**DeepHalo Documentation**

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

**Table of Contents**

1. [**Introduction**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#introduction)
2. [**Core Features**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#core-features)
3. [**Technical Advantages**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#technical-advantages)
4. [**Installation**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#installation)
5. [**Quickstart**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#quickstart)
6. [**Command-Line Usage**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#command-line-usage)
7. [**Dependencies**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#dependencies)
8. [**License**](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#license)

**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)**
  + Dual-branch Isotope Neural Network (IsoNN) architecture for Cl/Br detection.
  + Wide 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-based Anomaly Detection Model (ADM).

**3. Multi-Level Scoring (H-score)**

* A hierarchical scoring mechanism that combines predictions by leveraging isotope patterns at both the feature level and individual scan level.

**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**

* **High Throughput**: process unlimited samples in <30 seconds each on standard hardware (Core i9, 16GB RAM).
* **Accuracy**: >98.6% precision in halogen detection across experimental LC-MS datasets.
* **Integration with GNPS**: Enhance Molecular Network Annotation in the Element Dimension by Embedding DeepHalo Results into GNPS Output GraphML File
* **Efficient Dereplication**: Significantly Higher Efficiency Compared to Molecular Networking Alone in GNPS

**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

Replace **/path/to/mzml\_files** with the path to your .mzML file or directory, and **/output\_directory** with your desired output directory.

**2. Dereplication with GNPS and Custom Database**

halo dereplication -o **/output\_directory** -g **/path/GNPS\_results** -ud **/path/custom\_database.csv**

Replace **/output\_directory** with the **/output\_directory** in step 1, **/path/GNPS\_results** with **Unzipped GNPS results directory containing .GraphML format file**, **/path/custom\_database.csv** with **the director your own database in csv or JSON format**

**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

[-ms2] # Enable MS2 extraction (optional)

**2. Dereplication**

halo dereplication

-o <project\_path> # Project directory

-g <GNPS\_folder> # Unzipped GNPS results directory containing .GraphML format file

-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](https://chat.deepseek.com/a/chat/s/d3b7de1c-627f-43db-b551-a75fc99a7c87#dependencies).

**License**

Distributed under the [MIT License](https://github.com/xieyying/DeepHalo/blob/main/LICENSE).

*For methodology details and benchmarks, refer to the*[*GitHub repository*](https://github.com/xieyying/DeepHalo)*.*