다중 계층 Stacking Ensemble 기반 단백질 구조 품질 예측 모델

윤재하, 김나연, 김동혁, 전창재*
세종대학교

613jay@sju.ac.kr, nanakim0329@sju.ac.kr, kimdonghyuk@sju.ac.kr, *cchun@sejong.ac.kr

A Multi-level Stacking Ensemble for Protein Structure Quality Prediction

Jaeha Yoon, Nayeon Kim, Donghyuk Kim, Chang-Jae Chun*
Sejong Univ.

요 약

본 연구는 단백질 구조 품질 예측의 정확도 향상을 목표로 한다. 이를 위해, 제16회 단백질 구조 예측 평가 대회(CASP16)의 단백질 구조 품질 평가(Evaluation of Model Accuracy, EMA) 모델들의 예측 값을 결합하는 다중 계층 stacking ensemble 기법을 제안한다. 제안 모델은 개별 EMA 모델들의 예측 점수에 예측 분포와 통계적 특성을 추가 feature로 활용하여 최적의 예측 점수를 도출하도록 설계되었다. 제안 모델은 CASP16의 평가 방식을 그대로 재현했을 때, SCORE(전역 품질) 부문 1위 및 QSCORE(결합부 품질) 부문 2위를 달성하였으며, 세부적으로 두 부문 모두 loss 점수에서 1위를 기록하여 구조 선별 성능의 우수성을 입증하였다.

Ⅰ. 서론

단백질의 3차원 구조는 생체 내 기능과 작용 기작을 결정하는 핵심 요소이다. 이를 정확히 규명하는 것은 현대 구조 기반 신약 설계의 필수적인 선결 과제이다. 최근 AlphaFold와 같은 딥러닝 모델의 발전으로 단백질 구조 예측에 소요되는 비용과 시간이 크게 감소하였으나, 동시에 수많은 예측 결과 중 실제 생물학적 구조와 가장 일치하는 단백질 구조 모델을 선별해야 하는 난제를 야기했다. 특히, 정답 구조를 모르는 블라인드 예측 환경에서 단백질 모델의 품질을 정확하게 예측하는 EMA 기술의 중요성은 더욱 부각되고 있다. 이러한 흐름에 맞춰 단백질 구조 예측 분야에서 가장 공신력 있는 국제 대회인 CASP는 2022년 CASP15부터 복합체 품질 평가(EMA)를 정식 종목으로 채택했다. 그러나 2024년 CASP16의 최신 EMA 모델들은 단백질 모델 간의 유사성을 통해 품질을 예측하는 합의(Consensus) 방식의 한계로 인해 다양성이 부족할 때 성능이 급락하거나, 신약 개발의 핵심인 단백질 이종 복합체(Heteromer) 평가에서는 상대적으로 낮은 정확도를 보이는 등 한계를 드러냈다. 또한, 대규모 데이터셋에서 최적의 구조를 가려내는 선별 능력의 부족으로 고품질 구조가 존재함에도 이를 식별하지 못하는 문제가 지속되고 있다 [1].

따라서 본 연구는 이러한 한계를 극복하고자 앙상블(Ensemble) 기법인 다중 계층 stacking을 통해 CASP16의 주요 EMA 모델들의 예측 값을 효과적으로 결합한 모델을 제안한다. 궁극적으로, 개별 모델의 편향을 상호 보완하고 이종 복합체와 같은 고난도 타겟에서도 안정적이고 높은 평가 정확도를 달성하는 것을 목표로 한다.

Ⅱ. 본론

2.1. 데이터셋 및 특징 추출

본 연구는 모델 학습 및 검증에 PSBench의 CASP16 Community 데이터셋을 활용하였다. 데이터셋은 실제 CASP16에서 사용된 데이터로 이루어졌으며 총 39개 단백질 타겟 중 자원 제약으로 아미노산 시퀀스 길이가 5000이 넘어가는 H1217, H1227을 제외하고 37개 타겟이 포함됐다. 모든 타겟은 단백질 복합체이며 각 타겟에는 평균 332개의 예측된 단백질 모델이 존재하여 총 12294개의 샘플이 데이터셋에 포함된다 [2].

제안하는 모델은 개별 EMA 모델들이 예측한 점수를 메타 예측기의 feature로 사용하되 이를 단순히 결합하는 것을 넘어 집단 지성을 반영한 파생 변수를 생성해 EMA 모델들의 점수를 전략적으로 결합하도록 설계했다. 구체적으로 각 단백질 모델에 대해 모든 EMA 모델이 제출한 예측 값들의 평균, 표준편차, 최댓값, 최솟값, 왜도와 같은 기술 통계량을 산출하여 입력 벡터에 추가하였다. 이를 통해 메타 모델은 개별 EMA 모델의 예측 값이 전체의 경향성과 얼마나 일치하는지, 혹은 통계적으로 유의미한 편차를 보이는지를 학습하여 예측의 강건성을 확보한다. 각 EMA 모델은 단백질 모델의 3차원 구조를 나타내는 PDB 파일을 입력으로 받아 품질을 예측한다. 이때, EMA 모델 간 단백질 호환성 차이로 인해 일부 예측 값에 결측치가 발생할 수 있으며, 본 연구에서는 이를 중앙값(Median)으로 대체하여 메타 모델의 학습 안정성을 확보하였다.

2.2. 모델 구조 설계

모델 아키텍처는 일반화 성능을 극대화하기 위해 3단계 stacking 구조를 채택하였다 [3].

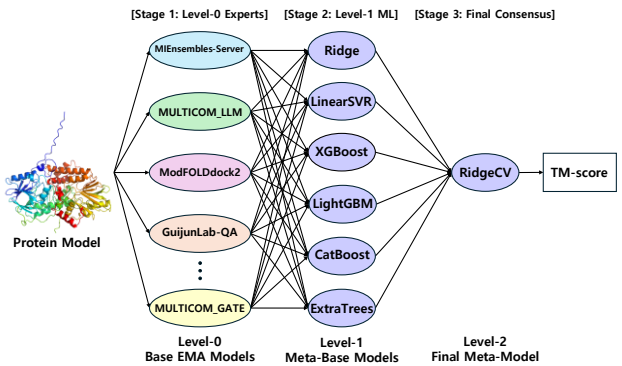

그림 1. 다중 계층 stacking 모델 구조도

1단계(Level-0) 기저 모델군(Base EMA Models)은 CASP16의 다양한 최신 모델을 앙상블하고자 그림 2의 상위 15개(2 ~ 16위) 모델로 구성됐다 [4]. 2단계(Level-1) 메타 기저 모델군(Meta-Base Learners)은 선형 모델(Ridge, LinearSVR)과 비선형적 패턴 및 변수 간

상호작용을 학습하기 위한 트리 기반 모델(XGBoost, LightGBM, CatBoost, ExtraTrees)을 포함한 총 6개의 모델로 구성되었다. 최종 예측을 수행하는 3단계(Level-2) 메타 모델(Final Meta Learner)로는 과적합 방지와 가중치 최적화에 유리한 RidgeCV를 선정했다 (그림 1).

2.3. 학습 전략 및 검증 방법

학습 과정에서 TM-score를 ground-truth 타겟으로 사용하였다. TM-score는 단백질 예측 구조와 정답 구조 간의 전반적인 topology 유사성을 정규화된 척도로 평가할 수 있으며, 단백질 구조의 실제 품질을 나타내는 지표로서 CASP 등에서 널리 활용되어 왔다 [1, 2]. 학습 시 TM-score의 분포가 정규성을 따르지 않을 경우, 일부 회귀 모델의 가정과 손실 최적화가 어려워지며 학습이 불안정해질 수 있다. 이를 완화하기 위해 본 연구에서는 타겟 값에 QuantileTransformer를 적용하여 정규분포에 가깝게 변환한 후 학습을 수행하였다. 또한 데이터 누출(Data Leakage)을 방지하고 일반화 성능을 공정하게 평가하기 위해, 5-fold 교차 검증 기반의 out-of-fold(OOF) 예측 방식을 적용하였다. 메타 모델은 각 기저 모델이 학습에 사용하지 않은 검증 데이터에서 생성된 예측 값만을 입력으로 사용한다.

성능 평가는 CASP16의 평가 방식을 그대로 재현하여 단백질-단백질 복합체 구조의 전역 품질을 다루는 SCORE 부문과 결합 부위 품질을 다루는 QSCORE 부문에서 RS(Ranking Score)를 산출해 비교하였다 [1]. RS는 식 (2)와 같이 Pearson 상관계수(P), Spearman 상관계수(S), AUC-ROC(R), 그리고 loss(L)의 Z-score를 선형 결합하여 계산된다. Pearson 상관계수의 Z-score는 단백질 타겟 t, 실제 품질 점수 지표 r, EMA 모델의 예측 p에 대해 식 (1)과 같이 정의되며, loss의 경우 분자에 -1을 곱하여 점수의 방향성을 통일한다. 여기서 ROC는 단백질 예측 구조 중 정확도 상위 25%를 참, 나머지를 거짓으로 하여 구해지며, loss는 실제 최적 구조와 모델이 예측한 최적 구조 간의 실제 품질 점수 차이를 의미한다.

$$P(r,p) = \sum_{t}^{targets} max\left(0, \frac{P(r,p,t) - \mu\big(P(r,t)\big)}{\sigma\big(P(r,t)\big)}\right) \quad (1)$$

$$RS(r,p) = 0.5 \times P(r,p) + 0.5 \times S(r,p) + R(r,p) + L(r,p) \quad (2)$$

단, SCORE 부문의 경우 일부 타겟에서 정답 구조가 비공개되어 Oligo-GDT_TS 산출이 제한되므로, 본 연구에서는 TM-score를 SCORE 부문의 단일 평가 지표로 사용하였다.

2.4. 실험 결과

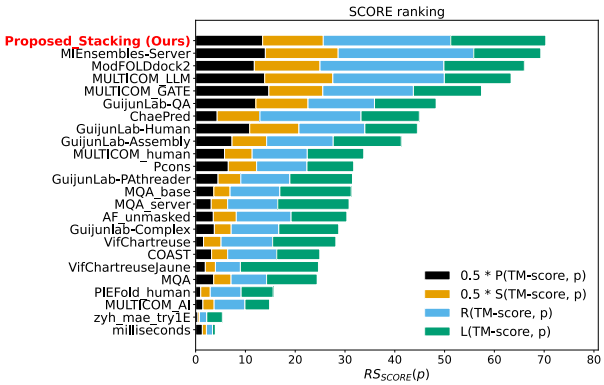

그림 2. CASP16 모델들과 비교한 SCORE 부문 RS 순위

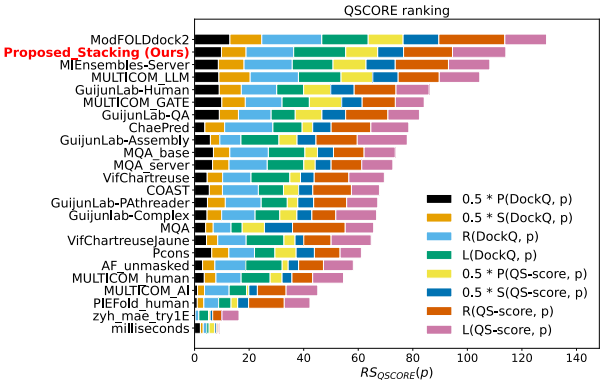

그림 3. CASP16 모델들과 비교한 QSCORE 부문 RS 순위

제안하는 모델과 CASP16 모델들의 성능을 비교한 결과, 제안 모델이 SCORE(전역 품질) 부문에서 1위, QSCORE(결합부 품질) 부문에서 2위를 달성했다. Loss 점수의 경우 두 부문 모두 1위를 기록했다 (그림 2와 3).

Ⅲ. 결론

제안 모델의 성능 평가 결과, 기존 CASP16의 SOTA 모델 대비 우수한 성능을 입증하였다. SCORE 부문에서 1위, QSCORE 부문에서는 2위를 하였으며 loss 점수에서는 두 부문 모두 1위를 하였다. Loss 점수에서 1위를 기록한 것은 제안 모델이 단순히 경향성을 예측하는 것을 넘어, 실제 최적 단백질 구조를 선별하는 과정에서 발생할 수 있는 성능 손실(Failure Cost)을 최소화할 수 있음을 시사하며, 실질적인 구조 예측 응용에 있어 가장 신뢰할 수 있는 모델임을 입증한다. 또한, 제안 모델은 전역 품질 지표(TM-score)를 예측 대상으로 설정했음에도, SCORE와 QSCORE 두 부문 모두 우수한 성능을 보여주었다. 이는 단일 지표에 편향되지 않고, 단백질 복합체의 전역 및 결합부 품질 모두에서 지표 간 일반화 성능이 뛰어났음을 보여준다.

이러한 결과는 다양한 EMA 및 메타 모델의 상호 보완적 결합과 집단 지성을 반영한 다중 계층 stacking 전략이 성능 향상과 예측 안정성 확보에 유의미하게 기여했음을 보여준다. 궁극적으로 제안 모델은 단백질 구조 예측 기반의 신약 설계 시 예측된 구조의 품질을 정확하게 평가함으로써, 부정확한 구조 선택으로 인한 시행착오와 비용 손실을 최소화할 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업 연구 결과로 수행되었습니다(IITP-2026-RS-2023-00254529).

참 고 문 헌

[1] Fadini, A., Studer, G., and Read, R. J., “Model quality assessment for CASP16,” *Proteins: Structure, Function, and Bioinformatics*, 2025.

[2] Neupane, P., Liu, J., and Cheng, J., “PSBench: a large-scale benchmark for estimating the accuracy of protein complex structural models,” *arXiv preprint*, arXiv:2505.22674, 2025.

[3] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[4] A. Fadini *et al.*, “Highlights of model quality assessment in CASP16,” *Proteins: Structure, Function, and Bioinformatics*, 2025.