UNIVERSITY of York

This is a repository copy of *Functional and informatics analysis enables* glycosyltransferase activity prediction.

White Rose Research Online URL for this paper: <u>https://eprints.whiterose.ac.uk/139200/</u>

Version: Accepted Version

Article:

Yang, Min, Fehl, Charlie, Lees, Karen V. et al. (7 more authors) (2018) Functional and informatics analysis enables glycosyltransferase activity prediction. NATURE CHEMICAL BIOLOGY. pp. 1109-1117. ISSN 1552-4450

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

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk https://eprints.whiterose.ac.uk/

1	Functional and informatics analysis enables glycosyltransferase activity
2	prediction
3	
4	Min Yang ^{#,†} ¶, Charlie Fehl [#] ¶, Karen V. Lees [‡] , Eng-Kiat Lim [§] , Wendy A. Offen [¶] ,
5	Gideon J. Davies [¶] , Dianna J. Bowles [§] , Matthew G. Davidson [¢] , Stephen J.
6	Roberts [‡] , and Benjamin G. Davis ^{#,} *
7	
8	[#] Chemistry Research Laboratory, Oxford University, Mansfield Road, Oxford,
9	OX1 3TA, UK
10	[‡] Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, UK.
11	[¶] York Structural Biology Laboratory, Department of Chemistry, University of York,
12	York, YO10 5DD, UK
13	[§] Center for Novel Agricultural Products, Department of Biology, University of York,
14	York, YO10 5DD, UK
15	[¢] Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY,
16	UK
17	[†] Current address: UCL School of Pharmacy, 29/39 Brunswick Square, London,
18	WC1N1AX, UK
19	¶ These authors contributed equally
20	*To whom correspondence should be addressed: ben.davis@chem.ox.ac.uk
21	
22	

page 1

23 Abstract (149 words)

24 The elucidation and prediction of how changes in a protein give altered 25 activities and selectivities remains a major challenge in chemistry. Two hurdles 26 have prevented accurate family-wide models: i) obtaining diverse datasets and ii) 27 suitable parameter frameworks that encapsulate activities in large sets. Here we 28 show that a relatively small but broad activity dataset is sufficient to train 29 algorithms for functional prediction over the entire glycosyltransferase 30 superfamily 1 (GT1) of the plant Arabidopsis thaliana. Whilst sequence analysis 31 alone fails for GT1 substrate utilization patterns, our chemical-bioinformatic 32 model, GT-Predict, succeeds by coupling physicochemical features with isozyme 33 recognition patterns over the family. GT-Predict identified GT1 biocatalysts for 34 novel substrates and allowed functional annotation for uncharacterized GT1s. 35 Finally, analyses of GT-Predict decision pathways revealed structural modulators 36 of substrate recognition, informing mechanism. This multifaceted approach to 37 enzyme prediction could guide streamlined utilization (and design) of biocatalysts 38 and discovery of other family-wide protein functions.

39

41 Introduction

42 Subtle evolutionary divergence within a protein family allows an enormous breadth of functional activities to occur within a versatile core scaffold.^{1,2} The 43 44 reutilization of common scaffolds in the design of *de novo* protein functions is 45 also a current major goal. Several large, architecturally-related protein families 46 are known amongst which the group-transfer enzyme proteins are of particular 47 interest since several utilize multiple modular domains upon which relevant functional groups are evolutionarily-selected.¹ Multiple group transfer enzyme 48 49 superfamilies, including certain acetyltransferases and glycosyltransferases (GTs), share a conserved β -sheet/ α -helical core upon which they exploit variable 50 51 domains to generate selectivity towards (in some cases thousands of) substrates.^{3,4} Some have binding sites that are readily understood by virtue of 52 53 their narrow substrate range (e.g. the lysine acetyltransferases that necessarily 54 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 55 56 vitro unite highly variable substrates and phylogenetic analyses have provided only limited insights into the evolution of substrate recognition and specificity.^{6,7} 57 This is despite high scaffold conservation among GTs,⁸ exploited in only select 58 examples,⁹ suggesting therefore that subtle mutations in the background of these 59 60 scaffolds have profound effects on chemical function. Thus, there remains a 61 general difficulty in understanding the basis for active site plasticity within many enzyme families¹⁰ and GTs in particular represent a striking example of this limit 62 63 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

67 As a primary hurdle, there remains no general informatics strategy to 68 accurately assess functional effects of changes between key features of 69 otherwise similar isoforms of biocatalysts equivalent, for example, to strategies 70 able to model and predict subtle stereoelectronic effects in homogeneous small molecule catalyst performance.¹² Notably *de novo* protein design methods, whilst 71 72 powerfully allowing the creation of rigid structural scaffolds for housing putative 73 function, still fail on the finer details associated with positioning of key catalytic residues.¹³ Therefore, bridging this gap between prediction and structure of 74 75 precise active site features might allow valuable additional insight into the 76 discovery of desired protein functional activities.

77 Here we show that functional profiling (Figure 1) using broad, unbiased 78 sampling methods of a full GT family present in a single species (the 107-79 member GT1 family of the plant Arabidopsis thaliana) allows construction of 80 chemical-bioinformatic models that encapsulate family-wide recognition patterns 81 for both electrophilic sugar donor and nucleophilic acceptor substrates. We 82 observe extreme scattering in activity patterns as scored by phylogenetic linkage 83 analysis alone, confirming that sequence-based assessments cannot explain 84 substrate recognition. However, by incorporating relevant physicochemical 85 parameters such as size, hydrophobicity, and nucleophilicity predictive

- 86 algorithms can be trained to annotate function with high accuracy for these87 promiscuous dual-substrate enzymes.
- 88
- 89

90 Results

91

92 Strategy for Functional Profiling of Enzyme Superfamily

93 To date, informatics or computational strategies for predicting GT1 enzyme 94 activity have made only limited progress, further exacerbated by the small numbers of solved 3-dimensional structures.¹¹ High-confidence phylogenetic 95 trees for a complete GT1 family were previously reported by some of us.⁶ 96 97 wherein a limited set of substrates was tested for common activity. Little 98 correlation was found between primary sequence alignment and enzymatic 99 function over a 39-enzyme/3-coumarin substrate panel probing gains, losses, 100 and regiochemical switching of activity even among closely-related subfamilies. A 101 screen of Medicago truncatula GT1s over 23 benzopyran(one) substrates, 102 similarly, gave only sporadically clustered activity throughout the 8-enzyme dataset.⁷ We reasoned therefore that any successful approach (**Figure 1**) would, 103 104 in essence, require sufficient threshold of unique activity patterns of individual 105 isoforms to be directly coupled with iterative ('learning') algorithms. This 106 functional-informatic method, in turn, would require a sufficiently diverse array of 107 chemical substrate recognition motifs to avoid bias *plus* a method allowing the 108 measurement of many (semi-)quantitative activity 'events' unencumbered ('label-109 free') by structural bias or perturbation (e.g. by virtue of installed chromo-/fluoro-110 phores^{6,7}). The resulting dataset would subsequently be tested for utility in its 111 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).

114 We reasoned that a diverse, unbiased substrate usage coupled with broad, a 115 priori examination of properties would allow the primary algorithmic focus to be 116 intentionally generated by protein sequence (Figure 2A). We employed a 117 decision tree (DT) learning approach, using a 'deviance' splitting criterion 118 implemented using a cross-entropy function (the optimal score function for 119 classification, being the (negative) log of the multi-nomial probability distribution 120 for correct/incorrect decisions into 1 or K categories). Such strategies 121 advantageously allow interpretable insight into the key parameters (i.e. for the 122 branching of the trees) for successful prediction, if any – essentially allowing us 123 to learn how our putative models learnt. Importantly, in such an approach any 124 lack of statistical power from insufficient breadth in substrate variation or poor 125 choice testing (chemo-/biological) correlate would also be directly revealed by 126 non-robustness or poor performance in the emergent algorithms.

127 We have previously demonstrated a potentially general, label-free HT/MS-128 based assay for (semi-)quantitative kinetic characterization of individual enzymes. ¹⁴⁻¹⁷ We considered that, in theory, combining the speed and broad, unbiased 129 130 detection capabilities of this HT/MS assay with proteins from an entire multigene 131 family of GTs, could, for the first time, feasibly catalog a sufficiently diverse 132 chemical dataset from a complete family to allow algorithmic correlation (Figure 133 **2B**), thereby allowing mechanistic and predictive insight to emerge regarding both substrates and sequences (Figure 2C). 134

135

136 Screening of Diverse Substrates Against an Enzyme Family

137 GT1 group-transfer enzymes couple two substrates through the transfer to 138 nucleophile 'acceptors' (1-91) of electrophilic glycosyl 'donor' moieties (92-104) 139 (Figure 2). Electrophilicity is generated in the donor by the presence of a 140 nucleotide diphosphate leaving group. Three corresponding modes of substrate 141 diversity, corresponding to three potential structural selectivity elements were 142 explored: (i) configurational and constitutional (i.e. hydroxyl replacement) 143 variation in glycosyl moiety of donor; (ii) nucleobase variation in the leaving group 144 molety of donor; and (iii) nucleophile heteroatom type (O, NH, S) and constitution 145 of scaffold (Figure 2A). Such an approach is consistent with the few structures of 146 GTs that reveal corresponding pockets and their primary engagement with substrates via these three distinct moieties in Michaelis complexes.^{18,19} In this 147 148 way we were able to create a broad substrate scope that would test sufficiency 149 for a predictive model for the GT1 enzyme superfamily (Supplementary Figure 150 1).

151 Configurational and constitutional alterations of the donor substrate library 152 (92-104, Figures 2B, 3 and Supplementary Figure 1) were designed to explore 153 the logical variation of the glycosyl moiety from a canonical Glc starting point 154 (Figure 3A). For example, Glc \rightarrow Man, Glc \rightarrow Gal allowed exploration of C-2 and 155 C-4 configuration, respectively; Glc \rightarrow GlcNAc, Glc \rightarrow Xyl, Glc \rightarrow 5-*S*-Glc allowed 156 exploration of altered functional groups OH-2 \rightarrow NHAc, CH₂OH-5 \rightarrow H, O-5 \rightarrow S; as 157 well multiply-combined alterations e.g. Glc \rightarrow Fuc and Glc \rightarrow Rha (OH-6 \rightarrow 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 \rightarrow Glc-dTDP purine (e.g. guanine (G)) usage Glc-UDP \rightarrow 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).

167 We designed the nucleophilic acceptor library (1-91) to probe chemical space 168 (molecular shape, solvent-excluded volumes), electronics (logP ranges, polarity, 169 lone-pair count), and reactivity (nucleophile type) (Supplementary Figure 1). 170 Systematic variations in molecular shape (e.g. via hybridization alterations / unsaturations $sp^3 \rightarrow sp^2$; acyclic vs fused/bridged polycyclic substrates) created a 171 172 systematically altered yet diverse range of 'sizes'. Substrate series to reveal 173 electronic effects included acidic, basic, and neutral variations of the same 174 molecular cores. Finally, various O-, NH-, and S-based nucleophiles were utilized 175 to evaluate heteroatom type. Accommodation of heteroatoms in active sites 176 appears, in particular, to be connected with subtle mutations that are not readily 177 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 178 moments of inertia analysis using energy-minimized structures,²⁰ confirmed a 179

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

182 We conducted a sequential screen to collect datasets for enzyme activity. 183 donor utilization patterns, and acceptor recognition (Figure 2B). First, we 184 established initial activity of the full family of 107 Arabidopsis GT1 enzymes using canonical, physiologically-relevant⁶ plant substrates UDP-D-glucose (Glc-UDP, 185 186 donor) with known endogenous plant acceptors 23 and 31 against a panel of 187 GT1 gene-derived lysates expressed in parallel under identical conditions⁶ 188 (Supplementary Figure 2). This initial survey revealed activity for 54 of the 107 189 at levels and under conditions that would allow functional screening.

190 Next, the systematically varied 13-member sugar donor library was screened 191 with the two optimal acceptors (23 and 31) that had shown full activity with Glc-192 UDP over the entire 54-enzyme panel. This revealed 'coarse-grain' interaction 193 patterns for the whole sugar/nucleoside library (Figure 3A): nucleoside 194 component was more stringently regulated, with dTDP utilization (addition of a 195 methyl group) at 25% and GDP (a purine) at only 7.4%. Alternative functional 196 groups at C6, C4, and C2 could be utilized by 28-48% of the GT1 library, 197 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.

206

207 Clustered Functional Trends Are Distinct From Phylogeny.

208 This diverse activity dataset was used as the basis for training chemical-209 bioinformatic classifiers to identify patterns useful for predictive modeling (Figure 210 **2C**). The data were parsed according to threshold activity levels determined by 211 product ion count signal-to-noise. Comparison of these data with the global 212 amino acid sequence alignment of each active enzyme revealed only extremely 213 scattered patterns for both donors and the acceptors (Figure 4A and 214 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 215 216 fitness of biochemical clustering methods for our dataset analysis, we recapitulated the GT1 familial phylogenetic arrangement⁶ for the aglycone 217 218 acceptor library (Figure 4A) and the sugar donor library (Supplementary Figure 219 **3A**). Confirming earlier reports, we observed major discrepancies between 220 related sequences and activities for both the sugar donors and acceptors (Figure 221 **4A** and **Supplementary Figure 3**). Given the suggested, structurally-related nature of sugar donor binding in plant GT1s via the so-called plant secondary 222 product glycosyltransferase (PSPG) motif,²¹ we expected ready clustering. The 223 224 failure to observe this within our initial phylogenetic analyses strikingly highlights 225 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.

232 Similarly scattered activity patterns were observed for acceptors (full acceptor 233 profile shown in Supplementary Figures 3B, 4). However, some pockets of 234 conserved function could be assigned, at least partially, to phylogenetic 235 groupings. First, polyphenolic flavonoids and coumarins were widely used 236 throughout the GT1 panel. Small aromatic acids also made up a significant 237 activity group, albeit scattered throughout the phylogenetic classes. For instance, 238 roughly half (9/17) of the tested Group E enzymes utilized acid-containing 239 substrates, but this was split into two subgroups over the tree rather than 240 localizing in one defined subgroup, suggesting that overall amino acid 241 conservation is not the major driver of substrate recognition. The Group D and 242 Group L enzymes, the only two groups to have subsets of enzymes that process 243 polar heterocyclic rings, were also divergent in overall sequence: the Group D 244 UGT73C6 (see **Online Methods** for nomenclature) and the Group L UGT84A2 245 have 26.5% identity, 48.5% similarity, and significant gaps (18.6% of the 246 sequence), for example. Our results thus bolster the earlier hypotheses⁶ that 247 parallel independent evolutionary events have led to both the frequent acquisition 248 and loss of substrate recognition patterns and that sequence alignment alone is

therefore not predictive for functional activity.

250 Next, a wholly sequence-naïve, stepwise analysis allowed activity-based 251 clustering of GT1 isoforms and elucidation of common functional patterns from 252 within the superfamily. First, threshold activities were used to assign activity 253 commonality (full, partial, or no-activity) between each enzyme for each substrate 254 molecule (Figure 4B, Supplementary Table 1 and Eqn. 1, Online Methods). 255 Average linkage clustering (Eqn. 2, Online Methods) was then implemented to 256 hierarchically arrange the interaction patterns for enzymes in a sequence-257 independent fashion (Figure 4B, horizontal axis). Notably, such 'activity 258 clustering', guided by each acceptor and donor substrates' interaction patterns 259 with GT1 proteins, allowed some manual classification of meaningful substrate-260 enzyme subtypes directly, where phylogenetic analysis had wholly failed (**Figure** 261 **4B**, horizontal axes). For each substrate library, clustering identified groups of 262 GT1s with, for example, promiscuous donor substrate scopes (towards the right-263 hand side of **Supplementary Figure 3**) that were unrelated to amino acid 264 similarity or acceptor promiscuity (c.f. the right side of **Supplementary Figure 5**). 265 Excitingly, robust substrate clusters also emerged for acceptor *nucleophiles* 266 (Figure 4B) along with substrates with singular recognition patterns that 267 suggested modes of GT1 isoform specialization towards e.g. N-heterocycles, 268 bulky fused aliphatic ring systems, and polar glycosides. This 'chemical 269 clustering', which emerged *without* the input of *any* physicochemical or structural 270 information, importantly revealed the strong influence of substrate chemical 271 properties as major drivers of substrate recognition in the GT1 superfamily.

272

273 Physicochemical Analyses Allow Algorithmic Prediction.

274 To correlate and appropriately weight such physicochemical features 275 rigorously, we developed an analytical process that would allow the discovery of 276 overall quantitative structure-activity relationship (QSAR)-based classifiers for the GT1 family. Decision tree-based²² algorithms were trained on systematically 277 278 varied combinations of physicochemical properties (cLogP, molecular volume, 279 pK_a) and structural parameters (functional group copy numbers: hydroxyl groups, 280 carboxylic acids, amines) (**Supplementary Table 2**). Emergent algorithms were 281 evaluated using a "leave one out cross-validation" (LOOCV) approach to rank the 282 various models' predictive abilities for each compound and GT1 enzyme (Figure 283 5, Supplementary Figure 6,7 and Online Methods). From these, DT4 used a 284 combination of physicochemical inputs (logP, molecular area, solvent-excluded 285 volume, and number/type of nucleophilic groups) and structural information 286 (scaffold type, mono/bi-cyclic variation (5-, 6-membered, [4.3.0], [4.4.0] bicycles, 287 functional groups) that allowed prediction of interactions with $90\% \pm 1.3\%$ 288 accuracy for our Arabidopsis GT1 dataset. Further statistical benchmarking using 289 the Matthews Correlation Coefficient (MCC, **Online Methods**), which analyzes 290 the quality of correlations between -1.0 and +1.0 based on true positive/negative 291 vs. false positive/negative for binary predictions gave an average value of 0.591 292 for the DT4 model over all 59 acceptor molecules with experimental and/or 293 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*.

296

297 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.

305 The clear driving influence of substrate features that we observed suggested 306 that a focused analysis of salient, corresponding protein features would allow 307 suitable influence of substrate-interacting regions in an unbiased manner. Local 308 sequence alignment can be used to rank short, highly-similar regions while 309 ignoring large gaps or regions of sequence divergence more effectively than global sequence alignment.²⁵ This, in principle would allow algorithmic focus 310 311 upon more relevant (e.g. substrate-interacting) protein regions. Thus, use of the Smith-Waterman algorithm for local sequence alignment²⁵ allowed us to 312 313 interrogate novel sequences of GT1 enzymes outside of our dataset using our 314 functionally-characterized enzyme library. To do this efficiently, we developed a 315 program to perform combined local alignment and BLOSUM50 scoring of the 316 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.

320 In this way, GT-Predict was first able to propose hypothetical activities for 321 putative gene products individually selected from other species (**Figure 6**). First, 322 four, individually-selected, GT1 gene sequences from legume Medicago 323 truncatula (UGT71G1, UGT78G1) and cereal Avena strigosa (UGT74H5, 324 UGT88C4) were analyzed, and the activities of the encoded enzymes 325 (mtUGT71G1, mtUGT78G1, asUGT74H5, asUGT88C4, respectively, see **Online** 326 **Methods** for use of nomenclature) predicted and then compared with results determined experimentally.^{26,27} These revealed (**Figure 6**) an 85-92% accuracy 327 328 (Supplementary Table 4) for GT-Predict when tested against the subset of 44 329 substrates that demonstrated robust activity in the Arabidopsis dataset 330 (Supplementary Figure 13); corresponding MCC values were between 0.518-331 0.910 (Supplementary Table 3), indicating very strong to excellent predictive 332 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 differentphyla, 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 sIMGT,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 sIMGT.

346

347 GT-Predict Guides Synthetically-Useful Transformations.

348 Next, we tested the predictive power of GT-predict on a model compound as 349 potential substrate. Resveratrol (105) is an antioxidant and pan-histone deacetylase inhibitor²⁹ currently in clinical trials for cancer prevention³⁰ and 350 neurodegenerative disease.³¹ Its poor solubility as free drug³² has prompted 351 352 investigation into the production of resveratrol glycosides to improve its pharmacological properties.^{33,34} Moreover, for the purposes of validating GT-353 354 Predict, resveratrol is endogenous only to berry-producing plant species, but is not found in Arabidopsis thaliana (At).³⁵ 355

356 Using GT-Predict we identified several GT1s in the At-GT superfamily 357 predicted to hypothetically glycosylate resveratrol as an acceptor nucleophile; 358 usefully these included GTs predicted to also be capable of utilizing a selection of 359 NDP-sugar donor electrophiles, allowing good diversity of elaboration. When 360 experimentally tested in vitro, predicted biocatalyst atUGT73C6 proved most 361 efficient from within the enzyme set, allowing regioselective and one-step 362 synthesis of mono-glycosylated resveratrol on a preparative scale 363 (**Supplementary Figure 12**). Notably and importantly, these *in vitro* results 364 confirmed elegant results previously determined when the *Arabidopsis* GTs were 365 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 antibiotic³⁶ biosynthesized by *Streptomyces niveus*, thereby demonstrating predictive activity discovery for not only non-endogenous substrates but even those outside of normal plant metabolism.

371

372 GT-Predict Shows Site Features Modulating Selectivity.

373 Structural guidance insight remains a vital aspect for hypothesis-driven insight into biocatalyst mechanism and enzyme engineering.¹⁹ Whilst GT-Predict is 374 375 founded on a comprehensive *functional* dataset, its use in conjunction with 376 structural approaches also allowed identification of possibly important structural 377 motifs and their roles within active sites. This was aided by a combined 378 visualization tool and graphical user interface that highlighted patterns based on 379 physicochemical property analyses (Supplementary Figure 14). In this way, for 380 example, given acceptor substrates for a particular GT1 enzyme could be related 381 to any two chosen chemical properties vs functional activity in three-dimensional 382 plots (**Supplementary Figure 14**) to allow interrogation of emergent correlations. 383 These, in turn, allowed discovery of intriguing observations and parameter 384 determinants related to possible structural origins for observed activities. For 385 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-Predictguided 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-meansquare deviations (RMSDs) of 0.73-1.25 Å (**Supplementary Table 7** and **Online Methods**)).

393 Unique chemical patterns were investigated to explore three hypothetical 394 'drivers' of substrate recognition for several isozymes. First, the breadth of 395 utilized substrate volume correlates with GT1 active site size (Supplementary 396 **Figure 14A,B**), as judged by mapping the Accessible Volume vs. LogP - a397 surrogate for molecular surfaces – in the crystallized (atUGT72B1) or modeled 398 (asUGT84A2) active sites. Second, selection of negatively-charged substrates 399 (at pH 8.0) involves either engagement by cationic active site residue motifs 400 and/or gating by anionic residue motifs **Supplementary Figure 14C,D**). For 401 example, in carboxylic acid-utilizing GT1 atUGT84A2 (Supplementary Figure 402 14D) this revealed a neutral active site cavity (Supplementary Figure 14B). 403 Conversely, this showed that in two GT1s not able to glycosylate acids, 404 atUGT72C1 and atUGT73C5, each displayed negatively-charged 'gates' 405 composed of two acidic residues near the proposed substrate access cleft: 406 D180/E187 of atUGT72C1 (Supplementary Figure 14C) and D92/E198 of 407 atUGT73C5 (Supplementary Figure 15). Third, the utilization of sugar donors is 408 modulated by the recognition of larger, polar substituents through hydrogen

page 19

409 bonding to polar amino acids in accommodating pockets (Supplementary 410 Figure 14E). For example, the use by atUGT71C4 of more bulky, polar UDP-411 GlcNAc donor substrate correlated with a unique arginine residue at position 292 412 (Supplementary Figure 14E), adjacent to the UDP-binding PSPG motif at a 413 distance of 7.4 Å from the C2 substituent nearly optimal for a hydrogen bonding 414 interaction with the N-acyl group of GlcNAc. A hydrophobic residue or glycine 415 occupies this position in the remaining Group E GT1s studied. Notably, this 416 arginine substitution was not found to be general among all other plant UDP-417 GlcNAc utilizing GT1s, highlighting that directed algorithmic functional annotation 418 can suggest rare but functional protein features, perhaps picking up on a unique 419 evolutionary direction taken by an individual isoform within the GT1 family. Other 420 structurally-characterized UDP-GlcNAc-utilizing enzymes also appear to exploit 421 arginine residues to mediate selectivity.^{37,38}

422 The residues pin-pointed by GT-Predict in these 'gating' interactions, namely 423 sites D180/E187 in atUGT72C1 and R292 in atUGT71C4, were experimentally 424 probed using site-directed mutagenesis (Supplementary Figure 15). Notably, 425 consistent with drivers implicated by GT-Predict, mutation of Asp/Glu→Ala in 426 atUGT72C1-D180A/E187A enabled activity towards acids (not present in WT) 427 and mutation of Arg \rightarrow Ala in atUGT71C4-R292A removed the ability to transfer 428 GlcNAc (but not Glc). These not only confirmed the importance of these residues 429 in controlling activity and but also directly highlighted the potential of GT-Predict 430 in rational enzyme engineering.

432 Discussion

433

434 Comprehensive predictive modeling of enzyme superfamilies has remained 435 an unsolved challenge despite advances in genomics, proteomics, and 436 metabolomic data gathering and analyses.³⁹ Certain predictive attempts have 437 found some success, such as a database of *in silico* docking data compiled for over 100 hydrolase enzyme structures⁴⁰ and in the development of a structure-438 439 guided metabolomic prediction system to annotate new protein functions.⁴¹ 440 However, these approaches to-date have been confined to proteins of known 441 structure and with relatively narrow substrate variation. Substrate utilization and 442 chemical properties have been linked to generate QSAR-based predictive models for individual proteins from large protein families^{42,43} and have long been 443 444 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.

452 As an example of informatically-encapsulating a full protein family, several 453 limitations to this approach should be recognized. First, regiochemical selectivity 454 was not strongly considered when designing GT-Predict, which was based 455 around presence vs absence of chemical groups but not their 3-dimensional 456 orientation. Some limitations can be noted when comparing seemingly highly-457 related substrates where the relative position of an additional putative 458 nucleophile may give rise to enhanced reactivity (e.g. kaempferol (23) >> 459 resveratrol (105)). Additional strategies to exploit such regiochemical bias ('substrate fit') might further enhance accuracy⁶ (see e.g. **Supplementary Figure** 460 461 4). Second, whilst our substrate library proved sufficiently broad for successful 462 training, predictive scope might also be further enhanced by adding database input, for example DrugBank⁴⁵ or metabolomic compound collections like the 463 Plant Metabolome Database (PMDB),⁴⁶ if sufficiently well curated and tested. 464 465 Third, GT-Predict now allows the accurate prediction of GT1 activities correlated 466 with local primary sequence alignment, in a manner not possible previously, with 467 greatest accuracy for plant proteins. More advanced secondary structure 468 prediction/alignment methods might be anticipated to extend this yet further (e.g. 469 for low sequence homology but high predicted structural similarity). Similarly, 470 validation of the mechanistic hypotheses suggested by GT-Predict using structural biology⁴⁷ would clearly be of direct benefit in augmenting the promising 471 472 mutagenic results we have obtained here. Given the existence of an excellent database for GTs (and other carbohydrate-processing enzymes) in CAZy,⁴ one 473 474 might even anticipate further refinements and implementations based on this 475 informatics environment.

476 Given the apparently related structural nature of sugar donors, then it still 477 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 478 479 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 480 481 enzymatic evolution of, at least this superfamily, of GTs and/or the guidance of 482 selectivity by other parameters that are not defined by ground-state (e.g. conformation⁴⁹). It is also 483 state clear transition that. nonetheless, 484 physicochemical parameters provide a strong guide that emerges through their 485 striking hierarchical influence upon clustering that we observe here, consistent with recent analyses of the evolution of function within certain conserved folds.⁵⁰ 486

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.

498

499

502 Acknowledgments

503 We gratefully acknowledge Prof. Anne Osbourne (JIC) for contribution of *Avena* 504 *strigosa* GT1 genes As08 (UGT74H5) and As09 (UGT88C4), Prof. Robert 505 Edwards and Dr. Melissa Brazier-Hicks for sharing activity data and Dr. Isobel 506 Mear for assistance with coding. This work was funded by the BBSRC 507 (EGA16205, EGA16206, EGA17763) and the EPSRC (The UK Catalysis Hub: 508 EP/K014668/1, EP/M013219/1).

509 510

511 Author Contributions

512 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.

513 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.

516

517 Competing Financial Interests

518 The authors declare that they have no competing financial interests.

519

520 521

522

524 Figure Legends

525 526

527 Figure 1. Challenges and solutions for the rational prediction of 528 **multisubstrate enzyme reactions.** (a) The glycosyltransferase GT1 superfamily 529 couples electrophilic sugars with nucleophilic acceptors. These reactions span 530 the full metabolome with many permutations, rendering current screening and 531 prior informatics approaches insufficient for comprehensive predictive modeling. 532 (b) Our function-based algorithmic learning approach, GT-Predict, utilizes a 533 diverse training set of enzymes, electrophiles, and nucleophiles to create a 534 physicochemical and local-sequenced based classifier for prediction of novel 535 transformations and functional annotation of glycosyltransferase group transfer 536 enzymes.

537

538

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.

546

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.

554

555 Figure 4: Comparison of clustering techniques for acceptor dataset. A 556 Phylogenetic global sequence analysis of the 54 active GT1s was coupled with 557 the Green-Amber-Red (GAR) screening data heatmap. Activity scores were 558 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 559 560 clustering via average linkage analysis according to Equation 1 and Equation 2 561 (**Online Methods**). Hierarchical clustering arrangement on the X-axis is arranged 562 by the similarity of individual GT1 activity patterns against all other GT1s. The 563 tree on the Y-axis is arranged via the association patterns of each substrate with 564 the overall GT1 enzyme library against the other substrates' patterns. Chemical 565 groupings refer to the emergent interaction similarity clusters as discussed in the 566 text. Full datasets available in **Supplementary Figures 3-5**; inactive acceptors 567 removed for clarity. All high throughput GAR screening experiments were 568 performed as single measurements.

- 569
- 570

571 Figure 5: GT-Predict development, validation and utilization. Diagram of the optimal decision tree (DT4) used to classify information (see Supplementary 572 573 Note). B Leave-one-out cross validation of all DT models. Shown is the % 574 accuracy of the trained model for each member of the sugar acceptor library. 575 Dotted error bars indicate the full range of the validation accuracy, with single 576 outliers shown in red crosses determined by ranking predicted vs experimental 577 results for each acceptor that showed activity with at least one GT1 enzyme. 578 Median % accuracy values are shown in red lines for 59 acceptors tested in 579 single measurements via high throughput GAR screening experiments (See 580 **Supplementary Table 3).** The interguartile range (25-75%) are shown in blue 581 boxes. The hashed lines indicate the full range of the dataset. Red crosses are 582 singleton outliers that were not included in the statistics of the box plot but are 583 shown here for completeness. DT1-DT5 are decision tree-based models (see 584 **Supplementary Note**, which includes further validation using Matthews 585 Correlation Coefficient analysis). C A subset of the GT-Predict results (in the bold 586 box) for compounds kaempferol (23) and application to prediction of enzymes for 587 new the substrate resveratrol (105) alongside GAR activity for 105 glycosylation 588 with various NDP-sugar substrates. Results confirmed predictions and allowed 589 use of atUGT73C6 for these transformations on a preparative scale (see 590 Supplementary Note). The variation in donor utilization by 23 and 105 highlights 591 the essential discovery from DT4 of acceptor hydroxyl functional group (circled) 592 presence (or not) as a key parameter for successful activity prediction for 593 alternative NDP-donor substrates. All GAR screening experiments were 594 performed as single measurements.

- 595
- 596

597 Figure 6: GT-Predict extends functional annotation to other species, 598 kingdoms, and GT families.

599 A Summary of *GT-Predict* prediction results for six selected individual enzymes 600 from differing species, including accuracy and Matthews Correlation Coefficient. 601 Further details and analysis are found in the **Supplementary Note**. For other 602 extensions to additional GT families from Avena strigosa and Lycium barbarum 603 see also Supplementary Figures 8-11. Images generated in Pymol from PDB 604 files (2ACB, 3HBF, 2IYF) or models created using I-TASSER.⁶² **B** Predicted vs. 605 actual experimental results for acceptor utilization for single enzyme mtUGT78G1 606 for 38 acceptors tested in singleton high throughput GAR screening experiments 607 (See Supplementary Figures 8, 9, 13). C Representation of successful 608 *PredictEnzymeInteraction* module, which combines the DT4 model for chemical 609 interaction pattern prediction and ranking with a k-Nearest Neighbor (k-NN) 610 algorithm for local sequence alignment matching. Coloured dots represent the 611 GT1 training set for the DT4/k-NN model. The bold/pink circle represents the 612 novel sequence of interest. The decision trees (DT) represent the activity sets 613 and physicochemical property space of the nearest two GT1s in the training set, 614 which are utilized for activity prediction.

615 **References**

- Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in
 protein superfamilies, from a structural perspective. *J Mol Biol* **307**, 11131143, doi:10.1006/jmbi.2001.4513 (2001).
- 619 2 Gerlt, J. A. & Babbitt, P. C. Mechanistically diverse enzyme superfamilies:
 620 the importance of chemistry in the evolution of catalysis. *Current opinion in chemical biology* 2, 607-612 (1998).
- Friedmann, D. R. & Marmorstein, R. Structure and mechanism of nonhistone protein acetyltransferase enzymes. *FEBS J* 280, 5570-5581,
 doi:10.1111/febs.12373 (2013).
- 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).
- Li, T. *et al.* Characterization and Prediction of Lysine (K)-AcetylTransferase Specific Acetylation Sites. *Molecular & Cellular Proteomics* 11, M111.011080-M011111.011080, doi:10.1074/mcp.M111.011080 (2012).
- 631 6 Lim, E.-K. *et al.* Evolution of substrate recognition across a multigene
 632 family of glycosyltransferases in Arabidopsis. *Glycobiology* 13, 139-145,
 633 doi:10.1093/glycob/cwg017 (2003).
- 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).
- 6378Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G.638Glycosyltransferases: Structures, Functions, and Mechanisms. Annual639Review of Biochemistry 77, 521-555,640doi:10.1146/annurev.biochem.76.061005.092322 (2008).
- 641 9 Cartwright, A. M., Lim, E.-K., Kleanthous, C. & Bowles, D. J. A Kinetic 642 Analysis of Regiospecific Glucosylation by Two Glycosyltransferases of 643 Arabidopsis thaliana. *J. Biol. Chem.* **283**, 15724-15731, 644 doi:10.1074/jbc.M801983200 (2008).
- 64510Todd, A. E., Orengo, C. A. & Thornton, J. M. Plasticity of enzyme active646sites. Trends Biochem Sci 27, 419-426 (2002).
- 647 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).
- 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).
- 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).
- 657 14 Yang, M., Brazier, M., Edwards, R. & Davis, B. G. High-throughput mass-658 spectroscopy monitoring for multisubstrate enzymes: Determining the

- 659 kinetic parameters and catalytic activities of glycosyltransferases. 660 *ChemBioChem* **6**, 346-357 (2005).
- 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).
- Kang, 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).
- 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).
- 670 18 Offen, W. *et al.* Structure of a flavonoid glucosyltransferase reveals the 671 basis for plant natural product modification. *EMBO J.* **25**, 1396-1405 672 (2006).
- 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).
- 677 20 McLeod, M. C. *et al.* Probing chemical space with alkaloid-inspired 678 libraries. *Nat Chem* **6**, 133-140, doi:10.1038/nchem.1844 (2014).
- 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).
- 682 22 Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification 683 and regression trees. Wadsworth & Brooks. *Monterey, CA* (1984).
- 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).
- 687 24 Sanchez-Rodriguez, A. *et al.* A network-based approach to identify
 688 substrate classes of bacterial glycosyltransferases. *BMC genomics* 15,
 689 349, doi:10.1186/1471-2164-15-349 (2014).
- 690 25 Smith, T. F. & Waterman, M. S. Identification of common molecular
 691 subsequences. *Journal of Molecular Biology* 147, 195-197,
 692 doi:10.1016/0022-2836(81)90087-5 (1981).
- 693 26 Shao, H. *et al.* Crystal Structures of a Multifunctional Triterpene/Flavonoid
 694 Glycosyltransferase from Medicago truncatula. *Plant Cell* **17**, 3141-3154,
 695 doi:10.1105/tpc.105.035055 (2005).
- 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).
- 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).

- 70429Venturelli, 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).
- 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).
- 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.0000000002035 (2015).
- 71532Tomé-Carneiro, J. *et al.* Resveratrol and clinical trials: the crossroad from716in vitro studies to human evidence. *Curr. Pharm. Des.* **19**, 6064-6093717(2013).
- 718 33 Pandey, R. P. *et al.* Enzymatic Biosynthesis of Novel Resveratrol
 719 Glucoside and Glycoside Derivatives. *Appl. Environ. Microbiol.* **80**, 7235720 7243, doi:10.1128/AEM.02076-14 (2014).
- 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).
- 72535Burns, J., Yokota, T., Ashihara, H., Lean, M. E. & Crozier, A. Plant foods726and herbal sources of resveratrol. J Agric Food Chem 50, 3337-3340727(2002).
- 72836Heide, L. The aminocoumarins: biosynthesis and biology. Natural Product729Reports 26, 1241-1250, doi:10.1039/B808333A (2009).
- 730 37 Peneff, C. *et al.* Crystal structures of two human pyrophosphorylase
 role of the alternatively
 role of the alternatively
 spliced insert in the enzyme oligomeric assembly and active site
 architecture. *Embo Journal* **20**, 6191-6202 (2001).
- 734 38 X-ray crystal of Unligil, U. Μ. et al. structure rabbit N-735 acetylglucosaminyltransferase I: catalytic mechanism and a new protein 736 superfamily. Embo Journal 19, 5269-5280 (2000).
- 737 39 Pearson, W. R. Protein Function Prediction: Problems and Pitfalls. *Curr* 738 *Protoc Bioinformatics* 51, 4.12.11-18, doi:10.1002/0471250953.bi0412s51
 739 (2015).
- Tyagi, S. & Pleiss, J. Biochemical profiling in silico—Predicting substrate
 specificities of large enzyme families. *Journal of Biotechnology* **124**, 108116, doi:10.1016/j.jbiotec.2006.01.027 (2006).
- 743 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).
- 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).

- 749 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).
- 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).
- 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).
- 75946Udayakumar, M. et al. PMDB: Plant Metabolome Database—A760Metabolomic Approach. Med Chem Res 21, 47-52, doi:10.1007/s00044-761010-9506-z (2012).
- 762 47 Schmid, J., Heider, D., Wendel, N. J., Sperl, N. & Sieber, V. Bacterial
 763 Glycosyltransferases: Challenges and Opportunities of a Highly Diverse
 764 Enzyme Class Toward Tailoring Natural Products. *Frontiers in*765 *Microbiology* 7, doi:10.3389/fmicb.2016.00182 (2016).
- 766 48 Osmani, S. A., Bak, S. & Møller, B. L. Substrate specificity of plant UDP767 dependent glycosyltransferases predicted from crystal structures and
 768 homology modeling. *Phytochemistry* **70**, 325-347,
 769 doi:10.1016/j.phytochem.2008.12.009 (2009).
- 770 49 Davies, G. J., Planas, A. & Rovira, C. Conformational analyses of the 771 reaction coordinate of glycosidases. *Acc Chem Res* **45**, 308-316, 772 doi:10.1021/ar2001765 (2012).
- 77350Newton, M. S. *et al.* Structural and functional innovations in the real-time774evolution of new ($\beta\alpha$)8 barrel enzymes. *Proc Natl Acad Sci USA* **114**,7754727-4732, doi:10.1073/pnas.1618552114 (2017).
- 776 777
- 778
- 779
- 780
- _
- 781
- 782

783 Online Methods

784

785 General Considerations.

786 Unless otherwise noted, chemical reagents, media, and bacterial cell stocks were 787 obtained from commercial suppliers (Sigma-Aldrich, Fluorochem, Carbosynth, 788 VWR, Alfa Aesar, Fisher Scientific) and used without further purification. 789 Sonication was performed using a Fisher Scientific Model 505 Sonic 790 Dismembrator. Proteins were purified using an Äkta FPLC System UPC-900 (GE 791 Healthcare, UK). High-throughput mass spectrometry (HT-MS) was performed 792 using either a Waters Quattro Micro API (ESI⁻ mode) or a Waters ZMD-MS (ESI⁻ 793 mode) detector, each equipped with a Waters 600 HPLC System and a Waters 794 2700 autosampler capable of 96-well sampling format. Gel electrophoresis was 795 performed using Invitrogen NuPAGE 4-12% Bis-Tris gels, Novex MiniCell tanks, 796 and a BioRad PowerPac controller. Western blotting was performed using an 797 iBlot gel transfer device from Thermo-Fisher. Thin layer chromatography was 798 performed using Silica Gel 60 F₂₅₄ plates (Merck) using 1-10% methanol in 799 dichloromethane. Nuclear magnetic resonance spectra were recorded on a 800 Bruker AVIII HD 400 nanobay (400MHz) spectrometer. Carbon nuclear magnetic 801 resonance spectra were recorded on a Bruker DQX 400(100 MHz) spectrometer. 802 All ¹H NMR chemical shifts are quoted in ppm using residual solvent as the 803 internal standard relative to TMS (d6-acetone: 2.09 ppm). All ¹³C NMR chemical 804 shifts are quoted in ppm using the central solvent peak as the internal standard 805 relative to TMS (d6-DMSO 39.3 ppm). Coupling constants (J) are reported in 806 Hertz (Hz). Infrared (IR) spectra were recorded on a Bruker Tensor 27 Fourier-807 Transform spectrophotometer. High-resolution mass spectra were recorded on a 808 Micromass LCT (resolution = 5000 RWHM) using a lock-spray source. Protein 809 crystal structures were analyzed and displayed using MacPyMOL v. 1.3 810 (Schrodinger, Inc.). Synthetic genes for *Medicago truncatula mtUGT71G1* and 811 mtUGT78G1 were obtained from GeneArt Gene Synthesis (Thermo-Fisher) 812 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).

818 UGT enzymes are named according to the UGT Nomenclature Committee's 819 latest guidelines⁵¹ as follows: *Arabidopisis thaliana* protein UGT73C6 encoded by 820 gene *UGT73C6* is written atUGT73C6.

821

822 Plant GT1 production.

823 Arabidopsis GT1 plasmids in pGEX-2T (as reported by Lim et a^{6}) were 824 transformed into Rosetta (DE3) pLysS Escherichia coli expression strains and produced essentially as reported.^{6,52} Cells were resuspended in glutathione S-825 826 transferase (GST) purification buffer (50 mM Tris, pH 7.4, 1 mM DTT), lysed, 827 centrifuged (10,000 × g, 10 min, 4 °C followed by centrifugation at 25,000 × g, 60 828 min, 4 °C) and either used as the crude supernatant or taken forward for 829 purification using a Sepharose 4B glutathione resin (GE Healthcare) as described.⁵² Western blotting was performed with mouse anti-GST (BD 830 831 Biosciences) (Supplementary Figure 2A). GT1 protein-containing lysates could 832 be flash-frozen and thawed once with activity remaining for up to 6 months' 833 storage at -80 °C.

834

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

Activity assays were conducted using reported MS methods¹⁴ on either a Waters 836 837 Quattro Micro API (ESI⁻ mode) or a Waters ZMD-MS (ESI⁻ mode), each equipped 838 with a Waters 600 HPLC System and a Waters 2700 autosampler capable of 96-839 well format. Reaction mixtures were composed of 93 µL reaction buffer (1 mM Tris, pH 7.8, 50 µM MgCl₂), 1 µL of NDP-Sugar (10 mg/mL stock), 1 µL of 840 841 aglycone (10 mg/mL stock), and 5 μ L cell supernatant or purified protein (ca. 1 842 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 843

844 reaction mixture was injected into the MS with 50:50 MeCN:H₂O (0.1 mL/min flow 845 rate, 5.5 min flush). Data was ranked Green (signal/noise > 10), Amber (s/n 1-10), 846 or Red (s/n < 1) from the total ion count integration of the full peak 847 (representative data shown in **Supplementary Figure 2B,C**). The acceptor 848 library is shown in **Supplementary Figure 1** and the full acceptor dataset is 849 shown in Supplementary Figure 3B. The full donor dataset is shown in 850 Supplementary Figure 3A. Regioselectivities were based on comparison of LC-MS elution time with internal standards as reported⁸ or as deduced from 851 852 substitution patterns within the same chemical families (Supplementary Figure 853 4).

854

855 Chemical Diversity Calculations.

856 Molecular shape calculations were used to design library features that sample a 857 broad range of 3-dimentional chemical space (**Supplementary Figure 1C**). Each 858 structure was energy minimized using the MM2 function of Chem3D 859 (CambridgeSoft) and converted to .sdf format. The principal moment of inertia 860 was calculated for the energy-minimized conformations of our library members using the Knime Analytics Platform⁵³ with the "SDF Reader" \rightarrow "PMI Calculation" 861 862 (Vernalis)→"JavaScript Scatter Plot" nodes and compared to reference 863 molecules for "rod" (octa-2,4,6-trivne), "sphere" (adamantane), and "disk" (benzene).⁵⁴ Our compounds were found to lie primarily along the rod-disk axis, 864 865 but sampled space well into the other principal chemical shape regions.

- 866
- 867

868 Clustering of activity based on phylogenetic alignment or functional patterning.

Phylogenetic analyses were performed with CLUSTAL_X⁵⁵ or Clustal Omega⁵⁶ 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 analysis⁵⁸ (further details
provided in **Supplementary Note**).

876

877 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**.

- 884
- 885 Equation 1

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

886

887 Equation 2

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

888

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**.

- 893
- 894

895 *GT-Predict* – *Classifying substrate interactions using quantifiable on* 896 *physicochemical properties.* A Decision Tree-based model was trained on 897 various combinations of each substrates' cLogP, molecular volume, solvent 898 accessible area, and carboxylate pKa. Additionally, structural information such as 899 number of hydroxyl groups or amines as well as substitution patterns on 900 coumarin, flavonoid, or phenylpropenoid scaffolds (the physicochemical 901 parameters, calculated using Chem 3D version 16.0, are listed in 902 Supplementary Tables 8, 9). GAR scores were input for each enzyme and 903 classifier programs were written in MATLAB as part of the GT-Predict 904 "PredictAcceptorInteraction" module. The cross-entropy function was used for the 905 splitting criterion for the branching of the tree. Models were evaluated by 906 determining the accuracy and Matthews correlation coefficient using leave-oneout cross validation.59,60 907

- 908
- 909

910 GT-Predict – Prediction of novel enzyme activities based on GAR dataset and 911 alignment.

Smith-Waterman²⁵/BLOSUM50⁶¹ 912 pairwise Α alignment algorithm was 913 implemented with the GAR scoring matrix in the GT-Predict module 914 "PredictEnzymeInteraction". A weighted k-nearest neighbor approach was used 915 to predict substrate interactions for novel GT1 FASTA amino acid sequences 916 using **Equation 3** to obtain weighted votes from the closest protein sequences in 917 our dataset and provide interaction predictions for novel sequences. The top two 918 sequences in our dataset for a novel GT1 amino acid sequence input are used in 919 a weighted vote for prediction, given a 1/"yes" for weighted votes (p_m) of over 0.5 920 or a 0/"no" for p_m less than 0.5 (**Equation 4**).

921

922 Equation 3

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

923

924 Equation 4

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

926 In **Equation 3**, x_{mj} is the interaction data for molecule m interacting with the 927 *j*th nearest neighbor of the enzyme, and equals 1 if there is an interaction or 0 if 928 there is not. Results of the prediction were tested against the interaction patterns 929 of experimental GAR screens.

930 We applied the GT-Predict module "PredictEnzymeInteraction" to two novel 931 GT1 enzymes from the legume *Medicago truncatula* and the cereal grain Avena 932 strigosa. Data for two "divergent" GT1 sequences from bacterial GT1 enzymes was adapted from our previous screen.²⁸ Prediction and experimental validation 933 934 data are shown in Supplementary Figure 13 with accuracies tabulated in 935 Supplementary Table 3. Parameters and data from bacterial enzymes saOleD and sIMGT were essentially those from previous studies.²⁸ For details and 936 937 validation see the Supplementary Note. Protein accession codes used for 938 prediction: M. truncatula mtUGT71G1 (UniProt Q5IFH7), M. truncatula mtUGT78G1 (UniProt A6XNC6), A. strigosa asUGT74H5 (GenBank EU496509), 939 940 A. strigosa asUGT88C4 (GenBank EU496511), S. antibioticus OleD (UniProt 941 Q53685), S. lividans MGT (UniProt Q94FR0). All alternative GTs were expressed 942 via our Plant GT1 production workflow.

943

944 GT-Predict – Exploration of Other Complete Families.

945 Two separate and complete GT1 families from Avena strigosa and Lycium 946 *barbarum*, respectively, containing candidates given as 'confirmed' in the CAZy "Glycosyltransferases" database⁴ were selected for further benchmarking of 947 948 "PredictEnzymeInteraction." Each contain ca. 20-25 validated isozymes. Amino 949 acid sequences were collected from Uniprot, DNA sequence-optimized for 950 production in Escherichia coli, and ordered as synthetic gene fragments (Twist 951 Bioscience, San Francisco, USA). GT1 sequences were flanked with restriction 952 sites (N-terminal BamHI and C-terminal EcoRI) for for subcloning into pGEX-2t 953 and a *C*-terminal hexahistadine tag was added for Western blotting and optional 954 purification, although these were used as crude lysates for screening purposes. 955 Fragments are listed in **Supplementary Table 5** (*Avena*) and **Supplementary** 956 Table 6 (Lycium). Synthetic gene adaptors: 5'-GGATCC-GT1 gene fragment957 GCAGCAGCACTGGAACATCATCATCATCATCATCAT_TAA_GAATTC_3' (BamHI
958 site - *GT1 sequence* - linker/hexahistidine tag - stop codon - EcoRI site) were
959 used for all sequences.

960 GT1 fragments were dissolved in Tris-EDTA buffer and digested using EcoRI 961 and BamHI (New England Biolabs) following recommended protocols and 962 purified using Qiagen PCR Purification Spin columns. The vector pGEX-2t was 963 digested with EcoRI and BamHI and purified on agarose gel and isolated using 964 Qiagen Gel Purification Spin columns. Ligation was performed with T4 DNA 965 ligase (New England Biolabs) following the standard overnight 16 °C protocol. All 966 sequences were verified. Note: a minor number of GT1 gene fragments failed 967 during DNA production or subcloning, but 16/18 Avena and 16/23 Lycium GT1 968 expression plasmid were verified. The expansion plant GT1s were produced in 969 Rosetta 2 (DE3) pLysS *E. coli* strains following our standard procedure (briefly, 970 250 mL Terrific Broth cultures grown at 37 °C to $OD_{600} \approx 0.6$, cooled to 20 °C, 971 and induced for overnight expression with 0.1 mM IPTG and 140 rpm shaking). 972 Cell pellets were isolated, sonicated, centrifuged at 12,000 \times g for 15 minutes at 973 4 °C and then 25,000 × g for 60-90 minutes at 4 °C. Gels and Western blots 974 (using anti-poly-histidine—alkaline phosphatase clone HIS-1, Sigma cat. number 975 A5588) are shown in **Supplementary Figure 8**.

976 "GT-Prediction" of EnzymeInteractions and confirmatory screening reactions
977 were performed as above. Aglycones were chosen as the ca. 40 substrates that
978 showed positive reactivity with at least one GT1 in the *Arabidopsis* collection.
979 The predicted/experimental datasets and summary are shown in **Supplementary**980 **Figures 9-11**.

981

982 Homology model construction for confirmation of chemical recognition 983 hypotheses.

984 Structurally-characterized Michaelis complexes of GT1 enzymes (either 985 UGT72B1, PDB ID: 2VCE¹⁹ or VvGT1, PDB ID: 2C1Z¹⁸) were input as templates 986 for homology model construction using the I-TASSER server.^{48,62} Models were 987 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 GTPredict "AcceptorGUI" module, which highlights associations for each enzyme by
property.

- 993
- 994 Site-Directed Mutagenesis of UGT71C4 and UGT72C1.
- 995 Enzyme engineering of the anionic substrate and UDP-GlcNAc activity was
 996 carried out using the Q5 Site Directed Mutagenesis kit (New England BioLabs)
 997 with the following primers:
- 998 UGT71C4 R292A
- 999 Forward: 5'- TTTCGGGAGCgcAGGAAGCGTTG-3'
- 1000 Reverse: 5'- CAGAGGAACACCACCGAT-3'
- 1001 UGT72C1 D180A
- 1002 Forward: 5'-CGGGCTCAAGcTCCGAGAAAATATAT-3'
- 1003 Reverse: 5'- CTCAAACTTAACCGGGCTG-3'
- 1004 UGT72C1 E187A
- 1005 Forward: 5'- TATATTCGGGcACTCGCTGAG -3'
- 1006 Reverse: 5'- TTTTCTCGGATCTTGAGC -3'
- 1007 UGT72C1 D180A:E187A
- 1008 Forward: 5'- tatattcgggcACTCGCTGAGTCTCAGCG -3'
- 1009 Reverse: 5'- ttttctcggagCTTGAGCCCGCTCAAACTTAAC -3'
- 1010 UGT72C1 G284R:
- 1011 Forward: 5'- TTTTGGGAGTagaGGGGCACTAAC-3'
- 1012 Reverse: 5'- GAAACATAAACCACTGACTC-3'
- 1013 Mutagenesis reactions were processed according to the manufacturer's protocol.
- 1014 All transformants were confirmed by nucleotide sequencing.
- 1015
- 1016 Biotransformation to prepare trans-resveratrol-4'-O- β -D-glucopyranoside.
- 1017 Reactions were carried out in aqueous buffer (20 mM Tris, pH 8.0, 40 mM NaCl,
- 1018 4 mM KCl, 2 mM MgCl₂). A 50 mL Falcon tube was charged with 5.7 mg (25

1019 μmol, 1 equiv.) resveratrol and 15.7 mg (25 μmol, 1 equiv.) UDP-glucose 1020 disodium salt. 50 mL of cold buffer was added (to 500 µM final concentration), 1021 followed by 500 µL of rapidly-thawed GST-UGT73C6 crude lysate, stored on ice. 1022 Reactions were placed in a 37 °C shaking incubator at 200 rpm and followed by 1023 an upright 50 mL Falcon tube is optimal. Too much t.l.c. (Note: 1024 headspace/shaking precipitates the GT1 catalyst.) Reactions were worked up by 1025 extracting 5 times with 10 mL EtOAc. The organic layer was washed with 50 mL 1026 brine, dried over MgSO₄, and purified by silica chromatography (2.5 g silica gel, 0%) 1027 MeOH/CH₂Cl₂ to 15% MeOH/CH₂Cl₂) to afford 3.0-3.8 mg product as a pale 1028 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/CH₂Cl₂. ¹H NMR (d6-1029 acetone, 400 MHz) δ = 8.27 (s, 1H, phenolic OH), 7.55 (d, J = 8.8 Hz, 2H, H2', 1030 1031 H6',), 7.10–7.02 (m, 3H, vinylic H, H3', H5'), 6.98 (d, J = 16 Hz, 1 H, vinylic H), 1032 6.59 (d, J = 2.0 Hz, 2H, H2, H6), 6.32 (s, 1H, H4), 5.01 (d, J = 7.2 Hz, 1H, H1"), 1033 4.64 (s, 1H, sugar OH), 4.38 (s, 1H, sugar OH), 4.32 (s, 1H, sugar OH), 3.93 (dd, 1034 J = 2.8 and 14 Hz, 1H, H6"A), 3.75 (dd, J = 2.4 and 13 Hz, 1H, H6"B), 3.48 (m, 1035 4H, H2", H3", H4", H5"). Common solvent impurities at δ = 2.88 (H₂O), 2.45 1036 (ethyl methyl ketone), 2.09 (acetone), 1.97 (ethyl acetate), 1.32 and 0.914 1037 ("grease"), and 0.17 (silicone grease) were found due to low sample 1038 concentration following repeated attempts by HPLC to remove. ¹³C-NMR (d6-1039 DMSO, 100 MHz) δ = 159.0 (C3, C5), 157.4 (C4'), 139.4 (C-1), 136.8 (C1'), 1040 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: 1041 calc for $C_{20}H_{21}O_8$ [M-H⁺]: 389.12419; found: 389.12442. IR (neat) \tilde{v} = 3361, 2980, 1042 1043 2402, 1601 cm⁻¹. The obtained spectroscopic data (**Supplementary Figure 16**) were in accordance with those reported in the literature.^{64,33} 1044

1045

1046 Statistical Analyses.

1047 Validation of all the predictive models in the paper considered all elements of the
 1048 *confusion matrix,* namely the number of Positives and Negatives predicted that
 1049 matched correctly the true categories (True Positives – TP, and True Negatives –

1050 TN, respectively) as well as Positive and Negative predictions that are incorrect 1051 (False Positives - FP and False Negatives – FN, respectively). The median % 1052 accuracy (the accuracy associated with the 50th percentile of the accuracies over 1053 all data) and the *Matthews Correlation Coefficient* (MCC, **Equation 5**) for each 1054 acceptor are plotted in the box-and-whisker plots in **Figure 5**; all data reported in 1055 **Supplementary Table 3** (DT4 model) and in the GT-Predict package is available 1056 online.

1057

1058 Equation 5

1059 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.

1063

1064 Data and Code Availability.

1065 Custom code for GT-Predict was packaged into an executable file compatible 1066 with Windows (XP, Windows 7, and Windows 10 tested), available as a 1067 supplementary file through the Oxford University Research Archive 1068 DOI: 10.5287/bodleian:zg5195kaE. Activity datasets, mass spectrograms, and 1069 the protein FASTA sequences used here are also included in this package.

1070

1072 1073

Online Methods References

- 107451Mackenzie, P. I. *et al.* Nomenclature update for the mammalian UDP1075glycosyltransferase (UGT) gene superfamily. *Pharmacogenetics and genomics* **15**, 677-685 (2005).
- 107752Lim, E.-K. et al. Identification of Glucosyltransferase Genes Involved in1078Sinapate Metabolism and Lignin Synthesis in Arabidopsis. Journal of1079Biological Chemistry 276, 4344-4349, doi:10.1074/jbc.M007263200 (2001).
- 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).
- 108554Sauer, W. H. B. & Schwarz, M. K. Molecular Shape Diversity of1086Combinatorial Libraries: A Prerequisite for Broad Bioactivity. Journal of1087Chemical Information and Computer Sciences 43, 987-1003,1088doi:10.1021/ci025599w (2003).
- 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).
- 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).
- 109657Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular1097Biology Open Software Suite. Trends Genet. 16, 276-277 (2000).
- 109858Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241-1099254 (1967).
- 110059Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy1101Estimation and Model Selection. (1995).
- 110260Matthews, B. W. Comparison of the predicted and observed secondary1103structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) -1104Protein Structure 405, 442-451, doi:10.1016/0005-2795(75)90109-91105(1975).
- 110661Pearson, W. R. Selecting the Right Similarity-Scoring Matrix. Curr Protoc1107Bioinformatics 43, 3.5.1-3.5.9, doi:10.1002/0471250953.bi0305s43 (2013).
- 110862Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for1109automated protein structure and function prediction. Nat Protoc 5, 725-738,1110doi:10.1038/nprot.2010.5 (2010).
- 1111
 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).
- 1114 64 Learmonth, D. A. A Novel, Convenient Synthesis of the 3-O-β-D- and 4'-O
 1115 β-D-Glucopyranosides of trans-Resveratrol. *Synthetic Communications* 34, 1565-1575, doi:10.1081/SCC-120030744 (2004).
- 1117

(a)

Modeling bi-substrate enzyme families:

200,000+ nucleophilic

> 50 known > - * sugar electrophiles

12,000+ GT1 sequences

ca.120 billion permutations: how to predict?

(b)

GT Predict: modeling a complete GT1 subfamily





novel transformations

50

mechanistic insight

hypothetical annotation of new GT1 genes

Output















Enzyme

mtUGT71G1

mtUGT78G1

Kingdom

Accuracy

MCC

Organism Medicago truncatula (barrelclover) Plant 92% 0.901

Medicago truncatula (barrelclover) Plant 85% 0.662

Α

3	DMSO + UDP-Glc only control	oleanodmycin	Baicalein	Jmbelliferone	4-Methyl-umbelliferone	Sinapic acid	4-hydroxyl-benzoic acid	a-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	uteolin	Quercetin	Fisetin	Kaempferol	Cinnamic acid	4-hydroxy cinnamic acid	3,4-dihyroxy cinnamic acid	4-hydroxy 3-methoxy cinnamate	2-hydroxy cinnamic acid	3-hydroxy cinnamic acid	7-hydroxy 6-methoxy coumarin	5,7-dihydroxy coumarin	Frans-dihydroquercetin	novobiocin	Catechin	Genistein	Olomoucine	spectinomycin	CHAPS	a-ecdysone	solanidine	solasodine	oeta-sitosterol	
Acceptor	0	1	2	3	4	5	6	7	8	9	10	13	14	15	16	18	19	20	21	22	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																																									
naccuracies				х	х		х																					х				х	х								





asUGT88C4



Avena strigosa (black oat) Plant 79% 0.49



asUGT88C4

Avena strigosa (black oat) Plant 82% 0.357

С





saOleD

sIMGT

Streptomyces antibioticus Bacteria 69% 0.373

Streptomyces lividans Bacteria 74% 0.414



Local alignment (Smith-Waterman/BLOSUM scoring)

