

낮은 신뢰도는 부재가 아니다: 정렬 깊이가 파지 역전사효소 파트너 단백질의 구조 예측에 대해 말해 주는 것

저자: Chulmin Pyo, AI3 Discovery. 교신: discovery@ai3.kr

공개 상태: 자체 공개 연구 원고. 사람의 독립적인 동료 심사를 거치지 않음.

범위 표기. 이 글이 다루는 것은 두 가지다. 하나는 시퀀싱 런 어셈블리에서 서열 검색으로 찾은 상동체이고, 다른 하나는 서열을 확보한 반복 배열 역전사효소 좌위의 파트너 단백질이며, 모두 예측 구조에 대한 기술이다. 대상은 *Listeria* 파지 LPJP1의 두 파라로그 파트너(원 기술에 ART_11로 실린 좌위이고, P2는 길이와 간격이 원 기술 도면의 파트너와 같다), 그 상동체 91개 체인 패밀리, 다른 좌위의 파트너 네 개(둘은 길이와 간격으로 원 기술의 도면과 맞춘 것, 둘은 같은 계통의 서로 다른 접근번호이며 좌위의 지문이 도식된 도면과 맞지 않은 것)이다. 원 기술의 구조 분석은 다른 좌위(ART_57, 접근번호 미공개)에서 이루어졌다. 신규성을 주장하지 않으며 효소 활성화에 대해서도 말하지 않는다.

초록

저신뢰 구조 예측은 흔히 "도메인이 없다"는 뜻으로 읽히지만, 배양하지 않은 파지의 단백질에서는 예측기가 기대는 공개 정렬이 너무 얇아, 그렇게 읽을 근거가 되지 못하는 경우가 많다. 최근 한 프리프린트는 반복 배열과 함께 나타나는 역전사효소(ART)의 Type I 파트너가 GNAT(GCN5-related N-acetyltransferase) 유사 폴드 두 개가 직렬로 이어진 구조라고 보고했다. 다만 분석 결과를 보여 준 파트너 하나에서 두 번째 영역은 tGCN5와의 유사성이 더 약한(TM 0.31) 91잔기의 C 말단 로브였다(Yoon et al. [1]). 우리는 부족한 정렬 깊이를 공개 데이터로 채웠을 때 구조 예측이 이런 파트너에 대해 무엇을 말해 주는지, 그리고 그 답이 한 좌위를 넘어서도 유지되는지를 물었다. 초기 판정 규칙은 예측에 앞서 날짜가 남는 로컬 git 기록에 고정했고, 마지막 단계에서는 Pfam 검색을 다른 판독을 읽은 뒤에 추가했다. 대상은 *Listeria* 파지 LPJP1의 두 파라로그 파트너, 그 상동체 91개 체인 패밀리, 서열이 완전한 다른 반복 배열 좌위의 파트너 네 개이며, 모두 AlphaFold2-multimer v3와 Boltz-2로 단독 체인을 예측했다. 입력 정렬은 ColabFold 서버 정렬(최대 18행)과 여기에 공개 Logan 단백질 클러스터의 상동체를 더한 정렬이다. 결과는 Foldseek, Pfam 클랜 모델, 이름을 쓰지 않는 구조 패널 검사로 읽었으며, 양성·음성 대조군과 행을 섞은 정렬 대조를 나란히 돌렸다. 상동체를 더하자 P2의 AlphaFold2 평균 pLDDT는 80–83으로 올랐지만 행을 섞은 정렬에서는 35였고, ART_69 파트너(594잔기)는 체인 대부분이 높은 신뢰도로 예측되었다(88.4–88.8, 섞으면 46–49). 반면 P1과 P2의 N 말단은 끝까지 저신뢰였다. 세 판독은 서로 어긋났다. Foldseek은 GNAT 유사 영역을 많아야 하나 찾았고 두 개를 찾은 체인은 없었다(P2와 패밀리 체인 91개 중 AlphaFold2 35개·Boltz-2 17개는 NONE으로 분류되었다). Pfam 클랜 모델은 패밀리 단계에서 GNAT 대조군 항목 8개(고유 단백질 7개)를, 마지막 단계에서는 3개를 모두 잡아냈지만 표적에서는 클랜 히트가 없었다. 구조 패널 검사의 사전 지정한 두 영역 기준은 패밀리 체인 91개 중 56개와 추가 파트너 4개의 모델 24개 모두에서 충족되었다(세 시드 모델 집합은 결과를 읽은 뒤에 기록했고, 3개 계통의 정렬을 공유하므로 독립된 관찰이 아니다). 이와 별도로 탐색적 두 번째 영역 점수로는 두 번째 원도를 채점할 수 있었던 패밀리 체인 71개 중 55개에서 맞닿은 원도 두 개를 찾았다(전체 91개 기준으로는 55개, 60%). 고정한 판정은 P1과 P2에서 UNKNOWN이었고, 추가 파트너에는 판정을 내리지 않았으며 셀 라벨 8개는 모두 unknown이었다. 따라서 얇은 정렬 위에서 얻은 저신뢰 모델은 "두 번째 도메인 없음"의 근거가 될 수 없고, 서열이 많이 달라진 GNAT 유사 폴드릴 가능성과 다른 알파/베타 폴드릴 가능성이 모두 남는다. 다만 소수 파트너를 예측한 결과일 뿐 효소 실험이 없고, 프리프린트가 분석한 좌위 자체는 구하지 못했으므로 그 주장을 뒷받침하지도 반박하지도 않는다.

1. 서론

구조 예측은 기능을 모르는 단백질의 폴드를 추정하는 데 일상적으로 쓰이며, 추정의 신뢰도는 예측기에 입력하는 다중 서열 정렬(MSA)에 좌우된다. 배양하지 않은 파지와 메타게놈 어셈블리에서 온 단백질은 공개 MSA 데이터베이스에 상동체가 거의 없을 수 있다. 그런 경우 모델의 신뢰도가 낮은 것은 단백질의 성질 탓이 아니라 정렬이 얇은 탓일 수 있다. 음성 결론("두 번째 도메인 없음")은 그 영역을 높은 신뢰도로 예측했을 때만 의미가 있다.

Yoon et al. [1]이 기술한 배열 연관 역전사효소(ART)는 반복 배열 및 전용 파트너 유전자와 짝을 이룬다. Type I 파트너는 "약 600잔기이고 두 개의 직렬 GNAT 유사 폴드를 가진다"고 적혀 있다. 근거는 저자들이 자체 메타게놈 데이터로 만든 좌위 쌍 패밀리 정렬을 넣은 AlphaFold2 공동 접힘(co-fold)이다. TM-align 연산(tGCN5를 체인 전체에, 이어서 나머지 부분에 정렬)으로 분석한 파트너 하나(ART_57 파트너, 약 464잔기)를 보면, 첫 영역(잔기 212-373)은 TM 0.58, 나머지 C 말단 91잔기(374-464)는 TM 0.31로 맞았다. 프리프린트 보충 Fig. 7C의 표는 이 구간을 "C-terminal lobe"라 적었고, 보충 Fig. 7B의 구조 그림에는 "GNAT-like lobe"라는 표시가 붙어 있다. P-loop도 strand beta5도 없지만 겹쳐 놓은 acetyl-CoA에서 2 Å 안에 드는 원자가 20개라고 적어 두었다(첫 영역은 32개). 본문은 직렬 GNAT 유사 폴드 두 개를 말한다. 이번 분석은 연구 발표문에서 연결한 [Anthropic 게시 원고](#)를 대상으로 했다. 서술은 이 원본 사본을 기준으로 하며, 해당 사본의 동료 심사 여부는 확인하지 못했다.

우리는 공개 게놈(*Listeria* 파지 LPJP1, 파라로그 P1과 P2)의 Type I 쌍에서 출발했다. LPJP1은 프리프린트에 ART_11로 올라 있고 P2는 거기 그려진 파트너와 같지만(3.9절), 프리프린트의 GNAT 분석은 접근번호가 공개되지 않아 우리가 구하지 못한 ART_57에서 이루어졌다. 질문은 처음에 셋이었다. (i) 공개 MSA 파이프라인이 원 기술을 재현하는가, (ii) 부족한 정렬 깊이를 공개 서열 데이터로 채울 수 있는가, (iii) 회수한 상동체 패밀리 전체를 같은 검출기로 예측하고 검색하면 무엇이 나오는가. 이 단계들이 끝난 뒤 네 번째 질문을 더했다. (iv) 서열이 완전한 다른 반복 배열 좌위의 파트너도 같은 양상을 보이는가. 우리가 쓴 구조 검색 검출기는 전체 크기의 GNAT 유사 영역(141-213 aa)을 판정하고, 여기에 더해 로브 분석(60-140 aa)도 따로 수행했다. 두 분석의 내용과, 그 분석으로 배제하지 못하는 것은 아래에 함께 적었다. Figure 1에 단계와 판독을 개관했다.

우리는 어떤 예측보다 먼저 판정 규칙을 사전 지정했고(방법에서 정의한 뜻으로, 로컬에 기록했을 뿐 외부 시각 증명은 없다), 대조군을 먼저 돌렸으며 모든 이탈을 기록했다. 이 연구는 예측에 근거한 음성이 어떻게 음성이 아니라 측정되지 않은 것일 수 있는지 보여 주는 실례이며, 일반적인 추정치가 아니다.

2. 방법

사전 지정과 검토. 첫 단계(D3)의 판정 규칙은 MSA 기반 예측이 시작되기 전에 날짜가 기록되고 해시된 문서로 고정했다(문서 본문은 공개하지 않으며, 그 해시와 커밋 식별자는 공개 보충자료에 기록했다. 데이터 및 코드 공개 참조). 수정안 A1은 대조군 채점 전에 기록했고, 문서화된 해석 A1.7은 채점이 시작된 뒤에 기록했다. 모든 채점 코드가 첫 대조군 모델보다 앞서 있었다는 그 문장은 정확하지 않았으며, 그 사실을 기록해 두었다. 이후 단계(D4-D6)는 분석 전에 같은 성격의 둘째 문서에 기록했다. 마지막 단계(D13, 다른 ART 좌위의 파트너)는 그 문서의 한 절과 수정안 A1을 표적의 Logan 히트를 읽기 전에 기록했다(첫 통과 스트림은 이미 시작되어 있었으나 히트 파일은 읽지 않았다고 A1에 적었다). 수정안 A2는 이 단독 체인들에 대한 D3 판정을 철회하고, 해상된 유닛만이 아니라 모든 주 유닛에 Foldseek과 구조 패널 검사를 돌리도록 정한 것이다. A2는 완료된 유닛의 해상도 분석을 읽은 뒤(그 코드는 먼저 커밋했다) Foldseek을 돌리기 전에 썼다. 수정안 A3는 Foldseek과 구조 패널 결과를 읽은 뒤에 쓴 것으로, 탐색적 Pfam 검색을 다루며 패널을 D8처럼 시드 0 모델만이 아니라 주 유닛의 세 시드 전부에 돌렸다는 점도 사후에 기록한다. 두 영역 기준은 사전 지정했지만 세 시드 모델 집합과 그 S2 값은 사후 분석이다. 각 단계의 지정 문서는 산출물을 읽기 전에 커밋했다. 예외는 편차 항목에 열거했다(수정안 D8-A2, A3, D8-D는 앞선 D8 산출물을 읽은 뒤에 기록했고, D8의 세 단계 코드는 결과와 함께 커밋했다). 탐색적 분석은 그렇게 표시했다. 여기서 "사전 지정(pre-registered)"은 외부로 푸시한 적 없는 로컬 git 저장소에 자체 생성 해시와 함께 커밋했다는 뜻이며, 외부 시각 증명은 없다. 읽기 전용 검토 세 건(D3, D4/D5, D6 단계)과 앞선 초고에 대한 검토 한 건은 사람이 아니라 별개의 AI 세션이 수행했다. 이들은 원본 행에서 판정을 다시 도출하고 원본 모델 파일에서 값을 다시 계산했으며, 수정 사항은 반영했다. 이 검토들이 다루지 않은 D13 부분은 읽기 전용 AI 세션 세 건(주장과 근거의 대조, 영한 의미 일치, 한글 자연스러움)과 스크립트 점검(공개하지 않음)으로 확인했다. 모든 핵심 수치는 결과 파일에서 NUMBERS.json으로 계산했다.

서열. P1(QXN67954.1, 661 aa), P2(QXN67952.1, 481 aa)와 짝을 이루는 역전사효소 RT1, RT2는 LPJP1 게놈(GenBank MZ422438.1)에서 가져왔다. 대조군은 Eis(직렬 GNAT, 구조 3R1K [24]), RimI(단일 GNAT, 실험 구조 없음), MshD(직렬 GNAT 유사, 1OZP, 1P0H, 2C27 [25]), Orf14(직렬 GNAT, 2WPW [26])였고, DnaK(4B9Q [38]), UPP(5E38 [39]), GlmU(1HV9 [30])는 GNAT이 아닌 대조군으로 썼다. 서열은 UniProt에서 가져와 사전 지정 문서에 고정했다(Eis P9WFK7, RimI P0A944, MshD P9WJM7, Orf14 Q8KRB5, DnaK P0A6Y8, UPP P9WFF3, GlmU P0ACC7).

다른 ART 좌위의 파트너(D13). 후보는 앞서 공개 어셈블리를 훑은 조사에서 역전사효소 바로 하류의 첫 유전자로 정했다. 프리프린트 보충 Fig. 3C의 좌위와 도면 지문(RT 길이, 파트너 길이, RT-파트너 간격, 반복 간격을 벡터 PDF에서 읽은 값)으로 맞춘 좌위 둘(ART_69, ART_34)이나, 반복 배열 판정이 있는 한 계통(L1)의 서로 다른 접근번호 가운데 좌위의 지문이 도식된 도면과 맞지 않은 것들이 후보였다. 프리프린트 전체 좌위 집합에 포함되지 않았는지는 확인하지 못했다. 도면과의 일치는 접근번호가 아니라 지문에 근거한다. RT 길이와 파트너 길이는 도면의 값과 같고(종결 코돈을 센 값), RT-파트너 간격은 3 nt 차이가 난다. 원래의 탐침은 반복 간격도 비교했지만 여기서는 다시 계산하지 않았다. 사전 지정에 앞서 파트너 ORF가 완전한지 검사했다. ORF 길이가 잔기 수의 세 배에 종결 코돈을 더한 값과 같아야 하고(contig 안에 종결 코돈이 있어야 하고), 메티오닌으로 시작해야 하며, 모호한 잔기가 없어야 한다. 검사와 결과는 모두 코드로 생성했다(결과는 numbers/NUMBERS_D13.json에 있다). 파이프라인은 앞선 단계와 같고, 등록된 변경은 다음과 같다(D13-A1, A2). 표적마다 패밀리 집합을 따로 만들었다(Logan 첫 통과만, $E \leq 1e-5$, 질의 커버리지 ≥ 0.5 , 90% 동일성 중복 제거). 단독 체인만 시드 0-2로 예측했다. Foldseek에는 보정한 분류기와 변경 없는 관문을 그대로 적용했고, 구조 패널 검사는 주 유닛에 사전 지정한 두 영역 기준을 적용한 서술적 판독이다(세 시드 전부를 쓴 범위는 D13-A3에서 사후에 기록했다). 고정된 시드 판정 규칙은 각 단독 체인에 적용하되 표적을 첫 GNAT이 알려지지 않은 파트너로 취급했다. D3 판정표는 공동 접합 라벨을 기준으로 정의했고 이 좌위의 역전사효소 서열은 동결한 표적에 들어 있지 않으므로, 판정은 내리지 않았다. 서버 정렬은 노트북에서 받았다(계산 짖은 서버에 접속할 수 없다). 그 과정에서 표적 서열이 공개 ColabFold 서버로 나갔다. Pfam 클랜 검색(HMM 98개, gathering 문턱, 양성 대조군으로 Eis, RimI, Orf14)은 Foldseek과 패널 결과를 읽은 뒤에 추가했으며 탐색적이다(D13-A3).

구조 예측. ColabFold 1.6.1 [2]로 AlphaFold2-multimer v3 [3,4] (5개 모델 x 3 시드, 3회 recycle, 템플릿 없음, relaxation 없음)와 Boltz-2 v2.2.1 [5] (recycling 3단계, 5 diffusion 샘플 x 3 시드, 커널 없음)를 NVIDIA B300(Blackwell) GPU에서 실행(D12의 Boltz-2는 H100 80 GB)했다. 두 엔진에 같은 a3m을 입력했다. 911잔기인 두 체인 대조군 공동 접합에서 CPU와 GPU를 비교하니, 체인별로 정렬된 부분은 웅스트롬 미만으로 일치했으나 체인 사이 상대 배향은 달랐다(탐색적). Chai-1 [7] (MSA 없음, ESM 임베딩)과 Protenix 0.5.5 [8] (MSA 없음, 대조군을 하나도 해상하지 못해 더 보고하지 않음)는 같은 체인에 대한 탐색적 교차 점검으로 돌렸다. 이들은 정렬이 없는 예측기를 시험할 뿐, 똑같이 깊은 MSA를 쓰는 예측기를 시험하지는 않는다.

서버 정렬과 패밀리 정렬. 서버 정렬은 ColabFold 서버에서 받았다(유일한 외부 서열 전송). 각 서버 정렬에는 질의 서열이 두 번 들어 있어 레코드 수는 두 사본을 모두 포함한다. P1은 고유 정렬 행 6개·고유 서열 6개이며 질의를 제외한 상동체는 5개다. P2는 고유 정렬 행 17개이지만 고유 서열은 16개이고 상동체는 15개다. A0A233VSS0가 서로 다르게 정렬된 두 행에 들어 있기 때문이다. Logan [17]의 50% 단백질 클러스터 대표 서열(공개 버킷 s3://logan-pub, SRA 재분석; 정확한 객체는 ETag 357cde7dd7847a6726500702e2e7125c-8087와 크기 271,342,209,328 B로 식별하며 이후 HEAD 요청에서 같은 값이 재현되었음. 릴리스 이름은 기록하지 않았음)은 익명 HTTPS 범위 다운로드로 스트리밍했다 (3,009,464,616개 레코드). 300-1000 aa 레코드만(110,770,355개, 4%) MMseqs2 [10]로 검색했다(한 번 통과, 민감도 7.5, $E \leq 1e-5$, 질의당 최대 2000 히트, 이어서 첫 통과 히트로 만든 프로파일로 두 번째 통과, 표류한 두 구성원의 프로파일 행은 사용하지 않음). 상동체는 질의 커버리지 ≥ 0.5 와 90% 동일성 중복 제거 기준으로 섰으며, 이 길이 필터는 단편과 1000 aa를 넘는 파트너를 제외한다. 패밀리 F(45개 서열)에는 사람(human) 유래 클러스터 파일의 히트와 프로파일 통과와 히트가 들어 있다. 이를 클러스터 대표 서열의 구성원(공개 비클러스터 단백질 파일)까지 넓힌 것이 F2(784개 구성원 레코드에서 나온 89개 서열)이다. 패밀리 행은 MMseqs2로 질의에 정렬했고(삽입 제거, 질의 열의 25% 미만인 행 제거) 서버 행에 이어 붙였으며, 중첩된 부분집합으로 용량-반응을 구했다. 센서스 MSA(각 체인을 91개 체인 가운데 자기 자신을 뺀 나머지에 정렬, $E \leq 1e-3$, 서버 행 없음)의 행 수는 1-83(중앙값 77)이다.

해상 기준. 슬롯은 141잔기 원도(보폭 10)이다. Tier 1과 센서스에 쓴 기술적 규칙에서는 슬롯의 절반 이상이 세 시드 중 두 시드 이상에서 평균 pLDDT ≥ 70 이면 체인이 해상되었다고 본다. Tier 1의 P1에서는 잔기 484-658을 슬롯에서 제외한다. 이미 GNAT 유사 피크가 알려진 영역이다. 어느 한 엔진에서라도 일치 단위 하나가 이 규칙을 충족하면 그 파트너를 "해상 가능(resolvable)"으로 센다. 고정된 D3 규칙은 음성을 판정하기 전에 첫 GNAT 바깥의 모든 슬롯이 예측기별 하한에 도달하고 슬롯 내 PAE ≤ 10 Å일 것을 추가로 요구한다. 이 규칙의 AlphaFold2 하한(pLDDT 90)은 드문 대조군 데이터에 근거하며 검증된 검출 한계가 아니다.

검출기. GNAT 유사 영역은 Foldseek [9] (213-aa 원도, 보폭 50, afdb50, pdb100, CATH50, $E \leq 1e-3$)로 검출했다. 사전 지정 문서의 수정 분류기(A1.1)를 썼다. 이 분류기는 GNAT 클랜의 GNAT/acetyltransferase로 이름이 붙은 표적, CATH 3.40.630 또는 Pfam 클랜 CL0257 accession이 있는 표적을 받아들이며, 이름에 느슨한 'N-acetyl' 문자열이 들어 있는 표적도 받아들인다. 이 분기가 아래에 기술한 GlimU 위양성의 원인이지만 SINGLE 체인의 판정은 이 분기에 의존하지 않는다. 피크 규칙은 141-213 aa의 정렬, 표적 커버리지 ≥ 0.6 , 서로 다른 표적 10개 이상, 피크 간격 ≥ 100 aa를 요구한다. 로브 분석은 60-140 aa, 커버리지 ≥ 0.4 , 같은 표적 수 문턱을 썼다. 검출기는 어떤 표적도 읽기 전에 대조군으로 관문을 통과했다(통과한 대조군 셀에서 Eis는 피크 두 개, RimI는 하나, GNAT이 아닌 대조군은 없음). 대조군 하나(Eis)의 깊이 사다리에서 시험한 GNAT 폴드 둘은 Neff80 17과 124에서 두 예측기 모두 검출했다. 가장 낮은 두 단계는 각각 Neff80 1이었고, AlphaFold2는 두 폴드 모두 검출하지 못했다. Boltz-2는 일부 폴드 사례를 검출했으며, 일부 시드에서는 두 폴드 모두 검출했다. Neff80 1과 17 사이의 단계는 시험하지 않았다. 깊이를 맞춘 대조군(D7)에서 검출기는 Eis, MshD, Orf14의 대조군-예측기-깊이 판정 12건 모두에서 TANDEM을 냈고(12건 중 12건), DnaK에는 NONE을 냈으며, GNAT이 아닌 아세틸전이효소 GlimU에도 TANDEM을 냈다(분류기가 주석 이름으로 표적을 받아들이므로 생긴 위양성). 대조군의 로브 크기 한계는 120 aa이므로, 검출기가 91-aa 로브를 판정할 수 있다고 보이지는 않았다.

조작적 정의(재계산용). Foldseek. 릴리스 10-941cd33, `easy-search -e 10 --max-seqs 1000 -s 9.5 --alignment-type 2`. 창 PDB는 모델 사슬에서 잘라 1번부터 다시 번호를 매겼고, 세 데이터베이스의 결과를 합쳤다. 수정 분류기(A1.1, 'VA')의 이름 검사는 대문자로 바꾼 표적 헤더만 읽는다(분류군 이름은 쓰지 않으며, 동결 원본은 분류군 이름도 읽었다). 정규식 `ACETYLTRANSFERF\bGNAT\b|GCN5|ACETYLASE|N-ACETYL`(따라서 DEACETYLASE도 일치한다), 문자열 3.40.630, 또는 Pfam clan CL0257에 속한 Pfam-A 패밀리의 accession(사용한 clan 목록에는 69개)이다. 봉우리 규칙에서 정렬 길이는 쿼리 구간($qend - qstart + 1$)이고 표적 coverage는 $(tend - tstart + 1)/\text{표적 길이}$ 이다. 봉우리. 각 정렬의 중점 $(qstart + qend)/2$ 를 25 aa bin에 넣고, 최대 bin 개수의 25% 이상인 bin을 봉우리로 본다(인접 bin은 병합). 봉우리 안에 중점이 있는 고유 표적이 10개 이상이면 유효하다(키: pdb100은 PDB entry id, 그 외는 데이터베이스와 헤더 첫 토큰). 모든 창의 유효 봉우리를 지지도 내림차순, 다음으로 최소 E 순으로 정렬하고, 중심 $(\text{median } qstart + \text{median } qend)/2$ 가 이미 수락한 모든 봉우리에서 100 aa 이상 떨어진 경우에 수락한다. 로브 후보는 span 60-140 aa, 표적 coverage 0.4 이상, 표적 10개 이상이며 중심이 수락한 봉우리에서 90 aa 이상 떨어진 봉우리이다. 사슬의 첫 GNAT는 지지도가 가장 큰 수락 봉우리이다(P1은 중심이 잔기 434-661에 있는 봉우리가 알려진 봉우리이고, 433 이하는 그 바깥이다). **채점하는 모델.** AlphaFold2는 시드마다 가장 높게 순위가 매겨진 모델(pLDDT 기준 ColabFold rank)이며 이 논문에서 'rank-1'이라 부른다. Boltz-2는 시드마다 model_0(신뢰도 점수 최고)이다. 데이터 보충자료의 파일 이름 `rank_001`은 한 실행의 AlphaFold2 모델 15개 전체에 대한 순위이므로 시드별 최상위와 같지 않을 수 있다. **D3 시드 호출.** 영역 R은 사슬에서 첫 봉우리 span(바깥쪽으로 반올림)을 뺀 부분이며 P1은 잔기 484-658도 뺀다. 슬롯(141 aa, 간격 10, 각 연속 구간의 끝에 맞춘 슬롯 하나 추가)은 R에서 141 aa 이상인 연속 구간에만 만들며, 슬롯이 없으면 UNRES이다. 슬롯의 평균 pLDDT가 하한(AlphaFold2 90, Boltz-2 60) 이상이고 슬롯×슬롯 블록의 평균 PAE가 10 Å 이하이면 해상된 것이다(PAE가 없으면 미해상). 시드 호출은 수락 봉우리가 둘 이상이면 POS(P1은 알려진 봉우리 바깥에 하나 이상), POS가 아니고 첫 GNAT를 회수했으며 모든 슬롯이 해상되고 로브 후보가 없고 파이프라인이 끝났으면 NEG_RESOLVED, 로브 후보만 NEG_RESOLVED를 막으면 LOBE, 그 밖에는 UNRES이다. 셀 라벨(파트너, 엔진)은 세 시드 중 둘 이상이 POS이면 'ok', 세 시드 모두 NEG_RESOLVED이면 'not_found', 그 밖에는 'unknown'이다. D3의 파트너 판정은 두 엔진을 합친다(ok/ok와 ok/unknown은 YES, not_found/not_found는 NO, 나머지는 UNKNOWN). D13의 단독 사슬에는 판정을 부여하지 않았다. **센서스 분류.** pLDDT 70 이상인 잔기의 비율이 0.5 이상인 시드가 세 시드 중 둘 이상이면 사슬이 해상된 것이다. 수락 봉우리가 둘 이상인 시드가 둘 이상이면 TANDEM, 정확히 하나이면 SINGLE, 없으면 NONE이고 그 밖은 MIXED이다(발생하지 않았다). **유효 서열 수.** Neff80: 비갭 열이 50% 이상인 행만 남기고(삽입 제거), 각 행의 가중치는 $1/n$ 이다. n은 두 행에서 모두 비갭인 열에서 동일성이 0.8 이상인 행의 수(자기 포함)이다. 쿼리와 모든 서버 행·패밀리 행을 포함한다. **패밀리 중복 제거:** `mmseqs easy-linclust --min-seq-id 0.9 -c 0.8`, 끝의 종결 문자 제거, 쿼리와 동일한 행 제거. **구조 패널.** 150잔기 창, 간격 25, 마지막 창은 사슬 끝에 맞춘다(150잔기 이하인 사슬은 창 하나). 기본 옵션의 TM-align이고 TM 점수는 창으로 정규화한다. 문턱 $t = \text{round}(\text{비GNAT 참조의 최고 점수} + 0.01, 2)$ 이며 자기 자신과 형제 항목은 제외한다. 비GNAT 참조(각 파일의 첫 사슬, 첫 모델): 3CLA, 1HV9, 1LXA, 1OCX, 3CHY, 1TIM, 4AKE, 2TRX, 2CMD, 1LDM, 1FLV. D11의 긴 비GNAT 구조: 1GPB, 1PKN, 1AON, 1ALK, 1BGV, 1PGJ, 2ACE, 1E7P, 1FCB. D12의 음성: 3EH0, 1NDB, 1KRR, 1QMG, 1ECF, 1YP2, 1PJQ, 2AEV, 1QOR, 1HZ4, 2ACE, 1PGJ. '두 영역'은 시작 위치가 150잔기 이상 떨어진 두 창이 모두 t에 도달한 경우이다. 음성 호출을 허용할지 정하는 대조군 보정은 결과 절에 보고했다. 여기 보고한 모든 판정은 슬롯 미해상 때문에 UNKNOWN이며, 이는 그 보정에 의존하지 않는다.

센서스. 91개 체인(P1, P2와 패밀리) 각각에 예측기별로 네 클래스 중 하나를 부여했다. TANDEM(세 시드 중 두 시드 이상에서 피크 두 개), SINGLE(피크 한 개), NONE(해상된 체인에서 피크 없음), UNRESOLVED(위 규칙으로 해상되지 않음)이다.

탐색적 분석. (a2) 행을 섞은 정렬 대조(D10): D5-B 패밀리 정렬에서 모든 패밀리 행의 잔기를 행 안에서 치환하고(gap, 행 수, 서버 행은 유지; 섞은 정렬과 그 결과 모델은 공개 보충자료에 보존) 같은 단독 체인 예측을 했다. (a3) 긴 GNAT 비해당 실험 구조 아홉 개를 D8 연산으로 채점했다(D11). D8이 쓴 패밀리 Boltz-2 모델은 시드 0 모델이고 AlphaFold2 모델은 시드 전체에서 순위 1인 모델이다. (a0) 깊이를 맞춘 대조군(D7): 같은 Logan 검색과 같은 파이프라인을 Eis, MshD, Orf14(직렬 GNAT)와 GlimU, DnaK(GNAT 아님)에 적용했고, 각각 45와 89개 상동체 행(F와 F2의 크기)의 중첩된 무작위 부분집합을 썼다. 또한 P1의 N 말단 원도(잔기 1-200)와 P2의 N 말단 원도(잔기 1-110)로 Logan 검색을 했다. (a1) 패밀리 체인의 metapredict 3.0.2 [18-20] 무질서 예측. (a) TM-align [12] (tGCN5 기준 정규화, 1QSR 체인 A, 시드 0 최상위 모델)으로 tGCN5를 체인 전체에, 이어서 나머지 부분에 정렬하고 대조 패널을 둔 것. (b) GNAT 클랜 CL0257을 쓴 Pfam-A 38.2 [13,14] (HMMER 3.3.2 [16])(D6). 이후(D8-B) InterPro에서 내려받은 그 클랜의 HMM 98개로 hmmscan을 수행했다(gathering 문턱, 전체 Pfam-A 다운로드가 멈췄으므로 클랜 히트만). 이어서 (D8-B2) 같은 검색을 전체 Pfam-A(HMM 30,134개, 운영자 제공)에 대해 서로 다른 GNAT 단백질 일곱 개(항목 여덟 개, Eis는 D7 모델과 3R1K로 각각 한 번씩)와 서로 다른 GNAT 비해당 단백질 열두 개(PDB 기준 1HV9이기도 한 GlimU, DnaK, 다른 PDB 기준 열 개)에 수행했다. (c) 모든 R/Q-x-x-G-x-G/A 구간의 CA RMSD(인덱스별 Kabsch, 28잔기)를 tGCN5 잔기 125-152(beta4, P-loop, alpha3)에 대해 구하고, 대조군 GNAT motif A 구간과 GNAT이 아닌 대조군으로 보정했다. (d) F와 서버 행의 중복.

편차와 정정(모두 프로젝트 기록에 남김): 첫 전체 Pfam 검색(D8-B2)에서는 스크립트의 문자열 치환이 소리 없이 빗나가, 클랜 필터가 적용되지 않은 채 GNAT이 아닌 히트가 클랜 히트로 출력되었다. 출력을 보고 이를 발견해 교정한 스크립트로 반복했다. D8 수정안 A2, A3, D8-D는 앞선 D8 산출물을 읽은 뒤에 기록했다(두 번째 영역 점수 S2와 AlphaFold2 및 원도 반복 실험은 그 산출물에서 동기를 얻었다). S2, 키메라, Pfam 단계의 코드는 결과와 함께 커밋했으며, D8-A2의 지정 문서는 첫 패널 출력을 잘못 기술했다(채점된 모든 체인이 문턱에 도달했다고 쓰였으나 결과 파일에는 문턱 아래인 체인이 둘 있다). 다른 세션이 같은 패밀리를 탐침 검색했지만 사전 지정한 규칙 아래에서 F2 외의 새 클러스터는 추가되지 않았다(D9. 300-1000 aa 후보 83개 중 66개는 $E \leq 1e-5$ 에서 P1 또는 P2에 대한 MMseqs2 히트가 없어 더 조사하지 않았다). Logan 서열 끝의 종결 코돈 문자 때문에 센서스의 첫 실행이 멈췄다(제거 후 재실행). 첫 구성원 확장 실행은 잘못된 인덱스 필드로 구성원을 찾았고(784개 중 724개 누락) 폐기한 뒤 반복했다. 마지막 Foldseek 배치가 끝나기 전에 결과 파일이 채점되어 다시 채점했다. 분석 하나가 튜플 풀기 오류로 한 번 중단되었고 CIF 파서가 체인 하나를 불완전하게 읽은 적이 한 번 있었다(둘 다 결과를 읽기 전에, 또는 같은 레이블로 수정). 첫 단계의 대조군은 보강한 표적과 MSA 출처를 공유하지 않았다. 이는 사전 지정한 단일 출처 규칙에서의 이탈이었고, 규칙의 깊이 구간으로 처리했으며 이후 깊이를 맞춘 대조군(D7)으로 보완했다. 패밀리가 서버 데이터베이스에 없었다는 앞선 진술은 틀렸다(3.2절). 끝으로, 마지막 단계에서 대화 중 정리한 후보 목록에는 contig 끝에서 잘린 파트너가 둘 있었다. 완전성 검사로 이를 알아내 예측 전에 제외했고, L1 계통에서 가장 긴 완전한 파트너도 처음 짐작한 674잔기가 아니라 약 850잔기였다.

3. 결과

3.1 공개 정렬로는 파트너가 해상되지 않는다

ColabFold 서버 정렬(P1 7행, P2 18행, Neff80은 둘 다 3)을 쓰면 파트너 체인은 두 예측기 모두에서 대부분 저신뢰였다. 공동 접힘 평균 pLDDT는 P1이 AlphaFold2에서 35–38, Boltz-2에서 46–47였고, P2는 31–32와 41–42였다. 고정된 규칙은 민감도 변형 9개 모두에서 P1은 UNKNOWN, P2는 UNKNOWN라는 판정이 나왔다. 첫 GNAT 바깥의 슬롯 34–36개 가운데 해상 하한에 도달한 것은 어느 유닛-시드에서도 최대 4개였으므로, 두 번째 GNAT이 없다고 주장할 수 없었다. 통과한 대조군 셀(Eis, RimI)에서는 체인이 높은 신뢰도로 예측되면 파이프라인이 GNAT 폴드를 검출한다.

3.2 Logan에는 서버 정렬이 돌려준 것보다 훨씬 많은 상동체가 있다

서버 정렬은 작았다. UniRef50(릴리스 2026_03)에서는 먼 서열 하나가 나왔고(질의 네 개 중 3개에서 보고), RefSeq 바이러스 단백질과 GenBank 파지 구획에서는 두 파라로그 말고는 상동체가 나오지 않았다. 반면 Logan 50% 단백질 클러스터 대표 서열에는 첫 통과에서 서로 다른 상동체가 43개 있었다(규칙에 따라 센 값: P2 36, P1 잔기 1-483 32, P1 잔기 29, P1 잔기 484-661 20, 패밀리 F 45). 이들은 서버 행과 서로소가 아니다. 패밀리 구성원 45개 중 14개는 서버 행 25개 중 하나와 동일성 90% 이상(커버리지 80%)이고 동일한 것은 없으며, 43개는 어떤 히트든 하나는 있다. 따라서 서버 데이터베이스에도 많은 구성원의 가까운 친척이 들어 있었다. 차이는 패밀리의 존재 여부가 아니라 행의 수와 다양성이다. 서열은 시퀀싱 런 어셈블리에서 나왔으며 원 기술의 좌위가 섞여 있을 수 있다(Logan이 그 연구 자신의 출처이며, 일부 히트 식별자는 이 프로젝트의 다른 곳에 기록된 배열 좌위의 파트너 단백질과 일치하나 검증하지 않았다).

3.3 정렬 깊이를 보충하면 파트너 상당 부분이 해상된다

패밀리 행을 더하자 단독 체인의 평균 pLDDT가 올랐다(Figure 2). AlphaFold2의 P2는 38-47(서버 행만)에서 80-83(60행)으로, P1은 50-54에서 78-79(47행)으로 올랐다. 상승은 $k = 8, 16, 32$ 와 추가한 전체 행에 걸쳐 P2와 AlphaFold2의 P1에서 단조로웠고, Boltz-2의 P1에서는 단조롭지 않았다($k = 32$ 가 $k = 16$ 과 같았다). 패밀리를 두 배로 늘려도(89개 서열) 변화는 작았다. P2 평균은 84.1-84.4(AlphaFold2)로 움직였고 P1 평균은 약간 떨어졌으며, Boltz-2 단독 P1 체인은 더 이상 해상 규칙을 충족하지 못했다. Neff80은 P1이 32에서 43까지, P2가 38에서 47까지만 올랐으므로, 여기서 "포화"는 이 정도로 다양성을 더해서는 더 이상 이득이 없다는 뜻이다. 사전 지정한 기술적 슬롯 규칙으로 보면 P2는 단독과 공동 접힘 모두에서 두 예측기 모두 해상되었고, P1은 경계에 있었다(Boltz-2 공동 접힘에서는 해상되지 않았으며, 고정된 엄격 규칙은 어떤 P1 단위도 충족하지 못했다).

MSA 없는 예측기 Chai-1은 대조군 단백질(Eis, RimI, DnaK, UPP)은 접었으나 파트너는 해상하지 못했다(평균 pLDDT P1 51.6-52.4, P2 52.0-52.7). 이는 앞선 실패가 처음 쓴 두 예측기에만 국한되지 않음을 보여 주지만, 똑같이 깊은 MSA를 쓰는 예측기도 실패하리라는 점은 보여 주지 못한다.

행을 섞은 대조(D10, 어떤 모델도 만들기 전에 기록, Figure 7)로는 상승이 행 수나 정렬 형식에서 비롯될 수 있는지 시험했다. 여기서는 패밀리 행의 행 수, gap, 조성은 그대로 두고 모든 패밀리 행 안에서 잔기를 치환했다. 그러자 단독 체인 평균 pLDDT는 서버 행만 쓴 수준 근처이거나 그 아래에 머물렀다(최대 약 5 pLDDT 위. AlphaFold2 P2는 실제 행에서 82인 데 비해 35, P1은 78인 데 비해 41, Boltz-2 P2는 77인 데 비해 41, P1은 69인 데 비해 49). 따라서 사전 지정한 규칙으로 파트너-엔진 네 셀 모두 내용 특이적이다(all_content_specific = True). 상승은 행 수나 정렬 형식으로 만들어지지 않는다. 이 대조로는 실제 행이 올바른 공진화 정보를 담고 있다고 말할 수 없다. 섞은 행에는 없는 정보를 담고 있다는 것까지만 보일 뿐이다. 섞은 행은 조성은 유지하지만 질의와의 동일성이 낮다(약 7%). 그래서 이 대조는 정렬 내용을 행 수 및 형식과 분리할 뿐, 보존성을 공변이와 분리하지는 못한다. 용량-반응은 체인 두 개와 시드 세 개에 근거한다.

3.4 고정된 판정은 UNKNOWN으로 남고, 해상되지 않은 영역은 N 말단이다

패밀리 정렬을 써도 고정된 D3 규칙은 모든 민감도 변형에서 P1은 UNKNOWN, P2는 UNKNOWN라는 판정이 나왔다(확장한 패밀리도 같다: UNKNOWN / UNKNOWN). 주된 이유는 N 말단 영역(P1의 처음 약 200잔기와 P2의 약 100잔기, 잔기별 프로파일은 Figure 8)이 모든 모델에서 저신뢰로 남았기 때문이다. 특히 Boltz-2의 P1에서는 N 말단 바깥의 일부 슬롯도 하한에 못 미쳤다. 해상된 부분에서 P1에는 C 말단 영역에 GNAT 유사 Foldseek 피크가 정확히 하나 있었고 P2에는 없었으며, 두 예측기와 모든 시드에서 같았다. N 말단은 본질적 무질서로 예측되지 않는다(metapredict: P1 잔기 1-200의 무질서 비율 0.0, P2 잔기 1-100의 무질서 비율 0.0, 패밀리 중앙값 0.0, AlphaFold2 N 말단 pLDDT와의 Spearman 상관 -0.168). N 말단 원도를 질의로 Logan을 검색하니 P2 원도에서는 상동체 1개가 나왔고 P1 원도에서는 하나도 나오지 않았다(MSA는 P2 원도 2행, P1 원도 1행). 원도 모델은 저신뢰로 남았고 Foldseek 피크도 없었다. 따라서 이 영역은 고아 영역이나 다름없다. 서열상으로는 정렬되어 있으나 검색한 자원에서 검출 가능한 상동체가 없다. 더 많은 행을 얻을 수 없었으므로 이 영역의 폴드에 대해서는 아무것도 말할 수 없다.

3.5 패밀리의 해상된 체인 중 수락된 GNAT 유사 Foldseek 피크가 두 개인 체인은 없다

패밀리 체인 91개 전체에서(Figure 3) 체인-엔진-시드 예측 546건 중 받아들여진 피크가 1개를 넘는 경우는 없었다. 450-750 aa 체인($n = 70$)에서 AlphaFold2는 SINGLE 34, NONE 32, UNRESOLVED 4를 냈고, Boltz-2는 SINGLE 34, NONE 17, UNRESOLVED 19를 냈다(모든 길이: UNRESOLVED 6, 24). 두 예측기는 이 체인 중 2개에서 판정이 달랐다.

이 깊이에서 검출기는 전체 크기 직렬 GNAT에 대해 충분하다. 상동체 행 45개 또는 89개에서 직렬 GNAT 대조군 Eis, MshD, Orf14는 모든 판정에서 TANDEM이었고(세 시드 중 두 시드 이상, 89행 AlphaFold2의 MshD는 한 시드에서 피크가 하나), DnaK는 NONE이었다(2절). 따라서 TANDEM 체인이 없는 것은 이 행 수에서 전체 크기 직렬 GNAT에 대한 민감도 손실로 설명되지 않는다. 대조군은 행 수가 패밀리와 같지만 잔기 1-150의 60-93%가 덮이는 반면, 패밀리의 NONE 체인에서는 커버리지가 훨씬 낮고, 검출기가 원 기술의 91잔기 두 번째 영역을 판정할 수 있다고 보이지도 않았다. 같은 대조군에서 검출기가 과대 판정을 하기도 했으므로(GlmU), SINGLE 체인에서 나온 것은 아세틸전이효소 유사 Foldseek 피크다. 아래의 TM-align 결과가 이를 GNAT 유사로 뒷받침하지만 확정하지는 않는다. 그러므로 "TANDEM 없음"이라는 진술은 해상된 체인 안에서만 성립한다(70개 중 AlphaFold2 66개, Boltz-2 51개).

단일 피크는 C 말단 쪽에 있다(단일 판정 100건 전체에서 중심이 체인 길이의 0.78-0.86, 중앙값 0.84). P1은 SINGLE 체인이고(그 피크는 이미 GNAT 유사로 알려진 영역이다) P2는 두 예측기 모두에서 NONE 체인이다. NONE 쪽은 일부가 측정되지 않은 것이다. AlphaFold2에서 잔기 1-150의 평균 pLDDT 중앙값은 NONE 체인에서 47, SINGLE 체인에서 89이고, 그 잔기의 MSA 커버리지 중앙값은 NONE 체인 0.17, SINGLE 체인 0.37이다. 로브 분석(60-140 aa)은 어떤 체인에서도 두 시드 이상에서 지지되는 로브 후보를 찾지 못했지만(AlphaFold2 0개, Boltz-2 0개 체인), 대조군의 로브 크기 한계가 120 aa이므로 이것이 91잔기 로브를 배제하지는 않는다.

3.6 TM-align 연산과 구조 검색 검출기는 일치하지 않는다

원 기술의 TM-align 연산에서 첫 영역의 TM은 SINGLE 체인에서 0.57, NONE 체인에서 0.58였고, 나머지 부분의 TM은 SINGLE 체인에서 0.55, NONE 체인에서 0.39였다(Figure 4). SINGLE 체인 50개 중 29개에서 Foldseek 피크가 TM-align의 "두 번째" 구간 안에 있으므로, 이 쌍은 Foldseek가 확인한 C 말단 영역과 Foldseek가 판정하지 않는 상류 영역으로 이루어진다. 이 상류 영역은 NONE 체인에서도 TM 0.57-0.58을 보이므로 tGCN5와의 유사성은 특이적이지 않다. 대조군 GNAT 단백질은 첫 영역 0.73(Eis), 나머지 부분 0.54를 보였고, GNAT이 아닌 대조군은 첫 영역에서 0.32-0.35 범위의 값을 보였으며 패밀리 값은 그 사이에 놓인다. 따라서 단일 기준에 대한 TM-align은 여기서 도메인 수를 결정하지 못한다.

3.7 패밀리 체인 중 서열 프로파일 GNAT 서명이나 motif A 골격을 가진 것은 없다

gathering 문턱(0개 체인)이나 $E \leq 1e-3$ (0개 체인)에서 Pfam GNAT 클랜(CL0257)에 히트한 체인은 없었다. 첫 실행에는 양성 대조군이 없었고, Foldseek가 확인한 C 말단 영역도 음성이었다. 이후의 대조군 실행(3.8절)에서는 GNAT 대조군 8개 중 8개에서 클랜 히트가 나왔다. GNAT 비해당 대조군에서는 없었고 패밀리에서도 여전히 없었다(91개 중 0개). 그러므로 이 음성은 정보를 담은 음성이다. 패밀리에는 Pfam GNAT 서명이 없다. 전체 Pfam 반복(D8-B2, 경쟁하는 패밀리 존재)은 같은 클랜 개수를 주었고(GNAT 대조군 8개 중 8개, 패밀리 체인 0개), 패밀리 체인 91개 중 88개는 gathering 문턱에서 어떤 Pfam 패밀리에도 히트하지 않았다. 나머지 체인은 서로 무관한 단일 패밀리(RVT_1, Rotamase, WD40, PFU)에 히트했으므로, 패밀리는 서열 프로파일로는 대체로 주석이 없다.

R/Q-x-x-G-x-G/A 구간은 27개 체인에서 나타났으나(31개 구간, 조성으로 기대되는 것보다 많음) 골격이 tGCN5 P-loop 구간에 맞지 않았다(CA RMSD 4.8-5.0-9.0 Å, 중앙값 5.0 Å). 대조군 GNAT 단백질의 motif A 구간(0.4-2.3 Å)과 대비된다. 패밀리 값은 GNAT 비해당 구간의 하단과 겹치는데(최솟값 5.0 Å), 이는 맞지 않는 골격에 대한 이 측정의 하한일 수 있다. 이 검색은 R 또는 Q로 시작하는 구간만 다룬다. 대조군의 진짜 GNAT motif A 구간 열 개 중 두 개(RMSD 2.2-2.3 Å, MshD 첫 도메인)도 2 Å 선 위에 있으므로 사전 지정한 문턱은 깨끗한 분리 기준이 아니며, 사전 지정한 motif 맥락 검사는 자체 양성 대조군에서 실패했다. 센서스 검토에서 지정한 반복 R-K-G-G-x-A 구간은 구조 패널 검사가 GNAT 유사로 표시하는 영역에 있다. 따라서 패밀리의 GNAT 유사성은 많아야 정석적 motif A가 없는, 폴드 수준의 먼 유사성이다.

3.8 이름을 쓰지 않는 구조 패널 검사는 GNAT 유사 원도를 찾는다(두 영역 기준은 사전 지정, 두 번째 영역 점수 S2는 탐색적)

기준은 실험 GNAT 구조 다섯 개(tGCN5 1QSR [21], Eis 3R1K [24], 효모 Gcn5 1YGH [27], 세로토닌 N-acetyltransferase 1CJW [28], 1BO4 [29]; PDB [31,32])와 GNAT 비해당 구조 열한 개였다. 체인의 모든 150잔기 원도(보폭 25)를 GNAT 패넬에 TM-align으로 정렬했다. 원도 점수는 원도로 정규화한 TM-score를 최적 기준에 대해 취한 값이며 질의 자신과 그 형제는 제외하고, 체인 점수(chain score)는 최고 원도 점수이다. 문턱 $t = 0.46$ 은 기준 구조만으로 정했다(GNAT 비해당 체인 점수 최댓값 0.45, leave-one-out GNAT 점수 0.71-0.75, 민감도 1.00). 패밀리가 예측 모델 이루어져 있으므로 깊이를 맞춘 예측 대조군에도 같은 연산을 적용했다. Eis, MshD, Orf14는 0.69-0.72, GlmU와 DnaK는 0.35-0.38를 얻었으므로 t 는 그대로 옮겨 쓸 수 있었다. 평균 pLDDT ≥ 70 인 원도가 있는 패밀리 체인 88개(Boltz-2, 시드 0) 중 86개가 t 에 도달했고, 여기에는 Foldseek NONE 체인 17개 전부가 포함된다(최소 점수 0.59). 체인 점수 중앙값은 0.62로 GNAT 기준의 0.71-0.75보다 낮았다. 사전 지정한 "두 영역" 규칙(두 번째 영역 점수 S2를 도입하기 전에 적용)은 91개 체인 중 56개가 충족했다.

두 번째 영역 점수 S2는 최고 원도에서 150잔기 이상 떨어진 원도 가운데 점수가 가장 높은 것의 값이다. S2를 계산할 수 있는 체인 71개 중 55개에서 주장 문턱(0.46, t 와 예측된 GNAT 비해당 S2 최댓값 중 큰 값)을 넘었다(전체 체인 91개 중 55개 = 60%, 전체 체인 비율의 실행 부트스트랩 95% 구간 51%-70%). 이는 SINGLE 체인 50개 중 50개, NONE 체인 3개, UNRESOLVED 체인 2개였다. P1에서 t 를 넘는 원도는 잔기 276-351과 451-512에서 시작하며, Foldseek가 확인한 C 말단 영역의 상류 블록이다. 체인 71개에서 두 번째 원도를 계산할 수 있었고, 나머지 체인에는 평균 pLDDT ≥ 70 인 두 번째 원도가 없으므로, 위의 71개 중 55개(그리고 전체 91개 중 55개, 60%)는 전체 체인에서의 유병률이 아니다. 거의 모든 SINGLE 체인(50개 중 49개에서 서로 200잔기 이내)에서 두 원도는 인접한 타일이며(둘 사이에 점수 골이 있다), 따라서 이 데이터로는 두 도메인과 하나의 긴 GNAT 유사 영역을 구별할 수 없다.

사전 지정한 강건성 기준은 충족했으나 비율 기준은 관대하다. 또 AlphaFold2와 Boltz-2가 같은 패밀리 정렬을 썼으므로 이 반복은 정렬과 독립이 아니다. AlphaFold2 순위 1 모델에서는 62%였고, 130잔기와 100잔기 원도에서는 68%와 81%였다(100잔기에서 NONE 체인이 13개인 데 비해 150잔기에서는 3개이므로 원도가 줄수록 특이도가 떨어진다). tGCN5와 DnaK의 키메라(800잔기)는 S2가 0.39로 문턱 아래였으나, 이 조건은 구성상 저절로 충족되므로(tGCN5가 최소 원도 간격의 두 배보다 짧다) 특이도의 증거가 되지 못한다.

음성 대조를 두 가지 더 수행했다. S2 규칙의 추가 음성 대조(D11, 실행 전에 기록)로서 긴 GNAT 비해당 실험 구조 9개(449-823 aa, 서로 다른 패밀리의 단백질 아홉 개, 그중 두 개는 Rossmann 탈수소효소)를 같은 연산과 문턱으로 채점했으며, t 에 도달한 것은 없었다(최대 체인 점수 0.43, 최대 S2 0.43, 9개 중 0개가 문턱 초과). 이들은 예측 구조가 아니라 실험 구조이다. 또 다른 대조(D12, 모델 생성 전에 기록, ColabFold 서버 정렬을 쓴 Boltz-2 모델)는 GNAT 비해당 단백질 12개(203-596 aa)를 예측했으며, 그중에는 GNAT이 아닌 아실전이효소 LpxD, carnitine acetyltransferase, galactoside acetyltransferase와 다른 알파/베타 효소 아홉 개가 있다. t 에 도달한 것은 없었다(최대 체인 점수 0.45, 최대 S2 0.41, 단백질 3개(203-374 aa)는 최고 원도에서 150잔기 이상 떨어진 원도가 없어 S2가 존재하지 않았고, 나머지 중 0개가 문턱을 넘었다). 이들의 모델은 신뢰도가 높아(평균 pLDDT 91-97) 패밀리보다 더 많은 원도가 채점되었다. 이들의 정렬은 패밀리 깊이가 아니라 깊은 서버 정렬이며, 단백질 하나(acetohydroxy-acid isomeroreductase)는 $t = 0.46$ 에 대해 0.45를 얻었다. 점수는 Figure 6에 있다. 클랜 HMM 대조(3.7절)는 서열 수준의 대응물이다. GNAT 대조군은 히트하고 패밀리는 히트하지 않는다.

이 검사는 150잔기 원도와 작은 GNAT 패넬의 유사도를 측정한다. 원도는 긴 기준 구조와 부분적으로 맞고, 도메인 경계는 이 검사로 추론하지 않는다. Foldseek 분류기의 이름 규칙을 쓰지 않으므로 GlmU 위양성의 영향을 받지 않지만, 자체 특이도는 위의 대조군으로만 보여 준다.

3.9 서열이 완전한 다른 ART 좌위의 파트너(사전 지정한 D13)

P2는 프리프린트 Fig. 2B에서 LPJP1 좌위(ART_11)의 파트너로 그려진 유전자이다. 도면에는 RT 1365 nt, 파트너 1446 nt, 간격 -20 nt(겹침)로 나와 있다. GenBank MZ422438.1에서는 RT2가 1365 nt, P2가 1446 nt이고 같은 겹침을 보인다. 도면에서 그 상류에 그려진 1986 nt 유전자가 P1이다. 우리가 벡터 PDF에서 읽은 바로는 보충 Fig. 3C가 ART_11, ART_69, ART_34의 파트너를 모두 Type I 부류로 그렸지만, 본문이 좌위마다 그렇게 적은 것은 아니고 GNAT 분석은 다른

좌위(ART_57)에서 이루어졌다. 그래서 서열을 완전히 확보한 다른 좌위의 파트너를 더했다. 후보 6개 중 4개가 완전성 검사를 통과했고, 나머지 둘(473잔기와 674잔기)은 contig 끝에서 잘려 있어 예측 전에 제외했다(표 1, numbers/NUMBERS_D13.json의 partner_qc 항목).

파트너	좌위의 근거	길이 (aa)	서버 행 (고유)	패밀리 행	Neff80	AlphaFold2 평균 pLDDT	Boltz-2 평균 pLDDT	해상 기준 충족 (AlphaFold2 / Boltz- 2)
ART_69 파트너	도면 지문(보충 Fig. 3C)	594	6	35	33	88.4–88.8	80.0–80.1	strict / strict
ART_34 파트너	도면 지문(보충 Fig. 3C)	890	6	18	13	78.1–81.1	72.2–73.0	strict / strict
L1 파트너 (ERR10896470)	서로 다른 접근번호; 도시된 도면 지문 불일치	852	6	15	13	67.6–68.9	58.0–61.1	RESOLVED_70 / no
L1 파트너 (SRR26805139)	서로 다른 접근번호; 도시된 도면 지문 불일치	850	5	14	9	68.0–69.5	59.0–60.2	RESOLVED_70 / no

표 1. D13에서 분석한 다른 ART 좌위의 파트너. 서버 행은 ColabFold 정렬의 고유 서열 수이며 질의는 한 번만 센다. 패밀리 행은 90% 동일성 중복 제거 뒤 남은 Logan 상동체이다. Neff80은 서버 행과 패밀리 행을 합친 정렬의 값이다. 평균 pLDDT는 체인 전체를 평균한 값이고, 표에는 세 시드의 범위를 적었다. "strict"는 141잔기 슬롯의 절반 이상이 평균 pLDDT 90 이상(AlphaFold2) 또는 60 이상(Boltz-2)이면서 슬롯 안 PAE가 10 Å 이하인 경우가 세 시드 중 두 시드 이상에서 성립한다는 뜻이다(D4 단계의 등록 규칙). "RESOLVED_70"은 2절의 기술적 슬롯 규칙만 충족했다는 뜻이고, "no"는 둘 다 충족하지 못했다는 뜻이다. 아래의 고정된 시드 판정 규칙은 이보다 엄격하다(첫 피크 바깥의 모든 슬롯을 요구한다). 이 파트너들에는 서버 행만 넣은 예측을 돌리지 않았으므로, 행을 더한 효과는 k 사다리와 섞은 행 대조로 보였다.

서버 정렬의 고유 행은 5-6개였고, Logan 검색은 파트너마다 패밀리 행 14-35개를 돌려주었다. LPJP1 패밀리의 45개보다 적다. 서버 행과 패밀리 행을 함께 넣으면 네 파트너 모두 사전 지정한 Tier 1 기준을 충족했다. 이 기준은 어느 한 엔진에서라도 규칙을 충족하면 해상 가능으로 세며, L1 파트너 둘은 AlphaFold2에서만 충족했다(표 1). 원 기술이 말한 크기 정도인 ART_69 파트너는 AlphaFold2 평균 pLDDT 88.4–88.8에 이르렀다. 141잔기 슬롯의 100%가 모든 시드에서 70을 넘었고, 시드 0 모델에서는 잔기의 88%가 pLDDT 70 이상이었다(Figure 10). 시험한 엔진과 설정에서 P1과 P2의 미해상 영역으로 남았던 N 말단 쪽도 여기서는 대체로 신뢰도가 높다. Foldseek 피크가 시작되기 전 잔기 중 70 미만인 것은 18%이고, 피크 이후에는 하나도 없다. 더 긴 세 파트너는 앞부분의 신뢰도가 훨씬 낮았다. 피크 앞 잔기 중 70 미만인 비율은 ART_34 파트너 37%, L1 파트너 57%와 64%이고, 피크 이후에는 각각 5%, 20%, 11%이다.

strict 기준을 두 엔진에서 모두 충족한 것은 ART_69와 ART_34 파트너뿐이었고, L1 파트너 둘은 AlphaFold2에서 기술적 규칙만 충족했다. 그래도 얻은 이득은 정렬 내용에 따른 것이었다. 패밀리 행의 잔기를 행 안에서 섞은 정렬(3.3절과 같은 조작)을 넣으면 네 파트너의 AlphaFold2 평균 pLDDT는 37-49, Boltz-2는 41-57로 떨어졌다. 실제 행은 섞은 행보다 AlphaFold2에서 최소 24.1, Boltz-2에서 최소 12.5 pLDDT 단위 높았다. ART_69 파트너의 AlphaFold2 평균 pLDDT는 패밀리 행 수가 늘수록 올랐다(k = 8, 16, 32, 전체 순으로 76.4–79.9, 81.7–85.1, 87.5–88.1, 88.4–88.8. Figure 9).

세 판독은 이번에도 서로 다른 답을 냈다. Foldseek(보정한 이름 기반 분류기, 변경 없는 관문. GlimU를 과대 판정한다)은 시드 판정 24건 모두에서 GNAT 유사 피크를 정확히 하나 받아들였고, 피크는 체인의 C 말단 쪽에 있었다(AlphaFold2 시드 0 구간: ART_69 파트너 405–557, ART_34 파트너 676–833, L1 파트너 641–805와 639–795). Pfam GNAT 클랜 모델(탐색적)은 어느 파트너에서도 히트가 없었고, 같은 모델이 양성 대조군 세 개에서는 히트했다. 구조 패널 검사(t = 0.46)는 모델 24개 중 24개에서 t를 넘는, 서로 겹치지 않는 원도 두 개를 찾았다(두 번째 영역 점수 S2 0.57–0.63). 원도 시작 위치는 서로 인접한 두 무리로 나뉘며(ART_69 파트너: 201–276과 376–445), 두 번째 무리에 Foldseek 피크가 들어 있다(Figure 10). 두 영역 기준은 사전 지정한 D8 규칙이지만, 세 시드 모델 집합과 S2 값은 사후 분석이다(D13-A3). 맞닿은 타일만으로는 도메인 둘과 긴 GNAT 유사 영역 하나를 구별할 수 없다(3.8절). 고정된 시드 판정 규칙은 24건 모두 UNRES였고(첫 피크 바깥의 체인 부분이 엄격한 하한이나 PAE 한계에 못 미쳤다) 셀 라벨 8개는 모두 unknown이었다. D3 판정표의 판정은 내리지 않았다.

네 파트너 중 셋은 원 기술의 "약 600잔기"보다 길고 패밀리 집합도 작다. 또 체인이 길수록 구조 패널 검사에서 문턱을 넘을 수 있는 원도가 많아진다. 이 검사의 특이도는 긴 GNAT 비해당 실험 구조 아홉 개(3.8절)로 확인했을 뿐, 이 길이들에 대해서는 확인하지 않았다. 또 네 파트너는 세 계통이다(L1 파트너 둘은 서로 가깝다). P1의 친척으로 찾은 것이며, 시드와 엔진이 파트너당 하나의 정렬을 공유하므로 모델 24개 중 24개는 서로 독립된 관찰이 아니다. 91개 체인 패밀리와 겹치는지는 확인하지 않았다. Eis 깊이 사다리에서 폴드 사례를 모두 검출한 것은 Neff80 17과 124에서였고, 깊이를 맞춘 대조군은 상동체 45-89행을 썼다. 추가 파트너 4개 중 3개는 Neff80 9-13, 패밀리 14-18행으로 시험하지 않은 깊이 구간에 있다. 이 파트너에서 Foldseek 피크가 정확히 하나라는 결과는 두 번째 전체 크기 영역의 부재를 입증한 음성이 아니다. 그래서 결과는 LPJP1 패밀리와 같고, 차이는 하나다. P1과 P2에서 해상되지 않은 N 말단은 이 패밀리의 모든 파트너가 가진 성질이 아니다.

4. 논의

방법론적으로 얻은 결론은 둘이다. 해상되지 않은 파트너 모델은 대체로 정렬 행이 너무 적은 탓이었고, 공개 메타게놈 자원에는 서버 정렬이 돌려준 것보다 훨씬 많은 상동체가 있었다. 저신뢰 모델에서 내린 예측 기반의 "두 번째 도메인 없음"은 여기서 아무 정보도 주지 못했을 것이다. 사전 지정한 규칙은 올바르게 UNKNOWN 판정을 냈다. 행 수에 따른 용량-반응(Figure 2)이 이를 뒷받침하는 증거이다. 그 한계는 3.3절에 밝혔다. 체인 두 개, 보존성을 공변이와 구별하지 못하는 행 섞기 대조, 시험한 엔진과 설정에서 미해상으로 남은 N 말단, 정렬 없는 예측기를 시험하는 Chai-1이다.

원 질문으로 돌아오면, 해상된 패밀리 구성원 가운데 GNAT 유사 Foldseek 피크 두 개가 재현된 것은 하나도 없다. 탐색적 두 번째 영역 분석에서는 이름을 쓰지 않는 구조 패널 검사(3.8절)로 SINGLE 클래스 체인 50개 전부에서 서로 겹치지 않는 GNAT 유사 원도 두 개가 나온다. 다만 대부분 맞닿은 타일이어서 도메인 둘과 긴 영역 하나를 구별하지 못한다. 문턱 설정에 사용한 기준 구조는 독립적인 음성 검증 자료가 아니다. 추가 시험에서는 긴 실험 GNAT 비해당 구조 9개와 예측된 GNAT 비해당 단백질 12개가 문턱 아래에 머물렀다. 두 집합은 서로 다른 대조 자료이며, 서로 독립적인 고유 단백질의 합계가 아니다. 단백질 둘이 실험 구조와 예측 구조로 두 집합에 함께 들어 있다. 예측된 GNAT 비해당 단백질의 최고 체인 점수는 0.45였고 문턱 t는 0.46였다. 한편 원 기술이 두 번째 영역으로 보고한 로브는 우리 검출기가 판정할 수 있다고 보인 크기 아래에 있다. 이 결과는 원 기술을 반박하지 않는다. 거기서 쓴 방법인 TM-align은 C 말단에 피크가 있는 체인에서 두 번째 영역을 돌려준다(특이적이지는 않지만). N 말단은 저신뢰였고, 원 정렬은 우리가 재현할 수 없었던 좌위 쌍 패밀리 블록이었다. 지금으로서 가장 무리 없는 해석은 이렇다. 두 번째 GNAT 유사 영역이 있더라도 서열 프로파일, motif A 구조, 구조 검색 어느 것으로도 정석적 GNAT은 아닐 것이며, 그것을 GNAT 유사라고 부를 수 있느냐는 폴드를 비교하는 데 쓴 방법에 달려 있다.

서열을 확보한 다른 좌위의 파트너(3.9절)도 같은 방향을 가리킨다. 단백질을 네 개 더 늘렸는데도 세 판독은 이번에도 어긋났고, 원 기술이 말한 크기 정도인 파트너(ART_69, 594잔기)는 체인 대부분이 높은 신뢰도로 예측되었다. 이는 해상되지 않은 N 말단이 이 패밀리의 모든 파트너가 가진 성질은 아님을 보여 줄 뿐, P1과 P2가 왜 다른지는 설명하지 못한다. ART_69 파트너의 정렬이 더 깊은 것도 아니고(Neff80 33, 패밀리 행을 넣은 P1은 32, P2는 38), 더 긴 세 파트너는 앞부분이 여전히 대체로 저신뢰이기 때문이다. Foldseek 피크 주변이 두 번째 GNAT의 일부인지도 이것으로 가려지지 않는다. 구조 패널이 GNAT 유사로 부르는 원도 무리는 피크 상류까지 뻗어 있는데, 그 구간은 신뢰도 높게 예측되면서도 Pfam 서명도 Foldseek 피크도 없다. 세 판독은 저마다 알려진 한계가 있어서(전체 크기 GNAT에서만 검증된 검출기, 서열이 많이 달라진 폴드를 놓칠 수 있는 클렌 모델, 도메인 경계가 없는 원도 점수) 합쳐도 서열이 많이 달라진 GNAT 유사 폴드와 다른 알파/베타 폴드를 구별하지 못한다.

한계. 이 체인은 시퀀싱 런 어셈블리에서 서열 검색으로 얻은 상동체다(키메라나 오염 가능성이 있고, 첫 통과 구성원 45개 중 네 개는 경계선상의 히트 하나로 들어왔다). 독립된 좌위가 아니라 한 패밀리의 클러스터 대표 서열이므로 RT 파트너와 게놈 맥락은 알 수 없다. 센서스 MSA는 패밀리 자체로 만들었고, P1과 P2를 다룰 때는 앞선 단계의 결과를 가리지 않았다. 주 분석의 구조는 시험한 두 엔진의 예측이다. 미해상 영역과 검출기 결과에 관한 서술은 이 모델과 설정에 한정하며 모든 예측기로 확대하지 않는다. AlphaFold 3는 평가하지 않았고, 더 높은 신뢰도의 모델이나 다른 구조 판독 결과를 낼 가능성이 있다. 그 가능성이 관찰한 예측기 내부의 정렬 비교를 무효로 만들지는 않지만, 다른 예측기로 해상할 수 있는 범위에 관한 주장은 제한된다. AlphaFold 3를 평가하려면 해당 예측기의 대조군을 별도로 확인해야 하며, 예측기별 신뢰도 하한이 그대로 옮겨진다고 가정하지 않는다. 고정된 규칙의 AlphaFold2 해상 하한은 검증된 한계가 아니며, "GNAT 없음" 쪽의 일부는 N

말단 영역이 잘 예측되지 않아 측정되지 않았다. 검출기의 표적 커버리지 규칙은 큰 다도메인 표적 안의 GNAT을 가릴 수 있고, 피크 간격 규칙은 100 aa보다 가까운 직렬 도메인을 합친다(한 엔진에서 MshD의 두 GNAT을 합쳤다). TM-align은 시드 0 모델만 썼다. 첫 단계 대조군은 보강한 표적과 MSA 출처를 공유하지 않았다(D7의 깊이를 맞춘 대조군은 공유했고 이름 기반 분류기의 위양성을 드러냈다). motif 검색은 R 또는 Q로 시작하는 구간만 다루며, 전체 Pfam 검색은 gathering 문턱만 썼고 서열은 모델 파일에서 읽었다. 구조 패널 검사는 150잔기 원도를 GNAT 기준 다섯 개에 대해 쓰며(원도가 작을수록 특이도가 떨어진다), S2 점수는 첫 패널 결과를 읽은 뒤에 추가했다(탐색적, 계산 전에 기록). 음성 대조군은 짧은 실험 구조 열한 개와 긴 실험 구조 아홉 개, 예측된 GNAT 비해당 단백질 열두 개(D12), 예측 단백질 둘(GlmU, DnaK, 각각 두 깊이), 그리고 구성상 실패할 수 없는 키메라 하나이며, D12의 예측 음성 열두 개 중 GNAT이 아닌 아실전이효소는 세 개뿐이다. 마지막 단계에는 그 단계만의 한계가 더 있다. 파트너가 네 개뿐이고 단독 체인만 예측했으며, 그중 셋은 우리가 쓴 450-750 aa 대역 밖이다(프리프린트는 '약 600잔기'라 쓰면서도 890잔기인 ART_34 파트너를 Type I로 그린다). 둘은 같은 계통이고, 패밀리 집합은 14-35행에 그친다. 좌위는 접근번호가 아니라 도면 지문으로 맞췄거나(ART_69, ART_34) 도시된 도면 지문과 맞지 않는 서로 다른 접근번호에서 가져왔고(L1; 프리프린트 전체 좌위 집합에 없는지는 확인하지 못함), Pfam 검색은 클랜 HMM만 썼으며 다른 판독을 읽은 뒤에 추가했다. Eis 깊이 사다리에서 폴드 사례를 모두 검출한 것은 Neff80 17과 124에서였고, 깊이를 맞춘 대조군은 상동체 45-89행을 썼다. 추가 파트너 4개 중 3개는 Neff80 9-13, 패밀리 14-18행으로 시험하지 않은 깊이 구간에 있다. 이 파트너에서 Foldseek 피크가 정확히 하나라는 결과는 두 번째 전체 크기 영역의 부재를 입증한 음성이 아니다. 잘려서 제외한 파트너 둘에서 보듯, 파트너 ORF는 쓰기 전에 완전한지 검사해야 한다. 효소 활성이나 acetyl-CoA 결합에 대해서는 어떤 진술도 하지 않는다. 직접 검증에는 실험(아세틸전이효소 활성 측정), GNAT 기준을 여럿 쓰는 추가 구조 검색 방법, 또는 원 기술의 패밀리 정렬이 필요하다.

5. 선언

AI 보조 사용. 이 연구는 AI3 Discovery가 개발한 과학적 발견 하네스를 운영자의 지시 아래 자율적 코딩·분석 에이전트로 작동한 AI 코딩 보조(Claude, Anthropic)와 함께 운용하여 수행했다. 하네스와 보조는 분석·작업 흐름 스크립트의 작성과 실행, 계산 장비에서의 구조 예측과 검색의 조직, 사전에 정해 둔 판정 규칙의 적용, 원고의 초안 작성과 수정, 그리고 결과 파일과 주장의 대조, 영한 일치 점검, AI 에이전트의 적대적 검토를 포함한 자동 검토에 쓰였다. 이 검토들은 AI가 수행했으며 사람의 독립적인 동료 심사가 아니다. 하네스 자체는 공개하지 않는다. 원고의 내용에 대한 책임은 저자에게 있다. **이해충돌 / 공개.** 시험 대상 연구(Yoon et al. [1])는 Anthropic 소속 직원들이 저술했고, 여기서 쓴 AI 시스템(Claude)은 Anthropic의 제품이다. 저자는 Anthropic이나 Yoon et al.의 저자들과 소속 관계 또는 연구 협업 관계가 없다. 앞서 교신저자들에게 좌위 접근번호와 파트너 서열을 요청하는 메일을 보냈으나 회신이나 비공개 데이터는 받지 못했다. 이번 분석에는 공개 자료를 사용했다. 이 연구에서 '독립'이라고 표시한 검토는 같은 모델 계열의 AI 세션이 수행했으며 사람의 동료 심사라는 의미에서 독립적이지 않다. **연구비.** 없음. 연구는 AI3 Discovery의 단독 연구자가 외부 연구비 없이 수행했다. **AlphaFold 3**는 당시 이 연구에 필요한 공식 모델 가중치를 확보하지 못해 사용하지 않았다. 공개 가중치의 접근에는 코드 라이선스와 별개로 기관 및 사용 목적에 따른 자격 조건이 적용된다. 이는 연구 당시 사용한 접근 경로에 관한 설명이며, 모든 상용 이용 경로가 없다는 뜻은 아니다. **데이터 및 코드 공개.** 이번 정정 원고와 그림은 [AI3 Discovery](#)를 통해 공개한다. 보고된 결과를 검산·재도출할 수 있는 데이터는 [Zenodo](#)에 CC BY 4.0으로 공개했다. 데이터 보충자료에는 단백질 서열, 입력 패밀리 정렬, 예측 모델, Foldseek 히트 표, 단위별 결과 표, 계산한 수치와 파일별 SHA-256 목록이 들어 있다. 본문에 적은 파일 이름(예: numbers/NUMBERS_D13.json, integrity/SHA256SUMS)은 이 보충자료 안의 경로다. 분할 압축파일을 다시 합치고 해시를 확인하는 방법은 보충자료의 README에 적었다. Zenodo 레코드에 함께 들어 있는 원고 PDF는 2026-10-03 판이며, AI3 Discovery 사이트의 최신 정정 원고가 이를 대체한다. 이번 수정은 서술과 Figure 1 배치에 한정되며, 공개한 연구 데이터 파일은 바뀌지 않았다. 분석·채점 코드, 보정한 문턱값 파일, 탐침 모델, 하네스 모듈, 사전 지정 문서의 본문은 공개하지 않는다. 대신 데이터 레코드에는 사전 지정 문서의 해시·크기·로컬 git 커밋 식별자와 비공개 파일 범주별 통합 해시를 무결성 기록으로 담았다. 이는 외부 시각 증명을 갖춘 사전 등록은 아니다. 공개 입력은 방법 절에 밝혔고, 수치·표 파일은 Logan 객체를 ETag와 크기로 식별한다. 데이터 문의를 discovery@ai3.kr로 보낼 수 있다. **참고문헌.** 프리프린트는 목록에 그렇게 표시했으며, 항목 1이 시험 대상이 된 연구이다.

1. Yoon PH, Athukoralage JS, Ameisen E, Kauderer-Abrams E, Perry NT, Durrant MG (2026). Autonomous AI agents discover reverse transcriptases with tandem repeat arrays. Manuscript hosted by Anthropic. <https://www-cdn.anthropic.com/22573675ada52a8ca8a97a1a4b4326b2f208a071.pdf>.

2. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022). ColabFold: making protein folding accessible to all. *Nature Methods* 19(6):679-682. doi:10.1038/s41592-022-01488-1.
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583-589. doi:10.1038/s41586-021-03819-2.
4. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. (2021). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. doi:10.1101/2021.10.04.463034.
5. Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler S, Somnath VR, et al. (2025). Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*. doi:10.1101/2025.06.14.659707.
6. Wohlwend J, Corso G, Passaro S, Getz N, Reveiz M, Leidal K, et al. (2024). Boltz-1 Democratizing Biomolecular Interaction Modeling. *bioRxiv*. doi:10.1101/2024.11.19.624167.
7. Chai Discovery, Boitreaud J, Dent J, McPartlon M, Meier J, Reis V, Rogozhnikov A, Wu K (2024). Chai-1: Decoding the molecular interactions of life. *bioRxiv*. doi:10.1101/2024.10.10.615955.
8. ByteDance AML AI4Science Team, Chen X, Zhang Y, Lu C, Ma W, Guan J, et al. (2025). Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. *bioRxiv*. doi:10.1101/2025.01.08.631967.
9. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* 42(2):243-246. doi:10.1038/s41587-023-01773-0.
10. Steinegger M, Söding J (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35(11):1026-1028. doi:10.1038/nbt.3988.
11. Steinegger M, Söding J (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* 9(1):2542. doi:10.1038/s41467-018-04964-5.
12. Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* 33(7):2302-2309. doi:10.1093/nar/gki524.
13. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research* 49(D1):D412-D419. doi:10.1093/nar/gkaa913.
14. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. (2024). The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research* 53(D1):D523-D534. doi:10.1093/nar/gkae997.
15. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, et al. (2024). InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* 53(D1):D444-D456. doi:10.1093/nar/gkae1082.
16. Eddy SR (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology* 7(10):e1002195. doi:10.1371/journal.pcbi.1002195.
17. Chikhi R, Lemane T, Loll-Krippelbein R, Montoliu-Nerin M, Raffestin B, Camargo AP, et al. (2024). Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. *bioRxiv*. doi:10.1101/2024.07.30.605881.
18. Emenecker RJ, Griffith D, Holehouse AS (2021). Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical Journal* 120(20):4312-4319. doi:10.1016/j.bpj.2021.08.039.
19. Emenecker RJ, Griffith D, Holehouse AS (2022). Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *bioRxiv*. doi:10.1101/2022.06.06.494887.
20. Lotthammer JM, Hernández-García J, Griffith D, Weijers D, Holehouse AS, Emenecker RJ (2024). Metapredict enables accurate disorder prediction across the Tree of Life. *bioRxiv*. doi:10.1101/2024.11.05.622168.

21. Rojas JR, Trievel RC, Zhou J, Mo Y, Li X, Berger SL, Allis CD, Marmorstein R (1999). Structure of Tetrahymena GCN5 bound to coenzyme A and a histone H3 peptide. *Nature* 401(6748):93-98. doi:10.1038/43487.
22. Vetting MW, de Carvalho LPS, Yu M, Hegde SS, Magnet S, Roderick SL, Blanchard JS (2005). Structure and functions of the GNAT superfamily of acetyltransferases. *Archives of Biochemistry and Biophysics* 433(1):212-226. doi:10.1016/j.abb.2004.09.003.
23. Dyda F, Klein DC, Hickman AB (2000). GCN5-Related N-Acetyltransferases: A Structural Overview. *Annual Review of Biophysics and Biomolecular Structure* 29(1):81-103. doi:10.1146/annurev.biophys.29.1.81.
24. Chen W, Biswas T, Porter VR, Tsodikov OV, Garneau-Tsodikova S (2011). Unusual regioversatility of acetyltransferase Eis, a cause of drug resistance in XDR-TB. *Proceedings of the National Academy of Sciences* 108(24):9804-9808. doi:10.1073/pnas.1105379108.
25. Vetting MW, Roderick SL, Yu M, Blanchard JS (2003). Crystal structure of mycothiol synthase (Rv0819) from *Mycobacterium tuberculosis* shows structural homology to the GNAT family of N-acetyltransferases. *Protein Science* 12(9):1954-1959. doi:10.1110/ps.03153703.
26. Iqbal A, Arunlanantham H, Brown T, Chowdhury R, Clifton IJ, Kershaw NJ, Hewitson KS, McDonough MA, Schofield CJ (2009). Crystallographic and mass spectrometric analyses of a tandem GNAT protein from the clavulanic acid biosynthesis pathway. *Proteins: Structure, Function, and Bioinformatics* 78(6):1398-1407. doi:10.1002/prot.22653.
27. Trievel RC, Rojas JR, Sterner DE, Venkataramani RN, Wang L, Zhou J, Allis CD, Berger SL, Marmorstein R (1999). Crystal structure and mechanism of histone acetylation of the yeast GCN5 transcriptional coactivator. *Proceedings of the National Academy of Sciences* 96(16):8931-8936. doi:10.1073/pnas.96.16.8931.
28. Hickman AB, Nambodiri MAA, Klein DC, Dyda F (1999). The Structural Basis of Ordered Substrate Binding by Serotonin N-Acetyltransferase. *Cell* 97(3):361-369. doi:10.1016/s0092-8674(00)80745-x.
29. Wolf E, Vassilev A, Makino Y, Sali A, Nakatani Y, Burley SK (1998). Crystal Structure of a GCN5-Related N-acetyltransferase. *Cell* 94(4):439-449. doi:10.1016/s0092-8674(00)81585-8.
30. Olsen LR, Roderick SL (2001). Structure of the *Escherichia coli* GlmU Pyrophosphorylase and Acetyltransferase Active Sites. *Biochemistry* 40(7):1913-1921. doi:10.1021/bi002503n.
31. Berman HM (2000). The Protein Data Bank. *Nucleic Acids Research* 28(1):235-242. doi:10.1093/nar/28.1.235.
32. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. (2022). RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research* 51(D1):D488-D508. doi:10.1093/nar/gkac1077.
33. The UniProt Consortium, Bateman A, Martin MJ, Orchard S, Magrane M, Adesina A, et al. (2024). UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* 53(D1):D609-D617. doi:10.1093/nar/gkae1010.
34. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium (2014). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6):926-932. doi:10.1093/bioinformatics/btu739.
35. Sayers EW, Cavanaugh M, Frisse L, Pruitt KD, Schneider VA, Underwood BA, Yankie L, Karsch-Mizrachi I (2024). GenBank 2025 update. *Nucleic Acids Research* 53(D1):D56-D61. doi:10.1093/nar/gkae1114.
36. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. (2020). CATH: increased structural coverage of functional space. *Nucleic Acids Research* 49(D1):D266-D273. doi:10.1093/nar/gkaa1079.

37. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. (2023). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* 52(D1):D368-D375. doi:10.1093/nar/gkad1011.
38. Kityk R, Kopp J, Sinning I, Mayer MP (2012). Structure and Dynamics of the ATP-Bound Open Conformation of Hsp70 Chaperones. *Mol.Cell* 48:863-. doi:10.1016/J.MOLCEL.2012.09.023 (PDB 4B9Q; record: RCSB primary citation).
39. Ghode P, Jobichen C, Ramachandran S, Bifani P, Sivaraman J (2015). Structural basis of mapping the spontaneous mutations with 5-fluorouracil in uracil phosphoribosyltransferase from *Mycobacterium tuberculosis*. *Biochem.Biophys.Res.Comm.* 467:577-582. doi:10.1016/j.bbrc.2015.09.133 (PDB 5E38; record: RCSB primary citation).

그림

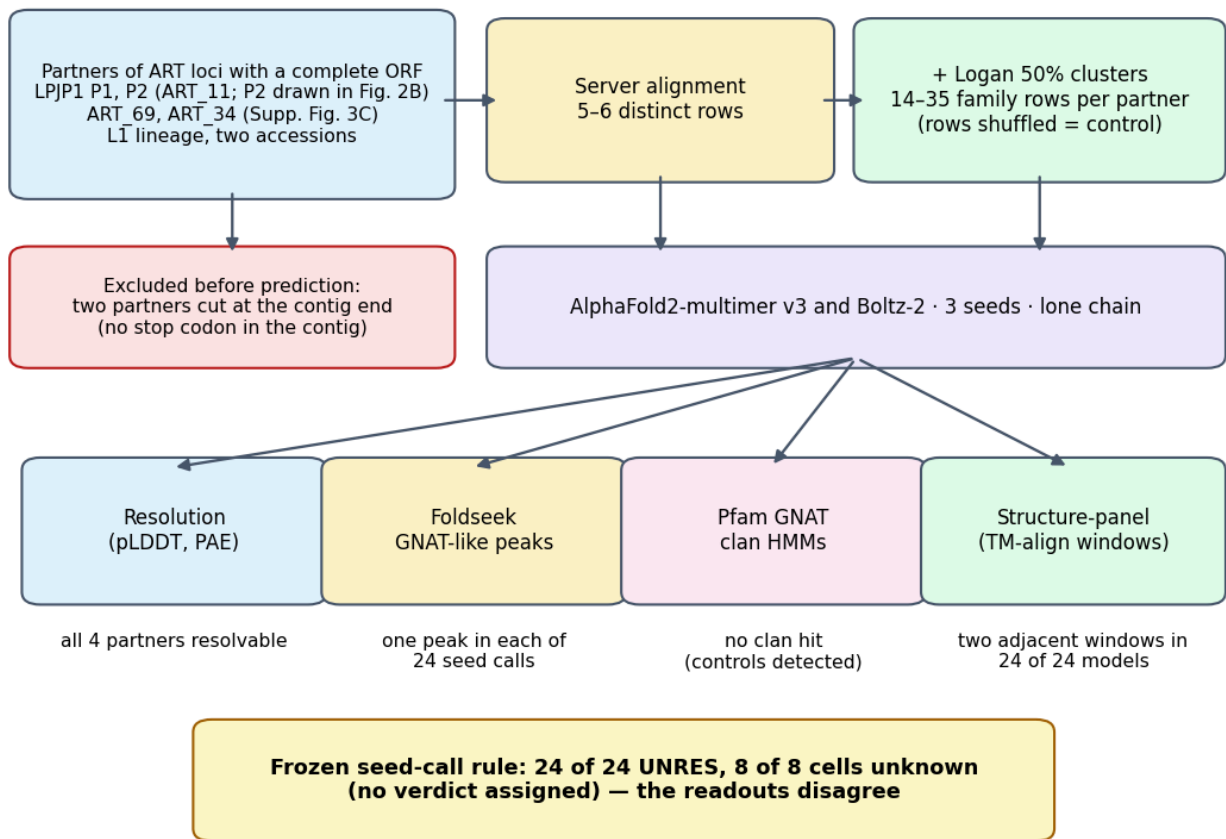


Figure 1. 마지막 단계와 판독의 개요. 완전한 ORF를 가진 ART 좌위의 파트너를 서버 정렬과 Logan 상동체를 더한 정렬(대조로 행 안에서 섞은 정렬 포함)에서 두 예측기로 예측하고, 네 가지 판독(해상도, Foldseek 피크, Pfam 클랜 모델, 구조 패널 윈도우)을 비교한다. 고정된 시드 판정 규칙은 모든 시드 판정이 UNRES이고 셀 라벨이 모두 unknown이다(판정은 내리지 않았다). 상자 속 숫자는 3.9절의 값이다.

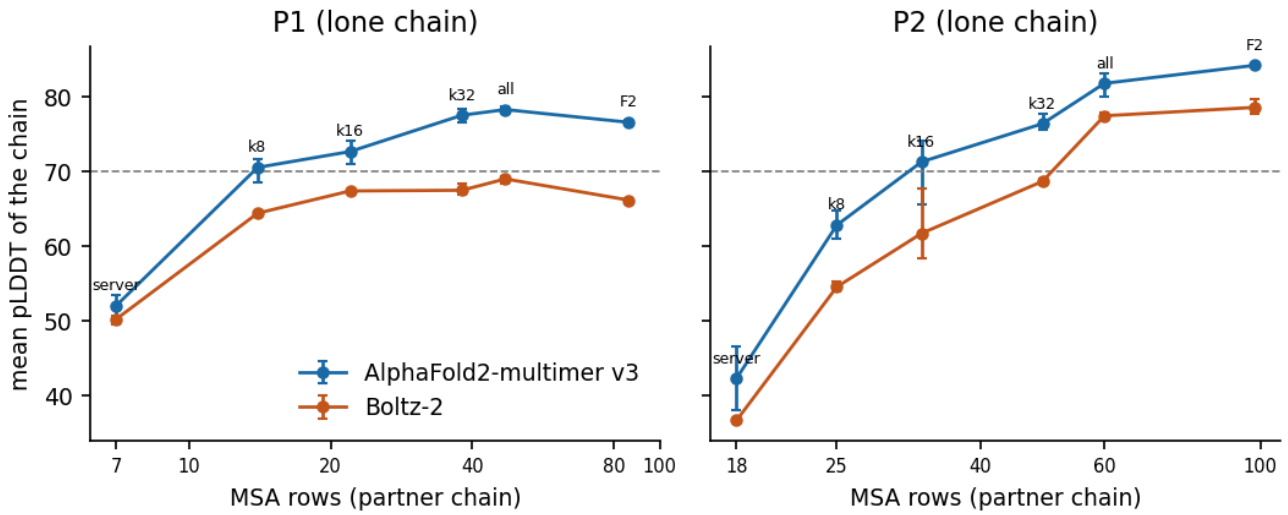


Figure 2. 파트너 정렬의 MSA 행 수에 따른 단독 P1과 P2 전체 체인의 평균 pLDDT(서버 행만, 서버 행에 요청한 패밀리 행 $k = 8, 16, 32$ 를 더한 것으로 서버 행과 합한 행 수가 그려져 있음, 서버 행에 패밀리 전체 행을 더한 것, 확장 패밀리), 세 시드의 평균과 범위. 70의 파선은 시각 보조선이며(해상 규칙은 슬롯 규칙이고 P1에서는 잔기 484-658을 제외한다).

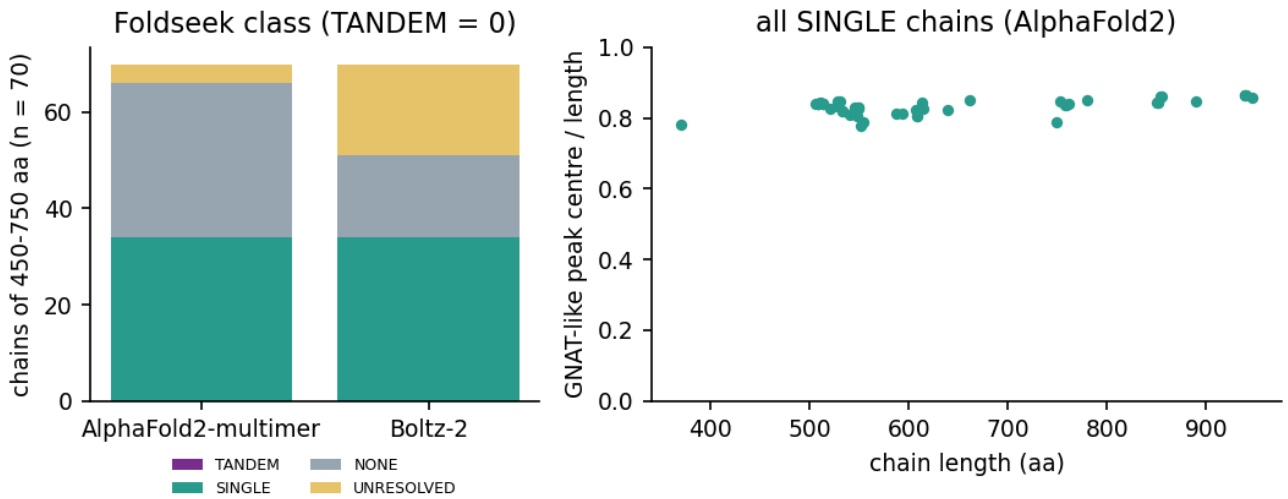


Figure 3. 왼쪽은 예측기별 450-750 aa 체인의 Foldseek 클래스(TANDEM: 없음), 오른쪽은 모든 SINGLE 체인(AlphaFold2)에서 체인을 따른 단일 GNAT 유사 피크의 위치.

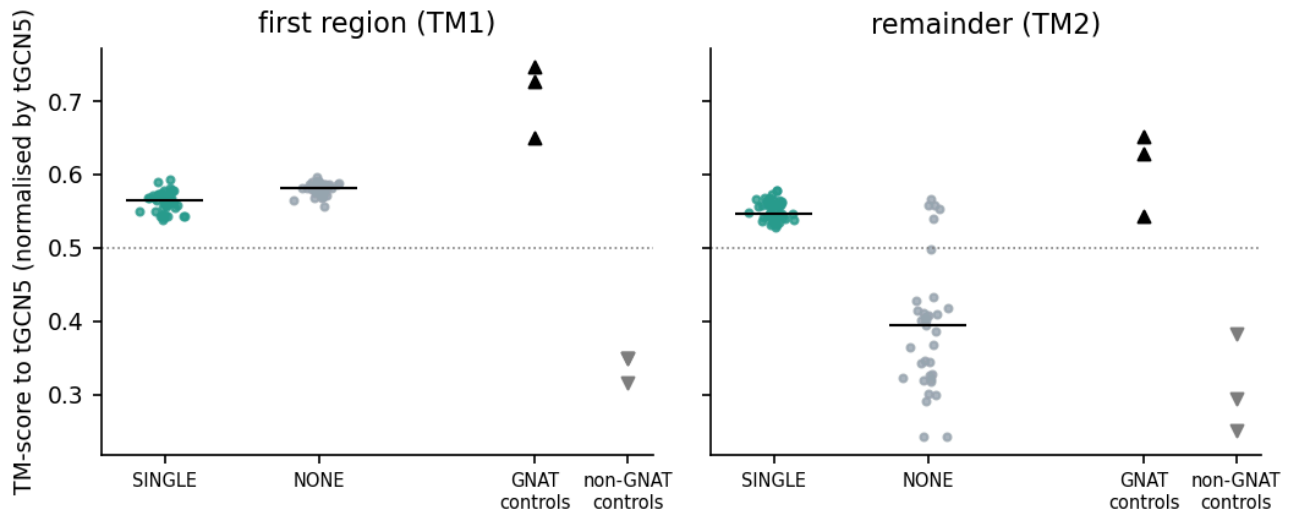


Figure 4. SINGLE 및 NONE 체인(AlphaFold2, 시드 0, 모든 길이)에서 첫 영역과 나머지 부분의 tGCN5에 대한 TM-score, 대조군 GNAT 단백질(Eis, MshD, Orf14, Rim1는 나머지 부분 없음)과 GNAT 비해당 대조군(GlmU, UPP, DnaK) 포함. 0.5의 점선은 시각 보조선이다.

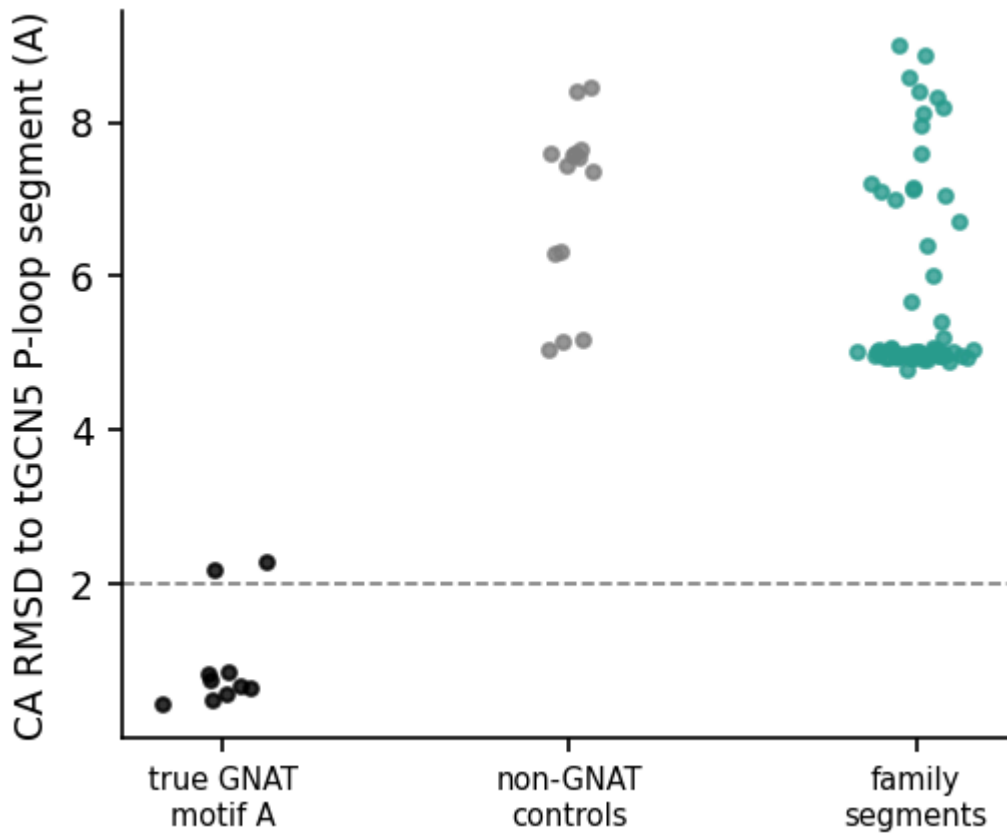


Figure 5. tGCN5 P-loop 구간에 대한 motif 구간의 CA RMSD. 대조군의 진짜 GNAT motif A 구간과 GNAT 비해당 대조군의 R/Q-x-x-G 구간(느슨한 패턴)을 패밀리의 R/Q-x-x-G-x-G/A 구간(엄격한 패턴)과 비교했다. 2 A의 파선은 사전 지정한 문턱이다.

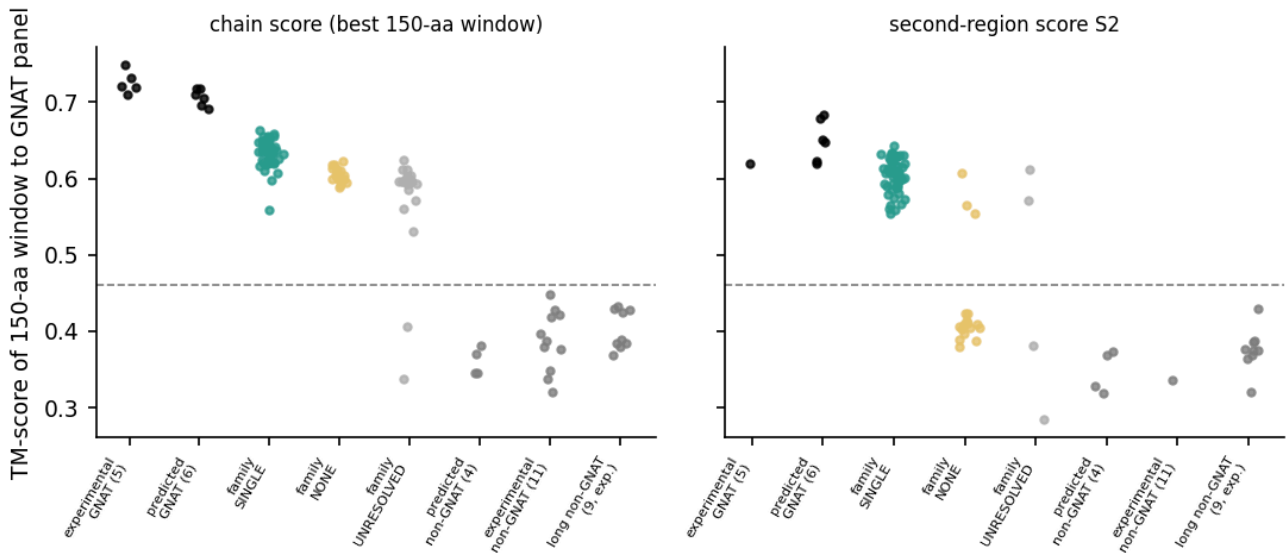


Figure 6. 구조 패널 점수(150잔기 윈도우의 GNAT 기준 다섯 개 중 최적에 대한 TM-score, 윈도우로 정규화, 파선 = 실험 기준 구조로 정한 문턱). 왼쪽은 체인 점수(최고 윈도우), 오른쪽은 두 번째 영역 점수 S2(최고 윈도우에서 150잔기 이상 떨어진 최고 윈도우, 두 번째 윈도우가 없거나 짧은 대조군처럼 있을 수 없는 경우에는 표시 없음). 집단: 실험 GNAT 기준(leave-one-out), 예측 GNAT 대조군(Eis, MshD, Orf14의 두 깊이), Foldseek 클래스별 패밀리 체인(Boltz-2 시드 0 모델), 예측 GNAT 비해당 대조군(GlmU, DnaK의 두 깊이), 실험 GNAT 비해당 기준, 긴 실험 GNAT 비해당 구조 아홉 개(D11).

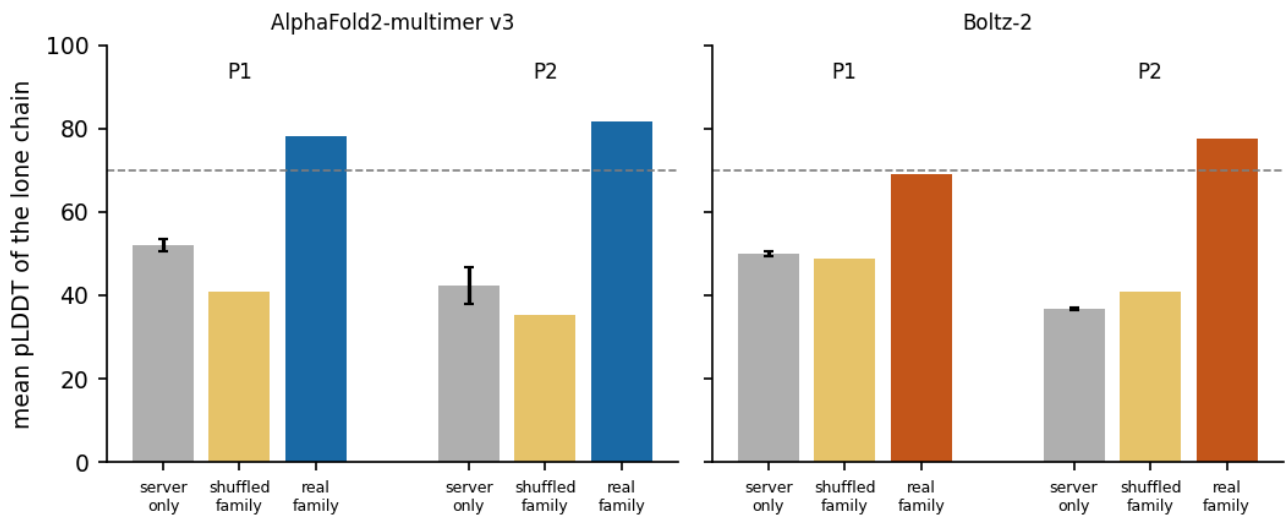


Figure 7. 행을 섞은 정렬 대조(D10): 서버 행만 넣은 단독 체인의 평균 pLDDT(막대: 시드 범위), 서버 행에 행 안에서 잔기를 치환한 패밀리 행을 더한 경우, 서버 행에 실제 패밀리 행을 더한 경우(시드 평균, 왼쪽 AlphaFold2, 오른쪽 Boltz-2, 70의 파선은 시각 보조선).

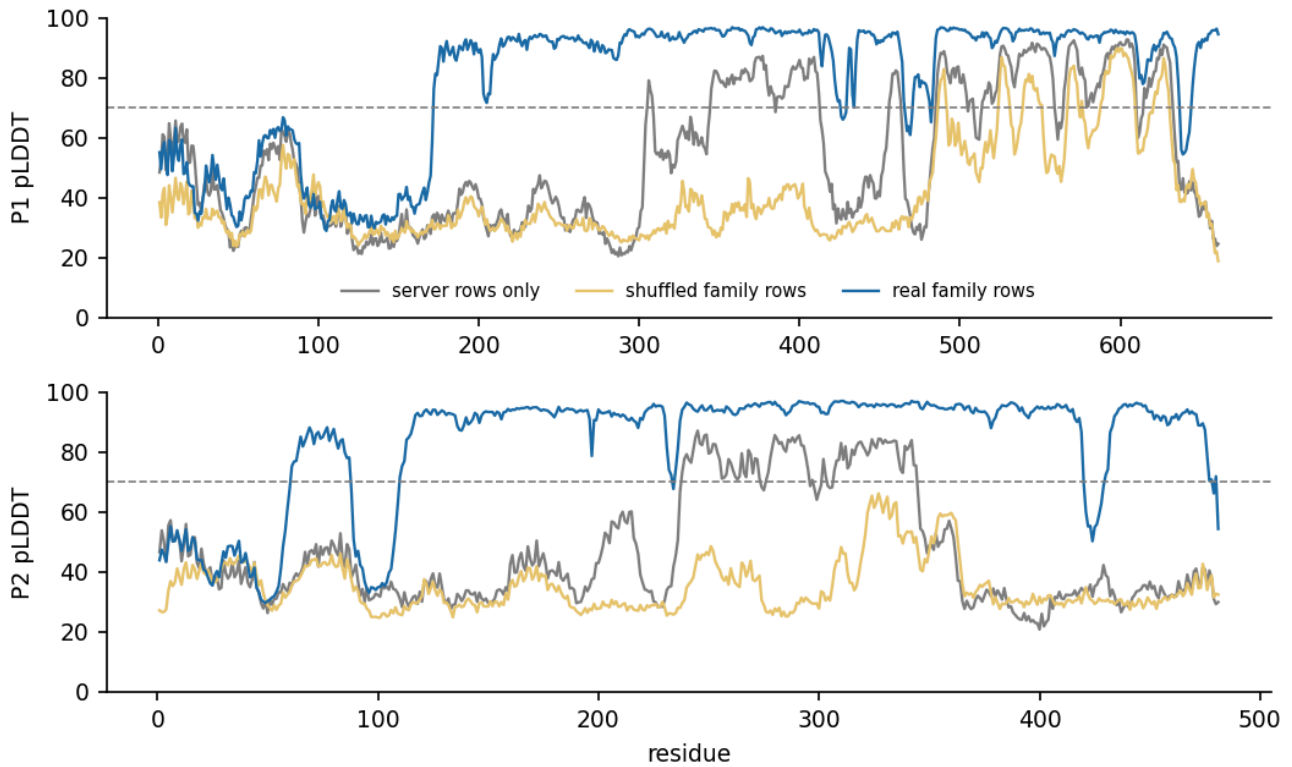


Figure 8. 서버 행만 넣은 경우, 서버 행에 행을 섞은 패밀리 행(D10)을 더한 경우, 서버 행에 실제 패밀리 행(D5-B)을 더한 경우의 단독 P1(위)과 P2(아래) AlphaFold2 순위 1 모델의 잔기별 pLDDT. 각각 시드 하나(수확물의 첫 파일이며 시드는 곡선 사이에 맞추지 않았다). 70의 파선은 시각 보조선이다. N 말단 영역은 실제 패밀리 행을 넣어도 낮게 머물고 섞은 행은 프로파일을 올리지 못한다.

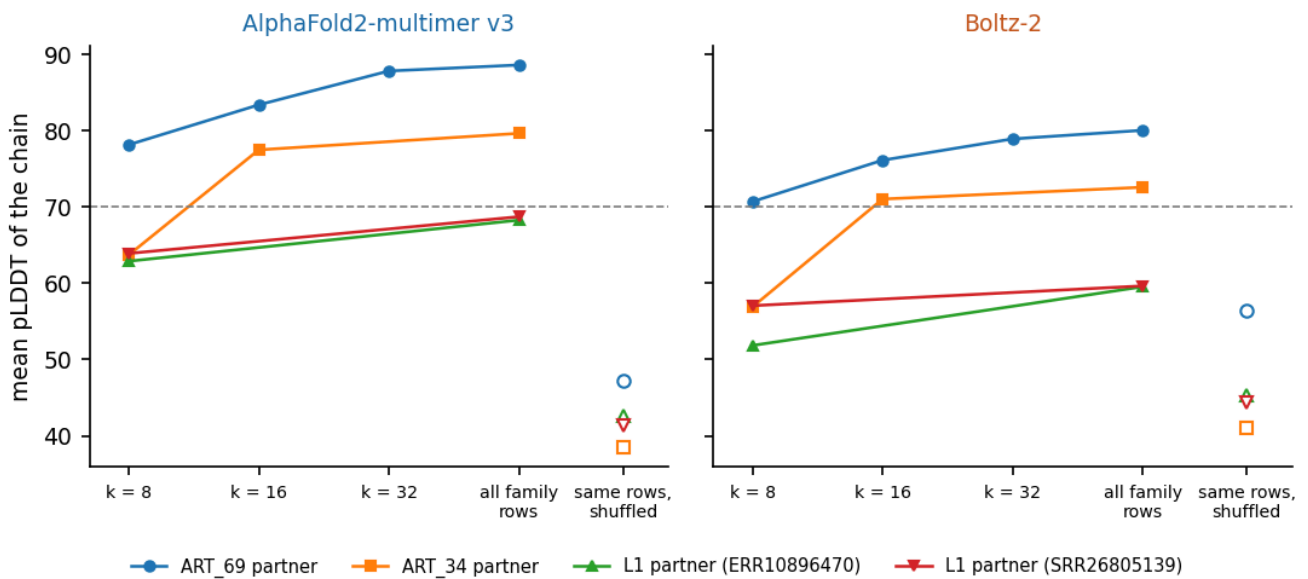


Figure 9. 3.9절 네 파트너의 단독 체인 평균 pLDDT를 서버 행에 더한 Logan 패밀리 행 수($k = 8, 16, 32$, 전체. 패밀리보다 큰 부분집합은 만들지 않았다)에 따라 그린 것과, 패밀리 행 전체를 행 안에서 섞은 경우(속이 빈 기호). 세 시드 범위의 평균이며 왼쪽이 AlphaFold2, 오른쪽이 Boltz-2이고 70의 파선은 시각 보조선이다.

orange bar: Foldseek GNAT-like peak · green bar: span of the structure-panel windows above the threshold
 grey: AlphaFold2 pLDDT (seed 0, rank 1 model of the server + Logan alignment) · Pfam GNAT clan: no hit in any partner

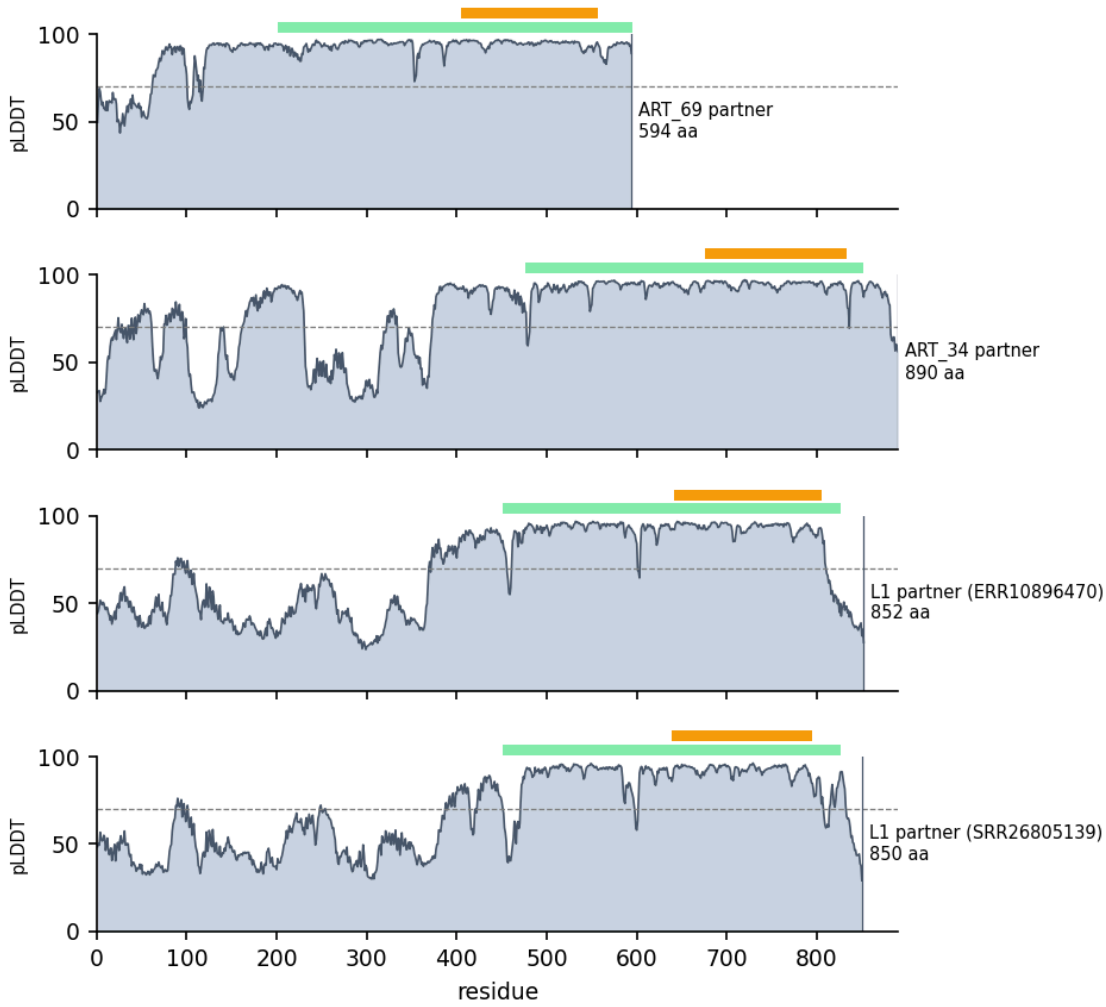


Figure 10. 네 파트너의 판독을 잔기별로 나타낸 지도. AlphaFold2 시드 0 순위 1 모델의 pLDDT(회색), Foldseek GNAT 유사 피크(주황 막대), 문턱을 넘은 구조 패널 윈도우의 합집합(초록 막대). Pfam GNAT 클랜 모델은 어느 파트너에서도 히트하지 않았다.