DPHO: Prospects of Optimizing DIA Phosphoproteomics Using Deep Learning Techniques

Chenyu Liang, Ruichao Nie, Xueying Bao, Yufan Chen, Ziqi Yang
School of Informatics Xiamen University
24520230157441, 24520230157435, 24520231154536, 23020231154174, 23020231154248

Abstract

In the realm of phosphoproteomics, Data-Independent Acquisition (DIA) outperforms Data-Dependent Acquisition (DDA) in accuracy and reproducibility of quantification. However, DIA's reliance on spectral libraries, typically derived from DDA analyses, limits its throughput and proteome coverage. To address this, we introduce DPHO, a novel deep learning framework for generating in silico phosphopeptide libraries. By circumventing the construction of DDA libraries, DPHO streamlines the phosphoproteome profiling process. It offers enhanced phosphoproteome coverage and facilitates the discovery of more signaling pathways compared to conventional DDA-based methods. Utilizing a synthetic phosphopeptide mixture from HeLa cell lysate and advanced mass spectrometry techniques, DPHO has demonstrated superior efficiency in site localization. The framework's adaptability is further evident in its compatibility with various fragmentation methods, including Multistage Activation (MSA), Electron Transfer Dissociation (ETD), and Higher Energy Collisional Dissociation (HCD). In conclusion, we aim to develop a phosphorylation site prediction tool to achieve faster and more comprehensive DIA phosphoproteome profiling.

Keywords: DIA; Phosphoproteome; Phosphorylation sites; Prediction; Deep learning

Introduction

Protein phosphorylation, a crucial post-translational modification, is integral to regulating almost all cellular signaling pathways. Phosphoproteomics, especially mass spectrometry (MS)-based techniques, have become pivotal in comprehensive studies of protein phosphorylation and the dynamics of cell signaling[1,2]. Traditionally, these studies have relied on data-dependent acquisition (DDA), which, despite its utility, often encounters limitations such as constrained throughput and inconsistent reproducibility due to the limitations of MS sequencing speed and the semi-random nature of DDA sampling[3].

The evolution of data-independent acquisition (DIA) methodologies has significantly transformed proteomic profiling[4]. DIA allows for the analysis of large sample sets with enhanced quantification accuracy and reproducibility. This method has shown great promise in various fields, including cellular signaling research, proteogenomic analysis of clinical cancer samples, and the discovery of antiviral drugs. A benchmark study by Olsen J and colleagues established that DIA phosphoproteomics achieves better dynamic range, identification reliability, and enhanced sensitivity and quantification accuracy compared to DDA-based methods[5–7].

Despite these advancements, DIA phosphoproteomics currently faces a critical challenge – the need for a high-quality spectral library built before data. Most DIA phosphoproteomic analyses require project-specific DDA libraries, typically constructed from extensively prefractionated or repeatedly injected samples. While these libraries offer broader proteome coverage, they demand considerable time, samples, and efforts, especially with prefractionation[8,9].

These challenges underscore the need for innovative approaches to streamline the DIA workflow and maximize its efficiency and effectiveness in phosphoproteomic studies. The development and implementation of such methods could revolutionize our understanding of protein phosphorylation and its role in cellular signaling, opening new avenues for research and therapeutic discovery[10,11].

An earlier study demonstrated the feasibility of constructing a DIA library directly from DIA data for extensive phosphoproteome profiling[12,13]. This method, however, relied on data from a large number of DIA runs, making it less practical. In contrast, in silico libraries, which predict fragment ion intensities and retention times using advanced machine learning techniques, offer an efficient alternative. These libraries, especially those generated by recent deep neural network technologies, can potentially achieve proteome coverage comparable to or better than traditional DDA libraries[10,14]. Despite their promise, in silico libraries remain underexplored in DIA phosphoproteomic analysis. Existing deep learning methods, typically employing LSTM or RNN architectures, are limited by their linear amino acid embedding approach[15]. To address these limitations, we developed DPHO, a novel deep learning framework specifically designed for phosphopeptides. DPHO-generated in silico libraries have shown superior performance in phosphoproteome profiling, outperforming conventional DDA libraries in terms of speed and depth[14–16].

Related Work

When conducting DIA phosphoproteomic studies, researchers quickly realize the critical importance of having a comprehensive spectral library. Traditionally, these libraries are painstakingly constructed through rigorous DDA experiments. While DDA libraries tailored to specific projects provide extensive coverage of the proteome landscape, their creation requires significant investments of time, effort, and resources. However, a promising development amidst these challenges is the emergence of computationally generated libraries[18]. Leveraging sophisticated machine learning techniques, these libraries have demonstrated remarkable potential in the broader field of proteomics[19].

Nevertheless, upon closer examination, it becomes apparent that the full potential of computationally generated libraries in DIA phosphoproteomic data analysis is still largely untapped. This represents a unique opportunity for our proposed solution, called Model. Model represents an innovative approach that seamlessly incorporates advanced deep learning methodologies to address the intricate and practical requirements of phosphoproteomics[20]. By leveraging deep learning techniques, we aim to overcome the limitations of current spectral libraries and unlock unprecedented precision, efficiency, and transformative insights in future research endeavors[21].

Through the integration of deep learning algorithms into phosphoproteomics, our visionary approach seeks to pave the way for exciting advancements in the field. By harnessing the power of these state-of-the-art methodologies, we aim to revolutionize phosphoproteomic data analysis, ultimately propelling the understanding of cellular signaling dynamics to new heights[23].

Result

Principle of DPHO

DPHO distinguishes itself in the realm of computational biology by pioneering a sophisticated approach to phosphopeptide prediction[25]. This advanced model leverages a progressively enriched peptide representation, adeptly capturing both the intricate local and expansive global structures of peptides for nuanced prediction. This is achieved through a cutting-edge hybrid network design, a significant departure from traditional methodologies.

The DPHO model utilizes a cutting-edge deep learning approach for predicting indexed retention time (iRT) and the intensity of fragment ions for a specific phosphopeptide. This process begins by taking the peptide sequence and its precursor charge as the inputs[26-27]. Initially, a bidirectional Long Short-Term Memory (bi-LSTM) network is employed to create preliminary representations for each amino acid in the sequence. These representations are further refined through a Transformer module. Subsequently, the refined, comprehensive features are processed through a linear regression network, which is responsible for producing the final predictions concerning both the fragment ion intensities and the iRT[28].

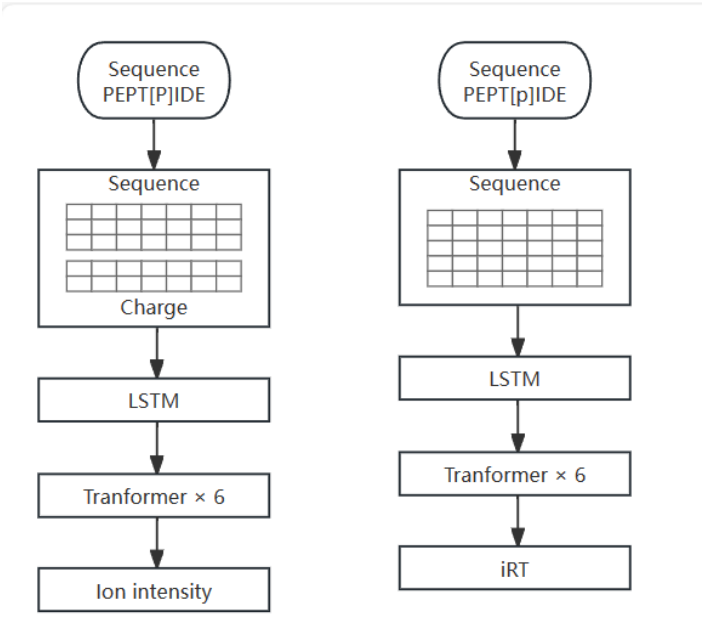

Figure 1: Model architecture of DPHO

At its core, DPHO is structured as a modular deep network, comprising three specialized sub-networks. The first is a recurrent network, specifically a bi-LSTM, tasked with initial encoding of the peptide sequences. This bi-LSTM lays the groundwork by embedding each amino acid into a detailed vector representation. Subsequently, these representations are refined through two layers of bidirectional LSTM units, allowing each amino acid to incorporate contextual information from its peers within the same peptide. Yet, it's notable that the bi-LSTM's context encoding may have limitations due to potential information loss over recurrent cycles.

To address this and to harness long-range dependencies within peptide sequences, DPHO introduces its second module: a Transformer network. This network takes the baton from the bi-LSTM, enhancing the peptide representations with a multi-head self-attention mechanism. This innovation enables simultaneous feature updates across all amino acids, facilitating the model's focus on multiple peptide regions, regardless of their spatial separation. The final stage of DPHO involves a linear regressor network. It receives the newly formulated peptide representation and is responsible for generating predictions for either fragment ion intensities or indexed retention times (iRT).

DPHO's specificity for phosphopeptide prediction is further honed by incorporating additional tokens to represent various phosphorylated amino acids. These tokens are learned in tandem with the base peptide embeddings. For fragment ion intensity predictions, the model employs a modified loss training approach, which adheres to the structural nuances of the peptides and selectively ignores nonexistent phosphate moieties[29].

DPHO's utilization of the Transformer network, a first in peptide fragmentation pattern prediction, marks a signif-

icant advancement, considering its extensive use in natural language processing. An ablative study, comparing DPHO with either bi-LSTM or Transformer models alone, as well as a CNN-Transformer combination, further underscores its efficacy. Using two phosphoproteomic datasets, DPHO consistently surpassed these alternatives, demonstrating its superior capability in capturing phosphopeptide features[30]. This suggests that the integration of bi-LSTM and Transformer models within DPHO is not only innovative but also synergistically effective in peptide representation learning.

Accurate prediction of fragment ion intensity and retention time for phosphopeptides.

To evaluate the effectiveness of the tool we developed, two datasets (RPE1 DDA and RPE1 DIA) both collected from RPE1 cells, one by DDA, the other by DIA acquisition methods, were searched by MaxQuant and Spectronaut respectively to yield phosphopeptide identification results. The data in each library was divided in a ratio of 8:1:1 for the purposes of training, validating, and testing the DPHO model, respectively. The performance of the trained DPHO model was impressive, showing high correlation between experimental and predicted fragment ion intensities for the test set. Specifically, the model achieved a median Pearson correlation coefficient (PCC) of 0.968 and a median spectral angle (SA) of 0.881 for the RPE1 DDA dataset, and a median PCC of 0.903 and a median SA of 0.791 for the RPE1 DIA dataset (Figure 2).

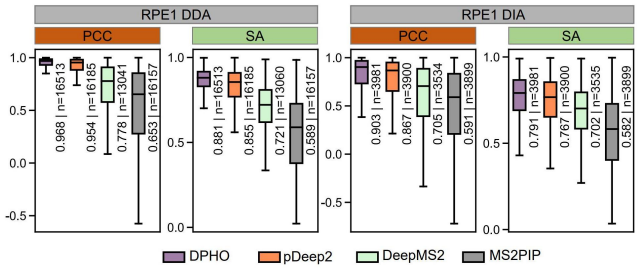

Figure 2: Compared with other models

Evaluation of DPHO and three other models based on the distribution of Pearson correlation coefficient (PCC) and spectral contrast angle (SA) calculated between predicted and experimental MSMS spectra from two datasets. Median PCC and SA are indicated; n is the number of phosphopeptides in the test set. Boxplot center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.

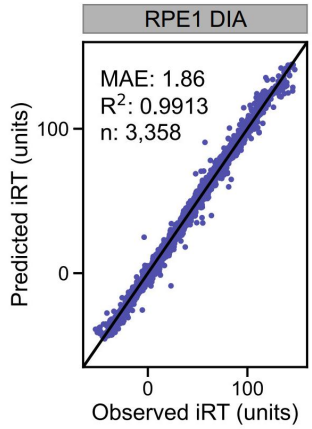

Figure 3: Performance of model

In addition, DPHO showcased its capability in accurately predicting indexed retention time (iRT) for the RPE1 DIA dataset, achieving a median absolute error (MAE) of 1.86 units. This performance highlights the tool's effectiveness in analyzing data from phosphoproteome profiling (Figure 2). Additionally, DPHO demonstrated proficiency in handling another dataset, U2OS DIA, by making precise predictions of fragment ion intensity and iRT. The model was particularly adept at predicting mono-phosphosite peptides and phosphopeptides containing phosphorylated serine (pS), likely due to the greater volume of data available for these peptide categories during training.

The assessment of DPHO's performance is conducted by examining the relationship between the predicted and experimentally determined indexed retention times (iRT). This evaluation includes the calculation of the correlation coefficient (R^2) from linear regression and the median absolute error (MAE).

Method

Processing of external DDA/DIA MS data

In the evaluation of DPHO's architecture, various datasets were utilized. Mouse brain DDA data, sourced from the PRIDE repository under the ID PXD006637, was applied in its original MaxQuant format for initial model assessment. Yeast R2P2 DDA data, also from PRIDE (ID PXD013453), underwent MaxQuant analysis using the Uniprot reference proteome for S. cerevisiae. The specific version of MaxQuant employed was v1.6.14.0, with parameters set to identify particular peptide modifications and an FDR threshold of 0.01 at both the PSM and protein levels[31].

The same yeast dataset was instrumental in refining the iRT prediction aspect of the model. Additionally, mouse brain DDA data was revisited for pre-training, complemented by Vero E6 DIA, yeast DIA (both retrieved from PRIDE, IDs PXD019113 and PXD013453, respectively), and human phosphopeptide RT datasets. The latter excluded phosphopeptides with low Ascore values from a previously

published study[32].

For DIA library construction, the Pulsar search engine within Spectronaut was utilized, referencing Uniprot proteomes for C. sabaeus and S. cerevisiae, aligning with dataset-specific proteomes. The methodology for creating these libraries is detailed under the section titled "Spectral Library Generation." Further validation of the model's prediction capabilities for phosphopeptides involved analyzing RPE1 DDA and DIA data (PRIDE ID PXD014525) and U2OS DIA data (PRIDE ID PXD017476). Modifications unsupported by DPHO were excluded, and searches were conducted to generate direct DIA libraries using the Uniport human reference proteome[33].

Reference spectra for phosphopeptides were gleaned from two DDA-based human studies (PRIDE IDs PXD017476 and PXD009227), and the model's predictive accuracy was assessed using these alongside RPE1 and U2OS datasets, in addition to a human/yeast two-proteome model (PRIDE ID PXD014525). The construction of direct DIA libraries followed the established protocol, including additional generation from the human/yeast dataset.

For inclusion in the model training and evaluation, phosphopeptides from these external datasets were required to have a localization score above 0.75, ensuring high-confidence site assignment[34].

model structure

DPHO stands out as a pioneering deep learning framework, uniquely designed for the intricate task of phosphopeptide prediction. Its core strength lies in its ability to progressively learn a rich and detailed representation of peptides, capturing both local and global structural nuances essential for precise predictions.

At the heart of DPHO is an innovative hybrid network architecture, distinct from traditional methods. This architecture synergistically integrates two distinct types of network structures to comprehensively encode various facets of peptide structure. The framework is composed of three principal sub-networks:

Recurrent Network (Bi-LSTM): This network forms the foundation of peptide encoding. Upon receiving the input peptide sequence, optionally alongside its charge state, the Bi-LSTM network initiates the process by generating a preliminary representation of each amino acid in the sequence. Through its dual-layer bidirectional LSTM units, each amino acid is embedded into a vector representation, enabling a context-sensitive portrayal enriched by the characteristics of neighboring amino acids. However, it's noteworthy that the context captured by the Bi-LSTM network can sometimes be constrained due to information loss in recurrent updates:

$$\begin{cases} \text{LSTM}(x_t, h_{t-1}) & \text{(forward pass)} \\ \text{LSTM}(x_t, h_{t+1}) & \text{(backward pass)} \end{cases} \tag{1}$$

Here, h_t represents the hidden state at time step t, x_t is the input at time step t, and LSTM denotes the LSTM function. The forward pass and backward pass capture the contextual information from both directions along the sequence[35].

Transformer Network: To overcome the limitations of the Bi-LSTM and capture long-range dependencies within peptide sequences, DPHO incorporates a Transformer network. This network refines the peptide representation formulated by the Bi-LSTM. Utilizing multi-head self-attention mechanisms, the Transformer network simultaneously updates all amino acid features, allowing the model to focus on multiple disparate peptide sites. This refined representation is then poised for further processing:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V_{rc} \tag{2}$$

Where Q, K and V are queries,keys and values matrices,respective,and d_k is the dimension of the vectors. The softmax function is applied to the rows of the matrix $QK^T/\sqrt{d_k}$,and this result is multiolied by V to get the final output.

Linear Regressor Network: As the final stage, the transformed peptide representation is fed into this network, tasked with predicting either fragment ion intensities or indexed retention times (iRT):

$$\text{y} = Wx + b \tag{3}$$

Metrics: In evaluating the accuracy of fragment ion intensity predictions, our approach involves calculating the Pearson correlation coefficient (PCC) for each peptide,comparing our predictions to the actual observed values. We then use the median value of these PCCs as the primary metric for assessment. Additionally, aligning with the methods used in Prosit11, we employ the normalized spectral angle (SA) as a secondary metric. For this, we again report the median value of the SAs calculated. The normalized spectral angle is defined in a specific manner for this purpose.

$$\text{SA}(y, y') = 1 - 2 \cdot \left(\frac{\cos^{-1}(y' \cdot y)}{\pi}\right) \tag{4}$$

For the definition of the normalized spectral angle, it involves comparing two vectors, each normalized to have an L2 norm of 1. The model selection is primarily based on the median Pearson correlation coefficient (PCC) metric derived from this comparison. Regarding the prediction of indexed retention time (iRT), we utilize the $\Delta t_{95\%}$ metric as the principal measure. This metric is defined as the smallest time interval that encompasses the discrepancies between the observed and predicted retention times (RTs) for 95% of the analyzed peptides.

$$\Delta t_{95\%} = 2 \cdot |z - z'|_{95\%} \tag{5}$$

References

[1] Ludwig, C. et al. Data-independent acquisition-based SWATH-MS for quanti tative proteomics: a tutorial. Mol. Syst. Biol. 14, e8126 (2018).

[2] Yang, Y., Horvatovich, P. & Qiao, L. Fragment Mass Spectrum Prediction Facilitates Site Localization of Phosphorylation. J. Proteome Res. 20, 634–644 (2021).

[3] Searle, B. C. et al. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. Nat. Commun. 11, 1548 (2020).

[4] Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat. Methods 16, 519–525 (2019).

[5] Humphrey, S. J., Karayel, O., James, D. E. & Mann, M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. Nat. Protoc. 13, 1897–1916 (2018).

[6] Humphrey, S. J., Azimifar, S. B. & Mann, M. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. Nat. Biotechnol. 33, 990–995 (2015).

[7] Lou, R. et al. Hybrid Spectral Library Combining DIA-MS Data and a Targeted Virtual Library Substantially Deepens the Proteome Coverage. iScience 23, 100903 (2020).

[8] Yang, Y. et al. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat. Commun. 11, 146 (2020).

[9] Li, C. et al. Integrated Omics of Metastatic Colorectal Cancer. Cancer Cell 38, 734-747.e9 (2020).

[10] Zeng, W.-F. et al. MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning. Anal. Chem. 91, 9724–9731 (2019).

[11] Wang, S. et al. NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. Nucleic Acids Res. 48, e83 (2020).

[12] Zhou, X.-X. et al. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. Anal. Chem. 89, 12690–12697 (2017).

[13] Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat. Methods 16, 509–518 (2019).

[14] Leutert, M., Rodríguez-Mias, R. A., Fukuda, N. K. & Villén, J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. Mol. Syst. Biol. 15, e9021 (2019).

[15] Bekker-Jensen, D. B. et al. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. Nat. Commun. 11, 787 (2020).

[16] Gillet, L. C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics MCP 11, O111.016717 (2012).

[17] Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 45, D1100–D1106 (2017).

[18] Humphrey SJ, Karayel O, James DE, Mann M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. Nat Protoc. 2018 Sep;13(9):1897-1916. doi: 10.1038/s41596-018-0014-9. PMID: 30190555.

[19] Humphrey SJ, Azimifar SB, Mann M. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. Nat. Biotechnol. 2015;33:990–995. doi: 10.1038/nbt.3327.

[20] Bekker-Jensen DB, et al. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. Nat. Commun. 2020;11:787. doi: 10.1038/s41467-020-14609-1.

[21] Gillet, L. C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell Proteomics11, O111 016717 (2012).

[22] Ludwig C, et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial.Mol. Syst. Biol. 2018;14:e8126. doi: 10.15252/msb.20178126.

[23] Leutert M, Rodriguez-Mias RA, Fukuda NK, Villen J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. Mol. Syst. Biol. 2019;15:e9021. doi: 10.15252/msb.20199021.

[24] Li C, et al. Integrated omics of metastatic colorectal cancer. Cancer Cell. 2020;38:734–747. doi: 10.1016/j.ccell.2020.08.002.

[25] Bouhaddou M, et al. The global phosphorylation landscape of SARS-CoV-2 infection.Cell. 2020;182: 685–712.e619. doi: 10.1016/j.cell.2020.06.034.

[26] Wang S, et al. NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. Nucleic Acids Res. 2020;48:e83. doi: 10.1093/nar/gkaa498.

[27] Searle BC, et al. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. Nat. Commun. 2020;11:1548. doi: 10.1038/s41467-020-15346-1.

[28] Gessulat S, et al. Prosit: proteomewide prediction of peptide tandem mass spectra by deep learning. Nat. Methods. 2019;16:509–518. doi: 10.1038/s41592-019-0426-7.

[29] Tiwary S, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat. Methods. 2019;16:519–525. doi: 10.1038/s41592-019-0427-6.

[30] Zhou XX, et al. pDeep: predicting MS/MS spectra of peptides with deep learning. Anal. Chem. 2017;89:12690–12697. doi: 10.1021/acs.analchem.7b02566.

[31] Yang Y, et al. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat. Commun. 2020;11:146. doi: 10.1038/s41467-019-13866-z.

[32] Lou R, et al. Hybrid spectral library combining DIA-MS data and a targeted virtual library substantially deepens the proteome coverage. iScience. 2020;23:100903. doi: 10.1016/j.isci.2020.100903.

[33] Luong, M.-T., Pham, H. & Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. In Empirical Methods in Natural Language Processing, Lisbon, Portugal. pp. 1412–1421. 10.18653/V1/D15-1166 (2015).

[34] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics, New Orleans, Louisiana, USA pp. 4171–4186. 10.18653/V1/N19-1423 (2018).

[35] Brown, T. B. et al. Language models are few-shot learners. In Neural Information Processing Systems, Vancouver Convention Center, Vancouver, Canada. pp. 1877–1901 (2020).