Deep Learning Explains the Biology of Branched Glycans from Single-Cell Sequencing Data

Rui Qin[1], Lara K. Mahal[1], Daniel Bojar[2,3,4,*]

[1] Department of Chemistry; University of Alberta; Edmonton, Alberta, T6G 2G2; Canada

[2] Department of Chemistry and Molecular Biology; University of Gothenburg; Gothenburg, 405 30; Sweden

[3] Wallenberg Centre for Molecular and Translational Medicine; University of Gothenburg; Gothenburg, 405 30; Sweden

[4] Lead Contact

* Correspondence: daniel.bojar@gu.se

Summary

Glycosylation is ubiquitous and often dysregulated in disease. However, the regulation and functional significance of various types of glycosylation at cellular levels is hard to unravel experimentally. Multi-omics, single-cell measurements such as SUGAR-seq, which quantifies transcriptomes and cell surface glycans, facilitate addressing this issue. Using SUGAR-seq data, we pioneered a deep learning model to predict the glycan phenotypes of cells (mouse T lymphocytes) from transcripts, with the example of predicting β1,6GlcNAc-branching across T cell subtypes (test set F1 score: 0.9351). Model interpretation via SHAP (SHapley Additive exPlanations) identified highly predictive genes, in part known to impact (i) branched glycan levels and (ii) the biology of branched glycans. These genes included physiologically relevant low-abundance genes that were not captured by conventional differential expression analysis. Our work shows that interpretable deep learning models are promising for uncovering novel functions and regulatory mechanisms of glycans from integrated transcriptomic and glycomic datasets.

Keywords

Glycosylation; Carbohydrates; Single-Cell Sequencing; Transcriptomics; Machine Learning; SHAP.

Introduction

Glycosylation is a ubiquitous post-translational modification of proteins. Approximately half of all proteins are glycosylated (Apweiler et al., 1999), influencing their physical properties and biological activities (A. Varki et al., 2017; Varki, 2017). Cells utilize glycosylation to control their function and fate. For example in CD8$^+$ T lymphocytes, β1,6-branched glycans of CD8$^+$ T cell surface proteins are upregulated following T cell activation to prevent overstimulation. Downregulation of sialylated core 1 *O*-linked glycans of CD8$^+$ T cell surface proteins induces apoptosis in the absence of activation, maintaining T cell homeostasis (Van Dyken et al., 2007; Priatel et al., 2000; Smith et al., 2018). Changes in glycosylation have been mechanistically implicated in cancer, infectious diseases, autoimmune diseases, metabolic disorders, and developmental defects (Demus et al., 2021; Ng and Freeze, 2018; Pinho and Reis, 2015; Qin and Mahal, 2021; Reily et al., 2019; Sun et al., 2016; Theodoratou et al., 2014; Vosseller et al., 2002).

Despite many observational studies reporting glycosylation changes in diseases, much remains unknown about the origin of these changes. Glycan biosynthesis is orchestrated in a non-templated manner by hundreds of enzymes encoded by "glycogenes", including glycosyltransferases (add sugar residues), glycosidases (remove sugar residues), sugar modifying enzymes (synthesize phosphorylated, sulfated, or acetylated glycans), enzymes of sugar metabolism pathways, and sugar transporters (A. Varki et al., 2017; Neelamegham and Mahal, 2016). Regulation of glycosylation includes transcriptional and post-transcriptional control of glycogenes, substrate availability, and intracellular trafficking of enzymes (Neelamegham and Mahal, 2016). Therefore, identifying factors driving observed changes in certain glycan structures can be a formidable challenge.

The functional significance of disease-associated glycosylation changes can also be difficult to ascertain, due to the multi-modal influences of glycans. Glycans can influence protein structure, interactions between proteins and receptors, recognition by carbohydrate-binding lectins, resistance to endocytosis and protease degradation, etc (A. Varki et al., 2017; Johannes et al., 2018; Mimura et al., 2018). One glycan feature can thus have multiple effects. For example, increased α2,3-sialylation of cancer cells contributes to cancer progression and metastasis through mechanisms including (i) immune system evasion by interacting with α2,3-sialic acid-binding, immunosuppressive Siglec receptors (e.g., Siglec-9), (ii) increasing metastatic potential via selectins (specific for α2,3-sialic acid-containing sialyl Lewis x and sialyl Lewis a antigens) displayed on circulating cells, and (iii) promoting angiogenesis and the epithelial-to-mesenchymal transition (EMT) process (Dobie and Skropeta, 2021; Natoni et al., 2016; Pietrobono and Stecca, 2021; Rodriguez et al., 2021).

Analyzing multi-omic data has recently emerged to identify glycosylation-related mechanisms in pathogenesis. Combining transcriptomic and glycomic data identified factors driving melanoma metastasis, pancreatic cancer, and HIV persistence (Agrawal et al., 2017; Colomb et al., 2020; Kurz et al., 2021). Agrawal et al. examined RNA-seq datasets of melanoma and found increased transcript levels of an enzyme synthesizing core fucosylated glycans in metastasized melanoma, matching melanoma glycosylation profiles. They also reported more transcripts of transcription factors that directly upregulate the expression of the core fucose-synthesizing enzyme (Agrawal et al., 2017).

State-of-the-art machine learning (ML) approaches, such as deep learning (DL) algorithms, are increasingly used to map transcriptomes onto phenotypic differences, such as tissue types, cancer stages and grades, drug responses, and disease outcomes, to understand and explain phenotypic outcomes on a molecular systems level (Hanczar et al., 2020; Jia et al., 2021; Smith et al., 2020; Yap et al., 2021). Model explanation methods such as Integrated Gradients (Dincer et al., 2018), LRP (layer-wise relevance propagation) (Bach et al., 2015), LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), DeepLIFT (Deep Learning Important FeaTures) (Shrikumar et al., 2017), and SHAP (SHapley Additive exPlanations) facilitated using DL models to shed light on biological questions. SHAP unifies and improves precedent methods (Lundberg and Lee, 2017). It assigns a "SHAP value" to each input feature, reflecting its impact on the expected output of a DL model for an input example. Using SHAP to identify biologically relevant transcripts contributing to phenotypic differences has only been explored recently (Huang et al., 2022; Withnell et al., 2021; Yap et al., 2021). Traditionally, these were identified by differential expression analysis (DEA), which is (i) prone to information loss due to arbitrariness regarding p-value and fold change thresholds (Bui et al., 2020; Yang et al., 2019) and (ii) biased towards highly expressed genes (Oshlack and Wakefield, 2009). Yap et al. found predictive genes identified by SHAP that were not identified by DEA (Yap et al., 2021), indicating that SHAP can detect subtle but important differences.

We hypothesized that we could develop a DL model to differentiate glycosylation states of single cells and use model interpretation approaches such as SHAP to provide meaningful biological insights into glycan biosynthesis and function (**Figure 1A**). To the best of our knowledge, DL algorithms using transcripts as inputs have not been employed to study differential glycosylation before. Our model is based on data from a new technology (SUGAR-seq) that simultaneously measures the transcriptome and glycosylation in single cells (Kearney et al., 2021). Technologies such as SUGAR-seq capture the microheterogeneity at single cell levels, which is inaccessible by bulk omics yet can be valuable for mechanistic interpretation, and also provide high volume, matched multi-omics data that enable ML modelling. Specifically, the SUGAR-seq data comprise single-cell transcriptomes of mouse T lymphocytes and the abundance data of surface β1,6-branched glycans on these cells. Using the transcriptomic data as input to predict the binarized glycosylation phenotype of the cells, our model achieved an average F1 score of 0.9351 on the test set, with comparable performance across cell types. It also outperformed alternative approaches, including Random Forest and Gradient Boosting. SHAP analysis identified highly predictive genes ("SHAP genes") that are involved in the biology of branched glycans in various ways, including their biosynthesis. SHAP genes were also enriched in immunosuppression pathways, matching known functions of branched glycans in T cells. Importantly, SHAP genes included low-abundance genes, such as glycogenes, which were not captured by DEA. Our work shows that explainable DL models are a promising tool for uncovering novel functions and regulatory mechanisms of glycans from paired single-cell transcriptomic and glycomic data.

Results

Differential Expression of Surface β1,6-branched Glycans in Mouse T Cells

Extracting mechanistic insights into the role of glycans within the context of a cell would be greatly aided by paired single-cell data combining multiple systems biology modalities. We thus used recently publicly available single-cell RNA- and lectin-seq data (Kearney et al., 2021), which included two datasets corresponding to mouse (i) tumor-infiltrating T lymphocytes (TIL) and (ii) lymph node T lymphocytes (LN) (**Figure 1B**). *Phaseolus vulgaris* leucoagglutinin (PHA-L) was used for lectin-seq. PHA-L is highly specific for branched *N*-glycans with β1,6-GlcNAc linkage ("β1,6-branched glycans"), a glycoform implicated in tumor progression, tumor metastasis, and immune cell development and functional regulation (Bojar et al., 2022; Demetriou et al., 2001; Granovsky et al., 2000; Morgan et al., 2004; Mortales et al., 2020). PHA-L binding is a proxy for the activity of alpha-1,6-mannosylglycoprotein 6-beta-*N*-acetylglucosaminyltransferase A (MGAT5), the enzyme synthesizing the β1,6-GlcNAc linkage. Based on transcriptomes, T cells were categorized into 9 major subtypes, including naïve/naïve-like CD4^{+} T cells, naïve/naïve-like CD8^{+} T cells (may include central memory T cells), early active CD8^{+} T cells, effector memory CD8^{+} T cells, terminally exhausted CD8^{+} T cells (Tex), precursor exhausted T cells (Tpex), regulatory T cells (Treg), T helper 1 cells (Th1), and follicular helper T cells (Tfh) (**Figure 1B; Figure 1C**). We validated this classification via marker gene expression (**Figure S1**). Additionally, TIL composition is highly heterogeneous, comprising activated and regulatory cells. Conversely, the LN pool mainly contained resting T cells (**Figure 1B; Figure 1C**). Observed compositions of T cells isolated from different sites are consistent with previous studies (Kumar et al., 2018; Szabo et al., 2019).

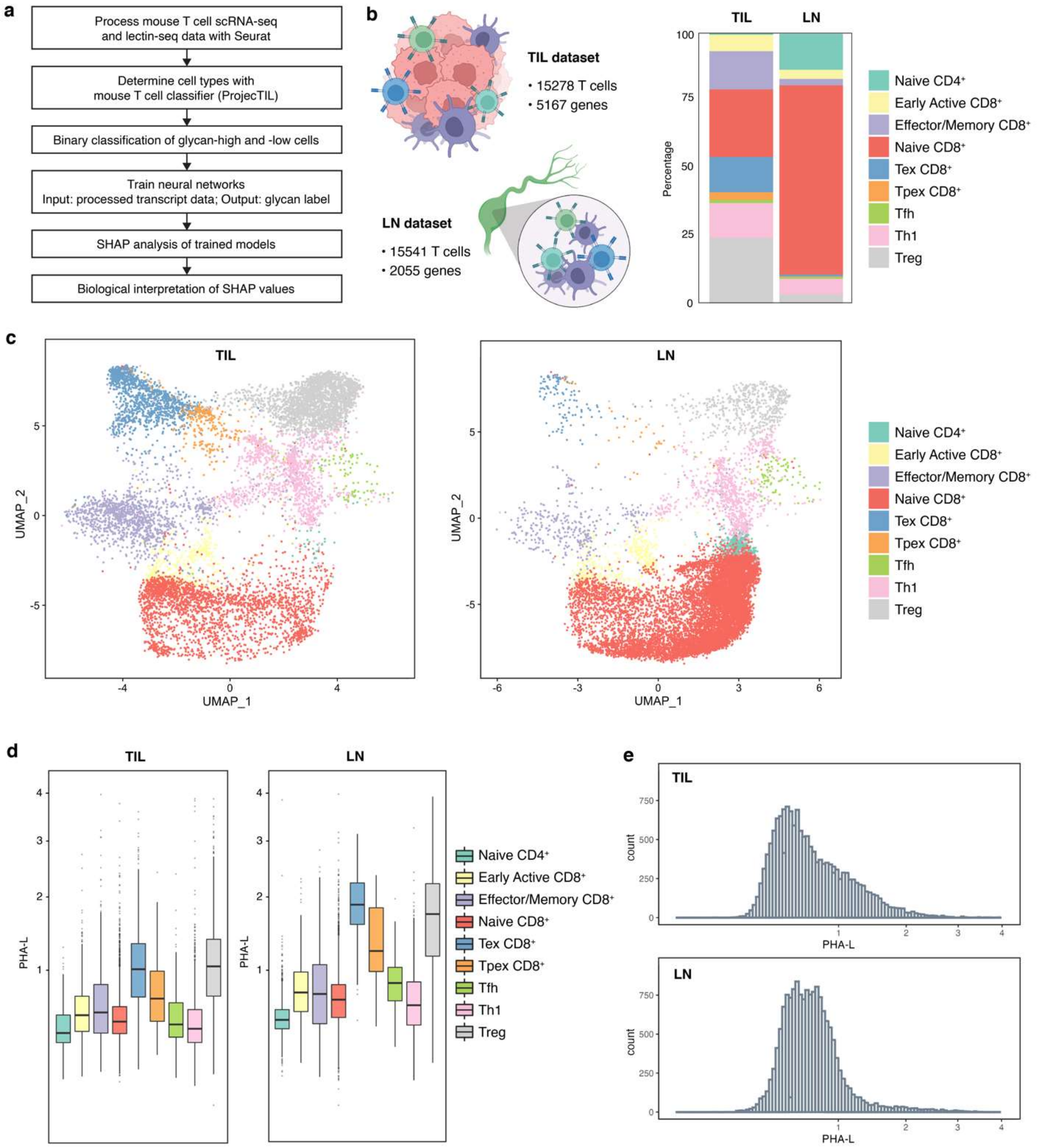


Figure 1. Different mouse T cells show distinct cell surface glycosylation patterns based on single cell RNA- and lectin-seq. (a) Graphical summary of workflow. (b) Composition of the TIL and LN datasets. Left: numbers of cells and genes in each processed dataset. Right: cell type composition as percentages in each dataset. (c) UMAP clustering of cells in each dataset.

(d) Boxplots of processed PHA-L data (β1,6-branched glycan abundance) by cell type in each dataset. **(e)** Histograms of processed PHA-L data in each dataset. Cell type annotations: Tex, terminally exhausted T cell; Tpex, precursor exhausted T cell. Tfh, follicular helper T cell; Th1, T helper 1 cell; Treg, regulatory T cell.

Surface glycosylation varies by cell type (Agrawal et al., 2014; Holst et al., 2016; Tao et al., 2008). We observed clear and reproducible differences in surface expression of β1,6-branched glycans across T cell subtypes in both datasets (**Figure 1D**). Treg and Tex consistently exhibited the highest expression of branched glycans, while naïve CD4$^+$, naïve CD8$^+$, and Th1 exhibited the lowest. Similarly, previous studies reported greater PHA-L binding to stimulated T cells than to their naïve counterparts (Cabral et al., 2017; Smith et al., 2018). Glycan expression was also more variable in TIL, consistent with its more diverse composition (**Figure 1E**). As surface branched glycans seemed to distinguish T cells of different characteristics and functions, we hypothesized that we could use DL to uncover the biology behind this differential glycosylation.

Training A Deep Learning Model to Predict Glycan Phenotypes from the Transcriptome

We set out to use DL to model surface glycosylation from transcriptome-wide gene expression data. We assigned binary labels to cells based on PHA-L data, corresponding to high and low ends of reads (top 25% "PHA-L^{high}", bottom 25% "PHA-L^{low}"). This robustly separated biologically distinct populations and enabled subsequent analyses. It also transformed our task into binary classification, with gene expression values as input, and the probabilities for the positive phenotype (PHA-L^{high}) of the corresponding cells as output.

We developed a neural network classifier comprising four hidden layers (**Figure 2A**), using data from either the TIL or the LN dataset. In the TIL hold-out test set, this classifier achieved a 92.17% prediction accuracy for the PHA-L^{high} phenotype and 95.01% for the PHA-L^{low} phenotype (AUC: 0.9359, F1 score: 0.9351, **Table 1**; **Figure 2B**). In the LN dataset, F1 score was lower, yet still exceeded 0.91. Prediction accuracies for both phenotypes remained greater than 90% (**Table 1**). This slight decrease in performance was most likely due to less variance in the LN data. Robustness of the classifier was also indicated by non-overlapping distributions for the predicted probabilities of PHA-L^{high} (**Figure 2C**). We further observed comparable performance across different cell subtypes, despite the inherent imbalance in T cell compositions (**Figure 2D**).

Using the TIL data, we trained three alternative models (Convolutional Neural Network (CNN), Random Forest, AdaBoost) on the same task and compared their performance to the abovementioned neural network model (**Table 2**). For all investigated metrics, our neural network outcompeted alternative models. Among the alternative models, Random Forest had the highest prediction accuracies (90.34% for PHA-L^{high}, 95.01% for PHA-L^{low}) and F1 score (0.9251). In contrast, the CNN model had the lowest prediction accuracies (89.82% for PHA-L^{high}, 91.60% for PHA-L^{low}) and F1 score (0.9065), although the cross-entropy loss was comparable to the standard network. We thus used the neural network model for all subsequent analyses.

Table 1. Performance metrics of the model.

	TIL			LN		
	Training set	Validation set	Test set	Training set	Validation set	Test set
Prediction Accuracy for PHA-L^{high}	95.21%	90.42%	92.17%	94.93%	93.15%	90.26%
Prediction Accuracy for PHA-L^{low}	97.85%	91.69%	95.01%	92.26%	93.94%	92.49%
Average Cross Entropy Loss	0.2031	0.2802	0.2485	0.3088	0.3323	0.3346
ROC Curve AUC (Positive class: PHA-L^{high})	0.9653	0.9106	0.9359	0.9360	0.9204	0.9137
F1 score (Positive class: PHA-L^{high})	0.9649	0.9102	0.9351	0.9371	0.9217	0.9131

[a] TIL: tumor-infiltrating T lymphocyte; LN: lymph node T lymphocyte.

[b] ROC Curve AUC: the area-under-the-curve of the receiver operating characteristic curve

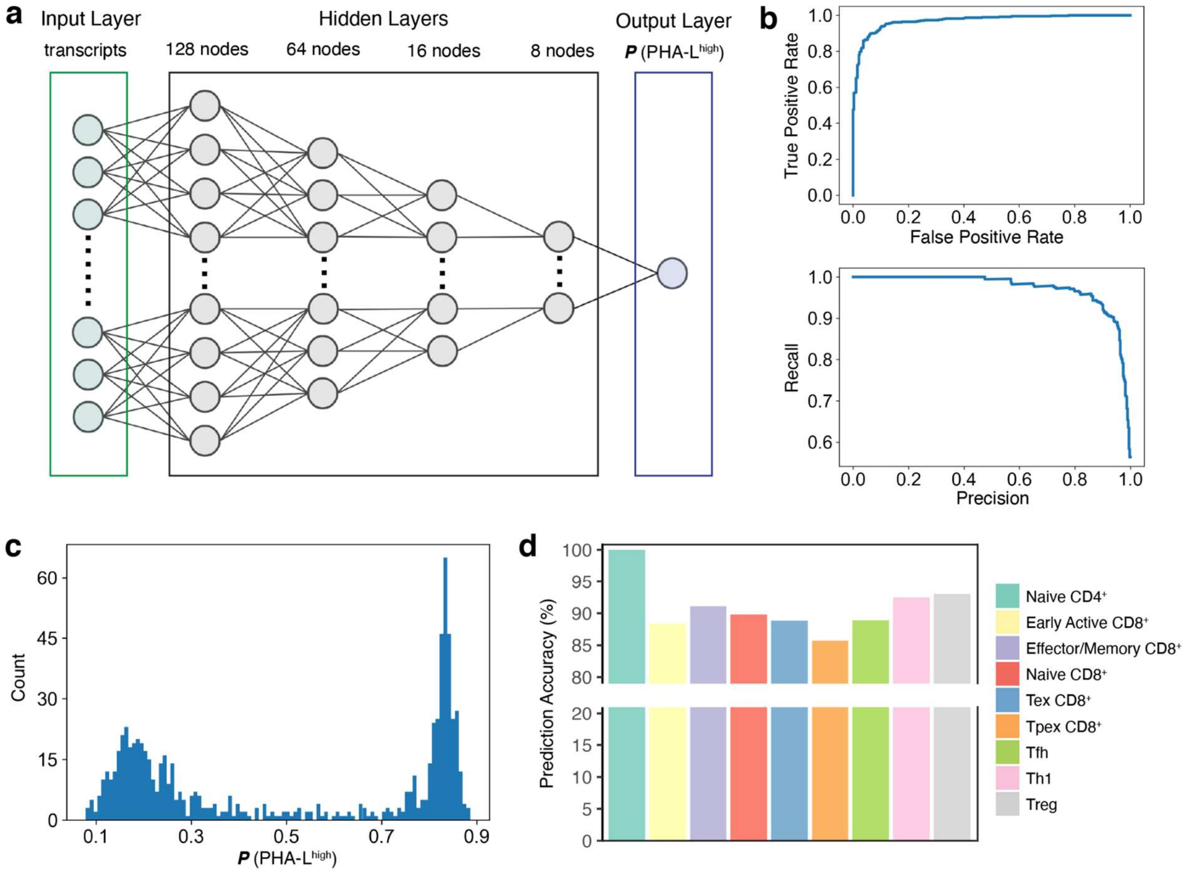


Figure 2. The deep learning model trained on TIL dataset is highly accurate in predicting glycan classes. (a) Graphical description of the neural network structure. **(b)** ROC curve (upper) and precision-recall curve (lower) of the model using the test set data. **(c)** Histogram of model output (probability for PHA-L^{high} class) using the test set data; **(d)** Prediction accuracies by cell types of the test set.

Table 2. Comparison of performance metrics of different models to predict glycan classes in the TIL dataset.

	Standard Neural Network		Convolutional Neural Network		AdaBoost		Random Forest	
	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
Prediction Accuracy for PHA-L^{high}	94.21%	92.17%	91.18%	89.82%	90.09%	90.86%	91.07%	90.34%
Prediction Accuracy for PHA-L^{low}	97.85%	95.01%	92.93%	91.60%	93.03%	92.91%	96.53%	95.01%
Average Cross Entropy Loss	0.2031	0.2485	0.2035	0.2499	0.6690	0.6556	0.2891	0.2901
ROC Curve AUC (Positive class: PHA-L^{high})	0.9653	0.9359	0.9205	0.9071	0.9156	0.9188	0.9380	0.9268
F1 Score (Positive class: PHA-L^{high})	0.9469	0.9351	0.9200	0.9065	0.9145	0.9182	0.9346	0.9251

[a] ROC Curve AUC: the area-under-the-curve of the receiver operating characteristic curve

Identification of Highly Predictive Genes Using SHAP

Next, we used SHAP to identify genes important for prediction. For each input sample (cell), the SHAP algorithm calculates a "SHAP value" for each feature (gene) that is reflective of the impact of this feature on the expected model output for this input (Lundberg and Lee, 2017). SHAP values can be positive or negative, corresponding to additive or subtractive effects on model output. The median absolute SHAP value is commonly used to assess global feature importance.

We averaged SHAP values derived from three identical models with comparable performance, separately trained using different seeds for random splitting of datasets. Then we calculated median absolute SHAP values for all genes (**Table S1**), resulting in gene rankings. For both the TIL and the LN model, only a small portion of genes were highly important (**Figure 3A**). We chose genes corresponding to the top 10% SHAP rankings ("SHAP genes"), yielding 516 genes for TIL and 205 genes for LN (see **Figure 3B** (TIL) and **Figure S2** (LN) for examples). For most SHAP genes (e.g., *Lag3*, *Ccl5*, *Foxp3*, **Figure 3B**), higher expression tended to favor a PHA-L^{high} prediction, whereas the opposite was true for a smaller subset (e.g., *Gm42418*, *Malat1*, *Klf2*, **Figure 3B**). As glycogenes directly control glycan biosynthesis (Neelamegham and Mahal, 2016), we also examined the top 30 glycogenes (**Figure 3C**), although not all were among the top 10% SHAP genes.

Next, we performed pathway enrichment analysis on the SHAP genes. For both the TIL and LN model, the most enriched pathways (biological processes) were involved in negative regulation of T cell activity and T cell differentiation (TIL: **Figure 3D**; LN: **Figure S3**), matching known roles of β1,6-branched glycans in T cells discussed further below. STRING protein interaction analysis also showed the top 10% SHAP genes to be highly interconnected via functional enrichment, co-expression, or direct interaction, arguing for concerted biology (**Figure 3E**).

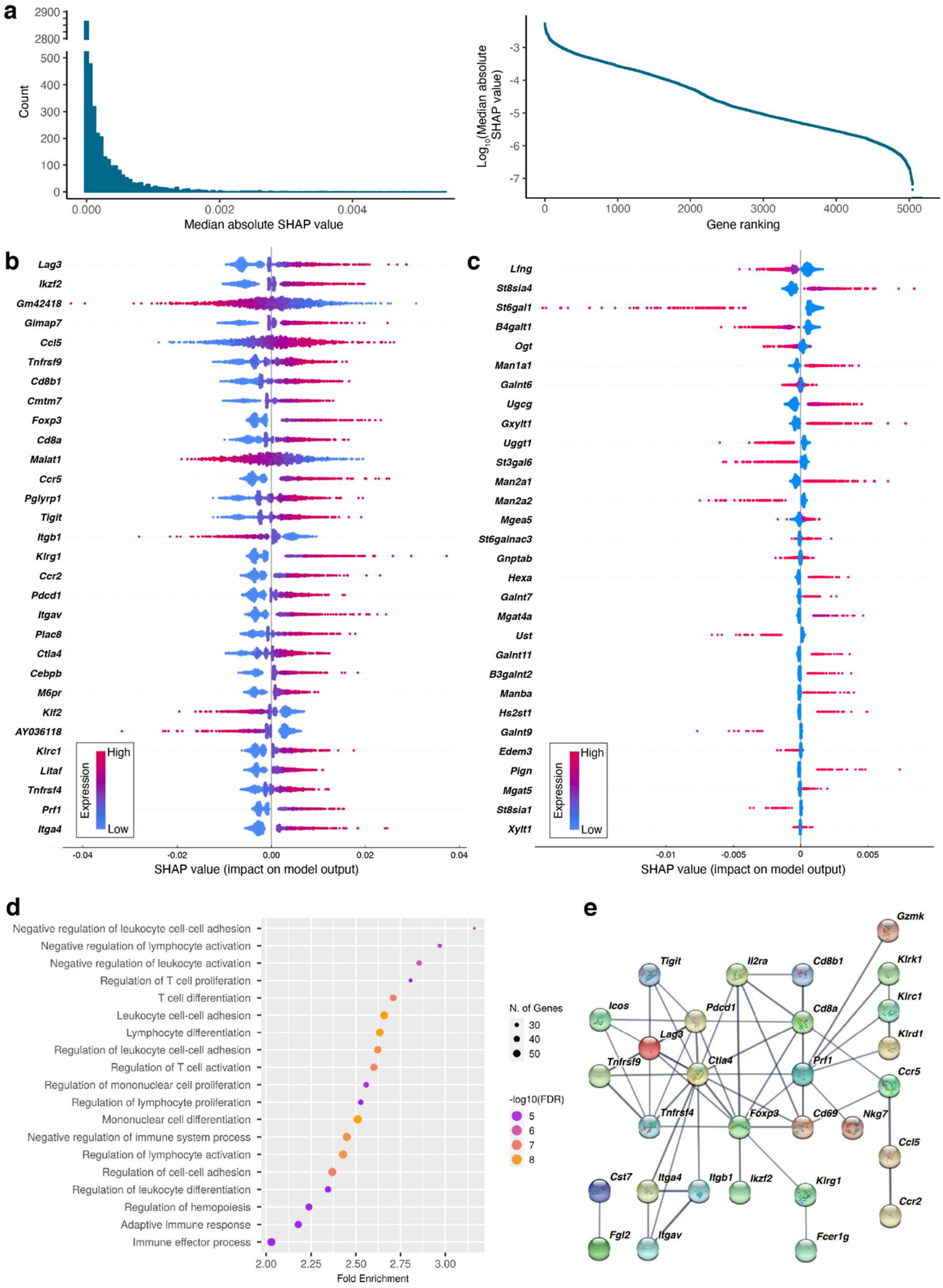
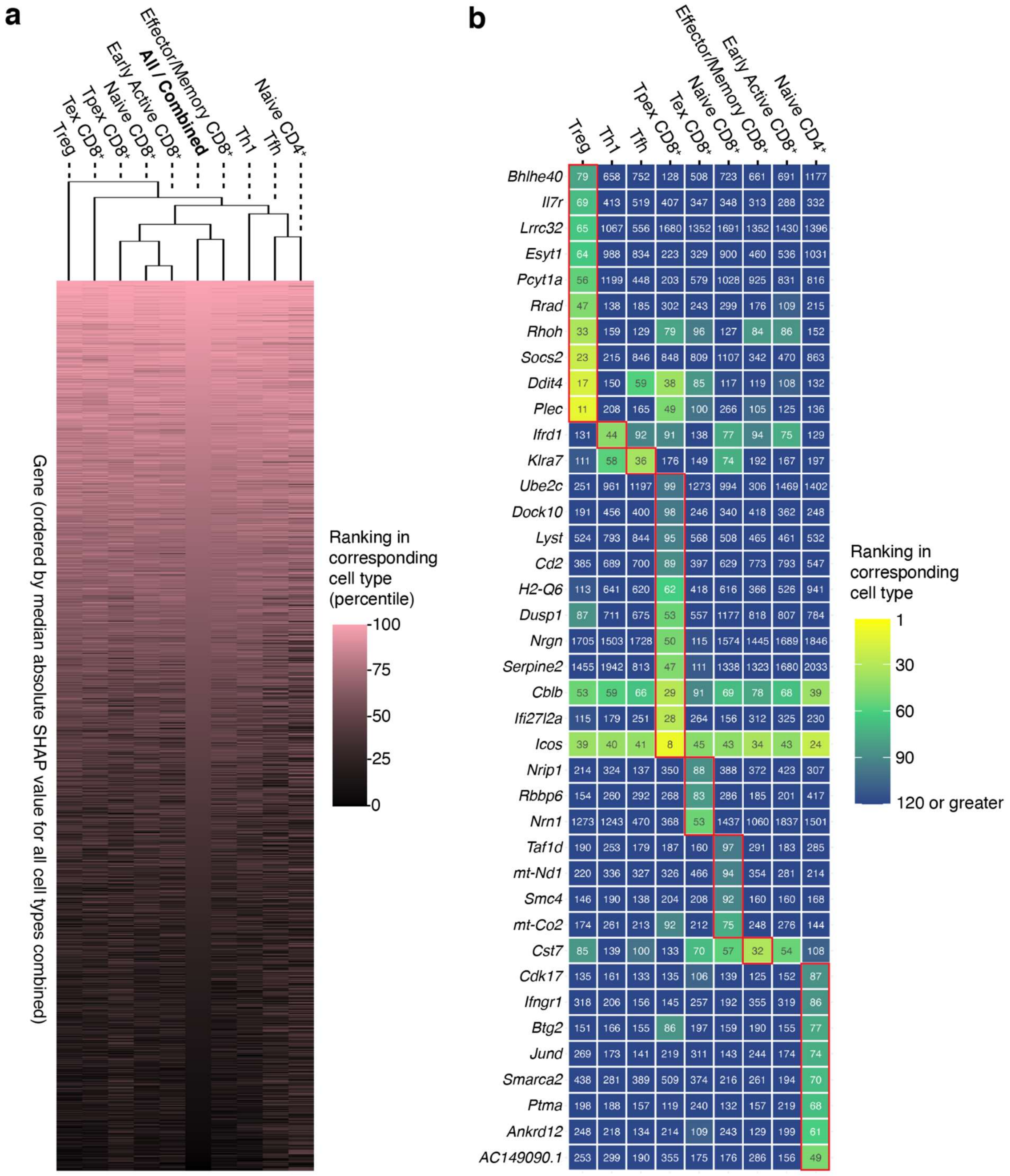

a
Count
2900
2800
500
400
300
200
100
0
0.000
0.002
0.004
Median absolute SHAP value
Log$_{10}$(Median absolute SHAP value)
-3
-4
-5
-6
-7
0
1000
2000
3000
4000
5000
Gene ranking
b
Lag3
Ikzf2
Gm42418
Gimap7
Ccl5
Tnfrsf9
Cd8b1
Cmtm7
Foxp3
Cd8a
Malat1
Ccr5
Pglyrp1
Tigit
Itgb1
Klrg1
Ccr2
Pdcd1
Itgav
Plac8
Ctla4
Cebpb
M6pr
Klf2
AY036118
Klrc1
Litaf
Tnfrsf4
Prf1
Itga4
Expression
High
Low
-0.04
-0.02
0.00
0.02
0.04
SHAP value (impact on model output)
c
Lfng
St8sia4
St6gal1
B4galt1
Ogt
Man1a1
Galnt6
Ugcg
Gxylt1
Uggt1
St3gal6
Man2a1
Man2a2
Mgea5
St6galnac3
Gnptab
Hexa
Galnt7
Mgat4a
Ust
Galnt11
B3galnt2
Manba
Hs2st1
Galnt9
Edem3
Pign
Mgat5
St8sia1
Xylt1
Expression
High
Low
-0.01
-0.005
0
0.005
SHAP value (impact on model output)
d
Negative regulation of leukocyte cell-cell adhesion
Negative regulation of lymphocyte activation
Negative regulation of leukocyte activation
Regulation of T cell proliferation
T cell differentiation
Leukocyte cell-cell adhesion
Lymphocyte differentiation
Regulation of leukocyte cell-cell adhesion
Regulation of T cell activation
Regulation of mononuclear cell proliferation
Regulation of lymphocyte proliferation
Mononuclear cell differentiation
Negative regulation of immune system process
Regulation of lymphocyte activation
Regulation of cell-cell adhesion
Regulation of leukocyte differentiation
Regulation of hemopoiesis
Adaptive immune response
Immune effector process
2.00
2.25
2.50
2.75
3.00
Fold Enrichment
N. of Genes
30
40
50
-log10(FDR)
5
6
7
8
e
Gzmk
Tigit
Il2ra
Cd8b1
Klrk1
Klrc1
Icos
Pdcd1
Cd8a
Klrd1
Lag3
Tnfrsf9
Ctla4
Prf1
Ccr5
Tnfrsf4
Foxp3
Cd69
Nkg7
Ccl5
Cst7
Itga4
Itgb1
Ikzf2
Klrg1
Ccr2
Fgl2
Itgav
Fcer1g

Figure 3. Model interpretation identifies genes important for predicting TIL cell surface glycosylation. **(a)** Histogram (left) and scatter plot (right) presentation of the median absolute SHAP values for all genes. **(b)** SHAP values of top 30 genes ranked by median absolute SHAP value. **(c)** SHAP values of top 30 glycogenes ranked by median absolute SHAP value. **(d)** Gene Ontology pathway enrichment analysis of using the SHAP genes. **(e)** STRING protein interaction network analysis of the top 10%SHAP genes. Only high confidence (strong evidence) interactions are shown, and thicker edges denote higher confidence. Genes/proteins without high confidence interactions with any other genes/proteins are not displayed.

Next, we investigated the variation of SHAP genes with T cell subtypes, generating cell type-specific lists of SHAP genes (**Table S2**). As shown in **Figure 4A**, the majority of highly predictive genes were shared across cell subtypes. For example, *Ctla4* (cytotoxic T-lymphocyte-associated protein 4) was consistently highly ranked in all cell subtypes (highest rank-naïve CD4^{+}: 13 or top 0.25%; lowest rank-Tex: 165 or top 3.19%). *Ctla4* is a Treg marker, with research primarily focusing on its roles in Treg (Rowshanravan et al., 2018; Sobhani et al., 2021; Zheng et al., 2013). Our analysis showed that *Ctla4* expression was predictive of glycan phenotype and might influence biological pathways mediated by branched glycans in T cells beyond Treg.



	Treg	Th1	Tfh	Tpex CD8+	Tex CD8+	Naive CD8+	Effector/Memory CD8+	Early Active CD8+	Naive CD4+
Bhlhe40	79	658	752	128	508	723	661	691	1177
Il7r	69	413	519	407	347	348	313	288	332
Lrrc32	65	1067	556	1680	1352	1691	1352	1430	1396
Esyt1	64	988	834	223	329	900	460	536	1031
Pcyt1a	56	1199	448	203	579	1028	925	831	816
Rrad	47	138	185	302	243	299	176	109	215
Rhoh	33	159	129	79	96	127	84	86	152
Socs2	23	215	846	848	809	1107	342	470	863
Ddit4	17	150	59	38	85	117	119	108	132
Plec	11	208	165	49	100	266	105	125	136
Ifrd1	131	44	92	91	138	77	94	75	129
Klra7	111	58	36	176	149	74	192	167	197
Ube2c	251	961	1197	99	1273	994	306	1469	1402
Dock10	191	456	400	98	246	340	418	362	248
Lyst	524	793	844	95	568	508	465	461	532
Cd2	385	689	700	89	397	629	773	793	547
H2-Q6	113	641	620	62	418	616	366	526	941
Dusp1	87	711	675	53	557	1177	818	807	784
Nrgn	1705	1503	1728	50	115	1574	1445	1689	1846
Serpine2	1455	1942	813	47	111	1338	1323	1680	2033
Cblb	53	59	66	29	91	69	78	68	39
Ifi27l2a	115	179	251	28	264	156	312	325	230
Icos	39	40	41	8	45	43	34	43	24
Nrip1	214	324	137	350	88	388	372	423	307
Rbbp6	154	260	292	268	83	286	185	201	417
Nrn1	1273	1243	470	368	53	1437	1060	1837	1501
Taf1d	190	253	179	187	160	97	291	183	285
mt-Nd1	220	336	327	326	466	94	354	281	214
Smc4	146	190	138	204	208	92	160	160	168
mt-Co2	174	261	213	92	212	75	248	276	144
Cst7	85	139	100	133	70	57	32	54	108
Cdk17	135	161	133	135	106	139	125	152	87
Ifngr1	318	206	156	145	257	192	355	319	86
Btg2	151	166	155	86	197	159	190	155	77
Jund	269	173	141	219	311	143	244	174	74
Smarca2	438	281	389	509	374	216	261	194	70
Ptma	198	188	157	119	240	132	157	219	68
Ankrd12	248	218	134	214	109	243	129	199	61
AC149090.1	253	299	190	355	175	176	286	156	49

Figure 4. Most of the highly predictive genes are shared across cell types in the TIL dataset. **(a)** Heatmap of the percentage rankings (by median absolute SHAP value) of genes in each cell type compared to all cell types combined. Cell types are clustered by euclidean distance. **(b)** Genes uniquely high ranked (median absolute SHAP value among top 2%) in each cell type. Values in tiles are the rankings of genes (by median absolute SHAP value) in the corresponding cell types. The highest rankings of genes among all cell types are boxed in red.

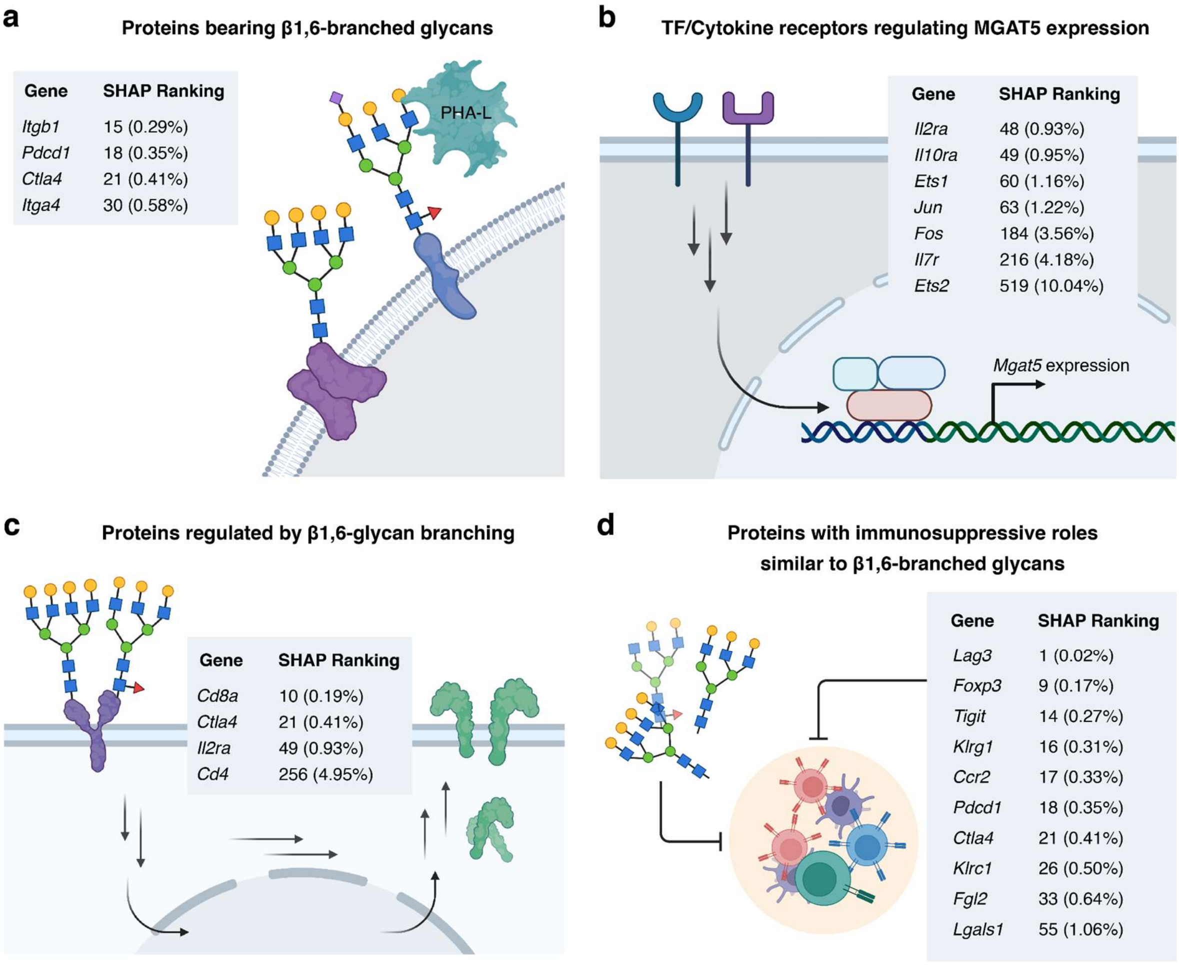
Nonetheless, we identified some genes that were highly predictive of glycan phenotype only in certain T cell subtypes. A gene is considered specifically important to a T cell subtype if its ranking percentile in this subtype is (i) among the top 2% and (ii) greater than the average ranking percentile in other subtypes by at least 1.25 standard deviations. This identified gene subsets for each cell subtype (**Figure 4B**, TIL). Treg, naïve CD4, and Tpex exhibited the highest numbers of type-specific highly predictive genes. Some genes play unique biological roles in their corresponding subtypes and are associated with the biology of branched glycans. For example, our cell-type specific comparison ranked *Lrrc32* (leucine rich repeat containing 32; also known as glycoprotein-A repetitions predominant, *Garp*) as particularly high in Treg. *Lrrc32* controls the expression of latent TGF-β (transforming growth factor β) in Treg, and TGF-β signaling is suppressed in *Mgat5* knockout mice (Lehmkuhl et al., 2021; Tran et al., 2009; Zhang et al., 2021). We speculate that *Lrrc32* may mediate this loss of TGF-β signaling, explaining its importance for prediction in this context. *Jund* (transcription factor JunD) and *Ifngr1* (interferon gamma receptor 1), highly ranked in naïve CD4$^+$ T cells, are important for the activation and differentiation of naïve CD4$^+$ cells, usually followed by upregulating β1,6-branched glycans (Afkarian et al., 2002; Meixner et al., 2004; Morgan et al., 2004). *Jund* and *Ifngr1* expression may reflect differential levels of activation in naïve CD4$^+$ cells, which could correlate with branched glycan expression. As most other cell subtype-specific genes lack well-characterized roles in the corresponding subtypes of T cells, their biological associations with branched glycans are unclear and constitute opportunities for future research.

SHAP Genes Explain the Biology of β1,6-branched Glycans

One of our major goals was to see whether the most predictive SHAP genes, identified by this data-driven approach, were implicated in the biology of branched glycans. A highly predictive gene can be biologically associated with β1,6-branched glycans in several ways: (i) encoding a cell surface protein bearing this glycan; (ii) regulating the expression of MGAT5, the enzyme biosynthesizing the β1,6-branch of *N*-glycans; (iii) being itself regulated by MGAT5/β1,6-branched glycan levels; (iv) being functionally synergistic with β1,6-branched glycans; (v) influencing the biosynthesis of β1,6-branched glycans through, e.g., changing substrate availability. Indeed, many highly predictive SHAP genes (i.e., SHAP ranking in top 2%), were already implicated in β1,6-branched glycan biology in at least one of these ways (**Figure 5**; **Figure 6**), arguing for high biological relevance of SHAP genes. We detail our findings below.

We primarily focused on interpreting SHAP genes from the TIL dataset because the dataset is more representative of a general T cell population, and the corresponding model exhibited better prediction performance. First, we investigated whether TIL SHAP genes encode T cell surface proteins bearing β1,6-branched glycans, as some glycoproteins are more prone to be modified with this glycoform (Breen, 2002; Li et al., 2013; Wu et al., 2019). Out of the top 50 SHAP genes, *Pdcd1* (programmed cell death protein 1, PD-1), *Ctla4*, *Itgb1* (integrin beta 1), and *Itga4* (integrin alpha 4) can be modified with β1,6-branched glycans (**Figure 5A**) (Liu et al., 2020; Przybyło et al., 2007; Zhu et al., 2014). Except for *Itgb1*, these genes had SHAP values increasing with expression, as expected. Others either lack experimental glycosylation data or encode non-membrane proteins. Identifying proteins modified with β1,6-branched glycans among the most predictive genes may inform future investigations on whether other SHAP genes also encode proteins with this glycoform.

Next, we investigated SHAP genes regulating MGAT5 expression. Only three transcription factors seem to directly upregulate MGAT5 expression: ETS-1, ETS-2, and AP-1 (dimer of c-Jun and c-Fos) (Chen et al., 1998; Ko et al., 1999; Wang et al., 2018). In our SHAP analysis, *Ets1*, *Jun*, and *Fos* were all ranked among the top 2% genes (**Figure 5B**), coinciding with their importance in regulating MGAT5 expression. The ranking of *Ets2* was lower, potentially due to lower expression (**Figure S4**), but still bordering the top 10% range (**Figure 5B**). Intriguingly, SHAP values of *Ets1* negatively correlated with expression, potentially due to cell-dependent regulation. Indeed, RNA-seq data showed increased *Mgat5* transcripts in T cells of *Ets1*$^{-/-}$ mouse (Kim et al., 2018), suggesting ETS-1-mediated downregulation in T cells. Beyond transcription factors, we also investigated more upstream cytokine regulators. Three cytokines, interleukin-2 (IL-2), interleukin-7 (IL-7), and interleukin-10 (IL-10), upregulate β1,6-branched glycans in T cells (Grigorian et al., 2012; Smith et al., 2018). Correspondingly, *Il2ra*, *Il10ra*, *Il2rg*, and *Il7r*, encoding components of the membrane receptors of IL-2, IL-10, IL-2/IL-7, and IL-7, are highly ranked at top 0.93%, 0.95%, 3.79%, and 4.18%, respectively (**Figure 5B**). They were also the top four ranked genes out of the 29 interleukin receptor genes here, indicating their strong association with glycan branching. In aggregate, all transcription factors and cytokine receptors regulating MGAT5/branched glycan levels were highly predictive SHAP genes, showing the potency of SHAP analysis to reveal the mechanisms behind the regulation of glycosylation.



a Proteins bearing β1,6-branched glycans

Gene	SHAP Ranking
Itgb1	15 (0.29%)
Pdcd1	18 (0.35%)
Ctla4	21 (0.41%)
Itga4	30 (0.58%)

b TF/Cytokine receptors regulating MGAT5 expression

Gene	SHAP Ranking
Il2ra	48 (0.93%)
Il10ra	49 (0.95%)
Ets1	60 (1.16%)
Jun	63 (1.22%)
Fos	184 (3.56%)
Il7r	216 (4.18%)
Ets2	519 (10.04%)

c Proteins regulated by β1,6-glycan branching

Gene	SHAP Ranking
Cd8a	10 (0.19%)
Ctla4	21 (0.41%)
Il2ra	49 (0.93%)
Cd4	256 (4.95%)

d Proteins with immunosuppressive roles similar to β1,6-branched glycans

Gene	SHAP Ranking
Lag3	1 (0.02%)
Foxp3	9 (0.17%)
Tigit	14 (0.27%)
Klrg1	16 (0.31%)
Ccr2	17 (0.33%)
Pdcd1	18 (0.35%)
Ctla4	21 (0.41%)
Klrc1	26 (0.50%)
Fgl2	33 (0.64%)
Lgals1	55 (1.06%)

Figure 5. Predictive genes identified by SHAP tend to be involved in the biology of MGAT5/β1,6-branched glycans. These genes encode proteins that: **(a)** bear PHA-L binding, β1,6-branched N-glycans that can be important to their protein functions; **(b)** regulate the expression of MGAT5/β1,6-branched glycans; **(c)** are regulated by β1,6-glycan branching; **(d)** have immunosuppressive functions that may be synergistic with β1,6-branched N-glycans, which are also immunosuppressive. Gene names, their rankings (by median absolute SHAP value) and relative rankings (ranking/number of all genes × 100%, indicated in parentheses) are shown.

Dysregulated T cell-mediated immunity has been identified in *Mgat5*$^{-/-}$ animals (Demetriou et al., 2001; Lee et al., 2007; Silva et al., 2020). Therefore, we hypothesized some SHAP genes could be regulated by MGAT5/branched glycans. Deleting *Mgat5* decreases surface expression of CTLA-4, CD8 alpha coreceptor, and CD4 coreceptor in T cells, while upregulating *Mgat5*/surface β1,6-branched glycans resulted in increased T cell surface expression of IL-2 receptor (Araujo et al., 2017; Lau et al., 2007; Zhou et al., 2014). These studies argue MGAT5 upregulate these proteins was by rescuing them from endocytosis. However, whether the mRNA levels of these proteins also changed was not investigated. In our analysis, the corresponding genes, *Ctla4*, *Cd8a*, *Cd4*, and *Il2ra*, were highly ranked at 0.41%, 0.19%, 4.95% and 0.93% (**Figure 5C**). This suggests that regulation of these proteins by glycan branching may involve both mRNA expression and endocytosis.

Next, we sought to interpret genes controlling glycan biosynthesis. We identified four glycogenes, *Lfng*, *St8sia4*, *St6gal1*, and *B4galt1*, among the top 10% SHAP rankings. ST6GAL1 (Beta-galactoside alpha-2,6-sialyltransferase 1) is the predominant enzyme adding terminal α2,6-linked sialic acid to galactose residues in glycans. While this step occurs downstream of *N*-glycan branching (Petit et al., 2010), α2,6-sialylation blocks the binding of β1,6-branched glycans to PHA-L and endogenous receptors such as galectins (**Figure 6A**) (Bojar et al., 2022; Stowell et al., 2008; Zhuo et al., 2008). Thus, SHAP values of *St6gal1* were negatively correlated with expression (**Figure 3C**). The same was found for *B4galt1* (beta-1,4-galactosyltransferase 1), the primary enzyme transferring β1,4-galactose to glycans (Bydlinski et al., 2018). B4GALT1 is indispensable for efficient α2,6-sialylation (**Figure 6A**), potentially explaining the decreasing SHAP values upon *B4galt1* expression (Khoder-Agha et al., 2019; Nguyen et al., 2021).

Expression of the fifth-ranking glycogene, *Ogt* (total rank: 10.06%), decreased with SHAP values (**Figure 3C**). OGT transfers GlcNAc onto intracellular proteins. The activated sugar donor UDP-GlcNAc is the shared substrate of MGAT5 and OGT. Supplementing GlcNAc, converted to UDP-GlcNAc in cells, upregulates both *O*-GlcNAcylation and β1,6-

glycan branching (Araujo et al., 2017; Taylor et al., 2009). Conversely, a drug-induced decrease in cellular UDP-GlcNAc levels affected both *O*-GlcNAcylation and β1,6-glycan branching (Ricciardiello et al., 2018). OGT thus influences substrate availability for MGAT5, altering abundances of β1,6-branched glycans (**Figure 6B**). A new study showed that shRNA knockdown of OGT indeed increased β1,6-glycan branching (Song et al., 2022).

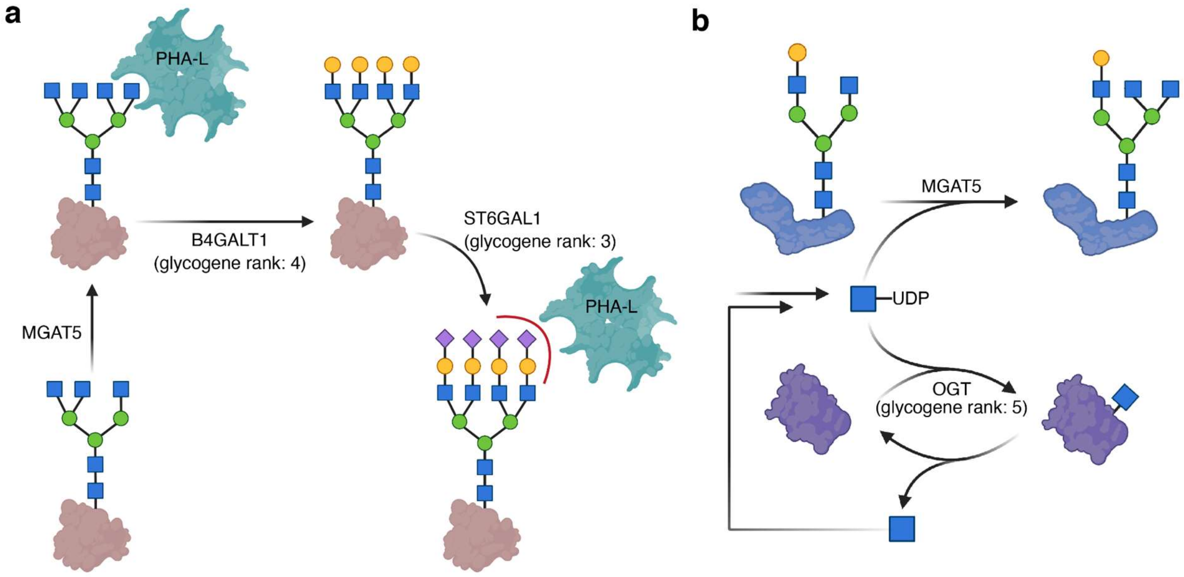


Figure 6. SHAP analysis identifies glycogenes that impact β1,6-branched N-glycan levels or PHA-L binding. (a) Partial biosynthetic route of α2,6-sialylated, β1,6-branched glycans and the involvement of B4GALT1 and ST6GAL1 identified by SHAP analysis in this process. α2,6-sialylation abrogates PHA-L binding to branched N-glycans. (b) Interplay between UDP-GlcNAc, O-GlcNAcylation and N-glycan branching, and the involvement of OGT in this process. UDP: uracil diphosphate group.

Having identified biological associations between highly ranked SHAP genes and branched glycans, we asked whether we could systematically identify functional roles of glycan branching in T cells. We based this on previous observations that cells utilize both proteins and protein glycosylation to fulfill their function. For example, fucosylated glycans and sialyl Lewis x on effector T cell surfaces facilitate homing to tumor sites, as do mechanisms mediated by chemokine receptors and integrins (Alatrash et al., 2019; Sackstein et al., 2017). Work over the past two decades has shown β1,6-glycan branching to play an immunosuppressive role in activated T cells. T cell receptor (TCR) activation upregulates MGAT5, yielding more β1,6-branched glycans on T cells. This promotes binding of the multimeric galectin-3 to T cell surface glycoproteins, forming a localized lattice that prevents T cell-activating protein-protein interactions (e.g., TCR-CD8 interaction) (Demetriou et al., 2001; Lau et al., 2007; Morgan et al., 2004). Mice deficient in *Mgat5* developed autoimmune disease due to dampened negative regulation of T cell activities (Grigorian and Demetriou, 2011; Silva et al., 2020). In Treg, surface β1,6-branched glycans were positively correlated with immunosuppressive marker expression and the suppressive potency of Treg (Cabral et al., 2017). Ye et al. identified *Mgat5* as one of the four hits in a CRISPR screen for targets that enhance T cell-based cancer therapy (Ye et al., 2019). Correspondingly, functional enrichment of SHAP genes showed enrichment in negative regulation of T cell activation (**Figure 3D**). Well-established suppressive immune checkpoint receptors were within the top 50 genes, such as *Lag3*, *Ctla4*, *Pdcd1*, and *Tigit* (**Figure 5D**). We also found other highly ranked SHAP genes known for predominantly immunosuppressive roles in T cells, including: (i) genes of the killer cell lectin-like receptor family such as *Klrc1* (0.31%) and *Klrg1* (0.50%) (Huot et al., 2021; Li et al., 2016); (ii) chemokines and chemokine receptors such as *Ccl5* (0.10%), *Ccr5* (0.21%), and *Ccr2* (0.33%) (Aldinucci and Casagrande, 2018; Matsuo et al., 2021; Tu et al., 2020; Zeng et al., 2022); (iii) other genes such as *Fgl2* (0.64%) and *Lgals1* (1.06%) (Corapi et al., 2018; Hou et al., 2021) (**Figure 5D**). Overall, SHAP genes had substantial overlap with genes known to be implicated in the biology of branched glycans.

Finally, we examined the SHAP genes of the LN dataset (**Table S1**; **Figure S2**). More than half (116 genes) of the 205 SHAP genes of the LN dataset were also found among the

516 SHAP genes of the TIL dataset. Examples include *Cd8b1* (0.04% in LN, 0.14% in TIL), *Ikzf2* (0.83% in LN, 0.04% in TIL), *Gm42418* (0.13% in LN, 0.06% in TIL), and *Ets1* (4.33% in LN, 1.16% in TIL). Notably, the negative correlation between the transcript abundance and SHAP values of *Ets1* was also seen in the LN dataset. Shared high-ranking genes could be associated with the biology of β1,6-branched glycan in similar ways as discussed above. Some genes had substantially different rankings, which was anticipated since the two datasets had different compositions of T cell subtypes. For example, the highest-ranking gene, *Lag3* (lymphocyte activation gene 3), in TIL was ranked at 746 (36.30%) in LN, due to the minimal expression of *Lag3* in resting T cells comprising most of the LN dataset (Grosso et al., 2007). The roles of β1,6-branched glycans in resting T cells are much less understood. *Mgat5*$^{-/-}$ mice displayed a lowered T cell activation threshold (Demetriou et al., 2001), suggesting that β1,6-branched glycans are also immunosuppressive in resting T cells, by desensitizing them to stimulus. Therefore, we again hypothesized that the SHAP genes of the LN dataset are primarily immunosuppressive, synergizing with β1,6-branched glycans. In line with this, SHAP genes of the LN dataset are functionally enriched in the pathway of negative regulation of T cell activity (**Figure S3**). The top 50 SHAP genes of the LN dataset also included immunosuppressive genes, such as *Cxcr6*, *Socs1*, *Cd7*, and *Klrd1* (Heesch et al., 2014; Pace et al., 2000; Sempowski et al., 2004; Sheu et al., 2005; Takahashi et al., 2011). *Stab1*, known for its anti-immunosuppression activity in T cells, was ranked at 31 and had SHAP values decreasing with expression (Beyer et al., 2011; Nüssing et al., 2019; Stephen et al., 2017). One study showed that IL-7 treatment reduced β1,6-glycan branching in resting T cells, in contrast to its effect in activated T cells (Mkhikian et al., 2011). Aligning with this, the SHAP values of *Il7r* (rank 2.29%) anti-correlated with mRNA expression in the LN dataset in contrast to the correlation observed in the TIL dataset, underscoring the context-aware nature of our approach.

SHAP Genes Include Low Abundance Genes That Are Not Captured by Differential Expression Analysis

SHAP identifies gene subsets that were not captured by differential expression analysis (DEA) (Yap et al., 2021). We wondered whether this could also be observed here, and whether any SHAP-exclusive genes were biologically relevant. Using the TIL data, we performed DEA between PHA-L^{high} and PHA-L^{low} populations. Applying a false discovery rate threshold of 0.05, we identified 381 differentially expressed genes (7.4% of all genes; **Table S3**), slightly less than the number of SHAP genes (516 genes or 10% of all genes). 267 genes were present in both DEA and SHAP genes, with many shared top-ranking genes (e.g., *Lag3*, *Tigit*, *Malat1*).

DEA is known for biasing towards highly expressed genes (Bui et al., 2020; Yang et al., 2019). Among the DEA genes, 18 (4.72%) were highly-transcribed ribosomal genes (Petibon et al., 2020; Zhao et al., 2018), with two even among the top 50 DEA genes. In contrast, only 8 (1.55%) ribosomal genes were found in the SHAP genes, none of which were in the top 50. Other high-abundance housekeeping genes were also found only in DEA genes, such as *Gapdh* (rank 23), *Pgk1* (rank 60), and *Actb2* (rank 71). These observations suggest that SHAP is less prone to biasing towards high-abundance genes. Pointedly, transcription factors *Jun* and *Ets2*, known to upregulate β1,6-glycan branching as discussed above, did not appear in DEA genes. Moreover, glycogenes in general were absent from DEA genes. Overall, our results indicate SHAP to be more powerful than the more conventional DEA in identifying low-abundance genes that are biologically important for phenotypic glycosylation differences.

Discussion

In contrast with the profusion of genomic, transcriptomic, and proteomic data, matching glycomic data remain scarce (Qin and Mahal, 2021). Consequently, glycosylation research has long been lacking integrated multi-omics analysis and has not benefited greatly from rapidly evolving computational tools for mapping molecular interaction networks in health and disease (Kellman and Lewis, 2021). State-of-the-art artificial intelligence technologies, used widely for drug response prediction or regulatory molecule identification (Adam et al., 2020; Kim et al., 2021; Zheng et al., 2020), remain rarely used in glycosylation studies. Single-cell RNA- and lectin-seq technologies such as SUGAR-seq have started to provide new opportunities to use DL for studying differential glycosylation, as demonstrated here.

It should be noted that the feasibility of predicting glycan features from single-cell transcriptomics data cannot be considered a foregone conclusion. Most genes that would seem relevant from a domain perspective (glycosyltransferases, sugar transporters, etc.) are typically lowly expressed and frequently absent from single-cell transcriptomics data (Nairn et al., 2008; Qiu, 2020). Nonetheless, we show that it is feasible to use the sparse nature of single-cell data to predict a glycan feature with high accuracy using a neural network model. Models predicting “cross-omics”, in this case from the realm of the transcriptome towards the glycome, will be important in the future to aid in multi-omics integration and are particularly relevant in the context of glycans, as they are technically outside the central dogma of molecular biology. Our results here, however, yet again demonstrate that glycans can be re-integrated with the rest of the central dogma, by using transcriptional information to successfully predict parts of glycan expression.

Using SHAP, we further showed that, next to predicting glycan feature abundance, our model can also be used to extract, at scale, compelling biological associations between the transcriptome and the glycome. Tellingly, the genes most important for glycosylation phenotype prediction significantly overlapped with genes involved in the biology of β1,6-branched glycans. A direct explanation for the high predictivity of some SHAP genes could be that they encode membrane proteins bearing β1,6-branched glycans. Although we only identified four genes with experimental evidence for β1,6-branching glycosylation among the top 50 TIL SHAP genes (*Pdcd1*, *Ctla4*, *Itgb1*, *Itga4*), branched glycans are involved in functions for two of them: In persisting exhausted T cells, PD-1 interacts with galectin-9, which binds to branched glycans displayed on PD-1, to inhibit galectin-9-mediated cell death (Yang et al., 2021). Integrin beta 1 is an essential component of a series of integrin complexes that are critical mediators of T cell adhesion and signaling, and its activities are regulated by β1,6-branching of glycans (Bellis, 2004; Jankowska et al., 2018). We expect that more highly-predictive SHAP genes will be discovered to bear β1,6-branched glycans in immune cells, influencing protein function. *Lag3*, a heavily glycosylated immune checkpoint receptor implicated in many diseases (Graydon et al., 2021), was the most predictive gene in the TIL dataset. Although the glycosylation profile of this protein has not yet been determined, LAG3 binds galectin-3 in T cells, an immune system lectin with similar binding specificity as PHA-L (Demetriou et al., 2001; Kouo et al., 2015). This indicates that LAG3 function may be regulated by β1,6-branched glycans. Thus, our DL-based approach can inform future investigations on whether other SHAP genes encode proteins with this glycoform and whether it might impact protein function.

While most SHAP genes do not encode surface glycoproteins, they can still be associated with the biology of β1,6-branched glycans in multiple ways, as discussed above. We note that all known transcription factor- and cytokine-regulators of MGAT5/β1,6-branched glycans, as well as some glycogenes impacting β1,6-branched glycan synthesis, were among the top ranked SHAP genes. This emphasizes the potential of a DL-based approach to identify regulatory mechanisms of glycosylation. Examining the top ranked SHAP genes also yielded new candidates for potential regulators of β1,6-branched glycans. The chemokine receptor CCR2 (C-C chemokine receptor type 2; rank 0.33%), when bound to its ligand CCL2 (C-C motif chemokine 2), upregulates the transcription factor AP-1 that increases MGAT5 expression (Fei et al., 2021; Lin et al., 2012). In turn, AP-1 also enhances CCL2 expression (Deng et al., 2013; Fei et al., 2021; Novoszel et al., 2021). Therefore, the CCL2/CCR2 axis may be an alternative mechanism to regulate MGAT5/β1,6-branched glycans in immune cells. In line with the immunosuppressive role of β1,6-branched glycans, CCL2 secreted by cancer cells contributes to an immunosuppressive tumor microenvironment, and blockade of CCR2 in mice improved the efficacy of immune checkpoint therapy (Matsuo et al., 2021; Tu et al., 2020), suggesting a direct biomedically relevant role of this glycan feature.

Our work showed that (i) glycosylation phenotypes can be modeled by neural networks from transcriptomic data, (ii) biological processes linked to post-translational modifications such as glycosylation can be deciphered by DL model-explaining methods such as SHAP, and (iii) this combined approach may facilitate the discovery of new regulators of branched glycans and downstream effectors of branched glycan-mediated pathways. We note that this single model essentially recapitulates decades of experimental work on this aspect of biology, including generated hypotheses for future work. While we report largely overlapping regulatory associations in our set of immune cell types, future work could also compare these results with the regulation and function of β1,6-branched glycans in other cell types or other species, to develop a global understanding of the diverse roles of this glycan feature.

We envision a pipeline in which this explainable DL approach is used to analyze data generated by SUGAR-seq or similar technologies, to decode the biology of less well-studied glycoforms (e.g., high-mannose glycans, sulfated glycans, I-branched glycans) and their importance in disease (Chuzel et al., 2021; Dimitroff, 2019; de Leoz et al., 2011; Loke et al., 2016; Sun et al., 2022). For this, future work needs to expand the capabilities of the associated lectin-seq, similar to the already reported Glycan-seq technology (Oinam et al., 2021). We also anticipate the integration of more data modalities besides transcriptomes. In RNA-seq, measurement of glycogenes and other low-abundance genes is less accurate, potentially skewing the interpretation of the importance of these genes (Tarazona et al., 2011). Adding proteomic or miRNA data (important regulators of low abundance genes) to the DL model input may provide a more accurate account of the regulation of glycosylation mediated by low-abundance genes (Schmiedel et al., 2015; Thu and Mahal, 2020). We envision that this combination of systems biology and artificial intelligence will provide fresh insights into the complex, interleaved biosynthesis and functional role of glycans in various biological contexts.

Acknowledgements

This work was funded by a Branco Weiss Fellowship – Society in Science awarded to D.B., by the Knut and Alice Wallenberg Foundation, and the University of Gothenburg, Sweden. Some graphical contents were created with Biorender.com.

Author contributions

Conceptualization, D.B.; Methodology, R.Q. and D.B.; Formal Analysis, R.Q. and D.B.; Investigation, R.Q. and D.B.; Writing – Original Draft, R.Q. and D.B.; Writing – Review & Editing, R.Q., L.K.M., and D.B.; Supervision, L.K.M. and D.B.

Declaration of interests

The authors declare no competing interests.

Data availability

The matrices of processed RNA and PHA-L reads, and the R and python codes for single-cell data processing, model training and SHAP analysis have been uploaded to https://github.com/BojarLab/scGlycomics_b16_branching.

Methods

Dataset, Data Processing, and T Cell Subtype Classification

10X sequencing data of mouse tumor infiltrating T lymphocytes (TIL) and mouse lymph node T lymphocytes (LN) were downloaded from Gene Expression Omnibus (GSE166325, GSE166326). Single-cell data processing was performed with the Seurat package (version 4.0.6) in R (Hao et al., 2021). For each dataset, downloaded raw data were initialized with the Read10X function of Seurat. Doublets were removed with the HTODemux function of Seurat. Next, RNA counts were normalized and scaled with the SCTransform function of Seurat, in which the mitochondrial genes, ribosomal genes, and cell cycle phase scores (computed with the CellCycleScoring function of Seurat) were regressed out. The default residual variance cut-off (1.3) of SCTransform was used and the low variance genes were removed. PHA-L reads were processed with the Normalize function of Seurat, using the centered log ratio transformation method.

T cell subtype classification was performed with the multi-dataset reference atlas-based mouse T cell classifier ProjecTIL (version 2.0.0) in R (Andreatta et al., 2021). ProjecTIL categorized the remaining T cells into 9 major subtypes, including naive and naïve-like CD4$^+$ T cells (naïve CD4$^+$), naïve and naïve-like CD8$^+$ T cells (naïve CD8$^+$), early active CD8$^+$ T cells (early active CD8$^+$), effector memory CD8$^+$ T cells (effector memory CD8$^+$), terminally exhausted CD8$^+$ T cells (Tex), precursor exhausted T cells (Tpex), regulatory T cells (Treg), T helper 1 cells (Th1), and follicular helper T cells (Tfh). Cell subtype was confirmed by comparing marker gene expressions (**Figure S1**).

Model Training

To generate input data from the exported matrix of transcript and PHA-L reads, cells of PHA-L reads within the upper quartile range (top 25%) were assigned the label 1 (PHA-L^{high}), and cells of PHA-L reads within the lower quartile range (bottom 25%) were assigned the label 0 (PHA-L^{low}). Other cells were removed. Input data was randomly split into a training set (72%), validation set (18%), and test set (10%). All models (standard neural network, convolutional neural network, Random Forest, and AdaBoost) were trained using the same training, validation, and test set in Python 3.8.

For the standard neural network, models were trained using the PyTorch framework. Models were trained in mini-batches of size 128. For the best performing model, the neural network consisted of 4 hidden layers (the 1st, 2nd, 3rd, and 4th hidden layer, in the direction of forward propagation), which had 128, 64, 16, and 8 nodes, respectively. The 1st, 2nd, and 3rd hidden layers were each followed by leaky ReLU activation layers (negative slope 0.01), dropout regularization layers (dropout probability 0.4, 0.4, and 0.2, respectively), and batch normalization layers. The fourth hidden layer was followed by a sigmoid activation layer.

Starting learning rate was 0.0001. Binary cross entropy loss function and cosine annealing learning rate scheduler were used for model optimization.

For the convolutional neural network, models were trained using the PyTorch framework, in mini-batches of size 128. For the best performing model, the neural network consisted of 4 hidden convolutional layers and 2 hidden fully connected layers. The first convolutional layer (6 filters; filter size:1; stride:1) was followed by a batch normalization layer. Other convolutional layers (6 filters; filter size:4; stride:4) were each followed by batch normalization layers, leaky ReLU activation layers (negative slope 0.01), and max pooling layers (filter size:2; stride:2), The first fully connected layer (16 nodes) was followed by a leaky ReLU activation layer (negative slope 0.01) and a batch normalization layer. The second fully connected layer was followed by a sigmoid activation layer. Starting learning rate was 0.001. Binary cross entropy loss function and cosine annealing learning rate scheduler were used for model optimization.

Random Forest and AdaBoost classifiers were trained with the scikit-learn (version 1.0.2) library (Pedregosa et al., 2012). Grid searches were performed via 5-fold cross-validation to optimize the accuracy. The best performing Random Forest classifier had 500 estimators and a maximum depth of 2. The best performing AdaBoost classifier had 300 estimators and was trained with a learning rate of 0.1.

Model Interpretation with SHapley Additive exPlanations (SHAP)

SHAP analysis was performed with the DeepExplainer function of the SHAP library (version 0.40.0) (Lundberg and Lee, 2017). We generated three models with similar performances, using different seeds for random splitting of the datasets. Based on our computational resources, we used these models to compute SHAP values of 1000 cells randomly selected from the whole dataset. SHAP values used for biological interpretation were computed by averaging the SHAP values derived from the three models. For cell subtype-specific SHAP analysis, we computed SHAP values of 1000 randomly selected cells corresponding to each cell subtype, or all cells of that subtype if their total number was less than 1000.

Pathway Enrichment Analysis and Protein Interaction Network Analysis

Gene Ontology pathway enrichment analysis was performed in the online portal of ShinyGo version 0.76 (http://bioinformatics.sdstate.edu/go/) (Ge et al., 2020). Genes with the top 10% (90th percentile) median absolute SHAP values were compared to all genes in the dataset (background genes; top 10% genes included). False discovery rate cut-off was set to 0.05.

Protein interaction network analysis was performed in the online portal of STRING version 11.5 (https://string-db.org/) (Szklarczyk et al., 2019). Genes with the top 1% (99th percentile) median absolute SHAP values were analyzed using the “Multiple proteins” entry. To select only the high confidence interactions, the minimum required interaction score was set to 0.7.

Differential Gene Expression Analysis

Differential gene expression analysis was performed with the FindMarkers function of Seurat. By default, absolute $\log_2$(fold change) threshold was 0.25 and Wilcoxon Ranked Sum test was used to determine *p* values.

Supplemental Excel Tables

Table S1. Ranked median SHAP values of all genes.

Table S2. Ranked cell subtype-specific median SHAP values of all genes.

Table S3. Differentially expressed genes between PHA-L^{high} and PHA-L^{low} cells in the TIL dataset.

References

A. Varki et al. (2017). Essentials of Glycobiology [Internet]. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. Npj Precis. Oncol. *4*, 1–10. https://doi.org/10.1038/s41698-020-0122-1.

Afkarian, M., Sedy, J.R., Yang, J., Jacobson, N.G., Cereb, N., Yang, S.Y., Murphy, T.L., and Murphy, K.M. (2002). T-bet is a STATI-induced regulator for IL-12R expression in naïve CD4+ T cells. Nat. Immunol. *3*, 549–557. https://doi.org/10.1038/ni794.

Agrawal, P., Kurcon, T., Pilobello, K.T., Rakus, J.F., Koppolu, S., Liu, Z., Batista, B.S., Eng, W.S., Hsu, K.L., Liang, Y., et al. (2014). Mapping posttranscriptional regulation of the human glycome uncovers microRNA defining the glycocode. Proc. Natl. Acad. Sci. U. S. A. *111*, 4338–4343. https://doi.org/10.1073/pnas.1321524111.

Agrawal, P., Fontanals-Cirera, B., Sokolova, E., Jacob, S., Vaiana, C.A., Argibay, D., Davalos, V., McDermott, M., Nayak, S., Darvishian, F., et al. (2017). A Systems Biology Approach Identifies FUT8 as a Driver of Melanoma Metastasis. Cancer Cell *31*, 804-819.e7. https://doi.org/10.1016/j.ccell.2017.05.007.

Alatrash, G., Qiao, N., Zhang, M., Zope, M., Perakis, A.A., Sukhumalchandra, P., Philips, A. V., Garber, H.R., Kerros, C., St John, L.S., et al. (2019). Fucosylation enhances the efficacy of adoptively transferred antigen-specific cytotoxic T lymphocytes. Clin. Cancer Res. *25*, 2610–2620. https://doi.org/10.1158/1078-0432.CCR-18-1527.

Aldinucci, D., and Casagrande, N. (2018). Inhibition of the CCL5/CCR5 axis against the progression of gastric cancer. Int. J. Mol. Sci. *19*. https://doi.org/10.3390/ijms19051477.

Apweiler, R., Hermjakob, H., and Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim. Biophys. Acta - Gen. Subj. *1473*, 4–8. https://doi.org/https://doi.org/10.1016/S0304-4165(99)00165-8.

Araujo, L., Khim, P., Mkhikian, H., Mortales, C.L., and Demetriou, M. (2017). Glycolysis and glutaminolysis cooperatively control T cell function by limiting metabolite supply to N-glycosylation. Elife *6*, 1–16. https://doi.org/10.7554/eLife.21330.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One *10*, 1–46. https://doi.org/10.1371/journal.pone.0130140.

Bellis, S.L. (2004). Variant glycosylation: An underappreciated regulatory mechanism for β1 integrins. Biochim. Biophys. Acta - Biomembr. *1663*, 52–60. https://doi.org/10.1016/j.bbamem.2004.03.012.

Beyer, M., Thabet, Y., Müller, R.U., Sadlon, T., Classen, S., Lahl, K., Basu, S., Zhou, X., Bailey-Bucktrout, S.L., Krebs, W., et al. (2011). Repression of the genome organizer SATB1 in regulatory T cells is required for suppressive function and inhibition of effector differentiation. Nat. Immunol. *12*, 898–907. https://doi.org/10.1038/ni.2084.

Bojar, D., Meche, L., Meng, G., Eng, W., Smith, D.F., Cummings, R.D., and Mahal, L.K. (2022). A Useful Guide to Lectin Binding: Machine-Learning Directed Annotation of 57 Unique Lectin Specificities. ACS Chem. Biol. https://doi.org/10.1021/acschembio.1c00689.

Breen, K.C. (2002). The role of protein glycosylation in the control of cellular N-sialyltransferase activity. FEBS Lett. *517*, 215–218. https://doi.org/10.1016/S0014-5793(02)02626-1.

Bui, T.T., Lee, D., and Selvarajoo, K. (2020). ScatLay: utilizing transcriptome-wide noise for identifying and visualizing differentially expressed genes. Sci. Rep. *10*, 1–11. https://doi.org/10.1038/s41598-020-74564-1.

Cabral, J., Hanley, S.A., Gerlach, J.Q., O'Leary, N., Cunningham, S., Ritter, T., Ceredig, R., Joshi, L., and Griffin, M.D. (2017). Distinctive surface glycosylation patterns associated with mouse and human cD4+ regulatory T cells and their suppressive function. Front. Immunol. *8*. https://doi.org/10.3389/fimmu.2017.00987.

Chen, L., Zhang, W., Fregien, N., and Pierce, M. (1998). The her-2/neu oncogene stimulates the transcription of N-acetylglucosaminyltransferase V and expression of its cell surface oligosaccharide products. Oncogene *17*, 2087–2093. https://doi.org/10.1038/sj.onc.1202124.

Chuzel, L., Fossa, S.L., Boisvert, M.L., Cajic, S., Hennig, R., Ganatra, M.B., Reichl, U., Rapp, E., and Taron, C.H. (2021). Combining functional metagenomics and glycoanalytics to identify enzymes that facilitate structural characterization of sulfated N-glycans. Microb. Cell Fact. *20*, 1–17. https://doi.org/10.1186/s12934-021-01652-w.

Colomb, F., Giron, L.B., Kuri-Cervantes, L., Adeniji, O.S., Ma, T., Dweep, H., Battivelli, E., Verdin, E., Palmer, C.S., Tateno, H., et al. (2020). Sialyl-LewisX Glycoantigen Is Enriched on Cells with Persistent HIV Transcription during Therapy. Cell Rep. *32*. https://doi.org/10.1016/j.celrep.2020.107991.

Corapi, E., Carrizo, G., Compagno, D., and Laderach, D. (2018). Endogenous Galectin-1 in T Lymphocytes Regulates Anti-prostate Cancer Immunity. Front. Immunol. *9*, 1–10. https://doi.org/10.3389/fimmu.2018.02190.

Demetriou, M., Granovsky, M., Quaggin, S., and Dennis, J.W. (2001). Negative regulation of T-cell activation and autoimmunity by Mgat5 N-glycosylation. Nature *409*, 733–739. https://doi.org/10.1038/35055582.

Demus, D., Jansen, B.C., Gardner, R.A., Urbanowicz, P.A., Wu, H., Štambuk, T., Juszczak, A., Medvidović, E.P., Juge, N., Gornik, O., et al. (2021). Interlaboratory evaluation of plasma N-glycan antennary fucosylation as a clinical biomarker for HNF1A-MODY using liquid chromatography methods. Glycoconj. J. *38*, 375–386. https://doi.org/10.1007/s10719-021-09992-w.

Deng, X., Xu, M., Yuan, C., Yin, L., Chen, X., Zhou, X., Li, G., Fu, Y., Feghali-Bostwick, C.A., and Pang, L. (2013). Transcriptional regulation of increased CCL2 expression in pulmonary fibrosis involves nuclear factor-κB and activator protein-1. Int. J. Biochem. Cell Biol. *45*, 1366–1376. https://doi.org/10.1016/j.biocel.2013.04.003.

Dimitroff, C.J. (2019). I-branched carbohydrates as emerging effectors of malignant progression. Proc. Natl. Acad. Sci. U. S. A. *116*, 13729–13737. https://doi.org/10.1073/pnas.1900268116.

Dincer, A.B., Celik, S., Hiranuma, N., and Lee, S.I. (2018). DeepProfile: Deep learning of cancer molecular profiles for precision medicine. BioRxiv 1–13. .

Dobie, C., and Skropeta, D. (2021). Insights into the role of sialylation in cancer progression and metastasis. Br. J. Cancer *124*, 76–90. https://doi.org/10.1038/s41416-020-01126-7.

Van Dyken, S.J., Green, R.S., and Marth, J.D. (2007). Structural and Mechanistic Features of Protein O Glycosylation Linked to CD8 + T-Cell Apoptosis . Mol. Cell. Biol. *27*, 1096–1111. https://doi.org/10.1128/mcb.01750-06.

Fei, L., Ren, X., Yu, H., and Zhan, Y. (2021). Targeting the CCL2/CCR2 Axis in Cancer Immunotherapy: One Stone, Three Birds? Front. Immunol. *12*, 1–20. https://doi.org/10.3389/fimmu.2021.771210.

Ge, S.X., Jung, D., Jung, D., and Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. Bioinformatics *36*, 2628–2629. https://doi.org/10.1093/bioinformatics/btz931.

Granovsky, M., Fata, J., Pawling, J., Muller, W.J., Khokha, R., and Dennis, J.W. (2000). Suppression of tumor growth and metastasis in Mgat5-deficient mice. Nat. Med. *6*, 306–312. https://doi.org/10.1038/73163.

Graydon, C.G., Mohideen, S., and Fowke, K.R. (2021). LAG3's Enigmatic Mechanism of Action. Front. Immunol. *11*, 1–7. https://doi.org/10.3389/fimmu.2020.615317.

Grigorian, A., and Demetriou, M. (2011). Mgat5 Deficiency in T Cells and Experimental Autoimmune Encephalomyelitis . ISRN Neurol. *2011*, 1–6. https://doi.org/10.5402/2011/374314.

Grigorian, A., Mkhikian, H., and Demetriou, M. (2012). Interleukin-2, Interleukin-7, T cell-mediated autoimmunity, and N-glycosylation. Ann. N. Y. Acad. Sci. *1253*, 49–57. https://doi.org/10.1111/j.1749-6632.2011.06391.x.

Grosso, J.F., Kelleher, C.C., Harris, T.J., Maris, C.H., Hipkiss, E.L., De Marzo, A., Anders, R., Netto, G., Getnet, D., Bruno, T.C., et al. (2007). LAG-3 regulates CD8+ T cell accumulation and effector function in murine self- and tumor-tolerance systems. J. Clin. Invest. *117*, 3383–

3392. https://doi.org/10.1172/JCI31184.

Hanczar, B., Zehraoui, F., Issa, T., and Arles, M. (2020). Biological interpretation of deep neural network for phenotype prediction based on gene expression. BMC Bioinformatics *21*, 1–18. https://doi.org/10.1186/s12859-020-03836-4.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Heesch, K., Raczkowski, F., Schumacher, V., Hünemörder, S., Panzer, U., and Mittrücker, H.W. (2014). The function of the chemokine receptor CXCR6 in the T cell response of mice against Listeria monocytogenes. PLoS One *9*. https://doi.org/10.1371/journal.pone.0097701.

Holst, S., Deuss, A.J.M., Van Pelt, G.W., Van Vliet, S.J., Garcia-Vallejo, J.J., Koeleman, C.A.M., Deelder, A.M., Mesker, W.E., Tollenaar, R.A., Rombouts, Y., et al. (2016). N-glycosylation profiling of colorectal cancer cell lines reveals association of fucosylation with differentiation and caudal type homebox 1 (CDX1)/Villin mRNA expression. Mol. Cell. Proteomics *15*, 124–140. https://doi.org/10.1074/mcp.M115.051235.

Hou, X.X., Wang, X.Q., Zhou, W.J., and Li, D.J. (2021). Regulatory T cells induce polarization of pro-repair macrophages by secreting sFGL2 into the endometriotic milieu. Commun. Biol. *4*, 1–16. https://doi.org/10.1038/s42003-021-02018-z.

Huot, N., Rascle, P., Tchitchek, N., Wimmer, B., Passaes, C., Contreras, V., Desjardins, D., Stahl-Hennig, C., Le Grand, R., Saez-Cirion, A., et al. (2021). Role of NKG2a/c+CD8+ T cells in pathogenic versus non-pathogenic SIV infections. IScience *24*. https://doi.org/10.1016/j.isci.2021.102314.

Jankowska, K.I., Williamson, E.K., Roy, N.H., Blumenthal, D., Chandra, V., Baumgart, T., and Burkhardt, J.K. (2018). Integrins modulate T cell receptor signaling by constraining actin flow at the immunological synapse. Front. Immunol. *9*, 1–19. https://doi.org/10.3389/fimmu.2018.00025.

Jia, P., Hu, R., Pei, G., Dai, Y., Wang, Y.Y., and Zhao, Z. (2021). Deep generative neural network for accurate drug response imputation. Nat. Commun. *12*, 1–16. https://doi.org/10.1038/s41467-021-21997-5.

Johannes, L., Jacob, R., and Leffler, H. (2018). Galectins at a glance. J. Cell Sci. *131*, 1–9. https://doi.org/10.1242/jcs.208884.

Kearney, C.J., Vervoort, S.J., Ramsbottom, K.M., Todorovski, I., Lelliott, E.J., Zethoven, M., Pijpers, L., Martin, B.P., Semple, T., Martelotto, L., et al. (2021). SUGAR-seq enables simultaneous detection of glycans, epitopes, and the transcriptome in single cells. Sci. Adv. *7*, 1–13. https://doi.org/10.1126/sciadv.abe3610.

Kellman, B.P., and Lewis, N.E. (2021). Big-Data Glycomics: Tools to Connect Glycan Biosynthesis to Extracellular Communication. Trends Biochem. Sci. *46*, 284–300. https://doi.org/https://doi.org/10.1016/j.tibs.2020.10.004.

Khoder-Agha, F., Harrus, D., Brysbaert, G., Lensink, M.F., Harduin-Lepers, A., Glumoff, T., and Kellokumpu, S. (2019). Assembly of B4GALT1/ST6GAL1 heteromers in the Golgi membranes involves lateral interactions via highly charged surface domains. J. Biol. Chem. *294*, 14383–14393. https://doi.org/10.1074/jbc.RA119.009539.

Kim, C.J., Lee, C.G., Jung, J.Y., Ghosh, A., Hasan, S.N., Hwang, S.M., Kang, H., Lee, C., Kim, G.C., Rudra, D., et al. (2018). The Transcription Factor Ets1 Suppresses T Follicular Helper Type 2 Cell Differentiation to Halt the Onset of Systemic Lupus Erythematosus. Immunity *49*, 1034-1048.e8. https://doi.org/10.1016/j.immuni.2018.10.012.

Kim, G.B., Gao, Y., Palsson, B.O., and Lee, S.Y. (2021). DeepTFactor: A deep learning-based tool for the prediction of transcription factors. Proc. Natl. Acad. Sci. U. S. A. *118*, 1–5. https://doi.org/10.1073/pnas.2021171118.

Ko, J.H., Miyoshi, E., Noda, K., Ekuni, A., Kang, R., Ikeda, Y., and Taniguchi, N. (1999). Regulation of the GnT-V promoter by transcription factor Ets-1 in various cancer cell lines. J. Biol. Chem. *274*, 22941–22948. https://doi.org/10.1074/jbc.274.33.22941.

Kouo, T., Huang, L., Pucsek, A.B., Cao, M., Solt, S., Armstrong, T., and Jaffee, E. (2015). Galectin-3 shapes antitumor immune responses by suppressing CD8 T Cells via LAG-3 and

Inhibiting Expansion of Plasmacytoid Dendritic Cells. Cancer Immunol. Res. *3*, 412–423. https://doi.org/10.1158/2326-6066.CIR-14-0150.

Kumar, B. V., Connors, T.J., and Farber, D.L. (2018). Human T Cell Development, Localization, and Function throughout Life. Immunity *48*, 202–213. https://doi.org/10.1016/j.immuni.2018.01.007.

Kurz, E., Chen, S., Vucic, E., Baptiste, G., Loomis, C., Agrawal, P., Hajdu, C., Bar-Sagi, D., and Mahal, L.K. (2021). Integrated Systems Analysis of the Murine and Human Pancreatic Cancer Glycomes Reveals a Tumor-Promoting Role for ST6GAL1. Mol. Cell. Proteomics *20*, 100160. https://doi.org/10.1016/j.mcpro.2021.100160.

Lau, K.S., Partridge, E.A., Grigorian, A., Silvescu, C.I., Reinhold, V.N., Demetriou, M., and Dennis, J.W. (2007). Complex N-Glycan Number and Degree of Branching Cooperate to Regulate Cell Proliferation and Differentiation. Cell *129*, 123–134. https://doi.org/10.1016/j.cell.2007.01.049.

Lee, S.U., Grigorian, A., Pawling, J., Chen, I.J., Gao, G., Mozaffar, T., McKerlie, C., and Demetriou, M. (2007). N-glycan processing deficiency promotes spontaneous inflammatory demyelination and neurodegeneration. J. Biol. Chem. *282*, 33725–33734. https://doi.org/10.1074/jbc.M704839200.

Lehmkuhl, P., Gentz, M., Garcia de Otezya, A.C., Grimbacher, B., Schulze-Koops, H., and Skapenko, A. (2021). Dysregulated immunity in PID patients with low GARP expression on Tregs due to mutations in LRRC32. Cell. Mol. Immunol. *18*, 1677–1691. https://doi.org/10.1038/s41423-021-00701-z.

de Leoz, M.L.A., Young, L.J.T., An, H.J., Kronewitter, S.R., Kim, J., Miyamoto, S., Borowsky, A.D., Chew, H.K., and Lebrilla, C.B. (2011). High-Mannose Glycans are Elevated during Breast Cancer Progression. Mol. Cell. Proteomics *10*, M110.002717. https://doi.org/10.1074/mcp.m110.002717.

Li, L., Wan, S., Tao, K., Wang, G., and Zhao, E. (2016). KLRG1 restricts memory T cell antitumor immunity. Oncotarget *7*, 61670–61678. https://doi.org/10.18632/oncotarget.11430.

Li, S., Mo, C., Peng, Q., Kang, X., Sun, C., Jiang, K., Huang, L., Lu, Y., Sui, J., Qin, X., et al. (2013). Cell Surface Glycan Alterations in Epithelial Mesenchymal Transition Process of Huh7 Hepatocellular Carcinoma Cell. PLoS One *8*. https://doi.org/10.1371/journal.pone.0071273.

Lin, Y.M., Hsu, C.J., Liao, Y.Y., Chou, M.C., and Tang, C.H. (2012). The CCL2/CCR2 Axis Enhances Vascular Cell Adhesion Molecule-1 Expression in Human Synovial Fibroblasts. PLoS One *7*, 1–10. https://doi.org/10.1371/journal.pone.0049999.

Liu, K., Tan, S., Jin, W., Guan, J., Wang, Q., Sun, H., Qi, J., Yan, J., Chai, Y., Wang, Z., et al. (2020). N-glycosylation of PD-1 promotes binding of camrelizumab. EMBO Rep. *21*, 1–12. https://doi.org/10.15252/embr.202051444.

Loke, I., Kolarich, D., Packer, N.H., and Thaysen-Andersen, M. (2016). Emerging roles of protein mannosylation in inflammation and infection. Mol. Aspects Med. *51*, 31–55. https://doi.org/https://doi.org/10.1016/j.mam.2016.04.004.

Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *2017-Decem*, 4766–4775. .

Matsuo, K., Yoshie, O., and Nakayama, T. (2021). Multifaceted roles of chemokines and chemokine receptors in tumor immunity. Cancers (Basel). *13*. https://doi.org/10.3390/cancers13236132.

Meixner, A., Karreth, F., Kenner, L., and Wagner, E.F. (2004). JunD regulates lymphocyte proliferation and T helper cell cytokine expression. EMBO J. *23*, 1325–1335. https://doi.org/10.1038/sj.emboj.7600133.

Mimura, Y., Katoh, T., Saldova, R., O'Flaherty, R., Izumi, T., Mimura-Kimura, Y., Utsunomiya, T., Mizukami, Y., Yamamoto, K., Matsumoto, T., et al. (2018). Glycosylation engineering of therapeutic IgG antibodies: challenges for the safety, functionality and efficacy. Protein Cell *9*, 47–62. https://doi.org/10.1007/s13238-017-0433-3.

Mkhikian, H., Grigorian, A., Li, C.F., Chen, H.L., Newton, B., Zhou, R.W., Beeton, C., Torossian, S., Tatarian, G.G., Lee, S.U., et al. (2011). Genetics and the environment converge to dysregulate N-glycosylation in multiple sclerosis. Nat. Commun. *2*, 313–334. https://doi.org/10.1038/ncomms1333.

Morgan, R., Gao, G., Pawling, J., Dennis, J.W., Demetriou, M., and Li, B. (2004). N - Acetylglucosaminyltransferase V (Mgat5)-Mediated N -Glycosylation Negatively Regulates Th1 Cytokine Production by T Cells . J. Immunol. *173*, 7200–7208. https://doi.org/10.4049/jimmunol.173.12.7200.

Mortales, C.L., Lee, S.U., Manousadjian, A., Hayama, K.L., and Demetriou, M. (2020). N-Glycan Branching Decouples B Cell Innate and Adaptive Immunity to Control Inflammatory Demyelination. IScience *23*, 101380. https://doi.org/10.1016/j.isci.2020.101380.

Nairn, A. V., York, W.S., Harris, K., Hall, E.M., Pierce, J.M., and Moremen, K.W. (2008). Regulation of glycan structures in animal tissues: Transcript profiling of glycan-related genes. J. Biol. Chem. *283*, 17298–17313. https://doi.org/10.1074/jbc.M801964200.

Natoni, A., Macauley, M.S., and O'Dwyer, M.E. (2016). Targeting selectins and their ligands in cancer. Front. Oncol. *6*, 1–12. https://doi.org/10.3389/fonc.2016.00093.

Neelamegham, S., and Mahal, L.K. (2016). Multi-level regulation of cellular glycosylation: from genes to transcript to enzyme to structure. Curr. Opin. Struct. Biol. *40*, 145–152. https://doi.org/https://doi.org/10.1016/j.sbi.2016.09.013.

Ng, B.G., and Freeze, H.H. (2018). Perspectives on Glycosylation and Its Congenital Disorders. Trends Genet. *34*, 466–476. https://doi.org/10.1016/j.tig.2018.03.002.

Nguyen, N.T.B., Lin, J., Tay, S.J., Mariati, Yeo, J., Nguyen-Khuong, T., and Yang, Y. (2021). Multiplexed engineering glycosyltransferase genes in CHO cells via targeted integration for producing antibodies with diverse complex-type N-glycans. Sci. Rep. *11*, 1–16. https://doi.org/10.1038/s41598-021-92320-x.

Novoszel, P., Holcmann, M., Stulnig, G., De Sa Fernandes, C., Zyulina, V., Borek, I., Linder, M., Bogusch, A., Drobits, B., Bauer, T., et al. (2021). Psoriatic skin inflammation is promoted by c-Jun/AP-1-dependent CCL2 and IL-23 expression in dendritic cells. EMBO Mol. Med. *13*, 1–18. https://doi.org/10.15252/emmm.202012409.

Nüssing, S., Koay, H.F., Sant, S., Loudovaris, T., Mannering, S.I., Lappas, M., d'Udekem, Y., Konstantinov, I.E., Berzins, S.P., Rimmelzwaan, G.F., et al. (2019). Divergent SATB1 expression across human life span and tissue compartments. Immunol. Cell Biol. *97*, 498–511. https://doi.org/10.1111/imcb.12233.

Oinam, L., Minoshima, F., and Tateno, H. (2021). Glycomic profiling of the gut microbiota by Glycan-seq.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct *4*, 1–10. https://doi.org/10.1186/1745-6150-4-14.

Pace, K.E., Hahn, H.P., Pang, M., Nguyen, J.T., and Baum, L.G. (2000). Cutting Edge: CD7 Delivers a Pro-Apoptotic Signal During Galectin-1-Induced T Cell Death. J. Immunol. *165*, 2331 LP – 2334. https://doi.org/10.4049/jimmunol.165.5.2331.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., et al. (2012). Scikit-learn: Machine Learning in Python. ArXiv.

Petit, D., Mir, A.M., Petit, J.M., Thisse, C., Delannoy, P., Oriol, R., Thisse, B., and Harduin-Lepers, A. (2010). Molecular phylogeny and functional genomics of β-galactoside α2,6-sialyltransferases that explain ubiquitous expression of st6gal1 gene in amniotes. J. Biol. Chem. *285*, 38399–38414. https://doi.org/10.1074/jbc.M110.163931.

Pietrobono, S., and Stecca, B. (2021). Aberrant sialylation in cancer: Biomarker and potential target for therapeutic intervention? Cancers (Basel). *13*. https://doi.org/10.3390/cancers13092014.

Pinho, S.S., and Reis, C.A. (2015). Glycosylation in cancer: Mechanisms and clinical implications. Nat. Rev. Cancer *15*, 540–555. https://doi.org/10.1038/nrc3982.

Priatel, J.J., Chui, D., Hiraoka, N., Simmons, C.J.T., Richardson, K.B., Page, D.M., Fukuda, M., Varki, N.M., and Marth, J.D. (2000). The ST3Gal-I sialyltransferase controls CD8+ T lymphocyte homeostasis by modulating O-glycan biosynthesis. Immunity *12*, 273–283. https://doi.org/10.1016/S1074-7613(00)80180-6.

Przybyło, M., Martuszewska, D., Pocheć, E., Hoja-Łukowicz, D., and Lityńska, A. (2007). Identification of proteins bearing β1-6 branched N-glycans in human melanoma cell lines from different progression stages by tandem mass spectrometry analysis. Biochim. Biophys. Acta - Gen. Subj. *1770*, 1427–1435. https://doi.org/10.1016/j.bbagen.2007.05.006.

Qin, R., and Mahal, L.K. (2021). The host glycomic response to pathogens. Curr. Opin. Struct. Biol. *68*, 149–156. https://doi.org/https://doi.org/10.1016/j.sbi.2020.12.011.

Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. Nat. Commun. *11*, 1–9. https://doi.org/10.1038/s41467-020-14976-9.

Reily, C., Stewart, T.J., Renfrow, M.B., and Novak, J. (2019). Glycosylation in health and disease. Nat. Rev. Nephrol. *15*, 346–366. https://doi.org/10.1038/s41581-019-0129-4.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess. 97–101. https://doi.org/10.18653/v1/n16-3020.

Ricciardiello, F., Votta, G., Palorini, R., Raccagni, I., Brunelli, L., Paiotta, A., Tinelli, F., D'Orazio, G., Valtorta, S., De Gioia, L., et al. (2018). Inhibition of the Hexosamine Biosynthetic Pathway by targeting PGM3 causes breast cancer growth arrest and apoptosis. Cell Death Dis. *9*. https://doi.org/10.1038/s41419-018-0405-4.

Rodriguez, E., Boelaars, K., Brown, K., Eveline Li, R.J., Kruijssen, L., Bruijns, S.C.M., van Ee, T., Schetters, S.T.T., Crommentuijn, M.H.W., van der Horst, J.C., et al. (2021). Sialic acids in pancreatic cancer cells drive tumour-associated macrophage differentiation via the Siglec receptors Siglec-7 and Siglec-9. Nat. Commun. *12*, 1–14. https://doi.org/10.1038/s41467-021-21550-4.

Rowshanravan, B., Halliday, N., and Sansom, D.M. (2018). CTLA-4: A moving target in immunotherapy. Blood *131*, 58–67. https://doi.org/10.1182/blood-2017-06-741033.

Sackstein, R., Schatton, T., and Barthel, S.R. (2017). T-lymphocyte homing: An underappreciated yet critical hurdle for successful cancer immunotherapy. Lab. Investig. *97*, 669–697. https://doi.org/10.1038/labinvest.2017.25.

Schmiedel, J., Klemm, S., Yannan, Z., Apratim, S., Nils, B., S., M.D., and Alexander, van O. (2015). MicroRNA control of protein expression noise. Science (80-.). *348*, 128–132. https://doi.org/10.1126/science.aaa1738.

Sempowski, G.D., Cross, S.J., Heinly, C.S., Scearce, R.M., and Haynes, B.F. (2004). CD7 and CD28 Are Required for Murine CD4 + CD25 + Regulatory T Cell Homeostasis and Prevention of Thyroiditis . J. Immunol. *172*, 787–794. https://doi.org/10.4049/jimmunol.172.2.787.

Sheu, B.C., Chiou, S.H., Lin, H.H., Chow, S.N., Huang, S.C., Ho, H.N., and Hsu, S.M. (2005). Up-regulation of inhibitory natural killer receptors CD94/NKG2A with suppressed intracellular perforin expression of tumor-infiltrating CD8 + T lymphocytes in human cervical carcinoma. Cancer Res. *65*, 2921–2929. https://doi.org/10.1158/0008-5472.CAN-04-2108.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. 34th Int. Conf. Mach. Learn. ICML 2017 *7*, 4844–4866. .

Silva, M.C., Fernandes, Â., Oliveira, M., Resende, C., Correia, A., de-Freitas-Junior, J.C., Lavelle, A., Andrade-Da-Costa, J., Leander, M., Xavier-Ferreira, H., et al. (2020). Glycans as immune checkpoints: Removal of branched N-glycans enhances immune recognition preventing cancer progression. Cancer Immunol. Res. *8*, 1407–1425. https://doi.org/10.1158/2326-6066.CIR-20-0264.

Smith, A.M., Walsh, J.R., Long, J., Davis, C.B., Henstock, P., Hodge, M.R., Maciejewski, M., Mu, X.J., Ra, S., Zhao, S., et al. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC Bioinformatics *21*, 1–18. https://doi.org/10.1186/s12859-020-3427-8.

Smith, L.K., Boukhaled, G.M., Condotta, S.A., Mazouz, S., Guthmiller, J.J., Vijay, R., Butler, N.S., Bruneau, J., Shoukry, N.H., Krawczyk, C.M., et al. (2018). Interleukin-10 Directly Inhibits CD8+ T Cell Function by Enhancing N-Glycan Branching to Decrease Antigen

Sensitivity. Immunity *48*, 299-312.e5. https://doi.org/10.1016/j.immuni.2018.01.006.

Sobhani, N., Tardiel-Cyril, D.R., Davtyan, A., Generali, D., Roudi, R., and Li, Y. (2021). CTLA-4 in regulatory T cells for cancer immunotherapy. Cancers (Basel). *13*, 1–18. https://doi.org/10.3390/cancers13061440.

Song, W., Isaji, T., Nakano, M., Liang, C., Fukuda, T., and Gu, J. (2022). O-GlcNAcylation regulates β1,4-GlcNAc-branched N-glycan biosynthesis via the OGT/SLC35A3/GnT-IV axis. FASEB J. *36*, 1–16. https://doi.org/10.1096/fj.202101520R.

Stephen, T.L., Payne, K.K., Chaurio, R.A., Allegrezza, M.J., Zhu, H., Perez-Sanz, J., Perales-Puchalt, A., Nguyen, J.M., Vara-Ailor, A.E., Eruslanov, E.B., et al. (2017). SATB1 Expression Governs Epigenetic Repression of PD-1 in Tumor-Reactive T Cells. Immunity *46*, 51–64. https://doi.org/10.1016/j.immuni.2016.12.015.

Stowell, S.R., Arthur, C.M., Mehta, P., Slanina, K.A., Blixt, O., Leffler, H., Smith, D.F., and Cummings, R.D. (2008). Galectin-1, -2, and -3 exhibit differential recognition of sialylated glycans and blood group antigens. J. Biol. Chem. *283*, 10109–10123. https://doi.org/10.1074/jbc.M709545200.

Sun, L., Konstantinidi, A., Ye, Z., Nason, R., Zhang, Y., Büll, C., Kahl-Knutson, B., Hansen, L., Leffler, H., Vakhrushev, S.Y., et al. (2022). Installation of O-glycan sulfation capacities in human HEK293 cells for display of sulfated mucins. J. Biol. Chem. *298*, 1–14. https://doi.org/10.1016/j.jbc.2021.101382.

Sun, Q., Zhang, Z., Zhang, H., and Liu, X. (2016). Aberrant IgA1 glycosylation in iga nephropathy: A systematic review. PLoS One *11*, 1–15. https://doi.org/10.1371/journal.pone.0166700.

Szabo, P.A., Levitin, H.M., Miron, M., Snyder, M.E., Senda, T., Yuan, J., Cheng, Y.L., Bush, E.C., Dogra, P., Thapa, P., et al. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat. Commun. *10*. https://doi.org/10.1038/s41467-019-12464-3.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613. https://doi.org/10.1093/nar/gky1131.

Takahashi, R., Nishimoto, S., Muto, G., Sekiya, T., Tamiya, T., Kimura, A., Morita, R., Asakawa, M., Chinen, T., and Yoshimura, A. (2011). SOCS1 is essential for regulatory T cell functions by preventing loss of Foxp3 expression as well as IFN-γ and IL-17A production. J. Exp. Med. *208*, 2055–2067. https://doi.org/10.1084/jem.20110428.

Tao, S.C., Li, Y., Zhou, J., Qian, J., Schnaar, R.L., Zhang, Y., Goldstein, I.J., Zhu, H., and Schneck, J.P. (2008). Lectin microarrays identify cell-specific and functionally significant cell surface glycan markers. Glycobiology *18*, 761–769. https://doi.org/10.1093/glycob/cwn063.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. Genome Res. *21*, 2213–2223. https://doi.org/10.1101/gr.124321.111.

Taylor, R.P., Geisler, T.S., Chambers, J.H., and McClain, D.A. (2009). Up-regulation of O-GlcNAc transferase with glucose deprivation in HepG2 cells is mediated by decreased hexosamine pathway flux. J. Biol. Chem. *284*, 3425–3432. https://doi.org/10.1074/jbc.M803198200.

Theodoratou, E., Campbell, H., Ventham, N.T., Kolarich, D., Pučić-Baković, M., Zoldoš, V., Fernandes, D., Pemberton, I.K., Rudan, I., Kennedy, N.A., et al. (2014). The role of glycosylation in IBD. Nat. Rev. Gastroenterol. Hepatol. *11*, 588–600. https://doi.org/10.1038/nrgastro.2014.78.

Thu, C.T., and Mahal, L.K. (2020). Sweet Control: MicroRNA Regulation of the Glycome. Biochemistry *59*, 3098–3110. https://doi.org/10.1021/acs.biochem.9b00784.

Tran, D.Q., Andersson, J., Wang, R., Ramsey, H., Unutmaz, D., and Shevach, E.M. (2009). GARP (LRRC32) is essential for the surface expression of latent TGF-β on platelets and activated FOXP3+ regulatory T cells. Proc. Natl. Acad. Sci. U. S. A. *106*, 13445–13450.

https://doi.org/10.1073/pnas.0901944106.

Tu, M.M., Abdel-Hafiz, H.A., Jones, R.T., Jean, A., Hoff, K.J., Duex, J.E., Chauca-Diaz, A., Costello, J.C., Dancik, G.M., Tamburini, B.A.J., et al. (2020). Inhibition of the CCL2 receptor, CCR2, enhances tumor response to immune checkpoint therapy. Commun. Biol. *3*, 1–12. https://doi.org/10.1038/s42003-020-01441-y.

Varki, A. (2017). Biological roles of glycans. Glycobiology *27*, 3–49. https://doi.org/10.1093/glycob/cww086.

Vosseller, K., Wells, L., Lane, M.D., and Hart, G.W. (2002). Elevated nucleocytoplasmic glycosylation by O-GlcNAc results in insulin resistance associated with defects in Akt activation in 3T3-L1 adipocytes. Proc. Natl. Acad. Sci. U. S. A. *99*, 5313–5318. https://doi.org/10.1073/pnas.072072399.

Wang, R., Fan, Q., Zhang, J., Zhang, X., Kang, Y., and Wang, Z. (2018). Hydrogen Sulfide Demonstrates Promising Antitumor Efficacy in Gastric Carcinoma by Targeting MGAT5. Transl. Oncol. *11*, 900–910. https://doi.org/10.1016/j.tranon.2018.04.008.

Wu, D., Li, J., Struwe, W.B., and Robinson, C. V. (2019). Probing: N -glycoprotein microheterogeneity by lectin affinity purification-mass spectrometry analysis. Chem. Sci. *10*, 5146–5155. https://doi.org/10.1039/c9sc00360f.

Yang, R., Sun, L., Li, C.F., Wang, Y.H., Yao, J., Li, H., Yan, M., Chang, W.C., Hsu, J.M., Cha, J.H., et al. (2021). Galectin-9 interacts with PD-1 and TIM-3 to regulate T cell death and is a target for cancer immunotherapy. Nat. Commun. *12*. https://doi.org/10.1038/s41467-021-21099-2.

Yang, W., Rosenstiel, P., and Schulenburg, H. (2019). AFold - Using polynomial uncertainty modelling for differential gene expression estimation from RNA sequencing data. BMC Genomics *20*, 1–17. https://doi.org/10.1186/s12864-019-5686-1.

Yap, M., Johnston, R.L., Foley, H., MacDonald, S., Kondrashova, O., Tran, K.A., Nones, K., Koufariotis, L.T., Bean, C., Pearson, J. V., et al. (2021). Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. Sci. Rep. *11*, 1–12.

https://doi.org/10.1038/s41598-021-81773-9.

Ye, L., Park, J.J., Dong, M.B., Yang, Q., Chow, R.D., Peng, L., Du, Y., Guo, J., Dai, X., Wang, G., et al. (2019). In vivo CRISPR screening in CD8 T cells with AAV–Sleeping Beauty hybrid vectors identifies membrane targets for improving immunotherapy for glioblastoma. Nat. Biotechnol. *37*, 1302–1313. https://doi.org/10.1038/s41587-019-0246-4.

Zeng, Z., Lan, T., Wei, Y., and Wei, X. (2022). CCL5/CCR5 axis in human diseases and related treatments. Genes Dis. *9*, 12–27. https://doi.org/10.1016/j.gendis.2021.08.004.

Zhang, J., ten Dijke, P., Wuhrer, M., and Zhang, T. (2021). Role of glycosylation in TGF-β signaling and epithelial-to-mesenchymal transition in cancer. Protein Cell *12*, 89–106. https://doi.org/10.1007/s13238-020-00741-7.

Zheng, J., Chan, P.L., Liu, Y., Qin, G., Xiang, Z., Lam, K.T., Lewis, D.B., Lau, Y.L., and Tu, W. (2013). ICOS regulates the generation and function of human CD4+ Treg in a CTLA-4 dependent manner. PLoS One *8*, 1–11. https://doi.org/10.1371/journal.pone.0082203.

Zheng, X., Fu, X., Wang, K., and Wang, M. (2020). Deep neural networks for human microRNA precursor detection. BMC Bioinformatics *21*, 1–7. https://doi.org/10.1186/s12859-020-3339-7.

Zhou, R.W., Mkhikian, H., Grigorian, A., Hong, A., Chen, D., Arakelyan, A., and Demetriou, M. (2014). N-glycosylation bidirectionally extends the boundaries of thymocyte positive selection by decoupling Lck from Ca2+ signaling. Nat. Immunol. *15*, 1038–1045. https://doi.org/10.1038/ni.3007.

Zhu, L., Guo, Q., Guo, H., Liu, T., Zheng, Y., Gu, P., Chen, X., Wang, H., Hou, S., and Guo, Y. (2014). Versatile characterization of glycosylation modification in CTLA4-Ig fusion proteins by liquid chromatography-mass spectrometry. MAbs *6*, 1474–1485. https://doi.org/10.4161/mabs.36313.

Zhuo, Y., Chammas, R., and Bellis, S.L. (2008). Sialylation of 1 Integrins Blocks Cell Adhesion to Galectin-3 and Protects Cells against Galectin-3-induced Apoptosis. J. Biol. Chem. *283*, 22177–22185. https://doi.org/10.1074/jbc.m8000015200.