CLI

Perform Exponential Feature Embedding on Biosynthetic Gene Clusters (BGCs)

Usage:

```
$ [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--install-completion`: Install completion for the current shell.
- `--show-completion`: Show completion for the current shell, to copy it or customize the installation.
- `--help`: Show this message and exit.

Commands:

- `extract-embeddings`: Extract embeddings from BGC probabilistic...
- `load-for-efe`: Load BGC matrix for EFE model input.
- `train-efe-model`: Train a probabilistic EFE model on...
- `calculate-dissimilarity`: Calculate dissimilarity scores for BGCs...
- `infer-reference-models`: Infer embeddings against a reference EFE...
- `calculate-gmm-novelty`: Calculate novelty scores for BGCs using a...

`extract-embeddings`

Extract embeddings from BGC probabilistic EFE models

Usage:

```
$ extract-embeddings [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--help`: Show this message and exit.

Commands:

- `extract-embeddings`: Extracts learned BGC context embeddings or...

`extract-embeddings extract-embeddings`

Extracts learned BGC context embeddings or specific domain embeddings from a trained EFE model.

Usage:

```
$ extract-embeddings extract-embeddings [OPTIONS]
```

Options:

- `--model-path PATH`: Path to the trained EFE model file [required]
- `--bgc-map-path PATH`: Path to the BGC index map JSON [required]
- `--domain-map-path PATH`: Path to the domain index map JSON [required]
- `--output-path PATH`: Directory to save the extracted embeddings [required]
- `--embedding-dim INTEGER`: Dimensionality of the embeddings [default: 64]
- `--data-source TEXT`: Set to 'bgc' to extract BGC context embeddings, or domain_, where value is 0, 85, 170, 255 for specific domain embeddings. [default: bgc]
- `--help`: Show this message and exit.

load-for-efe

Load BGC matrix for EFE model input.

Usage:

```
$ load-for-efe [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--help`: Show this message and exit.

Commands:

- `main`

load-for-efe main

Usage:

```
$ load-for-efe main [OPTIONS]
```

Options:

- `--input-tsv PATH`: Path to the input BGC-feature matrix [required]
- `--output-dir PATH`: Directory to save output files [required]
- `--help`: Show this message and exit.

train-efe-model

Train a probabilistic EFE model on long-form BGC feature data.

Usage:

```
$ train-efe-model [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--help`: Show this message and exit.

Commands:

- `main`

`train-efe-model main`

Usage:

```
$ train-efe-model main [OPTIONS]
```

Options:

- `--long-df-path PATH`: Path to the long-form DataFrame TSV [required]
- `--bgc-map-path PATH`: Path to the BGC index map JSON [required]
- `--domain-map-path PATH`: Path to the domain index map JSON [required]
- `--output-dir PATH`: Directory to save the model and training history [required]
- `--embedding-dim INTEGER`: [default: 64]
- `--batch-size INTEGER`: [default: 1024]
- `--epochs INTEGER`: [default: 30]
- `--learning-rate FLOAT`: [default: 0.001]
- `--help`: Show this message and exit.

`calculate-dissimilarity`

Calculate dissimilarity scores for BGCs using a trained EFE model.

Usage:

```
$ calculate-dissimilarity [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--help`: Show this message and exit.

Commands:

- `compute-dissimilar-scores`

`calculate-dissimilarity compute-dissimilar-scores`

Usage:

```
$ calculate-dissimilarity compute-dissimilar-scores [OPTIONS]
```

Options:

- `--input-tsv PATH`: Path to input BGC-feature TSV [required]
- `--model-path PATH`: Path to trained EFE model (.pt) [required]
- `--output-tsv PATH`: Path to save output TSV with novelty scores [required]
- `--help`: Show this message and exit.

infer-reference-models

Infer embeddings against a reference EFE model

Usage:

```
$ infer-reference-models [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--help`: Show this message and exit.

Commands:

- `infer-embeddings-cli`

infer-reference-models infer-embeddings-cli

Usage:

```
$ infer-reference-models infer-embeddings-cli [OPTIONS]
```

Options:

- `--reference-model-path PATH`: Path to the trained EFE reference model (.pt file) (e.g., MiBiG model). [required]
- `--domain-map-path PATH`: Path to the domain index map JSON from the reference model's data. [required]
- `--input-bgc-map-path PATH`: Path to the BGC index map JSON for the input (experimental) data. [required]
- `--reference-bgc-map-path PATH`: Path to the BGC index map JSON from the reference model's training. [required]
- `--input-matrix-path PATH`: Path to your raw input (e.g., experimental) BGC feature matrix TSV. [required]
- `--output-path PATH`: Path to save the inferred BGC embeddings TSV. [required]

- `--embedding-dim INTEGER`: Dimension of the embeddings (must match trained reference model). [default: 64]
- `--help`: Show this message and exit.

calculate-gmm-novelty

Calculate novelty scores for BGCs using a trained EFE model.

Usage:

```
$ calculate-gmm-novelty [OPTIONS] COMMAND [ARGS]...
```

Options:

- `--help`: Show this message and exit.

Commands:

- `calculate-gmm-novelty`: Calculates novelty scores for experimental...

calculate-gmm-novelty calculate-gmm-novelty

Calculates novelty scores for experimental BGCs based on a GMM fitted to MiBiG reference embeddings. This command will augment your original experimental matrix with a 'novelty_score' column and save it.

Usage:

```
$ calculate-gmm-novelty calculate-gmm-novelty [OPTIONS]
```

Options:

- `--mibig-embeddings-path PATH`: Path to the MiBiG BGC embeddings TSV (the reference anchor, output from 'extract-embeddings'). [required]
- `--experimental-embeddings-path PATH`: Path to your inferred experimental BGC embeddings TSV (output from 'infer-embeddings-cli'). [required]
- `--original-experimental-matrix-path PATH`: Path to the original experimental BGC feature matrix TSV. Novelty scores will be added to this output. [required]
- `--output-novelty-path PATH`: Path to save the augmented experimental matrix (with novelty scores) as TSV. [required]
- `--plot-output-path PATH`: Optional: Path to save a histogram of novelty scores (e.g., .png).
- `--gmm-n-components INTEGER`: Number of components for the Gaussian Mixture Model. If not provided, it will be auto-determined using BIC/AIC.
- `--gmm-n-components-min INTEGER`: Minimum number of components to test for GMM auto-selection. [default: 1]
- `--gmm-n-components-max INTEGER`: Maximum number of components to test for GMM auto-selection. [default: 20]

- `--gmm-covariance-type TEXT`: Type of covariance parameters ('full', 'tied', 'diag', 'spherical'). 'full' is most flexible. [default: full]
- `--gmm-n-init INTEGER`: Number of initializations to perform for GMM. Higher is more robust but slower. [default: 10]
- `--random-state INTEGER`: Random state for GMM reproducibility.
- `--gmm-selection-criterion TEXT`: Information criterion to use for GMM component auto-selection ('bic' or 'aic'). [default: bic]
- `--help`: Show this message and exit.