Pan-specific prediction of glycosyltransferase acceptors enabled by automated substrate feature generation of chemicals and large-scale experimental screening

ξ,‡,\*David Teze, ξ,‡,†Christian D. Madsen, ξTiia Kittilä, ξSören Maschmann, ξEvelyn Travnik, ξ,\*Ditte H. Welner

AUTHOR ADDRESS.ξ The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK- 2800 Kongens Lyngby, Denmark.

KEYWORDS. Glycosyltransferases. GT1. UGT. Glycosylation. Biocatalysis. Specificity. Prediction. Machine learning. Random forest.

# ABSTRACT

Glycosylation represents one of the major chemical challenges. While it is one of the most common reactions in nature, conventional chemistry struggles with stereochemistry, regioselectivity and solubility issues. Conversely, family 1 glycosyltransferase (GT1, UGT) enzymes can glycosylate virtually any given nucleophilic group with perfect control over stereochemistry and regioselectivity. However, the appropriate catalyst for a given reaction needs to be identified among the tens of thousands of available sequences. Here, we present a data-driven approach to find the needle in the haystack. We first assayed 28 GT1 enzymes against 30‒80 acceptors. Leveraging this dataset together with literature data, we obtained activity information on 147 GT1 enzymes, and used this to train several pan-specific random forest-based acceptor predictors. Importantly, we constructed a pipeline for automated generation of 153 chemical features for any given molecule, to input to the predictors. Further, we balanced the datasets with unreactive compounds chemically related to the acceptors in the dataset. We obtained predictors with AUC>0.8, that can be used to parse the >30.000 available GT1 sequences for any chemical. We constructed two top-performing predictors. The first was trained for prediction of both substrates and non-substrates of any GT1 enzyme, useful for understanding the acceptor profile of the enzyme. We demonstrate 80% accuracy in predicting the acceptor profiles of 8 GT1 enzymes selected to be phylogenetically far from our training dataset, against 47 polyphenols absent from the training dataset. The second was trained for highly accurate prediction of substrates only s, useful for biotechnological applications where identification of an efficient enzyme for a particular substrate is key. This predictor was used to predict the best GT1 enzymes for glycosylation of X acceptors absent from the training data, and predicted enzymes for both natural products and new-to-Nature glucosides with >90% accuracy. Finally, we experimentally measured and computationally evaluated another 1833 enzyme-substrate pairs comprising sequences in the training data and acceptors absent from the training data, resulting in an accuracy of, AUC of. Altogether, we present here the largest experimental dataset on GT1 specificity, together with predictors able to parse all available GT1 sequences and with high accuracy predict the most likely GT1 enzyme to glycosylate any given molecule.

# INTRODUCTION

Glycosylation is a crucial step to obtain many biologically and industrially relevant molecules, from proteins to natural products and artificial compounds.1 Accordingly, glycosylation is one of the most common reactions occurring in the biosphere. However, this reaction is particularly challenging for conventional chemistry because of the requirement for tight control of stereo- and regioselectivity. This leads to a succession of reactions, including protecting group manipulations and bond activations, amounting to low chemical yields, poor atom economy, and large amounts of waste. In Nature, these reactions are chiefly catalyzed by glycosyltransferases, which offer perfect stereoselectivity control in a single reaction with unprotected substrates.1 The control over acceptor scope and regioselectivity are however less stringent, and its chemical determinants are yet unknown, making it challenging to select an appropriate biocatalyst without extensive experimentation.2

Glycosyltransferases are phylogenetically organized into 114 families (as of June 24th, 2022) in the CAZy database.3 Glycosylation of natural products and secondary metabolites are primarily catalyzed by enzymes belonging to glycosyltransferase family 1 (GT1), which thus represents particularly important biocatalysts for biotechnological applications.1 GT1 enzymes present a so-called GT-B fold, catalyzing glycosylation at the interface of their two Rossmann domains, the N-terminal domain binding mainly acceptor substrates, and the C-terminal domain binding mainly α-glycosyl donors.4 This glycosyl donor is activated by uridine diphosphate, and thus GT1s are called UDP-dependent glycosyltransferases or UGTs.5 They catalyze *O*-, *N*- and *S*- glycosylation through a single-step reaction leading to an inversion of stereochemistry, leading to β-linked products (Figure 1A).6 However, while much is known about their structure and mechanisms – 55 different enzyme have at least one deposited crystallographic structure, and 330 are biochemically characterized as of June 29th, 2022 according to the CAZy database – little is known about their acceptor scope, except that it is tremendously varied, thousands of different acceptors having been reported, and that individual enzymes vary from highly specific to very promiscuous.7,8 Their activity is also hard to infer from biological data, as a single organism can contain over hundred different GT1 genes.

Machine learning is emerging as a powerful tool in enzymology, due to its strength in recognizing patterns in complex data.9,10 For general enzyme:substrate pairs prediction, the deep learning-based ESP algorithm is notable.11 Unfortunately, it performs well only for substrates within the training data, which do not cover the vast substrate space of GT1 enzymes. Interestingly, a decision tree-based algorithm – GT-Predict - has been developed specifically for GT1 enzymes.2

GT-Predict is trained on reactivity measurements of 53 *Arabidopsis thaliana* GT1 enzymes against 59 structurally diverse glycosylation acceptors. GT-predict is currently one of the most accurate enzyme-substrate prediction tools available, although – like ESP – it cannot accurately predict new substrates, and chemical features for new substrates need to be manually curated. The major limitation, however, is in predicting substrates for sequences absent from the training data, i.e. non-*Arabidopsis* GT1 enzymes; here, GT-Predict does not employ machine learning, but simply returns the substrate reactivity measured experimentally for the closest *Arabidopsis thaliana* homolog. Unfortunately, phylogeny has repeatedly been shown to be a relatively poor predictor of GT1 specificity8

Here, we attempt to resolve these limitations, by largely expanding the experimental data GT-predict is trained on, and by conceptually changing the architecture of the algorithm. For the algorithm, we naturally went from a decision tree to a random forest algorithm. We set up more drastic changes on the way both the enzyme sequences and the chemicals are encoded. We trained 8 different predictors, depending on encoding and features selections. We finally evaluated the obtained predictors with a large new dataset, demonstrating the generation of a generic and accurate pan-predictor, able to evaluate any GT1:acceptor pair.

# RESULTS

**Generation of experimental data for algorithm training.** An in-house library of 24 GT1 enzymes (attempted production on 60 sequences, table S1) were screened for activity on 30–80 natural product or xenobiotic acceptors with at least one glycosylation site (a hydroxyl, primary amine, or thiol) (Figure 1B, Table S2) using an enzyme-coupled assay developed to obtain apparent rates. Out of all 1021 assayed enzyme:acceptors pairs, 81 led to significantly higher rates than the background, 30 proved to be inconclusive (Figure S1), and 910 were considered unreactive. GT1 reactivity data is publicly available as both reactivity rates and Boolean reactivity values, *i.e*., classification of whether a specific GT1 enzyme and a specific acceptor are considered reactive or not. Rates were converted to reactivity Booleans, rather than discarding classification data in an effort to predict rates. Rates are converted to Boolean by assigning each datapoint as reactive if the measured rate is above a threshold. The rate data was classified as reactive by detecting outliers, due to the observation that most measurements are of non-reactivity, typically with a sharp contrast to a minor set of non-zero rates (Figure S1). The outlier detection is performed independently on each enzyme by assuming the measurements follows a normal distribution N(μ=0, σ=σ(measurements)), *i.e.* they are all non-reactive with non-zero rates occurring due to noise. From the distribution a p-value is calculated to quantify how extreme any of the measurements are. Adjusted p-values are calculated from the p-values with the Holm method (default for R function p.adjust) which aims to control the family-wise error rate, i.e. adjust p-values for multiple testing. Measurements that have both p-values > 0.05 and adjusted p-value > 0.05 are considered to fit the null-hypothesis and are therefore classified as non-reactive observations, while measurements with both p-value < 0.05 and adjusted p-value < 0.05 does not fit the null-hypothesis, so are classified as observations of reactivity. Some datapoints have p-value < 0.05 but adjusted p-value > 0.05 which we consider inconclusive evidence, so those datapoints are discarded. The resulting classifications are summarized in Table 1.



Figure S1 Reactivity classification of rate measurements. Measured rate plotted versus ranking among measurements from all enzymes. Standard significant threshold 0.05. 3 enzymes shown as examples, see supplementary figure for remaining enzymes.

**Negative dataset.** All the above-described experimental datasets, whether from the literature, our own experimental, or the one used to train GT-predict certainly contain only molecules that can be substrates of GT1 enzymes. A first attempt at a predicting algorithm thus did not discriminate and predicted that a significant fraction of GT1 enzymes could glycosylate benzene, which is a chemical impossibility. We then leveraged the mechanistic knowledge of the reaction to balance the datasets with molecules that cannot be substrate for GT1 enzymes. Indeed, some minimal requirements are known for the potential of reactivity: as GT1 enzymes catalyse *O*-, *N*- and *S*-glycosylations via a type 2 nucleophilic substitution (SN2), a nucleophilic group is required (Figure 1A).6 *C*-glycosylation are likely aromatic electrophilic substitution (SEAr), yet also require a proton bearing heteroatom to be deprotonated.12,13 occurring The most common nucleophilic group is an hydroxyl, yet amino, thiol, or carboxylic groups can also be reactive//ref. This simple rule can be learnt in training by adding generated negative datapoints without these reactive groups. We do so in the following steps:

Collect all positives, which are acceptors from our curated experimental data that are reactive for at least one GT1 enzyme.

Discard acceptors with R-NH, R-NH2, or R-NH-R’.

Replace any carboxyl (COOH) with COOMe.

Discard any acceptor without a hydroxyl group.

Remove (replace with H) any hydroxyl group or thiol group. Since these molecules contains no nitrogen, thiol, carboxyl or hydroxyl they will not be reactive with any GT1 enzyme, so we refer to them as negatives.

Make a copy of each of these negatives with either a flour, chlorine, OMe, or OAc substitution at a site where a hydroxyl or thiol group had been present in the positives. Only a single randomly selected site gets a substitution, which means exactly five negatives will be generated for each given positive (a version with H, Fl, Cl, OMe, and OAc).

Generate SMILES for each of the negatives.

Convert generated SMILES to CIDs using webchem.

Convert the SMILES listed on pubchem to canonical SMILES.

Check that the SMILES from pubchem and the generated ones match.



**Figure 1. GT1 enzymes acceptor scope.** A) GT1 mechanism, illustrated for the *O*-glycosylation of the polyphenol genistein (grey). Residues numbering according to PtUGT1. B) Various acceptors from the acceptor screen. Reactive groups are shown in red. C) Examples of unreactive acceptors derived from 3,4 dichloroaniline or 3,4 dichlorothiophenol. The replaced group is shown in red.

Table 1 Reactivity data. Reactivity data curated from Boolean sources and converted from rate datapoints. \*Public rate data was not converted to Booleans with the outlier detection approach. Instead, all 62 datapoints were assumed to be reactive observations.

|  |  |  |  |
| --- | --- | --- | --- |
| Raw type | Source | Reactive | Count |
| bool | GT-Predict | no | 2235 |
| bool | GT-Predict | inconclusive | 164 |
| bool | GT-Predict | yes | 653 |
| rate | HTS\_UGT | no | 485 |
| rate | HTS\_UGT | inconclusive | 18 |
| rate | HTS\_UGT | yes | 55 |
| rate | pTMH | no | 435 |
| rate | pTMH | inconclusive | 12 |
| rate | pTMH | yes | 26 |
| bool | public | no | 36 |
| bool | public | yes | 10 |
| rate\* | public | yes | 62 |
| bool | negatives | no | 15912 |

**GT-Predict assessment.** GT-Predict was tested on in-house data (see [Experimental data](#_Experimental_data)) to validate the feasibility of accurate prediction on unseen data using simple machine learning methods and limited by the scale of data availability for GT1 reactivity.

First the enzymes from the dataset HTS\_UGT (see [Experimental data](#_Experimental_data)) was given to GT-Predict, which when given enzyme sequences as input finds the nearest neighbor enzyme by Smith-Waterman alignment scoring and returns all Boolean reactivity values from GT-Predicts dataset. These scores were then filtered for overlap of acceptors measured in HTS\_UGT and compared (see Figure 1).



Figure S2 GT-Predict returns reactivity values consistent with HTS\_UGT data. Inset: zoom on y-axis.

The HTS\_UGT reactivity rates were highly consistent with the corresponding reactivity classifications from GT-Predict (AUC=0.814 and correlation=0.276) demonstrating that sequence similarity in GT1 is a reliable indicator of similar reactivity preference.

**Predictors development.** Discussions about the training, choices and internal validation/results. Explanations about biotechnological predictors (sharp start of the ROC curve) versus biological ones (minimizing AUC). One-hot vs Blossum62 encoding. Choice of data excluded for validation / preventing overfitting. Adding negative acceptors for balancing the dataset.

**Machine learning architecture.** Random Forest Boolean Classifiers were trained and tested since they are a natural extension to the previous work in GT-Predict and are quickly trained compared alternatives such as neural networks//ref. Random Forest prediction scores are averages of the decisions from each of multiple separate decision trees (the forest). Each tree is exposed to a different random subset of the input features. All tests were run with 10000 trees.

**Pan-specific prediction.** A predictor specific to a fixed set of acceptors and enzymes restricts prediction on unmeasured candidates. We aim to build a pan-specific predictor that is able to take any GT1 enzyme or potential acceptor as input. This is simply achieved by encoding both the acceptors and enzymes as numerical model inputs (features) which can differentiate any two acceptors or enzymes that have different reactivities. The domain of enzymes and acceptors for which it will perform adequately will depend on a number of factors, such as the chemical landscape of the acceptors, variation among GT1 subfamilies and the general complexity of the chemical reaction.

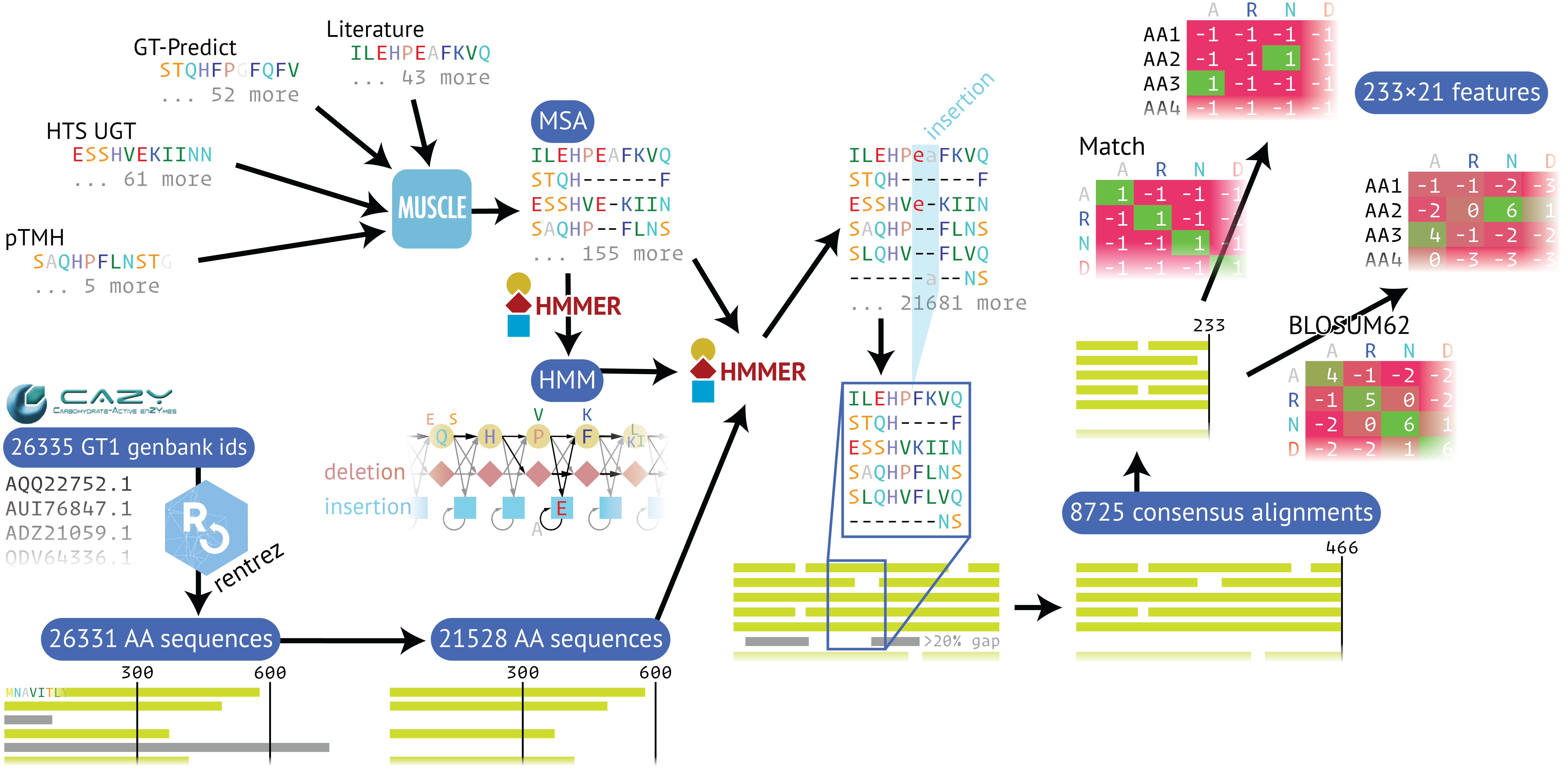
**Raw data curation.** Reactivity data (Table 1) was curated from //Publications found by Sebastian, from GT-Predict as well as measured experimentally in-house (see [Experimental data](#_Experimental_data)). As part of the GT-Predict publication a dataset was shared with 3052 measurements between 53 enzymes and 59 acceptors.

**Retrieving and curating GT1 sequences.** The CAZy database describes carbohydrate-active enzymes,3 and contains 32728 genbank ID entries for GT1 enzymes, which were curated for discovery of their potential reactivity with acceptors of interest. Amino acid sequences were curated for 26335 unique genbank IDs using the entrez REST API (rentrez//ref). 21528 GT1 sequences were found after discarding any sequence shorter than 300 amino acids or longer than 600.

**Enzyme features preparation.** Features generated from unaligned sequences can also be inputs for machine learning architectures that accepts varying input lengths, such as Recurrent Neural Networks, or some designs of Convolutional Neural Networks. However, we applied the following architecture, as such neural network approaches could eventually be more suitable for future work including more experimental data (Figure 2).

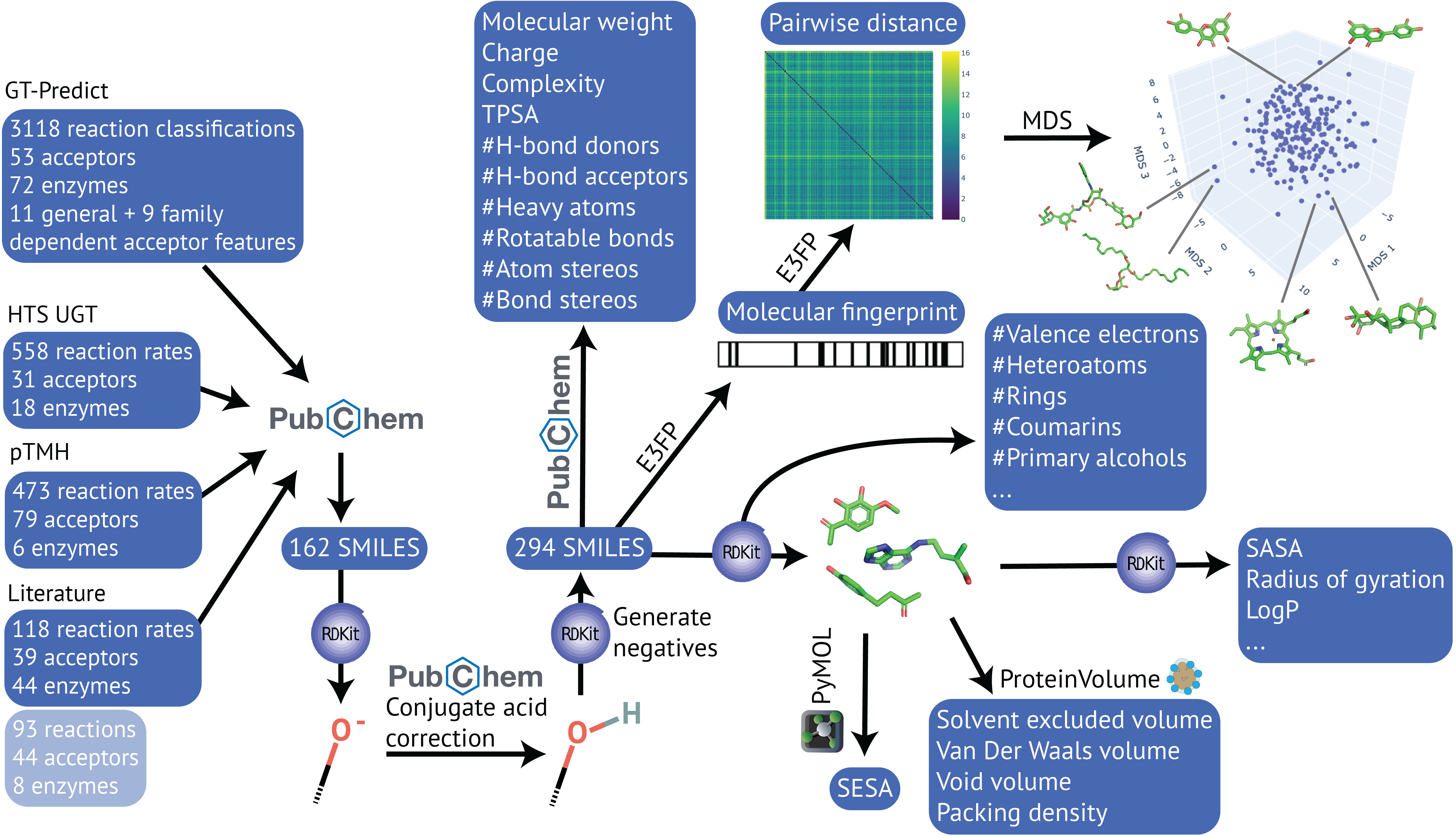
1. Amino acid sequences were curated for all GT1 enzymes with reactivity measurements then aligned using MUSCLE//ref
2. A Hidden Markov Model (HMM) was then built from the alignment using HMMER.
3. The HMM was used for aligning the full set of available sequences, which includes available GT1 sequences curated from CAZy.
4. Non-consensus positions of the resulting alignment were discarded, where a consensus location is identified as the majority of sequences containing the same letter for this location.
5. Low quality alignments were discarded, which were sequences with <80% match to the consensus sequence, which is the most frequent letter for each location.
6. Since the N-terminus is the region of GT1 expected to be most important for binding preference, the last half of each sequence was discarded.
7. Amino acids are encoded as numerical features using a substitution matrix (see sec. [Enzyme encoding](#_Enzyme_encoding)).

**Enzyme encoding.** Enzymes can be encoded in numerous ways, such as capturing the 3D coordinates of the backbone//ref. Typically, the amino acid sequence is encoded instead, which has far larger availability than 3D conformations from crystallized protein imaging. A protein sequence can be encoded for a machine learning model by representing each letter of its alphabet as a separate vector of numbers, where each vector has a length equal to the alphabet size. As such we are selecting columns from a, usually symmetric, encoding matrix where entries indicate a notion of similarity between two letters of the alphabet. The simplest such matrix is the identity matrix which is known as sparse or one-hot encoding. We used a slightly different encoding, referred to as “match” encoding from NCBI//ref, where off-diagonal elements are -1 instead of 0 and where the gap character has a similarity of zero with itself, as opposed to 1. Furthermore, encodings for ambiguous amino acids are included, where entries for the ambiguous amino acids sum to 1, e.g. the letter B symbolizes N or D, so the vector representation contains -1 for all entries except ½ for N and D. BLOSUM (BLOcks SUbstitution Matrix) encoding is another typical encoding matrix designed to capture amino acid similarities from studies of evolutionary divergence. Multiple versions exist to capture amino acid sequence similarity at different evolutionary scales. We used the standard BLOSUM62 which is created from sequences with less than 62% similarity.



**Figure 2.** Enzyme feature generation.

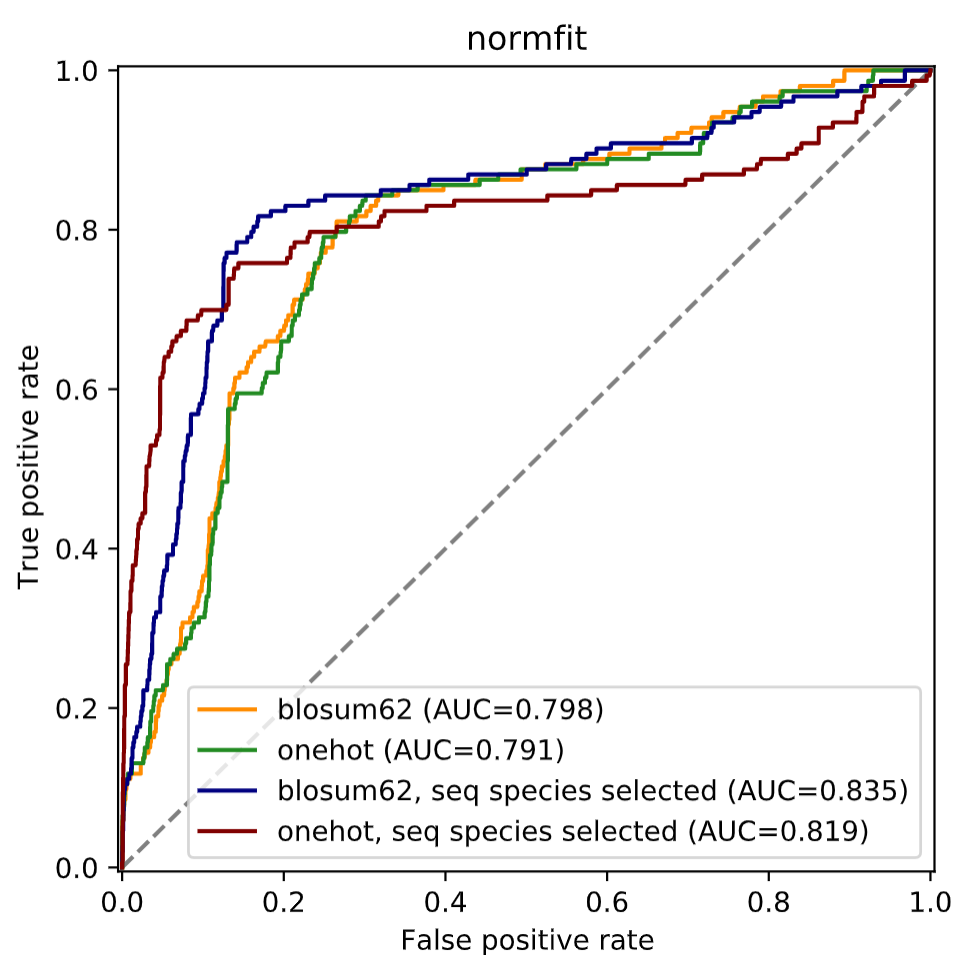
**Chemical feature generation pipeline.** A pipeline was designed to automatically generate features for acceptors to avoid manually curating data for each new candidate. This required that all features were general enough to exist for any family of acceptor that we may consider. Given a systematic identifier, such as PubChem CID, CAS number, INCHI, INCHI-key, or SMILES the pipeline curates and generates 153 chemical features (see supplementary table for full list). Features are curated from PubChem//ref, and generated from open source PyMol//ref, RDKIT//ref, ProteinVolume//ref, and E3FP//ref.



**Figure 3. C**hemical features generation pipeline.

1. Identifiers are converted to PubChem CIDs using webchem//ref.
2. SMILES and molecular properties on PubChem are generated from CIDs using webchem.
3. Acceptors that have been given as conjugate bases are converted to acids by finding CO- using RDKit and converting to COH, then finding the CIDs that correspond to the updated SMILES. Step 1 and 2 are then repeated once to account for the changes.
4. SMILES are given to RDKit which generates a molecular representation, including 3D conformers and from this a large number of features. The 3D conformers are written to PDBs.
5. Area features are generated using open source PyMol given the PDBs.
6. Volume features are generated using ProteinVolume given the PDBs.
7. Features are generated from molecular fingerprints.
   1. Molecular fingerprints are generated using E3FP given SMILES.
   2. All pairwise Euclidean distances are calculated between the molecular fingerprints using E3FP.
   3. Pairwise distances are converted to projected points in an n-dimensional space using MultiDimensional Scaling (MDS) which are dimensionally reduced features representing the molecular fingerprints.
8. Features from all steps above are concatenated.

**Training and evaluation of predictors.** For Christian.

****

**Figure 4.** Training and evaluation of predictors. For Christian. Figure1. Evaluation of several predictors during training.

Additionally to the evaluation on the test sets, we wanted to generate new data to evaluate in a truly unbiased fashion our predictors. Particularly, as the two known existing predictors (ESP11 and GT-predict2) fail to provide accurate predictions about molecules absent from their training datasets, we wanted to assess how our predictors performed towards molecules it was naïve towards.

**Parsing the sequence space to identify catalysts.** To assess the usefulness of the predictor to hunt for the right catalyst within all available sequences, we ranked all sequences from the CAZy database (xx.xxx as of date) for their likelihood to glycosylate indoxyl. Indeed, the glucoside of indoxyl, indican, is the natural precursor of indigo and a potential road for sustainable biotechnological denim dyeing. Among the top scorers was *Pt*UGT1, which is both the natural catalyst for this reaction in the indigo plant, and was demonstrated as a potent catalyst for indicant synthesis (ref BioRxiv, nat chem biol). Another top scorer among the enzymes of the training dataset was *Nt*UGT from the tobacco plant, which indeed glycosylated indoxyl and allowed for the synthesis of indican. Interestingly, *Nt*UGT and *Pt*UGT1 were predicted and found active on indoxyl, while both enzymes were inactive on its closest chemical analogue in the dataset, indole-3-carboxylic acid. While there was a clear bias favoring enzymes from the training dataset, the “biotechnological predictor” correctly identified useful catalysts.

**Assessment of the predictors on newly generated datasets.** Given that the first selected chemical feature was the presence of a phenolic oxygen, together with the fact that from both the GT-predict dataset and our dataset, all enzymes were active on some phenols or polyphenols, we decided to validate our data on such acceptors. Indeed, they represent more interesting biotechnological glycosylation targets, due to i) their numerous health benefits, ii) their low solubility without glycosides; and iii) the increase complexity for chemical glycosylation, due to the necessity to ensure regioselectivity. Moreover, it allows to assess whether the predictor could discriminate between rather subtle differences. We thus generated a new dataset, evaluating 47 (poly)phenolic acceptors absent from the training sets (Figure 5A), against 8 new GT1 sequences selected to be as different as possible from the one present in the training set. We also assessed 1833 enzyme-substrate pairs using the same 47 acceptors with enzymes from our in-house library (Figure 5C). Description of enzyme selection (clustering + soluprot), production, assay. Description of results.

|  |  |
| --- | --- |
|  |  |
|  |  |

**Figure 5.** Assessment of the predictor against a large, naïve dataset. A) Representative acceptors of the various families of structures used to produce the “naive” dataset. The full list is available in table SX. B) ROC curves on the algorithms on a “naïve” dataset, with both sequences far from the dataset and new acceptors with high similarity. C) experimental results. Yields of glycosylation for the conversion of 50 µM acceptor in presence of 60 µM UDP-Glc, and 20 µg/mL enzyme, during OVN reaction at 293 K in HEPES 25 mM NaCl 50 mM pH7. D) predictions of the same.

It is particularly interesting to observe that the predictor keeps a high accuracy despite the methodology used for the training dataset (detection of products by MS and apparent initial rates by an indirect assay) and in the validation dataset (glycosylation yields observed by reverse phase HPLC) being fundamentally different. Together with the variability in sequences and chemicals proposed between training and evaluation datasets, this result speak for a particularly robust predictor.

Also markedly different from the known one, being specifically for GT1s (can we actually try GT predict on our evaluation dataset, even if it makes little sense?) or for enzyme in general (here we put the data from the one from BioRxiv).

**Maybe?** Comments on the features selected, both on the chemical side, and on the sequence/structure (Fig. ¾). Comments on the encoding. FIGURES. Figure 1: 4 panels ROC curves, with GT predict, at some of our steps (let’s forget we did not set thresholds correctly at the beginning, but the improvement by balancing the dataset, and the two training modes (max AUC, max initial slopes) could be there.

From the training and using sequence selection \*insert by Christian\*, the most relevant features according to the algorithm can be highlighted. The first selected chemical feature was the presence of a phenolic oxygen, which is consistent with what is known in the literature, with all GT1s displaying activity against phenols, and most against polyphenols (ref GT predict + others). The other most selected features were related to the total volume of the acceptor, and the MSD (have to read more here, or insert by Christian).

On the structure, pinpointing residues: big question on Christian: are the one-hot predictors not barcoding the enzymes? I mean taking any feature that will pinpoint a specific sequence or group of sequences.

# Perspectives / conclusions

In this work, we demonstrated the synergistic effect of high-throughput data generation with chemically-informed predictors training. Indeed, we proposed the most accurate and robust enzyme specificity predictor to date, obtained through the generation of the largest experimental dataset on GT1s. Using knowledge about the enzymatic structure, we could focus on the relevant part of the enzymes’ sequences, and knowledge about the mechanism allowed for the generation of a balanced dataset by constructing artificial yet true negatives – contrarily to the assumption used by ESP that any non-descript molecule similar to a described substrate is not a substrate by definition. Particularly, the predictor is consistent despite the very different types of data presented in training set (various format from literature, product detection from GT-predict, apparent rates pTMH and HTS) and in the algorithm evaluation (reaction yields by HPLC). Importantly, the algorithm discriminate very well between close sequences and close molecules, and perform well on molecules absent from the training set. Moreover, the addition of new data is straightforward, thus the structure exist for predictors improvements in the future along with the availability of new data. Furthermore, we provide a pipeline for automated features generation on chemicals, which could be used for any other enzyme class.

It is also particularly interesting to observe that the algorithm is largely successful despite structurally necessarily misses: loops of varying length near the active site cannot be accounted for, while such loops are known to have a strong impact in CAZymes specificity, including in GT1s.14 Indeed, those are impossible to properly account for and evaluate within a large MSA. Yet, our predictor performs well despite missing these information. With the recent release of AlphaFold215 and the wealth of accurate structural model it provides, it might be feasible to actually account for both the protein tridimensional structure, and active site loops, similarly to what has been done for the predictions of binding parameters of cellulases.16 An important drawback of the predictors is that they cannot currently provide information over regioselectivity. Indeed, the regiochemical outcome of GT1 glycosylation are currently not informed, neither in our datasets or in the GT-predict one, nor in most of the literature, and thus predictors cannot be trained for it yet.

# TABLES

# ASSOCIATED CONTENT

Detailed data on biochemistry (all Tiia’s results, also all soeren as yields), I guess the algorithm as well?

# AUTHOR INFORMATION

Corresponding Author

\*(Word Style “FA\_Corresponding\_Author\_Footnote”). \* (Word Style “FA\_Corresponding\_Author\_Footnote”). Give contact information for the author(s) to whom correspondence should be addressed.

Present Addresses

†30 Royal Parade, Parkville VIC 3052, Australia..

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally.

# Funding Sources

The Novo Nordisk Foundation Grants NNF18OC0034744, NNF10CC1016517, and NNF20CC0035580..

# Notes

Any additional relevant notes should be placed here.

# ACKNOWLEDGMENT

The authors thank Folmer Fredslund for selecting sequences for the algorithm validation, and Sebastian for literature mining.

# ABBREVIATIONS

GT1, Glycosyltransferase family 1; AUC, Area Under the ROC Curve; ....

REFERENCES

(1) Nidetzky, B.; Gutmann, A.; Zhong, C. Leloir Glycosyltransferases as Biocatalysts for Chemical Production. *ACS Catal.* **2018**, *8* (7), 6283–6300.

(2) Yang, M.; Fehl, C.; Lees, K. V.; Lim, E. K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G. Functional and Informatics Analysis Enables Glycosyltransferase Activity Prediction. *Nat. Chem. Biol.* **2018**, *14* (12), 1109–1117.

(3) Lombard, V.; Ramulu, H. G.; Drula, E.; Coutinho, P. M.; Henrissat, B. The Carbohydrate-Active Enzymes Database (CAZy) in 2013. *Nucleic Acids Res.* **2014**, *42* (November 2013), 490–495. https://doi.org/10.1093/nar/gkt1178.

(4) Bidart, G. N.; Putkaradze, N.; Fredslund, F.; Kjeldsen, C.; Ruiz, A. G.; Duus, J. Ø.; Teze, D.; Welner, D. H. Family 1 Glycosyltransferase UGT706F8 from *Zea Mays* Selectively Catalyzes the Synthesis of Silibinin 7‑O‑β‑D‑Glucoside. *ACS Sustain. Chem. Eng.* **2022**. https://doi.org/10.1021/acssuschemeng.1c07593.

(5) Ross, J.; Li, Y.; Lim, E.-K.; Bowles, D. J. Higher Plant Glycosyltransferases. *Genome Biol.* **2001**, *2* (2), reviews3004.1-3004.6.

(6) Teze, D.; Coines, J.; Fredslund, F.; Dubey, K. D.; Bidart, G. N.; Adams, P. D.; Dueber, J. E.; Svensson, B.; Rovira, C.; Welner, D. H. *O*‑/*N*‑/*S*‑Specificity in Glycosyltransferase Catalysis: From Mechanistic Understanding to Engineering. *ACS Catal.* **2021**, *11* (11), 1810–1815.

(7) He, J.; Zhao, P.; Hu, Z.; Liu, S.; Kuang, Y.; Zhang, M.; Li, B.; Yun, H.; Qiao, X. Molecular and Structural Characterization of a Promiscuous *C*-Glycosyltransferase from *Trollius Chinensis*. *Angew. Chem. Int. Ed.* **2019**, *131* (131), 11637–11644. https://doi.org/10.1002/ange.201905505.

(8) Zhang, L.; Wang, D.; Zhang, P.; Wu, C.; Li, Y. Promiscuity Characteristics of Versatile Plant Glycosyltransferases for Natural Product Glycodiversification. *ACS Synth. Biol.* **2022**, *11* (11), 812–819. https://doi.org/10.1021/acssynbio.1c00489.

(9) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16* (August), 687–694. https://doi.org/10.1038/s41592-019-0496-6.

(10) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10* (10), 1210–1223. https://doi.org/10.1021/acscatal.9b04321.

(11) Kroll, A.; Ranjan, S.; Engqvist, M. K. M.; Lercher, M. J. The Substrate Scopes of Enzymes : A General Prediction Model Based on Machine and Deep Learning. *BioRxiv* **2022**, 1–28.

(12) Putkaradze, N.; Teze, D.; Fredslund, F.; Welner, D. H. Natural Product C-Glycosyltransferases – a Scarcely Characterised Enzymatic Activity with Bioyechnological Potential. *Nat. Prod. Rep.* **2020**, *00*, 1–12. https://doi.org/10.1039/D0NP00040J.

(13) Tegl, G.; Nidetzky, B. Leloir Glycosyltransferases of Natural Product C-Glycosylation : Structure, Mechanism and Specificity. *Biochem. Soc. Trans.* **2020**, *48* (48), 1583–1598.

(14) Brazier-Hicks, M.; Offen, W. A.; Gershater, M. C.; Revett, T. J.; Lim, E. K.; Bowles, D. J.; Davies, G. J.; Edwards, R. Characterization and Engineering of the Bifunctional N- and O-Glucosyltransferase Involved in Xenobiotic Metabolism in Plants. *Proc. Natl. Acad. Sci.* **2007**, *104* (51), 20238–20243.

(15) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(16) Schaller, K. S.; Kari, J.; Borch, K.; Peters, H. J.; Westh, P. Binding Prediction of Multi-Domain Cellulases with a Dual-CNN. *arXiv 2207 . 02698v1 [ Phys. . bio-ph ] 6 Jul 2022*.

BRIEFS (Word Style “BH\_Briefs”). If you are submitting your paper to a journal that requires a brief, provide a one-sentence synopsis for inclusion in the Table of Contents.

SYNOPSIS (Word Style “SN\_Synopsis\_TOC”). If you are submitting your paper to a journal that requires a synopsis, see the journal’s Instructions for Authors for details.

SI

Table 1

# Supplementary Table . Uniprot accession numbers and melting temperatures of the enzymes used in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Enzyme name | Organism of origin | Uniprot ID | *T*m (°C) |
| *Zm*\_UGT72G3 | *Zea mais* | B6SRY5 | 56.4±0.1 |
| *Zm*\_UGT72G4 | *Zea mais* | B4F9H1 | 61.2±0.4 |
| *Zm*\_UGT708A6 | *Zea mais* | A0A096SRM5 | 37.4±0.1 |
| *Zm*\_UGT88C10 | *Zea mais* | C0HFA0 | 42±0.3 |
| *Zm*\_UGT706F8 | *Zea mais* | B4FG90 | 47.8±0.5 |
| *Zm*\_71B1 | *Zea mais* | A0A1D6ICF2 | 57.5±0.2 |
| *Sl*\_UGT72B68 | *S. lycopersicum* | D7S016 | 33.1±0.2 |
| *Rh*\_GT1 | *Rosa hybrid cultivar* | Q4R1I9 | 40.9±0.2 |
| *Os*\_88C1 | *Oryza sativa (rice) japonica* | Q8LJ11 | 45±0.6 |
| *Lc*\_72B10 | *Lycium chinense* | B6EWZ3 | 52.1±0.2 |
| *Gm*\_88E3 | *Glycine max  (soybean)* | A6BM07 | 37.4±0.2 |
| *Fi*\_88A10 | *Forsythia x intermedia* | D2KY82 | 40.4±0.4 |
| *Fe*\_88J1 | *Fagopyrum esculentum* | A0A0A1H7N8 | 57.9±0.3 |
| *At*\_71D1 | *A. thaliana* | O82383 | NA |
| *At*\_71C1 | *A. thaliana* | O82381 | NA |
| *Pt*UGT1 | *Polygonum Tinctorium* | A0A2R2JFJ4 | 49.3±0.2 |
| *At*\_UGT72E2 | *A. thaliana* | Q9LVR1 | NA |
| *Mt*\_UGT78G1 | *Medicago truncatula* | A6XNC6 | NA |

**Supplementary Table 2. Enzyme:acceptor pair generated and used in the training dataset.** 

**Supplementary Table 3. Enzyme:acceptor pair generated and used to evaluate the predictors.** 

Supplementary Table: Chemical features

// in github GT/writing/SupplTable\_chemFeatures.tsv

Supplementary figure: Reactivity rate classification



M&M

Second dataset generation.

From a polyphenolic natural compound library (Target Molecule, U.S., 250 µL at 10 mM in DMSO-d6), 47 acceptors with molecular weights 200–400 kDa were selected. Reactions were performed without stirring, in presence of 50 µM acceptor, 60 µM UDP-Glc, and 20 µg/mL enzyme, during OVN reaction at 293 K in 100 μL HEPES 25 mM NaCl 50 mM pH7. Reactions are then diluted by the addition of 140 μL of milliQ H2O, and analyzed trough reversed-phase chromatography with an Ultra-High-Performance Liquid Chromatograph (UHPLC) Dionex UltiMate 3000 (Thermo Fisher Scientific, USA). Analysis is performed with a multistep gradient of acetonitrile (%B) and 0.1 % formic acid in milliQ H2O (%A), and at a flow rate of 1 mL/min: 0.00–0.50 min 2 %B, 0.51‒1.50 min 35 %B, 1.5–3.01 min 35–80 %B, 3.01–4.20 min 98 %B, 4.21–5.00 min 2 %B. A Kinetex 2.6 μm C18 100 Å 100x4.6 mm analytical column (Phenomenex, U.S.) is used at 40 °C. Absorption of every injection is measured at 280 nm and 300 nm and analyzed via the Chromeleon 7.2 software. Glycosylation products were measured under the inaccurate assumption that their extinction coefficient in acid was similar as the one of the acceptors.