令和 6 年度　　京都大学化学研究所　スーパーコンピュータシステム　利用報告書

うつ病とその症状の生物学的背景を解明するための機械学習アルゴリズムの開発

The development of machine learning algorithms to decipher the biological background of major depression and its symptoms

Bioinformatics Center, Institute for Chemical Research, Kyoto University,
Peter Petschner

研究成果概要

During FY2024/2025 we re-started training, fine tuning and assessment of our machine learning model on the UK Biobank derived dataset to identify genes behind major depressive disorder after discovering a serious bug in the 3rd party genetic software used for the quality control of the data. The full-scale dataset contained ca. 300,000 individuals, 2 environmental (phenotypic variables) factors, depression score as output and ~6M genetic factors (single-nucleotide polymorphisms [SNPs]). The dimension reduction of this dataset followed the aggregation of SNPs onto genes, which resulted in a 16 first layer, corresponding to ~1bn parameters. To run an analysis, thus, required ~1.3 TB of memory and more than 2 months computation time.

In the first half of the fiscal year we tested multiple runs. Unfortunately, these runs did not result in a meaningful increase in prediction performance compared to the baseline, in stark contrast with the wrong data used before the discovery of the bug in the QC software. Therefore, we began exploring the potential causes of the failure of the algorithm both in theory and in practice, and confirmatory runs were also started. Nonetheless, we could not obtain better performance on test set than baseline (Figure 1).

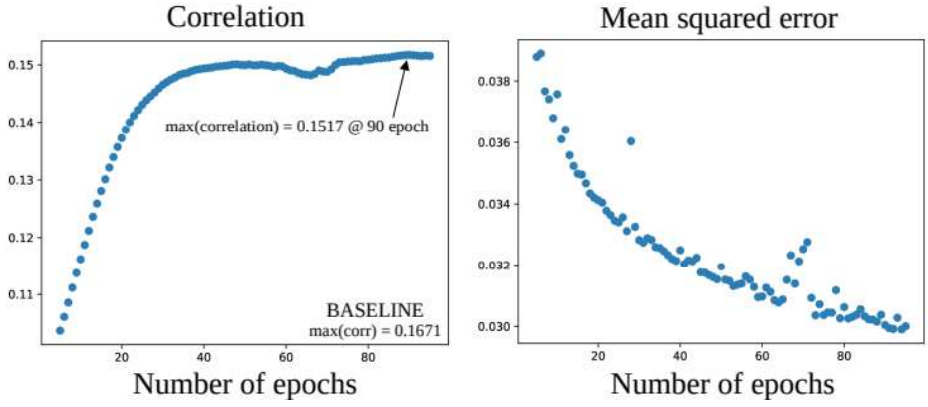

Figure 1. Results on a test set in a 10-fold cross validation setting.
Baseline value of correlation using only phenotype and no genetic features (BASELINE on figure) remained higher than the maximum of the correlation metric on the test set (indicated with an arrow).

We intend to continue our project and test another outcome, like migraine, a neurologic disorder, which may be more genetically determined (and, thus, predictable) and could validate the model architecture. Since UK Biobank currently imposed stricter control on the data use policy, and forces users to migrate to Amazon Web Services-based Research Analysis Platform, we intend to submit an exception, which would allow us to continue working on the data in the next fiscal year on the Supercomputer System.