LC-MS MS1 image map classification enables real-time sample quality control for nanoparticle-based deep untargeted proteomics

Biao Li*, Ryan Benz, Harendra Guturu, Iman Mohtashemi, Theo Platt, and Serafim Batzoglou

Quality controls are important to generate high-quality deep untargeted proteomics data

The Seer Proteograph™ workflow uses nanoparticles (NP) combined with LC-MS (Liquid Chromatography-Mass Spectrometry) to enable deep untargeted proteomics[1]. As with any proteomics workflow, good quality controls (QCs) are important to facilitate the generation of high-quality data, particularly in large cohort studies. Here, we present a new QC procedure for the Proteograph workflow that utilizes machine learning on MS1 image maps of the raw LC-MS data to help identify unexpected patterns and highlight potential issues for further investigation. These models provide real-time monitoring of data quality, facilitate troubleshooting analysis for root cause investigations, and ensure that only high-quality LC-MS data are used for downstream differential proteomics analysis.

Proteograph™ Product Suite

Proteograph workflow provides unbiased, deep, and rapid proteomics at scale

From sample to peptides, ready for analysis on most LC-MS instruments with a variety of proteomics methods including label-free workflows

Methods

- **Input data** was curated from thousands of LC-MS runs from internal database resulting in a set of high-quality, representative examples for model training, consisting of 1,313 unique MS runs across 5 nanoparticles, 3 biological sample types, 5 MS instruments and 4 SP100 Automation Instruments.
- **MS1 image maps** were created using MS1 scans extracted from raw *mzML* files, binned into high-resolution images projecting spectrum intensities into color maps along the m/z and retention time axes.
- **Modeling architectures** on a range of neural networks were used including off-the-shelf image recognition model EfficientNet[2] by Google Research with transfer learning, and customized sequential CNN (Convolutional Neural Networks) built from scratch.
- **Model training** was using >= 30 epochs, minimizing categorical cross entropy loss with an Adam optimizer. The final models can generate predictive probabilities, be employed to classify expected nanoparticle from MS1 image map of a given run, and be applied to highlight samples of possible QC issues including contamination (e.g. detergents) causing skewed TIC (Total Ion Current) chromatogram or operational errors resulting from sample swap or mis-annotations.
- **Performance evaluation** was based on overall classification accuracy, weighted average F1 score and error analysis of misclassified samples out of 20% hold-out test set while entire input data dataset was split into 60% for model training, 20% validation for parameter tuning and 20% for hold-out test.

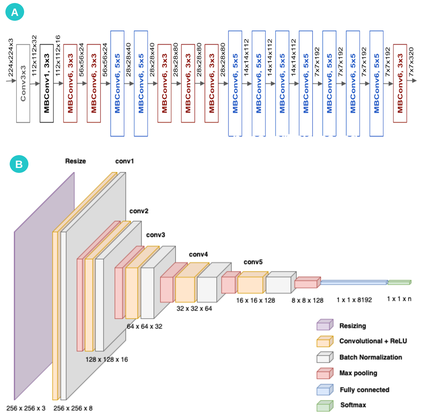

Figure 1. Modeling architectures. **(A)** EfficientNet architecture is an off-the-shelf pre-trained model capable of image classification tasks and good for transfer learning.

(B) Customized CNN is a simplified modeling architecture built from scratch consisting of 5 convolutional and pooling layers sequentially and ending with a fully connected layer and softmax.

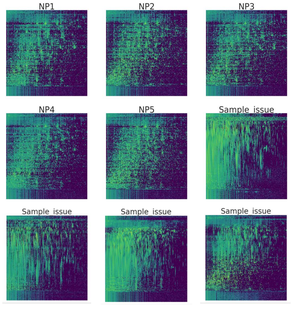

Figure 2. MS1 image maps of good vs poor MS runs. Training samples span 6 different classes, including MS1 image maps of nominally good MS runs from 5 different nanoparticle types (NP1 to NP5), and 6th class of maps of runs with sample contamination issues identified by manual retrospective investigation. The MS runs with sample contamination are evident from their MS1 image maps, including more subtle cases which may be difficult to identify by other means (e.g. TIC traces).

Applying image classification directly on raw MS1 data demonstrates great accuracy for Proteograph Assay quality control

A

	precision	recall	f1-score	support
NP1	1.00	0.92	0.96	50
NP2	1.00	1.00	1.00	50
NP3	0.98	0.98	0.98	50
NP4	0.91	1.00	0.95	49
NP5	1.00	0.98	0.99	51
Sample_issue	0.94	0.94	0.94	16
accuracy			0.97	266
macro avg	0.97	0.97	0.97	266
weighted avg	0.98	0.97	0.97	266

B

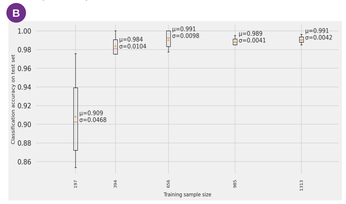

Figure 3. Sample size learning curve and model comparison. Performance in table **(A)** utilizing EfficientNet with transfer learning converged at 0.97 accuracy. The customized sequential CNN can attain >0.99 both in accuracy and weighted average F1 score. Experimentation on sample size learning curve **(B)** shows that >1K samples can achieve >0.99 accuracy of using customized CNN model.

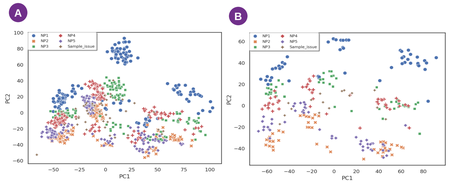

Figure 4. Visualization on embeddings of fully connected layer using PCA. Feature characteristics of both **(A)** training set and **(B)** hold-out test set are largely non-linear with top two principal components explaining only ~35% of total variance.

Results

Figure 5. Performance evaluation: customized CNN model.

Among 3 model training conditions, **(A)** 30 epochs of 48 batch size, **(B)** 40 epochs of 32 batch size and **(C)** 50 epochs of 24 batch size, best attained overall accuracy is 0.993 using 24 batch size. Left and middle panels are curves of training & validation accuracy and loss. Right panel shows confusion matrix of each classifier on the test set. At ~20 epochs performance starts to converge, but at >40 epochs, the models begin to overfit as training & validation loss diverge.

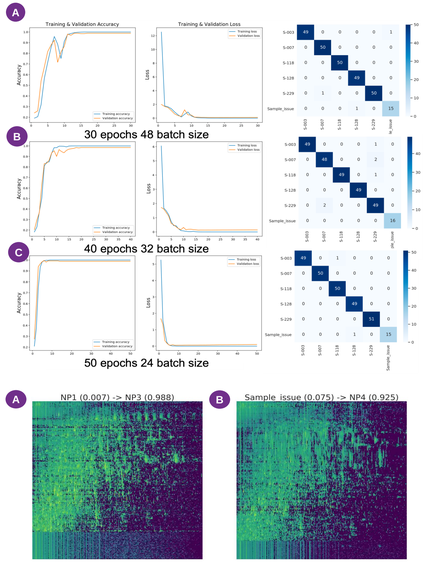

Figure 6. Error analysis on misclassified MS runs in hold-out test set from best performing model. There are only 2 cases of misclassifications out of 266 samples in hold-out test set. **(A)** NP1 was classified as NP3 with predictive probability of 0.988. **(B)** A sample deemed as having contamination by an expert MS reviewer was classified as a good performing NP4 with probability of 0.925. Data quality of functional metrics, e.g. peptide quantification, protein group counts, were further examined and OK for both. **(A)** is likely a true misclassification for which model can further improve on to avoid false positives alike, while **(B)** is debatable and perhaps overly stringent labeling as sample issue from its TIC. Discounting **(B)** can elevate model accuracy to > 0.995. As a result, employing MS1 classification can better and more quantitatively curate sample quality control over manually examining TIC chromatogram.

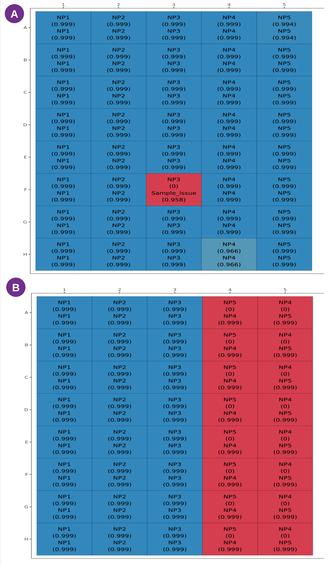

Figure 7. Applications of real-time sample QC and monitoring on assay plates. In **(A)**, the classifier highlights well(s) on the assay plate with lower-than-expected predictive probabilities for further QC follow-ups. In **(B)**, the classifier identifies an operational error (swaps or mis-annotations between NP4 and NP5). Each tile in figures corresponds to its sample's physical well position on the assay plate, with text annotations showing the original class labeling and its predictive probability, and class labeling with the highest predictive probability among all classes.

Conclusion

- Proteograph workflow provides unbiased, deep, and rapid proteomics at scale. Using good QC procedures is important to facilitate generation of high-quality data with Seer's technology particularly in large cohort studies.
- Machine learning models provides real-time sample QC for nanoparticle-based deep untargeted proteomics with great accuracy (> 0.99) detecting sample contamination and/or operational errors by applying image classification directly on raw MS1 data.
- Applications of this work facilitates fingerprinting of other signatures in MS1 data (e.g. disease etiology[3]) and potentially enabling directed analysis and comparison of unannotated spectra.

References

[1] Blume et al. *Nat. Comm.* (2020)
[2] Tan et al. *ICML.* (2019)
[3] Cadow, et al. *Bioinformatics* (2021)

Publications

Copyright Seer, Inc 2023 Seer, Inc., Redwood City, CA 94065, USA | *bli@seer.bio