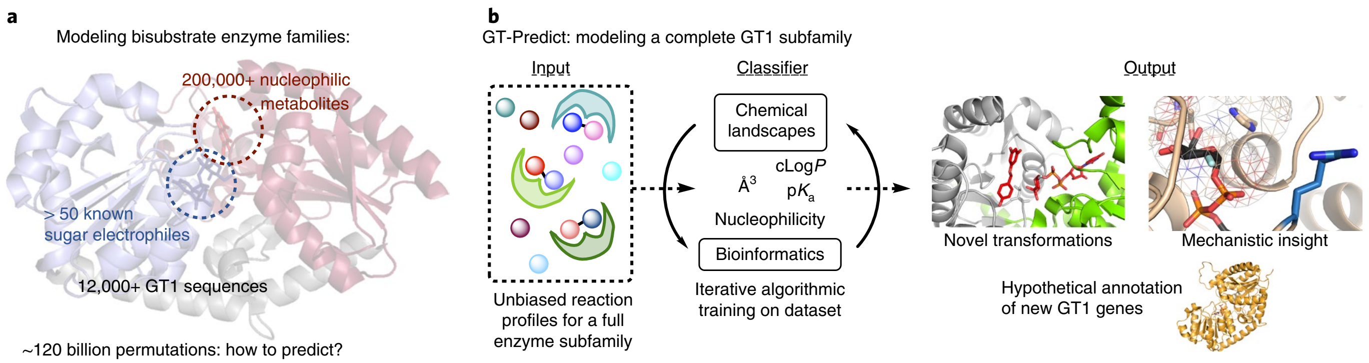

a
Modeling bisubstrate enzyme families:
200,000+ nucleophilic metabolites
> 50 known sugar electrophiles
12,000+ GT1 sequences
~120 billion permutations: how to predict?
b
GT-Predict: modeling a complete GT1 subfamily
Input
Classifier
Output
Chemical landscapes
cLogP
Å3
pKa
Nucleophilicity
Bioinformatics
Unbiased reaction profiles for a full enzyme subfamily
Iterative algorithmic training on dataset
Novel transformations
Mechanistic insight
Hypothetical annotation of new GT1 genes

Fig. 1 | Challenges and solutions for the rational prediction of multisubstrate enzyme reactions. a, The GT1 glycosyltransferase superfamily couples electrophilic sugars with nucleophilic acceptors. These reactions span the metabolome with many permutations, thus rendering current screening and prior informatics approaches insufficient for comprehensive predictive modeling. **b**, Our function-based algorithmic learning approach, GT-Predict, uses a diverse training set of enzymes, electrophiles, and nucleophiles to create a physicochemical and local-sequence-based classifier for prediction of novel transformations and functional annotation of GT group-transfer enzymes.

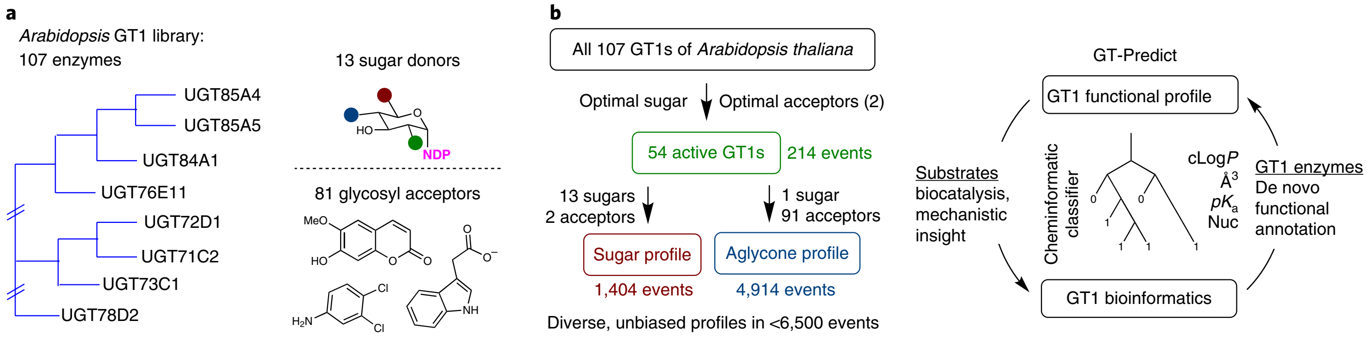


Fig. 2 | Strategy for function-based chemical–bioinformatic modeling of GT1 transformations. a, The complete GT1 library of *Arabidopsis* was screened for activity against 13 sugar electrophiles and 91 potential nucleophiles. **b**, This workflow identified 54 active GT1s, thus enabling dual-substrate library profiling by HT–MS in fewer than 6,500 events. **c**, This dataset was used to train DT models and validate cheminformatic and bioinformatic algorithms for functional prediction.

function over a 39-enzyme/three-coumarin substrate panel probing gains, losses, and regiochemical switching of activity even among closely related subfamilies. A screen of *Medicago truncatula* GT1s over 23 benzopyran(one) substrates similarly showed only sporadically clustered activity throughout the eight-enzyme dataset[7]. We therefore reasoned that any successful approach (Fig. 1) would, in essence, require a sufficient threshold of unique activity patterns of individual isoforms to be directly coupled with iterative (‘learning’) algorithms. This functional–informatic method, in turn, would require a sufficiently diverse array of chemical-substrate-recognition motifs to avoid bias, as well as a method permitting measurement of many (semi-)quantitative activity ‘events’ unencumbered (‘label free’) by structural bias or perturbation (for example, by virtue of installed chromo- or fluorophores[6,7]). The resulting dataset would subsequently be tested for utility in its ability to build and train classifier algorithms to correlate chemical and/or biological properties with the observed patterns for the protein library (here *A. thaliana* GT1 proteins).

We reasoned that a diverse, unbiased substrate usage coupled with broad a priori examination of properties would allow for the primary algorithmic focus to be intentionally generated by protein sequence (Fig. 2a). We used a decision tree (DT) learning approach, with a ‘deviance’ splitting criterion implemented through a cross-entropy function (the optimal-score function for classification, which was the (negative) log of the multinomial probability distribution for correct/incorrect decisions into one or *k* categories). Such strategies can advantageously yield interpretable insight into the key parameters (that is, for the branching of the trees) for successful prediction, if any, thus essentially allowing researchers to learn how their putative models learn. Importantly, in such an approach, any lack of statistical power from insufficient breadth in substrate variation or poor choice testing (chemical or biological) would also be directly revealed by nonrobustness or poor performance in the emergent algorithms.

We previously demonstrated a potentially general, label-free high-throughput MS (HT–MS)-based assay for (semi-)quantitative kinetic characterization of individual enzymes[14–17]. We considered that, in theory, combining the speed and broad, unbiased detection capabilities of this assay with proteins from an entire multigene family of GTs, could, for the first time, feasibly catalog a sufficiently diverse chemical dataset from a complete family to allow for algorithmic correlation (Fig. 2b), thereby permitting mechanistic and predictive insight to emerge regarding both substrates and sequences (Fig. 2c).

Screening of diverse substrates against an enzyme family. GT1 group-transfer enzymes couple two substrates through the transfer to nucleophilic ‘acceptors’ (**1**–**91**) of electrophilic glycosyl ‘donor’ moieties (**92**–**104**) (Fig. 2). Electrophilicity is generated in the donor by the presence of a nucleotide diphosphate leaving group.

Three corresponding modes of substrate diversity, corresponding to three potential structural-selectivity elements were explored: (i) configurational and constitutional (that is, hydroxyl replacement) variation in the glycosyl moiety of the donor; (ii) nucleobase variation in the leaving-group moiety of the donor; and (iii) nucleophile heteroatom type (O, NH, or S) and the constitution of the scaffold (Fig. 2a). Such an approach is consistent with the few structures of GTs that reveal corresponding pockets and their primary engagement with substrates via these three distinct moieties in Michaelis complexes[18,19]. In this way, we were able to create a broad substrate scope that could test the sufficiency of a predictive model for the GT1 enzyme superfamily (Supplementary Fig. 1).

Configurational and constitutional alterations of the donor-substrate library (**92**–**104**; Fig. 2b, Fig. 3 and Supplementary Fig. 1) were designed to explore the logical variation of the glycosyl moiety from a canonical D-glucose (Glc) starting point (Fig. 3a). For example, Glc→D-mannose (Man), Glc→D-galactose (Gal) permitted exploration of the C-2 and C-4 configurations, respectively, and Glc→*N*-acetyl-D-glucosamine (GlcNAc), Glc→D-xylose, and Glc→5-*S*-Glc permitted exploration of altered functional groups (OH-2→NHAc, CH_2OH-5→H, and *O*-5→S) as well multiply combined alterations, for example, Glc→L-fucose and Glc→L-rhamnose (OH-6→H combined with multisite configurational variation at C-2, 3, 4, and 5), which were intended to provide even greater structural diversity.

Second, the nucleobase moiety of the donor substrate was varied (for example, **92**, **99**, and **102**) from the canonical pyrimidine U in UDP to explore both other pyrimidines (for example, T), Glc-UDP→Glc-dTDP purine (for example, G) usage Glc-UDP→Glc-GDP (Fig. 3a). Consequently it was necessary to create unnatural-variant donor substrates designed to probe this nucleobase pocket in conjunction with natural variants (for example, Glc-GDP compared with Man-GDP, respectively) and variants that are species specific (for example, eukaryotic UDP compared with prokayrotic dTDP).

We designed the nucleophile-acceptor library (**1**–**91**) to probe the chemical space (molecular shape and solvent-excluded volumes), electronics (log*P* ranges, polarity, and lone-pair count), and reactivity (nucleophile type) (Supplementary Fig. 1). Variations in molecular shape (for example, via hybridization alterations or unsaturations sp^3→sp^2; acyclic versus fused/bridged polycyclic substrates) created a systematically altered yet diverse range of 'sizes'. Substrate series to reveal electronic effects included acidic, basic, and neutral variations of the same molecular cores. Finally, various O-, NH-, and S-based nucleophiles were used to evaluate the heteroatom type. The accommodation of heteroatoms in active sites appears, in particular, to be connected with subtle mutations that are not readily understood, and predictive understanding might enable creation of catalysts for the formation of new C-X bond types[19]. Diversity measures, based on the principal moments of inertia analysis of energy-minimized structures[20], confirmed a broad range of rod-like, disk-like, and spherical overall shapes (Supplementary Fig. 1c).

We conducted a sequential screen to collect datasets for enzyme activity, donor-utilization patterns, and acceptor recognition (Fig. 2b). First, we established the initial activity of the full family of 107 *Arabidopsis* GT1 enzymes, by using canonical, physiologically relevant[6] plant UDP-D-glucose substrates (Glc-UDP, donor) with known endogenous plant acceptors **23** and **31** against a panel of GT1-gene-derived lysates expressed in parallel under identical conditions[6] (Supplementary Fig. 2). This initial survey revealed activity for 54 of the 107 members at levels, and under conditions, that would permit functional screening.

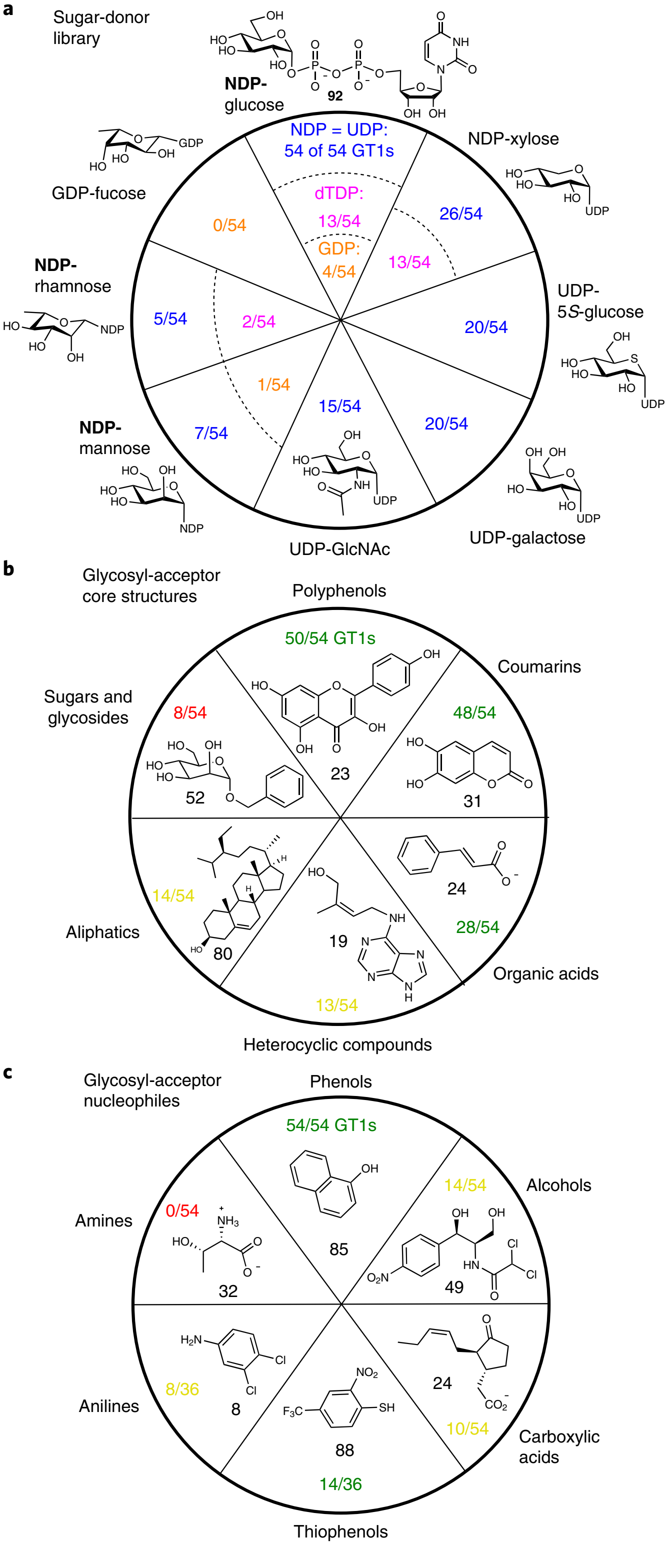


Fig. 3 | Overall donor- and acceptor-utilization patterns for the active GT1 library. **a**, Sugar-donor species, arranged by the total number of positive utilization patterns with acceptor **23** and/or **31**. The nucleotide in the NDP leaving group is listed according to color: blue, UDP; magenta, dTDP; orange, GDP. **b**, Acceptor utilization by chemical classification with donor **92**. **c**, Nucleophile-utilization examples from among the acceptor library.

Next, the systematically varied 13-member sugar-donor library was screened with the two optimal acceptors (**23** and **31**) that had shown full activity with Glc-UDP over the entire 54-enzyme panel. This procedure revealed 'coarse-grained' interaction patterns for the entire sugar/nucleoside library (Fig. 3a): the nucleoside component was more stringently regulated, with dTDP utilization (addition of a methyl group) at 25% and GDP (a purine) at only 7.4%. Alternative functional groups at C6, C4, and C2 could be used by 28–48% of the

GT1 library, including more bulky sugar 2-*N*-acetylglucosamine-UDP (GlcNAc-UDP).

Third, the canonical donor sugar Glc-UDP was used in an initial acceptor screen. Unguided manual classification of the dataset on the basis of some overall structural features (for example, aliphatics, heterocycles, and small aromatic acids; Fig. 3b) and nucleophilicity patterns (Fig. 3c) highlighted rough substrate functional group types with broad activity (for example, polyphenolic compounds) or lower activity (highly polar glycosides or amino acids). This process critically revealed that up to half of these GT1s can use a range of nucleophiles, including more unusual functional groups such as acids, anilines, and thiophenols.

Clustered functional trends are distinct from phylogeny. This diverse activity dataset was used as the basis for training chemical–bioinformatic classifiers to identify patterns useful for predictive modeling (Fig. 2c). The data were parsed according to threshold activity levels determined by the product-ion-count signal-to-noise ratio. Comparison of these data with the global amino acid sequence alignment of each active enzyme revealed only extremely scattered patterns for the both donors and acceptors (Fig. 4a and Supplementary Figs. 3–5), in agreement with the poor correlations of observed activity patterns in prior genomic and phylogenetic analyses[6,7,21]. To assess the fitness of biochemical clustering methods for our dataset analysis, we recapitulated the GT1 familial phylogenetic arrangement[6] for the aglycone acceptor library (Fig. 4a) and the sugar-donor library (Supplementary Fig. 3a). Confirming earlier reports, we observed major discrepancies between related sequences and activities for both the sugar donors and acceptors (Fig. 4a and Supplementary Fig. 3). Given the suggested structurally related nature of sugar-donor binding in plant GT1s via the so-called plant secondary-product glycosyltransferase (PSPG) motif[21], we expected ready clustering. The absence of clustering within our initial phylogenetic analyses strikingly highlighted the seemingly shallow influence of sugar type on the enzymatic evolution of at least this superfamily of GTs. Our results indicated that nucleotide diphosphate recognition, that is, for UDP, was conserved; while 25% of the GT1s surveyed here used the more structurally similar dTDP, only 7% used GDP sugars. These findings suggest that, although the PSPG motif is useful for identifying UDP-binding regions within GT1s, this motif may not account for the recognition events of the carbohydrate portions of sugar nucleotide diphosphates.

Similarly scattered activity patterns were observed for acceptors (full acceptor profile in Supplementary Figs. 3b and 4). However, some pockets of conserved function could be assigned, at least partially, to phylogenetic groupings. First, polyphenolic flavonoids and coumarins were widely used throughout the GT1 panel. Small aromatic acids also made up a significant activity group, albeit scattered throughout the phylogenetic classes. For instance, approximately half (9/17) of the tested group E enzymes used acid-containing substrates, but those enzymes were split into two subgroups over the tree rather than localizing in one defined subgroup, thus suggesting that overall amino acid conservation is not the major driver of substrate recognition. The group D and group L enzymes, the only two groups with subsets of enzymes that process polar heterocyclic rings, were also divergent in overall sequence: the group D UGT73C6 (nomenclature in Methods) and the group L UGT84A2 had 26.5% identity, 48.5% similarity, and substantial gaps (18.6% of the sequence), for example. Our results thus bolster the earlier hypotheses[6] that parallel independent evolutionary events have led to both the frequent acquisition and the loss of substrate-recognition patterns, and that sequence alignment alone is therefore not predictive of functional activity.

Next, a wholly sequence-naïve, stepwise analysis allowed for activity-based clustering of GT1 isoforms and elucidation of common functional patterns from within the superfamily. First, threshold activities were used to assign activity commonality (full, partial, or no activity) between each enzyme and each substrate molecule (Fig. 4b, Supplementary Table 1 and equation (1), Methods). Average linkage clustering (equation (2), Methods) was then implemented to hierarchically arrange the interaction patterns for enzymes in a sequence-independent fashion (Fig. 4b, horizontal axis). Notably, such 'activity clustering', guided by the individual acceptor and donor substrates' interaction patterns with GT1 proteins, permitted some manual classification of meaningful substrate–enzyme subtypes directly, whereas phylogenetic analysis wholly failed (Fig. 4b, horizontal axes). For each substrate library, clustering identified groups of GT1s with, for example, promiscuous donor-substrate scopes (Supplementary Fig. 3, right) that were unrelated to amino acid similarity or acceptor promiscuity (cf. Supplementary Fig. 5, right).

Excitingly, robust substrate clusters also emerged for acceptor nucleophiles (Fig. 4b) along with substrates with singular recognition patterns that suggested modes of GT1-isoform specialization toward for example, *N*-heterocycles, bulky fused aliphatic-ring systems, and polar glycosides. This 'chemical clustering', which emerged without the input of any physicochemical or structural information, importantly revealed the strong influence of substrate chemical properties as major drivers of substrate recognition in the GT1 superfamily.

Physicochemical analyses permit algorithmic prediction. To correlate and appropriately weight such physicochemical features rigorously, we developed an analytical process that would facilitate the discovery of overall quantitative structure–activity relationship (QSAR)-based classifiers for the GT1 family. DT-based[22] algorithms were trained on systematically varied combinations of physicochemical properties (cLog*P*, molecular volume, and pK_a) and structural parameters (functional-group copy numbers: hydroxyl groups, carboxylic acids, and amines; Supplementary Table 2). Emergent algorithms were evaluated with a leave-one-out cross-validation approach to rank the various models' predictive abilities for each compound and GT1 enzyme (Fig. 5, Supplementary Figs. 6 and 7, and Methods). From these, DT4 used a combination of physicochemical inputs (log*P*, molecular area, solvent-excluded volume, and number/type of nucleophilic groups) and structural information (scaffold type, mono/bi-cyclic variation (five- or six-membered, [4.3.0], [4.4.0] bicycles, and functional groups) that permitted prediction of interactions with 90% ± 1.3% accuracy for our *Arabidopsis* GT1 dataset. Further statistical benchmarking with the Matthews correlation coefficient (MCC; Methods), which analyzes the quality of correlations between −1.0 and + 1.0 on the basis of the true-positive/negative versus false-positive/negative rates for binary predictions yielded an average value of 0.591 for the DT4 model over all 59 acceptor molecules with experimental and/or predicted activity in this dataset (Supplementary Table 3). This procedure confirmed a strongly positive agreement between predicted and experimental results in a system that we termed GT-Predict.

GT-Predict guides functional annotation in other species. Putative annotation of gene function remains a dominant form of predictive biological analysis[23], yet many superfamilies, such as those containing GTs, remain essentially intractable to typical analyses[24]. The failure of global amino acid sequence alignment (described above) to cluster accurately and rationalize GT substrate–activity patterns, in striking contrast to the strong correlative success of our substrate physicochemical-feature analysis (described above), suggested that putative assignment would require alternative strategies.

The clear driving influence of substrate features that we observed suggested that a focused analysis of salient corresponding protein features would allow for suitable influence of substrate-interacting

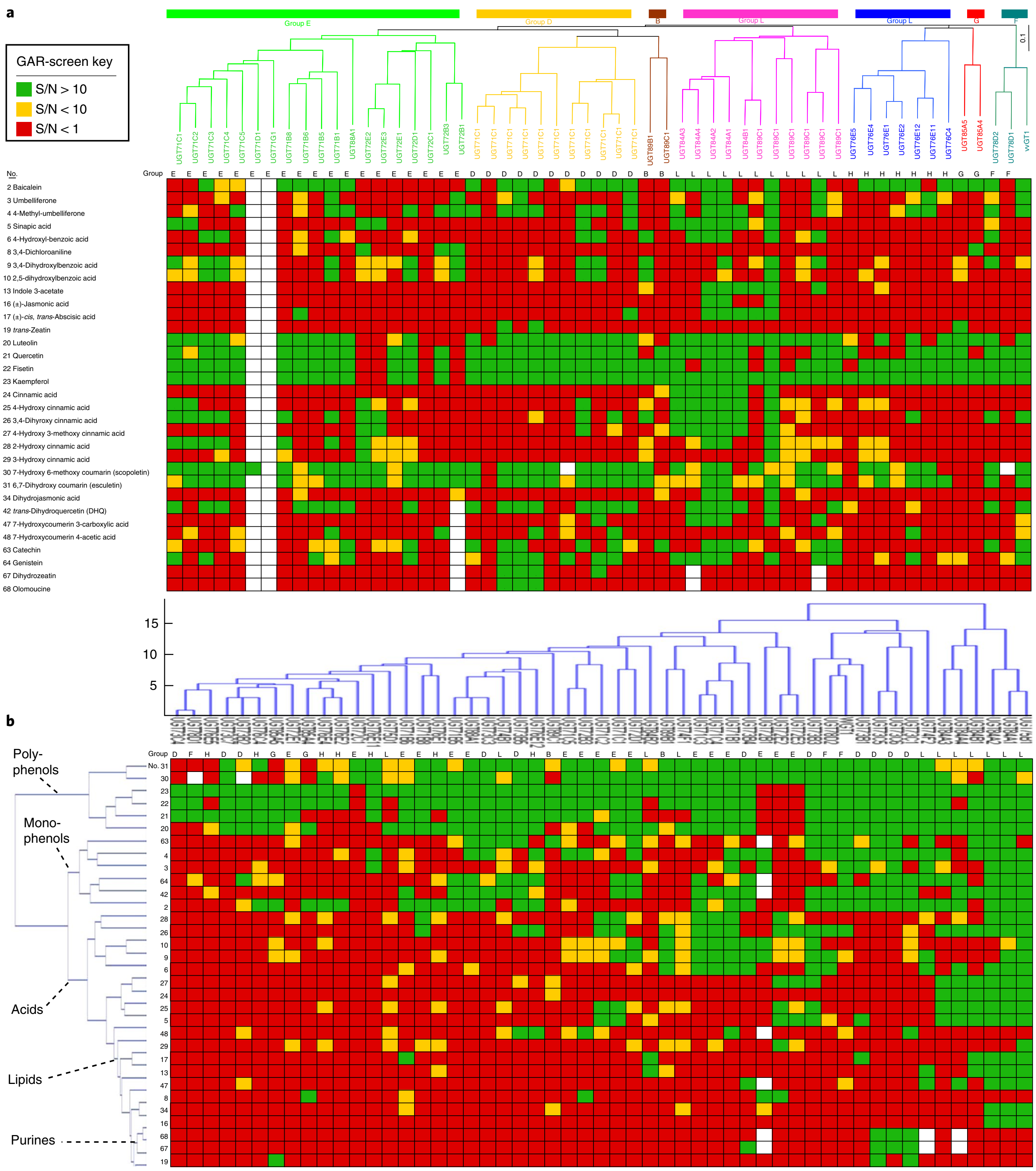


Fig. 4 | Comparison of clustering techniques for the acceptor dataset. a, Phylogenetic global sequence analysis of the 54 active GT1s was coupled with the green–amber–red (GAR)-screening-data heat map. Activity scores were judged by total ion counts of MS traces and classified according to the key. Groups indicate reported subfamilies of plant GT1 enzymes[21]. **b**, Hierarchical clustering via average linkage analysis according to equation (1) and equation (2) (Methods). The hierarchical-clustering arrangement on the *x* axis is arranged according to the similarity of individual GT1 activity patterns against all other GT1s. The tree on the *y* axis is arranged according to the association patterns of each substrate with the overall GT1 enzyme library against the other substrates' patterns. Chemical groupings refer to the emergent interaction-similarity clusters, as discussed in the text. Full datasets are available in Supplementary Figs. 3–5; inactive acceptors have been removed for clarity. All high-throughput GAR-screening experiments were performed as single measurements.

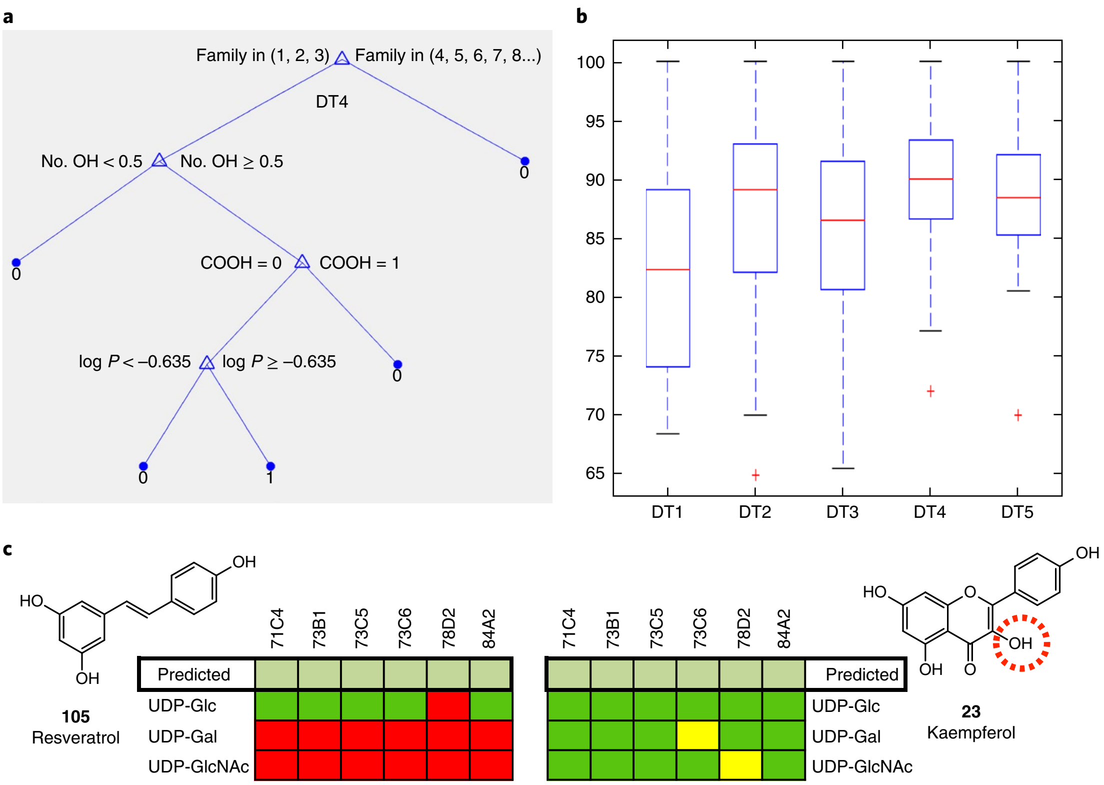


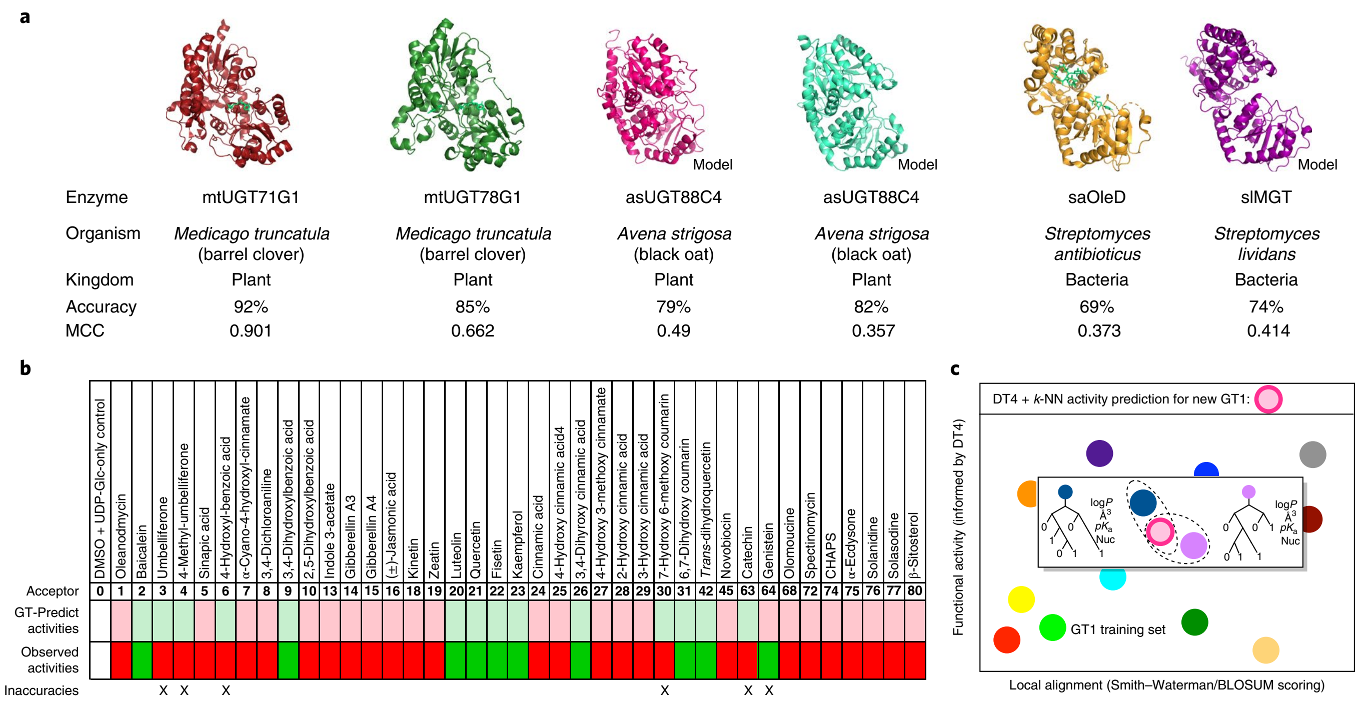
Fig. 5 | GT-Predict development, validation, and utilization. a, Diagram of the optimal DT (DT4) used to classify information (additional information in Supplementary Note). **b**, Leave-one-out cross-validation of all DT models. Shown is the percentage accuracy of the trained model for each member of the sugar-acceptor library. Dashed lines indicate the full range of the validation-accuracy dataset. Single outliers (red crosses) were determined by ranking the predicted versus experimental results for each acceptor that showed activity with at least one GT1 enzyme; they were not included in the statistics of the box plot but are shown for completeness. The median percentage-accuracy values are shown in red lines for 59 acceptors tested in single measurements via high-throughput GAR-screening experiments (additional data in Supplementary Table 3). The interquartile ranges (25–75%) are shown in blue boxes. DT1–DT5 are DT-based models (additional information, including further validation with MCC analysis, in Supplementary Note). **c**, A subset of the GT-Predict results (in the bold box) for kaempferol (**23**) and application to prediction of enzymes for the new substrate resveratrol (**105**) alongside GAR activity for glycosylation of **105** with various NDP-sugar substrates. The results confirmed predictions and permitted use of atUGT73C6 for these transformations on a preparative scale (Supplementary Note). The variation in donor utilization by **23** and **105** highlights the essential discovery from DT4 of the presence or absence of the acceptor hydroxyl functional group (circled) as a key parameter for successful activity prediction for alternative NDP-donor substrates. All GAR-screening experiments were performed as single measurements.

regions in an unbiased manner. Local sequence alignment can be used to rank short highly similar regions while ignoring large gaps or regions of sequence divergence more effectively than in global sequence alignment[25]. This process, in principle would enable algorithmic focus on more relevant (for example, substrate-interacting) protein regions. Thus, the use of the Smith–Waterman algorithm for local sequence alignment[25] allowed us to interrogate novel sequences of GT1 enzymes outside of our dataset, by using our functionally characterized enzyme library. For efficient interrogation, we developed a program to perform combined local alignment and BLOSUM50 scoring of the novel GT1 amino acid sequence against each of the GT1 sequences in our activity dataset. Merged use of the highest two 'scores' enabled predictive selection of the most likely set of substrates for the novel GT1 enzyme and hence putative functional assignment that could be tested experimentally.

In this way, GT-Predict was able to propose hypothetical activities for putative gene products individually selected from other species (Fig. 6). First, four individually selected GT1 gene sequences from the legume *M. truncatula* (mt; genes *UGT71G1* and *UGT78G1*) and the cereal *Avena strigosa* (as; genes *UGT74H5* and *UGT88C4*) were analyzed, and the activities of the encoded enzymes (mtUGT71G1 and mtUGT78G1, and asUGT74H5, asUGT88C4, respectively; nomenclature described in Methods) were predicted and then compared with experimentally determined results[26,27]. The comparison (Fig. 6) revealed an 85–92% accuracy (Supplementary Table 4) for GT-Predict when tested against the subset of 44 substrates that demonstrated robust activity in the *Arabidopsis* dataset (Supplementary Fig. 13); the corresponding MCC values were between 0.518 and 0.910 (Supplementary Table 3), thus indicating very strong to excellent predictive correlation.

Next, we extended the GT-Predict workflow to test prediction against all of CAZy-confirmed gene members of the two complete families from *A. strigosa* and *Lycium barbarum* (Supplementary Figs. 8–11 and Supplementary Tables 5 and 6). These tests again were successful, with accuracy rates of 79.0 (MCC + 0.338) and 78.8% (MCC + 0.319), respectively.

Finally, in addition to testing its utility against cognate-kingdom species from different phyla, we tested GT-Predict against far more divergent sequences from two different phyla within a different kingdom: the actinobacteria *Streptomyces antibioticus* and *Streptomyces lividans* GT enzymes (saOleD and slMGT[28], respectively; Fig. 6). Strikingly, despite the sequence divergence and the change of kingdom (plant→bacteria) from the *A. thaliana* GT1s in our dataset, GT-Predict was 69% accurate (with a positive MCC value of + 0.373) for saOleD and 74% (with a positive MCC value of + 0.414) for slMGT.


a
Model
Model
Model
b
c
DT4 + k-NN activity prediction for new GT1:
Functional activity (informed by DT4)
logP
Å3
pKa
Nuc
0
1
GT1 training set
Local alignment (Smith–Waterman/BLOSUM scoring)

Enzyme	mtUGT71G1	mtUGT78G1	asUGT88C4	asUGT88C4	saOleD	slMGT
Organism	*Medicago truncatula* (barrel clover)	*Medicago truncatula* (barrel clover)	*Avena strigosa* (black oat)	*Avena strigosa* (black oat)	*Streptomyces antibioticus*	*Streptomyces lividans*
Kingdom	Plant	Plant	Plant	Plant	Bacteria	Bacteria
Accuracy	92%	85%	79%	82%	69%	74%
MCC	0.901	0.662	0.49	0.357	0.373	0.414

	DMSO + UDP-Glc-only control	Oleanodmycin	Baicalein	Umbelliferone	4-Methyl-umbelliferone	Sinapic acid	4-Hydroxyl-benzoic acid	α-Cyano-4-hydroxyl-cinnamate	3,4-Dichloroaniline	3,4-Dihydroxylbenzoic acid	2,5-Dihydroxylbenzoic acid	Indole 3-acetate	Gibberellin A3	Gibberellin A4	(±)-Jasmonic acid	Kinetin	Zeatin	Luteolin	Quercetin	Fisetin
Acceptor	0	1	2	3	4	5	6	7	8	9	10	13	14	15	16	18	19	20	21	22
GT-Predict activities		🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩
Observed activities		🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩
Inaccuracies				X	X		X													

	Kaempferol	Cinnamic acid	4-Hydroxy cinnamic acid4	3,4-Dihyroxy cinnamic acid	4-Hydroxy 3-methoxy cinnamate	2-Hydroxy cinnamic acid	3-Hydroxy cinnamic acid	7-Hydroxy 6-methoxy coumarin	6,7-Dihydroxy coumarin	*Trans*-dihydroquercetin	Novobiocin	Catechin	Genistein	Olomoucine	Spectinomycin	CHAPS	α-Ecdysone	Solanidine	Solasodine	β-Sitosterol
Acceptor	23	24	25	26	27	28	29	30	31	42	45	63	64	68	72	74	75	76	77	80
GT-Predict activities	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Observed activities	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Inaccuracies								X				X	X							

Fig. 6 | GT-Predict extends functional annotation to other species, kingdoms, and GT families. a, Summary of GT-Predict prediction results for six selected individual enzymes from differing species, including accuracy and MCCs. Further details and analysis can be found in the Supplementary Note. Other extensions to additional GT families from *A. strigosa* and *L. barbarum* can be found in Supplementary Figs. 8–11. Images were generated in PyMOL from PDB 2ACV, 3HBF and 2IYF) or models created in I-TASSER[51]. **b**, Predicted versus actual experimental results for acceptor utilization for the single enzyme mtUGT78G1 for 38 acceptors tested in singleton high-throughput GAR-screening experiments (additional data in Supplementary Figs. 8, 9, and 13). **c**, Representation of the successful 'PredictEnzymeInteraction' module, which combines the DT4 model for prediction of chemical-interaction patterns and ranking with a *k*-nearest-neighbor (*k*-NN) algorithm for local sequence-alignment matching. Colored dots represent the GT1 training set for the DT4/*k*-NN model. The bold/pink circle represents the novel sequence of interest. The DTs represent the activity sets and physicochemical-property space of the nearest two GT1s in the training set, which are used for activity prediction.

GT-Predict guides synthetically useful transformations. Next, we tested the predictive power of GT-Predict on a model compound as a potential substrate. Resveratrol (**105**) is an antioxidant and pan–histone deacetylase inhibitor[29] currently in clinical trials for cancer prevention[30] and neurodegenerative disease[31]. Its poor solubility as a free drug[32] has prompted investigation into the production of resveratrol glycosides to improve its pharmacological properties[33,34]. Moreover, for the purposes of validating GT-Predict, resveratrol is endogenous only to berry-producing plant species but is not found in *A. thaliana*[35].

Using GT-Predict, we identified several GT1s in the *A. thaliana* (at) GT superfamily predicted to hypothetically glycosylate resveratrol as an acceptor nucleophile; usefully, these included GTs predicted to also be capable of using a selection of NDP sugar-donor electrophiles, thus allowing for good diversity of elaboration. When experimentally tested in vitro, the predicted biocatalyst atUGT73C6 proved most efficient from within the enzyme set, permitting regioselective and one-step synthesis of monoglycosylated resveratrol on a preparative scale (Supplementary Fig. 12). Notably and importantly, these in vitro results confirmed elegant results previously determined when the *Arabidopsis* GTs were used in whole-cell biocatalytic transformation to glucosylate **105** (ref. [34]).

In an essentially similar manner, asUGT88C4 was identified as a novel biocatalyst able to glycosylate novobiocin (Supplementary Fig. 13), a prenylated antibiotic[36] biosynthesized by *Streptomyces niveus*, thereby demonstrating predictive activity discovery for not only nonendogenous substrates but also those outside of normal plant metabolism.

GT-Predict shows site features modulating selectivity. Structural guidance remains a crucial aspect for hypothesis-driven insight into biocatalyst mechanisms and enzyme engineering[19]. Whereas GT-Predict is founded on a comprehensive functional dataset, its use in conjunction with structural approaches also allowed for the identification of possibly important structural motifs and their roles within active sites. This identification was aided by a combined visualization tool and graphical user interface that highlighted patterns on the basis of physicochemical property analyses (Supplementary Fig. 14). In this way, for example, the given acceptor substrates for a particular GT1 enzyme could be related to any two chosen chemical properties versus functional activity in three-dimensional plots (Supplementary Fig. 14), to permit interrogation of emergent correlations.

These activity plots, in turn, enabled the discovery of intriguing observations and parameter determinants related to possible structural origins of the observed activities. For example, the activity plots of acid-containing acceptors revealed distinct dichotomous 'allowed versus forbidden' utilization of anionic substrates by GT1 isoforms. These findings in turn prompted structural investigation through GT-Predict-guided identification of relevant homolog sequences for which useful structural information is available in combination with homology-guided modeling (all models mapped closely onto known structures, with minor overall r.m.s. deviations of 0.73–1.25 Å (Supplementary Table 7 and Methods)).

Unique chemical patterns were investigated to explore three hypothetical 'drivers' of substrate recognition for several isozymes. First, the breadth of the used substrate volume correlated with the GT1 active site size (Supplementary Fig. 14a,b), as judged by mapping the accessible volume versus log*P*—a surrogate for molecular surfaces—in the crystallized (atUGT72B1) or modeled (asUGT84A2) active sites. Second, selection of negatively charged substrates (at pH 8.0) involves either engagement by cationic active site–residue motifs and/or gating by anionic-residue motifs (Supplementary Fig. 14c,d). For example, in carboxylic acid–using GT1 atUGT84A2 (Supplementary Fig. 14d), this procedure revealed a neutral active site cavity (Supplementary Fig. 14b). In contrast, in two GT1s not able to glycosylate acids, atUGT72C1 and atUGT73C5, each displayed negatively charged 'gates' composed of two acidic residues near the proposed substrate-access cleft: D180/E187 of atUGT72C1 (Supplementary Fig. 14c) and D92/E198 of atUGT73C5 (Supplementary Fig. 15). Third, the utilization of sugar donors is modulated by the recognition of larger polar substituents through hydrogen-bonding to polar amino acids in accommodating pockets (Supplementary Fig. 14e). For example, the use by atUGT71C4 of more bulky polar UDP-GlcNAc donor substrate correlated with a unique arginine residue at position 292 (Supplementary Fig. 14e), adjacent to the UDP-binding PSPG motif at a distance of 7.4 Å from the C2 substituent, a configuration nearly optimal for a hydrogen-bonding interaction with the *N*-acetyl group of GlcNAc.A hydrophobic residue or glycine occupied this position in the remaining group E GT1s studied. Notably, this arginine substitution was not found to be general to all other plant UDP-GlcNAc using GT1s, thus highlighting that directed algorithmic functional annotation can suggest rare but functional protein features, perhaps by identifying a unique evolutionary direction taken by an individual isoform within the GT1 family. Other structurally characterized UDP-GlcNAc-using enzymes also appear to exploit arginine residues to mediate selectivity[37,38].

The residues pinpointed by GT-Predict in these 'gating' interactions, namely sites D180 and E187 in atUGT72C1, and R292 in atUGT71C4, were experimentally probed through site-directed mutagenesis (Supplementary Fig. 15). Notably, in agreement with drivers implicated by GT-Predict, the mutation of aspartate/glutamate→alanine in atUGT72C1 D180A/E187A enabled activity toward acids (not present in the wild type), and mutation of arginine→alanine in atUGT71C4 R292A removed the ability to transfer GlcNAc (but not Glc). These results not only confirmed the importance of these residues in controlling activity but also directly highlighted the potential of GT-Predict for use in rational enzyme engineering.

Discussion

Comprehensive predictive modeling of enzyme superfamilies has remained an unsolved challenge despite advances in genomics, proteomics, and metabolomic data-gathering and analysis[39]. Certain predictive attempts have found some success, such as a database of in silico docking data compiled for more than 100 hydrolase enzyme structures[40] and the development of a structure-guided metabolomics-prediction system to annotate new protein functions[41]. However, these approaches to date have been confined to proteins of known structure and with relatively narrow substrate variation. Substrate utilization and chemical properties have been linked to generate QSAR-based predictive models for individual proteins from large protein families[42,43] and have long been applied in inhibitor design[44].

Here, a structurally and phylogenetically naïve functional approach succeeded in a testing proof-of-concept family (the GTs) by using libraries designed to probe chemical space across enough members of a species-wide collection of enzymes to obtain a training set. In this way, the combination of an extensive functional dataset and a chemical–bioinformatic analytical method enabled accurate modeling of a full protein family and, indeed, prediction, testing, and validation of mechanistic hypotheses and synthetic activities.

As an example of informatic encapsulation of a full protein family, several limitations to this approach should be recognized. First, regiochemical selectivity was not strongly considered in designing GT-Predict, which was based on the presence versus absence of chemical groups but not their three-dimensional orientation. Some limitations can be noted when comparing seemingly highly related substrates in which the relative position of an additional putative nucleophile may give rise to enhanced reactivity (for example, kaempferol (**23**) » resveratrol (**105**)). Additional strategies to exploit such regiochemical bias ('substrate fit') might further enhance accuracy[6] (for example, Supplementary Fig. 4). Second, although our substrate library was found to be sufficiently broad for successful training, the predictive scope might also be further enhanced by adding database input, for example through DrugBank[45] or metabolomic compound collections such as the Plant Metabolome Database[46], if sufficiently well curated and tested. Third, GT-Predict now permits accurate prediction of GT1 activities correlated with local primary-sequence alignment, in a manner that was not previously possible, with the greatest accuracy for plant proteins. More advanced secondary-structure prediction/alignment methods might be anticipated to extend this method yet further (for example, for low sequence homology but high predicted structural similarity). Similarly, validation of the mechanistic hypotheses suggested by GT-Predict through structural biology[47] would clearly be of direct benefit in augmenting the promising mutagenic results obtained here. Because an excellent database for GTs (and other carbohydrate-processing enzymes) is available in CAZy[4], even further refinements and implementations based on this informatics environment might be anticipated.

Given the apparently related structural nature of sugar donors, it is surprising that direct phylogenetic clustering of their utility as substrates fails. Yet, our results, like those of other studies[7,47,48] clearly show that such analyses alone are not successful and are limited by, for example, sequence variability[47]. This finding strikingly highlights the shallow influence of sugar type on the enzymatic evolution of at least this superfamily of GTs and/or the guidance of selectivity by other parameters that are not defined by the ground state (for example, transition-state conformation[49]). Nonetheless, it is also clear that physicochemical parameters provide a strong guide that emerges through their striking hierarchical influence on clustering that we observed here, in agreement with the results of recent analyses of the evolution of function within certain conserved folds[50].

GT-Predict also allows for rational selection with some confidence of scaffolds for desired transformations and thus might complement some current de novo computational design algorithms, which have succeed at creating defined packing and active site cavities but may fail in terms of the finer points of active site residue identity and position[13]. For example, augmentation of computational and forced-evolution-based protein-design methods might also use starting points for a desired function identified from within a large protein superfamily.

Finally, the strategy presented here of algorithmically coupling chemical-interaction patterns with local sequence analysis might be readily extended to other protein superfamilies that remain currently intractable to predictive functional annotation and engineering.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41589-018-0154-9.

Received: 30 August 2017; Accepted: 19 September 2018;
Published online: 12 November 2018

References

1. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
2. Gerlt, J. A. & Babbitt, P. C. Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* **2**, 607–612 (1998).
3. Friedmann, D. R. & Marmorstein, R. Structure and mechanism of non-histone protein acetyltransferase enzymes. *FEBS J.* **280**, 5570–5581 (2013).
4. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
5. Li, T. et al. Characterization and prediction of lysine (K)-acetyl-transferase specific acetylation sites. *Mol. Cell. Proteomics.* **11**, M111.011080 (2012).
6. Lim, E.-K. et al. Evolution of substrate recognition across a multigene family of glycosyltransferases in A*rabidopsis*. *Glycobiology* **13**, 139–145 (2003).
7. Modolo, L. V. et al. A functional genomics approach to (iso)flavonoid glycosylation in the model legume *Medicago truncatula*. *Plant Mol. Biol.* **64**, 499–518 (2007).
8. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.* **77**, 521–555 (2008).
9. Cartwright, A. M., Lim, E.-K., Kleanthous, C. & Bowles, D. J. A kinetic analysis of regiospecific glucosylation by two glycosyltransferases of Arabidopsis thaliana: domain swapping to introduce new activities. *J. Biol. Chem.* **283**, 15724–15731 (2008).
10. Todd, A. E., Orengo, C. A. & Thornton, J. M. Plasticity of enzyme active sites. *Trends. Biochem. Sci.* **27**, 419–426 (2002).
11. Gloster, T. M. Advances in understanding glycosyltransferases from a structural perspective. *Curr. Opin. Struct. Biol.* **28**, 131–141 (2014).
12. Harper, K. C. & Sigman, M. S. Predicting and optimizing asymmetric catalyst performance using the principles of experimental design and steric parameters. *Proc. Natl. Acad. Sci. USA* **108**, 2179–2183 (2011).
13. Kries, H., Blomberg, R. & Hilvert, D. De novo enzymes by computational design. *Curr. Opin. Chem. Biol.* **17**, 221–228 (2013).
14. Yang, M., Brazier, M., Edwards, R. & Davis, B. G. High-throughput mass-spectrometry monitoring for multisubstrate enzymes: determining the kinetic parameters and catalytic activities of glycosyltransferases. *Chembiochem* **6**, 346–357 (2005).
15. Flint, J. et al. Structural dissection and high-throughput screening of mannosylglycerate synthase. *Nat. Struct. Mol. Biol.* **12**, 608–614 (2005).
16. Yang, M., Davies, G. J. & Davis, B. G. A glycosynthase catalyst for the synthesis of flavonoid glycosides. *Angew. Chem. Int. Edn Engl.* **46**, 3885–3888 (2007).
17. Backus, K. M. et al. Uptake of unnatural trehalose analogs as a reporter for *Mycobacterium tuberculosis*. *Nat. Chem. Biol.* **7**, 228–235 (2011).
18. Offen, W. et al. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. *EMBO J.* **25**, 1396–1405 (2006).
19. Brazier-Hicks, M. et al. Characterization and engineering of the bifunctional N- and O-glucosyltransferase involved in xenobiotic metabolism in plants. *Proc. Natl. Acad. Sci. USA* **104**, 20238–20243 (2007).
20. McLeod, M. C. et al. Probing chemical space with alkaloid-inspired libraries. *Nat. Chem.* **6**, 133–140 (2014).
21. Li, Y., Baldauf, S., Lim, E. K. & Bowles, D. J. Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. *J. Biol. Chem.* **276**, 4338–4343 (2001).
22. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Wadsworth & Brooks, Monterey, CA, 1984).
23. Kotera, M., Goto, S. & Kanehisa, M. Predictive genomic and metabolomic analysis for the standardization of enzyme data. *Perspect. Sci.* **1**, 24–32 (2014).
24. Sánchez-Rodríguez, A. et al. A network-based approach to identify substrate classes of bacterial glycosyltransferases. *BMC Genomics* **15**, 349 (2014).
25. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
26. Shao, H. et al. Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from *Medicago truncatula*. *Plant Cell* **17**, 3141–3154 (2005).
27. Modolo, L. V. et al. Crystal structures of glycosyltransferase UGT78G1 reveal the molecular basis for glycosylation and deglycosylation of (iso)flavonoids. *J. Mol. Biol.* **392**, 1292–1302 (2009).
28. Yang, M. et al. Probing the breadth of macrolide glycosyltransferases: in vitro remodeling of a polyketide antibiotic creates active bacterial uptake and enhances potency. *J. Am. Chem. Soc.* **127**, 9336–9337 (2005).
29. Venturelli, S. et al. Resveratrol as a pan-HDAC inhibitor alters the acetylation status of histone [corrected] proteins in human-derived hepatoblastoma cells. *PLoS ONE* **8**, e73097 (2013).
30. Kjaer, T. N. et al. Resveratrol reduces the levels of circulating androgen precursors but has no effect on, testosterone, dihydrotestosterone, PSA levels or prostate volume: a 4-month randomised trial in middle-aged men. *Prostate* **75**, 1255–1263 (2015).
31. Turner, R. S. et al. A randomized, double-blind, placebo-controlled trial of resveratrol for Alzheimer disease. *Neurology* **85**, 1383–1391 (2015).
32. Tomé-Carneiro, J. et al. Resveratrol and clinical trials: the crossroad from in vitro studies to human evidence. *Curr. Pharm. Des.* **19**, 6064–6093 (2013).
33. Pandey, R. P. et al. Enzymatic biosynthesis of novel resveratrol glucoside and glycoside derivatives. *Appl. Environ. Microbiol.* **80**, 7235–7243 (2014).
34. Weis, M., Lim, E.-K., Bruce, N. & Bowles, D. Regioselective glucosylation of aromatic compounds: screening of a recombinant glycosyltransferase library to identify biocatalysts. *Angew. Chem. Int. Ed. Engl.* **45**, 3534–3538 (2006).
35. Burns, J., Yokota, T., Ashihara, H., Lean, M. E. & Crozier, A. Plant foods and herbal sources of resveratrol. *J. Agric. Food. Chem.* **50**, 3337–3340 (2002).
36. Heide, L. The aminocoumarins: biosynthesis and biology. *Nat. Prod. Rep.* **26**, 1241–1250 (2009).
37. Peneff, C. et al. Crystal structures of two human pyrophosphorylase isoforms in complexes with UDPGlc(Gal)NAc: role of the alternatively spliced insert in the enzyme oligomeric assembly and active site architecture. *EMBO J.* **20**, 6191–6202 (2001).
38. Unligil, U. M. et al. X-ray crystal structure of rabbit N-acetylglucosaminyltransferase I: catalytic mechanism and a new protein superfamily. *EMBO J.* **19**, 5269–5280 (2000).
39. Pearson, W. R. Protein function prediction: problems and pitfalls. *Curr. Protoc. Bioinformatics* **51**, 12.1 –4.12.8 (2015).
40. Tyagi, S. & Pleiss, J. Biochemical profiling in silico: predicting substrate specificities of large enzyme families. *J. Biotechnol.* **124**, 108–116 (2006).
41. Zhao, S. et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* **502**, 698–702 (2013).
42. Nembri, S., Grisoni, F., Consonni, V. & Todeschini, R. In silico prediction of cytochrome P450-drug interaction: QSARs for CYP3A4 and CYP2C9. *Int. J. Mol. Sci.* **17**, E914 (2016).
43. Dong, D., Ako, R., Hu, M. & Wu, B. Understanding substrate selectivity of human UDP-glucuronosyltransferases through QSAR modeling and analysis of homologous enzymes. *Xenobiotica* **42**, 808–820 (2012).
44. Wang, T., Yuan, X. S., Wu, M.-B., Lin, J.-P. & Yang, L.-R. The advancement of multidimensional QSAR for novel drug discovery: where are we headed? *Expert Opin. Drug Discov.* **12**, 769–784 (2017).
45. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
46. Udayakumar, M. et al. PMDB: plant metabolome database: a metabolomic approach. *Med. Chem. Res.* **21**, 47–52 (2012).
47. Schmid, J., Heider, D., Wendel, N. J., Sperl, N. & Sieber, V. Bacterial glycosyltransferases: challenges and opportunities of a highly diverse enzyme class toward tailoring natural products. *Front. Microbiol.* **7**, 182 (2016).
48. Osmani, S. A., Bak, S. & Møller, B. L. Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. *Phytochemistry* **70**, 325–347 (2009).
49. Davies, G. J., Planas, A. & Rovira, C. Conformational analyses of the reaction coordinate of glycosidases. *Acc. Chem. Res.* **45**, 308–316 (2012).
50. Newton, M. S. et al. Structural and functional innovations in the real-time evolution of new $(\beta\alpha)_8$ barrel enzymes. *Proc. Natl. Acad. Sci. USA* **114**, 4727–4732 (2017).
51. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).

Acknowledgements

We gratefully acknowledge A. Osbourne (JIC) for contribution of *A. strigosa* GT1 genes As08 (*UGT74H5*) and As09 (*UGT88C4*); R. Edwards and M. Brazier-Hicks for sharing activity data; and I. Mear for assistance with coding. This work was funded by the BBSRC (EGA16205, EGA16206, and EGA17763) and the EPSRC (the UK Catalysis Hub, EP/K014668/1 and EP/M013219/1).

Author contributions

G.J.D., D.J.B., M.G.D., S.J.R., and B.G.D. designed the research; M.Y., C.F., and K.V.L. performed the research; M.Y., C.F., K.V.L., E.-K.L. W.A.O., G.J.D., S.J.R., and B.G.D. analyzed the data; G.J.D., D.J.B., M.G.D., S.J.R., and B.G.D. wrote the paper; all authors read and commented on the paper. M.Y. and C.F. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-018-0154-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.G.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

General considerations. Unless otherwise noted, chemical reagents, media, and bacterial cell stocks were obtained from commercial suppliers (Sigma-Aldrich, Fluorochem, Carbosynth, VWR, Alfa Aesar, or Fisher Scientific) and were used without further purification. Sonication was performed with a Fisher Scientific Model 505 Sonic Dismembrator. Proteins were purified with an Äkta FPLC System UPC-900 (GE Healthcare). HT–MS was performed with either a Waters Quattro Micro API (ESI$^-$ mode) or a Waters ZMD-MS (ESI$^-$ mode) detector, each equipped with a Waters 600 HPLC System and a Waters 2700 autosampler capable of 96-well sampling format. Gel electrophoresis was performed with Invitrogen NuPAGE 4–12% Bis-Tris gels, Novex MiniCell tanks, and a Bio-Rad PowerPac controller. Western blotting was performed with an iBlot gel-transfer device from Thermo Fisher. Thin-layer chromatography was performed with Silica Gel 60 F_{254} plates (Merck) with 1–10% methanol in dichloromethane. Proton NMR spectra were recorded on a Bruker AVIII HD 400 nanobay (400-MHz) spectrometer. Carbon NMR spectra were recorded on a Bruker DQX 400 (100-MHz) spectrometer. All ^{1}H NMR chemical shifts are shown in p.p.m. with residual solvent as the internal standard relative to TMS (d6-acetone, 2.09 p.p.m.). All ^{13}C NMR chemical shifts are shown in p.p.m. with the central solvent peak as the internal standard relative to TMS (d6-DMSO, 39.3 p.p.m.). Coupling constants (J) are reported in Hertz. Infrared (IR) spectra were recorded on a Bruker Tensor 27 Fourier-transform spectrophotometer. High-resolution mass spectra were recorded on a Micromass LCT (resolution = 5,000 RWHM) with a lock-spray source. Protein crystal structures were analyzed and displayed with MacPyMOL v. 1.3 (Schrödinger). Synthetic genes for *M. truncatula UGT71G1* and *UGT78G1* were obtained from GeneArt Gene Synthesis (Thermo Fisher) with *Escherichia coli* codon-optimized amino acid sequences, as reported by Wang et al[26,27]. and cloned into the pGEX2 T vector (Amersham Pharmacia Biotech) with T4 DNA ligase (New England BioLabs). Mutagenesis was performed with a Q5 Site-Directed Mutagenesis Kit (New England BioLabs). Nucleotide sequencing was confirmed by the Source Bioscience DNA Sanger sequencing services of Oxford University (UK).

UGT enzymes are named according to the UGT Nomenclature Committee's most recent guidelines[52] as follows: *A. thaliana* protein UGT73C6 encoded by gene *UGT73C6* is written atUGT73C6.

Plant GT1 production. *Arabidopsis* GT1 plasmids in pGEX-2T (as reported by Lim et al[6].) were transformed into Rosetta (DE3) pLysS *E. coli* expression strains and produced essentially as previously reported[6,53]. Cells were resuspended in glutathione *S*-transferase (GST) purification buffer (50 mM Tris, pH 7.4, and 1 mM DTT), lysed, and centrifuged (at 10,000*g* for 10 min at 4 °C, then at 25,000*g* for 60 min at 4 °C) and either used as the crude supernatant or purified with a Sepharose 4B–glutathione resin (GE Healthcare) as previously described[53]. Western blotting was performed with mouse anti-GST (Supplementary Fig. 2a). Catalog numbers for antibodies: anti–glutathione *S*-transferase mouse IgG1 (BD Biosciences, clone G172-1138, catalog number 554805, lot number 4163768, dilution 1:1000); rabbit anti–mouse IgG–alkaline phosphatase fusion (Sigma Aldrich, polyclonal, catalog number A3562, lot number SLBK3154V, dilution 1:20,000); and mouse monoclonal anti–polyhistidine IgG–alkaline phosphatase fusion (Sigma, clone HIS-1, catalog number A5588, lot number 085M4836V, dilution 1:5000). All antibodies used were commercially available with respective documentation. We found that the GT1-protein-containing lysates could be flash-frozen and thawed once, with activity remaining for up to 6 months of storage at −80 °C.

Green–amber–red HT–MS screening. Activity assays were conducted with previously reported MS methods[14] on either a Waters Quattro Micro API (ESI$^-$ mode) or a Waters ZMD-MS (ESI$^-$ mode) instrument, each equipped with a Waters 600 HPLC System and a Waters 2700 autosampler capable of 96-well format. Reaction mixtures were composed of 93 µL reaction buffer (1 mM Tris, pH 7.8, and 50 µM $MgCl_2$), 1 µL NDP sugar (10 mg/mL stock), 1 µL aglycone (10 mg/mL stock), and 5 µL cell supernatant or purified protein (~1 mg/mL). Glycosylation reactions were incubated at 37 °C overnight and monitored by MS full scans (150–1,100 Da). A direct infusion of 10 µL of each reaction mixture was injected into the MS with 50:50 MeCN/H_2O (0.1 mL/min flow rate, 5.5 min flush). Data were ranked green (signal/noise > 10), amber (signal/noise 1–10), or red (signal/noise < 1) from the total-ion-count integration of the full peak (representative data in Supplementary Fig. 2b,c). The acceptor library is shown in Supplementary Fig. 1, and the full acceptor dataset is shown in Supplementary Fig. 3b. The full donor dataset is shown in Supplementary Fig. 3a. Regioselectivities were based on comparison of LC–MS elution time with internal standards, as previously reported[8] or as deduced from substitution patterns within the same chemical families (Supplementary Fig. 4).

Chemical-diversity calculations. Molecular-shape calculations were used to design library features sampling a broad range of three-dimensional chemical space (Supplementary Fig. 1c). Each structure was energy minimized with the MM2 function of Chem3D (CambridgeSoft) and converted to.sdf format. The principal moment of inertia was calculated for the energy-minimized conformations of our library members with the Knime Analytics Platform[54] with the 'SDF Reader'→ 'PMI Calculation' (Vernalis)→'JavaScript Scatter Plot' nodes and compared against reference molecules for 'rod' (octa-2,4,6-triyne), 'sphere' (adamantane), and 'disk' (benzene)[55]. Our compounds were found to lie primarily along the rod–disk axis, but they sampled space well into the other principal chemical-shape regions.

Clustering of activity according to phylogenetic alignment or functional patterning. Phylogenetic analyses were performed with CLUSTAL_X[56] or Clustal Omega[57] and fully matched reported analysis for the *Arabidopsis* UGT family[21]. Pairwise alignment was performed with the EMBOSS Water program[58]. Functional activity analysis used hierarchical clustering to score and regroup the acceptors and donors according to GT1 interaction patterns (green, score of 1.0; amber, score of 0.5; red, score of 0.0). Clustering proceeded via average linkage analysis[59] (further details in the Supplementary Note).

Hierarchical clustering of activity. Functional activity analysis used hierarchical clustering to score and regroup the acceptors and donors on the basis of UGT interaction patterns (green, score of 1.0; amber/'unclear', score of 0.5; red, score of 0.0). With our interaction data for each donor or acceptor molecule and the full collection of enzymes, each pair of enzymes i and j was assigned a distance score based on equation (1) with parameters from Supplementary Table 1.

$$d(i,j) = \sum_{m=1}^{M} d_m(i,j) \tag{1}$$

$$D(A,B) = \frac{\sum_{i\in A}\sum_{j\in B} d(i,j)}{N_A N_B} \tag{2}$$

Hierarchical arrangement proceeded via average linkage-analysis clustering according to equation (2) in MATLAB. This process provided distance trees for each enzyme as well as each substrate, which were used to construct the arrangements used in Supplementary Fig. 5.

GT-Predict. *Classifying substrate interactions with quantifiable physicochemical properties.* A DT-based model was trained on various combinations of each substrate's cLogP, molecular volume, solvent accessible area, and carboxylate pK_a. Additionally, structural information such as the number of hydroxyl groups or amines as well as substitution patterns on coumarin, flavonoid, or phenylpropenoid scaffolds (physicochemical parameters, calculated with Chem 3D version 16.0, are listed in Supplementary Tables 8 and 9). GAR scores were input for each enzyme, and classifier programs were written in MATLAB as part of the GT-Predict 'PredictAcceptorInteraction' module. The cross-entropy function was used for the splitting criterion for the branching of the tree. Models were evaluated by determining the accuracy and MCC with leave-one-out cross validation[60,61].

Prediction of novel enzyme activities on the basis of the GAR dataset and alignment. A Smith–Waterman[25]/BLOSUM50 (ref. [62].) pairwise-alignment algorithm was implemented with the GAR scoring matrix in the GT-Predict 'PredictEnzymeInteraction' module. A weighted k-nearest-neighbor approach was used to predict substrate interactions for novel GT1 FASTA amino acid sequences with equation (3) to obtain weighted votes from the closest protein sequences in our dataset and to provide interaction predictions for novel sequences. The top two sequences in our dataset for a novel GT1 amino acid sequence input were used in a weighted vote for prediction, with a 1/'yes' for weighted votes (p_m) > 0.5 or a 0/'no' for p_m < 0.5 (equation (4)).

$$f_m = \frac{\sum_{j=1}^{k} w_j x_{mj}}{\sum_{j=1}^{k} w_j} \tag{3}$$

$$p_m = \left\{ \begin{array}{c} 0\ if\ f_m < 0.5 \\ \hline 1\ if\ f_m \geq 0.5 \end{array} \right\} \tag{4}$$

In equation (3), x_{mj} represents the interaction data for molecule m interacting with the jth nearest neighbor of the enzyme, and is equal to 1 if there is an interaction or 0 if there is not. The results of the prediction were tested against the interaction patterns of experimental GAR screens.

We applied the GT-Predict 'PredictEnzymeInteraction' module to two novel GT1 enzymes from the legume *M. truncatula* and the cereal grain *A. strigosa*. Data for two 'divergent' GT1 sequences from bacterial GT1 enzymes were adapted from our previous screen[28]. Prediction and experimental validation data are shown in Supplementary Fig. 13, and accuracies are tabulated in Supplementary Table 3. Parameters and data from the bacterial enzymes saOleD and slMGT were essentially those from previous studies[28]. Details and validation can be found in the Supplementary Note. Protein accession codes used for prediction were as follows: *M. truncatula* mtUGT71G1, UniProt Q5IFH7; *M. truncatula* mtUGT78G1, UniProt A6XNC6; *A. strigosa* asUGT74H5, GenBank EU496509; *A. strigosa* asUGT88C4, GenBank EU496511; *S. antibioticus* OleD, UniProt Q53685; and *S. lividans* MGT, UniProt Q54387). All alternative GTs were expressed via our Plant GT1 production workflow.

Exploration of other complete families. Two separate and complete GT1 families from *A. strigosa* and *L. barbarum*, respectively, containing candidates given as 'confirmed' in the CAZy 'glycosyltransferases' database[4] were selected for further benchmarking with the 'PredictEnzymeInteraction' module. Each contained ~20–25 validated isozymes. Amino acid sequences were collected from UniProt, DNA-sequence-optimized for production in *E. coli*, and ordered as synthetic gene fragments (Twist Bioscience). GT1 sequences were flanked with restriction sites (N-terminal BamHI and C-terminal EcoRI) for cloning into pGEX-2t) and a C-terminal hexahistidine tag for western blotting and optional purification, although these were used as crude lysates for screening purposes. Fragments are listed in Supplementary Table 5 (*Avena*) and Supplementary Table 6 (*Lycium*). The synthetic gene adaptors 5′-GGATCC–GT1 gene fragment–GCAGCAGCACTGGAACATCATCATCATCATCAT–TAA–GAATTC-3′ (BamHI site–GT1 sequence–linker/hexahistidine tag–stop codon–EcoRI site) were used for all sequences.

GT1 fragments were dissolved in Tris-EDTA buffer, digested with EcoRI and BamHI (New England BioLabs) according to the recommended protocols, and purified with Qiagen PCR Purification Spin columns. The vector pGEX-2t was digested with EcoRI and BamHI, purified on agarose gel, and isolated with Qiagen Gel Purification Spin columns. Ligation was performed with T4 DNA ligase (New England BioLabs) according to the standard overnight 16 °C protocol. All sequences were verified. Of note, a minor number of GT1 gene fragments failed during DNA production or cloning, but 16/18 *Avena* and 16/23 *Lycium* GT1 expression plasmids were verified. The expansion plant GT1s were produced in Rosetta 2 (DE3) pLysS *E. coli* strains according to our standard procedure (briefly, 250 mL terrific broth cultures were grown at 37 °C to OD$_{600}$ $\approx$ 0.6, cooled to 20 °C, and induced for overnight expression with 0.1 mM IPTG and shaking at 140 r.p.m.). Cell pellets were isolated, sonicated, and centrifuged at 12,000*g* for 15 min at 4 °C and then 25,000*g* for 60–90 min at 4 °C. Gels and western blots (with anti-polyhistidine–alkaline phosphatase clone HIS-1, Sigma, A5588) are shown in Supplementary Fig. 8.

GT-Prediction of enzyme interactions and confirmatory screening reactions were performed as above. Aglycones were chosen as the ~40 substrates that showed positive reactivity with at least one GT1 in the *Arabidopsis* collection. The predicted/experimental datasets and summary are shown in Supplementary Figs. 9–11.

Homology-model construction for confirmation of chemical-recognition hypotheses. Structurally characterized Michaelis complexes of GT1 enzymes (either UGT72B1, PDB 2VCE[19], or VvGT1, PDB 2C1Z[18]) were input as templates for homology-model construction with the I-TASSER server[48,51]. Models were aligned to the corresponding structure in COOT[63]. Structural images were created in PyMOL (Schrödinger, v 1.3). Model validations (r.m.s. deviation) are listed in Supplementary Table 7 and fell between 0.73 and 1.25 Å. The physicochemical properties of the acceptor libraries were visualized in the GT-Predict 'AcceptorGUI' module, which highlights associations for each enzyme by property.

Site-directed mutagenesis of UGT71C4 and UGT72C1. Enzyme engineering of the anionic substrate and UDP-GlcNAc activity assays were carried out with the Q5 Site Directed Mutagenesis kit (New England BioLabs) with the following primers. UGT71C4 R292A: forward, 5′-TTTCGGGAGCgcAGGAAGCGTTG-3′; reverse, 5′-CAGAGGAACACCACCGAT-3′. UGT72C1 D180A: forward, 5′-CGGGCTCAAGcTCCGAGAAAATATAT-3′; reverse, 5′-CTCAAACTTAACCGGGCTG-3′. UGT72C1 E187A: forward, 5′-TATATTCGGGcACTCGCTGAG-3′; reverse, 5′-TTTTCTCGGATCTTGAGC-3′. UGT72C1 D180A E187A: forward, 5′-tatattcgggcACTCGCTGAGTCTCAGCG3′; reverse, 5′-ttttctcggagCTTGAGCCCGCTCAAACTTAAC-3′. UGT72C1 G284R: forward, 5′-TTTTGGGAGTagaGGGGCACTAAC-3′; reverse, 5′-GAAACATAAACCACTGACTC-3′.

Mutagenesis reactions were processed according to the manufacturer's protocol. All transformants were confirmed by nucleotide sequencing.

Biotransformation to prepare *trans*-resveratrol-4′-*O*-β-ᴅ-glucopyranoside. Reactions were carried out in aqueous buffer (20 mM Tris, pH 8.0, 40 mM NaCl, 4 mM KCl, and 2 mM MgCl$_2$). A 50-mL Falcon tube was charged with 5.7 mg (25 μmol, 1 equiv.) resveratrol and 15.7 mg (25 μmol, 1 equiv.) UDP-glucose disodium salt. Then 50 mL of cold buffer (to 500 μM final concentration), followed by 500 μL of rapidly thawed GST-UGT73C6 crude lysate, was added, and the samples were stored on ice. The reactions were placed in a 37 °C shaking incubator at 200 r.p.m., then subjected to TLC. (An upright 50-mL Falcon tube is optimal. Too much headspace/shaking precipitates the GT1 catalyst.) Reactions were worked up by extraction five times with 10 mL EtOAc. The organic layer was washed with 50 mL brine, dried over MgSO$_4$, and purified by silica chromatography (2.5 g silica gel, 0% MeOH/CH$_2$Cl$_2$ to 15% MeOH/CH$_2$Cl$_2$) to afford 3.0–3.8 mg product as a pale beige solid (average 34% ± 4% yield over three attempts, $n = 3$) of m.p. 215–223 °C (lit, 210–215 °C). TLC $R_f = 0.22$ in 15% MeOH/CH$_2$Cl$_2$. ^{1}H NMR (d6-acetone, 400 MHz) $\delta = 8.27$ (s, 1 H, phenolic OH), 7.55 (d, $J = 8.8$ Hz, 2 H, H2′, H6′), 7.10−7.02 (m, 3 H, vinylic H, H3′, H5′), 6.98 (d, $J = 16$ Hz, 1 H, vinylic H), 6.59 (d, $J = 2.0$ Hz, 2 H, H2, H6), 6.32 (s, 1 H, H4), 5.01 (d, $J = 7.2$ Hz, 1 H, H1′′), 4.64 (s, 1 H, sugar OH), 4.38 (s, 1 H, sugar OH), 4.32 (s, 1 H, sugar OH), 3.93 (dd, $J = 2.8$ and 14 Hz, 1 H, H6′′A), 3.75 (dd, $J = 2.4$ and 13 Hz, 1 H, H6′′ B), 3.48 (m, 4 H, H2′′, H3′′, H4′′, H5′′). Common solvent impurities at $\delta = 2.88$ (H$_2$O), 2.45 (ethyl methyl ketone), 2.09 (acetone), 1.97 (ethyl acetate), 1.32 and 0.914 ('grease'), and 0.17 (silicone grease) were found, owing to low sample concentration after repeated attempts by HPLC to remove. ^{13}C-NMR (d6-DMSO, 100 MHz) $\delta = 159.0$ (C3, C5), 157.4 (C4′), 139.4 (C-1), 136.8 (C1′), 128.0 (vinylic C), 127.8 (C2′), 127.6 (vinylic C), 116.9 (C3′), 104.9 (C2), 102.5 (C4), 100.8 (C1′′), 77.5 (C2′′), 73.7 (C5′′), 70.2 (C4′′), 61.2 (C6′′). MS (ESI): *m*/*z*: calc for C$_{20}$H$_{21}$O$_8$ [M-H$^+$]: 389.12419; found: 389.12442. IR (neat) $\tilde{\nu} = 3{,}361, 2{,}980, 2{,}402, 1{,}601$ cm^{-1}. The obtained spectroscopic data (Supplementary Fig. 16) were in accordance with those reported in the literature[33,64].

Statistical analyses. Validation of all the predictive models in the paper considered all elements of the confusion matrix, namely the number of positives and negatives predicted that correctly matched the true categories (true positives and true negatives, respectively) as well as positive and negative predictions that are incorrect (false positives and false negatives, respectively). The median percentage accuracy (the accuracy associated with the fiftieth percentile of the accuracies over all data) and the MCC (equation (5)) for each acceptor are plotted in the box-and-whisker plots in Fig. 5; all data reported in Supplementary Table 3 (DT4 model) and in the GT-Predict package are available online.

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

Data and predictive analysis for new enzyme families for *A. strigosa* and *L. barbarum* GT1s can be found in Supplementary Figs. 13 and 14. **A**ll the GAR high-throughput-screening measurements were used as single data points.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Custom code for GT-Predict has been packaged into an executable file compatible with Windows (tested in XP, Windows 7, and Windows 10), which is available through the Oxford University Research Archive at https://doi.org/10.5287/bodleian:zg5195kaE. Installation and use of GT-Predict on Microsoft Windows 10 are shown in Supplementary Video 1.

Data availability

Activity datasets, mass spectrograms, and the protein FASTA sequences used herein are included in a package available through the Oxford University Research Archive at https://doi.org/10.5287/bodleian:zg5195kaE.

References

52. Mackenzie, P. I. et al. Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet. Genomics* **15**, 677–685 (2005).
53. Lim, E.-K. et al. Identification of glucosyltransferase genes involved in sinapate metabolism and lignin synthesis in Arabidopsis. *J. Biol. Chem.* **276**, 4344–4349 (2001).
54. Berthold, M. R. et al. in *Data Anal., Mach. Learn.Appl.: Proc. 31st Annu. Conf. Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007* (eds. Preisach, C. et al.) 319–326 (Springer, Berlin, 2008).
55. Sauer, W. H. B. & Schwarz, M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **43**, 987–1003 (2003).
56. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).
57. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
58. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
59. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
60. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in *IJCAI'95 Proc. 14th Int. Joint Conf. Artif. Intel.* Vol. 2, 1137–1143 (Morgan Kaufmann, San Francisco, 1995).
61. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
62. Pearson, W. R. Selecting the right similarity-scoring matrix. *Curr. Protoc. Bioinformatics* **43**, 5.1–3.5.9 (2013).
63. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
64. Learmonth, D. A. Novel convenient synthesis of the 3-O-β-D- and 4′-O β-D-glucopyranosides of trans-resveratrol. *Synth. Commun.* **34**, 1565–1575 (2004).

nature research

Corresponding author(s): Benjamin G. Davis

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.	No statistical methods were used to predetermine the sample size- we included the entire group of known GT1 enzymes from Arabidopsis thaliana. For revision experiments, all confirmed/validated GT1s from Avena strigosa and Lycium barbarum from the Carbohydrate Active EnZyme (CAZy) bioinformatic database were used. We reasoned that using ALL possible enzymes in this family would provide an unbiased window into their activity for rationalizing and predicting activities.

2. Data exclusions

Describe any data exclusions.	No data was excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.	All data were checked and reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.	All known A. thaliana glycosyltransferase family 1 enzymes were utilized. All confirmed A. strigosa and L. barbarum GT1s were used. No randomization of these samples were necessary, as we used the complete Arabidopsis dataset to train our models using leave-one-out cross validation (i.e. the entire dataset).

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.	Blinding was not necessary in this study. The data were collected and used as the entire grouping and were not separated for analysis. No animal/human participants were used; only enzyme assays.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
☒	□	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	□	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	□	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

ClustalX and ClustalOmega were used for sequence alignment.
A custom executable program was developed for function prediction: GT-Predict (this is shared along with datasets at the link provided in the Data Availability Statement).
COOT (version 0.8) was used to align enzyme structural coordinates and homology models.
MacPyMOL (version 2.0.6) was used to display structures in all Figures.
Knime (version 3.3.2, with the Vernalis plugin) was used for the cheminformatic calculation of principal moments of inertia for the acceptor dataset.
The I-TASSER server was used for homology model construction.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Chemicals were purchased from commercial resources. DNA sequences were originally from one author's laboratory which is now available to purchase by DNA sequence. Plasmids may be requested from Benjamin G. Davis for academic use.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Anti Glutathione S-Transferase Mouse IgG1 (BD Biosciences, clone G172-1138, catalog number 554805, lot number 4163768, dilution 1:1000)
Rabbit Anti-Mouse IgG-alkaline phosphatase fusion (Sigma-Aldrich, polyclonal, catalog number A3562, lot number SLBK3154V, dilution 1:20,000)
Mouse monoclonal anti-polyhistidine IgG--alkaline phosphatase fusion (Sigma, clone HIS-1, catalog number A5588, lot number 085M4836V, dilution 1:5000)
All antibodies used were commercially available with respective documentation.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No human participants were included in this study.