Auditable local SimpleFold-3B inference on Apple Silicon: an 8-target reproducibility benchmark

Raúl Chío León[1,2]

[1] Eigen Biotech (umbrella). [2] Tecnológico de Monterrey, Campus Guadalajara, Mexico.

Correspondence: raul.chio.leon@gmail.com.

Authorship note: this first manuscript is prepared as a solo-author submission by default. External biological or technical reviewers may be acknowledged, with permission and with their specific contribution named, but they are not listed as authors unless their contribution meets journal authorship criteria and they approve and accept accountability for the submitted work.

Abstract

Local use of open protein-structure models is attractive for researchers without CUDA cluster access, but hardware-viability claims need target-level evidence rather than anecdotes. We report an 8-target reproducibility benchmark of SimpleFold-3B on a 128 GB M3 Max workstation, using a direct local CLI route and public audit artifacts. Phase 1A froze the target set (13CZ:A, 13EE:A, 13FT:A, 11BC:C, 12OY:A, 9H8N:A, 9K9X:A, 9K9Y:A) before the four prospective targets were executed. All 8 targets produced predictions and scores. Across the frozen set, median wall-clock time was 395.96 seconds, median approximate peak RSS was 34.08 GB, median sequence-aligned lDDT was 0.8597, and median TM-score was 0.9218. The three long-bin targets (565-604 aa) completed without OOM at 34.02-34.64 GB peak RSS. Selective repeat controls now cover the 13CZ:A route/cache anchor plus the reviewer-sensitive 11BC:C hard/coverage target and 9K9Y:A long-bin target. This is not a model leaderboard, not a reproduction of the SimpleFold CASP14 headline benchmark, not an official AlphaFold 3 evaluation, and not a biological generality claim. Under this exact model, hardware, execution route, target set, and scorer, local SimpleFold-3B inference was operationally viable and auditable; the clean source snapshot and release bundle are archived at Zenodo DOI 10.5281/zenodo.20275055.

---

1. Introduction

Protein structure prediction has changed shape in 2024-2026. AlphaFold 3 (Abramson et al. 2024) introduced a diffusion-based all-atom architecture, while open and partially open systems such as Chai-1 (Chai Discovery 2024), Protenix-v1 (authors 2026), Boltz-2 (Wohlwend et al. 2025), and SimpleFold (Apple ML Research 2026) made parts of the ecosystem inspectable outside proprietary services. For small labs, the practical question is not whether Apple Silicon is generally sufficient for modern protein prediction. The defensible question is narrower: whether a specified model, route, machine, target set, and scorer can run with enough provenance for another researcher to audit.

SimpleFold is a useful first anchor because it provides an MLX path and model sizes up to 3B parameters. The M3 Max workstation (16-core CPU, integrated GPU, 128 GB unified memory) has enough memory headroom to attempt this route without a CUDA cluster. Community reports of such runs remain scattered and often omit exact checkpoints, route choice, scoring method, memory measurement, target context, or negative evidence. This paper therefore treats local inference as an empirical reproducibility problem rather than a hardware marketing claim.

Agent-assisted biology systems such as ProteinMCP (Xu et al. 2026), Agentomics (BioGeMT 2026), and Biomni (Snap-Stanford 2025) motivate machine-readable scientific workflows. This paper does not evaluate autonomous scientific discovery. LLM tooling is used here as documentation, curation, and verification support around a conventional benchmark; the scientific claim rests on executed predictions, scoring outputs, logs, cards, and environment provenance.

Contributions

1. **A frozen single-workstation viability protocol** for local protein-structure inference on a 128 GB M3 Max workstation, with Phase 1A frozen as an 8-target SimpleFold-3B benchmark and the broader multi-model design retained as later work [§3].
2. **A public reproducibility atlas pattern** that records target manifests, recipes, predictions, scoring outputs, model/runtime cards, environment JSON, checkpoint hashes, route choices, and failure modes [§4].
3. **A scoring-quality audit** showing that residue numbering can artifactually depress lDDT unless single-chain sequence alignment and aligned-residue coverage are reported [§4].
4. **A bounded empirical Apple Silicon result**: SimpleFold-3B completed all 8 frozen targets, including three 565-604 aa long-bin targets, under the direct local CLI route without OOM [§4].

Claim boundaries

Claim	Evidence	Limitation
SimpleFold-3B completed Phase 1A on Apple Silicon.	8/8 frozen targets produced predictions and scores in `pilot/data/simplefold3b_phase1a_summary.*`.	One model, one M3 Max machine, one direct CLI route.

Claim	Evidence	Limitation
Long-bin targets did not OOM under this route.	`9H8N:A`, `9K9X:A`, and `9K9Y:A` completed at 34.02-34.64 GB approximate peak RSS.	RSS is sampled process RSS, not full unified-memory profiling.
Selected structural scores are repeat-stable.	`13CZ:A`, `11BC:C`, and `9K9Y:A` each have n=3 controls with stable lDDT/TM-score and aligned-residue counts.	Five Phase 1A targets remain single-run observations.
Scoring is auditable.	Sequence-aligned lDDT, USalign TM-score, aligned-residue counts, cards, JSON summaries, logs, and 14-card score recomputation are public.	External biological interpretation remains limited without domain review.
The benchmark is reproducible as an artifact.	Environment card, model/runtime card, checkpoint hashes, target manifest, recipes, predictions, and scores are versioned.	Full inference reproduction requires comparable local hardware and SimpleFold checkpoint access; scoring/artifact reproduction is the lighter-weight path.

2. Related work

2.1 Open protein-structure model stack

AlphaFold 3 established the all-atom diffusion template for contemporary biomolecular structure prediction (Abramson et al. 2024), but the official model remains outside the scope of a fully reproducible local benchmark because of gated weights and licence constraints. The open ecosystem is therefore the relevant substrate for this study. Chai-1 exposes an open all-atom co-folding model and code path, but its upstream package documents a Linux/CUDA expectation; Protenix-v1 reports a fully open AlphaFold3-family system under an AlphaFold3-aligned cutoff and model scale; Boltz-2 adds affinity prediction but its primary path remains CUDA-oriented. These systems define the broader ecosystem, while their local portability remains uneven.

SimpleFold (Apple ML Research 2026) is the strongest initial anchor for this study because its repository supports both PyTorch and MLX inference and provides model sizes from 100M to 3B. Its architecture deliberately omits several domain-specific AF2/AF3 modules, including triangle attention and explicit pair representation biases,

while retaining competitive benchmark performance. That combination makes SimpleFold unusually suitable for separating two questions that are often conflated: model-method validity and workstation-substrate viability.

Community conversion paths such as chai-mlx and prospective MLX/OpenFold3 forks are treated as viability findings rather than assumptions. If a model has open weights but no stable Apple Silicon path, the correct result is not omission; it is a `port_unavailable`, `dependency_missing`, or `port_runtime_error` tag with exact installation evidence.

2.2 Agent-assisted reproducibility infrastructure

Agent-assisted biology systems such as ProteinMCP (Xu et al. 2026), Agentomics (BioGeMT 2026), and Biomni (Snap-Stanford 2025) motivate machine-readable interfaces for scientific workflows. In this paper, however, the assistant layer is deliberately secondary. The Phase 1A evidence does not evaluate autonomous scientific discovery, autonomous paper reading, or autonomous biological interpretation. Instead, LLM tooling is used as documentation, curation, and verification infrastructure around a conventional benchmark run. The benchmark claim therefore rests on frozen targets, executed predictions, scoring outputs, logs, cards, and environment provenance, not on an unevaluated autonomy claim.

2.3 Reproducibility infrastructure

Reproducibility work in this space must handle two different failure modes: scientific non-reproduction and engineering non-viability. A model can be scientifically strong while unavailable on a target substrate; conversely, a port can run while producing numerically divergent results. This study therefore treats installation, inference, scoring, and audit persistence as first-class observations, not just preliminary setup.

3. Methods

3.1 Candidate models

We pre-register the following candidate classes for the broader Phase 1B expansion. These models define the intended atlas direction; they are not evidence for the current Phase 1A claims. Phase 1A deliberately starts with SimpleFold-3B because it already has a native MLX route, an observed install path, and working runner telemetry.

Model	Source	Apple Silicon path	Parameters
SimpleFold-3B (Apple ML Research 2026)	Apple ML Research, ICLR 2026	Native MLX	3 B
Chai-1 via chai-mlx	Chai Discovery 2024 + community MLX weights/port	Native MLX (community)	contemporary all-atom candidate
ESMFold (ESM-2 backbone)	Meta / EvolutionaryScale	PyTorch + MPS	up to 15 B

Model	Source	Apple Silicon path	Parameters
Protenix-v1 (authors 2026)	Open AlphaFold3-family model	Unverified (subject to viability finding)	large biomolecular candidate
Boltz-2 (Wohlwend et al. 2025)	Open structure + affinity model	CUDA primary path; MPS community path is a caveat	structure + affinity candidate

We deliberately exclude AlphaFold 3 official (gated weights), AlphaProteo (closed), IsoDDE (closed), and Chai-2.x (closed). Boltz-2 is retained as an optional stress case only if the community MPS path is stable enough to avoid attribution confusion (colbyford 2025).

3.2 Test set

The broader candidate freeze was generated on 2026-05-13 using the RCSB Search and Data APIs and is stored as `pilot/data/test_set_v1_candidates.csv`. On 2026-05-17 we froze a smaller first publishable unit, `test_set_v1_phase1a`, by taking the 8 `provisional_keep` rows from the target-review triage. This prevents opportunistic target selection while keeping the first study feasible for a single workstation and a first paper.

Target	Bin	Length	Role	Pre-freeze evidence
`13CZ:A`	mid	161	repeat-control anchor	technical smoke observed
`13EE:A`	mid	161	benchmark slice	benchmark slice observed
`13FT:A`	mid	228	benchmark slice	benchmark slice observed
`11BC:C`	hard	116	alignment-coverage probe	benchmark slice observed
`12OY:A`	hard	304	membrane/complex context	prospective
`9H8N:A`	long	604	membrane/transport stress	prospective
`9K9X:A`	long	565	long enzyme context	prospective
`9K9Y:A`	long	565	long enzyme/ligand context	prospective

The freeze is computational rather than biological: final interpretation still requires domain review of the target set. The original 24-chain, multi-model design remains the Phase 1B expansion.

The target-set design used explicit inclusion and exclusion rules. Inclusion required an available RCSB reference structure, a single labelled chain suitable for a first-pass monomeric SimpleFold run, representation across mid/hard/long bins, and at least one prospective post-freeze long target. Exclusion removed entries that would require ligand placement, affinity evaluation, multi-chain co-folding, or biological mechanism interpretation to support the primary claim. The result is a computational viability set, not a biologically representative protein-family sample.

We therefore generated a structured biological pre-review packet from the frozen manifest, the Phase 1A result summary, and RCSB entry/entity metadata. All eight targets are acceptable for the bounded operational-viability claim, but two targets are explicitly blocked from biological generality: `11BC:C`, because only 63 aligned reference residues were scored, and `120Y:A`, because its membrane/complex cryo-EM context and 3.6 Å resolution make it a stress case. The paired DcrB structures (`13CZ:A`, `13EE:A`) are retained as controls but are not treated as evidence for broad protein-family generality, and `9K9Y:A` is not used for ligand-placement or affinity claims.

Target	Experimental context	Inclusion rationale	Interpretation limit
`13CZ:A`	X-ray, 2.136 Å; DcrB, *Salmonella enterica*; 161 aa.	Paired DcrB mid-length route/repeat anchor with prior CLI evidence.	Useful as route/repeat evidence; does not generalize beyond the paired DcrB family.
`13EE:A`	X-ray, 1.729 Å; DcrB, pH-variant pair; 161 aa.	Mid-bin DcrB benchmark slice paired with `13CZ:A`.	Same-family pairing prevents broad biological-family inference.
`13FT:A`	X-ray, 2.01 Å; STARD3 START-domain; 228 aa.	Clean mid-bin lipid-transfer domain quality anchor.	Single target only; no family-level claim.
`11BC:C`	Electron microscopy, 3.3 Å; HIV-1 Rev in complex; 116 aa label.	Hard-bin alignment-coverage and complex-derived scoring probe.	Only 63 reference residues aligned; use for scoring QA, not full-chain biological quality.
`120Y:A`	Cryo-EM, 3.6 Å; human MTCH2 membrane/complex; 304 aa.	Prospective hard-bin membrane/complex stress case.	Stress case only; not biological generality.
`9H8N:A`	Cryo-EM, 3.21 Å; BmrA transporter; 604 aa.	Prospective long membrane-transporter runtime/memory stress target.	Reports operational viability on this target, not transporter-general performance.

Target	Experimental context	Inclusion rationale	Interpretation limit
`9K9X:A`	X-ray, 2.03 Å; bicyclogermacrene synthase; 565 aa.	Prospective long enzyme target with high-quality X-ray reference.	Long-target runtime/quality evidence; no enzyme-family generality claim.
`9K9Y:A`	X-ray, 2.89 Å; bicyclogermacrene synthase with FsPP; 565 aa.	Prospective long enzyme/ligand-context target paired conceptually with `9K9X:A`.	Ligand context is provenance only; no ligand placement or affinity claim.

3.3 Metrics

- **lDDT** (local Distance Difference Test, Mariani 2013) computed via Bio.PDB on C-alpha atoms with 15 Å inclusion radius and (0.5, 1, 2, 4) Å thresholds.
- **TM-score** computed via USalign (Zhang and Skolnick 2024).
- **pLDDT distribution** captured from each model's native confidence head.
- **Wall-clock seconds** measured end-to-end per protein.
- **Peak resident-set memory (GB)** captured for the Phase 1A direct CLI route as approximate sampled process RSS, not full unified-memory profiling.

The lDDT scorer is versioned at `pilot/m3max-server/handlers/lddt.py`. For single-chain comparisons, it first extracts C-alpha records and residue identities from the predicted and reference PDBs, aligns the sequences with Biopython `PairwiseAligner`, maps aligned C-alpha coordinates into a shared index, and then computes the per-residue fraction of local reference-neighbour distances preserved within 0.5, 1, 2, and 4 Å thresholds under a 15 Å inclusion radius. If sequence alignment is not available, the scorer falls back to residue-index matching, and only then to a single-chain C-alpha-order fallback when too few residues overlap. Each score reports `chain_matching`, `sequence_identity`, and `n_residues_aligned` so coverage caveats are visible. A minimal regression smoke test in `pilot/scripts/test_lddt_scorer.py` checks that an identical four-residue structure with offset PDB residue numbering returns lDDT 1.0 under `single_chain_sequence_alignment`.

```
for each aligned residue i:
  local_pairs = residues j where reference_distance(i, j) <= 15 A
  preserved = count(|pred_distance(i, j) − reference_distance(i, j)| <
t
                    for each j and t in {0.5, 1, 2, 4} A)
  per_residue_lddt[i] = preserved / (4 ∗ number_of_local_pairs)
global_lddt = mean(per_residue_lddt)
```

3.4 Pipeline architecture

The benchmark is orchestrated by a hybrid Cloudflare + M3 Max stack with MCP-style services, configurable Reader/Validator agents, and an independent verifier path:

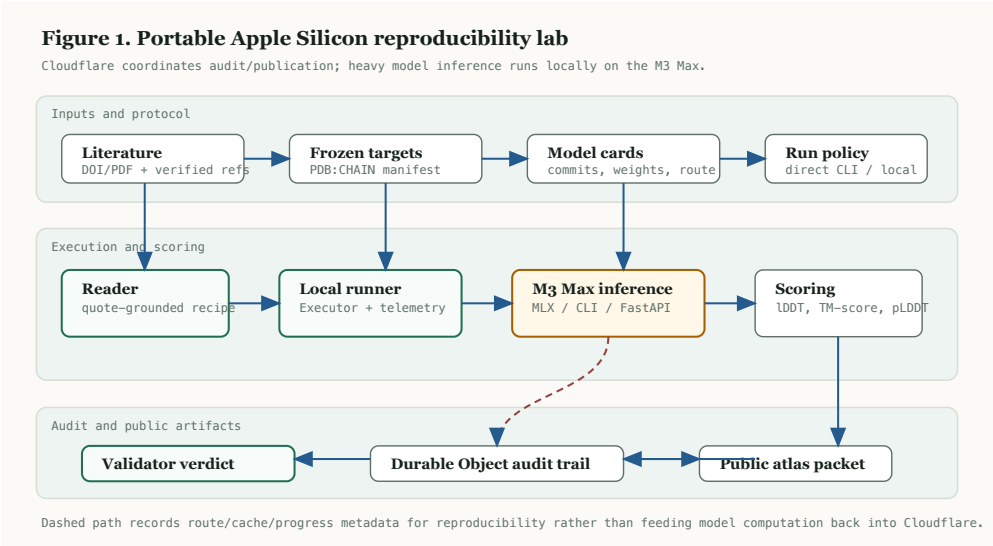

Figure 1. Portable Apple Silicon reproducibility lab. Cloudflare coordinates audit and publication; heavy model inference runs locally on the M3 Max.

The public services expose citation verification, PDF handling, PDB retrieval, scoring, and run-state persistence. Long-running inference is deliberately separated from public synchronous HTTP routes: the final benchmark policy runs model execution directly on the M3 Max, using either the local FastAPI server for short/medium jobs or direct CLI execution for longer SimpleFold-3B jobs. Cloudflare remains responsible for auditability, metadata, public result packets, and reproducibility surfaces rather than for heavy numerical inference.

Full implementation details and reproducibility instructions are maintained in `Theovia/eigen-reprod-atlas/docs/02_pipeline_architecture.md` and the public runbook.

3.5 Reproducibility constraints

- Phase 1A test set versioned at `pilot/data/test_set_v1_phase1a.csv` and `pilot/data/test_set_v1_phase1a.json`, with per-target fixtures in `pilot/fixtures/phase1a/`. The larger candidate file remains `test_set_v1_candidates.csv` for Phase 1B.
- Target-level biological pre-review versioned at `pilot/data/phase1a_biological_review.csv` and `pilot/data/phase1a_biological_review.json`; external domain review remains a final submission gate before biological generality claims.
- SimpleFold-3B Phase 1A install/runtime evidence is recorded in `pilot/model-cards/simplefold3b_phase1a.md` and `pilot/data/simplefold3b_phase1a_environment.json`, including M3 Max hardware, macOS build, package versions, SimpleFold source commit, USalign source commit, checkpoint sizes, and checkpoint SHA256 hashes.
- Model weights pinned to specific commits, package versions, and weight revisions where available; any Phase 1B model added later requires its own model/runtime card before comparison.
- All citations cleared through `cite-verify` MCP before inclusion in the final reference list.
- Audit log of Reader, Executor, Validator, scoring, and publication events persisted in Cloudflare Durable Object SQLite where the route supports it.

- Result packets include `recipe.json`, `predictions.json`, `scores.json`, `verdict.json`, and `card.json`.
- Direct CLI cards include execution metadata for route, seed, sample count, checkpoint files, auxiliary cache files, progress samples, and peak RSS when available.
- Random seeds set to 0 unless the model requires an ensemble or unsupported stochastic route, in which case the deviation is recorded in the card.

3.6 Statistical analysis

- Phase 1A descriptive summary: per-target lDDT, TM-score, pLDDT, wall-clock, peak RSS, failure tag, and aligned-residue coverage for the frozen SimpleFold-3B set.
- Phase 1A uses descriptive statistics only: per-target values, medians, ranges, aligned-residue coverage, and explicit single-run labels.
- Cross-model comparisons, bootstrapped confidence intervals, paired Wilcoxon tests, and reproduction-effect thresholds are reserved for Phase 1B and are not used to support the first-paper claim.
- Timing and memory summaries are reported as medians where repeated controls exist; otherwise they are explicitly marked as single-run estimates.
- pLDDT is treated as a model output under a stated route/cache condition, not as a stable calibration result unless repeated controls support that interpretation.
- Selective repeat controls are reported for three targets: the existing `13CZ:A` mid-bin repeat/cache anchor, a new `11BC:C` hard/coverage control, and a new `9K9Y:A` long-bin control.

4. Results

4.0 Phase 0 infrastructure validation

Before launching the Phase 1 benchmark, we ran a fixture-controlled public-tunnel smoke test to validate that the end-to-end architecture can produce real structures and real scores through the intended public route. The runner called deployed Cloudflare Workers, the `structure-pred` and `scoring` Workers reached the M3 Max server through `science.eigenatlas.com`, and the orchestrator persisted the run in a Durable Object.

This smoke used SimpleFold-100M on three intentionally small engineering targets (`1L2Y`, `1CRN`, `1UBQ`). It is not a biological benchmark and does not claim reproduction of the SimpleFold paper. It establishes that the infrastructure can produce auditable inference and scoring artifacts.

Target	pLDDT mean	Wall clock seconds	lDDT	TM-score	RMSD
1L2Y	97.2065	112.86	0.9708	0.6739	0.39
1CRN	90.5352	111.36	0.9867	0.9738	0.35
1UBQ	87.9837	110.86	0.9650	0.9659	0.81

The generated fixture verdict was `successful_prediction`, with median lDDT 0.9708 against the internal engineering fixture value 0.9695 (relative error 0.13%). The deployed Durable Object reported 41 audit entries, 3 prediction records, 3 score records, a saved

recipe, and a saved verdict. Public artifacts are available at `https://eigenatlas.com/eigen-reprod-atlas/results/phase0-public-tunnel-smoke/`.

The Phase 1A prospective results below are now complete and summarized from result cards.

4.1 Phase 1 technical rehearsal

After generating the 24-chain candidate manifest, we ran a three-target technical rehearsal using `simplefold-100m` on one short, one mid-length, and one long candidate chain (`9PFE:A`, `13CZ:A`, `13BB:A`). This rehearsal is not a final Phase 1 benchmark result and is not used for model-quality claims. It tests the candidate-target plumbing before biological target approval.

The short and mid targets completed through the public Worker/Tunnel route and produced predictions plus lDDT/TM-score records. The long target failed with HTTP 524 on the public synchronous route and is now tagged as `orchestration_timeout`. We then reran the same long target through direct local prediction (`DIRECT_M3MAX_URL=http://127.0.0.1:8001`), which completed in 280.38 s with sequence-aligned lDDT 0.8431 and TM-score 0.9567. This identifies a required engineering policy before the final long-bin benchmark: long jobs should use either a direct local M3 route or an asynchronous job/polling pattern rather than a single public synchronous fetch.

We also ran SimpleFold-3B smokes on one short and one mid-length candidate. The first short-target attempt downloaded the 3B checkpoint and outlived the runner client's request timeout; after checkpoint caching, the resumed direct-local run on `9PFE:A` completed in 285.08 s with pLDDT mean 96.1811, lDDT 0.7505, and TM-score 0.0464. A subsequent local HTTP attempt on the mid/long pair (`13CZ:A`, `13BB:A`) timed out while backend 3B subprocesses continued running, showing that even local synchronous HTTP is not a safe execution route for longer 3B jobs. We therefore added `DIRECT_PREDICTION_MODE=simplefold_cli` and reran the mid-length target `13CZ:A`; CLI mode completed in 609.38 s with pLDDT mean 81.8822, sequence-aligned lDDT 0.8627, and TM-score 0.9270. After adding progress logging and approximate RSS sampling, a repeated CLI telemetry run on `13CZ:A` completed in 378.92 s with peak RSS 31.4909 GB, pLDDT mean 85.2307, sequence-aligned lDDT 0.8613, and TM-score 0.9224. After adding structured `execution` metadata to `PredictionResult`, a third CLI repeat completed in 285.76 s with peak RSS 32.4891 GB, pLDDT mean 84.1826, sequence-aligned lDDT 0.8569, TM-score 0.9195, and machine-readable checkpoint/cache/progress evidence in the card. These runs are install, orchestration, telemetry, and metadata evidence for the native MLX 3B path, not final target-set performance.

The three `13CZ:A` CLI smokes were summarized by `pilot/scripts/summarize_simplefold3b_repeats.mjs` and converted into a repeat/cache-state protocol. Across these same-target runs, median pLDDT was 84.1826, median wall-clock time was 378.916 s, median peak RSS across telemetry-enabled runs was 31.99 GB, median sequence-aligned lDDT was 0.8613, and median TM-score was 0.9224. This makes the first methodological rule explicit: one-off pLDDT and timing values are preliminary unless they carry repeat controls or an explicit single-run limitation.

Finally, we removed the engineering recipe fixture from the Reader path for SimpleFold by adding a narrow manual-curation command adapter. This no-fixture dry run extracted the SimpleFold paper title, CASP14 target list, SimpleFold-3B model path, and headline CASP14 median LDDT value 0.709 from primary sources, producing `reader_raw.txt`, `recipe.json`, `verdict.json`, and `card.json` through the same runner output schema. The validator status is `successful_with_caveat`, because no inference was executed and the scores are dry-run schema values. This is sufficient for protocol curation in the benchmark paper; it is not yet evidence of autonomous Reader-agent extraction.

We then added an explicit benchmark-slice verdict mode and ran three provisional Phase 1 targets from the target-review triage with SimpleFold-3B. In benchmark mode, the runner preserves prediction, scoring, card generation, and audit persistence, but reports `successful_with_caveat` rather than a paper-reproduction verdict. The two mid-bin targets completed and scored: `13EE:A` produced pLDDT 81.6716, wall-clock 401.635 s, peak RSS 38.5429 GB, sequence-aligned lDDT 0.8582, TM-score 0.9212, and RMSD 1.68; `13FT:A` produced pLDDT 92.9916, wall-clock 390.276 s, peak RSS 31.2522 GB, sequence-aligned lDDT 0.8915, TM-score 0.9588, and RMSD 1.23. We then added the hard-bin complex-context target `11BC:C`, which completed with pLDDT 81.2384, wall-clock 337.455 s, peak RSS 41.4060 GB, sequence-aligned lDDT 0.7492, TM-score 0.6569, RMSD 1.97, and 63 aligned reference residues. Across the three provisional slices, median wall-clock time was 390.276 s, median peak RSS was 38.5429 GB, median sequence-aligned lDDT was 0.8582, and median TM-score was 0.9212. These target-level benchmark-slice results, together with the `13CZ:A` repeat-control target, are recorded as pre-freeze evidence inside the frozen Phase 1A set. The remaining four targets were then executed prospectively after the freeze.

Target	Bin	Outcome	pLDDT mean	Wall clock seconds	lDDT	TM-score	Failure
`9PFE:A`	short	predicted and scored	95.9711	117.69	0.7395	0.0482	n/a
`13CZ:A`	mid	predicted and scored	71.5778	115.44	0.8532	0.9004	n/a
`13BB:A`	long	public synchronous prediction timed out	n/a	n/a	n/a	n/a	`orchestratio`

Target	Model	Route	Outcome	pLDDT mean	Wall clock seconds	lDDT
`9PFE:A`	`simplefold-3b`	direct local HTTP, cached checkpoint	predicted and scored	96.1811	285.08	0.7505
`13CZ:A`	`simplefold-3b`	direct local CLI	predicted and scored	81.8822	609.38	0.8627

Target	Model	Route	Outcome	pLDDT mean	Wall clock seconds	lDDT
13CZ:A	simplefold-3b	direct local CLI, telemetry repeat	predicted, scored, peak RSS captured	85.2307	378.92	0.8613
13CZ:A	simplefold-3b	direct local CLI, execution metadata repeat	predicted, scored, checkpoint/cache metadata captured	84.1826	285.76	0.8569

Benchmark-slice target	Bin	Length	pLDDT mean	Wall clock seconds	Peak RSS GB	lDDT	TM-score	Alig resid
11BC:C	hard	116	81.2384	337.455	41.4060	0.7492	0.6569	
13EE:A	mid	161	81.6716	401.635	38.5429	0.8582	0.9212	
13FT:A	mid	228	92.9916	390.276	31.2522	0.8915	0.9588	

All negative and ambiguous observations are retained in a dedicated difficulty log (`docs/24_phase1_difficulty_log.md`) before being abstracted into manuscript limitations. Current logged issues include public Worker/Tunnel timeout on long inference, residue-numbering assumptions in lDDT scoring, checkpoint-download timeout, direct HTTP unsuitability for longer 3B jobs, CLI progress-observability gaps, per-run auxiliary cache materialization, repeat-run variability in confidence/timing, and the need for machine-readable execution metadata. The lDDT scorer now uses single-chain sequence alignment so model-numbered predictions can be compared to PDB-numbered references. The corresponding repeat/cache-state protocol is published as `docs/28_simplefold3b_repeat_cache_protocol.md`.

4.2 Phase 1A viability and quality

After the computational freeze, we ran the four prospective targets (`12OY:A`, `9H8N:A`, `9K9X:A`, `9K9Y:A`) through the same direct SimpleFold-3B CLI route. All four produced predictions and scores. Combined with the four pre-freeze targets, the frozen Phase 1A set is now complete: 8/8 targets scored, with median wall-clock time 395.96 s, median peak RSS 34.08 GB, median sequence-aligned lDDT 0.8597, and median TM-score 0.9218. The four prospective targets alone had median wall-clock time 726.63 s, median peak RSS 34.08 GB, median lDDT 0.8618, and median TM-score 0.9095.

Target	Stratum	Bin	Length	pLDDT mean	Wall clock seconds	Peak RSS GB	lDDT	TM scor
13CZ:A	pre-freeze	mid	161	84.1826	378.916	31.9900	0.8613	0.922
13EE:A	pre-freeze	mid	161	81.6716	401.635	38.5429	0.8582	0.921
13FT:A	pre-freeze	mid	228	92.9916	390.276	31.2522	0.8915	0.958

Target	Stratum	Bin	Length	pLDDT mean	Wall clock seconds	Peak RSS GB	lDDT	TM scor
11BC:C	pre-freeze	hard	116	81.2384	337.455	41.4060	0.7492	0.656
120Y:A	prospective	hard	304	87.7506	308.066	33.2323	0.8135	0.872
9H8N:A	prospective	long	604	88.1646	739.974	34.6408	0.8420	0.849
9K9X:A	prospective	long	565	95.8087	713.290	34.0239	0.8815	0.946
9K9Y:A	prospective	long	565	96.2908	753.399	34.1274	0.9157	0.982

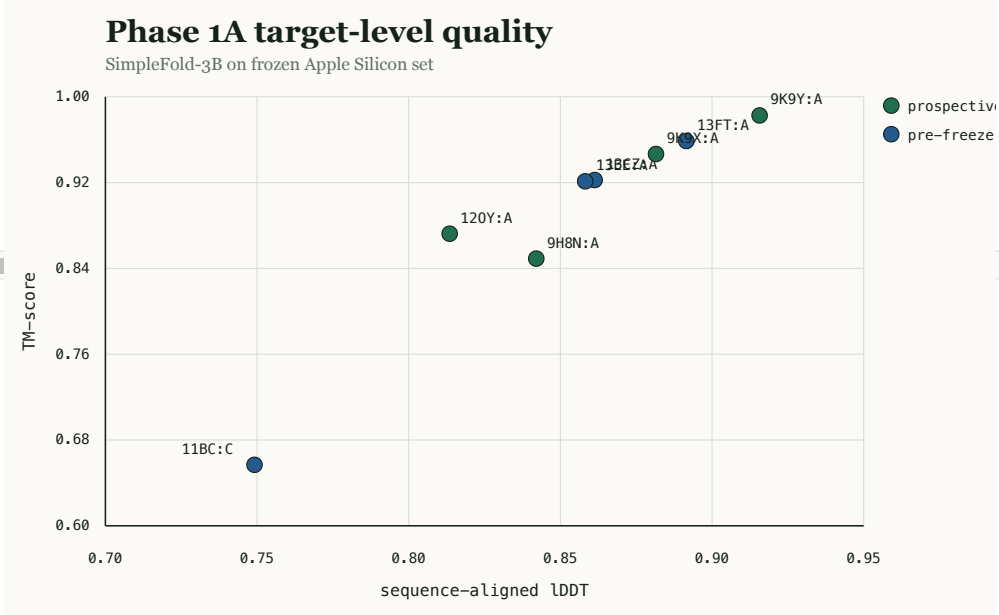

Figure 2. Phase 1A target-level quality for SimpleFold-3B on the frozen Apple Silicon set. Prospective targets are separated from pre-freeze evidence.

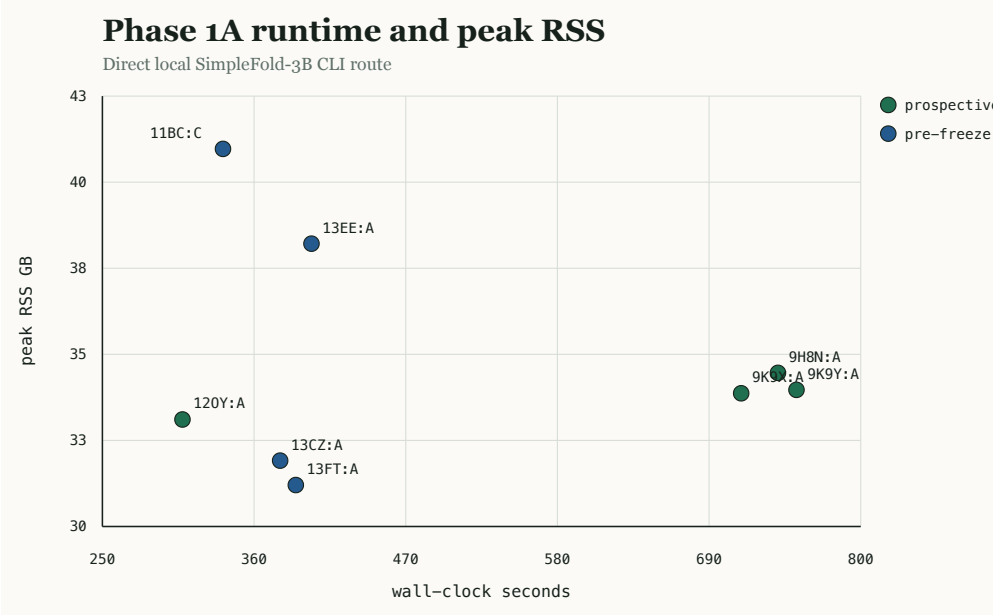

Figure 3. Phase 1A wall-clock time versus approximate peak RSS for the direct local SimpleFold-3B CLI route.

The long-bin targets (`9H8N:A`, `9K9X:A`, `9K9Y:A`) completed without OOM or route timeout. Their wall-clock range was 713.290-753.399 s and their peak RSS range was 34.0239-34.6408 GB. This supports a narrow viability finding for this hardware, route, model, and target set: long-target SimpleFold-3B inference on the M3 Max was runtime-bounded rather than memory-bounded.

We then added selective repeat controls for the two reviewer-sensitive cases: the hard/coverage target `11BC:C` and the long-bin target `9K9Y:A`. Combined with their original runs, each now has n=3. `11BC:C` preserved the coverage caveat across all repeats: aligned residues remained exactly 63, with median pLDDT 80.7909, median wall-clock time 299.255 s, median peak RSS 33.4526 GB, median lDDT 0.7492, and median TM-score 0.6520. `9K9Y:A` preserved high long-bin structural scores across all repeats: aligned residues remained 487, with median pLDDT 95.8859, median wall-clock time 806.112 s, median peak RSS 34.1274 GB, median lDDT 0.9131, and median TM-score 0.9826. These repeats strengthen the narrow viability claim while reinforcing the caution that wall-clock time varies more than structure-quality metrics. Only three reviewer-sensitive targets have n=3 controls; the remaining five Phase 1A targets support completion and target-level viability, not runtime-distribution estimates.

4.3 Reproduction of published headline

The Phase 1A benchmark is not reported as reproduction of the SimpleFold paper headline metric. The no-fixture manual Reader bridge extracted the SimpleFold-3B CASP14 median LDDT value 0.709, but Phase 1A uses a new frozen target set rather than the paper's CASP14 evaluation set. We therefore report target-level viability and quality directly, with `successful_with_caveat` verdicts where no published headline metric is being reproduced.

4.4 Failure modes and evidence packaging

No Phase 1A frozen target failed inference or scoring. The failure-mode result for this set is therefore not an OOM table but an evidence-quality finding: `12OY:A` and `9H8N:A` were preserved through primary result cards before terminal `tee` logging was added, while `9K9X:A` and `9K9Y:A` additionally have terminal logs under `pilot/evidence/phase1a/logs/`. Final numeric claims trace to `card.json`, `scores.json`, and `pilot/data/simplefold3b_phase1a_summary.*`; screenshots and terminal logs are supplemental reviewer aids.

The corresponding install/runtime evidence is also now public. The Phase 1A SimpleFold-3B card records the workstation model (`Mac15,9`), Apple M3 Max/128 GB hardware, macOS 26.5 build `25F71`, SimpleFold package version `0.1.0`, SimpleFold source commit `c7a5570a6be9f5c695126e27c804e77567209934`, USalign version `20260329`, and SHA256 hashes for `simplefold_3B.ckpt` and `plddt.ckpt`. This prevents the benchmark from depending on an implicit local setup.

5. Discussion

At the current stage, this study supports a bounded empirical claim: SimpleFold-3B can be run and scored across the frozen Phase 1A set on a 128 GB M3 Max workstation, with public machine-readable artifacts and explicit caveats. Practical model viability on Apple

Silicon cannot be inferred from model openness, headline accuracy, or even a successful short-target smoke. It must be measured through the combined chain of installation, route selection, inference, scoring, telemetry, and public artifact generation.

The most important early finding is that execution route is not incidental infrastructure. The same portable lab exposed three distinct operational regimes: public Worker/Tunnel execution, direct local HTTP execution, and direct CLI execution. Public synchronous prediction was adequate for small SimpleFold-100M smokes but produced an HTTP 524 on the long `13BB:A` rehearsal target. Direct local HTTP recovered that target, showing that the public failure was an orchestration limit rather than a model-viability result. For SimpleFold-3B, local HTTP still failed as a long-job wrapper, while direct CLI execution completed and allowed the runner to add process-level telemetry. These observations justify treating route as a recorded experimental variable.

The repeated `13CZ:A`, `11BC:C`, and `9K9Y:A` SimpleFold-3B CLI runs show why timing and confidence should not be overinterpreted from a single run. lDDT, TM-score, and aligned-residue coverage stayed close within each repeated target, but wall-clock time and sampled peak RSS varied enough to require repeat-aware reporting. Phase 1A therefore reports timing and memory as observed values or repeat medians/ranges, not as calibrated hardware constants.

The first benchmark slices extend that caution from repeated-run controls to scoring instrumentation. Their initial cards showed high TM-score but low lDDT; audit revealed that model predictions were numbered from residue 1 while PDB references retained source residue numbering. After switching to single-chain sequence-aligned lDDT, `13EE:A` corrected from 0.2368 to 0.8582 and `13FT:A` corrected from 0.2429 to 0.8915. The later `11BC:C` hard-bin slice added a different caution: even when sequence identity is 1.0, only 63 reference residues aligned because the reference is complex-derived and partially observed. This is a methodological result: final Phase 1 must treat metric implementation, sequence alignment, reference-state handling, and alignment coverage as part of the scientific apparatus.

The Phase 1A prospective runs add the first positive stress result: three long-bin targets from 565 to 604 residues completed without OOM at approximately 34-35 GB peak RSS. This does not mean Apple Silicon is generally sufficient for contemporary protein-structure inference. It means that, for this SimpleFold-3B route and this target set, the limiting practical factor was wall-clock time rather than memory collapse. The lower-scoring membrane/complex cases also show why “runs locally” and “biologically strong” are separate claims.

The paper’s contribution is therefore a viability map, not a model leaderboard. Native MLX paths, community ports, and CUDA-first repositories are expected to fail in different ways. Those failures are scientifically useful if they are recorded precisely enough that another researcher can distinguish install friction, route artifact, memory pressure, scoring mismatch, and genuine model-output drift. We explicitly do not claim a ranking of model quality or biological generality at this stage.

6. Limitations

- Community-grade ports may not produce numerics identical to CUDA reference implementations. The `port_quality_drift` failure tag was created for this exact ambiguity.

- Single-machine measurements are subject to confounds (background processes, thermal throttling).
- Direct CLI execution required extra runner telemetry after the first mid-target smoke; existing pre-telemetry cards must not be treated as peak-memory evidence.
- The initial repeat/cache-state control covers one target and one model path; it constrains reporting language but does not replace full repeated controls across the final benchmark.
- Five Phase 1A targets remain single-run observations. They support viability under stated conditions but not stable runtime or confidence calibration distributions; selected repeats now cover the `13CZ:A` mid-bin anchor, `11BC:C` hard/coverage control, and `9K9Y:A` long-bin control.
- Complex-derived targets can have limited observed-residue coverage; `11BC:C` aligned 63 residues despite a 116-aa target label, so final tables must report alignment coverage with lDDT.
- The Phase 1A target manifest is computationally frozen but not yet externally reviewed; final biological interpretation requires external domain review or an explicit limitation.
- Supplemental terminal logs were added after the first two prospective targets. `120Y:A` and `9H8N:A` still have primary card/scores/prediction evidence, but not `tee` terminal logs.
- No-fixture recipe extraction currently exists as manual curation rather than autonomous Reader-agent extraction; this is acceptable for benchmark protocol definition but not for a future assisted-methods claim.
- The empirical evidence is intentionally narrow: SimpleFold-3B local inference only. Conclusions do not generalise to other open structure models, official AlphaFold 3, design models (RFdiffusion3, BindCraft, etc.), or CUDA-only workflows.
- We did not benchmark gated proprietary models (AlphaFold 3 official, AlphaProteo, IsoDDE, Chai-2.x) and therefore cannot place our results on the same axis as commercial-tier benchmarks.
- Phase 1A benchmarks one model on eight frozen targets; the larger multi-model and 24-target design is Phase 1B, not part of the first submission claim.

7. Conclusion

This work frames Apple Silicon viability as an empirical question rather than an assumption. Phase 0 establishes that a single researcher can run, score, and publicly audit real SimpleFold-100M predictions through the proposed stack. Phase 1A answers a narrower first publishable question: SimpleFold-3B ran and scored across an 8-target frozen Apple Silicon benchmark, with quality, throughput, memory, alignment coverage, and evidence limitations reported as public artifacts.

Acknowledgements

External biological, technical, or academic review, if obtained before submission, will be acknowledged only with the reviewer's permission and with a specific description of the contribution. We thank the Apple ML Research team for releasing SimpleFold and the community maintainers who make local protein-structure inference paths inspectable. We acknowledge a mid-session mis-verification incident in our own earlier work and the subsequent design of the cite-verify MCP as motivation for the citation-verification

infrastructure included here. ChatGPT/Codex was used for coding assistance, drafting support, and artifact organisation; the human author reviewed the outputs and remains responsible for the accuracy, integrity, originality, and final wording of the manuscript.

Data and code availability

- Pipeline source snapshot: https://github.com/Theovia/eigen-reprod-atlas-phase1a/releases/tag/v0.1.2-phase1a. This public repository is a clean exported snapshot and does not expose private pre-release development history from the build repository.
- Zenodo archive DOI: https://doi.org/10.5281/zenodo.20275055.
- Public Phase 1A result packet: https://eigenatlas.com/eigen-reprod-atlas/results/phase1a-simplefold3b/.
- Public SimpleFold-3B runtime card: https://eigenatlas.com/eigen-reprod-atlas/model-card-simplefold3b.
- Machine-readable environment card: https://eigenatlas.com/eigen-reprod-atlas/data/simplefold3b_phase1a_environment.json.
- Public Phase 1A biological review packet: https://eigenatlas.com/eigen-reprod-atlas/phase1a-biological-review.
- Public Phase 1A repeat controls: https://eigenatlas.com/eigen-reprod-atlas/repeat-controls-results.
- Public score recomputation packet: https://eigenatlas.com/eigen-reprod-atlas/score-recompute/.
- Test set CSV + per-paper cards + source snapshot: archived in the Zenodo release bundle above.

References

Auto-generated from `paper/references.bib` *at PDF render. See file for entries.*

Abramson, Josh, Jonas Adler, et al. 2024. “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3.” *Nature*, ahead of print. https://doi.org/10.1038/s41586-024-07487-w.

Apple ML Research. 2026. “SimpleFold: Standard Transformer Flow-Matching for Protein Structure Prediction.” *ICLR*. https://github.com/apple/ml-simplefold.

authors, Protenix. 2026. “Protenix-V1: An Open AlphaFold3-Equivalent.” *bioRxiv*, ahead of print. https://doi.org/10.64898/2026.02.05.703733.

BioGeMT. 2026. “Agentomics: Autonomous Biomedical ML Agent.” *bioRxiv*, ahead of print. https://doi.org/10.64898/2026.01.27.702049.

Chai Discovery. 2024. *Chai-1: An Open All-Atom Co-Folding Model*. Https://chaidiscovery.com/blog/introducing-chai-1.

colbyford. 2025. “Apple Silicon (MPS) Support for Boltz Inference — PR #527.” *GitHub*. https://github.com/jwohlwend/boltz/pull/527.

Snap-Stanford. 2025. “Biomni: General Biomedical Agent over 105 Software Tools and 150 Specialised Algorithms.” *bioRxiv*, ahead of print. https://doi.org/10.1101/2025.05.30.656746.

Wohlwend, Jeremy et al. 2025. “Boltz-2: All-Atom Co-Folding and Binding Affinity Prediction.” *bioRxiv*, ahead of print. https://doi.org/10.1101/2025.06.14.659707.

Xu, Xin, Chen Feng, Chen Zha, et al. 2026. “ProteinMCP: An Agentic AI Framework for Autonomous Protein Engineering.” *bioRxiv*, ahead of print. https://doi.org/10.64898/2026.03.11.711149.

Zhang, Chengxin, and Jeffrey Skolnick. 2024. *USalign: Universal Structure Alignment of Monomeric and Complex Proteins*. https://github.com/pylelab/USalign.