

The emerging landscape of single-molecule protein sequencing technologies

Bohländer, Peggy; Filius, Mike; van Kooten, Xander F.; Pomorski, Adam; Schmid, Sonja; Dekker, Cees; Eelkema, Rienk; Kim, Sung Hyun; Joo, Chirlmin; More Authors

DOI

[10.1038/s41592-021-01143-1](https://doi.org/10.1038/s41592-021-01143-1)

Publication date

2021

Document Version

Accepted author manuscript

Published in

Nature Methods

Citation (APA)

Bohländer, P., Filius, M., van Kooten, X. F., Pomorski, A., Schmid, S., Dekker, C., Eelkema, R., Kim, S. H., Joo, C., & More Authors (2021). The emerging landscape of single-molecule protein sequencing technologies. *Nature Methods*, 18(6), 604-617. <https://doi.org/10.1038/s41592-021-01143-1>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

The emerging landscape of single-molecule protein sequencing technologies

Javier Alfaro^{1,3,†,*}, Peggy Bohländer^{2,†}, Mingjie Dai^{3,4,†}, Mike Filius^{5,†}, Cecil J. Howard^{6,†}, Xander F. van Kooten^{7,†}, Shilo Ohayon^{7,†}, Adam Pomorski^{5,†}, Sonja Schmid^{8,†}, Aleksei Aksimentiev⁹, Eric V. Anslyn⁶, Georges Bedran¹, Cao Chan¹⁰, Mauro Chinappi¹¹, Etienne Coyaud¹², Cees Dekker⁵, Gunnar Dittmar¹³, Nicholas Drachman¹⁴, Rien Eelkema², David Goodlett¹⁵, Sebastien Hentz¹⁶, Umesh Kalathiya¹, Neil L. Kelleher¹⁷, Ryan T. Kelly¹⁸, Zvi Kelman^{19,20}, Sung Hyun Kim⁵, Bernhard Kuster^{21,22}, David Rodriguez-Larrea²³, Stuart Lindsey²⁴, Giovanni Maglia²⁵, Edward M. Marcotte²⁶, John P. Marino¹⁹, Christophe Masselon²⁷, Michael Mayer²⁸, Patroklos Samaras²¹, Kumar Sarthak⁹, Lusia Sepiashvili²⁹, Derek Stein¹⁴, Meni Wanunu^{30,31}, Mathias Wilhelm²¹, Peng Yin^{3,4}, Amit Meller^{7,32,‡,*} and Chirlmin Joo^{5,‡,*}.

¹ International Centre for Cancer Vaccine Science, University of Gdańsk, Gdańsk, Poland

² Faculty of Applied Sciences, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft, The Netherlands

³ Wyss Institute for Biologically Inspired Engineering, Harvard University, 3 Blackfan circle, Boston, MA, 02115 USA

⁴ Department of Systems Biology, Harvard Medical School, Boston, MA, 02115 USA

⁵ Department of BioNanoScience, Kavli Institute of Nanoscience, Delft University of Technology, van der Maasweg 9, 2629 HZ Delft, The Netherlands

⁶ Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712, United States

⁷ Department of Biomedical Engineering, Technion – Israel Institute of Technology, Haifa, 32000, Israel

⁸ NanoDynamicsLab, Biophysics Chair, Wageningen University, Stippeneng 4, 6708WE Wageningen, The Netherlands.

⁹ Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801

¹⁰ Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

¹¹ Dipartimento di Ingegneria Industriale, Università di Roma Tor Vergata, via del Politecnico 1, 00133 Roma, Italy

¹² Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France

¹³ Proteomics of cellular signalling, Quantitative Biology Unit, Luxembourg Institute of Health, Strassen, Luxembourg.

¹⁴ Department of Physics, Brown University, Providence, RI 02912, USA.

¹⁵ Department of Microbial Pathogenesis, School of Dentistry, University of Maryland, Baltimore, Maryland 21201, United States

¹⁶ Université Grenoble Alpes, CEA, LETI, 38000 Grenoble, France.

¹⁷ Departments of Chemistry, Molecular Biosciences, and the Feinberg School of Medicine, Northwestern University, 2145 Sheridan Road, Box #101, Evanston, IL 60208-3113, United States

¹⁸ Department of Chemistry and Biochemistry, Brigham Young University, Provo, UT 84602, United States

¹⁹ Institute for Bioscience and Biotechnology Research, National Institute of Standards and Technology, University of Maryland, 9600, Gudelsky Drive, Rockville, MD, USA.

²⁰ Biomolecular Labeling Laboratory, IBBR, Rockville, MD, USA

²¹ Chair of Proteomics and Bioanalytics, Technische Universität München, Emil-Erlenmeyer Forum 5, 85354 Freising, Germany

²² Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany

²³ Biofisika Institute (CSIC, UPV/EHU), Department of Biochemistry and Molecular Biology (UPV/EHU), Leioa 48940, Spain

²⁴ Biodesign Institute, School of Molecular Sciences, Department of Physics, Arizona State University, Tempe, Arizona 85287, United States

²⁵ Groningen Biomolecular Sciences & Biotechnology Institute, University of Groningen, 9747 AG Groningen, The Netherlands

²⁶ Department of Molecular Biosciences, Center for Systems and Synthetic Biology, University of Texas, Austin, TX 78712, USA

²⁷ Université Grenoble Alpes, CEA, Inserm, BGE U1038, Grenoble, France

²⁸ University of Fribourg, Adolphe Merkle Institute, Chemin des Verdiers 4, CH-1700 Fribourg, Switzerland

²⁹ University of Toronto, Hospital for Sick Children, 555 University Ave., Rm 3606 Atrium, Toronto Ontario M5G 1X8

³⁰ Department of Physics, Northeastern University, Boston, Massachusetts 02115, United States

³¹ Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States

³² Russell Berrie Nanotechnology Institute, Technion – IIT, Haifa 32000, Israel

† These authors contributed equally

‡ These authors jointly directed this work.

* Corresponding Authors

* Correspondence should be addressed to: J.A. Alfaro- Javier Alfaro javier.alfaro@proteogenomics.ca, A.

Meller - Amit Meller ameller@technion.ac.il, and C. Joo - Chirlmin Joo c.joo@tudelft.nl

56 Abstract

57 Proteins are involved in the majority of structures and biochemical reactions of living cells.
58 New single-molecule protein sequencing and identification technologies alongside innovations
59 in mass spectrometry and antibody-based methods will eventually enable broad sequence
60 coverage in single-cell profiling. The ultimate precision and sensitivity of proteomes promised
61 by these technologies will create new directions in research and biomedical applications, from
62 global proteomics of single cells and bodily fluids to sensing and classifying low-abundance
63 protein biomarkers for disease screening and precision diagnostics.
64

65 Introduction

66 The emergence of Next Generation Sequencing (NGS) and single-molecule DNA sequencing
67 technologies have revolutionized genomics. Proteomics awaits a similar transformative wave
68 of protein-sequencing techniques that will allow for the examination of proteins at the single-
69 cell and ultimately single-molecule levels, even with low-abundant proteins. Such techniques
70 would allow routine global proteome profiling, like today's single-cell RNA sequencing studies,
71 creating opportunities for single-cell proteomics and potentially permitting real-time testing for
72 on-site medical diagnostics and disease screening. Importantly, whole proteome sequencing
73 and profiling of the vast repertoire of cell types is expected to fundamentally enhance our
74 understanding of all living systems.

75
76 While DNA sequencing technologies are routinely used for whole genome and transcriptome
77 profiling with extensive read depths and high sequence coverages, conventional bottom-up
78 mass-spectrometry (MS)-based proteomics assays (**Box 1**) fall short of providing the same
79 breadth of view for proteins. The analysis of complex protein mixtures is particularly
80 challenging since the >20,000 genes in the human genome¹ are translated into a diversity of
81 proteoforms that may include millions of variants as a result of post-translational modifications
82 (PTMs), alternative splicing and germline variants². In cancer, for example, the proteoform
83 landscape can be aberrant with many new protein variants resulting from non-canonical
84 splicing, mutations, fusions and PTMs. Characterizing such proteoforms is likely to benefit
85 from the improvements in current protein sequencing techniques and the emergence of new
86 methods.

87
88 MS remains a staple of protein identification and continues to develop towards single cell
89 methods (**Box 2**). Alongside, a diverse range of protein sequencing and identification
90 techniques have emerged that aim to increase the sensitivity of proteomics to the single-
91 molecule level. Many of these techniques rely on fluorescence and nanopores for single-
92 molecule sensing as an alternative means to sequence or identify proteins (**Figure 1**). The
93 landscape of emerging proteomics is already vast, with different approaches at various stages
94 of development, some of which have already secured industry investment^{3,4}, an important step
95 towards broad dissemination to the research community. Other technologies have shown
96 great promise and gained popularity among the single-molecule biophysics communities and
97 some are available as proofs of concept at just one or a few laboratories.

98

99 Here, we describe the prominent emerging protein sequencing and fingerprinting techniques
100 in the context of mature methods, such as MS-based proteomics, discuss challenges for their
101 real-world applications, and assess their transformative potential.

102 A Renaissance of Classical Techniques

103 Edman degradation, MS and Enzyme-Linked Immunosorbent Assay (ELISA) have been
104 broadly used for protein/peptide sequencing and identification for several decades, therefore
105 it is no surprise that further enhancements of these classical technologies are being sought.
106 The biophysics community has been developing methods to **increase the throughput⁵ and**
107 **sensitivity⁶ of single-molecule ELISA**, Edman degradation, single-particle MS, neutral-particle
108 nanomechanical MS, and single-particle electrospray. Even classical tools, which are
109 commonly used in materials sciences like electric tunneling and direct current measurements
110 can be repurposed for protein sequencing.

111 Massively parallel Edman degradation

112 Edman degradation⁷ was the first method to determine the amino acid sequence of a purified
113 peptide. The method relied on chemically modifying the N-terminal amino acid, cleaving it from
114 the peptide, and finally determining the sequence of the cleaved labeled amino acid using
115 high-performance liquid chromatography. Until recently, attempts to conduct sequencing of
116 this sort in a massively parallel fashion were not possible as the method relied on highly
117 purified peptides. However, recent multiplex strategies that employ peptide arrays and either
118 sequence chemically labeled peptides (“fluorosequencing”), or successively detect the N-
119 terminal amino acid are making breakthroughs (**Box 3**).

120
121 Fluorosequencing combines Edman chemistry, single-molecule microscopy, and stable
122 synthetic fluorophore chemistry (**Figure 2a**). Millions of individual fluorescently labeled
123 peptides can be visualized in parallel, while changing fluorescence intensities are monitored
124 as N-terminal amino acids are sequentially removed. The resulting fluorescence signatures
125 serve to uniquely identify individual peptides⁸. This method allows for millions of distinct
126 peptide molecules to be sequenced in parallel, identified, and digitally quantified on a
127 zeptomole-scale⁹. However, the technology is not without challenges, as the reagents used
128 for Edman degradation chemistry lead to increased rates of fluorescent dye destruction, which
129 in turn limits the read length. These reagents include slightly basic structures such as pyridine,
130 strong acids such as trifluoroacetic acid, and the electrophile phenyl isothiocyanate.
131 Furthermore, the reliance on chemical labelling leads to partial sequencing of the peptide, with

132 the unidentified remainder inferred by comparison to a reference proteome. In addition, the
133 inefficient labeling can lead to errors that must be modeled into the reference proteome
134 comparison, spurring the development of new protocols to increase yields¹⁰. **Exciting new**
135 **proposals could add the dimension of protonation-based sequencing. The pK_a of the N-**
136 **terminal amino acid could be used for identification by observing and interpreting the**
137 **protonation-deprotonation signal of the peptide at fixed pH through the Edman Degradation**
138 **process¹¹. Much like fluorosequencing, the signal observed would be for the whole peptide**
139 **and the decay pattern would be interpreted to derive a pK_a of each N-Terminal amino acid.**

140

141 Several natural proteins and RNA molecules recognize specific amino acids either as free
142 amino acids or as a part of a polypeptide chain¹². These proteins and nucleic acids provide
143 different solutions for N-terminal amino acid recognition. Each N-terminal amino acid binder
144 (NAAB) probe selectively identifies a specific N-terminal amino acid or an N-terminal amino
145 acid derivative. With each cycle, another amino acid is revealed in the sequence of the peptide.
146 However, further directed evolution and engineering of the NAAB probe is required to meet
147 the stringent affinity, selectivity and stability needed for error-free sequencing applications. In
148 addition, such probes would need to discriminate among all amino acids, including the same
149 amino acid in alternative positions in the peptide sequence. Probes that bind a class of N-
150 terminal amino acids (e.g., short aliphatic residues) could also be useful, but would introduce
151 ambiguities in the sequencing process. Different probes could also be designed to recognize
152 short N-terminal k-mers, which would increase the number of probes needed, but reduce the
153 ambiguity in the resulting sequencing information. To circumvent this limitation, it may be
154 possible to sequence the N-terminal amino acid by selective recognition using a plurality of
155 probes in each cycle of Edman degradation^{13,14} (**Figure 2b**).

156

157 Single-molecule mass spectrometry

158 MS is a century-old method that measures the mass-to-charge (m/z) ratio of ions, for example,
159 charged peptides/proteins and their assemblies. Single-ion detection has been possible since
160 the 1990's, for example, in Fourier-transform ion cyclotron resonance instruments¹⁵. Charge
161 detection MS (CDMS) is a single-ion method where charge assignment of each individual ion
162 is determined directly, enabling the conversion of mass-to-charge ratio into the neutral mass
163 domain. The approach has focused on the analysis of large biomolecular complexes,
164 especially viruses in the 1–100 MDa range¹⁶. While CDMS had been limited to specialized
165 instrumentation, the past year has seen breakthroughs built on early work for producing mass
166 spectra of single ions in Orbitrap mass analyzers^{17–19}. Today, these mass-analyzers can be

167 widely used to directly derive the charge states of single proteins and even their fragment
168 ions²⁰. Orbitraps are particularly useful since the readout of individual ions can be multiplexed
169 by 100-1000 fold in the Orbitrap-based CDMS²⁰. Individual ion MS has already shown
170 resolution of mixtures with ~1000 proteoforms that provided no data when standard MS was
171 used^{20,21}. This has greatly expanded the top-down approach to confirm DNA-inferred
172 sequences of whole proteins, with localization of their post-translational modifications²⁰⁻²².
173 Without extensive alteration, Orbitraps can therefore measure tens of thousands of proteins
174 in a matter of minutes. With these rapidly evolving technologies, the charting of the full human
175 proteoform atlas has already begun²³, making strides towards a comprehensive Human
176 Proteoform Project. However, a critical requirement for MS of proteins and peptides is
177 ionization, and not all ions are efficiently transmitted through the mass spectrometer. This
178 might restrict some of the proteoform mapping efforts providing a niche for the other
179 technologies in **Figure 1**.

180

181 For higher molecular weight species, the ionization of proteins and complexes yields a mixture
182 of macro ions with variable charge states, resulting in a net reduction of sensitivity, as the
183 signal distributes over multiple peaks in the mass-to-charge dimension. Moreover, charge
184 state distributions may overlap above a certain mass or in the case of mixtures, challenging
185 the species identification. Since their inception²⁴, nano-mechanical mass sensors have made
186 tremendous progress towards protein characterization²⁵. Such devices, which take the shape
187 of cantilevers or beams with lateral dimensions in the hundreds of nanometers, can detect
188 individual particles accreting onto their active surface through the changes in their vibration
189 frequency. Importantly, as the particle's inertial mass is determined directly from the frequency
190 change, these devices are insensitive to charge states²⁶. This realization prompted the
191 development of new MS instrument designs devoid of ion guides, which no longer depend on
192 electromagnetic fields to collect and transmit the analytes (**Figure 2c**). Such nano-mechanical
193 resonator-based MS system has recently shown the ability to characterize large protein
194 assemblies such as individual viral capsids above 100 MDa²⁷. Outside of proteomics, 1 Da
195 resolution has been demonstrated with carbon nanotubes²⁸. Moreover, recent reports
196 suggested the possibility to determine other physical parameters like the stiffness or shape of
197 the analyte by monitoring multiple vibrational modes^{29,30}. These previously inaccessible
198 metrics may open new avenues to discriminate peptides, proteins and their complexes.
199 Nonetheless, one of the challenges of the nano-resonator-MS lies in devising efficient ways
200 to bring individual proteins onto the resonator's active surface for mass sensing.

201

202 Ionization is commonly achieved by electrospray ionization of a solution containing the
203 compound(s) of interest. The use of ever-smaller electrospray ion source apertures has led to
204 significant improvements in the sensitivity of mass spectrometry^{31,32}. Mass spectrometers with
205 a nanopore ion source have been developed for the purpose of sequencing single proteins³³
206 (**Figure 2d**). A nanopore electrospray can potentially deliver individual amino acid ions directly
207 into a high-vacuum gas phase, where the ions can be efficiently detected by their mass-to-
208 charge ratios. This opens a path to sequencing peptides one amino acid at a time. The concept
209 makes use of the nanopore to guide the protein into a linear configuration so that its monomers
210 can be delivered into the mass spectrometer sequentially³⁴. Individual amino acids must be
211 cleaved from the protein molecule as they transit the nanopore, and this could potentially be
212 accomplished using photodissociation³⁵ or chemical digestion methods. The 100 MHz
213 bandwidth of the channeltron single ion detectors used in this setup is also sufficient to resolve
214 the arrival order of the ions. The high mass resolution makes this technique promising for
215 identifying post-translational modifications (PTMs), which change the masses of particular
216 amino acid residues by predictable amounts. One challenge on the path for this technology
217 will be achieving high throughput, which might require a strategy for parallelizing the mass
218 analysis.

219 Tunneling conductance measurements

220 The appearance of the scanning tunneling microscope in the 1980s opened a new way to
221 analyze molecules. Small organic molecules can be transiently trapped between two metal
222 electrodes with sub-nanometer separation, with the tunneling currents between the electrodes
223 reporting on the molecular signature of the analyte. Recently, several technical advances have
224 been made towards single-molecule amino acid and protein analysis. Extracting insightful
225 information from electron tunneling is complicated by the noise due to water and contaminants
226 reaching the electrode surfaces. To overcome these problems, recognition tunneling has been
227 developed. The electrodes are covalently modified with adaptor molecules that form transient,
228 but well-defined links to the target molecule³⁶. The rapidly fluctuating tunnel-current signals
229 are processed using machine learning algorithms, which makes it possible to distinguish
230 individual amino acids and small peptides³⁷. Moreover, smaller electrode gaps have been
231 made to obtain distinct signals from different amino acids and PTMs³⁸. Further development
232 of the technology will depend on a reliable source of **tunnel junctions with a defined gap to**
233 **replace the cumbersome scanning tunneling microscopy**, but it is clear that both the sequence
234 and PTMs of small peptides can be determined³⁷. Currently, tunneling conductance is a proof-
235 of-concept technology for fully sequencing short peptides that could one day be used for the
236 analysis of protein digests and expanded to PTM analysis (**Figure 1**).

237

238 Recently, it has been discovered that electrical charges can be transmitted through a protein
239 if electrodes are bridged by a protein via chemical bonding or ligand binding³⁹. The protein
240 conformation change upon nucleotide addition could be followed in real time from the direct
241 currents passing through a DNA polymerase⁴⁰. Although the observation was preliminary, the
242 electronic signatures were distinctive when the polymerase was associated with different DNA
243 sequences, enabling a new approach to label-free single-molecule DNA sequencing. A similar
244 approach could potentially be used for protein sequencing with enzymes, such as
245 proteasomes or glycopetidases that process substrates sequentially.

246 DNA Nanotechnologies for Protein Sequencing

247 DNA nanotechnologies, which utilize the ability to custom-design a large number of sequences
248 with prescribed pairing interaction and dynamic properties, have facilitated developments in
249 fields ranging from synthetic biology to diagnostics and drug delivery⁴¹. For example, the
250 programmable, transient binding between short DNA strands is central to the super-resolution
251 technique, DNA-based Point Accumulation for Imaging in Nanoscale Topography (DNA-
252 PAINT^{42–44}) (**Box 4**). Here we describe the application of DNA-PAINT and DNA-based local
253 and global pairwise distance measurement methods for single-molecule protein detection and
254 identification.

255 Fingerprinting via DNA PAINT

256 DNA-PAINT uses the repetitive binding between designed docking and imager DNA strands
257 to allow for imaging with molecular-level resolution (**Box 4**). This method provides a promising
258 way to fingerprint proteins on the level of single molecules. A simple way for characterizing
259 proteins could be through amino acid counting using quantitative DNA-PAINT (qPAINT⁴⁴). In
260 this technique, the total blinking rate of a region of interest is measured, which linearly reflects
261 the number of molecular targets in the region. It has been proposed that high-efficiency DNA
262 labeling of specific amino acids (**Figure 3a**) followed by qPAINT could lead to single-molecule
263 protein fingerprinting of intact proteins (**Figure 3b**)⁴⁵. More than 75% of the human proteome
264 can be identified (≤ 5 degeneracy) if the error in counting is less than 5% from detecting three
265 kinds of specific amino acids.

266

267 The recent development of DNA-PAINT has allowed discrete molecular imaging (DMI) of
268 individual molecular targets with < 5 nm spatial resolution⁴³. Therefore, protein identification by

269 the fingerprinting of amino acids along an extended protein backbone is a possibility. DMI was
270 achieved by combining a systematic analysis and optimization of DNA-PAINT super-resolution
271 workflow and a high-accuracy (<1 nm) drift correction method. To effectively unfold and extend
272 the protein backbone, N- and C-terminal specific modifications should be used to attach
273 surface and microbead anchors. The protein can then be subjected to mechanical or
274 electromagnetic extension force (**Figure 3c**). **Proposals to combine protein extension methods
275 with high-resolution DMI⁴⁵ indicate that with lysine labelling alone and 5-nm effective imaging
276 resolution, more than 50% of the human proteome could be uniquely identified, even with up
277 to 20% amino acid imaging error. Labeling lysine and cysteine would allow the proteome
278 coverage to increase to more than 75%.**

279

280 Protein fingerprinting using DNA-PAINT single-molecule imaging combines the ultrahigh
281 imaging resolution and quantitative capacity of the technique, and the inherent throughput of
282 wide imaging-based methods. qPAINT can produce signals linearly (within <5% deviation),
283 based on the amino acid composition of a particular protein. The proposed methods will be
284 particularly useful for global proteomic analysis of complex protein mixtures, PTM patterns
285 and combinatorial analysis at the single molecule level.

286 DNA proximity recording

287 **An alternative method for DNA-based protein identification attaches DNA probes to specific
288 amino-acids on a protein and uses enzymatic DNA-amplification between nearby probes to
289 generate DNA 'records' that vary in length according to pairwise distances within a protein. An
290 example is auto-cycling proximity recording⁴⁶ (APR) (**Figure 3d**).** The distribution of lengths of
291 these molecular records is then analysed to decode the pairwise distance between the two
292 DNA tags. It is possible to use unique molecular identifier barcoding and repetitive enzymatic
293 recording, such that each lysine and cysteine residue can be studied and a pairwise distance
294 map can be constructed among them, allowing for single-molecule protein identification^{47,48}.
295 DNA proximity recording takes advantage of high-throughput next generation DNA
296 sequencing methods for efficient protein fingerprinting analysis, and will be useful for both the
297 analysis of purified proteins and complex protein mixtures.

298 Protein fingerprinting using FRET

299 A different approach that allows for global pairwise distance measurement is combining DNA
300 technology with single-molecule Förster resonance energy transfer (FRET)⁴⁹. The current
301 state of the art of single-molecule FRET analysis allows us to deal with only one or two FRET

302 pairs⁵⁰. The new high-resolution FRET using transient binding between DNA tags allows for
303 probing one FRET pair at a time when many of them are collectively present on a single
304 protein⁴⁹. Similar to the above-mentioned approaches, specific amino acids (e.g. lysines,
305 cysteines, etc) required for fingerprinting have to be labelled with a set of different DNA
306 docking strands. Furthermore, a fixed position on the protein (either the N or the C terminus)
307 is labelled with the acceptor fluorophore. Only a single FRET pair forms at a time by using
308 DNA strands that are complementary to only a single docking strand. The measurements are
309 then repeated to probe the remaining docking strands and thus the amino acids. The output
310 of this approach will be a FRET histogram containing information on the position (referred to
311 as FRET fingerprints) of each detected amino acid relative to one of the reference points. This
312 information is compared to a database consisting of predicted 'FRET fingerprints' and allows
313 for the identification of the protein species (**Figure 3e**). The proposed high-resolution FRET
314 approach (named high resolution FRET using DNA eXchange, or FRET X) benefits from the
315 immobilization of the protein molecules, allowing users to probe each protein multiple times to
316 obtain fingerprints with a high resolution. FRET X will be a particularly promising tool for
317 targeted proteomics or proteoform analysis as it is able to distinguish small structural changes.

318 Biological and Solid-State Nanopores

319 Nanopore-based DNA and direct RNA sequencing technologies have become key players in
320 the sequencing field, offering unprecedented read-lengths and portability. Since its first
321 demonstration as a single-biomolecule sensor⁵¹, nanopore sensing has progressively matured
322 reaching the goal of single-molecule, long-read DNA sequencing⁵². Many of the nanopore
323 sequencing applications to date have materialized using an ultra-small device⁵³, which
324 features vast arrays of biological nanopores, each coupled to its own current amplifier,
325 allowing readout of hundreds of DNA strands simultaneously. Nanopore sequencing involves
326 drawing biomolecules through the nanopore in a single file manner, hence partially blocking
327 the ionic current flowing through the pore, leading to time-dependent and sequence-specific
328 electrical signals. In the past two decades a variety of synthetic nanopore biosensors have
329 significantly progressed and are currently used in diverse applications beyond sequencing,
330 including applications in detecting epigenetic variations and enabling ultra-sensitive mRNA
331 expression⁵⁴, to name a few.

332
333 Just like gel electrophoresis, nanopores may serve as a generic tool to analyze biomolecules.
334 Therefore, as nanopore-based DNA sequencing continues to advance, this technique is
335 poised to extend to proteins, metabolites and to other analytes. But despite the remarkable
336 advances in DNA and RNA sequencing, nanopore-based protein sensing is still in its infancy,

337 facing challenges unique to proteins and proteomics. In particular, proteins span a large range
338 of sizes and have a stable three-dimensional folded structure. In contrast to nucleic acids,
339 peptides' backbones are not naturally charged, complicating the possibility of single-file
340 electrokinetic threading into nanopores. In addition, proteins are composed of 20 amino acids
341 instead of 4 nucleobases, further complicating the task of relating the ionic current signals to
342 the amino acid sequence.

343

344 While a significant progress in nanopore-based protein sensing have been made, to date the
345 development of full protein sequencers, or single-protein identification based on nanopores
346 remains to be a topic of intense focus. Here, we focus on three of the principal directions in
347 this field (**Figure 4**): (i) Single-file threading and direct sensing of the sequence of the
348 polypeptide's amino acids, analogous to the nanopore DNA sequencing principle. In this
349 approach, either the translocation of full-length proteins or shorter polypeptide digests of the
350 proteins may be targeted. (ii) Protein identification methods based on sensing unique
351 fingerprints in linearized proteins, without *de novo* amino acid sequencing. (iii) Protein
352 identification of folded proteins, based on specific patterns in their nanopore current
353 blockades. In the following sections, we provide short overviews of the current state of these
354 approaches and refer to additional methods.

355 Reading the amino acid sequence of linearized peptides

356 In this proposed approach, a single protein or peptide is linearized and threaded through a
357 nanopore and the resulting ion current interpreted to an amino-acid sequence (**Figure 4a**).
358 Theoretical work using all-atom MD simulations on alpha-hemolysin pores has demonstrated
359 a global correlation between the volume of an amino acid and the current blockade in homo-
360 polymers⁵⁵. Computationally efficient predictions using course-grained models have also
361 performed well compared to all-atom MD simulations for both solid-state and biological
362 pores⁵⁶.

363 Discrimination among peptides differing by one amino acid substitution (alanine to glutamate)
364 have been demonstrated using an engineered Fragaecetoxin C (FraC) nanopores⁵⁷.

365 Moreover, Piguet *et al.* resolved single amino acid differences within short poly-arginine
366 peptides with superb resolution, using the aerolysin protein pore in its wild-type
367 conformation⁵⁸. Combining MD simulations and single channel experiments, Cao *et al.* have
368 rationally determined specific point mutations in aerolysin to fine-tune the charge and diameter
369 of the pore, which enhanced its sensitivity and selectivity as showcased experimentally using
370 DNA and peptides⁵⁹. Notably, protein pore sensors were used for analysis of bodily fluids

371 (blood, sweat, etc.), indicating significant potential for applications in diagnostics⁶⁰. As an
372 alternative to nanopore sequencing of intact polypeptide chains, smaller digested fragments
373 can also be analyzed and minute differences in the amino acid composition can be detected⁶¹.
374 Even post-translational modifications can be detected including individual phosphorylations
375 and glycosylations using the protein pore FraC⁶².

376 An essential step in the development of nanopore based DNA sequencing, came with the
377 application of an enzymatic stepping motor (e.g. a helicase) that produces a nucleotide-by-
378 nucleotide progression of the DNA through the nanopore. A similar system is pursued for
379 single-molecule protein sequencing: Molecular motors of the Type II secretion system⁶³
380 (SecY) and the AAA family⁶⁴ (ClpX) are known to unfold and pull protein substrates through
381 pores in an ATP-dependent way. Nivala et al.^{65,66} employed ClpXP (or ClpX alone) to unfold
382 and translocate a multi-domain fusion protein through the hemolysin pore using the energy
383 derived from ATP hydrolysis. In this approach the motor is at the exit of the nanopore, and
384 therefore the step size of translocation is caused by the stable structural motifs that resist
385 translocation – rather than being controlled by the enzyme. This approach is currently being
386 expanded by several groups, who conjugated ClpXP covalently to alpha-hemolysin at the
387 entrance of the nanopore to form a combined sensor as well as a substrate delivery machine.
388 The Maglia lab genetically introduced a nanopore directly into an archaeal proteasome and
389 found that the assisted transport across the nanopore is not influenced by the unfolding of the
390 protein. These nanoscale constructs would also allow a *cut-and-drop* approach, in which
391 single proteins are recognized by the pattern of peptide fragments as they are sequentially
392 cleaved by the peptidase above the nanopore⁶⁷. Knyazev *et al.* introduced a protein-secreting
393 ATPase as an additional natural choice for a potential peptide translocating motor^{68,69}. Other
394 proteins have the potential to control the protein translocation through nanopores, beyond
395 secretases and unfoldases, including chaperones (Hsp70), via processes resembling protein
396 translocation into the mitochondrial matrix⁷⁰. Recently, Rodriguez-Larrea's group has
397 discussed how protein refolding at the entry or exit compartment can oppose or promote
398 protein translocation, respectively^{71,72} and the use of deep learning networks to analyze the
399 raw ionic current signals for accurate classification of single-point mutations in a translocating
400 protein. In addition, Cardozo *et al.* built a library of ~20 proteins that are orthogonally barcoded
401 with an intrinsic peptide sequence, and successfully read them by nanopore sensors⁷³.

402 Fingerprinting linearized proteins

403 An accurate quantification of different protein species in the proteome with single-molecule
404 resolution would already be a highly significant achievement. This can be realized by single-

405 molecule fingerprinting, i.e. by the identification of individual protein molecules based on prior
406 knowledge of their amino-acid sequence, or based on the specific signal patterns, recognized
407 by machine learning^{8,74,75} (**Figure 4b**). To that end several nanopore approaches have been
408 pursued: Restrepo-Pérez *et al.*⁷⁶ established a fingerprinting approach using six chemical
409 tags, which were placed on a dipolar peptide⁷⁷. Additionally, Wang *et al.* reported the ability to
410 distinguish individual lysine and cysteine residues in short polypeptides, coupled specifically
411 to fluorescent tags using a solid-state nanopore with low fluorescence background⁷⁸. In all
412 these approaches, separating the proteins by mass, prior to single molecule sensing may
413 highly facilitate the identification of proteins in complex samples containing many different
414 proteins⁷⁹.

415 Nanopore protein fingerprinting can make extensive use of advanced deep-learning artificial
416 intelligence (AI) strategies to identify patterns in noisy signals. Ohayon *et. al.* has recently
417 shown computationally that >95% of all the proteins in the human proteome can be identified
418 with high confidence, based on the labelling of three amino acids (lysine, cysteine and
419 methionine) and linear threading through a solid state nanopore⁷⁵. These simulations predict
420 that even partial labelling of the proteins will be sufficient to achieve a high degree of accurate
421 whole proteome identification, due to the ability of AI functions to correctly recognize partial
422 protein patterns. This identification method involves the incorporation of subwavelength light
423 localization in the proximity of the nanopore using plasmonic nanostructures⁸⁰. The work in
424 this field benefits from recent advances in nanofabrication and nanopatterning technologies,
425 allowing for the formation of complex metallic nanostructures to induce light localization and
426 plasmonics⁸¹.

427 Characterization and identification of folded proteins

428 To date, nanopores have been successfully employed to detect specific sets of folded proteins
429 and protein oligomers⁸² (**Figure 4c**) such as large globular proteins, various cytokines and
430 even low molecular weight proteins, such as Ubiquitin. Holding the proteins in their folded
431 state inside the nanopore for sufficiently long times is a key requirement. Early studies have
432 shown that globular proteins of about 5 nm in size can only be detected for a few tens of
433 microseconds or less⁸³, which is too short for characterization. Several approaches to
434 overcome this challenge have been devised. A lipid bilayer coating of a solid-state nanopore
435 can be used to tether the proteins for extended periods of time⁸⁴. Lipid tethered proteins⁸⁴, and
436 more recently also freely diffusing proteins (using a higher bandwidth sensing system)⁸⁵ have
437 been characterized based on their size, shape, charge, dipole, and rotational diffusion
438 coefficient⁸⁶. Various strategies are being pursued to ‘trap’ proteins in a nanopore. One

439 strategy is to use plasmonics to hold a protein in a nanopore for seconds or even minutes^{87,88}.
440 More recently single proteins have been demonstrated to be held at the nanopore's most
441 sensitive region for minutes to hours using the nanopore electro-osmotic trap (NEOtrap) that
442 exploits strong electro-osmotic water flows created in-situ by a charged, permeable objects,
443 such as a DNA origami structures⁸⁹. Another approach for slowing down the translocation of
444 proteins involves the use of smaller nanopores compared to earlier studies, in order to
445 increase the hydrodynamic drag, thus resulting in longer translocation dwell-times that are
446 easier to measure^{90,91}. In addition, high bandwidth measurements can resolve differential
447 conformational flexibility within folded proteins⁹⁰⁻⁹³, and even changes in conformational
448 flexibility⁹⁴. Biological nanopores with a diameter of 5.5 or 10 nm⁹⁵ can also be used to
449 measure folded proteins, including protein conformations⁹⁶ and post-translational
450 modifications⁹⁷ such as ubiquitination. Lastly, Aramesh *et al.*⁹⁸ used a combination of atomic
451 force microscopy and nanopore technology, to make the first steps at nanopore sensing
452 directly inside cells. Altogether, protein detection, identification, and even sequencing using
453 single nanopore approaches has become a highly active, thriving research field, with great
454 potential to revolutionize proteomics, medical diagnostics, and also fundamental biosciences.
455

456 Chemistry for Next-Generation Proteomics 457 Technologies

458 Single-molecule protein fingerprinting has underlined the need for innovative approaches for
459 attaching various functional groups onto peptides, such as fluorescent moieties. A high degree
460 of chemical specificity is required to avoid down-stream misidentification of amino acids, which
461 could lead to sequencing errors. Chemists are making headway on a suite of selective and
462 high-yield methods for labeling specific amino acid side chains, amino acid termini, and post-
463 translational modifications with minimal cross reactivity (**Box 5**).

464
465 Labelling stability and efficiency is paramount to the success of sequencing technologies, but
466 is also a challenge the chemists face. First, modification of most or all individual residues of
467 one amino acid type is desired for an explicit identification of a peptide sequence, which
468 requires selective and highly efficient reactions. Second, error-free sequence prediction
469 requires multiple chemical labels, but the stability of the chemical labels has been an issue in
470 some sequencing techniques. These issues have been best characterized for
471 fluorosequencing (**Box 5a**).

472

473 For many of the sequencing techniques, amino acids must be labeled with a chemical tag to
474 allow for the differentiation between the amino acids. While it is theoretically possible to get a
475 broad coverage of the proteome with a minimal set of amino acid labeling, specific
476 identification of peptides or broader sequence coverage requires a larger suite of labels.
477 Overall, there are twelve distinct side chain types in peptides ranging from highly reactive
478 amino acids like lysine and cysteine, to functional groups that are more challenging to modify,
479 such as amides (Gln/Asn) and alkanes (Ala, Gly, Ile, Leu, Pro, and Val). There are a large
480 number of methods to label amino acids, however some chemistries do not provide sufficiently
481 stable bonds for some single-molecule sequencing approaches. To date, only eight (Lys, Cys,
482 Glu/Asp, Tyr, Trp, His, and Arg) have thus far been shown to be stable, selective, and reactive
483 enough for the single-molecule fluorosequencing approach^{9,99}. Research is ongoing to test a
484 wide variety of other labeling conditions to cover all of the proteinogenic amino acids (**Box**
485 **5b**).

486

487 Chemical modification of protein termini is highly desired for several sequencing techniques
488 like fluorosequencing, nanopores and DNA-PAINT approaches where end labeling or ligation
489 is required (**Figure 1**). The terminus provides an attachment point for surface immobilization
490 and can offer a simple way to remove excess chemical reagents during procedures that
491 require multiple labeling steps. Two terminus-specific methods have found great promise for
492 single-molecule sequencing, C-terminal labeling using decarboxylative alkylation (**Box 5c**)
493 and modification of the N-terminus with 2-pyridinecarboxaldehyde (**Box 5d**).

494 The long-term goal of characterizing proteoforms requires methods to detect and differentiate
495 PTMs. They can be recognized by mass spectrometry through the mass shifts they cause on
496 a protein, peptide and their fragments^{100,101} and databases of the expected mass shifts like
497 Unimod are used to support the identification¹⁰². However, these databases show that there
498 can be significant overlaps between PTMs of the same or similar mass suggesting that
499 orthogonal methods are needed. Single-molecule protein sequencing methods rely on either
500 site-specific labeling or elimination and replacement chemistries (**Box 5e**).

501 Discussion: a spectrum of opportunities

502 An emerging landscape of single-molecule protein sequencing and fingerprinting technologies
503 is being developed (**Figure 1**) with the promise to resolve the full proteome of single cells with
504 single-protein resolution, opening up unprecedented opportunities in fundamental science and
505 medical diagnostics. Cellular tissues' composition could then be resolved with single-cell
506 resolution, opening up new research avenues from embryonic development to cancer

507 research. Diagnostics could benefit from the ultimate single-molecule resolution by resolving
508 very low amounts of protein in bodily samples. The detection of rare proteins with copy
509 numbers as low as one or a few may uncover new molecular regulatory networks within cells.
510 Some of the emerging technologies described here are still at their early proof-of-concept
511 stages of development, whereas others like sequencing by Edman degradation and nanopore
512 sequencing technologies have already attracted industry funding. Additional single-molecule
513 approaches are also promoted by commercial entities, and are out of scope for this review.

514

515 A real-world application of a technology that is not MS- or antibody-based for whole proteome
516 characterization is yet to be achieved. In the meantime, MS will continue to improve in its
517 capacity, to support single-ion detection²² and single-cell proteomics¹⁰³. Similarly, antibody-
518 based strategies such as immunoassays that rely on specific antigen-antibody interactions
519 have served as the standard methods for protein identification and quantification for the last
520 few decades. Single Molecule Array technologies (Simoa¹⁰⁴) commercialized by Quanterix is
521 one of the most sensitive single protein sensing antibody-based methods used for the
522 analysis of small analytical samples and clinical studies down to attomolar concentration
523 level¹⁰⁵. The SARS-Cov-2 pandemic has accelerated the development of a high-throughput
524 serological tests of clinical samples utilizing Simoa¹⁰⁶ based on ultrasmall blood samples.
525 These and other antibody based protein sensing method are likely to take greater share in the
526 biomedical sensing industry, in parallel to the emergence of other single molecule techniques
527 that will further permit comprehensive proteoform inference or differentiation.

528

529 The emerging landscape of alternative protein sequencing and fingerprinting technologies in
530 **Figure 1** could one day help to sequence human proteoforms in a more complete way. High-
531 throughput Edman degradation could pair with bottom-up MS strategies to improve current
532 sequence coverage limitations (**Box 1**). These bottom-up methods could benefit from
533 nanopore sequencing and DNA fluorescence-based methods that aim for long read
534 sequencing and structural fingerprinting of whole proteins. The integration of both existing and
535 emerging technologies promises to iteratively reveal an atlas of full length proteoforms, which
536 could itself assist these up-and-coming technologies to infer what cannot be directly measured
537 in terms of protein primary sequence and structure.

538

539 The far-reaching vision in single-molecule proteomics is in their applications for the analysis
540 of protein-protein interactions. A map covering a wide range of proteoforms and their PPIs is
541 an unmet milestone needed to finely understand protein networks in normal tissues and in
542 disease. Bottom-up MS-based approaches, such as cross-linking^{107,108} or affinity-purification

543 are implemented to identify physical¹⁰⁹ and proximal interactomes¹¹⁰. However, these
544 techniques present either biochemical or sample processing yield limitations, which brings
545 problems, including intra-protein cross-link over-representation, PPIs loss upon solubilization
546 and limitations inherent to MS analysis, hindering single-cell interactome analysis. As of today,
547 single-molecule analysis of PPIs has not reached main-stream proteomics, and single-cell
548 interactomics even more so. Achieving this goal would be of outstanding interest for
549 accurately defining e.g. protein organization within highly dynamic membraneless
550 organelles¹¹¹, such as resolving protein condensates, spatial and temporal organization at a
551 single organelle or single cell scale, which will provide an unprecedented resolution of PPIs
552 organization.

553 Challenges for next-generation protein sequencing

554 Two grand challenges await technological innovations that need to be solved to enable the
555 high-throughput sequencing of complex protein mixtures. Firstly, there is no method to amplify
556 the copy number of proteins, as is the case for nucleic acids. These new techniques focus on
557 characterizing individual proteins. The aim is to sequence proteomes starting from a low
558 number of cells or extremely minute samples often containing just a few or single copies of
559 specific proteins. This presents a second problem: A eukaryotic cell contains billions of
560 proteins. **While the presented methods may enable single molecule protein identification, in
561 order to profile all proteins in the cell they must reach an extremely high sensing throughput
562 to permit whole cell analysis within a reasonable time-scale. These two seemingly
563 contractionary requirements (single-protein molecule sensitivity and an extremely high
564 throughput) present one of the main challenges to the field and striking an optimal balance
565 among them will be key for all the technologies discussed.** Of the orthogonal methods
566 presented, nanopores, fluorosequencing, protein linear barcoding using DNA-PAINT to name
567 a few, stand a chance to eventually measure billions of proteins within a few hours.

568
569 To gain utility in both, research and clinical settings, emerging technologies will be evaluated
570 in terms of their sensitivity, specificity, **proteome coverage (number of proteins in the sample
571 covered), sequence coverage (average fraction of a protein sequence covered), peptide read
572 length (number of amino-acids covered by a single read), accuracy (error in calling an amino-
573 acid) and cost.** In this regard, additional research and validation will be required to
574 demonstrate the benefits of these orthogonal technologies. The formation of a dedicated
575 global academic/scientific community in single protein sequencing may catalyze further
576 development and implementation of these technologies for more widespread use.
577 Multidisciplinary conferences that bring together experts in chemistry, physics, biochemistry,

578 industry, computation, and other relevant areas of expertise (e.g. pathologists, clinicians) with
579 a clear vision of the most relevant problems and unmet needs, will need to be embraced.

580

581 **Acknowledgements**

582 The authors thank all the presenting delegates of the 2019 Single Molecule Protein
583 Sequencing conference (Jerusalem). Authors thank the PL-Grid Infrastructure, Poland for
584 providing their hardware and software resources. S.S. acknowledges the Postdoc Mobility
585 fellowship no. P400PB 180889 by the Swiss National Science Foundation. E.M.M. and E.V.A.
586 acknowledge funding from the NIH (R35 GM122480, R01 DK110520 to EMM), Welch
587 Foundation (F1515 to EMM and F-0046 to EVA), Army Research Office grant W911NF-12-1-
588 0390, and Erisyon, Inc. E.M.M. and E.V.A. are co-founders and shareholders of Erisyon. RTK
589 acknowledges funding from NIGMS (R01 GM138931). C.D. acknowledges the ERC Advanced
590 Grant Looping DNA (no. 883684) and the NWO programs NanoFront and Basyc. E.M.M.,
591 E.V.A., and C.J.H. are co-inventors on patents relevant to this work. S.O acknowledges
592 the support of the Azrieli fellowship foundation. NLK acknowledges funding from Paul G. Allen
593 Frontiers Program (11715), the NIH HuBMAP program (UH3 CA246635) and NIGMS (P41
594 GM108569). J.P.M. and Z.K acknowledge internal funding from NIST and are co-inventors on
595 patents relevant to this work. M.W. acknowledges funding from the NIH (HG009186). K.S. and
596 A.A. acknowledge funding from NSF (PHY-1430124). C. Joo, C.D, and R.E. acknowledge
597 funding from NWO-I (SMPS). Joo acknowledges funding from HFSP (RGP0026/2019). A.P.
598 acknowledges the Bekker fellowship no. PPN/BEK/2018/1/00296 from the Polish National
599 Agency for Academic Exchange. .M. and S.H. acknowledge funding from the European
600 Research Council (ERC “Enlightened”, GA 616251) and the CEA Transverse Program
601 “Instrumentation and Detection” (PTC-ID VIRIONEMS). Support from the Proteomics French
602 Infrastructure (PROFI) is also gratefully acknowledged. GD acknowledges funding from FNR
603 (C17/BM/11642138). M.M. acknowledges funding from the Adolphe Merkle Foundation, the
604 Michael J. Fox Foundation for Parkinson’s Research (Grant ID: 17924) and the Swiss National
605 Science Foundation (grant no. 200021-169304). A.M acknowledges funding from the
606 European Union’s Horizon 2020 research and innovation programme under grant agreement
607 No. 833399-ERC NanoProt-ID, and ISF award 3485/19. MC acknowledges the computational
608 resources from CINECA (NATWE project), and the Swiss National Super-computing Centre
609 (CSCS), projects ID sm11 and s865. E.C. acknowledges funding from I-Site Lille, Région
610 Hauts-de-France and European Union’s Horizon 2020 Marie Skłodowska-Curie No 843052.
611 The study was supported by the project “International Centre for Cancer Vaccine Science”
612 that is carried out within the International Agendas Programme of the Foundation for Polish

613 Science co-financed by the European Union under the European Regional Development
614 Fund. We thank Viktorija Globyte for critical reading.

615

616 **Author Contributions**

617 J.A. Alfaro., C. Joo, and A. Meller conceived of and initiated, coordinated, and supervised the
618 project. The first draft of the manuscript was written by J.A. Alfaro, C. Joo, A. Meller, P.
619 Bohländer, M. Filius, X. F. van Kooten , S. Ohayon, A. Pomorski, S. Schmid, C. J. Howard, M.
620 Dai, P. Samaras, G. Bedran, M. Wilhelm and L. Sepiashvili. Subsequently, the manuscript
621 was revised and approved by all authors.

622

623 **Competing Interests:**

624 S. Hentz and C. Masselon are co-inventors of the patent application EP14158255. E. Marcotte
625 is a co-inventor on patent 9625469. D. Stein is sponsored by Oxford Nanopore for his work on
626 nanotip mass spectrometry. E.M. Marcotte and E.V. Anslyn are co-founders and shareholders
627 of Erisyon, Inc. Some authors may be bound by confidentiality agreements that prevent them
628 from disclosing their competing interests in this work.

629 **References:**

- 630 1. Breuza, L. *et al.* The UniProtKB guide to the human proteome. *Database* **2016**, 1–10
631 (2016).
- 632 2. Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nat.*
633 *Methods* **10**, 186–187 (2013).
- 634 3. Seattle Times business staff. Seattle biotech startup Nautilus to get \$350 million, stock
635 listing in blank-check deal. *The Seattle Times*. (2021).
636 [https://www.seattletimes.com/business/seattle-biotech-startup-nautilus-to-get-350-](https://www.seattletimes.com/business/seattle-biotech-startup-nautilus-to-get-350-million-stock-listing-in-blank-check-deal/)
637 [million-stock-listing-in-blank-check-deal/](https://www.seattletimes.com/business/seattle-biotech-startup-nautilus-to-get-350-million-stock-listing-in-blank-check-deal/)
- 638 4. Reuters Staff. Protein sequencing firm Quantum-Si to go public via \$1.46 billion SPAC
639 merger. *Reuters* (2021). [https://www.reuters.com/article/us-quantum-si-m-a-high-cape-](https://www.reuters.com/article/us-quantum-si-m-a-high-cape-capital-idUSKBN2AI1HT)
640 [capital-idUSKBN2AI1HT](https://www.reuters.com/article/us-quantum-si-m-a-high-cape-capital-idUSKBN2AI1HT)
- 641 5. Cohen, L. & Walt, D. R. Single-Molecule Arrays for Protein and Nucleic Acid Analysis.
642 *Annu. Rev. Anal. Chem.* **10**, 345–363 (2017).
- 643 6. Aggarwal, V. & Ha, T. Single-molecule fluorescence microscopy of native
644 macromolecular complexes. *Curr. Opin. Struct. Biol.* **41**, 225–232 (2016).
- 645 7. Edman, P. A method for the determination of the amino acid sequence in peptides.
646 *Arch. Biochem.* **22**, 475–476 (1949).
- 647 8. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single
648 molecule peptide sequencing. *PLoS Comput. Biol.* **11**, 1076–1082 (2015).
- 649 9. Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in
650 zeptomole-scale mixtures. *Nat. Biotechnol.* **36**, 1076–1082 (2018).
- 651 10. Howard, C. J. *et al.* Solid-phase peptide capture and release for bulk and single-
652 molecule proteomics. *ACS Chem. Biol.* **15**, 1401–1407 (2020).
- 653 11. Miclotte, G., Martens, K. & Fostier, J. Computational assessment of the feasibility of
654 protonation-based protein sequencing. *PLoS One* **15**, e0238625 (2020).

- 655 12. Tullman, J., Marino, J. P. & Kelman, Z. Leveraging nature's biomolecular designs in
656 next-generation protein sequencing reagent development. *Appl. Microbiol. Biotechnol.*
657 1–11 (2020).
- 658 13. Rodrigues, S. G., Marblestone, A. H. & Boyden, E. S. A theoretical analysis of single
659 molecule protein sequencing via weak binding spectra. *PLoS One* **14**, e0212868
660 (2019).
- 661 14. Tullman, J., Callahan, N., Ellington, B., Kelman, Z. & Marino, J. P. Engineering ClpS for
662 selective and enhanced N-terminal amino acid binding. *Appl. Microbiol. Biotechnol.*
663 **103**, 2621–2633 (2019).
- 664 15. Smith, R. D., Cheng, X., Brace, J. E., Hofstadler, S. A. & Anderson, G. A. Trapping,
665 detection and reaction of very large single molecular ions by mass spectrometry.
666 *Nature* **369**, 137–139 (1994).
- 667 16. Keifer, D. Z. & Jarrold, M. F. Single-molecule mass spectrometry. *Mass Spectrom. Rev.*
668 **36**, 715–733 (2017).
- 669 17. Rose, R. J., Damoc, E., Denisov, E., Makarov, A. & Heck, A. J. High-sensitivity Orbitrap
670 mass analysis of intact macromolecular assemblies. *Nat. Methods* **9**, 1084–1086
671 (2012).
- 672 18. Makarov, A. & Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass
673 analyzer. *J. Am. Soc. Mass Spectr.* **20**, 1486–1495 (2009).
- 674 19. Kafader, J. O. *et al.* Measurement of individual ions sharply increases the resolution of
675 orbitrap mass spectra of proteins. *Anal. Chem.* **91**, 2776–2783 (2019).
- 676 20. Kafader, J. O. *et al.* Multiplexed mass spectrometry of individual ions improves
677 measurement of proteoforms and their complexes. *Nat. Methods* **17**, 391–394 (2020).
- 678 21. Wörner, T. P. *et al.* Resolving heterogeneous macromolecular assemblies by Orbitrap-
679 based single-particle charge detection mass spectrometry. *Nat. Methods* **17**, 395–398
680 (2020).

- 681 22. Kafader, J. O. *et al.* Individual ion mass spectrometry enhances the sensitivity and
682 sequence coverage of top down mass spectrometry. *J. Proteome Res.* **19**, 1346–1350
683 (2020).
- 684 23. Smith, L. *et al.* The Human Proteoform Project: A Plan to Define the Human Proteome.
685 *Preprints* (2020) doi:10.20944/preprints202010.0368.v1.
- 686 24. Ekinci, K. L., Huang, X. M. H. & Roukes, M. L. Ultrasensitive nanoelectromechanical
687 mass detection. *Appl. Phys. Lett.* **84**, 4469–4471 (2004).
- 688 25. Hanay, M. S. *et al.* Single-protein nanomechanical mass spectrometry in real time. *Nat.*
689 *Nanotechnol.* **7**, 602-608 (2012).
- 690 26. Sage, E. *et al.* Neutral particle mass spectrometry with nanomechanical systems. *Nat.*
691 *Commun.* **6**, 1–5 (2015).
- 692 27. Dominguez-Medina, S. *et al.* Neutral mass spectrometry of virus capsids above 100
693 megadaltons with nanomechanical resonators. *Science* **362**, 918–922 (2018).
- 694 28. Chaste, J. *et al.* A nanomechanical mass sensor with yoctogram resolution. *Nat.*
695 *Nanotechnol.* **7**, 301–304 (2012).
- 696 29. Hanay, M. S. *et al.* Inertial imaging with nanomechanical systems. *Nat. Nanotechnol.*
697 **10**, 339–334 (2015).
- 698 30. Malvar, O. *et al.* Mass and stiffness spectrometry of nanoparticles and whole intact
699 bacteria by multimode nanomechanical resonators. *Nat. Commun.* **7**, 1–8 (2016).
- 700 31. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal.*
701 *Chem.* **68**, 1–8 (1996).
- 702 32. El-Faramawy, A., Siu, K. M. & Thomson, B. A. Efficiency of nano-electrospray
703 ionization. *J. Am. Soc. Mass Spectr.* **16**, 1702–1707 (2005).
- 704 33. Bush, J. *et al.* The nanopore mass spectrometer. *Rev. Sci. Instrum.* **88**, 113307 (2017).
- 705 34. Maulbetsch, W., Wiener, B., Poole, W., Bush, J. & Stein, D. Preserving the sequence of
706 a biopolymer's monomers as they enter an electrospray mass spectrometer. *Phys.*
707 *Rev. Appl.* **6**, 054006-1–054006-9 (2016).

- 708 35. Brodbelt, J. S. Photodissociation mass spectrometry: new tools for characterization of
709 biological molecules. *Chem. Soc. Rev.* **43**, 2757–2783 (2014).
- 710 36. Chang, S. *et al.* Tunnelling readout of hydrogen-bonding-based recognition. *Nat.*
711 *Nanotechnol.* **4**, 297–301 (2009).
- 712 37. Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by
713 recognition tunnelling. *Nat. Nanotechnol.* **9**, 466–473 (2014).
- 714 38. Ohshiro, T. *et al.* Detection of post-translational modifications in single peptides using
715 electron tunnelling currents. *Nat. Nanotechnol.* **9**, 835 (2014).
- 716 39. Zhang, B. *et al.* Observation of giant conductance fluctuations in a protein. *Nano*
717 *Futures* **1**, 035002 (2017).
- 718 40. Zhang, B. *et al.* Engineering an enzyme for direct electrical monitoring of activity. *ACS*
719 *Nano* **14**, 1360-1368 (2020).
- 720 41. Seeman, N. C. & Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mat.* **3**, 1–23 (2017).
- 721 42. Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R.
722 Super-resolution microscopy with DNA-PAINT. *Nat. Protoc.* **12**, 1198–1228 (2017).
- 723 43. Dai, M., Jungmann, R. & Yin, P. Optical imaging of individual biomolecules in densely
724 packed clusters. *Nat. Nanotechnol.* **11**, 798–807 (2016).
- 725 44. Jungmann, R. *et al.* Quantitative super-resolution imaging with qPAINT. *Nat. Methods*
726 **13**, 439–442 (2016).
- 727 45. Dai, M. & Yin, P. Methods and compositions relating to super-resolution imaging and
728 modification. (2018).
- 729 46. Schaus, T. E., Woo, S., Xuan, F., Chen, X. & Yin, P. A DNA nanoscope via auto-cycling
730 proximity recording. *Nat. Commun.* **8**, 1–9 (2017).
- 731 47. Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable
732 autonomous synthesis of single-stranded DNA. *Nat. Chem.* **10**, 155–164 (2018).
- 733 48. Gopalkrishnan, N., Punthambaker, S., Schaus, T. E., Church, G. M. & Yin, P. A DNA
734 nanoscope that identifies and precisely localizes over a hundred unique molecular

- 735 features with nanometer accuracy. *bioRxiv* 2020.08.27.271072 (2020)
736 doi:10.1101/2020.08.27.271072.
- 737 49. Filius, M., Kim, S. H., Severins, I. & Joo, C. High-resolution single-molecule FRET via
738 DNA eXchange (FRET X). *bioRxiv* 2020.10.15.340885 (2020)
739 doi:10.1101/2020.10.15.340885.
- 740 50. Lerner, E. *et al.* Toward dynamic structural biology: Two decades of single-molecule
741 Förster resonance energy transfer. *Science* **359**, eaan1133 (2018).
- 742 51. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of
743 individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci.*
744 *U.S.A.* **93**, 13770–13773 (1996).
- 745 52. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat.*
746 *Biotechnol.* **34**, 518–524 (2016).
- 747 53. Loman, N. J. & Watson, M. Successful test launch for nanopore sequencing. *Nat.*
748 *Methods* **12**, 303–304 (2015).
- 749 54. Rozevsky, Y. *et al.* Quantification of mRNA expression using single-molecule nanopore
750 sensing. *ACS Nano* **14**, 13964–13974 (2020).
- 751 55. Di Muccio, G., Rossini, A. E., Di Marino, D., Zollo, G. & Chinappi, M. Insights into
752 protein sequencing with an α -Hemolysin nanopore by atomistic simulations. *Sci. Rep.*
753 **9**, 1–8 (2019).
- 754 56. Wilson, J., Sarthak, K., Si, W., Gao, L. & Aksimentiev, A. Rapid and accurate
755 determination of nanopore ionic current using a steric exclusion model. *ACS Sens.* **4**,
756 634–644 (2019).
- 757 57. Huang, G., Voet, A. & Maglia, G. FraC nanopores with adjustable diameter identify the
758 mass of opposite-charge peptides with 44 dalton resolution. *Nat. Commun.* **10**, 1–10
759 (2019).
- 760 58. Piguet, F. *et al.* Identification of single amino acid differences in uniformly charged
761 homopolymeric peptides with aerolysin nanopore. *Nat. Commun.* **9**, 1–13 (2018).

- 762 59. Cao, C. *et al.* Single-molecule sensing of peptides and nucleic acids by engineered
763 aerolysin nanopores. *Nat. Commun.* **10**, 1–11 (2019).
- 764 60. Galenkamp, N. S., Soskine, M., Hermans, J., Wloka, C. & Maglia, G. Direct electrical
765 quantification of glucose and asparagine from bodily fluids using nanopores. *Nat.*
766 *Commun.* **9**, 1–8 (2018).
- 767 61. Ouldali, H. *et al.* Electrical recognition of the twenty proteinogenic amino acids using an
768 aerolysin nanopore. *Nat. Biotechnol.* **38**, 176–181 (2020).
- 769 62. Restrepo-Pérez, L., Wong, C. H., Maglia, G., Dekker, C. & Joo, C. Label-free detection
770 of post-translational modifications with a nanopore. *Nano Lett.* **19**, 7957–7964 (2019).
- 771 63. Korotkov, K. V., Sandkvist, M. & Hol, W. G. The type II secretion system: biogenesis,
772 molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10**, 336–351 (2012).
- 773 64. Olivares, A. O., Baker, T. A. & Sauer, R. T. Mechanical protein unfolding and
774 degradation. *Annu. Rev. Physiol.* **80**, 413–429 (2018).
- 775 65. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through
776 an α -hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).
- 777 66. Nivala, J., Mulrone, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among protein
778 variants using an unfoldase-coupled nanopore. *ACS Nano* **8**, 12365–12375 (2014).
- 779 67. Zhang, S. *et al.* Bottom-up fabrication of a multi-component nanopore sensor that
780 unfolds, processes and recognizes single proteins. *bioRxiv* 2020.12.04.411884 (2020)
781 doi:10.1101/2020.12.04.411884.
- 782 68. Sachelaru, I. *et al.* YidC and SecYEG form a heterotetrameric protein translocation
783 channel. *Sci. Rep.* **7**, 1–15 (2017).
- 784 69. Knyazev, D. G., Kuttner, R., Zimmermann, M., Sobakinskaya, E. & Pohl, P. Driving
785 forces of translocation through bacterial translocon SecYEG. *J. Membrane Biol.* **251**,
786 329–343 (2018).

- 787 70. Backes, S. & Herrmann, J. M. Protein translocation into the intermembrane space and
788 matrix of mitochondria: mechanisms and driving forces. *Front. Mol. Biosci.* **4**, 83 1-11
789 (2017).
- 790 71. Feng, J. *et al.* Transmembrane protein rotaxanes reveal kinetic traps in the refolding of
791 translocated substrates. *Commun. Biol.* **3**, 1–11 (2020).
- 792 72. Rosen, C. B., Bayley, H. & Rodriguez-Larrea, D. Free-energy landscapes of membrane
793 co-translocational protein unfolding. *Commun. Biol.* **3**, 1–9 (2020).
- 794 73. Cardozo, N. *et al.* Multiplexed direct detection of barcoded protein reporters on a
795 nanopore array. *bioRxiv* 837542 (2019) doi:10.1101/837542.
- 796 74. Yao, Y., Docter, M., Van Ginkel, J., de Ridder, D. & Joo, C. Single-molecule protein
797 sequencing through fingerprinting: computational assessment. *Phys. Biol.* **12**, 055003
798 (2015).
- 799 75. Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S. & Meller, A. Simulation of single-
800 protein nanopore sensing shows feasibility for whole-proteome identification. *PLoS*
801 *Comput. Biol.* **15**, e1007067 (2019).
- 802 76. Restrepo-Pérez, L. *et al.* Resolving chemical modifications to a single amino acid within
803 a peptide using a biological nanopore. *ACS Nano* **13**, 13668-13676 (2019).
- 804 77. Asandei, A. *et al.* Placement of oppositely charged aminoacids at a polypeptide termini
805 determines the voltage-controlled braking of polymer transport through nanometer-
806 scale pores. *Sci. Rep.* **5**, 10419 (2015).
- 807 78. Wang, R. *et al.* Single-molecule discrimination of labeled DNAs and polypeptides using
808 photoluminescent-free TiO₂ nanopores. *ACS Nano* **12**, 11648–11656 (2018).
- 809 79. Zrehen, A., Ohayon, S., Huttner, D. & Meller, A. On-chip protein separation with single-
810 molecule resolution. *Sci. Rep.* **10**, 1–12 (2020).
- 811 80. Assad, O. N. *et al.* Light-enhancing plasmonic-nanopore biosensor for superior single-
812 molecule detection. *Adv. Mater.* **29**, 1605442 (2017).

- 813 81. Spitzberg, J. D., Zrehen, A., van Kooten, X. F. & Meller, A. Plasmonic-nanopore
814 biosensors for superior single-molecule detection. *Adv. Mater.* **31**, 1900422 (2019).
- 815 82. Houghtaling, J., List, J. & Mayer, M. Nanopore-based, rapid characterization of
816 Individual amyloid particles in solution: concepts, challenges, and prospects. *Small* **14**,
817 1802412 (2018).
- 818 83. Plesa, C. *et al.* Fast translocation of proteins through solid state nanopores. *Nano Lett.*
819 **13**, 658–663 (2013).
- 820 84. Yusko, E. C. *et al.* Controlling protein translocation through nanopores with bio-inspired
821 fluid walls. *Nat. Nanotechnol.* **6**, 253 (2011).
- 822 85. Houghtaling, J. *et al.* Estimation of shape, volume, and dipole moment of individual
823 proteins freely transiting a synthetic nanopore. *ACS Nano* **13**, 5231–5242 (2019).
- 824 86. Yusko, E. C. *et al.* Real-time shape approximation and fingerprinting of single proteins
825 using a nanopore. *Nat. Nanotechnol.* **12**, 360 (2017).
- 826 87. Pang, Y. & Gordon, R. Optical trapping of a single protein. *Nano Lett.* **12**, 402–406
827 (2012).
- 828 88. Verschueren, D., Shi, X. & Dekker, C. Nano-optical tweezing of single proteins in
829 plasmonic nanopores. *Small Methods* **3**, 1800465 (2019).
- 830 89. Schmid, S. & Dekker, C. Nanopores—a versatile tool to study protein dynamics. *Essays*
831 *Biochem.* EBC20200020 (2020).
- 832 90. Larkin, J., Henley, R. Y., Muthukumar, M., Rosenstein, J. K. & Wanunu, M. High-
833 bandwidth protein analysis using solid-state nanopores. *Biophys. J.* **106**, 696–704
834 (2014).
- 835 91. Nir, I., Huttner, D. & Meller, A. Direct sensing and discrimination among ubiquitin and
836 ubiquitin chains using solid-state nanopores. *Biophys. J.* **108**, 2340–2349 (2015).
- 837 92. Waduge, P. *et al.* Nanopore-based measurements of protein size, fluctuations, and
838 conformational changes. *ACS Nano* **11**, 5706–5716 (2017).

- 839 93. Varongchayakul, N., Hersey, J. S., Squires, A., Meller, A. & Grinstaff, M. W. A Solid-
840 state hard microfluidic–nanopore biosensor with multilayer fluidics and on-chip
841 bioassay/purification chamber. *Adv. Funct. Mater.* **28**, 1804182 (2018).
- 842 94. Hu, R. *et al.* Differential enzyme flexibility probed using solid-state nanopores. *ACS*
843 *Nano* **12**, 4494–4502 (2018).
- 844 95. Huang, G. *et al.* Electro-osmotic vortices promote the capture of folded proteins by
845 PlyAB nanopores. *Nano Lett.* **20**, 3819–3827 (2020).
- 846 96. Soskine, M., Biesemans, A. & Maglia, G. Single-molecule analyte recognition with ClyA
847 nanopores equipped with internal protein adaptors. *J. Am. Chem. Soc.* **137**, 5793–5797
848 (2015).
- 849 97. Wloka, C. *et al.* Label-free and real-time detection of protein ubiquitination with a
850 biological nanopore. *ACS Nano* **11**, 4387–4394 (2017).
- 851 98. Aramesh, M. *et al.* Localized detection of ions and biomolecules with a force-controlled
852 scanning nanopore microscope. *Nat. Nanotechnol.* **14**, 791–798 (2019).
- 853 99. Hernandez, E. T., Swaminathan, J., Marcotte, E. M. & Anslyn, E. V. Solution-phase and
854 solid-phase sequential, selective modification of side chains in KDYWEC and KDYWE
855 as models for usage in single-molecule protein sequencing. *New J. Chem.* **41**, 462–469
856 (2017).
- 857 100. Kong, A. T., Lprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I.
858 MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–
859 based proteomics. *Nat. Methods* **14**, 513 (2017).
- 860 101. Zhong, J. *et al.* Proteoform characterization based on top-down mass spectrometry.
861 *Brief. Bioinformatics*, bbaa015 (2020).
- 862 102. Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry.
863 *Proteomics* **4**, 1534–1536 (2004).
- 864 103. Marx, V. A dream of single-cell proteomics. *Nat. Methods* **16**, 809–812 (2019).

- 865 104. Rissin, D. M. *et al.* Single-molecule enzyme-linked immunosorbent assay detects
866 serum proteins at subfemtomolar concentrations. *Nat. Biotechnol.* **28**, 595–599 (2010).
- 867 105. Wu, C., Garden, P. M. & Walt, D. R. Ultrasensitive detection of attomolar protein
868 concentrations by dropcast single molecule assays. *J. Am. Chem. Soc.* **142**, 12314–
869 12323 (2020).
- 870 106. Norman, M. *et al.* Ultrasensitive high-resolution profiling of early seroconversion in
871 patients with COVID-19. *Nat. Biomed. Eng.* **4**, 1180–1187 (2020).
- 872 107. Liu, F., Rijkers, D. T., Post, H. & Heck, A. J. Proteome-wide profiling of protein
873 assemblies by cross-linking mass spectrometry. *Nat. Methods* **12**, 1179–1184 (2015).
- 874 108. Iacobucci, C., Götze, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer
875 view on protein interaction networks. *Curr. Opin. Biotechnol.* **63**, 48–53 (2020).
- 876 109. Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass
877 spectrometry: Basic principles and strategies. *Proteomics* **12**, 1576–1590 (2012).
- 878 110. Gentzel, M., Pardo, M., Subramaniam, S., Stewart, A. F. & Choudhary, J. S. Proteomic
879 navigation using proximity-labeling. *Methods* **164**, 67–72 (2019).
- 880 111. Zhao, Y. G. & Zhang, H. Phase separation in membrane biology: the interplay between
881 membrane-bound organelles and membraneless condensates. *Dev. Cell* (2020).
- 882 112. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol.*
883 *Cell Biol.* **5**, 699–711 (2004).
- 884 113. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in top-down proteomics and the
885 analysis of proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499–519 (2016).
- 886 114. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome
887 profiling of 10–100 mammalian cells. *Nat. Commun.* **9**, 1–10 (2018).
- 888 115. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of
889 single mammalian cells quantifies proteome heterogeneity during cell differentiation.
890 *Genome Biol.* **19**, 161 (2018).

- 891 116. Zhu, Y. *et al.* Proteomic analysis of single mammalian cells enabled by microfluidic
892 nanodroplet sample preparation and ultrasensitive NanoLC-MS. *Angew. Chem. Int. Ed.*
893 **57**, 12370–12374 (2018).
- 894 117. Kelly, R. T. Single-cell proteomics: Progress and prospects. *Mol. & Cell. Proteom.* **19**,
895 1739–1748 (2020).
- 896 118. Gavrilyuk, J., Ban, H., Nagano, M., Hakamata, W. & Barbas, C. F. Formylbenzene
897 diazonium hexafluorophosphate reagent for tyrosine-selective modification of proteins
898 and the introduction of a bioorthogonal aldehyde. *Bioconjugate Chem.* **23**, 2321–2328
899 (2012).
- 900 119. Ban, H., Gavrilyuk, J. & Barbas III, C. F. Tyrosine bioconjugation through aqueous ene-
901 type reactions: a click-like reaction for tyrosine. *J. Am. Chem. Soc.* **132**, 1523–1525
902 (2010).
- 903 120. Bach, K., Beerkens, B. L., Zanon, P. R. & Hacker, S. M. Light-activatable, 2, 5-
904 disubstituted tetrazoles for the proteome-wide profiling of aspartates and glutamates in
905 living bacteria. *ACS Cent. Sci.* **6**, 546–554 (2020).
- 906 121. Taylor, M. T., Nelson, J. E., Suero, M. G. & Gaunt, M. J. A protein functionalization
907 platform based on selective reactions at methionine residues. *Nature* **562**, 563–568
908 (2018).
- 909 122. Lin, S. *et al.* Redox-based reagents for chemoselective methionine bioconjugation.
910 *Science* **355**, 597–602 (2017).
- 911 123. Christian, A. H. *et al.* A physical organic approach to tuning reagents for selective and
912 stable methionine bioconjugation. *J. Am. Chem. Soc.* **141**, 12657–12662 (2019).
- 913 124. Jia, S., He, D. & Chang, C. J. Bioinspired thiophosphorodichloridate reagents for
914 chemoselective histidine bioconjugation. *J. Am. Chem. Soc.* **141**, 7294–7301 (2019).
- 915 125. Bloom, S. *et al.* Decarboxylative alkylation for site-selective bioconjugation of native
916 proteins via oxidation potentials. *Nat. Chem.* **10**, 205 (2018).

- 917 126. Rosen, C. B. & Francis, M. B. Targeting the N terminus for site-selective protein
918 modification. *Nat. Chem. Biol.* **13**, 697–705 (2017).
- 919 127. Busch, G. K. *et al.* Specific N-terminal protein labelling: use of FMDV 3C pro protease
920 and native chemical ligation. *Chem. Commun.* **29**, 3369–3371 (2008).
- 921 128. Bandyopadhyay, A., Cambray, S. & Gao, J. Fast and selective labeling of N-terminal
922 cysteines at neutral pH via thiazolidino boronate formation. *Chem. Sci.* **7**, 4589–4593
923 (2016).
- 924 129. Agten, S. M., Dawson, P. E. & Hackeng, T. M. Oxime conjugation in protein chemistry:
925 from carbonyl incorporation to nucleophilic catalysis. *J. Pept. Sci.* **22**, 271–279 (2016).
- 926 130. MacDonald, J. I., Munch, H. K., Moore, T. & Francis, M. B. One-step site-specific
927 modification of native proteins with 2-pyridinecarboxaldehydes. *Nat. Chem. Biol.* **11**,
928 326 (2015).
- 929 131. Matheron, L. *et al.* Improving the selectivity of the phosphoric acid β -elimination on a
930 biotinylated phosphopeptide. *J. Am. Soc. Mass Spectr.* **23**, 1981–1990 (2012).
- 931 132. Du, J. *et al.* Metabolic glycoengineering: Sialic acid and beyond. *Glycobiology* **19**,
932 1382–1401 (2009).
- 933 133. Tommasone, S. *et al.* The challenges of glycan recognition with natural and artificial
934 receptors. *Chem. Soc. Rev.* **48**, 5488–5505 (2019).
- 935

936 **Boxes:**

937

938 **Box 1: mass spectrometry-based global proteomics**

939 The last decade saw the maturation of mass spectrometry use in global proteomics. The
940 typical proteomics workflow is “bottom-up” in nature and involves digesting a protein sample
941 using a protease and characterizing the resulting peptides by mass spectrometry (MS)¹¹². Two
942 types of measurements are typically made in succession: (1) MS¹ spectra survey the masses
943 of a set of peptides present in the mass-spectrometer at a given moment and (2) MS² spectra
944 probe the structures of peptide ion species identified in the MS¹ survey by isolating,
945 fragmenting and measuring the fragment masses of one or a few of them. Peptides identified
946 from the MS² spectra are then mapped back to the proteins they come from to infer their overall
947 protein abundances.

948

949 Current mass spectrometers have drawbacks in terms of their dynamic range, the read-length
950 (peptide-length) of “sequenced” peptides, and biases in detectability arising from the ionization
951 mechanism, transmission and the mass analyzer used. Consequently, although “top-down”
952 proteomics methods capable of analyzing intact proteins exist¹¹³, most state-of-the-art
953 proteomics approaches characterize the proteome with high numbers of proteins but on
954 average the proteins are characterized with low unambiguous sequence coverage and low
955 sequencing depth. Different sample-preparation strategies, instruments and elution profiles
956 can improve the numbers and average sequence coverage of proteins identified in an
957 experiment. Summarizing the best single-sample run from 47 experiments (a summary of over
958 1000 distinct samples) in ProteomicsDB⁴⁹ (**Figure Box1**) reveals that even with complex
959 sample preparation, the average sequence coverage for a single sample reaches just 33%.

960 **Box 2: MS-based single-cell proteomics**

961 The dream of extending mass spectrometry (MS)-based proteomics to the single-cell level
962 eluded researchers for decades. Even as the sensitivity of MS instrumentation improved to
963 provide single-cell-compatible detection limits, samples comprising at least thousands of cells
964 were in practice required to obtain an in-depth proteome profile. Two recent advances have
965 made single-cell proteomics a reality. Miniaturized sample processing workflows such as
966 nanoPOTS¹¹⁴ (Nanodroplet Processing in One pot for Trace Samples) have dramatically
967 increased the efficiency of single cell sample preparation. NanoPOTS utilizes a robotic
968 nanopipettor to interface with a microfabricated nanowell plate. The reduced surface contact
969 and increased protein concentrations within the nanoliter droplets dramatically enhance

970 digestion kinetics and increase sample recovery for single cells and other trace samples.
971 Concurrently, multiplexed strategies (i.e., Single Cell Proteomics by Mass Spectrometry;
972 SCoPE-MS)¹¹⁵ have been developed in which proteins from single cells are labeled with
973 unique isobaric tags, and several cells are analyzed together in the presence of a larger carrier
974 sample. Single cells and carrier provide a combined MS signal for each protein, and unique
975 reporter ions released upon fragmentation enable protein quantification for each cell. While
976 nanoPOTS and SCoPE-MS originally enabled quantification of hundreds of proteins^{115,116}, the
977 combination of the two techniques, as well as advances in miniaturized liquid chromatography
978 and gas-phase separations, now enable >1000 proteins to be quantified from single
979 mammalian cells¹¹⁷.

980 Box 3 High-throughput Edman-Sequencing.

981 In high-throughput Edman fluorosequencing, proteins are digested to shorter peptides and
982 immobilized on a glass surface using the C-terminus⁸. Multiple rounds of Edman degradation
983 coupled to fluorescence microscopy are used for sequencing. Specific amino acids are
984 covalently labeled with spectrally distinguishable fluorophores, and the peptide fingerprint
985 comes from measuring the decrease in fluorescence of peptides following Edman
986 degradation⁹. Much like in mass spectrometry, the partial sequence is mapped back to a
987 reference proteome within a probabilistic framework. In another method, NAAB probes
988 specifically recognize each N-terminal amino acid of an unlabeled peptide for more complete
989 amino acid identification¹².

990 Box 4: DNA-PAINT

991 DNA-PAINT relies on the transient binding of dye-labeled DNA strands (imagers) to their
992 complementary target sequence (docking site) attached to a molecule of interest. The
993 transient binding of imager strands is detected as 'blinking' in an intensity versus time trace.
994 DNA-PAINT has a few unique advantages. First, the blinking kinetics (on- and off-rates) can
995 be tuned over a wide range, by altering the length and sequence of the imager strands, or
996 buffer conditions, making the method compatible with different sample conditions. Second,
997 the repetitive binding with different imager strands makes the target "non-bleachable",
998 collecting a large number of high-quality and high-precision blinking events, allowing for high-
999 sensitivity imaging on single-molecule targets, and with discrete molecular resolution (<5 nm).
1000 Finally, combined with orthogonal sequence labels, DNA-PAINT can be multiplexed by
1001 imaging with up to dozens of molecular species (exchange-PAINT).

1002 Box 5: Chemistry concepts in protein sequencing

1003 **(a) labelling efficiency and stability.** The challenges in labeling efficiency and stability are
1004 well characterized in fluorosequencing, which uses harsh conditions (including neat
1005 trifluoroacetic acid) which can lead to the reversal of maleimide-labeled cysteine residues. To
1006 circumvent this reversal, fluorosequencing instead utilizes the iodoacetamide chemistry which
1007 generates a more stable bond. Another point of complexity is that full conversion is dictated
1008 by solvent accessibility of targeted amino acid side chains and has an influence on the labeling
1009 efficiency. However, modeling suggests labeling efficiencies and stabilities significantly less
1010 than 100% can be compensated for computationally, at least to some degree, during the
1011 reference database matching process⁸.

1012

1013 **(b) Labelling side chains.** The most widely accessible labels are those that target lysine and
1014 cysteine residues using NHS esters, maleimide, and iodoacetamide reactive groups,
1015 respectively, (**Figure 5a** and **b**). Additionally, the phenol ring of tyrosine can be labeled using
1016 benzyl diazo groups¹¹⁸ (**Figure 5c**), however, the attachment of fluorescent molecules
1017 generally requires a two-step labeling procedure due to the cross-reactivity with fluorescent
1018 molecules. Another robust bioconjugation method to selectively target tyrosine side chains is
1019 an ene-like reaction with cyclic diazodicaboxamides in aqueous buffer¹¹⁹. Carboxylic acids
1020 have also been labeled on peptides, but due to the similar reactivities between Asp, Glu, and
1021 the C-terminus, this has primarily been used on synthetic peptides. The method makes use of
1022 a standard technique (EDC-coupling) for binding amines covalently to carboxylic acids,
1023 forming an amide bond (**Figure 5d**). A recently reported promising bioconjugation approach
1024 has shown that light-activated 2,5-disubstituted tetrazoles are able to convert glutamic and
1025 aspartic acid residues in high yields¹²⁰. Finally, tryptophan can be labeled at the C-2 position
1026 using sulfenyl chlorides (**Figure 5e**). However, this comes with limitations that the reaction is
1027 extremely water sensitive and the reactive group must be made *in situ*⁹⁹. There are also
1028 promising new methods that allow for chemical modifications of other amino acids. Methionine,
1029 for example, can either be elegantly labeled with hypervalent iodine reagents¹²¹ or by the use
1030 of urea-derived oxaziridines^{122,123}. Recently, a histidine-selective conjugation methodology
1031 was reported where thiophosphorodichloridates selectively form a covalent bond with
1032 histidines in proteins¹²⁴.

1033

1034 **(c) C-terminal labeling.** Labeling of the C-terminus brings a challenge in that it must be
1035 separated from aspartic and glutamic acid, which carry the same functionality. A photoredox
1036 reaction on the C-terminus of peptides and proteins by de-carboxylation of the C-terminal
1037 carboxylic acid followed by an alkylation step by a Michael acceptor has been recently

1038 reported¹²⁵. Due to their higher oxidation potential, the carboxylates of internal amino acid
1039 chains are less prone to this modification, making the method highly site-selective. This
1040 technique has been applied to a variety of peptide substrates as well as the C-terminus-
1041 specific alkylation of human insulin A (**Figure 5f**).

1042

1043 **(d) N-terminal labeling.** Several methods exist for modifying the N-terminus¹²⁶. Classic
1044 approaches like reductive amination with aldehydes or acylation with NHS-esters, which rely
1045 on pH control to increase the selectivity, are not sufficiently specific. Other strategies involve
1046 the side chain of the N-terminal amino acid. Native chemical ligation¹²⁷ or condensation
1047 reactions with aldehydes¹²⁸, could be used to label N-terminal cysteine, serine, threonine or
1048 tryptophan residues. Furthermore, oxidizing N-terminal serine or threonine residues to their
1049 corresponding aldehydes allows oxime conjugation with hydrazides or hydroxylamines¹²⁹. A
1050 more general methodology has emerged where the N-terminal amine condenses with the 2-
1051 pyridinecarboxaldehyde (2PCA), forming an imine structure, which further reacts in a
1052 cyclisation with the nearby amide nitrogen of the second amino acid to form the stable
1053 imidazolidinone product¹³⁰. This reaction has recently been shown to be useful for single-
1054 molecule peptide sequencing as a method for the immobilization of peptides onto a solid-
1055 phase resin, multiple chemical derivatization steps without purification, and subsequent
1056 traceless release prior to fluorosequencing¹⁰.

1057 **(e) PTMs.** As an example of elimination replacement chemistries, phospho-serine and
1058 phospho-threonine residues can be labeled by β -elimination followed by Michael addition
1059 (BEMA). In mass-spectrometry-based phospho-proteomics, it is used to introduce an
1060 additional trypsin cleavage site at the phosphorylated amino acid¹³¹, whereas at the single
1061 molecule level it can be utilized to site-specifically attach a fluorescent label. Such approach
1062 has been established for Edman degradation described above⁹.

1063 Protein glycosylation can be complex, featuring many different types of monomeric units
1064 bound in possibly branching polymer structures. Their full structural characterization often
1065 requires derivatization and is done on glycans that are released from the protein. Therefore,
1066 schemes for understanding site-specific and simple glycosylation events should be the current
1067 focus. N-glycan anchoring asparagine residue can be converted to aspartate by glycan
1068 removal with PNGase F enzyme practically for all protein sequencing approaches, reducing
1069 the complexity to the detection of this mutation. Another possibility to introduce site-selective
1070 labels is the incorporation of azide-tagged glycans by adding modified carbohydrates to the
1071 cell medium¹³². In other detection schemes the location could be also inferred using glycan-
1072 specific reporter molecules such as lectins, engineered proteins or aptamers¹³³.

1073
1074

1075 Figure Legends:

1076 **Figure Box1: Sequence coverage in global proteomics studies.** MS-based global
1077 proteomics studies identify and quantify the proteins with variable sequence coverage. The
1078 single best run from 47 publications present in proteomicsDB shows how sample-specific
1079 protein sequence coverage improves with sample preparation methods. Sequence coverage
1080 generally decreases with sample complexity and increases with time (cost) dedicated to
1081 studying the sample.

1082

1083 **Figure 1: The emerging landscape of single-molecule protein sequencing and**
1084 **fingerprinting technologies.** The new proteomics landscape can be understood in terms of
1085 the type of analyte that is being studied, the method of protein identification, and the target
1086 niches in proteomics. Various techniques, particularly those involving complex readout
1087 signals, are suitable to characterize short peptide sequences, while others are primed to
1088 characterize full-length proteins or larger complexes. Technologies may specialize in short
1089 peptides (Peptides in the figure), whole proteins (Proteins) or macromolecular complexes
1090 (Complexes). The method of protein identification may fingerprint certain classes of amino
1091 acids (aa-fingerprinting), reveal each amino acid down to its physiochemical class or better
1092 (aa-sequencing). Much like mass spectrometry, technologies might characterize proteins by
1093 their masses and/or the masses of their fragments (Mass spectrum). Other methods aim to
1094 characterize properties of folded proteins (Structural fingerprint). The target niches could
1095 include the study of specific PTMs or deciphering whole proteoforms (PTM/proteoform
1096 inference), analyzing purified proteins or complex mixtures of proteins (Complex mixtures).
1097 Other applications can include protein interaction inference (PPI-studies) or glimpsing insights
1098 into protein structure (Structure).

1099

1100 **Figure 2: The renaissance of classic techniques.** High-throughput fluorosequencing by
1101 Edman degradation featuring (a) amino acid-specific chemical modification of peptides with
1102 fluorophores and (b) N-terminal amino acid recognition using a plurality of probes. (c) Neutral
1103 particle mass spectrometry is a promising technique to characterize proteoforms. Currently,
1104 the technology can be used to characterize large megadalton-scale complexes using Si-based
1105 nanosensors. Graphene-nanosensors and further developments may push the technology
1106 towards smaller and smaller proteins and potentially lead to increased sequence coverage in
1107 global proteomics. Electrospray Ionization (ESI) (d) Nanopore electrospray is a marriage of

1108 nanopores, classical electrospray, and single-particle detection techniques to sequence single
1109 proteins by measuring the amino acids one at a time.

1110

1111 **Figure 3: DNA facilitated protein sequencing.** (a) Schematic of specific amino acid labelling
1112 on a denatured protein with DNA strands. Each DNA strands contains both a barcode for the
1113 specific amino acid, and (optionally) a unique molecular identifier (UMI). (b-e) Illustration of
1114 various readout strategies of DNA-labelled samples, for protein identification. (b) Protein
1115 kinetic fingerprinting using quantitative DNA-PAINT. (c) Protein linear barcoding using
1116 molecular-resolution DNA-PAINT. (d) DNA Proximity Recording. (e) Protein structural
1117 fingerprinting using DNA-FRET-PAINT.

1118

1119 **Figure 4: Three strategies of Nanopore-based protein sequencing and sensing.** In all
1120 cases, an electrical force is used to translocate either a linearized or a folded protein through
1121 a nanoscale aperture (red arrow). (a) Reading unlabeled proteins or peptides using a
1122 biological nanopore. (b) Identification of whole proteins and peptides by fingerprinting with
1123 deep learning algorithms. Residue-specific fluorescent labels (e.g. at K, C, M) can be used to
1124 fingerprint proteins and peptides alongside electrical current sensing. (c) Identification of
1125 folded proteins using lipid tethering. Other tethers might include DNA carriers, DNA origami
1126 anchors, or plasmonic trapping.

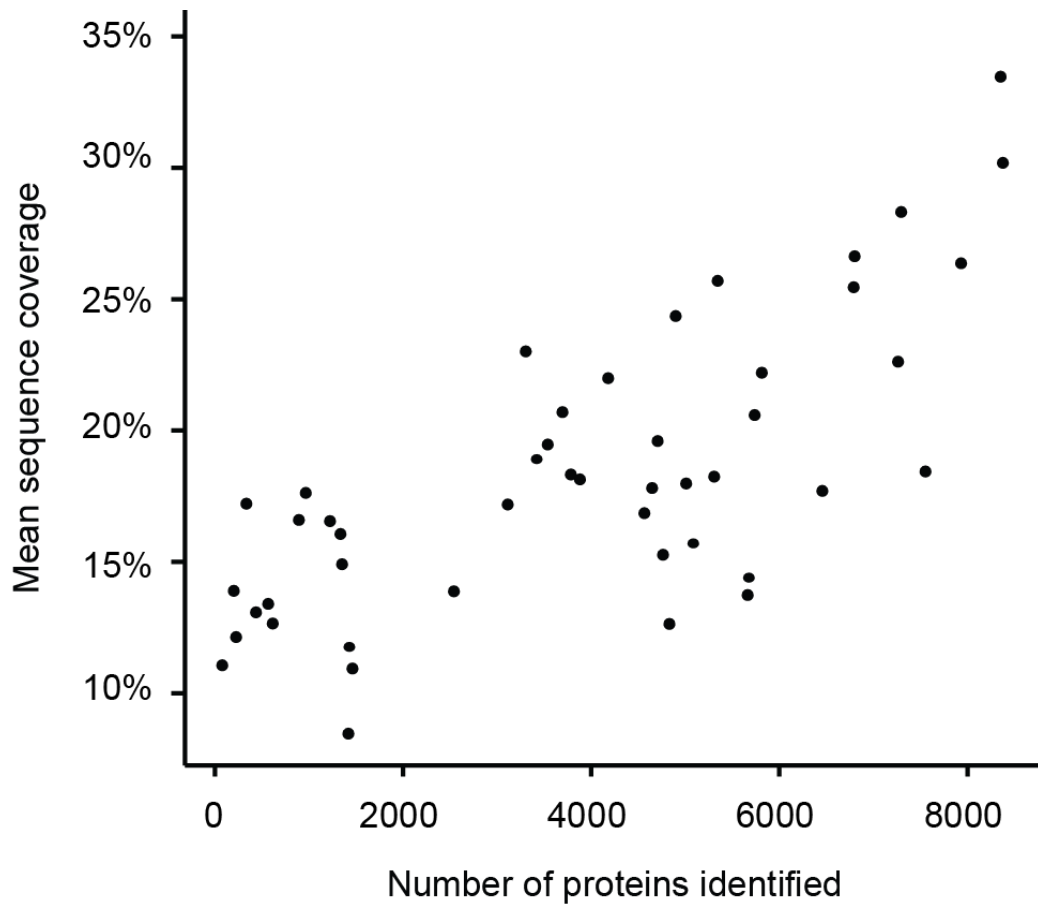
1127

1128 **Figure 5: Chemistry for protein sequencing.** (a) Lysine labeling with NHS esters (b)
1129 Cysteine labeling with iodoacetamide reactive groups (c) Strategies for labeling the phenol
1130 ring of tyrosine (d) Aspartate/Glutamate labeling (e) Tryptophan Labeling with sulfenyl
1131 chlorides. (f) C-terminal derivatization through Monoalkylation of A chain (41%).

1132

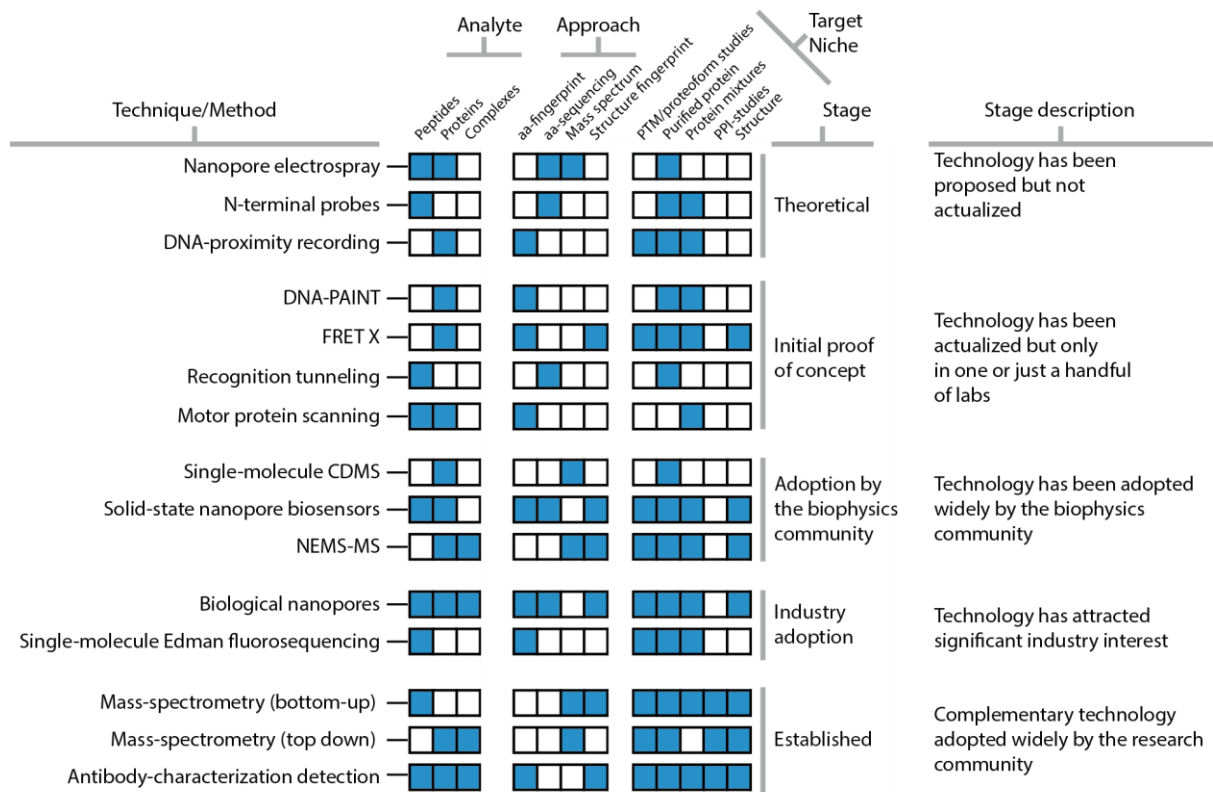
1133

1134 **Figure Box 1: Sequence coverage in global proteomics studies with MS**



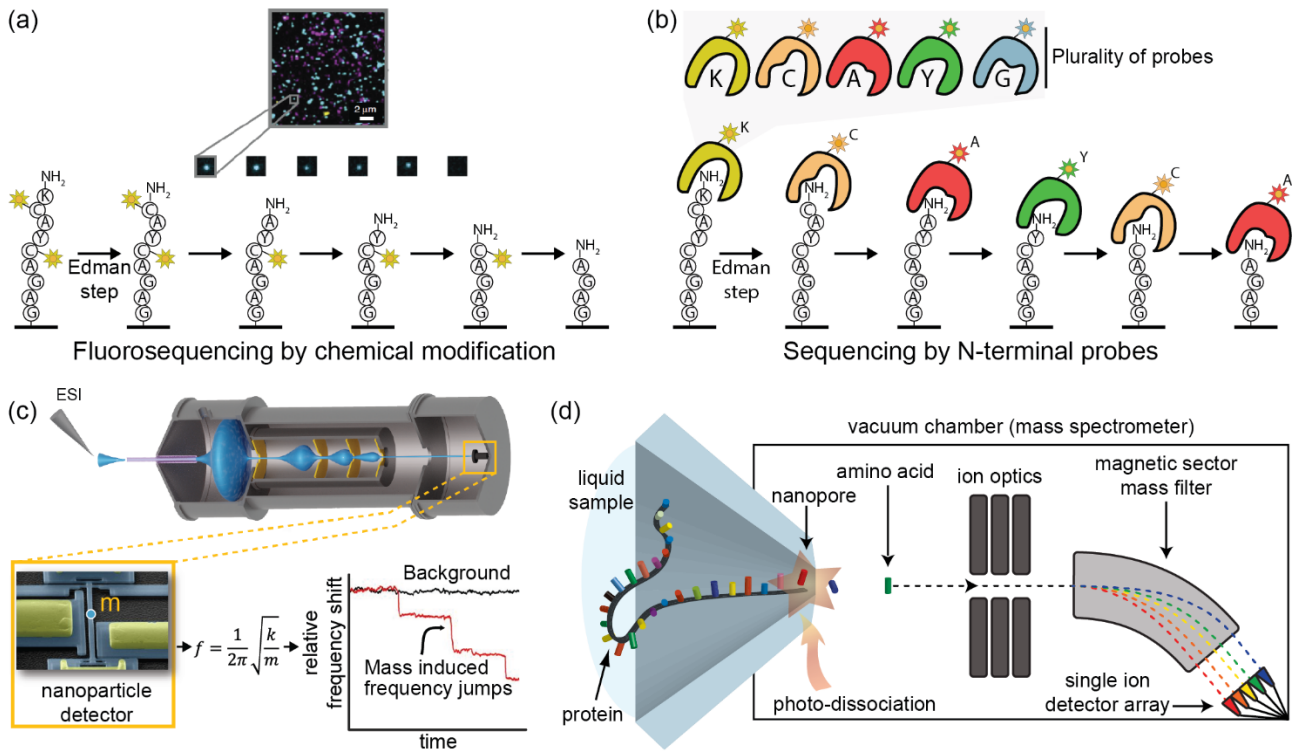
1135

1136 **Figure 1: The emerging landscape of single-molecule protein sequencing and**
 1137 **fingerprinting technologies.**



1138

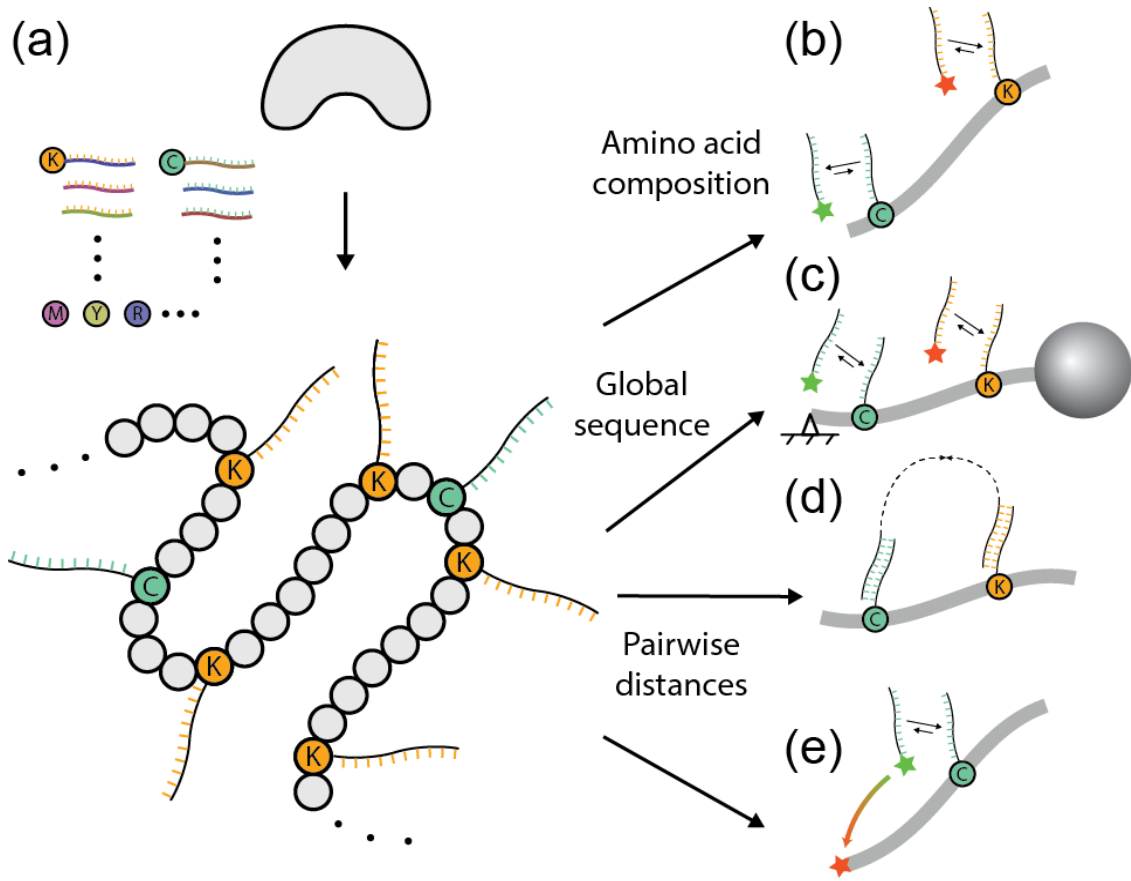
1139 **Figure 2: The renaissance of classical techniques**



1140 Neutral particle mass spectrometry with NEMS

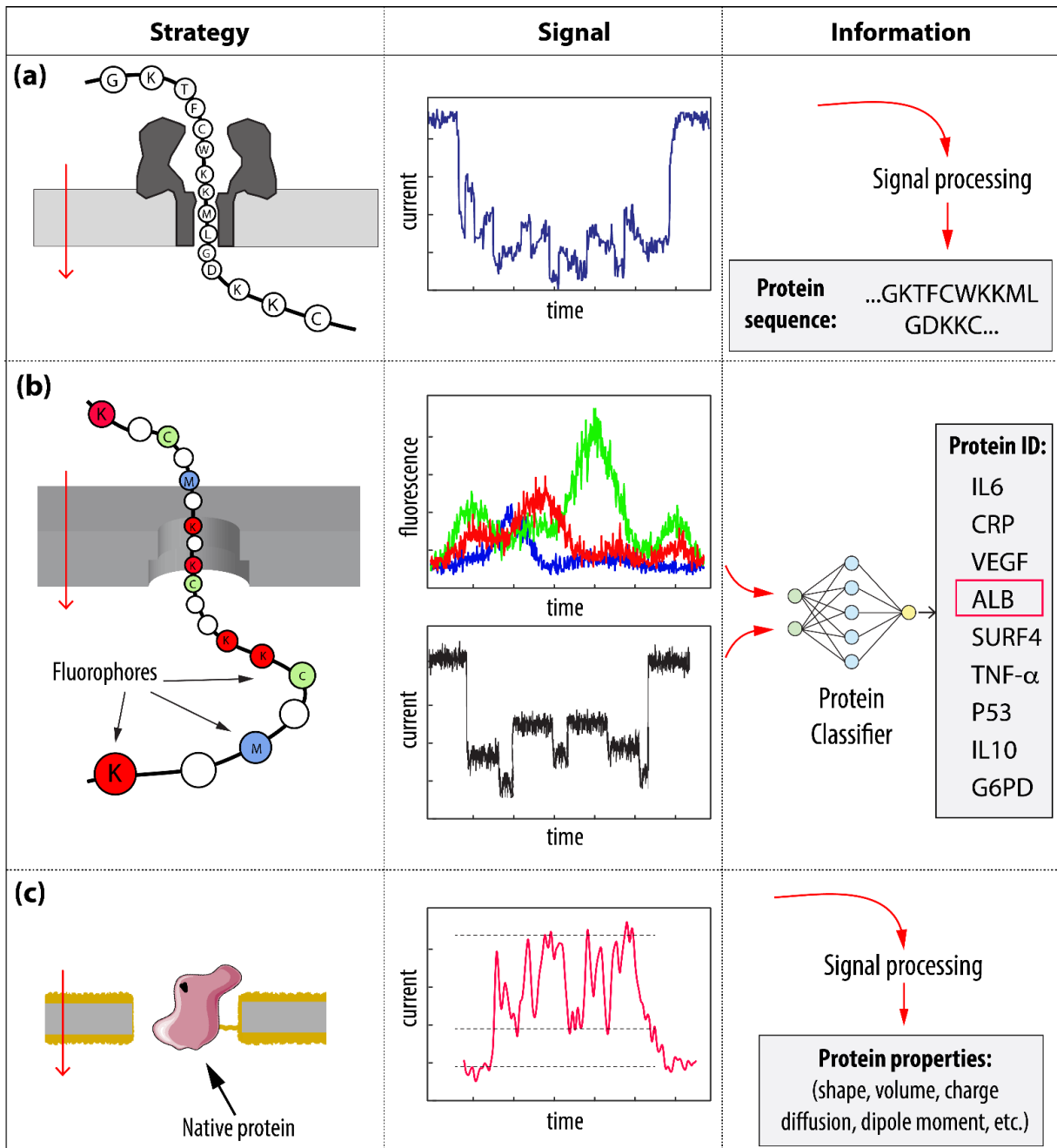
Nanopore electrospray

1141 **Figure 3: DNA-facilitated protein sequencing**

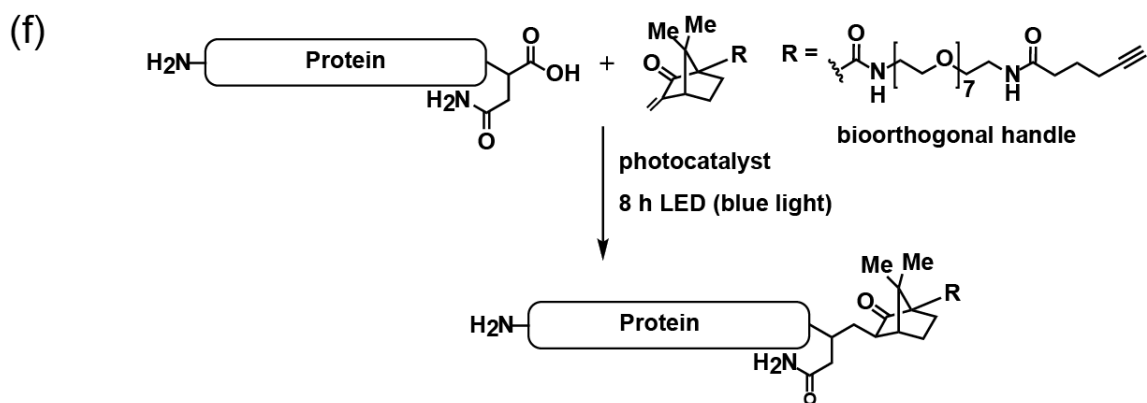
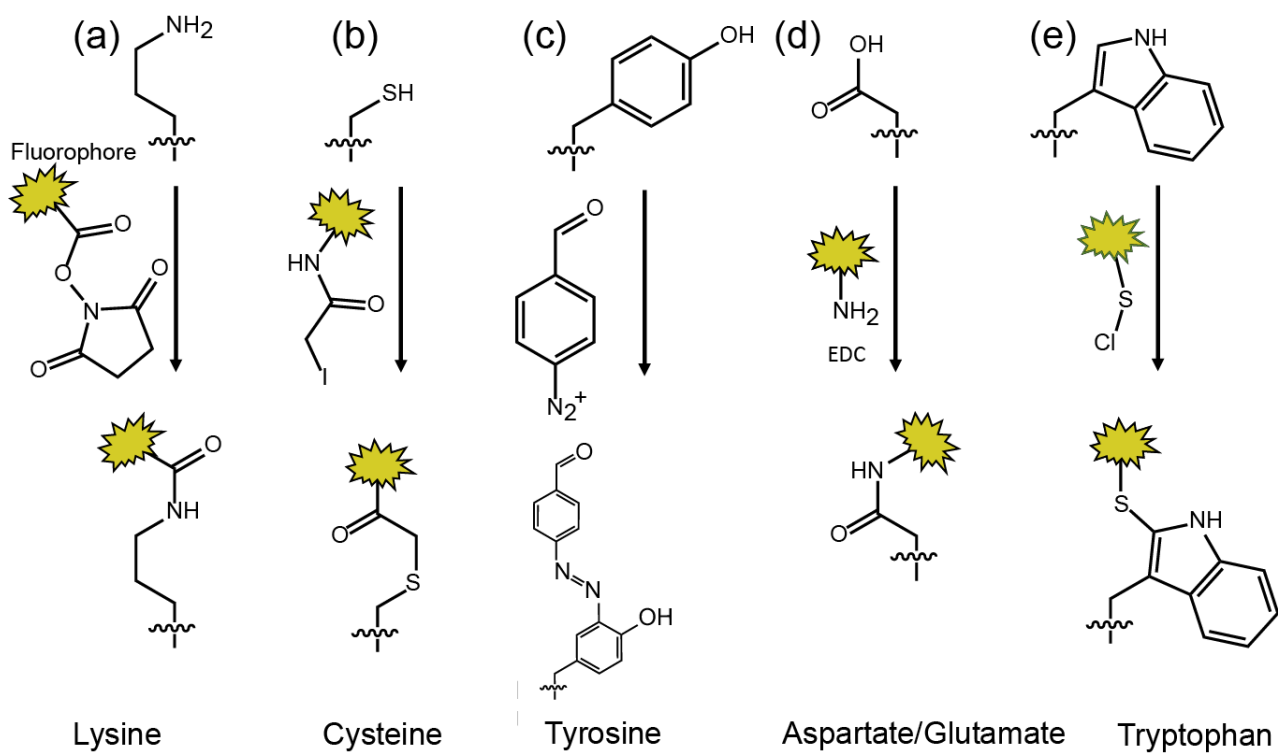


1142

1143 **Figure 4: Nanopore-based protein sequencing**



1144 **Figure 5: Chemistry for protein sequencing**



1145