Predicting Glycosylation Stereoselectivity Using Machine Learning

Sooyeon Moon,[1,2]† Sourav Chatterjee,[1]† Peter H. Seeberger,[1,2] Kerry Gilmore[1*]

[1] *Department of Biomolecular Systems, Max-Planck-Institute of Colloids and Interfaces, Am Mühlenberg 1, 14476 Potsdam, Germany*
[2] *Freie Universität Berlin, Institute of Chemistry and Biochemistry, Arnimallee 22, 14195 Berlin, Germany*

Corresponding Author
kerry.gilmore@mpikg.mpg.de
† These authors contributed equally to this work.

Abstract

Predicting the stereochemical outcome of chemical reactions is challenging in mechanistically ambiguous transformations. The stereoselectivity of glycosylation reactions is influenced by at least eleven factors across four chemical participants and temperature. A random forest algorithm was trained using a highly reproducible, concise dataset to accurately predict the stereoselective outcome of glycosylations. The steric and electronic contributions of all chemical reagents and solvents were quantified by quantum mechanical calculations. The trained model accurately predicts stereoselectivities for unseen nucleophiles, electrophiles, acid catalyst, and solvents across a wide temperature range (overall root mean square error 6.8%). All predictions were validated experimentally on a standardized microreactor platform. The model helped to identify novel ways to control glycosylation stereoselectivity and accurately predicts previously unknown means of stereocontrol. By quantifying the degree of influence of each variable, we discovered that environmental factors influence the stereoselectivity of glycosylations more than the coupling partners in this area of chemical space.

Predicting the outcome of an organic reaction generally requires a detailed understanding of the steric and electronic factors influencing the potential energy[1,2] surface[3] and intermediate(s).[4] Quantum mechanical calculations have significantly increased our ability to identify and quantify these factors. However, the correlation of these physical properties with reaction outcome becomes exceedingly challenging with each increase in dimensionality (*e.g.,* additional reaction participants, pathways). Layering onto this the additional and often subtle nuances impacting the regio- or stereoselectivity[5] of a reaction complicates proceedings.

Machine learning is a powerful tool for chemists[6,7] to identify patterns in complex datasets from composite libraries or high-throughput experimentation.[8] Chemical challenges including retrosynthesis,[9] reaction performance[10] and products,[11,12] the identification of new materials and catalysts,[13,14,15] as well as enantioselectivity[16,17] have been addressed. However, a significant challenge is predictability of reactions involving S_N1 or S_N1-type mechanisms[18] in the absence of chiral catalysts/ligands,[19] due to the potentially unclear mechanistic pathways resulting from the instability of the carbocationic intermediate.[16,17,20]

Glycosylation is one of the most mechanistically complex organic transformations,[20,21,22] where an electrophile (donor), upon activation with a Lewis or Brønsted-Lowry Acid, is coupled to a nucleophile (acceptor) to form a C-O bond and a stereogenic center. This reaction involves numerous potential transient cationic intermediates and conformations and can proceed via

mechanistic pathways spanning S_N1 to S_N2.[23] The stereochemical outcome is determined by more than eleven permanent (defined by the starting materials) or environmental factors (defined by the selected conditions/catalyst) whose degree of influence, interdependency, and relevance is poorly understood.[20,24,25] A systematic assessment of these factors on a flow platform allowed for the isolated interrogation of these variables. The empirical study identified general trends/influences of these factors (Figure 1) and hypothesized their relative rankings with respect to dominance.[24] However, a data sciences approach is required to fully understand and apply this knowledge for the accurate prediction of stereoselectivities of new coupling partners and conditions.

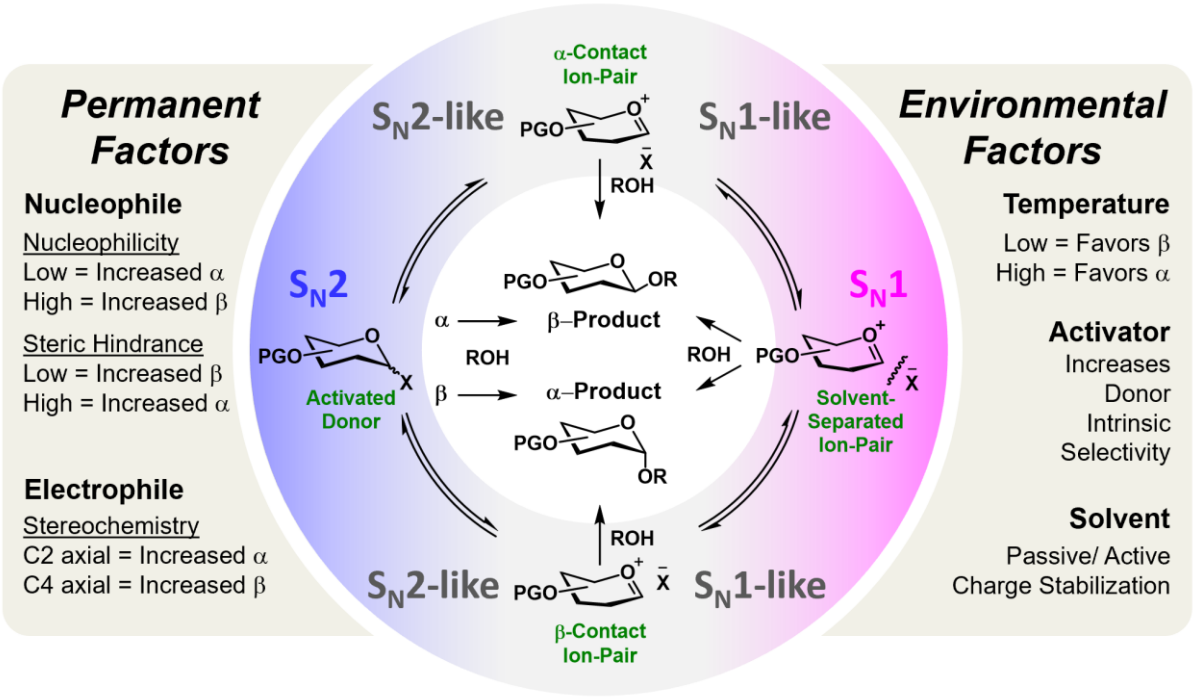

Figure 1. General representation of the potential mechanistic pathways of glycosylations leading to either the alpha (α) or beta (β) anomer of the formed C-O bond. The empirically-derived permanent and environmental factors and their influence on stereoselectivity are provided.[24]

We have trained a random forest algorithm using a dataset of glycosylation reactions with a variety of stereoselective outcomes to accurately predict the stereoselectivity of new glycosylations, varying coupling partners, acid catalyst, solvents, and temperature. Regression-based random forest algorithms have proven powerful in modeling chemical reaction performance.[10,26] This algorithm generates several weak models in the form of decision trees. The nodes of each of these decision trees are generated by random shuffling of the descriptors in the training set. The final model is an "ensemble" of a combined weighted sum of the decision trees, representing a collective decision of all individual trees that generate good predictions and reduces overfitting. The learning performance of the algorithm can be significantly enhanced by hyperparameter tuning (see Supporing Information).[27] Due to the heterogeneous nature of the descriptors in this work (*vide infra*), each tree was generated using the CART (classification and regression tree) algorithm with pruning, which does not require preprocessing or normalization.[28] An interaction-curvature algorithm was further utilized to reduce the selection bias of the split predictors of the standard CART algorithm (Figure 2).

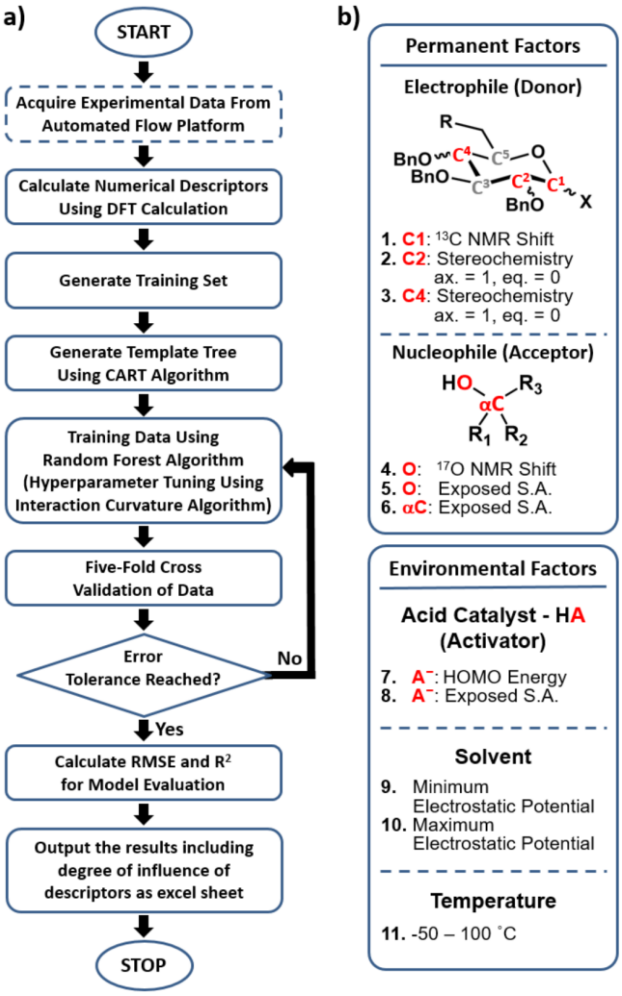

Figure 2. a) General workflow of the process from data input to prediction output. b) Calculated descriptors – either regressor or categorical – address the steric and electronic components of all chemical species in the reaction. S.A – surface area.

A set of numerical descriptors that accurately describe the relevant steric and electronic parameters of all reaction participants – starting materials, reagents, and solvent – is key to building an accurate, extrapolatable model to predict the subtle nuances of stereoselectivity. The concise nature of the training set (268 data points, see SI)[29,30] renders manual selection of descriptors – quantifying sterics/electronics – using chemical intuition[31] particularly important.[32] The training set is a lightly modified version of the dataset presented in our previous work,[24] removing two subsets of data (variance of the residence time and acceptor equivalents) and adding data for β-glucose donor (Table S1, lines 68-74, 101-106) and three additional solvents (Table S1, lines 238-268).

Structures of all starting compounds were optimized, and DFT calculations performed at the B3LYP 6-31G(d) or B3LYP 6-311G(d) levels of theory using SPARTAN (see Supporting Information). The lower level of theory was utilized for optimization of the donor molecules due to their size, and the values obtained were acceptable compared to those obtained at the more computationally expensive 6-311G(d) level of theory. The maximum number of potential descriptors per model was set to 18 to avoid overfitting.[33,34] The best-performing descriptors for each participant class were determined by the accuracy of the resultant trained models in predicting stereoselectivities of the entire validation dataset, containing variations in each participant class (electrophile, nucleophile, catalyst, solvent). Ten descriptors were identified

that, along with temperature, allow for the assignment of quantified values to the relevant steric/electronic properties of the chemicals involved.

The descriptors identified, described below, are either classified as regressors (intra-/extrapolatable values) or categorical (binary values). While the model can be developed solely using regressor values, it exhibits marginally poorer overall accuracy for the validation set tested and necessitates additional calculations (see Supporting Information and discussion below). The ability to interchange descriptors will facilitate the expansion of the developed model into adjacent or similar chemical subspaces as well as for multi-stage predictive algorithms, designing both reagents and environmental conditions to maximize the stereoselectivity of the desired transformation.

The key parameters needed to describe the electrophile were differences in the reactivity of the anomeric position and the orientations of the pyran ring substituents that may influence the selectivity through both conformational preferences[35] and hyperconjugative interactions.[36,37] The different leaving groups at the anomeric position were distinguished using the calculated ^{13}C NMR chemical shift,[38] which provided more clear distinctions between leaving groups than the ^{1}H NMR shift[39] of the anomeric proton. The relative orientations of the ether moieties around the pyran presented a challenge for descriptor selection, as our model performed well with both regressor and categorial descriptors. The accuracies of the three best performing descriptors (proton *J*-couplings around the ring, dihedral angles of the C-O bonds, and treating the relative axial/equatorial orientations of the substituents as binary) are shown in Figure 3. The binary classification is the most accurate and represents the simplest descriptor, and the loss of additional/more nuanced information provided by regressor values is, at present, acceptable.

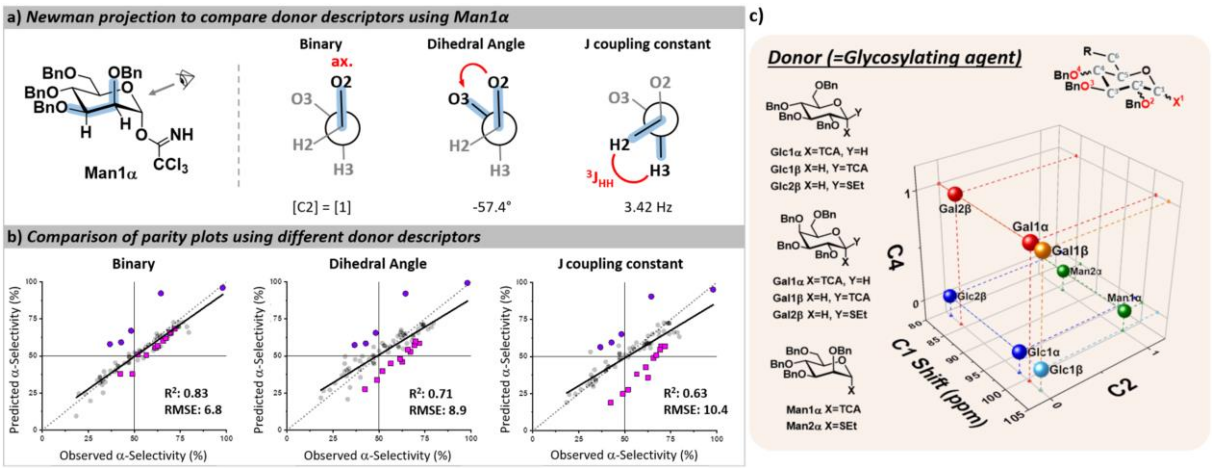

Figure 3. a) Three potential means of describing the stereochemistry of the ether groups around the pyran core. b) Parity plot of the resultant models using each set of descriptors for the donor (all also including the calculated ^{13}C NMR shift of C1). c) Three-dimensional map of the donor chemical subspace covered by the developed model, defined by the orientation of the C2 and C4 substituents on the pyran ring and the calculated ^{13}C NMR shift of C1. Glc – glucose, Gal – galactose, Man – mannose, Bn – Benzyl, TCA – trichloroacetimidate, SEt – ethylthio.

Observed nucleophile reactivity has been correlated with a range of parameters.[40,41,42] Where available, Mayr's nucleophilicity and Field inductive parameters correlate with glycosylation stereoselectivity.[43] To ensure general applicability, the ^{17}O NMR chemical shift of the oxygen nucleophile was calculated to capture the relevant hyperconjugative influences. The steric

environment of the nucleophile was described by the exposed surface areas of the oxygen and α-carbon in a space-filling model (Figure 4). While screening whether simple categorical descriptors can be utilized, specifically the whole values 0-3 to describe the substitution at the α-carbon, we found that the regressor value proved superior (see Supporting Information).

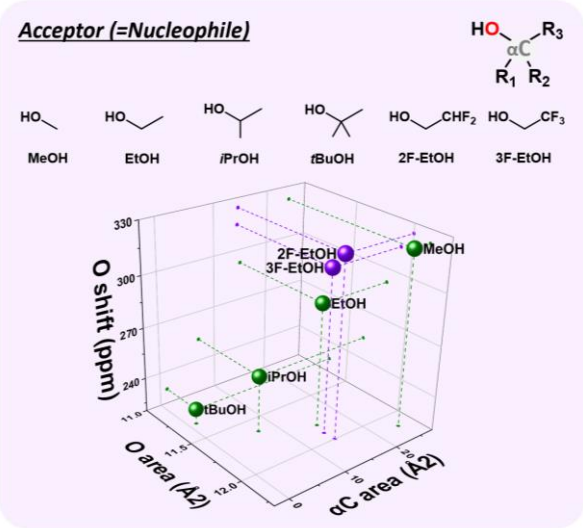

Figure 4. Three-dimensional map of the acceptor chemical subspace covered by the developed model, defined the by the exposed surface areas of the nucleophilic oxygen and the carbon alpha to the nucleophile as well as the calculated ^{17}O NMR shift. MeOH – methanol, EtOH – ethanol, iPrOH – isopropanol, tBuOH – *tert*-butanol, 2F-EtOH – 2,2-difluoroethanol, 3F-EtOH – 2,2,2-trifluoroethanol.

The chosen environmental conditions – solvent, acid catalyst, and temperature – are even more influential on the stereoselectivity than the intrinsic properties of the nucleophile and electrophile (*vide infra*). While regressor values for similar species have been calculated previously, the identification of the descriptors for acid catalysts relevant to this transformation was critical. The conjugate base of the acid catalyst has a significant impact on glycosylation stereoselectivity,[44] as evidenced by several studies observing an α-triflate intermediate[20,45] – the product of the conjugate base trapping the oxycarbenium ion.[46] Two values were identified that capture the nuanced role of this species (Figure 5a): the HOMO energy value of the conjugate base and the exposed surface area of the oxygen or nitrogen anion in a space-filling model.

While the influence of the solvent in glycosylations has been categorized by polarity and donicity (coordinating ability) values,[20] donicities are experimentally derived values and only available for select solvents. The calculated minimum and maximum electrostatic potentials describe the ability of the solvent to stabilize and interact with charged intermediates (Figure 5b). These descriptors perform well, such that even previously unreported means of solvent-control over stereoselectivity are accurately predicted (*vide infra*).

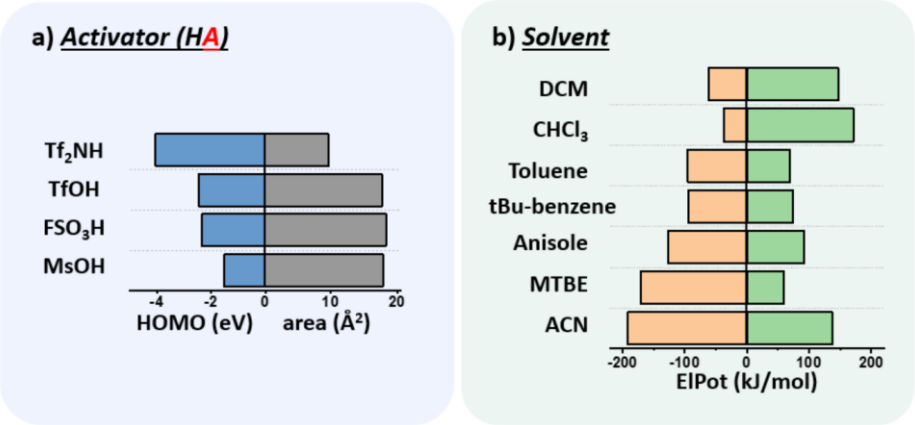

Figure 5. a) Plot of the descriptors used to quantify the relevant factors of the conjugate base of the activator. Area ($Å^2$) corresponds to the exposed surface area of the oxygen (O^-) or nitrogen anion (N^-) in a space-filling model. HOMO: highest occupied molecular orbital (eV). b) Plot of the descriptors used to quantify the relevant factors of the solvent. The maximum (MaxElPot), and minimum (MinElPot) values of the electrostatic potential(kJ/mol). Tf_2NH – bis(trifluoromethane)sulfonamide, TfOH – trifluoromethanesulfonic acid, FSO_3H – fluorosulfonic acid, MsOH – methanesulfonic acid, DCM – dichloromethane, $CHCl_3$ – chloroform, tBu-benzene – tert-butylbenzene, MTBE – methyl tert-butylether, ACN – acetonitrile.

The tuned random forest algorithm was trained using these descriptors on a dataset[24] containing systematic combinations of seven electrophiles, six nucleophiles, four acid catalysts, and seven solvents over a solvent-dependent temperature range of -50 to +100 ˚C (see Supporting Information). For comparison, three additional models were trained using gaussian process regression, support vector machine, and regression tree algorithms. Random forest proved superior (see Supporting Information). The model was then used to predict the stereoselectivities of a set of out-of-sample glycosylations varying each of the four chemical species in the reaction over the accessible temperature ranges. The predictions were validated experimentally using a microreactor platform.[24] The results of these predictions and validations are presented as the percentage of alpha product formed versus temperature. The corresponding parity plots for each of the out-of-sample sets are also provided (Figure 6).

The selectivity of electrophiles bearing phosphate leaving groups is accurately predicted to be similar[24] to those of glycosyl imidates and thioethers for glucose, galactose, and mannose donors, with a combined root mean square error (RMSE) of 2.0 (Figure 6a). The model can be applied to other pyran cores, such as L-fucose.[47] The predicted stereoselectivity of the fucose α-glycosyl imidate donor with isopropanol matches well with the experimental data (RMSE: 5.0), favoring the β-anomer at low temperatures and exhibiting a decrease in stereoselectivity with an increase in temperature (Figure 6b).

While the training set contains only simple alkyl alcohols as nucleophiles, the model accurately predicts the stereoselectivities of disaccharide formation. The predicted values for the coupling of α-galactose imidate with both glucose and mannose C6 alcohols matches well with the experimental data, albeit predicting a less α-selective process than observed (RMSE: 6.9 and 4.2, Figure 6d/e, respectively).

The model predicts more α-selective processes than experimentally observed in glycosylations using superacid 4,4,5,5,6,6-hexafluoro-1,3,2-dithiazinane-1,1,3,3-tetraoxide ($C_3F_6S_2O_4NH$) as acid catalyst. This deviation is seen at lower temperatures with galactose, however, the trend is

correct and has a low RMSE (5.5, Figure 6g). The weakest correlation of our model is observed for the $C_3F_6S_2O_4NH$-activated mannose coupling with isopropanol in DCM (RMSE: 19.3). Here, a stereoselective plateau is predicted at low temperatures with α–selectivity around 60% – as was observed experimentally for other activators with mannose.[24] However, experimentally the β-mannosylation product is mainly formed at low temperatures (-50 ˚C, 63% β-product). This finding is highly unexpected as β-mannosylation is challenging, generally requiring locked donor configurations.[21] With $C_3F_6S_2O_4NH$, the perbenzylated donor ranges from a 63% β-selectivity at -50 ˚C to 98% α-selectivity at 30 ˚C (Figure 6h).

Finally, the stereoselectivities of glucose and galactose α-imidate donors with isopropanol were predicted for two new solvents (Figure 6j/k). The strong influence of solvent[48] on the stereoselectivity of glycosylations is nicely captured by the descriptors chosen, and the model is accurate across a wide temperature range for both α,α,α-trifluorotoluene (RMSE: 6.2) and 1,4-dioxane (RMSE: 4.5).

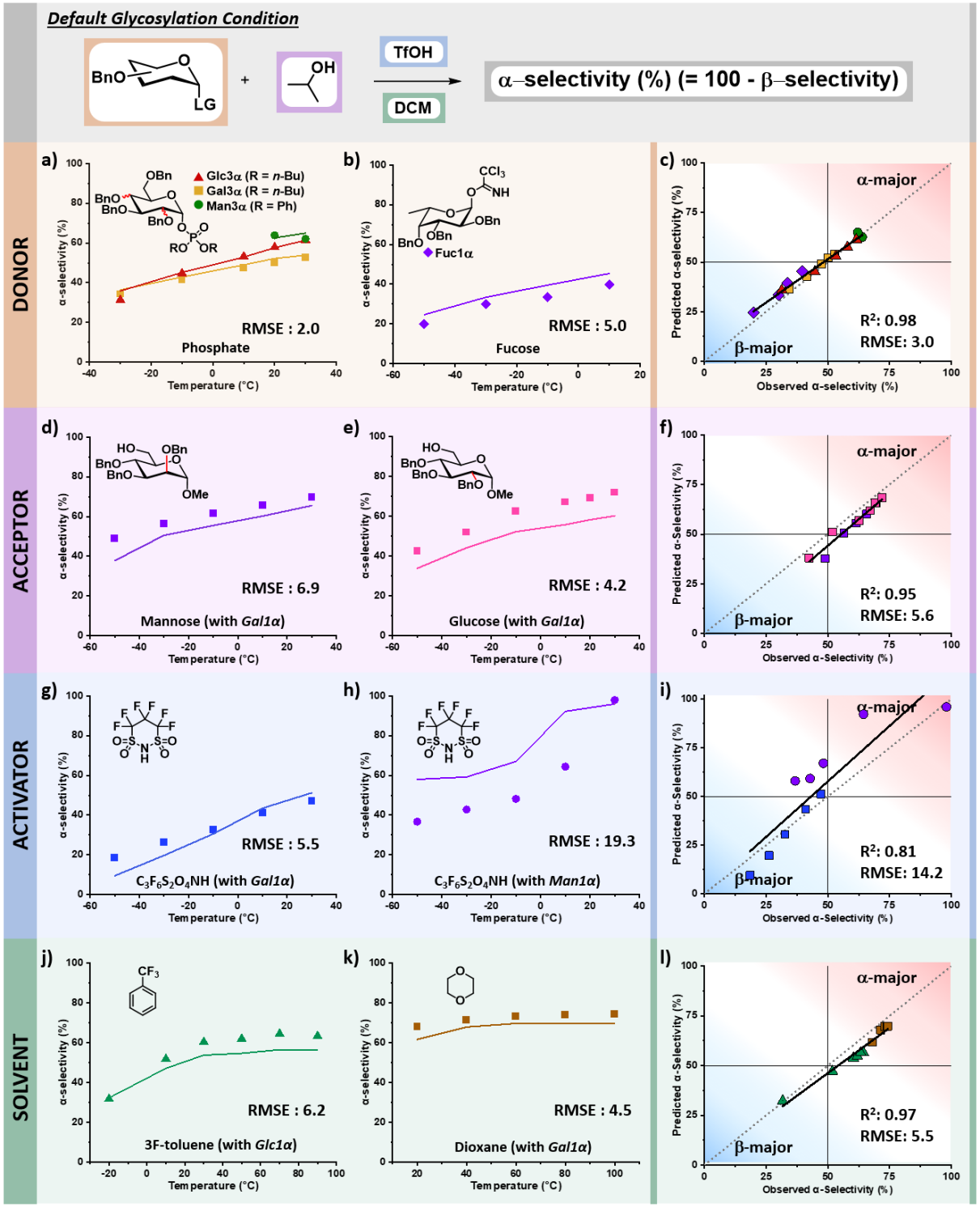

Figure 6. Prediction of stereoselectivity for glycosylations using different anomeric leaving groups, electrophiles, nucleophiles, activators, and solvents. a) Prediction of stereoselectivity for glycosylations involving a glycosyl phosphate leaving group. Bu – butyl, Ph – phenyl, RMSE – root mean square error. b) Prediction of stereoselectivity using a fucose (Fuc) donor with iPrOH in DCM. c) Parity plot of donor (electrophile) predictions. d,e) Prediction of mannose and glucose acceptor, respectively, with galactose α-imidate donor in DCM. f) Parity plot of acceptor (nucleophile) predictions. g) Prediction of 4,4,5,5,6,6-hexafluoro-1,3,2-dithiazinane 1,1,3,3-tetraoxide ($C_3F_6S_2O_4NH$) activator with galactose donor and iPrOH acceptor in DCM. h) Prediction of $C_3F_6S_2O_4NH$ with mannose donor and iPrOH in DCM. i) Parity plot of activator (acid catalyst) predictions. j) Prediction of α,α,α-trifluorotoluene (3F-toluene) solvent with glucose α-imidate donor and iPrOH in DCM. k) Prediction of 1,4-dioxane

solvent with galactose α-imidate donor and iPrOH in DCM. l) Parity plot of solvent predictions. Figure code: fucose (◆); glucose (▲); galactose (■); mannose (●); experimental (data points); predicted (solid colored line).

While the descriptors were chosen based on the current understanding of glycosylations, we wondered whether the model could also navigate newly discovered mechanistic peculiarities that influence stereoselectivity. One factor that is generally not considered significant while performing glycosylations is the orientation of the anomeric leaving group.[49] No influence of the α/β-orientation of the leaving group in dichloromethane was reported (Figure 7a),[24] and divergences in stereoselectivity based on this factor have sparingly been observed in the literature, *e.g.,* when phenylsilicon trifluoride ($PhSiF_3$) is used as a catalyst.[50]

The ability to use solvent to turn on and off the influence of leaving group orientation on glycosylation stereoselectivity has, to the best of our knowledge, not previously been reported. While essentially identical behavior is observed in DCM and chloroform, a slight divergence in MTBE at low temperatures is observed, with an 11% difference at -50 ˚C where the β-donor reaches 96% α-selectivity. This variable becomes important in toluene. Glucose β-imidate donor yields almost unchanged stereoselectivity (~60% alpha) over a 120 ˚C range! The orientation of the leaving group of the donor influences the stereoselectivity by more than 40% at -50 ˚C (Figure 7b).

With this limited data in our training set (Figure 7a/b), we tested the ability of our model to predict the influence of other factors on this to-date unreported phenomenon. The stereoselectivity of glucose α-imidate with ethanol as acceptor ranges from 10 – 54% α-product in toluene. The model predicts that the β-donor will behave differently, with a much less selective coupling overall (37% – 56% α-product). This prediction matches well with the experimental results, with an RMSE of 4.4 over the 120 ˚C range, though the process is less α-selective than predicted at low temperatures (Figure 7c). Conversely, the model predicts a less α-selective reaction at low temperatures than observed with *t*-BuOH as acceptor, though at higher temperatures, the prediction matches well with the experiment (RMSE: 6.4, Figure 7d).

Lastly, we sought to explore whether this additional mechanistic complexity exists for other electrophiles (Figure 7f/g). The model predicts that the α/β-galactose donors, when coupling with isopropanol, will give similar α-selectivity in DCM over the 80 ˚C temperature range, matching experimental results (RMSE 3.1, Figure 7f). In toluene, the model predicts a divergence in stereoselectivity at low temperatures, though not as large as what is observed with glucose. This prediction again aligns with experimental results (RMSE: 3.7, Figure 7g). Overall, the model correctly predicts the previously unknown ability to turn on and off the influence of the donor leaving group's orientation using solvents under otherwise identical conditions.

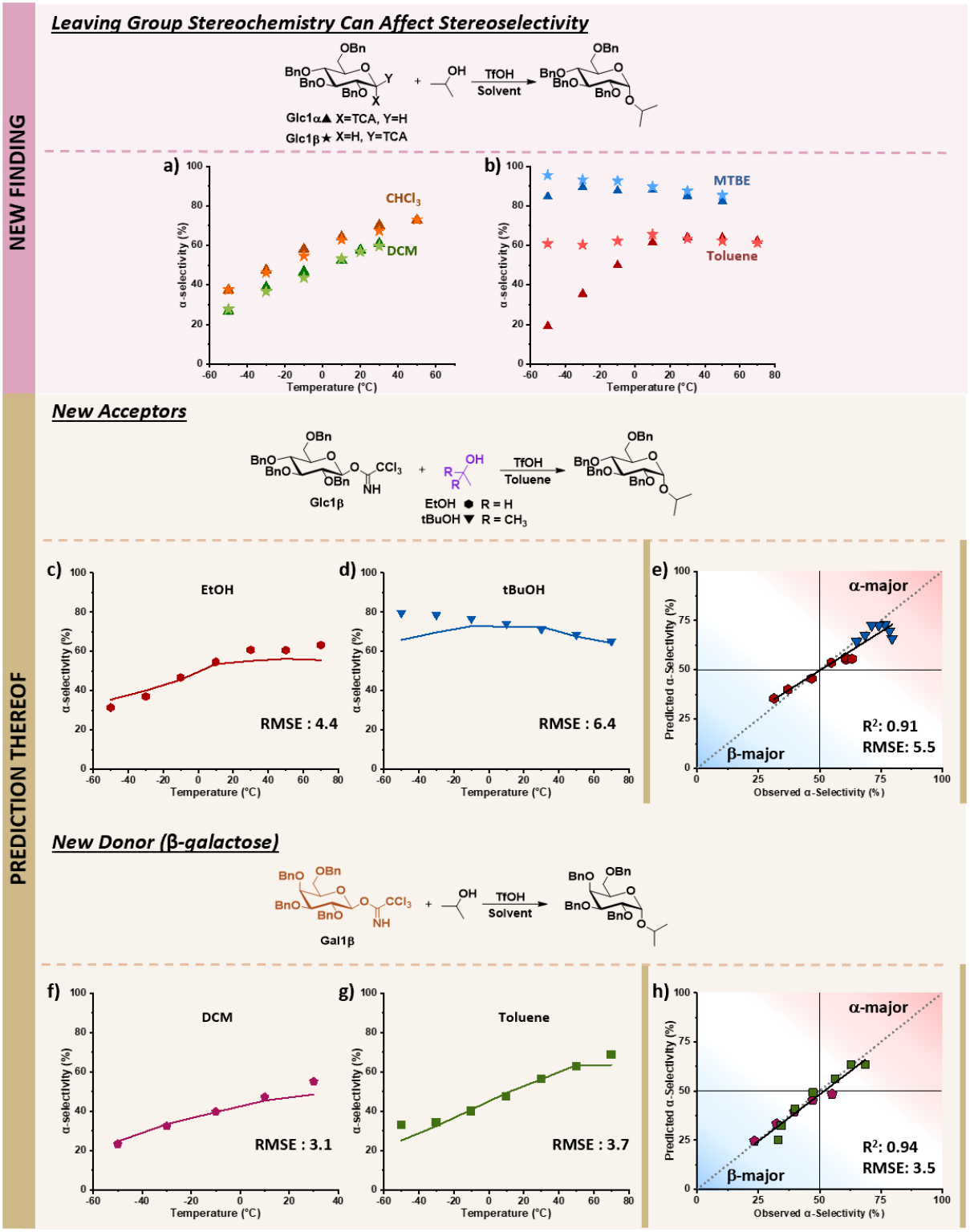

Figure 7. Prediction of novel mechanistic controls of glycosylation reactions, with experimental data shown as points and predicted data shown as lines. a) Experimental results of coupling α/β-glucose donors with iPrOH (**Glc1α** and **Glc1β**) in DCM and $CHCl_3$. b) Experimental results of coupling α/β-glucose donors with iPrOH (**Glc1α** and **Glc1β**) in toluene, and MTBE. c) Prediction and experimental results of β-glucose donor (**Glc1β**) with EtOH in toluene. d) Prediction and experimental results of β-glucose donor (**Glc1β**) with tBuOH in toluene. e) Parity plot of nucleophile predictions with α/β-glucose. f,g) Prediction and experimental results of β-galactose donor (**Gal1β**) with iPrOH in DCM and toluene, respectively. h) Parity plot for solvent predictions of α/β-galactose with iPrOH. Figure code:

Glc1α (▲); Glc1β (★); EtOH (⬢); tBuOH (▼); DCM (⬟); Toluene (■); experimental values (data points) and predicted values (solid colored lines).

Random forest algorithms help to quantify the influence of the variables within the model. Thus, values can be assigned to the identified factors influencing the stereoselectivity of a reaction (Figure 8). In the chemical subspaces covered by our model, 47% of the influence over a glycosylation's stereoselectivity is determined by the inherent properties of the coupling partners. The donor (27%) is more impactful than the acceptor (20%). Upon selection of the coupling partners, more than half of the stereoselectivity observed is controlled by the environmental conditions chosen. The most important environmental factors are the reaction temperature (19%) and the solvent (27%).

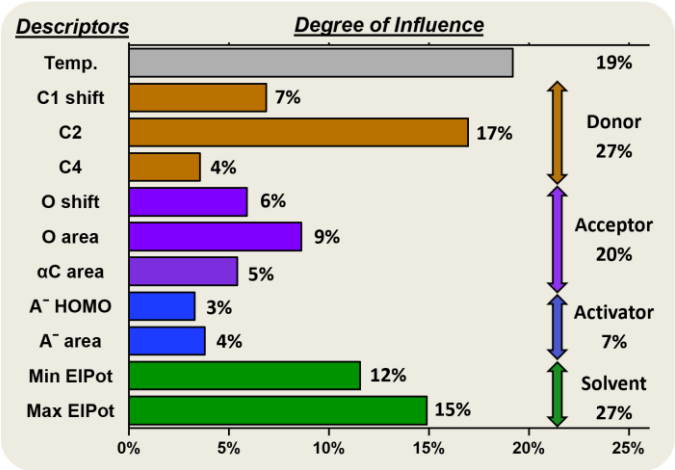

Figure 8. Degree of influence of the eleven factors (defined and described above) influencing the stereoselectivity of glycosylations, rounded to the nearest whole number.

In conclusion, a concise dataset generated on a continuous flow platform was utilized for training a random forest algorithm in an attempt to predict the stereoselectivity of glycosylations as an example for complex, mechanistically fluid transformations. Calculated descriptors were screened and assigned to quantify the individual influencing factors of the coupling partners, active species, and solvent. The predictions of out-of-sample glycosylations – testing nucleophiles, electrophiles, catalyst, solvents, and temperature – were validated experimentally and are highly accurate (overall RMSE: 6.8). Further, the model accurately predicts a previously unknown means of controlling glycosylation stereoselectivity. The approach will be applicable to better understand the stereoselectivity of other transformations based on reactions of nucleophiles and electrophiles.

Acknowledgments
We gratefully acknowledge the generous financial support of the Max-Planck Society and the DFG InCHeM (FOR 2177). We sincerely thank Ms. Tansitha Gupta of GlycoUniverse for providing the fucose precursor, Dr. Christoph Rademacher, Prof. Dr. Andrea Volkamer, and Prof. Bartosz Grzybowski for valuable discussions and Ms. Eva Settels for support.

Supporting Information:

Detailed experimental procedures, complete datasets, additional graphs and control studies, details regarding automation and instrumentation. Microsoft Excel worksheets listing of descriptors, the training set, and the validation set. This information is available free of charge.

Code availability. Software available @ https://github.com/DrSouravChemEng/GlyMecH.

Competing interests: None of the authors declare any competing interests.

References:

[1] Bahmanyar, S.; Houk, K. N.; Martin, H. J.; List, B., Quantum Mechanical Predictions of the Stereoselectivities of Proline-Catalyzed Asymmetric Intermolecular Aldol Reactions. *J. Am. Chem. Soc.* **2003**, *125*, 2475-2479.

[2] Houk, K.; Paddon-Row, M.; Rondan, N.; Wu, Y.; Brown, F.; Spellmeyer, D.; Metz, J.; Li, Y.; Loncharich, R., Theory and modeling of stereoselective organic reactions. *Science* **1986**, *231*, 1108-1117.

[3] Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O., Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016**, *49*, 996-1005.

[4] Hansen, T.; Lebedel, L.; Remmerswaal, W. A.; van der Vorm, S.; Wander, D. P. A.; Somers, M.; Overkleeft, H. S.; Filippov, D. V.; Désiré, J.; Mingot, A.; Bleriot, Y.; van der Marel, G. A.; Thibaudeau, S.; Codée, J. D. C., Defining the SN1 Side of Glycosylation Reactions: Stereoselectivity of Glycopyranosyl Cations. *ACS Cent. Sci.* **2019**, *5*, 781-788.

[5] Peng, Q.; Duarte, F.; Paton, R. S., Computing Organic Stereoselectivity–From Concepts to Quantitative Calculations and Predictions. *Chem. Soc. Rev.* **2016**, *45*, 6093-6107.

[6] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547-555.

[7] de Almeida, A. F.; Moreira, R.; Rodrigues, T., Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, 1-16.

[8] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway. A. M.; Aspuru-Guzik, A., The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241-2251.

[9] Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F., Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237-1245.

[10] Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186-190.

[11] Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F., Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434-443.

[12] Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A., Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725-732.

[13] Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360-365.

[14] Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E., Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631.

[15] Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D., Machine Learning Reactivity in the Chemical Space Surrounding Vaska's Complex. *ChemRxiv* Preprint. https://doi.org/10.26434/chemrxiv.10347566.v1.

[16] Reid, J. P.; Sigman, M. S., Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343-348.

[17] Zheng, F.; Zhang, Q.; Li, J.; Suo, J.; Wu, C.; Zhou, Y.; Liu, X.; Xu, L., Machine Learning Induction of Chemically Intuitive Rules for the Prediction of Enantioselectivity in the Asymmetric Syntheses of Alcohols. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 39-47.

[18] Wendlandt, A. E.; Vangal, P.; Jacobsen, E. N., Quaternary Stereocentres via an Enantioconvergent Catalytic SN1 Reaction. *Nature* **2018**, *556*, 447-451.

[19] Brak, K.; Jacobsen, E. N., E. N. Asymmetric Ion-Pairing Catalysis. *Angew. Chem. Int. Ed.* **2013**, *52*, 534-561.

[20] Bennett, C. S. *Selective Glycosylations: Synthetic Methods and Catalysts*. (John Wiley & Sons, 2017).

[21] Crich, D., Mechanism of a Chemical Glycosylation Reaction. *Acc. Chem. Res.* **2010**, **43**, 1144-1153.

[22] Frihed, T. G.; Bols, M.; Pedersen, C. M., Mechanisms of Glycosylation Reactions Studied by Low-Temperature Nuclear Magnetic Resonance. *Chem. Rev.* **2015**, *115*, 4963-5013.

[23] Phan, T. B.; Nolte, C.; Kobayashi, S.; Ofial, A. R.; Mayr, H., Can One Predict Changes from SN1 to SN2 Mechanisms? *J. Am. Chem. Soc.* **2009**, *131*, 11392-11401.
[24] Chatterjee, S.; Moon, S.; Hentschel, F.; Gilmore, K.; Seeberger, P. H., An Empirical Understanding of the Glycosylation Reaction. *J. Am. Chem. Soc.* **2018**, *140*, 11942-11953.
[25] Park, Y.; Harper, K. C.; Kuhl, N.; Kwan, E. E.; Liu, R. Y.; Jacobsen, E. N., Macrocyclic Bis-Thioureas Catalyze Stereospecific Glycosylation Reactions. *Science* **2017**, *355*, 162-166.
[26] (a) Chang, K. V.; Keiser, M. J., Comment on "Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning" *Science* **2018**, *362*, eaat8603; (b) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G., Response to Comment on "Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning" *Science* **2018**, *362*, eaat8763.
[27] Eggensperger, K.; Feurer, M.; Hutter, F.; Bergstra, J.; Snoek, J.; Hoos, H.; Leyton-Brown, K., Towards an Empirical Foundation for Assessing Bayesian Optimization of Hyperparameters. In *NIPS workshop on Bayesian Optimization in Theory and Practice.* **2013**; p 3.
[28] Steinberg, D. "Chapter 10: Classification and Regression Trees", **2009** Taylor & Francis Group, LLC.
[29] Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V., Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283-293.
[30] Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A., Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936-1949.
[31] Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A., High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324-7331.
[32] Guyon, I.; Elisseeff, A., An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157-1182.
[33] Theodoridis, S.; Koutroumbas, K., *Pattern Recognition*, Academic Press, Elsevier, **2009**.
[34] Harrell Jr, F. E.; Lee, K. L.; Califf, R. M.; Pryor, D. B.; Rosati, R. A., Regression Modelling Strategies for Improved Prognostic Prediction. *Stat. Med.* **1984**, *3*, 143-152.
[35] Lucero, C. G.; Woerpel, K., Stereoselective C-Glycosylation Reactions of Pyranoses: the Conformational Preference and Reactions of the Mannosyl Cation. *J. Org. Chem.* **2006**, *71*, 2641-2647.
[36] Alabugin, I. V.; Manoharan, M., Effect of Double-Hyperconjugation on the Apparent Donor Ability of σ-Bonds: Insights from the Relative Stability of δ-Substituted Cyclohexyl Cations. *J. Org. Chem.* **2004**, *69*, 9011-9024.
[37] Alabugin, I. V.; Gilmore, K. M.; Peterson, P. W., Hyperconjugation. *Wiley Interdiscip. Rev. Comput. Mol.* **2011,** *1*, 109-141.
[38] Gordon, C. P.; Raynaud, C.; Andersen, R. A.; Copéret, C.; Eisenstein, O., Carbon-13 NMR Chemical Shift: A Descriptor for Electronic Structure and Reactivity of Organometallic Compounds. *Acc. Chem. Res.* **2019**, *52*, 2278-2289.
[39] Zhang, Z.; Ollmann, I. R.; Ye, X.-S.; Wischnat, R.; Baasov, T.; Wong, C.-H., Programmable One-Pot Oligosaccharide Synthesis. *J. Am. Chem. Soc.* **1999**, *121*, 734-753.
[40] Edwards, J. O., Correlation of Relative Rates and Equilibria with a Double Basicity Scale. *J. Am. Chem. Soc.* **1954**, *76*, 1540-1547.
[41] Ritchie, C. D., Nucleophilic Reactivities Toward Cations. *Acc. Chem. Res.* **1972**, *5*, 348-354.
[42] Mayr, H.; Patz, M., Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions. *Angew. Chem. Int. Ed.* **1994**, *33*, 938-957.
[43] Van der Vorm, S.; Hansen, T.; Overkleeft, H.; Van der Marel, G.; Codee, J., The Influence of Acceptor Nucleophilicity on the Glycosylation Reaction Mechanism. *Chem. Sci.* **2017**, *8*, 1867-1875.
[44] Hosoya, T.; Kosma, P.; Rosenau, T., Contact Ion Pairs and Solvent-Separated Ion Pairs from d-Mannopyranosyl and d-Glucopyranosyl Triflates. *Carbohydr. Res.* **2015**, *401*, 127-131.
[45] Crich, D.; Sun, S., Are Glycosyl Triflates Intermediates in the Sulfoxide Glycosylation Method? A Chemical and 1H, 13C, and 19F NMR Spectroscopic Investigation. *J. Am. Chem. Soc.* **1997**, *119*, 11217-11223.
[46] (a) Mucha, E.; Marianski, M.; Xu, F.-F.; Thomas, D. A.; Meijer, G.; von Helden, G.; Seeberger, P. H.; Pagel, K., Unravelling the Structure of Glycosyl Cations via Cold-Ion Infrared Spectroscopy. *Nat. Comm.* **2018**, *9*, 1-5. (b) Marianski, M.; Mucha, E.; Greis, K.; Moon, S.; Pardo, A.; Kirschbaum, C.; Thomas, D.; Meijer, G.; von Helden, G.; Gilmore, K.; Seeberger, P.; Pagel, K., Direct Evidence for Remote Participation in Galactose Building Blocks during Glycosylations Revealed by Cryogenic Vibrational Spectroscopy. *Angew. Chem. Int. Ed.* **2020**, *132*, 1-7.
[47] Lubineau, A.; Drouillat, B., Lithium Triflate as a New Promoter of Glycosylation Under Neutral Conditions. *J. Carbohyd. Chem.* **1997**, *16*, 1179-1186.
[48] Kafle, A.; Liu, J.; Cui, L., Controlling the Stereoselectivity of Glycosylation via Solvent Effects. *Can. J. Chem.* **2016**, *94*, 894-901.
[49] Baek, J. Y.; Lee, B.-Y.; Jo, M. G.; Kim, K. S., β-Directing Effect of Electron-Withdrawing Groups at O-3, O-4, and O-6 Positions and α-Directing Effect by Remote Participation of 3-O-Acyl and 6-O-Acetyl Groups of Donors in Mannopyranosylations. *J. Am. Chem. Soc.* **2009**, **131**, 17705-17713.
[50] Kumar, A.; Geng, Y.; Schmidt, R. R., Silicon Fluorides for Acid-Base Catalysis in Glycosidations. *Adv. Synth. Catal.* **2012**, *354*, 1489-1499.