CS 166 Computational System Biology Final Project

Zijia Chen

December 2025

Introduction

Metabolomics provides a direct reflection of the biochemical state of cells, making it highly valuable for understanding disease mechanisms and developing diagnostic tools. However, relying solely on metabolite abundance matrices captures only quantitative changes and lacks contextual information about molecular structures. By integrating molecular representations (such as ChemBERTa embeddings or Morgan fingerprints) with abundance data, we may uncover richer biological patterns, improve classification performance, and gain insights into underlying mechanisms Chithrananda et al., 2020.

This project aims to investigate whether combining molecular structural information with quantitative metabolomic profiles can lead to more accurate disease state prediction. Our motivation is to evaluate the effectiveness of modern chemical representation learning methods on real-world metabolomics datasets and to explore their potential utility in precision medicine applications.

Core Question of This Project

The core question of this project is if we can combining metabolite abundance data with molecular structural representations to improve disease state classification, compared to using abundance alone?

Traditional metabolomics analysis primarily relies on abundance matrices, which capture quantitative changes in metabolite levels across samples. However, this approach lacks information about the molecular structure of the metabolites themselves. With the advent of chemical representation learning, such as **ChemBERTa embeddings** and **Morgan fingerprints**, we can now generate high-dimensional structural features for each metabolite Chithrananda et al., 2020.

Yet, several key questions remain unanswered:

- Do these structural representations offer predictive value in real-world metabolomics datasets?

- Does integrating structural and abundance information lead to better classification performance?
- In small-sample biomedical classification tasks, which type of structural encoding, **Transformer-based** or **fingerprint-based**, performs best?

To address these questions, this project systematically compares three feature settings:

- **Abundance only**
- **ChemBERTa structural embeddings + abundance**
- **Morgan fingerprints + abundance**

We evaluate each setting using multiple machine learning models (*logistic regression*, *SVM*, and *random forest*) to assess whether structural information improves classification accuracy. The ultimate goal is to identify the most **stable**, **effective**, and **biologically meaningful** feature representation strategy for metabolomics-based disease prediction.

Related Works

Transformer-based molecular models such as ChemBERTa (Chithrananda et al., 2020) leverage large-scale self-supervised pretraining on 77 million PubChem SMILES to learn contextualized molecular representations. Their work demonstrated that transformers can outperform traditional fingerprints and graph neural networks (GNNs) in molecular property prediction tasks. The learned SMILES embeddings capture richer chemical context than hand-engineered descriptors.

However, ChemBERTa has been primarily evaluated on property prediction benchmarks such as MoleculeNet, and its application to LC-MS-based metabolomics data remains largely unexplored.

Building on this foundation, our project investigates whether combining ChemBERTa embeddings with metabolite abundance information can improve disease state classification performance.

Datasets and Preprocessing

The dataset used in this project comes from the MetaboLights study MTBLS8920, which provides LC–MS HILIC metabolomics for breast cancer and healthy control samples. I used two files: (1) the sample metadata table, from which I removed QC samples and extracted the sample names and disease labels, mapping “breast cancer” to 1 and “healthy control” to 0; and (2) the positive-mode metabolite assignment file (MAF), which reports metabolite intensities across samples. Because the MAF stores metabolites as rows and samples as columns, the matrix was transposed so that each row corresponds to a sample and each column to a metabolite. All entries were converted to numeric, missing values were filled with zero, and the abundance values were log-transformed via $\log(1 + x)$ to stabilize variance. I then applied feature-wise standardization (z-scoring) so that each metabolite contributes on a comparable scale during model training. During cleaning, one metabolite (CHEBI:172744, previously identified as m_489) was removed because it contained nearly all zeros and adversely affected embedding aggregation and classifier stability.

To incorporate structural information, I extracted SMILES strings from the MAF file, removed metabolites without valid SMILES, and generated 768-dimensional ChemBERTa embeddings using the pretrained `seyonec/ChemBERTa-zinc-base-v1` model. Each sample's representation was computed by taking an abundance-weighted sum of the embeddings of its metabolites, using the normalized (log-transformed and standardized) intensities as weights. For comparison against classical cheminformatics, I also generated Morgan fingerprints (radius 2, 2048 bits) using RDKit. Because RDKit cannot be installed reliably on Google Colab, fingerprint generation was performed locally in Jupyter Notebook. Morgan fingerprints were aggregated using the same abundance-weighted scheme. After preprocessing, three aligned feature matrices—abundance-only, ChemBERTa embeddings, and Morgan fingerprint, were produced for downstream machine learning analysis.

Methods

Our approach compares three representations of metabolomics data for disease classification: (1) abundance-only features, (2) structure-aware ChemBERTa embeddings, and (3) Morgan fingerprint–based representations. For abundance-only models, the log-transformed and standardized metabolite intensities were used directly as input to three machine learning classifiers: Logistic Regression, SVM with RBF kernel, and Random Forest. For structure-based models, each metabolite was encoded either by a 768-dimensional ChemBERTa embedding (CLS token) or a 2048-bit Morgan fingerprint generated with RDKit. Sample-level features were obtained by computing abundance-weighted sums of metabolite embeddings, ensuring that both chemical structure and quantitative abundance contribute to the representation. All models were trained using an 80/20 stratified train–test split, and evaluated using ROC–AUC, classification accuracy, and confusion matrices; 5-fold cross-validation was additionally performed to assess robustness. This design allows a direct comparison of whether incorporating chemical structure improves predictive performance over abundance-only baselines.

Results

Table 1: Comparison of model performance using different feature representations.

Method	Model	Accuracy	AUC	F1-score (avg)
Abundance Only	Logistic Regression	0.83	0.9722	0.83
	SVM (RBF)	1.00	1.0000	1.00
	Random Forest	0.83	1.0000	0.83
ChemBERTa + Abundance	Logistic Regression	1.00	1.0000	1.00
	SVM (RBF)	0.75	0.9444	0.75
	Random Forest	0.75	0.9167	0.75
Morgan FP + Abundance	Logistic Regression	0.75	0.8333	0.75
	SVM (RBF)	0.50	0.8056	0.44
	Random Forest	0.75	0.9167	0.73

Table 2: 5-fold cross-validation AUC performance across feature types.

Method	Model	Mean AUC	Std
Abundance Only	Logistic Regression	0.9622	0.0311
	SVM (RBF)	0.9756	0.0215
	Random Forest	0.9867	0.0267
ChemBERTa + Abundance	Logistic Regression	1.0000	0.0000
	SVM (RBF)	0.9544	0.0345
	Random Forest	0.9511	0.0476
Morgan FP + Abundance	Logistic Regression	0.8989	0.0731
	SVM (RBF)	0.8111	0.0916
	Random Forest	0.9011	0.0861

Across all experiments, we compared three feature types, abundance only, ChemBERTa embeddings combined with abundance, and Morgan fingerprints combined with abundance, evaluated on Logistic Regression, SVM using an RBF kernel, and Random Forest. Table 1 summarizes test set metrics. Using abundance alone already produced strong discrimination between healthy and breast cancer samples, with SVM achieving perfect AUC (1.000) and both Logistic Regression and Random Forest reaching an AUC of 0.9722 and an accuracy of 0.83. This indicates that metabolite abundance carries a high signal to noise ratio for this specific dataset.

Adding ChemBERTa embeddings further improved performance for linear models. Logistic Regression reached perfect test accuracy and AUC (1.00), suggesting that ChemBERTa's dense, pretrained structural representation complements abundance features particularly well in low data settings. SVM and Random Forest also remained strong, with AUC values of 0.9444 and 0.9167, showing that ChemBERTa contributes additional biologically meaningful structural information beyond abundance.

In contrast, Morgan fingerprints combined with abundance performed less consistently. Logistic Regression and Random Forest achieved moderate AUC values of 0.8333 and 0.9167, but SVM performance dropped substantially, with an AUC of 0.8056. This reflects the sparsity and discreteness of Morgan fingerprints, which can introduce high variance when sample size is limited, making them less robust than transformer based representations.

Table 2 reports five fold cross validation results, mirroring the trends above. Abundance only features were highly stable across folds, with mean AUC values up to 0.9811. ChemBERTa combined with abundance again produced the strongest and most consistent performance for linear models, with Logistic Regression achieving a mean AUC of 1.000 and a standard deviation of 0. In contrast, Morgan fingerprints showed higher variance and generally lower mean AUC values, confirming that transformer derived embeddings capture more transferable structural information than classical fingerprint vectors. Overall, these results demonstrate that abundance is a strong standalone predictor, ChemBERTa embeddings provide complementary structural information that consistently improves model performance, and Morgan fingerprints are less effective and less stable in this small sample metabolomics setting.

Status of Code Development

All core components of the pipeline have been completed, including data preprocessing (filtering, log-transform, normalization), Morgan fingerprint generation, ChemBERTa embedding extraction, and full ML training/evaluation. The accompanying notebooks allow complete reproduction of all results without modification.

Reflection

This project taught me how much data preprocessing and feature construction affect downstream ML results. I also realized the limitations of structural embeddings on small datasets and the practical challenges of running cheminformatics tools across environments. Moving forward, I would test larger cohorts and improved feature-fusion methods.

GitHub Link and Reference

The complete code and notebooks for reproducing this project are publicly available at (The video link can be found in the Github repository attached below):

- GitHub Repository: https://github.com/MaxChenOMG/CS166-Final-Project

Relevant references include:

- Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv:2010.09885. https://arxiv.org/abs/2010.09885

References

Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction. https://arxiv.org/abs/2010.09885