Green Chemistry

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: V. Blay, L. T. Tran, S. Luang, C. Eurtivong, S. Choknud, H. Gonzalez-Diaz and J. R. Ketudat Cairns, *Green Chem.*, 2019, DOI: 10.1039/C9GC00621D.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/green-chem

Engineering Faster Transglycosidases and their Acceptor Specificity

View Article Online
DOI: 10.1039/C9GC00621D

Linh T. Tran[a,b,†,§], Vincent Blay[c,d,§,*], Sukanya Luang[e], Chatchakorn Eurtivong[f], Sunaree Choknud[a,b], Humbert González-Díaz[g,h], James R. Ketudat Cairns[a,b,g,*]

[a]School of Chemistry, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.
[b]Center for Biomolecular Structure, Function and Application, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.
[c]Fisher College of Business, The Ohio State University, Gerlach Hall, 2108 Neil Ave. Columbus, OH 43210, United States.
[d]Molecular Topology & Drug Design Research Unit, Departamento de Química Física, Universitat de València, Av. V. A. Estellés, s/n, 46100 Burjassot, Spain.
[e]Department of Biochemistry, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand.
[f]Program of Chemical Biology, Chulabhorn Graduate Institute, Chulabhorn Royal Academy of Science, Bangkok 10210, Thailand.
[g]Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.
[h]IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.
[g]Laboratory of Biochemistry, Chulabhorn Research Institute, Bangkok 10210, Thailand.

ABSTRACT. Transglycosidases are enzymes that have the potential to catalyze the synthesis of a wide range of high-value compounds starting from biomass-derived feedstocks. Improving their activity and broadening the substrate range are important goals to enable the widespread application of this family of biocatalysts. In this work, we engineered 20 mutants of the rice transglycosidase Os9BGlu31 and evaluated their catalysis in 462 reactions over 18 different substrates. This allowed us to identify mutants that expanded their substrate range and showed high activity, including W243L and W243N. We also developed double mutants that show very high activity on certain substrates and exceptional specificity towards hydrolysis, like L241D/W243N. In order to guide a more general use of Os9BGlu31 variants as transglycosylation catalysts, we built cheminformatics models based on topological descriptors of the substrates. These models showed useful predictive potential on the external validation set and are allowing the identification of efficient catalytic routes to novel phytohormone and antibiotic glucoconjugates of interest.

Keywords: Catalysis; Transglycosylation; Protein Engineering; Biocatalysis; Regioselectivity; Machine Learning; Neural Networks; Docking.

View Article Online
DOI: 10.1039/C9GC00621D

1. INTRODUCTION

Glycosylation is an essential mechanism for building structural and bioactive components of cells and their matrix and for providing blocking groups that restrain reactive and bioactive molecules from disrupting the metabolism of the cells in which they are synthesized.[1] Understanding the roles of the various glycosides in nature requires their synthesis or purification from natural sources, where they are often found in minute quantities. Moreover, the glycosylation of some bioactive compounds may modulate their uptake and bioactivity, which holds interest for the fine chemical and pharmaceutical industries. The large-scale synthesis of these glucoconjugates may require complicated strategies or expensive starting materials, while their extraction and purification may require large amounts of solvents. Therefore a simple and efficient system to produce such compounds in aqueous solvents is highly desirable.

Generally, glycosylation begins with the activation of a sugar by its attachment to a nucleotide, yielding nucleotide sugars, such as UDP-glucose (uridine diphosphate-α-D-glucopyranoside). The sugar can then be transferred by a Leloir-type glycosyltransferase (GT) to a carbohydrate or noncarbohydrate molecule to make a glycoside.[2] Glycosides and carbohydrates are broken down by cleavage of the glycosidic bond created by the GTs by glycoside hydrolases (GHs), usually by hydrolysis. However, many GHs are able to catalyze transglycosylation, that is, the transfer of the glycosyl moiety to another carbohydrate or aglycon moiety (Figure 1). Some enzymes related to retaining GH act as transglycosidases (TGs, also referred to as non-Leloir-type GTs), which preferentially catalyze transglycosylation rather than hydrolysis.

The transglycosidases that have been studied to date belong to families of retaining GHs and are believed to catalyze transglycosylation by the same mechanism used by the related GHs, shown in Figure 1. The retaining mechanism of GH (and GT) starts with a nucleophilic residue in the active site attacking the anomeric carbon. At the same time, the departure of the aglycon moiety is facilitated by a general acid in the active site. Since the glucosyl moiety becomes covalently bound to the nucleophilic residue, this first step is called glycosylation.[3-4] In the second step (deglycosylation), a nucleophilic acceptor molecule can attack from the same side as the aglycon departed. This is a ping-pong bi-bi mechanism, in which two S_N2 reactions, with inversion of stereochemistry at each step, return the anomeric carbon to its original stereochemical configuration. In the case of hydrolysis, water acts as the nucleophile in the deglycosylation step, whereas in transglycosylation another molecule acts as the nucleophile.

View Article Online
DOI: 10.1039/C9GC00621D

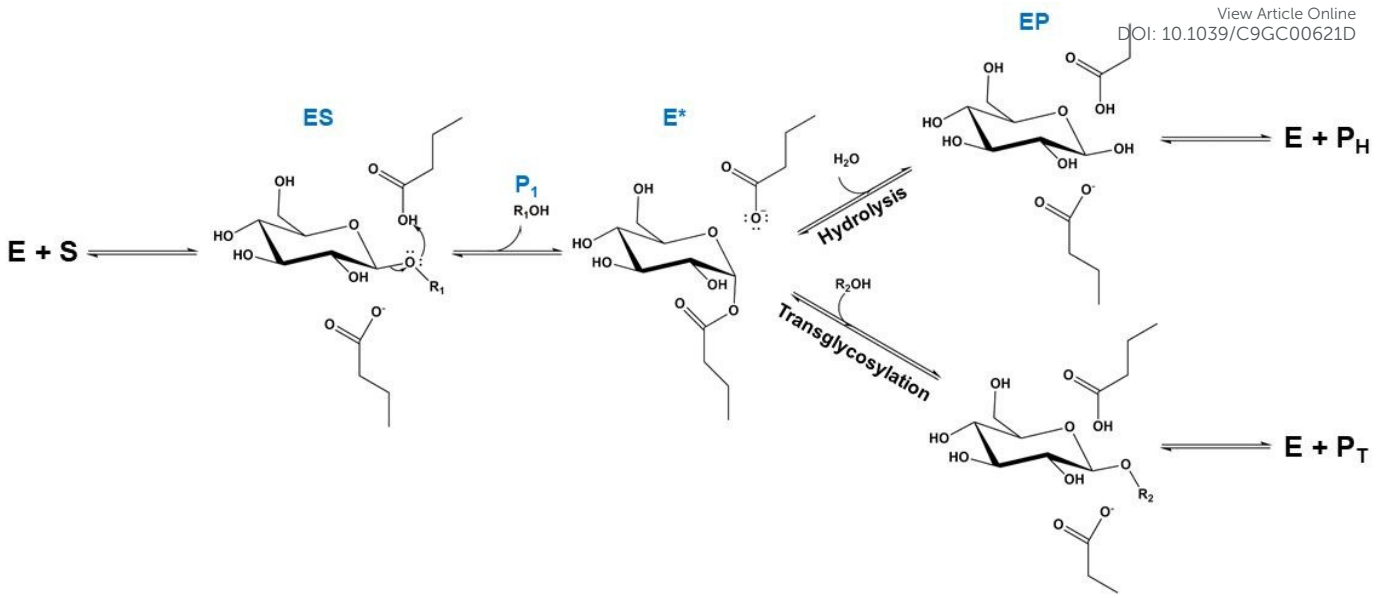

Figure 1. Mechanism of transglycosylation and hydrolysis reactions in retaining GH and GT. The enzyme complexes are labeled with the symbols used in the general nomenclature for a ping-pong bi-bi mechanism. The glycosyl intermediate is labeled E*. The second substrate is water in the case of hydrolysis or a hydroxyl group (R_2OH) in the case of transglycosylation. The glycon moiety shown is glucose, in line with the transglucosidase activity described in the current paper.

As noted above, TGs are related to GHs and have been included in the GH families grouped by amino acid sequence similarity at the carbohydrate active enzyme database (www.cazy.org). Most of the transglycosidases that have been described catalyze synthesis and remodeling of oligo- and polysaccharides. For instance, xyloglucan endotransferases (XET) belong to the family GH16, which also contains xyloglucan endohydrolases. Differences in a surface loop in the XET and the xyloglucan endohydrolases have been shown to contribute to the transferase versus hydrolase specificity[5], but this principle cannot be readily transferred to other families. Several attempts have been made to engineer GH to make TG, including the production of a β-transglucosidase from a bacterial GH1 β-glycosidase[6], a *trans*-α-*L*-arabinofuranosidase from GH51[7], as well as α-glucosyl transferases for the production of oligosaccharide antigens[8,9]. In each of these cases, the target products were oligosaccharides, whereas relatively little work has been done on the use of TG to synthesize glycosides.

The use of directed evolution approaches allowed the identification of changes that led to higher transglycosylation to hydrolysis ratios in the GH that were converted to TG.[6,7] While it was anticipated that decreasing the binding of water and increasing sugar binding could increase transglycosylation, it was

surmised that several of the mutations actually worked by decreasing the function of the catalytic nucleophile or acid/base. This was hypothesized to prolong the lifetime of the intermediate by destabilizing the transition states of the two reaction steps (Figure 1) and by decreasing the ability of water, a weaker nucleophile that requires deprotonation, to compete with sugars as the acceptor in the deglycosylation step. This strategy resembles the action of glycosynthase and thioglycoligase mutations of retaining GH, in which the nucleophile and acid/base, respectively, are mutated to non-nucleophilic, non-ionizable residues, thereby allowing the transfer of the glycon from activated donor glycosides to suitable sugar acceptors without hydrolysis[10,11].

In the last several years, some GH1 enzymes that catalyze the transglycosylation of lipids, anthocyanins, and other noncarbohydrate small molecules have been described[12-16]. The *sensitive to freezing 2* (*SFR2*) gene in *Arabidopsis thaliana* was found to encode a galactolipid/galactolipid galactosyltransferase. Several anthocyanin glucosyltransferases were found to actually be GH1 TG[13,15,16], whereas rice (*Oryza sativa*) Os9BGlu31 was found to be a general TG that can transfer glucose between phenolic acids, phytohormones, and flavonoids and it also appears to deglycosylate fatty acids and other substrates in the plant[14,17]. Little engineering of these types of enzymes has been reported, although we recently demonstrated that mutation of the active site cleft residue Trp243 (W243) to Asn (W243N) increased the activity and broadened the specificity of Os9BGlu31[18].

Promiscuous GT have been used to glycosylate several medicinal compounds to modulate their solubility and bioactivity[19]. For instance, several microbial antibiotics are glycosides and these promiscuous GTs have allowed the transfer of unusual sugars in glycodiversification to develop new or more robust activities[20,21]. From an industrial point of view, the development of transglycosidase (bio)catalysts would be desirable because they do not require a nucleotide sugar substrate or intermediate in the transfer of sugars to compounds of interest. Ideally, the sugars could be obtained from biomass[22,23]. Recently, cyclodextrin glucosyltransferases from *Thermoanaerobacterium sp.* were reported to transfer glucose onto aryl glucopyranosides and furanosides, achieving an unusual substrate specificity toward alkyl furanosides[24].

Although most of the transglycosidases studied to date show a high specificity for the transfer of the glucosyl moiety to certain sugars or aglycon groups, the promiscuity of Os9BGlu31 makes it especially promising to aid in the synthesis of a wide range of compounds[14,18]. Nonetheless, wild-type Os9BGlu31 cannot efficiently transfer glucosyl groups to all acceptors, so new variants with broader or different specificity are of interest to increase its potential as a biosynthetic tool. The development of catalysts that work under mild conditions

View Article Online
DOI: 10.1039/C9GC00621D

with environmentally friendly reagents would greatly facilitate the production of compounds for studies of phytohormone metabolism[1] and antibiotic glycosides with modified pharmacokinetic properties[19].

In this work, we have explored the potential of rice Os9BGlu31 active site cleft mutations to broaden its acceptor potential by engineering and evaluating 14 different amino acids at residue position 243. Moreover, we combined the mutation W243N with hydrophilic mutations of residues in positions that could interact with water during hydrolysis and evaluated their effect on the hydrolysis and transglycosylation specificity. Since the structure for Os9BGlu31 has not been solved yet, homology models of the covalent glycosyl-enzyme intermediate were made to evaluate the interaction of acceptor substrates with the active site. Although docking into such models provides insight into possible modes of binding, it could not quantitatively predict activity differences, due to the low resolution of the models. Therefore, we have applied cheminformatics tools to obtain further insights and anticipate the behavior of the different transglycosidases when presented new substrates.

There are two main possible approaches to build a cheminformatics model on experimental results. In the absence of a mechanistic hypothesis, it is possible to compute many molecular descriptors and then select statistically a *phenomenological route* that may be possible to interpret *a posteriori*.[25,26] Alternatively, one can hypothesize a *mechanistic route* that selects a reduced number of physicochemical descriptors beforehand to build a model with them.[27,28] In this work, we explored both approaches and built artificial neural network (ANN) models on a subset of physicochemical descriptors that can predict the activity of multiple enzymes against new substrates of interest.

2. MATERIALS AND METHODS

2.1 Construction of pET32a/DEST/TEV/Os9BGlu31

The wild-type pET32a/DEST/TEV/Os9BGlu31 expression vector was constructed as follows. The cDNA fragment encoding the mature Os9BGlu31 gene was amplified with the AK121679F and AK121679R primers indicated in Table S1 by *Pfu* DNA polymerase with the AK121679 clone plasmid as the template. The PCR product (~1.5 kb) was purified from the agarose gel and cloned into pENTR/TEV/D-TOPO Gateway entry vector (Invitrogen) and incubated at 22 °C for 18 h. The entry clone size was checked by *Sac*I restriction endonuclease digestion. Then, the pENTR/TEV/D-TOPO/Os9BGlu31 was recombined into the pET32a/DEST destination vector[29] by Gateway LR Clonase (Invitrogen) reaction and cloned in *Escherichia*

coli strain DH5α, selected on 50 μg ml^{-1} ampicillin LB agar. The recombinant expression vector clone was verified to contain the inserted gene by digestion with *EcoR*I restriction endonuclease and DNA sequencing (Macrogen Corp.).

2.2 Site-directed mutagenesis of Os9BGlu31

The tryptophan residue W243 in wild-type Os9BGlu31 was changed to 14 different residues by QuikChange site-directed mutagenesis (Agilent) with the pET32DEST/TEV/Os9BGlu31 vector as the template and the primers shown in Table S1. These mutants comprise cysteine (C), glutamic acid (E), glycine (G), histidine (H), isoleucine (I), lysine (K), leucine (L), asparagine (N), proline (P), glutamine (Q), arginine (R), serine (S), threonine (T) and valine (V). Furthermore, double mutants I172T/W243N, L183Q/W243N and L241D/W243N were prepared by introducing one of the mutations I172T, L183Q or L241D in the vector, in addition to W243N. The sequences of all mutant clones were confirmed by DNA sequencing **(**Macrogen Corp**.).**

2.3 Expression and purification of Os9BGlu31 and its mutants

The pET32a/DEST/TEV/Os9BGlu31 wild-type plasmid and all the mutants were transformed into Origami B**(**DE3**)** and the proteins expressed, extracted, and purified by an initial immobilized metal ion affinity chromatography **(**IMAC**)** step as previously described for the pET32a/DEST/Os9BGlu31 expression vector[14,18]. All of the purified fractions were assayed in 150 μl total volume with 2 mM 4-nitrophenyl β-D-glucopyranoside (4NPGlc) in 50 mM acetate buffer, pH 4.5, at 30 °C for 30 min. The reactions were stopped by adding 75 μl of 2 M Na_2CO_3 and the absorbance at 405 nm was then measured. The fractions showing activity with 4NPGlc were pooled and the imidazole was removed by buffer exchange with equilibration buffer (150 mM NaCl in 20 mM Tris HCl, pH 8.0) in 30 kDa molecular weight cutoff (MWCO) centrifugal filters. The N-terminal fusion tag was removed by cleavage with 1 mg TEV protease per 50 mg of the fusion protein at 4 °C for 16 h. Then, the digested proteins were loaded onto a second IMAC column in equilibration buffer, and the flow-through fractions showing activity to cleave 4NPGlc were pooled and concentrated with 30 kDa MWCO centrifugal filters.

2.4 Relative activities of Os9BGlu31 wild type and mutants

The activities of Os9BGlu31 and its mutants were compared upon a range of acceptor substrates for transglycosylation and water as the acceptor for hydrolysis in 50 mM citrate buffer (pH 4.5). Unless stated otherwise, the enzymatic assays were set up with 0.25 mM acceptor, 2.5 vol.% dimethyl sulfoxide

(DMSO), 5 mM 4NPGlc as the donor, and varying amounts of wild-type Os9BGlu31 and its mutants so that conversion of the limiting substrate was below 10% (Supporting Information files SI01.xlsx and SI02.xlsx). The reactions were conducted at 30 °C for 15 min, during which the concentration of products evolved linearly with time, and they were stopped by the addition of 1% formic acid. The reaction mixtures were centrifuged at 10,000 g for 10 min to remove the enzymes, and the supernatants were evaluated by reverse-phase UPLC, as previously described[18]. Briefly, 2 µL of the reaction mixtures were injected into a ZORBAX SB-C18 (1.8 µm, 2.1 x 150 mm) column equilibrated in 95% solvent A (0.2% formic acid in water) and 5% solvent B (0.2% formic acid in acetonitrile) in an Agilent 1290 UPLC with a diode-array detector (DAD). The compounds were eluted by a linear gradient from 5% to 50% B (v/v) for 13 min, 50% to 70% B (v/v) for 1 min, and 70% to 5% B (v/v) for 2 min, at a flow rate of 0.3 ml min^{-1}. Relative activities were evaluated from the absorbance at 360 nm of the 4-nitrophenol (4NP) released, which eluted at 10.5 min.

2.5 Transglycosylation of substrates to multiple glucoconjugates by Os9BGlu31 variants

To identify differences in glycosylation on one acceptor that has multiple nucleophilic groups, the products of transglycosylation by Os9BGlu31 wild type and W243 mutants (C, I, L, N, T, V) were determined in the reaction with 5 mM 4NPGlc as the donor, 0.5 mM acceptor, and 5 µg of proteins in 50 mM citrate buffer (pH 4.5). The reactions were conducted at 30 °C for 1 hour and then stopped with 1% formic acid. The mixtures were injected into the UPLC-DAD as described in section 2.4. Absorbance was measured at wavelengths between 190 and 500 nm to detect the glucoconjugate compounds.

2.6 Linear cheminformatics model. The glucosyl acceptors were characterized by 1D, 2D, and 3D molecular descriptors of different classes (constitutional, molecular format, autocorrelation, Basak, burden, connectivity, topological, charge descriptors, *etc.*). 1132 descriptors were computed using the software ChemDes[30], which integrates multiple state-of-the-art packages (Pybel, CDK, RDKit, Chemopy, *etc.*). 1824 additional molecular descriptors were computed with the newly released software Mordred[31], which computes 3D conformers by MM optimization and implements fully revised algorithms for popular indices found in Dragon and PaDEL.

To select the descriptors that will participate in the model, we first removed any index with zero variance across the range of substrates studied. We then reduced the set of descriptors under consideration by looking at the pairwise correlation between all of them. The correlation coefficient matrix was processed by the *findCorrelation* function in the *caret* R package[32]. We established the cutoff that no pair of descriptors had

a correlation coefficient > 0.8. When a pair of descriptors has a correlation above the cutoff, *findCorrelation* computes the mean absolute correlation of each descriptor in the pair to all others and removes the one with the largest mean absolute correlation. This step reduces the redundancy and the number of descriptors under consideration. In a second step, we ranked the descriptors using as the criterion the information gain as defined in the *FSelector* package[33] and the Pearson correlation of each individual descriptor to the average enzymatic activity measured in triplicate experiments. The top descriptors were then considered by a forward stepwise selection algorithm (*StepAIC*, *MASS* package) to build the multiple linear regression model[34]. A scheme of the pipeline devised is shown in Figure S1.

2.7 Nonlinear ANN models. The molecular descriptors used in the nonlinear models are D_1 = ALOGP (an estimate of logP), D_2 = MR (molar refractivity), and D_3 = TPSA (topological polar surface area), which can be found in file SI03.xslx. STATISTICA 10 was used to implement Artificial Neural Networks (ANN) as nonlinear ML models. The ANNs tested are Multi-Layer Perceptrons (MLP) with one dense hidden layer. The MLPs models have up to 7 hidden neurons in the hidden layer. Different activation functions were evaluated for the hidden and output layers. The models were trained within 200 cycles of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) backpropagation algorithm.

2.8 Docking study. The Os9BGlu31 model was calculated in MODELLER 9.19[35] based on the structure of Os3BGlu6 in covalent complex with 2-fluoroglucoside (PDB ID: 3GNR)[36]. To understand the structure of Os9BGlu31 in complex with glucose, the structure of Os3BGlu6 in complex with 2-fluoroglucoside structure was modified by replacing the fluorine at the C′2 with oxygen. Five models each of Os9BGlu31 and Os9BGlu31/glucose complexes were generated from the Os3BGlu6 covalent complex with 2-fluoroglucoside. The models with the lowest scores of the MODELLER Objective Function (MOF)[37] were selected. Os9BGlu31 mutants with and without glucose were generated from the Os9BGlu31 wild type models using FoldX[38] after insertion of the covalently linked α-D-glucoside coordinates into the models with glucose. The final models were evaluated for quality using ProSa2003[39] and PROCHECK software[40].

The acceptor substrate 3D structures were created with the drawing and manipulation tools in the Scigress 3.3.2 software suite. The homology models were protonated. The acceptor substrate structures and the newly added protons, glucoside, and mutated residues were then subjected to energy minimization using the MM2 force-field. Genetic Optimization for Ligand Docking (GOLD) software version 5.6.2 was used for the docking calculations. The center of binding was defined at coordinates (x = -0.581, y = -6.381, z = -21.102) with a 10 Å radius. Thirty docking runs were allowed for each ligand with default search efficiency (100%).

View Article Online
DOI: 10.1039/C9GC00621D

Lysine and arginine were defined as protonated; aspartic and glutamic acid were considered deprotonated. The GoldScore scoring function was used to evaluate the predicted binding modes[41].

3. RESULTS AND DISCUSSION

3.1 Activity of wild-type Os9BGlu31 and mutants

Previously, we showed that wild-type Os9BGlu31 and a several mutants are able to catalyze transglycosylation, transferring a glucosyl moiety from a donor substrate such as 4NPGlc to an acceptor[14,18]. In this study, we studied the effect of amino acid substitutions in the position W243 that complete the set of 20 natural amino acids at this position, when combined with our previous work[18]. These included the W243N mutant that had highest activity in our previous study and 13 new mutations (W243C, E, G, H, I, K, L, P, Q, R, S, T, and V). We also introduced the second point mutations I172T, L183Q, and L241D in the W243N mutant with the goal to test whether multiple hydrophilic substitutions increase the hydrolysis to transglycosylation ratio and broaden the substrate/product specificity to offer novel biosynthetic possibilities.

The Os9BGlu31 wild type and its mutants were evaluated in the glycosylation of 22 acceptors, including water (for hydrolysis). The relative activities of Os9BGlu31 wild type and its mutants are measured by quantifying the release of 4NP from the reaction of 4NPGlc with the different acceptors. The activity results are represented in Figure 2. Details on the activities are provided in the Supporting Information file SI01.xlsx.

Among the compounds tested, the preferred acceptor for the wild-type Os9BGlu31 is ferulic acid, while caffeic acid and 1-naphthaleneacetic acid are the substrates preferred by other Os9BGlu31 variants (Figure 2). Most of the Os9BGlu31 variants generated yield higher rates of 4NP release with phenolic acceptors than with buffer alone, which results in hydrolysis by water. Among the single mutants, W243L is arguably the most active overall. W243N also displays an increased activity with most of the substrates, although none is particularly favored, which agrees with previous results[18]. These mutations increase the activity of the enzyme and its potential use for the glycosylation of substrates of interest.

The mutants I172T, L183Q, and L241D show lower activity and stability than the wild type or W243N. The less soluble L241D mutant, in particular, shows the lowest performance over multiple acceptors (Figure 2). Remarkably, the introduction of L241D in the Os9BGlu31 W243N variant caused a significant activity increase for most substrates but the flavonoids (Table 1). Epistasis is the non-linear interaction between

mutations such that their effects over function, in this case transglycosidation activity, are significantly higher or lower than expected by simple addition of their inidividual effects. In Table 1, we observe several examples of positive epistasis[42-44] (higher improvements in function than expected) against several substrates. Notice that the double mutant L241D/W243N also exhibits the highest transglycosylation rate observed in this work using syringic acid as the acceptor. This is an example of sign epistasis[42-44] (the point mutation L241D has an adverse effect on the activity of the single mutant, but it has a positive effect on the double mutant L241D/W243N).

Another second mutation was introduced to generate the mutant L183Q/W243N. Although the second mutation barely affected the transglycosylation rate of phenolic compounds over W342N, it increased the activity with other acceptors and positive epistasis was observed in several cases (Table 1). Very prominent sign epistasis was also observed in this double mutant against the substrate 1-naphthol. Lastly, the introduction of the second mutation W243N in the mutant I172T could barely restore any of the activity lost upon the first mutation, suggesting that conservation of the amino acid I172 is important to catalyze the transglycosylation reaction. The entries in Table 1 with no annotation on epistasis indicate that those mutations may be purely additive or that the epistasis was not strong given the expected uncertainty in the data. Details on how the thresholds for the epistatic interactions were defined are provided in the Supporting Information file SI01.xlsx.

View Article Online
DOI: 10.1039/C9GC00621D

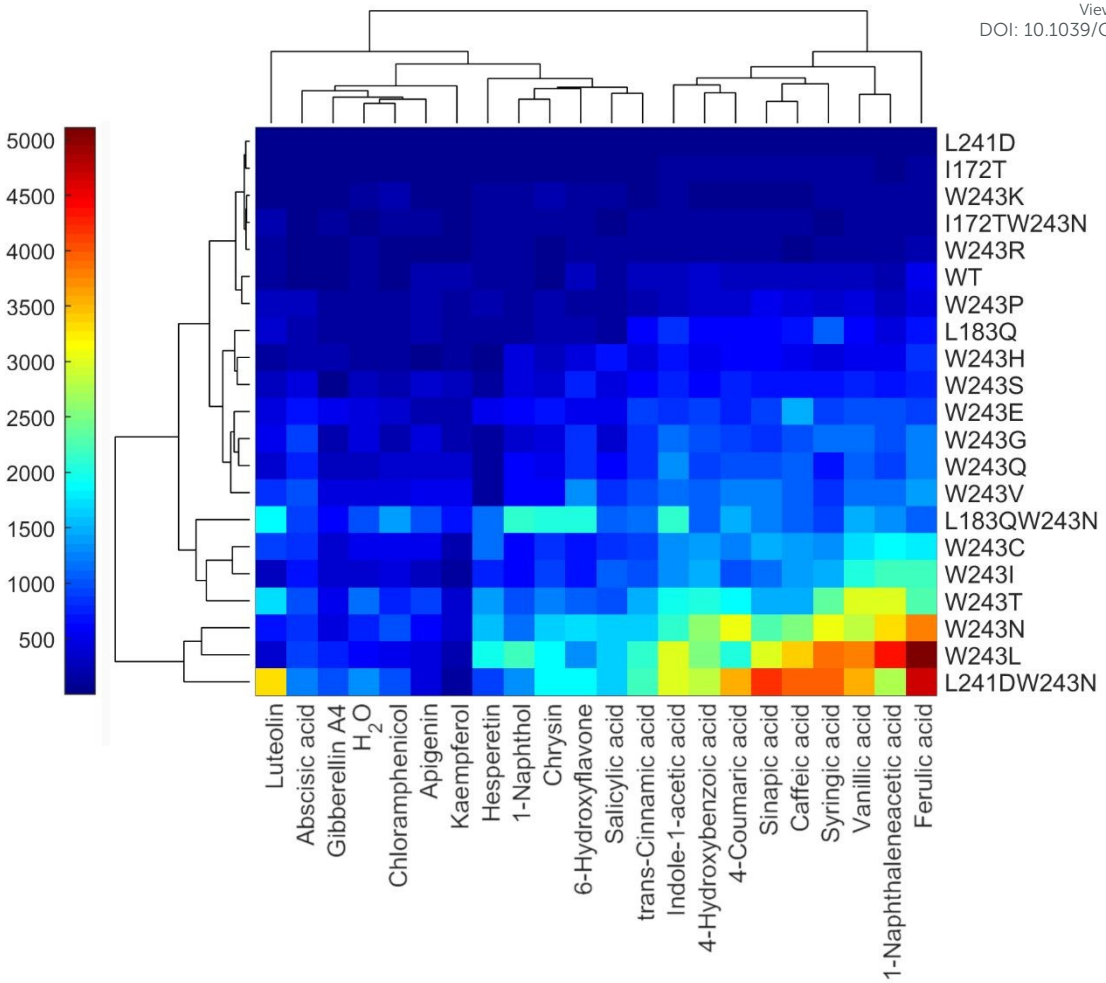

Figure 2. Heatmap of the 4NP release (in nmol min^{-1} mg^{-1}) catalyzed by 21 Os9BGlu31 variants when glycosylating 22 different acceptors (462 reactions). The dendrograms were constructed using the average Euclidian distance between clusters as the linkage method. For reaction conditions see SI01.xlsx.

Table 1. Reaction rates (in nmol min^{-1} mg^{-1}) and types of epistatic interactions observed for the double mutants in this work. For reaction conditions see SI01.xlsx.

Acceptor	Wild type	L183Q	L241D	I172T	W243N	L183Q/W243N / Epistasis	L241D/W243N / Epistasis	I172T/W243N / Epistasis
1-Naphthaleneacetic acid	185	40	16	68	3300	1344 / -	2750	108
1-Naphthol	84	12	4	33	1186	2104 / + (sign)	1316	88
4-Coumaric acid	275	26	15	82	3108	1482 / -	3515	85
4-Hydroxybenzoic acid	324	29	15	82	2592	1066 / -	2869	85
6-Hydroxyflavone	286	14	14	43	1703	1996	1877	134

Abscisic acid	60	17	15	45	861	911	1208 / + (sign)	78
Apigenin	165	12	0	33	603	1004 / + (sign)	407	83
Caffeic acid	252	29	15	114	2544	1081 / -	3957 / + (sign)	94
Chloramphenicol	55	10	0	61	1005	1401 / + (sign)	993	156
Chrysin	67	18	14	43	1655	1996	1877	134
Ferulic acid	550	37	24	99	3759	1088 / -	4674 / +	115
Gibberellin A4	77	8	15	34	479	591 / +	1029 / + (sign)	110
Hesperetin	92	24	0	34	1521	1167	957 / -	84
Indole-1-acetic acid	295	27	0	89	2152	2092	2977 / + (sign)	135
Kaempferol	210	12	0	25	321	653 / + (sign)	105	64
Luteolin	141	19	0	53	708	1913 / + (sign)	3348 / + (sign)	225
Salicylic acid	138	17	15	35	1674	1050 / -	1638	78
Sinapic acid	251	32	0	82	2276	1198 / -	4220 / + (sign)	82
Syringic acid	312	28	6	81	3051	897 / -	3933 / + (sign)	62
trans-Cinnamic acid	275	20	6	72	1617	1131	2188 / + (sign)	92
Vanillic acid	318	36	15	92	2812	1482 / -	3515	85
Water	103	11	6	41	727	968 / + (sign)	1316 / + (sign)	75

In addition to the mutants I172T, L183Q, and L241D, five W243 mutants (H, K, P, Q, and R) also showed lower transglycosylation and hydrolysis activities than the wild type. The rest of the mutants showed intermediate activities on the different substrates (Figure 2). Notably, the function of Os9BGlu31 could be affected by both single and double mutants, but none of the mutants studied showed a higher activity on all the substrates. Based on the results obtained, we selected the following 8 Os9BGlu31 variants with high activity for further study, along with the wild type as the reference: W243C, W243I, W243L, W243N, W243T, W243V, L183Q/W243N, and L241D/W243N.

The preceding results were obtained with different concentrations of DMSO, which were adjusted based on the substrate solubility (0.25 vol.% for phenolic compounds, 1.25 vol.% for phytohormones and flavonoids, 0 vol.% for water). We suspected that the DMSO used as co-solvent might have a slight effect on the catalytic activity of the enzymes. Therefore, we performed the activity study of the 9 high activity enzymes keeping the concentration of DMSO fixed at 2.5% by volume. These reactions were repeated in triplicate (file SI02.xlsx). The results are presented in Figure 3 and the activity trends are in agreement with the previous discussion.

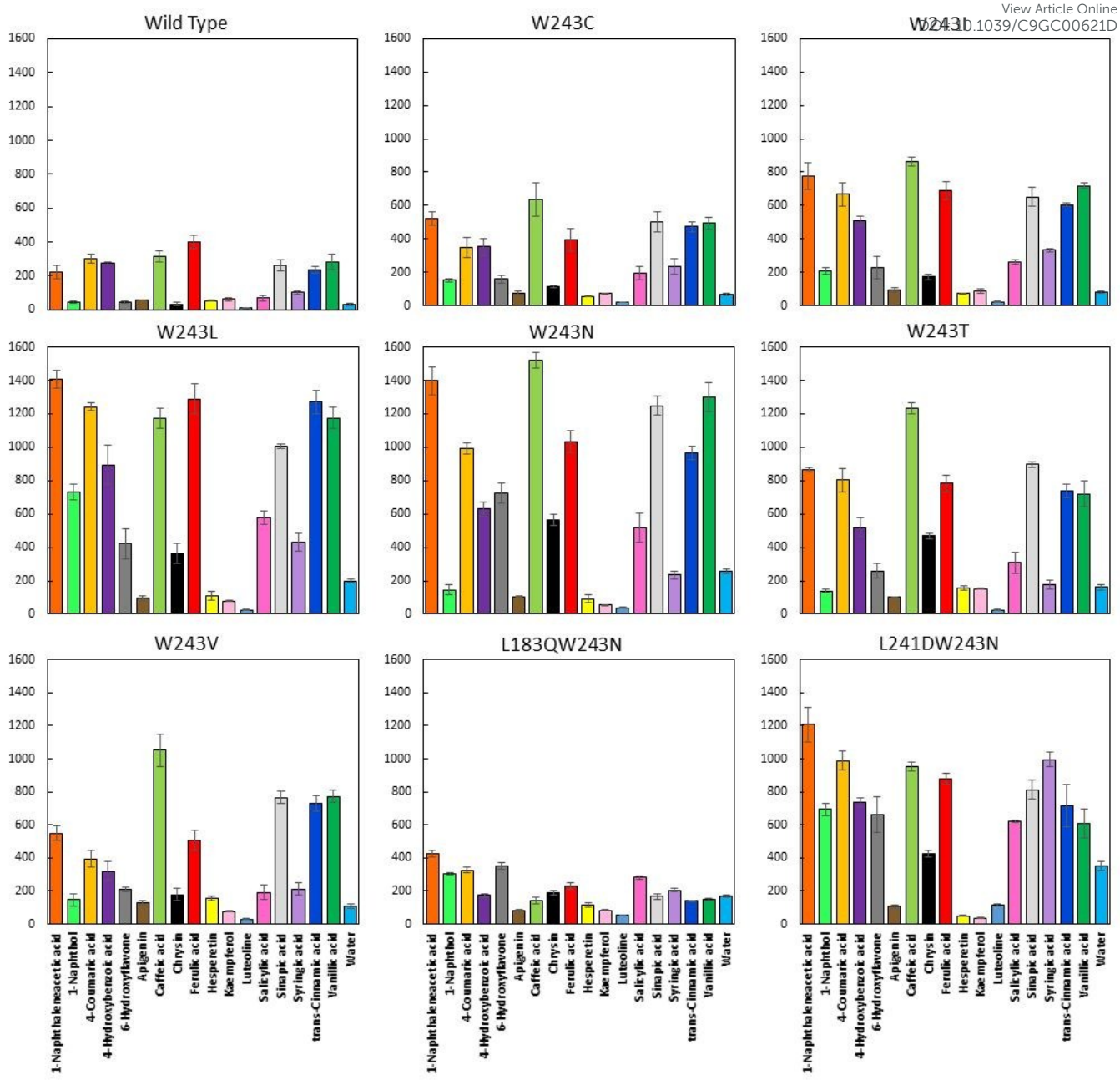

Figure 3. Relative rates of 4NP release by Os9BGlu31wild-type enzyme, W243 mutants, and L183Q/W243N and L241D/W243N double mutants against 22 different acceptors, including water. For reaction conditions, see SI02.xlsx.

It was found that the conversion of several hydrophilic residues did not improve substantially the hydrolysis activity of Os9BGlu31[18]. Remarkably, in the present study we have identified several high activity mutants that can hydrolyze 4NPGlc faster than the wild-type (Figure 4). In particular, Os9BGlu31 W243N shows the highest hydrolysis activity overall among the single mutant enzymes. This suggested that the mutation could be used in combination with additional hydrophilic mutations to further increase the hydrolysis rate of the

View Article Online
DOI: 10.1039/C9GC00621D

enzyme. In fact, the double mutants engineered catalyze both reactions, hydrolysis and transglycosylation, with the L183Q/W243N variant showing the highest overall ratio of hydrolysis to transglycosylation and L241D/W243N showing the second highest. These findings may validate the idea that hydrophobic residues near the putative water binding site contribute to transglycosylation activity, which was not apparent in previous studies[6,7,18]. The selectivity to each of the two reactions can be determined by monitoring the products being synthesized (section 3.2).

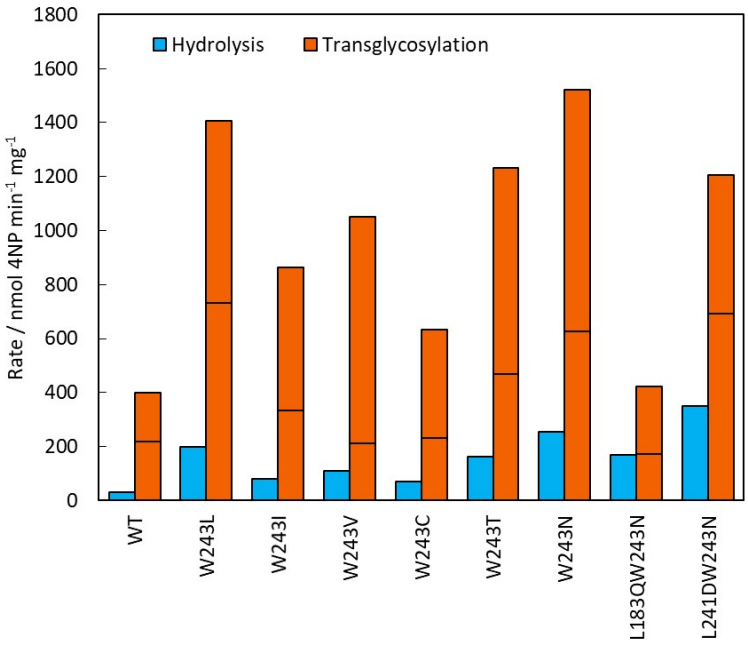

Figure 4. Rate of release of 4NP with the best substrate and with water for Os9BGlu31 wild type and high-activity mutants. The line inside the bar indicates the median activity over 18 different substrates. For reaction conditions see SI02.xlsx.

3.2 Effect of substrate on product distribution

To further understand the function of Os9BGlu31, we analyzed the different transglycosylation products during longer reaction times with a diverse group of substrates and enzyme mutants. Interestingly, several acceptors containing more than one hydroxyl group can accept glucose from the W243 mutants in different positions to produce multiple glucoconjugate compounds. These acceptors are the phenolic acids 4-hydroxybenzoic acid, ferulic acid, and vanillic acid and the flavonoids apigenin, luteolin, and kaempferol (Figure 5).

View Article Online
DOI: 10.1039/C9GC00621D

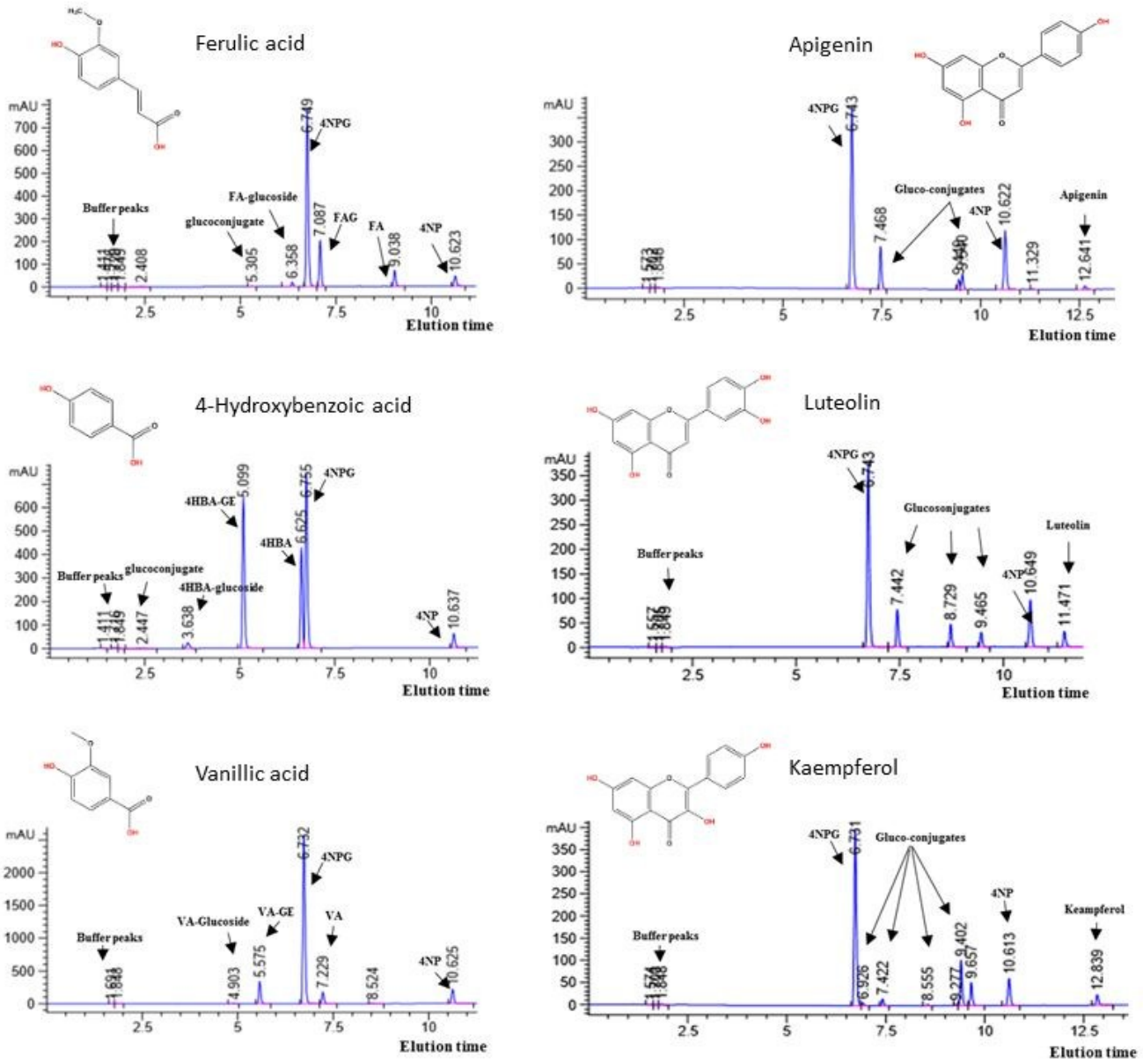

Figure 5. UPLC chromatograms of products of reactions catalyzed by Os9BGlu31 W243L variant using acceptors with multiple -OH groups and 4NPGlc as glucose donor (absorbance at λ=360 nm). The structures of the acceptors are shown at the top of the corresponding chromatogram. The reaction mixtures contained 5 mM 4NPGlc as the donor, 0.5 mM acceptor, and 5 µg of Os9BGlu31 W243L in 50 mM citrate buffer (pH 4.5), and were incubated at 30 °C for 1 hour and then stopped with 1% formic acid. The components were separated over a C18 UPLC column with the solvent and gradient conditions described in the Methods. The large peaks of 4NPGlc at 6.7 min and 4NP at 10.6 min were used to assess the total enzymatic activity. Other

View Article Online
DOI: 10.1039/C9GC00621D

peaks are marked according to their previous identification[18]. FA is ferulic acid, FAG is ferulic acid 1-O-glucose ester, 4NP is 4-nitrophenol, 4NPG is short for 4NPGlc (4NP-beta-D-glucoside), 4HBA is 4-hydroxybenzoic acid, 4HBA-GE is 4-HBA 1-O-glucose ester, VA is vanillic acid, VA-GE indicates VA glucose ester.

Notably, the wild-type enzyme transferred the sugar exclusively to one hydroxyl group position of the phenolic acids (Figures S2 to S4) or the flavonoids, which is a good example of regioselectivity. Mutants W243 C, I, L, and N produced more than two products when provided with ferulic acid, including the bis-glucoside at 5.3 min, the glucoside at 6.4 min, and the glucose ester at 7.1 min (based on our previous characterization[18]), while only one more product (ferulic acid glucose ester, FAG) was formed by W243T and W243V (Figure S2). The W243 mutants produced different amounts of glucoconjugates, with the glucosyl moiety being transferred to different positions to a variable extent. For instance, while the wild type and mutants yielded significant amounts of the 4HB-glucose ester, these yields were even higher for the mutants W243N and W243L. Since both glucosides and glucose esters of these phenolic acids have been reported in nature[45], the ability to synthesize them will be useful to study their biological roles.

We measured the product range formed from flavonoid substrates, some of which contain several hydroxyl groups (apigenin, luteolin, and kaempferol). Kaempferol and luteolin have similar structures with four hydroxyls (positions 3, 5, 7, and 4' in kaempferol and 5, 7, 3', and 4' in luteolin). W243N was shown to produce multiple kaempferol glucosides and bis-glucosides by transglycosylation[18]. In this study, we also observed various glycosides with the newly generated mutants (Figure S5) and were able to extend this to other flavonoids with different hydroxyl group positions.

The chromatograms in Figures S6 and S7 show the presence of multiple products in reactions catalyzed by Os9BGlu31 mutants. Four additional products were eluted in most of the reactions with apigenin catalyzed by Os9BGlu31 mutants. The product eluting at 9.45 min (apigenin 7-O-glucoside)[18] was synthesized by both wild-type and mutant enzymes but no other products were produced by the wild type, while additional peaks at 7.5, 9.55 and 11.3 min were seen in the mutants. The peak at 7.5 min that was particularly prominent in the reactions with the W243L and W243N variants is likely to be a bis-glucoside, based on the more hydrophilic elution position and previous observations with kaempferol glucosides[18]. This suggests that Os9BGlu31 W243L and W254N are efficient at transglycosylating two positions on apigenin, resulting in high bis-glucoside production. Besides, the presence of three products eluting in the range of monoglucosides (9-12 min) suggests that the selected mutants can glycosylate all three hydroxyl groups on apigenin.

Luteolin could also accept the transfer of glucose by Os9BGlu31 transglucosidase variants. W243L appears more active than the other mutants, based on the abundance of new compounds, especially the one eluting at 7.4 min. As noted for the corresponding apigenin glucoconjugate, this compound is likely to be a bis-glucoside, based on its relatively early elution, which might result from the high activity of W243L towards two hydroxyl positions. W243L also showed a high yield to the product eluting at 9.4 min, in addition to the peak at 8.7 min, which was high in most variants. W243T showed a lower yield of the peak at 8.7 min, as it is more selective to the compound eluting at 9.4 min. The relatively high peak at 10.6 min in the W243T reaction and relatively low hydrolysis rate in Figures 3 and 4 suggest it may be producing another glucoside hidden in the *p*-nitrophenol peak, but we were unable to resolve this putative glucoside. This data suggest that most variants could be used to produce the glycoside at 8.7 min, while W243L and W243T could be used to produce the bis-glucoside and other luteolin glucosides.

3.3 Structural model of the ternary complex including the acceptor substrate

Previously, we produced a homology model of the Os9BGlu31 wild type structure[18], but the apo structure is not appropriate to evaluate the reactivity toward acceptor substrates, since they react with the glucosyl-enzyme intermediate (Figure 1). To obtain a better understanding of the glycosyl transfer mechanism and visualize the binding orientations of the acceptor substrates within the Os9BGlu31 active site, we modeled the wild type and mutant proteins as their covalent glucosyl intermediates and docked the potential acceptors substrates on them. In these models, the glucoside is deeply buried in the active site and the access path is narrow, but small molecules can access it. This could be related to the lower activities of larger flavonoid substrates which were evident in Figure 3, as the narrow binding site would difficult the access of very large substrates. Molecular docking revealed the substrates are oriented on the opposite face of the glucose from the catalytic nucleophile, E387. The binding mode of the substrates is exemplified by ferulic acid and 1-naphthaleneacetic acid in Figure 6. Although ferulic acid was positioned somewhat distant from the anomeric carbons in the wild type active site, the models shown are in line with the predicted S_N2 mechanism proposed, as this requires the nucleophilic substrates to be positioned on the opposite face of the leaving group, E387, which ultimately results in the re-inversion of stereochemistry to form the β-anomeric products. Moreover, as shown in Table S2, the binding scores for the acceptor substrates on the glycosylated enzyme intermediate are generally higher for the W243L mutant compared to the wild type, suggesting that a higher free energy of binding of the acceptor may be contributing to lower the activation energy of the deglycosylation step and increase the kinetic rate.

View Article Online
DOI: 10.1039/C9GC00621D

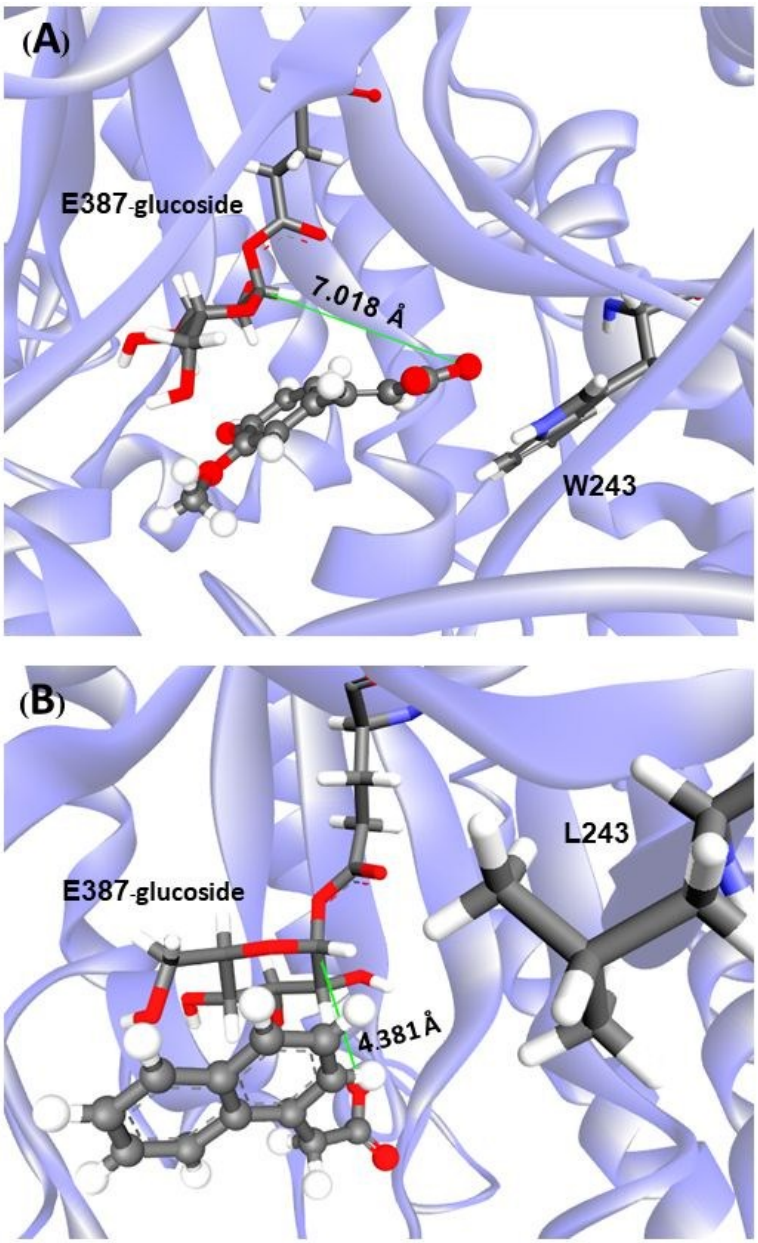

Figure 6. Graphical representations of the binding modes of **(**A**)** ferulic acid to wild-type Os9BGlu31 and **(**B**)** 1-naphthaleneacetic acid to W243L mutant covalent glucosyl intermediates. The blue ribbons depict the secondary structure of the proteins. The acceptor substrates are shown in a ball-and-stick representation. E387-glucoside, W243, and L243 are shown as sticks only. The solid green lines indicate the distances measured between the nucleophilic carboxylate and the electrophilic C1 on the glucoside.

View Article Online
DOI: 10.1039/C9GC00621D

3.4 Cheminformatics

3.4.1 Linear, phenomenological model

In order to generalize the findings of this work to other possible substrates, we constructed different cheminformatics models. To build a proper model we require a set of descriptors that is *relevant* (*i.e.*, is highly related to the enzymatic activity that we want to predict) and at the same time minimizes the *redundancy* of information captured by the different indices under consideration (*i.e.*, the descriptors are not totally correlated to each other).

We considered the results with all 9 high-activity enzymes altogether on all substrates except water. Catalysis on water is different from the rest of substrates and its small size prevents the calculation of several molecular descriptors. For instance, *RotRatio* is undefined for molecules that do not have a rotatable bond; *HybRatio* cannot be computed for molecules with no carbon atoms; *AATSp*, *AATSCp*, *MATSp,* and *GATSp* are undefined when p > number of atoms, *etc.*

Firstly, we computed a large number of molecular features describing each of the substrates (2956 descriptors). We then applied a redundancy-reduction filter so that no pair of molecular descriptors had a correlation above 0.8. This step reduced the number of descriptors under consideration to 100, which suggests that some molecular descriptors that have been proposed in the literature along the years may be more correlated to others than one might desire[46]. In a second step, we ranked the descriptors using as the criterion the Pearson correlation of each individual descriptor to the average enzymatic activity. The 12 most relevant descriptors according to this criterion are presented in Figure 7a. Their names are indicated in Table S3. We also considered another criterion for the ranking: in Figure 7b, the descriptor importance is measured in terms of information gain, an entropy-based metric.

In Figure 7, we can observe some interesting results. The descriptor most correlated to the enzymatic activity on its own is *largestChain*. This is a 1D constitutional descriptor from the RDK package and, as its names suggests, it quantifies the number of atoms in the largest chain on the molecular graph. *MATS3dv* is the Moran coefficient of lag 3 weighted by valence electrons. This is a 2D centered autocorrelation descriptor computed by the Mordred software. This descriptor being selected as important suggests that relatively symmetrical acceptors, such as aromatic-containing substrates[47], may favor faster transglucosylation rates.

View Article Online
DOI: 10.1039/C9GC00821D

ATSC5dv and ATS5c are the centered Moreau-Broto autocorrelation descriptors of lag 5 weighted by valence electrons and by charges, respectively.

If we look at the maximum information gain as the criterion for selecting descriptors, we also find multiple autocorrelation descriptors: *MATSv5* and *MATSp6* are the Moran autocorrelation coefficients of lag 5 weighted by van der Waals atomic volumes, and of lag 6 weighted by atomic polarizability, respectively. *GATSp6* and *GATS1Z* are the Geary autocorrelation coefficients of lag 6 weighted by atomic polarizability and of lag 1 weighted by atomic number, respectively. These two descriptors are considered relevant by both filtering algorithms. The fact that autocorrelation descriptors with different lags and different weightings are considered relevant by the algorithm suggests that the symmetry of the substrate, both in terms of connectivity and atomic makeup, would favor higher transglucosylation rates.

It is also interesting to see that, thanks to the pairwise-correlation reduction step, descriptors of diverse nature are considered in the ranking. For instance, *Mor29* and *Mor13* are the unweighted 3D-MoRSE descriptors of distance d = 29 and 13, as computed by Mordred. The large value of the scattering parameter d should make these descriptors insensitive to atom pairs situated at large distance but highly sensitive to short distances (<3 Å)[48]. This index may thus be capturing the effect of heteroatoms and multiple bonds (and thus bond distances) on the ability of the substrates to accept the glucosyl moiety. *Mor06m* is another 3D-MoRSE descriptor that is considered relevant. Notably, the topological charge indices *JGI7* and *JGI10* are also considered important by the information-based criterion. These indices have been shown to capture the charge distribution (*e.g.*, the dipolar moment) within a molecule based solely on its topology, thus bypassing the complexity of 3D-optimizing the molecular structure[49].

A major limitation of most filter methods is the fact that they barely consider the correlation between features[50]. Thanks to our filter strategy, however, this limitation could be greatly reduced. In Table S4, we can see that in a few cases the correlation between some of the indices shortlisted using the correlation to the enzymatic rate is close to the cutoff of 0.8. On average, however, the correlation between features is much lower, around 0.5, thanks to the design of *findCorrelation* algorithm. Interestingly, the correlation among features when using the information gain as the criterion is even lower, but the correlation to the enzymatic rate is more diverse, in the range R = 0.05-0.39 *vs*. R = 0.49-0.61 when using the correlation criterion (Table S4).

View Article Online
DOI: 10.1039/C9GC00621D

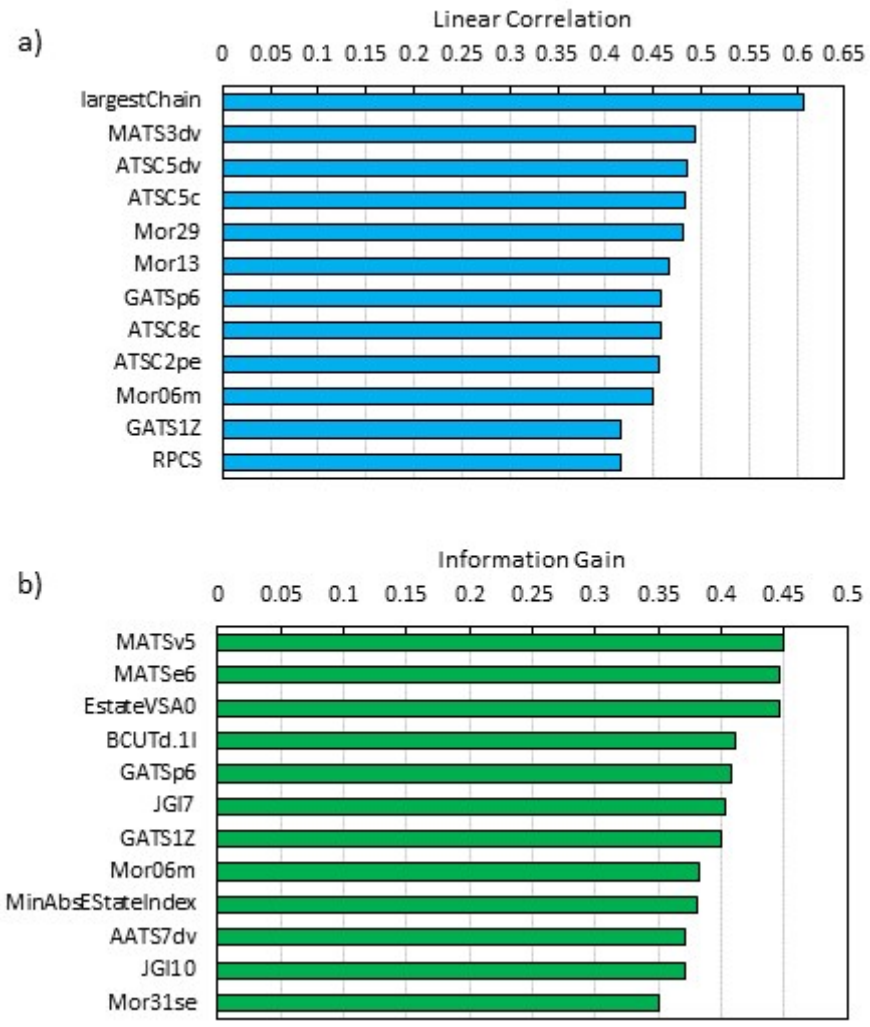

Figure 7. Top feature selection results using as criteria the linear correlation and the information gain between the descriptors and the average enzymatic activity measured across all enzymes.

With the subset of 21 different molecular descriptors shortlisted in Figure 7, it becomes possible to train a variety of models to seek a relationship between the molecular descriptors, the type of enzyme, and the enzymatic activity over a given acceptor substrate. One possibility is to build a linear model. The filtering algorithm has reduced the dimensionality of the problem by rejecting variables that are not very informative. Next, we conduct a *variable selection* to build a model (notice that we still have 22 variables in our consideration set: 21 continuous molecular descriptors and 1 categorical variable or factor representing the enzyme type). One way to select variables is to use a stepwise selection method. To this end, we used the function *stepAIC* in the *MASS* R package. This algorithm adds or removes one variable at a time so that the new model leads to the highest reduction in the Akaike Information Criterion (AIC) relative to the previous

model considered. AIC is a statistic that allows evaluating the quality of a model: it is closely related to the correlation coefficient R, but it introduces a penalty for every degree of freedom consumed upon introduction of a new variable[34]. This way, we selected the following model:

$$\text{Rate}\ (\text{nmolmin}^{-1}\,\text{mg}^{-1}) = \text{Intercept(Enzyme)} + 89.44 \cdot \text{largestChain} + 2.31 \cdot \text{ATSC5dv}$$

$$\text{n} = 153 \quad R^2 = 0.68 \quad R^2{}_{\text{adj}} = 0.65 \quad F(10, 142) = 29.5 \quad \text{AIC} = 1665.6 \quad p < 0.01$$

in which the independent term takes the values indicated in Table 2 depending on the query enzyme.

Not surprisingly, the algorithm has identified the type of enzyme and the *largestChain* descriptor as the most relevant variables, and it has also selected *ATSC5dv*. One advantage of linear models is their ease of interpretability. On the one hand, the molecular descriptors *largestChain* and *ATSC5dv* have a positive effect on the turnover of a substrate, with *largestChain* having predominant effect. On the other hand, differences between values in Table 2 correspond to the expected differences in activity between enzymes when presented a given substrate. It thus becomes clear that mutants W243L and W234N are much more active on average than the wild-type enzyme.

Table 2. Coefficients for the independent term as a function of the enzyme in the linear cheminformatics model.

Level	Intercept (nmol min^{-1} mg^{-1})
WT	8.69
L183Q/W243N	79.74
W243C	121.46
W243V	216.75
W243I	248.40
W243T	330.22
L241D/W243N	462.87
W243N	519.52
W243L	562.16

The linear model does not overfit the data, as indicated by the close values of R^2 and adjusted R^2, the relatively low value of the AIC, and the model and every variable are considered highly significant by the corresponding F-tests. Remarkably, this simple linear model yields a high determination coefficient, R^2 = 0.68. One factor that may contribute to this is that, in our study, most of the enzyme variants considered are single mutants at the same position 243. It seems thus conceivable that these mutations have a localized

View Article Online
DOI: 10.1039/C9GC00621D

steric effect within the active site cleft of the enzyme. This effect can thus be represented reasonably well by a constant factor dependent on the specific residue introduced and a variable related to the substrate chain size, which is in agreement with the docking study in the previous section. On the other hand, this model has been built and trained considering all the data (except for the hydrolysis reactions). To increase the predictive power of the model, in the next section we explore more flexible models and assess their performance against validation data.

3.4.2. Nonlinear, mechanistic models

In this section, we report a model based only on the molecular descriptors classically used in Hansch's method. The method has the advantage of using a few pre-selected descriptors to seek the models, making unnecessary the exploration of large spaces of chemical descriptors[51]. Hansch's model is based on an extra-thermodynamic linear free energy relationship (LFER). The LFER model decomposes a biological mechanism into a series of several steps. Then, it seeks a linear combination or weighted summation (additive model) of the free energies of the different steps. It considers as free energy contributions parameters like the negative logarithm of the ionization constant ($pK_i = -\log K_i$) and the logarithm of the partition coefficient (logP). These parameters are related to the free energy of the ionization process and to the posterior membrane transport (partition) steps of the biological compound, respectively. Specifically, P is the partition coefficient of the biological compound in the system *n*-octanol/water and it measures the compound's lipophilicity. Hansch's model also considers other steric and electronic properties of the compounds, like their polar surface area or molar refractivity, which increases the flexibility of the method to describe enzymatic processes like the one investigated. The molecular descriptors used in this work are D_1 = ALOGP (an estimate of logP), D_2 = MR (molar refractivity), and D_3 = TPSA (topological polar surface area). Moreover, as shown in Table S4, these different parameters show a limited correlation between them.

We evaluated multiple non-linear models using Artificial Neural Network (ANN) algorithms. The models use as input variables the type of enzyme and the 3 molecular descriptors calculated for the query compound. The networks built are Multi-Layer Perceptrons (MLP) with one hidden layer comprising 3, 6 or 9 neurons. Different combinations of identity, logistic, and exponential activation functions were considered. Table 3 summarizes the results obtained. On the one hand, we can see that the performance of the ANN is satisfactory (the coefficient of variation of experimental repeats is around 5-15%), with values of the correlation coefficient around R = 0.75-0.86 for the training and validation sets. In this case, 4/17 of the compounds were assigned randomly to the validation set across all the enzymes (see SI03.xlsx). Thus, the activity results for 1-naphtol, 4-coumaric acid, kaempferol, and *trans*-cinnamic acid were not used in training the ANNs.

View Article Online
DOI: 10.1039/C9GC00621D

Notably, the predictions for these compounds achieved a similar accuracy to those used to train the networks (Table 3, validation).

On the other hand, we observe that there is not a great advantage in using a very flexible model over using a linear model with appropriate descriptors in our problem. In fact, the ANN that only uses identity functions shows a limited performance, which stems from the descriptors specified. Abruzzo et al. also reported that variable selection had a greater impact than the selection of the model in a comparative study of transcriptomics data classifiers[52]. Furthermore, we observed that the training of larger dense hidden layers is more prone to encounter local optima and it may also increase the chances of undesired overfitting. Taken together, these results also suggest that the true relationship between inputs and output is not highly nonlinear.

Table 3. Performance of 10 different artificial neural networks predicting transglycosylation activity with multiple substrates and enzymes (for details, see SI03.xlsx).

Network			Training		Validation	
Topology	Hidden activation	Output activation	RSE	R	RSE	R
MLP 12-3-1	Logistic	Identity	33438	0.7264	33494	0.7397
MLP 12-5-1	Logistic	Identity	24856	0.8063	23629	0.8241
MLP 12-7-1	Logistic	Identity	30873	0.7554	27619	0.7786
MLP 12-3-1	Exponential	Identity	36041	0.7016	31832	0.7381
MLP 12-5-1	Exponential	Identity	20831	0.8403	27899	0.7996
MLP 12-7-1	Exponential	Identity	18715	0.8574	30130	0.7978
MLP 12-3-1	Exponential	Identity	24504	0.8170	23676	0.8017
MLP 12-5-1	Exponential	Identity	17788	0.8665	23239	0.8120
MLP 12-7-1	Exponential	Identity	18145	0.8639	25942	0.7997
MLP 12-7-1	Identity	Identity	46663	0.5822	32346	0.7488

Importantly, the models built can be used in practice to estimate the enzymatic activity of new acceptors over the different enzymes and thus used to screen libraries of potential substrates computationally. Given the reasonable performance of the different networks in Table 3, we averaged the individual predictions by all of them to reach an ensemble or consensus prediction. In Table S5, we report the results of exploring *in silico* the transglucosylation rates of a new set of putative acceptors by the wild-type Os9BGlu31 and different mutants. Although several of these compounds are expected to exhibit low activitiy, the results suggest that

View Article Online
DOI: 10.1039/C9GC00621D

3,4-dihydroxybenzoic acid, 6-benzylaminopurine, indole-3-acetic acid, and serotonin may be particularly reactive transglucosylation substrates over the enzyme variants W243L, W243N, and L241D/W243N. Aided by these predictions, we are already advancing the production of novel antibiotic and phytohormone glucoconjugates of interest, such as abscisic acid glucoside (preliminary results shown in Figure S8), on which we recently submitted a patent application.

In these results, glucose was transferred from 4NPGlc to a variety of compounds and acceptor positions, allowing us to prepare a broad array of valuable glucoconjugates. Other glucose donors have also been proposed[53]. For example, β-glucosyl fluoride was recently shown to be an effective donor for certain Leloir-type glycosyltransferases in the presence of catalytic amounts of UDP[54]. With regards to the large-scale production of glucoconjugates, more atom-efficient donors could be pursued, which opens up another area of potential research. The cheminformatics pipeline proposed in this work could also be extended to new combinations of substrates and catalysts, thus enabling increasing possibilities for transglycosidases and other enzymes.

4. CONCLUSION

Glucoconjugate molecules have great interest to the fine chemistry and phytochemistry industries, and many of these compounds could be synthesized from feedstocks derived from biorefineries. However, the production of complex glucoconjugates using conventional synthetic approaches may not be sustainable or economical, as multiple synthetic steps of limited yield would be involved. In this work, we studied 16 new Os9BGlu31 variants and compared them to the wild-type Os9BGlu31 to identify new high-activity catalysts, like W243L, and to correlate their activities with properties of the substrates being converted. Furthermore, the combination of the mutations designed to increase the hydrophilicity of the active site, L183Q and L241D, with the high-activity W243N mutation led to catalysts with high hydrolytic activity compared to their transglycosylation activity. These mutations exhibited positive epistasis on other substrates as well, which may be related to their proximity in the protein fold. We went a step further and used computational methods to facilitate a greener catalyst selection depending on the target substrate with a pipeline that can be used or adapted by other researchers. Cheminformatics models can help minimize the resources consumed to investigate new substrates and maximize the throughput of new catalytic processes. Our models suggest that steric constraints in the active site have a crucial effect on the activity of a substrate, which agrees with the docking simulations. Careful variable selection can afford linear models with reasonable performance,

although higher accuracy could be attained with nonlinear artificial neural networks. These models are proving useful to identify substrates and efficiently produce new glucoconjugates of interest, such as phytohormone glucosides for enzymatic and plant physiology studies or antibiotic glycosides with improved solubility and pharmacological properties.

Supporting Information

Figure S1 illustrates the data analysis pipeline used. Chromatograms in Figures S2 to S7 show the transfer of glucose to different acceptors by Os9BGlu31 wild type and its mutants. Figures S8 demonstrates the synthesis of the phytohormone glucoconjugate abscisic acid glucose ester enabled by this study. Table S1 presents the sequences of the oligonucleotides used as primers in the site-directed mutagenesis of Os9BGlu31. Table S2 indicates the docking scores on the wild type and W243L mutant. Table S3 indicates the names of the descriptors shortlisted. Table S4 is the correlation matrix between selected descriptors. Table S5 presents ensemble activity predictions for potential new transglucosylation acceptors. File SI01.xlsx contains the activity results in the screening of all mutants and the epistasis analysis. File SI02.xlsx presents the data from the high activity-mutants used to train the ML models. File SI03.xlsx contains cheminformatics data (descriptors and results of the ANNs).

Author Information:

*E-mail: blayroger.1@osu.edu (V.B.); cairns@sut.ac.th (J.R.K.C).

[§]These authors contributed equally.

[†]Current Address**:** Research Institute for Interdisciplinary Science, Okayama University, Okayama, Japan**.**

Notes

The authors declare no competing financial interests.

View Article Online
DOI: 10.1039/C9GC00621D

Acknowledgements

This work was supported by grants from Suranaree University of Technology (SUT) and the National Research University Project from the Commission on Higher Education (Thailand) to SUT. Ministerio de Economía y Competitividad (FEDER CTQ2016-74881-P and CTQ2013-41229-P) and Basque Government (IT1045-16) are gratefully acknowledged for their financial support. V. B. thanks the support from The Ohio State University and the Center for Biomolecular Structure, Function and Application (SUT).

REFERENCES

1. J. R. Ketudat Cairns, B. Mahong, S. Baiya and J. S. Jeon, *Plant. Sci.,* 2015, **241**, 246–259.
2. L. L. Lairson and S. J. Withers, *Chem. Comm.,* 2004, **20**, 2243–2248.
3. D. E. Koshland, *Biol*. *Rev*., 1953, **28**, 416–436.
4. S. G. Withers, R. A. J. Warren, I. P. Street, K. Rupitz, J. B. Kempton and R. Aebersold, *J*. *Am*. *Chem*. *Soc*., 1990, **112**, 5887-5889.
5. M. J. Baumann, J. M. Eklöf, G. Michel, A. M. Kallas, T. T. Teeri, M. Czjzek and H. Brumer, *Plant Cell,* 2007, **19**, 1947-1963.
6. D. Teze, J. Hendrickx, M. Czjzek, D. Ropartz, Y. H. Sanejouand, V. Tran, C. Tellier and M. Dion, *PEDS,* 2014, **27**, 13–19.
7. B. Bissaro, J. Durand, X. Biarnés, A. Planas, P. Monsan, M. J. O'Donohue and R. Faure, *ACS Catal*. 2015, **5**, 4598–4611.
8. E. Champion, I. André, C. Moulis, J. Boutet, K. Descroix, S. Morel, P. Monsan, L. A. Mulard and M. Remaud-Siméon, *J*. *Am*. *Chem*. *Soc*., 2009, **131**, 7379−7389.
9. E. Champion, F. Guérin, C. Moulis, S. Barbe, T. H. Tran, S. Morel, K. Descroix, P. Monsan, L. Mourey, L. A. Mulard, S. Tranier, M. Remaud-Siméon and I. André, *J. Am. Chem. Soc*., 2012, **134**, 18677–18688.
10. L. F. Mackenzie, Q. P. Wang, R. A. J. Warren and S. G. Withers, *J*. *Am*. *Chem*. *Soc*., 1998, **120**, 5583–5584.
11. M. Jahn, J. Marles, R. A. J. Warren and S. G. Withers, *Angew. Chem. Int. Ed*., 2003, **42**, 532–534.
12. E. R. Moellering, B. Muthan, C. Benning, *Science,* 2010, **330**, 226-228.
13. Y. Matsuba, N. Sasaki, M. Tera, M. Okamura, Y. Abe, E. Okamoto, H. Funabashi, M. Takatsu, M. Saito, H. Matsuoka, K. Nagasawa and Y. Ozeki, *Plant Cell,* 2010, **22**, 3374-3389.
14. S. Luang, J. L. Cho, B. Mahong, R. Opassiri, T. Akiyama, K. Phasai, J. Komvongsa, N. Sasaki, Y. Hua, Y. Matsuba, Y. Ozeki, J. S. Jeon and J. R. Ketudat Cairns, *J*. *Biol*. *Chem*., 2013, **288**, 10111–10123.
15. T. Miyahara, R. Sakiyama, Y. Ozeki and N. Sasaki, *J. Plant Physiol*. 2013, **170**, 619–624.

16. Y. Nishizaki, M. Yasunaga, E. Okamoto, M. Okamoto, Y. Hirose, M. Yamaguchi, Y. Ozeki and N. Sasaki, *Plant Cell,* 2013, **25**, 4150–4165.

17. J. Komvongsa, B. Mahong, K. Phasai, Y. Hua, J. S. Jeon and J. R. Ketudat Cairns, *J. Agric. Food Chem*., 2015, **63**, 9764–9769.

18. J. Komvongsa, S. Luang, J. V. Marques, K. Phasai, L. B. Davin, N. G. Lewis and J. R. Ketudat Cairns, *Biochim. Biophys. Acta,* 2015, **1850**, 1405–1414.

19. R. W. Gantt, P. Peltier-Pain and J. S. Thorson, *Nat*. *Prod*. *Rep*., 2011, **28**, 1811–1853.

20. C. Zhang, B. R. Griffith, Q. Fu, C. Albermann, X. Fu, I.-K. Lee and L. Li, *Science,* 2006, **313**, 1291–1294.

21. V. Křen and T. Řezanka, *FEMS Microbiol. Rev*., 2008, **32**, 858–889.

22. I. Finore, A. Poli, P. Di Donato, L. Lama, A. Trincone, M. Fagnano, M. Mori, B. Nicolaus and A. Tramice, *Green Chem*., 2016, **18**, 2460–2472.

23. R. A. Sheldon, *ACS Sustainable Chem. Eng.,* 2018, **6**, 4464–4480.

24. A. Pennec, L. Legentil, L. Herrera-Estrella, V. Ferrières, A.-L. Chauvin and C. Nugier-Chauvin, *Green Chem.,* 2014, **16**, 3803–3809.

25. V. Blay, R. Garcia-Domenech and J. Galvez, *ChemTexts,* 2017, **3**, 2.

26. V. Blay, J. Gullon-Soleto, M. Galvez-Llompart, J. Galvez and R. Garcia-Domenech, *ACS Sustain*. *Chem*. *Eng*., 2016, **4**, 4224–4231.

27. B. Bhhatarai, R. Garg and P. Gramatica, *Mol. Inform*., 2010, **29**, 511–522.

28. C. B. Santiago, J.-Y. Guo and M. S. Sigman, *Chem*. *Sci*., 2018, **9**, 2398–2412.

29. R. Opassiri, B. Pomthong, T. Onkoksoong, T. Akiyama, A. Esen and J. R. Ketudat Cairns, *BMC Plant Biol*., 2006, **6**, 33.

30. J. Dong, D. S. Cao, H. Y. Miao, S. Liu, B. C. Deng, Y. H. Yun, N. N. Wang, A. P. Lu, W. B. Zeng and A. F. Chen, *J. Cheminform*., 2015, **7**, 60.

31. H. Moriwaki, Y. S. Tian, N. Kawashita and T. Takagi, *J Cheminform*., 2018, **10**, 4.

32. M. Kuhn, *J. Stat. Softw., 2008*, **28**, 1–26.

33. T. Cheng, Y. Wang and S. H. Bryant, *Bioinformatics,* 2012, **28**, 2851–2852.

34. T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second ed., Springer, 2009.

35. A. Sali and T. L. Blundell, *J. Mol*. *Biol*., 1993, **234**, 779–815.

36. S. Seshadri, T. Akiyama, R. Opassiri, B. Kuaprasert and J. K. Cairns, *Plant Physiol*., 2009, **151**, 47–58.

37. M. Y. Shen and A. Sali, *Protein Sci*., 2006**, 15**, 2507–2524.

38. J. W. Schymkowitz, F. Rousseau, I. C. Martins, J. Ferkinghoff-Borg, F. Stricher and L. Serrano, *Proc*. *Natl*. *Acad*. *Sci*. *USA,* 2005**, 102**, 10147–10152.

39. M. J. Sippl, *Proteins*, 1993, **17**, 355–362.

40. R. A. Laskowski, M. W. Macarthur, D. S. Moss and J. M. Thornton, *J. Appl*. *Cryst*., 1993**, 26**, 283–291.

41. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, J. Mol. Biol., 1997, 267, 727–748.
42. J. F. Storz. *Curr. Opin. Struc. Biol*., 2018, **50**, 18-25.
43. T. N. Starr and J. W. Thornton. *Protein Sci*., 2016, **25**, 1204-1218.
44. M. Camps, A. Herman, E. Loh and L. A. Loeb, *Crit. Rev. Biochem. Mol. Biol*., 2007, **42**, 313–326.
45. S. Klick and K. Herrmann, *Phytochemistry*, 1988, **27**, 2177-2180.
46. R. Todeschini and V. Consonni. Handbook of Molecular Descriptors. Wiley 2008. p. 692. ISBN 9783527613106.
47. B. Hollas, *J. Math. Chem.,* 2003, **33**, 91–101.
48. O. Devinyak, D. Havrylyuk and R. Lesyk, *J. Mol. Graph. Model.,* 2014, **54**, 194–203.
49. J. Galvez, R. Garcia, M. T. Salabert and R. Soler, *J. Chem*. *Inf. Comput*. *Sci*., 1994, **34**, 520–525.
50. G. Chandrashekar and F. A. Sahin, *Comput. Electr. Eng.,* 2014, **40**, 16–28.
51. P. Abeijon, X. Garcia-Mera, O. Caamano, M. Yanez, E. Lopez-Castro, F. J. Romero-Duran, H. Gonzalez-Diaz, *Curr. Drug. Targets,* 2017, **18**, 511–521.
52. L. V. Abruzzo, L. L. Barron, K. Anderson, R. J. Newman, W. G. Wierda, S. O'Brien, A. Ferrajoli, M. Luthra, S. Talwalkar, R. Luthra, D. Jones, M. J. Keating and K. R. Coombes, *J. Mol. Diagn.,* 2007, **9**, 546–555.
53. S. Shoda, *Proc. Jpn. Acad., Ser. B,* 2017, **93**, 125-145.
54. A. Lepak, A. Gutmann and B. Nidetzky, *ACS Catal*., 2018, **8**, 9148-9153.