# LATTE: 단백질 서열 임베딩 및 검색을 위한 구조-인지 잠재 모델

분야 : 생물융합
안성민(대구과학고등학교 2학년)
문시현(대구과학고등학교 2학년)
정주영(대구과학고등학교 2학년)
이민재(대구과학고등학교 2학년)

지도교사: 박정수

## 요 약

우리는 UniRef50에 대해 학습한 구조 인지 인코더 LATTE를 제안한다. 이 모델은 서열 재구성 항, Kullback-Leibler(KL) 정규화 항, 그리고 잔기 간 유사성을 감독하는 구조 항을 결합한 복합 손실 함수를 사용한다.

LATTE는 학습하지 않은 UniRef50 서열을 97.17%의 정확도로 재구성하며, 잠재 공간에 노이즈를 주입해도 강인한 성능을 유지한다. 256차원 잠재 공간 위에 학습된 가우시안 프로세스(GP) 모델은 형광 단백질(FP)과 비-FP를 구분하는 이진 분류에서 5-fold 교차 검증 정확도 0.987을 달성하고, 여기/방출 파장을 예측할 때 각각 2.70 nm, 3.80 nm의 RMSE를 보였다. 이는 ESM-2 (650M) 임베딩을 사용한 기준선과 동등한 수준이며, 훨씬 적은 파라미터로 달성한 결과이다.

동일한 데이터셋에서 LATTE의 쌍별 코사인 거리(pairwise cosine distance)는 ESM-2보다 더 넓고 꼬리가 두꺼운(heavy-tailed) 분포를 보이지만, 두 공간 사이의 순위 상관(Spearman $\rho = 0.761$)은 높고, k=3 클러스터링 분할 역시 유사하게 나타난다. 이는 LATTE가 이웃 순서를 보존하면서도 거리의 동적 범위를 확장한다는 것을 의미한다.

이 표현을 바탕으로, 우리는 BLAST 이전 단계의 프리필터로 사용되는 잠재 트리 인덱스 LLAST(LATTE Latent Alignment Search Tool)를 구축하였다. 약 $10^6$개의 단백질 서열로 구성된 데이터베이스에서, LLAST는 쿼리당 BLAST에 전달되는 후보 서열 수를 최대 $10^4$개(전체의 약 0.97%)로 제한하면서도, 기준선 BLAST의 상위 50개 히트 중 약 55%를 회수한다. 이는 재현율과 연산량 감소 사이의 공격적이지만 조절할 수 있는 트레이드오프를 보여준다.



연구 핵심 키워드: 단백질 언어 모델; 구조 손실; 형광단백질; 임베딩 기하; 코사인 거리; 생성 모델링; 생물정보학

# Ⅰ. 연구의 동기 및 목적

트랜스포머 아키텍처가 서열 내 숨겨진 정보를 포착할 수 있는 능력이 있음이 밝혀졌다. 또한, 거대 트랜스포머 모델을 훈련하는 대신, 사전 훈련된 모델을 임베딩으로 사용하는 것이 효과적임이 입증되었다. 그러나 서열만으로 훈련하는 것은 일반화, 기능, 다재다능함에 명확한 한계가 있다. 이러한 한계를 극복하기 위해 구조적 정보의 반영이 필요하다.

또한, NCBI 서열 검색 시스템인 BLAST는 매우 유익하지만, 쿼리당 실행 시간은 데이터베이스 크기에 따라 거의 선형적으로 증가한다. 실제로 BLAST는 각 쿼리에 대해 라이브러리의 상당 부분을 스캔하며, 거의 중복되거나 매우 유사한 서열을 반복적으로 재방문한다. AI가 발전하며 연구가 자동화됨에 따라서, 서열 데이터베이스는 기하급수적으로 증가하고 있기 때문에, 서열 검색 시스템의 변화의 필요성이 대두된다.

# Ⅱ. 이론적 배경

## 1. 단백질 (Protein)

단백질은 펩타이드 결합으로 연결된 아미노산의 선형 중합체이며, 그 3차원 구조가 생화학적 기능을 결정한다. 1차 구조, 즉 20가지 표준 아미노산의 특정 서열은 수소 결합, 소수성 패킹, 정전기적 인력, 반데르발스 힘과 같은 분자 내 상호작용을 통해 접힘 경로를 지시한다. 이러한 상호작용은 2차 모티프($\alpha$ 나선, $\beta$ 시트), 3차 접힘, 그리고 다중 복합체의 4차 조립을 발생시킨다.

## 2. 딥러닝 (Deep Learning)

딥러닝은 대규모로 전이 가능한 서열 표현을 학습함으로써 단백질 모델링을 가속화했다. 여기서는 아키텍처 세부 사항은 방법(Methods) 섹션으로 미루고 설계에서 잠재 공간의 동기와 역할에 초점을 맞춘다. 딥러닝은 대규모 비라벨 코퍼스로부터 구조적 및 기능적 제약을 암묵적으로 인코딩하는 고차원 서열 표현을 학습함으로써 단백질 모델링을 변화시켰다. 마스크 언어 모델 PLM(예: ESM-2)은 장거리 잔기 결합을 포착하고, ESMFold를 통해 정확한 단일 서열 구조 예측을 가능하게 하며, ESM-3는 편집/설계를 위해 서열, 구조, 기능을 더욱 결합한다. 그러나 이러한 파운데이션 모델들은 대규모 스크리닝 및 반복 설계를 방해하는 지연 시간과 메모리 비용을 수반한다. 이를 보완하기 위해, 우리는 잠재 변수가 활성(active) 상태를 유지하며 정보량이 풍부한 콤팩트 VAE 스타일 인코더-디코더를 채택하고, 기하학적 구조가 실제 구조와 정렬되도록 ESMS에 대한 지각 손실(perceptual losses)로 정규화한다. 결과적으로 생성된 잠재 변수들은 빠른 클러스터링/검색 및 다운스트림 속성 모델을 지원하면서 적은 연산량으로 경쟁력 있는 성능을 유지하며, 해석 가능성을 희생하지 않고 검색을 줄이는 정렬 전 가지치기(LLAST)의 기초 역할을 한다.

## 3. ESM-2와 ESMS

사전 훈련된 ESM-2 임베딩을 사용하여, 우리는 높은 임베딩 충실도를 유지하면서 ESM-2의 서열당 0.5초의 추론 시간을 줄이는 ESMS를 개발했다. Rives 등이 소개한 ESM-2는 단거리 및 장거리 잔기 상호작용을 포착하는 마스크 언어 모델이며, AlphaFold 2에 필적하는 성능을 가진 3차원 구조 예측기인 ESMFold의

기반이 되고, 2차 구조 및 열안정성 예측과 같은 다운스트림 작업에 널리 사용된다. 다중 파라미터 규모(650M 파라미터가 가장 일반적임)로 제공되지만, ESM-2의 지연 시간은 더 가벼운 대안을 모색하게 한다. 보류된 테스트 세트에서 ESMS는 코사인 유사도 0.9647과 RMSE 1.2998을 달성했다.

## 4. BLAST

BLAST(Basic Local Alignment Search Tool)는 시드 및 확장(seed and extend) 휴리스틱을 통해 국소적 유사성을 찾아내어 높은 점수의 세그먼트 쌍(HSPs)을 빠르게 형성한다. 정렬의 통계적 유의성은 Karlin-Altschul 프레임워크에 의해 모델링되어 E-값(E-value)과 비트 점수(bit score)를 산출한다. 최적 동적 프로그래밍 정렬보다는 민감도가 낮지만, BLAST는 대규모 데이터베이스에 효과적으로 확장되며 실질적인 표준으로 남아 있다. BLAST+ 재아키텍처는 성능과 모듈성을 더욱 간소화했다.

BLAST의 시간 복잡도는, BLAST의 데이터셋이 매우 크다는 것을 고려하면, $O(N)$과 유사하게 나온다.

## 5. 단백질 검색을 위한 딥러닝 기반 탐색 알고리즘

트랜스포머 기반 단백질 언어 모델은 입력 서열 x를 임베딩 $z = f_\theta(x) \in \mathbb{R}^d$로 매핑한다. 데이터베이스 $D = \{z_i\}_{i=1}^N$와 쿼리 $z_q$가 주어졌을 때, 단백질 검색은 코사인 유사도 또는 최대 내적 탐색(MIPS) 설정에서의 내적과 같은 유사성 점수 $s(z_q, z_i)$하에서 최근접 이웃 검색으로 공식화된다. 단순한 top-k 검색은 모든 $N$개 항목에 대해 $s(z_q, z_i)$를 계산해야 하므로 쿼리당 시간 복잡도가 $O(Nd)$가 되어, $N$이 커지면 계산 비용이 급격히 증가한다.

학습 기반 검색 알고리즘은 임베딩 공간의 기하학을 활용하는 인덱스를 $D$ 위에 구축함으로써 이 비용을 완화한다. 그래프 기반 방법은 각 데이터베이스 포인트를 소수의 이웃 세트에 연결하고, 더 높은 유사도 노드를 향해 이동하며 제한된 프론티어만 탐색하는 탐욕적 검색(greedy search)을 수행하므로, 유사성 평가 횟수는 $N$에 직접적으로 비례하기보다 그래프 차수와 검색 깊이에 의해 지배된다. 클러스터 기반 방법은 대신 거친 양자화기(coarse quantizer)를 통해 공간을 $K$개의 셀로 분할하고 $z_q$에 가장 가까운 수의 셀만 탐색(probe)하여, 스캔되는 후보를 N에서 $(\frac{N}{K}) * n_{probe}$정도로 줄인다. 변형들은 거리 계산을 더욱 가속화하기 위해 곱 양자화(product quantization) 또는 해싱을 결합한다.

대규모 검색 시스템에서 이러한 인덱스는 일반적으로 계단식(cascade)의 첫 단계로 사용된다. 인덱스는 $M \ll N$개의 후보 집합 $S \subset D$를 반환하고, 더 정확하지만 느린 채점기(scorer)가 $S$에만 적용된다. 단백질 검색의 경우, 이는 임베딩 기반 인덱스를 사용하여 서열 공간의 명백히 관련 없는 영역을 폐기한 다음 축소된 후보 세트에서 BLAST와 같은 정렬 도구를 실행하는 것에 해당한다. 그래프 차수, 프로브 예산(probe budget), 검색 깊이와 같은 하이퍼파라미터는 top-k에서의 재현율과 평균 및 꼬리 지연 시간(tail latency)의 균형을 맞추기 위해 조정된다.

## 6. 시간복잡도

시간 복잡도는 데이터의 양이 증가함에 따라 알고리즘의 처리 시간이 어떻게 변화하는지를 나타내는 척도이다. $O(n)$과 같이 Big-O 표기법으로 나타낸다. 예를 들어, $O(n^2)$은 데이터의 제곱과 시간이 비례함을 뜻하고, $O(\log n)$은 데이터 개수의 로그값에 시간이 비례한다는 것을 나타낸다. 이 경우, $O(\log n)$이 $O(n^2)$에 비해 대용량 데이터에 대해서는 더 효율적이다.

# Ⅲ. 연구 방법

## 1. 잠재공간 형성 (Shaping the latent space)

우리는 인코더 메모리와 잠재 벡터를 받는 훈련 신호 생성기 역할을 하는 의사 디코더(pseudo decoder)로 LATTE를 훈련했다. 또한, 전체 구조 예측기보다 훨씬 낮은 비용으로 잔기 위치 규칙성을 포착하는 사전 훈련된 ESMS 임베딩에서 계산된 지각 손실(perceptual loss)을 통해 구조적 일관성을 강제한다. 원본 서열 $x_{orig}$과 재구성 서열 $x_{recon}$에 대해 다음 두 항을 정의한다.

$$L_{COS} = [1 - \cos(\mathrm{ESMS}(\mathrm{x}_{orig}), ESMS(x_{recon}))],$$

$$L_{MSE} = |\mathrm{ESMS}(\mathrm{x}_{orig}) - ESMS(x_{recon})|_2^2$$

이 구조적 항들은 다음 토큰 교차엔트로피 $L_{CE}$(교사 강요, teacher forcing) 및 KL 발산 $L_{KL}$과 결합하여 다음과 같이 정의된다.

$$L_1 = \lambda(\mathrm{L_{COS}} + \mathrm{L_{MSE}}) + \alpha \mathrm{L_{CE}} + \beta \mathrm{L_{KL}}$$

여기서 $\lambda = 5$ 이고, $\alpha$는 처음 100 Epoch에 걸쳐 30에서 0.1로 선형적으로 감소하고, $\beta$는 동일한 간격 동안 0에서 0.1로 선형적으로 증가한다. 코사인 항은 유사하게 가능성 있는 잔기들 간의 치환을 허용하는 반면, MSE 항은 큰 편차를 처벌한다. 이들은 함께 정보가 풍부한 잠재 변수에 의존하여 정확한 재구성을 하도록 함으로써 사후 붕괴(posterior collapse)를 방지한다.

LATTE는 5.5M 파라미터의 경량 트랜스포머로, 4-층 트랜스포머 인코더로 구성된다(표1, 그림 1). 최적화에는 Adam 옵티마이저가 사용되었다.

**표 1.** LATTE의 하이퍼파라미터.

| | Vocab Size | d_model | Latent Dim | n_heads | Feed Forward | Dropout |
|---|---|---|---|---|---|---|
| LATTE Hyperparameter | 33 | 256 | 256 | 4 | 512 | 0.3 |

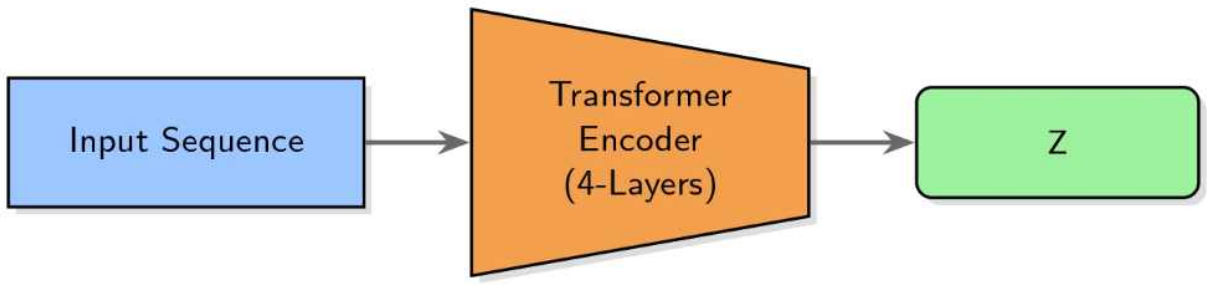
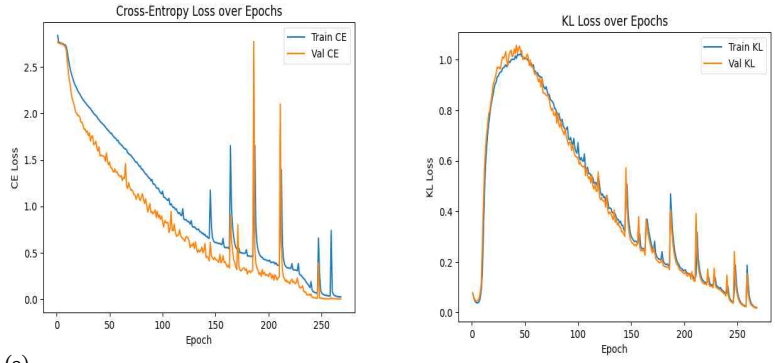



**그림 1.** LATTE 구조: 입력 서열 → 4-층 인코더 → 잠재 분포(평균/로그분산)

## 2. 모델 선택 (Model Selection)

모델은 Kaggle에서 제공하는 두 개의 T4 GPU 세션을 사용하여 UniRef50 데이터셋의 무작위 하위 샘플로 훈련되었다. 모니터링된 학습 곡선은 과적합의 징후를 보이지 않았다.(그림2, 그림3) Epoch 500 모델(LATTE-500)은 검증 손실 Val CE = 0.000, COS = 0.003, MSE = 0.007, KL = 0.002로 99.976%의 재구성

률에 도달했지만, 매우 낮은 KL 값은 잠재적인 사후 붕괴를 시사했다.(그림 4) 잠재 공간에 노이즈를 추가해도 재구성 성능에 큰 영향을 미치지 않는다는 것이다. 반면, Epoch 380이 선택된 이유는 KL 발산(0.048)이 권장 활성 값인 0.05에 더 가깝고 가장 낮은 Val CE를 보였기 때문이다.(그림 5) Epoch 380에서 검증 손실은 다음과 같다. Val CE = 0.072, COS = 0.010, MSE = 0.020, KL = 0.048




(a) (b)

**그림 2.** 구조 손실이 있을 때, Epoch에 따른 CE와 KL의 변화, (a): CE, (b): KL

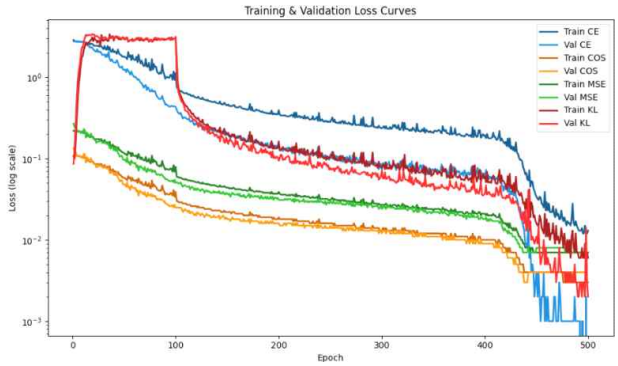
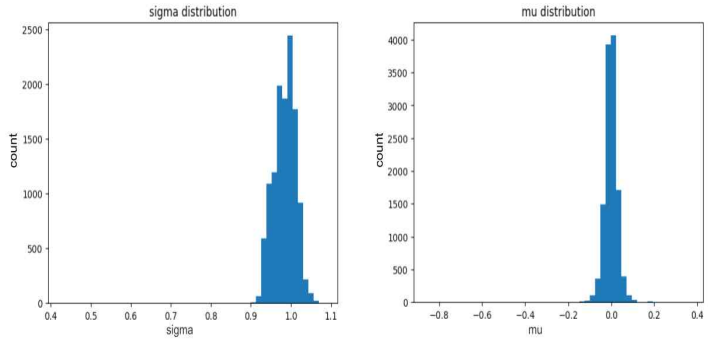
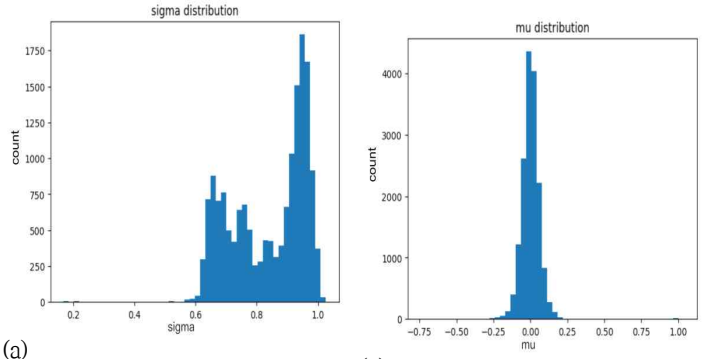



**그림 3.** 학습/검증 손실함수(Epoch 500 까지의 손실함수)




(a) (b)

**그림 4.** Epoch 500에서의 잠재분포 (a): sigma, (b): mu




(a) (b)

**그림 5.** Epoch 380에서의 잠재 분포 (a): sigma, (b): mu

## 3. LLAST

고정된 임베딩에서 작동하는 지역성 민감 해싱(locality-sensitive hashing)이나 곱 양자화(product quantization)와 같은 일반적인 근사 최근접 이웃(ANN) 방법과 달리, LLAST는 구조 정보가 포함된 LATTE 잠재 공간 위에 작업별 계층적 인덱스를 구축한다. 이를 통해 생성적 설계, 속성 예측, 서열 검색에 동일한 잠재 표현을 재사용할 수 있다.

우리는 LATTE Latent Alignment Search Tool(LLAST)을 도입하였다. 우리는 코사인 거리(L2 행 정규화 후)를 사용하여 잠재 공간의 단백질 서열을 k-평균(k-means) 클러스터링한 다음, 병합 계층적 클러스터링을 적용하여 트리 인덱스를 얻었다. (그림 6) 이 트리는 프리필터 역할을 하여 다운스트림 정렬 전에 잠재 벡터가 다른 서열을 가지치기한다. 클러스터 수 K는 평균 코사인 k-평균 비용 곡선에서 Kneedle 무릎 탐지(knee detection)를 통해 선택되었으며, 엘보(elbow) 주변에서 적응형 10방향 간격 개선이 수행되었다.(그림 7)

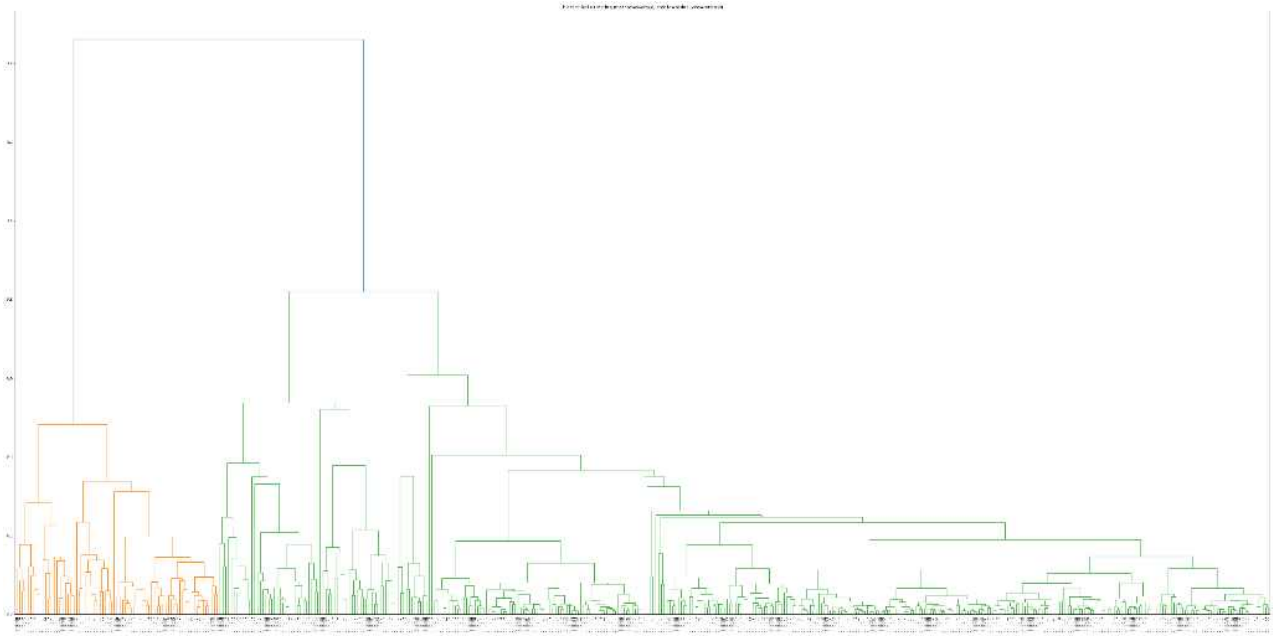


**그림 6.** 잠재 트리 인덱스 구조

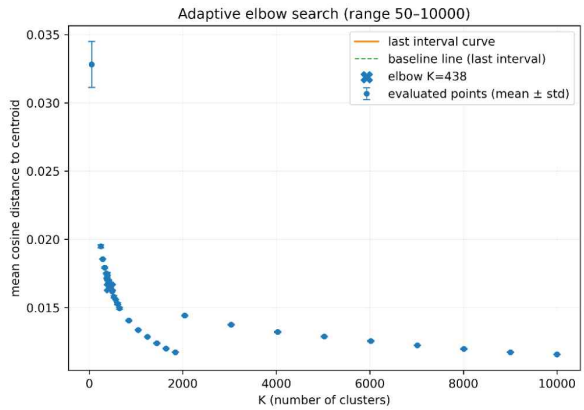



**그림 7.** 적응형 엘보 탐색(K=2000부터 근사해서 계산되었다)

# Ⅳ. 연구 결과

## 1. 사례 연구: 형광단백질(FP) 분석 및 생성

우리는 사전 훈련된 ESM-2 임베딩을 사용했다. FPbase에서 수작업으로 큐레이팅한 FP 서열을 사용하여, (1) FP 대 비-FP를 분류하고 (2) 스펙트럼 피크를 회귀 분석하기 위해 가우시안 프로세스(GP) 모델을 훈련했다. GP 분류기는 5-fold 교차 검증에서 0.987의 정확도를 달성했고, GP 회귀기는 최대 흡수 파장 $\lambda_{abs}$ 2.70 nm / 최대 방출 파장 $\lambda_{em}$은 3.80 nm 평균 제곱근 오차(RMSE)를 달성했다. 동일 파이프라인에서 ESM-2(650M)는 5-fold AUC 0.997, 2.70/3.80 nm RMSE를 보였다. LATTE 256-차원 잠재 임베딩은 정확도 0.987로, ≈100배 작은 파라미터 수와 낮은 추론비용으로 경쟁력 있는 다운스트림 신호를 제공한다.(표 2)

**표 2.** 서로 다른 임베딩에서의 GFP 스펙트럼 예측 성능

| Embedding | Classifier metric (5-fold) | $\lambda_{abs}$ RMSE (nm) | $\lambda_{em}$ RMSE (nm) |
|---|---|---|---|
| ESM-2(650M) | AUC 0.997 | 2.70 | 3.80 |
| LATTE latent (256-d, ~ 5.5M) | AUC 0.987 | 2.70 | 3.80 |

훈련/테스트 분류 리포트는 과적합 징후 없이 강한 성능을 확인한다(표 3, 4). t-SNE 시각화는 FP/비-FP의 명확한 분리를 보이고, 방출/흡수 파장에 따른 연속 그라디언트 가 잠재 다양체에 부드럽게 분포되어 있음을 보여 구조 정보가 효과적으로 인코딩 되었음을 시사한다.

**표 3.** 훈련 세트에 대한 분류 리포트

| Embedding | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non FP(0) | 0.9920 | 0.9940 | 0.9930 | 501 |
| FP (1) | 0.9880 | 0.9840 | 0.9860 | 250 |
| Accuracy | | | 0.9907 | 751 |
| Macro Avg | 0.9900 | 0.9890 | 0.9895 | 751 |
| Weighted Avg | 0.9907 | 0.9907 | 0.9907 | 751 |

**표 4.** 테스트 세트에 대한 분류 리포트

| Embedding | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non FP(0) | 0.9840 | 0.9840 | 0.9840 | 125 |
| FP (1) | 0.9683 | 0.9683 | 0.9683 | 63 |
| Accuracy | | | 0.9787 | 188 |
| Macro Avg | 0.9761 | 0.9761 | 0.9761 | 188 |
| Weighted Avg | 0.9787 | 0.9787 | 0.9787 | 188 |

t-SNE를 통한 잠재 공간 시각화는 형광 단백질과 비형광 단백질을 명확하게 분리하고, 스펙트럼 특성을 연속적인 그래디언트로 인코딩하여 구조적 정보가 효과적으로 포착되었음을 확인시켜 준다 (그림 8). 비형광 단백질은 두 곡선의 안쪽에 군집하는 반면, 형광 단백질은 바깥 영역을 차지한다. 또한, 방출 및 흡광 파장에 해당하는 색상 그래디언트는 유사한 스펙트럼 피크를 가진 단백질들이 함께 그룹화됨을 나타낸다.

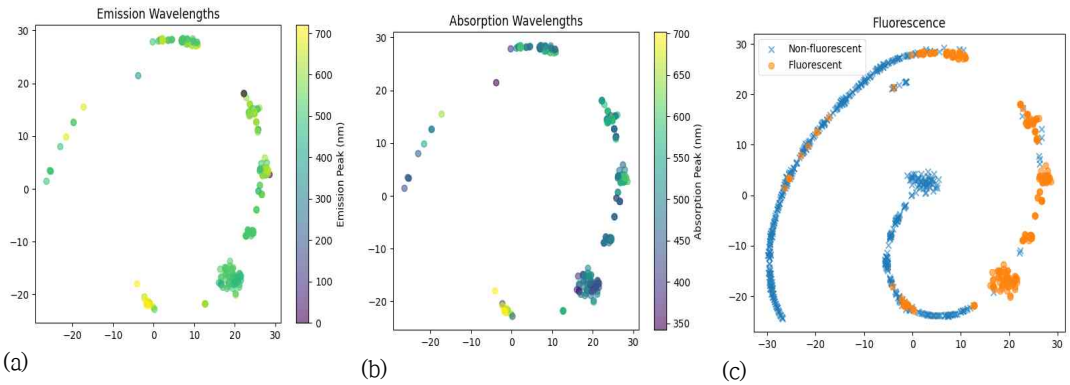



**그림 8.** 형광 단백질의 특성 (a): 방출 스펙트럼 (b): 흡수 스펙트럼 (c): 형광 세기

**표 5.** k-means로 식별된 FP 군집 통계

| 임베딩 | 샘플 수 | 평균 길이 | 길이 표준편차 |
|---|---|---|---|
| 0 | 22 | 316 | 5.28 |
| 1 | 318 | 234 | 10.00 |
| 2 | 14 | 118 | 29.71 |

k-평균 클러스터링으로 식별된 FP 클러스터 통계에서, 잠재 공간의 가까운 벡터들은 AlphaFold 예측으로 확인된 바와 같이 유사한 3차원 구조를 가진 단백질을 인코딩한다. 클러스터 1 단백질은 pLDDT 점수가 90 이상인 고전적인 $\beta$ 배럴 접힘을 보였고, 클러스터 0 단백질은 $\alpha$ 나선으로 감싸진 세 개의 내부 $\beta$ 가닥으로 구성된 두 그룹을 형성했으며, 클러스터 2 단백질은 $\beta$ 배럴을 채택하지 않았음에도 불구하고 $\alpha$ 나선으로 둘러싸인 세 개의 $\beta$ 가닥이라는 공통된 접힘을 공유했다.(그림 9)

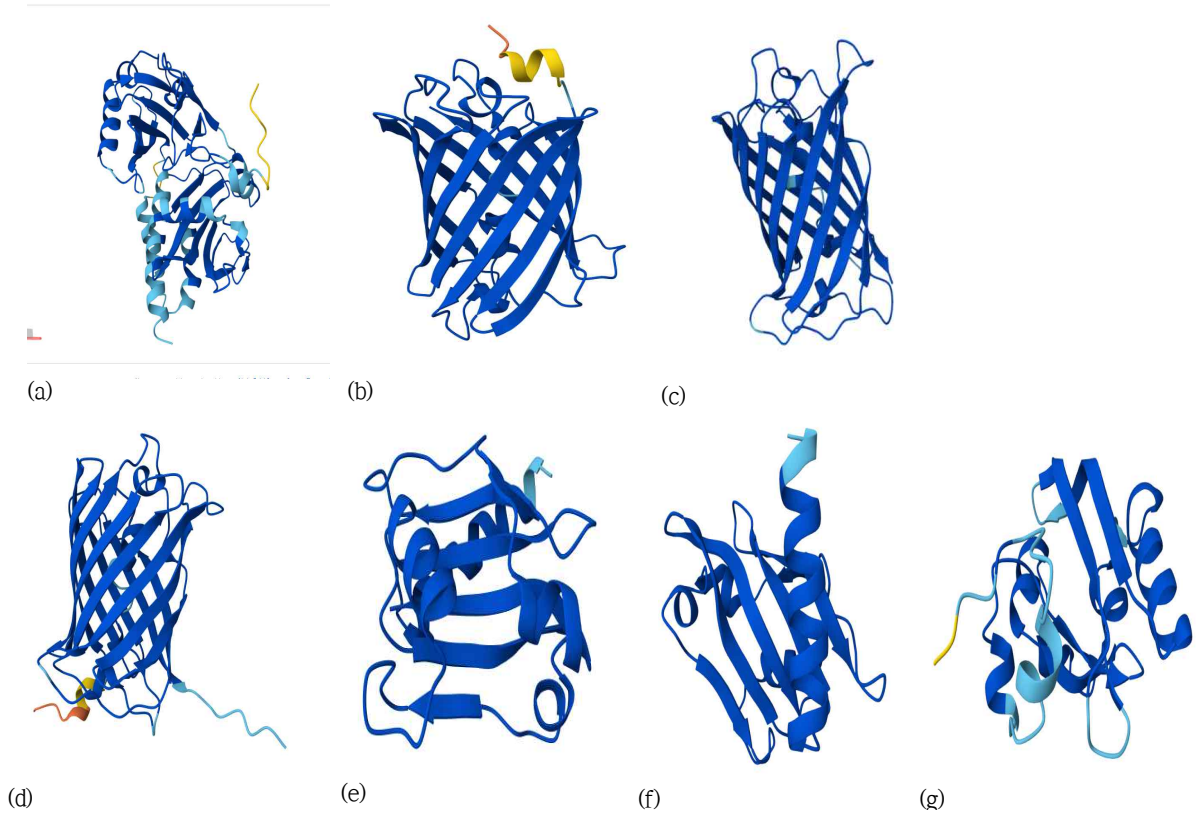



**그림 9.** 형광 단백질의 구조(Alphafold 예측) (a): Cluster 0, (b)~(d): Cluster 1, {e}~{g}: Cluster 2

## 2. LATTE vs ESM-2 임베딩 기하(코사인 거리)

임베딩 공간의 전역 구조를 비교하기 위해, 우리는 일치하는 하위 집합에 대해 pairwise cosine distance (1 − 코사인 유사도)를 계산하고 분포와 경험적 누적 분포 함수(ECDFs)를 모두 요약했다. LATTE 잠재 거리는 더 넓고 꼬리가 더 두껍다(평균 = 0.1694, 표준편차 = 0.3428, p50 = 0.0141, p90 = 0.9006; n = 49,847), 반면 ESM-2 거리는 더 좁다(평균 = 0.0381, 표준편차 = 0.0822, p50 = 0.00953, p90 =

0.1133; n = 49,847) (그림 10, 표 6). 이는 ESM-2가 점들을 더 조밀하게 군집시키는 반면, LATTE는 원거리 이웃에 대한 재현율(recall)을 높일 수 있는 더 확산된 기하학을 산출함을 시사한다. 실제로는 해석 가능성과 정밀도를 회복하기 위해 LATTE의 구조 인식 사전 필터와 서열 정렬("Deep BLAST")을 결합한다. 더 밝은 수평/수직 밴드는 다른 많은 서열들과 거리가 더 먼 서열 또는 클러스터를 표시하며, 전역 구조와 잠재적 이상치를 강조한다. (그림 11)

pairwise cosine distance를 직접 비교하기 위해, 우리는 일치하는 쌍에 대한 ESM(y) 대 LATTE(x) 코사인 거리를 플로팅했다(그림 12). ESM에서 거리가 체계적으로 더 작지만(기울기 b=0.125; 절편 a=0.017), 순위 상관관계는 높게 유지되어(Spearman $\rho$=0.761), LATTE가 동적 범위를 확장하면서 ESM의 이웃 순서를 보존한다는 것을 뒷받침하며, 이는 우리가 Deep BLAST의 잠재적 사전 필터링에 활용하는 속성이다.

우리는 LATTE와 ESM-2 임베딩을 비교하기 위해 k=3 clustering을 재실행한 결과를 보고한다. 실루엣 점수(코사인)는 LATTE를 선호하며, 파티션 간 교차 일치 지표는 두 표현 간의 강한 일관성을 나타낸다.

**표 6.** LATTE와 ESM-2 임베딩의 쌍별 코사인 거리 요약.

| 군집 | n | 평균 | 표준편차 | p10 / p50 / p90 |
|---|---|---|---|---|
| LATTE (잠재) | 49,847 | 0.1694 | 0.3428 | 0.00257 / 0.01406 / 0.90065 |
| ESM-2 (650M) | 49,847 | 0.0381 | 0.0822 | 0.00365 / 0.00953 / 0.11329 |

**표 7.** 코사인 실루엣 점수 (k=3).

| | LATTE | ESM-2 |
|---|---|---|
| Silhouette (cosine) | 0.9431 | 0.9022 |

**표 8.** k=3 클러스터링 일치도 요약.

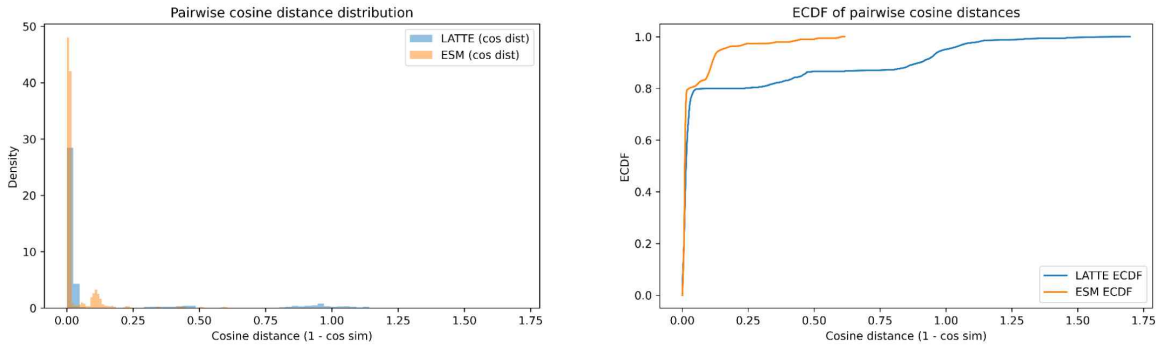
| Cross partition agreement (LATTE vs. ESM-2) | |
|---|---|
| Adjusted Rand Index (ARI) | 0.7373 |
| Adjusted Mutual Information (AMI) | 0.6055 |
| Fowlkes-Mallows Index (FMI) | 0.9596 |
| Variation of Information (VI) | 0.3529 |
| Purity (LATTE → ESM-2) | 0.9633 |

**표 9.** 혼동 행렬

| 군집 | ESM-2-0 | ESM-2-1 | ESM-2-2 |
|---|---|---|---|
| LATTE-0 | 0 | 0 | 1 |
| LATTE-1 | 0 | 23 | 1 |
| LATTE-2 | 6 | 5 | 318 |

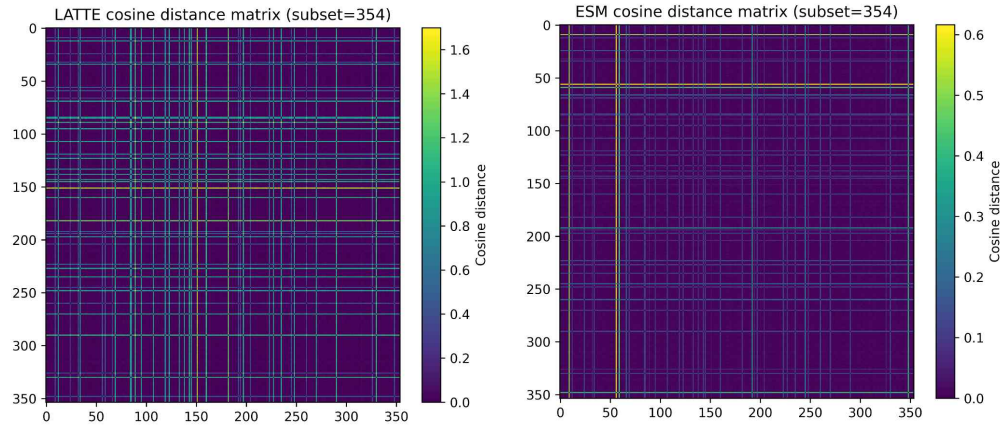
표 10. 클러스터별 크기와 평균값.

| Cluster | size | LATTE $\mu$ | ESM-2 $\mu$ |
|---|---|---|---|
| 0 | 6 | 0.000 | 0.728 |
| 1 | 28 | 0.904 | 0.737 |
| 2 | 320 | 0.949 | 0.920 |




(a) (b)

**그림 10** LATTE latent 임베딩과 ESM-2 임베딩 간의 pairwise cosine distance (a): 거리 분포의 밀도 히스토그램 (b): ECDF




(a) (b)

**그림 11.** 두 임베딩 공간의 354개 서열 하위 집합에 대한 pairwise cosine distance (a): LATTE (b): ESM

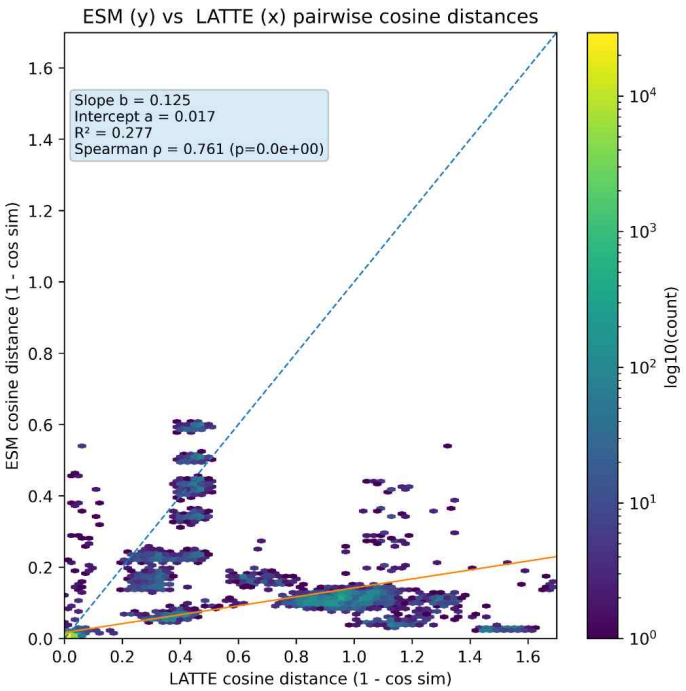



**그림 12.** 일치하는 쌍에 대한 ESM (y) 대 LATTE (x) pairwise cosine distance. 실선은 OLS 적합선이다 (b=0.125, a=0.017; $R^2$ =0.277). Spearman $\rho$=0.761은 ESM의 스케일 압축에도 불구하고 강한 순위 일치도를 나타낸다.

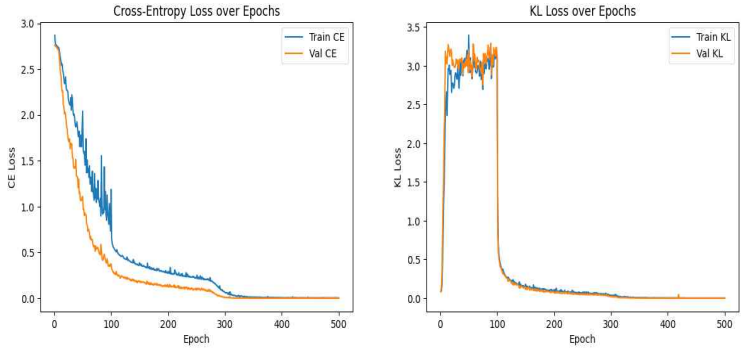
## 3. 어블레이션 연구

손실 함수를 다음과 같이 설정하여 어블레이션 연구를 수행했다.

$$L = \alpha L_{CE} + \beta L_{KL}$$

가중치는 Epoch< 30 동안 $\alpha$ = 30, $\beta$ = 0으로 설정되고, 그 이후에는 $\alpha$ = 0.1, $\beta$ = 0.1로 설정되었다. 이 설정에서 KL 소실(vanishing)이 Epoch 100 직후에 발생하여 구조적 손실 항의 중요한 역할을 입증했다.

그림 13은 구조적 손실이 사용되지 않았을 때의 CE 및 KL 곡선을 보여주고, 그림 2는 구조적 손실이 통합되었을 때의 해당 곡선을 보여준다. 후자만이 KL 붕괴를 방지하고 더 안정적인 수렴을 산출한다.




(a) (b)

**그림 13.** Structural loss가 없을 때, Epoch에 따른 CE와 KL의 변화, (a): CE, (b): KL

## 4. LLAST Results(LLAST 결과)

기존 BLAST 앞에 배치된 잠재 트리 프리필터의 기능을 평가하기 위해, 우리는 커버리지(coverage)와 효율성(efficiency)을 모두 정량화했다. 커버리지는 프리필터링 후 회수된 기준선 BLAST 상위 50개 고유 히트의 비율(SeqRecall@50)로 측정되었다. 효율성은 쿼리당 프리필터에서 BLAST로 실제로 전달된 서열의 수, 즉 유효 후보 세트 크기로 평가되었다. 기본 데이터베이스에는 약 N ≈ $10^6$개의 서열이 포함되었다. 잠재 트리 인덱스는 클러스터당 약 S ≈ 2000개 서열이 포함되도록 구성되었으며, 쿼리 시에는 트리에서 점수 상위 K개의 클러스터만 탐색한다. 우리는 BLAST로 전달될 후보 수의 상한 $N_{\max}$를 2000, 4000, 6000, 8000, 10000으로 변화시키며 다양한 운영점을 평가하였다.

991개 쿼리에서 N_max = 10000(전체의 약 0.97%)이라는 가장 보수적 설정은 평균 SeqRecall@50 ≈ 0.55를 달성하면서 BLAST 검색 공간을 약 100배 줄였다. N_max = 4000(약 0.38%)과 같은 더 과감 설정에서도 평균 SeqRecall@50 ≈ 0.53으로 커버리지 감소는 3-4% 수준에 그치며, 후보 집합은 약 260배 축소되었다.(표 11 그림 14 그림 15) N_max = 2000(약 0.18%)은 후보를 500배 이상 줄였으나 일부 어려운 쿼리에서 0-recall이 발생했다. 이러한 하드 케이스는 BLAST 기준선에서도 top-50에서 고유 서열 수가 극히 적은, 즉 데이터베이스 내 연관성이 거의 없는 '희박한(orphan-like)' 단백질들에 해당하였다. N_max ≥ 4000에서는 0-recall 사례가 대부분 사라져 실제 적용 가능한 균형점을 형성했다.

계산 복잡도 관점에서 프리필터는 BLAST의 데이터베이스 크기 의존성을 강하게 완화한다. 데이터베이스 총 서열 수를 N, 리프 용량을 S, 쿼리당 방문되는 리프 수를 K라 하면, 잎의 개수는 대략 C ≈ N/S, 트리 깊이는 log(C)에 비례한다. 프리필터는 단일 쿼리 인코딩 T_encode 이후, O(log(N/S))의 비용으로 트리를 탐색하고, 선택된 K개의 잎에 대해 최대 KS개의 서열만 BLAST에 전달하므로 전체 시간은

$$T_{prefilter}(N) = O(T_{encode} + \log(\frac{N}{S}) + KS)$$

로 표현된다. S와 K가 고정되어 있을 때 지배항은 KS이며, N에 대한 의존성은 로그 수준으로만 남는다. 반면 프리필터가 없는 BLAST에서는

$$T_{BLAST(N)} = O(N)$$

으로 N에 선형 비례한다. 따라서 LLAST는 데이터베이스가 $10^7$-$10^8$ 수준까지 증가해도 BLAST 단계의 연산량을 사실상 거의 일정하게 유지할 수 있으며, 이 점이 대규모 단백질 탐색에서 LLAST의 핵심적 효율성을 제공한다.

**표 11.** 약 10^6개 서열(991개 쿼리) 데이터베이스에서의 LLAST 운용 지점. Nmax는 쿼리당 BLAST로 전달되는 후보 서열 수의 상한값이다.

| TopK | $N_{max}$ | Mean SeqRecall@50 | Mean candidates/query | Approx. reduction |
|---|---|---|---|---|
| 1 | 2000 | 0.45 | 1.8k | ~ 540배 |
| 2 | 4000 | 0.53 | 3.8k | ~ 260배 |
| 3 | 6000 | 0.54 | 5.6k | ~ 180배 |
| 4 | 8000 | 0.55 | 7.7k | ~~ 130배 |
| 5 | 10000 | 0.55 | 9.7k | ~100배 |

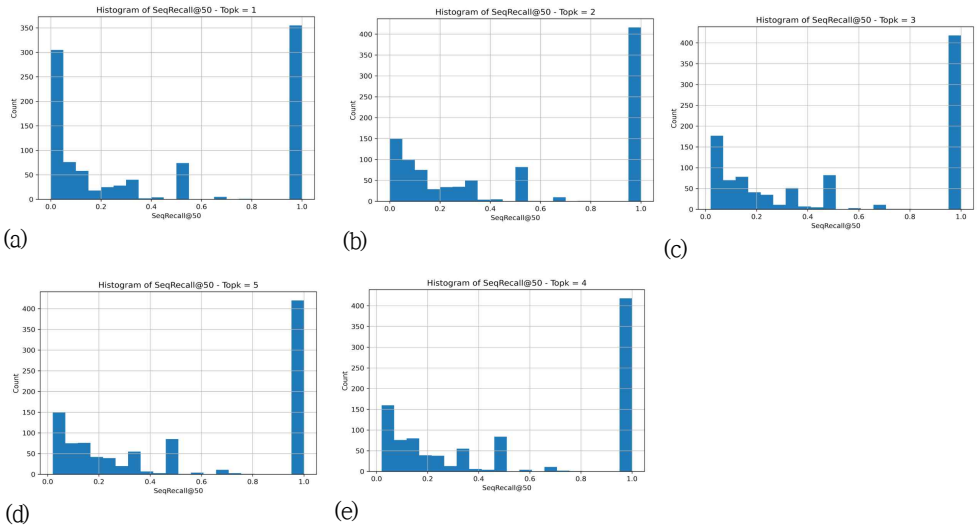



**그림 14.** 서로 다른 TopK 값(1-5)에 대한 프리필터 단계의 시퀀스 재현율(SeqRecall@50) 히스토그램: (a) TopK=1, (b) TopK=2, (c) TopK=3, (d) TopK=4, (e) TopK=5.

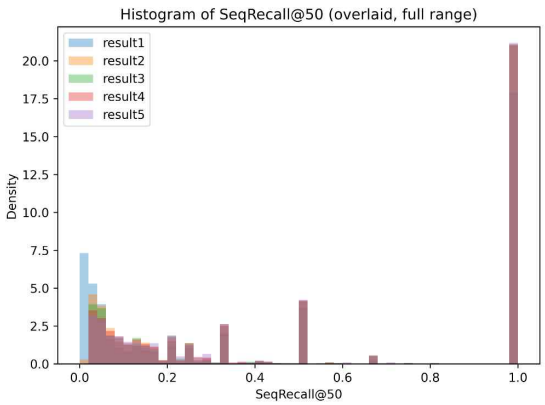



**그림 15** TopK 조건별 시퀀스 재현율 히스토그램을 겹쳐서 그린 비교.

# V. 결론 및 논의

요약하자면, 우리는 콤팩트하고 구조를 인지하는 잠재 모델(LATTE)을 제안하고, 결과적으로 생성된 256차원 공간이 재구성, 속성 예측, 그리고 BLAST를 위한 확장 가능한 잠재-트리 프리필터(LLAST)를 동시에 지원함을 보인다.

## 1. LATTE

수백만 개의 서열 규모에서 AlphaFold 기반 손실이나 접촉 지도(contact maps)로 구조적 일관성을 직접 강제하는 것은 비실용적이다. 종단간(end-to-end) 구조 예측은 서열당 몇 분이 소요되고, 접촉 지도는 상당한 메모리와 큐레이팅된 구조를 요구하며, 두 접근 방식 모두 주석이 없는 영역에 대해 정렬 및 커버리지 문제를 야기한다. 본 연구에서 우리는 대신 더 가벼운 단백질 언어 모델(PLM)을 구조적 교사로 활용하여, ESM-2에서 파생된 ESMS 임베딩을 사용하여 LATTE의 잠재 공간을 형성한다.

경험적 결과는 이 전략이 콤팩트한 잠재 공간에 유용한 기하학을 부과하기에 충분함을 시사한다. ESM-2와 비교할 때, LATTE는 높은 Spearman 순위 상관관계와 두 공간 간의 k=3 분할의 강력한 일치로 반영되듯이 이웃 순서와 클러스터링 구조를 보존하면서 더 넓고 꼬리가 두꺼운 코사인 거리 분포를 산출한다. 구조적 손실을 제거하면 빠른 KL 붕괴와 퇴화된 잠재 변수로 이어졌던 어블레이션 연구와 함께, 이는 ESMS 기반 지각 손실이 사후 붕괴를 적극적으로 방지하고 단순히 재구성 통계를 일치시키는 것이 아니라 정보가 풍부한 잠재 변수를 장려한다는 견해를 뒷받침한다.

GFP 사례 연구는 이러한 해석을 강화한다. 0.05 근처의 활성 KL 발산을 갖는 256차원 잠재 벡터는 가우시안 프로세스 모델이 형광 단백질과 비형광 단백질을 분리하고 여기 및 방출 피크를 예측하는 데 충분하며, LATTE가 훨씬 더 적은 파라미터 예산과 낮은 추론 비용을 가짐에도 불구하고 ESM-2 임베딩 기준선과 비슷한 오차를 보인다. 이는 PLM을 통한 구조적 감독이 직접적인 서열 생성에 의존하지 않고도 다운스트림 기능 작업에 경쟁력 있는 저차원, 구조 정렬 표현을 생성할 수 있음을 나타낸다.

앞으로 동일한 프레임워크를 다른 구조적 교사(예: ESM-3 또는 중간 크기 PLM)와 접촉 확률 또는 무질서(disorder) 예측과 같은 추가 감독 신호로 확장할 수 있을 것이다. 이러한 방향은 LATTE 스타일의 인코더가 검색 및 설계에 적합한 잠재 공간을 만드는 일반적인 기하학적 속성을 유지하면서 특정 기능적 패밀리에 맞게 조정될 수 있는지 테스트할 것이다.

## 2. LLAST와 기준선 BLAST 간의 불일치 해석

한 LLAST가 명시적으로 BLAST의 프리필터 역할을 하도록 설계되었지만, LLAST와 기준선 BLAST에 의해 회수된 쿼리당 상위 50개 히트 세트는 동일하지 않다. 우리는 이것을 LLAST가 BLAST를 복제하는 데 실패한 것으로 보지 않고, 두 절차 간의 여러 구조적 및 알고리즘적 차이의 결과로 본다.

첫째, LLAST는 명시적인 예산 제약 하에서 작동한다. 각 쿼리에 대해 잠재 트리 검색은 고정된 수의 잎(Top-K)으로 제한되며 최대 $N_{\max}$개의 후보 서열을 BLAST로 전달한다. 방문하지 않은 노드에 포함된 관련 서열은 설계상 BLAST에서 얼마나 높은 점수를 받을지와 관계없이 후보 풀에서 제외된다. 대조적으로, 기준선 BLAST는 개념적으로 전체 데이터베이스를 검색하며 비교 가능한 클러스터 수준 방문 제약을 받지 않는다.

둘째, LLAST는 구조 인식 잠재 공간에서 후보의 순위를 매기는 반면, BLAST는 국소 정렬 점수로 순위를 매긴다. LATTE는 유사한 전체 접힘과 잔기-잔기 유사성 패턴을 가진 서열을 서로 가깝게 배치하도

록 훈련되며, LLAST는 256차원 잠재 벡터 간의 코사인 거리를 사용하여 이 공간을 검색한다. 대조적으로, BLAST는 국소 고득점 세그먼트 쌍과 치환 행렬에 의해 구동되며, 전체 접힘이나 전역적 맥락이 다르더라도 짧은 모티프, 저복잡성 세그먼트 또는 부분 도메인 일치를 공유하는 서열에 높은 점수를 할당할 수 있다. 결과적으로 LLAST는 구조 감독 훈련에서 학습된 잠재 기하학과 일치하는 후보를 우선시하는 경향이 있는 반면, BLAST는 유사성의 다른 측면을 강조할 수 있다.

셋째, 우리가 사용하는 평가 지표는 의도적으로 보수적이다. 우리의 SeqRecall@50 지표는 기준선 BLAST 상위 50개 세트를 평면적인 참조로 취급하며, 동일한 주제 서열에 대한 거의 중복된 정렬과 점수가 약하거나 커버리지가 낮은 한계 히트를 포함하여 LLAST+BLAST 히트에 나타나지 않는 모든 기준선 히트에 대해 누락(miss)으로 계산한다. 많은 쿼리에서 BLAST 상위 50개에는 50개보다 훨씬 적은 고유 표적(unique targets)이 포함되지만, SeqRecall@50은 중복 히트를 놓치는 것과 진정으로 별개의 상동체를 놓치는 것에 대해 LLAST에 동일하게 불이익을 준다. 이는 SeqRecall@50이 더 허용적인 검색으로 LLAST가 달성할 수 있는 상한선이라기보다는, 공격적인 후보 예산 하에서 의미 있는 표적의 재현율에 대한 하한선으로 해석되어야 함을 의미한다.

넷째, LLAST는 이산화된 트리 인덱스를 통해 근사 최근접 이웃 검색을 수행한다. 잠재 트리 순회는 중심 거리(centroid distances)와 계층적 가지치기 결정에 의존하는데, 이는 양자화 효과를 도입한다. 즉, 클러스터 경계 근처에 있는 서열은 검색된 후보 중 일부보다 잠재 공간에서 약간 더 가깝더라도 방문 지역 바로 바깥에 떨어질 수 있다. 이러한 근사화는 확장 가능한 트리 기반 인덱스에 내재되어 있으며, BLAST로 전달되는 서열 수의 상당한 감소를 위해 재현율의 통제된 손실을 교환한다.

종합하면, 이러한 요인들은 LLAST가 기준선 BLAST 상위 50개 세트를 정확하게 재현하지 못하는 이유를 설명한다. LLAST는 BLAST의 전체 데이터베이스 동작을 복제하는 것이 아니라, 고정된 예산 하에서 콤팩트하고 구조를 인지하는 후보 풀을 제공하는 것을 목표로 한다. 이런 의미에서 관찰된 불일치는 잠재 인덱스의 의도적인 제한(제한된 클러스터 방문 및 근사 검색)과 구조 감독 잠재 모델 및 국소 정렬 통계에 의해 포착된 유사성의 서로 다른 개념을 모두 반영한다.

실제로 이러한 불일치는 우리가 관찰하는 운영 포인트를 고려할 때 수용 가능하다. $N_{\max} = 4000 \sim 10000$인 구성은 대략 0.53–0.55의 SeqRecall@50을 유지하면서 평균 BLAST 작업량을 ~$10^6$개 서열 데이터베이스에서 100–260배 줄여, 달성 가능한 효율성 이득과 비교할 때 기준선 BLAST 히트의 정확한 복제는 열악한 트레이드오프가 되게 한다.

## 3. MMseqs2 및 서열 기반 프리필터와의 관계

LLAST의 자연스러운 비교 대상은 대규모 서열 검색에서 BLAST의 고도로 최적화된 프리필터이자 독립형 대안으로 널리 사용되는 MMseqs2이다. MMseqs2는 갭 정렬(gapped alignment) 전에 검색 공간을 빠르게 가지치기하기 위해 k-mer 기반 필터와 갭 없는 확장의 계단식(cascade)을 사용하며, 그 구현은 대규모 데이터베이스에서 거의 선형에 가까운 확장을 달성하기 위해 SIMD 벡터화 및 병렬 실행으로 신중하게 최적화되었다. LLAST는 MMseqs2를 대체하거나 새로운 최첨단 서열 전용 필터로 의도된 것이 아니다.
개념적으로 LLAST는 두 가지 주요 측면에서 MMseqs2와 다르다. 첫째, LLAST는 LATTE에 의해 학습된 콤팩트하고 구조를 인지하는 잠재 공간에서 작동하며, 트리 순회는 원시 서열의 k-mer 일치가 아닌 256차원 잠재 벡터 간의 코사인 거리에 의해 구동된다. 동일한 잠재 표현이 재구성, 다운스트림 속성 예측 및 검색에 재사용되는 반면, MMseqs2는 빠른 서열 비교에 특화되어 있으며 다운스트림 모델에 직접 통합할 수 있는 공유 잠재 기하학을 노출하지 않는다. 둘째, LLAST는 명시적이고 사용자 제어 가능한 후보

예산($N_{\max}$ 및 TopK를 통해)을 강제하고 최종 점수 계산을 BLAST에 위임하여, 사실상 BLAST를 학습된 인덱스 위의 예산이 책정된 재순위화(re-ranking) 계층으로 바꾼다. 대조적으로, MMseqs2는 자체 필터링 및 정렬 단계를 통합하며 일반적으로 전체 검색 파이프라인으로 사용되거나 고처리량 워크플로에서 BLAST의 대체품으로 사용된다.

이러한 관점에서 LLAST는 MMseqs2와 같은 도구의 직접적인 경쟁자가 아니라 보완적인 것으로 간주되어야 한다. 주요 목표가 최대 처리량으로 초대형 데이터베이스를 스캔하는 시나리오에서는 MMseqs2와 같은 서열 기반 엔진이 선택의 방법으로 남는다. 대신 본 연구는 단일 구조 정렬 잠재 공간이 기능적 모델링과 명시적 후보 예산 하에서 BLAST 작업량을 상당히 줄이는 잠재-트리 프리필터를 모두 지원할 수 있음을 보여준다. 공유 하드웨어 및 데이터베이스에서의 MMseqs2와의 체계적인 실제 시간(wall-clock) 비교는 가치가 있겠지만, 본 연구의 범위를 벗어난다.

## 4. 결론

우리는 ESMS 기반 지각 손실을 사용하여 활성 KL 발산을 유지하면서 256차원 잠재 공간을 단백질 구조적 제약 조건과 정렬하는 경량 변분 인코더인 LATTE를 소개했다. 형광 단백질에서 LATTE 임베딩으로 훈련된 가우시안 프로세스 모델은 높은 정확도로 FP와 비-FP를 분리하고 ESM-2 임베딩 기준선에 필적하는 RMSE로 여기 및 방출 피크를 예측한다. LATTE가 훨씬 더 적은 파라미터와 훨씬 낮은 추론 비용을 가짐에도 불구하고 말이다. 더 넓고 꼬리가 두꺼운 코사인 거리 분포, ESM-2와의 강력한 순위 일치도, 구조적 손실에 대한 어블레이션 연구와 함께, 이러한 결과는 LATTE의 저차원 잠재 변수가 다운스트림 기능 모델링에 적합한 정보가 풍부하고 구조 정렬된 표현임을 나타낸다.

이 표현을 바탕으로 우리는 BLAST를 위한 프리필터 역할을 하는 잠재-트리 인덱스인 LLAST를 구축했다. 약 $10^6$개 서열 데이터베이스에서 LLAST는 쿼리당 BLAST로 전달되는 후보 세트를 최대$N_{\max} = 10000$개(데이터베이스의 약 0.97%)로 제한하면서 기준선 BLAST 상위 50개 고유 히트의 약 55%를 회수한다. $N_{\max} = 4000$인 더 공격적인 구성은 평균 SeqRecall@50을 약 0.53으로 유지하면서 평균 BLAST 작업량을 약 260배 줄여, 재현율과 후보 세트 크기 간의 조정 가능한 트레이드오프를 보여준다. 고정된 잎 용량 및 검색 예산 하에서 결합된 프리필터+BLAST 파이프라인은 데이터베이스 크기에 대해 로그 의존성만을 나타내어 실제 규모에서 쿼리당 비용을 효과적으로 제한한다.

종합하면, LATTE와 LLAST는 콤팩트하고 구조를 인지하는 잠재 공간이 정확한 재구성, 속성 예측, 그리고 확장 가능한 서열 검색을 동시에 지원할 수 있음을 보여준다. 생성적 인코더와 검색 인덱스를 별도의 시스템으로 보는 대신, 우리의 결과는 이를 동일한 잠재 기하학의 다른 용도로 취급할 것을 제안한다. 이 접근 방식을 더 큰 데이터베이스로 확장하면 수천만 또는 수억 개의 서열 규모에서도 효율적으로 유지되는 단백질 검색 시스템을 구축할 수 있다.

마지막으로, 본 연구에서 제안된 기본 설정 외에도 LATTE 모델의 성능을 극대화하기 위한 하이퍼파라미터 정밀 조정(tuning)이 중요한 향후 과제로 남아있다. 인코더의 층 수나 잠재 차원의 크기뿐만 아니라, 손실 함수 내 구조적 항과 정규화 항 사이의 가중치 균형($\lambda, \alpha, \beta$)을 체계적으로 최적화함으로써 모델의 표현력을 더욱 강화할 수 있을 것이다. 이러한 조정은 LATTE가 다양한 단백질 패밀리와 다운스트림 작업에 유연하게 적응하도록 돕고, 재구성 정확도와 검색 속도 사이의 트레이드오프를 더욱 개선할 수 있을 것으로 전망된다.

# Ⅵ. 참고문헌

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol.

1990;215(3):403–410.

Anfinsen CB. Principles that Govern the Folding of Protein Chains. Science. 1973;181(4096):223–230.

Brändén CI, Tooze J. Introduction to Protein Structure. 2nd ed. Garland Science; 1999.

Camacho C, Coulouris G, Avagyan V, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:421.

Dill KA, MacCallum JL. The Protein-Folding Problem, 50 Years On. Science. 2012;338(6110):1042–1046.

FPbase contributors. FPbase: A community-editable fluorescent protein database. FPbase.

Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model. Science. 2025;387(6736):850–858.

Indyk P, Motwani R. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. STOC. 1998:604–613.

Jégou H, Douze M, Schmid C. Product Quantization for Nearest Neighbor Search. IEEE TPAMI. 2011;33(1):117–128.

Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. ECCV. 2016:694–711.

Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.

Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS. 1990;87(6):2264–2268.

Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980. 2014.

Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv:1312.6114. 2014.

Lambert TJ. FPbase: a community-editable fluorescent protein database. Nat Methods. 2019;16(3):239–244.

Lin Z, Akin H, Rao R, et al. ESMFold: End-to-end single-sequence protein structure prediction. bioRxiv. 2022.

Lloyd SP. Least Squares Quantization in PCM. IEEE Trans Inf Theory. 1982;28(2):129–137.

Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classification. 2014;31(3):274–295.

Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.

Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS. 2021;118(15):e2016239118.

van der Maaten L, Hinton G. Visualizing data using t-SNE. JMLR. 2008;9:2579–2605.

Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. NIPS. 2017:5998–6008.

Vieira LC, Handojo ML, Wilke CO. Medium-sized protein language models perform well at transfer learning on realistic datasets. Sci Rep. 2025;15(1):21400.