Full-length single-molecule protein fingerprinting

Filius, M.; van Wee, R.G.; de Lannoy, C.V.; Westerlaken, I.; Li, Zeshi; Kim, S.H.; de Agrela Pinto, C.; Wu, Yunfei; Boons, Geert-Jan; Pabst, Martin

DOI
10.1038/s41565-023-01598-7

Publication date
2024

Document Version
Final published version

Published in
Nature Nanotechnology

Citation (APA)
Filius, M., van Wee, R. G., de Lannoy, C. V., Westerlaken, I., Li, Z., Kim, S. H., de Agrela Pinto, C., Wu, Y., Boons, G.-J., Pabst, M., de Ridder, D., & Joo, C. (2024). Full-length single-molecule protein fingerprinting. *Nature Nanotechnology*, *19*(5), 652-659. https://doi.org/10.1038/s41565-023-01598-7

nature nanotechnology

Article

https://doi.org/10.1038/s41565-023-01598-7

Full-length single-molecule protein fingerprinting

Received: 21 November 2022

Accepted: 22 December 2023

Published online: 13 February 2024

Check for updates

Mike Filius[1], Raman van Wee[1,7], Carlos de Lannoy[1,2,7], Ilja Westerlaken[1], Zeshi Li[1], Sung Hyun Kim[1,3], Cecilia de Agrela Pinto[1], Yunfei Wu[4], Geert-Jan Boons[4,5], Martin Pabst[6], Dick de Ridder[2] & Chirlmin Joo[1,3]

Proteins are the primary functional actors of the cell. While proteoform diversity is known to be highly biologically relevant, current protein analysis methods are of limited use for distinguishing proteoforms. Mass spectrometric methods, in particular, often provide only ambiguous information on post-translational modification sites, and sequences of co-existing modifications may not be resolved. Here we demonstrate fluorescence resonance energy transfer (FRET)-based single-molecule protein fingerprinting to map the location of individual amino acids and post-translational modifications within single full-length protein molecules. Our data show that both intrinsically disordered proteins and folded globular proteins can be fingerprinted with a subnanometer resolution, achieved by probing the amino acids one by one using single-molecule FRET via DNA exchange. This capability was demonstrated through the analysis of alpha-synuclein, an intrinsically disordered protein, by accurately quantifying isoforms in mixtures using a machine learning classifier, and by determining the locations of two *O*-GlcNAc moieties. Furthermore, we demonstrate fingerprinting of the globular proteins Bcl-2-like protein 1, procalcitonin and S100A9. We anticipate that our ability to perform proteoform identification with the ultimate sensitivity may unlock exciting new venues in proteomics research and biomarker-based diagnosis.

Protein synthesis is a highly regulated process, and much of this regulation occurs beyond the genome and transcriptome level. Via mechanisms such as alternative splicing and post-translational modifications (PTMs), a single protein-encoding gene can produce hundreds of unique protein products, or proteoforms[1]. Even subtle differences between proteoforms can markedly alter their biological functioning, and their aberrant expression is implicated in many diseases, including neurodegenerative diseases, metabolic disorders and a variety of cancers[2–4]. It is increasingly appreciated that protein functionalities in a given biological context need to be analysed at proteoform level, rather than at coding gene level.

Proteoform information can only be obtained without fault when the protein of interest is studied intact, for example using affinity-based approaches with probes (for example, antibodies)[5,6]. However, these approaches may suffer from low specificity and are limited by the number of probes that are available. Recently, high-resolution native mass spectrometry (MS) has been shown to be a powerful approach to investigate proteoform profiles[7,8]. However, exact information on

[1]Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Delft, The Netherlands. [2]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. [3]Department of Physics, Ewha Womans University, Seoul, Republic of Korea. [4]Department of Chemical Biology and Drug Discovery, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands. [5]Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands. [6]Department of Biotechnology, Delft University of Technology, Delft, The Netherlands. [7]These authors contributed equally: Raman van Wee, Carlos de Lannoy. e-mail: c.joo@tudelft.nl

the sequence of co-occurring modifications cannot be determined by the widely employed bottom-up approaches. Alternative top-down fragmentation experiments require large sample quantities, purity and substantial data interpretation, and may not be applicable to isobaric proteoforms without additional separation efforts[9]. Analysing full-length proteins at single-molecule resolution will offer a powerful solution to issues with the existing approaches.

While single-molecule sequencing of DNA and RNA is omnipresent[10,11], the nature of proteins creates several challenges that have thus far precluded their sequencing at the single-molecule level[12–15]. The increased number of building blocks in the polymer backbone from 4 nucleobases to 20 different amino acids complicates their discrimination and hinders specific labelling. The protein sequencing task is further impeded by the absence of a polymerase-like enzyme that can replicate proteins. Third, protein folding and interactions are much less predictable than nucleic acid base pairing. As a workaround for these challenges, multiple groups have proposed protein fingerprinting, in which partial sequence information is used to generate a unique protein fingerprint[16–20]. By mapping this fingerprint against a reference database, a protein can be identified. Thus far, proof-of-concept studies for protein fingerprinting have been limited to small model peptides[19,21–23], as their feasibility for full-length proteins is often hampered by their resolution, throughput or experimental complexity.

Here we present a single-molecule protein fingerprinting technology, termed fluorescence resonance energy transfer by DNA exchange, or FRET X, in which the distances from multiple specific amino-acid residues to a reference point on an intact protein are measured via FRET[19,24,25]. These nanoscale distances are inferred from the FRET efficiency and constitute the unique protein fingerprint, allowing for the identification of the protein analytes. Central to this technology is the use of fluorescently labelled short DNA oligonucleotides that transiently bind to the complementary sequence conjugated to specific amino-acid residues of the protein. The use of short DNA strands for protein fingerprinting has four main advantages: (1) the transient binding of the DNA probes allows for the detection of a single FRET pair at a time, even when multiple points of interest (for example, amino acids or PTMs) are present, which is not possible by direct labelling with fluorophores; (2) the highly specific and programmable nature of DNA hybridization allows for the specific and controlled targeting of each target residue, much like the super-resolution imaging technique DNA-PAINT[26,27]; (3) the repeated interrogation of the same FRET pairs on a protein increases the fingerprinting precision; (4) the pool of fluorescently labelled DNA probes is constantly replenished, eliminating concerns over photobleaching and enabling indefinite signal collection.

We demonstrate that full-length single proteins can be analysed with FRET X, reaching an average classification accuracy of 84% on alpha-synuclein (aSyn) isoforms. Harnessing the high resolving power, we demonstrate the ability to quantify protein abundances in a mixture and map threonine *O*-GlcNAcylation. We further show that, with site-specific amino-terminal bioconjugation, FRET X can be extended to non-recombinantly tagged proteins, a critical step towards analysing native samples.

Results

Single-molecule fingerprinting

To demonstrate the concept of protein fingerprinting using FRET X, we designed a single-molecule FRET assay where a DNA-labelled protein is immobilized on a surface passivated with polyethylene glycol (PEG) in a microfluidic device through biotin–streptavidin conjugation (Fig. 1a). Besides immobilization, this single-stranded DNA at the protein terminus also functions as a docking site for transient binding of complementary acceptor (Cy5)-labelled imager strands. The cysteine residues introduced at different positions were labelled with an orthogonal DNA sequence to allow transient binding of donor (Cy3)-labelled imager strands (Fig. 1a and Supplementary Fig. 1a). The donor and acceptor imager strands were designed to have mean dwell times ($\Delta\tau$) of 0.5 ± 0.1 s and 2.1 ± 0.1 s (Supplementary Fig. 1b,c), respectively. Binding of both imager strands was sufficiently weak to ensure dissociation and thereby repetitive, transient binding, but it was long enough to allow precise determination of the FRET efficiency for acquisition lasting several minutes (Supplementary Fig. 1d)[24]. Furthermore, to increase the probability of the presence of the acceptor imager strand upon donor imager strand binding and thus allow for FRET, we injected a fivefold molar excess of the acceptor imager strand over the donor imager strand.

We constructed six human aSyn model proteins (Fig. 1b,c). Each variant contains a genetically introduced cysteine at a different location and has a carboxy-terminal aldehyde encoding sequence for immobilization[28]. We constructed a kymograph with the FRET events for each single protein molecule (Fig. 1d and Supplementary Figs. 2 and 3), where the lines indicate the FRET efficiency (E) for each data point and the dots are the mean FRET efficiency for each event. The mean FRET efficiencies were fitted with a Gaussian mixture model (GMM) and we used the Bayesian information criterion to select the appropriate number of distributions to fit each histogram. The Gaussian function was used to resolve the centre of each peak with high precision and generate the protein fingerprint (Fig. 1d,e, bottom panels, and Supplementary Fig. 1e,f). The precision of the fingerprint depends on the number of binding events, and we can experimentally determine the fingerprint with a precision of $\Delta E \approx 0.03$ after ten binding events (Supplementary Fig. 1g), underscoring the benefit of our DNA hybridization scheme, in which the impact of stochastic photophysical effects is mitigated through repeated probing. It should be noted that the standard integration time of our measurement is 100 ms, which is several orders of magnitude slower than the typical timescale of protein conformational dynamics[29], and thus we expect a single FRET peak for each point of interest.

The different aSyn variants yield distinct distributions and fingerprints, with the FRET efficiency monotonically decreasing as the distance from the C-terminal reference point increases (Fig. 1e). This experiment shows that FRET X has a range of ~100 amino acids and that target amino acids whose locations differ by 5 amino acids (Cys 124 and Cys 129) are still discernible. We next sought to determine the classification accuracy of different aSyn constructs, all added in equal proportions to a mixture. To accomplish this, a support vector machine (SVM) classifier was first trained on FRET values obtained from four separate experiments, each containing a single aSyn mutant to learn the characteristic FRET value distribution for each (see Methods for an extensive description). The trained SVM was then used to classify individual molecules within the mixture on the basis of their respective FRET values, enabling us to determine the relative concentrations of each of the constructs (Supplementary Fig. 4). We demonstrate that we are able to retrieve the initial relative abundance of each construct with high reproducibility (Fig. 1f).

Single-molecule fingerprinting of disordered proteins

As a single type of amino acid can recur multiple times in a protein sequence, FRET X fingerprinting requires the detection of multiple FRET pairs in a single protein. To demonstrate that the transient and repetitive nature of binding events in FRET X facilitates fingerprinting of species with multiple FRET pairs, we designed two aSyn constructs, each containing two cysteines. The distances between the reference point and the first cysteine (Cys 124) are identical for the two constructs, while the distances to the second cysteine differ by 21 amino acids (Cys 78 for Fig. 2a, and Cys 99 for Fig. 2b). We observed a high FRET peak reporting on the relative position of Cys 124, which was similar for the two constructs (Fig. 2a,b and Supplementary Fig. 5a,b), and as expected the average FRET efficiency of the second cysteine differs between Cys 78 (0.32, Fig. 2a) and Cys 99 (0.43, Fig. 2b). Furthermore, the FRET efficiencies for the double-cysteine constructs are similar to the FRET efficiencies found in our experiments with the single-cysteine

https://doi.org/10.1038/s41565-023-01598-7

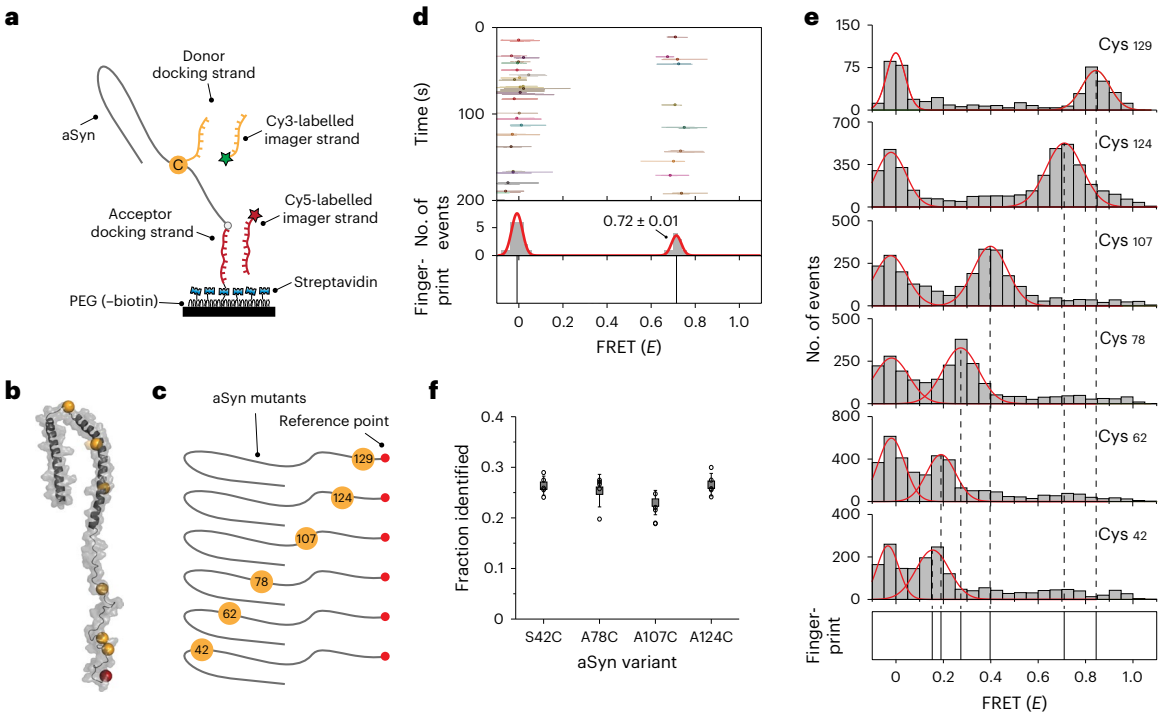

Fig. 1 | Repetitive binding of short DNA imager strands allows for high-resolution protein fingerprinting. a, Schematic representation of the single-molecule assay. The model protein, aSyn, is conjugated to a biotinylated single-stranded DNA strand (red) to facilitate immobilization of the target protein to the PEGylated quartz surface. The donor (Cy3)-labelled imager strand binds to the DNA docking site on the cysteine, while the acceptor (Cy5)-labelled imager strand hybridizes to the docking site at the C terminus of the protein. Simultaneous binding generates short FRET events and these are observed with total internal reflection microscopy. **b**, Three-dimensional conformation for micelle-bound aSyn [PDB 1XQ8] with the location of the six cysteines probed (orange) and the C terminus (red) indicated. **c**, Schematic representation of the six aSyn constructs. Each construct contains a single cysteine (orange circle), whose position relative to the N terminus is indicated, and a reference point at the C terminus (red circle). **d**, Representative kymograph from a single aSyn protein with a cysteine (Cys 124). The FRET efficiency for each data point in a binding event (lines) and the mean FRET efficiency from all data points in a binding event (dots) are indicated over the course of an experiment. The distribution of the average FRET efficiencies for each FRET event is fitted with a Gaussian function. The mean values of the Gaussian fits are plotted in a separate panel (bottom) and are referred to as the FRET X fingerprint of the protein. The population on the left ($E \approx 0$) originates from events where the acceptor fluorophore was absent. The mean ± s.e.m. values of the Gaussian fit of the ten FRET binding events are indicated in the plot. **e**, Ensemble FRET X histograms for each of the aSyn constructs shown in **c** (single-molecule and ensemble kymographs shown in Supplementary Fig. 2); the mean ± FWHM (full-width at half-maximum) FRET efficiencies were 0.84 ± 0.13 for Cys 129, 0.71 ± 0.18 for Cys 124, 0.40 ± 0.17 for Cys 107, 0.27 ± 0.17 for Cys 78, 0.19 ± 0.14 for Cys 62 and 0.15 ± 0.12 for Cys 42. The dashed line represents the centre of the Gaussian fit. **f**, Relative frequencies of detection for equimolar mixtures of four aSyn constructs, as determined by the trained SVM. The mean (grey squares) ± s.d. (whiskers) for each fraction measured is reported, and is derived from five individual measurements (open circles).

constructs, demonstrating the reproducibility of FRET X protein fingerprinting (Fig. 1e,f).

Single-molecule fingerprinting of globular proteins

To effectively fingerprint cellular proteins, our FRET X platform should be able to cope with folded structures. To demonstrate this ability, we sought to fingerprint the human apoptosis regulator Bcl-2-like protein 1 (Bcl) isoform Bcl-X_L, which has a single cysteine that is located close to its C terminus (Fig. 2c). For identification of a folded protein with FRET X, an experimentally obtained fingerprint should be mapped against a database consisting of computationally generated protein fingerprints using their three-dimensional structures available online. Hence, we used our previously developed FRET X fingerprint prediction tool, which takes the effect of the DNA tags on the protein structure into account[19], to predict the fingerprint of Bcl-X_L. The fingerprint prediction consisted of a single high FRET peak, and this was in line with experimental data (Fig. 2d,e and Supplementary Fig. 5c). Taken together, these results show that our FRET X fingerprinting approach is capable of obtaining reproducible fingerprints for both intrinsically disordered and folded human proteins, underscoring that the introduction of additional DNA tags and the labelling procedure itself do not interfere with our fingerprinting approach.

PTM mapping

O-GlcNAcylation is an essential process in mammalian cells involving the addition of a single *N*-acetylglucosamine (GlcNAc) to the hydroxyl side chain of serine and threonine residues by *O*-GlcNAc transferase (OGT)[30]. Dysregulation of *O*-GlcNAcylation has been implicated in many human pathologies, such as cancer, diabetes and neurodegenerative diseases, where the PTM site on the protein substrate is decisive for its outcome[30,31]. However, for *O*-GlcNAcylation, obtaining such information remains challenging, especially as there is no consensus motif for predicting the potential sites of the PTM[32]. As a result, mapping *O*-GlcNAc sites relies largely on the use of synthetic peptide fragments derived from the protein of interest[33,34], which may not reflect the bona fide PTM sites on the intact protein. We hypothesized that the high resolving power of FRET X could be leveraged to map potential *O*-GlcNAc sites of a full-length protein.

We incubated aSyn, which is known to undergo *O*-GlcNAcylation, with OGT in the presence of uridine diphosphate-linked 6-azido-GlcNAc[35]. The modified aSyn was subjected to copper click chemistry to attach the donor docking strands to the PTMed residues (Fig. 3a) and then immobilized at the C terminus in a similar fashion to before (Figs. 1 and 2). We observed a main FRET peak with an efficiency of 0.12 and a second FRET peak with an efficiency of 0.23 (Fig. 3b,c), indicating

https://doi.org/10.1038/s41565-023-01598-7

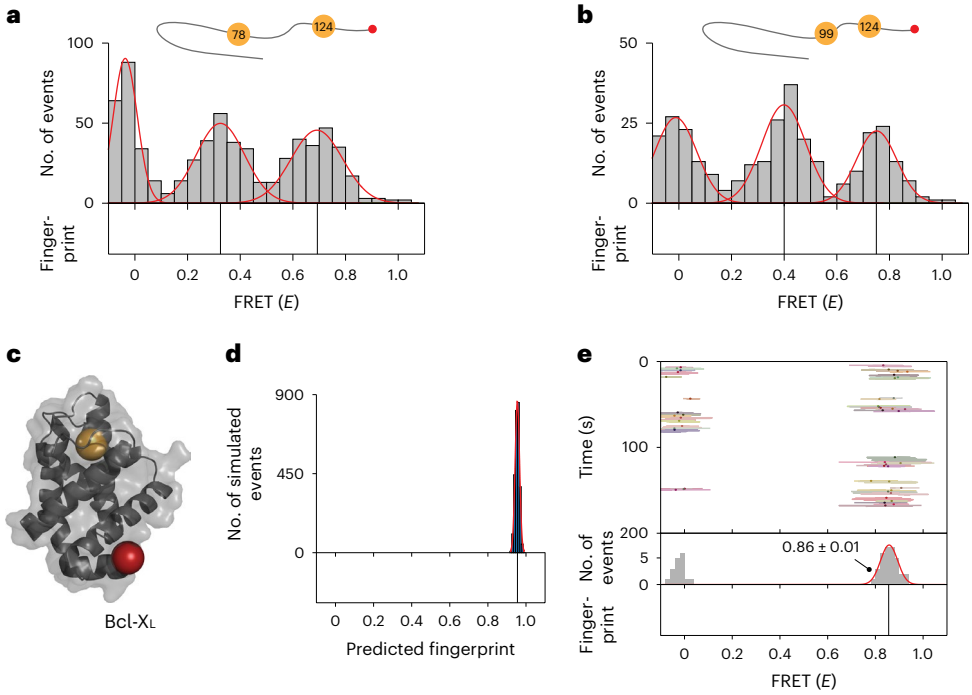

Fig. 2 | Single-molecule protein fingerprinting of disordered and folded proteins. a,**b**, Top, schematic representations of the double-cysteine variants of the aSyn model substrate. The cysteines (orange circles) are labelled with a DNA donor docking strand, and the C terminus (red circle) is labelled with a DNA acceptor docking strand. Both constructs contain a cysteine at position 124, with the second cysteine at different positions. Middle, ensemble distributions of the observed FRET events. Bottom, FRET X histograms and fingerprints reporting on the relative distances of the cysteines from the reference point. The mean ± FWHM values of the Gaussian fits for aSyn Cys 78 + Cys 124 are 0.32 ± 0.10 and 0.70 ± 0.18 (**a**) and those for aSyn Cys 99 + Cys 124 are 0.43 ± 0.14 and 0.77 ± 0.16 (**b**). **c**, Three-dimensional conformation for Bcl-X_L [PDB 1R2D] with the cysteine indicated in orange. **d**, Predicted fingerprint (mean ± FWHM) for Bcl-X_L (0.95 ± 0.03). The predicted fingerprint histograms were built from simulated FRET efficiencies from 200 individual molecules each with ten FRET events. **e**, Representative single-molecule kymograph with its determined fingerprint (mean ± s.em.) of 0.86 ± 0.01 for Bcl-X_L.

that labelling was successful. We compared the FRET efficiencies of the *O*-GlcNAcylated aSyn proteins with those that we obtained for the single-cysteine constructs (Fig. 1) and found that the FRET efficiency for *O*-GlcNAc modified residues is close to those of aSyn Cys 42 and Cys 78, suggesting that the *O*-GlcNAc is attached to residues in that region (Fig. 3d, blue shading). Consistent with these data, MS revealed that the *O*-GlcNAcylation had occurred at Thr 54 and Thr 64 (Fig. 3d, blue spheres, Supplementary Fig. 6), underscoring the predictability, accuracy and reproducibility of FRET X and the utility of FRET X for PTM mapping.

A universal approach for protein fingerprinting

Finally, we focused on bringing FRET X fingerprinting to natural proteins by circumventing recombinantly expressed tags for immobilization. Such universality is a crucial step towards analysing specific biomarkers from natural sources.

We developed a labelling approach that allows for the site-selective attachment of a bifunctional linker to the N terminus of any protein substrate, making use of the previously reported pyridinecarboxaldehyde (PCA) chemistry[36] and copper-free click chemistry (Fig. 4a). To demonstrate site-selective N-terminal modification, we fingerprinted the aSyn constructs, and as expected we observed a monotonic decrease in FRET efficiency as the distance to the N terminus increases (Fig. 4b, black squares). This shows that reliable fingerprints can be obtained for residues that are ~100 amino acids away from the reference point (Fig. 4b). By combining these fingerprints with those that were obtained with the reference point at the C terminus, we were able to probe every region of the aSyn protein (Fig. 4b and Supplementary Fig. 7).

Next, we reasoned that the ability to probe proteins from more than one reference point should increase prediction accuracy. To validate this, we generated a simulated dataset of constructs containing both C- and N-terminal reference points, by pairing the experimental fingerprints from the C- and N-terminal measurements for each construct. This allowed us to classify the seven aSyn constructs with an accuracy of >80% when combining C- and N-terminal fingerprints for the same molecule, which is higher than when a single reference point is used (Fig. 4c and Supplementary Fig. 7g,f).

To demonstrate the general applicability of FRET X, we used the N-terminal modification approach on two inflammatory disease biomarkers, the S100A9 protein, informative for severe forms of COVID-19 (Fig. 4d)[37], and procalcitonin (PCT), used for diagnosis of bacterial infections (Fig. 4e)[38]. The fingerprinting simulations predicted a single high FRET peak for S100A9 protein (Fig. 4f), which is in good agreement with our experimental findings (Fig. 4h). For PCT we anticipated that resolving the locations of the two cysteines would be challenging, as their distances to the reference point differ by only ~0.2 nm (Supplementary Fig. 7a). However, when we predicted the structure, with the DNA labels attached, using our lattice model prediction tool (Supplementary Fig. 8b,c), we observed two clearly distinguishable FRET peaks for PCT (Fig. 4g). We hypothesize that the larger difference in FRET efficiency between the cysteines is a result of the DNA docking strands increasing the resolvability. This hypothesis is further supported by the experimental data, showing two clear FRET peaks for the PCT biomarker (Cys 85 and Cys 91) (Fig. 4i and Supplementary Fig. 8d,e). We speculate that the discrepancy in peak position for the predicted and experimental fingerprints is caused by the low prediction power of AlphaFold for the intrinsically disordered regions within PCT (predicted local distance difference test (pLDDT) < 50)[39,40]. To demonstrate the ability of FRET X to identify different proteins with similar fingerprints in a mixture, we trained a classifier on experimental data and determined its protein identification accuracy. We observed a mean classification accuracy of 80%

https://doi.org/10.1038/s41565-023-01598-7

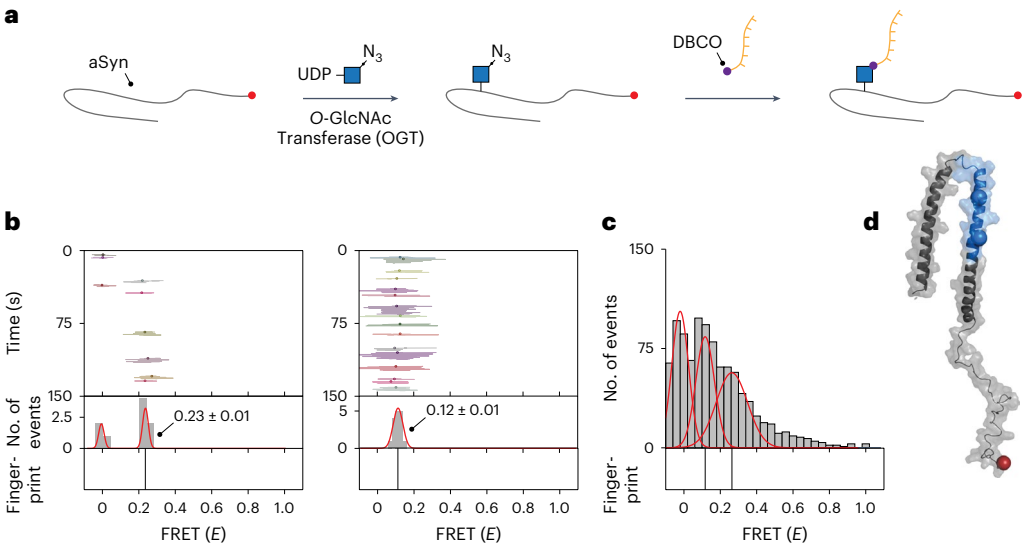

Fig. 3 | PTM mapping using FRET X. a, Schematic representation of the PTM labelling scheme. In a first step, the UDP-linked 6-azido-GlcNAc is conjugated to the aSyn substrate via the OGT enzyme. Next, the docking strands are conjugated to the *O*-GlcNAcylated aSyn protein via dibenzocyclooctyne (DBCO) click chemistry. **b**, Representative kymographs from individual aSyn molecules reporting on the distance of the *O*-GlcNAc from the reference point. The mean FRET ± s.e.m. is reported for each molecule (left panel, $n = 7$ FRET events; right panel $n = 14$ FRET events). **c**, The FRET X histogram and fingerprint for all molecules in a single field of view. We observed two FRET peaks, indicating the attachment of two *O*-GlcNAc residues on aSyn, with FRET efficiencies of 0.12 ± 0.12 and 0.23 ± 0.19. These values report mean FRET efficiency ± FWHM of the Gaussian fit. **d**, Three-dimensional conformation for micelle-bound aSyn [PDB 1XQ8] with the C terminus shown in red. The proposed region for the PTM sites on the basis of the FRET (*E*) of the cysteines probed in Fig. 1f is indicated with the blue shading, while the exact PTM locations are indicated with the blue spheres.

(Supplementary Fig. 8f) for a protein mixture of Bcl-X_L, S100A9 and aSyn A124C on the basis of single-cysteine fingerprints. This indicates that protein identification is more efficient when a database of experimental fingerprints is constructed and proteins are identified on the basis of this database.

Discussion

We have introduced FRET X, a single-molecule fingerprinting approach for protein identification and PTM mapping. FRET X enables discrimination of proteins with only subtle differences, owing to its ability to localize residues on intact proteins. Because the positional information is preserved, structures of the same mass that are generally not discernible in MS can be readily differentiated from each other by their distinct FRET efficiencies.

By using short fluorescently labelled DNA strands and their transient binding, FRET X allowed repeated examination of an amino-acid residue in a single protein, increasing the localization precision (to 5 amino acids) and thereby the overall accuracy of the protein fingerprint, reaching an identification accuracy of 84% (refs. 19,24). FRET X can fingerprint full-length proteins, such as the intrinsically disordered protein aSyn and folded proteins such as Bcl-X_L, S100A9 protein and PCT, and quantify protein abundance in mixtures. Furthermore, our FRET X fingerprinting approach benefits from the programmable and predictable kinetics of DNA hybridization, which allows for further speed optimization and for multiple target residues to be probed in sequential imaging cycles[26]. This sequential probing allows us to probe different residues (for example amino acids or PTMs) separately by flushing in orthogonal imager strands[24,41]. Such a strategy avoids crowding of the FRET spectrum, thereby allowing us to resolve the FRET fingerprint for each of the residues with high precision. We have previously shown that by probing either cysteines and lysines or cysteines, lysines and arginines the uniqueness of a protein fingerprint increases substantially, enhancing the proportion of human proteins that can be identified to 82% or 95%, respectively[19].

At the current acquisition speed, high-resolution protein fingerprints of several thousand proteins can be obtained within a few minutes, which is several orders of magnitude faster than other single-molecule fluorescence protein fingerprinting methods[22]. The ability to fingerprint full-length proteins avoids the need for additional sample preparation steps such as digestion into peptides or protein translocation, which are often required for other single-molecule protein identification approaches[12–15]. Since the average protein diameter is estimated to be 5 nm (ref. 42), the typical protein is well within the range of the Cy3–Cy5 FRET pair, while for proteins of different sizes other FRET pairs may be selected. Furthermore, we have shown that proteins can be immobilized using an N-terminal labelling strategy, thereby removing the need for genetic or synthetic tags, which opens up avenues for the analysis of proteins from natural sources (for example body fluids or single cells). While the N terminus might not always be accessible for labelling[43], the C terminus may instead be targeted[44] for conjugation of the reference and immobilization strands. Additionally, by combining N- and C-terminus reference points, we are able to expand sequence coverage within a protein and improve identification accuracy (Fig. 4c).

We further demonstrated that FRET X can be readily exploited to map *O*-GlcNAc sites of aSyn. The use of intact proteins is critical in this use case because it better mimics how an enzyme encounters a substrate in vivo, compared to synthetic protein constructs[30]. Moreover, it also takes into account the crosstalk between *O*-GlcNAc residues at adjacent sites, which is generally neglected when using peptides as substrate. Our results on aSyn *O*-GlcNAcylation were consistent with a previous report in which aSyn expressed in mammalian cells contained up to two *O*-GlcNAc residues[45]. Although we have demonstrated PTM detection using in vitro attachment of the modified *O*-GlcNAc, we envision that this approach can be further developed for in vivo analysis of *O*-GlcNAcylated protein using metabolic labelling experiments.

Using existing chemoenzymatic labelling approaches, FRET X-based PTM analysis can be expanded to detect acetylation[46], ribosylation[47] and fucosylation[48]. Additionally, by combining endoglycosidases and galactosyltransferases, click handles[49,50] can be incorporated into *N*-glycans to allow for DNA labelling and analysis of full-length

https://doi.org/10.1038/s41565-023-01598-7

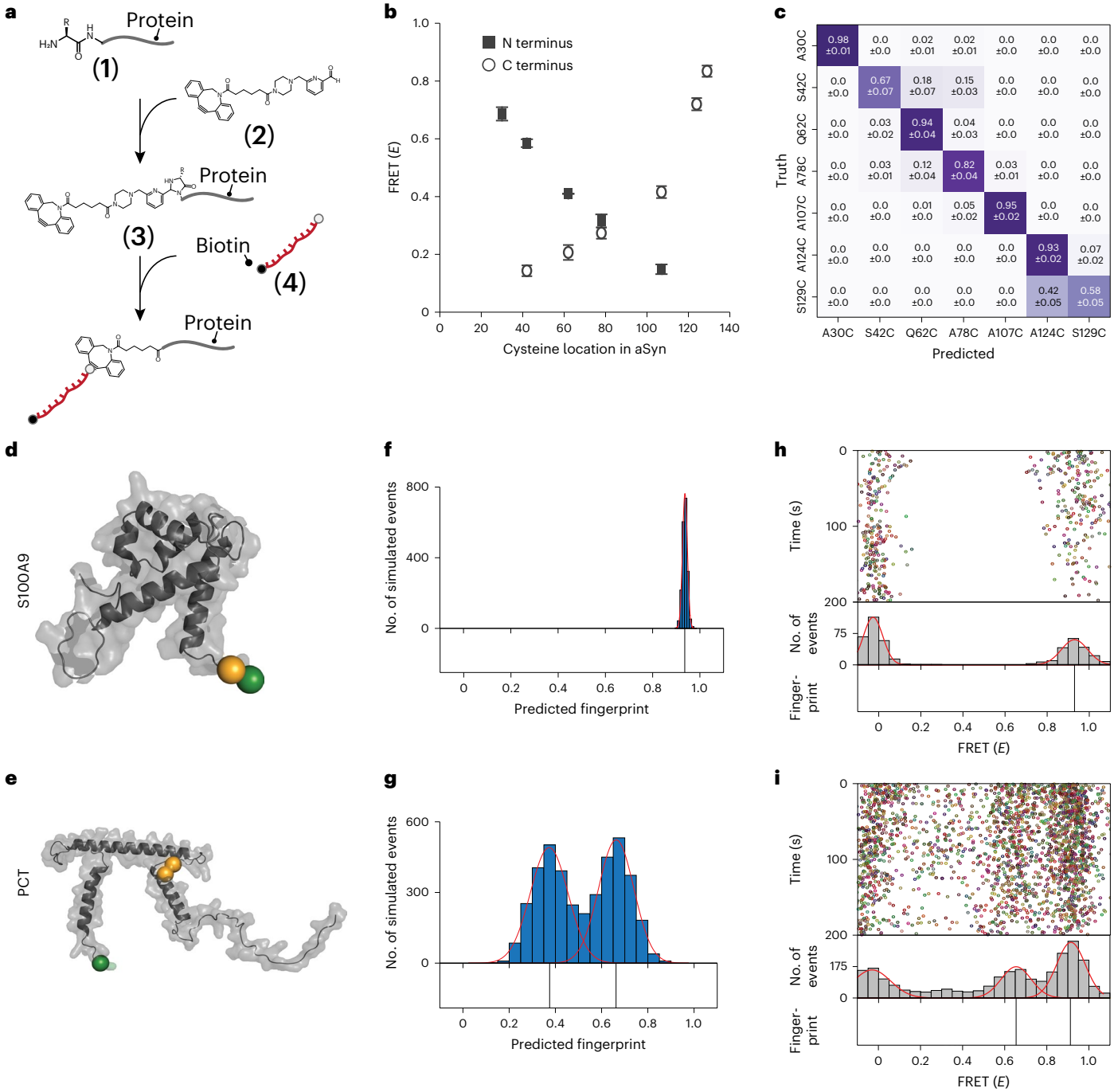

Fig. 4 | Single-molecule protein fingerprinting using N-terminal labelling. **a**, Schematic representation of the labelling procedure. The N terminus of any target protein (**1**) is labelled with a 2PCA–DBCO derivative (**2**). The product of this reaction is a protein that has its N terminus functionalized with a unique DBCO group (**3**) that allows for the attachment of a biotinylated DNA reference point (**4**). **b**, The FRET efficiency as a function of the location of the cysteine in aSyn. A monotonic decrease in FRET efficiency is observed for the cysteine relative to the N terminus (grey squares). The values are reported as mean ± s.d. of three independent experiments. **c**, Support vector classifier performance on fingerprints for seven aSyn mutants probed from both termini. Heatmaps showing how each mutant is classified (mean ± s.d. over ten cross-validation folds), whereby the diagonal positions indicate correct classifications. Classifiers were trained and tested on experimental data from separate experiments, conducted on different days to avoid batch effects. **d**,**e**, Three-dimensional structures of two inflammatory disease biomarkers, S100A9 [AlphaFold: AF-P06702-F1] (**d**) and PCT [AlphaFold:AF-P01258-F1] (**e**) with the cysteines (orange spheres) and N terminus (green sphere) highlighted. **f**,**g**, The predicted FRET fingerprints (mean ± FWHM) for S100A9 protein (**f**, 0.94 ± 0.02) and PCT (**g**, 0.37 ± 0.19 and 0.66 ± 0.18). **h**,**i**, Experimentally obtained fingerprints reporting on the location of the cysteines relative to the N-terminal reference point. **h**, Ensemble kymograph for S100A9 protein with a single high FRET peak (0.93 ± 0.15). **i**, Two high FRET peaks (0.65 ± 0.17 and 0.91 ± 0.16), both mean ± FWHM, obtained for PCT.

glycoproteins. Furthermore, PTM detection using FRET X may go beyond metabolic labelling using other chemical biology strategies to attach orthogonal DNA docking strands in phosphorylation[22,23] or lipidation[51].

One of the main challenges for single-molecule proteomics lies in the varying abundances of different protein species in the cell. The dynamic range of the proteome spans several orders of magnitude[1,52], due to which low-abundance species can easily be masked by more

abundant ones. Owing to its single-molecule sensitivity and the ability to fingerprint several thousand proteins in a single field of view, our method detects even the sparsest proteins, and future optimizations including automated acquisition and scanning stages might increase throughput and thereby sensitivity even further. Alternatively, we may address the challenge posed by the large dynamic range by adopting protein enrichment strategies for a targeted approach[53]. In the current study, less than a femtomole of labelled protein was needed for fingerprinting, but sensitivity may increase by one or two additional orders of magnitude using existing microfluidics technology.

To conclude, we envision that our full-length single-molecule protein fingerprinting approach may allow researchers and analysts to finally investigate proteoform variety with ultimate sensitivity.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41565-023-01598-7.

References

1. Aebersold, R. et al. How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
2. Kim, H. K., Pham, M. H. C., Ko, K. S., Rhee, B. D. & Han, J. Alternative splicing isoforms in health and disease. *Pflügers Arch.* **470**, 995–1016 (2018).
3. Paronetto, M. P., Passacantilli, I. & Sette, C. Alternative splicing and cell survival: from tissue homeostasis to disease. *Cell Death Differ.* **23**, 1919–1929 (2016).
4. Lin, H. & Caroll, K. S. Introduction: posttranslational protein modification. *Chem. Rev.* **118**, 887–888 (2018).
5. Carbonara, K., Andonovski, M. & Coorssen, J. R. Proteomes are of proteoforms: embracing the complexity. *Proteomes* **9**, 38 (2021).
6. Benson, M. D., Ngo, D., Ganz, P. & Gerszten, R. E. Emerging affinity reagents for high throughput proteomics: trust, but verify. *Circulation* **140**, 1610–1612 (2019).
7. Yang, Y. et al. Hybrid mass spectrometry approaches in glycoprotein analysis and their usage in scoring biosimilarity. *Nat. Commun.* **7**, 13397 (2016).
8. Čaval, T., Tian, W., Yang, Z., Clausen, H. & Heck, A. J. R. Direct quality control of glycoengineered erythropoietin variants. *Nat. Commun.* **9**, 3342 (2018).
9. Siuti, N. & Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **410**, 817–821 (2007).
10. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
11. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **46**, 2159–2168 (2018).
12. Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nat. Nanotechnol.* **13**, 786–796 (2018).
13. Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* **18**, 604–617 (2021).
14. Floyd, B. M. & Marcotte, E. M. Protein sequencing, one molecule at a time. *Annu. Rev. Biophys.* **51**, 181–200 (2022).
15. Timp, W. & Timp, G. Beyond mass spectrometry, the next step in proteomics. *Sci. Adv.* **6**, eaax8978 (2020).
16. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11**, e1004080 (2015).
17. Rodriques, S. G., Marblestone, A. H. & Boyden, E. S. A theoretical analysis of single molecule protein sequencing via weak binding spectra. *PLoS ONE* **14**, e0212868 (2019).
18. Yao, Y., Docter, M., Van Ginkel, J., De Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: computational assessment. *Phys. Biol.* **12**, 10–16 (2015).
19. de Lannoy, C. V. et al. Evaluation of FRET X for single-molecule protein fingerprinting. *iScience* **24**, 103239 (2021).
20. Yu, L. et al. Unidirectional single-file transport of full-length proteins through a nanopore. *Nat. Biotechnol.* **41**, 1130–1139 (2023).
21. van Ginkel, J. et al. Single-molecule peptide fingerprinting. *Proc. Natl Acad. Sci. USA* **115**, 3338–3343 (2018).
22. Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **36**, 1076–1082 (2018).
23. Shrestha, P. et al. Single-molecule mechanical fingerprinting with DNA nanoswitch calipers. *Nat. Nanotechnol.* **16**, 1362–1370 (2021).
24. Filius, M., Kim, S. H., Severins, I. & Joo, C. High-resolution single-molecule FRET via DNA exchange (FRET X). *Nano Lett.* **21**, 3295–3301 (2021).
25. Filius, M., van Wee, R. & Joo, C. in *Single Molecule Analysis: Methods and Protocols* (eds Heller, I. et al.) 203–213 (Springer, 2024).
26. Van Wee, R., Filius, M. & Joo, C. Completing the canvas: advances and challenges for DNA-PAINT super-resolution imaging. *Trends Biochem. Sci.* **11**, 918–930 (2021).
27. Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. *Nat. Protoc.* **12**, 1198–1228 (2017).
28. Shi, X. et al. Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. *Nat. Methods* **9**, 499–503 (2012).
29. Schuler, B. & Hofmann, H. Single-molecule spectroscopy of protein folding dynamics—expanding scope and timescales. *Curr. Opin. Struct. Biol.* **23**, 36–47 (2013).
30. Yang, X. & Qian, K. Protein *O*-GlcNAcylation: emerging mechanisms and functions. *Nat. Rev. Mol. Cell Biol.* **18**, 452–465 (2017).
31. Vellosillo, P. & Minguez, P. A global map of associations between types of protein posttranslational modifications and human genetic diseases. *iScience* **24**, 102917 (2021).
32. Mauri, T. et al. *O*-GlcNAcylation prediction: an unattained objective. *Adv. Appl. Bioinform. Chem.* **14**, 87–102 (2021).
33. Shi, J., Ruijtenbeek, R. & Pieters, R. J. Demystifying *O*-GlcNAcylation: hints from peptide substrates. *Glycobiology* **28**, 814–824 (2018).
34. Shen, D. L. et al. Catalytic promiscuity of *O*-GlcNAc transferase enables unexpected metabolic engineering of cytoplasmic proteins with 2-azido-2-deoxy-glucose. *ACS Chem. Biol.* **12**, 206–213 (2017).
35. Mayer, A., Gloster, T. M., Chou, W. K., Vocadlo, D. J. & Tanner, M. E. 6′-Azido-6′-deoxy-UDP-*N*-acetylglucosamine as a glycosyltransferase substrate. *Bioorg. Med. Chem. Lett.* **21**, 1199–1201 (2011).
36. Macdonald, J. I., Munch, H. K., Moore, T. & Francis, M. B. One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes. *Nat. Chem. Biol.* **11**, 326–331 (2015).
37. Wang, S. et al. S100A8/A9 in inflammation. *Front. Immunol.* **9**, 1298 (2018).
38. Vijayan, A. L. et al. Procalcitonin: a promising diagnostic marker for sepsis and antibiotic therapy. *J. Intensive Care* **5**, 51 (2017).
39. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

40. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
41. Jungmann, R. et al. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nat. Methods* **11**, 313–318 (2014).
42. Erickson, H. P. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol. Proced. Online* **11**, 32–51 (2009).
43. Ree, R., Varland, S. & Arnesen, T. Spotlight on protein N-terminal acetylation. *Exp. Mol. Med.* **50**, 1–13 (2018).
44. Bloom, S. et al. Decarboxylative alkylation for site-selective bioconjugation of native proteins via oxidation potentials. *Nat. Chem.* **10**, 205–211 (2018).
45. Ramirez, D. H. et al. Engineering a proximity-directed *O*-GlcNAc transferase for selective protein *O*-GlcNAcylation in cells. *ACS Chem. Biol.* **15**, 1059–1066 (2020).
46. Yang, Y.-Y., Ascano, J. M. & Hang, H. C. Bioorthogonal chemical reporters for monitoring protein acetylation. *J. Am. Chem. Soc.* **132**, 3640–3641 (2010).
47. Westcott, N. P., Fernandez, J. P., Molina, H. & Hang, H. C. Chemical proteomics reveals ADP-ribosylation of small GTPases during oxidative stress. *Nat. Chem. Biol.* **13**, 302–308 (2017).
48. Rabuka, D., Hubbard, S. C., Laughlin, S. T., Argade, S. P. & Bertozzi, C. R. A chemical reporter strategy to probe glycoprotein fucosylation. *J. Am. Chem. Soc.* **128**, 12078–12079 (2006).
49. Boeggeman, E. et al. Direct identification of nonreducing GlcNAc residues on *N*-glycans of glycoproteins using a novel chemoenzymatic method. *Bioconjugate Chem.* **18**, 806–814 (2007).
50. van Geel, R. et al. Chemoenzymatic conjugation of toxic payloads to the globally conserved *N*-glycan of native mAbs provides homogeneous and highly efficacious antibody–drug conjugates. *Bioconjugate Chem.* **26**, 2233–2242 (2015).
51. Tate, E. W., Kalesh, K. A., Lanyon-Hogg, T., Storck, E. M. & Thinon, E. Global profiling of protein lipidation using chemical proteomic technologies. *Curr. Opin. Chem. Biol.* **24**, 48–57 (2015).
52. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteom.* **1**, 845–867 (2002).
53. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12**, 483–490 (2008).

Methods

Protein expression and purification

All proteins were codon optimized for *Escherichia coli* BL21(DE3) and inserted into a pET-52b(+) vector for aSyn or pET-15b for BCL2L1-X_L (see Supplementary Table 1 for full list of protein sequences). All proteins were engineered to contain a C-terminal aldehyde encoding sequence. The cysteine in this motif is converted in vivo into formylglycine by co-expression of the formylglycine-generating enzyme[28]. The plasmids and protein encoding genes were synthesized and prepared using GenScript.

The proteins were expressed in *E. coli* BL21(DE3) cells. Cultures were grown at 37 °C in lysogeny broth medium supplemented with 50 μg ml^{-1} kanamycin and 50 μg ml^{-1} ampicillin until an optical density at 600 nm of 0.5 was reached. Expression of the formylglycine-generating enzyme was induced through addition of 1% L-arabinose at 37 °C, and after 30 min the expression of the model proteins was induced by 1 mM isopropyl-β-D-thiogalactoside (IPTG). The cultures were transferred to 26 °C to allow for expression of the proteins for 5 h, after which the cells were collected at 4,000 *g*. The cells were lysed by resuspending the pellet in 10 ml lysis buffer (50 mM HEPES–KOH pH 7.5, 500 mM NaCl, 0.5% Triton X-100). For the aSyn proteins, the cells were lysed by boiling the cell suspension for 15 min. The cells containing BCL2L1-X_L proteins were lysed by tumbling the cell suspension for 2 h at room temperature, followed by sonication on ice during six cycles of 30 s on and 1 min off at 30% amplitude. Next, the cell lysate of each model protein was centrifuged at 30,000 *g* for 30 min at 4 °C. The proteins were purified from the cell-free extract using HisPur Ni-NTA (nitrilotriacetic acid) resin according to the manufacturer's manual and buffer exchanged into storage buffer (50 mM HEPES–KOH pH 7.5, 150 mM NaCl, 10% glycerol, 25 mM tris(2-carboxyethyl)phosphine, TCEP) using 10 kDa Amicon Ultra centrifugal filters. All proteins were aliquoted and stored at −80 °C.

Cysteine labelling

Proteins were labelled without chemical, temperature or mechanical denaturation to preserve their structure. First, cysteine residues were reduced with 40-fold molar excess TCEP for 30 min and then labelled with 25-fold molar excess monoreactive maleimide–DBCO in 50 mM HEPES pH 7.5, 150 mM NaCl, 1% Triton X-100 buffer overnight at room temperature (23 ± 1 °C). Excess maleimide–DBCO and TCEP were removed with Zeba spin desalting columns, 7 kDa molecular weight cut-off (Thermo Fisher), and the reaction buffer was changed to 0.3 M NaAc pH 5.5 for the aSyn proteins and 50 mM HEPES pH 6.9, 150 mM NaCl, 1% Triton X-100, for Bcl-X_L. Then monoreactive azidobenzoate-(5′) functionalized DNA was added in 10-fold molar excess and incubated overnight at room temperature. The formylglycine residues were acceptor-labelled with 10-fold excess biotinylated and hydrazide-functionalized DNA for 46 h at room temperature for aSyn and 96 h at 4 °C for Bcl-X_L in a rotary shaker. Free hydrazide–DNA–biotin was removed with Ni-NTA magnetic agarose beads (Qiagen) according to the manufacturer's protocol. See Supplementary Table 2 for the full list of substrates.

O-GlcNAcylation labelling

The aSyn constructs were *O*-GlcNAcylated using a recombinant human OGT (Novus Biologicals). The reaction was performed by adding a 20-fold excess of aSyn over OGT in a reaction buffer (25 mM Tris pH 7.5, 10 mM $CaCl_2$, 10 mM $MgCl_2$) supplemented with 10 mM UDP-azido-*O*-GlcNAc, and incubated overnight at 37 °C. The next day, excess UDP-azido-*O*-GlcNAc was removed with Zeba spin desalting columns and the reaction buffer was changed to 0.3 M NaAc pH 5.5. The *O*-GlcNAc residues were labelled with 10-fold excess DBCO-functionalized DNA and the formylglycine residue was labelled with 10-fold excess biotinylated and hydrazide-functionalized DNA for 48 h at room temperature (23 ± 1 °C) in a rotary shaker. Free DNA was removed with Ni-NTA magnetic agarose beads (Qiagen) according to the manufacturer's protocol.

N-terminal modification

To make a 2PCA–DBCO bifunctional linker, we incubated 100 mM of a 2PCA intermediate (6-(piperazin-1-ylmethyl)-2-pyridinecarboxaldehyde HCl salt, Sigma Aldrich 808571) with twofold excess *N*-hydroxysuccinimide–DBCO and threefold excess of triethylamine in dimethylsulfoxide for 24 h at room temperature, while shaking. Next, we quenched the *N*-hydroxysuccinimide by adding tenfold excess dimethylamine and incubated for 4 h at room temperature. The reaction mixture was dried using a SpeedVac and dissolved in dimethylsulfoxide to a concentration of 100 mM 2PCA–DBCO.

The target proteins (aSyn, human recombinant PCT (Bio-Techne) and human recombinant S100A9 protein (Novus Biologicals)) were dissolved in PBS at a concentration of 5 μM. To this solution, we added 400-fold excess of the 2PCA–DBCO linker and incubated for 24 h at 37 °C while shaking. The next day, free 2PCA–DBCO was removed with Zeba spin desalting columns, 7 kDa molecular weight cut-off. Next the proteins were labelled with twofold excess biotinylated and azide-functionalized DNA for 48 h at room temperature (23 ± 1 °C) in a rotary shaker. Free DNA was removed with Ni-NTA magnetic agarose beads (Qiagen) according to the manufacturer's protocol. Finally, the eluted proteins were reduced with 40-fold molar excess TCEP for 30 min and then labelled with 20-fold molar excess monoreactive maleimide–DNA for 24 h.

Single-molecule set-up

All experiments were performed on a custom-built microscope set-up. An inverted microscope (IX73, Olympus) with prism-based total internal reflection was used, in combination with a 532 nm diode-pumped solid-state laser (Compass 215M/50 mW, Coherent). A ×60 water immersion objective (UPLSAPO60XW, Olympus) was used for the collection of photons from the Cy3 and Cy5 dyes on the surface, after which a 532 nm long-pass filter (LDP01-532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal, which is then projected onto an electron-multiplying CCD (charge-coupled device) camera (iXon Ultra, DU-897U-CS0-# BV, Andor Technology). Our pixel size is 107 nm × 107 nm and the complete field of view is 512 pixels × 256 pixels (54.8 μm × 27.4 μm) and contains roughly 200 molecules. A series of electron-multiplying CCD images was recorded using a custom-made program in Visual C++ (Microsoft).

Single-molecule data acquisition

Single-molecule flow cells were prepared as previously described[54]. In brief, to avoid non-specific binding, quartz slides (Finkenbeiner) were acidic piranha etched and PEGylated twice. The first round of PEGylation was performed with methoxy PEG–succinimidyl valerate (Laysan Bio) and PEG–biotin (Laysan Bio), followed by a second round of PEGylation with MS(PEG)4 (Thermo Fisher). After assembly of a microfluidic chamber, the slides were incubated with 20 μl of 0.1 mg ml^{-1} streptavidin (Thermo Fisher) for 2 min. Excess streptavidin was removed with 100 μl T50 (50 mM Tris-HCl, pH 8.0, 50 mM NaCl). Next, 50 μl of 75 pM DNA-labelled protein was added to the microfluidic chamber. After 2 min of incubation, unbound protein was washed away with 200 μl T50. Then, 50 μl of 10 nM donor-labelled imager strands and 50 nM acceptor-labelled imager strands in imaging buffer (50 mM Tris-HCl, pH 8.0, 500 mM NaCl, 0.8% glucose, 0.5 mg ml^{-1} glucose oxidase (Sigma), 85 μg ml^{-1} catalase (Merck) and 1 mM 6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid (Sigma)) was injected. All single-molecule FRET experiments were performed at room temperature (23 ± 1 °C). See Supplementary Table 2 for the full list of docking and imager strands.

Data analysis

Fluorescence signals are collected at 0.1 s exposure time. During the acquisition of the video, a green laser is used to excite the Cy3 donor

fluorophores. The fluorescence images were analysed using a custom script written in IDL. The script collects the individual intensity hotspots in the acceptor channel and pairs them with intensity hotspots in the donor channel, after which the time traces are extracted. The details of the automated detection of individual imager strand binding events from the fluorescence time traces are described elsewhere[24,55]. Briefly, a two-state *k*-means clustering algorithm was applied to the sum of the donor and acceptor fluorescence intensities of individual molecules to find an intensity threshold, with which the traces were divided into high- or low-intensity segments. The high-intensity segments that lasted for more than three consecutive frames were selected for further analysis. If there were abrupt donor and acceptor intensity changes within a high-intensity segment, possibly due to photobleaching or imager dissociation, the data points that came after the transition moment were removed from the segment. The gamma and leakage factors were determined from acceptor bleaching events and donor-only events, respectively[56]. Average FRET efficiencies from each selected segment were used to build the FRET kymograph and histogram. Populations in the FRET histogram are automatically classified using a GMM. GMMs containing one to five distributions are fitted, after which the best-fitting GMM is selected using the Bayesian information criterion. Peaks with weights lower than 0.2 are discarded, as these were found to capture background noise. The automated analysis code in Python is freely available at https://github.com/kahutia/transient_FRET_analyzer2.

aSyn mutant classification

aSyn mutant molecules were filtered to retain only those for which the FRET histogram contained one or no donor-only peak (E(FRET) < 0.1) and one FRET peak (0.1 < E(FRET) < 9.0) to remove junk molecules and aggregates (molecules that contain multiple FRET states). FRET values for these molecules were used to train and test an SVM classifier, a commonly used machine learning approach[57]. Briefly, training an SVM classifier automates the definition of class boundaries in a feature space (here FRET efficiency), choosing boundaries such that training examples with known classification are correctly classified most often. Distance between samples may be determined using an arbitrary function, the kernel function, which effectively transforms feature space before linear class boundaries are drawn. In the kernel function, the gamma parameter (γ) determines how flexible the boundaries are allowed to be to accommodate the training data. Here we used an SVM with radial basis function implemented in the scikit-learn package (v.1.2.1)[58]:

$$K(x_1, x_2) = \exp(-\gamma \times \|x_1 - x_2\|^2)$$

where γ was set to the inverse of the number of features.

Single-feature SVM classifiers were trained and tested on FRET values acquired from experiments on samples containing a single mutant, as the mutant—the class in machine learning terms—for these samples is known. This was done separately for mutants measured from the C terminus and from the N terminus. To simulate experiments in which molecules were measured from both ends, FRET values from C-terminally and N-terminally measured molecules were randomly paired, again taking care to keep test and training experiments separated. A two-feature SVM was then trained and tested on these data.

Finally, equimolar mixtures of four aSyn mutants (S42C, A78C, A107C and A124C) from five separate experiments were classified using an SVM classifier trained on all classified (that is, derived from single-mutant samples) data. Raw data, intermediate data and analysis code are freely available and documented at https://github.com/cvdelannoy/FRET_X_fingerprinting_simulation/tree/main/FRET_X_proteoform_fingerprinting.

Protein fingerprint simulation

DNA-labelled protein fingerprinting simulations were performed using lattice models, previously described[19]. The procedure starts with a fully atomistic native structure predicted by AlphaFold2 (refs. 39,40), which is converted to a lattice structure with tagged residues marked. On the basis of residue interactions and secondary structure formations, this structure is assigned a modelling energy, which is then minimized using a Markov chain Monte Carlo process, by repeatedly applying random perturbations to the structure and accepting or rejecting them on the basis of the incurred energy change. Further Markov chain Monte Carlo iterations are used to generate hundreds of slightly different structures, from which distances between donor and acceptor dye positions are deduced. These values are then translated to FRET efficiencies E_{FRET} as follows:

$$E_{\mathrm{FRET}} = \frac{1}{1 + (R/R_0)^6}.$$

Here R is the modelled inter-dye distance and R_0 is the Förster radius, which characterizes the used FRET dye pair (R_0 assumed constant at 54 Å for the Cy3–Cy5 FRET pair). Simulation and analysis code for the protein fingerprints are freely available at https://github.com/cvdelannoy/FRET_X_fingerprinting_simulation, while simulation data are available at https://git.wur.nl/lanno001/fret_x_proteoform_sim_data.

In-gel proteolytic digestion using Glu-C

For proteomic analysis of the *O*-GlcNAc-modified aSyn proteins, we performed a conventional SDS–PAGE followed by in-gel proteolytic digestion and mass spectrometric analysis as previously described[59]. The aSyn proteins were analysed using a 4–12% NuPAGE Bis–Tris (Invitrogen) gel and stained with Instant Blue protein stain (Sigma) according to the manufacturer's instructions. The stained gel bands were cut from the gel and destained using Coomassie destaining solution (100 mM ammonium bicarbonate buffer in 40% acetonitrile) for 15 min at 300 r.p.m. at 37 °C. The supernatant was removed and the gel pieces were dehydrated using acetonitrile for 10 min at room temperature. The supernatant was removed and the dehydrated protein-containing gel pieces were reduced using 200 µl reducing reagent solution (10 mM dithiothreitol) for 30 min at 56 °C. Next, the supernatant was removed and the samples were cooled to room temperature and alkylated for 30 min at room temperature using 200 µl alkylation reagent (55 mM iodoacetamide in ammonium bicarbonate buffer). After this, the alkylation solution was removed and the samples were washed with 200 µl of Coomassie destaining solution for 5 min at room temperature on a shaker. The supernatant was removed, and the samples were dehydrated using 200 µl acetonitrile for 10 min. Finally, 2 µl Glu-C protease stock solution (100 ng µl^{-1} in H_2O, Pierce, MS grade) was mixed with 98 µl 100 mM ammonium bicarbonate buffer, added to the dehydrated gel pieces and incubated overnight at 37 °C under gentle shaking (300 r.p.m.). The next day, the supernatant of each digest was collected and 150 µl of extraction solution was added to each sample and incubated for 15 min at 37 °C. The supernatant was combined with the first fraction, and 100 µl of acetonitrile was added and incubated for 15 min at 37 °C, and this extract was again combined with the earlier fractions. Finally, 100 µl of 10:90 acetonitrile:H_2O was added to each sample, incubated for 15 min at 37 °C and combined with the earlier factions in the new Eppendorf tube. The pooled extracts from every sample were then dried using a SpeedVac concentrator at elevated temperature (50–60 °C).

β-elimination

To approximately 10 µl of glycopeptide/peptide extract, 300 µl of 26% dimethylamine solution was added. The reaction was carried out at 55 °C for 6 h under careful mixing, and subsequently stopped by removing the reagent under vacuum. The residue was dissolved in 150 µl Milli-Q water and stored at −20 °C before analysis.

Proteomic analysis

The SpeedVac-dried peptide fractions (digested or additionally beta-eliminated) were resuspended in H_2O containing 3% acetonitrile and 0.01% trifluoroacetic acid under careful vortexing. An aliquot corresponding to approximately 100 ng digest was analysed using a one-dimensional shotgun/parallel reaction monitoring proteomics approach. Briefly, samples were analysed using a nano-liquid-chromatography separation system consisting of an EASY nano LC 1200, equipped with an Acclaim PepMap RSLC RP C18 separation column (50 µm × 150 mm, 2 µm and 100 Å), and a QE Plus Orbitrap mass spectrometer (Thermo, Germany). The flow rate was maintained at 350 nl min^{-1} with solvent A H_2O containing 0.1% formic acid, and solvent B 80% acetonitrile in H_2O and 0.1% formic acid. Either a short gradient, consisting of a linear increase of solvent B from 5 to 30% within 38 min and finally to 60% over 15 min, or a longer gradient, consisting of a linear increase of solvent B from 5 to 25% over 88 min and to 55% over an additional 60 min, was used. In either case, the Orbitrap was operated in data-dependent acquisition mode acquiring spectra at 70,000 resolution from 385 to 1,250 m/z (mass to charge ratio), where the top ten signals were isolated with a window of 2.0 m/z for fragmentation using a normalized collision energy of 28. Fragmentation spectra were acquired at 17,000 resolution, with an automated gain control target of 2×10^5, at a maximum injection time of 75 ms. Unassigned, singly charged, ×6 and higher charge states were excluded from fragmentation. Alternatively, additional (confirmatory) parallel reaction monitoring scans were included targeting the potentially HexNAc-modified peptides (inclusion list m/z = 621.3, 519.8, 705.15 and 654.35).

Processing of mass spectrometric raw data

Mass spectrometric raw data were analysed using PEAKS Studio X (Bioinformatics Solutions) allowing 20 ppm parent ion and 0.02 Da fragment ion mass error tolerance, considering three missed cleavages, carbamidomethylation as fixed and methionine oxidation and asparagine/glutamine deamidation as variable modifications. The mass spectrometric raw data were furthermore analysed using a protein sequence database covering the aSyn protein sequence (synthetic, AGJ51950.1) and the Global Proteome Machine common Repository of Adventitious Proteins contaminant protein sequences (https://www.thegpm.org/crap/). Additionally, decoy fusion was used to estimate false discovery rates. Peptide spectrum matches were filtered against 1% false discovery rate and a minimum of two unique peptides per protein. Relative protein abundances were correlated to protein molecular weight normalized spectral counts. *O*-HexNac modified peptides were identified by including HexNAc (+203.08 Da) modifications in the variable modification search. Serine and threonine modification sites were determined by β-elimination, where the β-elimination products were determined by including dehydration (−18.01 Da) as variable modification search. Correct annotation of HexNAc-modified peptides and β-elimination sites was ensured by additional manual investigation of identified spectra—for example, by confirming the presence of the HexNAc oxonium ion (20.4.0872 m/z) and confirming the presence of the respective y/b peptide fragment ions from data-dependent acquisition and additional parallel reaction monitoring experiments.

Synthesis of UDP-linked 6-azido-GlcNAc

6-Azido-6-deoxy-*N*-acetyl-glucosamine-1-phosphate disodium salt was prepared as previously reported (508 mg, 1.24 mmol, 13.4% yield over eight steps)[60]. The monophosphate was dissolved in MeOH (28 ml) and acidified to pH 5–6 by addition of Dowex-H^+ resin. Resin was removed by filtration, and triethylamine (4 ml) and H_2O (12 ml) were added. The mixture was stirred at room temperature for 18 h and then the solvents were evaporated to obtain the triethylammonium salt of 6-azido-6-deoxy-*N*-acetyl-glucosamine-1-phosphate. To this compound was added trioctylamine (2.48 mmol, 1.08 ml), and the mixture was co-evaporated with pyridine (3 × 3 ml). UMP-morpholidate (1.38 g, 1.98 mmol) was added and the mixture was co-evaporated again with pyridine (3 × 3 ml). The mixture was diluted with pyridine to a total volume of 12 ml, tetrazole (347.2 mg, 4.96 mmol) was added and the resulting reaction mixture was stirred at room temperature for 3 d. The reaction mixture was concentrated in vacuo after no more starting material was observed by thin-layer chromatography analysis (ethyl acetate:MeOH:H_2O; 4:2:1 v/v/v, staining with 10% H_2SO_4 in MeOH followed by charring). The crude product was purified by flash silica column chromatography (ethyl acetate:MeOH:H_2O, 4:2:1 v/v/v) and fractions containing carbohydrate were identified, pooled and concentrated in vacuo. The resulting solid was dissolved in a minimal amount of H_2O, loaded on a Bio-Gel P2 size exclusion column and eluded with H_2O. Fractions containing carbohydrate were identified, pooled and lyophilized, resulting in a white crystalline solid (217 mg, 0.34 mmol, 27% overall yield). ^{1}H NMR (600 MHz, D_2O) δ 7.97 (d, J = 8.0 Hz, 1H), 6.00–5.95 (m, 2H), 5.53–5.49 (m,1H), 4.36 (dd, J = 5.3, 4.5 Hz, 2H), 4.29 (m, 1H), 4.28–4.22 (m, 2H), 4.09–4.00 (m, 2H), 3.80 (app.t, J = 9.8 Hz, 1H), 3.75 (dd, J = 4.0, 2.0 Hz, 1H), 3.58 (app.d, J = 9.1 Hz, 2H), 2.07 (s, 3H). ^{31}P NMR (400 MHz, D_2O): d −11.55 (d, J = 21.4 Hz), −13.39 (d, J = 21.4 Hz). Electrospray ionization time of flight MS m/z calculated for $C_{17}H_{25}N_6O_{16}P_2$ $(M-H)^-$ exact 631.0802, found 631.0306.

Statistics and reproducibility

No data were manually excluded from the analyses; however, data for some individual molecules have been rejected on quality, number of events (fewer than ten) and FRET efficiency during data processing and filtering, as described in 'Data analysis' and 'aSyn mutant classification'. These filtering steps can be reproduced using the code provided for those sections. The processed datasets were sufficiently large (~200–600 molecules) to ensure that distribution parameters could be determined with statistical significance. In classification tasks, the classifier was tested on sample data acquired from experiments different from those that produced training data, conducted on different days to avoid batch effects. Bootstrapping and tenfold cross validation were used to determine confidence intervals and prediction intervals respectively.

Data availability

The data supporting the main finding of this study are available at the publicly accessible online repository Zenodo with identifier https://doi.org/10.5281/zenodo.10179066. Any additional data are available from the corresponding author upon request.

Code availability

The algorithms for the codes supporting the main findings of this study are available at https://doi.org/10.5281/zenodo.10156504. Any additional information concerning the code is available from the corresponding author upon request.

References

54. Filius, M. et al. High-speed super-resolution imaging using protein-assisted DNA-PAINT. *Nano Lett.* **20**, 2264–2270 (2020).
55. Kim, S. H., Kim, H., Jeong, H. & Yoon, T. Y. Encoding multiple virtual signals in DNA barcodes with single-molecule FRET. *Nano Lett.* **21**, 1694–1701 (2021).
56. McCann, J. J., Choi, U. B., Zheng, L., Weninger, K. & Bowen, M. E. Optimizing methods to recover absolute FRET efficiency from immobilized single molecules. *Biophys. J.* **99**, 961–970 (2010).
57. Cristianini, N. & Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, 2000).
58. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

59. Pabst, M. et al. A general approach to explore prokaryotic protein glycosylation reveals the unique surface layer modulation of an anammox bacterium. *ISME J.* **16**, 346–357 (2022).
60. Chuh, K. N., Zaro, B. W., Piller, F., Piller, V. & Pratt, M. R. Changes in metabolic chemical reporter structure yield a selective probe of *O*-GlcNAc modification. *J. Am. Chem. Soc.* **136**, 12283–12295 (2014).

Acknowledgements

C.J. and D.d.R. acknowledge funding from NWO-I (SMPS). C.J. acknowledges funding from NWO (Vici, VI.C.202.015), the Basic Science Research Program (NRF-2023R1A2C2004745) and Frontier 10-10 (Ewha Womans University). Y.W. was funded by a Chinese Scholarship Council (CSC) grant.

Author contributions

M.F. and C.J. initiated and designed the project. M.F. designed and performed the protein labelling procedures. M.F. and R.v.W. performed the single-molecule FRET X experiments. I.W. and C.d.A.P. expressed and purified the proteins. C.d.L. wrote the software for and performed the fingerprinting predictions and protein classification. D.d.R. supervised the fingerprinting prediction simulations. S.H.K. wrote the automated peak-finding code for single-molecule FRET X analysis. M.F. and Z.L. conceptualized the *O*-GlcNAc site mapping, and designed the chemoenzymatic and N-terminal labelling strategies. Y.W. and G.-J.B. synthesized UDP-azido-*O*-GlcNAc for protein *O*-GlcNAcylation. M.P. performed proteomic analysis. M.F., S.H.K., C.d.L. and C.J. analysed and discussed the data. M.F., R.v.W. and C.J. prepared the initial draft of the manuscript. All authors read and approved the manuscript.

Competing interests

C.J., M.F., C.d.L. and D.d.R. hold a patent on single-molecule FRET for protein characterization (patent number WO2021049940). C.J., M.F. and Z.L. have filed a patent for the bifunctional linker for N-terminal protein modification. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41565-023-01598-7.

Correspondence and requests for materials should be addressed to Chirlmin Joo.

Peer review information *Nature Nanotechnology* thanks Carlos Penedo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.