Computers and Chemical Engineering 194 (2025) 108937

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Linear and neural network models for predicting N-glycosylation in Chinese Hamster Ovary cells based on B4GALT levels

Pedro Seber, Richard D. Braatz *

Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States of America

ARTICLE INFO

Dataset link: https://github.com/PedroSeber/CHO_N-glycosylation_prediction

Keywords:
N-glycosylation
Glycosylation
Biopharmaceuticals
Machine learning
Deep learning
Chinese hamster ovary

ABSTRACT

Glycosylation is an essential modification to proteins that has positive effects, such as improving the half-life of antibodies, and negative effects, such as promoting cancers. Despite the importance of glycosylation, data-driven models to predict quantitative N-glycan distributions have been lacking. This article constructs linear and neural network models to predict the distribution of glycans on N-glycosylation sites. The models are trained on data containing normalized B4GALT1–B4GALT4 levels in Chinese Hamster Ovary cells. The ANN models achieve a median prediction error of 1.59% on an independent test set, an error 9-fold smaller than for previously published models using the same data, and a narrow error distribution. We also discuss issues with other models in the literature and the advantages of this work's model over other data-driven models. We openly provide all of the software used, allowing other researchers to reproduce the work and reuse or improve the code in future endeavors.

1. Introduction

Glycosylation is a form of co-translational and post-translational modification that involves adding a glycan or glycans to proteins. When a glycan is added to the nitrogen of an asparagine or arginine, this process is called N-linked glycosylation. Glycosylation has many important functional and structural roles (Imperiali and O'Connor, 1999; Patterson, 2005; Schjoldager et al., 2020). Improper glycosylation or deglycosylation, on the other hand, is associated with multiple diseases such as cancers (Stowell et al., 2015), infections (Bhat et al., 2019), and congenital disorders (Jaeken, 2013). Many enzymes participate in the glycosylation process, and B4GALT1–B4GALT4 have been shown to be key contributors in multiple independent studies, such as Bydlinski et al. (2018).

Due to its potential for diagnoses and treatments, glycosylation has been of significant interest to the biomedical and pharmaceutical industry, physicians, and patients. For example, increases in fucosylation, branching, and sialylation occur in many types of carcinoma (Almeida and Kolarich, 2016). Disialoganglioside is expressed by almost all neuroblastomas, and anti-disialoganglioside monoclonal antibodies have been successful against high-risk neuroblastoma in Phase I–III studies (Ho et al., 2016; Ahmed and Cheung, 2014). Another example is poly-α2,8-sialylation, which increases the half-lives of antibodies without introducing tolerance problems (Van Landuyt et al., 2019). On the other hand, the presence of glycans not produced by humans can be detrimental to a therapeutic. N-glycolylneuraminic acid, which is in some CHO-cell-derived glycoproteins (Hokke et al., 1995), is immunogenic to humans (Padler-Karavani et al., 2008).

Despite the importance of glycosylation for biotherapeutics and the many advances made in the field, such as the genetic engineering of CHO cells to increase the sialylation of glycoproteins (Bork et al., 2007), some challenges remain. Proteins have multiple glycosylation sites, and analyses need to take into account not only the glycan compositions but also where these glycans are located (Almeida and Kolarich, 2016). This structural diversity makes it difficult to explore specific functions of glycosylation (Schjoldager et al., 2020). Clinical laboratories struggle with analyzing patient glycosylation samples due to the complex equipment needed, which has limited the progress of personalized medicine (Almeida and Kolarich, 2016).

Many computational models have been developed to assist researchers in better understanding and predicting glycosylation patterns. Mechanistic models use physical knowledge, typically in the form of differential equations, to make predictions, whereas data-driven models directly leverage experimental data to make predictions. Each kind of model has advantages and disadvantages, and this work focuses on data-driven models. Some of these models, such as DeepNGlyPred (Pakhrin et al., 2021) or SPRINT-Gly (Taherzadeh et al., 2019), are classifiers. In the context of glycosylation, classifier models may predict whether an amino acid is N-glycosylated, or whether it is O-GlcNAcylated, for example. Other models, such as by Moon et al.

* Corresponding author.
E-mail address: braatz@mit.edu (R.D. Braatz).

https://doi.org/10.1016/j.compchemeng.2024.108937
Received 1 August 2024; Received in revised form 17 November 2024; Accepted 17 November 2024
Available online 23 November 2024

(2021) and Liang et al. (2020), perform regression, that is, the models attempt to predict numerical values. In the context of glycosylation, a regression model may predict the numerical glycan distribution at a given glycosylation site. The numerical glycan distribution is, for a known glycosylation site X, what percentage of proteins have glycan A attached to that site, what percentage of proteins have glycan B attached to that site, and so forth for every glycan to obtain the complete glycan distribution for a site, then so forth for every site to obtain the complete glycan distribution for all sites of a protein. In particular, Liang et al. (2020) attempted to answer, using a different modeling method, a similar problem to this work. While their model and the models in this work cannot be directly compared because the datasets and input types are different, advantages of this work's models are analyzed in the Discussion (Section 4).

In this work, we construct linear models and artificial neural network (ANN) regression models to predict glycan distributions in different glycosylation sites of fusion Fc-proteins based on changes in B4GALT1–4 enzyme levels in Chinese Hamster Ovary cells. The model construction procedures employ nested cross-validation and rigorous unbiased prediction error estimation. This work's ANN models have 9-fold lower median prediction error than previously reported models. Moreover, Shapley values are used to interpret the models' predictions (Shapley, 1951). Open-source software is provided so that other researchers can reproduce the work or retrain the models as additional glycosylation data become available.

2. Materials and methods

This section describes the datasets and methods for constructing the data-driven models. Details on how to run the software to make glycan distribution predictions or recreate the results in this article are provided in Supplementary Information.

2.1. Datasets

The models in this article are constructed using an experimental dataset that comprises distributions of glycans in nine different glycosylation sites of fusion Fc-proteins in response to changes in the enzyme levels of B4GALT1–B4GALT4 due to five types of knockouts (Supplemental Data of Bydlinski et al. (2018)). Data are preprocessed using our own code, which includes standard Python packages (Harris et al., 2020; Hunter, 2007; McKinney, 2010). The levels of B4GALT1–B4GALT4 are provided to the model as an $N\times4$ numerical matrix. This matrix is scaled by subtracting the mean of the training data and then dividing by the standard deviation of the training data. Thus, the scaled training data (but not the scaled validation and test data) have mean = 0 and standard deviation = 1. 20% of the data for each glycosylation site are separated for testing, with the remaining 80% used for cross-validation. The data are split into five groups according to the knockouts that were performed.[1] During each cross-validation fold and testing, the validation/testing groups are selected such that an entire knockout group was held out, and the other knockout groups are used for training. This procedure avoids test set leakage by ensuring the test set is sufficiently different from the training set.

Note that the glycans in Bydlinski et al. (2018)'s dataset use non-standard names. This work converted these names into the Oxford Notation. For more details, please view Section S4.

2.2. Linear data-driven models

Linear data-driven models (specifically, elastic net (EN), ridge regression (RR), and partial least squares (PLS), which together are called EN/RR/PLS) are constructed using a modified implementation of the Smart Process Analytics software (Sun and Braatz, 2021)[2] to serve as a baseline for comparison. A list of hyperparameters tested for each model is available in Section S3. For each position and glycan, the best hyperparameters for the linear models are selected through cross-validation on four of these groups, leading to one model per combination of position and glycan. The model and hyperparameters with the lowest cross-validation average loss (as determined by the mean squared error loss function) are selected and the last group used to test that model's performance.

2.3. Artificial Neural Networks (ANNs)

Artificial neural network models (specifically, multilayer perceptrons) are constructed for both datasets using PyTorch (Paszke et al., 2019). We constructed models for 7 different layer configurations, 6 different learning rates (5e−2, 1e−2, 5e−3, 1e−3, 5e−4, 1e−4), 4 different activation functions (ReLU, SeLU, tanh, tanhshrink), and plateau or cosine scheduling (Loshchilov and Hutter, 2017), as further detailed in Section S3. The best hyperparameters for each glycan are determined by a grid search, leading to one model per combination of position and glycan. The combination with the lowest cross-validation average loss (as determined by the mean squared error loss function) is selected and, for each position and glycan combination, its performance is reported for an independent test dataset as the "This work's ANN" model. Lastly, the predictions of selected models are made interpretable using Shapley values (Shapley, 1951) as implemented by Lundberg and Lee (2017).

2.4. Shapley values

Originally created for game theory, Shapley values are a unique set of values that represent the value (in economic terms, the surplus) generated by a coalition of players (Shapley, 1951). Given a game with N players, a coalition S of players, and a value function for a coalition $v(S)$, the Shapley value for a player i, $\varphi_i(v)$ is defined as (Shapley, 1951):

$$\varphi_i(v) = \frac{1}{N} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to the coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

$$\varphi_i(v) = \sum_{\text{coalitions including } i} \frac{\text{synergy of the coalition}}{\text{members in the coalition}}$$

Definitions for S and v can vary significantly depending on the context and application. In a machine learning context, i represents each feature, S is a subgroup of all the features present in the data and v is the loss function used (Lundberg and Lee, 2017). Models are trained with sets of missing features and compared to the model trained with all features, allowing the calculation of a Shapley value φ_i for each feature i. To avoid need to train multiple versions of the model, Shapley values may also be estimated well via sampling (Lundberg and Lee, 2017). Finally, once the Shapley values have been generated, a prediction $\hat{y}$ can be made with a simple linear model, vastly simplifying the interpretation of a nonlinear model's output.

$$\hat{y} = \varphi_0 + \sum_{i=1}^{N} \varphi_i x_i$$

[1] Group labels are available in the `group_names.txt` file on the GitHub repository.

[2] The software version used in this article is available at github.com/PedroSeber/SmartProcessAnalytics.

Table 1
Mean and median percent relative errors (PRE) and RMSEs for different N-glycosylation models. Models "ANN-2" and "ANN-3" are from Kotidis and Kontoravdi (2020); their PRE data are obtained from Figs. 5 and 6 of Kotidis and Kontoravdi (2020). Models "EN/RR/PLS" and "ANNs" come from this work. "Train mean" is the mean of the training data (Bydlinski et al. (2018)).

Statistic	ANN-2 (From Kotidis and Kontoravdi (2020))	ANN-3 (From Kotidis and Kontoravdi (2020))	EN/RR/PLS (Baseline)	ANNs (This work)	Train mean (From Bydlinski et al. (2018))
Mean PRE ($\pm\sigma$)	17.11 ± 19.45	15.51 ± 10.40	26.72 ± 20.49	2.45 ± 2.86	38.94 ± 17.68
Median PRE	15.67	13.83	26.17	1.59	43.43
RMSE (%)	4.08	4.81	8.90	0.72	13.91
Times Best Model	1	4	1	25	0

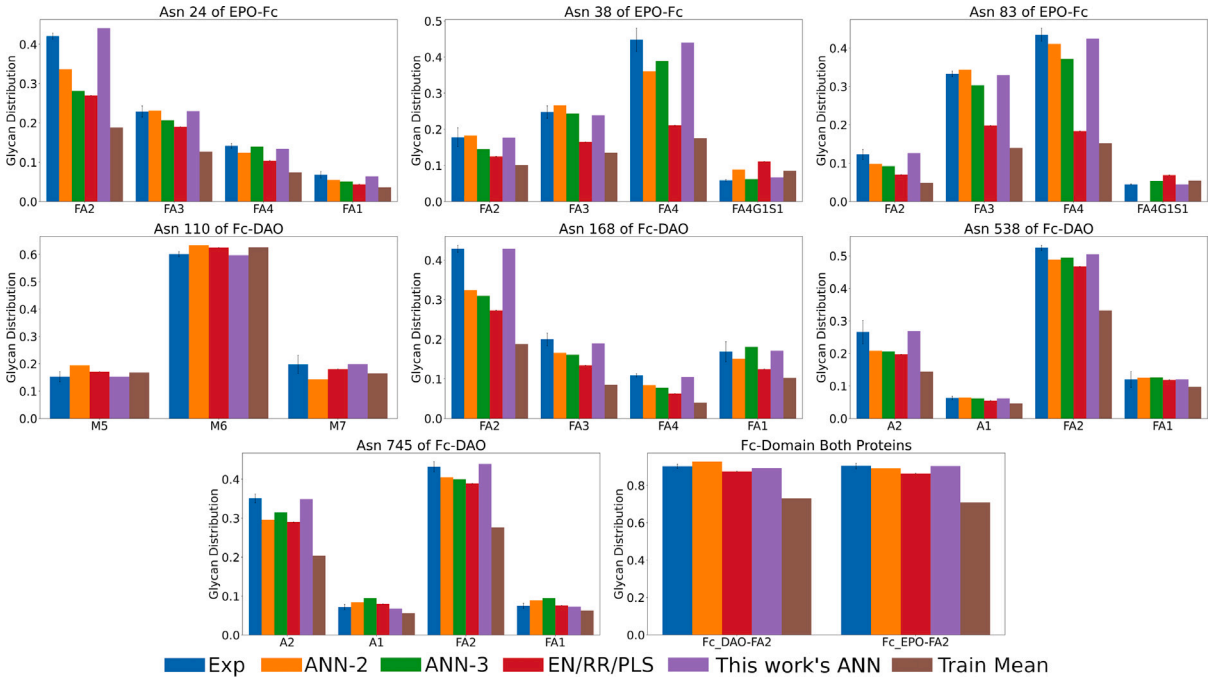

Fig. 1. Glycan distributions predicted by different models from test data. Experimental results ("Exp") came from Bydlinski et al. (2018). Models "ANN-2" and "ANN-3" came from Kotidis and Kontoravdi (2020); their prediction data are obtained directly from Figs. 5 and 6 of Kotidis and Kontoravdi (2020). Models "EN/RR/PLS" and "This work's ANN" come from this work; the former are used as a baseline. "Train mean" is the mean of the training data (Bydlinski et al. (2018)). For each subplot, the bar orders and colors are the same, and each x-axis position represents a different glycan.

3. Results

3.1. Properly trained ANN models display a 9-fold reduction in median prediction error on these datasets

All of the data-driven models constructed in this study for prediction of N-glycosylation are trained with hyperparameters selected by cross-validation. This section compares this work's models with the models in Kotidis and Kontoravdi (2020) (denoted by ANN-2 and ANN-3), which trained ANNs using the same data but selected the hyperparameters based on test set performance, a procedure that leads to overly optimistic model errors and issues with generalization (Kapoor and Narayanan, 2022; Liao et al., 2021). It should be noted that Kotidis and Kontoravdi (2020) makes many statements to the contrary; this is further discussed in Section S5 in the Supplemental Information. Thus, the real relative errors for the models from Kotidis and Kontoravdi (2020) are higher than what is reported in this work or Kotidis and Kontoravdi (2020). Moreover, the training procedure of Kotidis and Kontoravdi (2020) had some restrictions, such as on the learning rate and activation function, that reduced the potential prediction accuracy of its models.

The predictions of the data-driven models are compared with experimental data in the test sets in Fig. 1, and the corresponding percent relative errors (PRE) are summarized in Table 1. All predictions, PRE, and root mean square error (RMSE) data for the ANN-2 and ANN-3 models are obtained from Figs. 5 and 6 of Kotidis and Kontoravdi (2020). The median prediction error for this work's ANNs is 9-fold lower than for the previously published models. The mean and median prediction errors for ANN-2 and ANN-3 are higher than for this work's ANNs, in spite of being selected based on their performance on the test set instead of on a validation set, indicating issues with the ANN training in Kotidis and Kontoravdi (2020).

The median and mean prediction errors for the linear models are higher than for any of the ANN models. On the other hand, the linear models have the lowest prediction errors for one of the glycans, which suggests that any nonlinearity in the true relationship for that glycan is low enough that the bias of assuming linearity is small relative to the increase in variance associated with having more degrees of freedom in the training of the ANN models (Sun and Braatz, 2021).

The PRE for each model for each glycan is reported in Table 2. All of the models have low prediction errors for some glycans (e.g., Fc-Domain of Fc-DAO — FA2), but other glycans tend to have higher prediction errors (e.g., Asn38 of EPO-Fc — FA4G1S1). These prediction errors are not meaningfully correlated with the Train Mean prediction error (all $R^2 < 0.10$, except for the EN/RR/PLS models, which have $R^2 = 0.47$). This work's ANNs are the best model 25/29 times and produce lower prediction errors than the training mean for all glycans. In contrast, the predictions for ANN-2 and ANN-3 are worse than the training mean for some glycans.

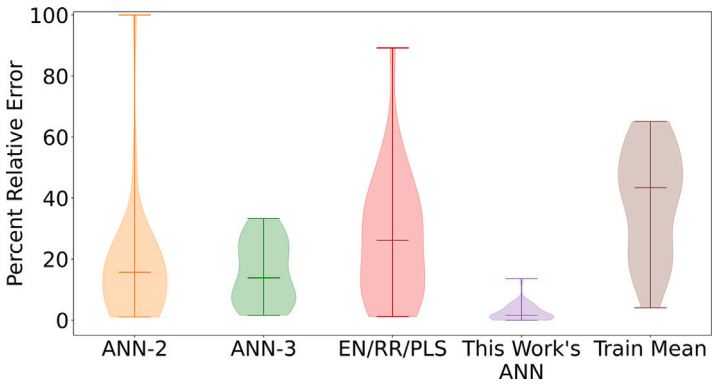

Fig. 2. Violin plots showing the distribution of test set PREs for different models. Model labels are as in Fig. 1. The horizontal bars within the distributions are median PREs.

Table 2
Test set percent relative errors (PREs) on the main glycans for different N-glycosylation models. Models "ANN-2" and "ANN-3" came from Kotidis and Kontoravdi (2020); their PRE data are obtained from Figs. 5 and 6 of Kotidis and Kontoravdi (2020). Models "EN/RR/PLS" and "ANNs" come from this work; the former are used as a baseline. For each glycan, the "EN/RR/PLS" column reports the PRE for the model type with the best cross-validation score. "Train mean" is the PRE when using the mean of the training data (Bydlinski et al. (2018)). For each glycan, the lowest PRE(s) among the models is in bold. "N/A" marks predictions that are not reported in Kotidis and Kontoravdi (2020).

Site	Glycan	ANN-2 (From Kotidis and Kontoravdi (2020))	ANN-3 (From Kotidis and Kontoravdi (2020))	EN/RR/PLS (Baseline)	ANNs (This work))	Train mean (From Bydlinski et al. (2018))
Asn24 of EPO-Fc	FA2	19.95	33.25	35.87	**4.75**	55.40
Asn24 of EPO-Fc	FA3	1.02	9.48	16.91	**0.44**	44.39
Asn24 of EPO-Fc	FA4	12.88	**1.64**	26.93	5.63	47.89
Asn24 of EPO-Fc	FA1	19.51	25.37	35.17	**5.88**	47.80
Asn38 of EPO-Fc	FA2	3.00	18.39	29.64	**0.56**	43.43
Asn38 of EPO-Fc	FA3	7.11	**1.74**	33.56	3.63	45.54
Asn38 of EPO-Fc	FA4	19.36	13.10	52.87	**1.79**	60.82
Asn38 of EPO-Fc	FA4G1S1	51.70	**5.68**	89.20	13.56	45.45
Asn83 of EPO-Fc	FA2	20.11	25.00	42.61	**2.44**	59.92
Asn83 of EPO-Fc	FA3	3.41	8.92	40.48	**0.90**	57.84
Asn83 of EPO-Fc	FA4	5.59	14.55	57.73	**2.30**	65.07
Asn83 of EPO-Fc	FA4G1S1	100.00	20.00	52.22	**0.00**	22.78
Asn110 of Fc-DAO	M5	27.07	N/A	12.01	**0.00**	10.26
Asn110 of Fc-DAO	M6	5.43	N/A	3.94	**0.67**	4.02
Asn110 of Fc-DAO	M7	27.78	N/A	8.54	**0.51**	16.54
Asn168 of Fc-DAO	FA2	24.53	27.79	36.41	**0.00**	56.19
Asn168 of Fc-DAO	FA3	17.12	19.62	33.10	**5.00**	57.61
Asn168 of Fc-DAO	FA4	23.15	28.64	42.36	**3.67**	63.86
Asn168 of Fc-DAO	FA1	10.81	6.91	26.17	**1.18**	39.56
Asn538 of Fc-DAO	A2	21.80	22.56	25.94	**1.13**	45.80
Asn538 of Fc-DAO	FA2	6.80	5.65	10.99	**3.81**	36.82
Asn538 of Fc-DAO	A1	**1.59**	**1.59**	12.64	**1.59**	27.65
Asn538 of Fc-DAO	FA1	3.88	4.71	1.11	**0.00**	19.04
Asn745 of Fc-DAO	A2	15.67	10.26	17.38	**0.57**	41.93
Asn745 of Fc-DAO	FA2	6.32	7.48	10.02	**1.62**	36.12
Asn745 of Fc-DAO	A1	17.21	32.56	11.35	**5.56**	21.28
Asn745 of Fc-DAO	FA1	19.20	27.23	**2.19**	2.67	15.29
Fc-Domain of Fc-DAO	FA2	2.73	N/A	3.03	**1.00**	19.12
Fc-Domain of EPO-Fc	FA2	1.36	N/A	4.46	**0.11**	21.71

The distribution of prediction errors for this work's ANNs are concentrated near low values (Fig. 2). At the other extreme, the distribution of prediction errors for ANN-2 has a wide spread with very high prediction error (≥50%) for two glycans. The linear models also have a long tail, although not to the same degree. ANN-3 has a lower range of prediction errors, but it is still significantly higher than that of this work's ANNs (Fig. 2).

3.2. Nested validation highlights the robustness of this work's models

This work uses the same test set as Kotidis and Kontoravdi (2020) to allow a fair comparison between models. To ensure the models constructed in this study are not biased by that choice of test set, nested validation was performed. In each round, a group within the complete dataset is selected as the test set, and the rest is used for training and cross-validation. These rounds repeat until all groups have been used as the test set. In this work, five rounds of nested validation are performed, (as per Section 2.1).

The difference is negligible between the PREs of the models selected with and without nested validation (Fig. 3). The Jensen–Shannon distance between the EN/RR/PLS distribution is 0.124 and that between the ANN distributions is 0.387. As such, the selection of test set did not introduce any significant bias to the models, and the models are robust to changes in the training and testing data.

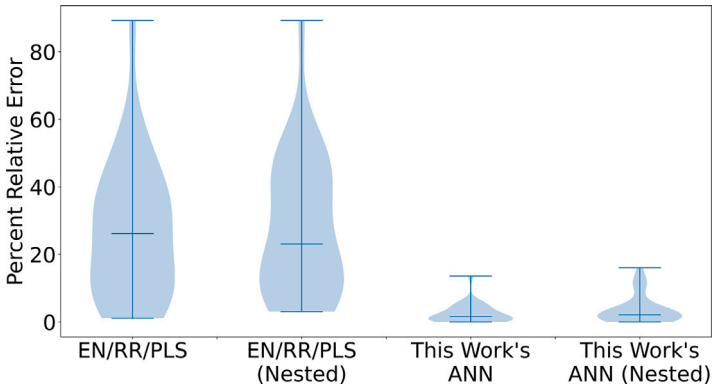

Fig. 3. Violin plots showing the distribution of test set PREs before and after nested validation. Model labels are as in Fig. 1. The horizontal bars within the distributions are median PREs.

3.3. Shapley values for interpretation of the model predictions

As detailed in Section 2.4, Shapley values assign a linear coefficient to each input of a model (in this work, the levels of B4GALT1–4). These coefficients allow for interpretation of the effects of each enzyme on a given position/glycan combination. As these coefficients can vary depending on what enzymes are present or absent, the coefficients allow the elucidation of interactions and dependencies among the enzymes. Two examples are provided in Fig. 4, but these Shapley values may be obtained for any glycosylation site and glycan in this work. In the position Asn538 of Fc-DAO, the levels of A2 are always reduced in the presence of any B4GALT. B4GALT3 by itself has a considerably negative effect on the levels of A2, but this effect is considerably diminished in the presence of B4GALT1, and B4GALT3 even increases the levels of A2 in the presence of B4GALT1 and B4GALT2 or 4. In the position Asn745 of Fc-DAO, B4GALT4 has a marginally positive effect on FA1, and this effect becomes larger in the presence of B4GALT1 or 3, but smaller in the presence of B4GALT2. Conversely, B4GALT2 has a marginally negative effect on FA1, and that effect becomes more negative in the presence of B4GALT4. These predictions are in accordance with what was previously reported in previous studies (such as Bydlinski et al. (2018), Lee et al. (2001), Yang et al. (2015) and McDonald et al. (2014)), corroborating the quality of the models of this work and the usefulness of Shapley values in providing interpretable results.

4. Discussion

This study constructs models from literature data on antibody and fusion protein glycan distribution based on normalized CHO cell B4GALT1–B4GALT4 levels. Two types of models are trained to predict the glycan distribution based on user-input B4GALT levels. Linear models (EN/RR/PLS) provided a 1.7-fold reduction in median error relative to the train mean on predictions of glycan distribution despite the linear models' simplicity. Artificial neural network models are constructed that have significantly improved prediction performance compared to published models. The median PRE of the new ANN models is 1.59% (Table 1), with the error distribution concentrated near low values (Fig. 2). This median PRE is 9-fold lower than for the other models. A nested validation study shows that the prediction performance of the linear and ANN models was insensitive to the split of the data into training/validation and test sets (Section 3.2 and Fig. 3).

The models trained in this work are compared to the ANN models of Kotidis and Kontoravdi (2020), which did not properly separate the experimental data into training, validation, and test sets (Section S5). This issue effectively ensured Kotidis and Kontoravdi (2020)'s models are overfit, resulting in overly optimistic estimates of the prediction errors. Kotidis and Kontoravdi (2020) manually set the ANN structures and fixed the choice of some hyperparameters, namely, the learning rate and activation function. The models in Kotidis and Kontoravdi (2020) also were not trained using learning rate scheduling. These choices limit the potential prediction performance of the models in Kotidis and Kontoravdi (2020), and explain why the PREs of their models are high despite being overfitted to the test data.

As previously mentioned, Liang et al. (2020) attempted to predict N-glycosylation distributions using a different architecture (a hidden Markov model, HMM). While the PREs of this work and the (absolute) RMSEs from Liang et al. (2020) cannot be compared directly because the datasets and input types are different, it should be noted that the model from Liang et al. (2020) attains an absolute RMSE equal to 1.91% and 1.26% on a mixed knockout set (Fig. 5 of Liang et al. (2020)) and equal to 5.85% and 4.89% on *de novo* predictions (Fig. 6 of Liang et al. (2020)). This work's ANNs have an absolute error of 0.50%, about 3.2-fold lower than that from the Liang et al. (2020) HMM on a mixed knockout set and 10.7-fold lower than that from the same HMM on a *de novo* prediction. Absolute errors are not ideal metrics because many glycans are present at very low percentages (or even not present), so absolute errors are biased downwards. The mean PRE of Liang et al. (2020)'s model is equal to 115.2% and 64.5% on a mixed knockout set and 59.0% and 34.9% on *de novo* predictions, 38.2- and 20.0-fold (respectively) higher than this work's ANNs' mean PRE (2.35%). Using only glycans with a distribution ≥3% to calculate the PREs (to avoid minor perturbations having a disproportional effect on the PRE), the mean PRE of Liang et al. (2020)'s model is equal to 40.4% and 43.4% on a mixed knockout set and 36.9% and 45.8% on *de novo* predictions, 17.8- and 19.5-fold (respectively) higher than this work's ANNs' mean PRE.

According to their protocol (Liang et al. (2023)), a user would need 17.1 h to train Liang et al. (2020)'s model on one sample using one copy of MATLAB. This work's ANNs, on the other hand, can be cross-validated (using 5 folds) and trained in about 5 min. Moreover, their protocol states that the prediction of a mixed knockout set took 15 min, and a *de novo* prediction on another protein took about 30 h with a single MATLAB license. This work's ANNs can generate predictions in about 2 s. The HMM model from Liang et al. (2020) can model only wild-type levels or full enzyme knockouts, whereas this work's ANNs can handle any numerical values for the enzyme models, including partial knockouts and upregulations. Finally, due to its HMM architecture, the model from Liang et al. (2020) will return different outputs if ran multiple times on the same input, whereas this work's models always return the same output when given the same input.

The software used in this work is publicly available, allowing other researchers to reproduce this work and reuse or improve the code in future studies. The software provides a simple way to install and run the models to predict the glycan distribution of antibodies and fusion proteins made in CHO cells given normalized B4GALT1–B4GALT4 levels.

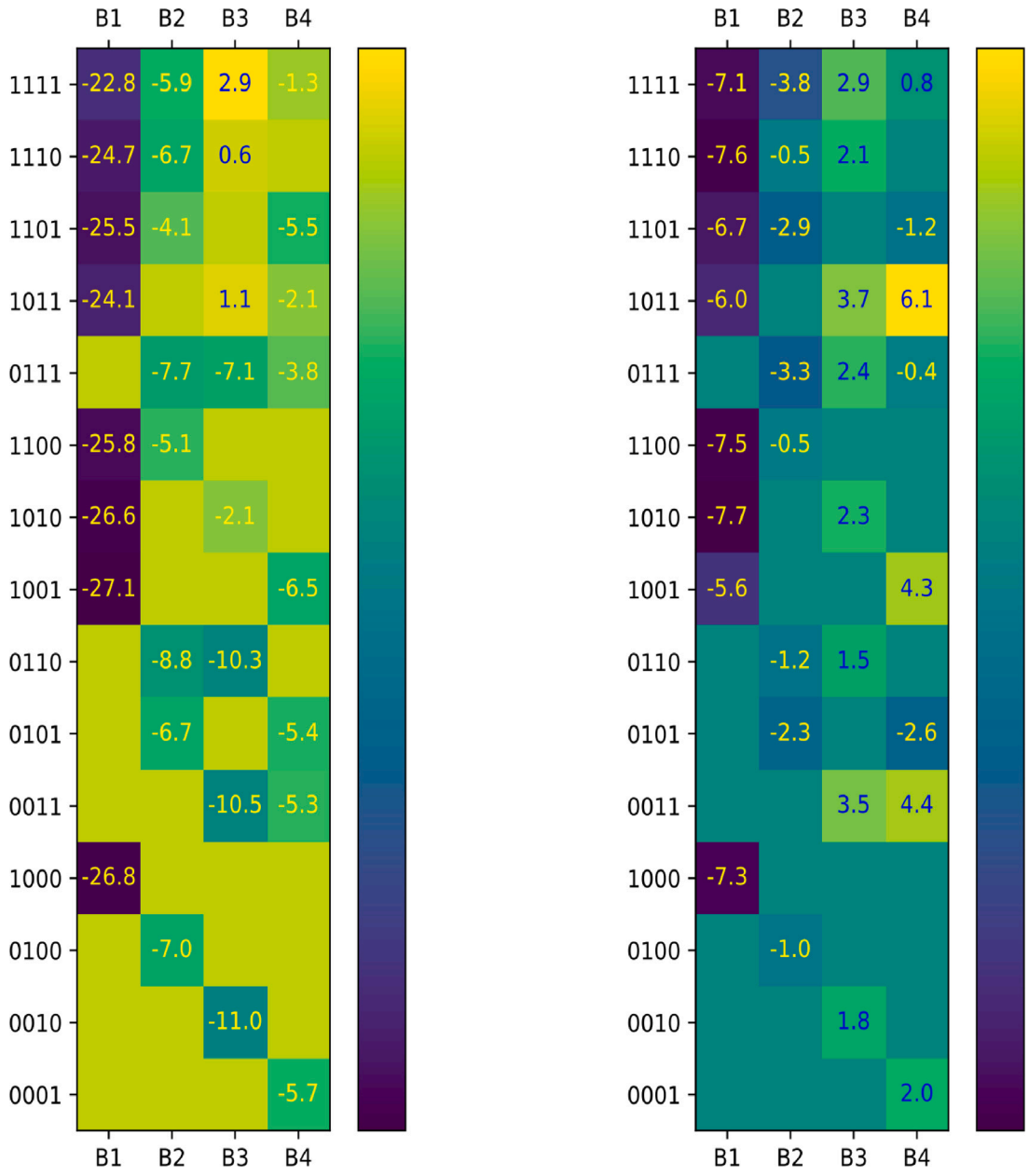

Fig. 4. Predicted Shapley values for Asn538 of Fc-DAO — A2 (left) and Asn745 of Fc-DAO — FA1 (right). B1 to B4 represent B4GALT1–4 respectively. The *y*-axis labels represent what enzymes are present (e.g.: 1111 is the wild-type, 1101 indicates only B4GALT3 was knocked out, 0001 indicates only B4GALT4 is present). Coefficients not shown are equal to 0 (as their respective enzymes have been knocked out in that sample).

CRediT authorship contribution statement

Pedro Seber: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Richard D. Braatz:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was financially supported by a Project Award Agreement from the National Institute for Innovation in Manufacturing Biopharmaceuticals (NIIMBL), U.S., with financial assistance from awards 70NANB17H002 and 70NANB20H037 from the U.S. Department of Commerce, National Institute of Standards and Technology. P.S. was partially supported by a MathWorks Engineering Fellowship.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compchemeng.2024.108937.

Data availability

The raw data, code used to process these data, and models underlying this article are available in a GitHub repository at https://github.com/PedroSeber/CHO_N-glycosylation_prediction.

References

Ahmed, M., Cheung, N.-K.V., 2014. Engineering anti-GD2 monoclonal antibodies for cancer immunotherapy. FEBS Lett. 588 (2), 288–297.

Almeida, A., Kolarich, D., 2016. The promise of protein glycosylation for personalised medicine. Biochim. Biophys. Acta 1860 (8), 1583–1595.

Bhat, A.H., Maity, S., Giri, K., Ambatipudi, K., 2019. Protein glycosylation: Sweet or bitter for bacterial pathogens? Crit. Rev. Microbiol. 45 (1), 82–102.

Bork, K., Reutter, W., Weidemann, W., Horstkorte, R., 2007. Enhanced sialylation of EPO by overexpression of UDP-GlcNAc 2-epimerase/ManAc kinase containing a sialuria mutation in CHO cells. FEBS Lett. 581 (22), 4195–4198.

Bydlinski, N., Maresch, D., Schmieder, V., Klanert, G., Strasser, R., Borth, N., 2018. The contributions of individual galactosyltransferases to protein specific N-glycan processing in Chinese hamster ovary cells. J. Biotech. 282, 101–110.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362.

Ho, W.-L., Hsu, W.-M., Huang, M.-C., Kadomatsu, K., Nakagawara, A., 2016. Protein glycosylation in cancers and its potential therapeutic applications in neuroblastoma. J. Hematol. Oncol. 9 (1), 100.

Hokke, C.H., Bergwerff, A.A., Dedem, G.W.K., Kamerling, J.P., Vliegenthart, J.F.G., 1995. Structural analysis of the sialylated N- and O-linked carbohydrate chains of recombinant human erythropoietin expressed in Chinese hamster ovary cells. Sialylation patterns and branch location of dimeric N-acetyllactosamine units. Eur. J. Biochem. 228 (3), 981–1008.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95.

Imperiali, B., O'Connor, S.E., 1999. Effect of N-linked glycosylation on glycopeptide and glycoprotein structure. Curr. Opin. Chem. Biol. 3 (6), 643–649.

Jaeken, J., 2013. Chapter 179 – congenital disorders of glycosylation. In: Dulac, O., Lassonde, M., Sarnat, H.B. (Eds.), Pediatric Neurology Part III. In: Handbook of Clinical Neurology, vol. 113, Elsevier, Amsterdam, pp. 1737–1743.

Kapoor, S., Narayanan, A., 2022. Leakage and the reproducibility crisis in ML-based science. arXiv:2207.07048.

Kotidis, P., Kontoravdi, C., 2020. Harnessing the potential of artificial neural networks for predicting protein glycosylation. Metab. Eng. Commun. 10, e00131.

Lee, J., Sundaram, S., Shaper, N.L., Raju, T.S., Stanley, P., 2001. Chinese hamster ovary (CHO) cells may express six β4-galactosyltransferases (β4GalTs): Consequences of the loss of functional β4galt-1, β4galt-6, or both in CHO glycosylation mutants. J. Biol. Chem. 276 (17), 13924–13934.

Liang, C., Chiang, A.W., Hansen, A.H., Arnsdorf, J., Schoffelen, S., Sorrentino, J.T., Kellman, B.P., Bao, B., Voldborg, B.G., Lewis, N.E., 2020. A Markov model of glycosylation elucidates isozyme specificity and glycosyltransferase interactions for glycoengineering. Curr. Res. Biotechnol. 2, 22–36.

Liang, C., Chiang, A.W., Lewis, N.E., 2023. GlycoMME, a Markov modeling platform for studying N-glycosylation biosynthesis from glycomics data. STAR Protocols 4 (2), 102244.

Liao, T., Taori, R., Raji, I.D., Schmidt, L., 2021. Are we learning yet? A meta review of evaluation failures across machine learning. In: Vanschoren, J., Yeung, S. (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1, Curran Associates Inc., Red Hook, New York.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts. arXiv:1608.03983.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

McDonald, A.G., Hayes, J.M., Bezak, T., Gluchowska, S.A., Cosgrave, E.F., Struwe, W.B., Stroop, C.J., Kok, H., van de Laar, T., Rudd, P.M., Tipton, K.F., Davey, G.P., 2014. Galactosyltransferase 4 is a major control point for glycan branching in N-linked glycosylation. J. Cell Sci. 1.

McKinney, W., 2010. Data structures for statistical computing in Python. In: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference. pp. 56–61.

Moon, S., Chatterjee, S., Seeberger, P.H., Gilmore, K., 2021. Predicting glycosylation stereoselectivity using machine learning. Chem. Sci. 12 (8), 2931–2939.

Padler-Karavani, V., Yu, H., Cao, H., Chokhawala, H., Karp, F., Varki, N., Chen, X., Varki, A., 2008. Diversity in specificity, abundance, and composition of anti-Neu5Gc antibodies in normal humans: Potential implications for disease. Glycobiology 18 (10), 818–830.

Pakhrin, S.C., Aoki-Kinoshita, K.F., Caragea, D., KC, D.B., 2021. DeepNGlyPred: A Deep Neural Network-Based Approach for Human N-Linked Glycosylation Site Prediction. Molecules 26 (23).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Patterson, M.C., 2005. Metabolic mimics: The disorders of N-linked glycosylation. Semin. Pediatr. Neurol. 12 (3), 144–151.

Schjoldager, K.T., Narimatsu, Y., Joshi, H.J., Clausen, H., 2020. Global view of human protein glycosylation pathways and functions. Nat. Rev. Mol. Cell Biol. 21 (12), 729–749.

Shapley, L.S., 1951. Notes on the n-person game – II: The value of an n-person game. In: U.S. Air Force Project RAND.

Stowell, S.R., Ju, T., Cummings, R.D., 2015. Protein glycosylation in cancer. Annu. Rev. Pathol.: Mech. Dis. 10 (1), 473–510.

Sun, W., Braatz, R.D., 2021. Smart process analytics for predictive modeling. Comput. Chem. Eng. 144, 107134.

Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y., Campbell, M.P., 2019. SPRINT-Gly: predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. Bioinformatics 35, 4140–4146.

Van Landuyt, L., Lonigro, C., Meuris, L., Callewaert, N., 2019. Customized protein glycosylation to improve biopharmaceutical function and targeting. Curr. Opin. Biotechnol. 60, 17–28.

Yang, Z., Wang, S., Halim, A., Schulz, M.A., Frodin, M., Rahman, S.H., Vester-Christensen, M.B., Behrens, C., Kristensen, C., Vakhrushev, S.Y., Bennett, E.P., Wandall, H.H., Clausen, H., 2015. Engineered CHO cells for production of diverse, homogeneous glycoproteins. Nat. Biotechnol. 33, 842–844.