Chapter 10
Machine-Learning and the Future of HPC for MS-Based Omics

Fahad Saeed and Muhammad Haseeb

To date, MS proteomics data is identified using database search algorithms based purely on numerical techniques or some denovo techniques that allow peptide identification without using databases. Currently, there is no single strategy from database search or denovo techniques that can claim as the most accurate strategy. Substantial work has been carried out toward developing computational techniques for identification of peptides using database search [1], as well as denovo algorithms [2]. However, peptide identification problems are well-known and prevalent [3] including but not limited to misidentifications or no identification for peptides, statistical accuracy (FDR) and inconsistencies between different search engines [4].

Most of the algorithms applied to MS data have been limited to traditional numerical algorithms and can be categorized as database search algorithms and denovo algorithms. Comparison across literature indicates decreased average accuracy of denovo algorithms (38.1–64.0%) [4] relative to database search algorithms (30–80%) [5]. However, within-study direct comparisons of database verses denovo machine learning (ML) approaches have revealed modest gains [6], indicating further formal evaluation is warranted. Moreover, ML methods use different validation metrics and the lack of standard metrics and/or data-benchmarks can lead to overly optimistic assessment for machine learning algorithms [7]. Overall, prior literature demonstrated limited accuracy and generalizability [4] identifying peptides using current limited ML methods.

Previous work suggests that numerical algorithms and the use of traditional ML algorithms may not be able to capture and integrate the multidimensional features of MS data [8]. However, deep learning methods [4, 8] may offer an improved approach for identifying peptides in noisy high-dimensional MS data and peptides that are very similar to each other [9]. Preliminary progress assessing deep learning methods in peptide deduction applied to MS data has yielded an average accuracy of 82–95% on selected data sets but with limited precision (amino acid level 72%)

F. Saeed and M. Haseeb, *High-Performance Algorithms for Mass Spectrometry-Based Omics*, Computational Biology 34,
https://doi.org/10.1007/978-3-031-01960-9_10

and recall (peptide—level 39.24%) [8]. However, large volumes of data with large number of possible parameters are needed for deep learning training, particularly for MS data, which has resulted in a technical hurdle in developing such strategies. We have previously shown that this can result in overfitted deep learning models [10] with limited increase in accuracy due to noise feature-integration. Such overfitting leads to limited generalizability [11], and contributes to the ongoing reproducibility crisis [12–14]. One Deep-learning algorithm when used on another's data set leads to like 30% accuracy which suggests that there is a generalizability problem [15, 16]. Further, existing ML algorithms are computationally expensive and subsequently have limited scalability in training and application. Our prior work has demonstrated that this is particularly true for MS applications that scale poorly with increasing size of data sets [17]. Computational scaling and management are needed for these machine learning algorithms as this is currently a significant challenge for proteomics practitioners interested in applying these techniques.

In the future, we foresee the integrated use of image-processing, machine learning, including deep learning for MS data, to identify peptides from MS data in a highly accurate manner. To this end, there is some literature that has focused on processing MS data using machine learning and deep learning techniques. Importantly, image-processing, deep learning strategies, and fusing of multi-modal features have not been applied to MS data even though these techniques have the potential to radically change how MS data is processed with highly accurate peptide identifications. However, to make the proposed deep learning training and solutions scalable, HPC algorithms are needed.

10.1 Why HPC is Essential for Machine-Learning Models

Deep-learning models have unmatched expressive power as compared to traditional machine learning solutions. However, this expressive power comes from very large number of trainable parameters which can capture complex relationships between the data. In general, bigger and deeper Convolutional Neural Networks (CNN) models are used for various applications that are successful. The objective of this aim is to design and develop high-performance computing strategies which can process big MS data and accelerate the proposed deep learning models. CPU-GPU-based methods can give superior speeds in the processing of MS data, and of training, and inferring of deep learning models. Successful completion of such CPU-GPU-based pipelines is likely to contribute fundamental HPC techniques to our base of knowledge, without which the complex training and inferring of deep learning networks cannot be accomplished in reasonable timeframes. Upon completion of this research challenge, it is our expectation that we will have developed a HPC framework for the proposed DL solutions for MS-based omics computational strategies developed by us/others. Such tools would be important because they would likely aid in much-needed approaches to study human gut/environments microbiomes in a scalable fashion.

As part of our preliminary studies (Haseeb et al. Nature 2021), we explored our recently proposed HPC framework (called HiCOPS) [18] for efficient acceleration of database peptide search algorithms on large-scale symmetric multiprocessor distributed-memory supercomputers. HiCOPS exhibits orders-of-magnitude improvement in speed compared with several existing shared- and distributed-memory database peptide search tools, allowing several gigabytes of experimental MS/MS data to be searched against terabytes of theoretical databases in a few minutes compared with the several hours required by existing algorithms. The proposed HiCOPS parallel design implements an unconventional approach in which the (massive) theoretical databases are distributed across parallel nodes in a load-balanced fashion followed by asynchronous parallel execution of the database peptide search. On completion, the locally computed results are merged into global results in a communication-optimal manner. We have demonstrated an extensive performance evaluation in which we report between 70 and 80% strong-scale efficiency and less than 25% overall performance overheads (load imbalance, I/O, interprocess communication, pipeline halt); collectively depicting a near-optimal parallel performance. This overhead cost-optimal design, along with several optimizations, allows HiCOPS to maximize resource utilization and alleviate performance bottlenecks. Since our HPC framework is search-algorithm oblivious it will be a natural extension to incorporate meta-proteomics deep learning model into the HPC framework.

10.2 Preliminary Data and Findings

In our recent paper [19], we have designed and implemented a Deep Cross-Modal Similarity Network called SpeCollate. This is a deep learning network that tries to learn the scoring function between the spectra and peptides by mapping the different modalities of the data into a shared Euclidean subspace. This is achieved by learning fixed sized embeddings, and training the network using sextuplets of positive and negative examples. SpeCollate also uses a custom-designed SNAP-loss function and hardest negative mining for appropriate negative examples to improve the training performance. In order to train the network 4.8 million sextuplets obtained from the NIST and MassIVE peptide libraries were used and which allowed our deep learning model to perform better than Crux and MSFragger in terms of the number of peptide-spectrum matches (PSMs) and unique peptides identified under 1% FDR for real-world data. To the best of our knowledge, our deep learning network is the first model that can determine the cross-modal similarity between peptides and mass spectra for MS-based omics.

Despite all the superior accuracy of the deep learning model one thing that was a severe bottleneck was the time it takes to train the network. Since there is no theoretical framework that would allow us to predict the performance of the model, one has to *fully* train the model and run validation and testing before any judgement can be made. Our single preliminary design shows that it takes approx. 283 days of compute time to train a single deep learning network with 5 hyper-parameters optimizations. The intractability of a single model demonstrates that continued search for better DL models is huge technical hurdle in studying MS-based meta-omics. Each possible deep-network, even though carefully selected, has to be completely trained to assess its feasibility. Apart from the time it takes for the deep learning model training and inferences; the workflows that process MS data are also shown to be inefficient, both theoretically [20] and in real-world experiments [18, 21]. We believe, the HPC frameworks, that can accelerate the MS data analysis as well as the training and inference of deep learning will be essential to tract any kind of MS-based omics data analysis.

We believe that HPC will be at the forefront of these scientific investigations in the future.

References

1. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL (2015) Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. PROTEOMICS-Clin Appl 9(7–8):745–754
2. Vyatkina K, Wu S, Dekker LJ, VanDuijn MM, Liu X, Tolic N, Dvorkin M, Alexandrova S, Luider TM, Pasa-Tolic L et al (2015) De novo sequencing of peptides from top-down tandem mass spectra. J Proteome Res 14(11):4450–4462
3. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, del Toro N, Rurik M, Walzer M, Kohlbacher O, Hermjakob H et al (2016) Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat Methods 13(8):651
4. Tran NH, Zhang X, Xin L, Shan B, Li M (2017) De novo peptide sequencing by deep learning. Proc Natl Acad Sci 114(31):8247–8252
5. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 33(7):743
6. Chi H, Liu C, Yang, H Zeng W-F, Wu L, Zhou W-J, Niu X-N, Y-H Ding, Zhang Y, Wang R-M et al (2018) Open-pfind enables precise, comprehensive and rapid peptide identification in shotgun proteomics, bioRxiv 285395
7. Obermeyer Z, Emanuel EJ (2016) Predicting the future-big data, machine learning, and clinical medicine. N Engl J Med 375(13):1216
8. Qiao R, Tran NH, Li M, Xin L, Shan B, Ghodsi A Deepnovov2: better de novo peptide sequencing with deep learning. arXiv:1904.08514
9. Zhou X-X, Zeng W-F, Chi H, Luo C, Liu C, Zhan J, He S-M, Zhang Z (2017) pdeep: predicting ms/ms spectra of peptides with deep learning. Anal Chem 89(23):12690–12697
10. Eslami T, Saeed F (2018) Similarity based classification of adhd using singular value decomposition. In: Proceedings of ACM international conference on computing frontiers (CF'18), pp 19–25
11. Zou L, Zheng J, Miao C, Mckeown MJ, Wang ZJ (2017) 3d CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. IEEE Access 5:23626–23636
12. Allison DB, Shiffrin RM, Stodden V (2018) Reproducibility of research: issues and proposed remedies. Proc Natl Acad Sci 115(11):2561–2562
13. Hutson M (2018) Artificial intelligence faces reproducibility crisis. Science (New York, NY) 359(6377):725
14. Berrar D, Dubitzky W (2017) On the Jeffreys-Lindley paradox and the looming reproducibility crisis in machine learning. In: 2017 IEEE international conference on data science and advanced analytics (DSAA). IEEE pp 334–340

15. Gabriels R, Martens L, Degroeve S (2019) Updated ms^{2}pip web server delivers fast and accurate ms^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. Nucleic Acids Res 47(W1):W295–W299
16. Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J, Knaute T, Rechenberger J, Delanghe B, Huhmer A et al (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 16(6):509
17. Haseeb M, Afzali F, Saeed F (2019) Lbe: A computational load balancing algorithm for speeding up parallel peptide search in mass-spectrometry based proteomics. In: IEEE international parallel and distributed processing symposium workshops (IPDPSW). IEEE, pp 191–198
18. Haseeb M, Saeed F (2021). Source data: high performance computing framework for terascale database search of mass spectrometry data. https://doi.org/10.5281/zenodo.5076575
19. Tariq MU, Saeed F (2021) Specollate: deep cross-modal similarity network for mass spectrometry data based peptide deductions. PLoS ONE 16(10):e0259349
20. Saeed F, Haseeb M, Iyengar S Communication lower-bounds for distributed-memory computations for mass spectrometry based omics data. arXiv:2009.14123
21. Kumar S, Saeed F (2021) Communication-avoiding micro-architecture to compute xcorr scores for peptide identification. In: 2021 31st international conference on field-programmable logic and applications (FPL). IEEE, pp 99–103