Lab on a Chip

PAPER

View Article Online
View Journal | View Issue

Cite this: *Lab Chip*, 2021, **21**, 2922

Machine learning-aided protein identification from multidimensional signatures†

Yuewen Zhang ,‡§[a] Maya A. Wright,§[a] Kadi L. Saar,§[ab] Pavankumar Challa,§[a] Alexey S. Morgunov,§[ac] Quentin A. E. Peter,[a] Sean Devenish,[c] Christopher M. Dobson¶[a] and Tuomas P. J. Knowles *[ab]

The ability to determine the identity of specific proteins is a critical challenge in many areas of cellular and molecular biology, and in medical diagnostics. Here, we present a macine learning aided microfluidic protein characterisation strategy that within a few minutes generates a three-dimensional fingerprint of a protein sample indicative of its amino acid composition and size and, thereby, creates a unique signature for the protein. By acquiring such multidimensional fingerprints for a set of ten proteins and using machine learning approaches to classify the fingerprints, we demonstrate that this strategy allows proteins to be classified at a high accuracy, even though classification using a single dimension is not possible. Moreover, we show that the acquired fingerprints correlate with the amino acid content of the samples, which makes it is possible to identify proteins directly from their sequence without requiring any prior knowledge about the fingerprints. These findings suggest that such a multidimensional profiling strategy can lead to the development of a novel method for protein identification in a microfluidic format.

Received 13th November 2020,
Accepted 16th March 2021

DOI: 10.1039/d0lc01148g

rsc.li/loc

The diverse nature of proteins and their central role in a multitude of biological processes[1–3] necessitates a requirement for highly specific and sensitive approaches for protein detection and analysis. Indeed, protein detection and characterisation approaches have been of fundamental importance for a range of biological and medical research fields and have provided valuable information for better understanding the onset of a multitude of diseases, including various forms of cancer and neurodegenerative disorders.[4–10] In particular, at the centre of the discovery of novel protein-based disease biomarkers lies the ability to identify proteins.[11–16] In this context, protein microarrays are currently one of the most widely used techniques. By providing a high spatial density array of solid-phase supported affinity reagents, such as antibodies, protein microarrays allow proteins of interest to be selectively captured and subsequently detected through the introduction of a second, frequently fluorescently labelled affinity reagent.[17,18] As such, protein microarray based approaches usually require access to two distinct antibodies each targeting a different epitope of a single protein and, similarly to any other affinity reagent mediated system, their performance is sensitive to potential undesired cross-reactivity events. On a fundamental level, such an affinity-reagent mediated strategy is inherently limited to detecting known targets for which a suitable affinity reagent was consciously included in the library and does not allow for the detection and the discovery of hitherto unknown markers.

The possibility to detect the presence of hitherto unknown targets and perform explorative screening arises when affinity-reagent free protein analysis approaches are used. In this context, various forms of mass-spectrometry have been widely used for protein identification for many decades due to their high sensitivity, resolution, accuracy and dynamic range.[19,20] In a typical experiment, fragments of proteins are formed and separated through approaches, such as liquid chromatography before their injection to a mass-spectrometer.[21–24] While top-down identification has allowed characterising a number of different protein species, its application becomes challenging in the limit of high molecular weight and low solubility species. Due to these limitations, less than 10% of mammalian proteome can be accessed through these techniques.[25] For the analysis of higher molecular weight species, bottom-up sequencing approaches have been developed, which usually involve

[a] *Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. E-mail: tpjk2@cam.ac.uk*
[b] *Cavendish Laboratory, Department of Physics, University of Cambridge, J J Thomson Ave, Cambridge CB3 0HE, UK*
[c] *Fluidic Analytics Ltd., Cambridge, UK*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0lc01148g

‡ Current address: Department of Chemistry, Lanzhou University, Lanzhou, Gansu, 730000, P. R. China.
§ These authors contributed equally.
¶ Deceased September 8, 2019.

proteolysis of a complex mixture of proteins followed by a chromatographic separation of the peptides prior to their sequencing through tandem mass spectrometry (MS/MS). Whether the analysis is performed in a top-down or bottom-up manner, mass-spectrometry generally requires extensive sample preparation steps, often resulting in significant losses, and long experimental analysis time. Moreover, the presence of less abundant species is usually masked by more abundant ones, which prevents it effective use for detecting targets that are present at low concentrations, as is the case for biomarkers during the onset and early stages of diseases. Last but not least, its operation in gas-phase, has made it challenging to extend the analysis to protein complexes that are held together through transient interactions.

Recently, different approaches that would enable overcoming some of the challenges encountered with mass spectrometry have been demonstrated and proposed. For instance, Swaminathan *et al.*[26] have demonstrated the possibility of immobilising peptides onto a glass slide and measuring their fluorescence through total internal reflection microscopy in consecutive cycles of Edman degradation after selectively labelling lysine and cystine residues. While demonstrating the first steps towards the feasibility of single molecule peptide fluorosequencing,[27] the approach involves a number of consecutive Edman steps, setting a limit on the speed at which the analysis can be performed.

To open up the possibility of minute-scale liquid-phase protein identification, here, we devised and demonstrated a microfluidic platform that permits the identification of protein samples on a single device by relying on obtaining its characteristic multidimensional physicochemical signature. Specifically, by using a multi-wavelength detection system, we obtained readouts describing the tryptophan (Trp), tyrosine (Tyr) and lysine (Lys) content of the protein sample together with an estimate for their hydrodynamic radius (Fig. 1). By obtaining such multidimensional signatures for a total of ten proteins and using machine learning approaches for identifying the origin of a set of validation proteins, we showed that such a strategy can be used for identifying proteins with a high confidence. The characterisation and identification process is performed on unlabelled protein samples and on a minute timescale.

Materials and methods

Preparation of protein samples and the labelling solution

Bovine serum albumin (BSA), β-lactoglobulin (β-lac), glucose oxidase, α-lactalbumin (α-lac), ovalbumin, human transferrin, thyroglobulin (thglb), β-casein and ubiquitin (ubiq) were obtained from Sigma-Aldrich, and alcohol dehydrogenase (alc. dehydr) from Alfa Aesar. All the proteins were dissolved in 25 mM phosphate buffer at pH 8.0 to a micromolar concentration range. Precise values for the concentrations used in each experiments are listed in ESI† Table S1. The solution used for labelling the lysine residues (Fig. 2a) included 12 mM *o*-phthaldialdehyde (OPA), 18 mM β-mercaptoethanol (BME) and 4% wt/vol sodium dodecyl sulfate (SDS) in 200 mM carbonate buffer at pH 10.5.

UV-LED microscope

The schematic of the optical layout is shown in Fig. 2b. The sample was excited using either a 280 nm LED (Thorlabs M280L3, UK) or a 365 nm LED (Thorlabs M365L2, UK) light source with a flip mirror used to switch between the two sources. The light from either of the LEDs was passed through an aspherical lens of focal length 20 mm to get a

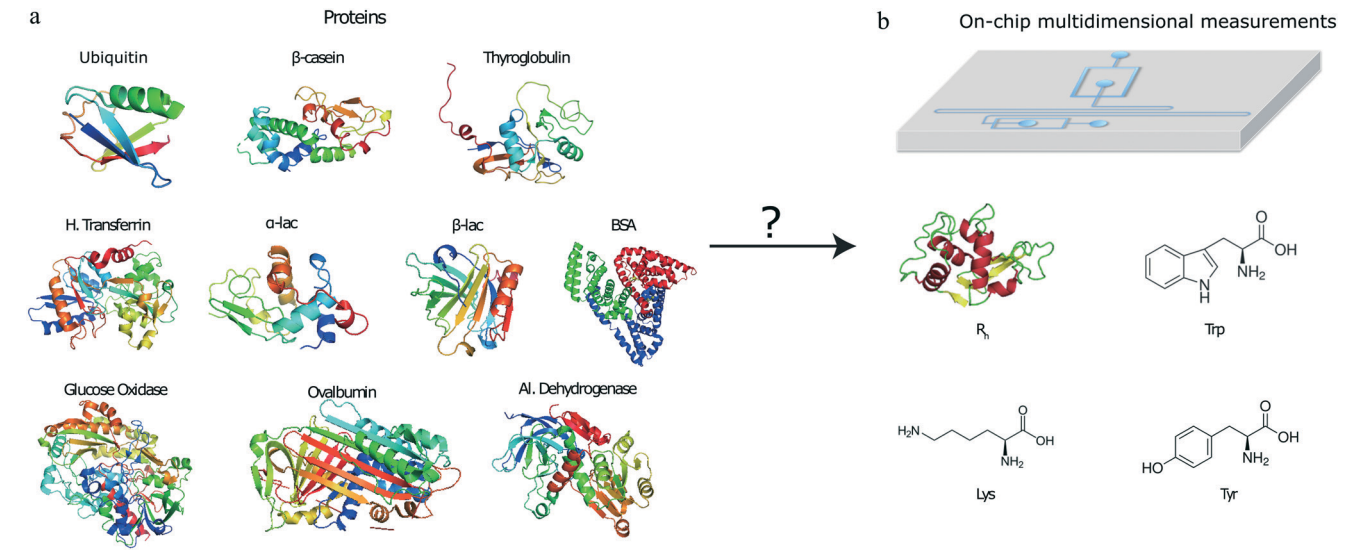

Fig. 1 Protein identification from multidimensional signatures on a microfluidic platform. (a) The proteins and (b) the microfluidic device used in this study. The device allows obtaining multi-dimensional fingerprints of protein samples that include information about their tryptophan, tyrosine and lysine content as well as the hydrodynamic radius, R_h.

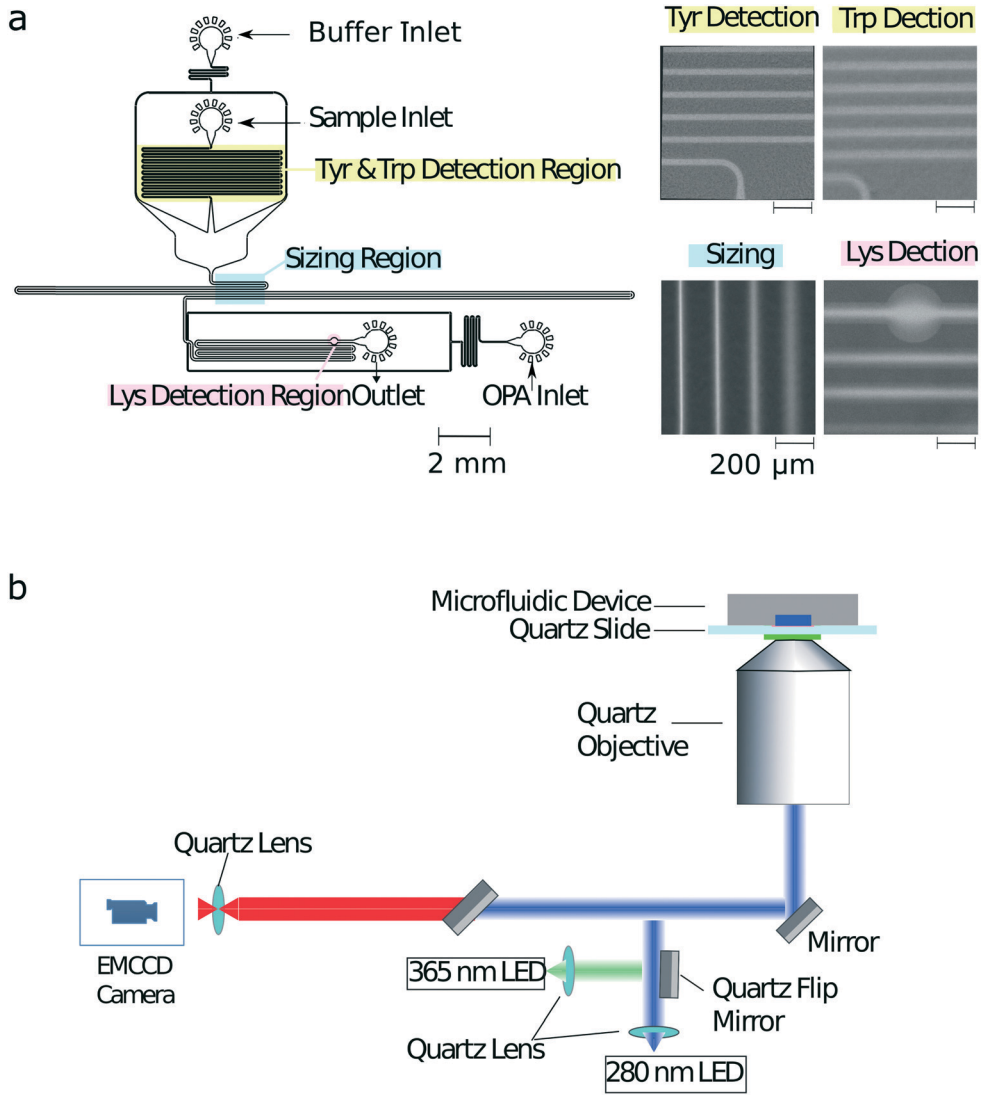

Fig. 2 The platform used for obtaining multidimensional signatures for proteins. (a) The microfluidic device used in this study allowed extracting a multidimensional characteristic signature of each analysed sample describing its tryptophan (Trp) and tyrosine (Tyr) residues (yellow highlighted region), its hydrodynamic radius R_h obtained by monitoring the diffusion of the sample molecules into a co-flowing buffer (blue highlighted region) and its lysine (Lys) content (pink highlighted region). The scale bars on all insets are 200 μm. (b) Schematic representation of the home-built inverted fluorescence microscope used. The two light sources (280 nm and 365 nm) and emission filters can be switched readily to record the characteristic fluorescent signals.

collimated output beam. The beam was passed through a dichroic filter cube, which consisted of an excitation filter (Semrock FF01-280/20-25) and a dichroic mirror (Semrock FF310-Di01-25 × 36). The light reflected by the dichroic mirror was then focussed onto the sample flowing in the microfluidic device by an infinity corrected UV objective lens (Thorlabs LMU-10X-UVB, UK) of numerical aperture NA = 0.25.The emitted fluorescent light from the sample was collected through the same objective and an emission filter (Semrock FF01-357/44-25 for a characteristic tryptophan, FF01-302/10-25 for a characteristic tyrosine and FF01-452/45-25 for a characteristic lysine signal) with an air-spaced achromatic doublet lens of focal length 20 mm (Thorlabs ACA254-200-UV) focussing it onto the camera (Rolera EMC2). All the optics used in the set-up were made out of fused silica for high transmission in the UV region.[28]

Microfluidic device fabrication

The microfluidic devices were cast using polydimethylsiloxane (PDMS) (Sylgard 184 kit, Dow Corning, USA) from a silicon wafer master imprinted with 50 μm high structures based using standard single layer soft-lithography techniques.[29] The precise height of the photoresist structures on different locations across the master mould were measured by a profilometer (DektakXT, UK) to correct the analysis for any variations in structure height across the master. Carbon black nanopowder (Sigma-Aldrich, UK) was

added to the PMDS to minimise undesired autofluorescence from the PDMS devices under UV illumination during the measurements. The devices were bonded to a quartz slide (Alfa Aesar, 76.2 × 25.4 × 1.0 mm, UK) using plasma treatment (Electronic Diener Femto plasma bonder; 15 seconds at 40% of the full power). The PDMS–glass microfluidic devices were then exposed to an additional extended plasma treatment step (500 seconds at 80% of the full power) to render channel surfaces more hydrophilic with the inlets and outlets blocked with water-filled gel-loading tips immediately after the exposure to maintain their hydrophilic character.

Device operation

To obtain a multidimensional signature for a sample, the channels of the microfluidic were first filled from the common outlet using a glass syringe (Hamilton, 500 μL, UK), 27 gauge needle (Neolus Terumo, 25 gauge, 0.5 × 16 mm, UK), and polyethene tubing (Scientific Laboratory Supplies, inner diameter 0.38 mm, outer diameter 1.09 mm, UK). Gel loading tips filled with the relevant solutions were then inserted into the device inlets (Fig. 2a). The fluid flow through of the solutions into the microfluidic channels was controlled using neMESYS syringe pumps (Cetoni GmbH, Germany) that was set to withdraw the solutions at a total flow rate of 200 μL h^{-1}. As described previously,[28] in order to increase the accuracy of the diffusional sizing process, the gel loading tip in the sample inlet was first filled with the auxiliary buffer and a background image of the diffusional sizing area recorded. This micrograph was later used for subtracting the static background arising from the autofluorescence of the PDMS device. The gel loading tip in the sample inlet was then carefully exchanged to a tip including the protein sample with care taken not to introduce any air bubbles in the process. For both images, an exposure time of 500 ms was used.

Finally, in order to account for any potential fluctuations in the power output of the LEDs, the intensities of standard calibration solutions (10 μM L-tryptophan and 10 μM 4-methylumbelliferone both in 400 mM potassium borate buffer at pH 9.7) were recorded in a channel adjacent to the identification device itself. The measured characteristic tryptophan and tyrosine fluorescence values were then normalised by the former of this calibration readings and the lysine value by the latter of the two calibration readings.

Results and discussion

Microfluidic multidimensional protein characterisation strategy

To facilitate the acquisition of multidimensional physicochemical signatures of proteins directly in solution, we designed a microfluidic device that allowed simultaneously obtaining four characteristic parameters of an unlabelled protein sample. Specifically, after introducing a sample from its corresponding inlet (Fig. 2a), first, the characteristic fluorescence intensities indicative of the tryptophan and tyrosine content of the sample were recorded in the yellow highlighted area by exciting the microfluidic device with a UV wavelength (280 nm) LED (Fig. 2b) and collecting the emitted fluorescent light using two distinct filters. The filters were chosen such that the collected light originated either predominantly from its tryptophan or from its tyrosine residues (Materials and methods).

The protein sample was then surrounded by a co-flowing buffer in order to monitor the lateral diffusion of the protein sample into an auxiliary carrier medium in space and in time. Such a strategy has been previously shown to yield the diffusion coefficients of protein samples.[30] In particular, the device we used in this study was designed for the camera field of view (800 μm × 1000 μm) to include four distinct sections of this channel (blue highlighted region), so that a single image could be used to extract the diffusion coefficient as described earlier.[31] The channels were imaged using the 280 nm excitation LED in combination with the tryptophan filter as the signal from latter residue was stronger than the signal from tyrosine residues. The diffusion profiles on the micrographs were then fitted to simulated basis functions for particles of known radii and each of the simulated profiles were compared to the measured profiles in order to extract the hydrodynamic radius of each sample.[30–33]

Finally, downstream the sizing unit, an on-chip latent labelling strategy was used to conjugate the lysine residues in each protein to *o*-phthaldialdehyde (OPA) dye molecules[32,34] (Materials and methods). The characteristic fluorescence intensity from the OPA labelled lysine residues was measured (pink highlighted region) by switching the UV-LED light source to an LED light source with excitation at 365 nm wavelength (Materials and methods) at which unconjugated OPA molecules have been observed to show only minimal background fluorescence. The dimensions of the labelling channel were chosen such that the OPA dye and the protein sample would be able to mix for over a 3 second long time period before the measurement was taken, a time scale that we had previously shown allows quantitative insight into the abundance of lysine residues in proteins.[32] The devices worked reliably with no major failure modes noted. To limit potential contamination between samples, each new measurement was performed in a separate device.

In summary, this strategy allowed us to obtain a four-dimensional signature for each protein sample using a single microfluidic platform and a dual-wavelength excitation system. One of the four measured parameters was later used for normalising the obtained fluorescent signals. This process ensured that the obtained signatures were independent of the sample concentration.

Multidimensional signatures of a set of ten proteins

We analysed a set of ten different proteins (Fig. 1a) and used the platform described above to obtain multidimensional signatures for each of them. In particular, we performed n =

4 repeats on all the ten proteins using a different microfluidic device for each experiment. We noted that the measured R_h values of all the proteins were consistent with the values reported in the literature (ESI† Table S1).

In order to eliminate concentration dependence, the measured signals in the tryptophan and tyrosine imaging channels were normalised by the signal in the lysine filter. This reduced the data structure to a three-dimensional signature but ensured that the obtained values were independent of the concentration of the protein that was used for analysis. Moreover, the measured intensities were corrected for fluctuations in the laser power by also measuring the fluorescence intensities of calibration solutions in a neighbouring channel, involving L-tryptophan and 4-methylumbelliferone molecules for the 280 nm and the 365 nm LED, respectively (Materials and methods).

The characteristic spaces that each of the analysed ten proteins occupied in a three-dimensional plot are shown in Fig. 3d with the 1D projections shown in Fig. 3a–c and the underlying data summarised in ESI† Table S1. In particular, the three-dimensional visualised ellipsoids (Fig. 3d) were defined by the centres being the average of the four measurement points and their radii corresponding to the standard deviation of the four measurements. We noted that the ten analysed proteins varied in their physiochemical signatures with Fig. 3d illustrating that it is likely that across a three-dimensional landscape each of the protein acquires a different signature.

We note that all the experiments were performed with protein concentration in the micromolar range (ESI† Table S1). Substantial decreases in these values would be possible. Specifically, we have previously demonstrated the possibility to detect proteins down to concentrations of around 100 nM using UV-fluorescent detection in microfluidic devices.[28] This sensitivity limit could be improved when an advanced detection mechanism (*e.g.* confocal-detection) or a higher laser powers is used, or if the device is fabricated from a material that shows a lower degree of autofluorescence[35] or in a manner where the autofluorescence from the microfluidic device would be suppressed.[36]

Protein classification

In order to evaluate whether our demonstrated platform is capable of distinguishing proteins reliably and uniquely based on their signatures, we developed two models to perform sample classification.

First, using the full data set of 10 classes of proteins with 4 experimental repeats for each class, leave-one-out cross-validation was used to assess the likelihood that a particular sample is classified as the correct protein. In particular, we argued that the errors in the measurements are likely to be Gaussian distributed and set out to use a multivariate Gaussian model for developing the classification algorithm. Specifically, multivariate Gaussian distributions were fitted to each of the ten protein classes with the means computed from the four repeats within each class, or from the three remaining repeats for the class from which the validation sample was removed. The covariance matrices were computed by combining the group variance (using either four or three repeats similarly to the means) with the global variance involving the full dataset of 39 data points excluding the validation sample. A weighting factor of 0.9 was used for the group variance and a weighting factor of 0.1 for the global variance to avoid singular covariance matrices and ensure computational stability while simultaneously taking advantage of the extra information about the system as the variances in the same dimension between the different classes are likely to be similar. Finally, the likelihood of each of the validation samples belonging to each of the protein classes was calculated by estimating the probability density function of the individual multivariate Gaussians at that point.

For each protein class, the likelihood was averaged across the four experimental repeats and the resulting values were normalised to one. Fig. 3e shows a heatmap of the calculated likelihoods for assigning proteins into available classes with the actual protein being measured on the vertical axis and the protein it is likely to be identified as on the horizontal axis. We observed that, individually, 33 out of 40 samples were classified correctly. Moreover, it can be seen that on average proteins are likely to be assigned to the correct class with high confidence.

The above estimates were arrived at by assuming that the errors in the measurements in each dimension were normally distributed, so that the protein classes can be approximated by multivariate Gaussian distributions. In order to improve our analysis and devise an analysis strategy that is not making an assumption about the distribution of the errors, we constructed a random forest classifier. As before, leave-one-out cross-validation was used on all 40 samples. In order to reduce variance, each random forest was trained with 1000 decision trees that were built using bootstrapping and with only 2 out of 3 variables selected at random to build each tree. The classification was performed using predictions by these ensemble models and, subsequently, predictions by all individual trees in the ensembles were collected to quantify the confidence of the ensemble model in making the predictions. For each group of four samples corresponding to the same protein class, the average number of trees in the ensemble predicting each target class were taken and normalised to sum to one for each protein.

Finally, a heatmap summarising the results was constructed, similarly showing the actual protein being measured on the vertical axis and the protein it is likely to be identified as on the horizontal axis (Fig. 3f). The results illustrate that the model predicts the correct class of proteins with high confidence. Moreover, on the individual level, the random forest model misclassified only 4 out of 40 samples, demonstrating a superior performance to the multivariate Gaussian model. This shows that highly accurate identification of proteins is possible even when no

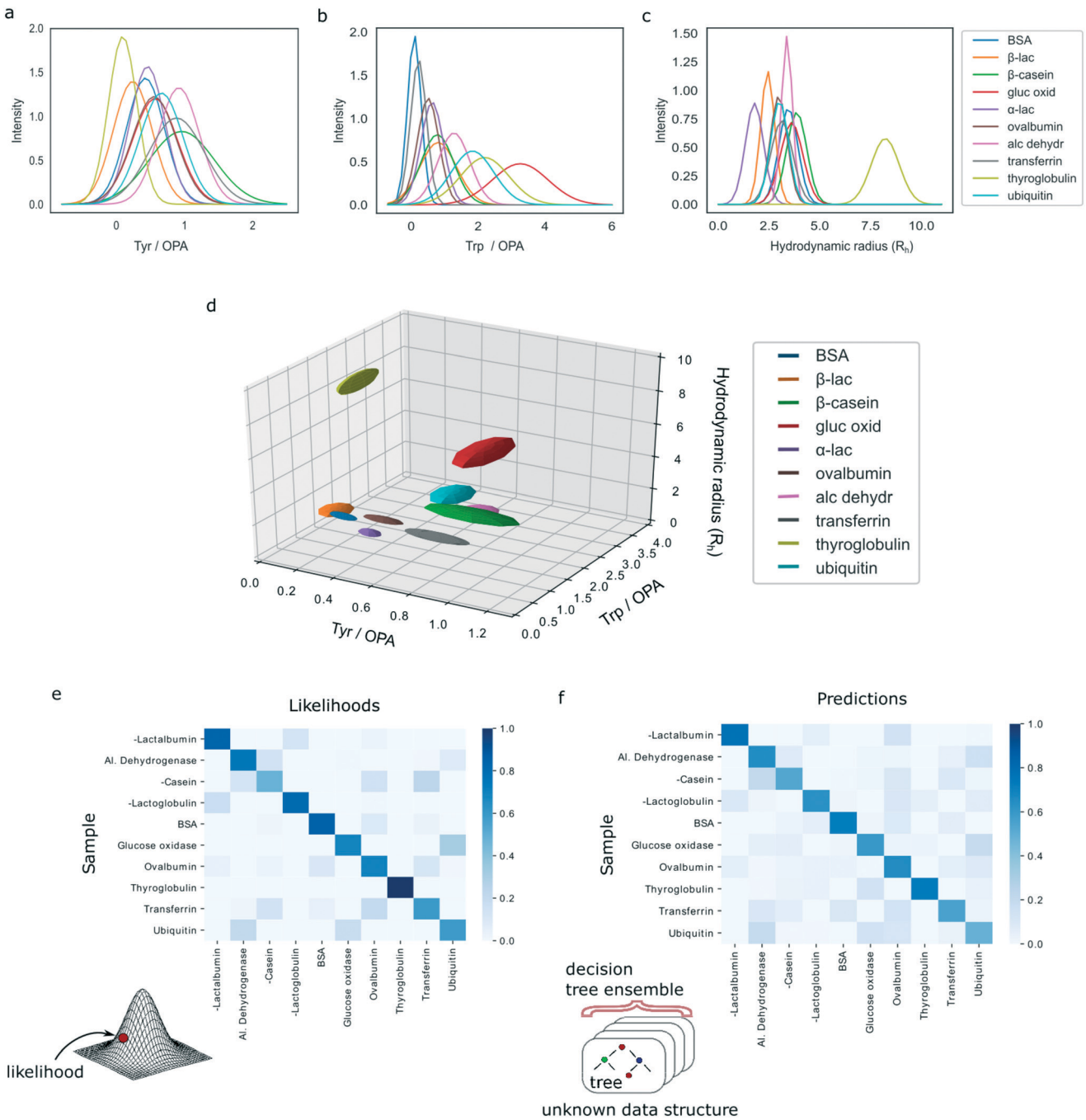

Fig. 3 Protein classification from their multidimensional fingerprints. A set of ten proteins was profiled by acquiring their three-dimensional fingerprints described by (a) the ratio of the signals measured at the wavelengths where tyrosine and OPA fluoresce (Materials and methods; dimension 1), (b) the ratio of the signals measured at the wavelengths where tryptophan and OPA fluoresce (dimension 2) and (c) the hydrodynamic radius, R_h (dimension 3). All these parameters are concentration independent. (d) Multidimensional signatures of the proteins in a 3D space. The radii of the ellipsoids correspond to one standard deviation. (e) The likelihoods of protein identification and misidentification in the 3D space showed in panel (d) assuming multivariate Gaussian model. (f) The confidence levels of identification process using a random forest classifier approach that assumes no underlying data distribution. The models identified correctly 82% (multivariate Gaussian) and 90% (random forest classifier) of the tested samples.

assumptions are made about the underlying distributions of measurement errors or data structure. We thank the referee for the suggestion to investigate the errors of the models in more detail. First, in the classification task only 4 misclassification events occurred, two of which corresponded to ubiquitin, one to β-casein and one to glucose oxidase. The remainder of the four samples for these proteins get identified correctly, which ensured that, on average, all proteins get identified correctly as seen in Fig. 3f. The misclassification of these events can be explained by their

relatively similar multidimensional signatures (Fig. 3d). The four misclassification events corresponded to two ubiquitin, one β-casein and one glucose oxidase sample, likely originating from the close similarity of their multidimensional signatures (Fig. 3d). The remainder of the four samples for these three proteins were identified correctly, which ensured that, on average, all proteins get identified correctly as seen in Fig. 3f.

Collectively, these results suggest that a model that makes no assumptions about the underlying data structure performs more accurately than the model that assumed that measurement noise was Gaussian distributed. Given the limited amount of training data for each test case (n = 39 samples), one possible strategy to improve the accuracy of the current model further is to train the model multiple times by each time sampling only some of the training points and create an ensemble classifier.[37] We employed this strategy and trained 100 classifiers each time randomly sampling 80% of the data. The models were then combined by setting the final prediction to be the mode of these 100 independently trained models. Using this strategy, we observed a small improvement to 37 proteins identified correctly. Even though the improvement in the current data set was small, more generally, such an approach provides a valuable strategy for reducing uncertainty in predictions and their sensitivity to outliers.[37]

Protein identification

Having confirmed the possibility to classify an unknown protein sample by evaluating which of the pre-determined multidimensional fingerprints it resembles the most (Fig. 3), we next set out to explore if it is possible to determine the origin of each of the test samples simply by performing an

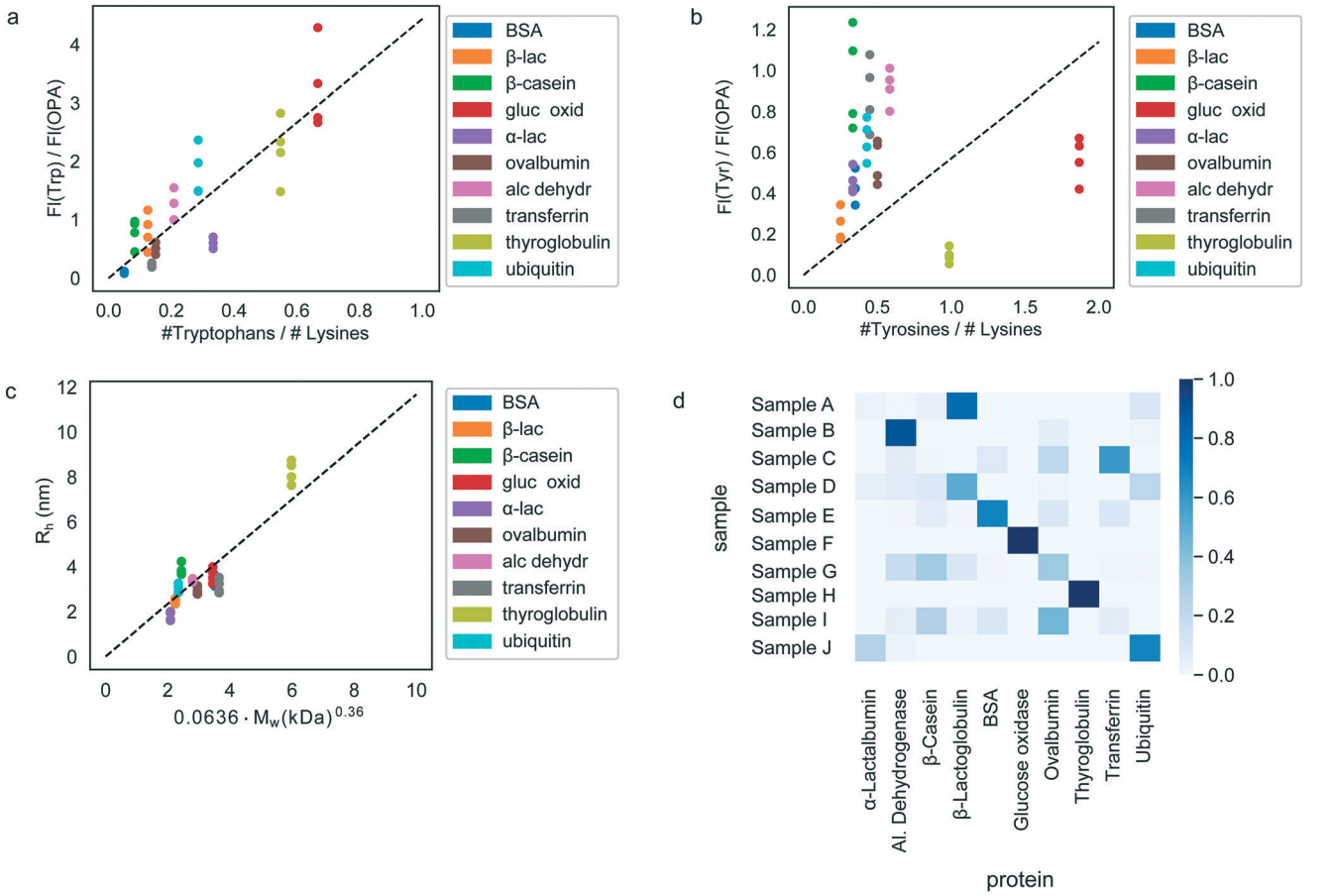

Fig. 4 Protein identification from their sequence. The correlations between the measured signals and the sequences of the analysed proteins, specifically (a) the ratio of the measured tryptophan and OPA signals as a function of the tryptophan and lysine composition of the proteins, (b) the ratio of the measured tyrosine and OPA fluorescence signals as a function of their tyrosine and lysine composition and (c) the measured hydrodynamic radius, R_h, as a function of the molecular weight. In all cases, the dotted line shows the best fit linear regression function with the intercept set to 0. We note that the fits shown here included all the proteins that were part of the study. The identification of an unseen protein was performed by excluding all the proteins of that particular sample, so that a slightly different fit was obtained each time. (d) The measured signals for each of the ten samples (A–J) were converted to estimates of their sequence-composition using the relationships outlined in panels (a)–(c) and the latter estimates were used to evaluate the probabilities of each of the ten samples being any one of the ten proteins in our dataset by using Gaussian mixture models. The data are shown such that the correct sample appears on the diagonal of the matrix. Individual samples were identified correctly on 21 out of 40 occasions. When averaging the results over n = 4 repeats, 7 out of 10 proteins were identified correctly.

identification process on the sample without requiring prior knowledge of the fingerprints.

To this effect, we first derived relationships that could be used to predict the sequence composition of each sample from its fingerprint. Indeed, the measured fluorescent signals were constructed in a manner where they can be expected to predominantly originate from the tyrosine, tryptophan and lysine residues of the proteins (Materials and methods) or, in the case of the hydrodynamic radius, be linked to its molecular weight. The observed correlation between the measured fluorescent signals and the amino acid content of the proteins are shown in Fig. 4a and b. As before, in order to eliminate concentration dependence, ratios between the measured signals were used. We note that these panels include all the 40 samples that were studied with all the points also used when estimating the best-fit line (dotted line). However, when the actual identification process was performed (next paragraph), each time, the test protein as well as all other proteins of the same type were excluded from the fit to ensure that there is no information leakage. Fig. 4c additionally outlines the relationship between the measured hydrodynamic radius and the molecular weight of the proteins. We modelled all the three relationships as linear regressions with zero-intercepts and estimated the gradient of the line by minimising the ordinary least squares (Fig. 4a–c, dotted lines). It is possible that when more abundant data is available, more nuanced relationships between the measured signals and the sequence-specific quantities can be learned. However, for the current analysis we used the prior of a linear relationship as the use of a model with a larger number of parameters may have resulted in overfitting.

Next, the derived relationships (Fig. 4a–c) were used to convert the measured three-dimensional signature of our test samples into their predicted $\frac{\text{Trp}}{\text{Lys}}$ and $\frac{\text{Tyr}}{\text{Lys}}$ ratios and molecular weights. To eliminate any information leakage, we re-fitted the linear regression after excluded all the measurements that involved the test sample, reducing the size of the data that was used for fitting down to 36 points. Following this step, the *z*-score of the measured sample being a particular protein in the database was calculated by using the estimated sequence-specific properties of the protein that the sample was assumed to be as the mean value and the measurement noise as the standard deviation when defining the clusters. The heatmap describing the probability of the test sample being a specific protein is shown in Fig. 4d with the data arranged such that the correct sample appeared on the diagonal of the matrix. These data show that the samples that corresponded to proteins that followed the linear relationship closely (*e.g.* sample F; ovalbumin) achieved high identification accuracy while samples corresponding to proteins that did not follow the relationship so well (*e.g.* sample C, β-casein) showed substantially poorer identification performance (Fig. 4d). Individually, 21 out of the 40 samples were identified correctly. Moreover, when the probability estimates from the four repeats performed on each protein were combined, 7 out the 10 proteins were identified correctly (Fig. 5). These results illustrate that not only can the multidimensional signatures used for classifying proteins into pre-determined clusters (Fig. 3e and f), it is also possible to convert the measured signals into absolute sequence-specific parameters and through this process identify the test samples.

Analysing the misidentification events in more detail, we noticed that two proteins, β-casein and transferrin were never identified correctly. This effect likely originated from these proteins being among the most significant outliers from the approximated linear relationships (Fig. 4a–c). While we had chosen to use these linear models for simplicity, it is clear that this strategy cannot capture the full nuance. The effect can be particularly pronounced for the UV-fluorescent signals, where not all tryptophan and tyrosine residues contribute to the emitted fluorescent signal equally but their contribution is defined by the local environment and expose to the solvent. As such, the relationships between the actual amino acid ratios and measured fluorescent signals are likely more complex functions also involving the protein fold and prior insight into such relationships is one of the parameters that would allow us to achieve an enhanced identification accuracy.

Discussion

The protein classification and identification platform developed here that relied on measuring multi-dimensional signatures for ten different protein samples indicated that

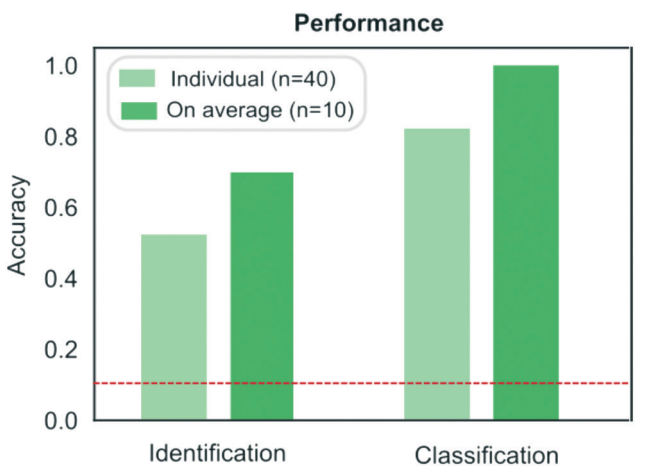

Fig. 5 Comparison of the performance of the protein classification (Fig. 3) and identification (Fig. 4) strategies. When identifying a measured sample directly from its sequence, samples were identified correctly on 53% of the occasions or on 70% of the occasions when the results were averaged across the four repeats performed on each samples. When pre-determined fingerprints were used, proteins were classified correctly on 83% of the occasions or on 100% of the occasion when the results were averaged across the repeats. The red dotted line corresponds to the case where the classification or identification was performed by a process of random guessing.

within our dataset, the platform had a high capability to both, classify an unseen protein sample from its measured signature and identify it directly from sequence (100% and 70%, respectively). Using our current data, we set out to explore the theoretical limitation of the platform. Specifically, by focussing the analysis on the proteins that were listed as expressed with immunohistochemistry level of evidence in at least one tissue in the Human Protein Atlas,[38,39] we first estimated that 99% of these proteins had their Tyr/Lys amino acid ratio between 0.01 and 3.50, Trp/Lys ratio between 0.01 and 2.40 and hydrodynamic radius (estimated from $R_h = 0.06358 \cdot M_w^{0.36}$, where M_w was the molecular weight of the proteins in kDa) between 1.71 and 5.97 nm. The standard deviation estimates for these parameters averaged across the 10 protein classes were 0.10, 0.27 and 0.23 nm. Thus, as a first order approximation, our platform with this level of measurement error would allow distinguishing between $\frac{3.49}{2 \cdot 0.10} \cdot \frac{2.39}{2 \cdot 0.27} \cdot \frac{4.26}{2 \cdot 0.23} = 715$ **proteins**, requiring the signatures of the samples to be separated by at least two times the standard deviation along the axes describing the three dimensions.

As in a representative practical example a random set of proteins does not lie in regularly spaced intervals, this estimate can be viewed as an upper bound for the resolution capacity. We therefore speculate that our demonstrated multidimensional profiling strategy proves the most useful when a handful of proteins are present. This is the case, for instance, when a complex mixture has been first purified on a 2D-gel that allows additional information about the protein to be obtained, such as its electrophoretic mobility. We further note that significant improvements on the currently demonstrated resolution capability are possible. This objective could be achieved either by ensuring that the accuracy of individual measurements is increased or by incorporating additional dimensions, such as the charge of the protein that could be accessed by an on-chip electrophoresis step.[40,41]

Conclusions

We developed a strategy for obtaining multidimensional physicochemical signatures of individual proteins on a single microfluidic platform and showed that this strategy can be used for protein classification as well identification. Specifically, we achieved this objective by designing a device on which the hydrodynamic radius of a protein sample could be obtained simultaneously with signals describing its tryptophan, tyrosine and lysine content. We showed that this approach generated unique fingerprints for all the ten proteins in our test set and, moreover, that the signatures can be used to identify proteins through their multidimensional signatures. Our results suggest that an on-chip multidimensional protein characterisation strategy could serve as a powerful probe-free approach for on-chip profiling of protein samples from microlitre sized volumes.

Author contributions

Y. Z., M. A. W., K. L. S., P. C. and T. P. J. K. designed the study and conceived the experiments, P. C. built the experimental hardware, Y. Z. and M. A. W. acquired the data and processed the micrographs for which Q. A. E. P. contributed relevant software. K. L. S. and A. S. M. analysed and interpreted the data and built the machine learning models. Y. Z., K. L. S., A. S. M. and T. P. J. K. wrote the original manuscript and M. A. W., P. C., S. D., Q. A. E. P. and C. M. D. reviewed it.

Conflicts of interest

Parts of this work have been the subject of a patent application filed by Cambridge Enterprise Limited, a fully owned subsidiary of the University of Cambridge.

Acknowledgements

This is a contribution from the Cambridge Centre for Misfolding Diseases. The research leading to these results has received funding from the ERC under the European Union Seventh Framework Programme (FP7/2007-2013) through the ERC grant PhysProt (agreement no 337969 T. P. J. K.), the EPSRC (K. L. S), the Schmidt Science Fellows program, in partnership with the Rhodes Trust (K. L. S), the BBSRC (C. M. D, T. P. J. K), and the Frances and Augustus Newman Foundation (T. P. J. K.).

Notes and references

1 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Mol. Biol. Cell*, 2002, 53–80.
2 J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*, 2002, 84–137.
3 B. Alberts, *Cell*, 1998, **92**, 291–294.
4 F. S. Collins, E. D. Green, A. E. Guttmacher and M. S. Guyer, *Nature*, 2003, **422**, 835.
5 F. Coscia, K. Watters, M. Curtis, M. Eckert, C. Chiang, S. Tyanova, A. Montag, R. Lastra, E. Lengyel and M. Mann, *Nat. Commun.*, 2016, **7**, 12645.
6 F. Genovese, A. Gualandi, L. Taddia, G. Marverti, S. Pirondi, C. Marraccini, P. Perco, M. Pelà, R. Guerrini, M. R. Amoroso, F. Esposito, A. Martello, G. Ponterini, D. D'Arca and M. P. Costi, *J. Proteome Res.*, 2014, **13**, 5250–5261.
7 S. Kang, H. Maeng, B. G. Kim, G. M. Qing, Y. P. Choi, H. Y. Kim, P. S. Kim, Y. Kim, Y. H. Kim, Y. D. Choi and N. H. Cho, *J. Proteome Res.*, 2012, **11**, 4567–4574.
8 Z. Oláh, J. Kálmán, M. E. Tóth, Á. Zvara, M. Sántha, E. Ivitz, Z. Janka and M. Pákáski, *J. Alzheimer's Dis.*, 2015, **44**, 1303–1312.
9 M. Puchades, S. F. Hansson, C. L. Nilsson, N. Andreasen, K. Blennow and P. Davidsson, *Mol. Brain Res.*, 2003, **118**, 140–146.
10 R. Sultana, D. Boyd-Kimball, J. Cai, W. M. Pierce, J. B. Klein, M. Merchant and D. A. Butterfield, *J. Alzheimer's Dis.*, 2007, **11**, 153–164.
11 J. P. Savaryn, A. D. Catherman, P. M. Thomas, M. M. Abecassis and N. L. Kelleher, *Genome Med.*, 2013, **5**, 53.

12 R. Chen and M. Snyder, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2013, **5**, 73–82.
13 E. Petricoin, J. Wulfkuhle, V. Espina and L. A. Liotta, *J. Proteome Res.*, 2004, **3**, 209–217.
14 H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein and M. Snyder, *Science*, 2001, **293**, 2101–2105.
15 L. A. Liotta, E. C. Kohn and E. F. Petricoin, *JAMA, J. Am. Med. Assoc.*, 2001, **286**, 2211–2214.
16 H. Zhu, J. F. Klemic, S. Chang, P. Bertone, A. Casamayor, K. G. Klemic, D. Smith, M. Gerstein, M. A. Reed and M. Snyder, *Nat. Genet.*, 2000, **26**, 283.
17 G. MacBeath, *Nat. Genet.*, 2002, **32**, 526.
18 J. LaBaer and N. Ramachandran, *Curr. Opin. Chem. Biol.*, 2005, **9**, 14–19.
19 A. G. Marshall, C. L. Hendrickson and G. S. Jackson, *Mass Spectrom. Rev.*, 1998, **17**, 1–35.
20 G. A. Valaskovic, N. L. Kelleher and F. W. McLafferty, *Science*, 1996, **273**, 1199–1202.
21 M. Mann, R. C. Hendrickson and A. Pandey, *Annu. Rev. Biochem.*, 2001, **70**, 437–473.
22 A. M. Frank, M. M. Savitski, M. L. Nielsen, R. A. Zubarev and P. A. Pevzner, *J. Proteome Res.*, 2007, **6**, 114–123.
23 K. Vyatkina, S. Wu, L. J. Dekker, M. M. VanDuijn, X. Liu, N. Tolic, M. Dvorkin, S. Alexandrova, T. M. Luider, L. Pasa-Tolic and P. A. Pevzner, *J. Proteome Res.*, 2015, **14**, 4450–4462.
24 G. E. Reid and S. A. McLuckey, *J. Mass Spectrom.*, 2002, **37**, 663–675.
25 K. A. Resing and N. G. Ahn, *FEBS Lett.*, 2005, **579**, 885–889.
26 J. Swaminathan, A. A. Boulgakov, E. T. Hernandez, A. M. Bardo, J. L. Bachman, J. Marotta, A. M. Johnson, E. V. Anslyn and E. M. Marcotte, *Nat. Biotechnol.*, 2018, **36**, 1076.
27 B. C. Collins and R. Aebersold, *Nat. Biotechnol.*, 2018, **36**, 1051.
28 P. K. Challa, Q. Peter, M. A. Wright, Y. Zhang, K. L. Saar, J. A. Carozza, J. L. Benesch and T. P. Knowles, *Anal. Chem.*, 2018, **90**, 3849–3855.
29 D. C. Duffy, J. C. McDonald, O. J. Schueller and G. M. Whitesides, *Anal. Chem.*, 1998, **70**, 4974–4984.
30 P. Arosio, T. Müller, L. Rajah, E. V. Yates, F. A. Aprile, Y. Zhang, S. I. Cohen, D. A. White, T. W. Herling, E. J. De Genst, S. Linse, M. Vendruscolo, C. M. Dobson and T. P. J. Knowles, *ACS Nano*, 2015, **10**, 333–341.
31 K. L. Saar, Q. Peter, T. Müller, P. K. Challa, T. W. Herling and T. P. Knowles, *Microsyst. Nanoeng.*, 2019, **5**, 33.
32 E. V. Yates, T. Müller, L. Rajah, E. J. De Genst, P. Arosio, S. Linse, M. Vendruscolo, C. M. Dobson and T. P. Knowles, *Nat. Chem.*, 2015, **7**, 802.
33 Y. Zhang, T. W. Herling, S. Kreida, Q. A. Peter, T. Kartanas, S. Törnroth-Horsefield, S. Linse and T. P. Knowles, *Lab Chip*, 2020, **20**, 3230–3238.
34 Y. Zhang, E. V. Yates, L. Hong, K. L. Saar, G. Meisl, C. M. Dobson and T. P. Knowles, *Chem. Sci.*, 2018, **9**, 3503–3507.
35 A. Piruska, I. Nikcevic, S. H. Lee, C. Ahn, W. R. Heineman, P. A. Limbach and C. J. Seliskar, *Lab Chip*, 2005, **5**, 1348–1354.
36 K. L. Saar, T. Muller, J. Charmet, P. K. Challa and T. P. Knowles, *Anal. Chem.*, 2018, **90**, 8998–9005.
37 J. Ko, S. N. Baldassano, P.-L. Loh, K. Kording, B. Litt and D. Issadore, *Lab Chip*, 2018, **18**, 395–405.
38 M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester and S. Hober, *et al.*, *Nat. Biotechnol.*, 2010, **28**, 1248.
39 M. Uhlen, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt and A. Asplund, *et al.*, *Science*, 2015, **347**, 1260419.
40 T. W. Herling, D. J. O'Connell, M. C. Baeuer, J. Persson, U. Weininger, T. P. Knowles and S. Linse, *Biophys. J.*, 2016, **110**, 1957–1966.
41 K. L. Saar, Y. Zhang, T. Müller, C. P. Kumar, S. Devenish, A. Lynn, U. Łapińska, X. Yang, S. Linse and T. P. Knowles, *Lab Chip*, 2018, **18**, 162–170.