International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 78-87
https://doi.org/10.70917/ijcisim-2026-3745
Article

TMT-SpecGen: A Computational Utility for Boosting Peptide Identification via TMT Spectral Library Augmentation

Shraddha Kumar[1], Anuradha Purohit[2], Sunita Varma[3]

[1] Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore, M. P., India.
Email: shraddhakumar@sgsits.ac.in
[2] Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore, M. P., India.
Email: apurohit@sgsits.ac.in
[3] Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore, M. P., India.
Email: sverma19@sgsits.ac.in

Abstract: — Multiplexed Tandem Mass Tag (TMT) proteomics is a well-established technique widely employed for large-scale peptide quantification and identification. However, traditional sequence database searches for peptide identification are obstructed by the effects of TMT labeling. Spectral library searches offer improved detection capability (true positives) and selectivity (true negatives) for peptide identification in TMT-labeled mass spectrometry data. In this work, we aim to increase peptide identification by enhancing spectral library. The fundamental principle underlying the proposed work is that a peptide's tandem MS fragmentation pattern, under consistent conditions, is reproducible. This reproducibility enables the identification of unknown spectra acquired under similar conditions through spectral matching. To achieve this, we introduce TMT-SpecGen (Tandem Mass Tag- Spectrum Generator), a biologically inspired utility that generates annotated spectra using a deep learning model. It draws concepts from the way the human brain processes information, learns from experience, and adapts to new data. The model is trained on labeled spectra obtained from clustering and consensus generation. Utilizing consensus spectra for training reduces training time by 40% while improving accuracy of spectrum prediction by 5%. From 163,615 predictions, we constructed a high-quality simulated TMT spectral library containing 96,631 robust peptides derived from millions of peptide-spectrum matches. The simulated library demonstrates its efficacy in enhancing peptide identification, as 14% of previously unidentified spectra, when searched in the library, were successfully identified.
Keywords: — Deep learning, TMT-label, spectral library, annotated spectra, spectrum prediction, peptide identification

1. Introduction

Cancer remains the primary contributor to global deaths and presents a complex public health challenge, necessitating ongoing and extensive scientific research [1]. Over the past few years, substantial advancements have been made in understanding cancer, largely driven by the advent of omics-based technologies. Among these, proteomics—the systematic in-depth examination of the proteome, representing all proteins produced within a cell, tissue, or organism—has emerged as a pivotal field of study [2]. Dysregulation of the proteome is frequently associated with cancer, often resulting in unusual protein expression [3]. Proteomic analysis typically involves the examination of peptide fragments generated through enzymatic digestion, most commonly with trypsin [4]. To optimize peptide quantification and identification, chemical labeling techniques, including tandem mass tags (TMTs) and isobaric tags for relative and absolute quantification (iTRAQs), are employed. These TMTs generate distinct reporter ions upon fragmentation, facilitating the simultaneous multiplexed analysis. This capability has established TMTs as a widely employed technique in protein biomarker discovery [5]. For peptide identification, the two commonly used methods are traditional database searching and spectral library matching.

Traditional database search engines match experimental tandem mass spectra (MS/MS) against theoretical spectra for peptide identification. It is typically performed using specialized software such as Proteome Discoverer, which annotates spectra using peptide reference databases (e.g., PeptideAtlas, PepQuery, and UniProt) [6], [7], [8]. Database searching is significant for the identification of novel peptides. However, it is limited by challenges such as unidentified spectra resulting from low signal-to-noise ratios, the absence of certain peptides in reference databases, and unanticipated modifications, including post-translational modifications (PTMs). As a result, 70–75% of spectra (known as 'dark proteome') in typical experiments remain unidentified [9], [10].

Spectral library searching is another method for peptide identification that relies on experimentally derived spectra rather than in silico predictions. Spectral library searching is efficient, reusing previously identified spectra and requiring less computational time. It has demonstrated higher sensitivity compared to traditional database searching; however, a significant limitation of current spectral libraries is their incompleteness, often due to a scarcity of high-quality spectral data, and it is challenging to integrate spectral library searches into existing workflows. Peptides lacking corresponding spectra in the reference library remain unidentified [9], [11]. Constructing a TMT-specific spectral library is essential for accurately analyzing TMT-labeled MS data [12]. Subsequent advancements focused on large-scale data clustering millions of spectra from the theoretical database, validating peptide identifications, and creating spectral libraries [13]

Deep learning offers powerful alternatives for peptide identification in mass spectrometry-based proteomics and has significantly advanced mass spectrometry spectrum prediction [14]. Accurate prediction of fragmentation patterns in MS/MS spectra improves peptide identification in proteomics. It works on the concept that peptide fragmentation patterns under fixed experimental conditions are reproducible. This allows the identification of unknown spectra through spectral matching [15]

Deep learning models excel in predicting and simulating MS/MS spectra due to their accuracy and ability to handle complex scenarios. Algorithms such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been effectively utilized for spectrum prediction tasks. CNNs are adept at identifying local patterns within spectra, while RNNs excel at capturing sequential dependencies. By leveraging a combination of these architectures, cutting-edge performance in spectral prediction can be achieved, including the identification of previously unannotated spectra within the dark proteome. A major concern with DNN is the training time and quality spectra required to train the model [14], [16]. Additionally, MS/MS experiments produce numerous highly similar spectra of the same peptide across different laboratories. To address this, spectral clustering techniques can be employed as an effective solution. Spectral clustering algorithms are designed to efficiently and accurately classify large sets of spectra based on their similarity, ensuring that all spectra within a given cluster correspond to the same peptides. These algorithms generate representative or consensus spectra for each resulting cluster [17]. The consensus spectra can subsequently be used to train deep-learning models. This significantly reduces the amount of data to be analysed, thereby saving training time and increasing the quality of predicted spectra [18].

This research work proposes the TMT-SpecGen model for high-quality Tandem Mass Tag Spectra generation to augment in silico spectral libraries and improve peptide identification. The results demonstrate that TMT-SpecGen successfully identified 14% of previously unidentified spectra from the dark proteome. Thus, TMT-SpecGen reduces the dark cancer proteomes critical for the identification of cancer-specific proteins, paving the way for the discovery of novel therapeutic targets and biomarkers.

2. Proposed TMT-SpecGen Model

In this study, we present TMT-SpecGen: A Deep Learning-Based Solution for High-Quality Tandem Mass Tag Spectra Generation to augment in silico spectral libraries and improve peptide identification. The model utilizes spectrum clustering and consensus generation to reduce dataset complexity and enhance spectral quality before network training. The deep learning model generates annotated spectra, which are validated and incorporated into spectral libraries. A schematic of the proposed TMT-SpecGen model is demonstrated in Fig. 1, where (A) TMT experimental data is input into the model. (B) The Clustering and Consensus Generation module is used to generate representative spectra. (C) Preprocessed and representative spectra are searched in the sequence database tool (D). TMT spectra are bifurcated into identified and Unidentified spectra. (E) Identified spectra are labelled and paired with unlabelled spectra from the label-free spectral library. These spectra are used to train a Deep Neural Network. Some identified spectra are left unpaired for testing. (F) A Deep Neural Network is used to predict annotated spectra. It uses convolutional neural networks to learn features and LSTM with attention

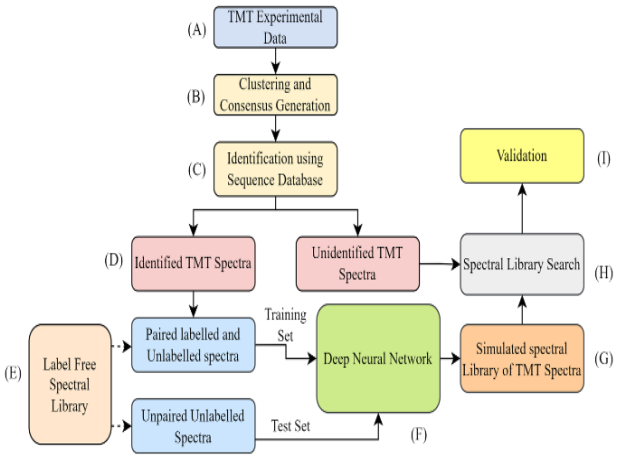

Fig. 1. Schematic of the TMT-SpecGen model.

mechanism to preserve TMT peaks. (G) Robust high quality

Spectra are stored in the TMT spectral library. (H) The simulated spectral library was then searched for previously unidentified spectra. (I) Newly identified spectra are then validated to check the authenticity of the identification.

A. TMT Experimental Dataset

In this study, we have used the COREAD (COloRectal cAncer cell Dataset) of mass spectra to train and evaluate the performance of the proposed model. The dataset contains a comprehensive functional characterization of cancer cell lines. These cell lines are instrumental in biochemical experimental designs, where their specific biological characteristics can influence experimental outcomes. The COREAD data sets contain 1 million spectra stored in Mascot Generic Format (MGF) with a file size of approximately 2.62 GB.

B. Clustering and Consensus Generation

The first spectral clustering tool, msCRUSH [19], [20], is used to cluster similar types of spectra. It involves Locality Similarity Hashing (LSH) and cosine similarity to check the similarity between each pair of spectra of the same charge. Here, the outputs of clustering are specified with a separate text file each for different charges, such as 2+ to 10+. For consensus generation, we have performed a two-step process as follows:

1) Preprocess Workflow

   i) Input clustered spectrum files.

   ii) Store all peaks (mz and intensity values) in the TMT range.

   iii) Split the spectrum into bins of size X (e.g., 100) according to the mz values of peaks.

   iv) Sort peaks in Bins in descending order of intensity values.

   v) Retain the top N peaks from each bin.

   vi) Take the square root of the intensity values of the retained peaks.

   vii) Normalize TMT peaks.

   viii) Rescale all peaks to a maximum intensity of 1000.

   ix) Write processed spectra to a new MGF file.

2) Consensus Workflow

   i) Input processed spectra.

   ii) Retrieve and store current spectrum's scan number, charge state, retention time, and peptide mass.

iii) Extract peaks in the TMT range.

iv) For non-TMT peaks, find peak match (i.e., error = current mz – consensus mz, and error < minimum tolerance).

v) If the peaks match, sum the intensities of the two peaks to generate a new consensus peak; else, copy the current peak to the new peak.

vi) Get the mean mz for the two peaks.

vii) Insert extracted TMT peaks into the spectra.

viii) Repeat steps from (i) to (vii) to generate consensus spectra.

ix) Write consensus spectra to a new MGF file, sorted by the m/z value of peaks.

C. Identification using a sequence database search

Following clustering and consensus-spectrum generation, each dataset's mass spectra were analyzed using the Proteome Discoverer v2.4 tool. We matched consensus spectra to reference peptide sequences derived from known proteins by employing the software's built-in consensus and processing workflows (see Fig. 1). The resulting output comprised peptide–spectrum matches (PSMs), which served as the basis for our subsequent analyses.

1) Identified TMT spectra

To create labelled (annotated) spectra, 'First Scan' numbers are extracted from the PSMs and matched with the 'Scan' number present in the consensus file, and the corresponding peptide sequence is assigned to the spectra. The annotated spectra with Posterior Error Probability (PEP) values less than 0.01 were considered to train the deep learning model.

2) Unidentified TMT spectra (Spectra NOT identified in PD search)

Analysis of all spectra not identified in the PD search shows that some of these have high PEP scores and some have unexpected modifications due to oxidation, indicating that these spectra are probably of poor quality. The sequences that are not identified but with low PEP score were used during validation to test the simulated spectral library.

D. Label-free spectral libraries

Based on the spectral composition of each cluster, we further stratified them into two groups: paired labeled (annotated) spectra and unpaired unlabeled (unannotated) spectra. This classification enabled the identification and subsequent analysis of unannotated cancer-related clusters.

1) Paired Labelled and Unlabelled spectra:

The labelled spectra were then paired with unlabelled spectra, and these spectra were used in training the deep learning model.

2) Unpaired Unlabelled spectra:

Along with paired spectra, unpaired and unlabelled spectra are used to test the deep learning model and to create a more generalized trained model for spectrum prediction.

E. Deep Neural Network

- *Training the new TMT model*

1) Creating annotated consensus spectra: The consensus spectra are matched with the spectra in the theoretical database to get the corresponding peptide sequence to annotate spectra, which is then used as training input.

2) Training new hybrid model using annotated spectra: A hybrid CNN-LSTM deep neural network (Fig. 2.) with attention mechanisms constructed, featuring custom layers, residual blocks with dropout, and sequential dependency modeling. CNN layers with coordinate channels are used for feature extraction. CNN layers are then integrated with LSTM layers to capture sequential dependencies and output dense layers for effective spectrum reconstruction. The model is compiled with Adam optimizer and cosine similarity loss. Finally, the model is trained and validated using preprocessed data, leveraging callbacks for optimization, and the trained model is saved for future use.

The deep learning model architecture includes [A] Annotated spectra as input. [B] Conv1D (64 filters, kernel sizes 2 to 9): for extracting features from sequential data, followed by concatenation(axis=-1) to combine the outputs of different convolutional layers. It also includes Conv1D (512 filters, kernel size=1): point-wise convolution. The layer will learn 512 different feature representations of the input. Each filter is applied to the input sequence, extracting different kinds of features. A kernel size of 1 means that each filter looks at only a single element at a time. [D] 8-Residual Blocks: to train deeper networks by addressing the vanishing gradient problem along with additional 3-Residual Blocks without squeeze-and-excitation to learn deeper representations without degradation and help improve the convergence speed, acting as a form of regularization. Dropout layers are included to prevent overfitting. [E] Attention Mechanisms incorporate attention layers, especially in LSTM blocks, to allow the model to focus on important parts of the sequence. [F] Combines multiple attention outputs to improve performance. [G] Long Short-Term Memory (LSTM) networks capture sequential dependencies in the spectral data. It combines time distributed two time-distributed LSTM blocks, each of 256 units, into a dense layer. [H] The Output layer generates the predicted spectrum, which is compared to the actual spectrum. In the output layer, Conv1D (K=1) with sigmoid activation serves as a transformation layer that effectively decides which features to pass on to the next stage. Global Average Pooling in the output layer reduces the dimensionality of the feature map, condensing spatial information and reducing model size and complexity. Isobaric labels (chemical masses) of the TMT peaks are included in training the model by adding additional masses to TMT-labeled peaks. The prediction task serves to inform the neural network of the relevance and impact of TMT labeling in the analysis process by pairing labeled and unlabeled spectra for the training set and unpaired, unlabeled spectra for validation (around 10%), and enforce the model to learn to balance between the m/z and intensity.

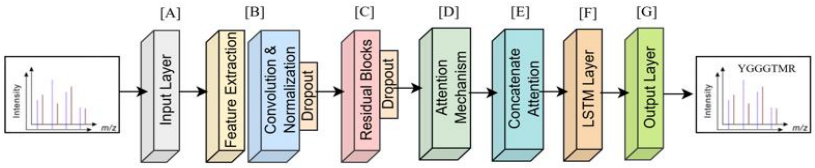

Fig. 2. CRNNA model to predict TMT-labelled annotated spectrum prediction.

3) Specifications of the training module:

The residual convolutional neural network (CNN) model using the Keras 3.0 framework and Tensorflow 3.1 at the back end is designed to predict MS/MS spectra. From a total of 522,212 spectra, the model has trained for a total of 443,879 (16plex) spectra, 212,736 charge 2+, and 186,463 charge 3+ spectra, and 15% of the total spectra are used for testing. It has been ensured that the training and testing spectra are non-overlapping. Moreover, since normalized collision energy (NCE) values are not available for all training data, we assume a default NCE value of 25% for all unlabeled data.

- *Spectra prediction and validation*

Once the training of the DNN is completed, the new model is used to predict TMT-labelled annotated spectra. The model has generated 163,615 spectra (Charge 2+ = 63486 and Charge 3+ = 85844). Experimental and predicted spectra are then compared using cosine similarity (normalized dot product). As the spectra are in the form of vectorscosine similarity captures directional alignment, so it is most suitable to compare spectra.

F. Simulated spectral library of TMT spectra

Predicted high-quality spectra are stored to create a simulated spectral library and can be used further for spectral library enhancement.

G. Spectral Library search

The unidentified spectra from step ‘C’ are then searched in the simulated spectral library for peptide identification.

H. Validation

Finally, validation of identified peptides is conducted by again searching corresponding spectra on the theoretical database and analysing the PEP values.

3. Results and Discussion

The spectral clustering and consensus generation module of TMT-SpecGen generates 22,728 consensus and 501,067 pre-processed spectra. Fig. 3 [A] show the results of clustering, excluding the clusters containing a single spectrum. These spectra, when searched in the theoretical database (Proteome Discoverer), produce 165,846 peptide-spectrum matches. The results of the Proteome Discoverer search shows that 32% (165,846) of the overall spectra are identified and 68% (356,366) remain unidentified, which we wanted to identify in our further investigation. The distribution of identified (57%) and unidentified (43%) consensus spectra according to PEP values

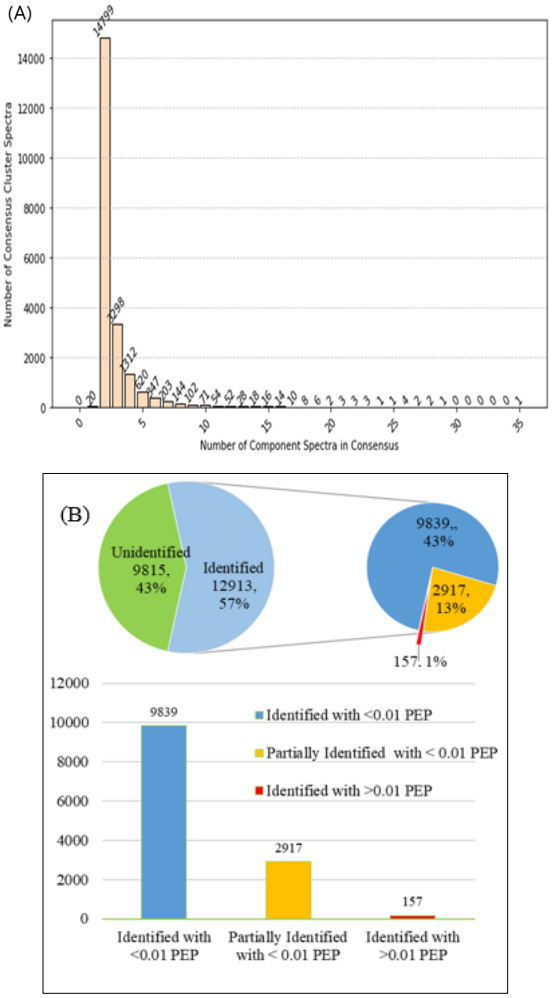

Fig. 3. (A) Distribution of spectra among different clusters. (B) Distribution of identified and unidentified consensus spectra according to PEP values

is shown in Fig. 3[B].

The identified consensus spectra include fully identified with a PEP score less than 0.01, partially identified with a PEP score less than 0.01, and identified with PEP score greater than 0.01. Identified spectra with a PEP score less than 0.01 are used further for training and testing. The consensus generation module processes the clustered spectra as mentioned above in section 3 to improve the quality of the spectra. Fig. 4 shows a comparison of PEP scores of processed and unprocessed spectra. -10log(x) of the PEP values has been taken for better visualization. The proposed TMT-SpecGen model is trained on the 123,522 paired labelled spectra. It is observed that the residual convolutional network structure is important for attaining

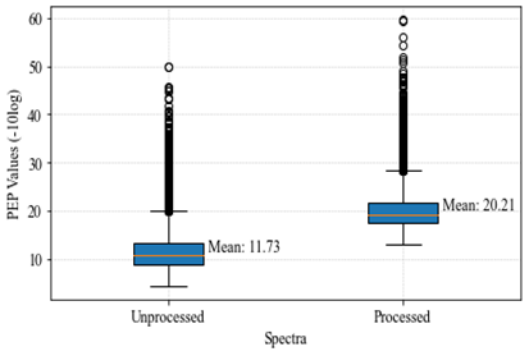

Fig. 4. Comparison of PEP values of unprocessed and processed spectra

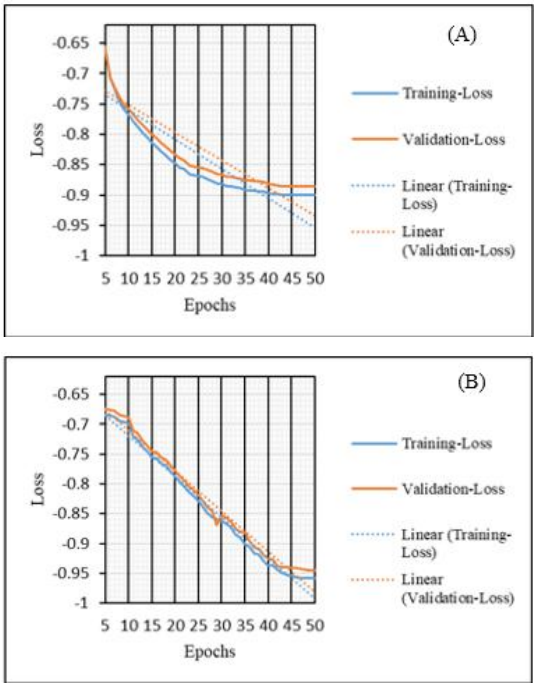

Fig. 5. Training and validation loss curve. (A) without clustering and consensus generation module [18]. (B) With the clustering and consensus generation module

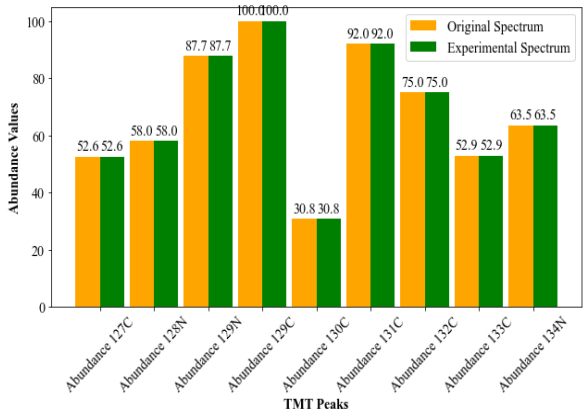

Fig. 6. Comparison of abundance values of TMT peaks of the predicted spectrum TTHVEGGDAGNREDQLNR

strong performance, as stated by the COREAD dataset training results. As shown by the training and validation loss curve in Fig. 5, the accuracy of the prediction consistently increases as the number of epochs increases and reaches the

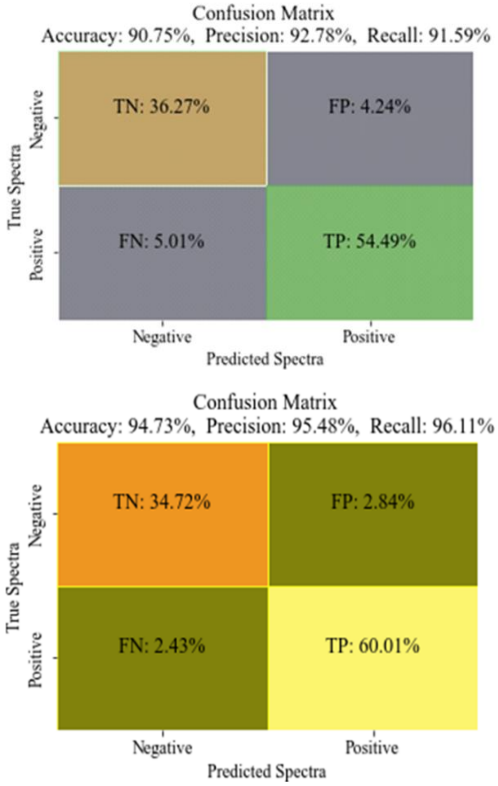

Fig. 7. Confusion matrix to evaluate the performance of the TMT-SpecGen model. (A) without clustering and consensus generation modules [18]. (B) With the clustering and consensus generation module

TABLE I. RESULT STATISTICS OF THE PROPOSED TMT-SPECGEN MODEL

Dataset: COREAD **Abundant modifications (Da):** Deamidation, oxidation				
Number of MS/MS Consensus spectra	**Un-identified spectra**	**Identified spectra**		**Identification rate raised by the Enhanced Spectral Library**
		Sequence Database	Enhanced Spectral Library	
22728	9815	12913	+1373 (14% of unidentified)	57% ⟶63%

maximum value of 0.95 in the 47th epoch, and it is not improved further by the additional epochs.

The abundance values of TMT peaks in the predicted and experimental MS/MS spectra are compared. As shown in Fig. 6, the predicted spectrum preserved all TMT peaks present in the original spectrum. Moreover, our learning algorithm successfully captures fragmentation patterns and accurately predicts full MS/MS spectra. In summary, the proposed model enforces stricter input constraints (like fixed peptide length) and is particularly well-suited for labeled proteomics (with TMT tagging). Finally, the performance evaluation of the proposed model is shown through the confusion matrix in Fig. 7. For assessment, around 78,330 spectra (mostly charge 2+ and 3+) are used, as they cover a major part of the total spectra. The cosine similarity function is used for comparison between the predicted and experimental MS/MS spectra (including TMT peaks). 96,631 predicted spectra are then used to create a simulated spectral library, which complements the existing spectral library. Unidentified consensus spectra, when searched in the enhanced spectral library, identify 14% of the previously unassigned spectra (as shown in Table 1).

4. Conclusion

The TMT-SpecGen model successfully addresses key limitations of existing deep learning methods for spectra prediction, particularly the preservation of TMT peaks and training time efficiency. By integrating spectral clustering and consensus generation techniques, the model effectively filters out low-quality spectra and generates representative spectra for clusters of similar spectra, thereby enhancing spectral quality and reducing training time by 40% (21.6 hrs as compared to 37.7 hrs of the previous state-of-the-art model [18]). The proposed workflow, which includes clustering, consensus generation, sequence database search, and spectral library construction, demonstrates a comprehensive approach to spectral enhancement. The clustering and consensus module processes large-scale mass spectrometry data, leading to the identification of high-quality spectra that contribute to an improved spectral library. The model leverages deep learning, specifically a hybrid CNN-LSTM neural network with attention mechanisms, to predict TMT-labeled spectra with high accuracy (0.95). Results have shown that 14% of the previously unidentified spectra are identified through the simulated spectral library. It indicates a significant improvement in peptide identification as compared to traditional approaches. The consensus generation module refines spectral data, resulting in a higher proportion of spectra with lower PEP scores, thus improving confidence (lower Type I and Type II error) in peptide identification. Additionally, the creation of a simulated spectral library from predicted high-quality spectra provides a valuable resource for using peptides in cancer research.

Future work

Future work could explore further optimizations in clustering techniques and neural network architectures to enhance predictive performance and expand the model's applicability to diverse proteomics datasets. The presented work can be carried further to integrate proteomic data with other omics layers, such as genomics and transcriptomics. This would enable healthcare providers to tailor therapies to the specific needs of patients based on their proteomic profiles, improving the chances of successful outcomes in cancer treatment and beyond.

Declaration of competing interest

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgment

The authors would like to thank Dr. James Wright for his constructive guidance and the Institute of Cancer Research (U.K.) for providing the dataset required for the training and validation of the deep learning model.

References

1. H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA Cancer J Clin, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.
2. R. P. Horgan and L. C. Kenny, "'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics," The Obstetrician & Gynaecologist, vol. 13, no. 3, pp. 189–195, Jul. 2011, doi: 10.1576/toag.13.3.189.27672.
3. Y. W. Kwon et al., "Application of Proteomics in Cancer: Recent Trends and Approaches for Biomarkers Discovery," Front Med (Lausanne), vol. 8, Sep. 2021, doi: 10.3389/fmed.2021.747333.
4. S. Sikdar, R. Gill, and S. Datta, "Improving protein identification from tandem mass spectrometry data by one-step methods and integrating data from other platforms," Brief Bioinform, vol. 17, no. 2, pp. 262–269, Mar. 2016, doi: 10.1093/bib/bbv043.
5. X. Chen, Y. Sun, T. Zhang, L. Shu, P. Roepstorff, and F. Yang, "Quantitative Proteomics Using Isobaric Labeling: A Practical Guide.," Genomics Proteomics Bioinformatics, vol. 19, no. 5, pp. 689–706, Oct. 2021, doi: 10.1016/j.gpb.2021.08.012.
6. E. W. Deutsch, H. Lam, and R. Aebersold, "PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows," EMBO Rep, vol. 9, no. 5, pp. 429–434, May 2008, doi: 10.1038/embor.2008.56.
7. B. Wen, X. Wang, and B. Zhang, "PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations," Genome Res, vol. 29, no. 3, pp. 485–493, Mar. 2019, doi: 10.1101/gr.235028.118.
8. C. Uniprot, "UniProt: the universal protein knowledgebase in 2021," Nucleic Acids Res, vol. 49, 2021.
9. Y. Perez-Riverol, J. A. Vizcaíno, and J. Griss, "Future Prospects of Spectral Clustering Approaches in Proteomics," Proteomics, vol. 18, no. 14, Jul. 2018, doi: 10.1002/pmic.201700454.
10. J. Griss et al., "Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets," Nat Methods, vol. 13, no. 8, pp. 651–656, Aug. 2016, doi: 10.1038/nmeth.3902.

11. J. Griss, “Spectral library searching in proteomics,” Proteomics, vol. 16, no. 5, pp. 729–740, Mar. 2016, doi: 10.1002/pmic.201500296.
12. J. Shen, V. R. Pagala, A. M. Breuer, J. Peng, Bin Ma, and X. Wang, “Spectral Library Search Improves Assignment of TMT Labeled MS/MS Spectra,” J Proteome Res, vol. 17, no. 9, pp. 3325–3331, Sep. 2018, doi: 10.1021/acs.jproteome.8b00594.
13. W. Xu, J. Kang, W. Bittremieux, N. Moshiri, and T. Rosing, “HyperSpec: Ultrafast Mass Spectra Clustering in Hyperdimensional Space,” J Proteome Res, vol. 22, no. 6, pp. 1639–1648, Jun. 2023, doi: 10.1021/acs.jproteome.2c00612.
14. B. Wen et al., “Deep Learning in Proteomics,” Proteomics, vol. 20, no. 21–22, Nov. 2020, doi: 10.1002/pmic.201900335.
15. H. Lam, “Building and Searching Tandem Mass Spectral Libraries for Peptide Identification,” Molecular & Cellular Proteomics, vol. 10, no. 12, p. R111.008565, Dec. 2011, doi: 10.1074/mcp.R111.008565.
16. J. G. Meyer, “Deep learning neural network tools for proteomics,” Cell Reports Methods, vol. 1, no. 2, p. 100003, Jun. 2021, doi: 10.1016/j.crmeth.2021.100003.
17. A. M. Frank et al., “Clustering Millions of Tandem Mass Spectra,” J Proteome Res, vol. 7, no. 1, pp. 113–122, Jan. 2008, doi: 10.1021/pr070361e.
18. S. Kumar, A. Purohit, and S. Varma, “TMT-Labeled Annotated Spectra Prediction using Deep Neural Network,” in 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, Dec. 2024, pp. 1271–1277. doi: 10.1109/CICN63059.2024.10847325.
19. L. Wang, S. Li, and H. Tang, “msCRUSH: Fast Tandem Mass Spectral Clustering Using Locality Sensitive Hashing,” J Proteome Res, p. acs.jproteome.8b00448, Dec. 2018, doi: 10.1021/acs.jproteome.8b00448.
20. S. Kumar, A. Purohit, and S. Varma, “Review of spectral clustering algorithms used in proteomics,” International Journal of Data Science, vol. 8, no. 1, p. 16, 2023, doi: 10.1504/IJDS.2023.129449.