DeepHalo Documentation
Open-Source Pipeline for High-Confidence Detection of Halogenated Natural Products in HRMS Data
Version 0.9

Table of Contents

Introduction

DeepHalo is a high-throughput computational pipeline designed for the detection and dereplication of halogenated natural products (HNPs) in high-resolution mass spectrometry (HRMS) data. It integrates deep learning models, statistical validation, and dual dereplication strategies to achieve high accuracy and efficiency. Key applications include natural product discovery and halogenated metabolite annotation.

Core Features

1. Halogen Prediction

- **Element Prediction Model (EPM)**
  - Bimodal Deep Neural Network (DNN) architecture for Cl/Br detection.
  - Broad mass range: 50–2000 Da (resistant to interference from B, Se, Fe, and dehydro isomers).

2. Isotope Pattern Validation

- **Dual Validation System**
  - *Mass Dimension*: Statistical rule-based correction.
  - *Intensity Dimension*: Autoencoder Deep Model (ADM) for anomaly detection.

3. Multi-Level Scoring (H-score)

- Combines feature centroid analysis and scan-level validation to eliminate oversaturation and peak overlap errors.

4. Dereplication

- **Dual Strategy**
  - *Custom Database Matching*: Validates exact mass, halogen patterns, and isotope intensity similarity.
  - *MS2 Networking*: Integration with GNPS for spectral similarity analysis.

Technical Advantages

- **Throughput**: Processes samples in <30 seconds each on standard hardware (Core i9, 16GB RAM).
- **Accuracy**: >98.6% precision in halogen detection across experimental datasets.
- **Interoperability**:
  - Input: .mzML files.
  - Output: Cytoscape-compatible network files.
  - *MS2 Networking*: Integration with GNPS for spectral similarity analysis.

Technical Advantages

- **Throughput**: Processes samples in <30 seconds each on standard hardware (Core i9, 16GB RAM).
- **Accuracy**: >98.6% precision in halogen detection across experimental datasets.
- **Interoperability**:
  - Input: .mzML files.
  - Output: Cytoscape-compatible network files.

Installation

Prerequisites

- Python 3.10 (Verify with python --version).

Installation Methods

From PyPI

pip install DeepHalo

From Local Wheel

pip install path/to/DeepHalo-xxx.whl

From Source

git clone https://github.com/xieyying/DeepHalo.git

cd DeepHalo

pip install -e .

Quickstart

1. Detect Halogenated Compounds in mzML Files

halo analyze-mzml -i /path/to/mzml_files -o /output_directory -ms2

2. Dereplication with GNPS and Custom Database

halo dereplication -o /output_directory -g /path/GNPS_results -ud /path/custom_database.csv -udk Formula

---

Command-Line Usage

General Help

```
halo --help                    # List all commands
halo [command] --help          # Show options for a specific command (e.g., `halo analyze-mzml --help`)
```

Commands

1. Analyze mzML Files

```
halo analyze-mzml \
  -i <input_path> \            # Input .mzML file or directory
  -o <project_path> \          # Output directory
  [-c <config_file>] \         # Custom configuration (optional)
  [-b <blank_samples_dir>] \ # Blank samples for subtraction (optional)
  [-ms2]                       # Enable MS2 extraction
```

2. Dereplication

```
halo dereplication \
  -o <project_path> \          # Project directory
  -g <GNPS_folder> \           # GNPS results directory
  -ud <user_database.csv> \    # Custom database (CSV/JSON)
  -udk <formula_column>        # Column name for formula matching
```

3. Create Training Dataset

halo create-dataset <project_path> [-c <config_file>]

4. Train Model

halo create-model <project_path> \

[-c <config_file>] \ # Custom configuration (optional)

[-m <manual/search>] # Training mode (default: manual)

Dependencies

- pandas == 2.0.3
- numpy == 1.22.0
- tensorflow == 2.10.1
- scikit-learn == 1.3.1
- pyopenms == 3.1.0
- Full list: See README.md.

License

For methodology details and benchmarks, refer to the GitHub repository.