Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2021, **12**, 2931

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 11th November 2020
Accepted 24th December 2020

DOI: 10.1039/d0sc06222g

rsc.li/chemical-science

Predicting glycosylation stereoselectivity using machine learning†

Sooyeon Moon, ‡[ab] Sourav Chatterjee, ‡[a] Peter H. Seeberger [ab] and Kerry Gilmore §*[a]

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 begin to gain a better general understanding of the transformation, for example that environmental factors influence the stereoselectivity of glycosylations more than the coupling partners in this area of chemical space.

Introduction

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,[13–15] 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,[20–22] 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 numerous permanent (defined by the starting materials) or environmental factors (defined by the selected conditions/catalyst) whose degree of influence, interdepency, 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 indicated general trends/influences of these factors (Fig. 1) and hypothesized their relative rankings with respect to dominance.[24] However, a data sciences approach is required to positively identify, quantify, and apply this knowledge for the accurate prediction of stereoselectivities of new coupling partners and conditions. While transfer learning has been applied to machine learning models

[a]*Department of Biomolecular Systems, Max-Planck-Institute of Colloids and Interfaces, Am Mühlenberg 1, 14476 Potsdam, Germany. E-mail: kerry.m.gilmore@uconn.edu*

[b]*Freie Universität Berlin, Institute of Chemistry and Biochemistry, Arnimallee 22, 14195 Berlin, Germany*

† Electronic supplementary information (ESI) available: 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 holdout datasets 1 and 2. Code availability: software available at https://github.com/DrSouravChemEng/GlyMecH. See DOI: 10.1039/d0sc06222g

‡ These authors contributed equally to this work.

§ Current address: Department of Chemistry, University of Connecticut, 55 N. Eagleville Rd, Storrs, CT, USA.

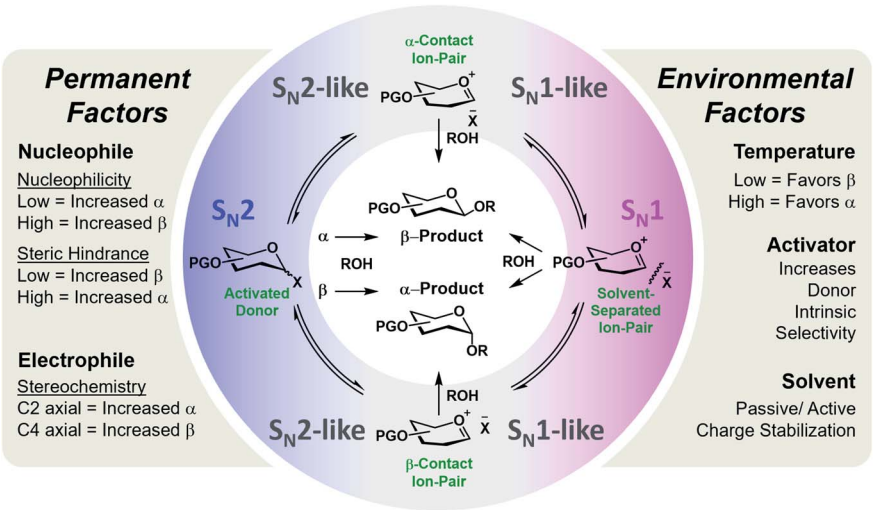

Fig. 1 General representation of the potential mechanistic pathways of glycosylations leading to either the alpha (α) or beta (β) anomer of the formed C–O bond. The empirically-derived permanent and environmental factors and their influence on stereoselectivity are provided.[24]

for the prediction of selectivities of glycosylations (reported between preprint and publication of this work), the stereoselectivity of couplings predicted were controlled by the C-2 acyl protecting group that provide a well-established, highly reproducible means of stereocontrol in these reactions.[26]

Results and discussion

Algorithm training and description of datasets

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 (pS5–S9, Table 1 of ESI†). Regression-based random forest algorithms have proven powerful in modeling chemical reaction performance.[10,27] 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 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 (pS35 of ESI†).[28] 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.[29] An interaction–curvature algorithm was further utilized to reduce the selection bias of the split predictors of the standard CART algorithm (Fig. 2).

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, Table S1 (pS5–S9), ESI†)[30,31] renders manual selection of descriptors – quantifying sterics/electronics – using chemical intuition[32] particularly important.[33]

The training dataset 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 nucleophile equivalents) and adding data for β-glucose electrophile (pS6, lines 68–74 and 101–106 of Table S1, ESI†) and three additional solvents (pS9, lines 238–268 of Table S1, ESI†). Two holdout datasets were experimentally generated (HD1, HD2). The first was comprised of new electrophiles, nucleophiles, acid catalysts, and solvents. Holdout dataset 2 was comprised of examples probing the influence of electrophile leaving group stereochemistry.

Descriptor generation

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 (pS37–S49 of ESI†). The lower level of theory was utilized for optimization of the electrophile 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 by keeping the ratio of datapoints : descriptors >10 : 1.[34,35] The best-performing descriptors for each participant class were determined by the accuracy of the resultant trained models in predicting stereoselectivities of the relevant portions of holdout dataset 1 (*e.g.* determining the accuracy of predicting the novel electrophiles in HD1 with systematic screening of electrophile descriptors). 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 identified descriptors, described below (see potential descriptors excel sheet for a list of all descriptors screened), are either classified as regressors (intra-/extrapolatable values) or

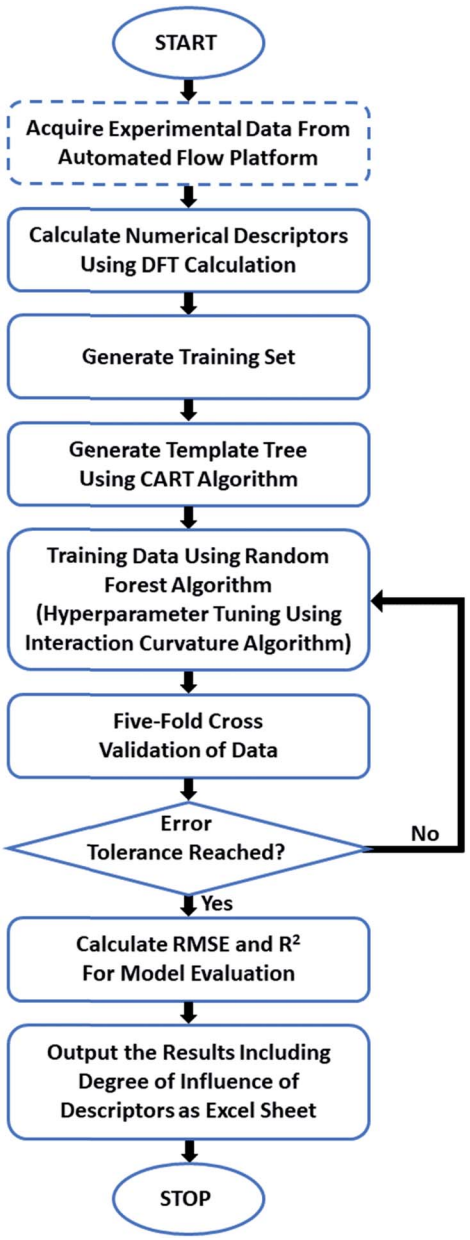

Fig. 2 General workflow of the process from data input to prediction output.

categorical (binary values). While the model can be developed solely using regressor values, it exhibits marginally poorer overall accuracy for holdout dataset 1 and necessitates additional calculations (*vide infra*). 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[36] and hyperconjugative interactions.[37,38] The different leaving groups at the anomeric position were distinguished using the calculated ^{13}C NMR chemical shift,[39] which provided more clear distinctions between leaving groups than the ^{1}H NMR shift[40] 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 Fig. 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 – *e.g.* the influence and nature of the leaving groups – is, at present, acceptable.

Observed nucleophile reactivity has been correlated with a range of parameters.[41–43] Where available, Mayr's nucleophilicity and field inductive parameters correlate with glycosylation stereoselectivity.[44] 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 (Fig. 4). While screening whether simple categorical descriptors can be utilized, specifically the whole values 0–3 to describe the substitution at the α-carbon (as opposed to the exposed surface area), we found that the regressor value proved superior (see ESI†).

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,[45] as evidenced by several studies observing an α-triflate intermediate[20,46] – the product of the conjugate base trapping the oxycarbenium ion.[47] Two values were identified that capture the nuanced role of this species (Fig. 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[48,49] 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 (Fig. 5b). These descriptors perform well, such that even previously unreported means of solvent-control over stereoselectivity are accurately predicted (*vide infra*).

Model training and algorithm comparison

The tuned random forest algorithm was trained using these descriptors on the training dataset[24] containing systematic

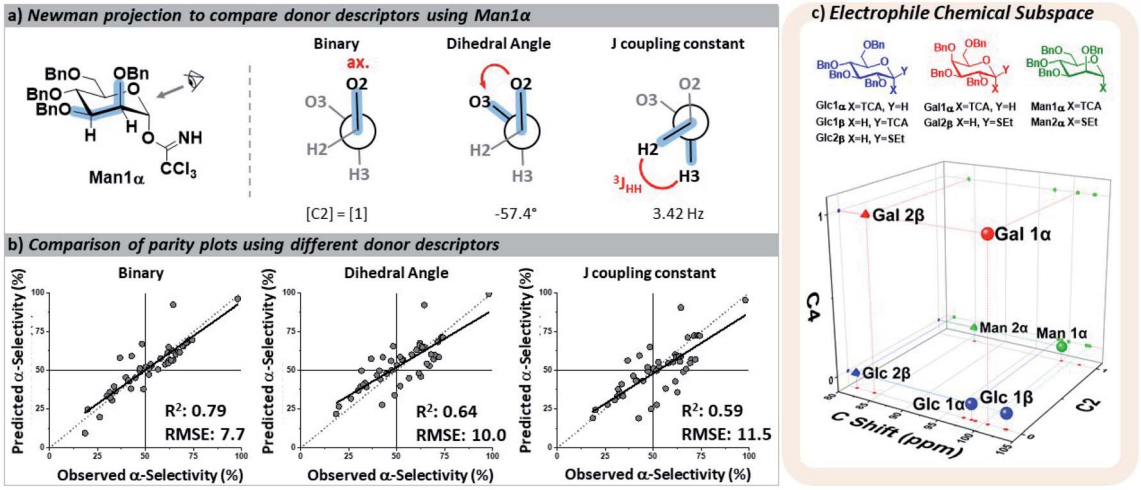

Fig. 3 (a) Three potential means of describing the stereochemistry of the ether groups around the pyran core. (b) Parity plot of the resultant models using each set of descriptors for the electrophile (all also including the calculated ^{13}C NMR shift of C1). Predictions were made of holdout dataset 1. (c) Three-dimensional map of the electrophile 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.

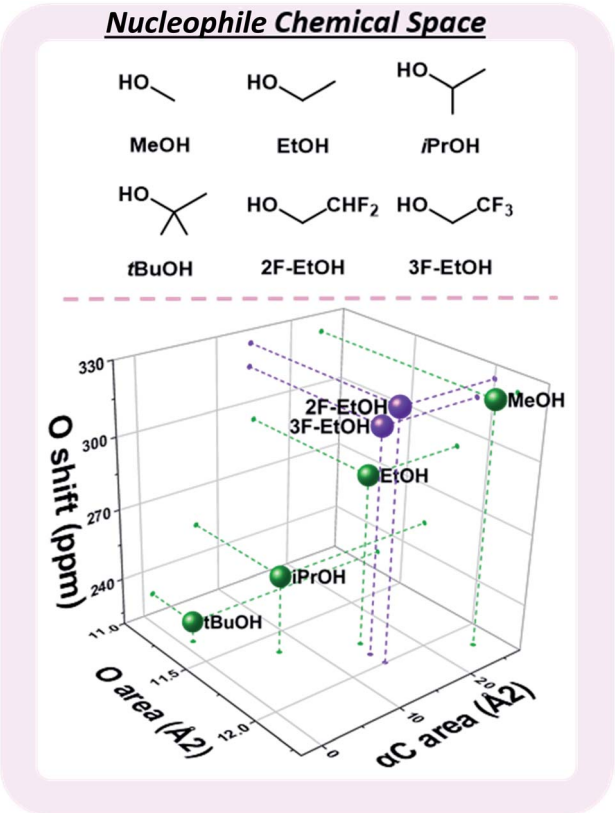

Fig. 4 Three-dimensional map of the nucleophile chemical subspace covered by the developed model, defined by the exposed surface areas of the nucleophilic oxygen and the carbon alpha to the nucleophile, as well as the calculated ^{17}O NMR shift. MeOH – methanol, EtOH – ethanol, iPrOH – isopropanol,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

combinations of seven electrophiles, six nucleophiles, four acid catalysts, and seven solvents over a solvent-dependent temperature range of −50 to +100 °C (pS5–S9, Table S1, ESI†). For comparison, three additional models were trained using Gaussian process regression (GPR), support vector machine (SVM), and regression tree (RT) algorithms. While for some specific predictions different algorithms would have lower RMSEs, random forest (RF) proved superior. The average RMSE of the four models were: RF – 5.9%, RT – 11%, GPR – 7.9%, SVM – 10%. In general terms, RT tended to overestimate the preference for beta-product formation at low temperatures, GPR captured the trend of stereoselectivity change with respect to temperature but lacked precision, and SVM often predicted no influence of temperature yielding a racemic mixture of products (see pages S14–S28 of the ESI†).

The trained RF model was then used to predict the stereoselectivities of the entirety of holdout dataset 1, containing unseen variants of each of the four chemical species in the reaction over the accessible temperature ranges (defined by the solvent and reactor). Holdout dataset 1 (see holdout dataset 1 excel sheet of ESI†) was generated using the same reproducible microreactor platform[24] as the training dataset. The results of these predictions, as compared to the experimentally observed selectivities, are presented as the percentage of alpha product formed *versus* temperature. The corresponding parity plots for each are also provided (Fig. 6).

Validation of descriptors and prediction accuracy of holdout dataset 1

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 electrophiles, with a combined root mean square error (RMSE) of 2.0 (Fig. 6a). The model can be applied to other pyran cores,

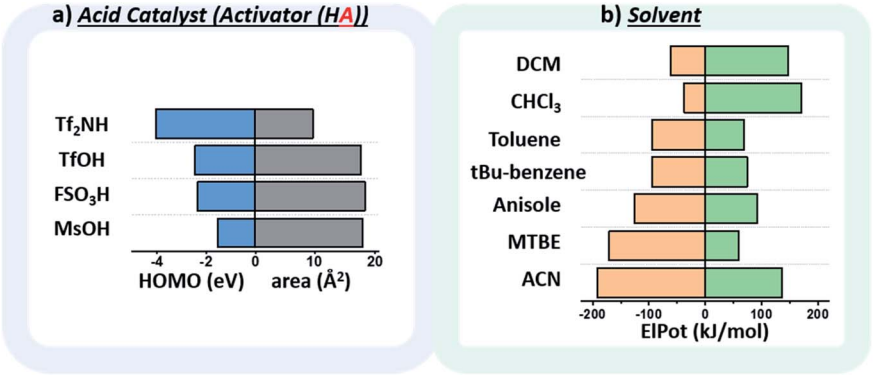

Fig. 5 (a) Plot of the descriptors used to quantify the relevant factors of the conjugate base of the activator. Area ($Å^2$) corresponds to the exposed surface area of the oxygen (O^-) or nitrogen anion (N^-) in a space-filling model. HOMO: highest occupied molecular orbital (eV). (b) Plot of the descriptors used to quantify the relevant factors of the solvent, the maximum (MaxElPot), and minimum (MinElPot) values of the electrostatic potential (kJ mol^{-1}). Tf_2NH – bis(trifluoromethane)sulfonamide, TfOH – trifluoromethanesulfonic acid, FSO_3H – fluorosulfonic acid, MsOH – methanesulfonic acid, DCM – dichloromethane, $CHCl_3$ – chloroform, *t*Bu-benzene – *tert*-butylbenzene, MTBE – methyl *tert*-butyl ether, ACN – acetonitrile.

such as L-fucose.[50] The predicted stereoselectivity of the fucose α-glycosyl imidate electrophile 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 (Fig. 6b).

While the training dataset 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, Fig. 6d and 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, Fig. 6g). The weakest correlation of our model is observed for the $C_3F_6S_2O_4NH$-activated mannose coupling with *tert*-butanol 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 electrophile configurations.[21] With $C_3F_6S_2O_4NH$, the perbenzylated electrophile ranges from a 63% β-selectivity at −50 °C to 98% α-selectivity at 30 °C (Fig. 6h).

Finally, the stereoselectivities of glucose and galactose α-imidate electrophiles with isopropanol were predicted for two new solvents (Fig. 6j and k). The strong influence of solvent[48,51] 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).

Model validation of unreported influences on stereoselectivity (holdout dataset 2)

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.[52,53] No influence of the α/β-orientation of the leaving group in dichloromethane was reported (Fig. 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.[54]

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-electrophile reaches 96% α-selectivity. This variable becomes important in toluene. Glucose β-imidate electrophile yields almost unchanged stereoselectivity (∼60% alpha) over a 120 °C range! The orientation of the leaving group of the electrophile influences the stereoselectivity by more than 40% at −50 °C (Fig. 7b).

With this limited data in our training dataset (Fig. 7a and b), we tested the ability of our model to predict the influence of other factors on this to-date unreported phenomenon in holdout dataset 2, whose experimental values were obtained on the same microreactor platform as TD and HD1 (see holdout dataset 2 excel sheet of ESI†). The stereoselectivity of glucose α-imidate with ethanol as nucleophile ranges from 10–54% α-product in toluene. The model predicts that the β-electrophile will behave differently, with a much less selective coupling overall (37–56% α-product) and a 27% difference in selectivity at low temperature compared to the α-electrophile. This prediction matches well with the experimental results, with an RMSE

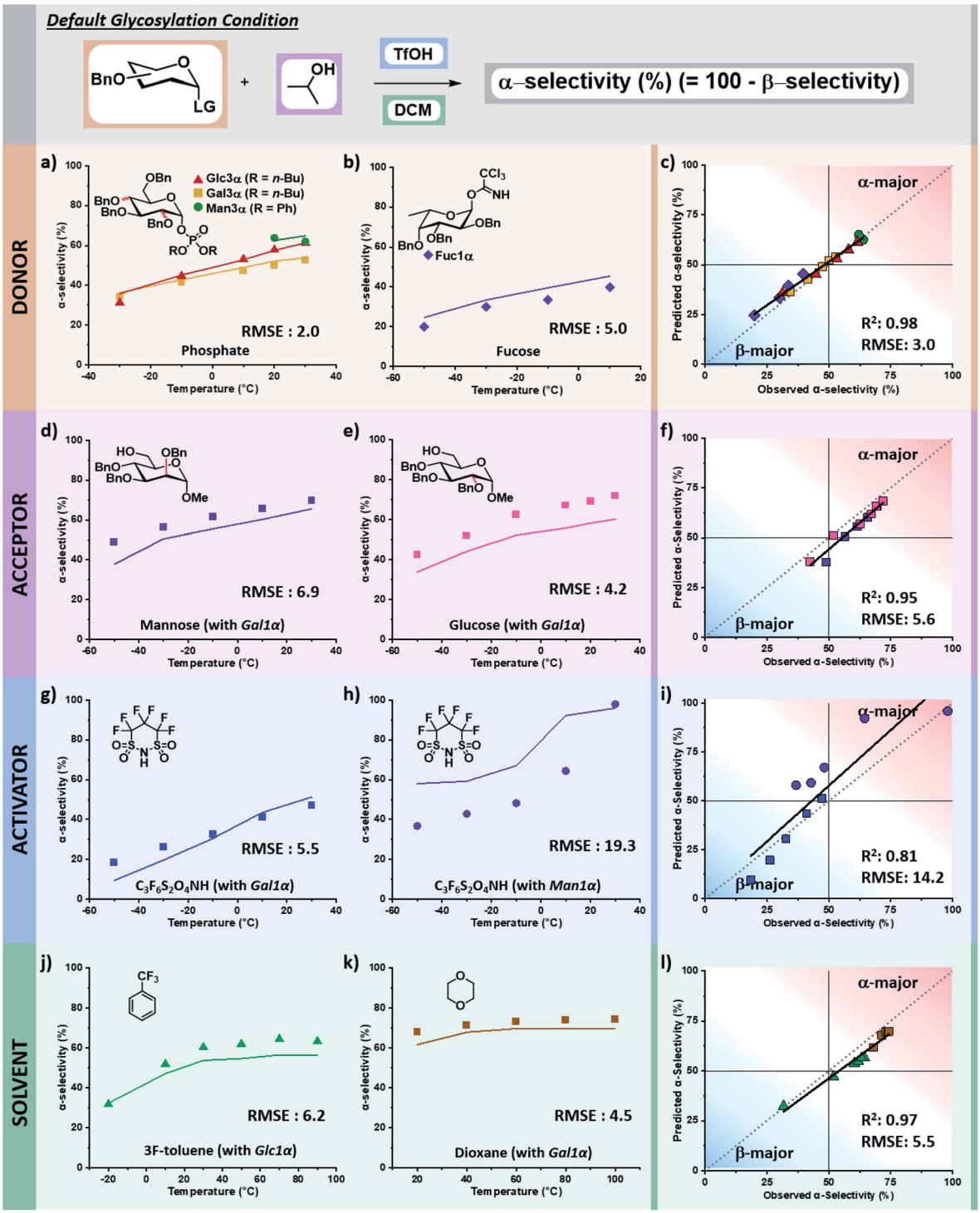

Fig. 6 Prediction of stereoselectivity for glycosylations using different anomeric leaving groups, electrophiles, nucleophiles, activators, and solvents. Descriptors used were: electrophile (C1: ^{13}C NMR shift, C2: stereochemistry axial = 1 or equatorial = 0, C4: stereochemistry axial = 1 or equatorial = 0), nucleophile (O: ^{17}O NMR shift, O: exposed surface area, αC: exposed surface area), acid catalyst (A^-: HOMO energy, A^-: exposed surface area), solvent (minimum electrostatic potential, maximum electrostatic potential), and temperature (−50 to 100 °C). (a) Prediction of stereoselectivity for glycosylations involving a glycosyl phosphate leaving group. Bu – butyl, Ph – phenyl, RMSE – root mean square error. TMSOTf was used as acid catalyst, which has the same descriptors as TfOH. (b) Prediction of stereoselectivity using a fucose (Fuc) electrophile with iPrOH in DCM. (c) Parity plot of electrophile (electrophile) predictions. (d and e) Prediction of mannose and glucose nucleophile, respectively, with galactose α-imidate electrophile in DCM. (f) Parity plot of nucleophile (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 electrophile and *t*BuOH nucleophile in DCM. (h) Prediction of $C_3F_6S_2O_4NH$ with mannose electrophile 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 electrophile and iPrOH. (k) Prediction of 1,4-dioxane solvent with galactose α-imidate electrophile and iPrOH. (l) Parity plot of solvent predictions. Figure code: fucose (◆); glucose (▲); galactose (■); mannose (●); experimental (data points); predicted (solid colored line).

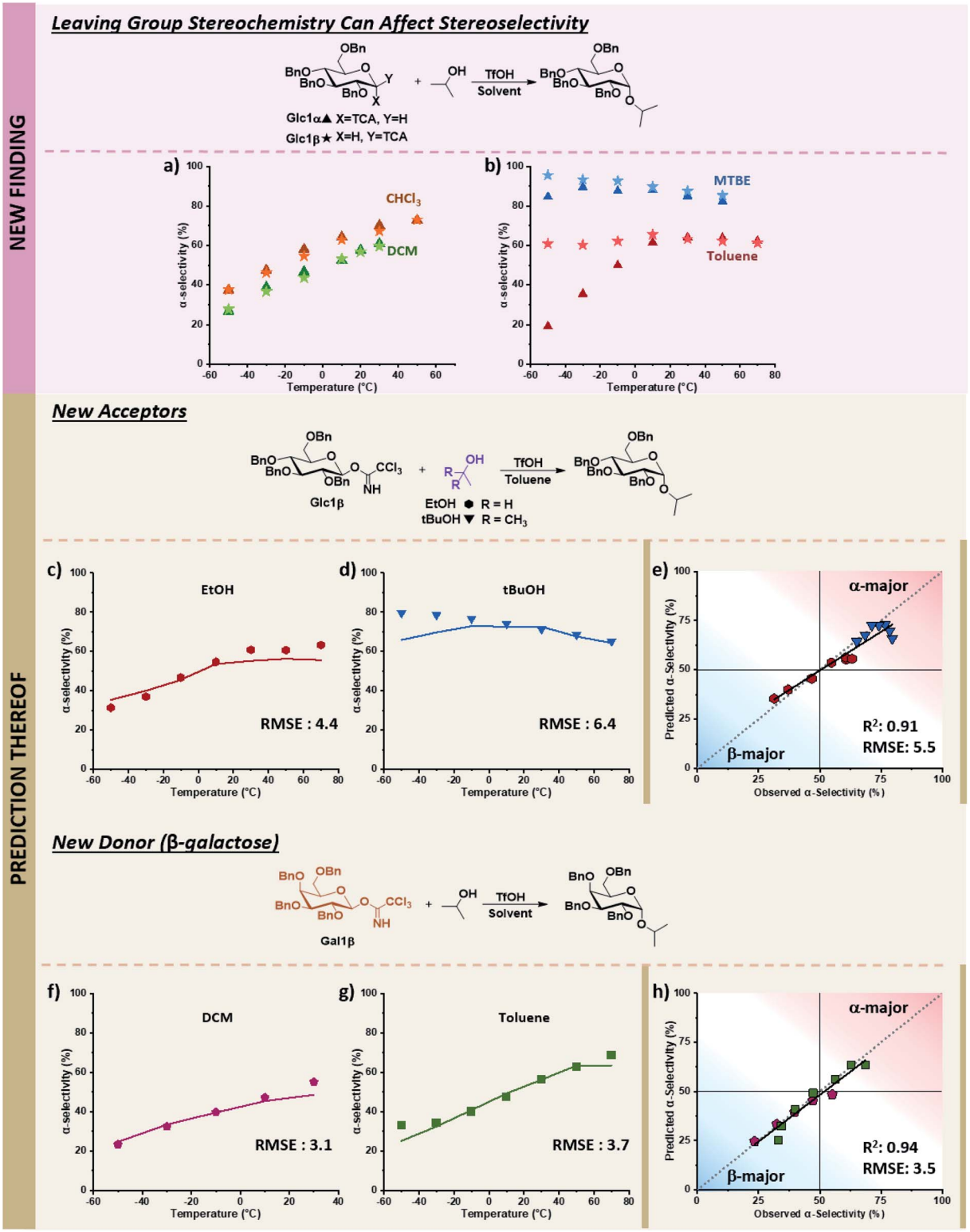

Fig. 7 Prediction of novel mechanistic controls of glycosylation reactions using holdout dataset 2, with experimental data shown as points and predicted data shown as lines. The relevant experimental data for the α-electrophiles can be found in Table S1 of the ESI.† (a) Experimental results of coupling α/β-glucose electrophiles with iPrOH (**Glc1α** and **Glc1β**) in DCM and $CHCl_3$. (b) Experimental results of coupling α/β-glucose electrophiles with iPrOH (**Glc1α** and **Glc1β**) in toluene, and MTBE. (c) Prediction and experimental results of β-glucose electrophile (**Glc1β**) with EtOH in toluene. (d) Prediction and experimental results of β-glucose electrophile (**Glc1β**) with tBuOH in toluene. (e) Parity plot of EtOH and *t*-BuOH nucleophile predictions with the β-glucose electrophile. (f and g) Prediction and experimental results of β-galactose electrophile (**Gal1β**) with iPrOH in DCM and toluene, respectively. (h) Parity plot for DCM and toluene solvent predictions of the β-galactose electrophile with iPrOH. Figure code: Glc1α (▲); Glc1β (★); EtOH (●); *t*BuOH (▼); DCM (♠); toluene (■); experimental values (data points) and predicted values (solid colored lines).

of 4.4 over the 120 °C range (Fig. 7c). The model predicts a less α-selective reaction at low temperatures than observed with *t*-BuOH as nucleophile (similar to what is observed using the α-electrophile, pS6, lines 82–88 of Table S1, ESI†), though at higher temperatures, the prediction matches well with experimental values (RMSE: 6.4, Fig. 7d).

Lastly, we sought to explore whether this additional mechanistic complexity exists for other electrophiles (Fig. 7f and g). In DCM, the coupling of α-galactose with isopropanol moderately favors the formation of the β-product (19–51% α-product from −50 to 30 °C, (pS7, lines 119–124, of ESI†)). The model predicts that the β-galactose electrophiles will give similar α-selectivity in DCM over the 80 °C temperature range (24–49% α-product), matching experimental results (RMSE 3.1, Fig. 7f). In toluene, the α-galactose electrophile exhibits a wide range of selectivities with isopropanol, from 10–69% α-product across the 130 °C range (pS7, lines 142–148, of ESI†). The model predicts a slight divergence (15%) in stereoselectivity at low temperatures when the β-galactose electrophile is used (25–64% α-product, −50 to 70 °C), though not as large as what is observed with β-glucose. This prediction again aligns with experimental results (RMSE: 3.7, Fig. 7g). Overall, the model correctly predicts the previously unknown ability to turn on and off the influence of the electrophile leaving group's orientation using solvents under otherwise identical conditions. We hypothesize the decrease of stereoselectivities for β-electrophiles when using toluene may result from an increase in the S_N1-type pathways. The π-system of the solvent can more easily induce solvolysis of the more planar equatorial leaving group from both faces (as compared to the axial orientation), leading to an accessible oxonium ion instead of an α-triflate intermediate. Additional detailed mechanistic studies are required to discern the degree and nature of mechanistic control.

Overall influences of permanent and environmental factors on stereoselectivity

Random forest algorithms can quantify the influence of the variables within the model. Thus, values can be assigned to the identified factors influencing the stereoselectivity of a reaction (Fig. 8), allowing for some cautious generalizations to be made. 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 electrophile (27%) is more impactful than the nucleophile (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 solvent (27%) and the reaction temperature (19%).

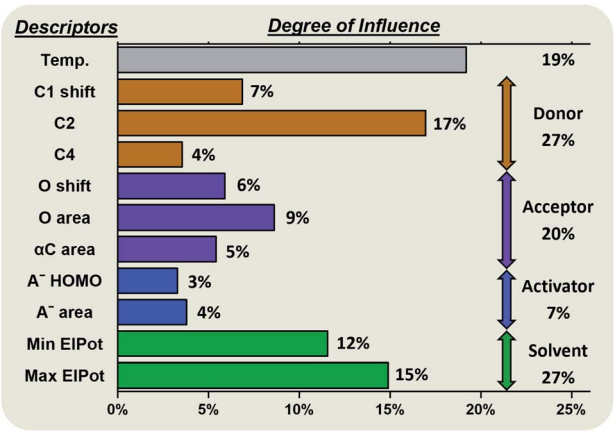

Fig. 8 Degree of influence of the eleven factors (defined and described above) influencing the stereoselectivity of glycosylations, rounded to the nearest whole number.

Conclusion

A concise training dataset generated on a continuous flow platform was utilized to train a random forest algorithm 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 glycosylation stereoselectivities were made of two holdout datasets – testing nucleophiles, electrophiles, catalyst, solvents, and temperature – containing data obtained experimentally on a microreactor platform. The model is highly accurate (overall RMSE: 6.8) in the chemical subspaces explored. 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.

Conflicts of interest

None of the authors declare any competing interests.

Acknowledgements

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.

References

1 S. Bahmanyar, K. N. Houk, H. J. Martin and B. List, *J. Am. Chem. Soc.*, 2003, **125**, 2475–2479.
2 K. Houk, M. Paddon-Row, N. Rondan, Y. Wu, F. Brown, D. Spellmeyer, J. Metz, Y. Li and R. Loncharich, *Science*, 1986, **231**, 1108–1117.
3 E. Hansen, A. R. Rosales, B. Tutkowski, P.-O. Norrby and O. Wiest, *Acc. Chem. Res.*, 2016, **49**, 996–1005.
4 T. Hansen, L. Lebedel, W. A. Remmerswaal, S. van der Vorm, D. P. A. Wander, M. Somers, *et al.*, *ACS Cent. Sci.*, 2019, **5**, 781–788.
5 Q. Peng, F. Duarte and R. S. Paton, *Chem. Soc. Rev.*, 2016, **45**, 6093–6107.

6 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
7 A. F. de Almeida, R. Moreira and T. Rodrigues, *Nat. Rev. Chem.*, 2019, 1–16.
8 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, *et al.*, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
9 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.
10 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
11 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
12 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
13 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
14 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
15 P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584–4601.
16 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.
17 F. Zheng, Q. Zhang, J. Li, J. Suo, C. Wu, Y. Zhou, *et al.*, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 39–47.
18 A. E. Wendlandt, P. Vangal and E. N. Jacobsen, *Nature*, 2018, **556**, 447–451.
19 K. Brak and E. N. Jacobsen, *Angew. Chem., Int. Ed.*, 2013, **52**, 534–561.
20 C. S. Bennett, *Selective Glycosylations: Synthetic Methods and Catalysts*, John Wiley & Sons, 2017.
21 D. Crich, *Acc. Chem. Res.*, 2010, **43**, 1144–1153.
22 T. G. Frihed, M. Bols and C. M. Pedersen, *Chem. Rev.*, 2015, **115**, 4963–5013.
23 T. B. Phan, C. Nolte, S. Kobayashi, A. R. Ofial and H. Mayr, *J. Am. Chem. Soc.*, 2009, **131**, 11392–11401.
24 S. Chatterjee, S. Moon, F. Hentschel, K. Gilmore and P. H. Seeberger, *J. Am. Chem. Soc.*, 2018, **140**, 11942–11953.
25 Y. Park, K. C. Harper, N. Kuhl, E. E. Kwan, R. Y. Liu and E. N. Jacobsen, *Science*, 2017, **355**, 162–166.
26 G. Pesciullesi, P. Schwaller, T. Laino and J.-L. Reymond, *Nat. Commun.*, 2020, **11**, 4874–4882.
27 (*a*) K. V. Chang and M. J. Keiser, *Science*, 2018, **362**, eaat8603; (*b*) J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher and A. G. Doyle, *Science*, 2018, **362**, eaat8763.
28 K. Eggensperger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. Hoos and K. Leyton-Brown, in *NIPS workshop on Bayesian Optimization in Theory and Practice*, 2013, p. 3.
29 D. Steinberg, *Classification and Regression Trees*, Taylor & Francis Group, LLC, 2009, ch. 10.
30 H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 283–293.
31 G. Subramanian, B. Ramsundar, V. Pande and R. A. Denny, *J. Chem. Inf. Model.*, 2016, **56**, 1936–1949.
32 A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig and A. Mar, *Chem. Mater.*, 2016, **28**, 7324–7331.
33 I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
34 S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, Elsevier, 2009.
35 F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor and R. A. Rosati, *Stat. Med.*, 1984, **3**, 143–152.
36 C. G. Lucero and K. Woerpel, *J. Org. Chem.*, 2006, **71**, 2641–2647.
37 I. V. Alabugin and M. Manoharan, *J. Org. Chem.*, 2004, **69**, 9011–9024.
38 I. V. Alabugin, K. M. Gilmore and P. W. Peterson, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 109–141.
39 C. P. Gordon, C. Raynaud, R. A. Andersen, C. Copéret and O. Eisenstein, *Acc. Chem. Res.*, 2019, **52**, 2278–2289.
40 Z. Zhang, I. R. Ollmann, X.-S. Ye, R. Wischnat, T. Baasov and C.-H. Wong, *J. Am. Chem. Soc.*, 1999, **121**, 734–753.
41 J. O. Edwards, *J. Am. Chem. Soc.*, 1954, **76**, 1540–1547.
42 C. D. Ritchie, *Acc. Chem. Res.*, 1972, **5**, 348–354.
43 H. Mayr and M. Patz, *Angew. Chem., Int. Ed.*, 1994, **33**, 938–957.
44 S. Van der Vorm, T. Hansen, H. Overkleeft, G. Van der Marel and J. Codee, *Chem. Sci.*, 2017, **8**, 1867–1875.
45 T. Hosoya, P. Kosma and T. Rosenau, *Carbohydr. Res.*, 2015, **401**, 127–131.
46 D. Crich and S. Sun, *J. Am. Chem. Soc.*, 1997, **119**, 11217–11223.
47 (*a*) E. Mucha, M. Marianski, F.-F. Xu, D. A. Thomas, G. Meijer, G. von Helden, *et al.*, *Nat. Commun.*, 2018, **9**, 1–5; (*b*) M. Marianski, E. Mucha, K. Greis, S. Moon, A. Pardo, C. Kirschbaum, *et al.*, *Angew. Chem., Int. Ed.*, 2020, **132**, 1–7.
48 A. Kafle, J. Liu and L. Cui, *Can. J. Chem.*, 2016, **94**, 894–901.
49 H. Satoh, H. S. Hansen, S. Manabe, W. F. van Gunteren and P. H. Hünenberger, *J. Chem. Theory Comput.*, 2010, **6**, 1783–1797.
50 A. Lubineau and B. Drouillat, *J. Carbohydr. Chem.*, 1997, **16**, 1179–1186.
51 J. C. Kendale, E. M. Valentin and K. A. Woerpel, *Org. Lett.*, 2014, **16**, 3684–3687.
52 J. Y. Baek, B.-Y. Lee, M. G. Jo and K. S. Kim, *J. Am. Chem. Soc.*, 2009, **131**, 17705–17713.
53 K. Greis, E. Mucha, M. Lettow, D. A. Thomas, C. Kirschbaum, S. Moon, *et al.*, *ChemPhysChem*, 2020, **21**, 1905–1907.
54 A. Kumar, Y. Geng and R. R. Schmidt, *Adv. Synth. Catal.*, 2012, **354**, 1489–1499.