SUPPLEMENTAL FIGURE AND TABLE LEGENDS

SUPPLEMENTAL FIGURES LEGENDS

S1. Immunocytology of SimpleCell Lines

The panels illustrate immunofluorescence staining of SimpleCell and corresponding WT cell lines with monoclonal antibodies (MAbs) 5F4 (anti-Tn), 3F1 (anti-STn), 3C9 after neuraminidase (anti-T) and 2G8 (anti-C1GalT1). Note that MDA-231 and T47D SimpleCells showed some reactivity with 3C9 even though a strong staining was observed for Tn and no staining was seen for C1GalT1. PCR sequencing did not show any WT alleles in these clones, and it is currently unclear what 3C9 is reacting with.

S2. Protein Coverage

The number of glycopeptides identified per protein listed as percentage of the total number of glycoproteins.

S3. WebLogo Plots of S and T O-Glycosites Without Adjacent PST Residues as well as all S/T O-Glycosites together.

A/B) WebLogo plots based on 74 glycosites without S, T, or P within 5 amino acids N- or C-terminal from the site.

C) WebLogo plots based upon all the SimpleCell glycosites.

S4. Validation Process for NetOGlyc4.0 Predictor

In order to better characterize the performance of the new predictor, a 4-fold cross-validation procedure was undertaken.

A) To ensure independence between training, testing and independent testing segments of the data, CD-HIT is used to reduce redundancy. Sequence windows are generated around each of the sites in the data sets, and then clustered against each other. Removing sites with windows that cluster together in other sets, one can iteratively remove redundancy between pairs of sets. Concretely, starting with SimpleCell (α) as a base, the negative data are made independent with respect to α (β-α). Similarly, the subset of O-GLYCBASE used to train NetOGlyc3.1 (referred to from here as Julenius) (Julenius et al., 2005) (δ) data are made independent with respect to α (δ-α). Finally, the UniProt mammalian dataset (γ) is made independent firstly with respect to α (γ-α) and then to (δ-α) (γ-δ-α).

B) Surrounding the data redundancy reduction, a structure for four-fold cross-validation was created. SimpleCell data were used primarily as training data, whilst UniProt mammalian (γ-α in A) as well as 20% of the SimpleCell negative data were used as an independent testing set. The figure illustrates the flow and segmentation of data regarding the number of sites and proteins used at each stage of the validation. Given the random nature of some of the sub-setting, the 4-fold cross-validation procedure is repeated three times, and the average MCC calculated for both cross-validation testing, as well as independent testing.

C) Similar to B, a more complex cross-validation is performed to evaluate if NetOGlyc4.0 has a baseline performance at least equivalent to NetOGlyc3.1. Positive and negative sites from the Julenius training set (set δ in A) are injected

into the cross-validation process. The 3 subset segmentation of the Julenius training set is maintained when performing the cross validation. Here, the independent training set again is the UniProt mammalian data set, but it has also been made independent with respect to Julenius (set γ-δ-α in A).

SUPPLEMENTAL TABLE LEGENDS

Table S1. ZFN-mediated *COSMC* Inactivation in SimpleCells
Deletions and insertions in SimpleCells detected by PCR sequencing.

Table S2. List of O-Glycoproteins, Glycopeptides and Glycosites Identified in SimpleCells (excel file)
Sheet (A) Glycoproteins identified in SimpleCells (+ indicates which cell lines each protein was identified in; chymotrypsin data (Capan-1 and HEK293) are listed in separate columns); (B) identified glycopeptide sequences listed by cell line (sequences shared to 100% identity with bovine are marked with an asterisk); (C) complete list of glycosites identified in each protein (residues are numbered according to SwissProt reviewed full length protein sequence including signal peptide; + indicates possible O-GlcNAc site); (D and E) list divided into unambiguously and ambiguously assigned sites, respectively, with corresponding cell lines; (F) summary list of degenerate peptides (shared between more than one protein); (G) list of identified tyrosine-O-GalNAc sites. Sheets H and I are summaries of all GalNAc sites and potential GlcNAc sites, respectively.

Table S3. GalNAc-T Expression Analysis in SimpleCells by Immunocytology
The expression of 12 GalNAc-Ts (T1-6, T11, T12, T14 and T16) was tested in the 12 SimpleCell lines by immunofluorescence using specific MAbs and mono-specific polyclonal antibodies in the form of murine eyebleeds. The staining was scored from negative (-) to very strong (+++). *MAb reviewed in Bennett et al., 2012, ** unpublished mAb, *** T10 eyebleed was positive in a lung cancer cell line (A549) and the T13 eyebleed was positive in a bone marrow derived cell line (BM).

Table S4. GalNAc-T Isoform Substrate Specificities (excel file)
The Table contains sequences of the 181 peptide substrates used for in vitro glycosylation along with results for the individual GalNAc-Ts. Numbers indicate the number of GalNAc residues added to each peptide.

Supplemental Table S1

ZFN mediated COSMC deletions in cell lines

```
WT COSMC        : AATGACTTATCACCCCAACCAGGTAGTAGAAGGCTGTTGTTCAAATATGGCTGTTACTTTTAA

HaCat           : AATGACTTATCACCCCAACCAGGTAG          TTGTTCAAATATGGCTGTTACTTTTAA

HeLa*^          : AATGACTTATCACCCCAACCAGGTAGTAGAAG CTGTTGTTCAGATATGGCTGTTACTTTTAA

Ovcar3^         :  AATGACTTATCACCCCAACCAGGTAGT          TGTTCAGATATGGCTGTTACTTTTAA

T47D^           : AATGACTTATCACCCCAACCAGGTAGT        GTTGTTCAGATATGGCTGTTACTTTTAA

IMR             : AATGACTTATCACCCCAACCAGGTAGTAG          TTCAGATATGGCTGTTACTTTTAA
                                                &
                  AATGACTTATCACCCCAACCAGGTAGTAG       TTGTTCAGATATGGCTGTTACTTTTAA

HEK293*^        : AATGACTTATCACCCCAACCAGGTAGTAT    CTGTTGTTCAAATATGGCTGTTACTTTTAA
                                                &
                  AATGACTTATCACCCCAACCAGGTAGTAGAAGGC       CAAATATGGCTGTTACTTTTAA

MDA^(89bp insert) :AATGACTTATCACCCCAACCAGGTAGTAGAAGGCTGTTGTTCAGATATGGCTGTTACTTTTAA
                  (ccggagaggctattcggctatgactgggcacaacagacaatcggctgctctgatgccgccgtgttcc
                   ggctgtcagcgcaggggcgccc)

MCF-7(5bp insert) : AATGACTTATCACCCCAACCAGGTAGTAGAAGGCTGTTGTTCAGATATGGCTGTTACTTTTAA
                                                    (aaggc)
                                                &
                                    -91bp deletion-            TCAGATATGGCTGTT
```

ZFN binding sites in wt sequence are underlined.

Positions of insertion are indicated in between italicized bases shown in bold, inserted sequences are parenthesized in lower case letters.

*Predominant alleles detected.

^Larger ZFN mediated deletions beyond the site of analysis remain undetected.

Supplemental Table S3

mAb / Cell Line	T1	T2	T3	T4	T5	T6	T11	T12	T14	T16	T10	T13
	4D8*	4C4*	2D10*	4G2*	5F11**	2F3*	1B2*	1F9*	3D2*	4C6-D10**	EB***	EB***
K562	+++	+++	-	+	+++	++	+	+	+	-	(+)	-
COLO-205	++	++	+++	+++	+++	+	-	+	-	-	-	-
Capan-1	++	+++	+++	-	+	+	-	(+)	-	-	-	-
HepG2	++	+++	-	(+)	-	-	-	-	-	-	(+)	-
OVCAR-3	+	+++	+++	-	-	-	-	-	-	-	ND	ND
T47D	++	+++	+++	+	(+)	++	+	(+)	-	-	ND	ND
MCF-7	++	+++	+++	++	-	+++	(+)	+	+	+	-	-
HeLa	+++	+++	++	-	-	-	-	-	-	-	+++	-
HaCaT	++	+++	+++	-	++	ND	-	(+)	-	-	-	-
MDA-MB-231	+++	+++	+++	+++	+++	+	+	+	-	-	-	-
IMR-32	++	+++	+++	++	++	(+)	-	-	-	-	-	-
HEK293	+	+++	+++	-	-	-	+	(+)	-	++	-	-

- negative
(+) very weak
+ weak
++Positive
+++very strong
ND not determined
EB eyebleeds of mice immunized with recombinant GalNAc-Ts exhibiting specific IgG reactivity with recombinant GalNAc-T10 or T13 expressed in insect cells and not other GalNAc-T isoforms.

EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture and SimpleCell Generation

Knock-out of *COSMC* was achieved using a ZFN-targeting construct (Sigma-Aldrich) with the presented binding and (cutting) site; 5'-CCCAACCAGGTAGT(AGAAGGCT)GTTGTTCAGATATGGCTGTT-3'. Human cell lines, including three breast cancer cell lines (non-metastatic ductal carcinoma T47D, adenocarcinoma MCF-7, and metastatic adenocarcinoma MDA-MB-231), kidney (embryonic kidney HEK293), ovary (adenocarcinoma OVCAR-3), cervix (adenocarcinoma HeLa), neuroblastoma (IMR32), and skin (immortalized keratinocyte HaCat), were transfected with 4 μg of compoZr® C1GalT1C1 DNA using an Amaxa^{TM} Nucleofector^{TM} according to the manufacture's cell lines specific protocols (Lonza). Clones were screened on acetone-fixed slides by immunocytology using monoclonal antibodies to Tn (5F4 and 1E3), STn (3F1 and TKH2), T (3C9 and HH8) and C1GalT1 enzyme (5B6 or2G8) with and without pretreatment with neuraminidase (Sigma) as previously described (Steentoft et al., 2011; Steentoft et al., 2013). PCR sequencing was performed on selected clones with the following primer pair; 5'-AGGGAGGGATGATTTGGAAG-3' and 5'-TTGTCAGAACCATTTGGAGGT-3'. Fluorescence microscopy was performed using a Zeiss Axioskop 2 plus. Staining for GalNAc-Ts was performed on acetone-fixed slides with the MAbs listed in Table S3. All cell lines were routinely screened with the MycoAlertTM kit (Lonza), and a positive result was observed for T47D WT and also in the T47D SimpleCell clone.

LWAC Isolation of Tn O-glycopeptides From Total Cell Lysates

Enrichment of O-glycopeptides by lectin weak affinity chromatography with immobilized *Vicia villosa* agglutinin (VVA-LWAC) was performed as previously described (Steentoft et al., 2011; Steentoft et al., 2013). In brief, media were removed (and saved, see following section) and cells were harvested from two T175 (175 cm^2) flasks by scraping after a washing step with PBS. Cell pellets were lysed in 2 ml 0.05 % RapiGest (Waters) and sonicated. The lysate was reduced, alkylated with iodoacetamide, and digested ON with trypsin or chymotrypsin (20 μg), and then for two hours more with additional (5 μg) protease. The digest was treated with acid (conc. TFA), purified by C18 Sep-Pak chromatography (Waters), and concentrated by Speedvac. The digest was then treated with neuraminidase (only for cells expressing STn), diluted in 2 ml LAC A buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 M Urea, 1 mM $CaCl_2$, $MgCl_2$, $MnCl_2$, and $ZnCl_2$) and injected onto a pre-equlibrated 2.6 m long VVA agarose (Vector Laboratories) column, similar to the system described previously (Chalkley et al., 2009; Darula & Medzihradszky, 2009; Steentoft et al., 2011; Schjoldager et al., 2012; Steentoft et al., 2013). The flow was set to 100 μl min^{-1} and 1000 μl fractions were collected. The column was washed with 0.4 M glucose in LAC A buffer and eluted with 2 × 2 ml 0.2 M GalNAc and 1x 2 ml 0.4 GalNAc.

Two-stage LWAC Enrichment of O-glycopeptides from Culture Medium

Isolation of secreted O-glycoproteins was carried out as previously described (Schjoldager et al., 2012)(Vakhrushev et al., submitted) with some modifications. A sample of SimpleCell culture supernatant (~70 mL), reserved as noted above, was cleared by centrifugation and dialyzed. The dialysis retentate was treated with neuraminidase (5 U *Clostridium perfringens* neuramidase Type VI (Sigma)) if necessary and loaded twice onto a short 0.3 ml VVA agarose (Vector laboratories) column in order to enrich the glycoproteins prior to digestion. The column was washed with 10 CV (3ml) of 0.4 M Glucose in LAC A buffer followed by 1 ml 50 mM AmBic. The glycoproteins were then eluted by 4x 500 µl 0.05% RapiGest with heating to 90°C for 10 min. Elution with RapiGest was performed for secretomes from MDA-213, MCF7, IMR32, HEK293, HeLa and HaCat SimpleCell lines. Secretomes from SimpleCell lines published previously (K562, Capan-1, COLO-205, and HepG2), as well as from T47D, were eluted with GalNAc as previously described (Schjoldager et al., 2012). RapiGest elution was introduced in order to avoid an extra dialysis step, since the RapiGest fractions could be pooled and then directly reduced, alkylated, proteolyzed, and subjected to LWAC on the long VVA column as described above for the cell pellet samples.

Peptide Isoelectric Focusing

Following VVA-LWAC enrichment, those fractions most enriched for glycopeptides were pooled together, dried by vacuum centrifugation, reconstituted in IPG rehydration buffer, and submitted to IEF fractionation, as described (Vakhrushev et al., submitted). Isoelectric focusing was performed on a 3100 OFFGEL fractionator (Agilent) using OFFGEL Low Res Kit, pH 3-10 (Agilent) according to manufacturer's instructions. Typically, 12 fractions were collected, desalted by StageTips (Empore 3M) (Rappsilber et al., 2002), and submitted to nLC-MS and -HCD/ETD-MS/MS as described below.

Liquid Chromatography - Mass Spectrometry

O-glycopeptide enriched samples were analyzed using a system composed of an EASY-nLC II (Thermo Fisher Scientific) interfaced via a nanoSpray Flex ion source to an LTQ-Orbitrap XL hybrid spectrometer (Thermo Fisher Scientific), equipped for both HCD- and ETD-MS2 modes, enabling peptide sequence analysis without and with retention of glycan site-specific fragments, respectively. The conditions of LC analysis were essentially as described previously (Steentoft et al., 2011), with some important additional workflow modifications. Briefly, the nLC was operated in a single analytical column set up using PicoFrit Emitters (New Objectives, 75 µm inner diameter) packed in-house with Reprosil-Pure-AQ C18 phase (Dr. Maisch, 3-µm particle size, 19 cm column length). Gradient elution times of 90 or 120 min were employed, depending on sample complexity, which was first assessed by subjecting a 5% aliquot of each IEF fraction to preliminary nLC-ESI-MS screening. This was carried out using 90 min gradient elution and a data-dependent acquisition (DDA) routine consisting of a precursor MS1 scan of intact peptides acquired in the Orbitrap (*m/z* 350–1,700; nominal resolution 30,000), followed by Orbitrap HCD-MS2 (*m/z* of 100–2,000; nominal resolution 30,000) of the five most abundant multiply charged precursors above 5,000 counts in

each MS1 spectrum ("top five method"). Only fractions exhibiting significant abundance of peptide precursors yielding a HexNAc fragment at *m/z* 204.086 were subjected to further nLC-ESI-MS analysis. For this subsequent analysis, the remaining 95% of each sample was subjected to one of two alternative DDA methods for acquiring "linked" HCD/ ETD-MS2 sets from the same precursor. These were: (i) a three-step DDA routine including HCD triggering of a subsequent ETD scan, whereby, similar to the method above, each Orbitrap MS1 scan was followed by Orbitrap HCD-MS2 acquisition from the three most abundant multiply charged precursors above a threshold of 50,000 counts ("top three" triggered method); in this case, however, the appearance of a HexNAc fragment at *m/z* 204.086 (± *m/z* 0.15) in the HCD-MS2 spectrum triggered a subsequent ETD-MS2 acquisition from the same precursor (*m/z* of 100–2,000; nominal resolution 15,000) (Steentoft et al., 2011); or (ii) each Orbitrap MS1 was followed by alternating acquisition of HCD-MS2 and ETD-MS2 spectra from each of the three most abundant multiply charged precursors ("top three"alternating method). The latter routine was adopted for analysis of fractions exhibiting a high degree of enrichment, where there was little concern for unnecessary acquisition of ETD-MS2 spectra of non-glycosylated peptides. In general, activation times were 30 msec and 100-200 msec for HCD and ETD fragmentation, respectively; isolation width was 4 mass units, and usually 1 microscan was collected for each spectrum. Automatic gain control (AGC) targets were 100,000 ions for Orbitrap MS1 and 10,000 for MS2 scans, and the AGC for fluoranthene ion used for ETD was 300,000. Supplemental activation (20%) of the charge-reduced species was used in the ETD analysis to improve fragmentation. Dynamic exclusion for 30 sec was used to prevent repeated analysis of the same components. Polysiloxane ions at *m/z* 445.12003 were used as a lock mass in all runs.

MS Data Analysis

The raw data were processed, in a manner similar to previous publications (Steentoft et al., 2011), using Proteome Discoverer 1.2 software (Thermo Fisher Scientific) and searched against the human-specific UniProt KB/SwissProt-reviewed database downloaded on July 8, 2010. The SEQUEST search engine was used for HCD and ETD data; in addition, the ZCore search engine was used for ETD data. In all cases the precursor mass tolerance was set to 10 ppm and fragment ion mass tolerance to 50 mmu. Carbamidomethylation on cysteine residues was used as a fixed modification. Methionine oxidation and HexNAc attachment to serine and threonine were used as variable modifications. As a further pre-processing procedure, all HCD data showing the presence of fragment ions at *m/z* 204.08 were extracted into a single .mgf file (signal intensity threshold, 1.5), and the exact mass of 1x, 2x, 3x, and 4x HexNAc units was subtracted from the corresponding precursor ion mass, generating four distinct files. For this purpose, a script written in Microsoft Visual Basic 6.5 was used. These pre-processed data files were submitted to a SEQUEST database search under the same conditions mentioned above, again considering HexNAc attachment. All spectra were searched against a decoy database using target false discovery rates (FDRs) of 1% and 5% and the results merged together into a final protein list.

In order to search efficiently for GalNAc O-glycosylation of tyrosine, a separate processing run on Proteome Discoverer was carried out on the data set with HexNAc allowed as a modification of Y. High scoring hits were accepted after careful validation of ETD-MS2 spectra to make sure that sufficient fragments were clearly observable above noise to form a self-consistent and unambiguous set defining sequence and HexNAc position (see below).

Validation of Computational Search Results

The computer assignments of all spectra from candidate glycopeptides were validated by a manual inspection process which in effect rendered FDR a less critical parameter than scoring rank. In this process, only assignments of rank 1 were considered, regardless of FDR, although for obvious reasons peptide matches at the extreme low end of the probability scale were automatically dropped from the validation list before proceeding. Higher probability, low (1%) FDR picks were generally examined first, wherever they appeared; however, careful examination of linked HCD and ETD spectra showed that a significant number of 5% and even some <5% FDR matches provided adequate fragmentation to specify the sequence, particularly with relatively short peptides; this was especially apparent where an HCD/ETD pair from the identical precursor provided an additional factor of congruence not taken into account in the SEQUEST/ZCore scoring in Proteome Discoverer. In other words, a relatively low confidence ETD-MS2 might provide just enough fragmentation to specify the O-glycosylation site on the peptide sequence specified by its associated high confidence HCD-MS2 match. In other cases, a high confidence ETD-MS2 provided all the necessary fragmentation to bolster a low-confidence (or even absent) HCD-MS2 sequence match. In some cases two low-confidence but complementary HCD/ETD matches could add up to a high confidence sequence match when considered together. On the other hand, ETD-MS2 spectral matches at 5% FDR in particular were examined carefully, since one or more of the O-glycosites specified by Proteome Discoverer were likely to be incorrect, or at least less specific than initially assigned.

In cases where the spectral data were sufficient to identify a definite peptide sequence bearing a specific number of HexNAc residues, but insufficient to yield their precise locations, these were designated as “ambiguously identified” sites. In general, this situation occurred where a quality HCD-MS2 spectrum was acquired, but the ETD-MS2 was either missing or exhibited insufficient fragmentation to confirm all O-glycosite locations. An exception to this was made when the number of HexNAc residues detected was found to agree exactly with the number of potential O-glycosites in the sequence; then the sites were considered to be specified. Partially identified sites based on HCD-MS2 only are accounted for in the results listed in Table S2.

In order to identify cases where peptides from bovine proteins could potentially contaminate the data set, a text string search of all proposed human peptide sequences was carried out against the bovine UniProt database (UniProtKB reviewed, September 2012 release). Although it is considered unlikely that O-glycosylation of natural bovine proteins would be found with the truncated structures as found in

SimpleCells, any peptide sequences found to be 100% identical between the two species is marked as such in Table S2.

NetOGlyc Version 4.0

Development of the predictor involved collection and preparation of data sets, calculation and encoding of features to be used, and an iterative process of training, testing and finally evaluating a support vector machine (SVM) to build the predictor.

Data Sets

We developed a two-class predictor during this project - where the predictor can classify unknown data points into two classes: positive and negative, representing glycosylation and non-glycosylation, respectively. Four sources of data points have been used to develop the predictor – the curated subset of UniProt for mammalian proteins, SimpleCell, SimpleCell negative, and the Julenius subset of O-GLYCBASE used to train NetOGlyc3.1. UniProt, SimpleCell and Julenius can provide positive data points, while SimpleCell negative and Julenius are providers of negative data points. The Julenius negative data are only used in evaluation of the predictor.

All sites from the different data sets were filtered so that any positive or negative site annotation occurred only on serine or threonine residues. For each of the data sets, UniProt identifiers, as well as complete FASTA files encompassing all the proteins represented in the set were retrieved from UniProt. Retrieval of the FASTA files is necessary to perform batch calculation of features as well as for the process of encoding.

Uniprot positive data points set – Using the bulk data retrieval utility for UniProt (The UniProt Consortium, 2012) mammalian (Taxonomy ID 40674) proteins were retrieved for UniProt release 2012_07, selecting only curated entries. Further, PTM annotations corresponding to O-GalNAc modifications with experimental evidence were extracted to a set of positive data points that can be used as the independent data set.

Julenius positive and negative data points sets - Downloads of the original definitions for training data sets for NetOGlyc 3.1 (Julenius et al., 2005) were available online. These original definitions were retrieved, and used to select sites from O-GLYCBASE 6.0 (Gupta et al., 1999), so that sets of O-GalNAc sites for each of the training and testing sets in NetOGlyc3.1 were obtained. Negative data is defined as the sites on a protein that do not belong to any of the NetOGlyc3.1 positive data points. There is a very large imbalance between the number of points in the positive data point set and the negative data point set for the Julenius sets. When used, in training and testing, the Julenius data sets were balanced to a ratio of 0.75 negative sites for every positive site.

SimpleCell positive data points set – Data for the positive data points was obtained experimentally, as described above.

SimpleCell negative data points set – A mass spectrometry-based proteomic analysis of the LWAC flow through (containing predominantly non-glycosylated peptides) of selected SimpleCell lines has been

performed, yielding 24,000 distinct peptide sequences to be used as negative data. Any negative peptides that contained a site that had been found to be positively glycosylated in the SimpleCell dataset were discarded, as well as those from proteins without a predicted signal peptide or any serine or threonine residues. The sequences of each of the remaining peptides were processed so that a list of all serines and threonines, and their positions, were obtained. The negative data point set comprises this list of serine and threonine sites.

Features - TMHMM, DisEMBL, NetSurfP, sparse encoding

To satisfy the constraints of the support vector machine implementation, all features were encoded as decimal values between 0 and 1.

To obtain features such as surface accessibility and disorder for each protein, a feature selection process has to be undertaken (Jensen et al., 2002). Usually this entails execution of the third party feature generation executable to obtain the results, which are then encoded for consumption by the support vector machine. Reference files, containing the output of each of the feature generation programs were run on the proteins obtained from this master data set of sites. TMHMM (Krogh et al., 2001) version 2.0c was run on local servers, to obtain predictions of transmembrane regions on proteins. DisEMBL (Iakoucheva & Dunker, 2003) was compiled and run locally to obtain complementary disorder predictions. Similarly, NetSurfP output (version 1.1) was obtained for all the data sets. Data was also generated for NetTurnP (Petersen et al., 2010), as well as generating positional sequence scoring matrices (Altschul et al., 1997) but these features have not been selected for use in the final predictor.

Encoding

TMHMM feature encoding – For a given site on a protein, three features are generated – encoding whether TMHMM predicts the site occurs in a transmembrane region or inside or outside. These three conditions are encoded in a simple binary fashion (1 if true or 0 if false).

DisEMBL feature encoding – For a given site on a protein, three features are generated. Each of the features encodes the propensity for a given site to be in a particular class of disordered region. The output from DisEMBL already provides values from 0 to 1, so the output value for the given amino acid is simply looked up in the data. The features encoded correspond to the coil, hotloops and rem465 prediction values.

NetSurfP feature encoding – Three main properties are calculated in the NetSurfP feature encoding – Coil, beta-strand and alpha-helix. The properties are in the range of 0 to 1 and represent the propensity for each of these properties. In order to encode this into a feature, a window size 15 was used. This resulted in 96 features (3x32) - 15 amino acids N and C terminal from the site each, the site itself, and an average value across the 31 amino acids. If the window is overlapping with the N or C terminus of the protein, values of 0 are substituted in for the property value.

Sparse encoding – Sparse encoding for each amino acid type is used on a window size of 10 amino acids up and downstream of a residue. If the window overlaps with N and C terminals, the whole site was removed

from the data set. The net effect of this is that no sites within 10 amino acids of the ends of proteins are not encoded.

Redundancy Reduction and Data Partitioning

A common trap in machine learning is to partition data sets incorrectly when performing cross-validation. Reported cross-validation performance has little weight if maximal independence between training and testing is not ensured. To reduce redundancy between training and test sets, a multi-stage approach was taken. Sources for the training data sets (SimpleCell positive data points, Julenius positive data points and SimpleCell negative data points) as well as the testing set (UniProt positive data points) were processed to reduce redundancy. Using CD-HIT (Li & Godzik, 2006), clusters were generated by performing a series of pairwise comparisons from SimpleCell to SimpleCell negative, SimpleCell to Julenius, SimpleCell to UniProt and Julenius to UniProt (Figure S4A). Clustering was performed on windows around the site from each data point. Data points from a set that clustered with data points from other data sets were removed, preferring to keep data points from the SimpleCell set. In addition, for the SimpleCell and SimpleCell negative data sets, data points were removed that clustered with other data points from another protein originating from the same data set. This ensured that at most, each data point had similarity only to other data points originating from the same protein. The SimpleCell negative data set was not redundancy reduced against Julenius and UniProt as those data sets do not carry negative information.

In detail, a two-stage clustering process involving CD-HIT was used to cluster the individual data points. For each site (data point), windows covering wide (25 amino acids N and C terminal from a site), and narrow (5 amino acids N and C terminal from a site) window sizes were generated. Clustering was then performed on the wide and narrow windows using CD-HIT. The wide window similarity cutoff is 0.5, while the narrow window similarity cutoff is 0.65. The described method is a closer fit to the process here compared to the commonly used (Hobohm et al., 1992) data redundancy algorithm due to the specific attention to homology of regions within proteins.

The Julenius data had already had data redundancy applied to it, and so rather than re-apply redundancy reduction to it, the original three data sets were kept, and their structure taken into account when designing the cross-validation evaluation procedure.

Support vector machine

In contrast to NetOGlyc3.1, a support vector machine was used to train the data set. The support vector machine is a margin classifier, machine learning approach. LibSVM (Chang & Lin, 2012) was chosen as the implementation of the support vector machine. By changing the value of the cost parameter, and evaluating the performance by plotting MCC learning curves, optimal parameters were selected for each set of feature tried. Feature selection was performed by evaluating the learning curves for each individual feature and then selecting features to be combined into a final predictor.

Evaluation

Two levels of evaluation of the predictor took place - first at the training stage, during feature development, to characterise the nature of the improvement in predictor performance delivered by various features, and a second evaluation that took place after the development of the predictor, where the overall performance of the predictor was evaluated and its ability to generalise established. The Matthews Correlation Coefficient (MCC) (Matthews, 1975) is used as a general-purpose measure of the performance of the predictor.

A useful tool to characterise the efficacy of machine learning algorithms is the learning curve. By varying the percentage of total training data that is used to train the machine, an estimate can be made as to what the performance bottleneck is for the predictor. Both the cross-validation and training MCC can be plotted. For predictors where the limitation of the predictor is that there is not enough data, the (naive) cross-validation and training MCC should both be decreasing with increasing amounts of data, with the cross-validation curve approaching the training curve. Features that lack enough resolving power to distinguish between classes will not improve the performance of the predictor even with increasing amounts of data. These learning curves will appear more flat, and the cross-validation curve will appear asymptotic to the cross-validation curve (with a significant gap between the two curves). During development of the predictor, learning curves were generated for each feature examined, as well as for combinations of features, varying both the amount of positive and negative data used for training.

Evaluation of the predictor was undertaken by performing a 4-fold cross-validation (Figure S4B,C) splitting the data sets by protein, additionally testing against an independent test set, and repeating three times. A training set, testing set and independent testing set were established from the SimpleCell data set, SimpleCell negative data set and UniProt data set. To establish the independent test set, sites for 20% of the proteins contained in the SimpleCell negative data set were used as an independent test sites. Since internal redundancy was removed from the SimpleCell negative data set, independence is not violated when this occurs. In a similar procedure, the SimpleCell data and SimpleCell negative data was divided into 4 sets so that three of the four sets were used to train the predictor, whilst the fourth was used to test. This process is repeated 4 times, so that each set can be used once for testing. On each iteration of the process, the trained SVM is tested against the independent data set. True and false positive and negative statistics are collected for the whole process, and the cross-validation MCC as well as independent testing set MCC were calculated. A further evaluation of the predictor took place, ensuring that the performance of the algorithm is at least comparable to NetOGlyc3.1. To achieve this, an evaluation of the SVM was performed on a combination of the NetOGlyc3.1 training data as well as the SimpleCell data.

It should be noted that further interrogation of proteins predicted to be O-glycosylated by the current algorithms is required in order to rule out sites predicted in the cytoplasmic domains of transmembrane proteins. These regions would not be exposed to the ER-Golgi glycosylation machinery and Ser/Thr-rich regions in cytosolic domains are often phosphorylated as well as O-GlcNAcylated by the cytosolic O-GlcNAc transferase (Hart & Akimoto, 2009). Using the GlycoDomainViewer we identified several

glycopeptides in such regions, which could represent contaminating O-GlcNAc glycopeptides (Table S2) although some cases could be due to faulty domain predictions (e.g., LSR).

In vitro GalNAc-T Enzyme Assay

In vitro glycosylation assays for purified GalNAc-Ts were conducted with 10 µg of acceptor peptides and 4 mM UDP-GalNAc as the sugar donor in 25 µL Cacodylate buffer containing 25 mM cacodylic acid sodium pH 7.4, 10 mM $MnCl_2$, 0.25% Triton X-100. After 4hr incubation at 37°C all the products were analyzed by MALDI-TOF-MS. Positive control peptides were included for each enzyme as follows; IgA-hinge VPSTPPTPSPSTPPTPSPSK for GalNAc-T1, T2, T3, and T11; IgA-hinge-2xGalNAc for T12; MUC2 peptide PTTTPITTTTTVTPTPTPTGTQTPTTTPISTTC for T5; ACV8 peptide PEVTYEPPPTAPTLLTVLAYSL for T14; and CA9 peptide GSLKLEDLPTVEAPGDPQEP for T16. Results were graded as positive (more than 10% product with one or more GalNAc incorporated) or negative (less than 10% product).

Reference List

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402

Bennett EP, Mandel U, Clausen H, Gerken TA, Fritz TA, and Tabak LA (2012) Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22:** 736-756

Chalkley RJ, Thalhammer A, Schoepfer R, Burlingame AL, (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci U S A* **106**: 8894-8899

Chang Chih-Chung & Lin Chih-Jen LIBSVM: A Library for Support Vector Machines 2012
Ref Type: Online Source

Darula Z & Medzihradszky KF (2009) Affinity enrichment and characterization of mucin core-1 type glycopeptides from bovine serum. *Mol Cell Proteomics* **8**: 2515-2526

Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE (1999) O-GLYCBASE version 40: a revised database of O-glycosylated proteins. *Nucleic Acids Res* **27**: 370-372

Hart GW & AkimotoY (2009) The O-GlcNAc Modification. In *Essentials of Glycobiology.* Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart G and Etzler ME pp.263-279. Cold Spring Harbor (NY)

Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of representative protein data sets. *Protein Sci* **1**: 409-417

Iakoucheva LM & DunkerAK (2003) Order disorder and flexibility: prediction from protein sequence. *Structure* **11**: 1316-1317

Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S (2002) Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* **319**: 1257-1265

Julenius K, Molgaard A, Gupta R, Brunak S (2005) Prediction conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* **15**: 153-164

Krogh A, Larsson B, von HG, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567-580

Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**: 442-451

Petersen B, Lundegaard C, Petersen TN (2010) NetTurnP--neural network prediction of beta-turns by use of evolutionary information and predicted protein sequence features. *PLoS One* **5**: e15079

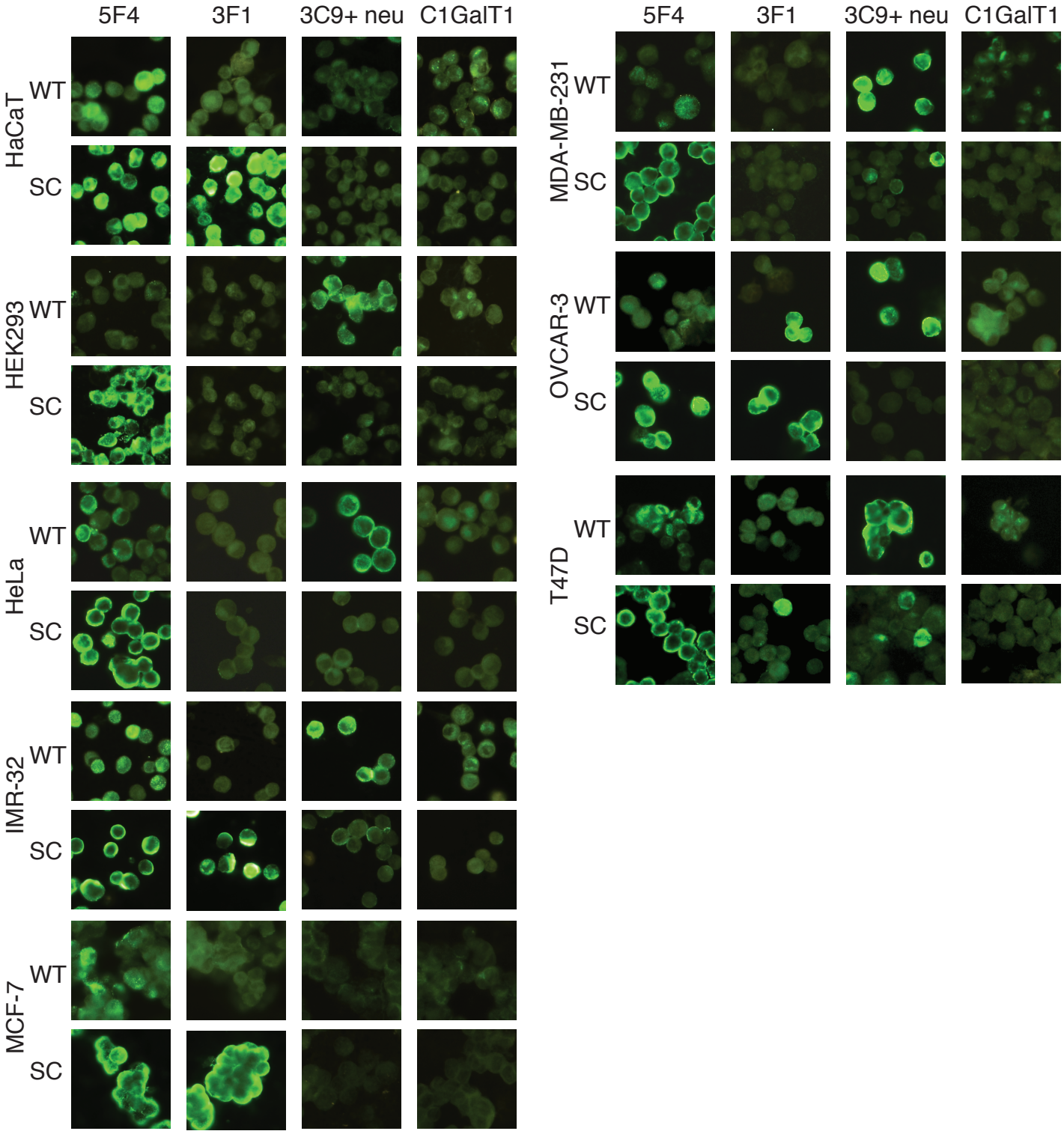
Rappsilber J, Ishihama Y, Mann M (2002) Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization Nanoelectrospray and LC/MS Sample Pretreatment in Proteomics. *Anal Chem **75***: 663-670

Schjoldager KT, Vakhrushev SY, Kong Y, Steentoft C, Nudelman AS, Pedersen NB, Wandall HH, Mandel U, Bennett EP, Levery SB, Clausen H (2012) Probing isoform-specific functions of polypeptide GalNAc-transferases using zinc finger nuclease glycoengineered SimpleCells. *Proc Natl Acad Sci U S A* **109**: 9893-9898

Steentoft C, Bennett EP, Clausen H (2013) Glycoengineering of human cell lines using zinc finger nuclease gene targeting - SimpleCells with homogeneous GalNAc O-glycosylation allow isolation of the O-glycoproteome by one-step lectin affinity chromatography *Methods Mol Biol*
Ref Type: In Press

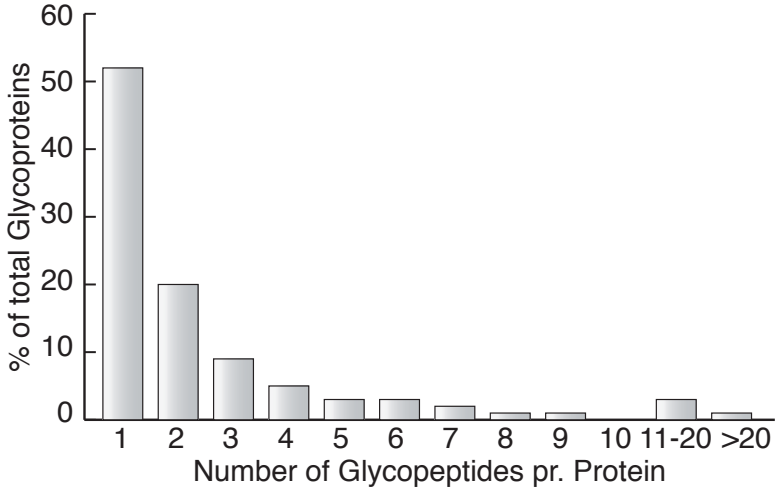
Steentoft C, Vakhrushev SY, Vester-Christensen MB, Schjoldager KT, Kong Y, Bennett EP, Mandel U, Wandall H, Levery SB, and Clausen H (2011) Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines. *Nat Methods* **8**: 977-982

The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**: D71-D75

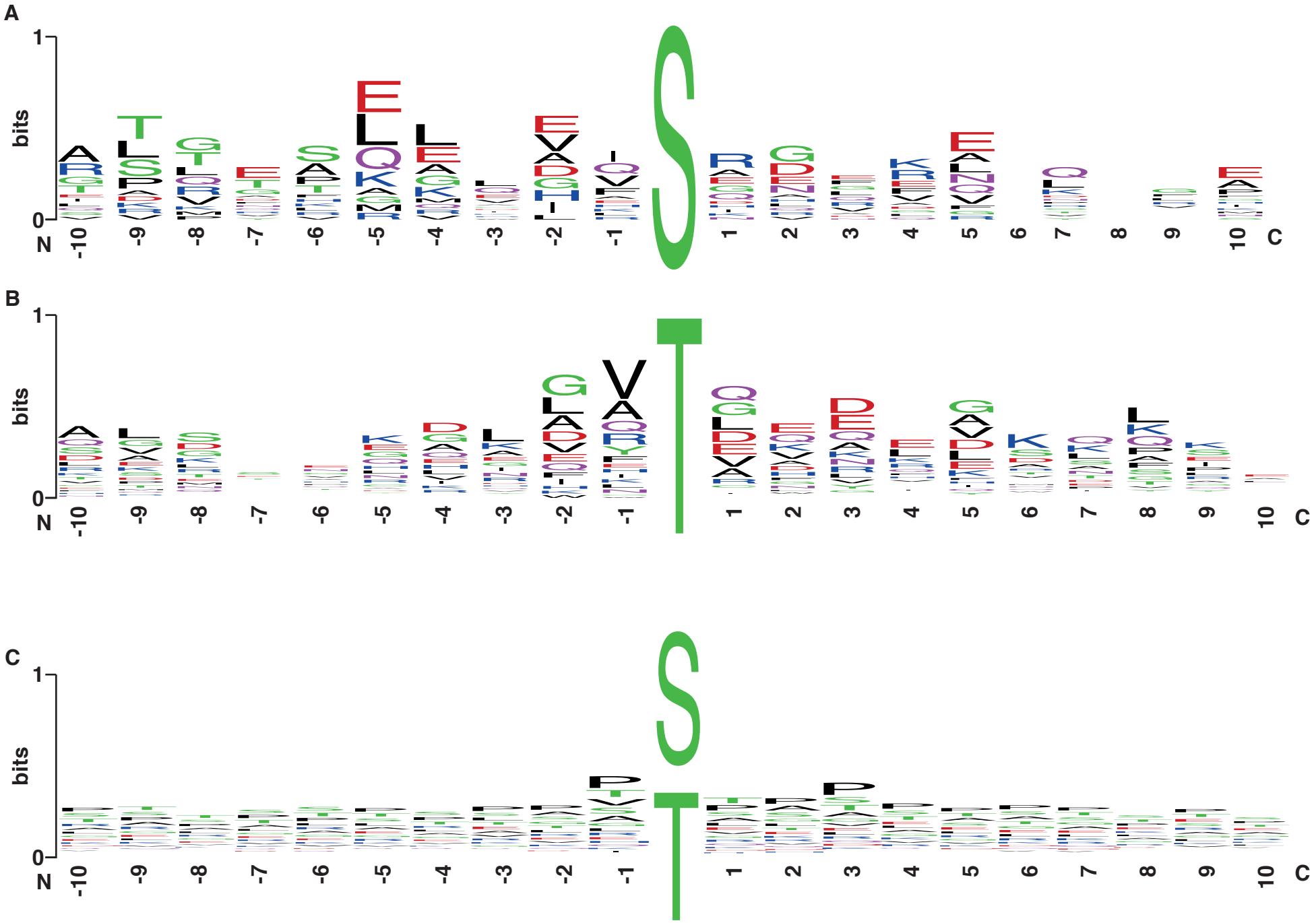


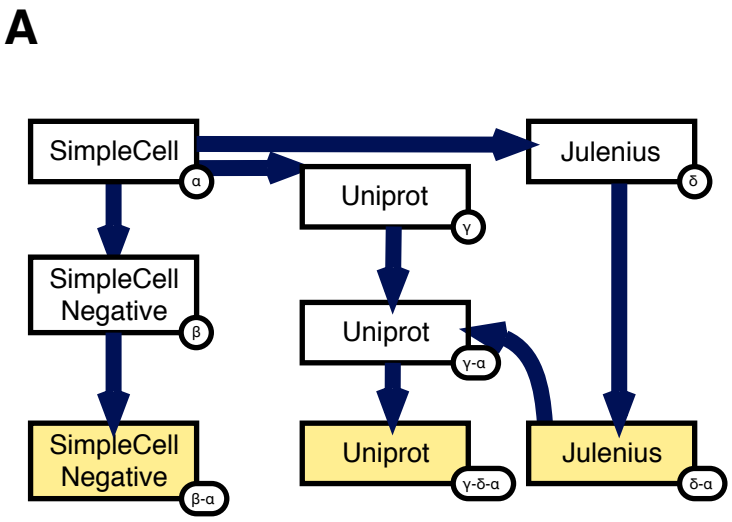
Steentoft *et al.* Supplemental Figure 1.

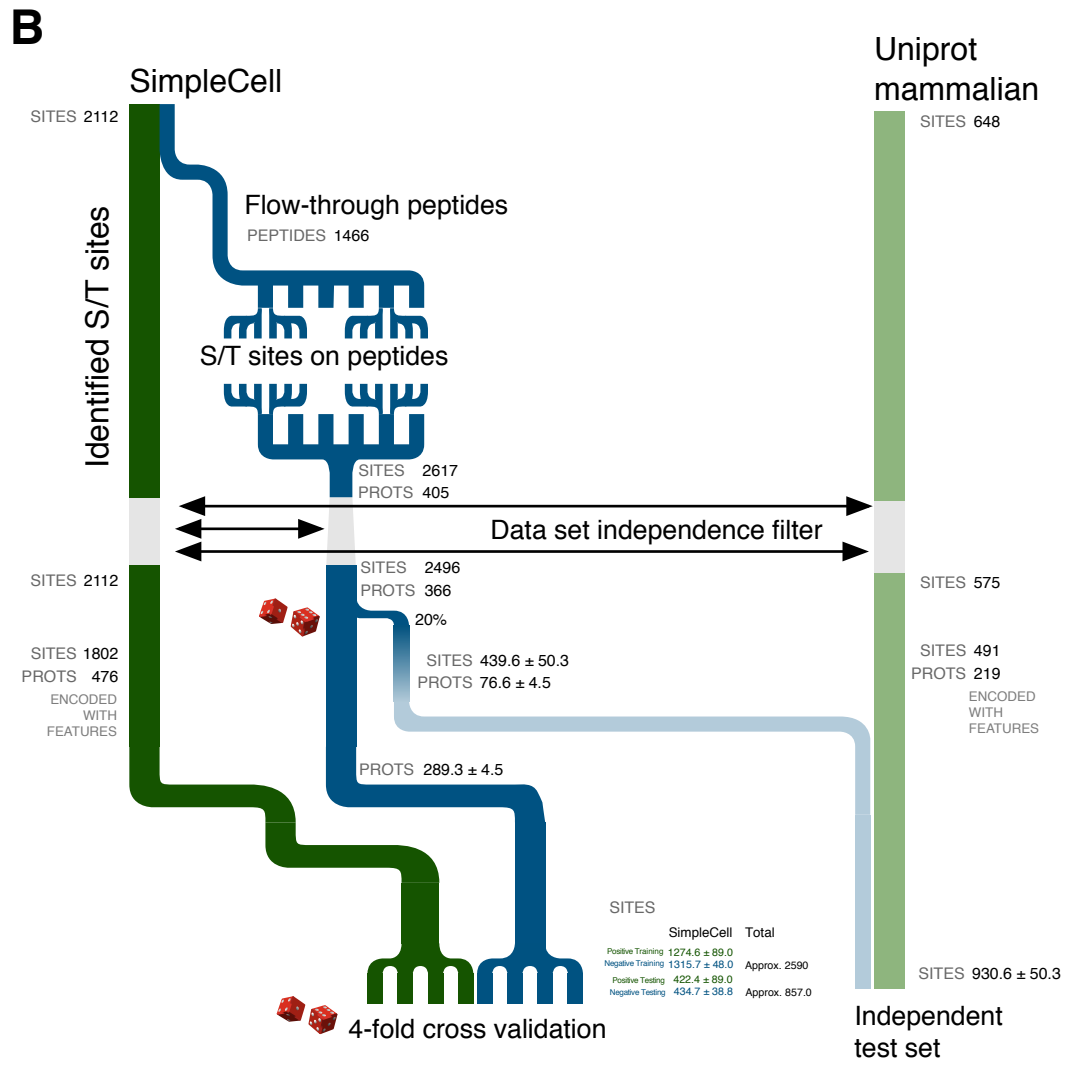
Steentoft *et al.* Supplemental Figure 2.

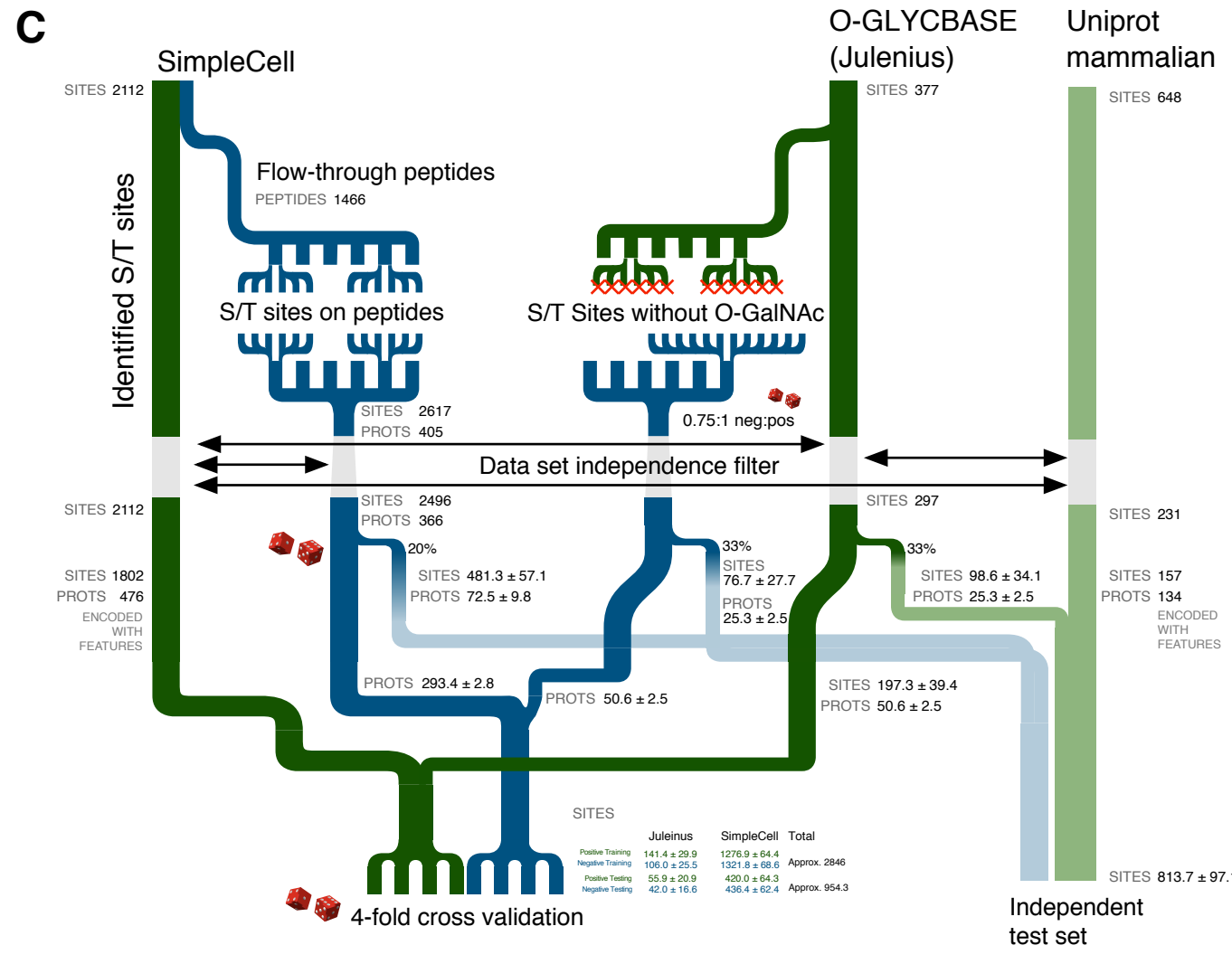


Steentoft *et al.* Supplemental Figure 3.









Steentoft *et al.* Supplemental Figure 4.