Layer-wise Compression Tolerance in Variational Information Bottleneck Networks

Feng Yuhan

August 31, 2026

Abstract

The Variational Information Bottleneck (VIB) is typically trained with a *single, global* compression strength β shared by all stochastic layers, implicitly assuming that every layer tolerates the same amount of compression. We test this assumption directly. Using a *single-layer bottleneck* protocol that isolates the effect of compressing one layer at a time, we measure per-layer accuracy degradation curves $A_l(\beta)$ on two systems: an MNIST multilayer perceptron and a CIFAR-10 convolutional network. Both systems exhibit a stable, depth-dependent tolerance ordering in which deep bottlenecks sustain compression that severely degrades or catastrophically collapses shallow ones (on CIFAR-10 the second layer collapses to chance level with full posterior collapse, KL $\to 0$, while the final layer retains 91% of baseline accuracy at the same β). We show that every stochastic layer pays a *sampling-noise floor* — a layer-dependent accuracy cost that is present even when the KL penalty is numerically negligible — and that on CIFAR-10 this floor is large enough to make the standard per-layer critical compression strength β_l^{crit} undefined for three of four layers. Comparing layers by the β-free relative rate $r = \mathrm{KL}(\beta)/\mathrm{KL}(\beta \to 0)$ preserves the ordering, confirming that it is a property of depth rather than of the control parameter. We further show that multi-layer compression costs are not additive: compressing the two shallow layers together causes catastrophic collapse at a β where each layer alone is nearly unharmed, supporting the hypothesis that tolerance arises partly from compensation by unconstrained layers. The absolute scale of tolerance does not transfer across systems, and neither β, $\beta \cdot \mathrm{KL}$, nor dimension-normalized KL is sufficient to predict it. The robust, transferable finding is the ordering itself; the practical consequence is that a global β is a misspecified control for multi-layer VIB.

1 Introduction

The information bottleneck (IB) principle [1, 2] formulates representation learning as a trade-off between compression and prediction: a representation Z should retain only the information about the input X that is relevant to the target Y. Its variational formulation [3] makes the objective trainable end-to-end,

$$\mathcal{L} = \mathbb{E}[\ell(y, \hat{y})] + \beta \,\mathrm{KL}\big(q(z \mid x) \,\|\, p(z)\big), \tag{1}$$

and has become a standard tool for studying and regularizing deep representations [4, 5].

When the bottleneck is applied at several layers of a deep network, the scalar β is almost universally shared across layers, or annealed globally over training time [6]. This convention embeds an unexamined assumption: *all layers have the same tolerance to compression.* Representations at different depths, however, play different functional roles. Shallow layers retain much of the input structure, including task-irrelevant detail; deep layers encode increasingly task-oriented summaries [2, 7]. There is no reason a priori for a single β to be equally appropriate at every depth, and if

tolerance varies systematically with depth, then a global β simultaneously under-compresses some layers and over-compresses others.

We ask a simple question that the global-β convention leaves unanswered:

$$\textbf{RQ:}\quad \text{Does compression tolerance vary across network depth?} \tag{2}$$

Answering it requires a measurement protocol that attributes degradation to a specific depth. If several layers are compressed at once, a drop in accuracy cannot be assigned to any of them. We therefore introduce a *single-layer bottleneck* protocol: exactly one stochastic VIB layer is inserted at position l, all other layers remain deterministic, and the accuracy–β curve $A_l(\beta)$ is swept. From these curves we define a per-layer critical compression strength β_l^{crit}, the largest β at which accuracy stays within a fixed margin of the no-bottleneck baseline.

Our contributions are as follows.

- We design the single-layer bottleneck protocol and introduce three per-layer statistics: the critical compression strength β_l^{crit}, the sampling-noise floor ϕ_l, and the relative rate r_l; these separate compression cost from the cost of stochasticity and make the comparison axis explicit.
- On an MNIST MLP we show, with three seeds, that tolerance is strongly layer-dependent: shallow layers degrade an order of magnitude earlier in β than deep ones, and one layer collapses catastrophically under strong compression.
- On a CIFAR-10 CNN we replicate the depth ordering (shallow $<$ deep), including a clean posterior-collapse event at a middle layer. We show that the absolute scale of β_l^{crit} does not transfer, that the sampling-noise floor is large enough to make β_l^{crit} undefined for most layers, and that the floor and the tolerance ordering are distinct phenomena.
- We show that the depth ordering is invariant to whether one compares layers by β, by achieved KL, or by relative rate r; only the magnitude of the gap depends on the axis, which identifies the ordering as a robust structural property rather than an artefact of the control parameter.
- We test the additivity of multi-layer compression and find that it fails: compressing two shallow layers together produces a super-additive collapse at a β where each layer alone is nearly unharmed, directly supporting the compensation hypothesis.
- We analyze candidate mechanisms — input-information burden, noise propagation, and compensation by unconstrained layers — and state precisely which claims the data support and which remain hypotheses.

2 Related Work

Information bottleneck and its variational forms. The IB principle [1] and its deep learning interpretation [2] motivated a line of work on compression as a lens on deep learning, including debates on what the information plane actually shows [7] and on estimating information flow in networks [8]. Deep VIB [3] made the objective scalable by variational bounding; β-VAE [4] established the β-weighted KL as a general compression control; Information Dropout [5] showed that multiplicative noise with learned statistics implements a closely related, locally adaptive compression. Our work differs from all of these in object of study: rather than proposing a new bottleneck or analyzing total compression, we measure how the *cost* of compression is distributed across depth under a fixed protocol.

Compression strength and β scheduling. The sensitivity of VAE/VIB models to β is well known, and annealed or staged β schedules are common practice [6]. Such schedules vary β over *time*, uniformly across layers. Per-layer compression strength has received little controlled study; where multiple stochastic layers are used, they typically share one β [3]. Our single-layer protocol provides the missing per-layer measurement on which any depth-dependent schedule would have to be based.

Depth-dependent compressibility in pruning and quantization. A related phenomenon is known from network compression: different layers tolerate pruning and quantization very differently, with fully connected and late layers typically surviving aggressive pruning while early convolutional layers are fragile [9]. Those results concern weight removal, not information removal, but the qualitative parallel — deep/late representations are more redundant and more compressible — is precisely what an information-level account would predict. Our results establish the parallel at the level of representational information, with a controlled single-variable protocol.

Posterior collapse. In variational autoencoders, posterior collapse denotes the degenerate regime in which the approximate posterior ignores the input and equals the prior, so that the latent variable carries no information [10]. We observe an analogous, layer-localized event in a *discriminative* VIB network under strong compression: a middle representation collapses onto the prior ($\mathrm{KL} \to 0$) and the network drops to chance accuracy. To our knowledge this is the first documentation of posterior collapse as a *layer-dependent* failure mode of compressed discriminative models.

3 Method

3.1 VIB layer

Each VIB layer maps a hidden state $h \in \mathbb{R}^{d_l}$ to a stochastic representation $z = \mu(h) + \sigma(h) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, against a standard Gaussian prior $\mathcal{N}(0, I)$. The per-sample KL term is

$$\mathrm{KL}_l(x) = \tfrac{1}{2} \sum_{i=1}^{d_l} \left(\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2\right). \tag{3}$$

For numerical stability we clamp $\log \sigma^2 \in [-10, 10]$ and $\mathrm{KL} \geq 0$. Without the clamp, strong compression of the final layer drives $\log \sigma^2 \to -\infty$ and produces NaN losses; with it, every configuration trains stably (Section 3.6). We stress throughout that KL_l is a *variational proxy* for the true mutual information $I(X; Z_l)$: it upper-bounds it under the variational family, and we make no claim about true information quantities [7, 8].

Two parameterizations are used. For fully connected layers, $\mu, \log \sigma^2$ are linear maps to a bottleneck of dimension $d_l^b \leq d_l$ followed by a linear projection back to d_l. For convolutional feature maps, $\mu, \log \sigma^2$ are 1×1 convolutions, i.e. channel-wise bottlenecks that preserve spatial dimensions; there d_l counts channel×spatial units.

3.2 Single-layer bottleneck protocol

To attribute degradation to a single depth, each run inserts **exactly one** VIB layer at position $l \in \{1, \ldots, 4\}$ and leaves all other layers deterministic. The training loss is

$$\mathcal{L} = \mathrm{CE} + \beta_l \, \overline{\mathrm{KL}_l}, \tag{4}$$

with $\beta_l = \beta$ at the compressed layer and 0 elsewhere. Because uncompressed layers are deterministic, they inject no sampling noise; any degradation is attributable to the single compressed

layer. (Section 5.1 reports one auxiliary experiment in which all four VIB modules are instantiated simultaneously, which we use only to motivate the protocol and report with an explicit confound caveat.)

3.3 Critical compression strength

Given a no-bottleneck baseline accuracy A_0 and a tolerance margin $\Delta = 1\%$, we define

$$\beta_l^{\text{crit}} = \max\big\{\beta : A_l(\beta) \geq A_0 - \Delta\big\}, \tag{5}$$

estimated over a discrete β grid. β_l^{crit} is a coarse summary of the tolerance curve $A_l(\beta)$; we always report the full curves alongside it. With multiple seeds, $A_l(\beta)$ and A_0 are per-seed quantities and the threshold is applied per seed against that seed's own baseline.

3.4 Decomposing the cost of a stochastic layer

Inserting a stochastic layer at position l changes the network in two ways that Eq. 4 does not separate. It adds the penalty $\beta_l \mathrm{KL}_l$, and it replaces a deterministic computation by a sampled one. Only the first is compression in the IB sense; the second is the price of the reparameterized sampler itself, and it is paid even when the penalty is numerically negligible. Since the two are confounded at every finite β, we separate them by a limit.

Sampling-noise floor. Define

$$\phi_l \;=\; \lim_{\beta \to 0^+} A_l(\beta) - A_0, \tag{6}$$

the *sampling-noise floor* of layer l: the accuracy cost of making layer l stochastic, with no compression pressure at all. The limit is estimated empirically from the lowest grid points, where the *effective penalty*

$$P_l(\beta) \;=\; \beta\, \overline{\mathrm{KL}}_l(\beta) \tag{7}$$

is of order 10^{-3}–10^{-2} nats per sample, three orders of magnitude below the cross-entropy it is added to (Section 5.4); the two lowest points agree to within 0.35 pp, so the limit is attained on our grid and ϕ_l is measured, not extrapolated. Degradation beyond ϕ_l as β grows is then attributable to compression. Note that ϕ_l is a property of the *layer*, not of the optimizer's compression trade-off, and that it is invisible to any analysis that only reports accuracy at finite β.

Relative compression axis. Total KL is a rate, not a compression level: the same 100 nats is a severe constraint on a layer whose unconstrained rate is 200 nats and a mild one on a layer whose unconstrained rate is 4000. We therefore also report the *relative* rate

$$r_l(\beta) \;=\; \overline{\mathrm{KL}}_l(\beta) \big/ \overline{\mathrm{KL}}_l(\beta \to 0) \;\in\; (0, 1], \tag{8}$$

the fraction of the layer's unconstrained rate that survives; $r_l = 1$ is an uncompressed layer and $r_l \to 0$ is complete posterior collapse. Comparing layers at matched r asks a question that contains no β at all: for the same *fraction* of information removed, which depth pays more accuracy? Reporting A_l against β, against $\overline{\mathrm{KL}}_l$, and against r_l gives three answers of decreasing dependence on the control parameter, and their agreement or disagreement is itself informative.

Additivity of compression cost. The single-layer protocol measures costs one layer at a time, but the practice it is meant to inform — training a VIB network with several stochastic layers — applies them together. We therefore also train configurations in which a set S of layers is compressed simultaneously at the same β, and compare the joint drop $\Delta A_S(\beta)$ with the sum of the single-layer drops measured at that β and against the same baseline:

$$e_S(\beta) \;=\; \Delta A_S(\beta) \;-\; \sum_{l \in S} \Delta A_l(\beta). \tag{9}$$

$e_S < 0$ (super-additive) means the joint drop is larger in magnitude than the sum of the single-layer drops, i.e. the layers were relying on a defence that is exhausted once several of them are constrained at once; $e_S > 0$ (sub-additive) means their costs overlap. In both cases the sign of e_S is a direct constraint on mechanism M2 below (Section 6.2).

3.5 Dimension-normalized KL and the effective penalty

The KL in Eq. 3 is a sum over d_l bottleneck units, so its magnitude scales with the representation dimension. The gradient pressure exerted by the regularization is therefore governed by the *effective penalty* $\beta \cdot \mathrm{KL}_l$, and a fixed β imposes very different total pressure on a 32-dimensional bottleneck than on a 4096-unit feature map. We therefore report, alongside total KL, the dimension-normalized quantity $\overline{\mathrm{KL}}_l / d_l$ (nats per bottleneck unit), and we use the comparison of total vs. normalized KL across systems to assess whether β or $\beta \cdot \mathrm{KL}$ is the transferable quantity.

3.6 Numerical stability

An early version of our implementation, without the $\log \sigma^2$ clamp, reproduced a dramatic failure: concentrating the full compression budget on the last layer produced NaN losses on most seeds. Tracing the dynamics showed $\log \sigma^2$ diverging to large negative values, so that $\sigma = \exp(0.5 \log \sigma^2)$ underflowed and gradients exploded. This is an implementation artifact, not evidence about last-layer tolerance: with the clamp, the last layer trains stably across the entire grid on both systems. We report this because the artifact could easily be misread as "the last layer cannot be compressed," a conclusion our stable runs contradict.

4 Experimental Setup

System M (MNIST, MLP). Architecture $784 \to 256 \to 256 \to 128 \to 64 \to 10$ with ReLU activations; VIB bottleneck dimensions $d^b = (128, 128, 64, 32)$ at positions $l = 1..4$ (after each hidden layer), with linear projection back to the layer width. Training: Adam, learning rate 10^{-3}, batch size 256, 5 epochs, CPU. Seeds $\{42, 1, 2\}$ for the sweep; seeds $\{1, \ldots, 5\}$ for the allocation experiment (Section 5.1).

System C (CIFAR-10, CNN). Three conv/pool blocks ($3 \to 16 \to 32 \to 64$ channels, 3×3 convs, 2×2 max-pooling) followed by $1024 \to 128 \to 10$. VIB slots after each conv block (channel-wise 1×1 parameterization, dimensions 4096/2048/1024) and after fc1 (linear, dimension 128). Training: Adam, 10^{-3}, batch 256, 3 epochs (CPU budget), seed 42 for the main grid plus seeds $\{1, 2\}$ at selected configurations.

Grids. Both systems use the single-layer protocol of Section 3.2 and margin $\Delta = 1\%$. Write $\mathcal{B}_{\text{main}} = \{10^{-4}, 5{\times}10^{-4}, 10^{-3}, 3{\times}10^{-3}, 10^{-2}, 3{\times}10^{-2}\}$ and $\mathcal{B}_{\text{low}} = \{10^{-6}, 3{\times}10^{-6}, 10^{-5}, 3{\times}10^{-5}\}$. System M sweeps $\mathcal{B}_{\text{main}}$ at three seeds $\{42, 1, 2\}$ and $\mathcal{B}_{\text{low}}$ at seed 42. System C sweeps $\mathcal{B}_{\text{main}}$ at seed 42 and $\mathcal{B}_{\text{low}}$ at all three seeds, and adds seeds $\{1, 2\}$ to System C at the four main-grid configurations bracketing the deepest layer's threshold crossing. The low grid exists because on System C the main grid lies entirely above β_l^{crit} for three of the four layers (Section 5.3); on System M it supplies the $\beta \to 0$ reference rate needed by Eq. 8.

Multi-layer configurations. Four subsets are compressed simultaneously at a shared β: $S \in \{\{1,2\}, \{3,4\}, \{1,4\}, \{1,2,3,4\}\}$. System M runs these at $\beta \in \{10^{-3}, 3{\times}10^{-3}, 10^{-2}\}$ with three seeds; System C at $\beta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ with seed 42. These configurations are used only for the additivity statistic of Eq. 9; all β_l^{crit} and tolerance orderings come from the single-layer sweeps.

Reporting conventions. Every run reports end-of-training test-set KL, so that the regularization can be verified as active rather than assumed. Where three seeds exist we report mean±std and apply the $\Delta = 1\%$ threshold *per seed* against that seed's own baseline, taking β_l^{crit} to be the largest grid β at which a strict majority of seeds pass; this avoids the inconsistency of comparing a seed-averaged accuracy to a seed-averaged threshold. Where only seed 42 exists (System C main grid, multi-layer) we report the single number and treat differences below the System C baseline spread (1.08 pp, Section 5.3) as not established. In total the paper reports approximately 250 trained networks. No configuration was selected after seeing test results: the grids, seeds, and margin were fixed before the sweeps, and every run performed is reported.

5 Results

5.1 Fixed-budget allocation motivates the protocol

As a preliminary, we fixed the total budget $\sum_l \beta_l = 10^{-3}$ and compared five spread schedules (uniform, ascending, descending, random) against single-layer concentrations, on System M with 5 seeds (Table 1).

Schedule	Test acc. (%)	Note
NoIB	97.71 ± 0.08	baseline
Third-only	97.35 ± 0.27	best VIB config
Second-only	97.31 ± 0.19	
Ascending	96.90 ± 0.25	budget grows with depth
Uniform	96.45 ± 0.27	
Random	96.21 ± 0.20	
Descending	95.84 ± 0.33	
First-only	77.36 ± 33.78	seed-dependent collapse
Last-only†	44.32 ± 42.28	NaN instability (pre-clamp)

Table 1: Fixed total budget $\sum \beta_l = 10^{-3}$, System M, 5 seeds. †Run with the unclamped implementation (Section 3.6).

Concentrating the budget on a middle layer is nearly lossless (-0.4%), while concentrating it on the first layer collapses some seeds, and among spread schedules Ascending > Uniform >

Descending, i.e. placing more compression on deeper layers hurts less. The gaps among spread schedules are $< 1\%$ and we do not over-interpret them.

Confound and protocol consequence. In this experiment all four VIB modules were instantiated in every configuration, so layers with $\beta_l = 0$ still injected sampling noise; "First-only" therefore confounds compression at layer 1 with noise at layers 2–4. The clean single-layer sweep below partially revises these numbers (e.g. layer 1 at $\beta = 10^{-3}$ reaches 96.81% when it is the *only* stochastic layer). We retain Table 1 solely as motivation: with a fixed budget, the *location* of compression matters, which is what the single-layer protocol then measures without confounds.

5.2 System M: tolerance curves and critical strengths

Layer	10^{-4}	5×10^{-4}	10^{-3}	3×10^{-3}	10^{-2}	3×10^{-2}	β_l^{crit}
L1	97.48±0.17	97.01±0.10	96.83±0.25	96.44±0.29	94.23±0.43	14.61±4.62	3×10^{-3}
L2	97.41±0.17	97.36±0.22	97.21±0.29	97.06±0.31	96.64±0.10	95.83±0.11	3×10^{-3}
L3	97.50±0.34	97.46±0.09	97.41±0.08	97.43±0.05	97.09±0.09	96.45±0.13	10^{-2}
L4	97.58±0.16	97.68±0.03	97.40±0.08	97.38±0.14	96.97±0.31	96.30±0.38	3×10^{-2}

Table 2: System M: test accuracy (%, mean±std over 3 seeds) under single-layer compression; β_l^{crit} is the largest β at which per-seed accuracy stays above that seed's baseline minus 1%, on a majority of seeds.

Three seeds turn the single-seed picture of this sweep into a robust one. The critical strengths resolve to

$$\beta_1^{\text{crit}} \approx \beta_2^{\text{crit}} \approx 3\times10^{-3} \;<\; \beta_3^{\text{crit}} \approx 10^{-2} \;<\; \beta_4^{\text{crit}} \approx 3\times10^{-2}, \tag{10}$$

i.e. tolerance increases monotonically with depth over a ten-fold range in β, with the two shallow layers equally fragile and the final layer most tolerant. Standard deviations are small ($\leq 0.43\%$ outside the collapse regime), and the layer separation is clearest at $\beta = 10^{-2}$: L1 has already lost 3.2% (94.23 ± 0.43), L2 sits at the margin (96.64 ± 0.10), while L3 (97.09 ± 0.09) and L4 (96.97 ± 0.31) remain within it on most seeds; at $\beta = 3\times10^{-2}$ only L4 stays above threshold on a majority of seeds. The catastrophic L1 collapse at $\beta = 3\times10^{-2}$ ($14.6\% \pm 4.6\%$) reproduces on every seed, as does the graceful, monotone degradation of L2–L4. One single-seed artifact is corrected by the replication: at seed 42, L4 missed the threshold marginally at $\beta = 10^{-2}$ (96.59% vs. 96.62%), suggesting a penultimate-layer tolerance peak; the other two seeds pass L4 through $\beta = 3\times10^{-2}$, so the peak does not exist and tolerance is monotone in depth on System M as well.

The achieved KL falls monotonically with β in every layer (e.g. L1: $144 \to 1.2$ nats and L4: $81 \to 6.2$ nats across the grid at seed 42), confirming that the regularization is active and that the accuracy differences reflect genuine differences in the *cost* of compression, not differences in whether compression happened.

5.3 System C: replication and posterior collapse

NoIB baseline: $A_0 = 60.76\%$; threshold 59.76% (seed 42).

Three observations. **(i) The depth ordering reproduces.** At every β the final bottleneck retains by far the most accuracy (55.33% at $\beta = 3\times10^{-2}$, i.e. 91% of baseline), while L2 collapses to the 10% chance level for $\beta \geq 10^{-2}$ with KL = 0.03: the posterior there has collapsed onto the prior, z is pure noise, and the network is functionally severed at that depth. L1 degrades gradually ($52.9\% \to 25.4\%$) and the two shallow layers cross once between $\beta = 5\times10^{-4}$ and 10^{-3}; given the

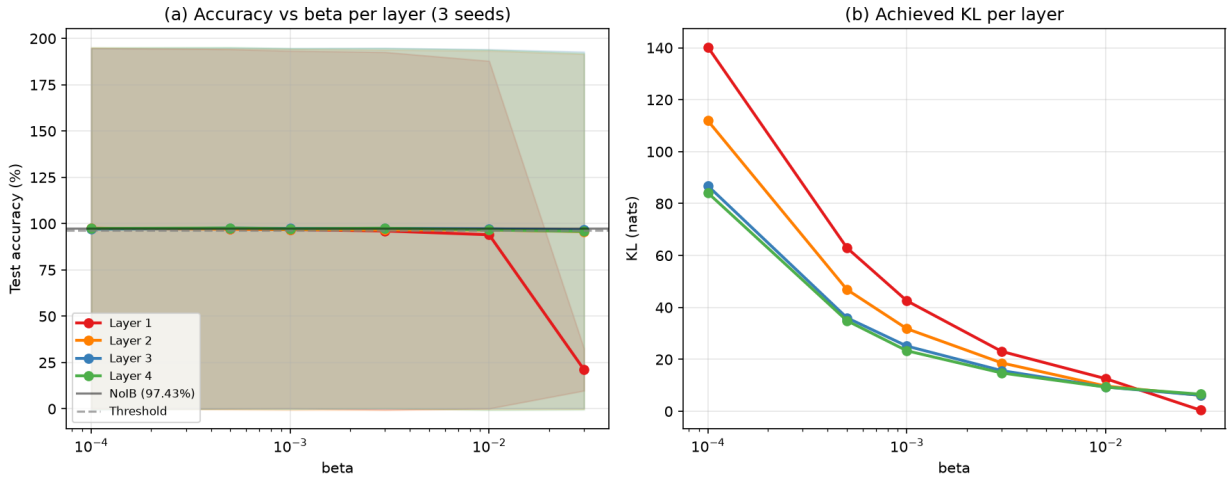

Figure 1: System M, single-layer bottleneck sweep, 3 seeds (mean line, shaded band: ±1 std). Left: test accuracy vs β; dashed line: failure threshold. Right: achieved KL per layer.

Layer	10^{-4}	$5{\times}10^{-4}$	10^{-3}	$3{\times}10^{-3}$	10^{-2}	$3{\times}10^{-2}$
L1 (conv1)	52.92	46.51	44.27	38.48	32.82	25.42
L2 (conv2)	53.90	47.26	42.03	30.16	10.00	10.00
L3 (conv3)	58.38	53.96	51.67	48.95	44.37	31.70
L4 (fc1)	58.60	59.44	59.25	58.63	58.34	55.33

Table 3: System C, seed 42: test accuracy (%) under single-layer compression. Every entry lies below the 59.76% threshold on the main grid; the ranking $L4 \gg L3 > \{L1, L2\}$ is stable, with L1 and L2 crossing once between $5{\times}10^{-4}$ and 10^{-3}.

baseline seed spread of 1.08 pp we treat them as tied. The reproducible statement is therefore $L4 \gg L3 > \{L1, L2\}$ rather than a strict ordering of the shallow pair, and the collapse at L2 remains the most dramatic failure (Section 6.4). **(ii) The absolute scale does not transfer, and a sampling-noise floor appears.** On the main grid no System C configuration passes the threshold, whereas the same grid brackets β_l^{crit} on System M. The low grid (Section 5.4) shows why: even as $\beta \to 10^{-6}$ the accuracy stays 0.8–4.5% below baseline, a layer-dependent *sampling-noise floor* — the cost of injecting stochasticity at all, before any compression pressure is applied — so that a threshold-based β_l^{crit} is undefined for three of four layers and the ordering must be read off the degradation curves. **(iii) The peak location agrees once seeds are accounted for.** At seed 42 alone, System M appeared to peak at the penultimate layer (L3) with a dip at L4; with three seeds this dip disappears (Section 5.2), and both systems show the same pattern: tolerance increases with depth and is maximal at the last bottleneck. The reproducible statement is therefore the full shallow<deep ordering, including the last-layer maximum.

5.4 System C: the low grid, β_l^{crit}, and the sampling-noise floor

The main grid does not bracket β_l^{crit} on System C, so we extended it downwards by two decades (Table 4).

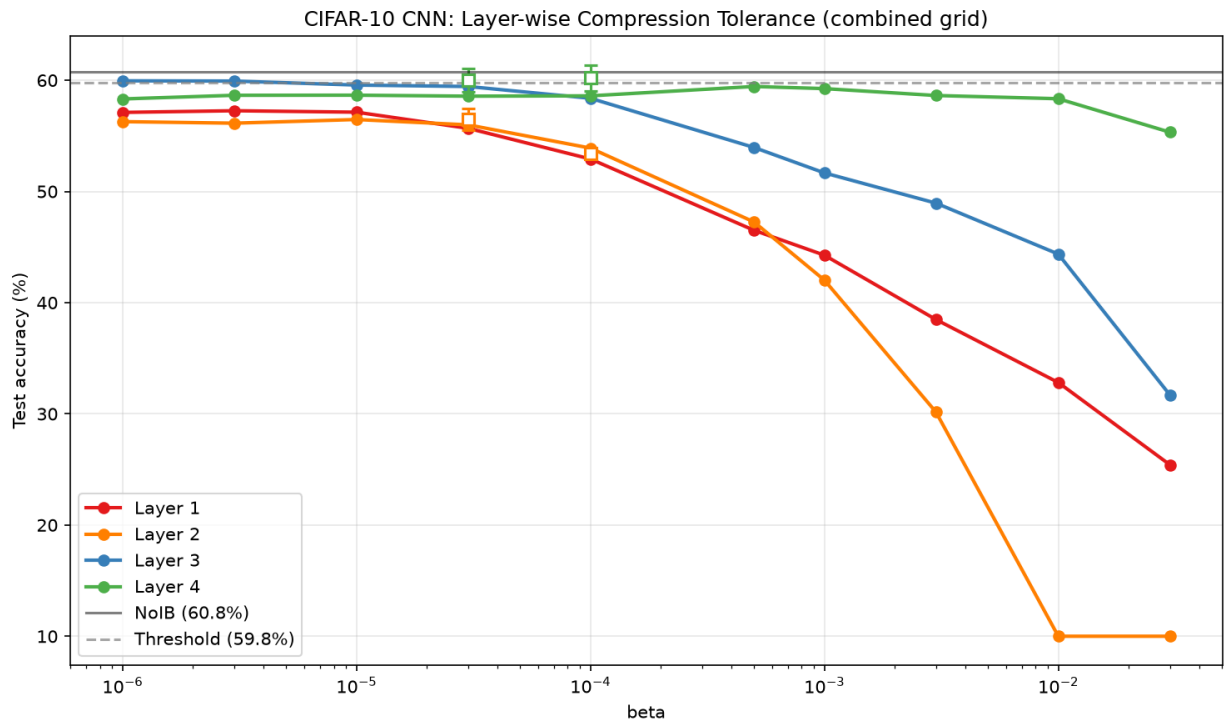

Figure 2: System C, single-layer bottleneck sweep over the combined grid (10^{-6}–$3{\times}10^{-2}$), seed 42, with multi-seed points where available. Dashed: threshold. The low grid locates each layer's crossing.

Compression is not yet acting at the bottom of the grid. Before interpreting the low-grid accuracies as a tolerance measurement we must establish that the $\beta \to 0$ limit of Eq. 6 has been reached, i.e. that the KL penalty is doing (almost) nothing there. Table 5 does this. At $\beta = 10^{-6}$ the effective penalty $P_l = \beta\overline{\mathrm{KL}}_l$ is $3{\times}10^{-4}$–$4{\times}10^{-3}$ nats per sample, against a cross-entropy of order 1.1–1.4 nats at these accuracies: the compression term is 10^{-3} of the term it is added to, smaller than the batch-to-batch noise in the cross-entropy itself. It is not possible for a 0.004-nat perturbation of the objective to cost 4.5 accuracy points.

The floor is large, and it is layer-dependent. Against $A_0 = 60.76\%$, the seed-42 floors are $\phi = (-3.65, -4.48, -0.81, -2.44)$ pp for L1–L4, i.e. simply making a convolutional block stochastic costs between 0.8 and 4.5 accuracy points before any compression is applied. This is not a small correction: it is larger than the entire $\Delta = 1\%$ margin for three of the four layers, and it is comparable to the accuracy differences that the fixed-budget allocation literature treats as meaningful (Table 1). On System M the corresponding floors at $\beta = 10^{-6}$ are -0.84, -0.57, $+0.08$, -0.13 pp for L1–L4 (seed 42): stochastic layers also pay a floor there, but it is an order of magnitude smaller, confirming that the floor is not an inevitable consequence of the reparameterized sampler itself.

Table 6 confirms, with three seeds, that the floor is real and layer-dependent. L3 is again the shallowest (-1.53 ± 1.24 pp at $\beta = 10^{-6}$), while L1 is the deepest on average (-5.35 ± 1.81 pp) and also the most seed-variable. The L1/L2 ranking swaps between seeds 1 and 2, so the robust statement is that the two shallow convolutional blocks have the deepest floors; the L3/L4 gap is smaller but consistent in sign.

Layer	10^{-6}	3×10^{-6}	10^{-5}	3×10^{-5}	β_l^{crit}
L1 (conv1)	57.11	57.26	57.13	55.68	—
L2 (conv2)	56.28	56.14	56.47	55.99	—
L3 (conv3)	59.95	59.93	59.57	59.44	3×10^{-6}
L4 (fc1)	58.32	58.65	58.66	58.57	—
NoIB	60.76 (threshold 59.76)				

Table 4: System C, seed 42, low grid $\mathcal{B}_{\text{low}}$: test accuracy (%). All four curves are flat between 10^{-6} and 3×10^{-6} (spread ≤ 0.33 pp) and begin to fall only at 10^{-5}–3×10^{-5}. Only L3 ever crosses the threshold.

Layer	10^{-6}	3×10^{-6}	10^{-5}	3×10^{-5}
L1 (conv1)	4.4×10^{-3}	1.1×10^{-2}	2.3×10^{-2}	3.5×10^{-2}
L2 (conv2)	3.6×10^{-3}	8.5×10^{-3}	1.9×10^{-2}	3.2×10^{-2}
L3 (conv3)	2.9×10^{-3}	7.1×10^{-3}	1.7×10^{-2}	3.4×10^{-2}
L4 (fc1)	3.4×10^{-4}	1.0×10^{-3}	3.1×10^{-3}	8.1×10^{-3}

Table 5: System C, seed 42: effective penalty $P_l = \beta\overline{\text{KL}}_l$ (nats per sample) on the low grid. Compare with a cross-entropy of $\sim$1.2 nats. Even at the top of the low grid, P_l is 40$\times$ smaller than the term it perturbs.

Two consequences follow immediately. First, β_l^{crit} as defined in Eq. 5 is *undefined* for L1, L2 and L4 on System C: no β, however small, restores them to within 1% of baseline, because the floor alone already exceeds the margin. Reporting "β_l^{crit} does not exist" would be the honest output of the statistic, and it is exactly the kind of statement that a reader would reasonably misread as "these layers tolerate no compression." The degradation curves in Figure 2, read relative to each layer's own floor, show that they do: from $\beta = 10^{-6}$ to $\beta = 3\times10^{-2}$ L1 loses a further 31.7 pp and L4 only 3.0 pp. We therefore use β_l^{crit} for cross-*layer* comparison only where the floor is under control (System M), and use the curves themselves elsewhere.

Second, the floor explains why the main grid looked uniformly bad on System C. The main grid starts at $\beta = 10^{-4}$, which is two decades above the only crossing we can locate and well above the point where the floor exhausts the margin; every main-grid entry in Table 3 therefore reflects floor *plus* substantial compression, and none of them can pass a baseline-referenced threshold. The non-transferability of β_l^{crit} across systems is thus partly an artifact of the baseline referencing, not only of the compression dynamics.

The floor is not ordered by depth. Unlike the tolerance ordering, ϕ_l does not increase monotonically with depth: L2 has the deepest floor (-4.48 pp) and L3 the shallowest (-0.81 pp), with L1 and L4 in between. The floor and the tolerance ordering are therefore different phenomena and must not be read as the same quantity measured at $\beta = 0$. We return to this in Section 6.3.

5.5 The compression–accuracy frontier: removing β from the comparison

Every comparison so far has been indexed by β — which is precisely the quantity this paper argues is not comparable across layers or systems. We now remove it. Each run produces the triple $(\beta, \overline{\text{KL}}_l, A_l)$, so the same measurements can be re-plotted as accuracy against the *achieved* rate rather than against the control parameter, giving a per-layer compression–accuracy frontier in

Layer	floor at $\beta = 10^{-6}$ (pp) s42	s1	s2	floor at $\beta = 10^{-5}$ (pp) mean±std	mean±std
L1 (conv1)	−3.65	−7.85	−4.55	-5.35 ± 1.81	-5.96 ± 2.36
L2 (conv2)	−4.48	−6.32	−3.13	-4.64 ± 1.31	-4.56 ± 1.43
L3 (conv3)	−0.81	−3.27	−0.51	-1.53 ± 1.24	-1.53 ± 0.87
L4 (fc1)	−2.44	−3.55	−0.37	-2.12 ± 1.32	-2.00 ± 1.26

Table 6: CIFAR-10 sampling-noise floor $\phi_l = A_l(\beta) - A_0$ (pp) at the lowest two grid points, measured per seed. The two lowest points agree within ∼1 pp for every layer except L1 (seed 1), so the $\beta \to 0$ limit is attained on the grid. L3 has the shallowest floor on all three seeds; L1 and L2 are the deepest and the most seed-variable.

which β does not appear. We use two axes: the absolute rate $\overline{\mathrm{KL}}_l$ in nats, and the relative rate r_l of Eq. 8.

Matched absolute rate (System C). Table 7 interpolates the System C curves at fixed achieved rates. At 15 nats the four layers retain 32.8 (L1), 29.4 (L2), 43.6 (L3) and 58.3 (L4) percent accuracy against a 60.8% baseline — a 29-point spread produced by exactly the same number of nats. At 100 nats the spread is still 14.7 points (44.6/45.6/51.8/59.3). Since β nowhere enters this comparison, the depth ordering cannot be an artifact of the dimension-dependence of β documented in Section 5.6; it is a property of where the information is removed, not of how hard the knob was turned. L1 and L2 are close and exchange places between the 100-nats and 15-nats rows (and exchange places once in the β plane as well, between $\beta = 5{\times}10^{-4}$ and 10^{-3}); with a baseline spread of 1.08 pp we read them as tied.

Layer	@ 500 nats	@ 100 nats	@ 15 nats
L1 (conv1)	52.6	44.6	32.8
L2 (conv2)	54.0	45.6	29.4
L3 (conv3)	57.8	51.8	43.6
L4 (fc1)	—	59.3	58.3
NoIB baseline 60.76%; L4's unconstrained rate is 342 nats, so 500 nats is outside its measured range.			

Table 7: System C, seed 42: test accuracy (%) interpolated at matched achieved rate. Reading down a column compares layers under identical compression with no reference to β.

Matched relative rate. Absolute rate is still not like-for-like, because the unconstrained rate $\overline{\mathrm{KL}}_l(\beta \to 0)$ that it is measured against varies by 13× across the layers of System C (4377 nats at L1 versus 342 at L4). The relative axis r_l removes this: $r_l = 0.1$ means "this layer retains 10% of the rate it would use if it were not constrained at all," which is a comparable statement across layers of very different width. Table 8 gives both systems on this axis.

On System C the ordering $L4 > L3 > \{L1, L2\}$ holds at every r, with a spread of 6.9 points at $r = 0.1$ growing to 19.0 points at $r = 0.01$. On System M the same ordering is visible only at the extremes: at $r = 0.01$, L1 has collapsed to 59.8% while L2–L4 remain at 95.8–96.6%. Over $r \in [0.1, 0.5]$ the System M curves lie within 1.1 pp of each other and of the baseline, which is the expected consequence of a saturated task rather than evidence about ordering — the entire

System	Layer	r=0.5	r=0.2	r=0.1	r=0.03	r=0.01
M	L1	97.12	97.24	96.79	93.75	59.77
	L2	96.90	97.42	97.02	96.60	95.72
	L3	97.61	97.07	97.51	97.30	96.60
	L4	97.71	97.45	97.59	96.85	95.83
C	L1	57.04	54.67	51.90	45.67	39.00
	L2	56.43	54.91	52.42	46.11	36.27
	L3	59.51	58.27	55.75	51.30	47.23
	L4	58.67	59.34	58.75	56.60	55.33

Table 8: Test accuracy (%) interpolated at matched *relative* rate r_l (Eq. 8); larger r = less compression, so each row should be read right-to-left as compression increases. Baselines: 97.62% (M, seed 42), 60.76% (C, seed 42).

dynamic range available on MNIST, from no compression to catastrophic collapse, is about 3 pp until the collapse itself. We therefore treat System C as the system that can resolve the frontier and System M as the system that resolves β_l^{crit}; the two agree on the ordering wherever both are informative.

Synthesis. Three parameterizations now give the same answer on System C: ordered by β (L1–L4 gap 29.9 pp at $\beta = 3{\times}10^{-2}$), by absolute rate (29.0 pp at 15 nats), and by relative rate (19.0 pp at $r = 0.01$). The *ordering* is invariant to the choice of axis; only the *magnitude* of the gap depends on it. This is the precise sense in which the paper claims the depth ordering to be a robust finding, and it is a stronger claim than stability under reseeding alone, because reseeding does not test the choice of control variable.

5.6 Cross-system comparison: KL scaling with dimension

	Layer	total KL (nats)	KL / d_l (nats/unit)	Δ acc (%)
System M	L1 (d^b=128)	143.9	1.12	−0.12
	L2 (d^b=128)	112.5	0.88	−0.10
	L3 (d^b=64)	89.5	1.40	−0.59
	L4 (d^b=32)	80.7	2.52	−0.23
System C	L1 (d=4096)	529.0	0.13	−7.84
	L2 (d=2048)	480.8	0.23	−6.86
	L3 (d=1024)	587.2	0.57	−2.38
	L4 (d=128)	182.8	1.43	−2.16

Table 9: Achieved KL and accuracy drop at $\beta = 10^{-4}$ (seed 42). Total KL is several-fold larger on System C and coarsely tracks d_l; per-unit KL is within one order of magnitude across systems and increases with depth on both. Baselines: 97.62% (M), 60.76% (C).

Table 9 compares the two systems at matched $\beta = 10^{-4}$. Total end-of-training KL on System C exceeds System M by 2–7× while raw dimensions differ by up to 32×; per-unit KL differs by at most ∼9× and, on both systems, *increases with depth*: the optimizer preserves more nats per unit in deeper bottlenecks under identical β. But dimension normalisation is only the first step; the key question is which quantity, if any, predicts the accuracy drop.

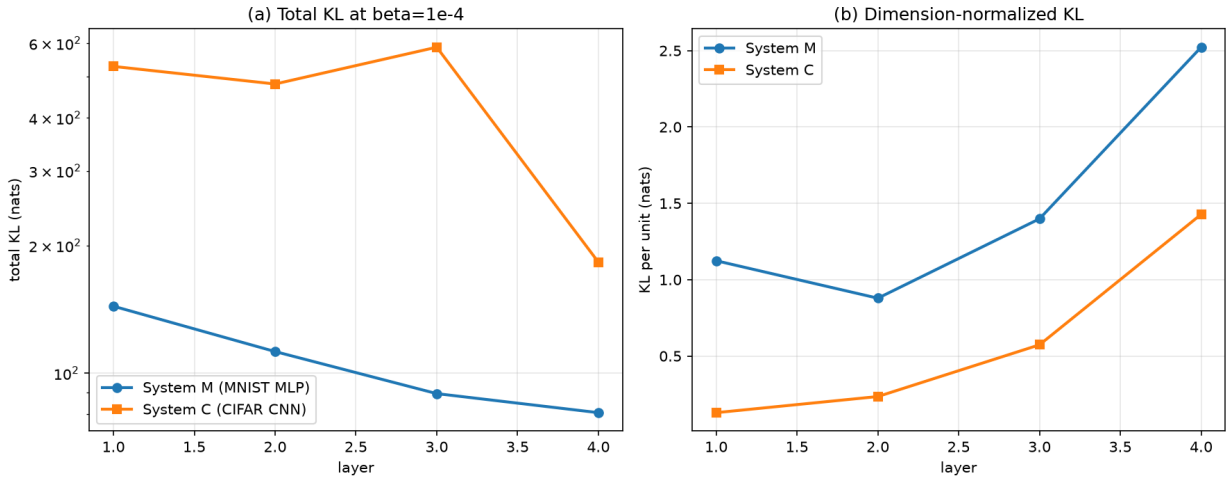

Figure 3: Achieved KL at $\beta = 10^{-4}$ (seed 42) as a function of layer depth. (a) Total KL: System C exceeds System M several-fold. (b) Per-unit KL: the two systems come within one order of magnitude, and both increase with depth.

Table 9 rejects the simplest candidate. At matched β the System C drop is 4–65× larger than the System M drop (geometric mean ∼22×). That ratio narrows when we move to the more layer-intrinsic relative-rate axis r of Section 5.5: at $r = 0.1$ the per-layer drops on C are 7.7× (L1), 8.1× (L2), 7.7× (L3) and only 1.8× (L4) larger than on M. The effective penalty β KL is the term that physically enters the objective, so any comparison that ignores it compares different training problems; yet it is not by itself a transferable tolerance measure. We therefore draw three conclusions: (i) raw β is a poor control variable across layers or systems; (ii) per-unit KL is a necessary normalisation because total KL scales with dimension; (iii) the relative rate r is the most layer-intrinsic axis we can construct, but the absolute scale of accuracy drop still varies across systems. The robust, transferable finding is the ordering, not the numerical value of any per-layer threshold.

5.7 Multi-layer compression costs are not additive

The single-layer protocol intentionally isolates one depth at a time. A multi-layer VIB network, however, compresses several layers together, and mechanism M2 (Section 6.2) predicts that the single-layer measurements are not a modular building block: once the network runs out of unconstrained layers to compensate for the bottleneck, the joint cost should exceed the sum of the parts. We tested this with the subsets $\{1, 2\}$, $\{3, 4\}$, $\{1, 4\}$ and $\{1, 2, 3, 4\}$ (System M, three seeds; System C, seed 42) and the excess-cost statistic of Eq. 9.

The shallow pair $\{1, 2\}$ on System M shows the strongest possible violation of additivity. At $\beta = 10^{-3}$, where each layer alone is essentially within the margin (L1 drop -0.81, L2 drop ≈ 0), compressing them together drops accuracy by 2.94 ± 0.32 pp — an excess cost of -2.13 pp. At $\beta = 3{\times}10^{-3}$, where each layer is still near 97%, the pair collapses to 11.35% on all three seeds, with KL $\to 0.001$ indicating posterior collapse at both bottlenecks simultaneously. The excess cost is then -84.7 pp: the network has been functionally severed at two depths at a β where neither single bottleneck does serious harm.

The deep pair $\{3, 4\}$ behaves differently. At $\beta = 10^{-3}$ the joint drop (-0.08 ± 0.32 pp) is indistinguishable from the sum of the single-layer drops (-0.05 pp): two tolerant layers together

System	Pair / set	β	ΔA_S (pp)	$\sum \Delta A_l$ (pp)
M	L1+L2	10^{-3}	-2.94 ± 0.32	-0.81
	L1+L2	$3{\times}10^{-3}$	-86.08 ± 0.29	-1.35
	L1+L2	10^{-2}	-86.08 ± 0.29	-3.98
	L3+L4	10^{-3}	-0.08 ± 0.32	-0.05
	L3+L4	$3{\times}10^{-3}$	-0.49 ± 0.23	-0.05
	L3+L4	10^{-2}	-1.29 ± 0.10	-0.80
	L1+L4	10^{-3}	-2.68 ± 0.29	-0.63
	L1+L4	$3{\times}10^{-3}$	-36.31 ± 35.49	-1.04
	L1+L4	10^{-2}	-86.68 ± 0.37	-3.66
	all four	10^{-3}	-86.38 ± 0.40	-0.86
	all four	$3{\times}10^{-3}$	-86.41 ± 0.20	-1.40
	all four	10^{-2}	-86.51 ± 0.41	-4.79
C	L1+L2	10^{-4}	-19.84	-14.70
	L1+L2	10^{-3}	-50.76	-35.22
	L1+L2	10^{-2}	-50.76	-78.70
	L3+L4	10^{-4}	-2.35	-4.54
	L3+L4	10^{-3}	-11.36	-10.60
	L3+L4	10^{-2}	-32.16	-18.81
	L1+L4	10^{-4}	-7.90	-10.00
	L1+L4	10^{-3}	-18.90	-18.00
	L1+L4	10^{-2}	-34.62	-30.36
	all four	10^{-4}	-30.88	-19.24
	all four	10^{-3}	-50.52	-45.82
	all four	10^{-2}	-50.79	-97.51

Table 10: Joint accuracy drop of multi-layer compression versus the sum of single-layer drops. Negative "excess" $e_S = \Delta A_S - \sum \Delta A_l$ means the joint cost is super-additive. System M averages over seeds $\{42, 1, 2\}$; System C uses seed 42 only.

are essentially additive when compression is weak. As β grows the joint cost becomes super-additive (-0.49 pp and -1.29 pp excess at $3{\times}10^{-3}$ and 10^{-2}), but the effect is small compared with the shallow pair. The mixed pair $\{1, 4\}$ is intermediate and highly seed-dependent at $3{\times}10^{-3}$: one seed collapses (-86.8 pp) while the other two lose only 6–16 pp, reflecting the fragility of pairing a compressible shallow layer with a tolerant deep one. Finally, compressing *all four* layers at $\beta = 10^{-3}$ collapses every seed (-86.4 ± 0.4 pp) even though the sum of the four single-layer drops at that β is only -0.86 pp. A global β that looks safe when applied to one layer is therefore catastrophic when applied to the whole network.

System C shows the same qualitative pattern but with larger numbers because every layer there has a deeper sampling-noise floor. The shallow pair $\{1, 2\}$ is super-additive at $\beta = 10^{-4}$ and 10^{-3}, collapsing at the latter; at $\beta = 10^{-2}$ both single layers are already at chance level, so the joint cost cannot exceed the sum and the excess flips sign. The deep pair $\{3, 4\}$ is actually sub-additive at $\beta = 10^{-4}$ (joint drop -2.35 pp versus sum -4.54 pp), suggesting that the two late bottlenecks share redundant information that a single deep bottleneck cannot fully exploit. At stronger compression the pair becomes super-additive. The mixed pair $\{1, 4\}$ and the all-four configuration follow the same broad trend: super-additive while at least one layer still carries information, bounded from below by chance once everything has collapsed. The central practical point is unchanged: the cost of applying the same β to several layers is not the sum of the single-layer costs, and a global β can produce a collapse that no single-layer measurement would predict.

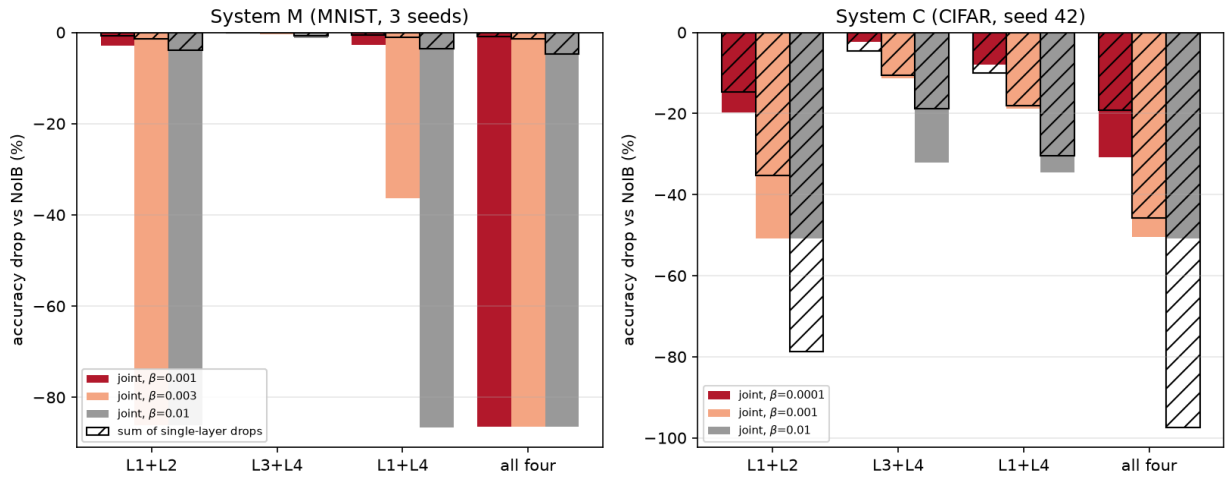

Figure 4: Multi-layer compression on System M (three seeds) and System C (seed 42). Solid bars: joint accuracy drop of the pair/set at a shared β; hatched bars: sum of the corresponding single-layer drops. Where the solid bar is lower (more negative), the cost is super-additive.

The L1+L2 collapse has a direct practical reading. A VIB practitioner who sets a single global β that is safe for each layer individually may still destroy the network when that β is applied to several layers at once, because the layers share a finite amount of "compensation budget." This is the central multi-layer misspecification implied by a global β: it treats compression as a one-dimensional resource when the relevant quantity is a budget that must be allocated across depths.

6 Discussion

6.1 What reproduces, and what does not

Across two architectures and datasets, three seeds on System M, and two grids on System C, the following is stable: (a) compression tolerance is layer-dependent, with differences of an order of magnitude in β; (b) tolerance increases with depth and is maximal at the last bottleneck, shallow bottlenecks being the fragile ones; (c) achieved KL is monotone in β everywhere, so the differences are differences in cost, not in compression effort; (d) the depth ordering is invariant to whether one indexes curves by β, by absolute rate, or by relative rate; (e) every stochastic layer pays a sampling-noise floor that is present before any compression pressure is applied, and the floor is large enough on System C to make the standard β_l^{crit} statistic undefined for three of four layers; (f) compressing two shallow layers together produces a super-additive collapse at a β where each layer alone is nearly unharmed.

What does *not* transfer is the absolute scale of tolerance. Matched β accuracy drops differ by an order of magnitude across systems; moving to the more layer-intrinsic relative-rate axis narrows but does not remove the gap. We therefore state as the paper's central claims the depth ordering, the layer-dependence of the sampling-noise floor, and the non-additivity of multi-layer compression. The numerical value of any single-layer β_l^{crit} should be treated as a system-specific snapshot, not as a transferable hyperparameter.

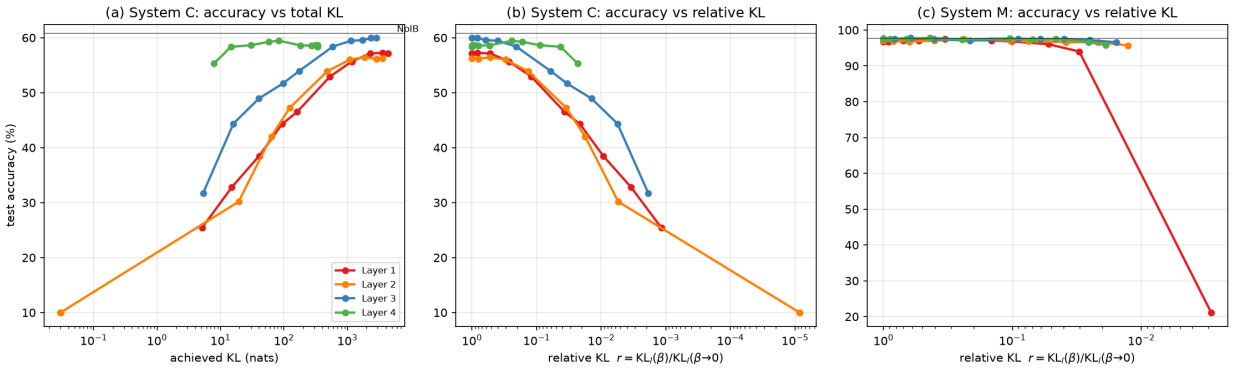

Figure 5: Compression–accuracy frontiers on System C (left and centre) and System M (right). (a) Accuracy vs achieved total KL; (b) accuracy vs relative rate $r = \mathrm{KL}(\beta)/\mathrm{KL}(\beta \to 0)$ on System C; (c) the same on System M. In each plot the accuracy axis is the one that matters, and the horizontal axis is a compression axis that does not contain β.

6.2 Candidate mechanisms

We consider three non-exclusive mechanisms for the shallow<deep ordering, now informed by the floor and additivity results.

(M1) Input-information burden. Shallow representations must still carry the input information that all downstream computation needs, including task-irrelevant detail that a bottleneck cannot selectively discard; deep representations have already shed irrelevant variation, so isotropic compression there removes proportionally more redundancy and less signal. This is the information-level analogue of the redundancy invoked to explain why late and fully connected layers prune well [9]. It predicts the monotone depth trend seen on both systems, the higher per-unit KL retained by deep layers at matched β (Table 9), and — crucially — the higher per-unit rate even at the noise floor: System C L3 and L4 retain 2.7–2.8 nats per unit at $\beta = 10^{-6}$, while L1 retains only 1.1 nats per unit. Deep layers thus carry more information per unit even when no compression is enforced, so losing the same fraction of their rate costs less accuracy.

(M2) Compensation by unconstrained layers. In the single-layer protocol, compressing layer l leaves all other layers free to adapt. If a *deep* layer is compressed, the shallow deterministic layers can pre-encode information redundantly to protect it; if layer 1 is compressed, no upstream capacity exists and the whole network must live with the degraded signal. The bottleneck's position thus determines how much compensating capacity is available, predicting the shallow<deep ordering. The additivity experiment of Section 5.7 tests this directly, and confirms the strongest prediction of M2: when both shallow layers are compressed together the cost is super-additive, and the network collapses at a β where neither single-layer curve has crossed its threshold. This can only happen if the unconstrained layers normally absorb some of the cost of each bottleneck and are overwhelmed when they must absorb both.

(M3) Noise propagation. Sampling noise injected at layer l passes through $L - l$ nonlinear maps before the readout. Early noise is transformed and potentially amplified along the feature hierarchy, whereas late noise enters the classifier nearly additively. M3 is naturally tied to the sampling-noise floor ϕ_l rather than to the tolerance ordering: on System C L2 has the deepest

floor (−4.48 pp) even though it is not the most intolerant layer under compression, which is the signature of a noise-propagation effect. Conversely, L3 has the shallowest floor (−0.81 pp) and is also the most tolerant under compression, so its robustness is partly a floor effect and partly a tolerance effect. We regard M1+M2 as the primary explanation of the shared tolerance ordering and M3 as the primary explanation of the floor pattern, with all three mechanisms contributing to the quantitative details.

Disentangling these mechanisms cleanly would require dedicated follow-up experiments (freezing subsets of layers to remove compensation, replacing the VIB sampler with matched additive noise to isolate propagation, or artificially inflating the dimension of shallow bottlenecks to test M1); our data establish the phenomenon and constrain the mechanism space but do not resolve it.

6.3 Sampling-noise floor versus tolerance ordering

A central methodological point of this paper is that the accuracy of a VIB layer at very small β is not the same as the accuracy of the corresponding deterministic layer. Every stochastic layer pays a sampling-noise floor ϕ_l, and on System C that floor is large enough to dominate the first several decades of the curve. This has two consequences for how the field should report and compare tolerance measurements.

First, a baseline-referenced statistic such as β_l^{crit} should be interpreted cautiously, or even avoided, when the floor itself exceeds the chosen margin. On System C, β_l^{crit} is undefined for L1, L2 and L4; reporting "does not exist" would be correct but easy to misread as "intolerant." We recommend reporting the floor-corrected degradation $A_l(\beta) - A_l(\beta \to 0)$ alongside $A_l(\beta) - A_0$ whenever the floor is non-negligible.

Second, the floor and the tolerance ordering are not the same thing and can disagree. On System C L2 has the deepest floor but L1 degrades more as β increases; L3 has the shallowest floor and is also the most tolerant. A mechanism that only explains the floor (e.g. noise propagation) is therefore not sufficient to explain the ordering, and a mechanism that only explains the ordering (e.g. information burden) is not sufficient to explain the floor. Both must be measured and modelled separately.

6.4 Posterior collapse as a layer-localized failure

The L2 collapse on System C (KL $\to$ 0.03, accuracy at chance) is a discriminative analogue of posterior collapse [10]: once the marginal value of information through that layer falls below the KL price set by β, the optimum jumps to the prior and the representation becomes pure noise. Three features are notable. First, the collapse is *abrupt* at L2 (42.0% $\to$ 30.2% $\to$ 10.0% over one step of the grid) but *gradual* at L1, suggesting a phase-transition-like threshold whose sharpness depends on how replaceable the layer's information is: a partially informative L1 code still helps, whereas a partially informative L2 code appears not to be exploitable by the downstream network trained end-to-end under that pressure. Second, collapse at a middle layer is catastrophic for the *whole* network even though three of four layers remain deterministic and uncompressed — a vivid illustration of why per-layer tolerance, not total budget, is the right object to control. Third, the same collapse can be induced on System M by compressing L1 and L2 together (Section 5.7), so collapse is not a property of a single layer at an extreme β but a property of how much compensating capacity the rest of the network has left.

6.5 Practical implication: the global-β convention is misspecified

The dimension-scaling of achieved KL (Table 9) has a direct practical reading: a global β imposes a dimension-proportional effective penalty, so wide layers are compressed far harder than narrow ones under nominally identical settings. But the problem is deeper than dimension scaling. The sampling-noise floor means that even very small β changes the network at every stochastic layer, and the additivity result means that a global β that is safe per layer can be catastrophic when applied to several layers at once. A single scalar therefore controls at least three different things: compression strength, noise injection, and multi-layer compensation budget.

A first-order correction is per-layer scaling designed to equalise the *effective* compression rather than the knob value. Two natural recipes are $\beta_l = \beta_0/d_l$ (matching per-unit KL pressure) or $\beta_l = \beta_0/\mathrm{KL}_l(\beta \to 0)$ (matching the relative rate r_l). Both are suggested by the data but neither is validated end-to-end here. More ambitiously, a multi-layer VIB schedule could treat the vector $(\beta_1, \ldots, \beta_L)$ as the object to tune, with the constraint that the sum of compensation demand (measured by $\beta_l \mathrm{KL}_l$ or by the floor-corrected drop) not exceed the network's capacity. The empirical design of such schedules is the most concrete application of this work, and the single-layer protocol introduced here is the measurement tool on which any schedule would have to be built.

6.6 Threats to validity

(i) KL is a variational proxy; true $I(X; Z_l)$ may rank layers differently [7, 8]. (ii) MNIST accuracy saturates near 98%, compressing the dynamic range; System C mitigates this but at 3 epochs the baseline is 60.8%. (iii) $\Delta = 1\%$ is conventional, not principled; near-threshold points are seed-sensitive, which is why we report full curves and per-seed thresholds. (iv) Two systems is two systems; architectures with residual connections or attention may redistribute compensation capacity (M2) and change the ordering. (v) CPU-scale training limits epochs and seeds; the multi-seed System C points bracket only the crossing region. (vi) The sampling-noise floor may be an artefact of the VIB sampler architecture (linear / 1×1-conv parameterization); a different encoder family could change its size but would not change the existence of a floor for a stochastic layer. (vii) The additivity experiment keeps the same β at every active layer; schedules with per-layer β_l could in principle cancel the super-additivity we observe.

7 Limitations

- Two systems; System C is trained for only three epochs and its main sweep is at seed 42.
- KL is a variational proxy; no direct MI estimation.
- The sampling-noise floor and the additivity result are measured only with the linear / 1×1-conv VIB parameterization used here; other encoder families may change their quantitative values.
- Mechanisms M1–M3 are hypotheses; the protocol establishes the phenomenon and constrains the mechanism space but does not decompose it.
- Additivity is tested at a small number of layer subsets and β values; the observed sub-additivity / super-additivity pattern may not generalise to other subsets.
- No end-to-end validation of dimension-normalized or per-layer β schedules (Section 6.5).

8 Conclusion

We introduced a single-layer bottleneck protocol and the per-layer critical compression strength β_l^{crit}, and used them to map how the cost of variational compression is distributed across depth. On two systems, tolerance increases with depth by roughly an order of magnitude in β, shallow layers degrade or collapse under compression that deep layers absorb almost losslessly, and middle-layer posterior collapse is a catastrophic, layer-localized failure mode. Three further findings sharpen the conclusion. First, every stochastic layer pays a sampling-noise floor that is present before any compression pressure is applied; on CIFAR-10 this floor is large enough to make the standard β_l^{crit} statistic undefined for three of four layers, so tolerance must be read from the degradation curves rather than from a single threshold. Second, the depth ordering is robust to whether one indexes curves by β, by absolute rate, or by relative rate; only the magnitude of the gap depends on that choice. Third, multi-layer compression costs are not additive: compressing the two shallow layers together triggers posterior collapse at a β where each layer alone is nearly unharmed, showing that a global β is not just a suboptimal control but a potentially catastrophic one.

These results turn the distribution of compression tolerance across depth — a quantity the global-β convention assumes away — into a measurable, reproducible object. They also suggest a research programme: move from global schedules to per-layer schedules that target either per-unit KL or relative rate, and validate those schedules against the single-layer tolerance curves that the protocol produces. Until that is done, the safest practical conclusion is that any multi-layer VIB network trained with a single β is silently solving different compression problems at different depths.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. Allerton Conference on Communication, Control and Computing*, 1999.

[2] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *Proc. IEEE Information Theory Workshop*, 2015.

[3] A. A. Alemi, I. Fischer, J. V. Davis, and K. Murphy. Deep variational information bottleneck. In *Proc. ICLR*, 2017.

[4] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proc. ICLR*, 2017.

[5] A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018.

[6] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-VAE. In *Proc. NeurIPS*, 2018.

[7] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019.

[8] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. Estimating information flow in deep neural networks. In *Proc. ICML*, 2019.

[9] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural networks. In *Proc. NeurIPS*, 2015.

[10] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *Proc. CoNLL*, 2016.