COMPRESSING PROTEIN LANGUAGE MODELS: A UNIFIED BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models are becoming central tools across a growing range of tasks, and their scale and adoption are increasing in tandem, making efficiency a main concern, with compression as a popular way to achieve it. Yet what it actually costs remains unmeasured. We present the first compression benchmark spanning model families, training objectives, and compression paradigms at a common protocol, evaluated at reconstruction, representation geometry, and downstream biology. For deployment, calibration-aware 4-bit is near-lossless at every scale, equivalence-tested with pre-declared margins; 3-bit is free only at the top of a family; and 2-bit collapses the zero-shot channel except at the very top of a family. How much compression costs depends on the readout: a probe retrained on the compressed features survives to 2-bit at scale, whereas a reused head or SAE does not. Underlying this is a structural dissociation: compression largely preserves the representation and shifts its coordinates, so a label-free realignment restores coarse readouts like the probe and SAE, each with a different correction, while the fine-grained LM head reads exactly the part no such map recovers.

1 INTRODUCTION

Protein language models (PLMs) are now standard tools in computational biology. Encoders like ESM2 (Lin et al., 2023) and ESM-C (Candido et al., 2026) are used for their embeddings and for zero-shot variant-effect prediction, and autoregressive models like ProGen2 (Nijkamp et al., 2023) generate and score sequences. The best of these models have grown to billions of parameters, which makes them expensive to run, especially for academic biology labs, and has turned compression into a practical necessity. But the methods we have for compression, and most of what we know about how it degrades a model, come from language models in different domains, and it is not obvious that any of it carries over. PLMs have a vocabulary of only twenty-odd amino acids; many are bidirectional rather than autoregressive; and they are judged by their own set of biological assays.

There is a second reason the question matters. PLMs are usually deployed behind readouts that stay fixed after pretraining. A variant-effect score reads the frozen LM head, and the community routinely reuses published sparse autoencoders, contact heads, and hidden layers on top of a frozen encoder. A readout that is never re-fit cannot absorb a change in the features it reads, so whatever compression does to the representation shows up immediately as lost accuracy. Existing work rarely addresses this: it usually compresses a single model with one method and reports aggregate accuracy. What we actually want to know is what compression does to the representation itself, and whether that damage can be undone without labels.

To study this we built a compression benchmark for PLMs that is unified across the axes that matter: three model families and two training objectives (ESM2, ESM-C, ProGen2), from 8M to 15B parameters; nineteen quantization and pruning configurations; and three evaluation levels, from reconstruction through representation geometry to downstream biology.

We use the benchmark to answer the deployment question directly. Calibration-aware 4-bit is near-lossless at every scale, 3-bit is free only at the top of a family, and 2-bit collapses the zero-shot channel except at the very top of a family; since skill saturates with scale, the memory-efficient move is the smallest model past the knee at 4-bit. What then breaks depends on the readout rather than on compression alone: a probe retrained on compressed features survives to 2-bit at scale, whereas a reused head or SAE does not.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Behind these results is one explanation, a *structural dissociation* (Sec. 5.3, Fig. 1). Compression largely preserves the representation and shifts its coordinates, so a label-free correction can restore a frozen readout; but the *minimum sufficient* correction differs by readout: the SAE recovers with the cheapest (sometimes just a rescale), the probe needs a basis alignment or refit, and no single map serves both (a fitness readout flips sign under a map that helps others). The zero-shot head is the case no label-free correction reaches, even a capacity-verified nonlinear adapter. That different readouts of the same activations are damaged along different axes is what a single-readout study cannot see.

Contributions. *(1)* The first compression benchmark spanning PLM families, training objectives, and compression paradigms at a common protocol (13 models $\times$ 19 configurations $\times$ 3 evaluation levels), with a runnable harness and per-cell records. *(2)* Two deployment findings: an equivalence-tested 4-bit rule with pre-declared margins and frozen-readout heterogeneity. *(3)* The explanation, a structural dissociation across readouts of the same activations, with the zero-shot LM head the case no label-free repair reaches.

2 Related Work

Protein language models and their evaluation. Transformer PLMs trained by masked-language modeling (Rives et al., 2021; Lin et al., 2023; Candido et al., 2026) and their successors learn transferable representations, while autoregressive models (Nijkamp et al., 2023) additionally generate sequences. Two evaluation modes dominate: a *zero-shot* mode reading the likelihood surface (masked-marginals (Meier et al., 2021); autoregressive log-likelihood), benchmarked by ProteinGym (Notin et al., 2023); and a *supervised* mode training a probe on frozen embeddings, FLIP (Dallago et al., 2021), DeepLoc (Almagro Armenteros et al., 2017), and secondary structure against CB513 (Klausen et al., 2019). We evaluate both.

Compressing large neural networks. Post-training quantization: GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024b), bitsandbytes INT8/NF4 (Dettmers et al., 2022; 2023), HQQ (Badri & Shaji, 2023). Pruning: magnitude, Wanda (Sun et al., 2024), SparseGPT (Frantar & Alistarh, 2023), and 2:4 semi-structured sparsity (Mishra et al., 2021). Distillation (Hinton et al., 2015) with modern depth/width reduction (Muralidharan et al., 2024).

Aligning a moving representation. Rotation-based quantization inserts orthogonal transforms that leave the function unchanged (for orthogonal Q, $Wx = (WQ^\top)(Qx)$), so the quantizer sees an outlier-free basis (QuaRot, SpinQuant, QuIP, DuQuant (Ashkboos et al., 2024b; Liu et al., 2025; Chee et al., 2023; Lin et al., 2024a); SliceGPT (Ashkboos et al., 2024a) applies the same function-preserving rotation for structured pruning), *exact* invariances applied *before* quantization where ours is *approximate* and fit *after*. Classical bias correction fixes the per-channel mean shift (Nagel et al., 2019); label-free feature distillation matches compressed to full-precision activations as a recovery *method* (Elthakeb et al., 2020; Tu et al., 2026); and embedding-drift adapters rescue a frozen retrieval index across a model upgrade with exactly our rungs: Procrustes, low-rank affine, a residual MLP, and a diagonal-scaling ablation (Vejendla, 2025). The construct is a ladder of label-free maps to rescue a frozen readout, and there is precedent for treating such representation change as structured rather than arbitrary: fine-tuning distorts pretrained features within the subspace spanned by the training data while leaving orthogonal directions largely intact (Kumar et al., 2022). Concurrent work independently decomposes compression drift into scale, shape, and head terms (Lin & Sun, 2026). However, prior work fits one map for one readout, so it cannot see that the *minimum sufficient correction differs across readouts of the same activations*, the falsifications we establish in Sec. 5.3; in model-stitching terms (Bansal et al., 2021) the informative statistic is the weakest stitch that works *per readout*.

Unified studies and the PLM gap. A few recent studies compare pruning, quantization, and distillation for LLMs under one protocol and report descriptive regularities: pruning hurts knowledge-intensive tasks long before perplexity notices while quantization stays comparatively benign (Jaiswal et al., 2024; von Rad et al., 2026; Yang et al., 2024); calibration data matters far more for pruning than for quantization (Williams & Aletras, 2024); and compression hits the long tail hardest while top-line metrics hold (Hooker et al., 2019; 2020). We build the PLM counterpart of these unified studies and reproduce the protein analogs of the first two (Secs. 5.1 and 5.4). Our benchmark

catalogues which capability survives across model families, training objectives, and compression paradigms, and we additionally ask why the split happens, giving a *structural dissociation* account (Sec. 5.3). The nearest prior framing is the "knowledge bias" of von Rad et al. (2026), who see knowledge preserved and reasoning degraded in LLMs; in PLMs the same capability survives or fails with the readout, and we pin down its axes and add a proper equivalence test on the readout that is actually deployed.

Compressing protein language models. Prior work on PLM compression exists but is narrow along exactly the axes we vary. On the quantization side, Peng et al. (2023) study post-training quantization of a single ESMFold model, and Hu et al. (2026) target 3-bit KV-cache quantization for inference; on the distillation side, Airas & Zhang (2026) distill ESM3 into a task-specific solvent model. Each treats one model family, one compression paradigm, and one downstream use, and none asks what compression does to the *representation* or whether the damage is recoverable. To our knowledge no prior work compresses PLMs across families, objectives, and paradigms at a common protocol, nor gives a recalibration-based account of what compression does to their representations.

3 Benchmark Design

Models (13). ESM-C 300M/600M/6B (Candido et al., 2026); ESM2 8M/35M/150M/650M/3B/15B (Lin et al., 2023); ProGen2 small (151M)/base (764M)/large (2.7B)/xlarge (6.4B) (Nijkamp et al., 2023). Two training objectives (masked-LM for the encoders, autoregressive for ProGen2), $\sim$3 orders of magnitude of scale.

Methods. *Quantization* (13, weight-only unless noted): RTN, bitsandbytes int8/NF4, HQQ, GPTQ, and AWQ across an 8/4/3/2-bit ladder. *Pruning* (6): Wanda / SparseGPT / magnitude at 50% unstructured and 2:4 semi-structured. App. E describes every method; calibrated methods use the authors' reference solvers run block-sequentially, with PAD-masked Hessians, a uniform per-model skip set for group-64-incompatible heads, and weight memory counted as analytic bits$\times$params or real packed bytes (full details in App. A). A matched-magnitude Gaussian-noise control (`gauss-int2/3`) is used only in the recoverability/severity analyses. bnb-int8 is the one W8A8 (activation-quantized) point and is near-lossless (supervised 0.974, zero-shot 0.992), indicating PLM activations tolerate int8 without SmoothQuant-style repair.

Evaluation data. L1/L2 on a frozen 2048-sequence probe set disjoint from calibration (verified). L3: ProteinGym substitutions (a deterministic 40-assay subset chosen for cross-family tractability, *not* leaderboard-comparable; we use within-model retention), FLIP, DeepLoc, and secondary structure (homology-reduced, exact-leakage-filtered). Encoders score variants by masked-marginals (Meier et al., 2021); ProGen2 by *bidirectional full-LL*, the mean of the full autoregressive sequence log-likelihood scored N$\rightarrow$C and on the reversed sequence C$\rightarrow$N, the two orientations ProGen2 is trained on, both at a common 4000-mutant cap; absolute zero-shot is therefore not cross-family comparable and we use within-family retention.

4 Metrics

L1: pseudo-perplexity, logit-KL(FP$\|$compressed), top-1 agreement, from *fp32* references (fp16 storage floors high-fidelity methods). **L2:** linear CKA (Kornblith et al., 2019), kNN-conservation, per-layer CKA. **L3:** ProteinGym within-protein-then-across Fisher-z Spearman; FLIP Spearman; DeepLoc/SSP accuracy; bootstrap CIs. **Retention** = compressed/FP per metric. Accuracy-based retention (DeepLoc, SSP) is floor-compressed toward 1 by the task's majority-class baseline ($\approx$0.1 DeepLoc, $\approx$0.44 SSP Q3). **SAE functional retention** is the median over live features (those firing on $>0.5\%$ of residues) of the compressed mean feature activation divided by the full-precision mean, a magnitude proxy for functional-feature preservation rather than a direct functional assay (we corroborate the severity conclusion with a non-circular functional assay, concept-detection AUC and a scale-invariant dictionary match, in Table 14). The main matrix scores ProteinGym on a deterministic 40-assay subset (chosen for cross-family tractability); the 6B equivalence test (Sec. 5.2) uses a 72-protein expanded set. Mechanism statistics use partial Spearman and leave-one-family-out ridge cross-validation.

5 RESULTS

5.1 THE CHANNEL ASYMMETRY, AS OBSERVED

The most visible pattern across the matrix is a sharp asymmetry between the two channels (Table 3, Fig. 2). Aggressive compression collapses zero-shot variant-effect prediction while leaving frozen-representation performance almost untouched. On ESM-C-300M, ProteinGym Spearman falls from 0.44 at full precision to 0.16 (2:4 Wanda) and 0.11 (2:4 magnitude), yet the *same* models retain 0.60–0.64 on FLIP versus 0.65 at full precision. The pattern is not ESM-C-specific: it holds for ESM2 and for ProGen2, where the zero-shot channel is likewise fragile while FLIP/DeepLoc stay within noise (per-family figures, Figs. 6, 7; per-task supervised breakdown, Table 23). Averaged over all 19 configurations and 13 models the split is ≈ 0.96 supervised vs ≈ 0.70 zero-shot; excluding the three collapsed 2-bit quantizers, which drag the zero-shot mean down, it is ≈ 0.98 vs ≈ 0.79. Practically, a frozen encoder feeding a trained probe tolerates aggressive compression; likelihood-based uses do not. Section 5.3 shows that most of this gap is not the two channels losing different amounts of information: a label-free realignment of the shared representation (a per-channel rescale or rotation, which cannot add information) brings the SAE and the re-fit probe back toward FP, whereas the zero-shot LM head is the one readout no such correction restores.

5.2 SCALE BUYS ROBUSTNESS

At 6B, 3-bit quantization is free for zero-shot variant effect. On a paired bootstrap over 72 matched ProteinGym proteins, GPTQ-int3 is statistically equivalent to full precision: mean difference -0.003 Spearman, TOST $p{=}0.006$ against an equivalence margin $\delta{=}0.02$; GPTQ-int4 is indistinguishable at $+0.000$, $p{<}10^{-3}$. The floor is sharp: GPTQ-int2 collapses to -0.236 and is not equivalent. Structured pruning is not free even here: 2:4-Wanda costs a real 0.068 Spearman, roughly 17% of the 6B FP zero-shot, far less than at 300M but well outside the margin. We choose δ to match the full-precision model's own across-assay noise, test equivalence with TOST (Lakens, 2017) at $\alpha{=}0.025$, and Holm-correct across the four methods within each model. App. C gives the margin justification, per-method intervals, and the full margin-sensitivity analysis.

The small models take a real 3-bit hit. Running the same paired test at every scale (Fig. 4, Table 1; ESM-C and ESM2 on the 72-protein matched set, ProGen2 on its in-context subset), the 3-bit penalty shrinks with scale as a *trend*, reaching TOST-equivalence to FP only at the top of each family, ESM-C-6B and ESM2-15B (both $\Delta = -0.003$). The approach is not strictly monotone. Two-bit stays significantly negative at every scale, though its magnitude softens as models grow. Scale thus buys robustness on the deployed readout, in line with larger models' greater capacity (Huang et al., 2026) and the larger-models-compress-better trend UniComp reports for LLMs (von Rad et al., 2026). One caveat on the mechanism: the models that tolerate compression best are also furthest *past* the accuracy plateau, so this may be *redundancy* rather than raw capacity; the smallest model on the plateau is by construction the least redundant, which is why it wants 4-bit rather than 3-bit.

The trend continues upward: at 15B (ESM2) even 2-bit is tolerated on a *frozen* readout, the transferred probe and the zero-shot head retain 0.87 and 0.76 of full precision with *no* recalibration (Table 12), so the deployable-precision frontier keeps dropping with scale.

big-compressed vs small-FP at matched bytes. The real choice is rarely between 3- and 4-bit in the abstract; it is between a large model compressed and a small model at full precision for a fixed memory budget. Fig. 5 plots absolute zero-shot Spearman against analytic weight memory for every model$\times$method, and two things stand out. First, the cheapest way to hold quality is calibration-aware 4-bit: near-lossless at *every* scale and a $4\times$ shift left, whereas 3-bit only preserves quality once a model is large. Second, zero-shot skill shows strong diminishing returns past $\sim$600M along the parameter axis, a <0.01 FP-Spearman gain over a $\sim 20\times$ parameter increase on this matched set. So the memory-optimal choice is the smallest model past the knee at 4-bit, ahead of a compressed big model. We say diminishing returns rather than a hard plateau because the public full-ProteinGym leaderboard and data-axis scaling still improve modestly (Spinner et al., 2025); the claim is only about the parameter axis on this set. One real limitation: the rule concerns *weight storage*, and says nothing about wall-clock latency. Low-bit points are fake-quantized to isolate quality from kernel maturity, and fast kernels exist at 4-bit and int8 but not for int3/int2, so a memory-optimal choice

Table 1: *Per-scale paired zero-shot test across all 13 models.* FP *absolute* Spearman (so ratios are interpretable) and paired ΔSpearman (method−FP); † marks TOST-equivalence to FP at δ=0.02, α=0.025 (positive evidence of equivalence, not mere non-significance). int3 reaches equivalence at the top of both families (ESM-C-6B -0.003, p=0.006; ESM2-15B -0.003, $p<10^{-3}$) while small models take a real 3-bit hit; int2 stays negative throughout. The lone small-model daggers (ProGen2-small, ESM2-8M) reflect near-chance baselines (wide CIs) rather than genuine robustness. ESM-C/ESM2 use the 72-protein matched set; ProGen2 uses its in-context subset (n=64/69/64/64), agreeing on its subset at every scale.

Model	FP abs. (Spearman)	int4	int3	int2	w2:4
		Δ(method–FP); † = TOST-equiv. (δ=0.02)			
EC.3	0.393	-0.013	-0.131	-0.376	-0.228
EC.6	0.392	-0.005†	-0.093	-0.349	-0.188
EC6B	0.400	+0.000†	-0.003†	-0.236	-0.068
E8	0.184	-0.013†	-0.054	-0.172	-0.184
E35	0.265	-0.015	-0.082	-0.261	-0.243
E150	0.335	-0.009†	-0.072	-0.318	-0.209
E650	0.392	-0.009†	-0.054	-0.352	-0.192
E3B	0.397	-0.004†	-0.022	-0.265	-0.084
E15B	0.400	-0.002†	-0.003†	-0.107	-0.055
PGs	0.265	+0.051	+0.001†	-0.260	-0.175
PGb	0.382	-0.001†	-0.016	-0.279	-0.203
PGl	0.325	-0.009†	-0.018	-0.108	-0.041
PGx	0.385	-0.002†	-0.003†	-0.124	-0.192

need not be time-optimal, especially for 3-bit. A first real-kernel point (Table 2) confirms weight-only 4-bit is near-lossless and latency-neutral at 3B/15B while cutting peak memory 3.2–3.6×, and that at matched bytes a smaller full-precision model is faster than a larger 4-bit one.

Unsupervised divergence is a correlational damage proxy. A cheap unsupervised divergence over one forward pass, L1 logit-KL and L2 (1−linear-CKA), correlates with downstream loss and underlies the ranking pre-screen (App. F; full per-cell L1/L2 in Tables 6–9); the damage concentrates in the late layers tasks read from (Fig. 8). Because L1 and L2 are collinear ($r \approx 0.83$) and CKA on 2048 sequences is a coarse instrument, we read this as *correlation* only, and do not claim geometry *drives* the damage over output divergence; full partial-correlation numbers are in App. F.

5.3 Readouts of the same representation dissociate along different axes

Different readouts need different corrections to undo the same compression. Compression moves the representation; for each frozen readout we ask whether a cheap *label-free* map can move it back. We try a ladder of increasingly expressive maps, none → a per-channel rescale → an orthogonal basis alignment → a full linear re-fit, and measure how much of each readout returns across ESM-C-300M/600M/6B and int4/int3/int2 (Fig. 1; per-cell numbers in Tables 10, 11). Two things follow. First, above a scale-dependent floor (below) a fixed map restores most readouts, so for them compression re-expresses the representation rather than destroying it; the zero-shot head is the exception (below). Second, the *dissociation*: the readouts differ in *which* map they need. The SAE needs the least, usually just a per-channel rescale, a second-moment fix that refines classical bias correction (Nagel et al., 2019; App. C); we do *not* claim a rescale suffices for the SAE in general, only that its correction is reliably cheaper than the probe's. The DeepLoc probe needs more, a basis alignment beyond a rescale (a ridge fit does as well, Table 16), so what it sees is a general change of basis that goes past pure rotation. The recovery is clearest for structured pruning, whose larger 6B damage is almost entirely a coordinate change: 2:4-Wanda lifts the SAE $0.67 \to 1.01$ and the probe $0.63 \to 0.98$ (Table 10). A separate scale-invariant assay (Table 14) shows the SAE feature *identities* likewise survive through 3-bit, with real loss only at 2-bit, reconciling gloomier results from uncalibrated RTN (Duan, 2026): calibrated GPTQ keeps feature identity well past their floor.

Is the channel asymmetry just an artifact of re-fitting? The result has a confound to rule out: supervised probes are re-fit on the compressed model while the head is left frozen, so the retrained readout could look healthy simply because it was retrained, whatever compression did to the features. The real question is whether the surviving supervised performance reflects information the compressed features still hold, or merely the act of re-fitting. To decide, we compare three readouts on the *same* compressed features: (i) the original probe applied *frozen*; (ii) that frozen probe after a *label-free* step that only realigns the compressed activations back to their full-precision counterparts, fit on a generic corpus with no task labels and its strength set by reconstruction error; and (iii) a probe fully *re-fit with labels*. Averaged over the nine encoders, the frozen probe (i) degrades at 2-bit and 2:4; the label-free realignment (ii) undoes a substantial part of that drop *without any labels* — roughly half under 2:4 pruning, less at 2-bit — and adds little at 3–4-bit, where little is broken; the labelled re-fit (iii) recovers more still (Table 22). That a purely label-free step already recovers a meaningful fraction shows the surviving structure is genuinely present in the compressed features rather than produced by re-fitting; the remainder is what labels add. The realignment is computed against the full-precision model, so it is a mechanism result and not a deployable recipe.

Recovery is full only above a bit-width floor, and the floor drops with scale. At 3–4 bits the SAE and probe return to full precision; at 2-bit a real residual remains that no label-free map removes, so recovery there is only *partial*. Where the floor sits depends on size: ESM-C-300M already needs correction at 3-bit while 6B holds uncorrected, and in a second family ESM2-3B breaks at 2-bit where ESM2-15B tolerates it (Tables 11, 12), the scale-buys-robustness of Sec. 5.2.

The zero-shot LM head is the one readout no label-free map fully restores. At 300M-int3 the best correction of its input lifts it only from 0.68 to ≈ 0.72, while a probe reading the *same* trunk returns to near-FP (Table 11). This is not a weakness of our maps: quantizing the head's own weights barely hurts (Table 18), and even a capacity-verified nonlinear adapter lifts int3 only to 0.66 (App. A), so the deficit lives in the trunk the head reads, upstream of the head and the optimizer. The reason is *sensitivity*: the head scores fine-grained per-token margins that an MSE-fit map does not preserve (a logit-level objective can (Lee et al., 2026)), so the same residual a pooled probe shrugs off breaks the head. Injecting that residual into full-precision activations, against a norm-matched random control, confirms the damage is set by its *direction* rather than its size (App. F).

Recoverability tracks error size; collapse tracks its direction. *Recoverability* — how much of the damage the correction ladder undoes — depends only on the *size* of the perturbation: Gaussian noise matched in Frobenius norm to the int-b quantization error recovers at 3-bit and floors at 2-bit, as real quantization does (Table 13). *Catastrophic collapse* is a different phenomenon, governed by the error's *direction* rather than its *size*. Naive RTN-int4 collapses ProGen2-xlarge to 0.004, yet at the *same* error magnitude calibrated int4 holds ~ 0.39 and Gaussian noise at RTN's mean norm (0.18) leaves it healthy (0.374): identical size, opposite outcome. The reason is that RTN's error is *signal-correlated distortion* rather than additive noise, and it concentrates in one place: quantizing the outlier-heavy *tied output embedding* at int4 is what drives the collapse (sparing that one matrix restores pseudo-perplexity to ≈ 13; App. F). Calibration-aware methods and bitsandbytes never touch it — they quantize only `nn.Linear` modules, leaving embeddings full precision — which is exactly why they hold; a scalar magnitude cannot see the difference, and matched-norm Gaussian noise is harmless.

The break is in the trunk, not the head weights. Quantizing trunk and head *separately* (head = the output modules `lm_head`/decoder/contact, with the decoder tied to the embedding in ESM2; Table 18) localizes the collapse: trunk-only reproduces it in full, head-only leaves most intact (0.390 vs FP 0.441 at 300M-int2). We make no precision-budget claim: the head is only ≈ 0.1–0.3% of parameters, so there is nothing to reallocate, and we leave it in full precision because it is tiny, not because it is precision-critical.

One map does not fix every readout at once. Even inside the recoverable regime a *single* label-free map does not restore readouts together: at the deciding cell (300M-int3) one shared map leaves DeepLoc below its per-readout alignment and collapses AAV (Table 21). The correcting map is readout-specific, real per readout rather than one global rotation, which is why the ladder fits a fresh map per readout: readouts of differing sensitivity need different correction directions.

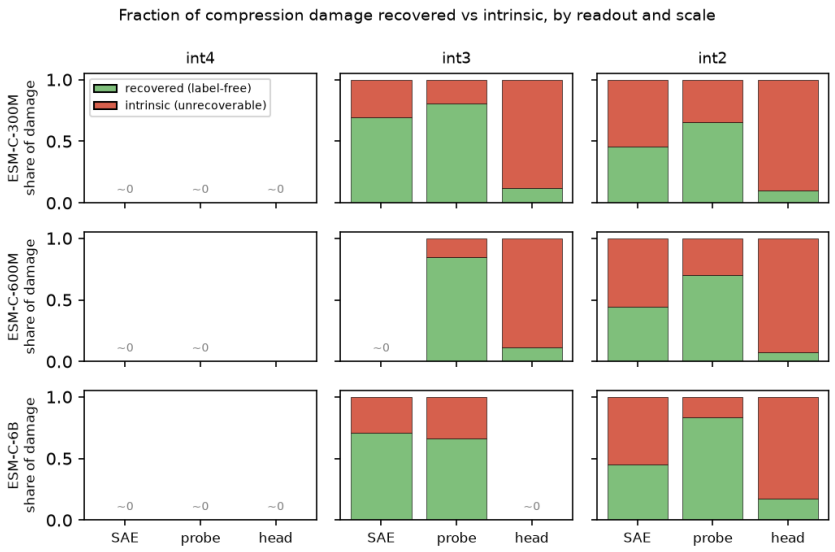


Figure 1: *Compression damage is mostly a recoverable change of coordinates, except for the LM head.* For each readout in each cell, the bar is normalized to that cell's *damage* (100%): the fraction a *label-free* map recovers (green; best of per-dim rescale / basis alignment / refit) versus the *intrinsic* part it cannot (red); cells with negligible damage are marked $\sim$0. Rows are ESM-C scale, columns bit-width. Two readings. **Dissociation** (within a cell): the SAE and probe are mostly green (their damage is a recoverable change of coordinates), while the zero-shot LM head is mostly red (intrinsic loss no label-free map removes). **The floor drops with scale** (down a column): the absolute damage shrinks toward 6B and more cells fall to $\sim$0, and int4 is near-lossless everywhere. The head uses the FP head on the aligned trunk. Per-cell numbers and the refit rung in App. C (Tables 10, 11).

5.4 Calibration-domain shift matters most for pruning

Matching the calibration domain to deployment helps pruning most and quantization least. On a general-vs-viral crossover, the match-gain (KL reduction from the matched domain) is largest for SparseGPT and Wanda and smallest for GPTQ (ESM-C-300M $0.042/0.013/0.012$; 6B $0.022/0.012/0.003$); within-domain composition is a null. This is the protein analog of UniComp's LLM finding (von Rad et al., 2026): calibration data matters more for pruning than for quantization.

Secondary findings (App. F). Three further results are detailed in the appendix: *(i)* a cheap pre-screen, one unlabeled forward pass ranks methods for a given model (within-family Spearman 0.87) but not across families (leave-one-family-out $R^2 < 0$); *(ii)* the familiar method ordering holds, with calibration-aware low-bit quantization dominating pruning at matched budget and naive 4-bit collapsing ProGen2-xlarge (RTN-int4 $\rightarrow$ 0.004) where calibrated 4-bit does not; and *(iii)* a shallow-MSA tail, where the low-N_{eff} third of assays degrades more under aggressive methods (Spearman$(N_{\text{eff}}, \text{retention})$ $+0.12$ to $+0.18$), a mild evolutionary-tail effect the average hides (Table 20).

6 Discussion

Why the readouts dissociate. The immediate reason is protocol: a probe is re-fit after compression and re-learns the new coordinates, while the frozen head keeps reading the old ones. But refitting is not the whole story, and that is the point: a *label-free* correction, which realigns the coordinates without any refit, restores the SAE and probe yet still leaves the head short. The deeper reason is *sensitivity*. The label-free map reconstructs the trunk only approximately (recon $R^2 \approx 0.68$ at int3); the pooled probe tolerates that residual, but the head reads a far finer signal, the per-position margin between two amino acids, which lives in exactly the part the reconstruction misses. The

same correction is thus good enough for the coarse readout while failing the fine one, which is also why no single map serves every readout (Sec. 5.3).

A decision rule. *(i)* For frozen encoders with probes *retrained* after compression, INT8 and calibration-aware 4-bit are near-lossless and 2:4 pruning holds supervised retention above 0.92; for a *fixed* probe, apply the label-free alignment above its recovery floor and retrain below it. *(ii)* For zero-shot or generation, calibration-aware 4-bit is the near-lossless default at every scale, 3-bit is free only at the top of a family, and naive 2-bit collapses this channel everywhere; since accuracy saturates with scale, pick the *smallest model on the plateau* at 4-bit rather than a compressed giant. *(iii)* If pruning, only 2:4 buys real memory, and match calibration to the deployment domain. *(iv)* "The floor" is two readout-specific thresholds: magnitude/probe readouts recalibrate down to $\approx$3-bit at 300M, the floor falling with scale to full 2-bit recovery only at the top (ESM2-15B; 6B-int2 remains a partial loss), but the deployed zero-shot head, already degraded at 300M-int3, where input correction leaves it at ≈ 0.72, with no tested label-free repair, should not go below 4-bit except at the top of a family.

7 Limitations

- *SAE coverage.* Public ESM-C SAEs exist only at 300M/600M/6B, so the SAE ladder is anchored there; ESM2/ProGen2 have no public SAE and use the two task rungs.
- *Probe-refit noise.* We do not average refits over seeds; bootstrap CIs bound the test-set component to ≈ 0.02–0.03, so we do not read sub-0.03 orderings within a ladder row.
- *Baseline and hard splits.* Accuracy retention is majority-baseline-compressed and the FLIP extrapolation splits are data-limited, so only their ratios are interpretable; we lead with the scale-clean FLIP-vs-ProteinGym Spearman contrast.
- *Latency.* We report analytic weight-memory plus one real-kernel snapshot (Table 2); a full throughput sweep across hardware, batch, and family is future work.
- *Objective and family.* ProGen2 is our only autoregressive family, so we attribute its behavior to ProGen2 specifically, not to autoregression (the two are confounded); cross-family absolute zero-shot is not comparable, so we use within-family retention.
- *Scope of the recovery result.* The alignment analysis covers the nine encoder models (ProGen2 is a decoder for which frozen probes are not the intended readout), needs FP activations (offline, not a deployable method), and is tested only within the readouts we use (no generation-quality or contact-map readout).

Ethics Statement

This benchmark involves no human subjects, no private data, and trains no new generative capability. The primary societal effect is efficiency: lowering the memory and energy cost of deploying PLMs for scientific use. As with any protein model there is a general dual-use consideration around sequence design, but compression neither adds generative capability nor removes existing safeguards, so we assess the marginal risk as negligible. All artifacts released are derived from public models and data.

Reproducibility Statement

All data and model checkpoints are public. We release the full codebase, the compression methods (RTN, bitsandbytes, HQQ, GPTQ, AWQ, and the pruners), the L1/L2/L3 metrics and cost meter, the run harness, and the analysis/figure scripts, together with the per-cell records and frozen evaluation sets.

LLM Usage Disclosure

We used generative AI tools (an LLM-based coding assistant) to implement analysis and figure-generation code, draft the reproducibility manifest, and help draft and edit prose, tables, and refer-

ences. We did not use generative AI tools to develop the theoretical framework, design the experimental methodology, or formulate the paper's core hypotheses. All AI-assisted text was reviewed and verified by the authors, including tracing every numerical claim back to its source record. We take responsibility for the final content of this work.

REFERENCES

Justin Airas and Bin Zhang. Knowledge distillation of a protein language model yields a foundational implicit solvent model. *Journal of Chemical Theory and Computation*, 2026. doi: 10.1021/acs.jctc.6c00574.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2401.15024.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. arXiv:2404.00456.

Hicham Badri and Appu Shaji. Half-quadratic quantization of large machine learning models. `https://mobiusml.github.io/hqq_blog/`, November 2023.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.07682.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology, 2026. URL `https://www.biorxiv.org/content/10.64898/2026.06.03.729735`. Preprint.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2307.13304.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. In *NeurIPS Datasets and Benchmarks*, 2021.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Evan Duan. Perplexity can miss SAE feature damage under quantization. *arXiv preprint arXiv:2606.03002*, 2026.

Ahmed Taha Elthakeb, Prannoy Pilligundla, Fatemeh Mireshghallah, Alexander Cloninger, and Hadi Esmaeilzadeh. Divide and conquer: Leveraging intermediate feature representations for quantized training of neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 2880–2891, 2020.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning (ICML)*, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Yue Hu, Junqing Wang, and Yingchao Liu. TurboESM: Ultra-efficient 3-bit KV cache quantization for protein language models with orthogonal rotation and QJL correction. *arXiv preprint arXiv:2603.26110*, 2026.

Jing Huang, Daniel Wurgaft, Rachit Bansal, Laura Ruis, Naomi Saphra, David Alvarez-Melis, Andrew Kyle Lampinen, Christopher Potts, and Ekdeep Singh Lubana. Why larger models learn more: Effects of capacity, interference, and rare-task retention. *arXiv preprint arXiv:2605.29548*, 2026.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. Compressing LLMs: The truth is rarely pure and never simple. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01382; introduces the LLM-KICK benchmark.

Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjærgaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Sønderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, and Paolo Marcatili. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2202.10054.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Jung Hyun Lee, June Yong Yang, Jungwook Choi, and Eunho Yang. LFQ: Logit-aware final-block quantization for boosting the generation quality of low-bit quantized LLMs. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.29756.

Chieh-Yen Lin and Shao-Hua Sun. PRISM: A geometric risk bound that decomposes drift into scale, shape, and head. *arXiv preprint arXiv:2605.11608*, 2026.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2406.01721.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024b.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.16406.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. Accelerating sparse deep neural networks. *arXiv preprint arXiv:2104.08378*, 2021.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *arXiv preprint arXiv:2407.14679*, 2024.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1325–1334, 2019.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023.

Shuang Peng, Fei Yang, Ning Sun, Sheng Chen, Yanfeng Jiang, and Aimin Pan. Exploring post-training quantization of protein language models. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 602–608, 2023.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Aviv Spinner, Erika DeBenedictis, and Corey M. Hudson. Scaling and data saturation in protein language models. In *ICML GenBio Workshop*, 2025. arXiv:2507.22210.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Fangbo Tu, Junhua Zhao, Chi Liu, Xin Chen, Haifeng Wu, Jian Wan, and Srinivasan Manoharan. Beyond output matching: Preserving internal geometry in NVFP4 LLM distillation. *arXiv preprint arXiv:2606.05682*, 2026.

Harshil Vejendla. Drift-adapter: A practical approach to near zero-downtime embedding model upgrades in vector databases. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Jonathan von Rad, Yong Cao, and Andreas Geiger. UniComp: A unified evaluation of large language model compression via pruning, quantization, and distillation. *arXiv preprint arXiv:2602.09130*, 2026.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10100–10118. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.544.

Ge Yang, Changyi He, Jinyang Guo, Jianyu Wu, Yifu Ding, Aishan Liu, Haotong Qin, Pengliang Ji, and Xianglong Liu. LLMCBench: Benchmarking large language model compression for efficient deployment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 87532–87544, 2024. arXiv:2410.21352.

A Experimental configuration

Quantization. RTN is symmetric per-output-channel round-to-nearest (one fp16 scale per output row; no sub-row grouping). GPTQ and SparseGPT use the IST-DASLab reference solvers run *block-sequentially* (each transformer block calibrated on the already-quantized upstream), with 512 calibration sequences from a mixed UniRef50/MGnify set, group size 128 (symmetric fp16 scale), blocksize 128, and 1% Hessian damping; Hessians $H = X^\top X$ are accumulated over real residues only (padding masked). AWQ is the full llm-awq path (per-channel activation-aware scale search plus weight clipping), group 128. HQQ uses the authors' half-quadratic solver at 2/3/4-bit, asymmetric (fp16 scale + zero-point) at group size 64. bnb-int8 is bitsandbytes LLM.int8() (W8A8 with fp16 outlier decomposition); bnb-nf4 is 4-bit NormalFloat with a blockwise (block 64) fp32 absmax (no double-quant). All methods skip the same group-64-incompatible layer set per model (two odd-shaped heads in ESM2; none in ESM-C/ProGen2).

Weight-memory accounting ($\times$comp). Every low-bit point is fake-quantized, so the memory axis is analytic and we apply one formula uniformly: *effective bits per weight* = raw bits + (scale [+ zero-point] metadata) / group size. Concretely, RTN $\approx b + 16/d_{\text{in}}$ (one fp16 scale per row, essentially b); GPTQ and AWQ $b + 16/128$ (fp16 scale, group 128); HQQ $b + 32/64$ (fp16 scale + zero, group 64); bnb-nf4 $4 + 32/64 = 4.5$ (fp32 absmax, block 64). The $\times$comp column in Table 3 is 16/(effective bits) blended with the fp16 non-quantized remainder (embeddings, norms, skipped heads); e.g. GPTQ-int2 $\rightarrow 7.5$, HQQ-2bit $\rightarrow 6.4$, RTN-int2 $\rightarrow 7.9$ (coarser per-row scale, so less metadata than a grouped 2-bit method, the finer-grouped methods buy quality with metadata), bnb-nf4 $\rightarrow 3.5$ once its absmax is counted. Earlier drafts omitted the bnb-nf4 absmax (reported 4.0); the corrected value is 3.5, and it does not change any ordering or conclusion.

Pruning. Wanda uses masked activation norms; SparseGPT uses the block-sequential Hessian solver; magnitude is calibration-free. We evaluate 50% unstructured and 2:4 semi-structured in the main matrix, and 75–95% unstructured for the severity dose-response (Fig. 3).

Two evaluation sets. The 13×19 breadth matrix (Table 3) uses the deterministic 40-assay subset (the only set scored on every model$\times$method); the per-scale paired analysis and equivalence tests use the larger 72-protein matched set, now including ProGen2 (Table 1). The two give consistent verdicts; the smaller paired subset carries a small-n fluctuation (6B GPTQ-int3 $+0.014$) that the 72-protein set resolves to a clean -0.003 equivalence (App. C). The $0.79/0.98$ channel-asymmetry ratios (Table 3) are the 40-assay mean; the equivalence verdicts are the 72-protein set.

Downstream probes (L3). SSP: per-residue logistic regression (C=1, max-iter 2000), trained on NetSurfP, tested on CB513 (5,000 train sequences, exact-leakage filtered). DeepLoc: pooled logistic regression (8,000/2,000). FLIP: pooled ridge (α=1, 12,000/4,000) on the hard splits (meltome clustered, GB1 three_vs_rest, AAV seven_vs_many). All probes standardize features. Zero-shot: masked-marginal scoring (encoders) or full-sequence log-likelihood (ProGen2), a 4,000-mutant cap; the main matrix uses a deterministic 40-assay ProteinGym subset, the paired equivalence test (Table 17) the 72-protein expanded set.

Recoverability ladder. The SAEs are the public BioHub ESM-C sparse autoencoders (released for the 300M/600M/6B backbones; source for reproduction); the 6B one is at layer 60 (k=64, codebook 16,384; live features fire on $>0.5\%$ of residues). Label-free alignments (Sec. 5.3) are fit on held-out unlabeled activation pairs and applied to disjoint test data: diagonal (per-dimension standardize to the FP distribution), orthogonal Procrustes, and ridge (full linear map, penalty swept 1–10^5, fit on $\leq$50k sampled residue pairs); "refit" retrains a fresh probe on compressed features. All three alignments are *label-free*: they are fit to reconstruct FP activations, so their one hyperparameter (ridge λ) is selectable by held-out *reconstruction* MSE on unlabeled pairs, using no task labels, and the main-text "linear" rung uses the parameter-free Procrustes map, so no rung in Fig. 1 requires any tuning. The ridge sweep (Table 16) is a supplementary check that a general-linear map matches the orthogonal one; the curve is flat over $\lambda \in [1, 10^3]$, so the conclusion is insensitive to λ regardless. Each alignment is fit *offline*, while the FP checkpoint is still available, from paired FP/compressed activations on task-independent calibration sequences, and adds a readout-specific $d \times d$ map: it is a *mechanistic recovery* experiment rather than an inference-time deployment method (it requires the FP activations and extra per-token compute; we make no latency claim for it). Equivalence uses a paired bootstrap over proteins with Holm correction; retention CIs use $10{,}000$ bootstrap resamples over models, per-item CIs 400.

A real-kernel deployment point. The body reports analytic weight memory, which is kernel-independent but an upper bound on realizable savings. Table 2 measures peak GPU memory and forward latency for full precision, weight-only 4-bit (bnb-nf4), and W8A8 int8 on a single H100 (batch 8, length 512), alongside absolute zero-shot Spearman on the 40-assay subset. Three points sharpen the deployment rule of Sec. 5.2. First, weight-only 4-bit is near-lossless (within noise of FP at every scale) and roughly *latency-neutral* at 3B/15B ($0.93\times$ FP throughput) while cutting peak memory 3.2–$3.6\times$, so the analytic 4-bit frontier is realized in practice. Second, W8A8 int8 saves memory but runs 1.4–$2.0\times$ *slower*, so "compressed" is not automatically "faster"; the effect is precision- and kernel-specific. Third, the matched-bytes trade favors the smaller model: ESM2-3B at 4-bit (1.93 GB) and ESM2-650M at full precision (1.55 GB) sit at a comparable memory budget and matched quality ($\rho\,0.456$ vs 0.447), yet 650M-FP is $2.7\times$ faster (28.8 vs 78.2 ms), and 650M at 4-bit is smaller still (0.59 GB) at 38.8 ms. This is a single-GPU, single-batch snapshot on ESM2 (stock-HF, kernel-friendly); a full throughput sweep across hardware, batch, and family remains future work.

Table 2: *Real-kernel deployment footprint* (ESM2, single H100, batch 8, length 512): measured peak GPU memory and forward latency, with absolute zero-shot ProteinGym Spearman (ρ) on the 40-assay subset. Weight-only 4-bit (bnb-nf4) is near-lossless and roughly latency-neutral at 3B/15B while cutting peak memory 3.2–$3.6\times$; W8A8 int8 is slower. At matched memory the smaller full-precision model wins on latency (ESM2-3B@nf4 vs ESM2-650M@FP: equal quality, $2.7\times$ latency gap).

model	precision	peak mem (GB)	ms/fwd	zero-shot ρ
ESM2-650M	FP (bf16)	1.55	28.8	0.447
ESM2-650M	nf4 (4-bit)	0.59	38.8	0.442
ESM2-650M	int8 (W8A8)	0.91	56.5	0.444
ESM2-3B	FP (bf16)	6.14	72.5	0.451
ESM2-3B	nf4 (4-bit)	1.93	78.2	0.456
ESM2-3B	int8 (W8A8)	3.32	128.0	0.450
ESM2-15B	FP (bf16)	31.16	263.3	0.441
ESM2-15B	nf4 (4-bit)	8.71	282.5	0.443
ESM2-15B	int8 (W8A8)	16.06	369.7	0.439

LM-head refit rung and its positive control. The head "refit" distills the FP head's per-residue log-probabilities onto a fresh linear head ($d \to V$, i.e. the model width to the 33-token vocabulary) reading the compressed trunk activations, with *no* biological labels, the same information the probe rung is given, applied to the head. We fit by minimizing forward cross-entropy to the FP soft targets with Adam (lr 5×10^{-3}, 600–800 steps, batch 8192 residues, on $\leq$50k sampled residues); the loss plateaus by $\sim$400 steps (final-vs-step-400 change $<0.5\%$), so the budget is not

the binding constraint. *Positive control (and why the refit is inconclusive):* the identical procedure run on the *uncompressed* trunk does *not* reach 1.0 for any head architecture we tried, bare-linear 0.885, LayerNorm+projection 0.888, and a tied-embedding LayerNorm+projection head 0.849. A from-scratch head simply cannot reproduce ESM-C's masked-marginal variant scores at this budget, so we could not build a capacity-unbounded refit and treat this route as *inconclusive*: the compressed refits (0.63 at 300M-int3, 0.09 at int2) fall below even these ceilings, but their absolute level is not diagnostic. The unconfounded evidence for the head exception is the input-correction ladder (rescale/Procrustes/ridge/KL and an output-side temperature+per-token affine), all applied to the *unchanged* FP head, which has no such ceiling. *Ceiling-free nonlinear test.* To rule out a linearity limit we also freeze the exact FP head and train an identity-initialized residual adapter $g(h) = h + \mathrm{MLP}(h)$ (two layers, last layer zero-init, so g starts as the identity and the FP-trunk control is exactly 1.0 by construction), fit label-free by distilling FP logits over the compressed trunk. It lifts 300M-int3 only from 0.64 to 0.66 and leaves int2 collapsed. Two positive controls make this diagnostic: the identity-initialization gives an exact 1.0 FP-trunk control (no capacity ceiling), and a *learnable-distortion* control shows the same recipe can learn a nonlinear correction, applying a known fixed nonlinear distortion (0.3 relative Frobenius) to the FP head input and training the identical adapter drives the head-output KL from 0.022 to 0.001 and top-1 agreement from 0.87 to 0.97. So the head's non-recovery on the compressed trunk is not an optimizer/capacity artifact; it is failure of the linear and nonlinear label-free repairs we test (we do not claim unrecoverability in principle). *Reconstruction R^2.* The best label-free linear map (ridge) reconstructs $0.90/0.68/0.33$ of FP per-residue activation variance at int4/int3/int2 (300M, held-out residues); the coarse probe tolerates this residual, the sensitive head does not. *The head exception replicates across families and scales.* No label-free input map (rescale, Procrustes, ridge, output-calibration) nor the identity-preserving adapter beats the uncalibrated head at any scale we tested, for ESM2 (650M/3B/15B) as well as ESM-C (300M/600M/6B); the adapter often *hurts*. Its severity drops with scale, mirroring the trunk floor: int3 head retention rises $0.85 \to 0.95 \to 1.00$ across ESM2-650M/3B/15B (near-lossless already at ESM-C-6B), while int2 collapses everywhere except ESM2-15B (0.76).

B FULL PER-MODEL RETENTION

Table 3: *Per-method retention* (all 13 models, mean [95% bootstrap CI]). Zero-shot = ProteinGym; supervised = mean(FLIP, DeepLoc, SSP); med. KL = median fp32 logit-KL; ×comp = weight-memory compression. The asymmetry sharpens at low precision: supervised stays ≥ 0.73 at 2-bit while zero-shot collapses (calibrated 2-bit 0.23, naive ≈ 0); 3-bit is mid-fidelity (0.86–0.89 at $\sim 5\times$). Because a mean can be dominated by one collapsed model (RTN-int4's 0.82 is a ProGen2-xlarge collapse averaged with twelve healthy cells), we add a per-model *median* (RTN-int4 0.90). These ratio means are *descriptive only*; inferential claims use the paired *absolute* ΔSpearman test (Table 1).

Method	Zero-shot ret.	(med.)	Supervised ret.	med. KL	×comp
rtn-int8	0.999 [0.99,1.00]	1.001	0.998 [1.00,1.00]	0.000	2.0
bnb-int8	0.992 [0.99,1.00]	0.996	0.974 [0.96,0.99]	0.006	2.0
rtn-int4	0.824 [0.67,0.93]	0.902	0.987 [0.98,0.99]	0.110	4.0
bnb-nf4	0.974 [0.95,0.99]	0.989	0.999 [1.00,1.00]	0.020	3.5
hqq-4bit	0.963 [0.94,0.98]	0.975	0.995 [0.99,1.00]	0.020	3.6
gptq-int4	0.991 [0.96,1.03]	0.983	0.998 [0.99,1.00]	0.020	3.9
awq-int4	0.987 [0.96,1.01]	0.992	0.996 [0.99,1.00]	0.019	3.9
hqq-3bit	0.884 [0.78,0.97]	0.935	0.987 [0.98,0.99]	0.095	4.6
gptq-int3	0.855 [0.77,0.93]	0.929	0.984 [0.98,0.99]	0.089	5.1
awq-int3	0.887 [0.81,0.97]	0.866	0.992 [0.99,1.00]	0.085	5.1
hqq-2bit	0.307 [0.14,0.48]	0.199	0.881 [0.85,0.91]	1.253	6.4
gptq-int2	0.233 [0.10,0.38]	0.081	0.843 [0.80,0.88]	0.461	7.5
rtn-int2	−0.003 [−0.08,0.07]	−0.028	0.737 [0.70,0.78]	3.829	7.9
wanda-50	0.704 [0.58,0.82]	0.771	0.985 [0.98,0.99]	0.157	1.0
sparsegpt-50	0.783 [0.69,0.87]	0.818	0.991 [0.98,1.00]	0.129	1.0
magnitude-50	0.504 [0.33,0.67]	0.566	0.935 [0.91,0.96]	0.440	1.0
wanda-2:4	0.461 [0.30,0.62]	0.461	0.962 [0.95,0.98]	0.330	1.9
sparsegpt-2:4	0.588 [0.45,0.72]	0.578	0.978 [0.97,0.99]	0.214	1.9
magnitude-2:4	0.314 [0.16,0.48]	0.260	0.923 [0.90,0.94]	0.516	1.9

Tables 4 and 5 give the complete 13-model $\times$ 19-method retention matrix underlying the model-averaged Table 3 (model codes: EC=ESM-C, E=ESM2, PG=ProGen2). Three patterns are visible per model rather than only in aggregate. First, the channel asymmetry is not an averaging artifact: in essentially every column the supervised matrix (Table 5) stays near 1.0 down to 2-bit and 2:4, while the zero-shot matrix (Table 4) spreads widely for the same cells, the asymmetry holds model-by-model. Second, *naive* low-bit quantization fails catastrophically at scale on ProGen2: RTN-int4 holds through ProGen2-large but collapses ProGen2-xlarge zero-shot (to ≈ 0.004), whereas calibration-aware 4-bit keeps it near 0.39, calibration rather than bit-width is what averts collapse in ProGen2's likelihood channel at scale. Third, damage from aggressive methods *recovers with scale*: across the ESM-C scale, 2:4-Wanda's *absolute* zero-shot Spearman rises from 0.16 at 300M to 0.37 at 6B, the same scale-buys-robustness trend that the equivalence test makes precise at 6B (Sec. 5.2). The matrix is complete, all 13×19 model$\times$method cells are populated.

Table 4: *Zero-shot retention per model* (ProteinGym, compressed/FP). Ratios on near-chance baselines (ProGen2-small, ESM2-8M, whose FP Spearman is ≈ 0.2) are noise-inflated and can exceed 1 or go negative from a ~ 0.05 swing; read the paired FP-absolute test of Table 1 for those cells, not the ratio.

Method	EC.3	EC.6	EC6B	E8	E35	E150	E650	E3B	E15B	PGs	PGb	PGl	PGx
rtn-int8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.01
bnb-int8	1.00	1.00	1.00	0.99	0.99	1.00	0.99	1.00	1.00	0.97	0.99	0.98	1.00
rtn-int4	0.89	0.90	0.98	0.67	0.78	0.87	0.92	0.97	0.98	0.82	0.94	0.97	0.01
bnb-nf4	0.97	0.97	1.01	0.85	0.94	0.97	0.99	1.01	1.01	0.95	1.00	1.00	1.01
hqq-4bit	0.97	0.97	1.00	0.88	0.91	0.97	0.97	0.99	1.00	0.89	0.99	0.97	0.99
gptq-int4	0.96	0.98	1.02	0.91	0.93	0.97	0.98	0.99	1.00	1.20	0.99	0.95	1.02
awq-int4	0.95	0.96	1.01	0.92	0.95	0.94	0.98	1.00	1.01	1.10	1.00	1.01	0.99
hqq-3bit	0.79	0.89	1.00	0.43	0.69	0.80	0.88	0.98	0.99	1.15	0.96	0.93	1.00
gptq-int3	0.63	0.76	1.02	0.64	0.65	0.70	0.85	0.94	1.00	1.03	0.93	0.93	1.03
awq-int3	0.80	0.86	1.02	0.67	0.64	0.81	0.87	0.99	1.00	0.74	0.96	1.17	1.01
hqq-2bit	0.12	0.18	0.69	-0.04	0.02	0.05	-0.22	0.46	0.90	0.46	0.20	0.60	0.60
gptq-int2	-0.04	0.08	0.35	0.05	-0.03	0.02	0.06	0.27	0.75	0.02	0.23	0.64	0.63
rtn-int2	0.20	0.06	-0.12	-0.06	0.17	-0.03	-0.13	-0.08	0.05	0.03	-0.29	-0.04	0.22
wanda-50	0.64	0.77	0.94	0.24	0.48	0.65	0.80	0.95	0.97	0.49	0.81	0.93	0.48
sparsegpt-50	0.67	0.73	1.02	0.54	0.56	0.64	0.82	0.93	0.99	0.53	0.87	0.88	1.00
magnitude-50	0.57	0.66	0.72	0.52	0.15	0.21	0.57	0.75	1.01	0.47	0.02	0.86	0.04
wanda-2:4	0.37	0.46	0.83	-0.12	0.04	0.31	0.48	0.76	0.89	0.33	0.37	0.79	0.47
sparsegpt-2:4	0.41	0.44	0.90	0.18	0.21	0.38	0.58	0.81	0.94	0.40	0.78	0.76	0.86
magnitude-2:4	0.24	0.35	0.63	0.26	-0.02	0.08	0.00	0.64	0.95	0.38	-0.12	0.54	0.13

Table 5: *Supervised retention per model* (mean of FLIP/DeepLoc/SSP, compressed/FP).

Method	EC.3	EC.6	EC6B	E8	E35	E150	E650	E3B	E15B	PGs	PGb	PGl	PGx
rtn-int8	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99	0.99	1.00
bnb-int8	0.97	0.99	0.98	0.97	0.97	0.96	0.94	0.91	0.99	1.00	1.00	1.00	0.99
rtn-int4	0.99	0.99	0.97	0.99	0.99	1.00	1.00	1.00	0.97	1.01	1.00	0.97	0.96
bnb-nf4	1.00	1.00	1.00	0.99	1.00	1.00	1.01	1.00	0.99	1.00	1.00	0.99	1.00
hqq-4bit	0.99	0.99	0.99	1.00	0.99	1.00	1.00	1.00	0.97	1.01	1.01	0.99	1.00
gptq-int4	1.00	1.00	1.01	0.99	0.98	1.00	0.99	1.01	0.99	1.00	1.00	1.01	1.00
awq-int4	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99
hqq-3bit	0.99	0.99	1.00	0.97	0.98	0.99	0.99	0.99	0.98	0.98	1.00	0.99	0.99
gptq-int3	0.96	0.98	1.01	0.98	0.97	0.98	0.99	0.99	0.98	0.98	0.99	0.99	1.00
awq-int3	0.99	0.99	1.00	0.99	0.98	0.98	1.00	1.01	0.99	1.00	0.99	0.97	1.00
hqq-2bit	0.86	0.87	0.92	0.84	0.78	0.80	0.81	0.95	0.95	0.92	0.90	0.91	0.94
gptq-int2	0.77	0.77	0.89	0.76	0.74	0.80	0.79	0.91	0.95	0.89	0.89	0.89	0.92
rtn-int2	0.66	0.73	0.80	0.80	0.79	0.72	0.69	0.78	0.62	0.87	0.63	0.78	0.71
wanda-50	0.97	0.98	0.99	0.97	0.97	0.98	0.99	1.00	0.99	1.01	0.99	0.98	0.99
sparsegpt-50	0.97	0.99	0.99	0.99	0.98	0.99	1.00	1.00	0.97	1.03	1.00	0.99	0.99
magnitude-50	0.95	0.95	0.98	0.94	0.93	0.92	0.95	0.96	0.97	0.98	0.84	0.94	0.86
wanda-2:4	0.96	0.97	0.98	0.95	0.92	0.96	0.92	1.00	0.98	1.00	0.96	0.97	0.96
sparsegpt-2:4	0.96	0.98	0.99	0.96	0.96	0.96	0.97	1.00	0.95	1.02	0.98	0.99	0.99
magnitude-2:4	0.91	0.94	0.96	0.94	0.91	0.89	0.84	0.93	0.97	0.99	0.93	0.92	0.88

Full L1 and L2 per cell. For completeness we also give the reconstruction-fidelity (L1) and representation-geometry (L2) numbers behind the unsupervised-divergence proxy (Sec. 5.2) and the layer-damage analysis (Fig. 8), for every model$\times$method cell. All compare the compressed

model against full precision; the FP baseline (`identity`) is 0 (KL) or 1 (top-1, CKA, kNN) by construction and is omitted.

Two geometry patterns stand out. Global geometry (CKA) survives far better than local neighborhood structure (kNN). Calibrated 4-bit keeps CKA near 0.99 yet reshuffles about a fifth of each residue's nearest neighbors (kNN ≈ 0.80), and the gap widens as compression tightens. Pruning shows this most sharply: at $2{:}4$ sparsity CKA stays near 0.8 while kNN drops to 0.2–0.4, so pruning keeps the global shape and scrambles the neighbor graph. Quantization at matched fidelity keeps both higher. We therefore read CKA as an optimistic geometry metric and kNN conservation as the stricter test.

Table 6: *L1 logit-KL per model* (KL of the compressed vs FP next-token distributions; lower is better, $0 =$ identical).

Method	EC.3	EC.6	EC6B	E8	E35	E150	E650	E3B	E15B	PGs	PGb	PGl	PGx
rtn-int8	0.002	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.012	0.004	0.000	0.155
bnb-int8	0.007	0.006	0.022	0.001	0.002	0.002	0.005	0.017	0.006	0.020	0.007	0.002	0.179
rtn-int4	0.199	0.183	0.099	0.062	0.056	0.048	0.067	0.058	0.110	2.379	3.930	0.215	86.300
bnb-nf4	0.062	0.053	0.024	0.020	0.018	0.014	0.017	0.014	0.007	0.070	0.035	0.018	1.201
hqq-4bit	0.050	0.042	0.020	0.021	0.018	0.014	0.017	0.015	0.006	0.057	0.031	0.015	0.355
gptq-int4	0.067	0.055	0.025	0.013	0.017	0.015	0.020	0.017	0.008	0.035	0.025	0.015	0.208
awq-int4	0.058	0.050	0.023	0.017	0.019	0.014	0.017	0.014	0.007	0.043	0.027	0.017	0.257
hqq-3bit	0.232	0.185	0.095	0.137	0.082	0.069	0.076	0.059	0.025	0.227	0.139	0.068	1.641
gptq-int3	0.302	0.260	0.122	0.067	0.083	0.077	0.089	0.073	0.038	0.110	0.104	0.078	0.282
awq-int3	0.232	0.205	0.091	0.070	0.083	0.063	0.068	0.054	0.030	0.128	0.116	0.085	0.402
hqq-2bit	1.253	1.251	0.749	1.957	1.311	1.283	1.200	0.302	0.136	2.807	1.474	0.421	13.947
gptq-int2	0.825	0.808	0.788	0.883	0.680	0.441	0.455	0.372	0.223	0.399	0.461	0.389	0.690
rtn-int2	5.644	7.160	65.428	3.257	2.153	2.135	2.120	3.829	2.805	6.783	1.230	4.998	74.197
wanda-50	0.341	0.288	0.157	0.112	0.121	0.098	0.116	0.090	0.080	6.161	1.495	0.182	5.947
sparsegpt-50	0.283	0.238	0.133	0.063	0.079	0.076	0.091	0.084	0.053	0.159	0.181	0.129	0.353
magnitude-50	0.584	0.557	0.440	0.363	0.196	0.318	0.271	0.367	0.117	8.702	1.675	0.451	5.892
wanda-2:4	0.558	0.541	0.330	0.295	0.300	0.201	0.223	0.168	0.172	4.651	0.964	0.335	6.540
sparsegpt-2:4	0.494	0.459	0.301	0.108	0.134	0.138	0.159	0.151	0.101	0.214	0.267	0.224	0.478
magnitude-2:4	0.728	0.706	0.583	0.302	0.259	0.329	0.516	0.476	0.255	5.803	1.441	0.494	24.530

Table 7: *L1 top-1 agreement per model* (fraction of positions whose argmax matches FP; higher is better).

Method	EC.3	EC.6	EC6B	E8	E35	E150	E650	E3B	E15B	PGs	PGb	PGl	PGx
rtn-int8	0.95	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.90	0.97	0.76
bnb-int8	0.91	0.92	0.89	1.00	1.00	1.00	0.99	0.98	0.99	0.77	0.86	0.93	0.75
rtn-int4	0.59	0.57	0.74	0.98	0.98	0.98	0.96	0.96	0.96	0.08	0.07	0.54	0.01
bnb-nf4	0.74	0.76	0.86	0.99	0.99	0.99	0.98	0.98	0.99	0.63	0.68	0.84	0.25
hqq-4bit	0.78	0.81	0.88	0.99	0.99	0.99	0.98	0.98	0.99	0.55	0.75	0.84	0.67
gptq-int4	0.79	0.82	0.88	0.99	0.99	0.99	0.98	0.98	0.99	0.73	0.80	0.88	0.71
awq-int4	0.79	0.81	0.88	0.99	0.99	0.99	0.98	0.98	0.99	0.66	0.77	0.86	0.57
hqq-3bit	0.56	0.63	0.75	0.96	0.97	0.97	0.97	0.96	0.98	0.43	0.58	0.67	0.23
gptq-int3	0.59	0.63	0.76	0.97	0.97	0.97	0.96	0.96	0.97	0.58	0.67	0.73	0.64
awq-int3	0.61	0.65	0.78	0.98	0.97	0.98	0.97	0.97	0.98	0.51	0.60	0.72	0.64
hqq-2bit	0.09	0.22	0.35	0.22	0.45	0.50	0.58	0.94	0.96	0.27	0.02	0.26	0.01
gptq-int2	0.24	0.28	0.40	0.61	0.75	0.87	0.88	0.91	0.94	0.28	0.31	0.39	0.48
rtn-int2	0.08	0.08	0.00	0.03	0.07	0.06	0.05	0.04	0.06	0.14	0.02	0.00	0.08
wanda-50	0.53	0.59	0.73	0.97	0.96	0.97	0.96	0.96	0.95	0.01	0.12	0.60	0.03
sparsegpt-50	0.61	0.66	0.75	0.98	0.98	0.98	0.97	0.96	0.97	0.54	0.59	0.69	0.62
magnitude-50	0.30	0.35	0.54	0.90	0.97	0.92	0.92	0.87	0.94	0.02	0.09	0.34	0.03
wanda-2:4	0.38	0.42	0.62	0.92	0.93	0.96	0.96	0.95	0.94	0.00	0.07	0.45	0.01
sparsegpt-2:4	0.47	0.52	0.64	0.98	0.97	0.97	0.96	0.94	0.96	0.49	0.53	0.61	0.56
magnitude-2:4	0.21	0.28	0.46	0.95	0.96	0.96	0.88	0.87	0.89	0.01	0.13	0.26	0.01

C Recoverability ladder and severity: full numbers

Table 10 gives the per-readout, per-method retention behind Fig. 1. Reading each row left to right shows which correction each readout needs: the SAE row is flat from *rescale* onward (a per-dim gain is enough), the probe row only jumps at *linear*/*refit* (a basis change is needed), and the LM-head row stays low (no label-free correction of its input brings it back). The ladder is informative wherever a readout breaks, and where that happens tracks scale: at 6B the readouts sit near 1.0 through 3-bit

Table 8: *L2 linear-CKA per model* (representational similarity of compressed vs FP hidden states; higher is better, 1 = identical geometry).

Method	EC.3	EC.6	EC6B	E8	E35	E150	E650	E3B	E15B	PGs	PGb	PGl	PGx
rtn-int8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
bnb-int8	0.999	0.999	0.995	0.999	0.999	0.999	0.995	0.979	0.991	0.997	0.999	1.000	0.995
rtn-int4	0.977	0.972	0.989	0.887	0.932	0.972	0.967	0.979	0.691	0.901	0.926	0.823	0.652
bnb-nf4	0.994	0.995	0.998	0.972	0.983	0.992	0.992	0.995	0.994	0.990	0.992	0.998	0.986
hqq-4bit	0.994	0.996	0.998	0.967	0.978	0.991	0.993	0.993	0.999	0.991	0.994	0.998	0.987
gptq-int4	0.993	0.995	0.997	0.986	0.985	0.993	0.994	0.996	0.999	0.995	0.995	0.999	0.987
awq-int4	0.994	0.995	0.998	0.970	0.981	0.994	0.994	0.996	0.999	0.991	0.991	0.998	0.985
hqq-3bit	0.965	0.972	0.994	0.787	0.876	0.951	0.969	0.985	0.995	0.961	0.971	0.989	0.952
gptq-int3	0.910	0.921	0.985	0.913	0.868	0.947	0.966	0.972	0.996	0.975	0.977	0.985	0.950
awq-int3	0.958	0.962	0.989	0.877	0.857	0.959	0.968	0.982	0.989	0.948	0.951	0.978	0.905
hqq-2bit	0.529	0.490	0.853	0.286	0.206	0.360	0.181	0.632	0.916	0.732	0.491	0.734	0.736
gptq-int2	0.290	0.386	0.546	0.194	0.110	0.458	0.293	0.477	0.887	0.590	0.756	0.691	0.669
rtn-int2	0.148	0.170	0.162	0.268	0.256	0.312	0.041	0.196	0.088	0.072	0.049	0.064	0.016
wanda-50	0.918	0.940	0.968	0.717	0.742	0.902	0.928	0.950	0.977	0.875	0.865	0.891	0.807
sparsegpt-50	0.914	0.929	0.972	0.808	0.793	0.917	0.941	0.954	0.985	0.913	0.890	0.951	0.876
magnitude-50	0.834	0.885	0.948	0.534	0.439	0.564	0.726	0.860	0.959	0.719	0.065	0.558	0.208
wanda-2:4	0.736	0.770	0.893	0.526	0.486	0.734	0.529	0.843	0.898	0.762	0.545	0.597	0.538
sparsegpt-2:4	0.712	0.730	0.888	0.694	0.614	0.772	0.767	0.874	0.952	0.872	0.823	0.851	0.769
magnitude-2:4	0.708	0.736	0.913	0.511	0.406	0.433	0.238	0.609	0.877	0.564	0.355	0.443	0.271

Table 9: *L2 kNN neighborhood conservation per model* (fraction of a residue's FP nearest neighbors preserved after compression; higher is better).

Method	EC.3	EC.6	EC6B	E8	E35	E150	E650	E3B	E15B	PGs	PGb	PGl	PGx
rtn-int8	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.98	0.92	0.93	0.97	0.94
bnb-int8	0.93	0.93	0.91	0.93	0.93	0.90	0.86	0.71	0.88	0.79	0.85	0.93	0.82
rtn-int4	0.63	0.62	0.74	0.50	0.57	0.64	0.59	0.66	0.43	0.26	0.36	0.39	0.13
bnb-nf4	0.80	0.82	0.88	0.69	0.76	0.79	0.79	0.83	0.90	0.64	0.70	0.86	0.71
hqq-4bit	0.80	0.82	0.88	0.67	0.76	0.78	0.78	0.82	0.92	0.67	0.74	0.88	0.73
gptq-int4	0.78	0.79	0.86	0.79	0.80	0.80	0.79	0.84	0.91	0.71	0.75	0.88	0.71
awq-int4	0.79	0.81	0.87	0.74	0.77	0.80	0.81	0.84	0.91	0.65	0.70	0.87	0.71
hqq-3bit	0.56	0.61	0.76	0.39	0.50	0.57	0.59	0.70	0.84	0.43	0.51	0.75	0.50
gptq-int3	0.48	0.49	0.66	0.55	0.56	0.59	0.56	0.64	0.82	0.48	0.52	0.73	0.48
awq-int3	0.56	0.58	0.73	0.51	0.53	0.60	0.61	0.70	0.80	0.36	0.42	0.67	0.38
hqq-2bit	0.13	0.15	0.36	0.09	0.07	0.06	0.06	0.23	0.56	0.11	0.06	0.33	0.16
gptq-int2	0.08	0.12	0.21	0.05	0.03	0.12	0.10	0.21	0.46	0.06	0.12	0.33	0.13
rtn-int2	0.06	0.07	0.09	0.08	0.09	0.10	0.06	0.08	0.07	0.03	0.02	0.03	0.01
wanda-50	0.48	0.50	0.58	0.36	0.42	0.50	0.47	0.57	0.66	0.24	0.24	0.48	0.21
sparsegpt-50	0.48	0.50	0.58	0.47	0.52	0.55	0.53	0.62	0.74	0.25	0.27	0.61	0.29
magnitude-50	0.35	0.40	0.48	0.22	0.18	0.20	0.22	0.34	0.58	0.13	0.03	0.23	0.03
wanda-2:4	0.32	0.33	0.44	0.20	0.19	0.30	0.19	0.36	0.45	0.13	0.10	0.33	0.08
sparsegpt-2:4	0.32	0.32	0.42	0.32	0.32	0.37	0.34	0.45	0.60	0.19	0.20	0.51	0.18
magnitude-2:4	0.25	0.29	0.39	0.19	0.16	0.16	0.10	0.20	0.42	0.10	0.07	0.21	0.04

and break only at 2-bit (shown here), whereas at 300M they already break at 3-bit and the same corrections recover them there (Table 11). So this is not something specific to 2-bit; it is the same mechanism, surfacing at whatever precision a given model's fixed readouts fail.

Table 13 decomposes the SAE scale correction into its components, which pins down the *mechanism* of Level 1. Two contrasts matter: *mean*-only correction never helps (it tracks *raw*, because the SAE's firing threshold is calibrated to the FP mean, so shifting the mean without fixing the gain only moves features further off threshold), and per-dim *var* substantially beats a single global *scalar* (e.g. 2:4-Wanda $0.92 \rightarrow 1.01$), so the corrective gain is *anisotropic*, not a single rescale factor. That *full*≈*var* for every method confirms the SAE has no basis-change component (unlike the probe): it is a purely magnitude readout. Among the calibrated ladder methods through 3-bit, the corrections are complete; the residuals that a linear correction cannot remove appear at 2-bit and under aggressive pruning, GPTQ-int2 (0.74), and more severely RTN-int2 (0.25) and 75–95% pruning, the genuine Level-3 losses below the recovery floor.

Table 17 reports the paired equivalence test at two sample sizes, and the comparison is itself instructive: on the 30-protein subset, 3–4-bit appears to *beat* FP ($+0.008$ to $+0.014$), but the effect vanishes on the 72-protein set ($+0.000$, -0.003), a small-n artifact that the larger matched set removes, leaving a clean 3-bit equivalence and an unambiguous 2-bit collapse.

Equivalence testing: margin, sensitivity, and corrections. We fix the equivalence margin at δ=0.02 Spearman in advance. It matches the full-precision model's own across-assay bootstrap variability (FP self-SD 0.025 on the 72-protein set, 0.018–0.026 across models) and is below a rank-relevant difference for variant scoring; we read equivalence as a two one-sided test (TOST, Lakens, 2017) at α=0.025, the framing now urged for "same-as-baseline" efficiency claims that a non-significant p cannot support. Per-method results on the 72-protein set: GPTQ-int3, mean -0.003 Spearman, two-sided 95% CI $[-0.016, +0.010]$, TOST p=0.006; GPTQ-int4, $+0.000$, $[-0.007, +0.005]$, $p<10^{-3}$; GPTQ-int2, -0.236, $[-0.28, -0.19]$, TOST p=1.0 (not equivalent); 2:4-Wanda, -0.068, $[-0.089, -0.047]$, $\approx$17% of the 6B FP zero-shot of 0.400. The verdict is margin-dependent: 6B-int3 equivalence holds for $\delta \gtrsim 0.016$ and 15B-int3 for $\delta \gtrsim 0.013$, so our δ=0.02 clears both while a stricter δ=0.015 would fail for 6B. We do not call δ pre-registered, and it is anchored to the FP model's own noise rather than to a task-external effect size; a margin tied to, e.g., the ΔSpearman that reorders a top-k variant selection would be preferable, and the -0.003 equivalence at 15B does not lean on a generous margin. All verdicts survive Holm correction across the four methods (int4/int3/int2/2:4) tested against FP within each model, the correction family being per-model, four tests.

Table 10: *Recoverability ladder* (ESM-C-6B): retention at each correction strength per readout and method, the numbers behind Fig. 1. The SAE has no "refit" (a fixed magnitude readout); the LM-head "refit" rung distills FP logits onto a fresh head reading the compressed trunk (App. A), an *inconclusive* route: it caps at 0.85–0.89 even on the *uncompressed* trunk across three head architectures, so the claim rests on the capacity-unbounded input rungs. Retention > 1 is within bootstrap noise (undamaged / fully recovered).

Readout	Method	none	rescale	linear	refit
SAE	gptq-int4	0.986	0.997	0.996	–
SAE	gptq-int3	0.934	0.981	0.978	–
SAE	gptq-int2	0.530	0.743	0.736	–
SAE	wanda-2:4	0.667	1.014	0.984	–
LM-head	gptq-int4	1.017	1.016	1.018	–
LM-head	gptq-int3	1.025	1.037	1.033	0.907
LM-head	gptq-int2	0.360	0.470	0.447	0.401
LM-head	wanda-2:4	0.871	0.905	0.903	–
probe	gptq-int4	0.981	0.986	0.988	0.991
probe	gptq-int3	0.921	0.940	0.965	0.973
probe	gptq-int2	0.509	0.508	0.754	0.919
probe	wanda-2:4	0.629	0.787	0.939	0.975

D SUPPLEMENTARY RESULTS AND FIGURES

Performance–reliability decoupling. UniComp (von Rad et al., 2026) finds that for LLMs, retained aggregate performance after compression does not imply preserved reliability: safety, fairness, and privacy can degrade even when the average benchmark score looks fine. We look for the same kind of decoupling in PLMs. We stratify ESM-C-6B zero-shot skill by ProteinGym assay class and report the absolute ΔSpearman (compressed$-$FP) rather than a retention ratio, since the catch-all and small-n classes have near-chance full-precision scores and their ratios are just noise; we therefore keep only the well-powered classes ($\geq$ 3 assays and FP Spearman $\geq$ 0.2: Activity, OrganismalFitness, Expression; Table 19). The high-fidelity methods are uniform (mean and worst-class Δ both near zero), and the aggressive ones do decouple in the expected direction, with the worst class (OrganismalFitness) trailing the assay average and the gap widening as compression gets harsher (magnitude-2:4: -0.187 mean vs -0.203 worst). The effect is small, though, at most about 0.02 Spearman between mean and worst class. Unlike the sharp reasoning-vs-knowledge split UniComp finds in LLMs, PLM zero-shot damage is fairly even across assay classes, so the average is representative here in a way it is not for LLMs. That is a genuine difference between the two settings, not just a replication. One slice does move: shallow-MSA assays fare worse under the aggressive methods (Table 20). The low-N_{eff} third retains less than the deep third, and Spearman(N_{eff}, retention) is positive (gptq-int2 $+0.18$, magnitude-2:4 $+0.15$, wanda-2:4 $+0.12$; near zero for the near-lossless gptq-int3), a mild evolutionary-tail truncation the average hides, al-

Table 11: *Recoverability ladder* at ESM-C-**300M**: the two *task* rungs (transferred probe, LM-head zero-shot) replicate the 6B pattern (Table 10) at a $\sim 20\times$ smaller model, the rescale step does not restore them, a basis alignment does, confirming the ladder is not an ESM-C-6B artifact. The SAE rung also replicates here, using the public ESM-C-300M SAE (top rows): a per-dim rescale recovers it (int2 $0.52 \rightarrow 0.74$), the same magnitude-readout behaviour as at 6B. The LM-head "refit" column distills FP logits onto a fresh head on the compressed trunk; it is *inconclusive* (caps at 0.85–0.89 on the *uncompressed* trunk across three head architectures; App. A), so the head claim rests on the input rungs, not this column. Retention > 1 is within bootstrap noise (undamaged / fully recovered).

Readout	Method	none	rescale	linear	refit
SAE	gptq-int4	0.985	0.997	0.996	–
SAE	gptq-int3	0.910	0.973	0.971	–
SAE	gptq-int2	0.521	0.741	0.694	–
SAE	wanda-2:4	0.570	0.915	0.927	–
LM-head	gptq-int4	0.961	0.965	0.963	–
LM-head	gptq-int3	0.682	0.721	0.709	0.631
LM-head	gptq-int2	-0.016	0.086	0.080	0.093
LM-head	wanda-2:4	0.398	0.515	0.469	–
probe	gptq-int4	0.964	0.967	0.986	0.991
probe	gptq-int3	0.795	0.815	0.894	0.960
probe	gptq-int2	0.307	0.340	0.427	0.763
probe	wanda-2:4	0.505	0.498	0.773	0.961

Table 12: *Second-family ladder replication* (ESM2; no ESM2 SAE exists, so no SAE rung). Top: ESM2-3B, where the readouts break at 2-bit and the diagonal-fails / basis-alignment-recovers contrast holds. Bottom: ESM2-15B, where even 2-bit barely breaks the readouts, the recovery floor has moved below 2-bit with scale.

ESM2-3B

Readout	Method	none	rescale	linear	refit
LM-head	gptq-int4	0.987	0.987	0.986	–
LM-head	gptq-int3	0.948	0.951	0.936	–
LM-head	gptq-int2	0.275	0.396	0.263	–
LM-head	wanda-2:4	0.762	0.771	0.730	–
probe	gptq-int4	0.990	1.001	0.994	0.994
probe	gptq-int3	0.950	0.968	0.987	1.004
probe	gptq-int2	0.591	0.648	0.776	0.951
probe	wanda-2:4	0.820	0.879	0.966	0.975

ESM2-15B

Readout	Method	none	rescale	linear	refit
LM-head	gptq-int4	0.995	0.996	0.995	–
LM-head	gptq-int3	0.999	1.000	0.997	–
LM-head	gptq-int2	0.763	0.776	0.786	–
LM-head	wanda-2:4	0.899	0.886	0.926	–
probe	gptq-int4	0.993	0.995	0.997	1.001
probe	gptq-int3	0.992	0.992	0.998	1.005
probe	gptq-int2	0.874	0.902	0.969	0.979
probe	wanda-2:4	0.858	0.947	0.984	0.987

beit with wide CIs at this assay count. In short, "3-bit is free" is a statement about the average; anyone deploying on a shallow-MSA or otherwise unusual target should look at the tail.

E COMPRESSION METHODS

This section describes each configuration in the 19-method matrix; solver settings and provenance are in App. A. Quantization is weight-only unless noted. For pruning we count memory including

Table 13: *SAE scale-correction ladder* (ESM-C-6B), the numbers behind Fig. 3. For each method we show the activation norm ratio, the sd ratio, the relative mean-shift, and the SAE function retention under increasingly expressive corrections: none, mean-only, a single global scale, a per-dimension scale, and a full linear map. A per-dimension scale (*var*) clearly beats a single global one, so the needed correction is per-channel rather than one number; correcting the *mean* alone does not help, because the SAE's firing threshold is set to the full-precision mean.

Method	norm_r	sd_r	m-shift	raw	mean	scalar	var	full
rtn-int8	1.000	0.999	0.005	1.000	1.000	1.000	1.000	1.000
gptq-int4	0.992	0.989	0.046	0.986	0.984	0.994	0.997	0.996
awq-int4	0.981	0.981	0.036	0.977	0.975	0.996	0.998	0.996
hqq-4bit	0.994	0.994	0.045	0.992	0.991	0.999	0.998	0.997
gptq-int3	0.970	0.956	0.092	0.934	0.930	0.963	0.981	0.978
awq-int3	0.936	0.934	0.082	0.922	0.918	0.986	0.993	0.989
hqq-2bit	0.982	0.957	0.326	0.804	0.822	0.818	0.847	0.863
gptq-int2	0.815	0.738	0.223	0.530	0.514	0.652	0.743	0.736
rtn-int2	0.564	0.116	1.285	0.000	0.000	0.000	0.069	0.248
wanda-50	0.864	0.848	0.109	0.856	0.838	0.991	1.025	1.015
sparsegpt-50	0.907	0.879	0.094	0.901	0.894	0.995	1.038	1.037
magnitude-50	0.807	0.778	0.160	0.778	0.750	0.965	1.018	1.004
wanda-2:4	0.725	0.684	0.171	0.667	0.628	0.922	1.014	0.984
sparsegpt-2:4	0.809	0.751	0.118	0.771	0.743	0.954	1.051	1.039
magnitude-2:4	0.696	0.649	0.196	0.632	0.586	0.910	1.011	0.984
wanda-75	0.463	0.409	0.448	0.284	0.218	0.615	0.759	0.719
sparsegpt-75	0.656	0.534	0.207	0.509	0.445	0.778	0.991	0.975
magnitude-75	0.380	0.269	0.454	0.110	0.057	0.291	0.478	0.473
wanda-80	0.335	0.261	0.649	0.059	0.027	0.177	0.315	0.298
sparsegpt-80	0.587	0.428	0.269	0.333	0.263	0.569	0.843	0.811
magnitude-80	0.301	0.157	0.604	0.004	0.000	0.013	0.137	0.116
wanda-85	0.283	0.191	0.770	0.001	0.000	0.005	0.110	0.087
sparsegpt-85	0.524	0.336	0.371	0.124	0.086	0.238	0.486	0.463
magnitude-85	0.297	0.102	0.683	0.000	0.000	0.000	0.090	0.079
wanda-90	0.251	0.134	0.730	0.000	0.000	0.000	0.064	0.039
sparsegpt-90	0.472	0.275	0.493	0.005	0.002	0.010	0.097	0.085
magnitude-90	0.340	0.068	0.732	0.000	0.000	0.000	0.130	0.160
wanda-95	0.243	0.086	0.673	0.000	0.000	0.000	0.101	0.074
magnitude-95	0.360	0.049	0.685	0.000	0.000	0.000	0.141	0.201
gauss-int3	0.938	0.912	0.227	0.810	0.808	0.864	0.900	0.899
gauss-int2	1.200	0.896	1.773	0.014	0.212	0.012	0.026	0.237

all metadata: unstructured 50% sparsity saves nothing without a sparse format, whereas $2{:}4$ (two of every four weights zero) is hardware-accelerable and counted at $\approx 0.5\times$.

Quantization (13 configurations). *RTN* is the naive baseline, symmetric per-output-channel round-to-nearest (one fp16 scale per row, no sub-row grouping), no calibration, run at 8/4/2-bit. From *bitsandbytes* we take two library-standard points: `int8` (LLM.int8(), the single W8A8 point, with fp16 outlier decomposition) and `NF4`, a 4-bit NormalFloat code designed for normally-distributed weights. *HQQ* runs the half-quadratic solver at 4/3/2-bit, fitting robust per-group scales and zero-points without any calibration data. *GPTQ* does calibrated error-feedback rounding against an inverse-Hessian at 4/3/2-bit, and *AWQ* searches per-channel activation-aware scales and clips weights at 4/3-bit. Together these span an 8/4/3/2-bit ladder from naive to fully calibration-aware.

Pruning (6 configurations). Three criteria, each at 50% unstructured and $2{:}4$ semi-structured: *magnitude* drops the smallest-$|w|$ weights (calibration-free); *Wanda* drops by an $|w| \cdot \|x\|$ importance with no weight update; and *SparseGPT* prunes OBS-style, with inverse-Hessian error compensation.

Gaussian control. `gauss-int2`/`gauss-int3` add Gaussian weight noise matched in per-layer Frobenius norm to the int2 and int3 quantization errors (RTN in the matrix runs at 8/4/2-bit, so the int3 reference is a synthetic RTN-int3 error); this is a non-compression control used only in the recoverability and severity analyses, and is not one of the 19.

Table 14: *Non-circular SAE functional retention* (ESM-C-6B, layer 60). The main metric is the *dict-match cosine*: for each SAE feature, the cosine between its firing pattern at full precision and after compression. We report *concept-AUC*, the survival of the best per-class secondary-structure detector, only as a secondary check; it is noisier and slightly optimistic, since the best feature is picked on full precision without a held-out split. Both metrics are scale-invariant, so a per-dimension rescale cannot move them: they measure whether feature *identities* survive, which the ladder's magnitude-based SAE rung cannot. The pattern matches Fig. 3: near-full retention through 3-bit, a real drop at 2-bit, and collapse under aggressive pruning and data-free 2-bit. Randomly pairing features gives a dict-match near zero (null s.d. $\approx 1/\sqrt{d}$), so the reported values are far above chance.

Method	concept-AUC ret.	dict-match cos	frac. cos≥ 0.5
gptq-int4	0.98	0.98	1.00
gptq-int3	0.99	0.96	1.00
gptq-int2	0.72	0.74	0.89
rtn-int2	-0.25	0.00	0.00
wanda-2:4	0.92	0.92	0.99
wanda-90	0.30	0.00	0.02

Table 15: *The SAE result replicates across scales*: non-circular dict-match cosine at the three ESM-C scales with public SAEs (300M/600M/6B). The same dose-response holds at every scale, near-full through 3-bit and dropping at 2-bit, so the SAE rung and its recovery floor are not specific to 6B.

Method	EC.3	EC.6	EC6B
	dict-match cosine (FP vs compressed)		
gptq-int4	0.94	0.94	0.98
gptq-int3	0.84	0.86	0.96
gptq-int2	0.13	0.54	0.74
wanda-2:4	0.69	0.71	0.92
rtn-int2	0.00	0.00	0.00

F SECONDARY RESULTS

This appendix details three analyses referenced from the main text: the residual-injection sensitivity control (Sec. 5.3), the ranking pre-screen, and the scale/method-ordering laws (Sec. 5.4). The shallow-MSA tail (secondary finding iii) is in Table 20.

Residual injection: sensitivity is direction, not magnitude. To turn the sensitivity account from an association into a directly measured per-readout quantity, we take the map's unrecovered residual $r = X_{\mathrm{FP}} - \mathrm{map}(X_{\mathrm{cm}})$, inject it back into the full-precision activations as $X_{\mathrm{FP}} - \alpha r$, and read each *frozen* FP readout as α sweeps $0 \to 2$ (at $\alpha{=}1$ the readout sees exactly the aligned-compressed input, so $\alpha{=}1$ reproduces the ladder's recovered value). The control replaces r with a norm-matched *random-direction* residual (same per-example $\|r\|$). Across all five supervised readouts (DeepLoc, SSP, Meltome, GB1, AAV) at ESM-C-6B and ESM2-650M, the true and random residuals pull the readouts apart as α grows: the damage is set by the residual's direction (its intersection with the readout's sensitive subspace) rather than its Frobenius norm. This is the per-readout, causal counterpart of the model-level magnitude-vs-structure result (Table 13); the sign of the true$-$random gap is cell-dependent (e.g. the fine-grained fitness readouts GB1/AAV are driven negative by the true residual where an equal-norm random one is not), so we report the direction-not-magnitude conclusion rather than a single scalar predictor.

A cheap pre-screen for *ranking* methods. A single unlabeled forward pass ranks compression methods for a given model, though not across families. Estimating downstream damage from cheap unsupervised divergence (L1 logit-KL + L2 CKA/kNN) over one 2048-sequence forward pass works *for ranking*: predicted vs actual zero-shot degradation correlates at within-family Spearman 0.87 (pooled cross-family 0.82). It does *not* work as an absolute cross-family regressor, a leave-one-family-out ridge yields $R^2 < 0$, because the divergence$\to$degradation mapping is family-specific (ProGen2's logit-KL lives on a different scale). The pre-screen indicates *which* method to prefer for

Table 16: *Ridge-penalty sweep* for the DeepLoc probe recovery (ESM-C-6B, retention vs FP), with Procrustes and full refit for reference. A tuned general-linear (ridge) alignment matches or exceeds the orthogonal (Procrustes) map, and the curve is broad (stable across $\lambda \in [1, 10^3]$) rather than a test-selected spike, so the probe's distortion is a general-linear change of basis rather than specifically a rotation. The figure's "linear" rung uses the selection-free Procrustes value. At the most aggressive settings the label-free ridge can slightly edge out the full *refit* (e.g. int2 0.96 vs 0.92); this is not paradoxical, "refit" retrains the probe head on limited compressed data and can underperform a well-regularized alignment that inherits the FP probe, so we treat refit as a practical upper bound rather than a theoretical ceiling.

Method	10^0	10^1	10^2	10^3	10^4	10^5	Procr.	refit
gptq-int4	0.997	1.008	1.023	0.995	0.869	0.710	0.988	0.991
gptq-int3	0.982	0.998	1.005	0.990	0.879	0.697	0.965	0.973
gptq-int2	0.900	0.938	0.962	0.912	0.783	0.608	0.754	0.919
wanda-2:4	0.997	1.007	1.015	0.977	0.857	0.660	0.939	0.975

Table 17: *Paired zero-shot equivalence test* (ESM-C-6B): ΔSpearman (method$-$FP) over matched proteins, bootstrap 95% CI, Holm-corrected p. 3-bit is equivalent to FP on the 72-protein set.

Set	Method	Δ(m–FP)	95% CI	p_{Holm}
full (n=72)	gptq-int4	+0.0000	[-0.007,0.005]	1.0000
full (n=72)	gptq-int3	-0.0033	[-0.016,0.010]	1.0000
full (n=72)	gptq-int2	-0.2360	[-0.282,-0.189]	0.0000
full (n=72)	wanda-2:4	-0.0682	[-0.089,-0.047]	0.0000
subset (n=30)	gptq-int4	+0.0082	[0.003,0.014]	0.0008
subset (n=30)	gptq-int3	+0.0135	[-0.006,0.034]	0.1980
subset (n=30)	gptq-int2	-0.2718	[-0.347,-0.196]	0.0000
subset (n=30)	wanda-2:4	-0.0643	[-0.101,-0.027]	0.0006

a given model rather than the absolute score an unseen family will reach; the supervised channel is near-lossless and leaves little variance to model.

Scale and method laws. Calibration-aware low-bit quantization dominates pruning at matched budget, and naive 4-bit can fail catastrophically at scale. INT8 is effectively lossless; calibration-aware 4-bit (GPTQ/AWQ/NF4/HQQ-4bit) is near-lossless, whereas naive 4-bit/3-bit degrade small models and can collapse the largest ProGen2 model (RTN-int4 holds ProGen2 to 2.7B but collapses ProGen2-xlarge zero-shot to 0.004, while calibrated 4-bit keeps it at 0.38–0.39). That collapse is a genuine weight-level failure, coherent across fidelity, geometry, and biology (pseudo-perplexity diverges, top-1 agreement 1%, CKA 0.65), and specific to naive rounding: int8, calibrated int4 (GPTQ/AWQ), and norm-matched Gaussian all hold on the same model. An independent from-scratch RTN reproduces it (int4 pseudo-perplexity $\approx 10^{36}$, int8 healthy at ≈ 11), and it localizes to a single matrix: sparing the outlier-heavy *tied output embedding* restores pseudo-perplexity to ≈ 13. Calibration-aware methods and bitsandbytes quantize only `nn.Linear` modules, leaving that embedding at full precision, which is precisely why they hold where naive RTN, operating on raw weight tensors, does not. Calibrated 3-bit is a genuine mid-fidelity point (zero-shot 0.86–0.89 at $\sim 5\times$ compression); 2-bit collapses the zero-shot channel even with calibration (GPTQ-int2 0.23, RTN-int2 ≈ 0) while supervised retention holds (0.84/0.74). Damage from naive methods recovers with scale (ESM-C 2:4-Wanda zero-shot 0.16 at 300M $\rightarrow$ 0.37 at 6B). Unstructured 50% sparsity buys no memory without a sparse format; only 2:4 yields real storage reduction.

Table 18: *Head/trunk scoped-quantization ablation*: does the zero-shot break come from the trunk representation or from corrupting the head's own weights? We compare absolute zero-shot Spearman under full quantization (trunk and head) against quantizing only the trunk (FP head) and only the head (FP trunk). Quantizing only the head is near-lossless, while quantizing only the trunk reproduces the full drop, at both scales, so the collapse is trunk-driven.

ESM-C-300M

Bits	FP	full-quant (both)	trunk-only (#nohead, FP head)	head-only (#headonly, FP trunk)
int2	0.441	-0.017	-0.007	0.390
int3	0.441	0.280	0.300	0.435

ESM-C-6B

Bits	FP	full-quant (both)	trunk-only (#nohead, FP head)	head-only (#headonly, FP trunk)
int2	0.449	0.156	0.160	0.441
int3	0.449	0.459	0.461	0.448

Table 19: *Performance–reliability decoupling* (ESM-C-6B), our PLM version of UniComp's reliability check (von Rad et al., 2026). We report the absolute change in Spearman (compressed − FP) per assay class, using only well-powered classes (≥ 3 assays, FP ≥ 0.2) and differences rather than ratios so weak classes cannot inflate the effect. High-fidelity methods change every class by ≈ 0; aggressive methods keep the overall mean but push one class down further, a bottleneck the mean hides.

Method	mean Δ	worst class	worst Δ
	ΔSpearman (compressed−FP), well-powered classes		
gptq-int4	+0.009	Activity	+0.007
awq-int4	+0.002	Activity	+0.001
wanda-2:4	-0.088	OrganismalFitness	-0.101
sparsegpt-2:4	-0.061	OrganismalFitness	-0.069
magnitude-2:4	-0.187	OrganismalFitness	-0.203

Table 20: *Shallow-MSA proteins degrade more* (ESM-C, zero-shot). We split assays by MSA depth (N_{eff}, the number of related sequences) and compare mean retention on the shallow (low-N_{eff}) third against the deep third, plus the rank correlation Spearman$(N_{\text{eff}}, \text{retention})$ with a bootstrap 95% CI. A positive ρ means shallow-MSA proteins are hit harder; the effect shows up for aggressive methods and is near zero for near-lossless gptq-int3, though the CIs are wide at this assay count.

Method	low-N_{eff} ret.	high-N_{eff} ret.	$\rho(N_{\text{eff}}, \text{ret})$	95% CI
gptq-int3	0.73	0.78	-0.030	[-0.22,+0.16]
gptq-int2	-0.02	0.14	+0.182	[-0.01,+0.35]
hqq-2bit	0.20	0.32	+0.083	[-0.10,+0.26]
wanda-2:4	0.37	0.53	+0.115	[-0.09,+0.31]
magnitude-2:4	0.27	0.38	+0.149	[-0.04,+0.34]

Table 21: *One map, many readouts* (ESM-C, gptq-int2/int3). A *single* orthogonal map, fit once label-free on generic pooled activations (compressed→FP), applied *unchanged* to five readouts (retention vs FP without the map, *transfer*, and with it, *shared map*). Recovery is *partial and readout-dependent*, even inside the recoverable regime: at 300M-int3, where a per-readout Procrustes lifts DeepLoc to ≈ 0.89, the shared map leaves DeepLoc at 0.75 (below no map) and collapses AAV; at 300M-int2 it rescues GB1 ($-0.51 \rightarrow 0.48$) but degrades SSP; at 6B-int2 it helps localization/thermostability but hurts binding/structure. So no single map serves all readouts, which is why the ladder fits a fresh map per readout. Ratios < 0 are FP sign-flips.

Setting	Readout	transfer retention vs FP	shared map retention vs FP
300M int2	DeepLoc (loc.)	0.25	0.32
	Sec. structure	0.72	0.60
	Meltome (stab.)	0.26	0.27
	GB1 (bind.)	-0.51	0.48
	AAV (fit.)	0.07	0.25
300M int3	DeepLoc (loc.)	0.77	0.75
	Sec. structure	0.96	0.91
	Meltome (stab.)	0.77	0.74
	GB1 (bind.)	0.55	0.63
	AAV (fit.)	-0.01	-0.51
6B int2	DeepLoc (loc.)	0.51	0.60
	Sec. structure	0.79	0.68
	Meltome (stab.)	0.09	0.22
	GB1 (bind.)	0.60	-0.62
	AAV (fit.)	0.05	-0.15
6B int3	DeepLoc (loc.)	0.93	0.86
	Sec. structure	0.99	0.86
	Meltome (stab.)	0.85	0.67
	GB1 (bind.)	0.29	0.15
	AAV (fit.)	0.22	0.35

Table 22: *Label-free recovery of a frozen probe* (mean retention vs FP over the nine *encoder* models, ESM2 and ESM-C; ProGen2 is a decoder for which frozen probes are not the intended readout). **frozen** = FP-fit probe on compressed features, no change; **LF map** = a label-free map fit on a *generic corpus* with the task's train/test sequences held out and ridge λ chosen by held-out reconstruction MSE (no downstream labels, no test-metric selection), then the FP probe; **refit (lab.)** = probe retrained on compressed features with task labels. The label-free generic map recovers a meaningful part of the 2-bit / 2:4 gap (int2 $0.58 \rightarrow 0.70$, 2:4 $0.70 \rightarrow 0.85$) and little at 3–4-bit, but a labelled refit recovers more: recovery is partial and teacher-dependent. Per-model values are in the released records.

Method	DeepLoc			SSP		
	frozen	LF map	refit (lab.)	frozen	LF map	refit (lab.)
gptq-int4	0.97	0.94	0.99	1.00	1.00	1.00
gptq-int3	0.85	0.87	0.98	0.97	0.98	0.98
gptq-int2	0.41	0.58	0.81	0.75	0.81	0.83
wanda-2:4	0.53	0.77	0.94	0.87	0.94	0.94

Table 23: *Per-task supervised retention* (compressed/FP), mean over 13 models, the disaggregation of the supervised column of Table 3. Robustness is uniform on the cross-protein tasks whose probes are not data-limited (SSP, DeepLoc). The FLIP splits are hard extrapolation splits (meltome clustered; GB1 three_vs_rest; AAV seven_vs_many); GB1/AAV are additionally single-protein combinatorial libraries whose FP learning curves are still rising at full training size, so their retention *ratios* are interpretable but no absolute Level-3 threshold can be read from them. GB1/AAV values >1.0 reflect the small-n regularization regime and are not claimed as an effect.

Method	SSP	DeepLoc	meltome	GB1 (3vR)	AAV (7vM)
rtn-int8	1.000	1.000	1.000	0.987	1.000
bnb-int8	0.999	0.998	0.996	0.892	0.893
rtn-int4	0.983	0.982	0.970	1.013	1.008
bnb-nf4	0.998	0.994	0.993	1.007	1.017
hqq-4bit	0.998	0.997	0.994	0.993	0.983
gptq-int4	0.997	0.994	0.992	1.002	1.020
awq-int4	0.997	0.995	0.984	0.992	1.013
hqq-3bit	0.985	0.984	0.970	0.988	1.024
gptq-int3	0.982	0.985	0.960	0.987	1.018
awq-int3	0.988	0.986	0.967	1.018	1.023
hqq-2bit	0.856	0.881	0.815	0.963	0.930
gptq-int2	0.841	0.838	0.764	0.894	0.891
rtn-int2	0.716	0.733	0.755	0.761	0.783
wanda-50	0.979	0.984	0.945	1.002	1.029
sparsegpt-50	0.982	0.989	0.958	1.020	1.035
magnitude-50	0.913	0.921	0.877	1.011	1.023
wanda-2:4	0.932	0.955	0.902	1.035	1.060
sparsegpt-2:4	0.958	0.978	0.936	1.025	1.043
magnitude-2:4	0.886	0.904	0.853	1.017	1.068

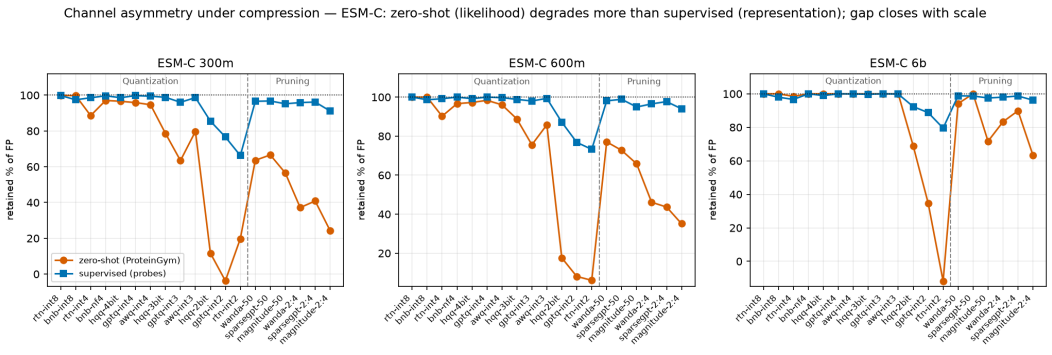


Figure 2: *Channel asymmetry* (ESM-C). Zero-shot (ProteinGym) retention falls with compression aggressiveness while supervised (FLIP/DeepLoc/SSP) retention stays near full precision; the gap narrows with model scale. This is the deployed special case of the recoverability ladder (Sec. 5.3). Retention is capped at 100%; values above full precision are bootstrap noise.

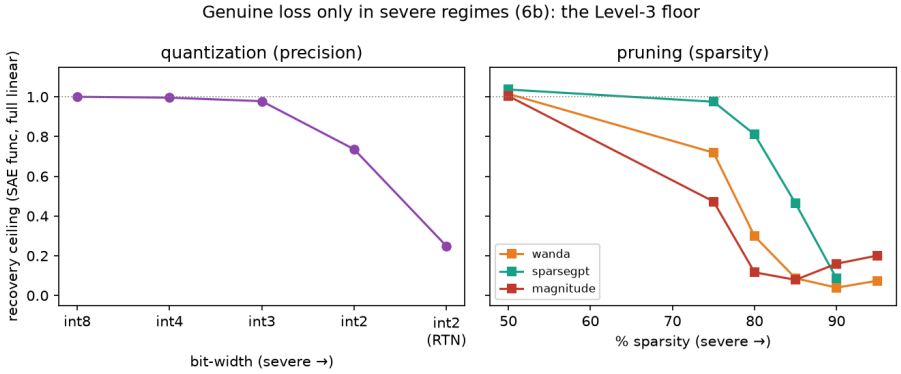


Figure 3: *Genuine loss appears only in severe regimes* (ESM-C-6B). SAE-function recovery ceiling (retention after the full label-free linear correction) versus perturbation severity, bit-width for quantization (left), % unstructured sparsity for pruning (right). Quantization is fully recoverable through 3-bit and drops only at 2-bit (0.74); pruning is recoverable through 50–2:4 and collapses across 75–90%. The same dose-response holds on the *non-circular* functional assay (concept-detection AUC and scale-invariant dictionary match; Table 14), so the floor is genuine functional loss, not an artifact of the magnitude-based SAE readout.

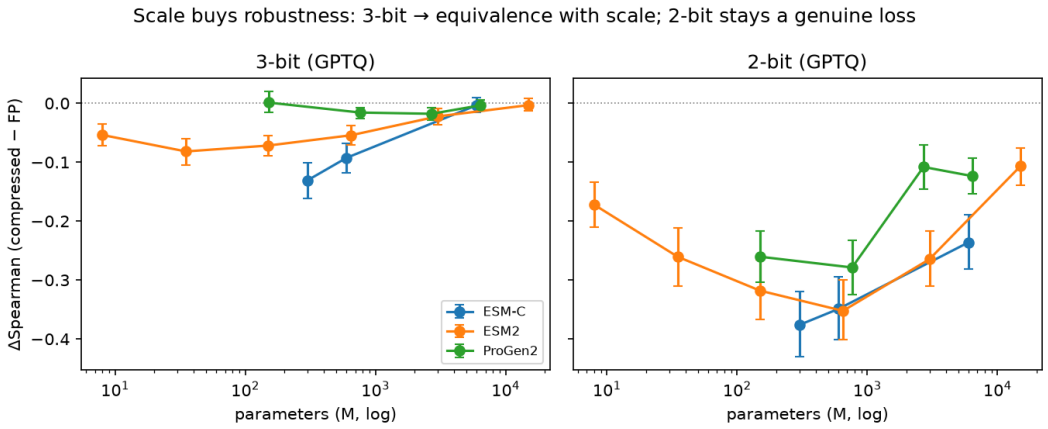


Figure 4: *Scale buys robustness on the deployed readout.* Paired zero-shot ΔSpearman (compressed − FP) vs. model size (log parameters), per family, for 3-bit (left) and 2-bit (right) GPTQ; error bars are paired bootstrap 95% CIs over proteins. The 3-bit gap reaches zero (equivalence) by the top of each family (not strictly monotone in between; see Sec. 5.2), whereas 2-bit stays a genuine loss throughout. Each point is a difference from that model's own full precision.

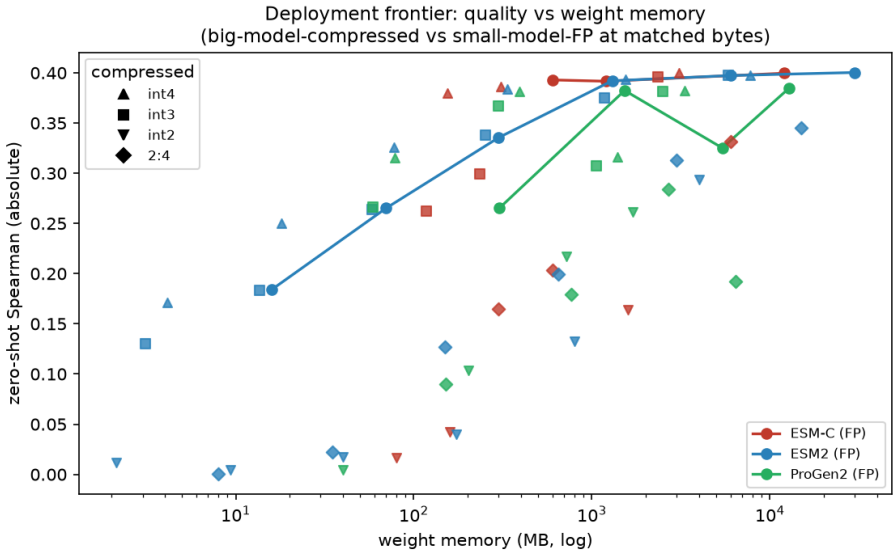


Figure 5: *Deployment frontier.* Absolute zero-shot Spearman vs analytic weight memory (log MB), one point per model×method; full-precision models joined per family. Calibration-aware 4-bit (triangles) is near-lossless at every scale and shifts each model $4\times$ left; 3-bit (squares) only preserves quality at the top of a family; 2-bit (down-triangles) drops off the frontier. Since zero-shot skill saturates by $\sim$600M, the frontier rewards the smallest model past the knee at 4-bit rather than a compressed giant. The x-axis uses the analytic effective-bits formula of Table 3 (App. A).

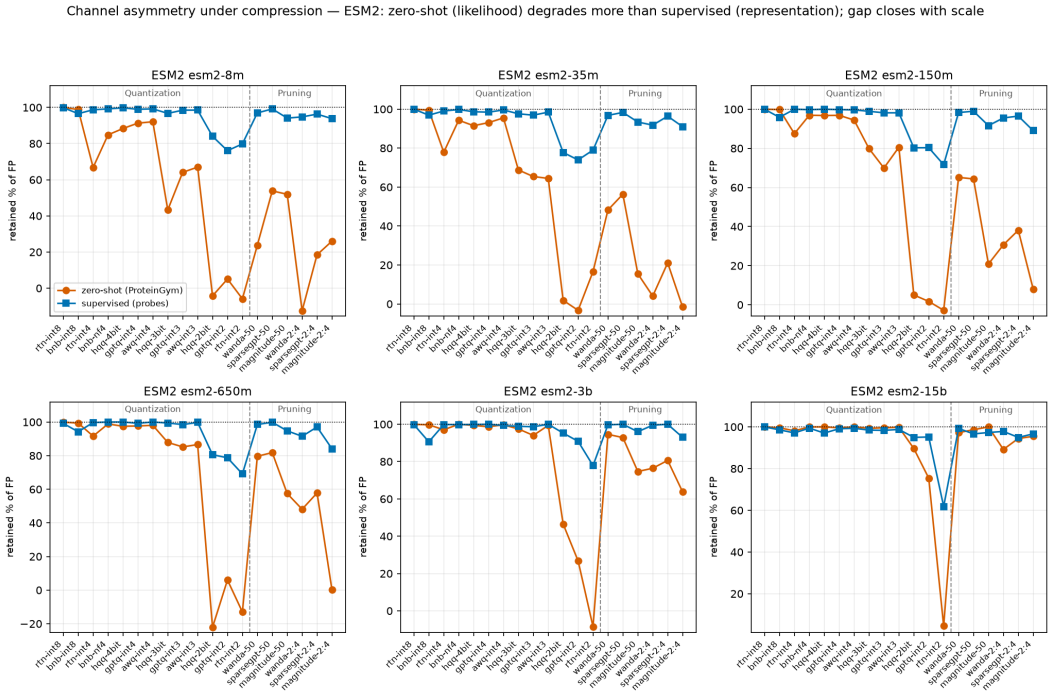


Figure 6: *Channel asymmetry* (ESM2 family, per-family version of Fig. 2). Zero-shot retention falls with compression while supervised retention stays near full precision; the gap narrows with scale. Retention is capped at 100%; values above full precision are bootstrap noise.

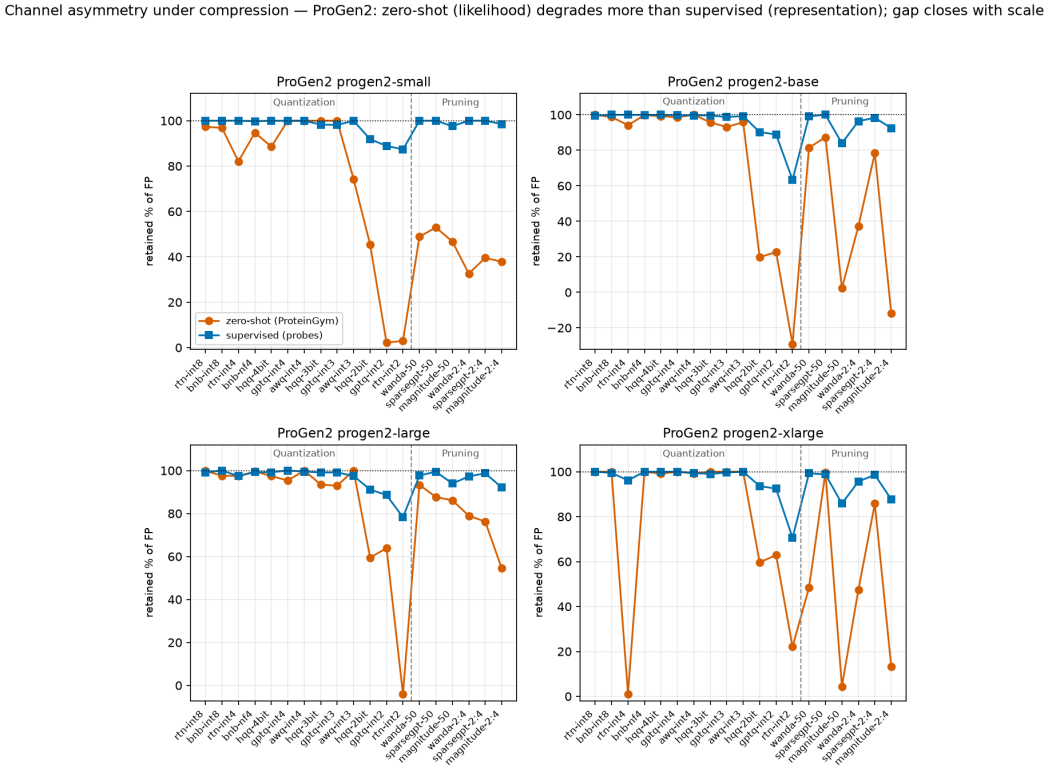


Figure 7: *Channel asymmetry* (ProGen2 family). The same asymmetry holds for ProGen2; per-residue autoregressive scoring additionally depresses absolute frozen-probe scores (a probe handicap rather than a compression effect). Retention is capped at 100%; values above full precision are bootstrap noise.

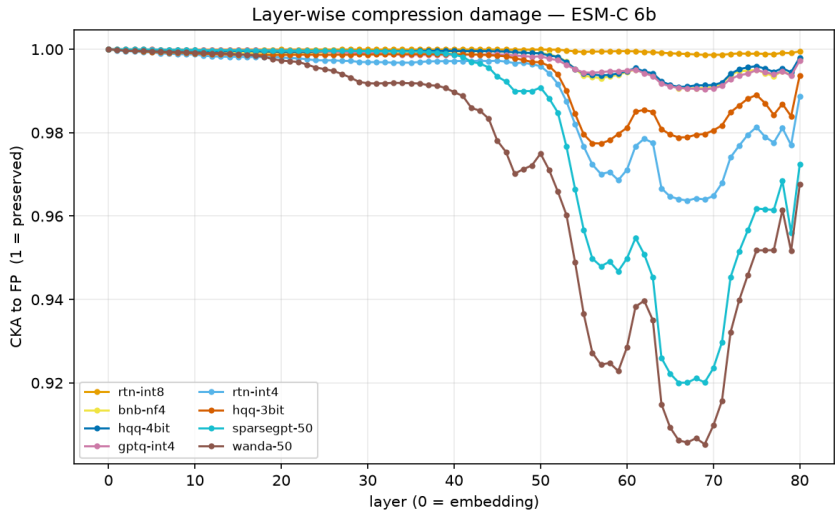


Figure 8: *Per-layer representation damage* (ESM-C-6B): CKA divergence concentrates in late layers, consistent with a late-layer representation change (Sec. 5.3).