Machine Learning-Based Glycan Structure Analysis and Synthesis

Jihoon Chung, Maede Maftouni, Zhenyu (James) Kong
Grado Department of Industrial and Systems Engineering
Virginia Tech, Blacksburg, VA

Introduction

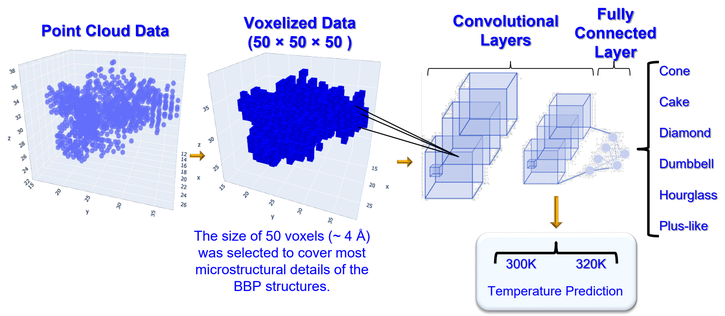
- Shape of the soft material such as Bottlebrush Polymers (BBPs) influences its unique properties and functions such as compactness and bending.
- High signal-to-noise ratio (SNR) is a desired property of the mass spectrum (MS) for accurate structure analysis of glycan data.
- Collecting MS of glycan takes a lot of cost and time, causing insufficient data.

Research Question

- Can multi-task learning improve representativeness and information richness of the learned features?
- How can we optimize the process parameters to achieve the target SNR with a few numbers of MS?
- How can we augment MS to overcome the insufficient data issue?

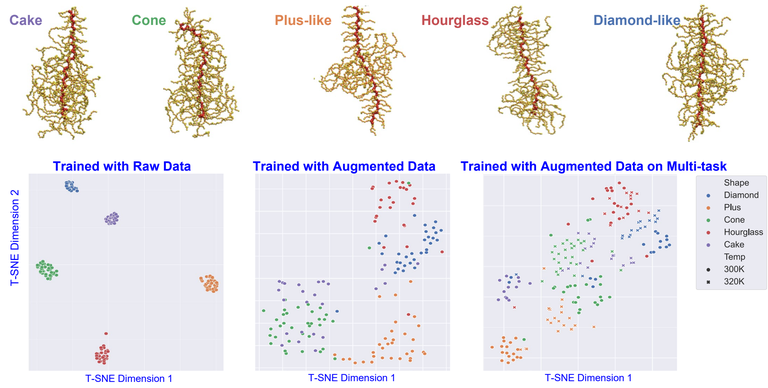
BBP Multi-Task Learning

- 3D object recognition network takes the voxelized data, learn its features and classifies it into one of 5 shape as well as 2 temperature categories.



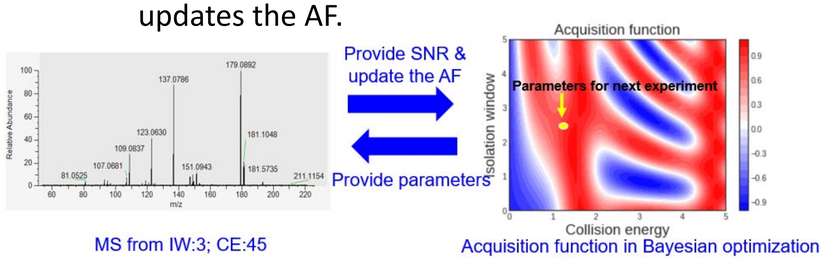
BBP Shape Recognition Interpretability

- More similar classes should cluster closer together in the T-SNE plot of the learned feature vectors.



Parameter Optimization in the MS

- The objective is to find optimal process parameters, including isolation window (IW) and collision energy (CE), to achieve the target SNR.
- Bayesian Optimization (BO) is a sequential optimization method and finds the optimal parameters based on the following two steps iteratively.
    1. BO provides parameters from the acquisition function (AF).
    2. SNR of MS acquired from the new parameters updates the AF.

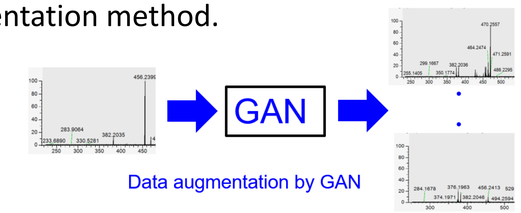


MS from IW:3; CE:45

Acquisition function in Bayesian optimization

Data Augmentation of MS

- The objective is to augment the MS to enable data-driven analysis, including regression and classification.
- Generative Adversarial Network (GAN) is the most widely used machine learning-based data augmentation method.



Data augmentation by GAN

Conclusion

- Multi-task learning improves the interpretability of learned features from BBP data in machine learning-based analysis.
- Bayesian optimization expects to find optimal parameters to achieve the target SNR with few numbers of MS.
- Effective data-augmentation method is required for the data-driven analysis of glycan data.

Acknowledgement

This work was supported by GlycoMIP, a National Science Foundation Materials Innovation Platform funded through Cooperative Agreement DMR-1933525.

References

- Joshi, Y., et al. Computational Materials 8, (2022).
- Goodfellow, I., et al. Advances in NIPS 27, (2014).