Functional and informatics analysis enables glycosyltransferase activity prediction

Min Yang[#,†¶], Charlie Fehl[#¶], Karen V. Lees[‡], Eng-Kiat Lim[§], Wendy A. Offen[¶], Gideon J. Davies[¶], Dianna J. Bowles[§], Matthew G. Davidson[¢], Stephen J. Roberts[‡], and Benjamin G. Davis[#,*]

[#]Chemistry Research Laboratory, Oxford University, Mansfield Road, Oxford, OX1 3TA, UK

[‡]Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, UK.

[¶]York Structural Biology Laboratory, Department of Chemistry, University of York, York, YO10 5DD, UK

[§]Center for Novel Agricultural Products, Department of Biology, University of York, York, YO10 5DD, UK

[¢]Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

[†]Current address: UCL School of Pharmacy, 29/39 Brunswick Square, London, WC1N1AX, UK

¶ These authors contributed equally

*To whom correspondence should be addressed: ben.davis@chem.ox.ac.uk

Abstract (149 words)

The elucidation and prediction of how changes in a protein give altered activities and selectivities remains a major challenge in chemistry. Two hurdles have prevented accurate family-wide models: i) obtaining 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*. Whilst sequence analysis alone fails for GT1 substrate utilization patterns, our chemical-bioinformatic model, GT-Predict, succeeds by coupling physicochemical features with isozyme recognition patterns over the family. GT-Predict identified GT1 biocatalysts for novel substrates and allowed functional annotation for uncharacterized GT1s. Finally, analyses of GT-Predict decision pathways revealed structural modulators of substrate recognition, informing mechanism. This multifaceted approach to enzyme prediction could guide streamlined utilization (and design) of biocatalysts and discovery of other family-wide protein functions.

Introduction

Subtle evolutionary divergence within a protein family allows 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 amongst which the group-transfer enzyme proteins are of particular interest since several utilize 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 towards (in some cases thousands of) substrates.[3,4] Some have binding sites that are readily understood by virtue of their narrow substrate range (e.g. the lysine acetyltransferases that necessarily bind acetyl CoA and lysine) and hence are easily 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 insights into the evolution of substrate recognition and specificity.[6,7] This is despite high scaffold conservation among GTs,[8] exploited in only select examples,[9] suggesting therefore 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 limit to our understanding 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 CAZy;[4] indeed, the curators of CAZy have highlighted functional prediction as an important future goal.[4]

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

Here we show that functional profiling (**Figure 1**) using broad, unbiased sampling methods of a full GT family present in a single species (the 107-member GT1 family of the plant *Arabidopsis thaliana*) allows construction of chemical-bioinformatic models that encapsulate family-wide recognition patterns for both electrophilic sugar donor and nucleophilic acceptor substrates. We observe extreme scattering in activity patterns as scored by phylogenetic linkage analysis alone, 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 Enzyme Superfamily

To date, informatics or computational strategies for predicting GT1 enzyme activity have made only limited progress, further exacerbated by the small numbers of solved 3-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 function over a 39-enzyme/3-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, gave only sporadically clustered activity throughout the 8-enzyme dataset.[7] We reasoned therefore that any successful approach (**Figure 1**) would, in essence, require 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 *plus* a method allowing the measurement of many (semi-)quantitative activity 'events' unencumbered ('label-free') by structural bias or perturbation (e.g. by virtue of installed chromo-/fluoro-phores[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 *Arabidopsis thaliana* GT1 proteins).

We reasoned that a diverse, unbiased substrate usage coupled with broad, *a priori* examination of properties would allow the primary algorithmic focus to be intentionally generated by protein sequence (**Figure 2A**). We employed a decision tree (DT) learning approach, using a 'deviance' splitting criterion implemented using a cross-entropy function (the optimal score function for classification, being the (negative) log of the multi-nomial probability distribution for correct/incorrect decisions into 1 or K categories). Such strategies advantageously allow interpretable insight into the key parameters (i.e. for the branching of the trees) for successful prediction, if any – essentially allowing us to learn how our putative models learnt. Importantly, in such an approach any lack of statistical power from insufficient breadth in substrate variation or poor choice testing (chemo-/biological) correlate would also be directly revealed by non-robustness or poor performance in the emergent algorithms.

We have previously demonstrated a potentially general, label-free 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 HT/MS 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 algorithmic correlation (**Figure 2B**), thereby allowing mechanistic and predictive insight to emerge regarding both substrates and sequences (**Figure 2C**).

Screening of Diverse Substrates Against an Enzyme Family

GT1 group-transfer enzymes couple two substrates through the transfer to nucleophile 'acceptors' (**1**-**91**) of electrophilic glycosyl 'donor' moieties (**92**-**104**) (**Figure 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 (i.e. hydroxyl replacement) variation in glycosyl moiety of donor; (ii) nucleobase variation in the leaving group moiety of donor; and (iii) nucleophile heteroatom type (O, NH, S) and constitution of scaffold (**Figure 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 would test sufficiency for a predictive model for the GT1 enzyme superfamily (**Supplementary Figure 1**).

Configurational and constitutional alterations of the donor substrate library (**92**-**104**, **Figures 2B, 3 and Supplementary Figure 1**) were designed to explore the logical variation of the glycosyl moiety from a canonical Glc starting point (**Figure 3A**). For example, Glc→Man, Glc→Gal allowed exploration of C-2 and C-4 configuration, respectively; Glc→GlcNAc, Glc→Xyl, Glc→5-*S*-Glc allowed exploration of altered functional groups OH-2→NHAc, CH_2OH-5→H, O-5→S; as well multiply-combined alterations e.g. Glc→Fuc and Glc→Rha (OH-6→H

combined with multisite configurational variation at C-2,3,4,5) intended to provide even greater structural diversity.

Second, the nucleobase moiety of donor substrate was varied (e.g **92, 99, 102**) from canonical pyrimidine uracil (U) in UDP to explore both other pyrmidines (e.g. thymine (T)), Glc-UDP→Glc-dTDP purine (e.g. guanine (G)) usage Glc-UDP→Glc-GDP (**Figure 3A**). This necessitated the creation of unnatural variant donor substrates designed to probe this nucleobase pocket in conjunction with natural variants (e.g. Glc-GDP *cf* Man-GDP, respectively) and variants that are species-specific (e.g. eukaryotic UDP *cf* prokayrotic dTDP).

We designed the nucleophilic acceptor library (**1**-**91**) to probe chemical space (molecular shape, solvent-excluded volumes), electronics (logP ranges, polarity, lone-pair count), and reactivity (nucleophile type) (**Supplementary Figure 1**). Systematic variations in molecular shape (e.g. via hybridization alterations / unsaturations $sp^3 \rightarrow sp^2$; acyclic *vs* 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 utilized to evaluate heteroatom type. Accommodation of heteroatoms in active sites appears, in particular, to be connected with subtle mutations that are not readily understood and predictive understanding might allow the creation of catalysts for the formation of new C–X-bond-types.[19] Diversity measures, based on principal moments of inertia analysis using energy-minimized structures,[20] confirmed a

broad range of rod-like, disk-like, and spherical overall shapes (**Supplementary Figure 1C**).

We conducted a sequential screen to collect datasets for enzyme activity, donor utilization patterns, and acceptor recognition (**Figure 2B**). First, we established initial activity of the full family of 107 *Arabidopsis* GT1 enzymes using canonical, physiologically-relevant[6] plant substrates UDP-D-glucose (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 Figure 2**). This initial survey revealed activity for 54 of the 107 at levels and under conditions that would allow functional screening.

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 revealed 'coarse-grain' interaction patterns for the whole sugar/nucleoside library (**Figure 3A**): 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 utilized 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 for an initial acceptor screen. Unguided, manual classification of the dataset based on some overall structural features (e.g. aliphatics, heterocycles, small aromatic acids, **Figure 3B**) and nucleophilicity patterns (**Figure 3C**) highlighted rough substrate functional group types with broad activity (e.g. polyphenolic compounds) or lower activity

(highly polar glycosides or amino acids). This critically revealed that up to half of these GT1s could use a range of nucleophiles that included 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 (**Figure 2C**)**.** The data were parsed according to threshold activity levels determined by product ion count signal-to-noise. Comparison of these data with the global amino acid sequence alignment of each active enzyme revealed only extremely scattered patterns for both donors and the acceptors (**Figure 4A** and **Supplementary Figures 3-5**), consistent 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 (**Figure 4A**) and the sugar donor library (**Supplementary Figure 3A**). Confirming earlier reports, we observed major discrepancies between related sequences and activities for both the sugar donors and acceptors (**Figure 4A** and **Supplementary Figure 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 failure to observe this within our initial phylogenetic analyses strikingly highlights the seemingly shallow influence of sugar type on the enzymatic evolution of at

least this superfamily of GTs. Our results indicate that nucleotide diphosphate recognition, i.e. for UDP, was conserved; whilst 25% of the GT1s surveyed here used the more structurally similar dTDP, only 7% utilized GDP sugars. This suggests that, while the PSPG motif is useful for identifying UDP-binding regions within GT1s, this motif may fail to account for the recognition events of the carbohydrate portion of sugar nucleotide diphosphates.

Similarly scattered activity patterns were observed for acceptors (full acceptor profile shown in **Supplementary Figures 3B**, **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, roughly half (9/17) of the tested Group E enzymes utilized acid-containing substrates, but this was split into two subgroups over the tree rather than localizing in one defined subgroup, 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 to have subsets of enzymes that process polar heterocyclic rings, were also divergent in overall sequence: the Group D UGT73C6 (see **Online Methods** for nomenclature) and the Group L UGT84A2 have 26.5% identity, 48.5% similarity, and significant 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 loss of substrate recognition patterns and that sequence alignment alone is

therefore not predictive for functional activity.

Next, a wholly sequence-naïve, stepwise analysis allowed 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 for each substrate molecule (**Figure 4B**, **Supplementary Table 1** and **Eqn. 1, Online Methods**). Average linkage clustering (**Eqn. 2, Online Methods**) was then implemented to hierarchically arrange the interaction patterns for enzymes in a sequence-independent fashion (**Figure 4B**, horizontal axis). Notably, such 'activity clustering', guided by each acceptor and donor substrates' interaction patterns with GT1 proteins, allowed some manual classification of meaningful substrate-enzyme subtypes directly, where phylogenetic analysis had wholly failed (**Figure 4B**, horizontal axes). For each substrate library, clustering identified groups of GT1s with, for example, promiscuous donor substrate scopes (towards the right-hand side of **Supplementary Figure 3**) that were unrelated to amino acid similarity or acceptor promiscuity (*c.f*. the right side of **Supplementary Figure 5**).

Excitingly, robust substrate clusters also emerged for acceptor *nucleophiles* (**Figure 4B**) along with substrates with singular recognition patterns that suggested modes of GT1 isoform specialization towards e.g. *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 Allow Algorithmic Prediction.

To correlate and appropriately weight such physicochemical features rigorously, we developed an analytical process that would allow the discovery of overall quantitative structure-activity relationship (QSAR)-based classifiers for the GT1 family. Decision tree-based[22] algorithms were trained on systematically varied combinations of physicochemical properties (cLogP, molecular volume, pK_a) and structural parameters (functional group copy numbers: hydroxyl groups, carboxylic acids, amines) (**Supplementary Table 2**). Emergent algorithms were evaluated using a "leave one out cross-validation" (LOOCV) approach to rank the various models' predictive abilities for each compound and GT1 enzyme (**Figure 5**, **Supplementary Figure 6**,**7** and **Online Methods**). From these, DT4 used a combination of physicochemical inputs (logP, molecular area, solvent-excluded volume, and number/type of nucleophilic groups) and structural information (scaffold type, mono/bi-cyclic variation (5-, 6-membered, [4.3.0], [4.4.0] bicycles, functional groups) that allowed prediction of interactions with 90% ± 1.3% accuracy for our *Arabidopsis* GT1 dataset. Further statistical benchmarking using the Matthews Correlation Coefficient (MCC, **Online Methods**), which analyzes the quality of correlations between -1.0 and +1.0 based on true positive/negative vs. false positive/negative for binary predictions gave 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 confirmed a

strongly positive agreement of predicted and experimental results in a system 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 (see above) to cluster accurately and rationalize GT substrate activity patterns, in striking contrast to the strong correlative success of our substrate physicochemical feature analysis (see 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 suitable influence of substrate-interacting 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 global sequence alignment.[25] This, in principle would allow algorithmic focus upon more relevant (e.g. substrate-interacting) protein regions. Thus, use of the Smith-Waterman algorithm for local sequence alignment[25] allowed us to interrogate novel sequences of GT1 enzymes outside of our dataset using our functionally-characterized enzyme library. To do this efficiently, 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' allowed 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 first able to propose hypothetical activities for putative gene products individually selected from other species (**Figure 6**). First, four, individually-selected, GT1 gene sequences from legume *Medicago truncatula* (*UGT71G1*, *UGT78G1*) and cereal *Avena strigosa* (*UGT74H5*, *UGT88C4*) were analyzed, and the activities of the encoded enzymes (mtUGT71G1, mtUGT78G1, asUGT74H5, asUGT88C4, respectively, see **Online Methods** for use of nomenclature) predicted and then compared with results determined experimentally.[26,27] These revealed (**Figure 6**) 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 Figure 13**); corresponding MCC values were between 0.518-0.910 (**Supplementary Table 3**), indicating very strong to excellent predictive correlation.

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

Finally, as well as its utility against cognate kingdom species from different phyla, GT-Predict was tested 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 (**Figure 6**). Strikingly, despite the sequence divergence and the change of kingdom (plant→bacteria) from the *At* GT1s in our dataset, GT-Predict was 69% (with a positive MCC value of +0.373) accurate for saOleD and 74% (with a positive MCC value of +0.414) for slMGT.

GT-Predict Guides Synthetically-Useful Transformations.

Next, we tested the predictive power of GT-predict on a model compound as 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 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 *Arabidopsis thaliana (At*).[35]

Using GT-Predict we identified several GT1s in the *At*-GT superfamily predicted to hypothetically glycosylate resveratrol as an acceptor nucleophile; usefully these included GTs predicted to also be capable of utilizing a selection of NDP-sugar donor electrophiles, allowing good diversity of elaboration. When experimentally tested *in vitro*, predicted biocatalyst atUGT73C6 proved most efficient from within the enzyme set, allowing regioselective and one-step synthesis of mono-glycosylated resveratrol on a preparative scale

(**Supplementary Figure 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**.[34]

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

GT-Predict Shows Site Features Modulating Selectivity.

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

These, in turn, allowed discovery of intriguing observations and parameter determinants related to possible structural origins for observed activities. For example, activity plots of acid-containing acceptors revealed distinct,

dichotomous ‘allowed *vs* forbidden’ utilization of anionic substrates by GT1 isoforms. These, 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 root-mean-square deviations (RMSDs) of 0.73-1.25 Å (**Supplementary Table 7** and **Online Methods**)).

Unique chemical patterns were investigated to explore three hypothetical ‘drivers’ of substrate recognition for several isozymes. First, the breadth of utilized substrate volume correlates with GT1 active site size (**Supplementary Figure 14A,B**), as judged by mapping the *Accessible Volume* vs. *LogP* – 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 Figure 14C,D**). For example, in carboxylic acid-utilizing GT1 atUGT84A2 (**Supplementary Figure 14D**) this revealed a neutral active site cavity (**Supplementary Figure 14B**). Conversely, this showed that 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 Figure 14C**) and D92/E198 of atUGT73C5 (**Supplementary Figure 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 Figure 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 Figure 14E**), adjacent to the UDP-binding PSPG motif at a distance of 7.4 Å from the C2 substituent nearly optimal for a hydrogen bonding interaction with the *N*-acyl group of GlcNAc. A hydrophobic residue or glycine occupies this position in the remaining Group E GT1s studied. Notably, this arginine substitution was not found to be general among all other plant UDP-GlcNAc utilizing GT1s, highlighting that directed algorithmic functional annotation can suggest rare but functional protein features, perhaps picking up on a unique evolutionary direction taken by an individual isoform within the GT1 family. Other structurally-characterized UDP-GlcNAc-utilizing enzymes also appear to exploit arginine residues to mediate selectivity.[37,38]

The residues pin-pointed by GT-Predict in these 'gating' interactions, namely sites D180/E187 in atUGT72C1 and R292 in atUGT71C4, were experimentally probed using site-directed mutagenesis (**Supplementary Figure 15**). Notably, consistent with drivers implicated by GT-Predict, mutation of Asp/Glu→Ala in atUGT72C1-D180A/E187A enabled activity towards acids (not present in WT) and mutation of Arg→Ala in atUGT71C4-R292A removed the ability to transfer GlcNAc (but not Glc). These not only confirmed the importance of these residues in controlling activity and but also directly highlighted the potential of GT-Predict 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 analyses.[39] Certain predictive attempts have found some success, such as a database of *in silico* docking data compiled for over 100 hydrolase enzyme structures[40] and in the development of a structure-guided metabolomic 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 also in inhibitor design.[44]

Here, a structurally- and phylogenetically-naïve *functional* approach succeeds 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 sufficient to obtain a training set. In this way, combination of an extensive functional dataset and a chemical-bioinformatic analytical method allowed accurate modeling of a full protein family and, indeed, prediction, testing and validation of mechanistic hypotheses and synthetic activities.

As an example of informatically-encapsulating a full protein family, several limitations to this approach should be recognized. First, regiochemical selectivity was not strongly considered when designing GT-Predict, which was based

around presence *vs* absence of chemical groups but not their 3-dimensional orientation. Some limitations can be noted when comparing seemingly highly-related substrates where the relative position of an additional putative nucleophile may give rise to enhanced reactivity (e.g. kaempferol (**23**) >> resveratrol (**105**)). Additional strategies to exploit such regiochemical bias ('substrate fit') might further enhance accuracy[6] (see e.g. **Supplementary Figure 4**). Second, whilst our substrate library proved sufficiently broad for successful training, predictive scope might also be further enhanced by adding database input, for example DrugBank[45] or metabolomic compound collections like the Plant Metabolome Database (PMDB),[46] if sufficiently well curated and tested. Third, GT-Predict now allows the accurate prediction of GT1 activities correlated with local primary sequence alignment, in a manner not possible previously, with greatest accuracy for plant proteins. More advanced secondary structure prediction/alignment methods might be anticipated to extend this yet further (e.g. for low sequence homology but high predicted structural similarity). Similarly, validation of the mechanistic hypotheses suggested by GT-Predict using structural biology[47] would clearly be of direct benefit in augmenting the promising mutagenic results we have obtained here. Given the existence of an excellent database for GTs (and other carbohydrate-processing enzymes) in CAZy,[4] one might even anticipate further refinements and implementations based on this informatics environment.

Given the apparently related structural nature of sugar donors, then it still remains surprising that direct phylogenetic clustering of their utility as substrates

fails. Yet, our results, like those of other studies[7,47,48] show clearly that such analyses alone are not successful and are limited by, for example, sequence variability.[47] This 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 ground-state (e.g. transition state conformation[49]). It is also clear that, nonetheless, physicochemical parameters provide a strong guide that emerges through their striking hierarchical influence upon clustering that we observe here, consistent with recent analyses of the evolution of function within certain conserved folds.[50]

GT-Predict also allows rational selection with some confidence of scaffolds for desired transformations and so might complement some current *de novo* computational design algorithms, which succeed at creating defined packing and active site cavities but can fail on 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 we present here of algorithmically coupling chemical interaction patterns with local sequence analysis might be readily extended to other protein superfamilies that remain currently intransigent toward predictive functional annotation and engineering.

Acknowledgments

We gratefully acknowledge Prof. Anne Osbourne (JIC) for contribution of *Avena strigosa* GT1 genes As08 (UGT74H5) and As09 (UGT88C4), Prof. Robert Edwards and Dr. Melissa Brazier-Hicks for sharing activity data and Dr. Isobel Mear for assistance with coding. This work was funded by the BBSRC (EGA16205, EGA16206, EGA17763) and the EPSRC (The UK Catalysis Hub: EP/K014668/1, EP/M013219/1).

Author Contributions

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

Competing Financial Interests

The authors declare that they have no competing financial interests.

Figure Legends

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

Figure 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, allowing dual substrate library profiling by HT-MS in under 6500 events. (**c**) This dataset was utilized to train decision tree models and validate cheminformatic and bioinformatic algorithms for functional prediction.

Figure 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 colour: blue for UDP, magenta for dTDP, and orange for GDP. (**b**) Acceptor utilization by chemical classification with donor **92**. (**c**) Nucleophile utilization examples from amongst the acceptor library.

Figure 4: **Comparison of clustering techniques for acceptor dataset**. **A** Phylogenetic global sequence analysis of the 54 active GT1s was coupled with the Green-Amber-Red (GAR) screening data heatmap. Activity scores were judged by total ion count (TIC) 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** (**Online Methods**). Hierarchical clustering arrangement on the X-axis is arranged by the similarity of individual GT1 activity patterns against all other GT1s. The tree on the Y-axis is arranged via 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 available in **Supplementary Figures 3-5**; inactive acceptors removed for clarity. All high throughput GAR screening experiments were performed as single measurements.

Figure 5: GT-Predict development, validation and utilization. Diagram of the optimal decision tree (DT4) used to classify information (see **Supplementary Note**). **B** Leave-one-out cross validation of all DT models. Shown is the % accuracy of the trained model for each member of the sugar acceptor library. Dotted error bars indicate the full range of the validation accuracy, with single outliers shown in red crosses determined by ranking predicted vs experimental results for each acceptor that showed activity with at least one GT1 enzyme. Median % accuracy values are shown in red lines for 59 acceptors tested in single measurements via high throughput GAR screening experiments (See **Supplementary Table 3**). The interquartile range (25-75%) are shown in blue boxes. The hashed lines indicate the full range of the dataset. Red crosses are singleton outliers that were not included in the statistics of the box plot but are shown here for completeness. DT1-DT5 are decision tree-based models (see

Supplementary Note, which includes further validation using Matthews Correlation Coefficient analysis). **C** A subset of the GT-Predict results (in the bold box) for compounds kaempferol (**23**) and application to prediction of enzymes for new the substrate resveratrol (**105**) alongside GAR activity for **105** glycosylation with various NDP-sugar substrates. Results confirmed predictions and allowed use of atUGT73C6 for these transformations on a preparative scale (see **Supplementary Note**). The variation in donor utilization by **23** and **105** highlights the essential discovery from DT4 of acceptor hydroxyl functional group (circled) presence (or not) as a key parameter for successful activity prediction for alternative NDP-donor substrates. All GAR screening experiments were performed as single measurements.

Figure 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 Matthews Correlation Coefficient. Further details and analysis are found in the **Supplementary Note**. For other extensions to additional GT families from *Avena strigosa* and *Lycium barbarum* see also **Supplementary Figures 8-11**. Images generated in Pymol from PDB files (2ACB, 3HBF, 2IYF) or models created using I-TASSER.[62] **B** Predicted vs. actual experimental results for acceptor utilization for single enzyme mtUGT78G1 for 38 acceptors tested in singleton high throughput GAR screening experiments (See **Supplementary Figures 8, 9, 13**). **C** Representation of successful *PredictEnzymeInteraction* module, which combines the DT4 model for chemical interaction pattern prediction and ranking with a *k*-Nearest Neighbor (*k*-NN) algorithm for local sequence alignment matching. Coloured dots represent the GT1 training set for the DT4/*k*-NN model. The bold/pink circle represents the novel sequence of interest. The decision trees (DT) represent the activity sets and physicochemical property space of the nearest two GT1s in the training set, which are utilized for activity prediction.

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, doi:10.1006/jmbi.2001.4513 (2001).

2 Gerlt, J. A. & Babbitt, P. C. Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Current opinion in chemical biology* **2**, 607-612 (1998).

3 Friedmann, D. R. & Marmorstein, R. Structure and mechanism of non-histone protein acetyltransferase enzymes. *FEBS J* **280**, 5570-5581, doi:10.1111/febs.12373 (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**, D490-495, doi:10.1093/nar/gkt1178 (2014).

5 Li, T. *et al.* Characterization and Prediction of Lysine (K)-Acetyl-Transferase Specific Acetylation Sites. *Molecular & Cellular Proteomics* **11**, M111.011080-M011111.011080, doi:10.1074/mcp.M111.011080 (2012).

6 Lim, E.-K. *et al.* Evolution of substrate recognition across a multigene family of glycosyltransferases in Arabidopsis. *Glycobiology* **13**, 139-145, doi:10.1093/glycob/cwg017 (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, doi:10.1007/s11103-007-9167-6 (2007).

8 Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annual Review of Biochemistry* **77**, 521-555, doi:10.1146/annurev.biochem.76.061005.092322 (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. *J. Biol. Chem.* **283**, 15724-15731, doi:10.1074/jbc.M801983200 (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. *Current Opinion in Structural Biology* **28**, 131-141, doi:10.1016/j.sbi.2014.08.012 (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 U S A* **108**, 2179-2183, doi:10.1073/pnas.1013331108 (2011).

13 Kries, H., Blomberg, R. & Hilvert, D. De novo enzymes by computational design. *Current opinion in chemical biology* **17**, 221-228, doi:10.1016/j.cbpa.2013.02.012 (2013).

14 Yang, M., Brazier, M., Edwards, R. & Davis, B. G. High-throughput mass-spectroscopy 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, doi:10.1038/nsmb950 (2005).

16 Yang, M., Davies, G. J. & Davis, B. G. A glycosynthase catalyst for the synthesis of flavonoid glycosides. *Angew Chem Int Ed Engl* **46**, 3885-3888, doi:10.1002/anie.200604177 (2007).

17 Backus, K. M. *et al.* Uptake of unnatural trehalose analogs as a reporter for Mycobacterium tuberculosis. *Nature Chemical Biology* **7**, 228-235, doi:doi:10.1038/nchembio.539 (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. *Proceedings of the National Academy of Sciences* **104**, 20238-20243, doi:10.1073/pnas.0706421104 (2007).

20 McLeod, M. C. *et al.* Probing chemical space with alkaloid-inspired libraries. *Nat Chem* **6**, 133-140, doi:10.1038/nchem.1844 (2014).

21 Li, Y., Baldauf, S., Lim, E. K. & Bowles, D. J. Phylogenetic analysis of the UDP-glycosyltransferase multigene family of Arabidopsis thaliana. *Journal Of Biological Chemistry* **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. *Perspectives in Science* **1**, 24-32, doi:10.1016/j.pisc.2014.02.003 (2014).

24 Sanchez-Rodriguez, A. *et al.* A network-based approach to identify substrate classes of bacterial glycosyltransferases. *BMC genomics* **15**, 349, doi:10.1186/1471-2164-15-349 (2014).

25 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195-197, doi:10.1016/0022-2836(81)90087-5 (1981).

26 Shao, H. *et al.* Crystal Structures of a Multifunctional Triterpene/Flavonoid Glycosyltransferase from Medicago truncatula. *Plant Cell* **17**, 3141-3154, doi:10.1105/tpc.105.035055 (2005).

27 Modolo, L. V. *et al.* Crystal Structures of Glycosyltransferase UGT78G1 Reveal the Molecular Basis for Glycosylation and Deglycosylation of (Iso)flavonoids. *Journal of Molecular Biology* **392**, 1292-1302, doi:10.1016/j.jmb.2009.08.017 (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. *Journal of the American Chemical Society* **127**, 9336-9337, doi:10.1021/ja051482n (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, doi:10.1371/journal.pone.0073097 (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, doi:10.1002/pros.23006 (2015).
31 Turner, R. S. *et al.* A randomized, double-blind, placebo-controlled trial of resveratrol for Alzheimer disease. *Neurology* **85**, 1383-1391, doi:10.1212/WNL.0000000000002035 (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, doi:10.1128/AEM.02076-14 (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. Intl Ed.* **45**, 3534-3538, doi:10.1002/anie.200504505 (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. *Natural Product Reports* **26**, 1241-1250, doi:10.1039/B808333A (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 Journal* **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 Journal* **19**, 5269-5280 (2000).
39 Pearson, W. R. Protein Function Prediction: Problems and Pitfalls. *Curr Protoc Bioinformatics* **51**, 4.12.11-18, doi:10.1002/0471250953.bi0412s51 (2015).
40 Tyagi, S. & Pleiss, J. Biochemical profiling in silico—Predicting substrate specificities of large enzyme families. *Journal of Biotechnology* **124**, 108-116, doi:10.1016/j.jbiotec.2006.01.027 (2006).
41 Zhao, S. *et al.* Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* **502**, 698-702, doi:10.1038/nature12576 (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**, doi:10.3390/ijms17060914 (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, doi:10.3109/00498254.2012.663515 (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 Opinion on Drug Discovery* **12**, 769-784, doi:10.1080/17460441.2017.1336157 (2017).
45 Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091-1097, doi:10.1093/nar/gkt1068 (2014).
46 Udayakumar, M. *et al.* PMDB: Plant Metabolome Database—A Metabolomic Approach. *Med Chem Res* **21**, 47-52, doi:10.1007/s00044-010-9506-z (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. *Frontiers in Microbiology* **7**, doi:10.3389/fmicb.2016.00182 (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, doi:10.1016/j.phytochem.2008.12.009 (2009).
49 Davies, G. J., Planas, A. & Rovira, C. Conformational analyses of the reaction coordinate of glycosidases. *Acc Chem Res* **45**, 308-316, doi:10.1021/ar2001765 (2012).
50 Newton, M. S. *et al.* Structural and functional innovations in the real-time evolution of new (βα)8 barrel enzymes. *Proc Natl Acad Sci USA* **114**, 4727-4732, doi:10.1073/pnas.1618552114 (2017).

Online 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, Fisher Scientific) and used without further purification. Sonication was performed using a Fisher Scientific Model 505 Sonic Dismembrator. Proteins were purified using an Äkta FPLC System UPC-900 (GE Healthcare, UK). High-throughput mass spectrometry (HT-MS) was performed using 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 using Invitrogen NuPAGE 4-12% Bis-Tris gels, Novex MiniCell tanks, and a BioRad PowerPac controller. Western blotting was performed using an iBlot gel transfer device from Thermo-Fisher. Thin layer chromatography was performed using Silica Gel 60 F_{254} plates (Merck) using 1-10% methanol in dichloromethane. Nuclear magnetic resonance spectra were recorded on a Bruker AVIII HD 400 nanobay (400MHz) spectrometer. Carbon nuclear magnetic resonance spectra were recorded on a Bruker DQX 400(100 MHz) spectrometer. All ^1H NMR chemical shifts are quoted in ppm using residual solvent as the internal standard relative to TMS (d6-acetone: 2.09 ppm). All ^{13}C NMR chemical shifts are quoted in ppm using the central solvent peak as the internal standard relative to TMS (d6-DMSO 39.3 ppm). Coupling constants (*J*) are reported in Hertz (Hz). Infrared (IR) spectra were recorded on a Bruker Tensor 27 Fourier-Transform spectrophotometer. High-resolution mass spectra were recorded on a Micromass LCT (resolution = 5000 RWHM) using a lock-spray source. Protein crystal structures were analyzed and displayed using MacPyMOL v. 1.3 (Schrodinger, Inc.). Synthetic genes for *Medicago truncatula mtUGT71G1* and *mtUGT78G1* were obtained from GeneArt Gene Synthesis (Thermo-Fisher) using *Escherichia coli* codon-optimized amino acid sequences as reported by

Wang *et al*.[26,27] and sub-cloned into the pGEX2T vector (Amersham Pharmacia Biotech, Chalfont St. Giles, UK) using T4 DNA Ligase (New England BioLabs, Inc.). Mutagenesis was performed with a Q5® Site-Directed Mutagenesis Kit (New England BioLabs). Nucleotide sequencing was confirmed by Source Bioscience DNA Sanger sequencing services of Oxford University (UK).

UGT enzymes are named according to the UGT Nomenclature Committee's latest guidelines[51] as follows: *Arabidopisis 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 *Escherichia coli* expression strains and produced essentially as reported.[6,52] Cells were resuspended in glutathione S-transferase (GST) purification buffer (50 mM Tris, pH 7.4, 1 mM DTT), lysed, centrifuged (10,000 ×*g,* 10 min, 4 °C followed by centrifugation at 25,000 ×*g*, 60 min, 4 °C) and either used as the crude supernatant or taken forward for purification using a Sepharose 4B glutathione resin (GE Healthcare) as described.[52] Western blotting was performed with mouse anti-GST (BD Biosciences) (**Supplementary Figure 2A**). GT1 protein-containing lysates could be flash-frozen and thawed once with activity remaining for up to 6 months' storage at −80 °C.

Green-Amber-Red (GAR) HT-MS Screening.

Activity assays were conducted using reported MS methods[14] on either a Waters Quattro Micro API (ESI^- mode) or a Waters ZMD-MS (ESI^- mode), 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, 50 μM $MgCl_2$), 1 μL of NDP-Sugar (10 mg/mL stock), 1 μL of aglycone (10 mg/mL stock), and 5 μL cell supernatant or purified protein (ca. 1 mg/mL). Glycosylation reactions were incubated at 37 °C overnight and monitored by MS full scan (150-1100 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 was ranked Green (signal/noise > 10), Amber (s/n 1-10), or Red (s/n < 1) from the total ion count integration of the full peak (representative data shown in **Supplementary Figure 2B,C**). The acceptor library is shown in **Supplementary Figure 1** and the full acceptor dataset is shown in **Supplementary Figure 3B**. The full donor dataset is shown in **Supplementary Figure 3A**. Regioselectivities were based on comparison of LC-MS elution time with internal standards as reported[8] or as deduced from substitution patterns within the same chemical families (**Supplementary Figure 4**).

Chemical Diversity Calculations.

Molecular shape calculations were used to design library features that sample a broad range of 3-dimentional chemical space (**Supplementary Figure 1C**). Each structure was energy minimized using 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 using the Knime Analytics Platform[53] with the "SDF Reader"→"PMI Calculation" (Vernalis)→"JavaScript Scatter Plot" nodes and compared to reference molecules for "rod" (octa-2,4,6-triyne), "sphere" (adamantane), and "disk" (benzene).[54] Our compounds were found to lie primarily along the rod-disk axis, but sampled space well into the other principal chemical shape regions.

Clustering of activity based on phylogenetic alignment or functional patterning.

Phylogenetic analyses were performed with CLUSTAL_X[55] or Clustal Omega[56] and fully matched reported analysis for the *Arabidopsis* UGT family.[21] Pairwise alignment was performed using the EMBOSS Water program.[57] Functional activity analysis used hierarchical clustering to score and re-group the acceptors and donors based on GT1 interaction patterns (Green: score of 1.0, Amber: 0.5,

Red: 0.0). Clustering proceeded via average linkage analysis[58] (further details provided in **Supplementary Note**).

Hierarchical Clustering of Activity.

Functional activity analysis used hierarchical clustering to score and re-group the acceptors and donors based on UGT interaction patterns (Green: score of 1.0, Amber/'Unclear': 0.5, Red: 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**.

Equation 1

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

Equation 2

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

Hierarchical arrangement proceeded via average linkage analysis clustering according to **Equation 2** in MATLAB. This provided distance trees for each enzyme as well as each substrate, which were utilized to construct the arrangements used in **Supplementary Figure 5**.

GT-Predict – Classifying substrate interactions using quantifiable on physicochemical properties. A Decision Tree-based model was trained on various combinations of each substrates' cLogP, molecular volume, solvent accessible area, and carboxylate pKa. Additionally, structural information such as number of hydroxyl groups or amines as well as substitution patterns on

coumarin, flavonoid, or phenylpropenoid scaffolds (the physicochemical parameters, calculated using Chem 3D version 16.0, are listed in **Supplementary Tables 8, 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 Matthews correlation coefficient using leave-one-out cross validation.[59,60]

GT-Predict – Prediction of novel enzyme activities based on GAR dataset and alignment.

A Smith-Waterman[25]/BLOSUM50[61] pairwise alignment algorithm was implemented with the GAR scoring matrix in the GT-Predict module "PredictEnzymeInteraction". A weighted k-nearest neighbor approach was used to predict substrate interactions for novel GT1 FASTA amino acid sequences using **Equation 3** to obtain weighted votes from the closest protein sequences in our dataset and provide interaction predictions for novel sequences. The top two sequences in our dataset for a novel GT1 amino acid sequence input are used in a weighted vote for prediction, given a 1/"yes" for weighted votes (p_m) of over 0.5 or a 0/"no" for p_m less than 0.5 (**Equation 4**).

Equation 3

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

Equation 4

$$p_m = \begin{cases} 0 & if \ \ f_m < 0.5 \\ 1 & if \ \ f_m \geq 0.5 \end{cases}$$

In **Equation 3**, x_{mj} is the interaction data for molecule m interacting with the *j*th nearest neighbor of the enzyme, and equals 1 if there is an interaction or 0 if there is not. Results of the prediction were tested against the interaction patterns of experimental GAR screens.

We applied the GT-Predict module "PredictEnzymeInteraction" to two novel GT1 enzymes from the legume *Medicago truncatula* and the cereal grain *Avena strigosa*. Data for two "divergent" GT1 sequences from bacterial GT1 enzymes was adapted from our previous screen.[28] Prediction and experimental validation data are shown in **Supplementary Figure 13** with accuracies tabulated in **Supplementary Table 3**. Parameters and data from bacterial enzymes saOleD and slMGT were essentially those from previous studies.[28] For details and validation see the **Supplementary Note**. Protein accession codes used for prediction: *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), *S. lividans* MGT (UniProt Q94FR0). All alternative GTs were expressed via our Plant GT1 production workflow.

GT-Predict – Exploration of Other Complete Families.

Two separate and complete GT1 families from *Avena strigosa* and *Lycium barbarum*, respectively, containing candidates given as 'confirmed' in the CAZy "Glycosyltransferases" database[4] were selected for further benchmarking of "PredictEnzymeInteraction." Each contain ca. 20-25 validated isozymes. Amino acid sequences were collected from Uniprot, DNA sequence-optimized for production in *Escherichia coli*, and ordered as synthetic gene fragments (Twist Bioscience, San Francisco, USA). GT1 sequences were flanked with restriction sites (*N*-terminal BamHI and *C*-terminal EcoRI) for for subcloning into pGEX-2t and a *C*-terminal hexahistadine tag was added 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*). 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 and digested using EcoRI and BamHI (New England Biolabs) following recommended protocols and purified using Qiagen PCR Purification Spin columns. The vector pGEX-2t was digested with EcoRI and BamHI and purified on agarose gel and isolated using Qiagen Gel Purification Spin columns. Ligation was performed with T4 DNA ligase (New England Biolabs) following the standard overnight 16 °C protocol. All sequences were verified. Note: a minor number of GT1 gene fragments failed during DNA production or subcloning, but 16/18 Avena and 16/23 Lycium GT1 expression plasmid were verified. The expansion plant GT1s were produced in Rosetta 2 (DE3) pLysS *E. coli* strains following our standard procedure (briefly, 250 mL Terrific Broth cultures grown at 37 °C to OD_{600} ≈ 0.6, cooled to 20 °C, and induced for overnight expression with 0.1 mM IPTG and 140 rpm shaking). Cell pellets were isolated, sonicated, centrifuged at 12,000 × *g* for 15 minutes at 4 °C and then 25,000 × *g* for 60-90 minutes at 4 °C. Gels and Western blots (using anti-poly-histidine—alkaline phosphatase clone HIS-1, Sigma cat. number A5588) are shown in **Supplementary Figure 8**.

“GT-Prediction” of EnzymeInteractions and confirmatory screening reactions were performed as above. Aglycones were chosen as the ca. 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 Figures 9-11**.

Homology model construction for confirmation of chemical recognition hypotheses.

Structurally-characterized Michaelis complexes of GT1 enzymes (either UGT72B1, PDB ID: 2VCE[19] or VvGT1, PDB ID: 2C1Z[18]) were input as templates for homology model construction using the I-TASSER server.[48,62] Models were aligned to the corresponding structure in COOT.[63] Structural images were

created in PyMOL (Schrodinger, LLC, Version 1.3). Model validations (RMSD) are listed in **Supplementary Table 7** and fell between 0.73 and 1.25 Å. 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 was carried out using 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'- tatattcgggcACTCGCTGAGTCTCAGCG -3'

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-D-glucopyranoside.

Reactions were carried out in aqueous buffer (20 mM Tris, pH 8.0, 40 mM NaCl, 4 mM KCl, 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. 50 mL of cold buffer was added (to 500 μM final concentration), followed by 500 μL of rapidly-thawed GST-UGT73C6 crude lysate, stored on ice. Reactions were placed in a 37 °C shaking incubator at 200 rpm and followed by t.l.c. (Note: an upright 50 mL Falcon tube is optimal. Too much headspace/shaking precipitates the GT1 catalyst.) Reactions were worked up by extracting 5 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_2Cl_2 to 15% MeOH/CH_2Cl_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). T.L.C. R_f = 0.22 in 15% MeOH/CH_2Cl_2. ^{1}H NMR (d6-acetone, 400 MHz) δ = 8.27 (s, 1H, phenolic OH), 7.55 (d, *J* = 8.8 Hz, 2H, H2', H6',), 7.10–7.02 (m, 3H, vinylic H, H3', H5'), 6.98 (d, *J* = 16 Hz, 1 H, vinylic H), 6.59 (d, *J* = 2.0 Hz, 2H, H2, H6), 6.32 (s, 1H, H4), 5.01 (d, *J* = 7.2 Hz, 1H, H1''), 4.64 (s, 1H, sugar OH), 4.38 (s, 1H, sugar OH), 4.32 (s, 1H, sugar OH), 3.93 (dd, *J* = 2.8 and 14 Hz, 1H, H6''A), 3.75 (dd, *J* = 2.4 and 13 Hz, 1H, H6''B), 3.48 (m, 4H, H2'', H3'', H4'', H5''). Common solvent impurities at δ = 2.88 (H_2O), 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 due to low sample concentration following repeated attempts by HPLC to remove. ^{13}C-NMR (d6-DMSO, 100 MHz) δ = 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) $\tilde{\nu}$ = 3361, 2980, 2402, 1601 cm^{-1}. The obtained spectroscopic data (**Supplementary Figure 16**) were in accordance with those reported in the literature.[64,33]

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 matched correctly the true categories (True Positives – TP, and True Negatives –

TN, respectively) as well as Positive and Negative predictions that are incorrect (False Positives - FP and False Negatives – FN, respectively). The median % accuracy (the accuracy associated with the 50th percentile of the accuracies over all data) and the *Matthews Correlation Coefficient* (MCC, **Equation 5**) for each acceptor are plotted in the box-and-whisker plots in **Figure 5**; all data reported in **Supplementary Table 3** (DT4 model) and in the GT-Predict package is available online.

Equation 5

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

Data and predictive analysis for new enzyme families for *Avena strigosa* and *Lycium barbarum* GT1s is found in **Supplementary Figures 13,14**. All the GAR high-throughput screening measurements were utilized as single data points.

Data and Code Availability.

Custom code for GT-Predict was packaged into an executable file compatible with Windows (XP, Windows 7, and Windows 10 tested), available as a supplementary file through the Oxford University Research Archive DOI: 10.5287/bodleian:zg5195kaE. Activity datasets, mass spectrograms, and the protein FASTA sequences used here are also included in this package.

Online Methods References

51 Mackenzie, P. I. *et al.* Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenetics and genomics* **15**, 677-685 (2005).

52 Lim, E.-K. *et al.* Identification of Glucosyltransferase Genes Involved in Sinapate Metabolism and Lignin Synthesis in Arabidopsis. *Journal of Biological Chemistry* **276**, 4344-4349, doi:10.1074/jbc.M007263200 (2001).

53 Berthold, M. R. *et al.* in *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007* (eds Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, & Reinhold Decker) 319-326 (Springer Berlin Heidelberg, 2008).

54 Sauer, W. H. B. & Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *Journal of Chemical Information and Computer Sciences* **43**, 987-1003, doi:10.1021/ci025599w (2003).

55 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).

56 Sievers, F. *et al.* Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539, doi:10.1038/msb.2011.75 (2011).

57 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

58 Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241-254 (1967).

59 Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. (1995).

60 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442-451, doi:10.1016/0005-2795(75)90109-9 (1975).

61 Pearson, W. R. Selecting the Right Similarity-Scoring Matrix. *Curr Protoc Bioinformatics* **43**, 3.5.1-3.5.9, doi:10.1002/0471250953.bi0305s43 (2013).

62 Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-738, doi:10.1038/nprot.2010.5 (2010).

63 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486-501, doi:10.1107/S0907444910007493 (2010).

64 Learmonth, D. A. A Novel, Convenient Synthesis of the 3‐O‐β‐D‐ and 4′‐O‐β‐D‐Glucopyranosides of trans‐Resveratrol. *Synthetic Communications* **34**, 1565-1575, doi:10.1081/SCC-120030744 (2004).