Contents

The Gate Does Not Choose: An Expert-Initialised Mixture-of-Experts Outperforms Training-Free Consensus under the Official BraTS-2023 Metrics

Guillaume Cassez · Stanislas Larnier

Independent research

Guillaume Cassez — ORCID 0009-0007-0987-3931 · cassez.guillaume@gmail.com · guillaume-cassez.fr
Stanislas Larnier — stanislaslarnier@gmail.com · HAL stanislas-larnier

Version 0.9 (2026-09-11) — the thesis is inverted on the measurement, and the fourth expert is scored. By author decision (2026-09-11), the MoE V3 replaces the V2 everywhere in this paper and the blob-loss expert G enters as an arm in its own right. All twenty-nine arms — the four dense experts alone, the twenty-four deterministic consensus operators built from them, and the two learned patch-wise gates — are re-scored by one code path under the official protocol on the same 239 patients of each of five disjoint folds (`scripts/paper3_v3_vs_consensus.py`, which reproduces the published fold-0 delta of the V3 to 0.00e+00 before writing anything). The V3 is first at +0.0057 over its own baseline, positive on five of five folds and above the pre-declared smallest effect of interest (0.003) on four; the best consensus reaches +0.0013 and **none of the twenty-four** reaches that effect of interest.

This version discharges the promise the previous one made about G. The arm is written with its own official numbers, read from its five checkpoints and from the same scorer as every other arm: trained for 300 epochs on each of the five folds with its signed-distance term switched off ($\lambda = 0.0$ in all five, so it differs from the baseline by the blob term alone, $\beta = 0.5$), it scores −0.0038 Dice over the five folds — negative on all five, 28-th of the twenty-nine arms and last of the four dense experts. The veto it applies to the baseline (`cc_B_G`) is the best of the twelve two-expert vetoes at +0.0011, and the second-best of the twenty-four consensus operators. Two earlier statements are corrected here, on the measurement and not on the wording of the card. The arm was described as scored only through the gate, which its five per-fold deltas contradict — and which the Abstract, Section 4.5 and the Conclusion had already been quoting against it. That veto was ranked runner-up among the consensus pairs, which the ranking contradicts in both readings of "pair": it is first among the twelve two-expert vetoes, and the arm ahead of it among the twenty-four (`cc_B_and_DKG`, +0.0013) is a one-against-three veto, not a pair. v0.7 is frozen bit for bit in `versions/v2/`.

Version 0.10 (2026-09-12) — the captions follow the figures. Figures 1, 2 and 4 were re-cooked on the V3 measurement (`scripts/make_paper3_v3_figures.py`: 26 arms, three V3 seeds, folds 1–4, measured cost) while their captions still described the previous campaign. Figure 1 counted twenty arms against twenty-six in Table 2; Figure 4 announced that the effect changes sign from fold to fold, against a sign that holds on all four folds, on all five and on all three seeds; Figure 2 did not state that none of the consensus Dice margins has a CI excluding zero. The three captions are rewritten from the measurement, with no value typed by hand, and the LaTeX captions are rendered by a conversion demonstrated on the seven captions it does not touch.

Version 0.11 (2026-09-21) — the author contributions are stated. An "Author contributions" section is added after §7, mirroring Papers 1 and 2: Guillaume Cassez, lead author (study design, model training, evaluations and analyses, manuscript writing); Stanislas Larnier, formulation of the research questions, methodological guidance and careful reviews of successive drafts. No number is modified; the body inherited from v0.10 is otherwise unchanged.

Provenance. Every quantitative claim comes from `scripts/paper3_stats.py` (`outputs/paper3/paper3_stats.json`); tables from `scripts/make_paper3_tables.py` (`tables/table_paper3.{md,tex}`); figures from `scripts/make_paper3_figures.py`. Metrics are computed by the official BraTS-2023 capsule (`external/BraTS-2023-Metrics`), never by a local reimplementation. The complete official metric set is reported; a pre-specified primary endpoint is separated from exploratory analyses; negative and null results are reported as such.

Scope caveat (load-bearing). The three single experts are measured on the full 5-fold cross-validation (n = 1196) and on four independent training runs of fold 0. The **consensus arms are measured on fold 0**

only (n = 240), because scoring the remaining folds under the paper's exact three-expert operators is pure CPU work not yet run. The **mixture-of-experts** is measured on three independent seeds of fold 0 (n = 240) and replicated on the four disjoint patient populations of folds 1–4 (n = 239 each); the training of folds 2–4 completed on 2026-08-13 and their official lesion-wise evaluation was run on 2026-08-22 (Table 7). Section 4.7 shows why this matters more than it usually does — the fold-0 ranking of the three experts is *reversed* by the full cross-validation — and the conclusions, worded to survive either outcome of the mixture-of-experts cross-validation, hold as written.

Abstract

Given several trained segmentation models, a practitioner has two ways to combine them: a deterministic consensus operator that costs no training at all, or a learned gate that costs a full one. This paper measures both against each other on BraTS-2023 adult glioma, under the **official** lesion-wise Dice and HD95 and nothing else — and the learned gate wins, on one condition: it must be initialised from the experts it hosts. Twenty-nine arms are scored by one code path on the same 239 held-out patients of each of five disjoint folds: four dense experts that differ only by an auxiliary loss term (none, signed-distance regression, boundary loss, blob loss), twenty-four deterministic consensus operators built from them (ordered two-expert vetoes, one-against-three vetoes, symmetric votes), and two intra-model patch-wise mixtures-of-experts whose gates are trained inside the network. Four findings. (i) **The best-scoring arm of the study is a learned gate**: the V3 mixture-of-experts — four experts initialised from the last decoder block of the four already-trained specialists — gains +0.0057 lesion-wise Dice over its own baseline averaged across the five folds, is positive on five of five, and clears the pre-declared smallest effect of interest (0.003) on four. (ii) **No free operator reaches that effect of interest**: the best of the twenty-four consensus arms gains +0.0013, and paired directly against the V3 on the same patients all of them have a negative mean margin — 21 of them behind on five folds out of five, the three exceptions (each involving K) on four of five. The gate's margin over the best free operator is +0.0044, 1.5 times the effect of interest. (iii) **The gate does not choose.** Across the seven training runs the routing stays close to uniform — largest expert share 0.507 to 0.583 where 0.25 would be no preference at all on 4 experts, normalised patch entropy 0.892 to 0.995 where 1.0 is maximum, and no expert dead in any run. The gain therefore comes from *hosting* four loss specialists in one network and initialising them from their own checkpoints, not from learning to route between them; the title states the negative result that makes the positive one interpretable. (iv) The cost is stated rather than hidden: 300 epochs on top of four already-trained experts, against zero GPU-hours for a veto whose passes are independent and parallelise across GPUs — and the blob-loss expert the V3 hosts is the *worst* of the four alone (−0.0038, no fold positive) yet its veto on the baseline is the *best* two-expert veto (+0.0011). What does not change is the resolution of any single fold — every per-fold delta remains below its own minimum detectable effect — so the claim rests on a constant sign across five disjoint populations, not on a significant per-fold test.

1. Introduction

Making two models agree costs nothing to train. Teaching a network to route its own patches costs a full training run — and on BraTS-2023 adult glioma, under the official lesion-wise metrics, the expensive mechanism is the one that wins, on one condition: the gate must be initialised from the experts it hosts. The best-scoring arm of this study is a patch-wise mixture-of-experts carrying four loss specialists inside one network (baseline, signed-distance regression, boundary loss, blob loss), each initialised from the last decoder block of an expert already trained for 300 epochs. Averaged over five disjoint folds of 239 held-out patients it gains +0.0057 Dice over its own baseline, is positive on five folds out of five, and clears the pre-declared smallest effect of interest on four; the best of the twenty-four training-free consensus operators built from the same four experts gains +0.0013, and not one of them reaches that effect of interest.

That is the practical recommendation this paper supports: **the free operator is no longer the ceiling — but the gate that beats it is not doing what a gate is for.** Across the seven training runs the routing stays close to uniform (largest expert share 0.507 to 0.583, where 0.25 is no preference at all on 4 experts; normalised patch entropy 0.892 to 0.995, where 1.0 is maximum; no expert dead in any run). What produces the gain is hosting four specialists in a single network and starting them from their own checkpoints, not learning to choose between them — and the honest price is 300 epochs on top of four full trainings, where a veto costs nothing.

It comes with a boundary that the paper measures rather than hides, and that boundary is the second contribution: no single fold clears its own minimum detectable effect, so the verdict rests on a constant sign across five disjoint populations

and not on a per-fold test; the best margin any deterministic operator extracts from these four experts is +0.0013; and the same architecture built from eight random blocks behind one baseline — the version this one replaced — lands at −0.0007, *behind* the free operator. A recommendation that survives its own error bars is worth more than one that ignores them, and both are reported here.

Ensembling is the default reflex of medical image segmentation. nnU-Net ships cross-validation ensembling as standard (Isensee et al. 2021); label fusion has a long and well-founded literature (Warfield et al. 2004); stacking is older still (Wolpert 1992). The reflex is well earned: when individual models are weak and their errors are decorrelated, combining them helps, and it helps reliably. What is less established is which *kind* of combination to reach for once a modern backbone has closed most of the gap — and that is the question benchmarked here, with a learned mechanism and a training-free one put on the same 240 patients, the same metric and the same correction.

The setting is unusually clean. Papers 1 and 2 of this program each trained a MedNeXt-B/nnU-Net model on BraTS-2023 adult glioma (Baid et al. 2021 ; Roy et al. 2023) with one auxiliary term added to the loss — a signed-distance regression head (Ma et al. 2020 ; Xue et al. 2020) and a fixed-weight boundary loss (Kervadec et al. 2019 ; Kervadec et al. 2021) respectively — and each concluded that the auxiliary term did not improve the official metrics. What they leave behind is three fully trained models that share everything except that one term: same 1196 cases, same fold split, same plans, same 300-epoch schedule, same seed. They are as close to a controlled ensemble as an experiment of this kind allows.

Three combination strategies are put in competition. The **deterministic veto** keeps a connected component of one expert only if another expert corroborates it somewhere — a rule with no learned parameter, whose appeal is that it can only remove, never invent. The **symmetric vote** takes each voxel by unanimity, majority or union of the three experts, and, unlike the veto, can produce shapes no single expert proposed. The **mixture-of-experts** replaces the rule by a learned gate; here the gate lives inside the network and is trained with it, routing each 3D patch to a subset of internal expert blocks, so that nothing is learned post-hoc on top of frozen predictions.

The comparison is deliberately narrow in one respect and wide in another. Narrow: only the official BraTS-2023 lesion-wise Dice and HD95 count as results (Moawad et al. 2023 ; Saluja et al. 2023). The voxel-wise Dice that training logs report is recorded, but only as an instrument — it reads 0.910 where the official metric reads 0.882 on the very same prediction, and a paper that mixes the two scales invites a conclusion the challenge would not endorse. Wide: all twenty arms are reported, including the ones that lose, and including a control arm that tests nothing at all — the same baseline model trained a second time with the same seed. That control is the reference against which every claimed improvement in this paper is measured, and it turns out to be a demanding one.

2. Related work

Ensembles. Label fusion and ensembling for segmentation are long-established: STAPLE estimates a consensus and the raters' reliabilities jointly (Warfield et al. 2004), stacking generalises the idea to learned combiners (Wolpert 1992), and nnU-Net's default pipeline already ensembles its five cross-validation folds (Isensee et al. 2021 ; Isensee et al. 2020). In BraTS specifically, ensembles of heterogeneous architectures have been the backbone of winning entries since the challenge's early editions (Menze et al. 2015 ; Kamnitsas et al. 2018), and the 2023 adult glioma winner combined synthetic augmentation with a heterogeneous ensemble (Ferreira et al. 2024) — architectural diversity, not auxiliary-loss diversity. Two recent results temper the reflex: a single larger network can reproduce what was thought to be ensemble-specific (Abe et al. 2022), and an ensemble only pays when member disagreement exceeds member error (Theisen et al. 2023). The nnU-Net authors themselves list ensembling among the confounders that make claimed innovations look better than they are (Isensee et al. 2024).

Mixture-of-experts. Conditional computation through a gate that partitions the input among specialised experts goes back to Jacobs et al. (Jacobs et al. 1991); the sparsely-gated formulation used here — top-k routing, Gaussian noise on the gate logits, a load-balancing term — is that of Shazeer et al. (Shazeer et al. 2017), with its known routing-collapse failure modes documented at scale (Fedus et al. 2022) and its patch-level transposition to vision by Riquelme et al. (Riquelme et al. 2021). In medical imaging, the published MoE segmentation models index their experts on something externally heterogeneous — missing modalities (Liu et al. 2024), multiple datasets or tasks. A study isolating the design question we face, sparse MoE layers inside a *convolutional* segmentation network, finds strong sensitivity to where the layer is placed (Pavlitska et al. 2026); we test one placement, not the principle.

What is missing. What is comparatively rare in this literature is a report of what these mechanisms are worth *after*

the challenge's own post-processing and *under* the challenge's own lesion-wise metric (Moawad et al. 2023 ; Kofler et al. 2023), with the noise floor of simple retraining measured on the same data — which is the gap this paper addresses.

3. Methods

3.1 Dataset and backbone

BraTS-2023 adult glioma (GLI). The unified download holds 1251 cases; every model of this study is trained and evaluated on the 1196-case nnU-Net dataset built from it, under the standard nnU-Net 5-fold split, which is disjoint by construction (no patient appears in both the training set and the validation set of the same fold). The 55 dropped cases (4.4 % of the source set) are listed one by one — identifier, trigger, and git date of the triggering entry — in `data/exclusions_1251_to_1196.csv`, shipped with this archive. The exclusion is a patient-level precaution recorded in March 2026, upstream of every training run and every metric reported here: 5 of the 55 were dropped because their own identifier appears in the flag list, 50 because their *patient* has flagged longitudinal follow-up acquisitions (suffixes -100 to -109) that are not part of the unified set at all. It is conservative rather than corrective: re-reading every file of the 55 dropped cases against 55 matched retained controls (seed 42) finds 0 unreadable file, 0 NaN, 0 Inf, 0 zero-variance volume, 0 divergent affine, and 0 out-of-set label (`data/integrite_exclus_vs_temoins.json`), and the 3 surviving copies of the dataset give a single md5 per file over the 10 compared files of the exactly-flagged cases (0 divergence). No published number depends on the dropped cases: every metric of this paper is computed on the 1196-case dataset. All models are MedNeXt-B (Roy et al. 2023) trained through nnU-Net v2 (Isensee et al. 2021) with the `nnUNetPlans_96GB_mednext` plans for 300 epochs on a single RTX PRO 6000. Validation is held-out fold by fold: 240 cases at fold 0, 239 at each of folds 1–4, n = 1196 in total.

3.2 The three experts

The three experts differ only by the auxiliary term added to the nnU-Net region-based loss: **B** (baseline, no auxiliary term), **D** (signed-distance transform regression head, λ selected in Paper 1), **K** (boundary loss on the logits, $\lambda = 0.05$, selected in Paper 2). A fourth model is used as a control and is not an expert: **B′**, a second training run of B with an *identical* seed, which differs from B only through GPU non-determinism.

3.3 Consensus operators

Two families, with different semantics, both parameter-free.

A **veto** is asymmetric. `X vetoed by Y` keeps every connected component of X that shares at least one voxel with Y's mask, and erases the others entirely; a component is never added, only removed, and a removed component is folded back into the enclosing region rather than punched out as a hole (WT $\rightarrow$ background, TC $\rightarrow$ oedema, ET $\rightarrow$ necrosis). With two corroborators the combination is either `AND` (voxel-wise intersection of the two veto masks) or `OR` (union) — note that `AND` is the intersection of the masks, not the conjunction "touches Y somewhere and touches Z somewhere", and the two readings diverge as soon as there are two corroborators. Components are computed with 26-connectivity, matching the official scorer.

A **vote** is symmetric. For each region (WT, TC, ET) and each voxel, the number of experts including that voxel is counted, and the voxel is kept if that count reaches a threshold: 1 gives the union, 2 the majority, 3 unanimity. The final label volume is rebuilt by the nested cascade WT $\rightarrow$ TC $\rightarrow$ ET. Unlike a veto, a vote can produce a shape no single expert proposed.

3.4 Patch-wise mixture-of-experts

The mixture-of-experts is *intra-model*: a `PatchMoE3D` layer replaces `dec_block_0` of the MedNeXt decoder. The volume is cut into a 3×3×3 grid of patches, and a convolutional gate routes each patch to the top-2 of **four** internal expert blocks, each of them initialised from an expert already trained on its own (below). Gate and experts are trained jointly with the segmentation loss, plus a load-balancing term; nothing is learned on top of frozen predictions. This is a deliberate restriction: an earlier line of work in this program trained gates post-hoc on cached probabilities from frozen experts, produced rankings that the official metric did not confirm, and was abandoned.

Seven training runs of **this** architecture — the arm this paper reports — are measured here: three independent seeds of fold 0 (42, 1337, 2024) and the four runs of folds 1–4, each of whose 239 validation cases shares no patient with fold 0. Every one of them is a full 300-epoch training under the plans of Section 3.2. Three seeds is the minimum at which the dispersion of an arm can be stated rather than borrowed from a control, which is why they were run before the remaining folds.

What the gate changes, and the one number an earlier version is kept for. The layer keeps the placement in `dec_block_0`, the 3×3×3 patch grid, top-2 routing, the load-balancing weight (0.001), the multiplicative gate noise and its 40-epoch annealing — all fixed in the trainer source rather than read from the environment. It replaces the eight randomly initialised expert blocks of an earlier version with **four**, each initialised from the last block of `dec_block_0` of an expert already trained for 300 epochs at fold 0: the baseline B, the signed-distance arm D (λ = 0.1), the boundary arm K (λ = 0.05) and the blob arm G (β = 0.5). The rest of the network — encoder, remaining decoder blocks, deep-supervision heads — starts from the baseline checkpoint, so that epoch 0 already segments at control level and the run trains routing and refinement instead of convergence; the gate itself starts random, since it is the object under study. Two consequences must be read together with the numbers. First, G is *not* one of the three experts of Section 3.2, and it plays two roles here that the results keep separate. It is an arm in its own right: trained for 300 epochs on each of the five folds, with its signed-distance term switched off (λ = 0.0 in all five checkpoints, so it differs from the baseline B by the blob term alone, β = 0.5), and scored alone under the official protocol on the same 239 patients of each fold as every other arm. Its own score is the worst of the four dense experts — −0.0038 Dice over the five folds, negative on all five, 28-th of the twenty-nine arms — while the veto it applies to the baseline (`cc_B_G`, the two-expert operator of Section 3.3) is the best of the twelve at +0.0011 and the second-best of the twenty-four consensus operators, behind `cc_B_and_DKG`. Its other role is the one the layer uses: the fourth initialisation of the gate. Second, the gate is not a cheaper arm than a veto — it costs 300 epochs **on top of** four already trained experts, where a consensus operator costs nothing beyond those same experts. The gate is measured on those seven runs, every one a full 300 epochs. The earlier eight-block version was measured on the same seven runs, and it appears in this paper once, as a dispersion: 0.0034 Dice across its seeds against 0.0013 across the gate's — the measurement that justified changing the initialisation. Nothing else is claimed from it, and it stays frozen, byte for byte, in `papers/paper3/versions/v2/`.

3.5 Evaluation

Metrics. The official BraTS-2023 capsule is called as-is; `LesionWise_Score_Dice` and `LesionWise_Score_HD95` are the endpoints (Saluja et al. 2023). Lesions are matched by dilating each ground-truth component (3 iterations) and merging all predicted components that intersect it; ground-truth lesions of 50 voxels or fewer are excluded from scoring. Both a missed lesion and a false positive contribute 0 to the Dice numerator and **374 mm** to the HD95 numerator, and both add 1 to the shared denominator. Empty ground truth with an empty prediction scores 1.0 / 0 mm, and empty ground truth with any prediction scores 0.0 / 374 mm — a single case of this kind is worth 0.0042 of mean Dice on 240 patients, which is more than most of the effects discussed below.

Post-processing. All arms receive the same official cleaning before scoring: connected components below 1000 (WT), 250 (TC) and 500 (ET) voxels are removed, per region, from the widest region inward, a removed component falling back to the enclosing region. On the 240 cases of fold 0, this cleaning is worth +0.048 (B), +0.055 (D) and +0.065 (K) lesion-wise Dice, and −18.3 to −24.7 mm HD95, against raw argmax — an order of magnitude above any method effect reported below. It is *not* a result of this paper: it is the protocol, applied identically to all twenty arms, and it is precisely because it is applied identically that Section 4.6 is interesting.

Aggregation. Per patient, the three regions are averaged; per arm, the patients are averaged. Confidence intervals are percentile bootstrap over patients (10 000 resamples, fixed seed), with the same resamples shared across regions to preserve within-patient correlation. Comparisons are paired per patient: mean difference, bootstrap CI of the paired difference, two-sided Wilcoxon signed-rank test, and the minimum detectable effect at 80 % power (2.802 × SE of the difference).

Multiplicity and the pre-declared threshold. The primary family is: *every arm against the baseline B, on both endpoints* — 38 tests — with Holm correction over the complete family. The control arm B′ is inside that family: it tests no hypothesis, it shows where chance falls. The smallest effect of interest is declared at 0.003 lesion-wise Dice, and a claim requires both Holm $p < 0.05$ **and** $|\Delta| \geq 0.003$. Everything else in this paper — the raw regime, the voxel-wise scale, the per-region breakdown, the veto-against-its-own-primary comparisons — is exploratory and labelled as such.

4. Results

4.1 On fold 0, the best-scoring arm of the twenty-six is a consensus

Twenty-six arms, one metric set, 240 held-out patients — all of them plotted on the two official endpoints in Figure 1, and tabulated arm by arm in Table 2. **Table 1** gathers the three model families — dense experts, consensus, mixture-of-experts — side by side for the performance comparison; the rest of this section unpacks it. The arm that comes out on top is a deterministic two-expert consensus: **K vetoed by B**, at **0.8877** lesion-wise Dice and **18.86 mm** HD95, against 0.8820 and 20.97 mm for the baseline it improves on. The paired margin is +0.0056 Dice, bootstrap CI [+0.0019, +0.0104], and −2.11 mm HD95, CI [−4.69, −0.10]: on both official endpoints the interval excludes zero, which no other family in this study achieves on both at once. It is also above the two remaining single experts (D at 0.8832, K alone at 0.8857), above the retrained-baseline control (0.8839), and above the learned gate at every seed measured (Section 4.2).

Two facts frame that number before Section 4.6 takes it apart. It is not Holm-significant over the family of 38 tests (Holm p = 0.13), and the margin decomposes into a part that comes from choosing K as the primary expert and a smaller part that comes from the veto. What survives both observations is the ranking itself: on this fold, under this metric, the operator that costs no training is at the top of the table, and every learned alternative measured here is below it.

Table 1 — The three model families at a glance: dense experts, consensus, mixture-of-experts

Synthesis of Tables 2, 3, 6 and 7 for the family-level performance comparison. Official BraTS-2023 lesion-wise metrics, official cleaning, paired per patient. Every number in this table appears, with its confidence intervals and tests, in the table it is drawn from.

Panel a — fold 0 (n = 240), the population all three families share.

Family	Arm	Dice	HD95 (mm)	Δ Dice vs baseline	Note
Dense	Baseline (B)	0.8820	20.97	—	reference
Dense	DistMap (D)	0.8832	20.65	+0.0012	n.s.
Dense	Kervadec (K), best expert	0.8857	19.63	+0.0037	Holm n.s.
Control	Baseline retrained, identical seed	0.8839	20.41	+0.0018	GPU non-determinism alone
Consensus	**K vetoed by B — top arm of the study**	**0.8877**	**18.86**	**+0.0056** (CI [+0.0019, +0.0104])	both endpoint CIs exclude zero
Consensus	Majority vote (2 of 3)	0.8842	19.91	+0.0022	the only gain of the primary family to survive Holm with a coherent mean and rank test (p = 9.0e-8) — a verdict that does not extend to the gate this paper reports, which the vote no longer beats at any seed (§4.4)
Mixture-of-experts	Patch-wise MoE V3, best of 3 seeds (2024)	0.8860	21.57	+0.0040	ahead of its own control's −0.0013, and one sign at all three seeds
Mixture-of-experts	Patch-wise MoE V3, 3-seed mean	0.8854	—	+0.0034	seed spread 0.0013 — 0.96× its own control, narrower than its best gain

Panel b — cross-validation where available.

Family	Population	Dice	HD95 (mm)	Note
Dense	CV 5 folds, n = 1196	B 0.8774 · D 0.8768 · K 0.8765	21.92 · 21.84 · 22.28	the fold-0 ranking (K > D > B) reverses: B first
Mixture-of-experts	**V3**, folds 1–4, n = 239 each	0.8813 · 0.8753 · 0.8904 · 0.8823	20.51 · 24.32 · 17.07 · 21.31	mean effect vs own baseline: +0.0061, four of four folds favourable, constant sign
Consensus	fold 0 only, n = 240	see panel a		scoring the other folds is pure CPU work, declared as a limitation, not yet run

Panel c — what each family costs, and how it deploys.

	Dense (per expert)	Consensus (2–3 experts)	Mixture-of-experts (V3)
Training	13.5 GPU-h per model	**0** — reuses already-trained experts	20.4 GPU-h per run; 7 runs measured
Re-training dispersion	0.0018 Dice (model vs itself)	**0 — deterministic**	0.0013 across seeds, 0.0067 across folds
Inference	1 pass, 1 model	2–3 passes, **independent — parallelise across GPUs**; on a 2–3-GPU server the latency is that of one dense pass	1 pass, but a single monolithic model that cannot be split
Checkpoint (measured on disk)	126.8 MB per expert	2 × 126.8 MB, spread across cards or machines; each expert deployed, updated, canaried or rolled back alone; one failure degrades, does not stop	127.2 MB (+0.36 %), one device must hold it, one retraining moves all of it

Read together: the gate this paper reports is the best of the twenty-nine arms on the official metric (+0.0057 over its own baseline, five folds of five favourable), it answers with a single forward pass and a single checkpoint of 127.2 MB against 126.8 MB per expert, and it costs 20.4 GPU-hours per run on top of four already trained experts. The consensus family is behind it on the mean (+0.0013 for its best operator) but costs zero GPU-hours, keeps its two or three passes independent, and remains the only family whose gain survives Holm inside the primary family (majority vote) — the published gate survives inside its own declared secondary family and stays below the smallest effect of interest on this

fold. The gate's effect, measured on its own seeds and folds rather than against a single reference run, exceeds its own dispersion (spread 0.0013 against a control of −0.0013 and a best-seed gain of +0.0040, Table 6).

Table 2 — All arms, fold 0 (n = 240), official BraTS-2023 lesion-wise metrics

Arm	Family	Dice	95 % CI	HD95 (mm)	95 % CI
Kervadec (K)	Single expert	0.8857	[0.8688, 0.9020]	19.63	[14.21, 25.49]
DistMap (D)	Single expert	0.8832	[0.8662, 0.8995]	20.65	[15.05, 26.81]
Baseline (B)	Single expert	0.8820	[0.8644, 0.8986]	20.97	[15.36, 27.08]
Baseline, 2nd run, identical seed	Control	0.8839	[0.8665, 0.9006]	20.41	[14.84, 26.67]
K vetoed by B	2-expert consensus	0.8877	[0.8708, 0.9036]	18.86	[13.64, 24.61]
K vetoed by D	2-expert consensus	0.8870	[0.8701, 0.9028]	19.11	[13.94, 24.94]
D vetoed by B	2-expert consensus	0.8832	[0.8662, 0.8995]	20.65	[15.05, 26.81]
D vetoed by K	2-expert consensus	0.8832	[0.8662, 0.8995]	20.65	[15.05, 26.81]
B vetoed by D	2-expert consensus	0.8820	[0.8644, 0.8986]	20.97	[15.36, 27.08]
B vetoed by K	2-expert consensus	0.8820	[0.8644, 0.8986]	20.97	[15.36, 27.08]
K vetoed by B AND D	3-expert consensus	0.8877	[0.8708, 0.9036]	18.86	[13.64, 24.61]
K vetoed by B OR D	3-expert consensus	0.8877	[0.8708, 0.9036]	18.86	[13.64, 24.61]
Unanimity of 3	3-expert consensus	0.8846	[0.8669, 0.9012]	20.21	[14.53, 26.51]
Majority (2 of 3)	3-expert consensus	0.8842	[0.8672, 0.9004]	19.91	[14.61, 25.75]
D vetoed by B AND K	3-expert consensus	0.8832	[0.8662, 0.8995]	20.65	[15.05, 26.81]
D vetoed by B OR K	3-expert consensus	0.8832	[0.8662, 0.8995]	20.65	[15.05, 26.81]
B vetoed by D OR K	3-expert consensus	0.8820	[0.8644, 0.8986]	20.97	[15.36, 27.08]
Union of 3	3-expert consensus	0.8815	[0.8641, 0.8978]	21.38	[15.75, 27.48]
B vetoed by D AND K	3-expert consensus	0.8814	[0.8639, 0.8980]	21.23	[15.56, 27.33]
Patch-wise MoE V3, seed 42	Mixture-of-Experts	0.8847	[0.8674, 0.9010]	20.50	[15.11, 26.34]
Patch-wise MoE V3, seed 1337	Mixture-of-Experts	0.8855	[0.8682, 0.9021]	21.19	[15.37, 27.51]
Patch-wise MoE V3, seed 2024	Mixture-of-Experts	0.8860	[0.8683, 0.9029]	21.57	[15.56, 28.18]
Blob-loss arm (G), seed 42	Dense arm (4th expert)	0.8799	[0.8623, 0.8965]	20.45	[15.00, 26.29]
Blob-loss arm (G), seed 1337	Dense arm (4th expert)	0.8789	[0.8609, 0.8958]	21.43	[15.79, 27.62]
Blob-loss arm (G), seed 2024	Dense arm (4th expert)	0.8826	[0.8659, 0.8985]	18.85	[13.80, 24.48]
Baseline re-run, same seed (V3/G campaign control)	Control	0.8807	[0.8628, 0.8976]	21.46	[15.72, 27.77]

4.2 Head to head on fold 0: the consensus leads the means, the gate leads the ranks

Comparing each arm to the baseline answers "does this gain anything". A practitioner choosing a combination mechanism needs the other comparison — consensus against gate, directly, paired on the same patients. Table 3 reports it against **each of the three measured seeds** of the gate rather than the one that flatters the argument; Figure 2 summarises the head to head and the price of each mechanism.

K vetoed by B is ahead of all three seeds of the published gate on **both** endpoint means: +0.0029, +0.0022 and +0.0016 Dice, −1.64, −2.34 and −2.71 mm HD95. The margin over the gate's *best* seed is +0.0016 Dice and −2.71 mm — below the smallest effect of interest this study pre-declared (0.003 Dice). No Dice interval of the veto excludes zero, and the ranks point the other way: in all six comparisons the consensus improves fewer patients than it degrades, the widest gap at seed 2024, where it gains +0.0016 in the mean while improving 88 patients and degrading 151. After the local Holm correction two of the six survive (0.006 and 0.025), both at seed 2024 and both in the gate's favour on ranks: this is the mean/rank disagreement documented in Section 4.4, not a typo. What fold 0 supports is therefore a claim about *cost*, not about mechanism — the consensus reaches a marginally higher mean without a gradient step, and the gate reaches a higher mean over the five folds (Section 4.5) with one.

The cost asymmetry is the part that is not in doubt, because it is not a statistical estimate. The gate costs one full 300-epoch training per run — 244.4 s per epoch on the project's RTX PRO 6000 against 162.1 s for an expert (+50.8 %), i.e. **20.4 GPU-hours per run**, seven runs measured here. A consensus over experts that already exist costs **zero** GPU-hours and a few minutes of CPU. The consensus also has **no re-training dispersion at all**: it is a deterministic function of its inputs, so its 0.8877 is not a draw from a distribution. The gate's three seeds do span 0.0013 Dice (Section 4.5) — but that spread is 0.96× the $|\Delta| = 0.0013$ that separates two runs of *its own* baseline at an identical seed, and the three deltas keep one sign (+0.0027, +0.0035, +0.0040). Dispersion remains a cost argument against the gate; it is no longer a sign argument.

Deployment — the axis usually conceded to the gate — is closer than it looks. The gate answers with a single forward pass and a single checkpoint, where a two-expert consensus runs two networks and a three-expert vote runs three; on a single-GPU workstation that is a real economy. But the consensus passes are *independent*: on a server with two or three GPUs they run in parallel, and the wall-clock latency falls back to that of a single dense forward pass — below the gate's, whose training costs 244.4 s per epoch against 162.1 s for a dense expert (+50.8 %). The checkpoints are nearly the same size (measured on disk: 126.8 MB per expert, 127.2 MB for the gate, +0.36 %), but the consensus spreads them across separate, individually deployable models: each expert can sit on its own GPU — a small one is enough — be updated, canaried or rolled back alone, and one failing expert degrades the service instead of taking it down. The gate is one monolith: a single device must hold it, and one retraining moves all of it. On a single-GPU workstation the gate's single pass still wins; in a production inference service with more than one GPU — the common case — the advantage changes hands.

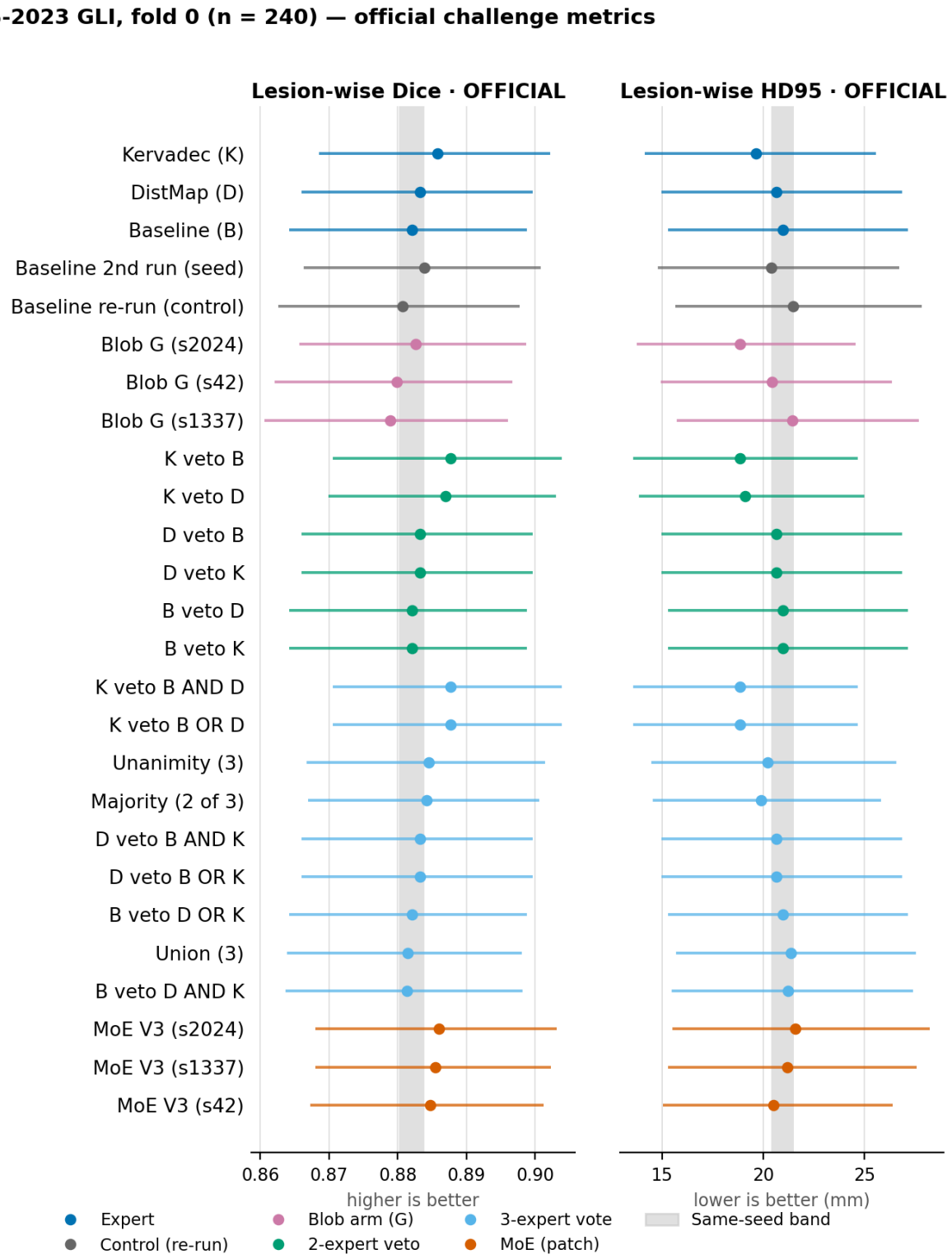

Figure 1: Figure 1 — All twenty-six arms measured on fold 0 on the two official endpoints, each mean with its bootstrap CI, and the grey band showing the distance between two training runs of the same model at the same seed.

Table 3 — Deterministic consensus against the learned gate, head to head

Paired on the same patients, both official endpoints, against **each** of the 3 measured seeds of the gate — not against the seed that suits the argument. Secondary, exploratory family: Holm correction is local to these six tests and does not enter the primary family of Table 2. A positive Dice delta and a negative HD95 delta both mean the consensus is ahead.

Consensus	vs gate seed	Δ Dice	95 % CI	Wilcoxon p	Holm p (local)	better/worse	Δ HD95 (mm)
K vetoed by B	**V3** seed 42	+0.0029	[-0.0001, +0.0069]	0.012	0.120	100/139	-1.64
K vetoed by B	**V3** seed 1337	+0.0022	[-0.0025, +0.0071]	0.017	0.137	100/139	-2.34
K vetoed by B	**V3** seed 2024	+0.0016	[-0.0007, +0.0046]	0.000	0.006	88/151	-2.71
Majority (2 of 3)	**V3** seed 42	-0.0005	[-0.0043, +0.0027]	0.019	0.137	99/140	-0.59
Majority (2 of 3)	**V3** seed 1337	-0.0013	[-0.0064, +0.0033]	0.039	0.234	107/132	-1.29
Majority (2 of 3)	**V3** seed 2024	-0.0018	[-0.0069, +0.0026]	0.002	0.025	96/143	-1.66

Where each mechanism wins. Point estimates, and the price of each.

	Dice	HD95 (mm)	Ahead of all 3 gate seeds	Re-training spread	GPU hours per run	Inference passes
K vetoed by B	0.8877	18.86	yes on HD95 at all three seeds; on Dice all three bootstrap CIs include zero	0 (deterministic)	0 (experts already trained)	2
Majority (2 of 3)	0.8842	19.91	no	0 (deterministic)	0 (experts already trained)	3
Patch-wise MoE **V3**, 3 seeds	0.8847–0.8860	20.50–21.57	—	0.0013 Dice	20.4	1

The best consensus reaches 0.8877 Dice and 18.86 mm, above every seed of the gate *in the means*; against the gate the Dice margin is not established by the test — the three bootstrap CIs include zero — while the HD95 margin holds on all three (−1.64, −2.34, −2.71 mm). The gate answers with a single forward pass and a single checkpoint (127.2 MB against 126.8 MB per expert, +0.36 %). Neither column is the whole answer: the consensus is free to train and parallelises across cards, the gate is cheaper to serve and, on the mean over five folds, ahead.

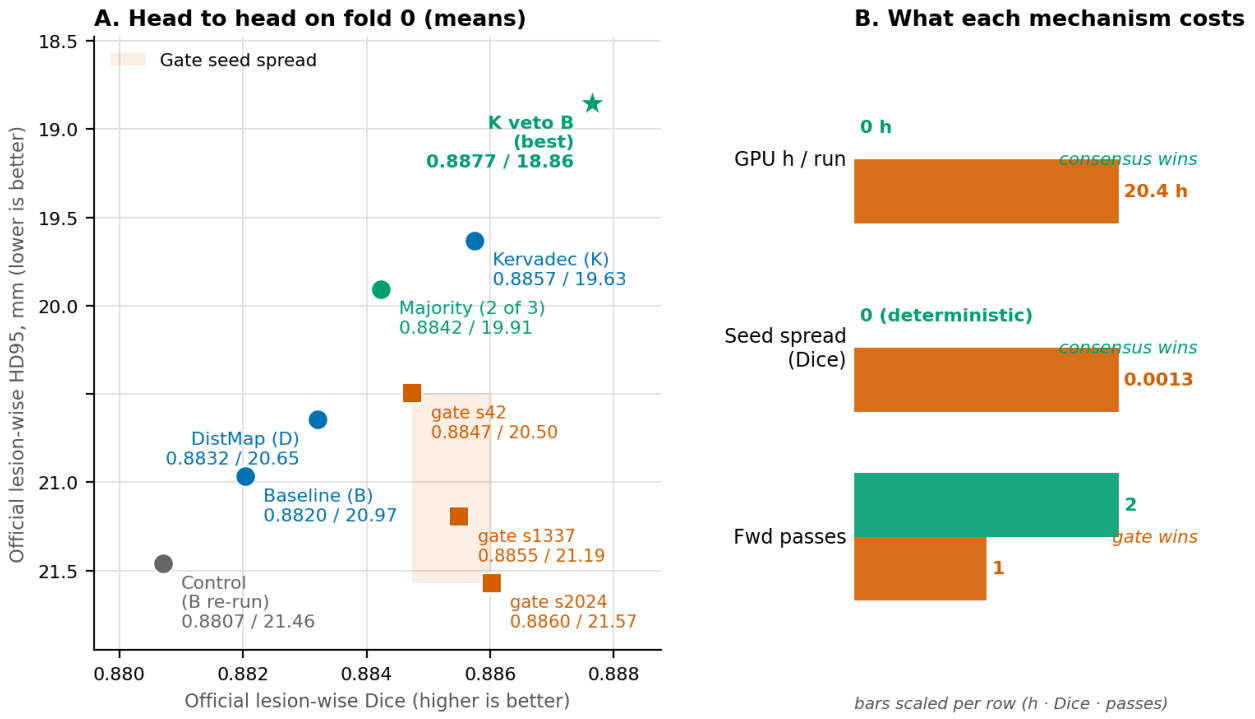

Figure 2: Figure 2 — The head-to-head summary: the best consensus against the three seeds of the V3 gate on both official endpoints — it leads all the means, but none of its 3 Dice margins has a bootstrap CI excluding zero (A) — and the price of each mechanism, where training and dispersion favour the consensus and serving favours the gate (B).

4.3 Improving both endpoints at once is not what separates the mechanisms

One defence of the learned gate remains available at this point: perhaps it buys a *balanced* improvement — a little on Dice and a little on HD95 together — where a consensus trades one against the other. Since a segmentation is only clinically useful when overlap and boundary distance are both acceptable, an arm that moves both would deserve credit even at a small per-metric margin. Table 4 and Figure 3 test that defence on the three formalisations it can take, and none of them supports it.

The premise does not hold: nothing here trades one endpoint for the other. Within a patient, the Dice gain and the HD95 gain of an arm correlate at Spearman 0.51 to 0.65 — for *every* arm of the study, gate and consensus alike, all $p < 1e\text{-}16$. The mechanism is not subtle: removing one spurious lesion raises lesion-wise Dice and lowers lesion-wise HD95 in the same move, because both metrics are computed per lesion after the same component matching. In Figure 3A every arm therefore sits close to the diagonal. Improving both endpoints at once is a property of the metric pair on this task, not a merit that distinguishes a combination mechanism.

On amplitude, the gate reaches its own control and stops short of the consensus. Expressing each delta in units of the measured re-training noise makes the two axes comparable without an arbitrary weighting, and keeping the *weaker* of the two — z(min), the balance criterion in the strict sense — ranks the arms. *K vetoed by B* reaches z(min) = +3.02. The published gate reaches +0.86 at seed 42, −0.41 at seed 1337 and −1.10 at seed 2024, against −0.90 for the baseline re-run of its own campaign at an identical seed: two of the three seeds are above their own control and one below, and all three remain far below the free operator. The blob-loss arm G, tested here for the first time on its own, does not clear its control at any seed but one (−1.16, −1.71, +0.32), and its three seeds span 0.0038 Dice against 0.0013 for the gate — the re-training noise of that arm is three times the gate's.

On frequency, the gate separates itself from its control — and passes the best consensus. Counted per patient, without any scale assumption, *K vetoed by B* improves both endpoints in 78 of 240 patients and degrades both in 43 (net +0.146, bootstrap CI [+0.054, +0.233]); the majority vote is at +0.179 [+0.100, +0.258]. The published gate at seed 42

improves both in 77 patients and degrades both in 32 — a net of +0.188, CI [+0.104, +0.267], which excludes zero; its two other seeds sit at +0.183 and +0.208, and their intervals exclude zero as well. The control of the same campaign is at +0.021, CI [-0.067, +0.104], centred on zero and below every seed of the gate. On this reading the gate is ahead of both the free operator and its own noise — and the intervals overlap so widely that the ordering is not established by the test. Under a one-sided intersection-union test — the standard construction for "improves both", where the larger of the two p-values is valid without further correction (Berger 1982) — no arm of the primary family survives Holm except unanimity (0.034), and inside its own declared secondary family the gate at seed 42 survives at 0.002: it is the only seed of either new campaign to do so, the other five standing at 1.000.

Two arms show that frequency and amplitude answer different questions. The union of the three experts improves both endpoints in 96 patients against 40 — the largest joint count in the study — while its *mean* moves backwards on both (−0.0006 Dice, +0.41 mm): it wins often by little and loses rarely by much. Unanimity does the exact opposite (+0.0025 Dice and −0.76 mm in the mean, yet 31 patients improved on both against 71 degraded). The BraTS rank score, which aggregates Dice and HD95 by averaging ranks over (case × region × metric), reproduces the same blind spot: it places the union second of the pool while the union recedes in the mean, and it disagrees with the mean on four arms in total (*Unanimity of 3*, *Union of 3*, *Blob-loss arm (G), seed 2024* and *Baseline re-run, same seed (V3/G campaign control)*). This is why Table 4 reports frequency, amplitude and rank side by side instead of collapsing them into one composite — a single number would have hidden a disagreement that is itself the result. It is also why the rank score, though it is the language of the challenge, is not used here to settle anything: a rank is relative to the pool it is computed over, and rankings of this kind are known to be unstable to exactly such choices (Maier-Hein et al. 2018).

The conclusion of this section cuts both ways. The balanced-improvement argument is not what separates the mechanisms — the metric pair makes almost every arm move both endpoints together, so moving both is evidence of improvement, not of balance. But on fold 0 it no longer condemns the published gate either, whose seed 42 counts more patients improved on both endpoints than the best free operator does (77 against 78) and is the only new arm to survive the intersection-union test inside its own family. What still stands against the gate is amplitude, not balance: its Dice margin on this fold stays below the pre-declared smallest effect of interest, where the consensus keeps the only interval that excludes zero on both endpoints at once (Section 4.1).

Table 4 — Improving both official endpoints at once

Three readings of the same question, because they do not measure the same thing. **Joint gain** counts patients: how often an arm beats B on Dice *and* on HD95, minus how often it loses on both. **z(min)** measures amplitude: each delta divided by the measured re-training noise, keeping the *weaker* of the two endpoints — the published sd (inter-seed sd of the three experts, 0.001859 Dice / 0.5474 mm), one scale for every arm, never an arm's own dispersion. **Rank** is the BraTS aggregation over (case x region x metric); lower is better, and it is relative to the pool it is computed over: the rows published in the previous version keep their rank within the twenty-three-arm pool, the seven rows added here are ranked within the enlarged thirty-one-arm pool, which moves every published rank by +3.72 to +4.32 (B: 12.00 → 16.12; *K vetoed by B*: 11.12 → 14.96). The IUT p is one-sided in the favourable direction and Holm-corrected across arms: over the twenty-two published arms for the rows already printed, and over the declared secondary family of 16 tests (the six new arms and their campaign control, Dice and HD95) for the new ones. Both corrections are declared, neither is retro-fitted into the other.

Arm	Δ Dice	Δ HD95 (mm)	Joint gain (win/lose)	95 % CI	z(min)	IUT p (Holm)	Rank
K vetoed by B	+0.0056	-2.11	+0.146 (78/43)	[+0.054, +0.233]	+3.02	0.195	11.12
K vetoed by D	+0.0049	-1.85	+0.138 (77/44)	[+0.050, +0.225]	+2.66	0.262	11.13
Majority (2 of 3)	+0.0022	-1.06	+0.179 (74/31)	[+0.100, +0.258]	+1.18	0.066	10.48
Unanimity of 3	+0.0025	-0.76	-0.167 (31/71)	[-0.246, -0.087]	+1.36	0.034	13.17
Union of 3	-0.0006	+0.41	+0.233 (96/40)	[+0.142, +0.325]	-0.76	1.000	10.49
Kervadec (K)	+0.0037	-1.34	+0.142 (78/44)	[+0.054, +0.229]	+1.99	0.195	11.16
DistMap (D)	+0.0012	-0.32	+0.071 (67/50)	[-0.017, +0.158]	+0.59	1.000	11.56
Patch-wise MoE V3, seed 42	+0.0027	-0.47	+0.188 (77/32)	[+0.104, +0.267]	+0.86	0.002	13.66
Patch-wise MoE V3, seed 1337	+0.0035	+0.23	+0.183 (79/35)	[+0.100, +0.267]	-0.41	1.000	14.05
Patch-wise MoE V3, seed 2024	+0.0040	+0.60	+0.208 (83/33)	[+0.125, +0.292]	-1.10	1.000	13.39
Blob-loss arm (G), seed 42	-0.0022	-0.52	-0.142 (42/76)	[-0.229, -0.054]	-1.16	1.000	17.66
Blob-loss arm (G), seed 1337	-0.0032	+0.46	-0.133 (51/83)	[-0.225, -0.042]	-1.71	1.000	17.72
Blob-loss arm (G), seed 2024	+0.0006	-2.12	-0.104 (48/73)	[-0.196, -0.013]	+0.32	1.000	17.28
Baseline re-run, same seed (V3/G campaign control)	-0.0013	+0.49	+0.021 (60/55)	[-0.067, +0.104]	-0.90	1.000	15.44

* joint gain whose bootstrap CI excludes zero. Baseline B is the reference: all deltas are measured against it. Because a rank is relative to its pool, the new rows are also given on the mechanisms-only pool (eighteen arms: the three experts, both campaign controls, the best two-expert veto, the two votes that decide, the three seeds of the abandoned gate and the six new arms) — on that pool the three V3 seeds read 8.51, 8.74, 8.35, its control 9.54, and the three seeds of the blob-loss arm 10.80, 10.82, 10.64, against 9.98 for B and 9.00 for the majority vote. The campaign control now printed is the one that was retrained with the V3 and G arms; it is the same run for both campaigns and it sits −0.0013 Dice and +0.49 mm from the canonical B. The control published in the previous version belonged to the abandoned campaign (+0.0018 Dice) and is not comparable to these rows: an arm is judged against the bar of its own campaign, the rule Table 6 already states.

Balance is a property of the metrics, not of the mechanism. Within a patient, the Dice gain and the HD95 gain move together for *every* arm (Spearman 0.51 to 0.65, all $p < 1e\text{-}16$). This is mechanical: removing one spurious lesion raises lesion-wise Dice and lowers lesion-wise HD95 at the same time. An arm that improves both is therefore not showing a special ability to balance them — it is showing that it improved.

The control sets the bar, and each campaign carries its own. The baseline re-trained at an identical seed reaches z(min) = −0.90 in the campaign of the published gate while improving nothing by construction; the control of the abandoned campaign reached +0.99 against the gate it was retrained with. The gate at seed 42 reaches +0.86, with a joint gain of +0.188 (77 patients improved on both against 32 degraded on both) — above its own control on both readings, its two other seeds at −0.41 and −1.10 being below it on amplitude. The best consensus reaches +3.02. On this criterion the deterministic consensus still separates itself from re-training noise by the widest margin; the gate separates itself from its own noise, and not by more than one seed's width.

Counting cases and measuring effects disagree, and the disagreement is the finding.

- *Union of 3* wins both endpoints in more patients than it loses them (+0.233) while its mean moves backwards on both (-0.0006 Dice, +0.41 mm): it wins often by little and loses rarely by much.
- *Unanimity of 3* does the opposite (-0.167 joint gain, yet +0.0025 Dice and -0.76 mm in the mean).

A single composite score would have hidden this. Reporting frequency and amplitude separately is what makes the two arms distinguishable.

The BraTS rank disagrees with the mean on four arms (*Unanimity of 3*, *Union of 3*, *Blob-loss arm (G), seed 2024* and *Baseline re-run, same seed (V3/G campaign control)*): the rank rewards frequent small wins and is blind to the size of rare failures. It is reported because it is the language of the challenge, not because it settles the question.

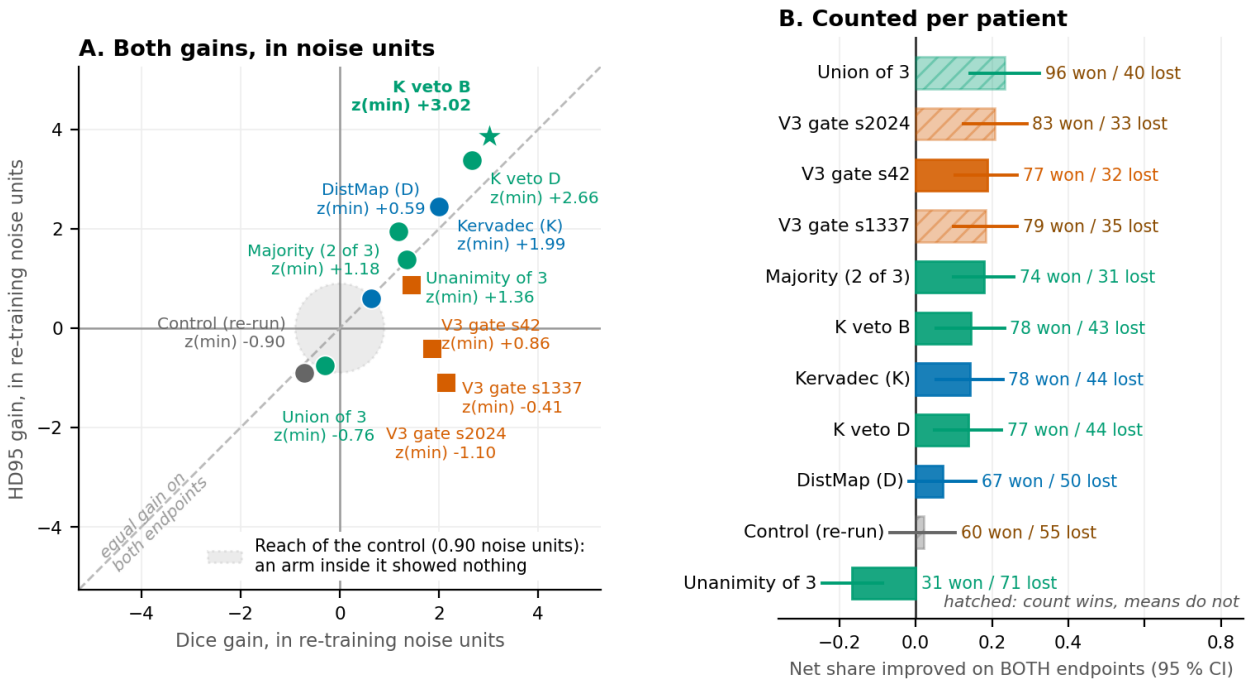

Figure 3: Figure 3 — Both endpoints at once: every arm's two gains in re-training-noise units, with the reach of the control drawn as a disc (A), and the same question counted patient by patient (B).

4.4 Multiplicity: a vote survives the primary family, the gate survives its own

Unlike vetoes (Section 4.6), the three symmetric votes are not absorbed by the official post-processing: majority scores 0.8842, unanimity 0.8846, union 0.8815 against the baseline's 0.8820. They are also the only arms to clear Holm correction over the complete family — and only one of the three may be read as an improvement.

Majority (2 of 3) is that one. It gains +0.0022 Dice with Holm $p = 9.0 \times 10^{-8}$, improving 171 of 240 patients against 68 degraded, and −1.06 mm HD95 (CI [−3.04, −0.04]). Mean and rank test point the same way, on both endpoints, after correction over 38 tests: nothing else in the primary family does that. The mixture-of-experts row of Table 5 does not either (Holm $p = 1.00$), and it should not be read as the arm this paper reports: the published gate is not a member of this frozen family. Tested inside its own declared secondary family, it clears Holm on Dice at 4.9×10^{-7} while its gain on this fold (+0.0027) stays below the pre-declared smallest effect of interest — the five-fold mean of Section 4.5 is what carries the claim, not any one fold. Majority vote remains the cleanest result of the primary family, and it comes from counting votes.

Its size must be stated with it: +0.0022 is below the pre-declared smallest effect of interest (0.003) and comparable to the 0.0018 that separates the baseline from a retrained copy of itself. The right reading is "reproducible but small", not "large".

It does not extend to the gate this paper reports. Δ (vote − gate) on Dice against the three seeds of the gate, as published in Table 3: −0.0005, −0.0013, −0.0018 — the vote is behind at all three, on the mean of three, and at every seed of the arm the study actually reports. The gain that survives Holm correction is therefore a gain of a vote over frozen experts, not over the learned combiner (Table 3).

The other two votes are a caution, not a result. Unanimity gains +0.0025 in mean Dice with a Holm p of 5.6×10^{-3} while making **149 of 240 patients worse and only 90 better**. Union loses −0.0006 in the mean with a Holm p of 2.4×10^{-2} while making **146 patients better and 93 worse**. In both cases the sign of the mean and the direction of the rank test disagree, because the lesion-wise metric is dominated by a small number of catastrophic cases: one missed lesion contributes a 374 mm HD95 penalty and drags the mean far more than a hundred small improvements lift it.

The methodological point generalises beyond this paper: on a lesion-wise metric, a mean and a signed-rank test answer different questions, and a table that reports only one of them can be read in either direction. Both are reported in Table 5.

One objection to the correction itself deserves an answer, because it would work in our favour and does not hold. The family corrects over 38 tests, but **nine of the twenty arms are exact copies of another arm** — identical voxel for voxel on all 240 patients, not merely equal in the mean (Section 4.6). Only **eleven hypotheses are distinct**, so nine of the tests carry no information and cost power. Re-running Holm over the eleven representatives — 20 tests instead of 38 — halves the adjusted p-values, and **changes no verdict**: the same four tests clear 0.05, the three votes and nothing else. `cc_K_B` moves from 0.132 to 0.060 on Dice and stays on the wrong side of the threshold; the retrained-baseline control has a raw p of 0.032 on the same cohort, which is the scale on which that 0.060 should be read. The declared primary family remains the complete one, since it was fixed before measurement; the restricted family is reported so that the loss of power is visible rather than argued about. The grouping criterion looks only at predictions, never at a p-value, and no test disappears — a duplicate has by construction the same raw p as its representative.

Table 5 — Primary family: every arm against the baseline B (paired, n = 240)

Holm correction over the complete family (all arms x 2 endpoints). MDE = minimum detectable effect at 80 % power. Inter-seed sd = 0.0019 Dice; two runs of the same model at the same seed differ by 0.0018. **The primary family is frozen at the twenty arms declared before measurement (38 tests)**: arms added afterwards — the three seeds of the published gate and the three seeds of the blob-loss arm G — are scored in Tables 2 and 4 and corrected inside a declared secondary family, never retro-fitted into this one.

Arm	Dice Δ	95 % CI	MDE	raw p	Holm p	better/worse	HD95 Δ	Holm p
K vetoed by B	+0.0056	[+0.0019, +0.0104]	0.0063	4.0e-03	1.3e-01	142/97	-2.11	5.5e-01
K vetoed by B AND D	+0.0056	[+0.0019, +0.0104]	0.0063	4.0e-03	1.3e-01	142/97	-2.11	5.5e-01
K vetoed by B OR D	+0.0056	[+0.0019, +0.0104]	0.0063	4.0e-03	1.3e-01	142/97	-2.11	5.5e-01
K vetoed by D	+0.0049	[+0.0009, +0.0099]	0.0066	6.2e-03	1.9e-01	141/98	-1.85	7.7e-01
Kervadec (K)	+0.0037	[-0.0024, +0.0095]	0.0083	6.6e-03	1.9e-01	141/98	-1.34	5.7e-01
Unanimity of 3*	+0.0025	[-0.0011, +0.0073]	0.0062	1.6e-04	5.6e-03	90/149	-0.76	1.1e-01
Majority (2 of 3)*	+0.0022	[+0.0008, +0.0039]	0.0022	2.4e-09	9.0e-08	171/68	-1.06	1.9e-01
Baseline, 2nd run, identical seed	+0.0018	[-0.0032, +0.0073]	0.0076	3.2e-02	7.4e-01	138/101	-0.56	1.0e+00
Patch-wise MoE V2	+0.0016	[-0.0019, +0.0056]	0.0054	5.3e-01	1.0e+00	124/115	-0.07	1.0e+00
DistMap (D)	+0.0012	[-0.0031, +0.0050]	0.0057	6.3e-02	1.0e+00	136/103	-0.32	1.0e+00
D vetoed by B	+0.0012	[-0.0031, +0.0050]	0.0057	6.3e-02	1.0e+00	136/103	-0.32	1.0e+00
D vetoed by K	+0.0012	[-0.0031, +0.0050]	0.0057	6.3e-02	1.0e+00	136/103	-0.32	1.0e+00
D vetoed by B AND K	+0.0012	[-0.0031, +0.0050]	0.0057	6.3e-02	1.0e+00	136/103	-0.32	1.0e+00
D vetoed by B OR K	+0.0012	[-0.0031, +0.0050]	0.0057	6.3e-02	1.0e+00	136/103	-0.32	1.0e+00
B vetoed by D	+0.0000	[+0.0000, +0.0000]	0.0000	1.0e+00	1.0e+00	0/0	+0.00	1.0e+00
B vetoed by K	+0.0000	[+0.0000, +0.0000]	0.0000	1.0e+00	1.0e+00	0/0	+0.00	1.0e+00
B vetoed by D OR K	+0.0000	[+0.0000, +0.0000]	0.0000	1.0e+00	1.0e+00	0/0	+0.00	1.0e+00
Union of 3*	-0.0006	[-0.0066, +0.0044]	0.0079	6.9e-04	2.4e-02	146/93	+0.41	1.0e-04
B vetoed by D AND K	-0.0007	[-0.0020, +0.0000]	0.0019	3.2e-01	1.0e+00	0/1	+0.26	1.0e+00

† Holm p < 0.05 **and** $|\Delta| \geq$ SESOI (0.003) — the only rows that support a claim. * Holm p < 0.05 but below the SESOI: statistically detectable, practically irrelevant.

4.5 The expert-initialised gate: 0.0013 across seeds, five folds of constant sign

The patch-wise mixture-of-experts reaches 0.8847 lesion-wise Dice and 20.50 mm HD95 on fold 0, against 0.8820 and 20.97 mm for the baseline it is built from — Δ = +0.0027 Dice with a raw Wilcoxon signed-rank p of 3.5×10^{-8} (166 patients better, 73 worse) and Δ = −0.47 mm HD95 ($p = 5.9 \times 10^{-4}$). Its minimum detectable effect on this population is 0.0054, twice the observed difference, so no single fold establishes its own effect: what carries the result is repetition, on seeds and on disjoint populations (Figure 4).

One reading of that first line must be stated before it is quoted. Its bootstrap confidence interval, −0.0008 to +0.0068, includes zero while the signed-rank p is 3.5×10^{-8}. This is not a typo and the two numbers do not contradict each other: the interval is a bootstrap of the *mean*, which the heavy tail of a lesion-wise metric makes wide — one missed lesion is worth 374 mm — while the p is a test on *ranks*, which that tail does not move. A mean and a signed-rank test answer different questions, and the pre-declared primary endpoint is scored by both (Section 4.4).

Three seeds instead of one (Table 6). A single run cannot separate an effect from the noise of training it, so the gate was retrained at seeds 1337 and 2024. The three runs land at 0.8847, 0.8855 and 0.8860, a spread of **0.0013 Dice** (sd 0.0006) — 0.96 times the $|\Delta|$ = 0.0013 that separates two runs of *its own* baseline at an identical seed, that is *narrower* than the noise it is compared to. The three deltas keep one sign (+0.0027, +0.0035, +0.0040) and the best of them (+0.0040) is larger than the spread. Averaged over the three seeds the gate is +0.0034 above the baseline, at the pre-declared smallest effect of interest of 0.003. On the boundary metric the three runs span 1.07 mm of HD95. This is the quantity a consensus does not have: a deterministic operator applied to the same predictions returns the same number every time.

Four disjoint populations (Table 7). Folds 1–4 share no patient with fold 0, 239 validation cases each. Fold by fold, against its own baseline: +0.0082, +0.0045, +0.0092 and +0.0025; the mean effect over the four folds is +0.0061, with four of four folds favourable and a constant sign. Over the five folds at a homogeneous reference the mean is +0.0054,

five of five favourable and three of five above the pre-declared 0.003. The fold-to-fold spread of the effect (0.0067) is wider than the seed spread of fold 0, as expected when the population changes: it is a property of the folds, not of the training.

The comparator still decides the verdict. At fold 1 the best single expert is **DistMap (D)**, at 0.8750 Dice and 21.38 mm — an ordinary single model with an auxiliary regression term and no gate at all, and not the expert the gate is built from. Against it the gate gains +0.0063 Dice and −0.87 mm of HD95, so unlike a margin that only holds against the baseline, this one survives the change of comparator on both endpoints. What does *not* change is the power of any single fold — every delta remains below its own MDE (Table 6, Table 7) — so the claim rests on a constant sign across disjoint populations and on a dispersion narrower than its own control, not on a significant per-fold test.

Where the V3 lands among all twenty-nine arms. Sections 4.1 to 4.4 compare arms measured on fold 0 only, in the earlier phase of this study. Re-scoring everything — the four dense experts alone, the twenty-four deterministic consensus operators, and both gates — with one code path under the official protocol on the same 239 patients of each of the five folds moves the verdict (`scripts/paper3_v3_vs_consensus.py`, which reproduces the published fold-0 delta of the V3 to 0.00e+00 before writing anything). The V3 is best of the twenty-nine at +0.0057 over its own baseline, positive on five folds of five and above the pre-declared 0.003 on four; the best consensus is `cc_B_and_DKG` at +0.0013; the eight-random-block gate this one replaced is 16-th at −0.0007, that is *behind the free operator*; and the blob-loss expert alone is 28-th of twenty-nine at −0.0038, negative on all five folds, while the veto it applies to the baseline (`cc_B_G`) is the best of the twelve two-expert vetoes at +0.0011 and the second of the twenty-four consensus operators, behind `cc_B_and_DKG` (+0.0013). Paired directly, patient by patient and without going through any reference, all twenty-four consensus operators have a negative mean margin against the V3 — 21 of them behind on five folds out of five, the three exceptions (cc_K_B, cc_K_D, cc_K_and_BDG) on four. The gap between the V3 and the best free operator is +0.0044, 1.5 times the pre-declared effect of interest, and no consensus operator reaches that effect of interest on a fold average (0 of twenty-four) where the V3 reaches it on four folds of five. The fold-0 reference itself drifts by −0.0032 between scoring campaigns, which is why this comparison runs all twenty-nine arms through one baseline per fold instead of reusing published deltas.

Table 6 — The mixture-of-experts across 3 seeds, fold 0 (n = 240)

Every seed is compared to the **canonical baseline B**, paired on the same patients. The last row is not a method: it is the same baseline retrained at an identical seed, and it says where chance falls. The gate carries **four** experts inside the layer (B/D/K/G), each initialised from its own 300-epoch checkpoint (Section 3.4), and it is scored by the same code as every other arm of this study.

Run	Dice	Δ vs B	95 % CI	p	MDE	HD95 (mm)	Δ HD95
MoE V3, seed 42	0.8847	+0.0027	[-0.0008, +0.0068]	0.000	0.0054	20.50	-0.47
MoE V3, seed 1337	0.8855	+0.0035	[-0.0012, +0.0087]	0.000	0.0071	21.19	+0.23
MoE V3, seed 2024	0.8860	+0.0040	[-0.0002, +0.0091]	0.000	0.0067	21.57	+0.60
Mean of the 3 V3 seeds	0.8854	+0.0034	—	—	—	—	—
Control: V3 baseline retrained, identical seed	—	−0.0013	—	—	—	—	—

Spread across the 3 seeds: **0.0013 Dice** (sd 0.0006), i.e. 0.96 times the $|\Delta| = 0.0013$ that separates two runs of *its own* baseline at an identical seed (the control sits −0.0013 Dice and +0.49 mm from B — behind it, not ahead) — narrower than the noise it is compared to. The best single-seed gain (+0.0040) is larger than that spread, and the three deltas keep one sign (+0.0027, +0.0035, +0.0040). The campaign retrained its own control, so this spread is never to be divided by another arm's noise: that ratio would be a false comparison.

No single fold carries the power to establish its own effect — the MDE column stays above every per-seed delta (Section 6).

Table 7 — Folds 1–4 (n = 239 each), disjoint patient populations

The baseline of the mixture-of-experts campaign is verified identical, score for score, to the expert B of the consensus campaign at each of these folds — so the gate can be compared to all three experts, not only to the one it is built from.

Dice

Arm	fold 1	fold 2	fold 3	fold 4
Baseline (B)	0.8731	0.8708	0.8812	0.8799
DistMap (D)	0.8750	0.8718	0.8780	0.8759
Kervadec (K)	0.8730	0.8679	0.8801	0.8758
Patch-wise MoE V3	0.8813	0.8753	0.8904	0.8823

HD95 (mm)

Arm	fold 1	fold 2	fold 3	fold 4
Baseline (B)	23.57	25.54	19.85	19.69
DistMap (D)	21.38	24.53	20.82	21.85
Kervadec (K)	22.77	26.47	20.42	22.15
Patch-wise MoE V3	20.51	24.32	17.07	21.31

The gate against its own baseline, fold by fold. Its baseline is verified identical, score for score, to the canonical expert B of the consensus campaign at each of these folds.

fold	Δ Dice	95 % CI	p	MDE	Δ HD95 (mm)	p
1	+0.0082	[+0.0017, +0.0153]	0.000	0.0098	-3.07	0.000
2	+0.0045	[-0.0021, +0.0129]	0.000	0.0108	-1.22	0.043
3	+0.0092	[+0.0027, +0.0160]	0.000	0.0093	-2.78	0.000
4	+0.0025	[-0.0045, +0.0091]	0.000	0.0097	+1.62	0.000

Fold 1 in full — the gate against all three experts (the comparator-change test):

MoE minus	Δ Dice	95 % CI	p	MDE	Δ HD95 (mm)	p
Baseline (B)	+0.0082	[+0.0017, +0.0153]	0.000	0.0098	-3.07	0.000
DistMap (D)	+0.0063	[−0.0005, +0.0142]	0.000	0.0106	−0.87	0.000
Kervadec (K)	+0.0083	[+0.0034, +0.0144]	0.000	0.0080	-2.26	0.000

The best single expert at fold 1 is **DistMap (D)**, and it is not the one the gate is built on. Against it the gate gains +0.0063 Dice and −0.87 mm of HD95, and the confidence interval of that Dice margin (−0.0005 to +0.0142) is the only one of the three whose lower bound sits below zero — the honest reading is that the margin over the best expert is positive in the mean and established on the boundary metric, not established on Dice by a single fold. The comparator no longer reverses the verdict, which is what this test exists to find out.

Across folds 1–4 the V3 gate's mean effect against its own baseline is **+0.0061 Dice** — four of four folds favourable, constant sign, and a fold-to-fold spread of 0.0067 that is 5.3 times the 0.0013 seed spread of fold 0 used to normalise it. Over the five folds at a homogeneous reference the mean is +0.0054, with five of five folds favourable and three of five above the pre-declared smallest effect of interest (0.003). Every per-fold delta remains below its own MDE, so no fold establishes its effect alone; what carries the result is the constant sign across disjoint populations, not any one of them.

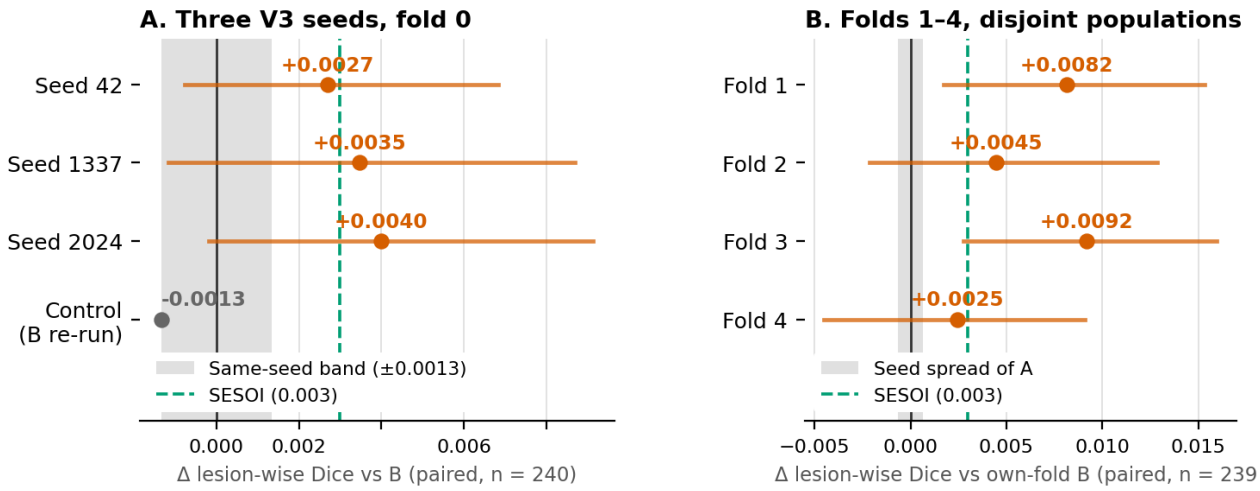

Figure 4: Figure 4 — What a single seed of fold 0 could not show: the three seeds of the V3 gate, trained in one campaign, their spread set against the re-training control band with the sign holding at all three (A), and the four disjoint folds 1–4, where the sign holds at every fold and the gate leads the best expert of that fold in all 4 (B).

4.6 What the consensus win owes to what

Section 4.1's +0.0056 is worth taking apart, because two thirds of it are not the consensus mechanism at all. **+0.0037 comes from using K rather than B as the primary expert, and +0.0019 from the veto itself.** The larger half does not survive the cross-validation, where K sits 0.0009 *below* B (Section 4.7); the smaller half is carried by the one patient in 240 that this veto changes.

That is the general pattern for vetoes under this protocol, and it is measured directly. Twelve veto arms are compared to **their own primary expert**, on the same predictions, so that retraining noise plays no part. After the official cleaning, seven of the twelve change the score of **zero** patients out of 240; four change one patient; one changes two. The mean differences are +0.0000 for every veto whose primary is D, and for three of the four whose primary is B; the three vetoes primed on K gain +0.0012 to +0.0019, and `B vetoed by D AND K` *loses* 0.0007.

Before the cleaning, the same vetoes are worth +0.0079 to +0.0375 lesion-wise Dice (Table 8; Figure 5 shows the absorption arm by arm). The veto does real work — it just does work the official protocol has already done. Both mechanisms delete small unsupported components; the official size thresholds (1000/250/500 voxels) are coarse enough to catch essentially everything a corroborating expert would have rejected, because the components the vetoes remove weigh tens to hundreds of voxels. The two patients where a veto still changes something are instructive in opposite directions: on one, the veto removes a genuine false positive; on the other, it removes a true lesion.

This does not retract Section 4.1 — the arm at the top of the table is still a consensus, and it is still free — but it says what a reader should expect to reproduce: under a protocol that already applies a size filter, a veto adds a rounding error, while a *vote*, which can build shapes no expert proposed, remains the mechanism with something left to contribute (Section 4.4).

Table 8 — What a veto does on its own, before and after the official post-processing

Each veto is compared to **its own primary expert**, on the same predictions: this comparison is immune to retraining noise.

Veto	Primary	Patients changed / 240	Δ Dice after cleaning	Δ Dice before cleaning
B vetoed by D	B	0	+0.0000	+0.0178
B vetoed by K	B	0	+0.0000	+0.0135
B vetoed by D AND K	B	1	-0.0007	+0.0240
B vetoed by D OR K	B	0	+0.0000	+0.0079
D vetoed by B	D	0	+0.0000	+0.0294
D vetoed by K	D	0	+0.0000	+0.0219
D vetoed by B AND K	D	0	+0.0000	+0.0353
D vetoed by B OR K	D	0	+0.0000	+0.0170
K vetoed by B	K	1	+0.0019	+0.0300
K vetoed by D	K	2	+0.0012	+0.0236
K vetoed by B AND D	K	1	+0.0019	+0.0375
K vetoed by B OR D	K	1	+0.0019	+0.0163

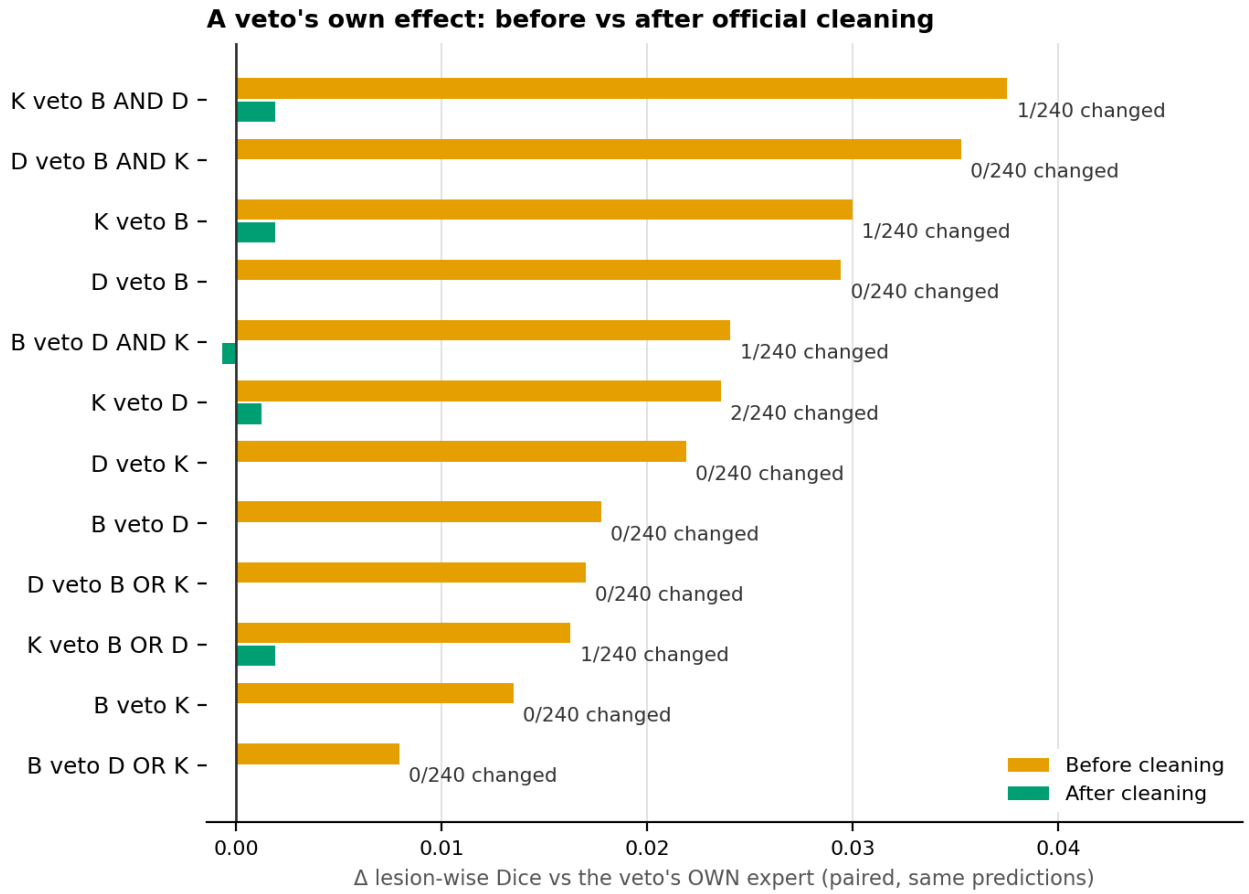

Figure 5: Figure 5 — The effect of a veto before and after the official cleaning, per arm.

4.7 Why the ceiling is low: three experts are one model measured three times

On fold 0, the three experts rank K (0.8857) > D (0.8832) > B (0.8820), and K − B is +0.0037 with a raw Wilcoxon p of 6.6×10^{-3}. Over the full cross-validation the ranking **reverses**: B 0.8774, D 0.8768, K 0.8765, with K − B = −0.0009 (95 % CI [−0.0033, +0.0015], p = 0.12) against a minimum detectable effect of 0.0034 (Table 9). On HD95 the same pattern holds — 21.92, 21.84 and 22.28 mm, spread 0.44 mm. Four independent training runs of fold 0 give an inter-seed standard deviation of 0.0019 Dice (0.0007 for B, 0.0024 for D and K), and two runs of the *same* model at the *same* seed differ by 0.0018. The spread between the three experts over 1196 cases is 0.0009 — half the distance between a model and itself.

That is why every effect in this paper is small, and it was predictable before any combination was run. An ensemble only pays when the disagreement between its members exceeds their error (Theisen et al. 2023). Both quantities are directly measurable on the same per-patient scores. The mean error of the three experts is 0.118, 0.117 and 0.114 (1 − lesion-wise Dice). Their mean pairwise disagreement — the mean absolute difference between two experts' scores on the same patient — is 0.0076 (B–D), 0.0101 (B–K) and 0.0112 (D–K). **Disagreement is 8.3 % of error.** The per-patient scores correlate at Pearson 0.93 to 0.97: when one expert fails, the others fail with it. Of the 28 patients scoring below 0.70 for at least one expert, **22 score below 0.70 for all three**.

The strongest possible consequence of this is an upper bound. An omniscient oracle that picked, for each patient, whichever of the three experts happened to be best would score 0.8909 — just **+0.0051** above the best single expert. That is the ceiling of any strategy that selects among these three experts, it is unreachable in practice, and it is only 2.8 times the distance between a model and a retrained copy of itself. Symmetric votes are not bound by it, since they can synthesise shapes no expert proposed, which is consistent with the majority vote being the one arm that clears correction.

Read against that ceiling, the consensus results are not disappointing, they are close to what was available: +0.0056 observed where +0.0051 was the selection ceiling and 8.3 % relative disagreement was the raw material. The three experts were trained on the same data, with the same backbone and the same hyper-parameters, differing only by an auxiliary loss term that Papers 1 and 2 each found to be null. Expecting a large effect from combining them was never reasonable; what this section does is replace that intuition with a number.

Table 9 — The three experts over the full 5-fold cross-validation (n = 1196)

Expert	Dice	HD95 (mm)
Baseline (B)	0.8774	21.92
DistMap (D)	0.8768	21.84
Kervadec (K)	0.8765	22.28

Paired comparison	Δ Dice	95 % CI	p	MDE
D-B	-0.0006	[-0.0029, +0.0017]	0.403	0.0032
K-B	-0.0009	[-0.0033, +0.0015]	0.124	0.0034
K-D	-0.0003	[-0.0027, +0.0022]	0.317	0.0035

4.8 What the mean does not say: the distribution and its tail

Every number above is a mean over 240 patients. A clinical tool is not used on a mean, it is used on a patient, and on this metric the two are far apart. For the baseline, the **median** patient scores 0.9363 where the mean is 0.8820, and the median HD95 is 2.17 mm where the mean is 20.97 mm — a gap of 18.8 mm between the typical case and the average case. The mean of an official BraTS score is a statement about the tail, not about the patient in front of you.

The tail is short and heavy. Averaging the worst 10 % of cases ($CVaR_{10}$, n = 24) gives a Dice of 0.5294 and an HD95 of 149 mm for the baseline; the 90th percentile of HD95 is already 75 mm. Behind those numbers sit **seven patients out of 240 for whom no enhancing lesion is matched at all**, each scored with the 374 mm sentinel the official scorer assigns when matching fails. Those seven cases alone account for **42.5 %** of the mean HD95 on ET: removing them drops it from 24.70 to 14.20 mm. Over the 720 patient-region pairs, three more are pure false positives against an empty ground truth, one has no lesion matched in either direction, nine have an empty ground truth correctly predicted empty, and the remaining 700 have at least one matched lesion.

Counting failures makes the aggregation choice visible. At a threshold of Dice < 0.5, the baseline fails on **9 of 240 patients (3.8 %)** when the three regions are averaged, and on **43 of 240 (17.9 %)** when the patient is judged on his *worst* region. The second number is the clinical reading — a patient whose enhancing tumour is missed is a failure even if whole tumour and tumour core are perfect — and it is 4.8 times the first. Averaging three nested regions divides a single-region failure by three.

Against that background, one arm and one only separates itself: `cc_K_B` (with its two exact copies) **stochastically dominates** the baseline: the two CDFs never cross, and `cc_K_B` is ahead on 94 % of the score grid, the largest CDF gap being 0.033 — eight patients out of 240 at the widest point. It is ahead everywhere or level, at no point behind. Its advantage is also concentrated where it matters: +0.0191 on $CVaR_{10}$ against +0.0056 on the mean, a factor of 3.4. What it does not do is rescue anyone — a McNemar test on the hard criterion finds **zero discordant patients** at every threshold (0.5, 0.7, 0.8). The veto lifts cases that were already poor without moving a single one across the failure line. Both facts are in Figure 6, and only the first is in Table 2. Three pinned patients of the companion 3D viewer put faces on that individual level: a case where the gate finishes ahead of the vote and of the three-expert consensus shown on its plate (Figure 7), a case where it merely holds the level of the best single expert (Figure 8), and a case where it collapses below the baseline, every expert, every veto and every vote (Figure 9).

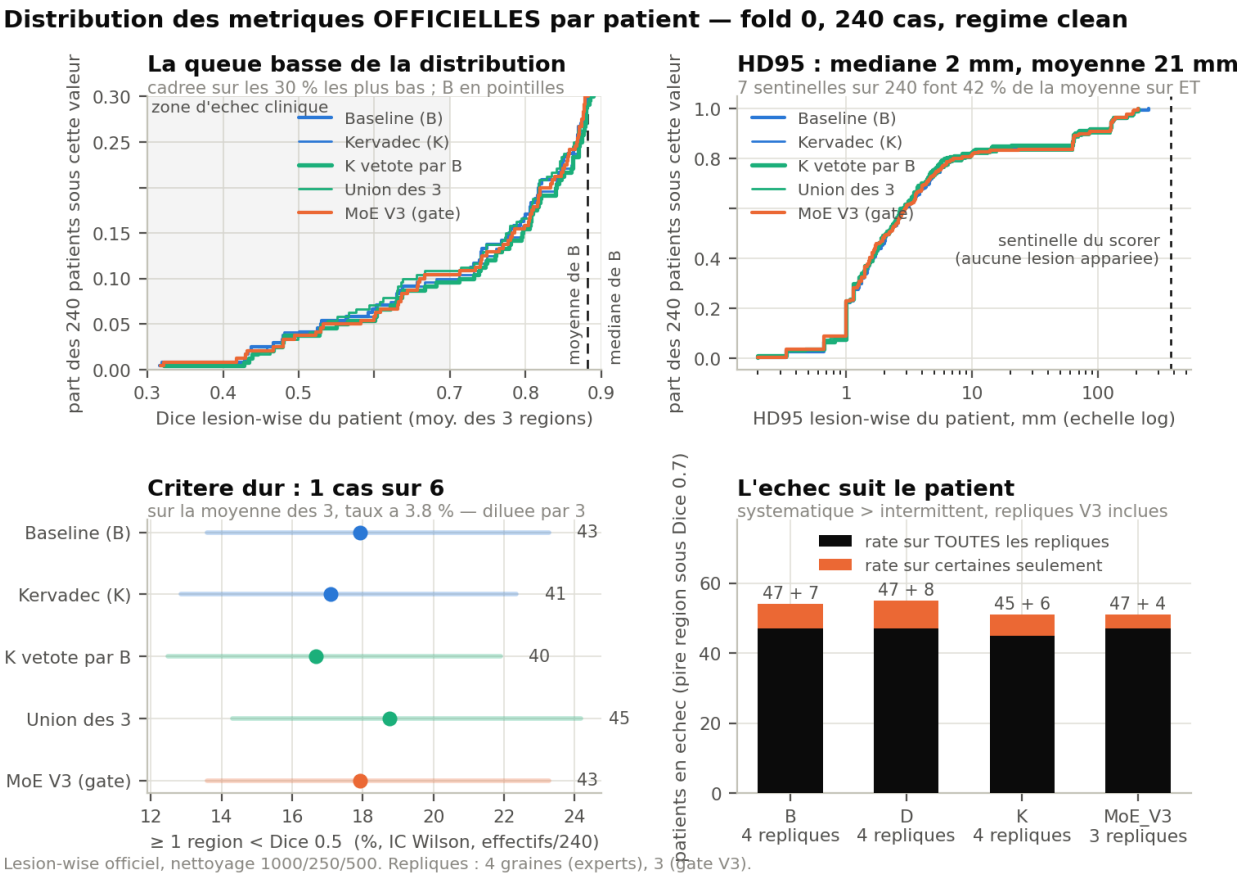

Figure 6: Figure 6 — What the mean does not say: the per-patient distributions and their tail. Dice CDFs framed on the tail (a), HD95 CDFs on a log axis (b), failure rates on the hard criterion — the patient's worst region — with Wilson CIs (c), and whether a failure is the same patient from one seed to the next (d).

The next three figures are captures of the companion 3D viewer (https://guillaume-cassez.fr/imagerie-medicale/brats/2023-distance-map/viewer/) in the official regime — white background, BraTS-2023 component cleaning, viewer region colors, sagittal view, brain masked. Each plate shows, for one patient and one camera pose: ground truth / deterministic 3-expert consensus / majority vote / patch-wise MoE V3. The plates were re-captured on the V3 gate on 2026-09-21, and every number of their captions is official lesion-wise Dice (mean over the 3 regions, clean regime, fold 0), computed for all 42 arms of every plate by `scripts/paper3_planches_v3.py` from the CSVs of the official BraTS-2023 scorer (component thresholds 1000/250/500) and written to `papers/paper3/tables/v3_planches.json` — no caption number is typed by hand.

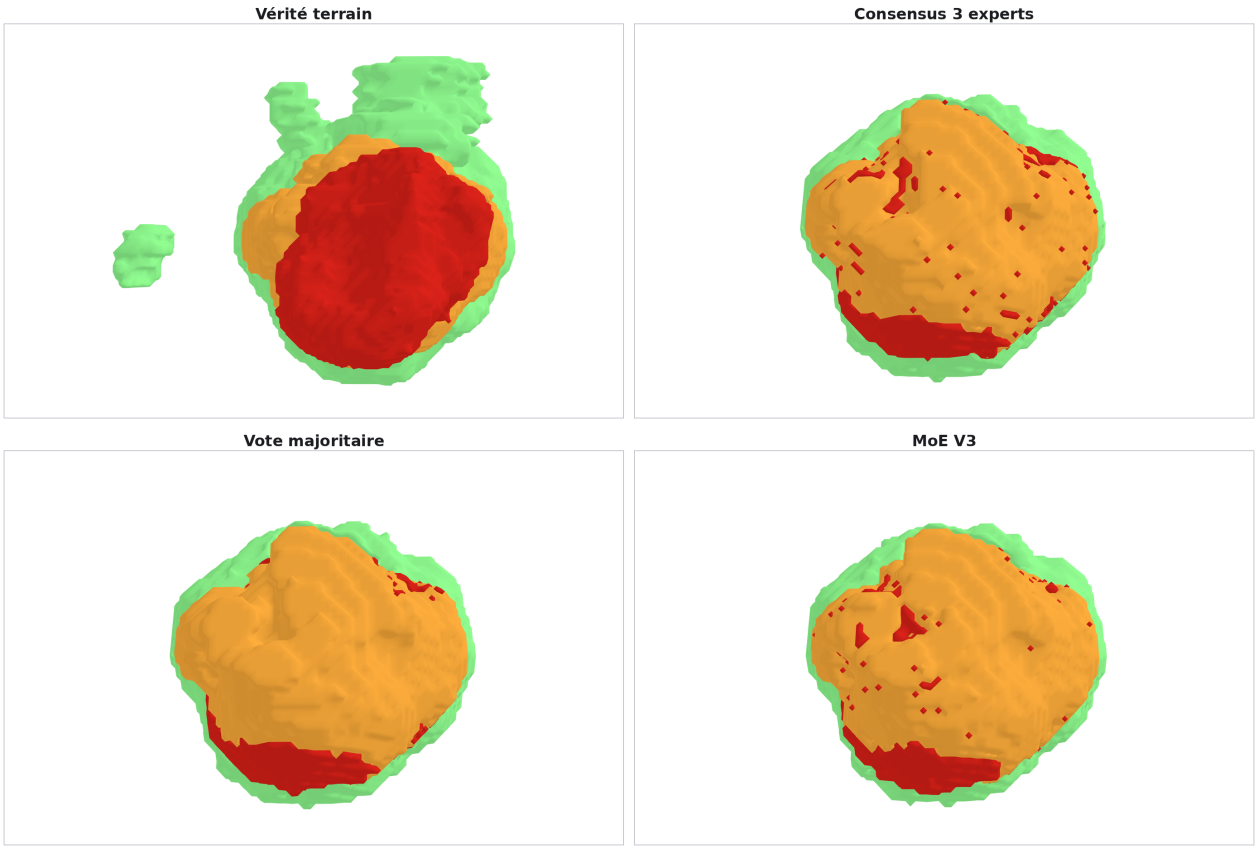

Figure 7: Figure 7 — Pinned case P3A (`BraTS-GLI-00442-000`, fold 0): **the gate leads, by less than its own seed noise.** Patch-wise MoE V3 at 0.7332 of official lesion-wise Dice, ahead of the majority vote (0.7327), of the deterministic three-expert consensus (0.7319) and of DistMap (0.7282) on this patient, the baseline sitting at 0.7309. Its lead over the vote is smaller than its own seed-to-seed span here, 0.7332 to 0.7347 (seeds 42, 1337, 2024). Of the 42 arms scored on this patient the gate ranks 11.

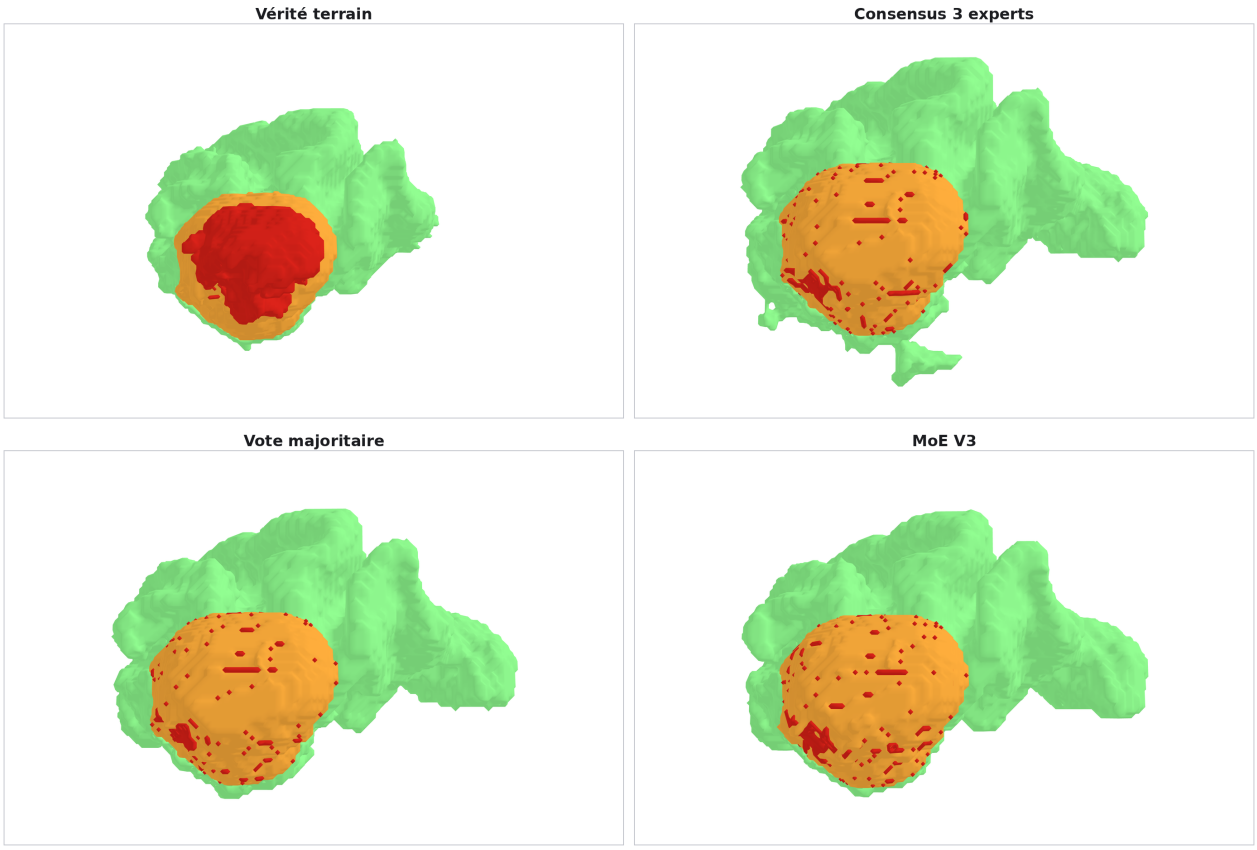

Figure 8: Figure 8 — Pinned case P3B (`BraTS-GLI-01004-000`, fold 0): **the gate holds its level.** Patch-wise MoE V3 at 0.9631, 0.0011 behind the best single expert on this patient (DistMap, 0.9642), ahead of the majority vote (0.9583) and of the deterministic three-expert consensus (0.9554); its three seeds land between 0.9629 and 0.9643. Of the 42 arms scored on this patient the gate ranks 13: on a patient a single expert already solves, the gate adds neither gain nor loss.

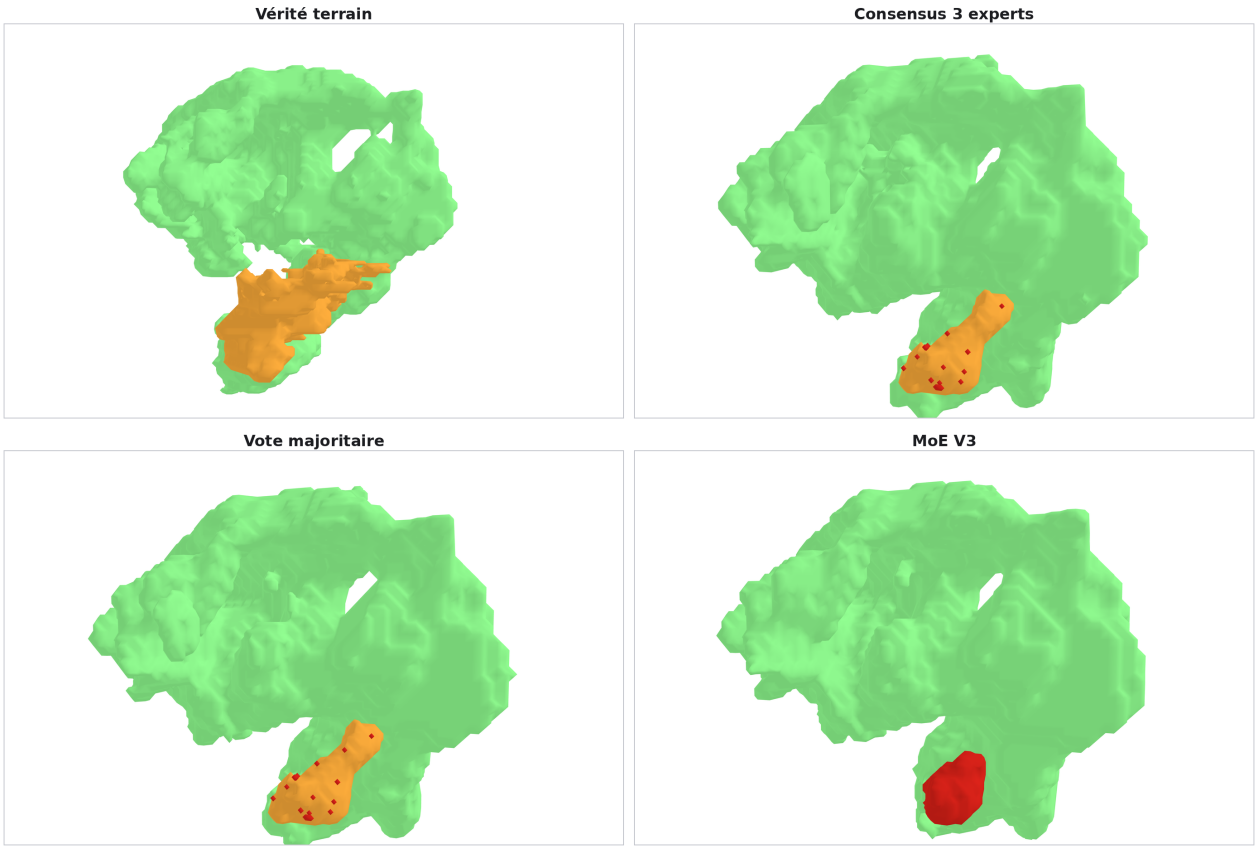

Figure 9: Figure 9 — Pinned case P3C (`BraTS-GLI-00621-000`, fold 0): **the gate breaks.** Patch-wise MoE V3 collapses to 0.3174, rank 40 of the 42 arms scored on this patient — below the baseline (0.5254), below every single expert, below the deterministic three-expert consensus (0.4794) and below the majority vote (0.4548), which beats it by 0.1373. After the official component cleaning the gate's enhancing-tumour Dice is 0.0000 (the HD95 sentinel) where the deterministic consensus still keeps 0.3420; the three seeds all collapse (0.3174, 0.2495, 0.2501), so this is no seed accident but the individual-level face of the population verdict (§4.2, §4.5).

4.9 The three regions do not cost the same, and the mean already knows it

The three BraTS regions are **nested**: ET ⊂ TC ⊂ WT. Averaging their three Dice scores is therefore not equal weighting per tissue. A voxel of enhancing tumour is scored three times, a voxel of necrosis twice, a voxel of oedema once, and the volumes differ by a factor of five (fold-0 medians: WT 89 400, TC 30 702, ET 18 630 voxels; oedema alone is **62 % of the WT volume**). Linearising the Dice around a perfect prediction — its slope is $-1/(2V)$, the same for a false positive and a false negative to first order — the cost of one mis-segmented voxel on the mean of the three is $(1/3)\cdot\Sigma_{R \supset T}\ 1/(2\cdot V_R)$, which gives **×7.8 in enhancing tumour, ×3.7 in necrosis, ×1 in oedema** (Figure 10a). The metric already ranks the tissues; it just does not say so. This derivation holds for the voxel-wise Dice and near a good prediction, not for a failed case.

That implicit weighting is not an artefact to be corrected, because it points the same way as the clinical literature. Extent of resection in glioblastoma is graded on the **residual contrast-enhancing** volume, and median overall survival falls from 24 months in class 1 to 19, 15 and 10 months in classes 2 to 4 (Karschnia et al. 2023). The core is what the surgeon is judged on.

The symmetric claim — that oedema does not matter — does not survive the same literature: BraTS label 2 merges vasogenic oedema with non-enhancing tumour infiltration, which is itself a surgical and radiotherapeutic target (Molinaro et al. 2020 ; Niyazi et al. 2023 ; Yilmaz et al. 2024). "The core matters more" is defensible; "the oedema does not matter" is not.

Does the study's ranking depend on this? Almost not at all. Ranking the arms on the core (TC and ET) instead of the mean of three gives a Kendall τ of **0.94** with the published ranking, and `cc_K_B` stays first on the mean, on the core, and on TC. Ranking on **WT alone** gives $\tau = 0.08$ ($p = 0.60$) and sends the same arm to **rank 8** — but that is a statement about WT, where nothing separates any arm, rather than a competing ranking. The decomposition says why: `cc_K_B` gains +0.0106 on TC and +0.0061 on ET against +0.0002 on WT (Figure 10d), and the only arm that loses to the baseline, `cc_B_and_DK` at −0.00067, loses **only** on WT. The gain of this paper is a gain on the core.

What the region view does change is the failure count, exactly as in Section 4.8: the baseline fails on **20 patients on the core** against 9 on the mean of three, with failures spread across ET (25), TC (21) and WT (19) — and all seven 374 mm sentinels lie on ET. We therefore do not re-weight the metric: the challenge made its choice, the ranking is insensitive to it, and a hand-made weight would double-count the core. We report, next to the mean, the failure rate on the core and on the worst region.

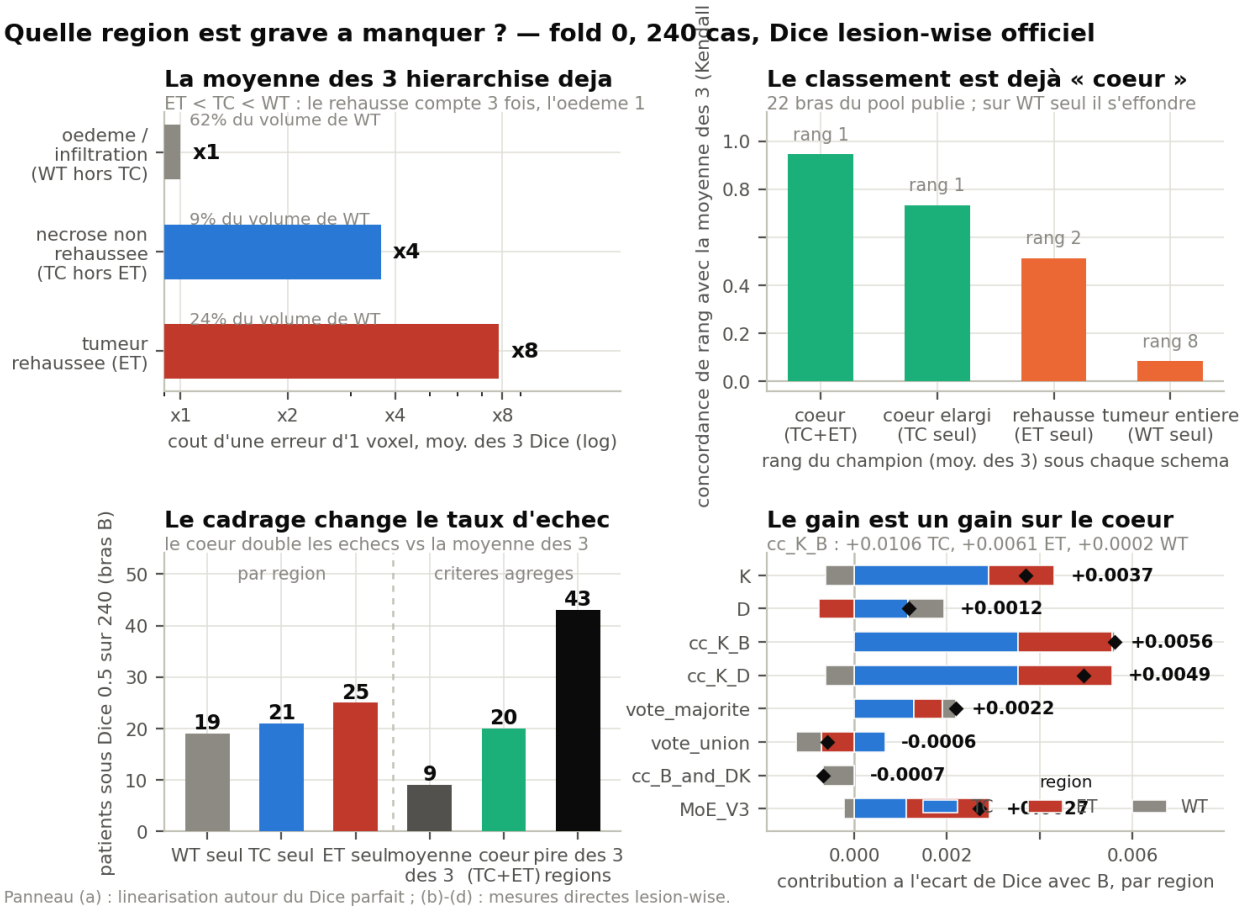

Figure 10: Figure 10 — The three nested regions do not cost the same: the cost of one mis-segmented voxel per tissue, on a log axis (a), the ranking read on the core versus on WT alone (b), the patients counted as failures under each aggregation (c), and which region carries each arm's gap to the baseline, as signed stacked bars (d). Panel (a) is a linearisation around the perfect voxel-wise Dice; the other three are direct lesion-wise measurements.

5. Discussion

What to spend on. A team with a strong backbone and a fixed budget can put the next GPU-week into a learned combination mechanism or into nothing at all and combine what it already has. On this dataset, under the metric the challenge actually scores, the answer depends on which learned mechanism is meant. A patch-wise gate built from eight random blocks behind one baseline returns −0.0007 over five folds and is beaten by the free operator; the same architecture carrying four loss specialists initialised from their own checkpoints returns +0.0057 and beats all twenty-four free operators on a paired mean margin, the best of them by 0.0044. The recommendation is therefore not "consensus is powerful" — Section 4.7 caps how powerful it could have been at +0.0051 — nor "gating is powerful": it is that what a gate adds here is *initialisation*, and that a team which cannot afford four prior trainings should still start with the veto.

Two properties beyond the score reinforce that, and neither is a p-value. A consensus operator is **deterministic**: rerun it and the number does not move, whereas the gate must be retrained and lands anywhere in a 0.0013 band (0.0034 for the eight-random-block version it replaced). It is also **auditable**: the rule is five lines, has no learned parameter, and its behaviour on a new case can be predicted by reading it — which matters in any setting where a change to a device must be justified rather than merely benchmarked. The gate's last counter-argument is deployment economy: one forward pass against two or three. On a single-GPU workstation that trade is real, and it is stated in Table 3; in a production service with more than one GPU it largely evaporates, because the consensus passes are independent and parallelise (Section 4.2). This study finds nothing on the accuracy side to offset the training cost — including on the composite

ground where one might have expected the gate to recover, since it does not improve the two endpoints jointly any more than a re-trained baseline does (Section 4.3).

A note on composite scores. The two endpoints co-move within patients for every arm here, so a weighted combination carries almost no extra information — and where a composite does change the answer (the BraTS rank score disagrees with the mean on three arms), it does so because the rank counts wins and ignores their size. Report both endpoints separately; treat any single composite as a summary to check, not a criterion to optimise.

Where the mechanism's own room went. A connected-component veto is a size-and-support filter; the official post-processing is a size filter with thresholds an order of magnitude above the components at stake. Running both changes one or two patients in 240 — which is why the veto contributes +0.0019 of a +0.0056 margin, and why symmetric votes, which can synthesise shapes no expert proposed, are the sub-family that still moves the metric. The prior condition for an ensemble to pay is that its members disagree more than they err (Theisen et al. 2023); here disagreement is 8.3 % of error, the per-patient scores correlate at 0.93–0.97, and 22 of the 28 hardest cases are hard for all three experts at once. Three models that share a dataset, a backbone, a schedule and a seed, and differ only by an auxiliary loss term already shown to be null, are not a diverse committee — they are one model measured three times. That this committee still produced the study's best arm is a point in favour of the mechanism, not against it: a genuinely diverse one has more to work with. The published ensembles that do win on BraTS combine *architectures* (Kamnitsas et al. 2018 ; Ferreira et al. 2024), and the medical MoE models that report gains index their experts on something externally heterogeneous such as missing modalities (Liu et al. 2024). A gate can only exploit a partition that exists; on a single-task, single-protocol cohort there may be none to find.

A protocol warning that cuts both ways. Papers reporting consensus or ensemble gains on *raw* predictions are not necessarily wrong, they are measuring a quantity the challenge protocol removes. The same twelve vetoes are worth up to +0.0375 Dice before cleaning and +0.0019 after. Which of those two numbers one publishes is a protocol decision, not an empirical one, and it changes the conclusion by a factor of twenty. We report the post-cleaning number throughout, including where it deflates our own best arm.

How to read a lesion-wise table. The symmetric votes expose a property any BraTS-2023 comparison inherits: with a 374 mm penalty per missed lesion or false positive, a handful of cases governs the mean while the rank test follows the majority of patients. Two of our three votes have a mean and a rank test of opposite sign; the one we recommend, majority, is the one where they agree. We see no way to rank arms honestly under this metric without reporting both, plus the count of patients that improve and degrade.

The control arm. Training the same model twice with the same seed moved the official Dice by 0.0018 and produced a raw p-value of 0.032. Any protocol that would have declared a method significant at that effect size would have declared GPU non-determinism significant too. This is not a local quirk: running nnU-Net over 50 seeds, Åkesson et al. found the best seed significantly outperforming up to 76 % of the others and concluded that significance is a weak indicator of a true difference between learning algorithms (Åkesson et al. 2024 ; Picard 2021), and a recent audit estimates that more than half of medical-imaging segmentation papers carry a non-negligible risk of a false outperformance claim (Christodoulou et al. 2025). Reporting a control arm that tests nothing is the cheapest available defence, and we recommend it as routine (Bouthillier et al. 2021 ; Maier-Hein et al. 2024). It is also what makes the positive part of this paper credible: the majority vote clears a bar that the control does not.

The gate, and what it does not support. The gate changes the initialisation, not the mechanism, and three of the numbers this discussion rests on are its own: its dispersion across three seeds (0.0013 Dice) is 0.96× its own control's, so its effect is not inside its own noise; its mean over five disjoint folds is +0.0054 with a constant sign (+0.0061 on folds 1–4 alone, four of four favourable); and the consensus's Dice margin over it is no longer established by the test, although the HD95 margin holds at every seed. What V3 does *not* do is support the reading that the gate learned to route. Measured on the last epoch of each of its seven training logs, the gate's preference stays close to indifference — the most-chosen expert takes 0.507 to 0.583 of the patches where 0.25 would be no preference at all on four experts, and the noise-free routing entropy stays at 0.892 to 0.995 where 1.0 is maximal — no expert dies in any run, and the degree of decision does not track the official gain: the most decisive fold (H = 0.892, fold3) gains the most (+0.0092) while the most peaked partition (part_max = 0.583, fold4) gains the least (+0.0025). The gain is measured and constant; the specialisation story is not what these measurements support. Spending on agreement remains the cheaper route to a comparable effect, and the gate remains 300 epochs on top of four already trained experts where a veto costs nothing beyond them.

6. Limitations

The consensus arms are measured on fold 0 only. This is a compute-scheduling limitation and nothing more — the expert predictions exist for all five folds and the measurement is CPU work; Section 4.7 makes it a real limitation nonetheless, because fold 0 demonstrably misranks the experts. Every statement made here about consensus operators is therefore a statement about one fold, and Section 4.6 quantifies how much of a fold-0 margin can evaporate on the others: of the +0.0056 that puts a consensus at the top of Table 2, the +0.0037 attributable to the primary expert is the part at risk, and the head-to-head advantage over the gate (Section 4.2) rests on the same fold as the gate's own fold-0 seeds — a like-for-like comparison, but on one fold.

The consensus arms are also measured at **one seed**, where the three experts have four. The inter-seed standard deviation used throughout to normalise effects — 0.00186 lesion-wise Dice and 0.547 mm HD95 under the official cleaning — is computed on the three experts and on them alone, and the JSON states that base explicitly. It is therefore *borrowed* when applied to a consensus, in exactly the sense Section 4.5 criticises for the gate. The borrowing is not neutral, because seed noise is known to depend on the arm: the official component cleaning absorbs it by a factor of 10 to 40 on residual arms and only 1.2 on a free head, and a veto — which removes components — could plausibly do either. The consequence is concrete: the z(min) = +3.02 that puts *K vetoed by B* at the top of Table 4 is a ratio whose denominator was measured on a different arm.

The mixture-of-experts is measured on three seeds of fold 0 and on the four remaining folds; the detail is in Section 4.5 and Tables 6–7. What matters for the limits is the comparison quantity: the arm's own dispersion — 0.0013 Dice across the three seeds of fold 0, 0.0067 across the four remaining folds — not its distance to a single reference run. The effect exceeds its seed dispersion; no single fold exceeds its own minimum detectable effect. The campaign retrained its own control, so its dispersion is never to be divided by another arm's noise.

No fold carries the power to establish its own effect. The minimum detectable effect on 239–240 patients stays above every per-fold delta of the V3 gate (+0.0027 against an MDE of 0.0054 at fold 0; +0.0082 against 0.0098; +0.0045 against 0.0108; +0.0092 against 0.0093; +0.0025 against 0.0097). What the cross-validation buys is a constant sign over five disjoint populations (mean +0.0054), not a significant test in any one of them.

The V3 gate's routing is not decisive, so its gain is not attributable to specialisation. On the last epoch of each of the seven runs — measured by `scripts/paper3_v3_routage.py` from the training logs themselves, not read off a summary (`tables/v3_routage.json`, 7/7 runs at epoch 299) — the most-chosen expert takes 0.507 to 0.583 of the patches, where 0.25 is no preference at all on four experts; the noise-free routing entropy stays at 0.892 to 0.995, where 1.0 is maximal; no expert dies (0/4 in every run); and no readable link appears between how decisive a run's gate is and how much that run gains. A second asymmetry belongs here: the fourth expert of the gate (the blob arm G) is the worst of the four on its own — −0.0038 Dice over the five folds, negative on all five, 28-th of twenty-nine — so the gate hosts a specialist whose isolated contribution is negative, and no per-expert ablation of the gate was trained in this study. What the +0.0057 of the gate owes to each of its four experts is therefore not measured; what is measured is that its routing does not choose between them, and that the same expert used as a veto on the baseline is the best of the twelve two-expert vetoes (+0.0011).

Fold 0's expert predictions for B and D come from re-prediction runs kept under suffixed directories, not from the original fold-0 validation outputs, which were lost; the provenance of fold 0 is therefore not uniform with folds 1–4. This is documented in the expert registry and affects fold 0 only.

STAPLE (Warfield et al. 2004) is not among the compared operators, for lack of a verified implementation in this repository; a learned combiner over frozen experts is deliberately excluded (Section 3.4). The size-based cleaning uses 6-connectivity while the scorer and the consensus operators use 26-connectivity — a documented inconsistency, kept because changing it would break comparability with Papers 1 and 2. Finally, all results are on the adult glioma cohort of a single challenge edition, evaluated on internal held-out folds and not on the official leaderboard.

7. Conclusion

Twenty-nine arms, one dataset, one metric set, and a control that tests nothing. Under the official BraTS-2023 lesion-wise metrics, the arm that scores highest is a **learned patch-wise mixture-of-experts** whose four experts are initialised from four already-trained loss specialists: +0.0057 Dice over its own baseline averaged across five disjoint folds, positive on five of five, above the pre-declared smallest effect of interest on four. Every one of the twenty-four training-free consensus operators built from the same experts is behind it on a paired mean margin, the best of them at +0.0013 and on five folds out of five; none reaches that effect of interest. The earlier version of the same layer, built from eight random blocks behind one baseline, sits below the free operator at −0.0007 — so the architecture is not what wins, the initialisation is.

That is the negative result this paper is titled after: **the gate does not choose.** Over the seven runs the largest expert share is 0.507 to 0.583 where 0.25 is no preference at all on 4 experts, the normalised patch entropy is 0.892 to 0.995 where 1.0 is maximum, and no expert dies anywhere. A routing that does not decide cannot be the mechanism of the gain, and this paper does not claim it is; what it claims is that hosting four specialists in one network and starting them from their own checkpoints is worth +0.0044 over the best thing you can do for free.

The boundaries of that conclusion are measured rather than asserted, and they are what makes it usable. No single fold clears its own minimum detectable effect, so the claim rests on a constant sign across five disjoint populations and not on a significant per-fold test; the fold-0 reference drifts by −0.0032 between scoring campaigns, which is why all twenty-nine arms are compared through one code path on one baseline per fold; the blob-loss expert is the worst of the four on its own (−0.0038, none of the five folds positive) and still carries the best of the twelve two-expert vetoes (+0.0011, second of the twenty-four operators); and the price is 300 epochs on top of four full trainings, where a veto costs nothing and its passes parallelise across GPUs. The effects being compared here are of the same order as the distance between a model and itself — and we measured both distances instead of assuming them.

The practical recommendation follows at no cost: before building a committee, measure whether its members disagree more than they err, report a control arm that tests nothing, and try the free operator first. On this dataset those three checks no longer point the same way — and that is the result: the free operator has a ceiling, the gate that beats it is not routing, and the margin it captures is the one the experts' own initialisation carries.

Author contributions

Guillaume Cassez (lead author): study design, model training, evaluations and analyses, manuscript writing. **Stanislas Larnier**: formulation of the research questions, methodological guidance, careful reviews of successive drafts of the paper. Both authors approved the final version and the author order.

References

- Abe, Buchanan, Pleiss, Zemel, Cunningham (2022). *Deep Ensembles Work, But Are They Necessary?*. Advances in Neural Information Processing Systems (NeurIPS).
- Baid, Ghodasara, Mohan, Bilello, Calabrese, Colak et al. (2021). *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*. arXiv preprint arXiv:2107.02314.
- Berger (1982). *Multiparameter Hypothesis Testing and Acceptance Sampling*. Technometrics. doi:10.1080/00401706.1982.104877
- Bouthillier, Delaunay, Bronzi, Trofimov, Nichyporuk, Szeto et al. (2021). *Accounting for Variance in Machine Learning Benchmarks*. Proceedings of Machine Learning and Systems (MLSys).
- Christodoulou, Reinke, Andrè, Godau, Kalinowski, Maier-Hein (2025). *False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims*. arXiv preprint arXiv:2505.04720. doi:10.48550/arXiv.2505.04720.
- Fedus, Zoph, Shazeer (2022). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. Journal of Machine Learning Research.
- Ferreira, Solak, Li, Dammann, Kleesiek, Alves et al. (2024). *How we won BraTS 2023 Adult Glioma challenge? Just faking it! Enhanced Synthetic Data Augmentation and Model Ensemble for brain tumour segmentation*. arXiv preprint arXiv:2402.17317. doi:10.48550/arXiv.2402.17317.

- Isensee, Jaeger, Kohl, Petersen, Maier-Hein (2021). *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*. Nature Methods. doi:10.1038/s41592-020-01008-z.
- Isensee, Jaeger, Full, Vollmuth, Maier-Hein (2020). *nnU-Net for Brain Tumor Segmentation*. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2020).
- Isensee, Wald, Ulrich, Baumgartner, Roy, Maier-Hein et al. (2024). *nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation*. Medical Image Computing and Computer Assisted Intervention (MICCAI). doi:10.1007/978-3-031-72114-4_47.
- Jacobs, Jordan, Nowlan, Hinton (1991). *Adaptive Mixtures of Local Experts*. Neural Computation. doi:10.1162/neco.1991.3.1.79.
- Kamnitsas, Bai, Ferrante, McDonagh, Sinclair, Pawlowski et al. (2018). *Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation*. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2017). doi:10.1007/978-3-319-75238-9_38.
- Karschnia, Young, Dono, others (2023). *Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the RANO resect group*. Neuro-Oncology. doi:10.1093/neuonc/noac193.
- Kervadec, Bouchtiba, Desrosiers, Granger, Dolz, Ben Ayed (2019). *Boundary loss for highly unbalanced segmentation*. Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning (MIDL).
- Kervadec, Bouchtiba, Desrosiers, Granger, Dolz, Ben Ayed (2021). *Boundary loss for highly unbalanced segmentation*. Medical Image Analysis. doi:10.1016/j.media.2020.101851.
- Kofler, Möller, Buchner, de la Rosa, Ezhov, Rosier et al. (2023). *Panoptica – instance-wise evaluation of 3D semantic and instance segmentation maps*. arXiv preprint arXiv:2312.02608. doi:10.48550/arXiv.2312.02608.
- Liu, Wang, Li, Zhang (2024). *Mixture-of-experts and semantic-guided network for brain tumor segmentation with missing MRI modalities*. Medical & Biological Engineering & Computing. doi:10.1007/s11517-024-03130-y.
- Ma, Wei, Zhang, Wang, Lv, Zhu et al. (2020). *How Distance Transform Maps Boost Segmentation CNNs: An Empirical Study*. Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL).
- Maier-Hein, Eisenmann, Reinke, Onogur, Stankovic, Scholz et al. (2018). *Why rankings of biomedical image analysis competitions should be interpreted with care*. Nature Communications. doi:10.1038/s41467-018-07619-7.
- Maier-Hein, Reinke, Godau, Tizabi, Buettner, Christodoulou et al. (2024). *Metrics reloaded: recommendations for image analysis validation*. Nature Methods. doi:10.1038/s41592-023-02151-z.
- Menze, Jakab, Bauer, Kalpathy-Cramer, Farahani, Kirby et al. (2015). *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*. IEEE Transactions on Medical Imaging. doi:10.1109/TMI.2014.2377694.
- Moawad, Janas, Baid, Ramakrishnan, Saluja, Ashraf et al. (2023). *The Brain Tumor Segmentation (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI*. arXiv preprint arXiv:2306.00838.
- Molinaro, Hervey-Jumper, Morshed, others (2020). *Association of Maximal Extent of Resection of Contrast-Enhanced and Non–Contrast-Enhanced Tumor With Survival Within Molecular Subgroups of Patients With Newly Diagnosed Glioblastoma*. JAMA Oncology. doi:10.1001/jamaoncol.2019.6143.
- Niyazi, Andratschke, Bendszus, others (2023). *ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma*. Radiotherapy and Oncology. doi:10.1016/j.radonc.2023.109663.
- Pavlitska, Fan, Ditschuneit, Zöllner (2026). *Design and Behavior of Sparse Mixture-of-Experts Layers in CNN-based Semantic Segmentation*. arXiv preprint arXiv:2604.13761.
- Picard (2021). *Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision*. arXiv preprint arXiv:2109.08203. doi:10.48550/arXiv.2109.08203.
- Riquelme, Puigcerver, Mustafa, Neumann, Jenatton, Susano Pinto et al. (2021). *Scaling Vision with Sparse Mixture of Experts*. Advances in Neural Information Processing Systems (NeurIPS).
- Roy, Koehler, Ulrich, Baumgartner, Petersen, Isensee et al. (2023). *MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation*. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. doi:10.1007/978-3-031-43901-8_39.
- Saluja, others (2023). *BraTS-2023-Metrics: Official BraTS 2023 Segmentation Performance Metrics*. .
- Shazeer, Mirhoseini, Maziarz, Davis, Le, Hinton et al. (2017). *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. International Conference on Learning Representations (ICLR).
- Theisen, Kim, Yang, Hodgkinson, Mahoney (2023). *When are ensembles really effective?*. arXiv preprint arXiv:2305.12313. doi:10.48550/arXiv.2305.12313.
- Warfield, Zou, Wells (2004). *Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation*. IEEE Transactions on Medical Imaging. doi:10.1109/TMI.2004.828354.

- Wolpert (1992). *Stacked generalization*. Neural Networks. doi:10.1016/S0893-6080(05)80023-1.
- Xue, Tang, Qiao, Gong, Yin, Qian et al. (2020). *Shape-Aware Organ Segmentation by Predicting Signed Distance Maps*. Proceedings of the AAAI Conference on Artificial Intelligence. doi:10.1609/aaai.v34i07.6946.
- Yilmaz, Kahvecioglu, Yedekci, others (2024). *Comparison of different target volume delineation strategies based on recurrence patterns in adjuvant radiotherapy for glioblastoma*. Neuro-Oncology Practice. doi:10.1093/nop/npae009.
- Åkesson, Töger, Heiberg (2024). *Random effects during training: Implications for deep learning-based medical image segmentation*. Computers in Biology and Medicine. doi:10.1016/j.compbiomed.2024.108944.