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 정규화 항, 그리고 잔기 간 유사성을 감독하는 구조 항을 결합한 복합 손실 함수를 사용한다. LATTE는 보지 못한 UniRef50 서열을 97.17% 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-겹 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 파라미터으로 달성한 결과이다. 동일한 데이터셋에서, LATTE의 pairwise cosine distance는 ESM-2보다 더 넓고 heavy-tailed(극단적인 값이 자주 발생하는 분포)한 모습을 보이지만, 두 공간 사이의 순위 상관(Spearman ρ = 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 10^6개의 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 조절 가능한 트레이드오프를 보여준다.

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

Ⅰ. 연구의 동기 및 목적

본 연구의 동기는 대규모 단백질 서열 데이터베이스에서 BLAST가 여전히 사실상의 표준임에도 불구하고, 쿼리당 계산 비용이 데이터베이스 크기에 거의 선형으로 증가한다는 구조적 한계에서 출발한다. 유사한 쿼리가 반복되면 이미 검토했던 유사 영역을 다시 넓게 훑는 중복 계산이 누적되고, 대형 단백질 언어모델(예: ESM-2)은 표현력은 크지만 추론 지연과 자원 소모가 커서 실시간 선별 단계에 직접 사용하기 어렵다. 단순 서열 유사도만으로는 기능적으로 의미 있는 근접성을 충분히 반영하기 어렵다는 점도 문제다. 따라서 정렬 수행 이전 단계에서 빠르게 후보를 가지치기해 BLAST의 입력 집합을 줄이는 정렬 전 사전 필터링이 필요하다.

Ⅱ. 이론적 배경

1. 단백질 (Protein)

단백질은 펩타이드 결합으로 연결된 아미노산의 선형 중합체로, 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 결정한다. 이러한 상호작용은 2차 구조(알파 나선, 베타 병풍), 3차 구조, 그리고 다중 소단위체 복합체의 4차 구조를 형성한다.

2. 딥러닝 (Deep Learning)

딥러닝은 대규모 데이터에서 전이 가능한 서열 표현을 학습함으로써 단백질 모델링을 가속해 왔다. 딥러닝은 대규모 비라벨 서열 코퍼스로부터 구조 및 기능 제약을 암묵적으로 인코딩하는 고차원 서열 표현을 학습함으로써 단백질 모델링을 변화시켰다. 마스크드 언어 모델(MLM) 계열 PLM(예: ESM-2)은 장거리 잔기 결합을 포착하고, ESMFold를 통해 단일 서열 기반의 고정도 구조 예측을 가능하게 한다. ESM-3는 더 나아가 서열, 구조, 기능을 밀접하게 결합하여 편집/설계 능력을 확장한다. 그러나 이러한 파운데이션 모델은 지연(latency)과 메모리 비용이 커서, 대규모 스크리닝과 반복적인 설계에 제약이 된다.

이에 보완책으로, 우리는 잠재 변수가 활성(active) 상태를 유지하며 정보량이 풍부한 VAE(variational autoencoder) 스타일의 콤팩트한 인코더-디코더를 채택한다. 이 잠재 변수는 ESMS에 대한 지각(perceptual) 손실로 정규화되어, 잠재 공간 기하가 구조와 정렬(aligned)되도록 한다. 이렇게 얻어진 잠재 표현은 빠른 클러스터링/검색과 다운스트림 특성 모델을 지원하며, 훨씬 적은 연산 비용으로도 전이 학습에서 경쟁력 있는 성능을 보인다. 또한, 이 잠재 표현은 정렬 이전 프루닝(pre-alignment pruning; LLAST)의 기반이 되어, 검색 범위를 줄이면서도 해석 가능성을 유지한다.

3. ESM-2와 ESMS

우리는 사전학습된 ESM-2 임베딩을 이용하여, ESM-2의 서열당 약 0.5초의 추론 시간을 줄이면서도 높은 임베딩 충실도를 보존하는 ESMS를 개발하였다. Rives 등은 짧은/긴 거리 잔기 상호작용을 포착하는 마스크드 언어 모델 ESM-2를 도입하였고, 이는 AlphaFold 2와 유사한 성능을 보이는 3차원 구조 예측기 ESMFold의 기반이 되며, 2차 구조와 열안정성 예측 등 다양한 다운스트림 작업에 널리 사용된다.

ESM-2는 다양한 파라미터 규모(그 중 650M 파라미터 버전이 가장 흔하다)로 제공되지만, 그 추론 지연

은 보다 경량 대안의 필요성을 자극한다. 보류된 테스트 세트에서, ESMS는 코사인 유사도 0.9647, RMSE 1.2998을 달성했다.

우리는 본 연구를 최근 대형 단백질 언어 모델의 진전 속에 위치시킨다. 특히 ESM-3는 서열, 구조, 기능 감독을 긴밀히 결합하여, 이전 ESM-2 기반 시스템을 넘어서는 설계/편집 능력을 보여준다. LATTE는 이러한 파운데이션 모델을 대체하기보다는 보완하는 경량 구조 인지 생성기로 설계되었으며, 제어 가능한 잠재 구조와 효율적인 학습에 중점을 둔다.

4. BLAST

BLAST (Basic Local Alignment Search Tool)는 시드-앤-익스텐드(seed-and-extend) 휴리스틱을 통해 국소적 유사성을 찾아내어 높은 점수를 갖는 세그먼트 쌍(HSPs)을 신속하게 형성한다. 정렬의 통계적 유의성은 Karlin-Altschul 프레임워크에 의해 모델링되어 E-값과 비트 점수를 산출한다. 최적의 동적 프로그래밍 정렬보다 민감도는 낮지만, BLAST는 대규모 데이터베이스에 효과적으로 확장되며 실용적인 표준으로 남아 있다. BLAST+ 재구축은 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

Transformer 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_q가 주어졌을 때, 단백질 검색은 보통 코사인 유사도나 최대 내적 탐색(MIPS)에서의 내적 등유사도 함수 $s(z_q, z_i)$를 기준으로 한 최근접 이웃 검색 문제로 정식화된다. 단순한 top-k 검색은 모든 N개 항목에 대해 $s(z_q, z_i)$를 계산해야 하므로 쿼리당 시간 복잡도가 $O(Nd)$가 되어, N이 커지면 계산 비용이 급격히 증가한다.

학습 기반 검색 알고리즘은 임베딩 공간의 기하 구조를 활용하는 인덱스를 구축함으로써 이 비용을 완화한다. 그래프 기반 방법은 각 데이터 포인트를 소수의 이웃과 연결하고, 탐색 깊이와 그래프 차수에 의해 유사도 계산 횟수가 제한되도록 탐욕적(greedy) 탐색을 수행한다. 클러스터 기반 방법은 공간을 K개의 셀로 분할하는 거친 양자화(coarse quantizer)를 만들고, 쿼리 z_q에 가장 가까운 소수의 셀만 탐색함으로써 스캔해야 할 후보를 N에서 $(\frac{N}{K}) \cdot n_{probe}$정도로 줄인다. 여기에 product quantization이나 해싱을 결합해 거리 계산 자체를 더 빠르게 하는 변형도 있다.

대규모 검색 시스템에서는 보통 이런 인덱스를 계단식(cascade)의 첫 단계로 사용한다. 인덱스는 $M \ll N$개의 후보 집합 $S \subset D$를 반환하고, 더 정확하지만 느린 점수 계산기는 이 후보 S에만 적용된다. 단백질 검색에서는 이를 단백질 임베딩 기반 인덱스로 서열 공간에서 명백히 무관한 영역을 버리고, 그 후 줄어든 후보 집합에 대해서만 BLAST와 같은 정렬 도구를 실행하는 것으로 생각할 수 있다. 그래프 차수, 탐색할 셀의 수(nprobe), 탐색 깊이와 같은 하이퍼파라미터는 top-k에서의 재현율과 평균·꼬리 지연(latency)의 균형을 맞추기 위해 조정된다.

Ⅲ. 연구 방법

1. 잠재공간 형성 (Shaping the latent space)

인코더 메모리와 잠재 벡터를 받아 훈련 신호 생성기 역할을 하는 의사-디코더(pseudo-decoder)를 사용하여 LATTE를 훈련했다. 또한, 전체 구조 예측기보다 훨씬 낮은 비용으로 잔기-위치 규칙성을 포착하는 사전 훈련된 ESMS 임베딩으로부터 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.

$L_{COS} = [1 - \cos(\mathrm{ESMS}(x_{orig}), ESMS(x_{recon}))]$, $L_{MSE} = |\mathrm{ESMS}(x_{orig}) - ESMS(x_{recon})|_2^2$

이 구조 항들은 교사강제(next-token) 교차엔트로피 L_{CE}와 KL 발산 L_{KL}과 결합하여

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

여기서 λ = 5 이고, α는 처음 100 Epoch에 걸쳐 30에서 0.1로 선형적으로 감소하고, β는 동일한 간격 동안 0에서 0.1로 선형적으로 증가한다. 코사인 항은 유사하게 가능성 있는 잔기들 간의 치환을 허용하는 반면, MSE 항은 큰 편차에 불이익을 준다. 이들은 함께 유용한 잠재 정보에 의존하는 정확한 재구성을 만듦으로써 후방 붕괴(posterior collapse)를 방지한다.

LATTE는 5.5M 파라미터의 경량 트랜스포머로, 4-층 트랜스포머 인코더로 구성된다(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

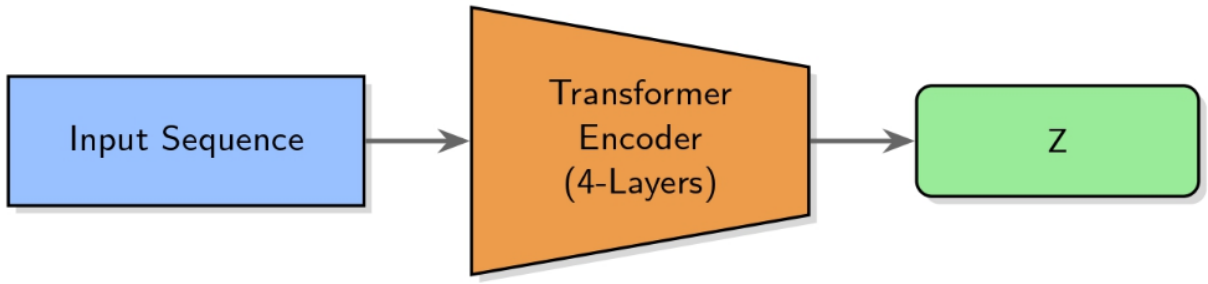
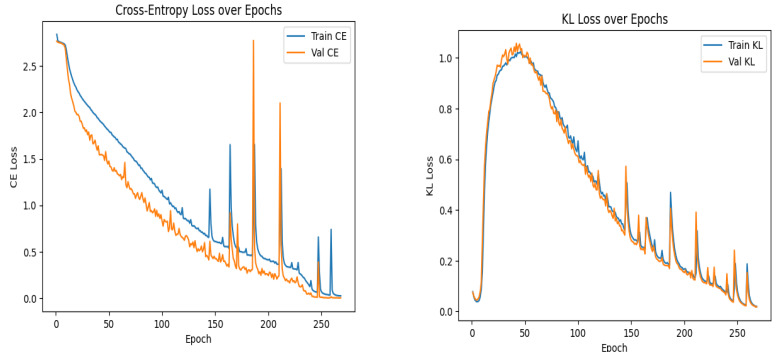


그림 1. LATTE 구조 개요: 입력 서열 → 4-층 인코더 → 잠재 분포(평균/로그분산)

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 재구성률에 도달했지만, 매우 낮은 KL 값은 잠재적인 후방 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

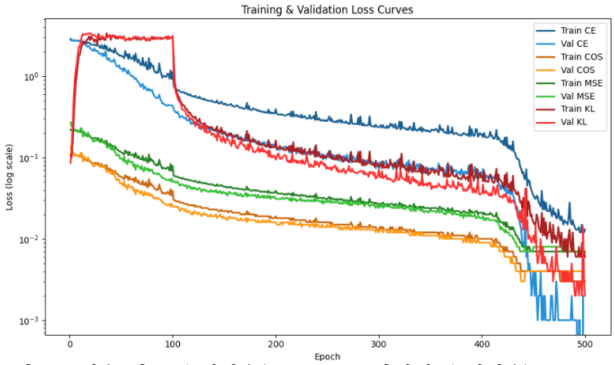
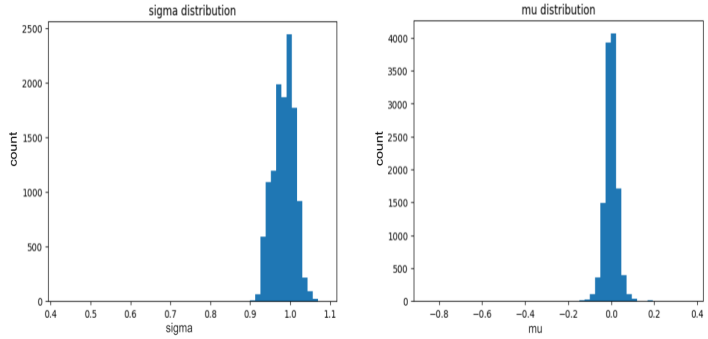
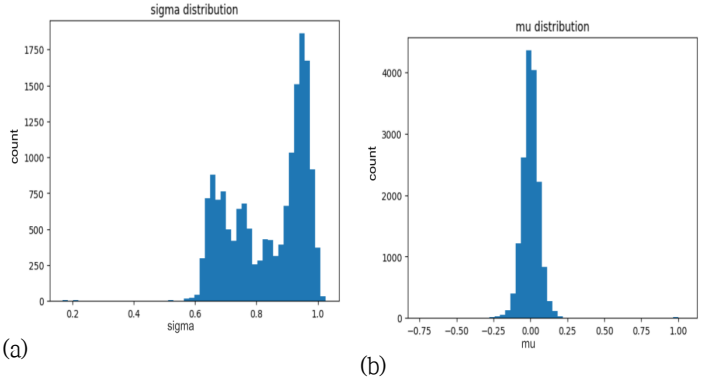


그림 3. 학습/검증 손실함수(Epoch 500 까지의 손실함수)



(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

Locality-sensitive hashing이나 product quantization과 같이 고정 임베딩에 대해 동작하는 일반적인 근사 최근접 이웃(ANN) 방법과 달리, LLAST는 구조 정보가 반영된 LATTE 잠재 공간 위에 태스크 특화 계층적 인덱스를 구축한다. 이를 통해 동일한 잠재 표현을 생성적 설계, 성질 예측, 서열 검색에 재사용할 수 있다.

우리는 LATTE Latent Alignment Search Tool(LLAST)을 도입하였다. 먼저 코사인 거리를 기준으로 L2 정규화된 잠재 벡터에 대해 k-평균 클러스터링을 수행한 뒤, 응집형 계층적 클러스터링을 적용하여 트리 인덱스를 생성하였다(그림 6). 이 트리는 프리필터로 사용되어, 잠재 벡터가 크게 다른 서열들을 정렬 전에 제거한다. 클러스터 개수 K는 평균 코사인 k-평균 비용 곡선에 대한 Kneedle 무릎(knee) 검출 방법과, 그 주변에 대한 적응적 10단계 인터벌 보정을 사용하여 선택하였다(그림 7).

.

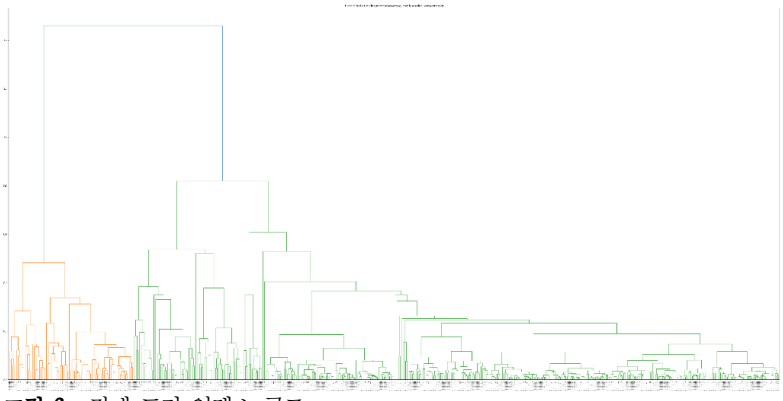


그림 6. 잠재 트리 인덱스 구조

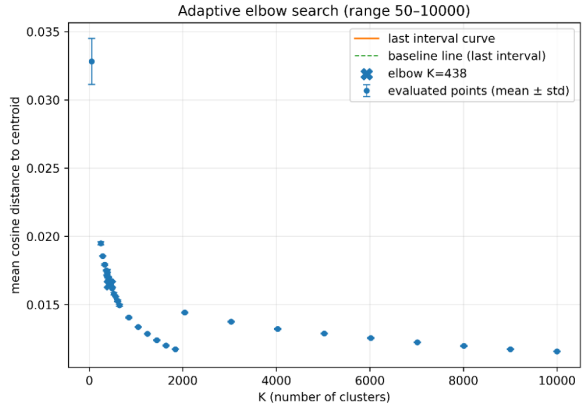


그림 7. 적응형 엘보 탐색

Ⅳ. 연구 결과

1. 사례 연구: 형광단백질(FP) 분석 및 생성

사전학습 ESM-2 임베딩 을 사용하고, FPbase에서 수집한 FP 서열로 가우시안 프로세스(GP) 를 학습해 (1) FP/비-FP 분류, (2) 스펙트럼 피크 회귀를 수행했다. GP 분류기는 5-fold CV 정확도 0.987, 회귀기는 λ_{abs} 2.70 nm / λ_{em} 3.80 nm RMSE 를 달성했다.

ESM-2 기준선: 동일 파이프라인에서 ESM-2(650M)는 5-fold AUC 0.997, 2.70/3.80 nm RMSE를 보였다. LATTE 256-차원 잠재 임베딩 은 정확도 0.987 로, ≈100배 작은 파라미터 수 와 낮은 추론비용으로 경쟁력 있는 다운스트림 신호 를 제공한다.

표 2. 서로 다른 임베딩에서의 GFP 스펙트럼 예측 성능

Embedding	Classifier metric (5-fold)	λ_{abs} RMSE (nm)	λ_{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 분포

되어 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.

투영된 FP 임베딩의 k-평균 클러스터링은 세 개의 뚜렷한 클러스터를 밝혔으며(그림 9, 이들의 컨센서스 벡터는 각 클러스터의 평균 길이로 잘린 새로운 서열로 디코딩되었다. 훈련된 GP 분류기를 사용하여 이러한 생성된 서열을 분류한 결과, 클러스터 1이 100%의 성공률을 달성했으며(그림 10), 이는 기능성 단백질을 생성하는 데 있어 샘플 크기보다 클러스터 내 샘플 품질과 일관성이 더 중요함을 시사한다. 생성된 단백질들은 FPbase 서열과 75% - 96%의 상동성을 보였으며, 이는 그들의 잠재 표현의 높은 유사성을 반영한다.

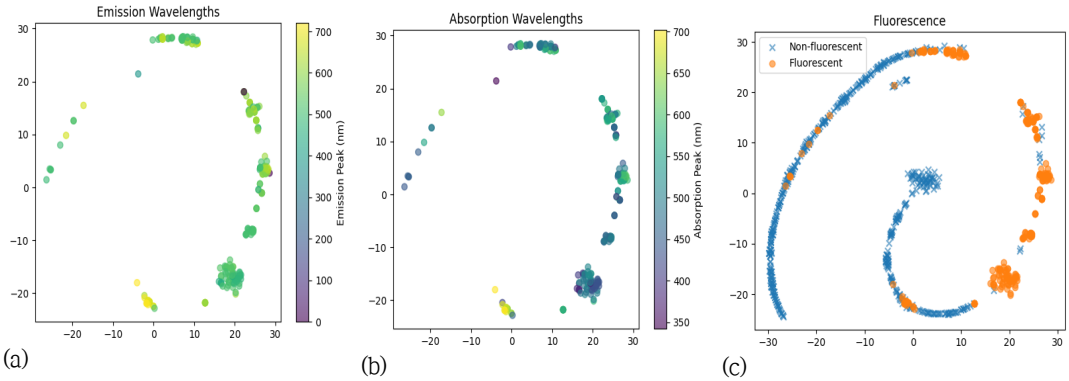


그림 8. 형광 단백질의 특성 (a): 방출 스펙트럼 (b): 흡수 스펙트럼 (C): 형광 세기

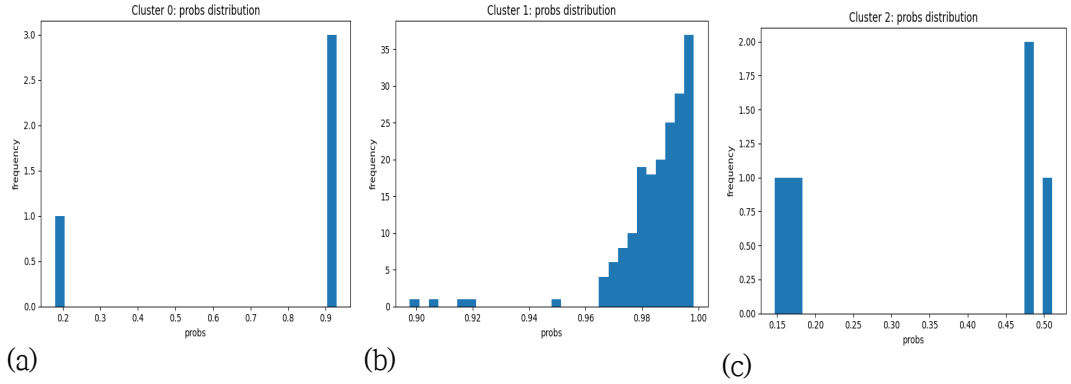


그림 9 클러스터별 예측 확률 분포 (a): Cluster 0, (b): Cluster 1 (c): Cluster 2

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

잠재 공간에서 가까운 벡터들은 AlphaFold 예측에 의해 확인된 바와 같이 유사한 3차원 구조를 가진 단백질을 인코딩한다. 클러스터 1의 단백질들은 pLDDT 점수가 90 이상인 전형적인 β-배럴 구조를 보였다. 클러스터 0의 단백질들은 α-나선으로 둘러싸인 세 개의 내부 β-가닥의 두 그룹을 형성했다. 클러스터 2의 단백질들은 β-배럴을 채택하지 않았지만, 그럼에도 불구하고 α-나선으로 둘러싸인 세 개의 β-가닥이라는 공통 구조를 공유했다.

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 ρ=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.

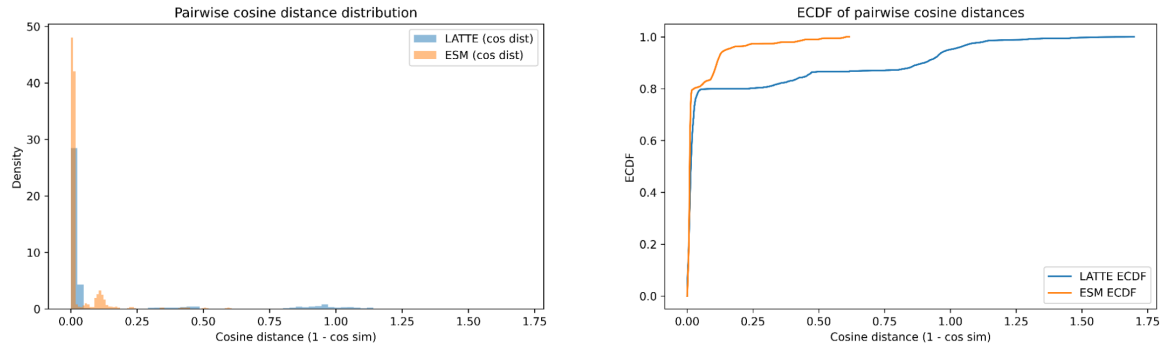
Cross partition agreement (LATTE vs. ESM-2)	
Adjusted Rand Index (ARI)	0.7373
Adjusted Mutual Information (AMI)	0.6055
Fowlkes-Mallows Index (FMI)	0.9596
Variation of Information (VI)	0.3529
Purity (LATTE → ESM-2)	0.9633

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

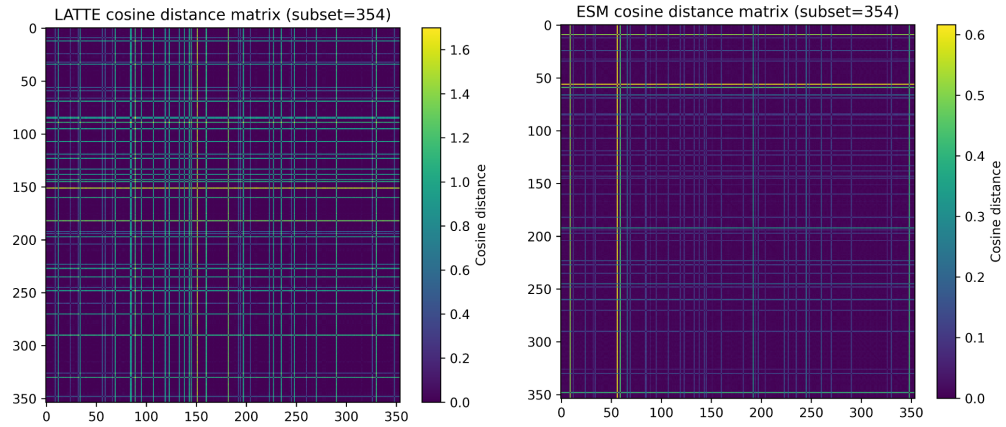
표 10. 클러스터별 크기와 평균값.

Cluster	size	LATTE μ	ESM-2 μ
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

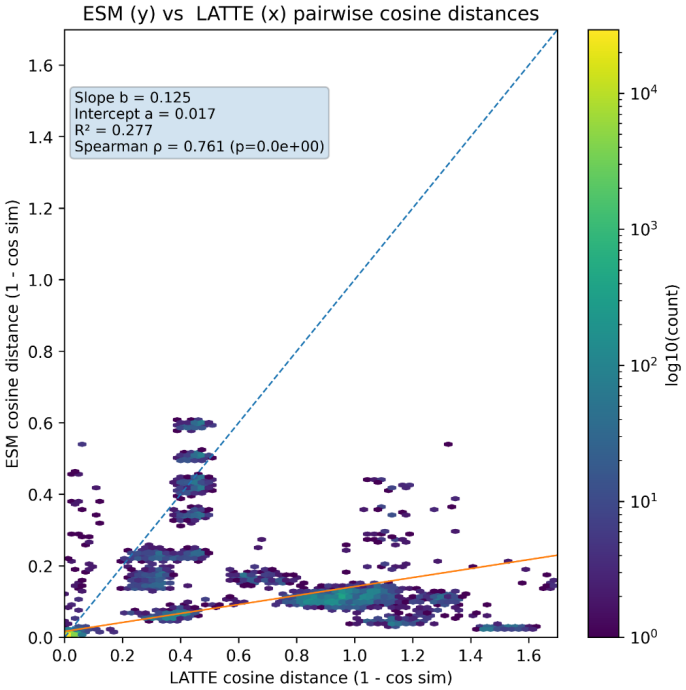
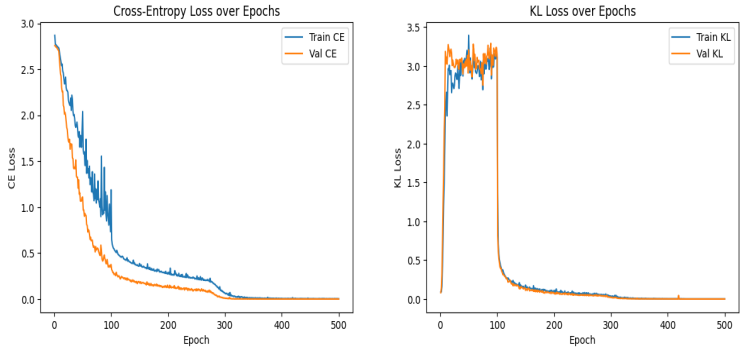


그림 12. 일치하는 쌍에 대한 ESM (y) 대 LATTE (x) pairwise cosine distance. 실선은 OLS 적합선이다 (b=0.125, a=0.017; R^2 =0.277). Spearman ρ=0.761은 ESM의 스케일 압축에도 불구하고 강한 순위 일치도를 나타낸다.

3. 어블레이션 연구

손실 함수를 다음과 같이 설정하여 어블레이션 연구를 수행했다. 가중치는 Epoch< 30 동안 α = 30, β = 0으로 설정되고, 그 이후에는 α = 0.1, β = 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)

LLAST를 전통적인 BLAST 앞단의 프리필터로 사용하기 위해, 우리는 커버리지(coverage)와 효율성(efficiency)을 기준으로 성능을 평가하였다. 커버리지는 프리필터 이후에도 기준선 BLAST top-50 중 고유한 서열이 얼마나 회수되는지를 나타내는 SeqRecall@50으로 정의하고, 효율성은 BLAST로 실제 전달되는 후보 서열 수(쿼리당 candidate size)로 측정하였다. 약 N ≈ 10^6개의 서열로 구성된 데이터베이스에 대해, 잠재 트리 인덱스는 잎(leaf)당 약 S ≈ 2000개 서열이 포함되도록 구성되었으며, 쿼리 시에는 트리에서 점수 상위 K개의 잎만 탐색한다. 우리는 BLAST로 전달될 후보 수의 상한 N_max를 {2000, 4000, 6000, 8000, 10000}으로 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 공격적인 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, 잎 용량을 S, 쿼리당 방문되는 잎 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 BLAST는 $T_{BLAST(N)} = O(N)$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

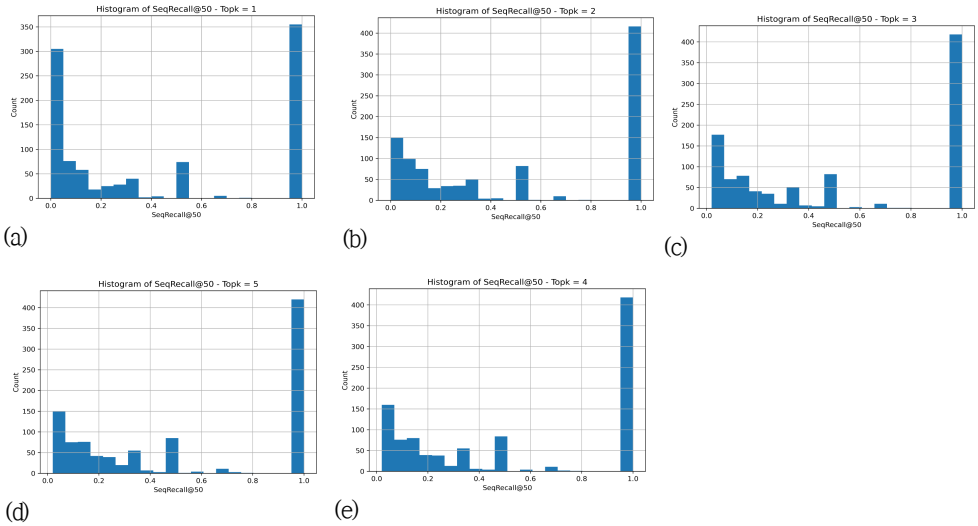


그림 14. 서로 다른 TopK 값(1-5)에 대한 프리필터 단계의 시퀀스 재현율(SeqRecall@50) 히스토그램: (a) TopK=1, (b) TopK=2, (c) TopK=3, (d) TopK=4, (e) TopK=5.

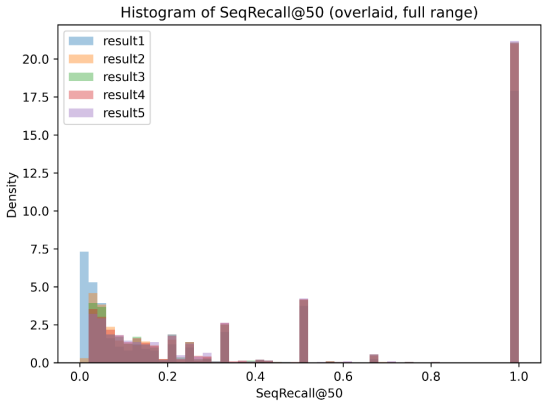


그림 15. TopK 조건별 시퀀스 재현율 히스토그램을 겹쳐서 그린 비교.

V. 결론 및 논의

향후 과제: 구조 감독을 직접적으로 적용하는 접근—예를 들어 AlphaFold 기반 손실이나 접촉 지도(contact map)를 통한 구조 정렬—은 계산 비용이 지나치게 크고, 대규모 서열 데이터에 적용하기에는 비

효율적이다. AlphaFold는 단일 서열 예측에 10-20분이 소요되고, 접촉 지도는 큰 메모리 요구량과 제한된 실측 구조 데이터 수 때문에 광범위한 일반화에 어려움이 있다. 이에 반해 ProtT5, ProtBert, ESM-2의 소형 변종 등 다양한 단백질 언어 모델의 중간 표현을 구조 감독을 위한 지각 신호(perceptual supervision)로 활용하는 방식은 훨씬 확장성이 높고 데이터 제약도 적다. 향후 연구에서는 다양한 PLM 기반 지각 손실 비교, 효소 활성·결합 친화도 같은 기능 레이블을 잠재 공간에 통합하는 확장, 그리고 더 큰 범주의 기능적 단백질 패밀리를 대상으로 LATTE 기반 잠재 인덱스를 구축하는 방향 등으로 발전할 수 있다.

GFP 사례 연구에서 LATTE의 256차원 구조 인지 잠재 공간은 형광 기능과 스펙트럼 특성을 재구성과 분리 없이도 안정적으로 포착함을 보였다. LATTE 임베딩으로 학습한 GP 분류기는 FP vs 비-FP를 5겹 교차 검증 정확도 0.987로 구분하며, GP 회귀 모델은 흡수·방출 최대파장을 각각 2.70 nm, 3.80 nm RMSE로 예측한다. t-SNE 시각화는 잠재 공간에서 FP와 비-FP의 분리 및 스펙트럼 특성의 연속적 변화(gradient)를 보여 주며, 이는 잠재 표현이 구조적·기능적 제약을 모두 인코딩하고 있음을 뒷받침한다. 종합하면 LATTE의 저차원 잠재 표현은 정보량이 높고 구조 정렬이 잘 되어 있어, 신뢰성 있는 성질 추정과 더 넓은 단백질 기능 패밀리로의 확장을 위한 일반적 경로를 제공한다.

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.