
Supplementary Materials

Dataset Documentation and Intended Uses

We provide a data card for this dataset, `DataCard-AsEP.md`, which can be downloaded using this link: https://drive.google.com/file/d/1fc5kFcmUdKhyt3WmS30oLLPgnkyEeUjJ/view?usp=drive_link

This dataset provides a unified benchmark for researchers to develop new machine-learning-based methods for the epitope prediction task.

Access to the Dataset

There are two alternative sources where users can download the dataset:

- The dataset can be downloaded using the Python interface provided by our GitHub Repository `AsEP-dataset`. Detailed instructions on how to download the dataset are provided in the README file. Briefly, after installing the provided Python module, `asep`, the dataset can be downloaded by running the following command in the terminal:

```
download-asep /path/to/directory AsEP
# For example, to download the dataset to the current directory, run
# download-asep . AsEP
```

- The dataset and benchmark are provided through Zenodo at <https://doi.org/10.5281/zenodo.11495514>.
- Code and Dataset interface is provided in our GitHub Repository `AsEP-dataset` at <https://github.com/biochunan/AsEP-dataset>.

Author Statement

The authors affirm that they bear all responsibility in case of violation of rights, etc., and confirm the data license. The dataset is licensed under the **CC BY 4.0** License (<https://creativecommons.org/licenses/by/4.0/>), which is provided through the Zenodo repository. The code is licensed under the **MIT License** (<https://opensource.org/licenses/MIT>).

Hosting, Licensing, and Maintenance Plan

The dataset is hosted on Zenodo, which provides a DOI (10.5281/zenodo.11495514) for the dataset. It also comes with a Python interface provided in our GitHub Repository, `AsEP-dataset` at <https://github.com/biochunan/AsEP-dataset>, where users can submit issues and ask questions. Future releases and updates will be made available through the same channels. As discussed in the main text, the future plan includes expanding the dataset to include novel types of antibodies, such as single-domain antibodies, and providing more sophisticated features for graph representations. The dataset will be maintained by the authors and will be available for a long time.

Links to Access the Dataset and Its Metadata

The dataset, benchmark, and metadata are provided through Zenodo.

The Dataset

The dataset is constructed using `pytorch-geometric Dataset` module. The dataset can be loaded using the following code:

```
from asepv1_data.asepv1_dataset import AsEPv1Evaluator

evaluator = AsEPv1Evaluator()

# example
torch.manual_seed(0)
```

```

y_pred = torch.rand(1000)
y_true = torch.randint(0, 2, (1000,))

input_dict = {'y_pred': y_pred, 'y_true': y_true}
result_dict = evaluator.eval(input_dict)
print(result_dict) # got {'auc-prc': tensor(0.5565)}
% \end{lstlisting}

```

We also provide detailed documentation of the dataset content on Zenodo and include a description below:

- `asepv1-AbDb-IDs.txt`: A text file containing the AbDb identifiers of the 1723 antibody-antigen pairs in the dataset.
- `asepv1_interim_graphs.tar.gz`: Contains 1723 `.pt` files, where each file is a dictionary with structured data:
 - `abdbid` A string representing the antibody AbDb identifier.
 - `seqres` A dictionary containing:
 - `ab` An `OrderedDict` mapping string chain labels H and L to their corresponding sequence strings, representing heavy and light chains respectively.
 - `ag` A dictionary mapping string chain labels to their corresponding sequence strings.
 - `mapping` Includes:
 - `ab` Contains:
 - `seqres2cdr` A binary numpy array indicating the CDR positions in the antibody sequence.
 - `ag` Contains:
 - `seqres2surf` A binary numpy array indicating the surface residues in the antigen sequence.
 - `seqres2epitope` A binary numpy array indicating the epitope residues in the antigen sequence.
 - `embedding` Comprises:
 - `ab` Includes embeddings computed using the AntiBERTy model and ESM2 model for the antibody sequences.
 - `ag` Includes embeddings for the antigen sequences computed using the ESM2 model.
 - `edges` Describes the interactions:
 - `ab` A sparse coo tensor representing the binary edges between the CDR residues.
 - `ag` A sparse coo tensor representing the binary edges between the surface residues.
 - `stats` Metadata about each antibody-antigen pair, including counts of CDR, surface, and epitope residues, and the epitope-to-surface ratio.
- `structures.tar.gz`: Contains 1723 `pdb` structures, each named using the AbDb identifier.
- `split_dict.pt`: Contains the `train/val/test` splits of the dataset, with splits based on the epitope ratio and epitope group of the antigen.

Long-term Preservation

The current version of the dataset and benchmark are provided through Zenodo, which provides long-term storage. Future versions will be made available through the same channel and users are encouraged to submit queries and issues through the issues channel on the GitHub repository.

Explicit License

The dataset is licensed under the **CC BY 4.0** license (<https://creativecommons.org/licenses/by/4.0/>), which is provided through the Zenodo repository. The code is licensed under the MIT License <https://opensource.org/licenses/MIT>.

Benchmarks

Detailed benchmark experiments and results are provided on Zenodo (<https://doi.org/10.5281/zenodo.11495514>), and the file `benchmark.zip` contains the instructions on how to reproduce the results. To run ESMFold, EpiPred, and MaSIF-Site, we provided docker images on the Zenodo repository or instructions on how to obtain them from DockerHub. The benchmark results are reproducible by following the instructions provided in the zip file. For the method WALLE and its variants used in the ablation studies, their configuration YAML files and the trained models are also provided on Zenodo.