Pilot evaluation of long-term reproducibility of capillary zone electrophoresis-tandem mass spectrometry for top-down proteomics of a complex proteome sample

Seyed Amirhossein Sadeghi,[1] Wenrong Chen,[2] Qianyi Wang,[1] Qianjie Wang,[1] Fei Fang,[1] Xiaowen Liu,[3] Liangliang Sun[1,*]

[1]Department of Chemistry, Michigan State University, 578 S Shaw Lane, East Lansing, Michigan 48824, USA

[2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 535 W Michigan Street, Indianapolis, IN 46202, USA

[3]Deming Department of Medicine, School of Medicine, Tulane University, 1441 Canal Street, New Orleans, LA 70112, USA

* Corresponding Author. Email: lsun@chemistry.msu.edu

Phone: 517-353-0498

Abstract

Mass spectrometry (MS)-based top-down proteomics (TDP) has revolutionized biological research by measuring intact proteoforms in cells, tissues, and biofluids. Capillary zone electrophoresis-tandem MS (CZE-MS/MS) is a valuable technique for TDP, offering high peak capacity and sensitivity for proteoform separation and detection. However, the long-term reproducibility of CZE-MS/MS in TDP remains unstudied, a crucial aspect for large-scale studies. This work investigated the long-term qualitative and quantitative reproducibility of CZE-MS/MS for TDP for the first time, focusing on a yeast cell lysate. Over 1000 proteoforms were identified per run across 62 runs using one linear polyacrylamide (LPA)-coated separation capillary, highlighting the robustness of the CZE-MS/MS technique. However, substantial decreases in proteoform intensity and identification were observed after some initial runs due to proteoform adsorption onto the capillary inner wall. To address this issue, we developed an efficient capillary cleanup procedure using diluted ammonium hydroxide, achieving high qualitative and quantitative reproducibility for the yeast sample across at least 23 runs. The data underscores the capability of CZE-MS/MS for large-scale quantitative TDP of complex samples, signaling its readiness for deployment in broad biological applications. The MS RAW files were deposited in ProteomeXchange Consortium with the dataset identifier of PXD046651.

KEYWORDS: top-down proteomics, capillary zone electrophoresis, mass spectrometry, proteoform, reproducibility, label-free quantification, yeast cell lysate

Introduction

Mass spectrometry (MS)-based top-down proteomics (TDP) is a powerful technique for the identification and quantification of proteoforms in biological samples[1]. During the last several years, TDP has been deployed widely to discover new proteoform biomarkers of various diseases, e.g., cancer[2-5], neurodegeneration[6-9], cardiovascular diseases[10], infectious disease[11-14], and immunobiology[15]. MS-based TDP is providing more and more new insights into the functions of proteins in modulating the cellular processes.

Due to the high complexity of the proteoforms in cells or tissues, high peak capacity separations of proteoforms before MS is crucial. Liquid chromatography (LC)-MS has been the widely used technique for TDP of complex samples[16,17]. Capillary zone electrophoresis (CZE) offers highly efficient separations of biomolecules according to electrophoretic mobility (μ_{ef}), which relates to their charge-to-size ratios[18]. CZE-MS has also been well recognized as an alternative technique to LC-MS for global TDP profiling of proteoforms in cells and tissues due to its high efficiency and sensitivity for proteoform separation and detection as well as its unique opportunity for accurate prediction of proteoform's μ_{ef}[19-21]. Several research groups have shown the early examples of CZE-MS for highly sensitive and global TDP of complex biological samples[22-25]. Our group has shown the identification of hundreds to thousands of proteoforms from complex samples by single-shot CZE-MS measurements via innovations in capillary coating, online proteoform stacking, and etc[19,20,26]. We further boosted the number of identified proteoforms from human cell lines to over 23,000 by coupling LC fractionation to CZE-MS[3]. Most recently, we developed online two dimensional FAIMS (high-field asymmetric waveform ion mobility spectrometry)-CZE-MS to benefit the identification of large proteoforms[27] and histone proteoforms[28]. We also showed the capability of CZE-MS for TDP of membrane proteins[29]. The Kelleher group documented the high sensitivity of CZE-MS for TDP and the reasonable complementarity between CZE-MS and LC-MS for proteoform identification[30]. The Ivanov group illustrated the potential of CZE-MS for TDP of single mammalian cells[31].

CZE-MS has made drastic progress in TDP and has been widely accepted as a useful tool for proteoform characterization. However, to use CZE-MS for large-scale TDP studies, we need to validate its long-term reproducibility for top-down MS measurement of complex samples. In this work, for the first time, we performed a pilot investigation of long-term reproducibility of CZE-MS for TDP of a complex sample (i.e., a yeast cell lysate) to achieve a better understanding of advantages, issues, and potential solutions of CZE-MS for large-scale TDP.

EXPERIMENTAL SECTION

Chemicals and materials

Ammonium bicarbonate (ABC), ammonium hydroxide (NH_4OH), 3-(trimethoxysilyl) propyl methacrylate, and Amicon Ultra (0.5 mL, 10 kDa cut-off size) centrifugal filter units and (Yeast Extract–Peptone–Dextrose) YPD Broth were ordered from Sigma-Aldrich (St. Louis, MO). LC/ MS grade water, acetonitrile (ACN), HPLC-grade acetic acid (AA), fused silica capillaries (50 mm i.d., 360 mm o.d., Polymicro Technologies) were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide was obtained from Acros Organics (Fair Lawn, NJ). Complete, mini protease inhibitor cocktail (provided in EASYpacks) was bought from Roche (Indianapolis, IN).

Sample preparation

Yeast growth in (Yeast Extract–Peptone–Dextrose) YPD Broth is meticulously cultivated using a well-defined procedure. To begin, 50 g of YPD Broth was blended with 1 liter of distilled water, ensuring a precise mixture. This suspension underwent autoclaving at 121°C for a duration of 15 minutes. Following this, yeast cultures are introduced into detergent-free containers. A brief vortexing was then carried out to uniformly disperse the yeast cells throughout the medium. The yeast cultures were subsequently nurtured in a shaking incubator at 300 rpm.

After yeast cell collection and cleanup with a PBS buffer, five gram of yeast cells were suspended into the lysis buffer containing 8 M urea, complete protease inhibitors and PhosSTOP (Roche), and 100 mM ammonium bicarbonate (pH 8.0), followed by

incubation on ice for 30 min with periodical vortexing. The cells were lysed for 3 min using a homogenizer (Fisher Scientific) and then sonicated under 50% duty cycle, level 10 output for 20 min on ice with a Branson Sonifier 250 (VWR Scientific). The yeast lysate was centrifuged at 14,000 g for 10 minutes at 4 °C to collect the supernatant containing extracted proteins. The concentration of total proteins was measured by a bicinchoninic acid (BCA) kit (Fisher Scientific) according to manufacturer's instructions and the sample was stored at -80°C.

Buffer exchange

In this study, an Amicon Ultra Centrifugal Filter (Sigma Aldrich) with a Molecular Weight Cut-Off (MWCO) of 10 kDa was utilized for buffer exchange to eliminate the urea effectively from protein samples. The procedure began with the initial wetting of the filter using 20 µL of 100 mM ammonium bicarbonate, followed by centrifugation at 14,000 g for 10 min. Subsequently, an aliquot of 200 µg proteins was added to the filter, and centrifugation was carried out for 20 min at 14,000 g. 200 µL of 100 mM ammonium bicarbonate was added to the filter, followed by centrifugation at 14,000 g for 20 min. This step was repeated twice to remove the urea and other small interferences completely. The final protein solution in 35 µL of 100 mM ammonium bicarbonate (protein concentration was 3 mg/mL) was collected for CZE-MS analysis. All centrifugation steps were performed at 4°C.

Preparation of linear polyacrylamide (LPA)-coated capillary

An LPA-coated capillary (1 meter, 50 µm i.d., 360 µm o.d.) was prepared according to our previous procedure with minor modifications.[32] First, 3 µL of ammonium persulfate (APS) solution (5% [w/v] in water) was added to 500 µL of acrylamide solution (4% [w/v] in water) and the mixture was degassed with nitrogen gas for 5 min to remove the oxygen in the solution. Then, the mixture was loaded into the pretreated capillary using a vacuum, followed by sealing both ends of the capillary with silica rubber, and incubating it in a water bath at 50 °C for 40 min. Finally, a small portion (~5 mm) of the capillary from both ends was removed with a cleaving stone and the unreacted solution (an agarose gel-like consistency) was pushed out of the capillary with water (200 µL), using the syringe pump.

One end of the separation capillary was etched by hydrofluoric acid to reduce its outer diameter to around 100 μm[33] .

CZE-ESI-MS/MS Analysis

The automated CE operation was performed using an ECE-001 CE autosampler from CMP Scientific (Brooklyn, NY). Through an electro-kinetically pumped sheath flow CE-MS interface (CMP Scientific, Brooklyn, NY), the CE system was coupled to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) [34,35]. For CZE separation, the LPA-coated capillary (50 μm i.d., 360 μm o.d., 1 meter in length) was used. A background electrolyte (BGE) of 5% (v/v) acetic acid (pH 2.4) was used for CZE. The sample buffer was 100 mM ammonium bicarbonate (pH 8). The dramatic difference of BGE and sample buffer in pH enabled online dynamic pH junction-based sample stacking.[26] The sheath buffer contained 0.2% (v/v) formic acid and 10% (v/v) methanol. The sample was injected into the capillary via applying pressure. The sample injection volume was calculated based on the pressure and injection time using Poiseuille's Law. In this study, 5 psi for a 20-s period was applied for sample injection, corresponding to about 100 nL of sample-loading volume for a 1-m-long separation capillary (50-μm i.d.). At the injection end of the separation capillary, a high voltage (30 kV) was applied for separation, and in the sheath buffer vial, a voltage of 2–2.2 kV was applied for ESI. With a Sutter P-1000 flaming/brown micropipet puller, ESI emitters were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., and 10 cm length). ESI emitters had an opening size of 25 to 35 μm.

All experiments were conducted using the Q-Exactive HF mass spectrometer. A Data-dependent acquisition (DDA) method was used for the yeast protein sample. MS parameters were 120,000 mass resolution (at m/z 200), three microscans, 3E6 AGC target value, 100 ms maximum injection time, and 600–2000 m/z scan range. For MS/MS, 60,000 mass resolution (at 200 m/z), 1 microscan, 1E6 AGC, 200 ms injection time, 4 m/z isolation window, and 20% normalized collision energy (NCE) were used. The top 8 most intense precursor ions in one MS spectrum were isolated in the quadropole and fragmented via higher energy collision dissociation (HCD). Fragmentation was performed only on ions with intensities greater than 1E4 and charge

states greater than 5. We enabled dynamic exclusion with a duration of 30 s. The "Exclude isotopes" function was enabled.

Data analysis

The complex sample data was analyzed using Xcalibur software (Thermo Fisher Scientific) to get intensity and migration time of proteins. For the final figures, the electropherograms were exported from Xcalibur and formatted using Adobe Illustrator.

Proteoform identification and quantification were performed on the yeast protein RAW files using the TopPIC (Top-down mass spectrometry-based Proteoform Identification and Characterization) pipeline[36]. In the first step, RAW files were converted into mzML files using the Msconvert tool[37]. The spectral deconvolution which converted precursor and fragment isotope clusters into the monoisotopic masses and proteoform features were then performed using TopFD (Top-down mass spectrometry Feature Detection, version 1.5.6)[38]. The resulting mass spectra and proteoform feature information were stored in msalign and text files, respectively. The database search was performed using TopPIC (version 1.5.6) against UniProt proteome database of Yeast (UP000002311, 6060 entries, version 11/14/2022) concatenated with a shuffled decoy database of the same size of the yeast database. The maximum number of unexpected mass shifts was one. The mass error tolerances for precursors and fragments were 15 parts-per-million (ppm). There was a maximum mass shift of 500 Da for unknown mass shifts. To estimate false discovery rates (FDRs) of proteoform identifications, the target-decoy approach was used and proteoform identifications were filtered by a 1% FDR at the proteoform-spectrum-match (PrSM) level and proteoform level[39,40]. The lists of identified proteoforms from all CZE-MS/MS runs are shown in **Supporting Information I**. The TopDiff (Top-down mass spectrometry-based identification of Differentially expressed proteoforms, version 1.5.6) software was used to perform label-free quantification of identified proteoforms by CZE-MS/MS using default settings[41]. The MS RAW files were deposited to the ProteomeXchange Consortium via the PRIDE[42]partner repository with the dataset identifier of PXD046651.

Capillary cleanup

To remove proteins adsorbed on the capillary inner wall, the capillary was cleaned periodically by flushing with 0.5% NH_4OH for 10 minutes at 30 psi, H_2O for 10 min at 20 psi, and the BGE (5% acetic acid) for 10 min at 20 psi successively.

Results and discussion

For the first time, we studied long-term reproducibility of CZE-MS/MS for TDP of a complex proteome sample, a yeast cell lysate, and developed an effective procedure for cleaning up the inner wall of LPA coated capillaries for reproducible CZE-MS/MS measurements of proteoforms. **Figure 1A** shows the experimental design of this project. Yeast cells were lysed by homogenization and sonication. The proteoform extract was analyzed by the dynamic pH junction-based CZE-MS/MS[26] after a simple buffer exchange with a 10-kDa cut-off centrifugal filter unit. The yeast cell lysate was diluted to 1 mg/mL with 100 mM ammonium bicarbonate (pH 8) for CZE-MS/MS. Finally, the TopPIC software developed by the Liu's group was used for database search to identify and quantify proteoforms. **Figure 1B** represents the clean-up procedure to remove the adsorbed proteoforms on the LPA polymer coating on the capillary inner wall.

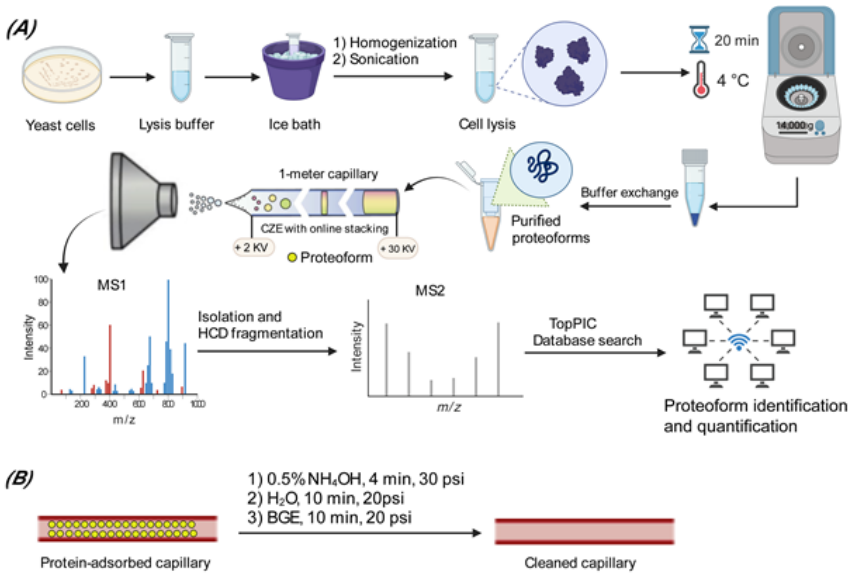


Figure 1. (A) Schematic of the experimental design of sample preparation, CZE-MS/MS analysis, and database search. (B) schematic of the capillary inner-wall clean-up

procedure using NH_4OH. The figure is created using the BioRender and is used here with permission.

Reproducibility of CZE-MS/MS for top-down proteomics of a complex sample

CZE-MS/MS with a fresh LPA-coated capillary generated reproducible measurements of the yeast cell lysate, which is evidenced by the example electropherograms and the number of proteoform identifications from the first roughly 10 runs, **Figures 2A** and **3A.** When we kept running the yeast cell lysate, we observed that the proteoform peaks were broadened gradually and proteoform intensity decreased accordingly, **Figure 2A**. The peak width of one proteoform doubled in run 14 compared to that in run 1 and the proteoform intensity decreased by a factor of two roughly. For runs 16, 22, and 23, the peak width of the example proteoform tripled and the proteoform intensity is only 20% of that in run 1. The number of proteoform and protein IDs decreased obviously from run 10 to run 24, **Figures 3A** and **3B**.

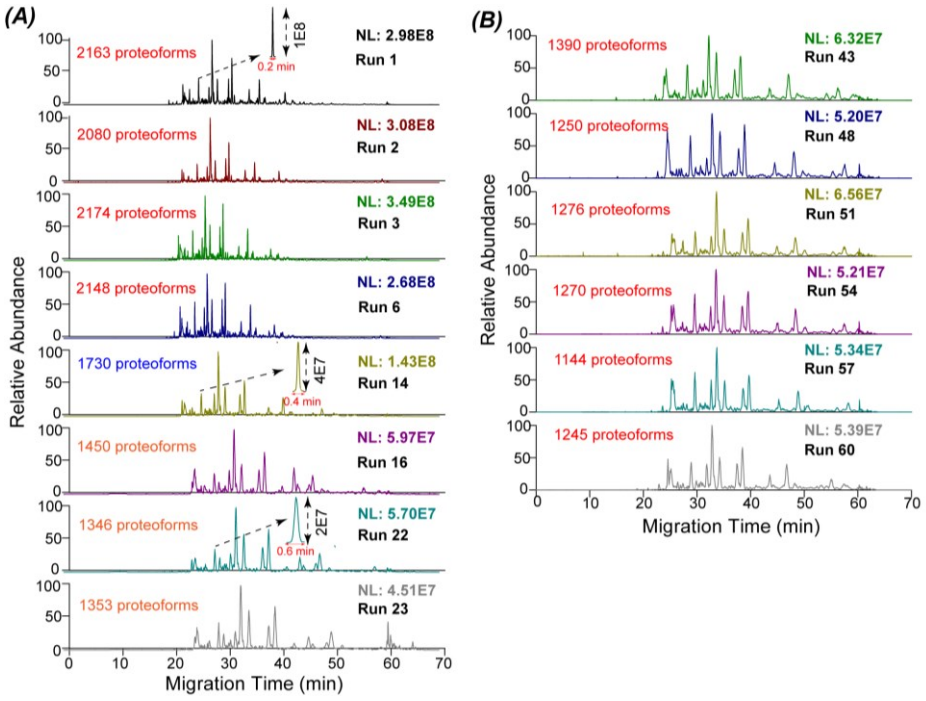


Figure 2. Electropherograms of a yeast cell lysate after analyses by CZE-MS/MS. (A) Example runs during the first 23 CZE-MS/MS measurements. (B) Six example CZE-MS/MS runs during the 40th to 62nd measurements.

We suspected that the phenomenon was due to proteoform adsorption onto the LPA polymer coating on the capillary inner wall. When more and more CZE-MS/MS runs are done, proteoforms are gradually adsorbed onto the capillary wall. The adsorbed proteoforms can have significant impacts on CZE separation. Proteoforms on the capillary inner wall are positively charged under the acidic BGE of 5% (v/v) acetic acid (pH 2.4), leading to a potential of the generation of low reversed electroosmotic flow (EOF) in the capillary. The reversed EOF slows down the migration of proteoforms in the cap usion and dispersion[43]. The reversed EOF could also affect the performance of dynamic pH junction stacking because it could negatively impact the migration of hydrogen protons from the BGE vial to the separation capillary for sample zone titration.

Figure S1B shows an example electropherogram of the yeast cell lysate after over 30 continuous CZE-MS/MS runs without capillary cleanup. Once we cleaned up the capillary inner wall using a procedure involving capillary flushing with 0.5% ammonium hydroxide, water, and the BGE, the separation profile and the number of proteoform IDs recovered back to nearly the original condition, **Figure S1A** and **S1C**. The data demonstrate that the cleaning up method can remove the adsorbed proteoforms efficiently.

After the 1st and 2nd capillary cleanup, we observed the repeated phenomenon as the fresh capillary. The number of proteoform and protein IDs declined as the runs continued, **Figures 3A** and **3B**. Interestingly, after the third cleanup, the capillary inner-wall condition became more stable, evidenced by the relatively more consistent numbers of proteoform and protein IDs (**Figures 3A** and **3B**) as well as more reproducible proteoform separations (**Figure 2B**). Our data suggest that to achieve reproducible top-down MS measurements of a complex proteome sample by CZE-MS/MS, we can perform the experiment either using a fresh LPA-coated capillary (Phase I) or using an LPA-coated capillary after an enough amount of protein adsorption and sufficient capillary cleanup with 0.5% ammonium hydroxide (Phase II). The phase II condition can provide reproducible CZE-MS/MS measurements for more than 23 runs.

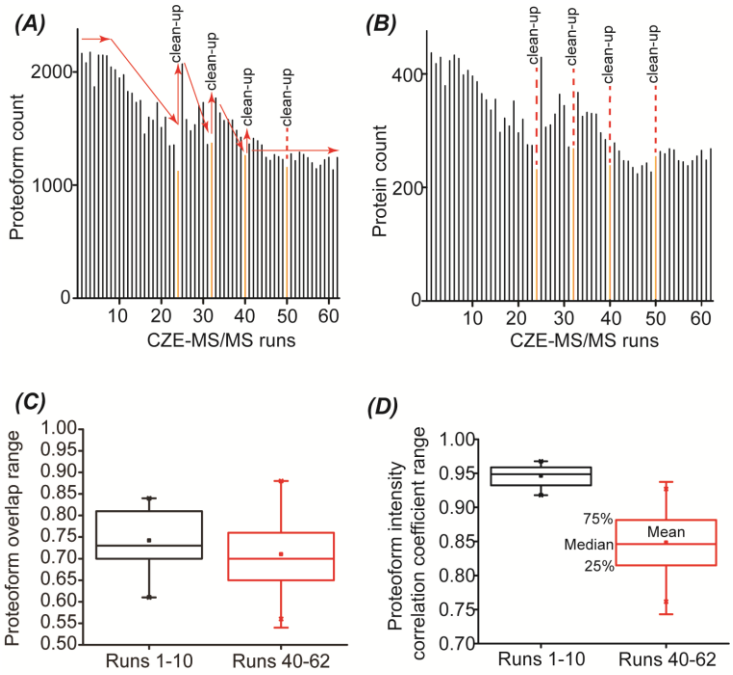


Figure 3. Summary of the identified proteoforms and proteins from 62 CZE-MS/MS runs. (A) The number of proteoform IDs as a function of the run number. (B) The number of protein IDs as a function of the run number. The trends of number of proteoform IDs and the time for capillary cleanup are marked. (C) Boxplots of pairwise proteoform overlaps for runs 1-10 and 40-62. (D) Boxplots of pairwise Pearson correlation coefficients of proteoform intensity for runs 1-10 and 40-62. Log2 transformed proteoform intensities were used to generate the Pearson correlation coefficients.

We further studied the pairwise overlap of identified proteoforms for phase I (runs 1-10) and phase II (runs 40-62) conditions, **Figure 3C**. The medians of proteoform overlap between any two CZE-MS/MS runs in phase I and phase II are both between 70%-75%, which is comparable to CZE-MS/MS data in literatures[44]. It documents that CZE-MS/MS in both conditions can repeatedly identify the same proteoforms from the yeast cell lysate. The small variations in the identified proteoforms are most likely due to the randomness of data dependent acquisition (DDA).

To investigate the quantitative reproducibility of CZE-MS/MS in both phase I and II conditions, we studied the pairwise proteoform intensity correlation coefficients for runs 1-10 and 40-62, **Figure 3D**. Label free quantification of proteoforms was performed by

the TopDiff software[41] . The intensities of overlapped proteoforms between any two runs were used to create the Pearson linear correlation and obtain the correlation coefficients. The median for the phase I runs is about 0.95 and the correlation coefficient has a narrow distribution, suggesting high quantitative reproducibility. The median for the phase II runs is about 0.85, indicating reasonable quantitative reproducibility. The much lower Pearson linear correlation coefficients in phase II runs than phase I runs are most likely due to drastically lower proteoform intensities in phase II runs, as shown in **Figures 2A** and **2B**.

To further confirm the possibility of CZE-MS/MS in the phase II condition for accurate label-free quantification of proteoforms in a complex sample, after the 62 CZE-MS/MS runs of the yeast cell lysate, we performed CZE-MS/MS analyses of a 3-times diluted yeast cell lysate in quintuplicate, **Figure 4**. The CZE-MS/MS produced reproducible measurements of the original and diluted yeast cell lysates in terms of separation profiles, the number of proteoform IDs (1204±49 for original vs. 753±33 for diluted, relative standard deviations (RSDs) as 4%, n=5), and the normalized level (NL) intensities (RSDs: 8%-9%, n=5), **Figures 4A** and **4B**. The average NL intensity of the diluted sample is about 3 times lower than that of the original sample (4.7E7 vs. 1.4E7), which agrees well with the dilution factor of 3, demonstrating that the CZE-MS/MS in the phase II condition performs well for relative quantification of proteoforms. We further analyzed the distribution of proteoform intensity ratios between original and diluted samples, **Figure 4C**. The median of the ratios is close to the theoretical ratio of 3. The number of matched fragment ions from the original sample is consistently higher than that from the diluted sample, most likely due to the much higher proteoform intensity, **Figure 4D**. Majority of the identified proteoforms have more than 10 matched fragment ions for the original and diluted samples, indicating reasonably high confidence of the proteoform IDs.

The results presented in this study are critically important for CZE-MS/MS for top-down proteomics of complex samples. First, the data document that CZE-MS/MS using one LPA-coated capillary can produce high-quality top-down proteomics data of a complex proteome sample for at least 78 hours (67 runs and 70 minutes per run), indicating the high robustness of the system. Second, the study provides us with rich experimental data that can be extremely useful for pursuing a better understanding of CZE-MS for

proteoform separation and characterization. Third, the results demonstrate that CZE-MS/MS with an appropriate operational procedure (i.e., capillary cleanup) can generate highly reproducible separation and identification of proteoforms in a complex sample across dozens of runs. The CZE-MS/MS is ready for some important biological applications to discover potentially critical proteoforms in biological processes and diseases in a quantitative manner. Fourth, the data also highlights some potential challenges of CZE-MS/MS for large-scale top-down proteomics studies in the next step and point out some important directions to work on. For example, we need to make more effort to create more consistent capillary inner-wall chemistry during CZE-MS/MS runs, which will eventually make CZE-MS/MS a powerful and highly reproducible technique for large-scale top-down proteomics studies.

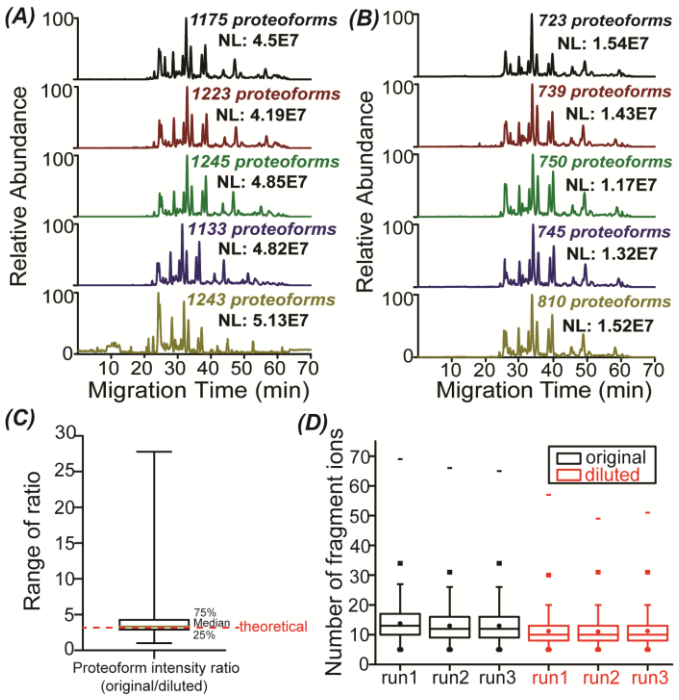


Figure 4. Comparisons of the original and diluted yeast cell lysate data from CZE-MS/MS analyses. (A) Base peak electropherograms of the original yeast cell lysate after CZE-MS/MS analyses in quintuplicate. (B) Base peak electropherograms of the 3-times diluted yeast cell lysate after CZE-MS/MS analyses in quintuplicate. (C) Boxplot of the intensity ratio of overlapped proteoforms between original and diluted yeast cell lysates.

(D) Boxplots of the number of matched fragment ions of identified proteoforms from original and diluted yeast samples.

Correlation of experimental and predicted electrophoretic mobility of proteoforms under different CZE-MS/MS conditions

We have shown that the electrophoretic mobility (μ_{ef}) of proteoforms in CZE can be predicted well using a simple semiempirical model[21,45]. Proteoforms' experimental and predicted μ_{ef} have high linear correlation coefficients. This feature is critically useful for validating the proteoform IDs and the PTMs (i.e., phosphorylation). Here we have multiple different CZE-MS/MS conditions, phase I (runs 1-10), phase II (runs 40-62), and transition period between them (runs 11-39). We are asking how those CZE-MS/MS conditions influence the correlation of experimental and predicted μ_{ef} of proteoforms.

For the experimental μ_{ef} ($cm^2 \cdot kV^{-1} \cdot s^{-1}$), we used the equation 1 for calculation.

$$\text{Experimental } \mu_{ef} = L / ((30-2) / L^* t_M) \quad \text{(eq. 1)}$$

Where L is the capillary length in cm, and t_M is the migration time in seconds. 30 and 2 are the separation voltage and electrospray voltage in kilovolts, respectively.

For the predicted μ_{ef} ($cm^2 \cdot kV^{-1} \cdot s^{-1}$), we utilized the equation 2.

$$\text{Predicted } \mu_{ef} = \ln (1 + 0.35^*Q) / M^{0.411} \quad \text{(eq. 2)}$$

Where M and Q represent molecular mass and charge number of each proteoform, respectively. We got the information of M directly from the database search results. We obtained Q by counting the number of lysine, arginine, and histidine amino acid residues in the proteoform sequence and added 1 for the N-terminus.

We only used proteoforms containing no PTMs and those having N-terminal acetylation or phosphorylation for this study. As shown in **Figures 5 A-C** (top panels), strong linear correlations between experimental and predicted μ_{ef} were observed for proteoforms without any PTMs (R^2=0.96). As shown in the middle panels, when we consider the

proteoforms with N-terminal acetylation or phosphorylation, those modified proteoforms fall off the main trend and have lower experimental μ_{ef} compared to the corresponding non-modified proteoforms. The reduction of experimental μ_{ef} is due to the charge (Q) reduction by one from the N-terminal acetylation or phosphorylation, considering the acidic BGE of CZE (i.e., 5% acetic acid, pH 2.4). After reducing the estimated net charge Q by one for the μ_{ef} prediction, we achieved strong linear correlation coefficients (R^2=0.95-0.96) for the non-modified proteoforms and proteoforms having N-terminal acetylation or phosphorylation, **Figures 5 A-C** (bottom panels). The results here suggest that the proteoforms identified in this study have high confidence because of the strong linear correlations between experimental and predicted μ_{ef}. In addition, the data indicate that the different CZE-MS/MS conditions do not have significant impact on the correlations between experimental and predicted μ_{ef}. We realized that the experimental μ_{ef} of proteoforms become lower from run 5 (≥0.15, **A**) to run 52 (≥0.1, **C**), which is due to the much longer migration times of proteoforms in run 52 compared to run 5, **Figure 2**.

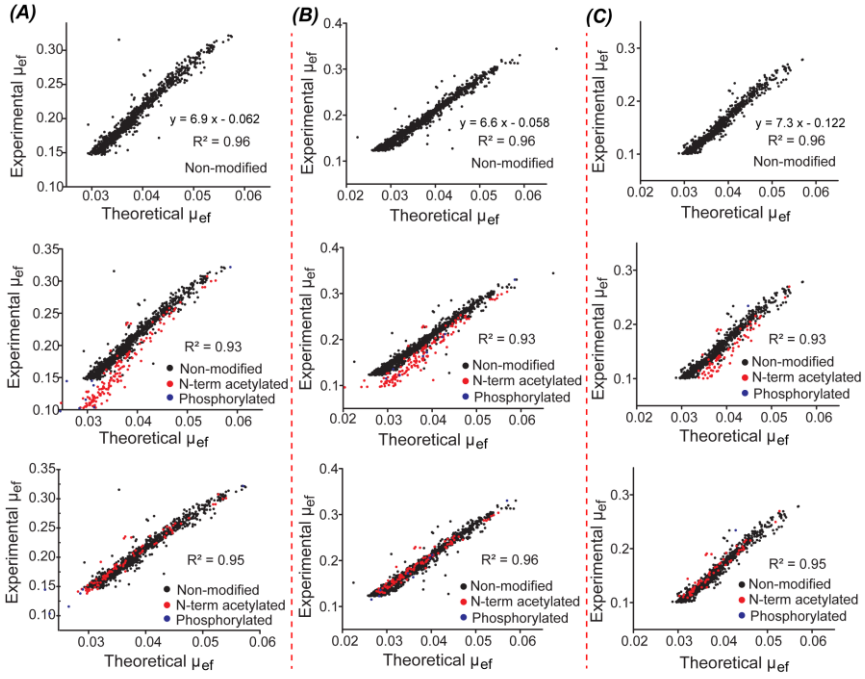


Figure 5. Linear correlations between predicted μ_{ef} and experimental μ_{ef} of proteoforms from the yeast cell lysate identified in CZE-MS/MS run 5 (A), 25 (B), and run 52 (C). The top figures show the correlations for proteoforms without any PTMs. The middle ones

indicate the correlations for all proteoforms without PTMs and with N-terminal acetylation or phosphorylation. The charge Q of those proteoforms were not corrected. The bottom ones show the correlations for the same proteoforms as the middle ones but with charge Q correction. For example, for one N-terminal acetylation or phosphorylation, the Q was reduced by one.

Conclusions

For the first time, long-term qualitative and quantitative reproducibility of CZE-MS/MS for a complex proteome sample was investigated. We revealed significant changes of proteoforms in migration time and intensity after about 10 CZE-MS/MS runs of the yeast cell lysate due to proteoform adsorption onto the capillary inner wall. We developed an efficient and simple capillary cleanup procedure via flushing the capillary with 0.5% NH_4OH, water, and the separation buffer successively. The capillary cleanup protocol can remove the adsorbed proteoforms efficiently. After several rounds of capillary cleanup, the capillary inner wall chemistry became more consistent, producing reproducible proteoform separation and identification across dozens of CZE-MS/MS analyses of the yeast cell lysate. The results in this work highlight that CZE-MS/MS is robust enough to create high-quality top-down proteomics measurement of a complex sample across dozens of runs, for example more than 60 runs of the yeast cell lysate. In addition, the measurement can be qualitatively and quantitatively reproducible across dozens of runs (i.e., at least 23 runs) under some specific condition with an appropriate operational procedure (i.e., regular capillary cleanup). We expect that it is time to apply CZE-MS/MS-based top-down proteomics to broad biological applications.

We have some recommendations about using CZE-MS for quantitative top-down proteomics. For label-free quantification, we should not combine the Phase I and Phase II conditions because of the dramatic shifts in migration time, making the data alignment challenging across CZE-MS runs for relative quantification. If a small-scale label-free quantification is performed, for example, comparing two samples with only about ten CZE-MS runs or fewer, the Phase I condition will be ideal. If a large-scale study is needed, for example, comparing multiple samples with more than 20 CZE-MS runs, the Phase II

condition should be considered. Alternatively, stable isotopic labeling techniques (e.g., tandem mass tags [46]) can be employed. In this case, we may do not need to worry about the Phase I or Phase II condition, because the relative quantification is performed based on the data within the same CZE-MS runs.

Supporting Information

The following supporting information is available free of charge at ACS website.

Supporting Information 1: The lists of identified proteoforms from all CZE-MS/MS runs (XLSX)

Supporting Information 2: Electropherograms of a yeast cell lysate by CZE-MS/MS in three instances (PDF)

Notes

The authors declare no competing financial interest.

Acknowledgments

We thank the supports from the National Cancer Institute (NCI) through the grant R01CA247863, the National Institute of General Medical Sciences (NIGMS) through grants R01GM125991 and R01GM118470, and the National Science Foundation through the grant DBI1846913 (CAREER Award).

References

[1] Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

[2] Adams, L. M.; DeHart, C. J.; Drown, B. S.; Anderson, L. C.; Bocik, W.; Boja, E. S.; Hiltke, T. M.; Hendrickson, C. L.; Rodriguez, H.; Caldwell, M.; Vafabakhsh, R.; Kelleher, N. L. Mapping the KRAS proteoform landscape in colorectal cancer identifies truncated KRAS4B that decreases MAPK signaling. *J. Biol. Chem.* **2023**, *299*, 102768.

[3] McCool, E. N.; Xu, T.; Chen, W.; Beller, N. C.; Nolan, S. M.; Hummon, A. B.; Liu, X.; Sun, L. Deep Top-down Proteomics Revealed Significant Proteoform-Level Differences between Metastatic and Nonmetastatic Colorectal Cancer Cells. *Sci. Adv.* **2022**, *8*, eabq6348.

[4] Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D.; Ellis, M. J. C.; Ruggles, K. V.; Fenyö, D.; Boja, E. S.; Rodriguez, H.; Townsend, R. R.; Kelleher, N. L. Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol. Cell. Proteomics* **2016**, *15*, 45-56.

[5] Ntai, I.; Fornelli, L.; DeHart, C. J.; Hutton, J. E.; Doubleday, P. F.; LeDuc, R. D.; Nispen, A. J. van.; Fellers, R. T.; Whiteley, G.; Boja, E. S.; Rodriguez, H.; Kelleher, N. L. Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4140-4145.

[6] Kandi, S.; Cline, E. N.; Rivera, B. M.; Viola, K. L.; Zhu, J.; Condello, C.; LeDuc, R. D.; Klein, W. L.; Kelleher, N. L.; Patrie, S. M. Amyloid β Proteoforms Elucidated by Quantitative LC/MS in the 5xFAD Mouse Model of Alzheimer's Disease. *J. Proteome Res.* **2023***, 22,* 3475-3488.

[7] Kellie, J. F.; Higgs, R. E.; Ryder, J. W.; Major, A.; Beach, T. G.; Adler, C. H.; Merchant, K.; Knierman, M. D. Quantitative measurement of intact alpha-synuclein proteoforms from post-mortem control and Parkinson's disease brain tissue by intact protein mass spectrometry. *Sci. Rep.* **2014**, *4*, 5797.

[8] Schmitt, N. D.; Agar, J. N. Parsing disease-relevant protein modifications from epiphenomena: perspective on the structural basis of SOD1-mediated ALS. *J. Mass Spectrom.* **2017**, *52*, 480-491.

[9] Forgrave, L. M.; Moon, K. M.; Hamden, J. E.; Li, Y.; Lu, P.; Foster, L. J.; Mackenzie, I. R. A.; DeMarco, M. L. Truncated TDP-43 proteoforms diagnostic of frontotemporal dementia with TDP-43 pathology. *Alzheimers Dement.* **2023**, doi: 10.1002/alz.13368.

[10] Ge, Y.; Rybakova, I. N.; Xu, Q.; Moss, R. L. Top-down High-Resolution Mass Spectrometry of Cardiac Myosin Binding Protein C Revealed That Truncation Alters Protein Phosphorylation State. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12658–12663.

[11] Chamot-Rooke, J.; Mikaty, G.; Malosse, C.; Soyer, M.; Dumont, A.; Gault, J.; Imhaus, A.-F.; Martin, P.; Trellet, M.; Clary, G.; Chafey, P.; Camoin, L.; Nilges, M.; Nassif, X.; Duménil, G. Posttranslational Modification of Pili upon Cell Contact Triggers N. Meningitidis Dissemination. *Science* **2011**, *331*, 778–782.

[12] Roberts, D. S.; Mann, M.; Melby, J. A.; Larson, E. J.; Zhu, Y.; Brasier, A. R.; Jin, S.; Ge, Y. Structural O-Glycoform Heterogeneity of the SARS-CoV-2 Spike Protein Receptor-Binding Domain Revealed by Top-Down Mass Spectrometry. *J. Am. Chem. Soc.* **2021**, *143*, 12014-12024.

[13] Gstottner, C.; Zhang, T.; Resemann, A.; Ruben, S.; Pengelley, S.; Suckau, D.; Welsink, T.; Wuhrer, M.; Domínguez-Vega, E. Structural and functional characterization of SARS-CoV-2 RBD domains produced in mammalian cells. *Anal. Chem.* **2021**, *93,* 6839-6847.

[14] Wilson, J. W.; Bilbao, A.; Wang, J.; Liao, Y. C.; Velickovic, D.; Wojcik, R.; Passamonti, M.; Zhao, R.; Gargano, A. F. G.; Gerbasi, V. R.; Pas A-Tolić, L.; Baker, S. E.; Zhou, M. Online hydrophilic interaction chromatography (HILIC) enhanced top-down mass spectrometry characterization of the SARS-CoV-2 spike receptor-binding domain. *Anal. Chem.* **2022**, *94*, 5909-5917.

[15] Melani, R. D.; Des Soye, B. J.; Kafader, J. O.; Forte, E.; Hollas, M.; Blagojevic, V.; Negrão, F.; McGee, J. P.; Drown, B.; Lloyd-Jones, C.; Seckler, H. S.; Camarillo, J. M.; Compton, P. D.; LeDuc, R. D.; Early, B.; Fellers, R. T.; Cho, B. K.; Mattamana, B. B.; Goo, Y. A.; Thomas, P. M.; Ash, M. K.; Bhimalli, P. P.; Al-Harthi, L.; Sha, B. E.; Schneider, J. R.; Kelleher, N. L. Next-generation serology by mass spectrometry: readout of the SARS-CoV-2 antibody repertoire. *J. Proteome Res.* **2021**, *21*, 274-288.

[16] Brown, K. A.; Melby, J. A.; Roberts, D. S.; Ge, Y. Top-down proteomics: challenges, innovations, and applications in basic and clinical research. *Expert Rev. Proteomics* **2020**, *17*, 719-733.

[17] Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19*, e1800361.

[18] Lukacs, K. D.; Jorgenson, J. W. Capillary Zone Electrophoresis. *Science* **1983**, *222*, 266-272.

[19] Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L. Recent Advances (2019–2021) of Capillary Electrophoresis-Mass Spectrometry for Multilevel Proteomics. *Mass Spectrom Rev* **2023**, *42*, 617–642.

[20] Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends Analyt. Chem.* **2019**, *120*, 115644.

[21] Chen, D.; Lubeckyj, R. A.; Yang, Z.; McCool, E. N.; Shen, X.; Wang, Q.; Xu, T.; Sun, L. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. *Anal Chem* **2020**, *92*, 3503–3507.

[22] Valaskovic, G. A.; Kelleher, N.L.; McLafferty, F. W. Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science* **1996,** *273*, 1199-1202.

[23] Akashi, S.; Suzuki, K.; Arai, A.; Yamada, N.; Suzuki, E.; Hirayama, K.; Nakamura, S.; Nishimura, Y. Top-down analysis of basic proteins by microchip capillary electrophoresis mass spectrometry. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1932-38.

[24] Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavallée-Adam, M.; Yates, J. R. Sheathless capillary electrophoresis-tandem mass spectrometry for top-down

characterization of Pyrococcus furiosus proteins on a proteome scale. *Anal. Chem.* **2014**, *86*, 11006-11012.

[25] Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. Coupling capillary zone electrophoresis to a Q exactive HF mass spectrometer for top-down proteomics: 580 proteoform identifications from yeast. *J. Proteome Res.* **2016**, *15*, 3679-3685.

[26] Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia Coli Proteoforms. *Anal Chem* **2017**, *89*, 12059–12067.

[27] Xu, T.; Wang, Q.; Wang, Q.; Sun, L. Coupling High-Field Asymmetric Waveform Ion Mobility Spectrometry with Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics. *Anal. Chem.* **2023**, *95*, 9497–9504.

[28] Wang, Q.; Fang, F.; Wang, Q.; Sun, L. Capillary zone electrophoresis-high field asymmetric ion mobility spectrometry-tandem mass spectrometry for top-down characterization of histone proteoforms. *Proteomics* **2023**, e2200389.

[29] Wang, Q.; Xu, T.; Fang, F.; Wang, Q.; Lundquist, P.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of Mouse Brain Integral Membrane Proteins. *Anal. Chem.* **2023**, *95*, 12590–12594.

[30] Drown, B. S.; Jooß, K.; Melani, R. D.; Lloyd-Jones, C.; Camarillo, J. M.; Kelleher, N. L. Mapping the Proteoform Landscape of Five Human Tissues. *J. Proteome Res.* **2022**, *21*, 1299–1310.

[31] Johnson, K. R.; Gao, Y.; Greguš, M.; Ivanov, A. R. On-Capillary Cell Lysis Enables Top-down Proteomic Analysis of Single Mammalian Cells by CE-MS/MS. *Anal. Chem.* **2022**, *94*, 14358–14367.

[32] Zhu, G.; Sun, L.; Dovichi, N. J. Thermally-Initiated Free Radical Polymerization for Reproducible Production of Stable Linear Polyacrylamide Coated Capillaries, and Their

Application to Proteomic Analysis Using Capillary Zone Electrophoresis-Mass Spectrometry. *Talanta* **2016**, *146*, 839–843.

[33] Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J. Ultrasensitive and Fast Bottom-up Analysis of Femtogram Amounts of Complex Proteome Digests. *Angew. Chem Int. Ed.* **2013***, 52, 13661-4.*

[34] Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. Simplified Capillary Electrophoresis Nanospray Sheath-Flow Interface for High Efficiency and Sensitive Peptide Analysis. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554–2560.

[35] Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. Third-Generation Electrokinetically Pumped Sheath-Flow Nanospray Interface with Improved Stability and Sensitivity for Automated Capillary Zone Electrophoresis-Mass Spectrometry Analysis of Complex Proteome Digests. *J. Proteome Res.* **2015**, *14*, 2312–2321.

[36] Kou, Q.; Xun, L.; Liu, X. TopPIC: A Software Tool for Top-down Mass Spectrometry-Based Proteoform Identification and Characterization. *Bioinformatics* **2016**, *32*, 3495–3497.

[37] Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, *24*, 2534–2536.

[38] Basharat, A. R.; Zang, Y.; Sun, L.; Liu, X. TopFD: A Proteoform Feature Detection Tool for Top–Down Proteomics. *Anal. Chem.* **2023**, *95*, 8189–8196.

[39] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74*, 5383–5392.

[40] Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4*, 207–214.

[41] Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L. Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 1435–1445.

[42] Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yılmaz, Ş.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res.* **2019**, *47*, D442–D450.

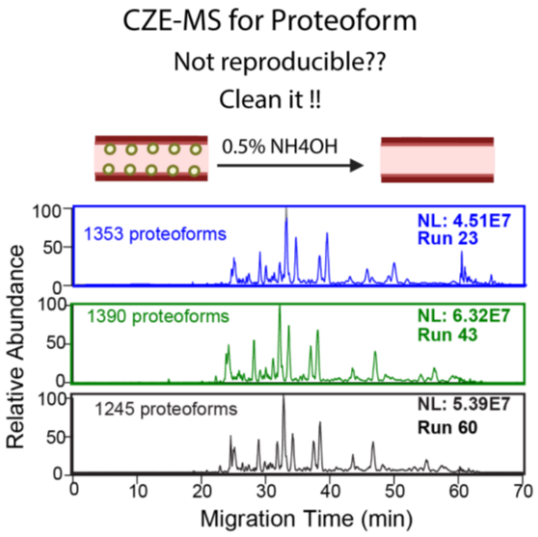
[43] Khodabandehloo, A.; Chen, D. D. Y. Electroosmotic Flow Dispersion of Large Molecules in Electrokinetic Migration. *Anal. Chem*. **2017**, *89*, 7823–7827.

[44] McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L. Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia Coli Proteome. *Anal. Chem.* **2018**, *90*, 5529–5533.

[45] Chen, D.; Yang, Z.; Shen, X.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry As an Alternative to Liquid Chromatography-Tandem Mass Spectrometry for Top-down Proteomics of Histones. *Anal. Chem.* **2021**, *93*, 4417–4424.

[46] Guo, Y.; Chowdhury, T.; Seshadri, M.; Cupp-Sutton, K. A.; Wang, Q.; Yu, D.; Wu, S. Optimization of Higher-Energy Collisional Dissociation Fragmentation Energy for Intact Protein-Level Tandem Mass Tag Labeling. *J. Proteome Res.* **2023**, *22*, 1406-1418.

For TOC Only


CZE-MS for Proteoform
Not reproducible??
Clean it !!
0.5% NH4OH
1353 proteoforms
NL: 4.51E7
Run 23
1390 proteoforms
NL: 6.32E7
Run 43
1245 proteoforms
NL: 5.39E7
Run 60
Relative Abundance
Migration Time (min)