TIMScore with PaSER: Exploiting the CCS-dimension

Precise and accurate validation of peptide-spectrum matches.

TIMScore empowers CCS-enabled peptide identifcation

Trapped ion mobility spectrometry (TIMS) provides major advancements in proteomics, simultaneously providing sensitivity, selectivity and speed to proteomics research. With each analyte measured, so is a collisional cross section (CCS) value. A CCS value is an intrinsic property unique to each analyte that is highly reproducible across different instruments and laboratories. TIMScore™ utilizes this CCS value in a database search algorithm boosting the number of peptide spectrum matches (PSMs), peptides and proteins identified in bottom-up proteomics measurements and is particularly sensitive toward phosphopeptide identification.

Challenge

One of the core challenges that exists in proteomics is converting complex LC-MS/MS datasets into tangible peptide spectrum matches (PSMs) and subsequently peptide identifications that can be used for protein inference, quantification, PTM analysis, and proteome sequence coverage within complex

Keywords: Software, PaSER, proteomics, CCS, ion mobility, database search, PTMs

Authors: Christopher Adams, Robin Park, Patrick Garrett, Sven Brehmer, Sebastian Wehner, Jonathan Krieger, Tharan Srikumar; Bruker Daltonics GmbH & Co. KG, Bremen, Germany.

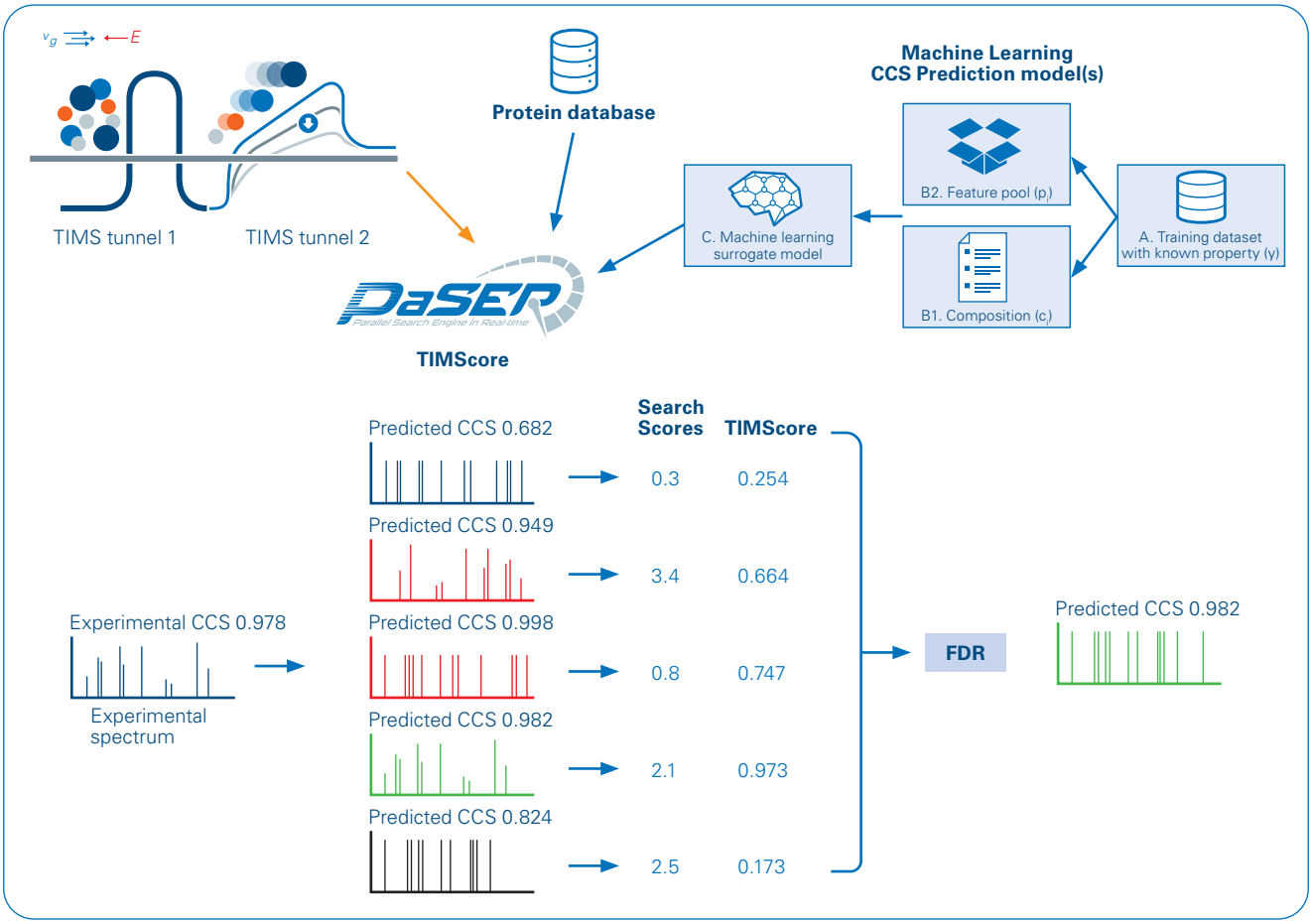

Figure 1: A CCS-enabled database search including TIMScore as an additional dimension. The trained machine learning model predicts the CCS values of tryptic and phosphorylated peptides. Experimental CCS is referenced against the predicted CCS to call the most probable assignment when a traditional search incorrectly, incompletely or is unable to assign a correct peptide match.

samples. Database search algorithms have extensively transformed biological and medical research yet, as instrumentation continues to become more sensitive, there remains room to improve these search algorithms.

In the simplest scenario database search algorithms rely on precursor and fragment ion spectra to be matched in-silico suggesting a best fit and assigning a probability score. In many instances the best fit of a particular PSM is of an equal probability score or just marginally better than the next best fit, yet a single assignment is delivered. Inferring only one PSM, which may be incorrect given the reliance purely on the information contained in a fragment ion spectra, can lead to increased false identifications and necessitate the need for stricter acceptance criteria such as lower false discovery rates, minimum increases in unique peptide counts, and more biological replicates. Incorporating CCS information via TIMScore: 1) assigns more PSMs, peptides and proteins 2) increases the confidence in the assignments 3) can be performed in real-time within PaSER™ 2022 to provide the most complete real-time feedback (Figure 1).

Solution

Machine Learning to Predict CCS

An essential component to TIMScore is defining the deviation between experimental and predicted CCS values. In order to accurately predict CCS values from a peptide's primary amino acid sequence machine learning was used. A training dataset of hundreds of thousands of tryptic and phosphorylated peptides was used, where the dataset included peptides of doubly, triply and quadruply charge states. A transformer model of peptide CCS was developed from this training set. The model was tested for accuracy against an independent dataset it had previously not seen. For doubly, triply and quadruply charged peptides the accuracy in predicting a peptide CCS from the primary amino acid sequence was 95% for tryptic peptides and 92% for phosphorylated tryptic peptides, respectively (Figure 2).

Applying the CCS-enabled Algorithm

With a CCS prediction model applicable to all tryptic and phosphorylated

peptides complete, TIMScore is deployed. Upon setting up the parameters file, in-silico peptide candidates are sent to the CCS prediction model to generate a predicted CCS value. The PaSER search algorithm is run as normal and the search algorithm compares the predicted and measured CCS values and calculates a correlation score, namely TIMScore for the top 5 peptide candidates for each spectra (Figure 1). The true benefit of TIMScore can be realized during the peptide-validation and False Discovery Rate (FDR) estimation steps of the proteomics pipeline. In a non CCS enabled algorithm, only two dimensions can be utilized to estimate the FDR rate, and so a discriminate line is fit to a 1% error (Figure 3A) to distinguish forward and reverse peptide candidates. With TIMScore, and the extra CCS dimension, the peptide-candidates can be vectorized in 3-dimensions (Figure 3B) allowing a discriminate contoured plane to be applied to achieve the same 1% error. Applying a discriminate plane provides increased accuracy and precision, helping to validate formerly poorly scoring PSMs in the standard two dimensions. Thus, the key effect of TIMScore is derived from the additional dimension of CCS and that it provides in assigning true positives from decoy peptide sequences as shown in Figure 3. TIMScore works in a bidirectional fashion, boosting the confidence of borderline peptides under strict FDR thresholds while simultaneously lowering the probability score of a peptide candidate such that it falls below the level of detection. Additionally, the probability score differentiates ambiguous PSMs where the traditional search score cannot distinguish between the 1st and 2nd (or more) best candidates.

Results

As a demonstration of the real-world application we present a previously published dataset analyzed with and without TIMScore for PSM, peptide and protein identification, as well as PTM localization.

TIMScore increases Proteins, Peptides, PSMs and Sequence Coverage

The published data set from the laboratory of Prof. Yasushi Ishihama titled "Effect of Phosphorylation on the Collision Cross Sections of Peptide Ions in Ion Mobility Spectrometry" was accessed from jPOST repository with the identifier PXD019746 [Ogata 2021]. In the paper, they systematically characterize the CCS values of 4433 pairs of mono-phosphopeptide and their corresponding unphosphorylated peptide ions using the timsTOF Pro. Interestingly, within this dataset one third of the enriched phosphopeptide pool was purposely dephosphorylated, further challenging the TIMScore CCS prediction model. We analyzed this dataset with and without TIMScore on the PaSER platform and also compared this to the published results which used an alternative search engine. Within Figure 4, we describe the number of PSMs, peptides, and proteins (Figure 4A&C) with, and without TIMScore. The use of TIMScore adds more than 110,000

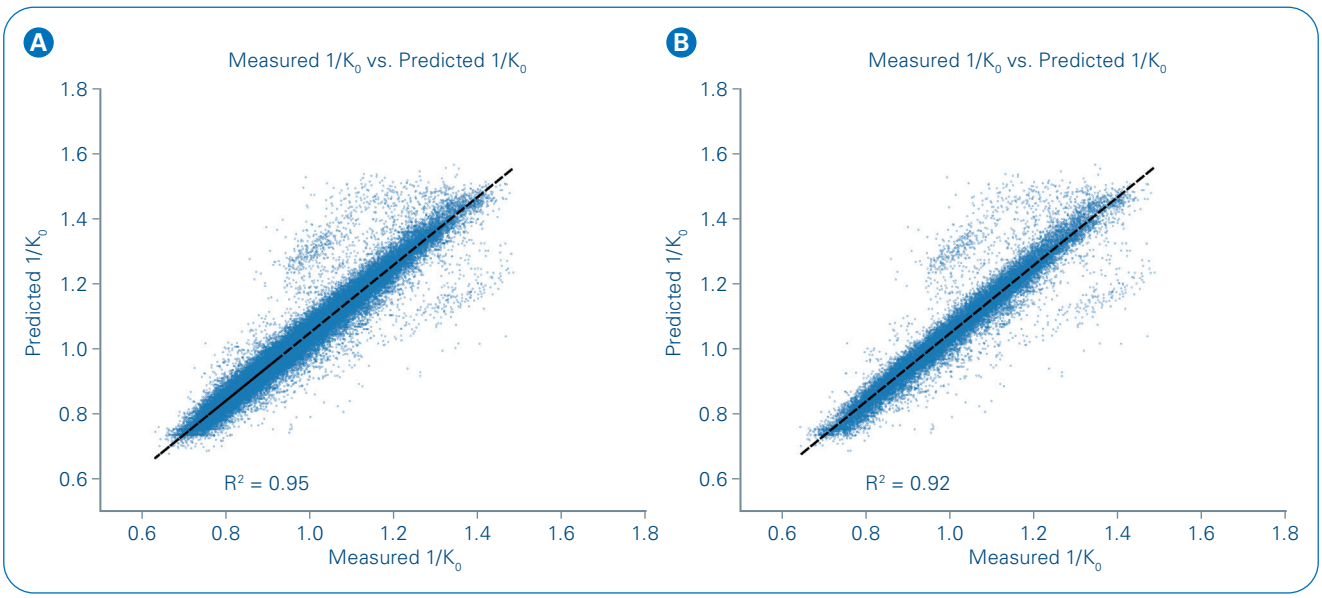

Figure 2: Scatter plots of the predicted ion mobility (CCS) values from the machine learned model and the experimentally derived values for (A) tryptic and (B) phosphorylated peptides.

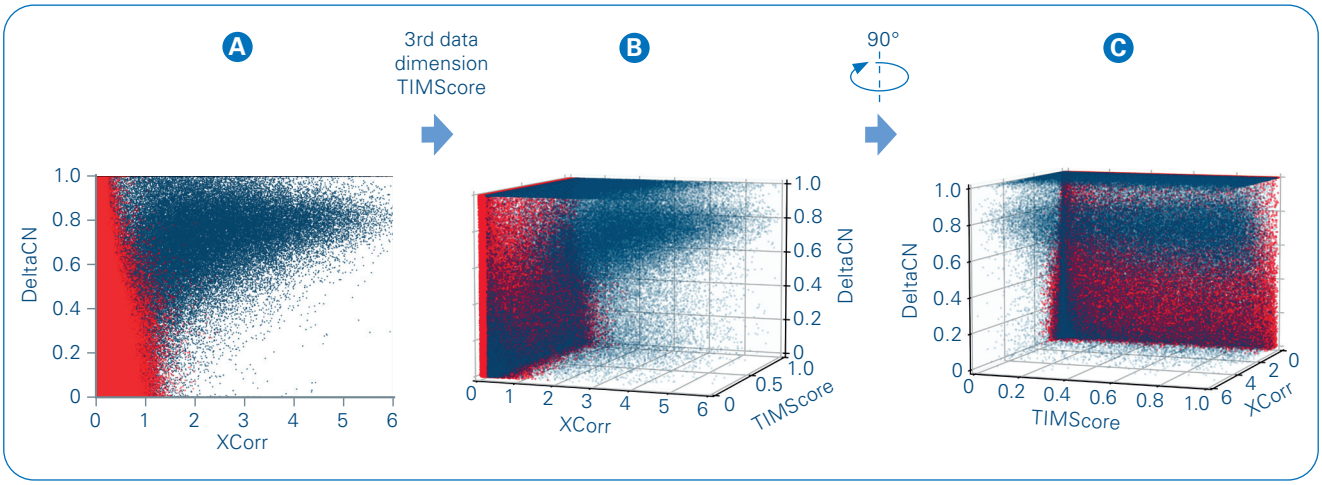

Figure 3: (A) 2D plot representing XCorr and DeltaCN. Red dots represent decoy peptides, and the blue dots represent matched sequences. XCorr is a cross correlation between experimental and predicted fragment ion spectra and whereas DeltaCN is a measure of specificity, describing how much better the assigned fit is as compared to the next best fit. To determine false discovery rate (FDR), a line is fit on the 2D plot that separates forward and reverse peptides with 1% error. (B) A 3-dimensional box plot representing XCorr, DeltaCN and TIMScore and (C) a counter-clockwise 90-degree rotation around the y-axis. TIMScore on the z-axis is the correlation between experimental and predicted CCS. To determine FDR with the addition of TIMScore, we visualize a contoured plane that separates forward and reverse sequences with 1% error. This allows for much more precise and accurate validation of peptide-spectrum matches.

PSMs and doubles the number of peptides observed from 42,930 to 98,949. The >98,000 peptides observed for this dataset is a 3.5 times increase compared to what was initially published (Figure 4B). Interestingly, TIMScore does boost the number of proteins identified (Figure 4C), however, the biggest contribution comes in a significant boost to protein sequence coverage. This helps provide quantitative accuracy and more expansive libraries for label free quantitation (LFQ), data independent acquisition (DIA) and parallel reaction monitoring (PRM) experiments. In these data the number of peptides identified per protein group was 15, 7.5 and 5.5 using TIMScore, without TIMScore or published data set, respectively. The addition of nearly 10 peptides by TIMScore per protein observed is currently being tested to the extended applications of dia-PASEF® and prm-PASEF®.

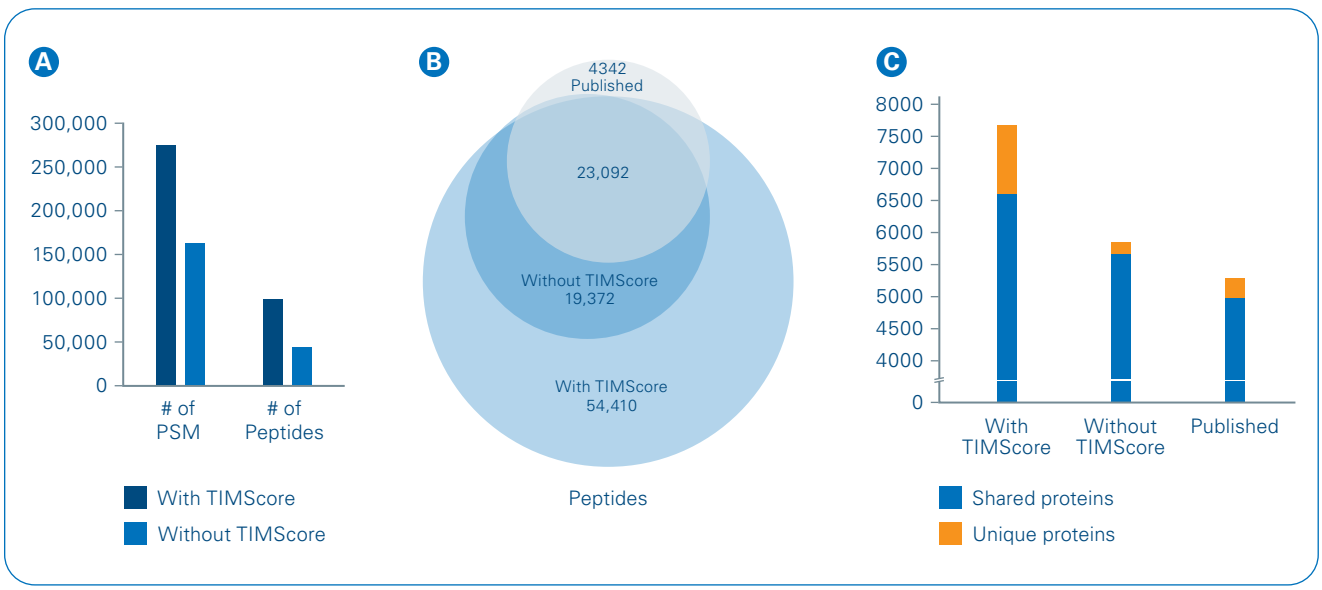

Figure 4: (A) Bar graph of the number of identified peptide-spectrum-matches in PaSER 2022 using the ProluCID engine with and without TIMScore and the CCS dimension, TIMScore identifies an additional >100,000 PSMs. Venn diagrams displaying (B) all peptide sequences and (C) protein groups identified. All results are from the published dataset PXD019746 as processed and presented without TIMScore, with TIMScore and as in the published research article.

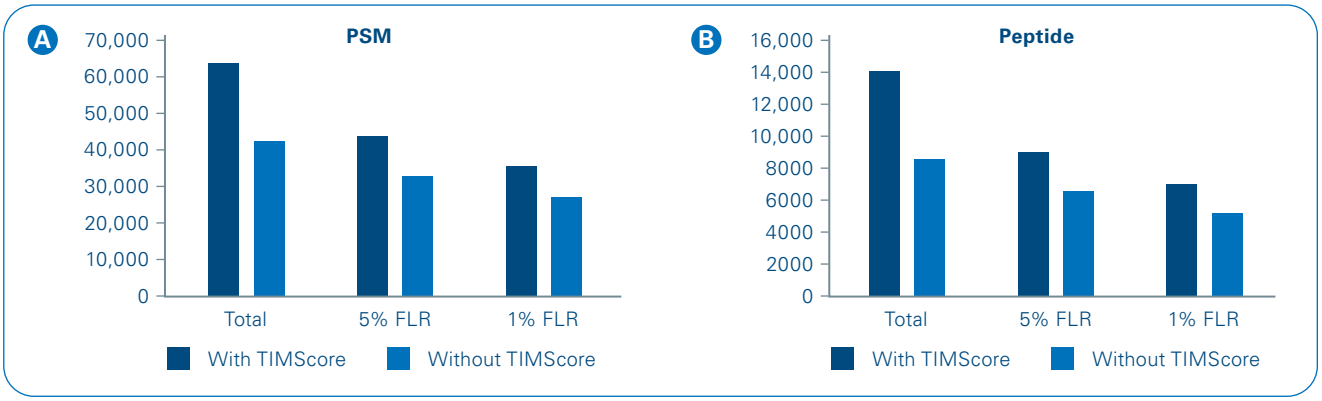

Figure 5: Bar charts of (A) Phosphorylated PSMs in PaSER 2022 using the ProLuCID search engine with and without TIMScore enabled. (B) Phosphorylated peptides identified using the ProLuCID search engine with and without TIMScore enabled. Phosphorylated PSM and Peptide sequences were further filtered applying a False Localization Rate (FLR) using LuciPHOr [Fermin et al. 2013] where even at the strictest FLR (1%) TIMScore identifies 57% more phosphorylated PSMs and Peptides.

TIMScore increases Phosphopeptide Identification and Phosphosite Localization

Having been trained on phosphopeptide data, TIMScore increases the sensitivity to detect phosphorylated peptides. In Figure 5A and B we demonstrate that TIMScore improves the number of both phosphorylated PSMs and peptides identified by on average 61% in the current dataset. Important in post translationally modified (PTM) proteins, or the epiproteome, is the ability to localize the site of the modification. Site localization can be difficult depending on the PTM present because of the lability of the PTM, or its preferential dissociation. Upon dissociation phosphopeptides often display a neutral loss, or the loss of the phosphate group with little backbone bond fragmentation which provides a deterministic location of the phospho group. Tools to assess the number and accuracy of site localization have been developed over the past 30 years, one such tool that generates a false localization rate (FLR) is LuciPHOr [Fermin 2013]. The FLR is the percent confidence interval of correct assignment of phosphorylated peptides. We applied LuciPHOr at a FLR of both 5% and the most stringent 1% intervals to the global phosphopeptide PSMs and peptides identified (Table 1). Consistent with the increased sensitivity TIMScore provides, an improvement of 57% (10,854 PSMs) in the number of localized phosphopeptide sites was observed at the 5% FLR cutoff. These percent improvements using TIMScore held consistent (57-62%) across both PSMs and peptides at both the 5% and 1% FLR settings. As in the publication, we also examined the number of phosphorylated peptides and their corresponding unphosphorylated peptide pairs as well as peptide ion pairs. In this analysis as well, TIMScore provided >55% increase compared to standard method (Table 1).

Table 1: Phosphorylated PSM and Peptide identifications with and without TIMScore and as filtered using a Phosphorylation False Localization Rate (FLR) of 5% and 1%. Unique peptide pairs are phosphorylated peptides and their corresponding unphosphorylated peptide. Unique peptide ion pairs are one or more phosphorylated peptides (or multi-phosphorylated) peptide and their corresponding unphosphorylated peptide sequence.

	PSM			Peptides			Unique Peptide Pairs	Unique Peptide Ion Pairs
	Total	5% FLR	1% FLR	Total	5% FLR	1% FLR		
Without TIMScore	42,427	33,065	27,095	8570	6594	5188	6038	8542
With TIMScore	63,823	43,919	35,638	13,993	8996	6939	9399	13,996
% GAIN	60%	57%	57%	62%	58%	57%	56%	64%

Conclusion

TIMScore on PaSER:

- Boosts the number of proteins, peptides and PSMs in complex datasets
- Increases protein sequence coverage substantially
- Provides an additional dimension for more precise and accurate peptide assignments
- Expands capabilities for library-based approaches in quantitative proteomics

Learn More

You are looking for further Information?
Check out the link or scan the QR code.

www.bruker.com/timstofpro

References

[1] Ogata K, Chang CH, Ishihama Y (2010). *Effect of Phosphorylation on the Collision Cross Sections of Peptide Ions in Ion Mobility Spectrometry.* Mass Spectrom (Tokyo);**10**:A0093. doi: 10.5702/massspectrometry.A0093. Epub 2021 Jan 30. PMID: 33552826; PMCID: PMC7843839.

[2] Fermin D, Walmsley SJ, Gingras AC, Choi H, Nesvizhskii AI (2013). *LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach.* Mol Cell Proteomics. Nov;**12**(11):3409-19. doi: 10.1074/mcp.M113.028928. Epub 2013 Aug 5. PMID: 23918812; PMCID: PMC3820951.

For Research Use Only. Not for use in clinical diagnostic procedures.

Bruker Daltonics GmbH & Co. KG

Bremen · Germany
Phone +49 (0)421-2205-0

Bruker Scientific LLC

Billerica, MA · USA
Phone +1 (978) 663-3660

ms.sales.bdal@bruker.com – www.bruker.com