

Supplementary Materials for Seq2DomML

1. Overview

1.1 Purpose of this supplementary guide

This supplementary guide describes the contents of the `supplementary_materials/` archive provided alongside the Seq2DomML manuscript. The purpose of this guide is to document the dataset files, ECOD annotation resources, train and held-out split identifiers, feature composition, and resources used throughout the Seq2DomML workflow. The directory structure of the archive is shown below in Section 1.2, and Sections 2–5 describe the contents of each of the four main folders within the `supplementary_materials/` folder.

1.2 Archive directory structure

```
supplementary_materials/
├── data_and_processing/
│   ├── raw_downloaded_sequences.csv
│   ├── target_encoded_sequences.csv
│   └── deduplicated_sequences.csv
├── ecod/
│   ├── domains_ecod.txt
│   └── sequences_ecod.txt
├── splits/
│   ├── train_entry_ids.txt
│   └── test_and_val_entry_ids.txt
└── features/
```

```
|— full_feature_set_names.txt
|— phase1_025_feature_names.txt
|— phase2_selected_feature_names.txt
|— full_feature_set_types.png
|— phase1_025_feature_types.png
|— phase2_selected_feature_types.png
|— aa_indices/
|   |— indices.txt
|   └— lookup_tables.txt
|— interaction_matrices/
|   └— *.csv
|— positional_feats/
|   └— pos_feats.txt
└— bin_feats/
    └— bin_feats.txt
```

2. Dataset and Processing Files ([data_and_processing/](#))

2.1 raw_downloaded_sequences.csv

The file [data_and_processing/raw_downloaded_sequences.csv](#) contains the initial collection of bacterial protein chain sequences used at the beginning of the Seq2DomML dataset construction workflow. These sequences were obtained from the Protein Data Bank (PDB) and were restricted to protein chains annotated within the Kingdom of Bacteria, solved

using X-ray diffraction, and associated with a refinement resolution of less than or equal to 2.5 Å. This file therefore represents the pre-deduplication sequence pool generated after the primary structural and taxonomic inclusion criteria were applied. The columns in this csv file are ‘Entry ID’ and ‘Sequence’, where the Entry ID is encoded via the four character PDB Entry ID followed by the chain letter, separated by an underscore like so: ‘1A0G_B’.

2.2 target_encoded_sequences.csv

The file [data_and_processing/target_encoded_sequences.csv](#) contains the residue-level target-labeled version of the raw bacterial protein sequence dataset. This file was generated from [raw_downloaded_sequences.csv](#) by mapping ECOD domain annotations onto each corresponding protein chain sequence using the ECOD files provided in the [ecod/](#) folder, including [domains_ecod.txt](#) and [sequences_ecod.txt](#). The target labels are stored in the ‘targ_binary’ column as a binary sequence whose length matches the amino acid sequence in the ‘Sequence’ column. Each residue position is therefore assigned a corresponding binary label based on whether that residue falls within or outside an ECOD-annotated domain region. In this encoding, label 0 represents residues located within annotated domain regions, while label 1 represents residues outside annotated domain regions semantically referring to non domain regions including terminal flanking regions, linkers, and domain boundaries. The file also includes the ‘Entry ID’ column as described in the [raw_downloaded_sequences.csv](#).

2.3 deduplicated_sequences.csv

The file [data_and_processing/deduplicated_sequences.csv](#) contains the target-encoded sequence dataset after sequence redundancy reduction. This file was generated from [target_encoded_sequences.csv](#) by removing highly similar protein chains using 90% global sequence identity clustering. During this step, sequences exceeding the 90% identity threshold were grouped into redundancy clusters, and one representative sequence was retained from each cluster to reduce overrepresentation of near-duplicate proteins and limit possible data leakage between downstream training, validation, and test sets. The file preserves the residue-level binary target labels in the ‘targ_binary’ column, the amino acid sequence in the ‘Sequence’ column, and the corresponding ‘Entry ID’ column for each retained protein chain.

3. ECOD Annotation Files ([ecod/](#))

3.1 domains_ecod.txt

The file [ecod/domains_ecod.txt](#) contains the ECOD domain annotation records used to identify the residue ranges corresponding to annotated protein domains. This file was used during target encoding to map ECOD domain ranges onto each matching PDB chain sequence

and generate the residue-level ‘targ_binary’ labels in [target_encoded_sequences.csv](#). The header of this file documents the ECOD database source information, including the database path/version entry, ECOD version develop292, domain list version 1.7, and the Grishin Lab ECOD source. Each non-header row corresponds to one ECOD domain annotation and includes fields such as the source PDB identifier, ECOD domain identifier, domain source type, topology identifier, family identifier, ECOD unique identifier, and residue range. The residue range field specifies the chain and residue positions assigned to the domain, such as A:4-330 for a continuous domain region or A:1-303,A:349-356 for a discontinuous domain made up of multiple annotated sequence segments.

3.2 sequences_ecod.txt

The file [ecod/sequences_ecod.txt](#) contains the ECOD domain sequences in FASTA format. Each entry begins with a header line containing the ECOD domain identifier, the annotated chain residue range, and the ECOD unique identifier, followed by the amino acid sequence assigned to that ECOD domain. This file was used alongside [domains_ecod.txt](#) during target encoding to support the mapping of ECOD annotated domain regions onto the corresponding sequences from [raw_downloaded_sequences.csv](#). The amino acid sequences in this file provide an additional sequence-level reference that are used to verify the alignment of ECOD domain annotations against the raw PDB chain sequences when generating the residue-level ‘targ_binary’ labels.

4. Dataset Split Files ([splits/](#))

4.1 train_entry_ids.txt

The file [splits/train_entry_ids.txt](#) contains the Entry IDs assigned to the training dataset. Each line corresponds to one protein chain using the same Entry ID format described previously, where the four character PDB Entry ID is followed by the chain letter. These identifiers define the subset of target-encoded, deduplicated sequences used for model training.

4.2 test_and_val_entry_ids.txt

The file [splits/test_and_val_entry_ids.txt](#) contains the Entry IDs assigned to the test set and validation dataset. The test set and validation set did not overlap, although they were retained in the same folder for brevity. Each line corresponds to one protein chain using the same Entry ID format described previously. These identifiers define the held out sequences used outside the training set for validation, model selection, final evaluation, and comparison against the external domain annotation tools.

5. Feature Set Overview ([features/](#))

5.1 full_feature_set_names.txt

The file `features/full_feature_set_names.txt` contains the complete list of residue-level feature names included in the original Seq2DomML feature space before feature reduction. This full feature set contains 2,389 total features, including positional features, binary amino acid identity features, amino acid index features, and interaction matrix derived features. The three positional features are listed as `rel_pos`, `dist_to_term`, and `sinusoidal_pos`. Binary amino acid identity features are indicated by the prefix `bin`, with one feature corresponding to each of the 20 standard amino acids. Amino acid index features are indicated by the prefix `pc_`, representing residue-level physicochemical and biochemical properties derived from amino acid index lookup tables. Interaction matrix derived features are named using the amino acid of interest followed by the matrix identifier, such as `'M_CSEM940101'` for a given residue's propensity to interact with a Methionine amino acid. Each interaction matrix contributes 20 features, one for each standard amino acid, representing matrix-derived residue interaction or propensity scores associated with that amino acid. This file therefore provides the complete naming reference for the full residue-level feature matrix used before correlation-based pruning and NSGA-III feature selection.

5.2 phase1_025_feature_names.txt

The file `features/phase1_025_feature_names.txt` contains the list of residue-level features retained after the first phase of feature reduction. This file represents the 25% subset of the original 2,389-feature space that was retained after correlation-based feature pruning. During this step, features from `full_feature_set_names.txt` were ranked according to their absolute correlation to the binary target labels, and the 25% most target-correlated features were retained to reduce the feature space before evolutionary feature selection. This produced a reduced feature set of 598 features, which was then used as the input feature pool for the NSGA-III feature selection stage. The feature names follow the same naming conventions described for the full feature set.

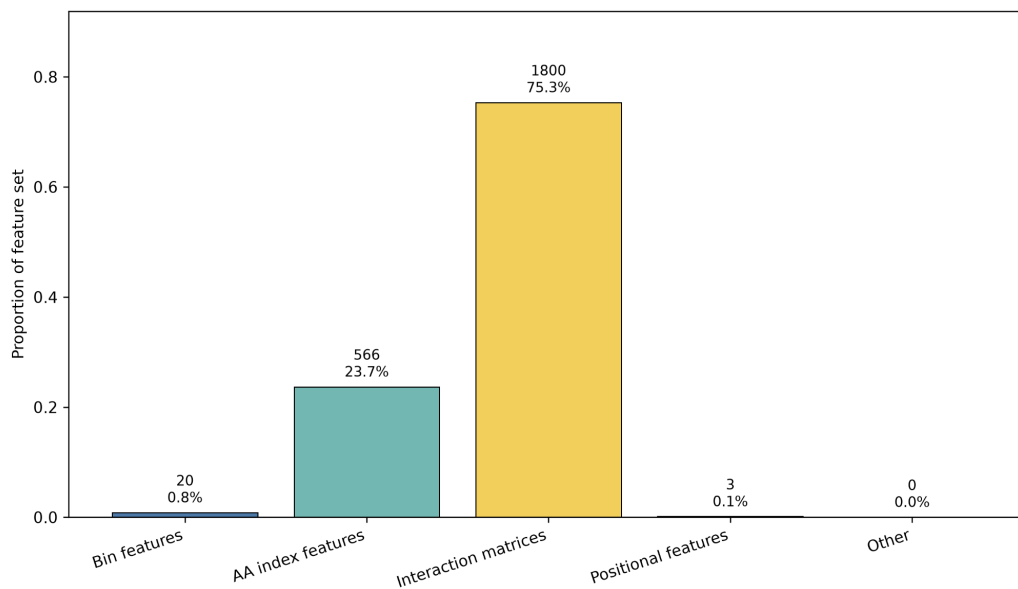
5.3 phase2_selected_feature_names.txt

The file `features/phase2_selected_feature_names.txt` contains the final list of residue-level features selected after the second phase of feature selection. This file was generated from the reduced feature pool in `phase1_025_feature_names.txt` through NSGA-III multi-objective evolutionary feature selection. During this step, different combinations of Phase 1 features were evaluated according to multiple validation objectives, including macro F1-score, non-domain precision, and non-domain recall. The final selected feature set contains 531 features and corresponds to the feature subset used for the selected Seq2DomML model configuration. The feature names follow the same naming conventions described for the full feature set.

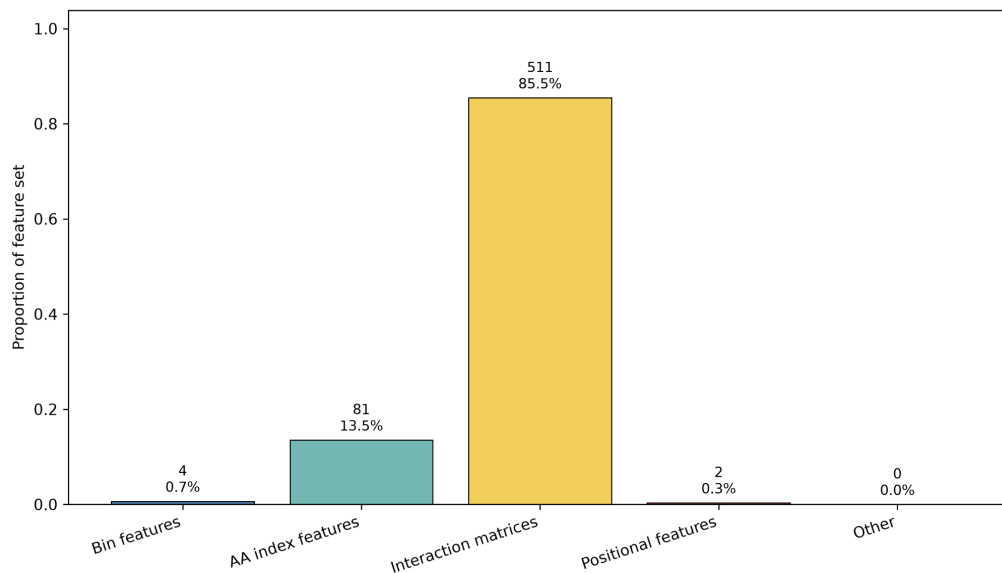
5.4 full_feature_set_types.png, phase1_025_feature_types.png,
phase2_selected_feature_types.png

The files [features/full_feature_set_types.png](#), [features/phase1_025_feature_types.png](#), and [features/phase2_selected_feature_types.png](#) provide visual summaries of feature type composition across the three major feature selection stages. The first figure summarizes the complete 2,389-feature input space before feature reduction, the second summarizes the 598-feature subset retained after Phase 1 correlation-based pruning, and the third summarizes the final 531-feature subset retained after Phase 2 NSGA-III feature selection. These figures are included to show how the relative contributions of positional features, binary amino acid identity features, amino acid index features, and interaction matrix derived features changed across each feature set variant in the feature reduction workflow.

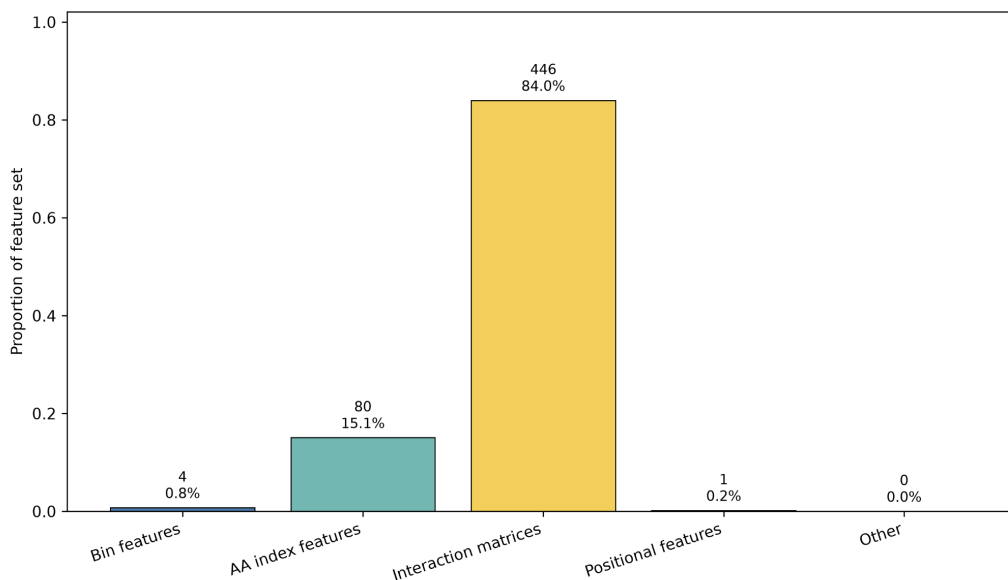
full_feature_set_types.png



phase1_025_feature_types.png



phase2_selected_feature_types.png



5.5 aa_indices

The folder [features/aa_indices/](#) contains the amino acid index resources used to generate the residue-level physicochemical feature group. These files define the lookup tables used to convert each amino acid residue into numerical values representing biochemical or physicochemical properties.

5.5.1 indices.txt

The file `features/aa_indices/indices.txt` contains the full amino acid index lookup tables used during feature extraction from aaindex.org. Each lookup table maps the 20 standard amino acids to numerical values for a specific residue property, such as hydrophobicity, flexibility, residue volume, or secondary structure tendency. The provided example represents one of the many indices present in this file:

```
_HYDROPHOBICITY_INDEX_ARGOS_ET_AL_1982 = {  
    "A": 0.61,  
    "L": 0.6,  
    "R": 0.06,  
    "K": 0.46,  
    "N": 1.07,  
    "M": 0.0,  
    "D": 0.47,  
    "F": 0.07,  
    "C": 0.61,  
    "P": 2.22,  
    "Q": 1.53,  
    "S": 1.15,  
    "E": 1.18,  
    "T": 2.02,  
    "G": 1.95,  
    "W": 0.05,  
    "H": 0.05,  
    "Y": 2.65,
```

```
"I": 1.88,  
"V": 1.32  
}
```

5.5.2 lookup_tables.txt

The file `features/aa_indices/lookup_tables.txt` contains the ordered list of amino acid indices feature names for reference in the feature extraction workflow. This file serves as a manifest for the available amino acid index features and links each lookup table name to its corresponding dictionary in `indices.txt`.

5.6 interaction_matrices/

The folder `features/interaction_matrices/` contains the interaction matrix CSV files used to generate the interaction matrix derived feature group, downloaded from aaindex.org. This folder contains 90 interaction matrix files, with each file storing a 20 by 20 amino acid matrix. Each matrix defines numerical scores for pairwise relationships between the 20 standard amino acids. During feature extraction, each matrix contributes 20 residue-level features, one for each amino acid column in the matrix. Across the 90 interaction matrices, this produces 1,800 total interaction matrix derived features in the full feature set. These features are named using the amino acid of interest followed by the matrix identifier, allowing each feature to be traced back to the corresponding CSV file in this folder.

5.6.1 Interaction Matrices table example from RIER950101.csv

The table below shows an example of one of 90 interaction matrices from this folder, specifically of `RIER950101.csv`. In this format, the row and column labels correspond to the 20 standard amino acids, and each cell contains the matrix-specific score for that amino acid pair. This example illustrates the structure used across the interaction matrix CSV files in `features/interaction_matrices/`.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	100.0	13.0	60.0	33.0	98.0	64.0	39.0	96.0	71.0	91.0	93.0	35.0	89.0	87.0	89.0	94.0	98.0	98.0	94.0	94.0
R	13.0	100.0	53.0	81.0	11.0	49.0	74.0	17.0	42.0	4.0	6.0	78.0	2.0	0.0	24.0	19.0	16.0	11.0	19.0	7.0
N	60.0	53.0	100.0	73.0	58.0	96.0	79.0	64.0	89.0	51.0	52.0	75.0	49.0	47.0	71.0	66.0	63.0	58.0	66.0	54.0
D	33.0	81.0	73.0	100.0	30.0	68.0	94.0	36.0	61.0	23.0	25.0	98.0	21.0	19.0	44.0	39.0	35.0	31.0	38.0	26.0

C	98.0	11.0	58.0	30.0	100.0	62.0	36.0	94.0	69.0	93.0	95.0	33.0	91.0	89.0	86.0	91.0	95.0	99.0	92.0	96.0
Q	64.0	49.0	96.0	68.0	62.0	100.0	74.0	68.0	93.0	55.0	57.0	71.0	53.0	51.0	76.0	71.0	67.0	63.0	70.0	58.0
E	39.0	74.0	79.0	94.0	36.0	74.0	100.0	43.0	68.0	29.0	31.0	96.0	27.0	26.0	50.0	45.0	41.0	37.0	44.0	33.0
G	96.0	17.0	64.0	36.0	94.0	68.0	43.0	100.0	75.0	87.0	89.0	39.0	85.0	83.0	93.0	98.0	99.0	94.0	98.0	90.0
H	71.0	42.0	89.0	61.0	69.0	93.0	68.0	75.0	100.0	62.0	64.0	64.0	60.0	58.0	83.0	78.0	74.0	69.0	77.0	65.0
I	91.0	4.0	51.0	23.0	93.0	55.0	29.0	87.0	62.0	100.0	98.0	26.0	98.0	96.0	79.0	84.0	88.0	93.0	85.0	97.0
L	93.0	6.0	52.0	25.0	95.0	57.0	31.0	89.0	64.0	98.0	100.0	27.0	96.0	94.0	81.0	86.0	90.0	94.0	87.0	99.0
K	35.0	78.0	75.0	98.0	33.0	71.0	96.0	39.0	64.0	26.0	27.0	100.0	24.0	22.0	46.0	41.0	38.0	33.0	41.0	29.0
M	89.0	2.0	49.0	21.0	91.0	53.0	27.0	85.0	60.0	98.0	96.0	24.0	100.0	98.0	78.0	83.0	86.0	91.0	83.0	95.0
F	87.0	0.0	47.0	19.0	89.0	51.0	26.0	83.0	58.0	96.0	94.0	22.0	98.0	100.0	76.0	81.0	84.0	89.0	81.0	93.0
P	89.0	24.0	71.0	44.0	86.0	76.0	50.0	93.0	83.0	79.0	81.0	46.0	78.0	76.0	100.0	95.0	91.0	87.0	94.0	83.0
S	94.0	19.0	66.0	39.0	91.0	71.0	45.0	98.0	78.0	84.0	86.0	41.0	83.0	81.0	95.0	100.0	96.0	92.0	99.0	88.0
T	98.0	16.0	63.0	35.0	95.0	67.0	41.0	99.0	74.0	88.0	90.0	38.0	86.0	84.0	91.0	96.0	100.0	96.0	97.0	91.0
W	98.0	11.0	58.0	31.0	99.0	63.0	37.0	94.0	69.0	93.0	94.0	33.0	91.0	89.0	87.0	92.0	96.0	100.0	93.0	96.0
Y	94.0	19.0	66.0	38.0	92.0	70.0	44.0	98.0	77.0	85.0	87.0	41.0	83.0	81.0	94.0	99.0	97.0	93.0	100.0	88.0
V	94.0	7.0	54.0	26.0	96.0	58.0	33.0	90.0	65.0	97.0	99.0	29.0	95.0	93.0	83.0	88.0	91.0	96.0	88.0	100.0

5.7 positional_features/pos_feats.txt

The file `features/positional_feats/pos_feats.txt` contains the feature manifest for the three positional features generated directly during Seq2DomML feature extraction. These features do not originate from an external lookup table and are instead calculated from each residue's zero based index and the total length of the protein sequence. These three features are `rel_pos`, `dist_to_term`, and `sinusoidal_pos`. The `rel_pos` feature represents the normalized linear position of a residue across the full sequence, `dist_to_term` represents the normalized distance from the residue to the nearest sequence terminus, and `sinusoidal_pos` provides a sinusoidal encoding of residue position across the sequence. These three values are calculated for every residue and concatenated into the final feature matrix before the binary amino acid identity features.

5.8 bin_feats/bin_feats.txt

The file `features/bin_feats/bin_feats.txt` contains the feature manifest for the binary amino acid identity features generated directly during Seq2DomML feature extraction. These features use one-hot encoding to represent the identity of each residue using the 20 standard amino acids. The file documents the 20 binary amino acid identity features included in the full residue-level feature set. These features are named from `bin_A` through `bin_Y`, with one feature corresponding to each accepted standard amino acid. For each residue, the feature

matching the residue's amino acid identity is assigned a value of 1, while all other binary amino acid identity features are assigned a value of 0. This feature block is calculated for every residue and concatenated into the final residue feature matrix after the positional feature block.