Predicting electrophoretic mobility of proteoforms for large-scale top-down proteomics

Daoyang Chen, Rachele Lubeckyj, Zhichang Yang, Elijah McCool, Xiaojing Shen, Qianjie Wang, Tian Xu, Liangliang Sun*

Department of Chemistry, Michigan State University, 578 S Shaw Ln, East Lansing, MI 48824, USA

* Corresponding author. Phone: 517-353-0498

Email: lsun@chemistry.msu.edu

Abstract

Large-scale top-down proteomics characterizes proteoforms in cells globally with high confidence and high throughput using reversed-phase liquid chromatography (RPLC)-tandem mass spectrometry (MS/MS) or capillary zone electrophoresis (CZE)-MS/MS. The false discovery rate (FDR) from the target-decoy database search is typically deployed to filter identified proteoforms to ensure high-confidence identifications (IDs). It has been demonstrated that the FDRs in top-down proteomics can be drastically underestimated. An alternative approach to the FDR can be useful for further evaluating the confidence of proteoform IDs after database search. We argue that predicting retention/migration time of proteoforms from the RPLC/CZE separation accurately and comparing their predicted and experimental separation time could be a useful and practical approach. Based on our knowledge, there is still no report in the literature about predicting separation time of proteoforms using large top-down proteomics datasets. In this pilot study, for the first time, we evaluated various semi-empirical models for predicting proteoforms' electrophoretic mobility (μ_{ef}) using large-scale top-down proteomics datasets from CZE-MS/MS. We achieved a linear correlation between experimental and predicted μ_{ef} of *E. coli* proteoforms (R^2=0.98) with a simple semi-empirical model, which utilizes the number of charges and molecular mass of each proteoform as the parameters. Our modeling data suggest that the complete unfolding of proteoforms during CZE separation benefits the prediction of their μ_{ef}. Our results also indicate that N-terminal acetylation and phosphorylation both decrease proteoforms' charge by roughly one charge unit.

Mass spectrometry (MS)-based top-down proteomics aims to delineate proteoforms in cells comprehensively with high confidence and throughput.[1-5] Proteoforms extracted from biological samples are typically separated by reversed-phase liquid chromatography (RPLC) or capillary zone electrophoresis (CZE), followed by electrospray ionization (ESI)-tandem mass spectrometry (MS/MS). Database search is then performed for the identification (ID) of proteoform spectrum matches (PrSMs), proteoforms, and proteins through comparing experimental and theoretical masses of proteoforms and their fragments. To improve the confidence of proteoform ID, the target-decoy database search approach is typically employed, [6,7] and the identified PrSMs and proteoforms were filtered by certain false discovery rates (FDRs). Recently, the Kelleher's group showed that the FDR estimation in top-down proteomics was complicated and the FDRs could be drastically under-reported.[8] High-confidence proteoform and protein IDs are vital. Therefore, after filtering the data with a specific FDR, we need to validate the data further using an alternative approach to the FDR.

The retention/migration time of proteoforms in LC/CZE can be useful information for improving the confidence of IDs. Some previous studies have deployed the retention/migration time of proteins and peptides to facilitate their IDs.[9-12] We believe that accurate prediction of the retention/migration time of proteoforms will push the use of separation time for ID forward drastically. By comparing the experimentally observed and accurately predicted separation time of proteoforms, we could further boost the confidence of identified proteoforms, determine wrong proteoform IDs, and even provide useful information to correct proteoform IDs.

Some work has been done in predicting migration time (electrophoretic mobility, μ_{ef}) of peptides from CZE separations.[13-21] It has been demonstrated that CZE outperformed RPLC regarding the prediction of migration/retention time of peptides for bottom-up proteomics.[21] One major reason is that the size and charge of peptides for CZE can be calculated relatively easily, by contrast, the interaction between peptides and beads for RPLC is complicated.[21] Krokhin *et al.* achieved a linear correlation (R^2=0.995) between predicted and experimental μ_{ef} of peptides in CZE using a large peptide dataset and an optimized semi-empirical model,[21] which was based on the model reported by Cifuentes

et al.,[19] Equation (1). Note: The equation (1) is the modified version from the reference [19], and Krokhin *et al.* started their optimization from this equation for peptides.

$$\mu_{ef} = 900 \times (\ln(1 + 0.35 \times Q)/M^{0.411}) \quad \text{Equation (1)}$$

In the modified Cifuentes's model, molecular weight (M) and charge (Q) were used as the parameters. The Charge (Q) was equal to the number of positively charged amino acid residues (K, R, H, and N-terminus) in the acidic background electrolyte (BGE) of CZE, for example, 5% (v/v) acetic acid (AA), pH 2.4.[21] More recently, we also applied the similar model for predicting the μ_{ef} of phosphorylated peptides and achieved a high correction (R^2=0.99) between the predicted and experimental μ_{ef} for mono-phosphorylated peptides from the HCT116 cell line.[22]

Great success has been achieved for predicting μ_{ef} of peptides, but much more effort need to be made on proteins/proteoforms. Some initial effort has been made using a handful of standard proteins.[17,23,24] However, there is no report about predicting μ_{ef} of proteins/proteoforms using large-scale proteoform datasets. There are two major reasons for that. First, large-scale top-down proteomics datasets from CZE-MS have been limited. Second, proteins/proteoforms are much larger than peptides, leading to potential difficulties in calculating their size and charge accurately. In the last 5 years, CZE-MS has been recognized as an important approach for large-scale top-down proteomics due to the improvement in CE-MS interfaces, capillary coatings, and online sample stacking techniques.[25-32] For instance, we identified nearly 600 proteoforms from an *E. coli* cell lysate in a single-shot CZE-MS/MS analysis.[27] In that study, we employed a commercialized electro-kinetically pumped sheath-flow CE-MS interface,[33,34] a 1-meter-long linear polyacrylamide (LPA)-coated capillary,[35] and a dynamic pH junction-based proteoform stacking method [36] to boost the sample loading capacity, separation window, and overall sensitivity of the CZE-MS system. In another study, we used a 1.5-meter-long LPA-coated capillary for CZE-MS/MS analysis of zebrafish brains and identified thousands of proteoforms in a single analysis with consumption of nanograms of protein material.[29] These large-scale proteoform datasets provide us great opportunities to push forward the prediction of μ_{ef} of proteoforms, which will be useful for improving the confidence of proteoform IDs in top-down proteomics.

Here, we applied previously reported semi-empirical mobility models in the prediction of proteoforms' μ_{ef} and evaluated their performance using large proteoform datasets from *E. coli* cells and zebrafish brains under different CZE conditions. For the zebrafish brain datasets, we used the published data from our group and the detailed experimental conditions are shown in reference [29]. Briefly, a 1.5-meters-long LPA-coated capillary (50/360 μm i.d./o.d.) was used for CZE separation. The BGE was 10% (v/v) AA, pH 2.2. For the *E. coli* datasets, we generated these data for the project. In brief, the *E. coli* proteins were denatured, reduced and alkylated, followed by desalting with a C4 trap column according to the procedure in the reference [27]. The lyophilized protein sample was redissolved in a 50 mM ammonium bicarbonate (NH_4HCO_3) buffer (pH 8.0) to get a 2 mg/mL protein solution for CZE-MS/MS. A 103-cm-long LPA-coated capillary (50/360 μm i.d./o.d.) was used for CZE. Three different BGEs were tested, including 5% (v/v) AA in water, 20% (v/v) AA in water, and 20% (v/v) AA in water containing 10% (v/v) isopropanol (IPA) and 15% (v/v) dimethylacetamide (DMA). Approximately 400 nL of the sample, equivalent to 800 ng of *E. coli* proteins was injected for analysis per CZE-MS/MS run. Technical triplicates were performed for each BGE. The commercialized electro-kinetically pumped sheath-flow CE-MS interface from CMP Scientific (Brooklyn, NY) was employed to couple CZE to MS.[33,34] For all the experiments, +30 kV was applied at the sample injection end, and +2 kV was applied at the interface for ESI. A Q-Exactive HF mass spectrometer was used. The raw files from *E. coli* cells were searched against the UniProt database (UP000000625) using TopPIC suite (version 1.2.6).[37,38] The identified PrSMs and proteoforms were filtered by a 0.1% FDR and a 0.5% FDR, respectively. The experimental details are described in the **Supporting Information I**.

The migration time (t_M) of each identified proteoform was obtained from the database search result. The number of charge (Q) of each proteoform equals the number of positively charged amino acid residues within their sequences (K, R, H, and N-terminus). The molecular mass (M) of each proteoform equals the adjusted mass reported by the TopPIC. The length (N) of each proteoform equals the number of amino acid residues within the sequence. Only proteoforms without post-translational modifications (PTMs) were used for calculation of experimental μ_{ef} and predicted μ_{ef}.

About 500-1100 proteoforms were used for the calculations. The molecular mass of proteoforms ranged from 1.5 kDa to 30 kDa. We also assumed that the electroosmotic flow (EOF) in an LPA-coated capillary with an acidic BGE was extremely low.[27] The proteoforms with their experimental and predicted μ_{ef} are listed in the **Supporting Information II**. The MS raw data have been deposited to the ProteomeXchange Consortium via the PRIDE [39] partner repository with the data set identifier PXD017265.

First, we calculated the experimental μ_{ef} using the Equation (2),

$$\text{Experimental } \mu_{ef} = L/((30-2)/L*t_M) \text{ (unit of } cm^2\ kV^{-1}s^{-1}) \qquad \text{Equation (2)}$$

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

Second, the predicted μ_{ef} of proteoforms from the *E. coli* datasets were calculated using six classical semi-empirical models,[14-16,18-20] **Table 1**. For the Cifuentes's model, we obtained the final equation (3) based on the equation (1) via omitting the prefactor 900.

$$\mu_{ef} = \ln(1+0.35*Q)/M^{0.411} \qquad \text{Equation (3)}$$

Where Q and M are the number of charge and molecular mass of each proteoform.

The Cifuentes's model produced the best linear correlation (R^2: 0.97-0.98) between the predicted and experimental μ_{ef} of proteoforms according to the R^2 values for the three CZE conditions, followed by the Offord's model (R^2: 0.92-0.94) and Kim's model (R^2: 0.82-0.90). The Reynolds's model generated the lowest correlation coefficient (R^2: 0.52-0.72). The Cifuentes's model obtained a drastically better linear correlation regarding the R^2 value than the Grossman's model (0.97 *vs.* 0.76 for the 5%AA BGE) and the two models have two differences, $M^{0.411}$ *vs.* $N^{0.435}$ and 0.35*Q *vs.* Q. After a more detailed study using the 5%AA BGE data, we figured out that the R^2 value of the Grossman's model could be boosted from 0.76 to 0.94 by simply changing the Q to 0.35*Q. Only a minor effect on the R^2 value was observed by changing $N^{0.435}$ to $M^{0.411}$. We note that the slopes of the linear correlation curves from the two best models (the Cifuentes's model and the Offord's model) are comparable for the different CZE conditions, *e.g.,* 0.22 *vs.* 0.25 for the 5%AA BGE, and are obviously smaller than that from other models,

suggesting that the predicted μ_{ef} from these two models are much smaller than that from other four models and significantly smaller than the experimental μ_{ef}. We can add a CZE condition-dependent prefactor to the Cifuentes's model to match the predicted and experimental μ_{ef}.

The data here represents the first try of predicting μ_{ef} of proteoforms using large-scale top-down proteomics datasets. The great correlation between experimental μ_{ef} and predicted μ_{ef} from the simple Cifuentes's model further implies that the μ_{ef} of proteoforms in CZE can be predicted easily. The predicted μ_{ef} of proteoforms discussed in the following parts were obtained from the Cifuentes's model.

We evaluated how the BGE of CZE influenced the μ_{ef} of proteoforms, **Figure 1A**. When the AA concentration in BGE increased from 5% to 20% and when 10% (v/v) IPA and 15% (v/v) DMA were added into the BGE, the experimental μ_{ef} of proteoforms decreased. Two possible reasons exist for that phenomenon. First, the lower pH of 20% (v/v) AA and the organic solvents unfold the proteoforms more completely, enlarging the size of proteoforms and reducing their mobility. It has been reported recently that in CZE protein size can increase significantly due to unfolding when the pH of BGE decreases.[40] Second, the lower pH of 20% (v/v) AA and the organic solvents further eliminate the residual EOF in the capillary. In addition, when 20% (v/v) AA with or without 10% (v/v) IPA and 15% (v/v) DMA was used as the BGE, a better linear correlation was observed compared to the 5% (v/v) AA (0.98 *vs.* 0.96). For the BGE containing 20% (v/v) AA, 10% (v/v) IPA, and 15% (v/v) DMA, the absolute value of predicted μ_{ef} is much closer to that of experimental μ_{ef} compared to the other two BGEs, indicated by the much larger slope of the linear correlation curve (0.51 *vs.* 0.20-0.25). The number of outliers from the BGE containing IPA and DMA is also much smaller compared to the other BGEs. The results suggest that adding some organic solvents to the BGE of CZE could benefit the prediction of μ_{ef} of proteoforms. There is also some evidence in the literature. For instance, in 2000, Katayama *et al.* demonstrated that the use of methanol in BGE could improve the correlation between predicted μ_{ef} and experimental μ_{ef} of peptides.[41] We speculate that the organic solvents (IPA and DMA) in the BGE facilitate the complete unfolding of proteoforms, leading to better prediction of

their μ_{ef}. It has been reported that certain types of polar solvents such as dimethyl sulfoxide (DMSO), dimethylformamide (DMF), and formamide have the ability to unfold proteins. [42,43]

We then tested the Cifuentes's model on our published zebrafish brain (optic tectum (Teo)) data and evaluated the performance of the model for predicting μ_{ef} of proteoforms with certain PTMs (*i.e.,* N-terminal acetylation and phosphorylation). When we only used nonmodified proteoforms, the predicted μ_{ef} and experimental μ_{ef} showed reasonably good linear correlations (R^2=0.96). We then further included the proteoforms with N-terminal acetylation and/or phosphorylation in the analysis. The zebrafish Teo data from one CZE-MS/MS run was used, which included 1163 nonmodified proteoforms, 92 proteoforms with only N-terminal acetylation, 3 proteoforms with one phosphorylation site, and 2 proteoforms with both N-terminal acetylation and one phosphorylation site. N-terminal acetylation and phosphorylation can reduce the proteoforms' charge by one charge unit in theory. **Figure 1B** shows the linear correlation between the experimental and predicted μ_{ef} for these post-translationally modified proteoforms (97 in total) regardless of the PTMs. First, the linear correlation is poor (R^2=0.76). Second, it is clear that the addition of one acetylation modification or one phosphoryl group to a proteoform can decrease its mobility significantly. After considering the effect of these PTMs on the proteoforms' charge, we corrected the charge (Q) in the Cifuentes's model. We achieved a linear correlation for the 97 proteoforms with PTMs (R^2=0.92) after we adjusted the Q by -1, -1 and -2 for proteoforms with N-terminal acetylation, proteoforms with one phosphorylation site, and proteoforms with both N-terminal acetylation and phosphorylation, respectively, **Figure 1C**. The results show that the proteoforms' charge shifts are very close to the theoretical contributions of N-terminal acetylation and phosphorylation. Additionally, the results suggest that the μ_{ef} of proteoforms with N-terminal acetylation and phosphorylation could be predicted as accurately as nonmodified proteoforms (R^2 0.92 *vs.* 0.96). We note that some outliers exist in Figure 1C due to two possible reasons. First, for these outliers, their experimental μ_{ef} values are larger than the predicted values, most likely due to the incomplete unfolding of these proteoforms in the BGE used in the experiment

(10% (v/v) AA, pH 2.2). Second, since the proteoform IDs were filtered by a 0.5% FDR, some of the outliers could be simply the wrong proteoform IDs.

In summary, in this work, for the first time, we evaluated various semi-empirical models for predicting proteoforms' μ_{ef} using large-scale top-down proteomics datasets. Using a simple semi-empirical model, we achieved a linear correlation between experimental μ_{ef} and predicted μ_{ef} of *E. coli* proteoforms (R^2=0.98). We note that some effort has been made on predicting retention time of proteins in RPLC using simple protein mixtures based on complicated models, producing reasonable correlations between predicted and experimental retention time (R^2=0.86-0.90).[11,44,45] We also note that our current study still has some limitations. First, the proteoforms used in this study have masses lower than 30 kDa. Top-down proteomics datasets of large proteoforms using CZE-MS/MS are required to expand the model into a wider range of proteoforms in mass. Second, the number of proteoforms with PTMs (*i.e.,* acetylation and phosphorylation) used here is small, less than 100. Larger numbers of proteoforms with PTMs are extremely important for improving the model for post-translationally modified proteoforms.

Acknowledgments

We thank Prof. Heedeok Hong's group at the Department of Chemistry of Michigan State University for kindly providing the *E. coli* cells for this project. We thank the support from the National Science Foundation (CAREER Award, DBI-1846913) and the National Institutes of Health (R01GM125991).

Supporting Information

Supporting information I. Experimental procedures (Docx)

Supporting information II. Lists of proteoforms used in the study from *E. coli* or zebrafish brain under different CZE conditions with experimental and predicted electrophoretic mobility (XLSX)

References

[1] Toby TK.; Fornelli L.; Kelleher NL Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem*. **2016**; *9(1)*: 499-519.

[2] Tran JC.; Zamdborg L.; Ahlf DR.; Lee JE.; Catherman AD.; Durbin KR.; Tipton JD.; Vellaichamy A.; Kellie JF.; Li M.; Wu C.; Sweet SM.; Early BP.; Siuti N.; LeDuc RD.; Compton PD.; Thomas PM.; Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*. **2011**; *480(7376)*: 254-8.

[3] Chen B.; Brown KA.; Lin Z.; Ge Y. Top-Down Proteomics: Ready for Prime Time? *Anal Chem.* **2018**; *90(1)*:110-127.

[4] Smith LM.; Kelleher NL.; Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat Methods*. **2013**;*10(3)*:186-7.

[5] Schaffer LV.; Millikin RJ.; Miller RM.; Anderson LC.; Fellers RT.; Ge Y.; Kelleher NL.; LeDuc RD.; Liu X.; Payne SH.; Sun L.; Thomas PM.; Tucholski T.; Wang Z.; Wu S.; Wu Z.; Yu D.; Shortreed MR.; Smith LM. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics*. **2019**;*19(10)*: e1800361.

[6] Keller A.; Nesvizhskii AI.; 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(20)*: 5383-92.

[7] Elias JE.; Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. **2007**; *4(3)*: 207-14.

[8] LeDuc RD.; Fellers RT.; Early BP.; Greer JB.; Shams DP.; Thomas PM.; Kelleher NL. Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. *Mol Cell Proteomics*. **2019**;*18(4)*: 796-805.

[9] Henneman AA.; Palmblad M. Retention time prediction and protein identification. *Methods Mol Biol*. **2013**; *1007*: 101-18.

[10] Strittmatter EF.; Kangas LJ.; Petritis K.; Mottaz HM.; Anderson GA.; Shen Y.; Jacobs JM.; Camp DG 2nd.; Smith RD. Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. *J Proteome Res*. **2004**; *3(4)*: 760-9.

[11] Tarasova IA.; Masselon CD.; Gorshkov AV.; Gorshkov MV. Predictive chromatography of peptides and proteins as a complementary tool for proteomics. *Analyst*. **2016**;*141(16)*: 4816-4832.

[12] Smith RD.; Anderson GA.; Lipton MS.; Pasa-Tolic L.; Shen Y.; Conrads TP.; Veenstra TD.; Udseth HR. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics. **2002**; *2(5)*:513-23.

[13] Mittermayr S.; Olajos M.; Chovan T.; BonnA G.K.; Guttman A. Mobility modeling of peptides in capillary electrophoresis. *Trends Analyt Chem*. **2008**; *27(5)*: 407-417.

[14] Offord RE. Electrophoretic mobilities of peptides on paper and their use in the determination of amide groups. *Nature*. **1966**; *211(5049)*: 591-3.

[15] Tanford C. Physical Chemistry of Macromolecules; Wiley: New York, U.S.A., **1961**.

[16] Kim J.; Zand R.; Lubman D. M. Electrophoretic mobility for peptides with post-translational modifications in capillary electrophoresis. *Electrophoresis.* **2003**; *24*, 782–793.

[17] Compton BJ. Electrophoretic mobility modeling of proteins in free zone capillary electrophoresis and its application to monoclonal antibody microheterogeneity analysis. *Journal of Chromatography A*. **1991**; *559(1-2)*: 357-366.

[18] Grossman PD.; Colburn JC.; Lauer HH. A Semiempirical Model for the Electrophoretic Mobilities of Peptides in Free-Solution Capillary Electrophoresis. *Anal. Biochem*. **1989**; *179:* 28–33.

[19] Cifuentes A.; Poppe H. Simulation and optimization of peptide separation by capillary electrophoresis. *J Chromatogr A*. **1994**; *680(1)*: 321-40.

[20] Adamson NJ.; Reynolds EC. Rules relating electrophoretic mobility, charge and molecular size of peptides and proteins. *J. Chromatogr., Biomed. Appl*. **1997**; *699:* 133–147.

[21] Krokhin OV.; Anderson G.; Spicer V.; Sun L.; Dovichi NJ. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem*. **2017**; *89:* 2000–2008.

[22] Chen D.; Ludwig KR.; Krokhin OV.; Spicer V.; Yang Z.; Shen X.; Hummon AB.; Sun L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Large-Scale Phosphoproteomics with the Production of over 11,000 Phosphopeptides from the Colon Carcinoma HCT116 Cell Line. *Anal Chem*. **2019**;*91(3)*:2201-2208.

[23] Chae KS.; Lenhoff AM. Computation of the electrophoretic mobility of proteins. *Biophys J*. **1995**; *68(3)*: 1120–1127.

[24] Rickard EC.; Strohl MM.; Nielsen RG. Correlation of Electrophoretic Mobilities from Capillary Electrophoresis with Physicochemical Properties of Proteins and Peptides. *Anal Biochem*. **1991**;*197(1)*:197-207.

[25] Shen X.; Yang Z.; McCool EN.; Lubeckyj RA.; Chen D.; Sun L. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends Analyt Chem*. **2019**;*120*. pii: 115644.

[26] Gomes FP.; Yates JR 3rd. Recent trends of capillary electrophoresis-mass spectrometry in proteomics research. *Mass Spectrom Rev.* **2019**;*38*(6):445-460.

[27] Lubeckyj RA.; McCool EN.; 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(22)*:12059-12067.

[28] McCool EN.; Lubeckyj RA.; 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(9)*:5529-5533.

[29] Lubeckyj RA.; Basharat AR.; 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(8)*:1435-1445.

[30] McCool EN.; Lodge JM.; Basharat AR.; Liu X.; Coon JJ.; Sun L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry with Activated Ion Electron Transfer Dissociation for Large-scale Top-down Proteomics. *J Am Soc Mass Spectrom*. **2019**.

[31] Shen X.; Kou Q.; Guo R.; Yang Z.; Chen D.; Liu X.; Hong H.; Sun L. Native Proteomics in Discovery Mode Using Size-Exclusion Chromatography-Capillary Zone Electrophoresis-Tandem Mass Spectrometry. *Anal Chem*. **2018**; *90(17)*: 10095-10099.

[32] Zhao Y.; Sun L.; Zhu G.; Dovichi NJ. 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(10)*: 3679-3685.

[33] Wojcik R.; Dada OO.; Sadilek M.; Dovichi NJ. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Commun Mass Spectrom*. **2010**; *24(17)*: 2554-60.

[34] Sun L.; Zhu G.; Zhang Z.; Mou S.; Dovichi NJ. 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(5)*: 2312-21.

[35] Zhu G.; Sun L.; Dovichi NJ. 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*: 839-43.

[36] Britz-McKibbin P.; Chen DD. Selective focusing of catecholamines and weakly acidic compounds by capillary electrophoresis using a dynamic pH junction. *Anal Chem*. **2000**;*72(6)*:1242-52.

[37] Kou Q.; Xun L.; Liu X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*. **2016**; *32(22)*: 3495-3497.

[38] Liu X.; Inbar Y.; Dorrestein PC.; Wynne C.; Edwards N.; Souda P.; Whitelegge JP.; Bafna V.; Pevzner PA. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol Cell Proteomics*. **2010**; *9(12)*: 2772-82.

[39] Perez-Riverol Y.; Csordas A.; Bai J.; Bernal-Llinares M.; Hewapathirana S.; Kundu DJ.; Inuganti A.; Griss J.; Mayer G.; Eisenacher M.; Pérez E.; Uszkoreit J.; Pfeuffer J.; Sachsenberg T.; Yilmaz S.; Tiwary S.; Cox J.; Audain E.; Walzer M.; Jarnuczak AF.; Ternent T.; Brazma A.; Vizcaíno JA. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**; *47(D1):* D442-D450.

[40] Zhang W.; Wu H.; Zhang R.; Fang X.; Xu W. Structure and effective charge characterization of proteins by a mobility capillary electrophoresis based method. *Chem Sci*. **2019**; *10(33)*: 7779-7787.

[41] Katayama H.; Ishihama Y.; Oda Y.; Asakawa N. Electrophoretic mobility-assisted identification of proteins by nanoelectrospray capillary electrophoresis/mass spectrometry under methanolic conditions. *Rapid Commun Mass Spectrom*. **2000**;*14(14)*: 1167-78.

[42] Knubovets T.; Osterhout JJ.; Klibanov AM. Structure of lysozyme dissolved in neat organic solvents as assessed by NMR and CD spectroscopies. *Biotechnol Bioeng*. **1999**; *63(2)*: 242-8.

[43] Mattos C.; Ringe D. Proteins in organic solvents. *Curr Opin Struct Biol*. **2001**;*11(6)*: 761-4.

[44] Champney WS. Reversed-phase chromatography of Escherichia coli ribosomal proteins. Correlation of retention time with chain length and hydrophobicity. *J Chromatogr*. **1990**; *522*: 163-70.

[45] Pridatchenko ML.; Perlova TY.; Ben Hamidane H.; Goloborodko AA.; Tarasova IA.; Gorshkov AV.; Evreinov VV.; Tsybin YO.; Gorshkov MV. On the utility of predictive chromatography to complement mass spectrometry based intact protein identification. *Anal Bioanal Chem*. **2012**; *402(8)*: 2521-9.

Table 1. Summary of the linear correlations between experimental μ_{ef} and predicted μ_{ef} of *E. coli* proteoforms using different semi-empirical models and under various CZE conditions.*

Semi-empirical model		BGE					
		5% (v/v) AA		20% (v/v) AA		10% (v/v) IPA 15% (v/v) DMA 20% (v/v) AA	
		R^2	Slope	R^2	Slope	R^2	Slope
$\ln(1+0.35*Q)/M^{0.411}$	**Cifuentes and Poppe [19,21]**	**0.97**	**0.22**	**0.98**	**0.26**	**0.98**	**0.51**
$\ln(1+Q)/N^{0.435}$	Grossman *et al.* [18]	0.76	1.72	0.82	2.1	0.82	4.4
$Q/M^{2/3}$	Offord [14]	0.93	0.25	0.94	0.29	0.92	0.58
$Q/M^{0.56}$	Kim *et al.* [16]	0.90	0.65	0.89	0.74	0.82	1.4
$Q/M^{1/2}$	Tanford [15]	0.86	1.1	0.84	1.2	0.74	2.3
$Q/M^{1/3}$	Reynolds *et al.* [20]	0.72	4.6	0.69	5.2	0.52	9.0

* Only proteoforms without PTMs were used. The R^2 and slope values were from the mean of the triplicate CZE-MS/MS runs, and the standard deviations of the R^2 values from the triplicate analyses were about 0.01.

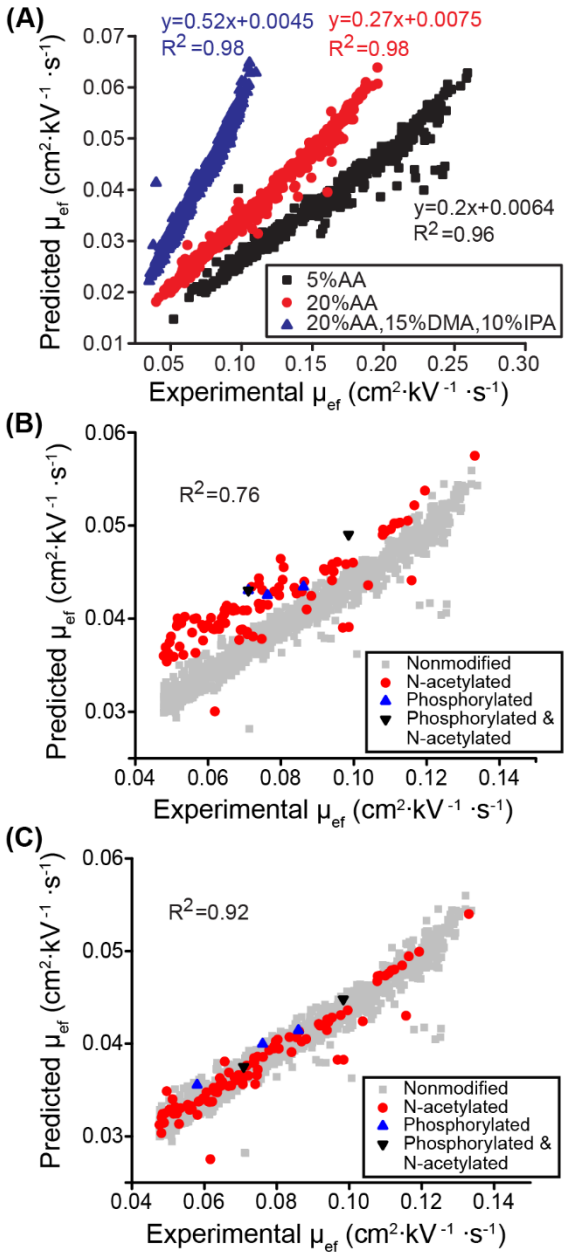
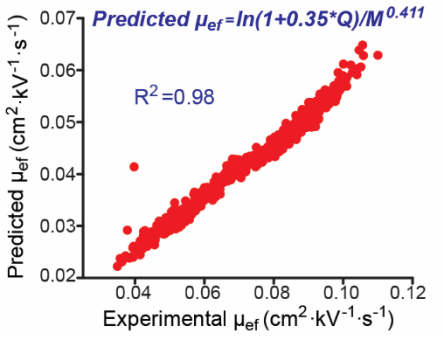


Figure 1. Linear correlations between predicted μ_{ef} and experimental μ_{ef} of proteoforms from *E. coli* cells under various CZE conditions (A) and proteoforms from zebrafish optic tectum (TEO) (B, C). For (A), only nonmodified proteoforms were used, and the data was from a single CZE-MS/MS run. For (B) and (C), nonmodified, N-terminal acetylated, and mono-phosphorylated proteoforms were employed. In (B), the charge of proteoforms in the BGE (Q) was calculated by counting the positively charged amino acid residues (K, R, H, and N-terminal) regardless of the PTMs. In (C), the charge of proteoforms (Q) was corrected based on their PTMs. For example, one charge reduction corresponded to one N-terminal acetylation or one phosphorylation.

For TOC only:


Predicted μ_{ef}=ln(1+0.35*Q)/M$^{0.411}$
R^2=0.98
Predicted μ_{ef} ($cm^2 \cdot kV^{-1} \cdot s^{-1}$)
Experimental μ_{ef} ($cm^2 \cdot kV^{-1} \cdot s^{-1}$)
0.07
0.06
0.05
0.04
0.03
0.02
0.04
0.06
0.08
0.10
0.12