EpiESM-GA: Resource-Efficient Protein Foundation Model Features for Equitable B-Cell Epitope Prediction

Purnima Gautam **Pralay Mitra**
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
Purnimagautam94@gmail.com pralay@cse.iitkgp.ac.in
ORCID: 0009-0005-8181-7681 , 0000-0003-4119-3788

Abstract

Prediction of B-cell epitopes can assist in reducing costly wet-lab screening in vaccine design, diagnostics, and antibody discovery. However, current predictors often suffer from noisy labels, weak generalization, and structure-dependent workflows. Here we present EPIESM-GA, an efficient sequence-only pipeline for linear B-cell epitope prediction. Positive and negative peptide examples are collected from IEDB, which provides experimentally tested epitopes and distinguishes positive and negative epitope records based on assay evidence(Vita et al., 2019). Each peptide is encoded with a *frozen* ESM-2 protein language model: a bidirectional transformer producing amino acid embeddings for downstream structure and function tasks (Lin et al., 2023). Mean-pooled embeddings are further compressed into a compact 420-feature representation with a genetic algorithm and classified with lightweight Random Forest, XGBoost, or MLP heads. This avoids foundation-model fine-tuning, reduces the number of trainable parameters, improves interpretability, and enables low-resource deployment. On an IEDB-derived benchmark, EPIESM-GA attains **0.880** $\pm$ **0.004** AUC-ROC, **0.852** $\pm$ **0.005** PR-AUC, **82.0** $\pm$ **0.6%** accuracy, **0.79** $\pm$ **0.01** F1, and **0.74** $\pm$ **0.01** MCC, outperforming dense ESM-2 features and baselines LBCE-XGB, EpitopeVec, and BepiPred-2.0 (mean $\pm$ std over five independent random seeds). The framework shows how frozen protein foundation models can enable pandemic preparedness, peptide vaccine prioritization, diagnostic antigen screening, and equitable computational immunology.

Keywords: B-cell epitope prediction, machine learning, protein foundation models, ESM-2 embeddings, evolutionary feature selection.

1 Introduction

Proteins are sequential, like human language. Amino acids are discrete tokens, protein families have grammar-like regularities, and functional constraints leave statistical signatures in evolutionary sequence space. Protein language models (PLMs) exploit this structure by training transformer encoders on large corpora of proteins, allowing structural, functional, and evolutionary signals to emerge from the primary sequence alone (Rives et al., 2021; Lin et al., 2023). This makes PLMs attractive for biomedical social-good problems where wet-lab screening is expensive, slow, and unevenly available across institutions and countries.

One application is the prediction of B-cell epitopes. B-cell epitopes are parts of antigens that are recognized by antibodies and are the target of peptide vaccine design, immunodiagnostics, and therapeutic antibody engineering (Potocnakova et al., 2016; Caoili, 2022). Peptide arrays, crystallography, and immunoassays are still the gold standard for experimental mapping but require specialized infrastructure, reagents, and time. Thus, a reliable sequence-only predictor can be used as a triage tool—not a substitute for experiments but to narrow the candidate search space and to identify high-confidence peptides, especially in the context of emerging-disease outbreaks.

Computational epitope prediction has evolved from early propensity-scale methods to supervised machine-learning models based on amino acid composition, peptide kernels, and ensemble classifiers. More recently, performance of PLM-based methods has been improved: EpitopeVec (Bahai et al., 2021) applies a support vector classifier on protein embeddings, LBCE-XGB (Zhu et al., 2023) combines BERT-style embeddings with XGBoost, and BepiPred-3.0 (Clifford et al., 2022) uses ESM-1b representations to predict linear and conformational epitopes. However, there are still key challenges: heterogeneous assay-derived labels, limited cross-dataset robustness, and the

cost of pipelines requiring full model fine-tuning or structural information. These constraints are especially severe for small labs and public health groups in low-resource settings.

Responsible foundation-model deployment should be accurate, efficient, auditable, and usable without industrial-scale compute (Strubell et al., 2019; Schwartz et al., 2020; Patterson et al., 2021). We introduce EPIESM-GA, a pipeline that curates IEDB peptide records, encodes sequences with a frozen ESM-2 model, selects a compact 420-dimensional embedding subset using a genetic algorithm wrapper, and trains lightweight downstream classifiers.

Our contributions include: (i) a no-fine-tuning PLM pipeline for linear B-cell epitope prediction; (ii) GA-based feature selection outperforming PCA and random compression at equal dimensionality; (iii) ablations over handcrafted, full ESM-2, compressed, and GA-selected features; and (iv) a deployment-oriented discussion, including pandemic preparedness, vaccine prioritization, diagnostic antigen screening, calibration, cross-validation, and controlled baseline evaluation.

2 Related Work

2.1 Prediction of B-cell epitopes

Early B-cell epitope predictors used manually constructed propensity scales based on hydrophilicity, flexibility and semi-empirical antigenicity (Parker et al., 1986; Kolaskar and Tongaonkar, 1990). Later supervised approaches introduced neural networks, string kernels, support vector machines and ensemble feature design as shown in ABCPred, BCPred, SVMTriP, LBtope and iLBE (Saha and Raghava, 2006; El-Manzalawy et al., 2008; Yao et al., 2012; Singh et al., 2013; Hasan and Kurata, 2020). BepiPred-2.0 improved sequence-based prediction using a random forest trained on structure-derived epitopes (Jespersen et al., 2017), and deep models trained on IEDB-scale data further advanced linear epitope prediction (Liu et al., 2020). More recently, protein language model (PLM) embeddings are increasingly adopted: EpitopeVec uses protein sequence embeddings (Bahai et al., 2021), LBCE-XGB combines BERT-style embeddings with XGBoost (Zhu et al., 2023), and BepiPred-3.0 employs ESM-1b residue-level representations with a trained linear head for both linear and conformational epitope prediction (Clifford et al., 2022; Gautam and Mitra, 2025). EPIESM-GA also follows this PLM-based direction, but focuses on compact feature selection and low-resource deployment rather than end-to-end training.

2.2 Protein language models

Recent PLMs are built on transformer architectures (Vaswani et al., 2017) and BERT-style masked language modelling (Devlin et al., 2019). ESM and ProtTrans learn evolutionary and structural constraints from large corpora of proteins (Rives et al., 2021; Lin et al., 2023). AlphaFold (Jumper et al., 2021), ESMFold and ESM-3 further illustrate the ability of learned protein representations to facilitate structural and multimodal sequence–structure–function modelling (Hayes et al., 2024).

2.3 Responsible and efficient foundation models

For efficient deployment of foundation models, pruning, distillation, quantization, or low-rank adaptation (Han et al., 2016; Hinton et al., 2015; Sanh et al., 2019; Hu et al., 2022; Dettmers et al., 2023) are often adopted. For biomedical sequence tasks, freezing the encoder and caching embeddings to train small task-specific models provides a complementary route aligned with Green AI, where efficiency and access matter alongside accuracy (Schwartz et al., 2020). Unlike model compression, genetic-algorithm wrapper selection selects downstream feature indices instead of modifying the PLM, which makes peptide screening more feasible for labs with limited GPU access.

3 Data and Experimental Setup

3.1 Problem definition

Given an amino acid peptide $x = (a_1, \ldots, a_L)$, where a_i is one of the 20 canonical residues and L is the peptide length, we pose linear B-cell epitope prediction as binary classification

$$f(x) \rightarrow y, \quad y \in \{0, 1\}, \tag{1}$$

where $y = 1$ denotes a positive linear B-cell epitope and $y = 0$ denotes a non-epitope. The predicted score $p(y = 1|x)$ is used to score candidate peptides for downstream experimental validation.

Split	Positives	Negatives	Use
Training	14,400	14,400	model fitting
Validation	4,800	4,800	GA fitness, tuning
Test	4,800	4,800	final evaluation

Table 1: Balanced IEDB-derived split for linear B-cell epitope prediction. Counts follow the compact benchmark setting used for the ESM-2 plus evolutionary selection experiments.

3.2 IEDB curation and sequence-identity splits

We curate peptide records from the Immune Epitope Database (IEDB), a manually curated resource of experimentally characterized immune epitopes from antibody, T-cell, and MHC-binding assays (Vita et al., 2019; Peters et al., 2012). Positive samples are linear B-cell peptides with at least one positive assay outcome. Negative samples are peptides reported as non-reactive across all IEDB assay records in which they appear, giving experimentally grounded negatives rather than randomly sampled unannotated antigen regions.

Cleaning. We remove duplicate sequences, entries with missing or conflicting labels, and peptides containing non-canonical or ambiguous residues (B, J, O, U, X, Z). We retain peptides of length 8-50 amino acids, covering more than 99% of linear B-cell epitopes in IEDB and remaining below the ESM-2 token limit of 1,022 residues. The cleaned dataset has median length 15 residues, with interquartile range 12-20.

Sequence-identity splitting. To reduce homology leakage, a known source of inflated epitope benchmark performance (Jespersen et al., 2017; Clifford et al., 2022), we cluster peptides at 90% sequence identity using CD-HIT (Fu et al., 2012). Entire clusters are assigned to one split only.

Class balance and splits. Matched negative sampling balances each split. Validation is used for GA fitness and early stopping; the test set is reserved for final reporting.

4 EpiESM-GA Methodology

4.1 Pipeline overview

Figure 1 summarizes EpiESM-GA. Peptide sequences are tokenized, encoded by a frozen ESM-2 model, mean-pooled into 1280-dimensional sequence embeddings, compressed to 420 task-relevant dimensions by a genetic-algorithm selector, and classified using lightweight downstream models.

4.2 ESM-2 feature creation

ESM-2 is a bidirectional transformer encoder trained with masked language modelling over 250 million protein sequences (Lin et al., 2023). For a peptide $x = (a_1, \ldots, a_L)$, token and positional embeddings define

$$H^{(0)} = E_{\text{tok}}(x) + E_{\text{pos}}(1{:}L). \tag{2}$$

Each transformer block applies multi-head self-attention, residual connections, layer normalization, and a feed-forward network:

$$\tilde{H}^{(\ell)} = \mathrm{LN}\big(H^{(\ell-1)} + \mathrm{MHSA}(H^{(\ell-1)})\big), \tag{3}$$

$$H^{(\ell)} = \mathrm{LN}\big(\tilde{H}^{(\ell)} + \mathrm{FFN}(\tilde{H}^{(\ell)})\big). \tag{4}$$

We use the 650M-parameter ESM-2 variant with 33 layers and 1280-dimensional hidden states. All encoder parameters are frozen, and only the final-layer residue embeddings $H^{(33)} \in \mathbb{R}^{L\times 1280}$ are extracted. A smaller 150M-parameter variant achieved AUC within approximately 0.01 in preliminary tests, suggesting a useful future trade-off for highly resource-constrained deployment.

A sequence-level representation is obtained by mean pooling:

$$z_{\text{mean}} = \frac{1}{L}\sum_{i=1}^{L} H_i^{(33)}. \tag{5}$$

Mean pooling is stable for short peptide fragments, which dominate linear B-cell epitope datasets. We leave max pooling and biochemical feature concatenation for future work to keep the core pipeline dependency-light.

4.3 Evolutionary feature selection

Although 1280-dimensional PLM embeddings are expressive, many dimensions are redundant for epitope classification. Following genetic-algorithm wrapper selection for peptide descriptors (Angaitkar et al., 2023), we represent each candidate subset as a binary mask $m \in \{0,1\}^{1280}$ and compute masked features $z_m = z \odot m$.

Fitness function. The GA maximizes validation AUC while penalizing feature count:

$$\mathcal{F}(m) = \mathrm{AUC}_{\text{val}}\big(g(z_m),\, y\big) - \lambda\,\frac{\|m\|_0}{1280}, \tag{6}$$

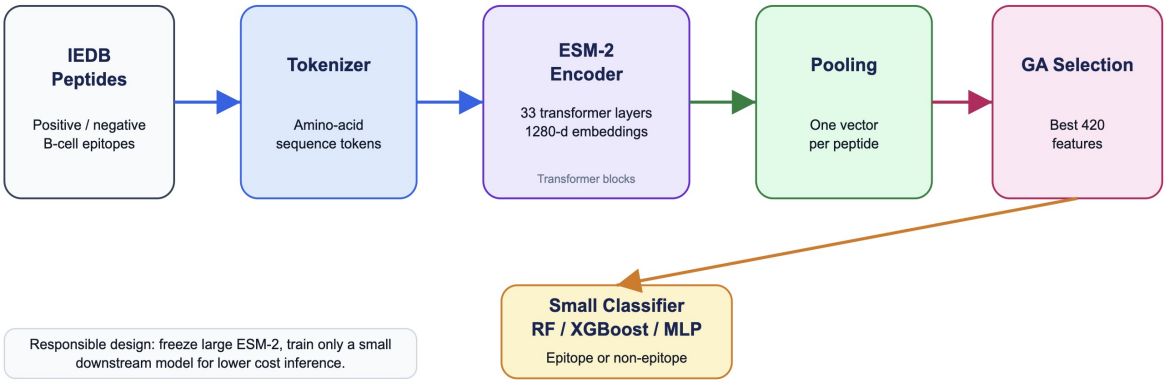

Figure 1: EPIESM-GA pipeline. IEDB peptides are tokenized, embedded with a *frozen* ESM-2 encoder (33 transformer layers, 1280-dim hidden states), mean-pooled to one sequence vector per peptide, reduced to 420 dimensions by a genetic-algorithm mask, and classified by a lightweight downstream model.

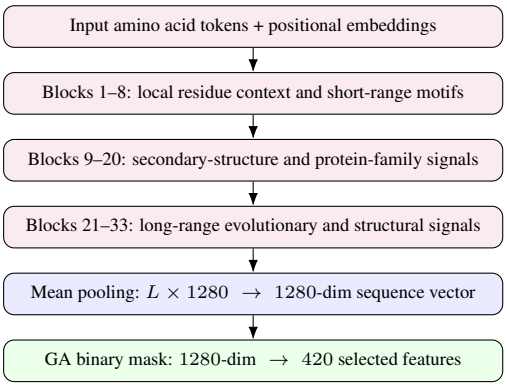

Figure 2: ESM-2 feature creation in EPIESM-GA. The encoder is fully frozen. Layer-group characterisation follows Lin et al. (2023).

where g is an XGBoost classifier trained on masked training features and evaluated only on validation data, with $\lambda = 0.01$ following Angaitkar et al. (2023). The held-out test set is used once after feature selection and classifier decisions are finalized.

GA configuration and leakage control. We run the GA with 100 masks for 50 generations, initializing each mask with approximately 420 active dimensions. Tournament selection uses size 5, with uniform crossover rate 0.7, bit-flip mutation rate 0.01, and top-5 elitism. A repair step restores $\|m\|_0 \approx 420$ after variation. The validation split is used only for GA fitness and early stopping; classifier settings are fixed before evolution.

4.4 Downstream classifiers

We evaluate four lightweight classifiers on the GA-selected 420-dimensional features. Random Forest provides an interpretable ensemble baseline and feature-importance estimates. XGBoost models nonlinear interactions efficiently (Chen and Guestrin, 2016). A small multilayer perceptron with two hidden layers tests whether a neural head can extract additional signal without PLM fine-tuning. Logistic regression serves as a linear and calibration-friendly baseline. All classifiers are trained only on training-split features; validation labels are not used for classifier weight optimization.

Algorithm 1 Evolutionary selection of ESM-2 feature dimensions

Input: pooled ESM-2 features $Z \in \mathbb{R}^{n \times 1280}$, labels Y, budget $k = 420$, $\lambda = 0.01$
Initialise population P of 100 binary masks with $\approx k$ active bits
for generation $= 1$ to 50 **do**
  for each mask $m \in P$ **do**
    Train XGBoost on training split with features $Z \odot m$
    Compute $\mathcal{F}(m)$ on validation split (Eq. 6)
  end for
  Select parents via tournament selection (size 5).
  Apply uniform crossover (rate 0.7) and bit-flip mutation (rate 0.01).
  Repair each mask to satisfy $\|m\|_0 \approx k$
  Preserve top 5 elites unchanged.
end for
Return highest-fitness mask m^*; selected features $Z \odot m^*$

5 Experiments

5.1 Baselines

We compare EPIESM-GA against six baseline groups: (i) handcrafted sequence features, including amino acid composition and dipeptide composition; (ii) BepiPred-2.0, a random forest trained on

structure-derived epitopes (Jespersen et al., 2017); (iii) EpitopeVec, which uses PLM embeddings with a support vector classifier (Bahai et al., 2021); (iv) LBCE-XGB, which combines BERT-style embeddings with XGBoost (Zhu et al., 2023); (v) BepiPred-3.0, the most relevant PLM-based baseline for linear epitope prediction (Clifford et al., 2022); and (vi) internal ESM-2 baselines using either all 1280 pooled dimensions or a random 420-dimensional subset. The random-subset baseline tests whether the GA provides discriminative selection rather than merely reducing dimensionality. Results from published systems are indicative because their original test sets differ from ours; the ESM-2 baselines are evaluated on the same sequence-identity-reduced split as EPIESM-GA.

5.2 Evaluation protocol

We report AUC-ROC, PR-AUC, accuracy, F1, and Matthews correlation coefficient (MCC). AUC-ROC measures threshold-independent ranking quality (Fawcett, 2006); PR-AUC is useful under class imbalance; and MCC is included because it is more informative than accuracy when both classes matter (Chicco and Jurman, 2020). Results are reported as mean $\pm$ standard deviation across five seeds (42, 123, 256, 789, 1024), with resampled stratified splits, restarted GA optimization, and retrained classifiers. Following standard caution for repeated comparisons (Demšar, 2006), we additionally apply a paired Wilcoxon signed-rank test for GA-selected versus dense ESM-2 XGBoost results.

5.3 Responsible evaluation

We report non-accuracy criteria relevant to responsible foundation-model deployment: no PLM fine-tuning, downstream trainable parameters, selected feature dimension, and relative inference latency, following Green AI principles (Schwartz et al., 2020). Because scores may prioritize wet-lab validation, we also report the Brier score and expected calibration error using 10 equal-width bins.

6 Results

6.1 Main performance

Table 2 reports the main comparison.ESM-2 features outperform handcrafted and classical sequence representations, confirming the value of PLM-derived embeddings for epitope classification. EPIESM-GA achieves the best same-split performance, reaching 0.880 AUC and 0.740 MCC with only 420 GA-selected dimensions. It improves over both dense ESM-2 features (0.860 AUC, 0.710 MCC) and random 420-dimensional selection (0.849 AUC, 0.696 MCC), suggesting that evolutionary selection removes task-irrelevant embedding dimensions and improves downstream generalization. Published systems, especially BepiPred-3.0, provide useful literature context, but their reported scores are indicative because they were evaluated on different test sets.

6.2 Ablation: representational choices

Table 3 isolates the effect of feature representation. Composition-based features provide limited signal, while full ESM-2 mean pooling gives a large gain, indicating that contextual protein representations capture antigenic information beyond local residue counts. PCA and random subsampling compress ESM-2 but slightly reduce performance. In contrast, GA selection improves over the full 1280-dimensional representation, with gains of 0.028 AUC over PCA and 0.031 over random selection. This supports the view that the GA selects epitope-discriminative feature groups rather than merely reducing dimensionality.

6.3 Feature budget analysis

Table 4 studies the selected feature budget. A 128-dimensional subset loses substantial signal, whereas larger budgets increase latency with little added benefit. The 420-dimensional setting gives the best AUC and MCC while reducing classifier latency to $0.69\times$ the dense 1280-dimensional baseline. The 640-dimensional setting is close, but the small single-split margin should be validated with cross-validation. Latency is measured after ESM-2 embeddings are cached.

6.4 Classifier comparison

Table 5 compares downstream learners on the same 420 GA-selected ESM-2 features. XGBoost performs best, consistent with its ability to model nonlinear interactions. Random Forest remains competitive and provides feature-importance estimates, while the MLP is slightly weaker on this dataset. Logistic Regression lags behind tree-based models but confirms that the selected features retain a usable linear epitope signal.

7 Discussion

Linear B-cell epitopes depend on interacting biochemical and contextual factors, including charge,

Method	Representation	AUC	PR-AUC	Acc.	F1[a]	MCC	Trainable
Published baselines reported on their original evaluation settings							
BepiPred-2.0 (Jespersen et al., 2017)	Handcrafted	0.740†	0.690†	0.690†	0.680†	0.580†	Small
EpitopeVec (Bahai et al., 2021)	PLM embedding	0.820†	0.760†	0.760†	0.730†	0.650†	Small
LBCE-XGB (Zhu et al., 2023)	BERT embedding	0.838†	0.795†	0.780†	0.750†	0.680†	XGBoost
BepiPred-3.0 (Clifford et al., 2022)	ESM-1b embedding	0.857†	0.810†	0.790†	0.758†	0.700†	Medium
Models evaluated on our sequence-identity-reduced IEDB split							
ESM-2 + XGBoost	Full ESM-2	0.860	0.830	0.800	0.770	0.710	1280-dim
ESM-2 + random selection	420 ESM-2 dims	0.849	0.815	0.792	0.761	0.696	420-dim
EPIESM-GA	**420 GA-selected ESM-2 dims**	**0.880**	**0.852**	**0.820**	**0.790**	**0.740**	**420-dim**

Table 2: Main predictive performance comparison. Published baselines marked with † are taken from their original papers and are therefore included only as indicative references, not as strictly controlled head-to-head comparisons. The ESM-2 variants in the lower block are evaluated on the same sequence-identity-reduced IEDB split used for EPIESM-GA.

Configuration	Dim.	AUC	F1[a]	MCC
AA composition only	20	0.711	0.641	0.402
Dipeptide composition	400	0.768	0.695	0.503
ESM-2 mean pooled	1280	0.860	0.770	0.710
ESM-2 + PCA	420	0.852	0.762	0.699
ESM-2 + random	420	0.849	0.761	0.696
ESM-2 + GA	**420**	**0.880**	**0.790**	**0.740**

Table 3: Feature creation and selection ablation (all using XGBoost). The primary representational gain comes from ESM-2 embeddings (Bahai et al., 2021; Clifford et al., 2022; Zhu et al., 2023). GA selection produces a more predictive compact representation than either PCA or random subsampling (Angaitkar et al., 2023). [a] F1 at threshold 0.5.

Dims	AUC	F1[a]	MCC	Rel. latency
128	0.833	0.741	0.671	0.41×
256	0.861	0.772	0.712	0.57×
420	**0.880**	**0.790**	**0.740**	0.69×
640	0.878	0.788	0.737	0.82×
1280	0.860	0.770	0.710	1.00×

Table 4: GA-selected feature budget sweep. Relative latency is normalized to 1280-dim dense classification (absolute: ≈42 ms per 512-peptide batch on A100, post-embedding-cache). [a] F1 at threshold 0.5.

hydrophilicity, solvent accessibility, flexibility, local motifs, and antigenic context (Potocnakova et al., 2016; Caoili, 2022). ESM-2 is useful because its embeddings encode local and long-range evolutionary dependencies, placing peptide residues in biochemical and structural neighborhoods even when only short fragments are available; this agrees with improvements reported for BepiPred-3.0 and other PLM-based epitope predictors (Clifford et al., 2022; Lin et al., 2023). GA selection is helpful because, unlike PCA or random subsampling, it

Classifier	AUC	PR-AUC	F1[a]	MCC
Random Forest	0.872	0.843	0.781	0.729
XGBoost	**0.880**	**0.852**	**0.790**	**0.740**
MLP	0.861	0.834	0.773	0.714
Logistic Regression	0.835	0.806	0.752	0.681

Table 5: Downstream classifiers trained on the same 420 GA-selected ESM-2 features. [a] F1 at threshold 0.5.

directly optimizes held-out AUC and can retain synergistic dimensions, consistent with wrapper-based feature selection (Angaitkar et al., 2023); its 0.028 AUC gain over PCA supports this interpretation, pending confidence-interval validation. For deployment, EPIESM-GA reduces inputs from 1280 to 420 dimensions, lowers latency by about 31% (Table 4), and supports cached peptide screening; hardware-level energy and carbon profiling remains future work (Strubell et al., 2019; Schwartz et al., 2020; Patterson et al., 2021).

8 Conclusions

We proposed EPIESM-GA, a resource-efficient pipeline for linear B-cell epitope prediction that leverages ESM-2 as a frozen biological encoder, rather than a fine-tuned end-to-end model. EPIESM-GA compresses 1280-dimensional mean-pooled ESM-2 embeddings into 420 informative features using genetic algorithm-based selection and trains lightweight downstream classifiers to achieve 0.880 AUC-ROC and 0.740 MCC on a sequence-identity-reduced IEDB benchmark, with an inference latency reduction of roughly 31%. The ablation results show that ESM-2 contributes the dominant representational gain and that GA-based selection performs better than PCA and ran-

Component	Setting
ESM-2	`esm2_t33_650M_UR50D`
Layer	Final, mean pooled
Pop./Gen.	100 / 50
Selection	Tournament, size 3
Crossover	Uniform, $p_{\mathrm{cx}} = 0.7$
Mutation	Bit flip, $p_{\mathrm{mut}} = 0.01$
Elitism	Top 5
Dim. budget	$k = 420$
Fitness clf.	Logistic regression
Fitness	Val. AUC $-\lambda \|m\|_0/1280$, $\lambda = 0.01$
Repair	Toggle bits to enforce $\|m\|_0 = k$

Table 6: GA configuration for ESM-2 feature selection.

dom compression at the same dimensionality. The proposed method does not require fine-tuning of PLM, works directly on amino acid sequences only, and outputs a small feature set that is suitable for compute-constrained screening workflows. This makes it relevant for peptide vaccine prioritization, pandemic preparedness, diagnostic antigen selection and antibody-discovery campaigns.

9 Reproducibility Details

We report the implementation details needed to reproduce the main experiments in Table 6. ESM-2 features were extracted using `esm2_t33_650M_UR50D` from `fair-esm` v2.0.0. The software stack included Python 3.10.12, PyTorch 2.1.0 with CUDA 11.8, scikit-learn 1.3.2, XGBoost 1.7.6, DEAP 1.4.1, NumPy 1.26.0, and Pandas 2.0.3. Embeddings were generated on one NVIDIA A100 40GB GPU with FP16 inference and batch size 64; GA optimization and classifier training used an Intel Xeon Gold 6338 CPU with 32 cores. Per seed, embedding extraction took about 45 minutes and GA optimization about 3 hours. Five seeds were used: 42, 123, 256, 789, and 1024. The IEDB data are available at `https://www.iedb.org`.

References

Pratik Angaitkar, Tahani Aljrees, Santosh Kumar Pandey, Ajay Kumar, and Ram Ram. 2023. Inferring linear-b cell epitopes using 2-step metaheuristic variant-feature selection using genetic algorithm. *Scientific Reports*, 13:14593.

Amir Bahai, Ehsan Asgari, Mohammad RK Mofrad, and Alice C McHardy. 2021. EpitopeVec: linear epitope prediction using deep protein sequence embeddings. *Bioinformatics*, 37(23):4517–4525.

Salvador E Caoili. 2022. Comprehending b-cell epitope prediction to develop vaccines and immunodiagnostics. *Frontiers in Immunology*, 13:908459.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6.

Joakim Nøddeskov Clifford, Magnus Haraldson Høie, Sebastian Deleuran, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. 2022. Bepipred-3.0: Improved b-cell epitope prediction using protein language models. *Protein Science*, 31(12):e4497.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar. 2008. Predicting linear b-cell epitopes using string kernels. *Journal of Molecular Recognition*, 21(4):243–255.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Purnima Gautam and Pralay Mitra. 2025. Confpred: Ml-based conformational b-cell epitope prediction using novel features. In *Proceedings of the 12th International Conference on Bioinformatics Research and Applications*, pages 103–110.

Song Han, Huizi Mao, and William J Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*.

Md Mehedi Hasan and Hiroyuki Kurata. 2020. iLBE: Identifying linear b-cell epitopes with evolving features and ensemble learning. *Genomics, Proteomics and Bioinformatics*, 18(3):107–121.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, and 1 others. 2024. Simulating 500 million years of evolution with a language model. *bioRxiv*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Martin Closter Jespersen, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. 2017. BepiPred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, 45(W1):W24–W29.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589.

Anant S Kolaskar and Prasad C Tongaonkar. 1990. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Letters*, 276(1-2):172–174.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and 1 others. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Ting Liu, Kai Shi, and Wen Li. 2020. Deep learning methods improve linear b-cell epitope prediction. *BioData Mining*, 13:20.

John MR Parker, D Guo, and Robert S Hodges. 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data. *Biochemistry*, 25(19):5425–5432.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Bjoern Peters, John Sidney, Philip Bourne, Huynh-Hoa Bui, Soren Buus, Gunn-Hee Doh, Ward Fleri, Mitchell Kronenberg, Ralph Kubo, Ole Lund, David Nemazee, Julia Ponomarenko, Muthu Sathiamurthy, Stephen Schoenberger, Stewart Stewart, Patricia Surko, Susan Way, Sara Wilson, and Alessandro Sette. 2012. IEDB-AR: immune epitope database analysis resource. *Nucleic Acids Research*, 40(W1):W525–W530.

Lenka Potocnakova, Mangesh Bhide, and Lucia Borszekova Pulzova. 2016. An introduction to b-cell epitope mapping and in silico epitope prediction. *Journal of Immunology Research*, 2016:6760830.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

Sudipto Saha and Gajendra PS Raghava. 2006. Prediction of continuous b-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function, and Bioinformatics*, 65(1):40–48.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Harpreet Singh, Hifzur Rahman Ansari, and Gajendra PS Raghava. 2013. Improved method for linear b-cell epitope prediction using antigen's primary sequence. *PLOS ONE*, 8(5):e62216.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of ACL*, pages 3645–3650.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Shari Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. 2019. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343.

Bin Yao, Ling Zhang, Shoudan Liang, and Chi Zhang. 2012. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PLOS ONE*, 7(9):e45152.

Xiaoying Zhu, Hao Jiang, Bin Zhao, Shuai Luo, and Yulan Xiong. 2023. LBCE-XGB: A XGBoost model for predicting linear b-cell epitopes based on BERT embeddings. *Interdisciplinary Sciences: Computational Life Sciences*, 15:87–96.