analytical chemistry

Open Access
This article is licensed under CC-BY-NC-ND 4.0

pubs.acs.org/ac
Article

iDeepLC: Chemical Structure Information Yields Improved Retention Time Prediction of Peptides with Unseen Modifications

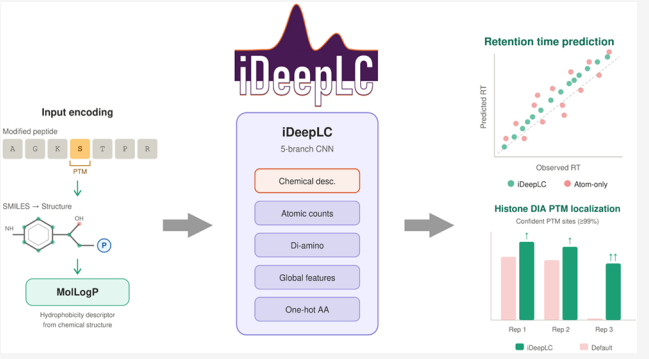
Alireza Nameni, Arthur Declercq, Ralf Gabriels, Robbe Devreese, Sven Degroeve, Amélie De Maesschalck, Maarten Dhaenens, Cristina Chiva, Eduard Sabidó, Lennart Martens,* and Robbin Bouwmeester

Cite This: *Anal. Chem.* 2026, 98, 17707−17717

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Deep learning has notably advanced the field of liquid chromatography−mass spectrometry-based proteomics. Accurate prediction of peptide retention times significantly enhances our ability to match LC-MS data with the correct peptides and proteins, especially for data-independent acquisition data. While numerous models predict peptide LC retention times with high accuracy, few can accurately predict the retention times of chemically modified peptides, particularly those with modifications not encountered during model training. In our previously developed DeepLC model, accurate predictions could be made for unseen modifications by leveraging the chemical compositions of (modified) residues. Here, however, we present a further enhancement of this model based on the chemical structural information. The resulting model, called iDeepLC, shows overall more accurate predictions and better generalization performance for predicting the retention time of modifications structurally defined as SMILES but unseen during training than DeepLC. iDeepLC is freely available as an open-source software under the Apache2 license and can be found at https://github.com/CompOmics/iDeepLC.



INTRODUCTION

In liquid chromatography−mass spectrometry (LC-MS) proteomics, retention time refers to the specific duration it takes for a peptide to traverse the chromatography column and reach the MS detector. Chromatographic separation and its associated retention time value serve as an additional dimension that separates peptides prior to analysis and thus substantially reduce sample complexity in the acquired MS dimensions. Deep-learning models have been built that can accurately predict this expected retention time for a given peptide sequence under various experimental conditions. These predictions enable the in silico creation of spectral libraries for the analysis of data-independent acquisition[1−4] data and enable the rescoring of peptide-spectrum-matches (PSMs) to increase identification sensitivity.[5−9] Furthermore, they are utilized to optimize experimental design,[10] to identify multiple peptides in chimeric fragmentation spectra,[11] to simulate LC-MS experiments,[12] and to provide for orthogonal validation of PSMs.[13]

However, accurate predictions of the retention time of modified peptides remain challenging. Most current predictors use a fixed encoding for modifications, so they either cannot predict retention times for modifications not seen during training or do so poorly due to limited training data. Examples of such models include architectures such as transformers (Pham et al.),[14] linear residual convolutional neural networks (CNNs) (Chronologer),[15] capsule CNNs (DeepRT),[16] neural networks with long short-term memory (LSTM) layers (Guan et al. and AutoRT),[17,18] combinations of an LSTM and a transformer (DeepPhospho),[19] and encoder-decoder models with gated recurrent units (Prosit).[20] In contrast, a smaller number of models employ a more sophisticated encoding method that includes the atomic composition of modifications, allowing the model to be generalized over the modifications in the training set. This approach facilitates the prediction of peptides carrying modifications not observed during training, as seen in models with LSTM layers (AlphaPeptDeep and pDeep3),[21,22] and our own branched CNN, DeepLC.[23]

Here, we have explored the extension of these atomic count models by providing molecular and structural information about the (modified) amino acid to the model in the form of

Received: December 18, 2025
Revised: May 7, 2026
Accepted: May 8, 2026
Published: June 8, 2026

© 2026 The Authors. Published by American Chemical Society

SMILES[24] and molecular descriptors. We hypothesized that providing this additional information could further enhance the accuracy of retention time predictions for modified peptides, particularly for modifications structurally defined as SMILES but unseen during training. More broadly, structural descriptions may also help the model capture physicochemical differences between modifications that share similar or identical elemental compositions. In small molecules, the issue of isomers has been addressed using quantitative structure-selectivity relationships (QSSR) and molecular descriptors.[25] Different sets of these molecular descriptor features have been used and compared in various machine learning models, with what is known as the MolLogP feature consistently emerging as the most discriminating.[26] This chemical descriptor represents the quantified propensity of a molecule to dissolve in nonpolar solvents such as octanol, compared to polar solvents such as water.[27]

Here, we introduce iDeepLC short for "improved DeepLC," a retention time predictor based on the DeepLC architecture, but that also incorporates structural, in form of SMILES, and molecular information through the chemical descriptor MolLogP, in a branch neural network architecture with five paths. Our findings demonstrate that iDeepLC achieves superior predictive accuracy and generalization capabilities than previous models, making it a valuable tool for proteomics research.

METHODS

Model Architecture

iDeepLC employs a multibranch neural network architecture similar to DeepLC,[23] with an additional, fifth branch that encodes chemical structural information. As explained in more detail in Input Encoding, the first three branches consist of convolutional layers[28] that encode chemical properties, diamino chemical properties, and atomic counts. A fourth branch is a fully connected layer that encodes the general peptide features. The fifth branch consists of convolutional layers for one-hot encoding of amino acids. The outputs of these five branches are then flattened and concatenated into several connected layers, finally yielding a single numerical output. A detailed schematic of the architecture is provided in Figure S1.

Incorporation of the Additional Chemical Descriptor

To represent the structure of each potentially modified amino acid, we obtained their corresponding SMILES representation. This SMILES representation is then used to compute the MolLogP chemical structure descriptor with the RDKit library[29] version 2023.3.2.

By incorporation of MolLogP, iDeepLC gains insights into how atomic bonding patterns and individual atomic properties contribute to the hydrophobicity of amino acids. As a result, it can more effectively distinguish among various amino acids and their modifications, including those with identical atom compositions.

Structural definition for each modified residue in the form of an SMILES string is required to calculate its associated MolLogP. Throughout this manuscript, we use the term unseen modification to denote a modification that is excluded from the training set but for which a SMILES structure is available at inference time and is therefore encoded during testing. Consequently, iDeepLC is not intended to model modifications that are structurally undefined (i.e., lacking a SMILES representation).

Input Encoding

The input of iDeepLC consists of, potentially modified, peptide sequences, each represented as a matrix with dimensions of 41 rows for the features described below, and 60 columns with each column representing a (modified) amino acid in a peptide with a maximum length of 60 amino acids. These features are categorized into five groups:

1. The **chemical descriptor, MolLogP**, is represented in the first row, encapsulating chemical characteristics of the peptides.
2. Rows two to seven encode per-residue **atomic** counts of six elements: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S) for the (modified) amino acids.
3. **Diamino chemical descriptor and diamino atoms** in rows eight and nine to 14, respectively. This branch sums the features of every two adjacent amino acids, enhancing the model's ability to generalize across different peptide sequences. For peptides with an odd number of amino acids, the last column repeats the previous column's features to ensure consistency without overlapping.
4. **Global features**, which span rows 15 to 21, provide supporting information about the peptide sequence. It includes the normalized length of the peptide relative to the maximum defined length, the atom composition of the first four and last four amino acids, and the total atom composition of the entire peptide.
5. Rows 22 to 41 are the **one-hot encoding** to provide a clear, binary representation of amino acids within the sequence. This path is particularly important to distinguish between isoleucine and leucine as they are structural isomers.[30]

To address the challenge of varying peptide lengths, peptides with less than 60 amino acids are padded to maintain uniformity of 60 columns. This facilitates the straightforward encoding of various (modified) peptides, regardless of their length. Note that in instances where modifications are present, the chemical descriptor, diamino chemical descriptor, and atomic composition are adjusted to reflect modified amino acid change.

Hyperparameter Tuning

The hyperparameters of the models were optimized on the HeLa HF[31] data set. This optimization process was conducted with the WandB machine learning platform.[32] In total, 18 different hyperparameters were tuned, and these hyperparameter values are available in Table S1. The tuned hyperparameters include the learning rate, batch size, dropout rate, kernel sizes, number of channels, and number of layers for the four convolutional layers.

Data Sets and Evaluations

We first evaluated the performance of iDeepLC on 20 proteomics data sets (Table S2), ranging from 3000 to 1,60,000 sequences. These data sets mainly contain unmodified peptides and common modifications such as oxidation on methionine and carbamidomethyl on cysteine. One exception is the ProteomeTools PTM[33] data set, which includes synthetic peptides with various modifications. To ensure a fair and accurate comparison, we used the same training, testing, and validation sets that are used in the evaluation of DeepLC.[23]

For comparison with external retention time predictors (AlphaPeptDeep[21] and Chronologer[15]), the training configurations indicated in their corresponding articles and GitHub repositories were used. When training hyperparameters were explicitly reported in the corresponding papers, we used those values. When specific hyperparameters were not reported, we used the default values set on the GitHub. For example, for AlphaPeptDeep, we applied the training schedule described by the authors (30 warm-up epochs followed by 300 training epochs), whereas for Chronologer, we used the default number of epochs specified in the published codebase.

Next, we assessed iDeepLC on 14 modifications from the ProteomeTools PTM data set, an experiment referred to as the *14PTMs experiment* that was originally described in DeepLC. In this experiment, we trained and optimized 14 iDeepLC models where each model was only trained on peptides that did not contain a specific modification. These models were then evaluated on the remaining peptides, which all contained modifications excluded during training. We created two test sets from these remaining

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

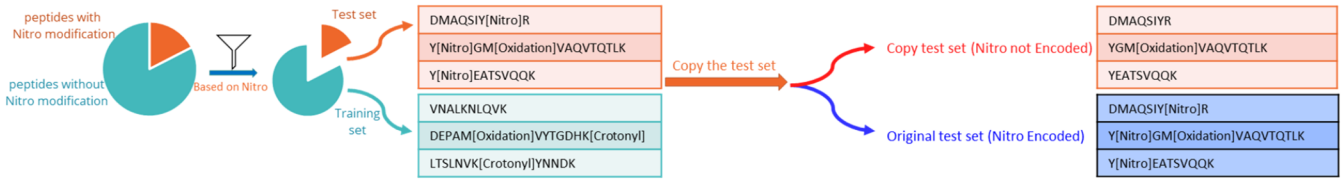


Figure 1. Process of generating data sets to evaluate the generalization of iDeepLC for peptides containing a specific peptide modification. In this example, we filtered all peptides carrying a nitro-oxidation modification and trained the model on peptides without nitro oxidation and used those with the modification as a test set. In the next step, the test set is duplicated, and all nitro-oxidation modifications were removed (nitro not encoded) while in the other one, we kept the peptides as they were (nitro encoded).

peptides to evaluate predictions: one where the excluded modification was encoded, and one where it was ignored (not encoded) (Figure 1).

In an additional evaluation that we call the *Glycine experiment*, we utilized all 20 data sets from Table S2. We trained 19 models for each data set, where each model excluded all peptides that contained a specific amino acid during training and was tested on peptides containing that amino acid. To assess the effect of the substitution, the test set was duplicated such that each peptide appeared in two versions: one retaining the original excluded amino acid, and one in which that residue was replaced by glycine. When substituting an excluded residue by glycine, any modifications present on that residue were also removed, this means the substituted position was always encoded as unmodified glycine. This evaluation method was originally designed and described in DeepLC and it was aimed at evaluating the generalization capabilities of iDeepLC. This evaluation assesses the model's ability to predict the retention times of peptides incorporating amino acids absent from the training set. Furthermore, for 17 out of the 20 data sets there are many more training examples compared to the *14PTMs experiment* evaluation and should thus reflect actual performance better. Finally, for the comparison with DeepLC, we focused on two specific data sets: the DIA HF data set,[4] which consists of approximately 1,13,000 sequences, and the HeLa DeepRT data set,[34] with around 3400 sequences (Table S2).

For all evaluations, three different metrics are used to assess performance: (i) Mean Absolute Error (MAE), (ii) relative MAE, and (iii) the Pearson correlation. The relative MAE is used to make the comparison between different data sets possible. Eq 1 is used for the calculation of the relative MAE:

$$\mathrm{MAE}_{\mathrm{relative}} = \frac{\mathrm{MAE}}{\mathrm{peptide}_{\mathrm{last}} - \mathrm{peptide}_{\mathrm{first}}} \times 100 \quad (1)$$

Here, the MAE is divided by the retention time difference between the first and last identified peptide in the respective data set.

Training Procedure

To initialize all models' parameters for learning, we utilized the Kaiming uniform initialization[35] method with a leaky ReLU[36] activation function and we initialized biases to zero. This initialization scheme was applied to both convolutional and fully connected layers within the models. All layers use ELU[37] as activation function except for fully connected layers that use ReLU[38] activation function. The training parameters are available in Table S1.

We trained the models for 1000 epochs for all 20 data sets and the *14PTMs evaluation*. For the *Glycine evaluation*, we trained for either 1000 for small data sets or for 300 epochs, when the size of the training data set was more than 1500 peptide sequences. The decision to run fewer epochs is based on the early convergence observed in training larger data set sizes. In the training of all models, the best model based on MAE was used for the validation analysis. The training was done on an NVIDIA GeForce RTX 3090, using Python version 3.10.11 and PyTorch version 2.2.1 with CUDA version 11.8.

Calf Thymus Sample

Histones from calf thymus (Roche #10223565001) were chemically derivatized and digested to generate a peptide mixture enriched for histone post-translational modifications (PTMs). Briefly, 10 μg of histones were derivatized in triplicate with propionic anhydride, digested with trypsin, and derivatized again with phenyl isocyanate following previously described protocols.[39] Samples were first dissolved in H_2O and buffered to pH 8.5 using HEPES. Propionic anhydride was added for lysine derivatization and the reaction was quenched with hydroxylamine. After tryptic digestion (3 h at 37 °C), peptides were derivatized with phenyl isocyanate, acidified with formic acid, vacuum-dried, and desalted using C18 ultramicrospin columns (The Nest Group, Inc., Southborough, USA). Peptides were analyzed using an LTQ-Orbitrap Eclipse mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1200 system. A 5 μg peptide aliquot was injected and separated by reversed-phase chromatography on a 50 cm × 75 μm column packed with 2 μm C18 particles using a 60 min gradient. Data were acquired in both data-dependent acquisition (DDA) and data-independent acquisition (DIA) modes. The DDA method employed a TopSpeed acquisition strategy, whereas the DIA method consisted of 25 staggered isolation windows of 24 m/z covering a precursor range from 350 to 1850 m/z.

Downstream Calf DDA/DIA Analysis

DDA runs were analyzed using Proteome Discoverer (v2.5, Thermo Fisher Scientific) with the Mascot[40] search engine (v2.6, Matrix Science). Searches were performed against the Swiss-Prot bovine database supplemented with common contaminants. In a DDA search, propionylation on lysine and phenyl isocyanate on peptide N-termini were considered variable modifications. Proteins identified in this DDA search were used to filter proteins in FASTA. This filtered FASTA was then used to create an in silico predicted library with DIA-NN where propionylation on protein N-termini and phenyl isocyanate on peptide N-termini were set as fixed modifications, while lysine propionylation, lysine dimethylation, lysine trimethylation, lysine propionyl+methylation, and lysine acetylation were included as variable modifications and up to five missed tryptic cleavages were allowed.[8]

MOLM-13 Cell Line Sample

MOLM-13 acute myeloid leukemia cells were cultured according to the supplier's recommendations and treated with 0.04 μM SNDX-5613 for 240 h before harvest. Histones were extracted by direct acid extraction, chemically derivatized with trimethylacetic anhydride, digested with trypsin, and derivatized again to modify peptide N-termini, followed by reversal of overderivatization on serine, threonine, and tyrosine residues. Histone peptides were analyzed by low-pH reversed-phase LC−MS/MS on an ACQUITY UPLC M-Class system coupled to a ZenoTOF 7600 mass spectrometer, with 400 ng loaded on column. Data were acquired in both data-dependent acquisition (DDA) and Zeno SWATH data-independent acquisition (DIA) modes.

Downstream MOLM-13 DIA Analysis

DIA data were analyzed with DIA-NN (v2.1.0) using an in silico predicted spectral library generated from a filtered FASTA database containing histone proteins and contaminant proteins. Searches were performed with cleavage after arginine residues, and up to two missed cleavages were allowed. Trimethylacetylation on peptide N-termini was specified as a fixed modification, whereas lysine methylation, dimethylation, trimethylation, acetylation, formylation, trimethylacetylation, and methyltrimethylacetylation were considered as variable modifications.

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

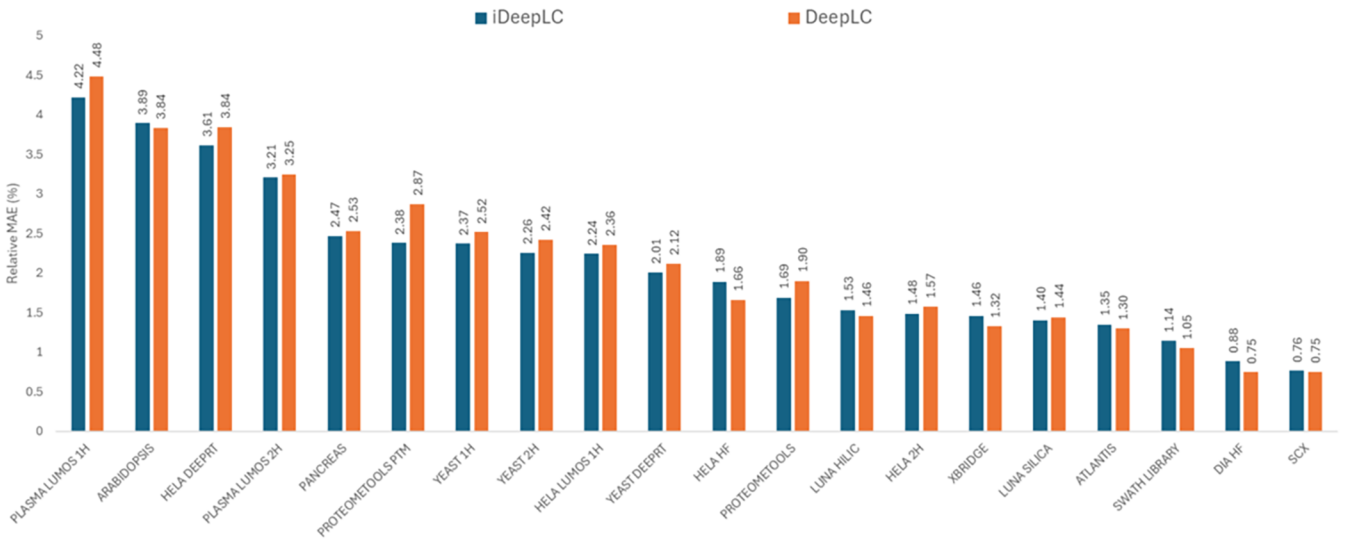


Figure 2. Performance comparison of iDeepLC (blue) and DeepLC (orange) for 20 data sets with the relative MAE as metric (lower is better).

To evaluate the impact of improved retention time prediction in a practical workflow, we analyzed the resulting spectral library using DIA-NN (v2.1.0). Two versions of the spectral library were used: the original library containing the experimentally observed retention times and a modified library in which the retention time entries were replaced with iDeepLC-predicted retention times. Apart from this modification, all library entries and DIA-NN analysis parameters were kept identical between both analyses. DIA data were processed independently for different replicate runs (PXD075498) by using identical search parameters and a precursor-level q-value threshold of 1%.

RESULTS

We first compare the performance of iDeepLC to DeepLC on 20 different data sets (Table S2), followed by an evaluation of iDeepLC's ability to generalize for (modified) peptides. This evaluation serves to assess its capability to generalize and is divided into two parts; first, iDeepLC's capability to accurately predict the retention time of peptides with a specific modification excluded from training of the model (*14PTMs evaluation*), and second, the prediction for peptides containing an amino acid not used for training (*glycine evaluation*). It is expected that the addition of the MolLogP descriptor improves the ability of iDeepLC to generalize for modifications, especially for those unseen during training.

Prediction Performance iDeepLC

We compared the prediction performance of iDeepLC with DeepLC on predicting the retention time of peptides from 20 LC-MS data sets that cover several organisms, instrument types, liquid chromatography columns, and gradient lengths (see Methods). Figure 2 shows that iDeepLC demonstrates on-par performance with DeepLC. The largest difference in the relative MAE among iDeepLC and DeepLC is observed for the ProteomeTools PTM data set,[33] where the relative MAE differs more than 0.5%. The second biggest difference is for the "Plasma 1h" data sets with a much smaller 0.28% relative MAE difference. It is important to emphasize that both iDeepLC and DeepLC were trained, validated, and evaluated using identical data splits across each data set. When iDeepLC is compared with the external tools AlphaPeptDeep and Chronologer, the performance in terms of relative MAE was comparable between the different models (Figure S2).

PTM Evaluation

The improved performance on the data set consisting primarily of modified peptides (*ProteomeTools PTM data set*) indicates that iDeepLC can generalize more effectively for modified peptides. Due to the additional chemical descriptor, iDeepLC should be able to understand the peptides' physicochemical properties better compared to DeepLC, where modified peptides are only represented with their atomic composition. This improvement of iDeepLC to predict for modified peptides is further investigated in the *14PTM evaluation*. In this evaluation, the ProteomeTools PTM[33] is used to assess how well iDeepLC understands the modified peptides' physicochemical properties, without explicitly training on observations for a specific amino acid modification.

Figure 3A,B highlights two modifications, nitro oxidation and deamidation, from the *14PTMs evaluation*. For these modifications, iDeepLC can predict the modification-induced retention time shift much better than DeepLC. For these two modifications, the MAE for deamidation and nitro oxidation are improved with 58% (from 257 to 109 s) and 52% (from 567 to 272 s), respectively. For completeness, the scatter plots of all modifications are available in Figure S3.

This trend of better performance for iDeepLC and DeepLC over the baseline continues for the remaining 12 modifications, where performance is either better or comparable (3C). Furthermore, these results show that the impact of encoding or not encoding for methyl group modifications is minimal. In contrast, there is a large difference between encoding and not encoding for acetyl, succinyl, propionyl, crotonyl, malonyl, oxidation, and deamidation. Also, these modifications are much more accurately predicted when they are encoded for both prediction models. The sole exception is the phosphorylated peptides, which are more accurately predicted when not encoded compared to when encoded. However, the prediction error of iDeepLC is still smaller than that of DeepLC when encoding the phosphorylation. This improved prediction accuracy of iDeepLC over DeepLC is also noticeably clear for carbamidomethylation, deamidation, oxidation, and nitro oxidation. The exact adjusted p values for each modification are provided in Table S3.

To further assess this effect, we examined the relationship between the MolLogP value of each modification and the

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

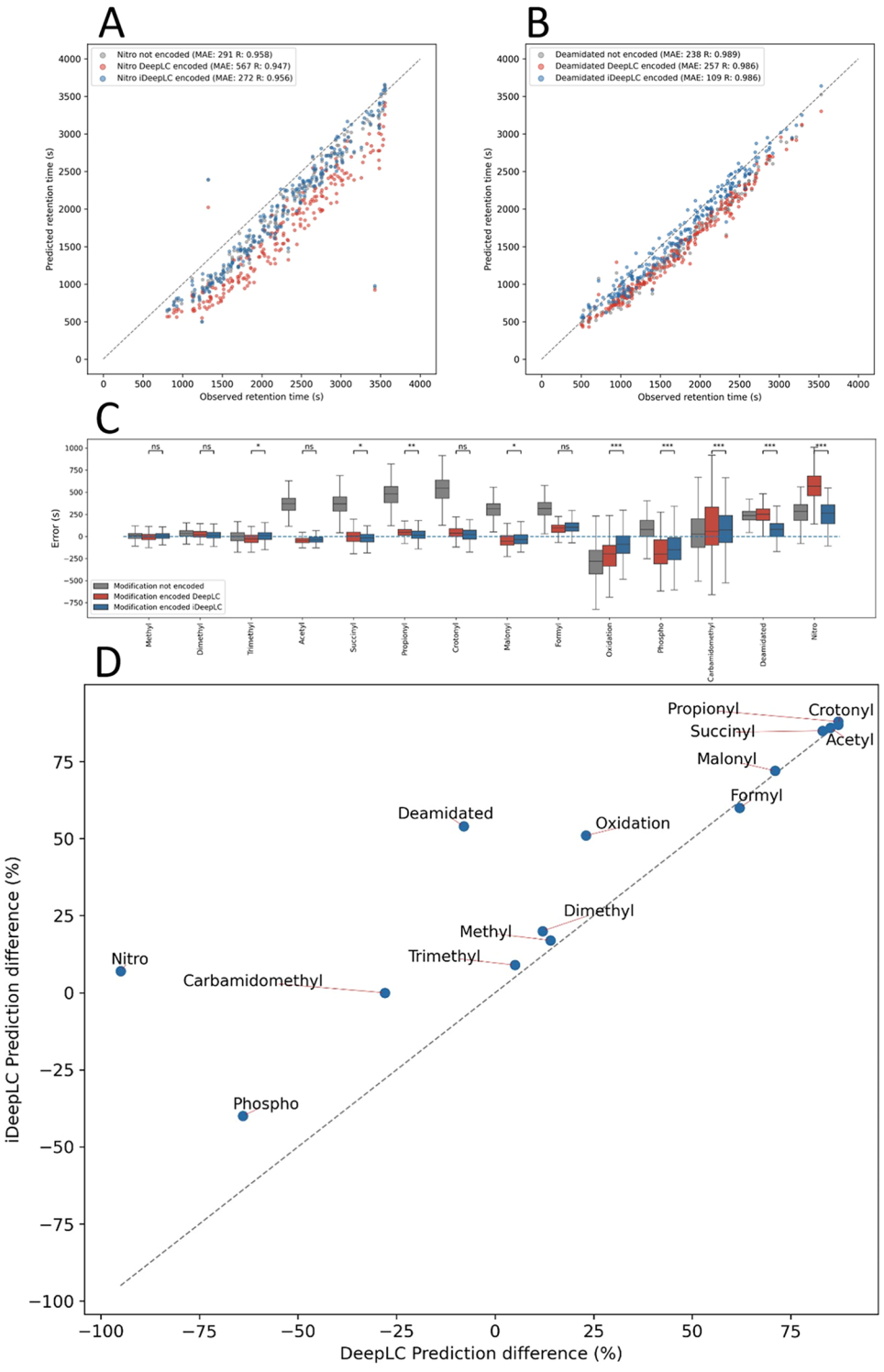


Figure 3. Evaluation of iDeepLC performance on peptides carrying modifications excluded during training (14PTMs evaluation). (A-B) Performance comparison of DeepLC and iDeepLC for models that were not trained on peptides containing nitro oxidation (A) and deamidation (B) but evaluated on their respective modifications. The blue dots and the red dots show the retention time predicted by iDeepLC and DeepLC respectively tested with encoded modifications, and the gray dots show the retention time on an iDeepLC model tested without encoding modifications. (C) Distribution of prediction errors (difference between predicted and observed retention time) for each modification. For every modification, three boxplots are shown corresponding to the baseline where the modification was not encoded (gray), predictions by DeepLC with the modification encoded (red), and predictions by iDeepLC with the modification encoded (blue). Significance markers denote the paired comparison between iDeepLC and DeepLC based on a one-sided Wilcoxon signed-rank test on per-peptide absolute errors, with Benjamini–Hochberg adjusted p values (* p_adj <0.05, ** p_adj <0.01, *** p_adj <0.001, ns = not significant). (D) Relative performance of iDeepLC and DeepLC across the 14PTMs evaluation, expressed as the percentage difference in MAE compared to the baseline where the modification was not encoded.

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

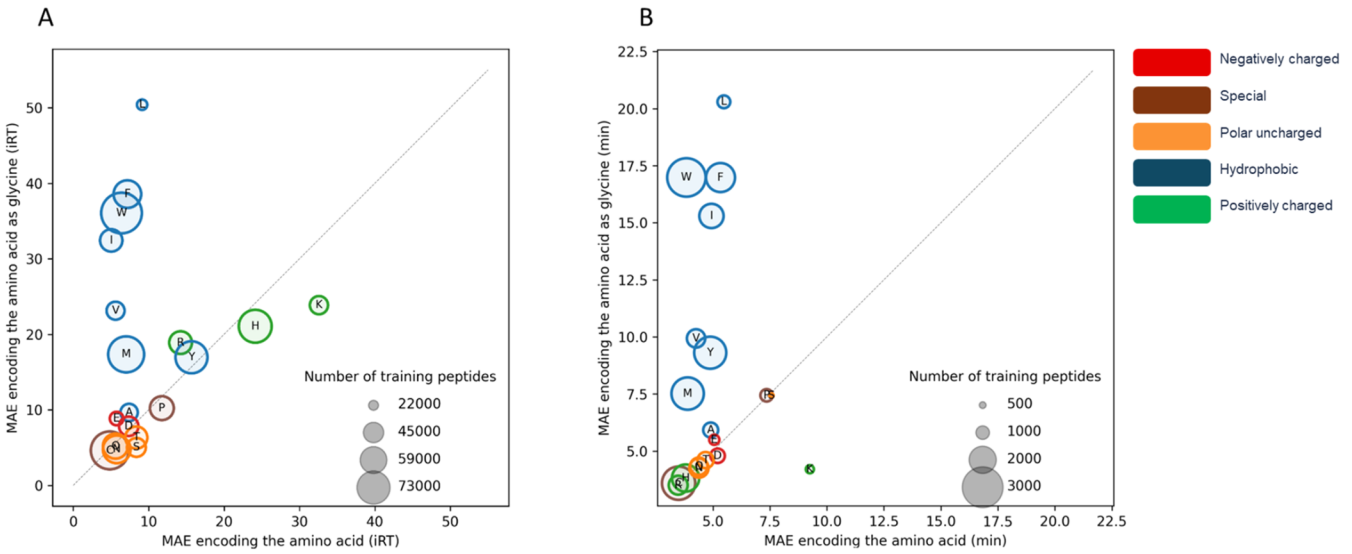


Figure 4. Glycine evaluation where each amino acid that was excluded is shown in a circle with its size determining the number of training peptides and its color shows the chemical property. An amino acid's position indicates the MAE for all peptides containing that amino acid and it is either encoded as glycine (vertical axis) or as its own atomic composition (horizontal axis). Everything above the diagonal line is predicted with a higher accuracy when the amino acid is encoded as itself. Two data sets are used: (A) DIA HF (bigger data set) and (B) HeLa DeepRT (smaller data set).

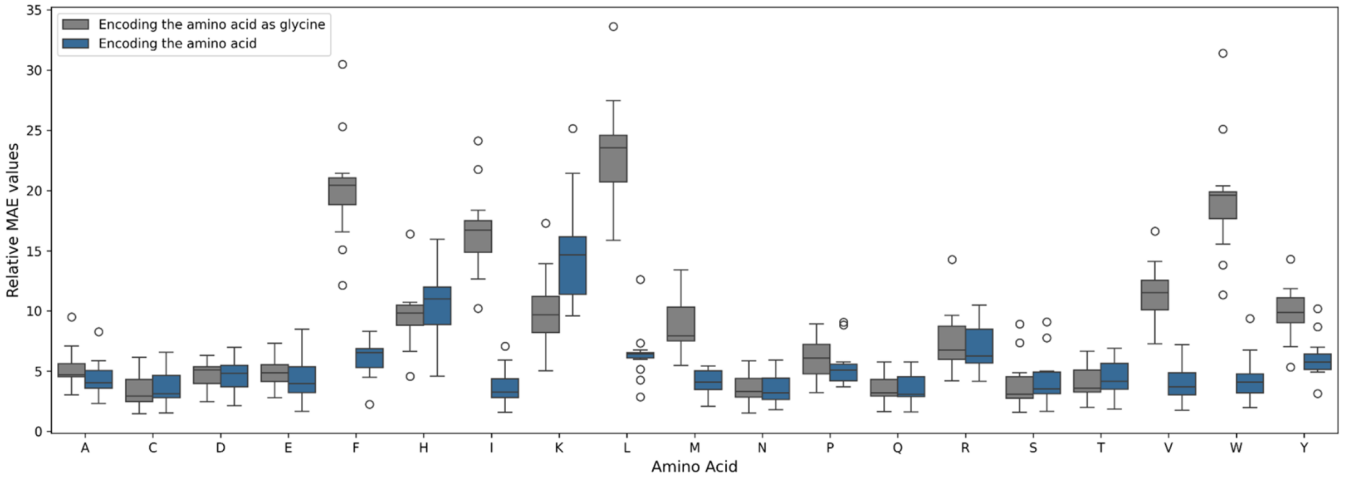


Figure 5. Relative MAE for each amino acid across all 15 reverse-phased data sets. The boxplots illustrate the comparison between encoding the amino acid as itself (blue) versus encoding it as glycine (gray).

improvement of iDeepLC over DeepLC (Supporting Information Figure 4). MolLogP reflects hydrophobicity, with higher values indicating more hydrophobic structures and lower values indicating more hydrophilic structures. Modifications associated with more extreme MolLogP values tended to show larger improvements in prediction accuracy, consistent with the idea that iDeepLC better models retention time shifts when a modification substantially changes peptide hydrophobicity. In contrast, modifications with MolLogP values closer to zero, such as methylation and dimethylation, showed only minor differences between iDeepLC and DeepLC, suggesting that these PTMs introduce limited hydrophobicity changes that are already sufficiently captured by the atomic composition alone.

The relative differences with the not-encoded baseline and DeepLC or iDeepLC are used to further investigate performance differences between these models (Figure 3D and Table S4). In Figure 3D, any modifications positioned above the diagonal line indicate a superior performance for iDeepLC. Any point that is below the diagonal indicates better performance for DeepLC. As observed before, iDeepLC outperforms DeepLC in predicting the retention times for phosphorylation, carbamidomethylation, deamidation, oxidation, and nitro-oxidation modifications. For the remaining nine modifications, the performance of iDeepLC is on par with DeepLC.

Finally, in comparison with retention time prediction models AlphaPeptDeep and Chronologer, iDeepLC performed substantially better on some modifications for which it had a lower relative MAE (e.g., Nitro and Phospho) or comparable performance on the remaining modifications (Figure S5).

Modified Glycine Evaluation

In the *glycine evaluation*, the same methodology as the *14PTMs evaluation* is followed, but instead of modifications, amino acids are excluded from training and tested. Furthermore, instead of not encoding the amino acid, a baseline is created by replacing the specific amino acid with a glycine.

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

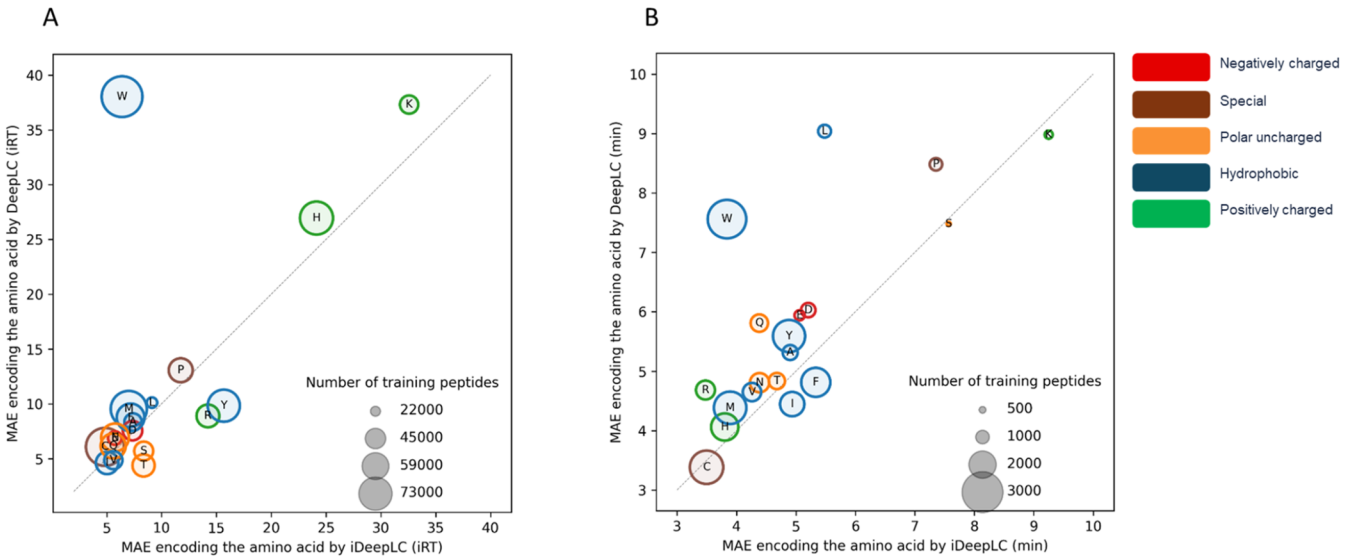


Figure 6. Glycine evaluation where each amino acid that was excluded is shown in a circle with its size determining the number of training peptides and its color shows the chemical property of that amino acid. These figures compare the error of predicted and actual retention time between DeepLC (vertical axis) and iDeepLC (horizontal axis). Everything above the diagonal line shows the better performance of iDeepLC compared with DeepLC. Two data sets are used: (A) DIA HF (bigger data set) and (B) HeLa DeepRT (smaller data set).

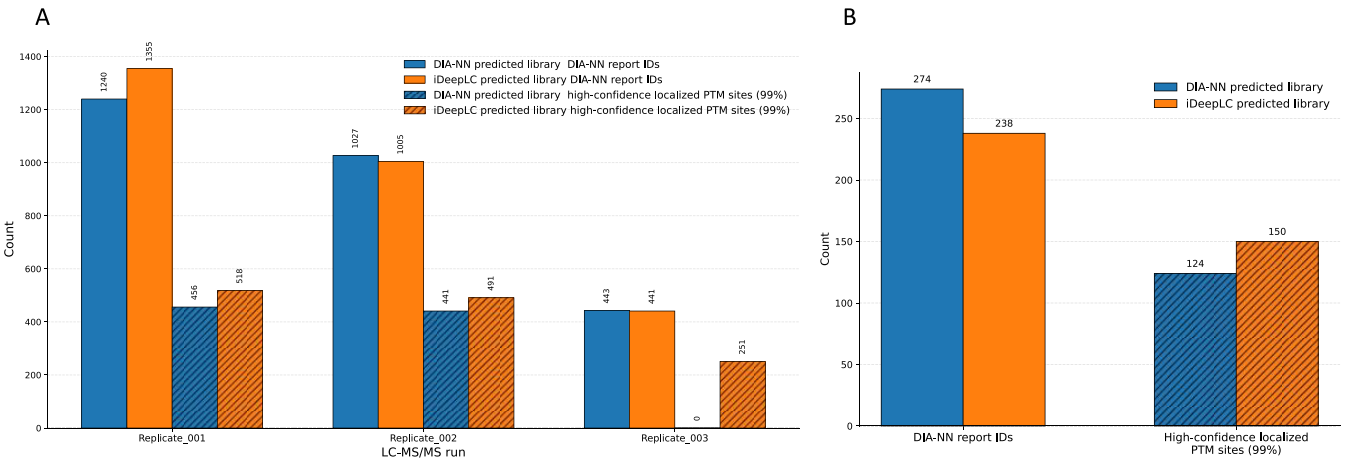


Figure 7. Impact of iDeepLC-predicted retention times on downstream calf DIA analysis. Comparison of DIA-NN analysis results obtained using the predicted spectral library retention times by DIA-NN (blue) and a library where retention times were replaced with iDeepLC predictions (orange). (A) Three replicate LC−MS/MS runs of calf histone data and (B) MOLM-13 histone data LC-MS/MS run. Solid bars represent the total number of precursor identifications and hatched bars indicate the number of high-confidence localized PTM sites (localization probability ≥99%).

As expected, this *glycine evaluation* shows that encoding amino acids as themselves results in a lower MAE in most cases (Figure 4). This improved performance can be seen for the individual amino acids in Figure 4 positioned above the diagonal line, where the distance to the diagonal line quantifies this improvement over the baseline where amino acids are encoded as glycine. This also means that amino acids below the diagonal line have the reverse conclusion, encoding the amino acid as glycine performs better. Notably, the largest improvements over the baseline are achieved for hydrophobic amino acids. Indeed, correctly encoding amino acids that have the largest contribution to a peptide's hydrophobicity, and thus retention time in reversed-phase chromatography, greatly impacts the prediction accuracy. Importantly, this large hydrophobic contribution of specific amino acids is very effectively captured in the iDeepLC model. iDeepLC does show slightly worse performance for amino acids like lysine (K), but this is likely due to its distinct polar properties and small training data set size after excluding tryptic peptides.

We summarized the results of the *glycine evaluation* for 15 data sets with a reverse-phase stationary phase in Figure 5. This figure shows the relative MAE for each amino acid in all reverse-phased data sets when encoded as themselves or when encoded as glycine. Among the 19 amino acids, encoding the amino acid as itself showed improvements for 11 amino acids (A, E, F, I, L, M, P, R, V, W, and Y). For three amino acids (C, H, and K), the results for encoding the amino acids were worse compared with the baseline. Finally, for the remaining five amino acids (D, N, Q, S, T), there is a minimal difference between encoding them as their respective amino acid and encoding them as glycine.

Instead of comparing iDeepLC to a baseline here, the model was also compared to DeepLC with the same *Glycine evaluation* (Figure 6). In the DIA HF data set, iDeepLC has

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

higher prediction accuracy in 12 out of 19 cases, while for the much smaller data set, HeLa DeepRT, a lower MAE was achieved in 14 out of 19 cases. Notably, in both data sets, most of the hydrophobic amino acids such as tryptophan (W) are much better predicted by iDeepLC. Furthermore, for the cases in which iDeepLC is outperformed by DeepLC, there is only a small gain in prediction accuracy.

Downstream Impact of iDeepLC Predictions

To assess whether improved retention time predictions translate into practical gains in proteomics data analysis, we compared DIA-NN results obtained with the built-in predicted spectral library and one obtained with iDeepLC for predicting the retention times (Figure 7A,B). For calf histone data (Figure 7A), the evaluation was performed using a data set of derivatized histone peptides generated with a PROP−PIC workflow, comprising highly modified peptides with complex PTM patterns, thus representing a challenging test case for retention time prediction models. Across the three DIA runs in Figure 7, the total number of precursor identifications remained comparable between the two analyses, with one run showing an increase of about 9% (from 1240 to 1355 identifications) when using iDeepLC predictions. More notably, the number of confidently localized PTM sites (localization probability ≥99%) increased substantially with iDeepLC from 456 and 441 to 518 and 491 for replicate 1 and replicate 2, respectively. In the third replicate, no PTM sites reached the 99% localization threshold using the default spectral library, whereas 251 confidently localized sites were obtained when using the iDeepLC-predicted retention times.

In the MOLM-13 histone analysis (Figure 7B), while the total number of report IDs was slightly lower with iDeepLC (238 versus 274), the number of high-confidence localized PTM sites (99%) increased from 124 to 150. These results indicate that improved retention time prediction can enhance confident PTM site localization even when the total number of identifications does not increase.

CONCLUSIONS & DISCUSSION

Our results show that adding structural information for each modified residue in the form of SMILES and the molecular characteristics with the MolLogP descriptor enables iDeepLC to learn the physicochemical characteristics of modifications. This understanding means that their impact on the LC retention time is more accurately predicted. Notably, iDeepLC delivers state-of-the-art predictive accuracy for unmodified peptides while also exhibiting superior performance for modified peptides. This was highlighted in two evaluations, where iDeepLC outperforms other retention time predictors for specific modifications and predicts retention times for previously unobserved amino acids.

While iDeepLC improves prediction for small- to medium-sized, well-defined modifications, large and heterogeneous modifications (such as glycosylation, ubiquitination, and lipidation) present a challenge for the current model. These modifications often involve complex and conformation-dependent interactions that cannot be adequately captured by the simplified encoding of modifications as a single feature. In addition to these benchmarking experiments, the practical value of improved retention time prediction was demonstrated in a downstream DIA analysis of a histone modification data set. Replacing the spectral library retention times with iDeepLC predictions resulted in comparable or improved numbers of precursor identifications for two independent histone LC-MS experiments. More importantly, iDeepLC increased the number of high-confidence-localized PTM sites. This suggests that improved retention time estimates can contribute directly to downstream proteomic analyses, especially for confident PTM localization in DIA workflows.

A practical consideration of iDeepLC is that it requires a structure, such as a SMILES representation, for amino acids and their modifications to calculate the MolLogP descriptor. Currently, iDeepLC supports 108 modifications, and users are able to add any custom modifications. As structural annotation of peptide modifications becomes increasingly accessible and more routinely incorporated into computational proteomics workflows, this requirement is expected to become less restrictive over time.

Overall, iDeepLC presents a promising method to generalize peptide modifications with the MolLogP. This allows the model to infer for many different modifications, even those not trained on, and thus substantially impact downstream analysis.

ASSOCIATED CONTENT

Data Availability Statement

All data used to train and evaluate iDeepLC are available on Zenodo at 10.5281/zenodo.15011301 and it contains the following projects: HeLa hf,[31] ProteomeTools,[41] SWATH library,[42] Plasma lumos 1h,[43] DIA HF,[4] HeLa lumos 2h,[43] Pancreas,[44] Xbridge,[45] ATLANTIS SILICA,[45] LUNA SILICA,[45] LUNA HILIC,[45] SCX,[46] Yeast 2h,[47] HeLa lumos 1h,[43] Yeast 1h,[47] *Arabidopsis*,[48] Yeast DeepRT,[49] ProteomeTools PTM,[33] Plasma lumos 2h,[43] and HeLa DeepRT.[34] In addition, the data generated for the calf histone DIA experiment have been deposited in PRIDE[50] under accession PXD075498. Raw data of the calf histone cell sample are available on Zenodo at 10.5281/zenodo.19230872. Raw data of the MOLM-13 cell line sample are available on Zenodo at 10.5281/zenodo.19230872.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.5c08017.

Hyperparameter distributions and final values used for model optimization (Table S1); overview of the 20 LC-MS data sets used for training, validation, and testing, including data set origin, peptide counts, repository identifiers, and column types (Table S2); paired statistical comparison of iDeepLC and DeepLC performance across the 14PTMs evaluation based on one-sided Wilcoxon signed-rank tests with Benjamini−Hochberg correction (Table S3); MAE comparison of iDeepLC and DeepLC on encoded test sets with percentage differences relative to the not-encoded baseline (Table S4); schematic overview of the five-branch iDeepLC neural network architecture (Figure S1); performance comparison of iDeepLC, AlphaPept-Deep, Chronologer, and DeepLC across 20 data sets using relative MAE (Figure S2); scatter plots observed against predicted retention times for iDeepLC and DeepLC models across all 14 modifications (Figure S3); relationship between modification hydrophobicity (MolLogP) and improvement of iDeepLC over DeepLC in the 14PTMs evaluation (Figure S4); performance comparison of iDeepLC, AlphaPeptDeep, Chronologer,

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

and DeepLC across the 14PTMs evaluation (Figure S5) (PDF)

AUTHOR INFORMATION

Corresponding Author

Lennart Martens − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium; BioOrganic Mass Spectrometry Laboratory (LSMBO), IPHC UMR 7178, University of Strasbourg, CNRS, Strasbourg 67000, France; Infrastructure Nationale de Protéomique ProFI−FR2048, Strasbourg 67087, France*;
Email: lennart.martens@vib-ugent.be

Authors

Alireza Nameni − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0002-7471-9195

Arthur Declercq − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0002-9376-1399

Ralf Gabriels − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0002-1679-1711

Robbe Devreese − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0002-3432-1502

Sven Degroeve − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0001-8349-3370

Amélie De Maesschalck − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Department of Pharmaceutics, Ghent University, Ghent 9000, Belgium*

Maarten Dhaenens − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Department of Pharmaceutics, Ghent University, Ghent 9000, Belgium;* orcid.org/0000-0002-9801-3509

Cristina Chiva − *Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona 08003, Spain; Universitat Pompeu Fabra, Barcelona 08003, Spain*

Eduard Sabidó − *Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona 08003, Spain; Universitat Pompeu Fabra, Barcelona 08003, Spain*

Robbin Bouwmeester − *CompOmics, VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0001-6807-7029

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.5c08017

Author Contributions

A.N., S.D., and R.B. conceived the study, designed the methodology and experiments, and analyzed the results. A.N. and R.B. wrote the manuscript text and made the tool available in a Python package. A.N. developed the graphical user interface. R.B. supervised the research. A.D., R.G., R.D., M.D., A.M., C.C., and E.S. reviewed and revised the manuscript. M.D., A.M., C.C., and E.S. curated the experimental data and acquired the LC-MS histone data. R.B. and L.M. validated, reviewed, and revised the manuscript. All authors reviewed the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.N. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement N° 956148. S.D. acknowledges funding from the European Union's Horizon 2020 Programme grant no. H2020-INFRAIA-2018-1. R.D., A.D., R.G., L.M., and R.B. acknowledge funding from the Research Foundation Flanders (FWO) [1SH9O24N, 12B7123N, 1SE3724N, G010023N, G028821N, 12A6L24N]. L.M. acknowledges funding from the Horizon Europe Projects BAXERNA 2.0 [101080544] and COMBINE [101191739], and from the Ghent University Concerted Research Action [BOF21/GOA/033]. L.M. is further supported by the CHIST-ERA project ODEEP-EU [G0GDV23N]. C.C. and E.S. acknowledge support from the Spanish Ministry of Science, Innovation, and Universities (PID2024-157336NB-I00, CEX2020-001049-S), and the Generalitat de Catalunya through the CERCA programme and The Departament de Recerca i Universitats de la Generalitat de Catalunya (2021SGR01225). The CRG/UPF Proteomics Unit is part of the Spanish Infrastructure for Omics Technologies (ICTS OmicsTech).

REFERENCES

(1) Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785−1793.

(2) Shishkova, E.; Hebert, A. S.; Coon, J. J. Now, More Than Ever, Proteomics Needs Better Chromatography. *Cell Systems* **2016**, *3*, 321−324.

(3) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput. *Nat. Methods* **2020**, *17* (1), 41−44.

(4) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Xuan, Y.; Sondermann, J.; Schmidt, M.; Gomez-Varela, D.; Reiter, L. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics* **2017**, *16* (12), 2296−2309.

(5) Zolg, D. P.; Gessulat, S.; Paschke, C.; Graber, M.; Rathke-Kuhnert, M.; Seefried, F.; Fitzemeier, K.; Berg, F.; Lopez-Ferrer, D.; Horn, D.; Henrich, C.; Huhmer, A.; Delanghe, B.; Frejno, M. INFERYS Rescoring: Boosting Peptide Identifications and Scoring Confidence of Database Search Results. *Rapid Commun. Mass Spectrom.* **2021**, *39*, No. e9128, DOI: 10.1002/rcm.9128.

(6) C Silva, A. S.; Bouwmeester, R.; Martens, L.; Degroeve, S. Accurate Peptide Fragmentation Predictions Allow Data Driven

Approaches to Replace and Improve upon Proteomics Search Engine Scoring Functions. *Bioinformatics* **2019**, *35* (24), 5243−5248.
(7) Declercq, A.; Bouwmeester, R.; Hirschler, A.; Carapito, C.; Degroeve, S.; Martens, L.; Gabriels, R. MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol. Cell. Proteomics* **2022**, *21* (8), No. 100266.
(8) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**, *26* (7), 966−968.
(9) Yang, K. L.; Yu, F.; Teo, G. C.; Li, K.; Demichev, V.; Ralser, M.; Nesvizhskii, A. I. MSBooster: Improving Peptide Identification Rates Using Deep Learning-Based Features. *Nat. Commun.* **2023**, *14* (1), No. 4539.
(10) Bertsch, A.; Jung, S.; Zerck, A.; Pfeifer, N.; Nahnsen, S.; Henneges, C.; Nordheim, A.; Kohlbacher, O. Optimal de Novo Design of MRM Experiments for Rapid Assay Development in Targeted Proteomics. *J. Proteome Res.* **2010**, *9* (5), 2696−2704.
(11) Dorfer, V.; Maltsev, S.; Winkler, S.; Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *J. Proteome Res.* **2018**, *17* (8), 2581−2589.
(12) Kösters, M.; Leufken, J.; Leidel, S. A. SMITER—A Python Library for the Simulation of LC-MS/MS Experiments. *Genes (Basel).* **2021**, *12* (3), No. 396.
(13) Gong, S.; Gaccioli, F.; Dopierala, J.; Sovio, U.; Cook, E.; Volders, P. J.; Martens, L.; Kirk, P. D. W.; Richardson, S.; Smith, G. C. S.; Charnock-Jones, D. S. The RNA Landscape of the Human Placenta in Health and Disease. *Nat. Commun.* **2021**, *12* (1), No. 2639.
(14) Pham, T. V.; Nguyen, V. V.; Vu, D.; Henneman, A. A.; Richardson, R. A.; Piersma, S. R.; Jimenez, C. R. A Transformer Architecture for Retention Time Prediction in Liquid Chromatography Mass Spectrometry-Based Proteomics. *Proteomics* **2023**, *23* (7−8), No. e2200041.
(15) Wilburn, D. B.; Shannon, A. E.; Spicer, V.; Richards, A. L.; Yeung, D.; Swaney, D. L.; Krokhin, O. V.; Searle, B. C.Deep Learning from Harmonized Peptide Libraries Enables Retention Time Prediction of Diverse Post Translational Modifications *bioRxiv* 2023 DOI: 10.1101/2023.05.30.542978.
(16) Ma, C.; Ren, Y.; Yang, J.; Ren, Z.; Yang, H.; Liu, S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **2018**, *90* (18), 10881−10888.
(17) Guan, S.; Moran, M. F.; Ma, B. Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. *Mol. Cell. Proteomics* **2019**, *18* (10), 2099−2107.
(18) Wen, B.; Li, K.; Zhang, Y.; Zhang, B. Cancer Neoantigen Prioritization through Sensitive and Reliable Proteogenomics Analysis. *Nat. Commun.* **2020**, *11* (1), No. 1759.
(19) Lou, R.; Liu, W.; Li, R.; Li, S.; He, X.; Shui, W. DeepPhospho Accelerates DIA Phosphoproteome Profiling through in Silico Library Generation. *Nat. Commun.* **2021**, *12* (1), No. 6685.
(20) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H. C.; Aiche, S.; Kuster, B.; Wilhelm, M. Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nat. Methods* **2019**, *16* (6), 509−518.
(21) Zeng, W. F.; Zhou, X. X.; Willems, S.; Ammar, C.; Wahle, M.; Bludau, I.; Voytik, E.; Strauss, M. T.; Mann, M. AlphaPeptDeep: A Modular Deep Learning Framework to Predict Peptide Properties for Proteomics. *Nat. Commun.* **2022**, *13* (1), No. 7238.
(22) Tarn, C.; Zeng, W. F. PDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning. *Anal. Chem.* **2021**, *93* (14), 5815−5822.
(23) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications. *Nat. Methods* **2021**, *18* (11), 1363−1369.
(24) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(25) Stefaniu, A.; Crippen, G. M.Molecular Descriptors and Properties of Organic Molecules *Symmetry (Group Theory) Math. Treat. Chem.* 2018 DOI: 10.5772/intechopen.72840.
(26) Bouwmeester, R.; Martens, L.; Degroeve, S. Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. *Anal. Chem.* **2019**, *91* (5), 3694−3703.
(27) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868−873.
(28) Fukushima, K. Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition. *Neural Network* **1988**, *1* (2), 119−130.
(29) Landrum, G. RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling. 2010 http://rdkit.sourceforge.net.
(30) Parker, J. M. R.; Guo, D.; Hodges, R. S. New Hydrophilicity Scale Derived from High-Performance Liquid Chromatography Peptide Retention Data: Correlation of Predicted Surface Residues with Antigenicity and x-Ray-Derived Accessible Sites. *Biochemistry* **1986**, *25* (19), 5425−5432.
(31) Kelstrup, C. D.; Bekker-Jensen, D. B.; Arrey, T. N.; Hogrebe, A.; Harder, A.; Olsen, J. V. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **2018**, *17* (1), 727−738.
(32) Biewald, L. Experiment Tracking with Weights and Biases, https://www.wandb.com/ (accessed June 19, 2024).
(33) Zolg, D. P.; Wilhelm, M.; Schmidt, T.; Médard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. ProteomeTools: Systematic Characterization of 21 Post-Translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (Lc-Ms/Ms) Using Synthetic Peptides. *Mol. Cell. Proteomics* **2018**, *17* (9), 1850−1863.
(34) Sharma, K.; D'Souza, R. C. J.; Tyanova, S.; Schaab, C.; Wiśniewski, J. R.; Cox, J.; Mann, M. Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Rep.* **2014**, *8* (5), 1583−1594.
(35) He, K.; Zhang, X.; Ren, S.; Sun, J. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015; pp 1026−1034.
(36) Nair, V.; Hinton, G. E. *Rectified Linear Units Improve Restricted Boltzmann Machines*, Proceedings of the 27th International Conference on Machine Learning (ICML-10); 2010; pp 807−814.
(37) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S.Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), arXiv:1511.07289. arXiv.org e-Print archive. https://arxiv.org/abs/1511.07289 2015.
(38) Agarap, A. F.Deep Learning Using Rectified Linear Units (ReLU), arXiv:1803.08375. arXiv.org e-Print archive. https://arxiv.org/abs/1803.08375 2018.
(39) Maile, T. M.; Izrael-Tomasevic, A.; Cheung, T.; Guler, G. D.; Tindell, C.; Masselot, A.; Liang, J.; Zhao, F.; Trojer, P.; Classon, M.; Arnott, D. Mass Spectrometric Quantification of Histone Post-Translational Modifications by a Hybrid Chemical Labeling Method. *Mol. Cell. Proteomics* **2015**, *14* (4), 1148−1158.
(40) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(41) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H. C.; Weininger, M.; Yu, P.; Schlegl, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717

ProteomeTools Based on a Complete Synthetic Human Proteome. *Nat. Methods* **2017**, *14* (3), 259−262.
(42) Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Collins, B. C.; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; Faini, M.; Schubert, O. T.; Faridi, P.; Ebhardt, H. A.; Matondo, M.; Lam, H.; Bader, S. L.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Tate, S.; Aebersold, R. A Repository of Assays to Quantify 10,000 Human Proteins by SWATH-MS. *Sci. Data* **2014**, *1*, No. 140031.
(43) Li, W.; Chi, H.; Salovska, B.; Wu, C.; Sun, L.; Rosenberger, G.; Liu, Y. Assessing the Relationship Between Mass Window Width and Retention Time Scheduling on Protein Coverage for Data-Independent Acquisition. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (8), 1396−1405.
(44) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; Frejno, M.; Schmidt, T.; Schnatbaum, K.; Wilhelm, M.; Ponten, F.; Uhlen, M.; Gagneur, J.; Hahne, H.; Kuster, B. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* **2019**, *15* (2), No. e8503.
(45) Krokhin, O. V.; Ezzati, P.; Spicer, V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Data Collection Methods and Features of Additive and Sequence-Specific Models. *Anal. Chem.* **2017**, *89* (10), 5526−5533.
(46) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O. V. Sequence-Specific Model for Peptide Retention Time Prediction in Strong Cation Exchange Chromatography. *Anal. Chem.* **2017**, *89* (21), 11795−11802.
(47) Jarnuczak, A. F.; Lee, D. C. H.; Lawless, C.; Holman, S. W.; Eyers, C. E.; Hubbard, S. J. Analysis of Intrinsic Peptide Detectability via Integrated Label-Free and SRM-Based Absolute Quantitative Proteomics. *J. Proteome Res.* **2016**, *15* (9), 2945−2959.
(48) Mucha, S.; Heinzlmeir, S.; Kriechbaumer, V.; Strickland, B.; Kirchhelle, C.; Choudhary, M.; Kowalski, N.; Eichmann, R.; Hückelhoven, R.; Grill, E.; Kuster, B.; Glawischnig, E. The Formation of a Camalexin Biosynthetic Metabolon. *Plant Cell* **2019**, *31* (11), 2697−2710.
(49) Nagaraj, N.; Alexander Kulak, N.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. System-Wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-Shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell. Proteomics* **2012**, *11* (3), No. M111.013722, DOI: 10.1074/mcp.M111.013722.
(50) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543−D552.

https://doi.org/10.1021/acs.analchem.5c08017
Anal. Chem. 2026, 98, 17707−17717