DeepCCS: Context-aware prediction of peptide collisional cross section increases peptide recovery for out-of-distribution data-independent acquisition (DIA) datasets

BIOGNOSYS
NEXT GENERATION PROTEOMICS

George Rosenberger[1], **Ignacio Jáuregui Novo**[2], **Alexandros Pachos**[3], **An-phi Nguyen**[3], **Tejas Gandhi**[3], **Dennis Trede**[4], **Lukas Reiter**[3]

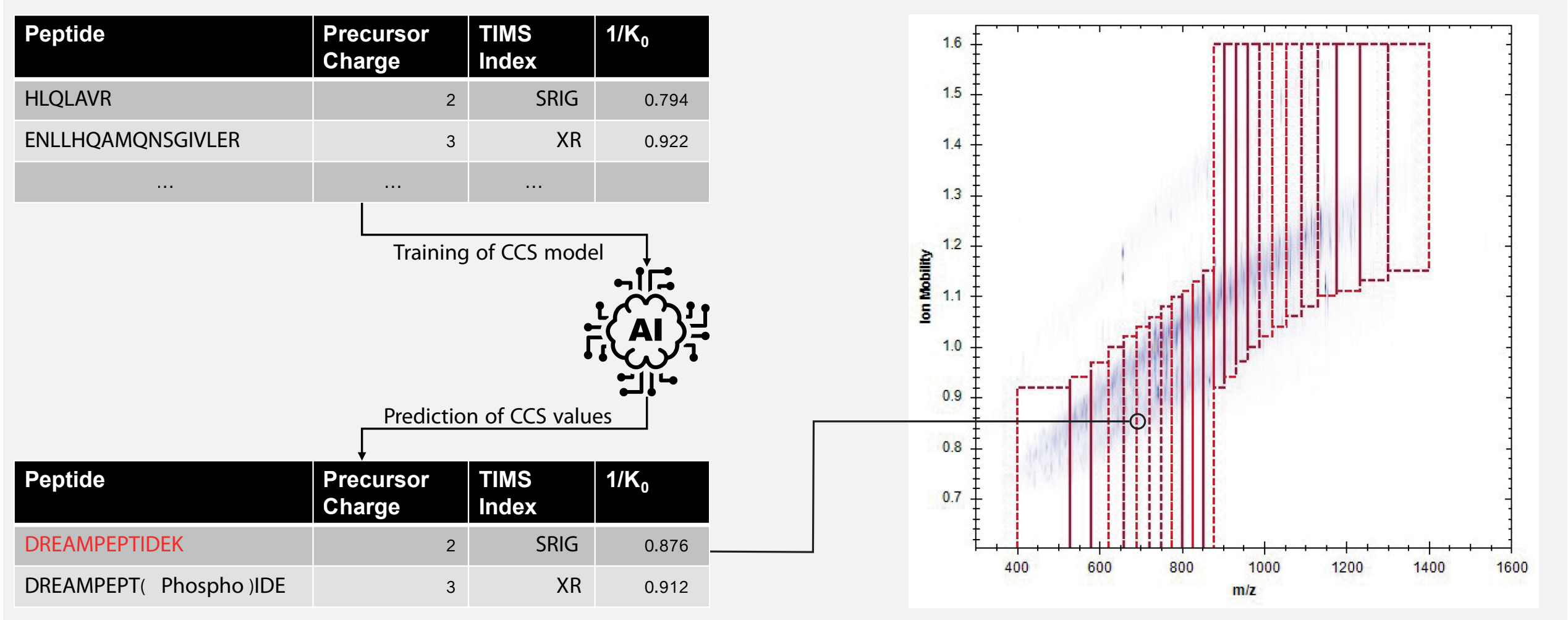
1) Bruker Switzerland AG, Faellanden, Switzerland; 2) Mestrelab Research S.L., Santiago de Compostela, Spain; 3) Biognosys AG, Schlieren, Switzerland; 4) Bruker Daltonics GmbH & Co. KG, Bremen, Germany
GR, IJ, DT are employees of subsidiaries of Bruker Corp. AP, AN, TG, and LR are employees of Biognosys AG. timsTOF is a trademark owned by Bruker Corp. Spectronaut® is a trademark owned by Biognosys AG

POSTER NUMBER: MP 435

INTRODUCTION

Trapped ion mobility spectrometry (TIMS)[1] extends conventional LC-MS/MS proteomic workflows with an additional ion mobility dimension. Next to the benefits of signal separation, the implementation of TIMS in Bruker timsTOF instruments has demonstrated that collisional cross section (CCS) values of peptides are highly reproducible and can be used as orthogonal coordinates to retention time and ion m/z values for targeted data extraction in data-independent acquisition (DIA) workflows[2], or as additional metric for rescoring in spectrum-centric database searches[3] **(Figure 1)**. However, established predictors typically underperform when applied to out-of-distribution datasets, e.g. those covering different PTM, sample or instrument types.

Here we tackle these challenges by two approaches: First, we use training data covering more diverse PTMs. Second, we encode different instrument configurations by a 'TIMS Index', which is later dynamically optimized during inference.



Peptide	Precursor Charge	TIMS Index	$1/K_0$
HLQLAVR	2	SRIG	0.794
ENLLHQAMQNSGIVLER	3	XR	0.922
…	…	…	

Training of CCS model

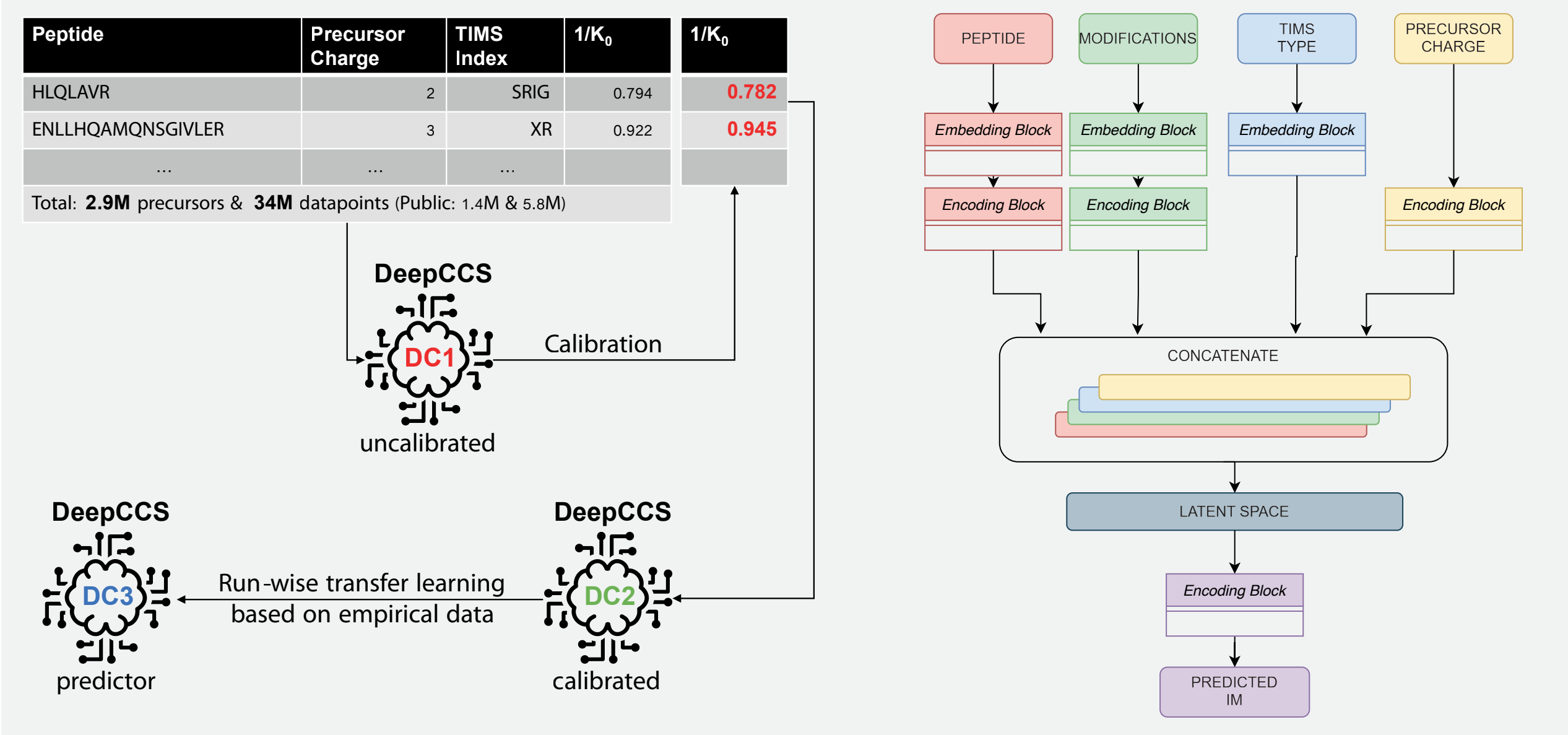
Prediction of CCS values

Peptide	Precursor Charge	TIMS Index	$1/K_0$
DREAMPEPTIDEK	2	SRIG	0.876
DREAMPEPT(Phospho)IDE	3	XR	0.912

Figure 1 Overview of Trapped Ion Mobility Spectrometry (TIMS). Bruker timsTOF instruments provide CCS-centric separation for both DDA and DIA data acquisition modes. Spectral libraries providing empirically measured peptide precursor CCS values support spectrum- or peptide-centric scoring of DDA/DIA spectra and peak groups. Machine learning allows to generalize this information as part of universal CCS predictors that can be applied to all theoretical peptide sequences.

METHODS

- DeepCCS implements a lightweight neural network architecture, encoding peptide sequence, modifications, charge state, as well as context-specific meta data efficiently **(Figure 2)**.The model is trained in a two-step process (DC1: uncalibrated, DC2: calibrated) on a diverse set of public[4,5,6] and internal timsTOF Pro, Pro 2 and HT datasets, which were acquired in DDA and DIA modes.
- In addition to peptide sequence, modifications and charge state, we explored the option to provide a set of new raw data-derived meta-scores to the model to capture instrument and sample dependencies of the measured CCS values. Using confidently identified peptide precursors from a first pass search, DeepCCS in Spectronaut leverages transfer learning to adapt the model to sample-specific conditions, e.g. differences in TIMS calibration between datasets (DC3).



Peptide	Precursor Charge	TIMS Index	$1/K_0$	$1/K_0$
HLQLAVR	2	SRIG	0.794	0.782
ENLLHQAMQNSGIVLER	3	XR	0.922	0.945
…	…	…		

Total: **2.9M** precursors & **34M** datapoints (Public: 1.4M & 5.8M)

Figure 2 DeepCCS workflow overview. Peptide sequences, PTMs and precursor charges are embedded via separate modules. A TIMS index is further concatenated to the input and used for DeepCCS, which creates a latent space via another module. A final block connects output CCS values. Three iterations of DeepCCS are trained; 1) an uncalibrated model; 2) a run-wise calibrated model with outliers removed; and 3) a final model for prediction refined with last-layer transfer learning.

RESULTS

DEEPCCS IMPROVES PERFORMANCE FOR OUT-OF-DISTRIBUTION PEPTIDES

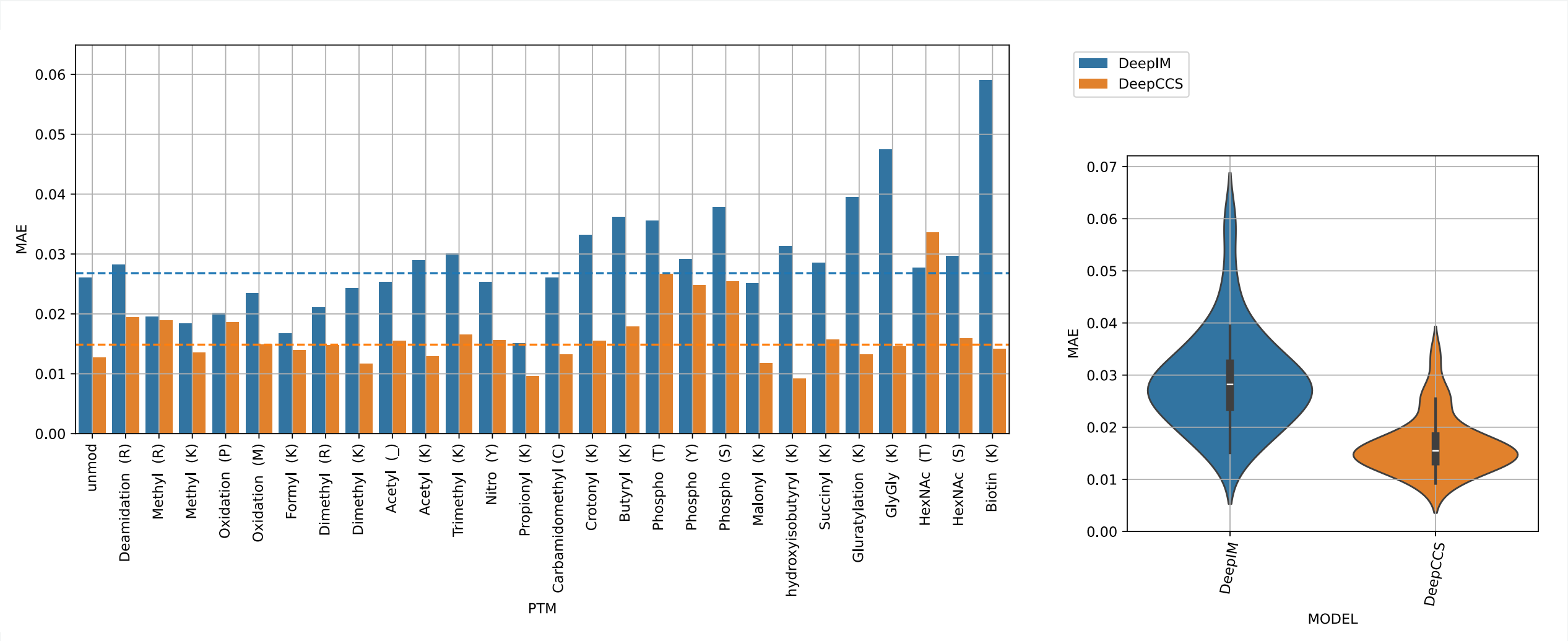


Figure 3 DeepCCS and DeepIM are applied to a dataset[6] with previously out-of-distribution peptide modifications. Mean Absolute Error (MAE) increases with increasing mass of PTM type (order of x-axis). MAE of DeepCCS improves substantially for unmodified peptides and most peptide modifications.

- DeepCCS substantially improves performance by predicting peptide precursor CCS values with a lower Mean Absolute Error (MAE) compared to DeepIM. When applied to a dataset of synthetic peptides covering 22 PTM types, DeepCCS substantially improves predictive performance for most PTM types **(Figure 3)**. In general, MAE correlates with the size of the PTM, with heavier PTMs generating a larger offset.
- Over a wide range of samples, instrument configurations and PTM types, DeepCCS shows substantial improvements compared to DeepIM, lowering median MAE values from ~0.027 to ~0.016.
- A critical component for the improvement is the encoding of TIMS type, either SRIG (timsTOF Pro, Pro 2, SCP) or XR (timsTOF HT, Ultra), because of different capacity limits of the TIMS cartridges. DeepCCS was trained with almost equal numbers of TIMS SRIG and XR data points, generating a stratified model.

DEEPCCS IMPROVES IDENTIFICATION PERFORMANCE IN SPECTRONAUT

- Employed within Spectronaut, DeepCCS substantially improves performance for identification of peptides, providing a median improvement of ~3% over all benchmark datasets **(Figure 4)**.

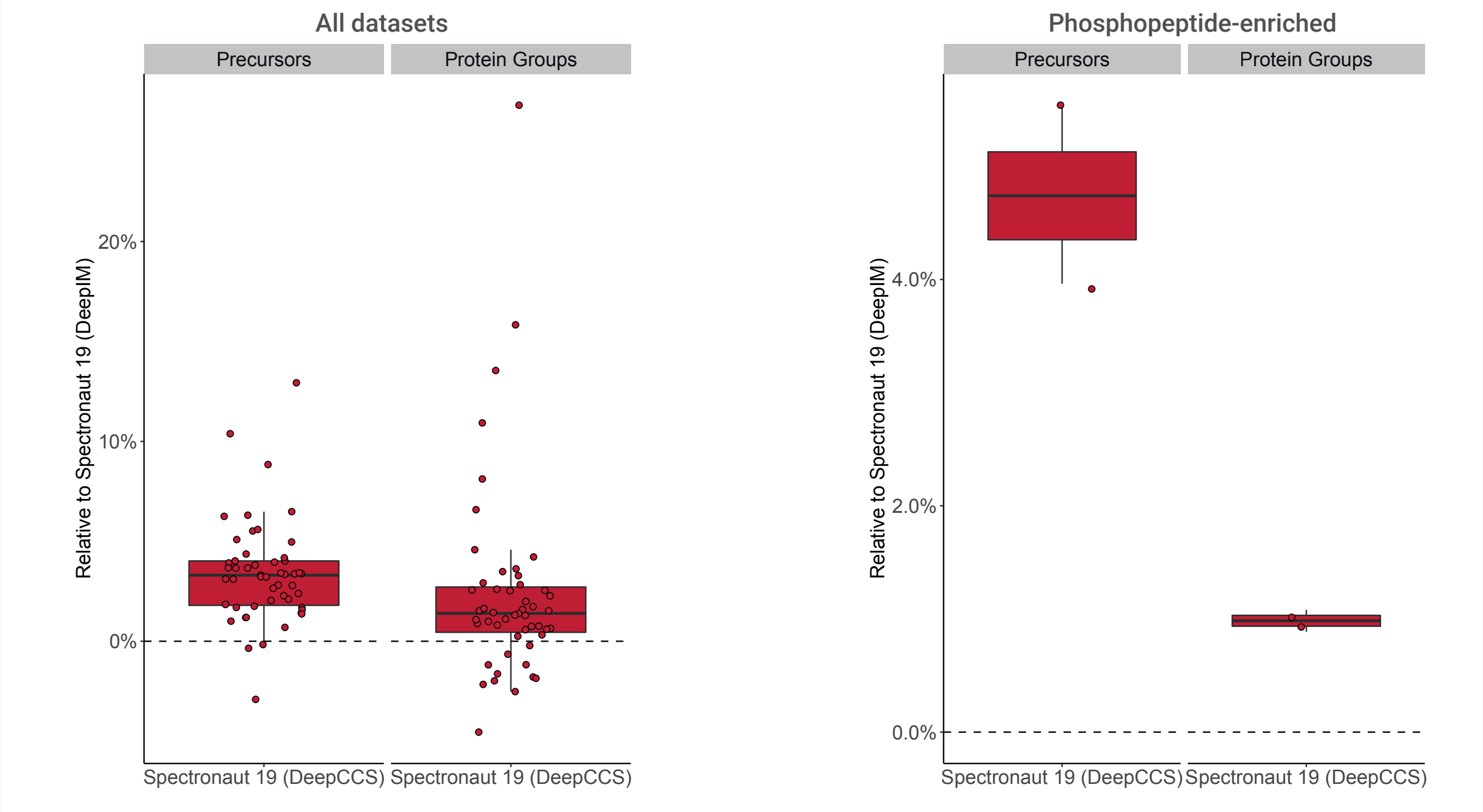
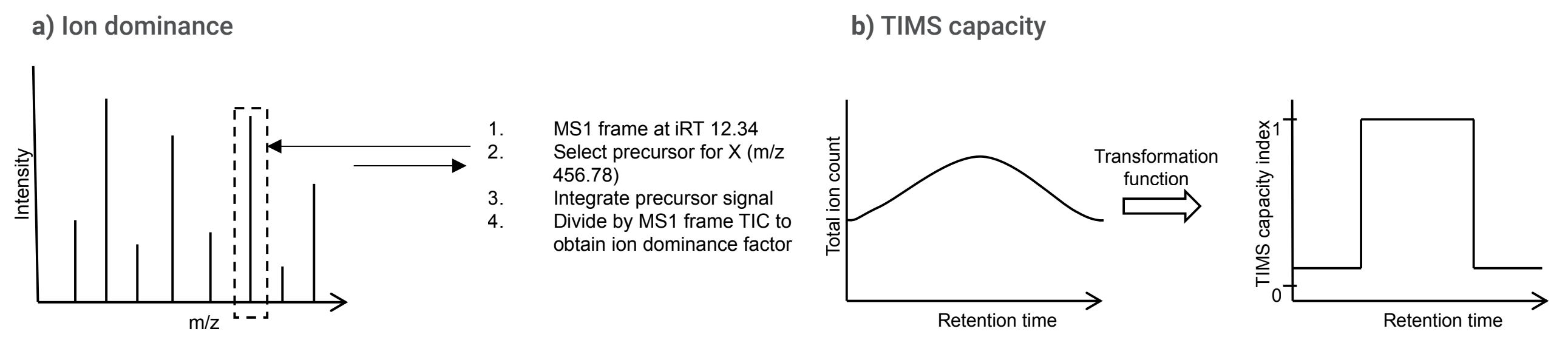


Figure 4 DeepCCS vs. DeepIM identification performance improvements in Spectronaut 19 are depicted. Typical identification rate improvements for standard whole proteome samples are in the order of ~3% on precursor- and ~1.5% on protein level. For phosphopeptides, which where not natively supported by DeepIM, improvements are higher, with ~5% on precursor-level.

ACCOUNTING FOR SAMPLE CONTEXT IN DEEPCCS: ION DOMINANCE AND TIMS CAPACITY



a) Ion dominance

1. MS1 frame at iRT 12.34
2. Select precursor for X (m/z 456.78)
3. Integrate precursor signal
4. Divide by MS1 frame TIC to obtain ion dominance factor

b) TIMS capacity

Figure 5 Schematics for computation of additional metadata scores used in DeepCCS-IDTC. (a) Ion dominance integrates precursor ion intensity over retention time. (b) TIMS capacity transforms TIC via a sliding window to a measure of TIMS capacity via index of dispersion.

- Although the TIMS cartridge in timsTOF instruments has a very high dynamic range of operation, some peptides of complex samples can reach the capacity limits temporarily. Frequently, this results in a drift of CCS values for the high and low abundant ions concurrently present in the TIMS cartridge. To investigate whether additional context-specific information could further boost the performance of DeepCCS and correct for this effect, we investigated whether providing two additional context-specific metrics would improve the model:
- Ion dominance: This metric measures how dominant a specific ion is compared to other ions concurrently being measured in the TIMS cartridge. It can be computed by integrating the precursor signal within a specific tolerance over the target MS1 frames and dividing this value by the MS1 TIC.
- TIMS capacity: When a TIMS cartridge operates at full capacity, total ion current typically reaches a local maximum. Over a sliding window in retention time, the index of dispersion is computed as a proxy measure of TIMS capacity at any given time.

PROTOTYPE EVALUATION OF DEEPCCS-IDTC ON OOD TISSUE SAMPLES

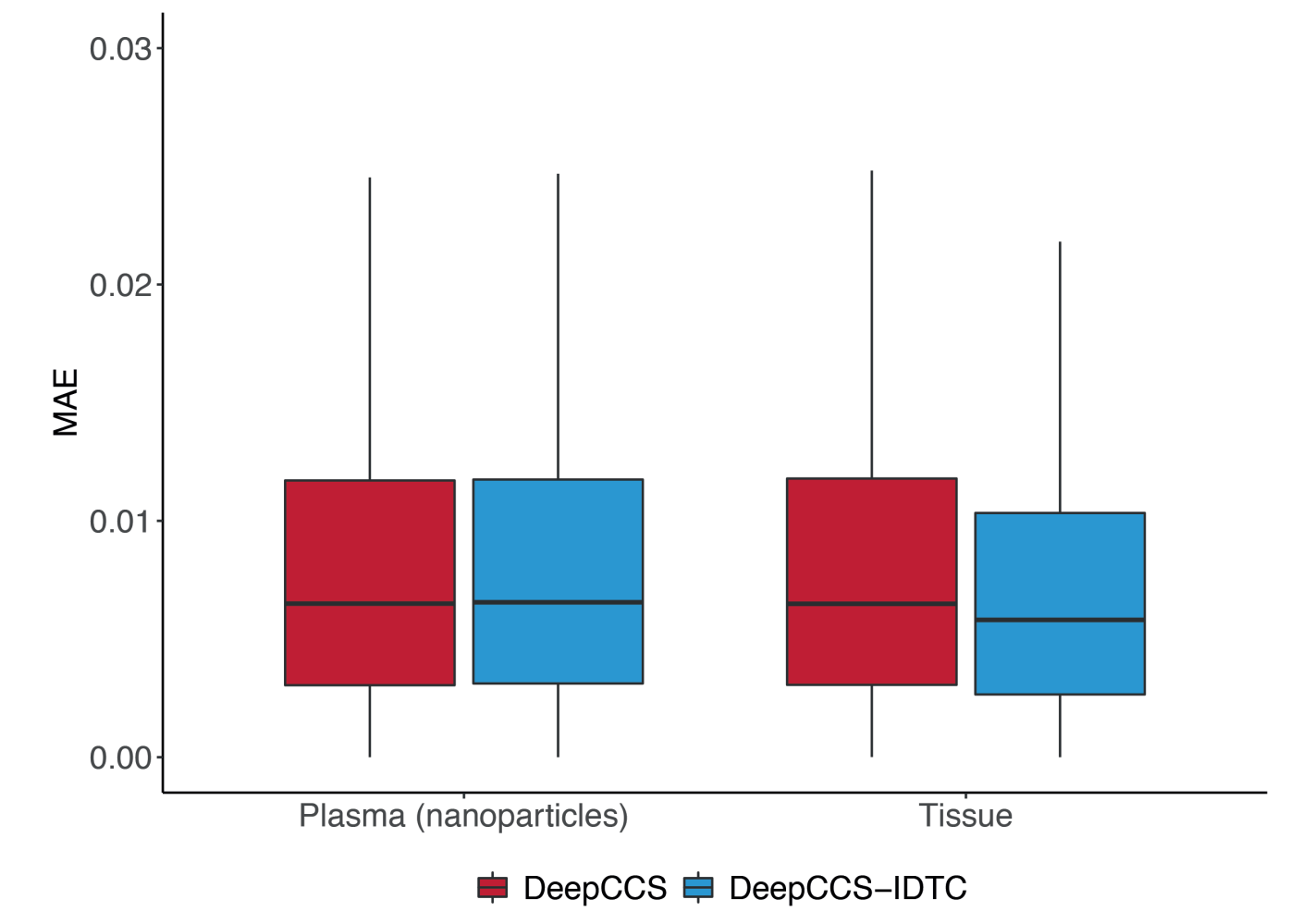


Figure 6 DeepCCS is compared to DeepCCS-IDTC on a selection of out-of-distribution tissue samples.

- Ion dominance (ID) and TIMS capacity (TC) metrics were computed for each data point and supplied to a modified model referred to as DeepCCS-IDTC as additional input. While DeepCCS-IDTC performs identical to DeepCCS on samples of low complexity, DeepCCS-IDTC achieves improved performance for complex samples like human tissue.
- DeepCCS-IDTC is currently a research project and not yet employed as predictor within Spectronaut 19.

SPECTRONAUT 19 CUMULATIVE IMPROVEMENTS FURTHER BOOST IDENTIFICATIONS

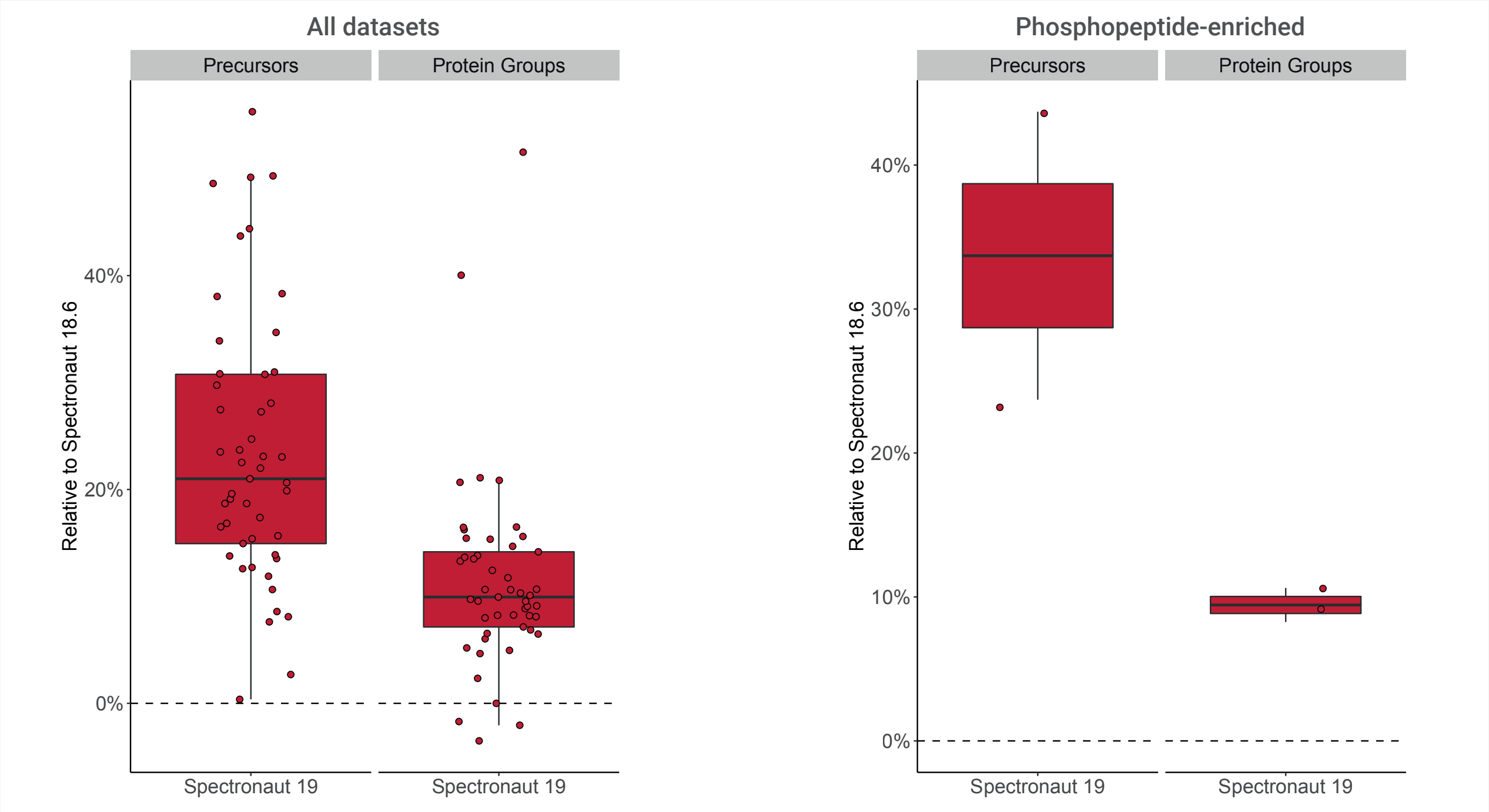


Figure 7 Cumulative Spectronaut 19 improvements over version 18 are depicted for all datasets (left) and phosphopeptide-enriched datasets (right).

- In combination with all other improvements of Spectronaut 19, the cumulative improvements on peptide precursor level range between 10–25% for different timsTOF datasets. Depending on the specific dataset, the contributions of DeepCCS to these improvements range from 5–15%.

REFERENCES

1. Meier et al.; Trapped Ion Mobility Spectrometry and Parallel Accumulation–Serial Fragmentation in Proteomics. Mol. Cell. Proteomics, 2021
2. Meier et al.; diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. Nat. Methods., 2020
3. Declercq et al.; MS^2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. Mol. Cell. Proteomics, 2022
4. Meier et al.; Deep learning the collisional cross sections of the peptide universe from a million experimental values. Nat Commun., 2021
5. Adams et al.; Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in TimsTOF Anal bioRxiv., 2023
6. Will et al.; Peptide collision cross sections of 22 post-translational modifications. Anal Bioanal Chem., 2023

CONCLUSIONS

- Compared to DeepIM, DeepCCS predicts CCS values more accurately and ~2x faster on CPU.
- DeepCCS represents a substantial improvement for CCS prediction in Spectronaut 19. In addition to native support for 22+ PTMS, DeepCCS leverages TIMS type indexing, two-pass model generation and transfer learning to substantially improve predictive performance, resulting in increased identification rates over DeepIM.
- With DeepCCS-IDTC, we investigated potential future extensions that could account for sample context specificity and further improve the predictive performance of DeepCCS.

CONTACT

For further information about this poster, please contact:

George Rosenberger, PhD
Software Project Manager, Machine Learning for MS-based Omics

T: +41 79 428 82 61
E: george.rosenberger@bruker.com