*2025 International Conference on Applied Intelligence*
*November 6-9, Nanning, China*
http://www.icai.org.cn/2025/index.php

# ProAttUnet: Protein Secondary-Structure Prediction Re-imagined via ESM2-Enhanced U-Net Dual-Fusion

Long Cheng, Zhiqiang Hui, Anchi Sun

Suzhou University of Science and Technology

**Abstract.** Protein secondary structure prediction remains a pivotal concern within the domain of bioinformatics. In this innovative research, we introduce a novel methodology to further enhance a protein prediction model grounded in single sequences. Our key contribution lies in integrating the state-of-the-art (SOTA) model ESM2, which hails from the field of universal protein language models. By leveraging ESM2, we are able to acquire residual embeddings and contact maps for the protein sequences under study. Regarding the model architecture, we employ a unique dual-way U-Net framework for effective feature fusion. This framework is complemented by the integration of a cross-attention mechanism, enabling the model to capture more comprehensive context information. To better capture both local and global characteristics of protein sequences, we introduce the GCU_SE module into both the encoder and decoder. This module cascades a Gated Convolutional Unit (GCU) with a Squeeze-and-Excitation (SE) block: the GCU employs a dual-branch convolutional structure—one branch with ReLU activation and the other with Sigmoid gating—to selectively extract local features; the subsequent SE block performs global channel-wise recalibration, emphasizing informative channels. Unlike the standard SE block that only reweights channels globally, GCU_SE synergizes local feature selection with global channel refinement, enabling the model to more effectively perceive complex structural motifs. These innovative enhancements enable the ProAttUnet model to outperform the benchmark model SPOT-1D-Single by 1.6%, 3.5%, 1.0%, 4.6%, and 7.2% for ss3, and by 5.5%, 7.8%, 4.1%, 8.1%, and 10.1% for ss8 across five test sets (SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ and TEST2018, respectively). This significant improvement vividly demonstrates the effectiveness and novelty of our proposed model.

## 1 Methods

The study leverages the 650M-parameter ESM2 pretrained protein language model (based on BERT with RoPE) to extract residual embeddings (1280-dimensional per amino acid) and contact maps of protein sequences. Its architecture is a dual-path U-Net, including an encoder (3 downsampling modules + TCN layer, with max pooling and residual connections) and a decoder (3 upsampling modules via 1D convolution, with feature concatenation). It integrates cross-attention (query from embeddings, key/value from contact maps; 128/256/512 kernels) and GCU_SE modules in both encoder and decoder. Each GCU_SE block first applies a Gated Convolutional Unit: the input is processed by two parallel 3×1 convolutions—one followed by ReLU to generate main features, the other by Sigmoid to produce a soft gate; their element-wise product suppresses noisy positions while preserving local motifs. The gated output is then fed into a Squeeze-and-Excitation layer that globally pools, excites channel-wise weights through a bottleneck MLP (reduction 16×) and recalibrates the feature maps, achieving synergistic local selection and global channel refinement. A dynamic sliding window (window size 10–100, step size calculated via formula) handles variable sequence lengths. Evaluation uses accuracy, precision, recall, F1-score, and SOV. Experiments run on an Intel Xeon Platinum 8362 CPU, NVIDIA RTX 3090 GPU (CUDA 11.2), with TensorFlow 2.5 and Python 3.8.

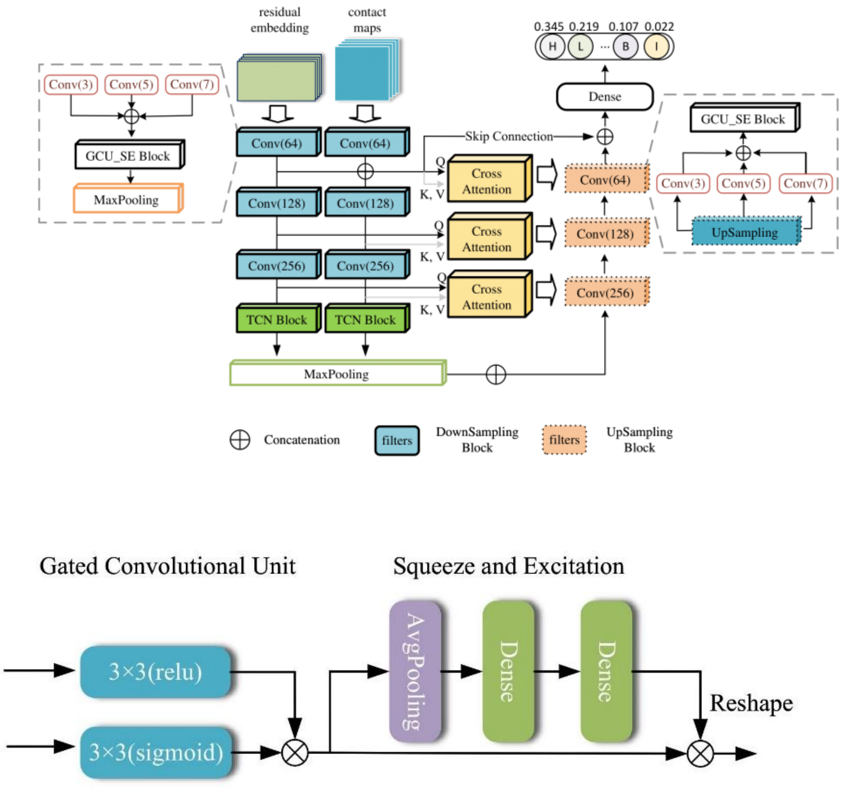

**Figure. 1.** The model's schematic diagram and the architecture of GCU_SE Block.

## 2 Datasets

Training relies on ProteinNet: initially 50,914 PDB entries (2016), 39,120 left after 95% truncation, 38,915 post-cleaning, and 25,166 selected (length 100–400, avoiding redundancy/noise). Evaluation uses 5 test sets: SPOT-2016 (2016–2020 PDB proteins, E-cutoff <0.1 excluded), SPOT-2016-HQ (SPOT-2016 with resolution <2.5Å, R-free

<0.25), SPOT-2018 (SPOT-2016-based, post-2018 proteins), SPOT-2018-HQ (SPOT-2018 with same HQ constraints), and TEST2018 (2018 proteins, 25% pre-2018 filter). Ablation uses CB513, TS115, CASP12 (sequence similarity >25%), and NEW364 (post-2019 proteins, length ⩾20, resolution <2.5Å). Dataset size experiments test 1000–15,000 sequences, with 100 sequences as validation.

**Table 1.** The 8 category labels in 4 testsets

| ss types\testsets | SPOT-2016 | SPOT-2016-HQ | SPOT-2018 | SPOT-2018-HQ |
|---|---|---|---|---|
| H | 99298 | 17486 | 34747 | 7183 |
| L | 93610 | 15251 | 36779 | 6476 |
| B | 2014 | 421 | 805 | 199 |
| E | 31884 | 10165 | 13503 | 4488 |
| G | 6516 | 1766 | 2466 | 772 |
| I | 299 | 10 | 19 | 5 |
| T | 23153 | 4908 | 8315 | 2105 |
| S | 23018 | 3585 | 8237 | 1596 |
| Overall Samples | 1473 | 295 | 682 | 125 |

## 3 Results

ProAttUnet outperforms SPOT-1D-Single (benchmark) across 5 test sets: for ss3, it is 1.6% (SPOT-2016), 3.5% (SPOT-2016-HQ), 1.0% (SPOT-2018), 4.6% (SPOT-2018-HQ), 7.2% (TEST2018) higher; for ss8, 5.5%, 7.8%, 4.1%, 8.1%, 10.1% higher (TEST2018 accuracy 72.2%). It has better precision/recall/F1 (e.g., superior AUC-PR for L/H/E/T on TEST2018) and SOV (10%/13%/8% higher for H/L/E vs SPOT-1D-Single; 17% higher for S, 2% for B). Removing GCU_SE reduces accuracy by 1.0–1.8% (e.g., TEST2018 -1.7%) and SOV. Dataset size ≥13,000 slows accuracy gain, 14,000 is optimal. Sliding window (25–150) boosts rare structures (B/G/S); 3 downsampling modules balance accuracy/error.

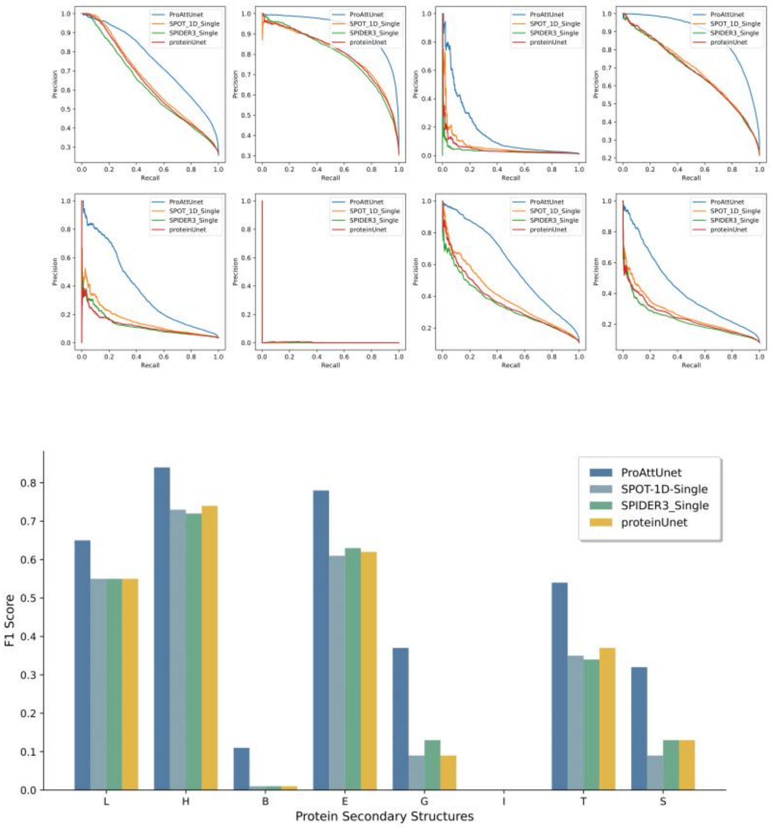

**Figure. 2.** The precision-recall curve(L,H,B,E,G,I,T,S in order) and F1 score on 8 categories and 4 models.

## 4 Conclusions

This study proposes ProAttUnet, a single-sequence-based model advancing protein secondary structure prediction. By integrating the ESM2 pretrained protein language model (650M parameters) for embeddings/contact maps, a dual-path U-Net with cross-attention for feature fusion, and GCU_SE modules for feature refinement, it outperforms the benchmark SPOT-1D-Single across five test sets (ss3: 1.0–7.2% higher; ss8: 4.1–10.1% higher). Ablation experiments confirm GCU_SE, dynamic sliding window (aiding rare structures like B/G/S), 3 downsampling modules, and 14,000-sequence dataset size optimize performance. Future work will explore knowledge graph multimodal learning and novel attention mechanisms to further enhance prediction accuracy.