Supplementary Material

Sparse Isotope Labeling for NMR of Glycoproteins using ^{13}C-Glucose

Supplemental methods:

Sequence of CEACAM1 construct:

```
UniprotP13688:
         39          49          59          69          79          89
GSGGAQLTTE SMPFNVAEGK EVLLLVHNLP QQLFGYSWYK GERVDGNRQI VGYAIGTQQA

         99         109         119         129         139
TPGPANSGRE TIYPNASLLI QNVTQNDTGF YTLQVIKSDL VNEEATGQFH VY
```

cyan = scar left from TEV cleavage

orange = potential N-glycosylation sites

NMR supporting data:
The ^{1}H-^{13}C HSQC-TOCSYs used the pulse sequence hsqcdietgpisp, which uses a DIPSI mixing scheme for isotropic mixing over a time of 40 ms for the u-^{13}C-glucose supplemented spectrum and 60 ms for the other samples. The sequence and parameters are otherwise equivalent to the HSQCs used.[think this figure will end up in supplement if anywhere, so maybe supplementary methods?]
The ^{13}C -^{13}C COSY pulse sequence was generated in-house based on a standard ^{1}H -^{1}H COSY. It is a direct ^{13}C detection experiment. The spectral width is 60 ppm centering in the sugar region at 60 ppm. The size of the fid is 1024 x 180. The relaxation delay is 2 sec.
The ^{1}H-^{13}C HMBC used the Bruker pulse sequence hmbcgpndqf, which uses gradient pulses for selection, optimizes long range couplings and has no decoupling during acquisition. It was taken over a spectral with of 13 ppm in the proton dimension and 220 ppm in the carbon dimension with offsets of 4.7 ppm and 100 ppm respectively. The size of the fid was 4096 x 256 and the relaxation delay was 1.5 sec.

MS Determination of relative proportions of CEACAM1-IgV glycoforms:
The N-terminal domain of CEACAM1 has three glycosylation sites. After enzymatic treatment with Endo-F each site can either contain a GlcNAc residue or an unmodified Asp, which means there are eight possible glycoforms: [000], [100], [010], [001], [110], [101], [011], and [111] (see results section for explanation of nomenclature). The first and last form have unique masses and their abundance can be determined directly from the intact mass spectrum, but there are three isomers with the mass of the amino acid sequence and one GlcNAc residue and also three isomers with the mass of the amino acid sequence and two GlcNAc residues. For these isomeric species further tandem mass spectrometry data is required.

In the case of the species containing one GlcNAc the following ions were detected and they inform on specific isomeric species:

C81	[010], [001]
C81 + GlcNAc	[100]
C85	[001]
C85 + GlcNAc	[100], [010]
Z31	[100]
Z31 + GlcNAc	[010], [011]
Z27	[100], [010]
Z27 + GlcNAc	[001]

If we assume that fragments are formed with the same efficiency, we can set up several equations relating the relative intensity of a fragment with or without GlcNAc and the relative abundance of the individual isomers, e.g.:

f_(C81 + GlcNac)= f_([100])

(I(C81 + GlcNAc))/(I(C81) + I(C81 + GlcNAc))=([100])/([100] + [010] + [001])

The assumption of equal fragmentation efficiency may be problematic. We attempt to minimize this by calculating the average using multiple fragments and then reporting the averaged result. In the case of the [010] species, no fragment can isolate this site by itself. Nonetheless, by subtracting the effect of the other site the relative abundance of this isomer can be determined, for example:

f_(C85 + GlcNAc)- f_(C81 + GlcNAc)= f_([010])

These relative abundances can then be related back to the intact mass spectrum allowing determination of the overall proportion of each glycoform.

Supplementary Table

Table S1. Top-Down ECD fragments used for glycosylation site mapping.

Precursor	Glycosylation Site(s)	Fragment Ion	Theoretical Monoisotopic Mass	Observed Mass	Mass Error (ppm)	S/N Ratio	Fractional Abundance[b] (%)
P + GlcNAc	N104	C81	8701.4325	8701.4335	0.12	7.9	79.8
		C81 + GlcNAc	8904.5119	N.D.[c]	N/A	N/A	20.2[d]
	N104, N111	C85	9143.6501	N.D.	N/A	N/A	5.1[d]
		C85 + GlcNAc	9346.7295	9346.7332	0.40	37.6	94.9
	N111, N115	Z31	3514.6887	N.D.	N/A	N/A	26.0[d]
		‘Z31 + GlcNAc	3718.7759	3718.7734	-0.66	5.7	74.0
	N115	Z27	3072.4711	3275.5520	0.47	43.7	89.4
		Z27 + GlcNAc	3275.5505	N.D.	N/A	N/A	10.6[d]
P + 2 GlcNAc	N104	C81	8701.4325	N.D.	N/A	N/A	7.4
		C81 + GlcNAc	8904.5119	8904.5117	-0.015	24.9	92.6
	N104, N111	C85 + GlcNAc	9346.7295	9346.7281	-0.14	29.6	61.7
		C85 + 2 GlcNAc	9549.8088	9549.8109	0.21	18.4	38.3
	N111, N115	Z31 + GlcNAc	3717.7681	3717.7688	0.21	37.2	93
		Z31 + 2 GlcNAc	3920.8474	3920.8485	0.27	2.6	7
	N115	Z27	3072.4711	3072.4696	-0.49	54.9	51.2
		Z27 + GlcNAc	3275.5505	3275.5504	-0.032	52.4	48.8

[a]Relative abundance is the intensity of a peak relative to the base peak intensity
[b]Fractional abundance is the proportion of intensity of a fragment in a certain glycan occupancy state relative to the total intensity of that same fragment in all glycan occupancy states.
[c]Not Detected
[d]In cases where are particular fragment was not observed these values were estimated based upon the signal to noise ratio of the observed peak, the reported values reflect a lower limit of the proportion of the observed GlcNAc occupancy state.

Supplemental Figures:

Figure S1. ^{1}H-^{13}C HSQC of u-^{13}C-glucose supplemented CEACAM1-IgV at a low threshold with a colored gradient filter for peak intensity. a) region including methyls of acetyl groups on GlcNAcs (extremely intense) as well as beta carbons of Gln and Glu, which are moderately intense, and Asp and Asn beta carbons, which are weak, and gamma carbons of Gln and Glu (weak). b) aromatic region, extremely limited natural abundance signal.

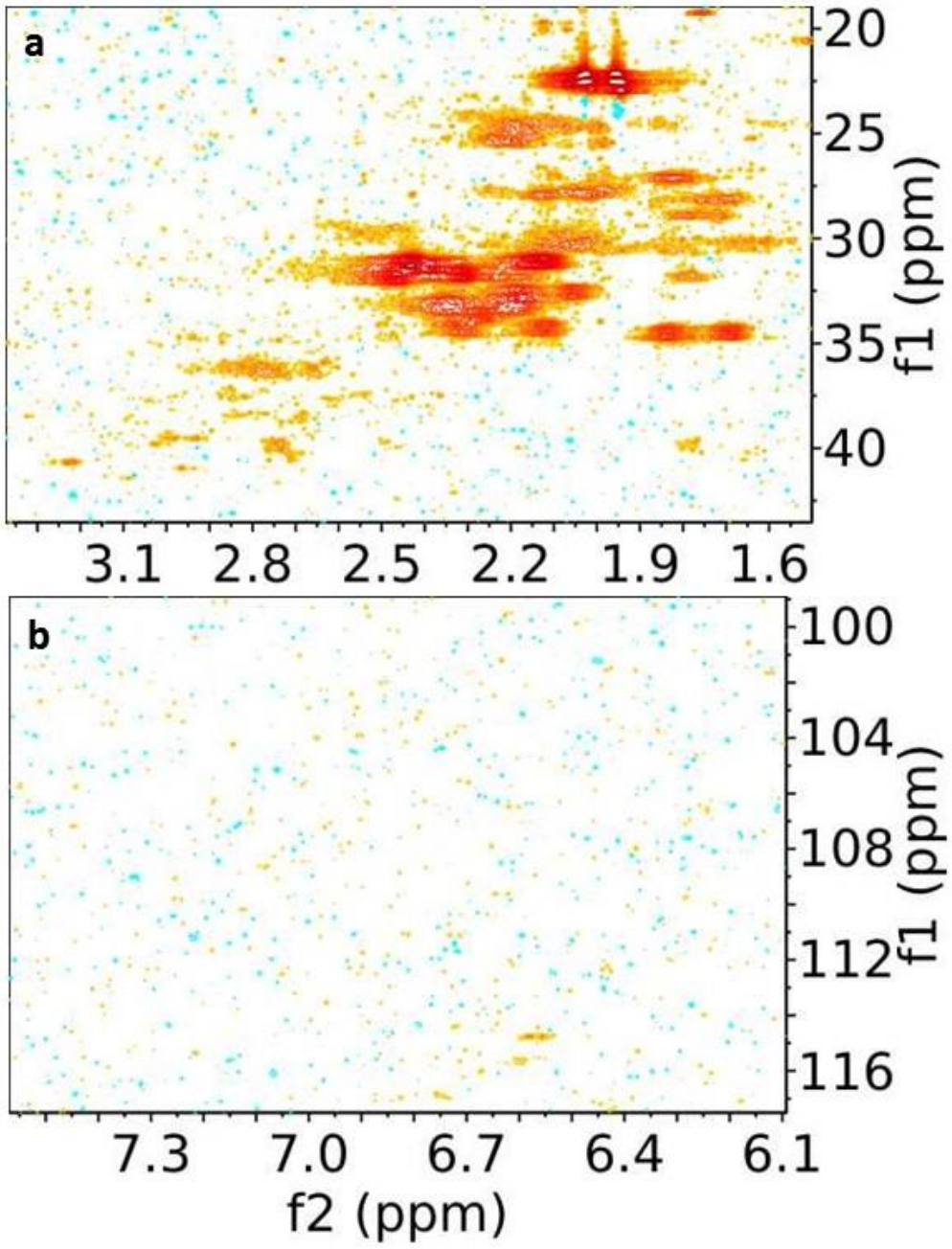


Figure S2. HSQC-TOCSY of sugar region of u-^{13}C-glucose supplemented CEACAM1-IgV.

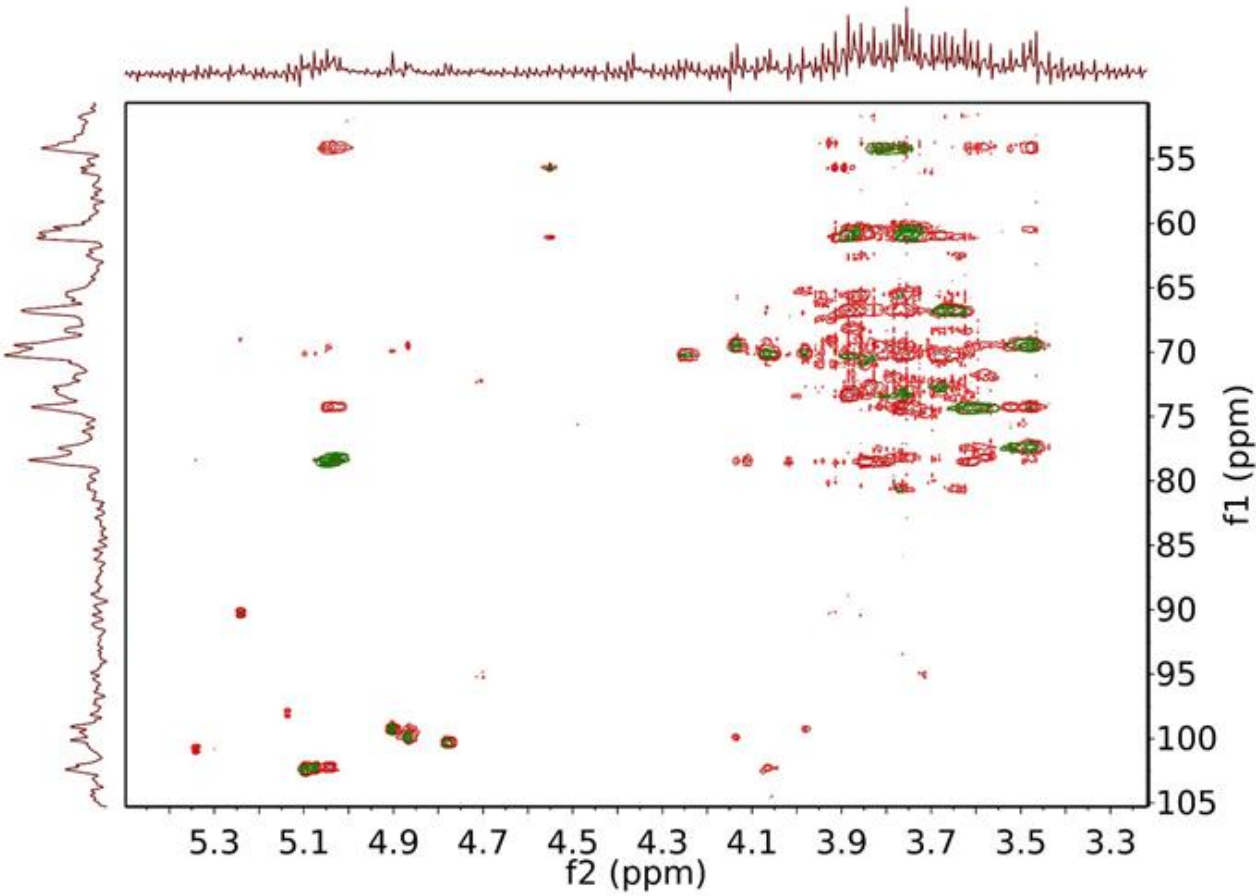
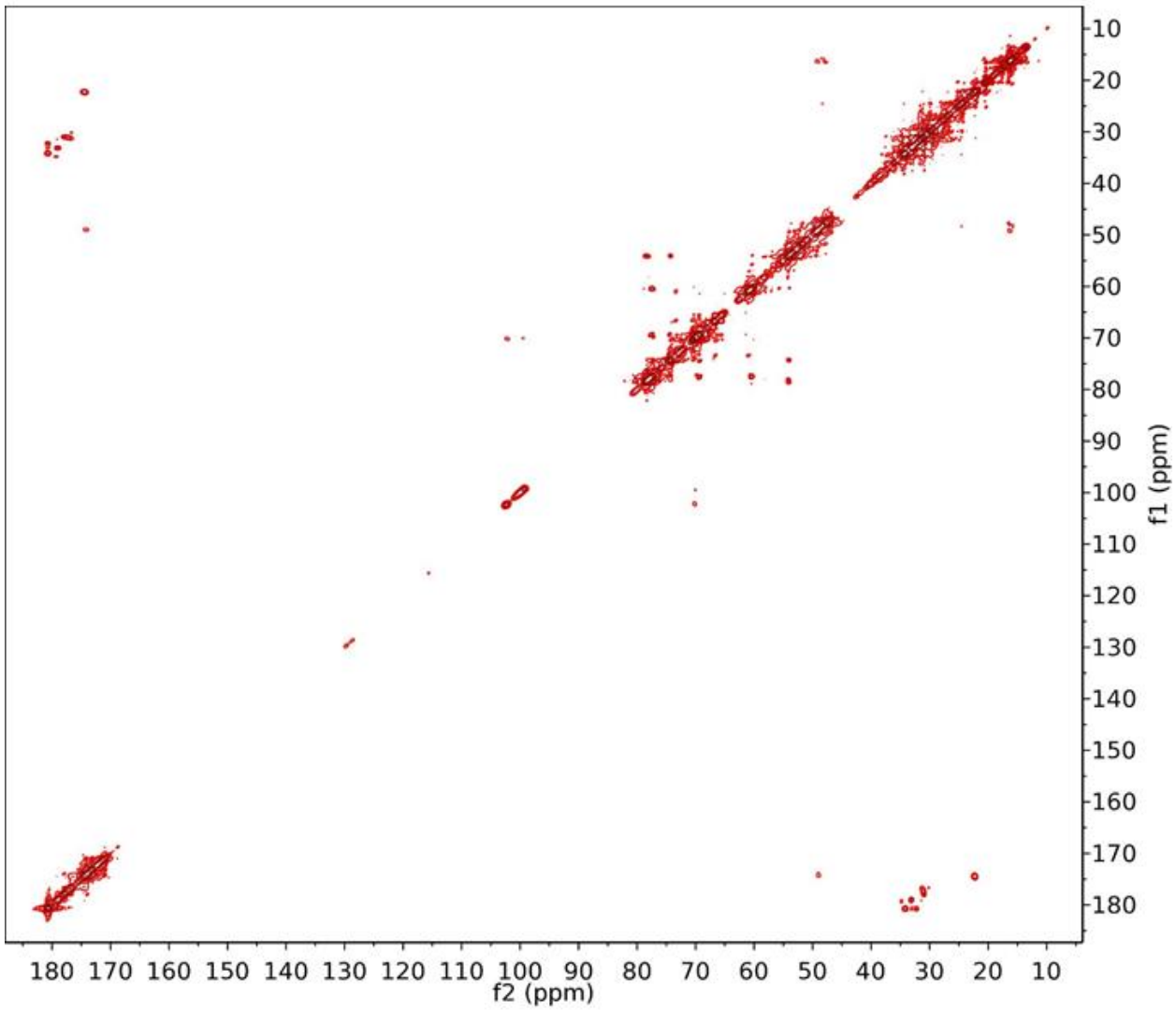


Figure S3. ^{13}C-^{13}C COSY of sugar region of u-^{13}C-glucose supplemented CEACAM1-IgV.


10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
f1 (ppm)
180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10
f2 (ppm)