단백질 도킹 구조 후보군 선별을 위한 Ensemble 기반 품질 평가 모델

황신혜, 윤재하, 전창재*
세종대학교

24010764@sju.ac.kr, 613jay@sju.ac.kr, *cchun@sejong.ac.kr

Ensemble-Based Quality Assessment Model for Protein Docking Candidate Structures

Shinhye Hwang, Jaeha Yoon, Chang-Jae Chun*
Sejong Univ.

요 약

본 연구는 단백질 도킹 구조 후보군 선별 단계에서 단일 AI 모델의 한계를 분석하고, Ensemble 기반 접근의 유효성을 검증하는 것을 목표로 한다. BM4(Protein-Protein Docking Benchmark 4.0) 데이터셋을 활용하여 단백질 구조 품질 평가(Evaluation of Model Accuracy, EMA) 모델들의 출력 점수를 결합한 Ensemble 모델을 구성하고, Logistic Regression, XGBoost, RandomForest 기반 모델들을 비교하였다. Top-N Success Rate(iRMSD ≤ 4 Å native 구조 포함 여부)를 기준으로 평가한 결과, Ensemble 기반 모델들은 전반적인 후보 선별 과정에서 단일 EMA 모델 대비 보다 안정적이고 일관된 Success Rate를 달성하였다. 이는 EMA 출력 간의 비선형 관계를 학습하는 Ensemble 기반 접근이 초기 후보 선별에서 효과적임을 확인하였다.

Ⅰ. 서론

신약 개발은 보통 10만 개 이상의 후보 물질 중 극히 일부만이 임상 단계에 도달하는 복잡한 고비용의 과정으로, 실제 치료 효능을 지닌 후보를 조기에 선별하는 것이 핵심 과제로 인식되고 있다. 이 과정에서 단백질 간 상호작용, 특히 항원-항체 결합과 같은 단백질 복합체 형성은 치료제 효능과 직결되는 중요한 단계 중 하나로, 도킹 구조의 품질을 신뢰성 있게 평가하는 것이 중요한 요소로 작용한다. 전통적인 실험 기반 선별 방법은 높은 시간적·경제적 비용을 요구하므로 단백질 도킹 시뮬레이션과 EMA, virtual screening과 같은 계산 기반 기법이 대규모 후보군 축소를 목적으로 활용되어 왔다. 그러나 단백질 도킹 구조 후보군은 타겟 단백질, 결합 형태, 생성 알고리즘에 따라 높은 이질성을 가지므로 단일 AI 모델만으로는 다양한 조건에서 일관된 성능을 확보하는 데 한계가 있다. 이러한 문제를 해결하기 위한 대안으로 Ensemble 접근법이 제안되어 왔으나, 단백질 도킹 구조 후보군 선별 문제를 중심으로 단일 모델과의 성능을 정량적으로 비교·분석한 연구는 아직 제한적이다.

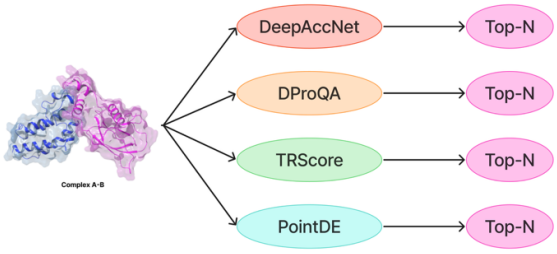
따라서 본 연구에서는 단백질 도킹 구조 후보군 선별을 주요 문제로 설정하고, 단일 EMA 모델과 Ensemble 기반 품질 평가 모델의 성능을 체계적으로 비교·분석함으로써 계산 기반 도킹 구조 품질 평가 과정에서 Ensemble 접근법이 효과적임을 검증하고자 한다.

Ⅱ. 본론

2.1. 데이터셋 및 특징 추출

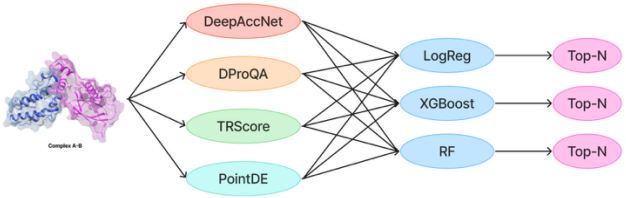
본 연구는 모델 학습 및 검증을 위해 BM4 데이터셋을 사용하였다. BM4는 단백질 타겟별 실제 결합 구조와 다양한 도킹 알고리즘으로 생성된 예측 도킹 구조들을 포함한다. 입력 feature로는 DeepAccNet [1], DProQA [2], TRScore [3], PointDE [4]와 같은 딥러닝 기반 EMA 모델의 출력과, MJ3H, PISA, SIPPER, TOBI, HADDOCK score, DFIRE2, DFIRE와 같은 전통적인 도킹 점수 함수들을 사용하였다. 반면, DockQ, LDDT, iRMSD 계열 지표 및 CAPRI classification은 실제 구조 기반 품질 지표이므로 입력 특징(feature)으로는 사용하지 않고 정답 라벨로만 활용하였다. 본 데이터셋에는 MetaScore에서 제안된 이진 라벨(is_native)이 포함되어 있으며, 이는 backbone iRMSD(iRMSD_bb) 4 Å 이하를 정답(is_native=1)으로 정의한 기준을 따른다. 훈련 데이터셋은 단백질별 클래스 불균형을 완화하기 위해 랜덤 언더샘플링이 적용된 데이터로 구성되었다. 또한, 테스트 데이터에서 is_native=1에 해당하는 샘플이 전혀 존재하지 않는 단백질(1RKE, 2X9A, 4IZ7)은 평가의 왜곡을 방지하기 위해 제외하였다. 이러한 데이터 구성과 전처리 설계를 통해 단일 모델과 Ensemble 기반 모델의 유효성을 공정하고 체계적으로 비교·분석할 수 있는 실험 환경을 구축하였다.

2.2. 모델 구조 설계



[그림 1] 단일 AI 모델 구조도

비교 실험을 위한 단일 AI 모델 구조에서는 각 EMA 모델을 독립적으로 사용하여 도킹 구조의 품질을 예측하였다. DeepAccNet, DProQA, TRScore, PointDE는 예측된 도킹 구조를 입력으로 받아 각각의 기준에 따라 native 가능성을 산출하며, 출력 점수를 기준으로 도킹 구조를 순위화한 후 상위 Top-N 후보를 선정하여 성능을 평가하였다. 그림 1과 같은 구조는 모델 간 정보 결합 없이 개별 EMA 모델의 예측 성능만을 기반으로 Top-N 성공률을 산출함으로써 Ensemble 구조와의 공정한 비교를 위한 기준(baseline)을 제공한다.



[그림 2] Ensemble 모델 구조도

Ensemble 모델 구조에서는 단일 모델 구조와 동일하게 DeepAccNet, DProQA, TRScore, PointDE를 사용하되, 각 EMA 모델의 출력 점수를 직접 순위화하지 않고 하나의 feature 집합으로 결합한다. 결합된 EMA 출력은 Logistic Regression(LogReg), XGBoost, RandomForest(RF) 모델의 입력으로 사용되며, 각 모델은 도킹 구조의 native 가능성을 예측하도록 학습된다. 이후 예측 확률을 기준으로 도킹 구조를 순위화하고 상위 Top-N 후보를 선정하여 성능을 평가한다. 그림 2와 같이 제안한 구조는 단일 AI 모델 기반 선별과 달리, 모델 간 보완적인 정보를 활용할 수 있다는 점에서 기존 방법 대비 구조적 차이를 가진다.

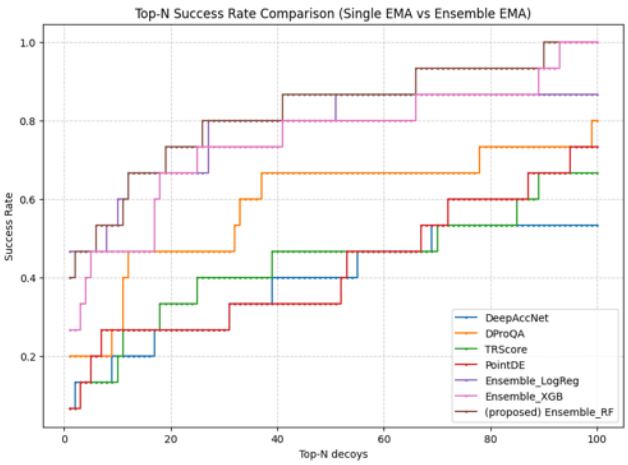
2.3. 학습 전략 및 검증 방법

본 연구에서는 Ensemble 구조에서 결합된 EMA 출력 점수를 입력 feature로 사용하여 LogReg, XGBoost, RF를 각각 독립적으로 학습하였다. 각 Ensemble 모델 학습에는 is_native 이진 라벨을 사용하였으며, 각 메타 모델은 테스트 데이터에 대해 도킹 구조가 native일 확률을 예측하도록 학습되었다.

모델 성능 평가는 Top-N Success Rate 곡선을 통해 수행하였다. 구체적으로, 각 단백질별로 예측 점수를 기준으로 도킹 구조(decoy)를 내림차순으로 정렬한 후, 상위 N개의 후보 중 iRMSD ≤ 4 Å에 해당하는 native 구조(is_native=1)가 하나라도 포함될 경우를 success로 정의하였다. Success Rate는 전체 단백질 타겟 중 성공으로 분류된 단백질의 비율로 계산되며, N을 1부터 100까지 증가시키며 Success Rate의 변화를 곡선 형태로 비교하였다.

동일한 평가 절차를 단일 EMA 모델과 각 메타 모델의 예측 결과에 모두 적용함으로써, 단일 AI 모델 기반 선별과 Ensemble AI 모델 기반 선별의 성능을 비교하였다.

2.4. 실험 결과



[그림 3] Top-N 후보들에 대한 Success Rate 비교

단일 EMA 모델 중에서는 DProQA와 TRScore가 상대적으로 높은 성능을 보였으나, 전반적인 구간에서 Success Rate의 증가 폭이 제한적이었으며, 특히 Top-N이 작은 구간에서는 Success Rate가 낮게 유지되는 경향이 관찰되었다. 이는 초기 후보 선별 단계에서의 신뢰성에 제약이 있음을 확인할 수 있다.

반면, 그림 3 에서와 같이 Ensemble 기반 모델들은 전반적인 Top-N 구간에서 단일 EMA 모델 대비 높은 Success Rate를 보였다. LogReg와 XGBoost 기반 모델 역시 성능 향상을 보였으나, RF 기반 모델이 전 구간에 걸쳐 가장 높은 Success Rate를 기록하며, 특히 중·소 Top-N 영역에서 뚜렷한 성능 우위를 나타냈다.

Ⅲ. 결론

본 연구에서는 단백질 도킹 구조 후보군 선별 문제를 대상으로, 단일 EMA 모델과 EMA 모델 출력 점수를 결합한 Ensemble 기반 모델의 성능을 Top-N Success Rate 관점에서 비교·분석하였다. 실험 결과, 단일 EMA 모델은 일부 Top-N 구간에서 제한적인 성공률을 보인 반면, Ensemble 기반 모델들은 전반적인 후보 선별 과정에서 보다 안정적이고 일관된 Success Rate를 달성하였다. 이는 대규모 단백질 도킹 구조 후보군 선별 단계에서 단일 모델만으로는 충분한 신뢰성을 확보하기 어렵다는 한계를 실험적으로 뒷받침한다.

특히 Ensemble 기반 모델 중 RandomForest를 활용한 경우, 다수 EMA 모델 출력 간의 비선형 관계와 변수 간 상호작용을 효과적으로 학습함으로써, 개별 EMA 모델이 독립적으로는 반영하지 못한 보완적인 정보를 활용할 수 있음을 확인하였다. 이러한 특성은 단순한 점수 예측 정확도 향상보다는, 상위 소수 후보를 선별해야 하는 실제 도킹 구조 후보군 선별 과정에서 발생할 수 있는 선별 실패 가능성을 완화하는 데 기여한 것으로 해석된다.

결과적으로 본 연구는 단백질 도킹 구조 후보군 선별 과정에서 EMA 모델 출력 점수를 결합한 Ensemble 기반 모델이 단일 EMA 모델 대비 효과적인 대안이 될 수 있음을 시사한다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., and Baker, D., “Improved protein structure refinement guided by deep learning-based accuracy estimation,” *Nature Communications*, vol. 12, Article no. 1340, 2021.

[2] Chen, X., Morehead, A., Liu, J., and Cheng, J., “DProQ: A Gated-Graph Transformer for Protein Complex Structure Assessment,” *arXiv preprint*, arXiv:2205.10627, 2022.

[3] Guo, L., He, J., Lin, P., Huang, S.-Y., and Wang, J., “TRScore: a 3D RepVGG-based scoring method for ranking protein docking models,” *Bioinformatics*, vol. 38, no. 9, pp. 2444–2451, 2022.

[4] Chen, Z., Liu, N., Huang, Y., Min, X., Zeng, X., and Ge, S., “PointDE: Protein Docking Evaluation Using 3D Point Cloud Neural Network,” *IEEE Access*, 2021.