M2F Python Package - API Reference

This document describes all public functions and classes available in the M2F package.

Logging Utilities

configure_logging(logs_dir: str, file_level: int = logging.DEBUG, console_level: int = logging.WARNING)

Configures a timed rotating file handler and a console handler for logging.

Args:

- logs_dir (str): Directory where log files will be stored.
- file_level (int): Logging level for the file handler. Default is logging.DEBUG.
- console_level (int): Logging level for the console handler. Default is logging.WARNING.

Usage:

```
from M2F import configure_logging
configure_logging("logs")
```

Mining Utilities

extract_accessions_from_humann(file_path: str) -> list

Extracts UniProt accessions from a HUMAnN output file.

extract_all_accessions_from_dir(directory: str) -> list

Extracts accessions from all HUMAnN files in a directory.

fetch_uniprotkb_fields(accessions: list, fields: list) -> dict

Fetches specific UniProtKB fields for given accessions.

fetch_save_uniprotkb_batches(accessions: list, fields: list, output_dir: str)

Fetches UniProtKB data in batches and saves to disk.

Usage:

```
from M2F import extract_all_accessions_from_dir, fetch_save_uniprotkb_batches
accs = extract_all_accessions_from_dir("humann_outputs")
fetch_save_uniprotkb_batches(accs, ["sequence", "go_id"], "uniprot_data")
```

Cleaning Utilities

clean_col(df: pd.DataFrame, col_name: str) -> pd.Series

Cleans and normalizes the content of a specific column.

clean_cols(df: pd.DataFrame, col_names: list) -> pd.DataFrame

Cleans multiple columns in a DataFrame.

Embedding Utilities

MultiHotEncoder

Encodes categorical variables into multi-hot vectors.

GOEncoder

Encodes Gene Ontology (GO) terms into fixed-length vectors.

FreeTXTEmbedder

Generates embeddings for free-text annotations.

AAChainEmbedder

Generates embeddings for amino acid sequences using ESM-2.

ECEncoder

Encodes Enzyme Commission numbers into structured embeddings.

Feature Engineering Utilities

embed_ft_domains(df: pd.DataFrame, embedder, inplace=True)

Embeds functional domains in protein sequences.

embed_AAsequences(df: pd.DataFrame, embedder, inplace=True)

Embeds amino acid sequences.

embed_freetxt_cols(df: pd.DataFrame, embedder, col_names: list)

Embeds free-text columns.

encode_go(df: pd.DataFrame, encoder)

Encodes GO terms.

encode_ec(df: pd.DataFrame, encoder)

Encodes EC numbers.

encode_multihot(df: pd.DataFrame, encoder)

Encodes categorical data into multi-hot vectors.

empty_tuples_to_NaNs(df: pd.DataFrame)

Replaces empty tuples with NaN values.

save_df(df: pd.DataFrame, name: str, metadata: dict = None)

Saves a heterogeneous DataFrame to Zarr format.

load_df(path: str) -> pd.DataFrame

Loads a DataFrame from a Zarr archive.

Utility Module

util

Helper functions for composing functions, suppressing warnings, and listing files.