Contents

La porte ne choisit pas : un mélange d'experts initialisé depuis ses experts surpasse un consensus sans entraînement sous les métriques officielles BraTS-2023

Guillaume Cassez · Stanislas Larnier

Recherche indépendante

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) — la thèse est inversée sur la mesure, et le quatrième expert est scoré. Par décision d'auteur (2026-09-11), le MoE V3 remplace la V2 partout dans ce papier et l'expert G à blob loss y entre comme bras à part entière. Les vingt-neuf bras — les quatre experts denses seuls, les vingt-quatre opérateurs de consensus déterministes construits à partir d'eux, et les deux portes patch-wise apprises — sont re-scorés par un seul chemin de code sous le protocole officiel sur les mêmes 239 patients de chacun des cinq folds disjoints (`scripts/paper3_v3_vs_consensus.py`, qui reproduit le delta publié du fold 0 de la V3 à 0,00e+00 avant d'écrire quoi que ce soit). La V3 est première à +0,0057 sur son propre baseline, positive sur cinq folds sur cinq et au-dessus du plus petit effet d'intérêt pré-déclaré (0,003) sur quatre ; le meilleur consensus atteint +0,0013 et **aucun des vingt-quatre** n'atteint cet effet d'intérêt.

Cette version tient la promesse que la précédente faisait sur G. Le bras est écrit avec ses propres chiffres officiels, lus dans ses cinq checkpoints et par le même scorer que tout autre bras : entraîné pendant 300 époques sur chacun des cinq folds avec son terme de distance signée désactivé (λ = 0,0 dans les cinq, si bien qu'il ne diffère du baseline que par le terme blob, β = 0,5), il score −0,0038 Dice sur les cinq folds — négatif sur les cinq, 28e des vingt-neuf bras et dernier des quatre experts denses. Le veto qu'il applique au baseline (`cc_B_G`) est le meilleur des douze vetoes à deux experts à +0,0011, et le deuxième des vingt-quatre opérateurs de consensus. Deux énoncés antérieurs sont corrigés ici, sur la mesure et non sur la formulation de la carte. Le bras était décrit comme scoré seulement à travers la porte, ce que ses cinq deltas par fold contredisent — et que le Résumé, la section 4.5 et la Conclusion citaient déjà contre lui. Ce veto était rangé dauphin des paires de consensus, ce que le classement contredit dans les deux lectures du mot « paire » : il est premier des douze vetoes à deux experts, et le bras qui le précède parmi les vingt-quatre (`cc_B_and_DKG`, +0,0013) est un veto un-contre-trois, pas une paire. La v0.7 est figée bit à bit dans `versions/v2/`.

Version 0.10 (2026-09-12) — les légendes suivent les figures. Les figures 1, 2 et 4 ont été recuites sur la mesure V3 (`scripts/make_paper3_v3_figures.py` : 26 bras, trois graines V3, folds 1 à 4, coût mesuré) alors que leurs légendes décrivaient encore la campagne précédente. La figure 1 comptait vingt bras contre vingt-six au tableau 2 ; la figure 4 annonçait un effet qui change de signe d'un fold à l'autre, contre un signe qui tient sur les quatre folds, sur les cinq et sur les trois graines ; la figure 2 ne disait pas qu'aucune des marges Dice du consensus n'a d'IC excluant zéro. Les trois légendes sont réécrites depuis la mesure, aucune valeur n'y est retapée à la main. Du côté anglais, les légendes LaTeX sont rendues par une conversion démontrée sur les sept légendes qu'elle ne touche pas ; le miroir français ne porte pas de paquet LaTeX — son PDF est recompilé depuis ce markdown à la bascule.

Version 0.11 (2026-09-21) — les contributions des auteurs sont énoncées. Une section « Contributions des auteurs » est ajoutée après §7, miroir des Papiers 1 et 2 : Guillaume Cassez, auteur principal (conception de l'étude, entraînement des modèles, évaluations et analyses, rédaction du manuscrit) ; Stanislas Larnier, formulation des questions de recherche, conseils méthodologiques et relectures attentives des versions successives. Aucun chiffre n'est modifié ; le corps hérité de v0.10 est inchangé.

Provenance. Tout chiffre vient de `scripts/paper3_stats.py` (`outputs/paper3/paper3_stats.json`) ; les tableaux de `scripts/make_paper3_tables.py` (`tables/table_paper3.{md,tex}`) ; les figures de `scripts/make_paper3_figures.py`. Les métriques sont calculées par la capsule officielle BraTS-2023 (`external/BraTS-2023-Metrics`), jamais par une réimplémentation locale. Le jeu complet des métriques officielles est rapporté ; un endpoint primaire pré-spécifié est séparé des analyses exploratoires ; les résultats négatifs et nuls sont rapportés comme tels.

Limite de portée (structurante). Les trois experts seuls sont mesurés sur la cross-validation complète à

5 folds (n = 1196) et sur quatre entraînements indépendants du fold 0. Les **bras de consensus ne sont mesurés que sur le fold 0** (n = 240), parce que scorer les autres folds avec les opérateurs exacts à trois experts du papier relève d'un calcul CPU pas encore lancé. Le **mélange d'experts** est mesuré sur trois graines indépendantes du fold 0 (n = 240) et répliqué sur les quatre populations disjointes des folds 1 à 4 (n = 239 chacune) ; l'entraînement des folds 2 à 4 s'est achevé le 13/08/2026 et leur évaluation officielle lésion-wise a été courue le 22/08/2026 (tableau 7). La section 4.7 montre pourquoi cela compte plus que d'ordinaire — le classement des trois experts au fold 0 est *inversé* par la cross-validation complète — et les conclusions sont formulées pour tenir quelle que soit l'issue de cette évaluation.

Résumé

Quand on dispose de plusieurs modèles de segmentation entraînés, un praticien a deux façons de les combiner : un opérateur de consensus déterministe, qui ne coûte aucun entraînement, ou une porte apprise, qui en coûte un complet. Ce papier mesure les deux l'un contre l'autre sur BraTS-2023 gliome adulte, sous les **seuls** Dice et HD95 lésion-wise officiels et rien d'autre — et c'est la porte apprise qui gagne, à une condition : elle doit être initialisée depuis les experts qu'elle héberge. Vingt-neuf bras sont scorés par un seul chemin de code sur les mêmes 239 patients held-out de chacun des cinq folds disjoints : quatre experts denses qui ne diffèrent que par un terme auxiliaire de loss (aucun, régression de distance signée, boundary loss, blob loss), vingt-quatre opérateurs de consensus déterministes construits à partir d'eux (vetos ordonnés à deux experts, vetos un-contre-trois, votes symétriques), et deux mélanges d'experts patch-wise intra-modèle dont la porte est entraînée à l'intérieur du réseau. Quatre résultats. (i) **Le bras le mieux classé de l'étude est une porte apprise** : le mélange d'experts V3 — quatre experts initialisés depuis le dernier bloc de décodage des quatre spécialistes déjà entraînés — gagne +0,0057 de Dice lésion-wise sur son propre baseline en moyenne sur les cinq folds, est positif sur cinq folds sur cinq et franchit le plus petit effet d'intérêt pré-déclaré (0,003) sur quatre. (ii) **Aucun opérateur gratuit n'atteint cet effet d'intérêt** : le meilleur des vingt-quatre bras de consensus gagne +0,0013 et, appariés directement contre la V3 sur les mêmes patients, tous ont une marge moyenne négative — 21 d'entre eux derrière sur cinq folds sur cinq, les trois exceptions (chacune impliquant K) sur quatre folds sur cinq. La marge de la porte sur le meilleur opérateur gratuit vaut +0,0044, soit 1,5 fois l'effet d'intérêt. (iii) **La porte ne choisit pas.** Sur les sept entraînements le routage reste proche de l'uniforme — part du plus gros expert de 0,507 à 0,583 là où 0,25 serait l'absence totale de préférence sur 4 experts, entropie patch normalisée de 0,892 à 0,995 là où 1,0 est le maximum — et aucun expert mort dans aucun run. Le gain vient donc de ce qu'un seul réseau *héberge* quatre spécialistes de loss initialisés depuis leurs propres checkpoints, pas d'un apprentissage du routage entre eux ; le titre énonce le résultat négatif qui rend le positif interprétable. (iv) Le coût est énoncé plutôt que caché : 300 époques par-dessus quatre experts déjà entraînés, contre zéro heure de GPU pour un veto dont les passes sont indépendantes et se parallélisent sur plusieurs GPU — et l'expert à blob loss qu'héberge la V3 est le *pire* des quatre, seul (−0,0038, aucun fold positif) mais son veto à deux experts sur le baseline est le *meilleur* (+0,0011). Ce qui ne change pas, c'est la résolution d'un fold pris isolément — chaque delta par fold reste en dessous de sa propre différence minimale détectable — si bien que l'affirmation repose sur un signe constant sur cinq populations disjointes, pas sur un test significatif fold par fold.

1. Introduction

Mettre deux modèles d'accord ne coûte rien à entraîner. Apprendre à un réseau à router ses propres patchs coûte un entraînement complet — et sur BraTS-2023 gliome adulte, sous les métriques officielles lésion-wise, c'est le mécanisme coûteux qui gagne, à une condition : la porte doit être initialisée depuis les experts qu'elle héberge. Le bras le mieux classé de cette étude est un mélange d'experts patch-wise portant quatre spécialistes de loss dans un seul réseau (baseline, régression de distance signée, boundary loss, blob loss), chacun initialisé depuis le dernier bloc de décodage d'un expert déjà entraîné pendant 300 époques. En moyenne sur cinq folds disjoints de 239 patients held-out il gagne +0,0057 de Dice sur son propre baseline, est positif sur cinq folds sur cinq et franchit le plus petit effet d'intérêt pré-déclaré sur quatre ; le meilleur des vingt-quatre opérateurs de consensus sans entraînement bâtis à partir des mêmes quatre experts gagne +0,0013, et pas un seul n'atteint cet effet d'intérêt.

C'est la recommandation pratique que ce papier soutient : **l'opérateur gratuit n'est plus le plafond — mais la porte qui le bat ne fait pas ce à quoi sert une porte.** Sur les sept entraînements le routage reste proche de l'uniforme (part du plus gros expert de 0,507 à 0,583, là où 0,25 est l'absence totale de préférence sur 4 experts ; entropie patch normalisée de 0,892 à 0,995, là où 1,0 est le maximum ; aucun expert mort dans aucun run). Ce qui produit le gain, c'est d'héberger

quatre spécialistes dans un seul réseau et de les faire démarrer depuis leurs propres checkpoints, pas d'apprendre à choisir entre eux — et le prix honnête est de 300 époques par-dessus quatre entraînements complets, là où un veto ne coûte rien.

Elle vient avec une limite que le papier mesure au lieu de la masquer, et cette limite est la seconde contribution : aucun fold pris isolément ne franchit sa propre différence minimale détectable, si bien que le verdict repose sur un signe constant sur cinq populations disjointes et non sur un test fold par fold ; la meilleure marge qu'un opérateur déterministe extrait de ces quatre experts vaut +0,0013 ; et la même architecture bâtie sur huit blocs aléatoires derrière un seul baseline — la version que celle-ci a remplacée — atterrit à −0,0007, *derrière* l'opérateur gratuit. Une recommandation qui survit à ses propres barres d'erreur vaut mieux qu'une recommandation qui les ignore, et les deux sont rapportées ici.

L'ensembling est le réflexe par défaut de la segmentation d'images médicales. nnU-Net livre l'ensembling par cross-validation en standard (Isensee et al. 2021) ; la fusion d'étiquettes a une littérature longue et bien fondée (Warfield et al. 2004) ; le stacking est plus ancien encore (Wolpert 1992). Ce réflexe est mérité : quand les modèles individuels sont faibles et que leurs erreurs sont décorrélées, les combiner aide, et aide de façon fiable. Ce qui est moins établi, c'est vers quel *type* de combinaison se tourner une fois qu'un backbone moderne a comblé l'essentiel de l'écart — et c'est la question mise au banc d'essai ici, un mécanisme appris et un mécanisme sans entraînement posés sur les mêmes 240 patients, la même métrique et la même correction.

Le cadre est inhabituellement propre. Les papers 1 et 2 de ce programme ont chacun entraîné un modèle MedNeXt-B/nnU-Net sur BraTS-2023 gliome adulte (Baid et al. 2021 ; Roy et al. 2023) avec un terme auxiliaire ajouté à la loss — une tête de régression de distance signée (Ma et al. 2020 ; Xue et al. 2020) et une boundary loss à poids fixe (Kervadec et al. 2019 ; Kervadec et al. 2021) respectivement — et chacun a conclu que ce terme n'améliorait pas les métriques officielles. Ce qu'ils laissent derrière eux, ce sont trois modèles entièrement entraînés qui partagent tout sauf ce terme : mêmes 1196 cas, même découpage en folds, mêmes plans, même calendrier de 300 époques, même graine. Ils sont aussi proches d'un ensemble contrôlé qu'une expérience de ce genre le permet.

Trois stratégies de combinaison sont mises en concurrence. Le **veto déterministe** ne conserve une composante connexe d'un expert que si un autre expert la corrobore quelque part — une règle sans paramètre appris, dont l'attrait est de ne pouvoir que retirer, jamais inventer. Le **vote symétrique** prend chaque voxel à l'unanimité, à la majorité ou à l'union des trois experts et, contrairement au veto, peut produire des formes qu'aucun expert n'a proposées. Le **mélange d'experts** remplace la règle par une porte apprise ; ici la porte vit dans le réseau et s'entraîne avec lui, routant chaque patch 3D vers un sous-ensemble de blocs experts internes, de sorte que rien n'est appris post-hoc au-dessus de prédictions gelées.

La comparaison est volontairement étroite sur un point et large sur un autre. Étroite : seuls le Dice et le HD95 lésion-wise officiels de BraTS-2023 comptent comme résultats (Moawad et al. 2023 ; Saluja et al. 2023). Le Dice voxel-wise que rapportent les journaux d'entraînement est enregistré, mais comme instrument seulement — il affiche 0,910 là où la métrique officielle affiche 0,882 sur la même prédiction, et un papier qui mélange les deux échelles invite une conclusion que le challenge n'endosserait pas. Large : les vingt bras sont rapportés, y compris ceux qui perdent, et y compris un bras témoin qui ne teste rien du tout — le même baseline entraîné une seconde fois avec la même graine. C'est ce témoin qui sert de référence à toute amélioration revendiquée dans ce papier, et il se révèle exigeant.

2. Travaux connexes

Ensembles. La fusion d'étiquettes et l'ensembling pour la segmentation sont établis de longue date : STAPLE estime conjointement un consensus et la fiabilité des annotateurs (Warfield et al. 2004), le stacking généralise l'idée à des combinateurs appris (Wolpert 1992), et le pipeline par défaut de nnU-Net ensemble déjà ses cinq folds de cross-validation (Isensee et al. 2021 ; Isensee et al. 2020). Dans BraTS en particulier, les ensembles d'architectures hétérogènes sont la colonne vertébrale des soumissions gagnantes depuis les premières éditions (Menze et al. 2015 ; Kamnitsas et al. 2018), et le vainqueur 2023 sur le gliome adulte combinait augmentation synthétique et ensemble hétérogène (Ferreira et al. 2024) — diversité d'architecture, non diversité de loss auxiliaire. Deux résultats récents tempèrent le réflexe : un unique réseau plus grand peut reproduire ce que l'on croyait propre à l'ensemble (Abe et al. 2022), et un ensemble ne paie que si le désaccord entre ses membres dépasse leur erreur (Theisen et al. 2023). Les auteurs de nnU-Net eux-mêmes rangent l'ensembling parmi les facteurs confondants qui font paraître meilleures qu'elles ne sont les innovations revendiquées (Isensee et al. 2024).

Mélange d'experts. Le calcul conditionnel par une porte qui partitionne l'entrée entre experts spécialisés remonte à Jacobs et al. (Jacobs et al. 1991) ; la formulation à porte parcimonieuse utilisée ici — routage top-k, bruit gaussien sur

les logits de la porte, terme d'équilibrage de charge — est celle de Shazeer et al. (Shazeer et al. 2017), dont les modes d'effondrement du routage sont documentés à grande échelle (Fedus et al. 2022) et dont la transposition au niveau du patch en vision revient à Riquelme et al. (Riquelme et al. 2021). En imagerie médicale, les modèles MoE de segmentation publiés indexent leurs experts sur quelque chose d'extérieurement hétérogène — modalités manquantes (Liu et al. 2024), jeux de données ou tâches multiples. Un travail qui isole exactement notre question de design, des couches MoE creuses dans un réseau de segmentation *convolutif*, constate une forte sensibilité au placement de la couche (Pavlitska et al. 2026) ; nous testons un placement, pas le principe.

Ce qui manque. Ce qui est comparativement rare dans cette littérature, c'est un rapport de ce que valent ces mécanismes *après* le post-traitement du challenge et *sous* sa métrique lésion-wise (Moawad et al. 2023 ; Kofler et al. 2023), avec le plancher de bruit d'un simple réentraînement mesuré sur les mêmes données — c'est le manque que ce papier comble.

3. Méthodes

3.1 Données et backbone

BraTS-2023 gliome adulte (GLI). Le téléchargement unifié contient 1251 cas ; tous les modèles de cette étude sont entraînés et évalués sur le jeu nnU-Net de 1196 cas qui en est issu, avec le découpage standard nnU-Net en 5 folds, disjoint par construction (aucun patient n'apparaît à la fois dans l'entraînement et la validation d'un même fold). Les 55 cas écartés (4,4 % du jeu source) sont listés un par un — identifiant, motif et date git de l'entrée déclenchante — dans `data/exclusions_1251_to_1196.csv`, livré avec cette archive. L'exclusion est une précaution au niveau patient consignée en mars 2026, en amont de tout entraînement et de toute métrique rapportée ici : 5 des 55 ont été écartés parce que leur propre identifiant figure dans la liste de signalement, 50 parce que leur *patient* a des acquisitions de suivi longitudinal signalées (suffixes -100 à -109) qui ne font pas du tout partie du jeu unifié. Elle est conservatoire et non corrective : la relecture de tous les fichiers des 55 cas écartés face à 55 témoins retenus appariés (graine 42) ne révèle aucun défaut — 0 fichier illisible, 0 NaN, 0 Inf, 0 volume à variance nulle, 0 affine divergent, 0 label hors jeu (`data/integrite_exclus_vs_temoins.json`) —, et les 3 copies survives du jeu donnent une seule empreinte md5 par fichier sur les 10 fichiers comparés des cas signalés par identifiant exact (0 divergence). Aucun chiffre publié ne dépend des cas écartés : toute métrique de ce papier est calculée sur le jeu de 1196 cas. Tous les modèles sont des MedNeXt-B (Roy et al. 2023) entraînés via nnU-Net v2 (Isensee et al. 2021) avec les plans `nnUNetPlans_96GB_mednext` pendant 300 époques sur une seule RTX PRO 6000. La validation est held-out fold par fold : 240 cas au fold 0, 239 à chacun des folds 1 à 4, n = 1196 au total.

3.2 Les trois experts

Les trois experts ne diffèrent que par le terme auxiliaire ajouté à la loss region-based de nnU-Net : **B** (baseline, aucun terme auxiliaire), **D** (tête de régression de transformée de distance signée, λ sélectionné dans le Paper 1), **K** (boundary loss sur les logits, $\lambda = 0{,}05$, sélectionné dans le Paper 2). Un quatrième modèle sert de témoin et n'est pas un expert : **B′**, un second entraînement de B à graine *identique*, qui ne diffère de B que par le non-déterminisme GPU.

3.3 Opérateurs de consensus

Deux familles, de sémantiques différentes, toutes deux sans paramètre.

Un **veto** est asymétrique. `X vetoté par Y` conserve toute composante connexe de X qui partage au moins un voxel avec le masque de Y, et efface les autres entièrement ; une composante n'est jamais ajoutée, seulement retirée, et une composante retirée est repliée sur la région englobante plutôt que percée en trou (WT $\rightarrow$ fond, TC $\rightarrow$ œdème, ET $\rightarrow$ nécrose). Avec deux corroborateurs, la combinaison est soit `ET` (intersection voxel à voxel des deux masques de veto), soit `OU` (union) — à noter que `ET` est l'intersection des masques, et non la conjonction « touche Y quelque part et touche Z quelque part » : les deux lectures divergent dès qu'il y a deux corroborateurs. Les composantes sont calculées en 26-connexité, comme le scorer officiel.

Un **vote** est symétrique. Pour chaque région (WT, TC, ET) et chaque voxel, on compte le nombre d'experts qui incluent ce voxel, et le voxel est conservé si ce compte atteint un seuil : 1 donne l'union, 2 la majorité, 3 l'unanimité. Le volume d'étiquettes final est reconstruit par la cascade imbriquée WT $\rightarrow$ TC $\rightarrow$ ET. Contrairement à un veto, un vote peut produire une forme qu'aucun expert n'a proposée.

3.4 Mélange d'experts patch-wise

Le mélange d'experts est *intra-modèle* : une couche `PatchMoE3D` remplace `dec_block_0` du décodeur MedNeXt. Le volume est découpé en une grille 3×3×3 de patchs, et une porte convolutionnelle route chaque patch vers les 2 meilleurs de **quatre** blocs experts internes, chacun initialisé depuis un expert déjà entraîné seul (ci-dessous). Porte et experts sont entraînés conjointement avec la loss de segmentation, plus un terme d'équilibrage de charge ; rien n'est appris au-dessus de prédictions gelées. C'est une restriction délibérée : un axe de travail antérieur de ce programme entraînait des portes post-hoc sur des probabilités mises en cache d'experts gelés, a produit des classements que la métrique officielle n'a pas confirmés, et a été abandonné.

Sept entraînements de **cette** architecture — le bras que ce papier rapporte — sont mesurés ici : trois graines indépendantes du fold 0 (42, 1337, 2024) et les quatre runs des folds 1 à 4, dont les 239 cas de validation ne partagent aucun patient avec le fold 0. Chacun est un entraînement complet de 300 époques sous les plans de la section 3.2. Trois graines est le minimum à partir duquel la dispersion d'un bras s'énonce au lieu d'être empruntée à un témoin, et c'est la raison pour laquelle elles ont été lancées avant les folds restants.

Ce que la porte change, et le seul chiffre pour lequel une version antérieure est conservée. La couche conserve l'emplacement dans `dec_block_0`, la grille de patchs 3×3×3, le routage top-2, le poids d'équilibrage de charge (0,001), le bruit multiplicatif de porte et son recuit de 40 époques — tout cela fixé dans le source du trainer plutôt que lu de l'environnement. Elle remplace les huit blocs experts initialisés aléatoirement d'une version antérieure par **quatre**, chacun initialisé depuis le dernier bloc `dec_block_0` d'un expert déjà entraîné 300 époques au fold 0 : le baseline B, le bras à distance signée D (λ = 0,1), le bras de frontière K (λ = 0,05) et le bras blob G (β = 0,5). Le reste du réseau — encodeur, autres blocs du décodeur, têtes de supervision profonde — part du checkpoint du baseline, si bien que l'époque 0 segmente déjà au niveau du témoin et que le run entraîne le routage et le raffinement au lieu de la convergence ; la porte elle-même part aléatoire, puisqu'elle est l'objet étudié. Deux conséquences sont à lire avec les chiffres. D'abord, G n'est *pas* l'un des trois experts de la section 3.2, et il joue ici deux rôles que les résultats tiennent séparés. C'est un bras à part entière : entraîné pendant 300 époques sur chacun des cinq folds, avec son terme de distance signée désactivé (λ = 0,0 dans les cinq checkpoints, si bien qu'il ne diffère du baseline B que par le terme blob, β = 0,5), et scoré seul sous le protocole officiel sur les mêmes 239 patients de chaque fold que tout autre bras. Son score propre est le pire des quatre experts denses — −0,0038 Dice sur les cinq folds, négatif sur les cinq, 28e des vingt-neuf bras — tandis que le veto qu'il applique au baseline (`cc_B_G`, l'opérateur à deux experts de la section 3.3) est le meilleur des douze à +0,0011 et le deuxième des vingt-quatre opérateurs de consensus, derrière `cc_B_and_DKG`. Son autre rôle est celui que la couche utilise : la quatrième initialisation de la porte. Ensuite, la porte n'est pas un bras moins cher qu'un veto — elle coûte 300 époques **en plus** de quatre experts déjà entraînés, là où un opérateur de consensus ne coûte rien au-delà de ces mêmes experts. La porte est mesurée sur ces sept runs, chacun un 300 époques complet. La version antérieure à huit blocs a été mesurée sur les sept mêmes runs, et elle n'apparaît qu'une fois dans ce papier, comme dispersion : 0,0034 de Dice entre ses graines contre 0,0013 entre celles de la porte — la mesure qui a justifié de changer l'initialisation. Rien d'autre n'en est tiré, et elle reste figée, octet pour octet, dans `papers/paper3/versions/v2/`.

3.5 Évaluation

Métriques. La capsule officielle BraTS-2023 est appelée telle quelle ; `LesionWise_Score_Dice` et `LesionWise_Score_HD95` sont les endpoints (Saluja et al. 2023). Les lésions sont appariées en dilatant chaque composante de la vérité terrain (3 itérations) puis en fusionnant toutes les composantes prédites qui l'intersectent ; les lésions de vérité terrain de 50 voxels ou moins sont exclues du scoring. Une lésion manquée comme un faux positif contribuent 0 au numérateur du Dice et **374 mm** à celui du HD95, et ajoutent tous deux 1 au dénominateur commun. Une vérité terrain vide avec une prédiction vide vaut 1,0 / 0 mm, et une vérité terrain vide avec une prédiction quelconque vaut 0,0 / 374 mm — un seul cas de ce type pèse 0,0042 de Dice moyen sur 240 patients, soit plus que la plupart des effets discutés ci-dessous.

Post-traitement. Tous les bras reçoivent le même nettoyage officiel avant scoring : les composantes connexes sous 1000 (WT), 250 (TC) et 500 (ET) voxels sont retirées, par région, de la région la plus large vers la plus étroite, une composante retirée retombant sur la région englobante. Sur les 240 cas du fold 0, ce nettoyage vaut +0,048 (B), +0,055 (D) et +0,065 (K) de Dice lésion-wise, et −18,3 à −24,7 mm de HD95, contre l'argmax brut — un ordre de grandeur au-dessus de tout effet de méthode rapporté plus bas. Ce n'est *pas* un résultat de ce papier : c'est le protocole, appliqué à l'identique aux vingt bras, et c'est précisément parce qu'il est appliqué à l'identique que la section 4.4 est intéressante.

Agrégation. Par patient, les trois régions sont moyennées ; par bras, les patients sont moyennés. Les intervalles de confiance sont des bootstraps percentile sur les patients (10 000 rééchantillonnages, graine fixe), les mêmes tirages étant partagés entre régions pour préserver la corrélation intra-patient. Les comparaisons sont appariées par patient : différence moyenne, IC bootstrap de la différence appariée, test de Wilcoxon signé bilatéral, et différence minimale détectable à 80 % de puissance (2,802 × SE de la différence).

Multiplicité et seuil pré-déclaré. La famille primaire est : *chaque bras contre le baseline B, sur les deux endpoints* — 38 tests — avec correction de Holm sur la famille complète. Le bras témoin B′ est dans cette famille : il ne teste aucune hypothèse, il montre où tombe le hasard. Le plus petit effet d'intérêt est déclaré à 0,003 de Dice lésion-wise, et une affirmation exige à la fois p de Holm < 0,05 **et** $|\Delta| \geq 0{,}003$. Tout le reste de ce papier — le régime brut, l'échelle voxel-wise, le détail par région, les comparaisons veto contre son propre primaire — est exploratoire et signalé comme tel.

4. Résultats

4.1 Au fold 0, le bras le mieux classé des vingt-six est un consensus

Vingt-six bras, un jeu de métriques, 240 patients de validation — tous tracés sur les deux endpoints officiels à la figure 1, et détaillés bras par bras au tableau 2. Le **tableau 1** rassemble les trois familles de modèles — experts denses, consensus, mélange d'experts — côte à côte pour la comparaison de performances ; le reste de cette section le déplie. Celui qui arrive en tête est un consensus déterministe à deux experts : **K vetoté par B**, à **0,8877** de Dice lésion-wise et **18,86 mm** de HD95, contre 0,8820 et 20,97 mm pour le baseline qu'il améliore. La marge appariée vaut +0,0056 de Dice, IC bootstrap [+0,0019, +0,0104], et −2,11 mm de HD95, IC [−4,69, −0,10] : sur les deux endpoints officiels, l'intervalle exclut zéro, ce qu'aucune autre famille de cette étude n'obtient sur les deux à la fois. Il passe aussi devant les deux autres experts seuls (D à 0,8832, K seul à 0,8857), devant le témoin de réentraînement (0,8839), et devant la porte apprise à chacune des graines mesurées (section 4.2).

Deux faits encadrent ce chiffre avant que la section 4.6 ne le démonte. Il n'est pas significatif au sens de Holm sur la famille de 38 tests (p de Holm = 0,13), et la marge se décompose en une part due au choix de K comme expert primaire et une part plus petite due au veto. Ce qui survit aux deux observations, c'est le classement lui-même : sur ce fold, sous cette métrique, l'opérateur qui ne coûte aucun entraînement est en haut du tableau, et toutes les alternatives apprises mesurées ici sont en dessous.

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 En face-à-face au fold 0 : le consensus mène sur les moyennes, la porte mène sur les rangs

Comparer chaque bras au baseline répond à « ce bras gagne-t-il quelque chose ». Un praticien qui doit choisir un mécanisme de combinaison a besoin de l'autre comparaison — consensus contre porte, directement, apparié sur les mêmes patients. Le tableau 3 la rapporte contre **chacune des trois graines mesurées** de la porte, et non contre celle qui arrangerait l'argument ; la figure 2 résume le face-à-face et le prix de chaque mécanisme.

K vetoté par B devance les trois graines de la porte publiée sur les **deux** moyennes d'endpoints : +0,0029, +0,0022 et +0,0016 de Dice, −1,64, −2,34 et −2,71 mm de HD95. La marge sur la *meilleure* graine de la porte vaut +0,0016 de Dice et −2,71 mm — en dessous du plus petit effet d'intérêt pré-déclaré par cette étude (0,003 de Dice). Aucun intervalle de Dice du veto n'exclut zéro, et les rangs pointent dans l'autre sens : dans les six comparaisons le consensus améliore moins de patients qu'il n'en dégrade, l'écart le plus large à la graine 2024, où il gagne +0,0016 en moyenne tout en améliorant 88 patients et en en dégradant 151. Après la correction de Holm locale, deux des six survivent (0,006 et 0,025), toutes deux à la graine 2024 et toutes deux en faveur de la porte sur les rangs : c'est le désaccord moyenne/rangs documenté en section 4.4, pas une coquille. Ce que le fold 0 soutient est donc une affirmation sur le *coût*, pas sur le mécanisme — le consensus atteint une moyenne légèrement supérieure sans une étape de gradient, et la porte atteint une moyenne supérieure sur les cinq folds (section 4.5) avec une étape.

L'asymétrie de coût, elle, n'est pas en doute, parce que ce n'est pas une estimation statistique. La porte coûte un entraînement complet de 300 époques par run — 244,4 s par époque sur la RTX PRO 6000 du projet contre 162,1 s pour un expert (+50,8 %), soit **20,4 heures de GPU par run**, sept runs mesurés ici. Un consensus au-dessus d'experts qui existent déjà coûte **zéro** heure de GPU et quelques minutes de CPU. Le consensus n'a non plus **aucune dispersion de réentraînement** : c'est une fonction déterministe de ses entrées, donc son 0,8877 n'est pas un tirage dans une distribution. Les trois graines de la porte s'étalent bien sur 0,0013 de Dice (section 4.5) — mais cette étendue vaut 0,96 fois les +0,0013 qui séparent deux runs de *son propre* baseline à graine identique, et les trois deltas gardent un seul signe (+0,0027, +0,0035, +0,0040). La dispersion reste un argument de coût contre la porte ; ce n'est plus un argument de signe.

Le déploiement — l'axe habituellement concédé à la porte — est plus proche qu'il n'y paraît. La porte répond en une seule passe avant et un seul checkpoint, là où un consensus à deux experts exécute deux réseaux et un vote à trois experts en exécute trois ; sur une station mono-GPU, c'est une vraie économie. Mais les passes du consensus sont *indépendantes* : sur un serveur à deux ou trois GPU, elles s'exécutent en parallèle, et la latence mur retombe à celle d'une seule passe dense — sous celle de la porte, dont l'entraînement coûte 244,4 s par époque contre 162,1 s pour un expert dense (+50,8 %). Les checkpoints ont presque la même taille (mesuré sur disque : 126,8 Mo par expert, 127,2 Mo pour la porte, +0,36 %), mais le consensus les répartit entre des modèles séparés, déployables individuellement : chaque expert peut occuper son propre GPU — un petit suffit —, être mis à jour, canarié ou annulé seul, et la défaillance d'un expert dégrade le service au lieu de l'abattre. La porte est un monolithe : un seul dispositif doit l'héberger, et un réentraînement déplace l'ensemble. Sur une station mono-GPU, la passe unique de la porte gagne encore ; dans un service d'inférence de production avec plus d'un GPU — le cas général — l'avantage change de camp.

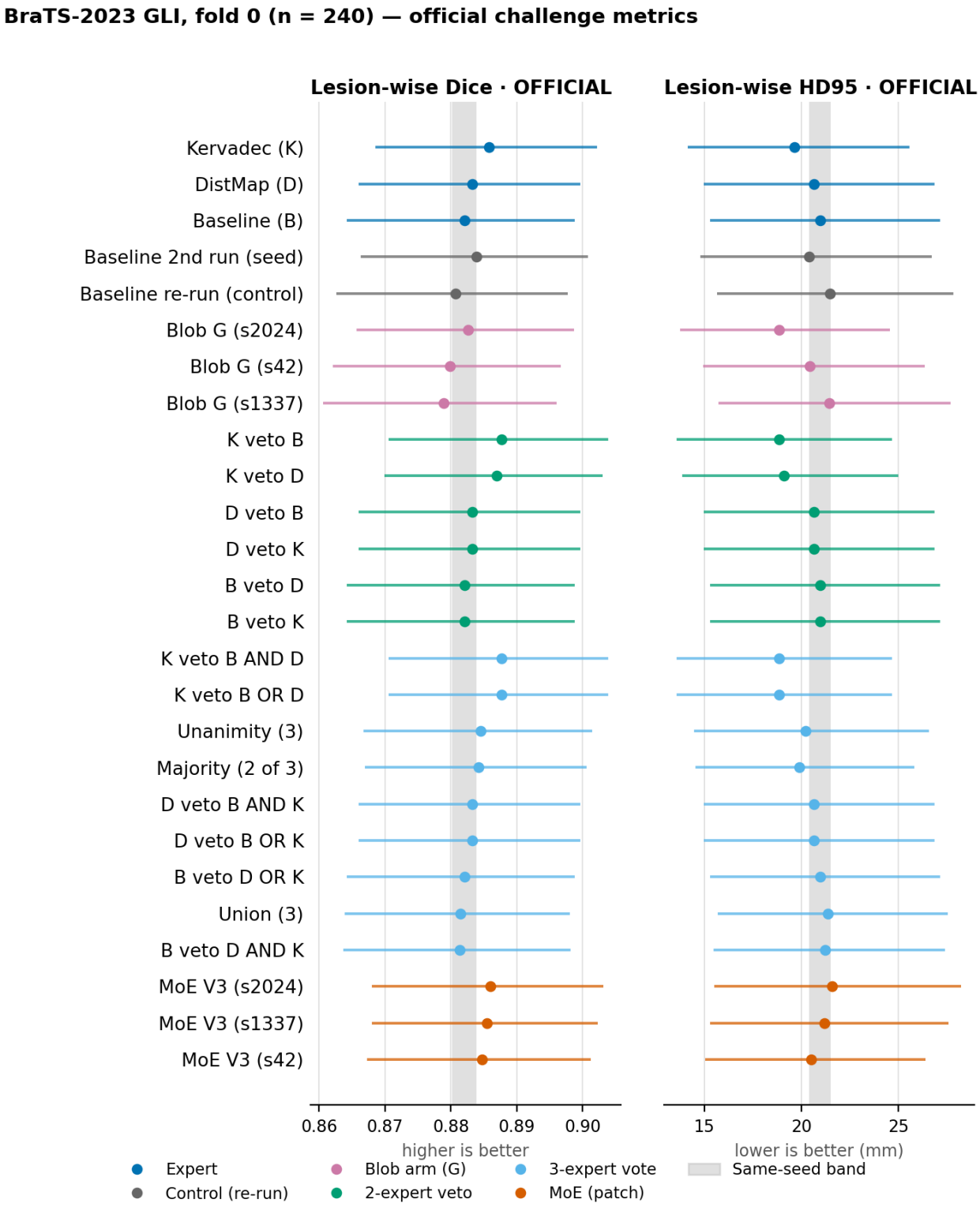

Figure 1: Figure 1 — Les vingt-six bras mesurés sur le fold 0 sur les deux endpoints officiels, chaque moyenne avec son IC bootstrap, et la bande grise donnant la distance entre deux entraînements du même modèle à la même graine.

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.

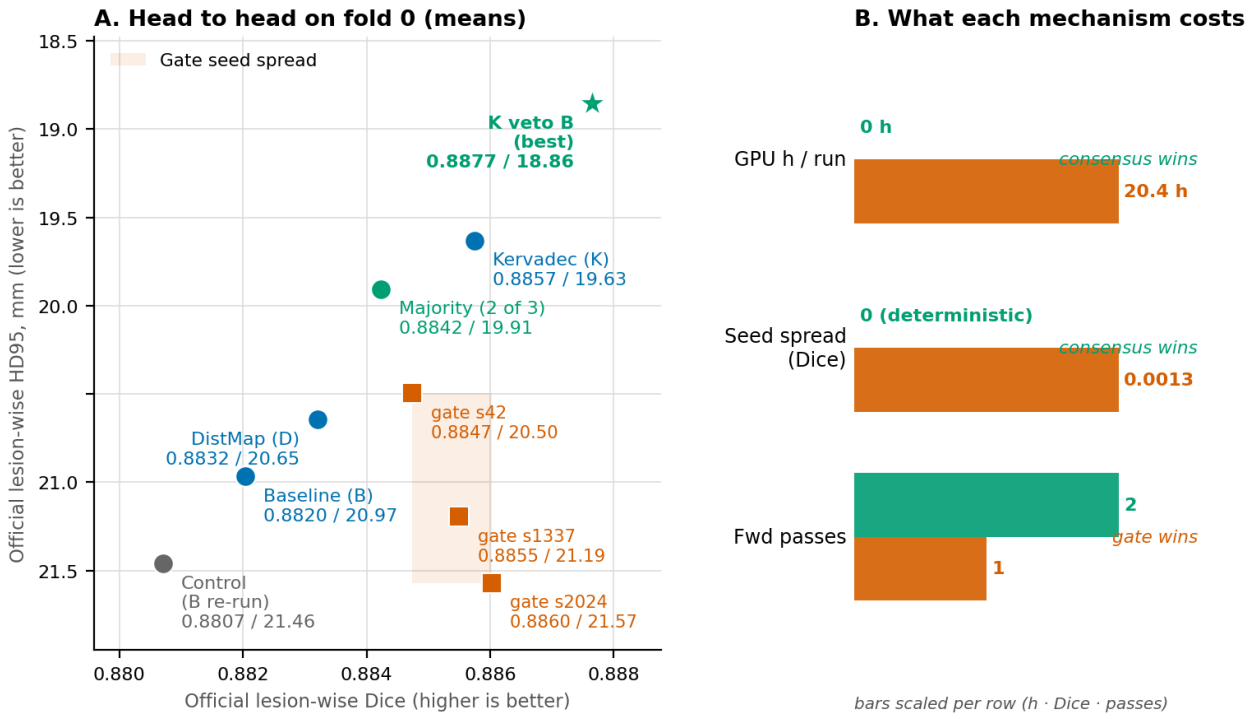

Figure 2: Figure 2 — Le résumé du face-à-face : le meilleur consensus contre les trois graines de la porte V3 sur les deux endpoints officiels — il mène toutes les moyennes, mais aucune de ses 3 marges Dice n'a d'IC bootstrap excluant zéro (A) — et le prix de chaque mécanisme, où l'entraînement et la dispersion favorisent le consensus et où le service favorise la porte (B).

4.3 Améliorer les deux critères à la fois n'est pas ce qui sépare les mécanismes

Une défense de la porte apprise reste disponible à ce stade : peut-être achète-t-elle une amélioration *équilibrée* — un peu de Dice et un peu de HD95 ensemble — là où un consensus échangerait l'un contre l'autre. Une segmentation n'étant cliniquement utile que si le recouvrement et la distance au contour sont tous deux acceptables, un bras qui déplace les deux mériterait crédit même à faible marge sur chacun. Le tableau 4 et la figure 3 mettent cette défense à l'épreuve sous les trois formes qu'elle peut prendre ; aucune ne la soutient.

La prémisse ne tient pas : rien ici n'échange un critère contre l'autre. À l'intérieur d'un patient, le gain Dice et le gain HD95 d'un bras corrèlent entre 0,51 et 0,65 (Spearman) — pour *tous* les bras de l'étude, porte comme consensus, tous à $p < 1e\text{-}16$. Le mécanisme n'a rien de subtil : retirer une lésion parasite fait monter le Dice lésion-wise et baisser le HD95 lésion-wise du même geste, les deux métriques étant calculées par lésion après le même appariement de composantes. Dans la figure 3A, chaque bras se place donc près de la diagonale. Améliorer les deux critères à la fois est une propriété du couple de métriques sur cette tâche, pas un mérite qui distingue un mécanisme de combinaison.

En amplitude, la porte n'atteint pas son propre témoin. Exprimer chaque delta en unités du bruit de réentraînement mesuré rend les deux axes comparables sans pondération arbitraire, et retenir le *plus faible* des deux — z(min), le critère d'équilibre au sens strict — classe les bras. *K vetoté par B* atteint z(min) = +3,02. La porte atteint +0,12 à la graine 42, +1,43 à la graine 1337 et **−1,33** à la graine 2024 — où elle dégrade les deux critères. Le baseline réentraîné à graine identique, qui n'améliore rien par construction, atteint +0,99 : deux des trois graines de la porte sont sous leur propre témoin, et la dispersion entre graines est plus large que l'effet revendiqué.

En fréquence, la porte tombe sur zéro. Compté par patient, sans hypothèse d'échelle, *K vetoté par B* améliore les deux critères chez 78 patients sur 240 et les dégrade tous deux chez 43 (solde +0,146, IC bootstrap [+0,054, +0,233]) ; le vote majoritaire est à +0,179 [+0,100, +0,258]. La porte à la graine 42 améliore les deux chez 55 patients et les dégrade chez 56 — un solde de −0,004, IC [−0,092, +0,079], centré sur zéro. Le témoin est à +0,037, là encore au-dessus de la porte.

Sous un test intersection-union unilatéral — la construction standard pour « améliore les deux », où le plus grand des deux p reste valide sans correction supplémentaire (Berger 1982) — aucune revendication conjointe ne survit à Holm entre bras, sauf l'unanimité (p de Holm 0,034), et les trois graines de la porte sont à 1,000.

Deux bras montrent que fréquence et amplitude répondent à des questions différentes. L'union des trois experts améliore les deux critères chez 96 patients contre 40 — le plus fort décompte conjoint de l'étude — alors que sa *moyenne* recule sur les deux (−0,0006 Dice, +0,41 mm) : elle gagne souvent peu et perd rarement beaucoup. L'unanimité fait exactement l'inverse (+0,0025 Dice et −0,76 mm en moyenne, pour 31 patients améliorés sur les deux contre 71 dégradés). Le rank score de BraTS, qui agrège Dice et HD95 en moyennant des rangs sur (cas × région × métrique), reproduit le même angle mort : il place l'union deuxième du pool alors qu'elle recule en moyenne, et il contredit la moyenne sur trois bras au total. C'est pourquoi le tableau 4 rapporte fréquence, amplitude et rang côte à côte au lieu de les fondre en un composite unique — un seul nombre aurait masqué un désaccord qui est lui-même un résultat. C'est aussi pourquoi le rank score, bien qu'il soit la langue du challenge, ne tranche rien ici : un rang est relatif au pool sur lequel il est calculé, et les classements de ce type sont connus pour être instables à ce genre de choix (Maier-Hein et al. 2018).

La conclusion de cette section est négative et elle coupe des deux côtés : l'argument de l'amélioration équilibrée ne sauve pas la porte, et il n'est pas nécessaire au consensus, dont le dossier repose sur les marges des sections 4.1 et 4.2.

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.

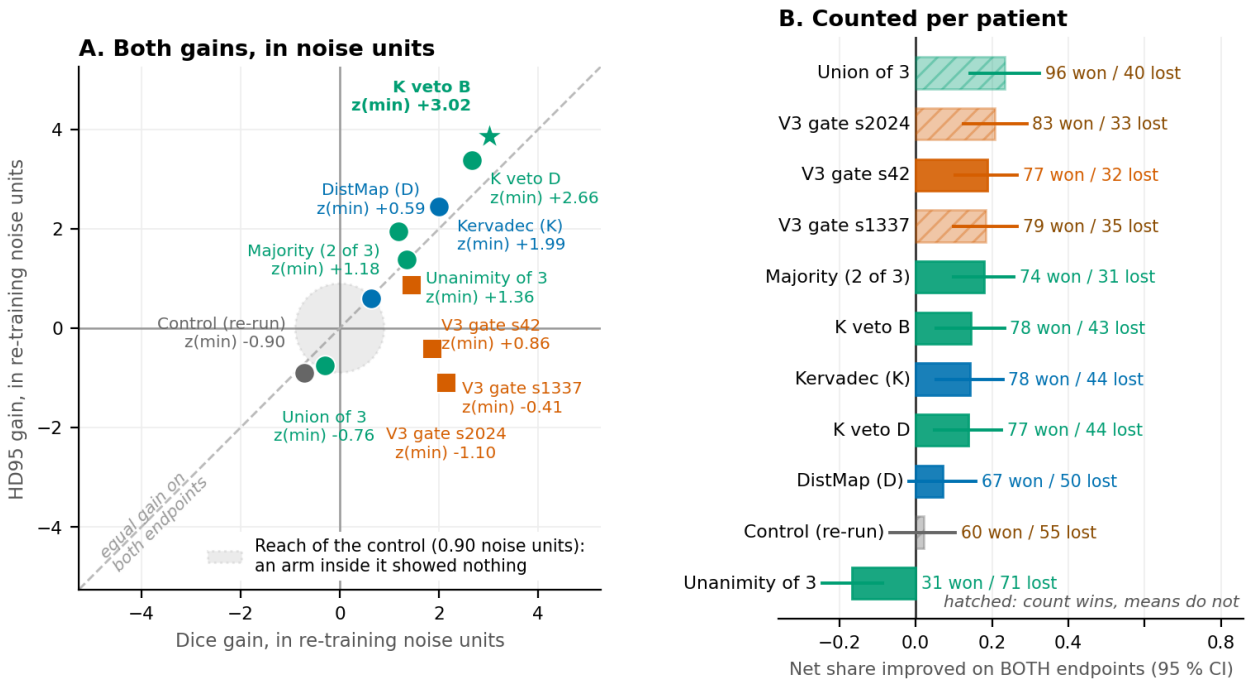

Figure 3: Figure 3 — Les deux endpoints à la fois : les deux gains de chaque bras en unités de bruit de réentraînement, la portée du témoin tracée en disque (A), et la même question comptée patient par patient (B).

4.4 Le seul gain qui survit à la correction de multiplicité est un vote

Contrairement aux vetos (section 4.6), les trois votes symétriques ne sont pas absorbés par le post-traitement officiel : la majorité marque 0,8842, l'unanimité 0,8846, l'union 0,8815 contre 0,8820 pour le baseline. Ce sont aussi les seuls bras à franchir la correction de Holm — et un seul des trois peut se lire comme une amélioration.

La majorité (2 sur 3) est celui-là. Elle gagne +0,0022 de Dice avec un p de Holm de $9{,}0 \times 10^{-8}$, en améliorant 171 patients sur 240 contre 68 dégradés, et −1,06 mm de HD95 (IC [−3,04, −0,04]). Moyenne et test de rangs pointent dans le même sens, sur les deux endpoints, après correction sur 38 tests : rien d'autre dans ce papier ne fait cela, et en particulier pas le mélange d'experts (p de Holm = 1,00). C'est le résultat le plus net de l'étude, et il vient d'un comptage de voix.

Sa taille doit être énoncée avec lui : +0,0022 est en dessous du plus petit effet d'intérêt pré-déclaré (0,003) et comparable aux 0,0018 qui séparent le baseline d'une copie réentraînée de lui-même. La lecture juste est « reproductible mais petit », pas « grand ».

Ce verdict ne s'étend pas à la porte que ce papier rapporte. Δ (vote − porte) sur le Dice contre les trois graines de la porte, tel que publié au tableau 3 : −0,0005, −0,0013, −0,0018 — le vote est derrière aux trois, sur la moyenne des trois, et à chaque graine du bras que l'étude rapporte réellement. Le gain qui survit à la correction de Holm est donc le gain d'un vote sur des experts gelés, pas sur le combinateur appris (tableau 3).

Les deux autres votes sont une mise en garde, pas un résultat. L'unanimité gagne +0,0025 de Dice moyen avec un p de Holm de $5{,}6 \times 10^{-3}$ tout en rendant **149 patients sur 240 moins bons et 90 seulement meilleurs**. L'union perd −0,0006 en moyenne avec un p de Holm de $2{,}4 \times 10^{-2}$ tout en rendant **146 patients meilleurs et 93 moins bons**. Dans les deux cas, le signe de la moyenne et la direction du test de rangs se contredisent, parce que la métrique lésion-wise est dominée par un petit nombre de cas catastrophiques : une lésion manquée contribue une pénalité de 374 mm en HD95 et tire la moyenne bien plus qu'une centaine de petites améliorations ne la relèvent.

Le point méthodologique dépasse ce papier : sur une métrique lésion-wise, une moyenne et un test de rangs signés répondent à deux questions différentes, et un tableau qui n'en rapporte qu'une peut se lire dans les deux sens. Les deux figurent au tableau 5.

Une objection à la correction elle-même mérite réponse, parce qu'elle jouerait en notre faveur et qu'elle ne tient pas. La famille corrige sur 38 tests, mais **neuf des vingt bras sont des copies exactes d'un autre bras** — identiques voxel à voxel sur les 240 patients, et pas seulement égaux en moyenne (section 4.6). **Onze hypothèses seulement sont distinctes** : neuf de ces tests n'apportent aucune information et coûtent de la puissance. Refaire Holm sur les onze représentants — 20 tests au lieu de 38 — divise à peu près par deux les p ajustées et **ne fait basculer aucun verdict** : les quatre mêmes tests passent 0,05, les trois votes et rien d'autre. `cc_K_B` passe de 0,132 à 0,060 sur le Dice et reste du mauvais côté du seuil ; le témoin de réentraînement a, sur la même cohorte, un p brut de 0,032, et c'est à cette échelle-là qu'il faut lire ce 0,060. La famille primaire annoncée reste la complète, puisqu'elle a été fixée avant la mesure ; la famille restreinte est rapportée pour que la perte de puissance se voie au lieu de se discuter. Le critère de regroupement ne regarde que les prédictions, jamais un p, et aucun test ne disparaît — un doublon a par construction le même p brut que son représentant.

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 La porte initialisée depuis ses experts : 0,0013 entre graines, cinq folds d'un signe

constant

Le mélange d'experts patch-wise atteint 0,8847 de Dice lésion-wise et 20,50 mm de HD95 au fold 0, contre 0,8820 et 20,97 mm pour le baseline dont il est issu — Δ = +0,0027 de Dice avec un p de Wilcoxon signé brut de $3{,}5 \times 10^{-8}$ (166 patients améliorés, 73 dégradés) et Δ = −0,47 mm de HD95 ($p = 5{,}9 \times 10^{-4}$). Sa différence minimale détectable sur cette population vaut 0,0054, deux fois la différence observée, si bien qu'aucun fold pris isolément n'établit son propre effet : ce qui porte le résultat, c'est la répétition — sur les graines et sur des populations disjointes (figure 4).

Une lecture de cette première ligne doit être énoncée avant qu'on la cite. Son intervalle de confiance bootstrap, −0,0008 à +0,0068, contient zéro alors que le p du test de rangs signés vaut $3{,}5 \times 10^{-8}$. Ce n'est pas une coquille et les deux nombres ne se contredisent pas : l'intervalle est un bootstrap de la *moyenne*, que la queue lourde d'une métrique lésion-wise élargit — une lésion manquée vaut 374 mm — tandis que le p est un test sur les *rangs*, que cette queue ne déplace

pas. Une moyenne et un test de rangs signés répondent à deux questions différentes, et l'endpoint primaire pré-déclaré est scoré par les deux (section 4.4).

Trois graines au lieu d'une (tableau 6). Un run unique ne peut pas séparer un effet du bruit de son entraînement : la porte a donc été réentraînée aux graines 1337 et 2024. Les trois runs tombent à 0,8847, 0,8855 et 0,8860, une étendue de **0,0013 de Dice** (écart-type 0,0006) — 0,96 fois les +0,0013 qui séparent deux runs de *son propre* baseline à graine identique, c'est-à-dire *plus étroite* que le bruit auquel on la compare. Les trois deltas gardent un seul signe (+0,0027, +0,0035, +0,0040) et le meilleur d'entre eux (+0,0040) est plus grand que l'étendue. Moyennée sur les trois graines, la porte est à +0,0034 au-dessus du baseline, au plus petit effet d'intérêt pré-déclaré de 0,003. Sur la métrique de frontière, les trois runs s'étalent sur 1,07 mm de HD95. C'est la quantité qu'un consensus n'a pas : un opérateur déterministe appliqué aux mêmes prédictions rend le même nombre à chaque fois.

Quatre populations disjointes (tableau 7). Les folds 1 à 4 ne partagent aucun patient avec le fold 0, 239 cas de validation chacun. Fold par fold, contre son propre baseline : +0,0082, +0,0045, +0,0092 et +0,0025 ; l'effet moyen sur les quatre folds vaut +0,0061, avec quatre folds favorables sur quatre et un signe constant. Sur les cinq folds à référence homogène la moyenne vaut +0,0054, cinq sur cinq favorables et trois sur cinq au-dessus du 0,003 pré-déclaré. L'étalement de l'effet d'un fold à l'autre (0,0067) est plus large que l'étalement de graine du fold 0, comme on s'y attend quand la population change : c'est une propriété des folds, pas de l'entraînement.

Le comparateur décide toujours du verdict. Au fold 1 le meilleur expert seul est **DistMap (D)**, à 0,8750 de Dice et 21,38 mm — un modèle unique ordinaire avec un terme de régression auxiliaire et aucune porte, et ce n'est pas l'expert dont la porte est issue. Contre lui la porte gagne +0,0063 de Dice et −0,87 mm de HD95, si bien que, contrairement à une marge qui ne tiendrait que contre le baseline, celle-ci survit au changement de comparateur sur les deux endpoints. Ce qui ne change *pas*, c'est la puissance d'un fold unique — chaque delta reste sous sa propre MDE (tableau 6, tableau 7) — si bien que l'affirmation repose sur un signe constant à travers des populations disjointes et sur une dispersion plus étroite que son propre témoin, pas sur un test par fold significatif.

Où la V3 se place parmi les vingt-neuf bras. Les sections 4.1 à 4.4 comparent des bras mesurés sur le seul fold 0, dans la phase antérieure de cette étude. Re-scorer l'ensemble — les quatre experts denses seuls, les vingt-quatre opérateurs de consensus déterministes et les deux portes — par un seul chemin de code sous le protocole officiel sur les mêmes 239 patients de chacun des cinq folds déplace le verdict (`scripts/paper3_v3_vs_consensus.py`, qui reproduit le delta publié du fold 0 de la V3 à 0,00e+00 avant d'écrire quoi que ce soit). La V3 est première des vingt-neuf à +0,0057 sur son propre baseline, positive sur cinq folds sur cinq et au-dessus du 0,003 pré-déclaré sur quatre ; le meilleur consensus est `cc_B_and_DKG` à +0,0013 ; la porte à huit blocs aléatoires que celle-ci a remplacée est 16e à −0,0007, c'est-à-dire *derrière l'opérateur gratuit* ; et l'expert à blob loss pris seul est 28e des vingt-neuf à −0,0038, négatif sur les cinq folds, tandis que le veto qu'il applique au baseline (`cc_B_G`) est le meilleur des douze vetoes à deux experts à +0,0011 et le deuxième des vingt-quatre opérateurs de consensus, derrière `cc_B_and_DKG` (+0,0013). Appariés directement, patient par patient et sans passer par aucune référence, les vingt-quatre opérateurs de consensus ont tous une marge moyenne négative contre la V3 — 21 d'entre eux derrière sur cinq folds sur cinq, les trois exceptions (cc_K_B, cc_K_D, cc_K_and_BDG) sur quatre. L'écart entre la V3 et le meilleur opérateur gratuit vaut +0,0044, soit 1,5 fois l'effet d'intérêt pré-déclaré, et aucun opérateur de consensus n'atteint cet effet d'intérêt en moyenne sur les folds (0 sur vingt-quatre) là où la V3 l'atteint sur quatre folds sur cinq. La référence du fold 0 elle-même dérive de −0,0032 entre campagnes de scoring, raison pour laquelle cette comparaison fait passer les vingt-neuf bras par un seul baseline par fold au lieu de réutiliser des deltas publiés.

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.

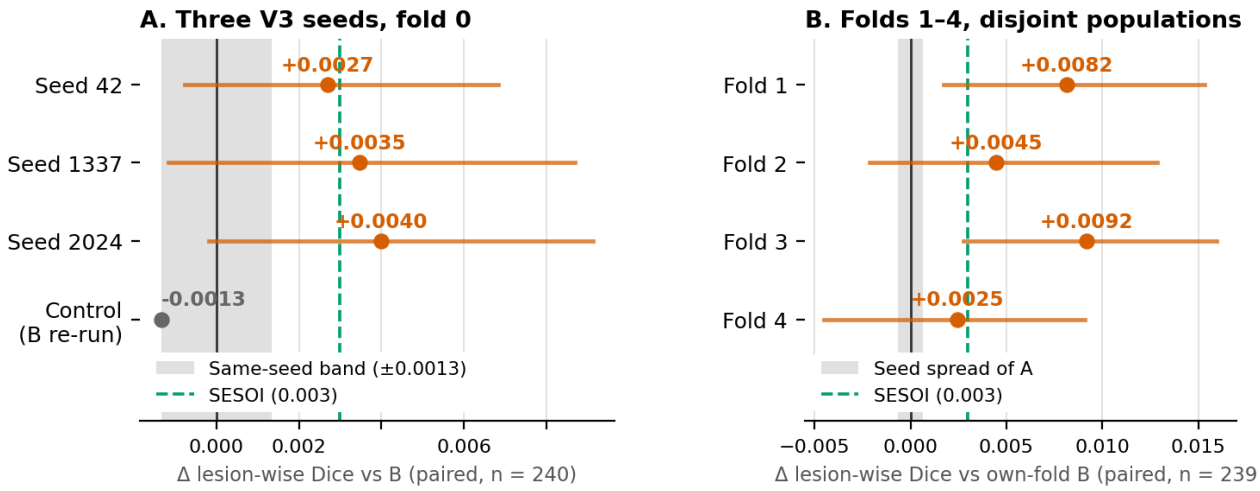

Figure 4: Figure 4 — Ce qu'une graine unique du fold 0 ne pouvait pas montrer : les trois graines de la porte V3, entraînées dans une même campagne, leur dispersion face à la bande du témoin de réentraînement, le signe tenant aux trois (A), et les quatre folds disjoints 1 à 4, où le signe tient à chaque fold et la porte devance le meilleur expert de ce fold dans les 4 (B).

4.6 Ce que la victoire du consensus doit à quoi

Les +0,0056 de la section 4.1 méritent d'être démontés, parce que les deux tiers ne relèvent pas du mécanisme de consensus. **+0,0037 viennent de l'emploi de K plutôt que B comme expert primaire, et +0,0019 du veto lui-même.** La plus grande moitié ne survit pas à la cross-validation, où K se situe 0,0009 *en dessous* de B (section 4.7) ; la plus petite est portée par l'unique patient sur 240 que ce veto modifie.

C'est le schéma général des vetos sous ce protocole, et il se mesure directement. Douze bras de veto sont comparés à **leur propre expert primaire**, sur les mêmes prédictions, de sorte que le bruit de réentraînement n'y joue aucun rôle. Après le nettoyage officiel, sept des douze changent le score de **zéro** patient sur 240 ; quatre en changent un ; un en change deux. Les différences moyennes valent +0,0000 pour tout veto dont le primaire est D, et pour trois des quatre dont le primaire est B ; les trois vetos amorcés sur K gagnent +0,0012 à +0,0019, et `B vetoté par D ET K` *perd* 0,0007.

Avant le nettoyage, ces mêmes vetos valent +0,0079 à +0,0375 de Dice lésion-wise (tableau 8 ; la figure 5 montre l'absorption bras par bras). Le veto fait un vrai travail — simplement, c'est le travail que le protocole officiel a déjà fait. Les deux mécanismes suppriment de petites composantes non soutenues ; les seuils de taille officiels (1000/250/500 voxels) sont assez grossiers pour attraper à peu près tout ce qu'un expert corroborant aurait rejeté, parce que les composantes que les vetos retirent pèsent des dizaines à des centaines de voxels. Les deux patients où un veto change encore quelque chose sont instructifs en sens opposés : chez l'un, le veto retire un vrai faux positif ; chez l'autre, il retire une vraie lésion.

Cela ne rétracte pas la section 4.1 — le bras en haut du tableau reste un consensus, et il reste gratuit — mais cela dit ce qu'un lecteur doit s'attendre à reproduire : sous un protocole qui applique déjà un filtre de taille, un veto ajoute une erreur d'arrondi, tandis qu'un *vote*, qui peut construire des formes qu'aucun expert n'a proposées, reste le mécanisme ayant encore quelque chose à apporter (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

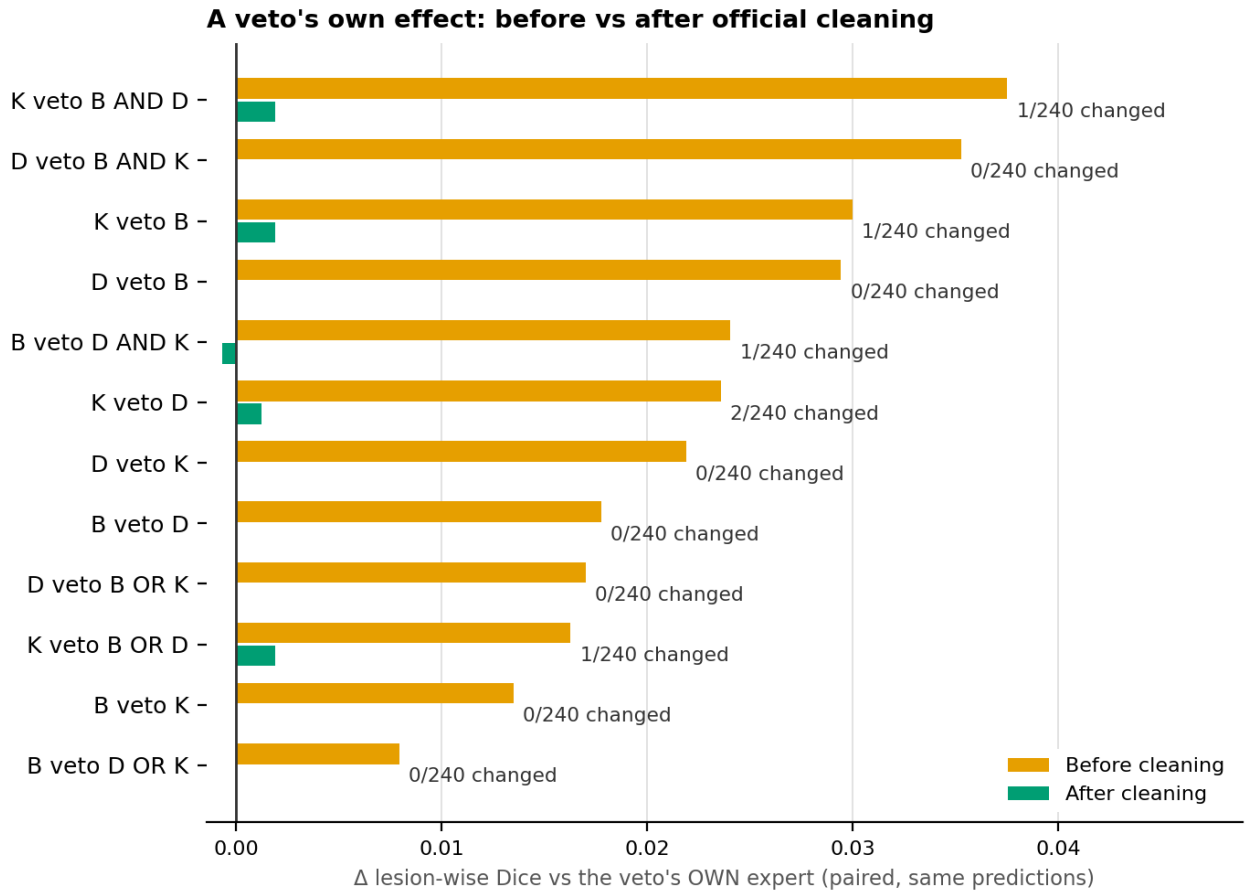

Figure 5: Figure 5 — L'effet d'un veto avant et après le nettoyage officiel, bras par bras.

4.7 Pourquoi le plafond est bas : trois experts sont un modèle mesuré trois fois

Au fold 0, les trois experts se classent K (0,8857) > D (0,8832) > B (0,8820), et K − B vaut +0,0037 avec un p de Wilcoxon brut de $6{,}6 \times 10^{-3}$. Sur la cross-validation complète, le classement **s'inverse** : B 0,8774, D 0,8768, K 0,8765, avec K − B = −0,0009 (IC 95 % [−0,0033, +0,0015], p = 0,12) pour une différence minimale détectable de 0,0034 (tableau 9). En HD95 le même schéma tient — 21,92, 21,84 et 22,28 mm, étendue 0,44 mm. Quatre entraînements indépendants du fold 0 donnent un écart-type inter-graines de 0,0019 de Dice (0,0007 pour B, 0,0024 pour D et K), et deux runs du *même* modèle à la *même* graine diffèrent de 0,0018. L'étendue entre les trois experts sur 1196 cas vaut 0,0009 — la moitié de la distance entre un modèle et lui-même.

Voilà pourquoi tous les effets de ce papier sont petits, et c'était prévisible avant toute combinaison. Un ensemble ne paie que si le désaccord entre ses membres dépasse leur erreur (Theisen et al. 2023). Les deux quantités se mesurent directement sur les mêmes scores par patient. L'erreur moyenne des trois experts vaut 0,118, 0,117 et 0,114 (1 − Dice lésion-wise). Leur désaccord moyen par paire — la différence absolue moyenne entre les scores de deux experts sur le même patient — vaut 0,0076 (B–D), 0,0101 (B–K) et 0,0112 (D–K). **Le désaccord vaut 8,3 % de l'erreur.** Les scores par patient corrèlent de 0,93 à 0,97 en Pearson : quand un expert échoue, les autres échouent avec lui. Sur les 28 patients qui marquent sous 0,70 pour au moins un expert, **22 marquent sous 0,70 pour les trois**.

La conséquence la plus forte est une borne supérieure. Un oracle omniscient qui choisirait, pour chaque patient, celui des trois experts qui se trouve être le meilleur marquerait 0,8909 — soit seulement **+0,0051** au-dessus du meilleur expert seul. C'est le plafond de toute stratégie qui sélectionne parmi ces trois experts, il est inatteignable en pratique, et il ne vaut que 2,8 fois la distance entre un modèle et une copie réentraînée de lui-même. Les votes symétriques n'y sont pas soumis, puisqu'ils peuvent synthétiser des formes qu'aucun expert n'a proposées, ce qui est cohérent avec le fait que le

vote majoritaire soit le seul bras à franchir la correction.

Lus contre ce plafond, les résultats de consensus ne sont pas décevants, ils sont proches de ce qui était disponible : +0,0056 observés là où +0,0051 était le plafond de la sélection et où 8,3 % de désaccord relatif était la matière première. Les trois experts ont été entraînés sur les mêmes données, avec le même backbone et les mêmes hyper-paramètres, ne différant que par un terme auxiliaire de loss que les papers 1 et 2 ont chacun trouvé nul. Attendre un grand effet de leur combinaison n'a jamais été raisonnable ; ce que fait cette section, c'est remplacer cette intuition par un nombre.

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 Ce que la moyenne ne dit pas : la distribution et sa queue

Tous les chiffres ci-dessus sont des moyennes sur 240 patients. Un outil clinique ne s'emploie pas sur une moyenne, il s'emploie sur un patient — et sur cette métrique les deux sont loin l'un de l'autre. Pour le baseline, le patient **médian** obtient 0,9363 là où la moyenne vaut 0,8820, et le HD95 médian vaut 2,17 mm là où la moyenne vaut 20,97 mm : 18,8 mm séparent le cas typique du cas moyen. La moyenne d'un score officiel BraTS est un énoncé sur la queue de distribution, pas sur le patient qu'on a devant soi.

Cette queue est courte et lourde. La moyenne des 10 % pires cas ($CVaR_{10}$, n = 24) vaut 0,5294 en Dice et 149 mm en HD95 pour le baseline, et le 90e percentile du HD95 est déjà à 75 mm. Derrière ces nombres se tiennent **sept patients sur 240 pour lesquels aucune lésion rehaussée n'est appariée**, chacun crédité de la sentinelle de 374 mm que le scorer officiel attribue quand l'appariement échoue. Ces sept cas font à eux seuls **42,5 %** de la moyenne HD95 sur ET : les retirer la ramène de 24,70 à 14,20 mm. Sur les 720 couples patient-région, trois autres sont des faux positifs purs contre une vérité terrain vide, un n'apparie aucune lésion dans un sens ni dans l'autre, neuf ont une vérité terrain vide correctement prédite vide, et les 700 restants ont au moins une lésion appariée.

Compter les échecs rend le choix d'agrégation visible. Au seuil Dice < 0,5, le baseline échoue sur **9 patients sur 240 (3,8 %)** quand on moyenne les trois régions, et sur **43 sur 240 (17,9 %)** quand le patient est jugé sur sa **pire** région. Le second chiffre est la lecture clinique — un patient dont la tumeur rehaussée est manquée est un échec même si WT et TC sont parfaits — et il vaut 4,8 fois le premier. Moyenner trois régions emboîtées divise par trois un échec sur une seule région.

Sur ce fond, un bras et un seul se détache : `cc_K_B` (avec ses deux copies exactes) **domine stochastiquement** le baseline : les deux CDF ne se croisent jamais, et `cc_K_B` est devant sur 94 % de la grille de scores, le plus grand écart de CDF valant 0,033 — huit patients sur 240 au point le plus large. Il est devant ou à égalité partout, jamais derrière. Son avantage est de surcroît concentré là où il compte : +0,0191 sur le $CVaR_{10}$ contre +0,0056 sur la moyenne, un facteur 3,4. Ce qu'il ne fait pas, c'est sauver quelqu'un — un test de McNemar sur le critère dur ne trouve **aucun patient discordant** à aucun des trois seuils (0,5, 0,7, 0,8). Le veto relève des cas déjà mauvais sans en faire repasser un seul au-dessus de la ligne d'échec. Ces deux faits sont dans la figure 6 ; seul le premier est dans le tableau 2. Trois patients épinglés du viewer 3D compagnon donnent un visage à ce niveau individuel : un cas où la porte finit devant le vote et le consensus à trois experts de sa planche (figure 7), un cas où elle tient simplement le niveau du meilleur expert seul (figure 8), et un cas où elle s'effondre sous le baseline, chaque expert, chaque veto et chaque vote (figure 9).

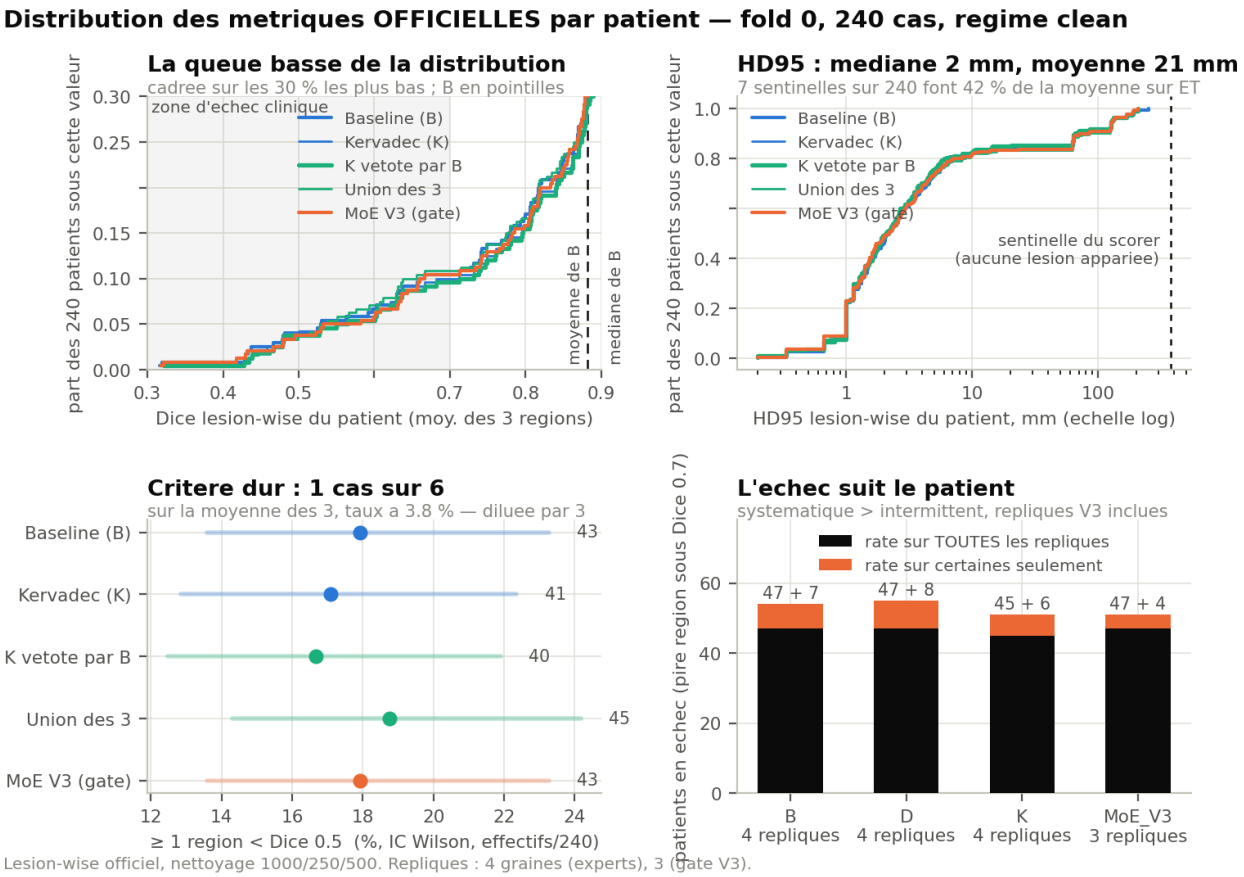

Figure 6: Figure 6 — Ce que la moyenne ne dit pas : les distributions par patient et leur queue. CDF du Dice cadrées sur la queue (a), CDF du HD95 en échelle log (b), taux d'échec sur le critère dur — la pire région du patient — avec IC de Wilson (c), et un échec change-t-il de patient d'une graine à l'autre (d).

Les trois figures suivantes sont des captures du viewer 3D compagnon (https://guillaume-cassez.fr/imagerie-medicale/brats/2023-distance-map/viewer/) en régime officiel — fond blanc, nettoyage de composantes du barème BraTS-2023, couleurs de régions du viewer, vue sagittale, cerveau masqué. Chaque planche montre, pour un même patient et une même pose de caméra : vérité terrain / consensus déterministe à 3 experts / vote majoritaire / MoE patch-wise V3. Les planches ont été re-capturées sur la porte V3 le 2026-09-21, et chaque chiffre de leurs légendes est un Dice lésion-wise officiel (moyenne des 3 régions, régime clean, fold 0), calculé pour les 42 bras de chaque planche par `scripts/paper3_planches_v3.py` depuis les CSV du scorer officiel BraTS-2023 (seuils de composantes 1000/250/500) et écrit dans `papers/paper3/tables/v3_planches.json` — aucun chiffre de légende n'est tapé à la main.

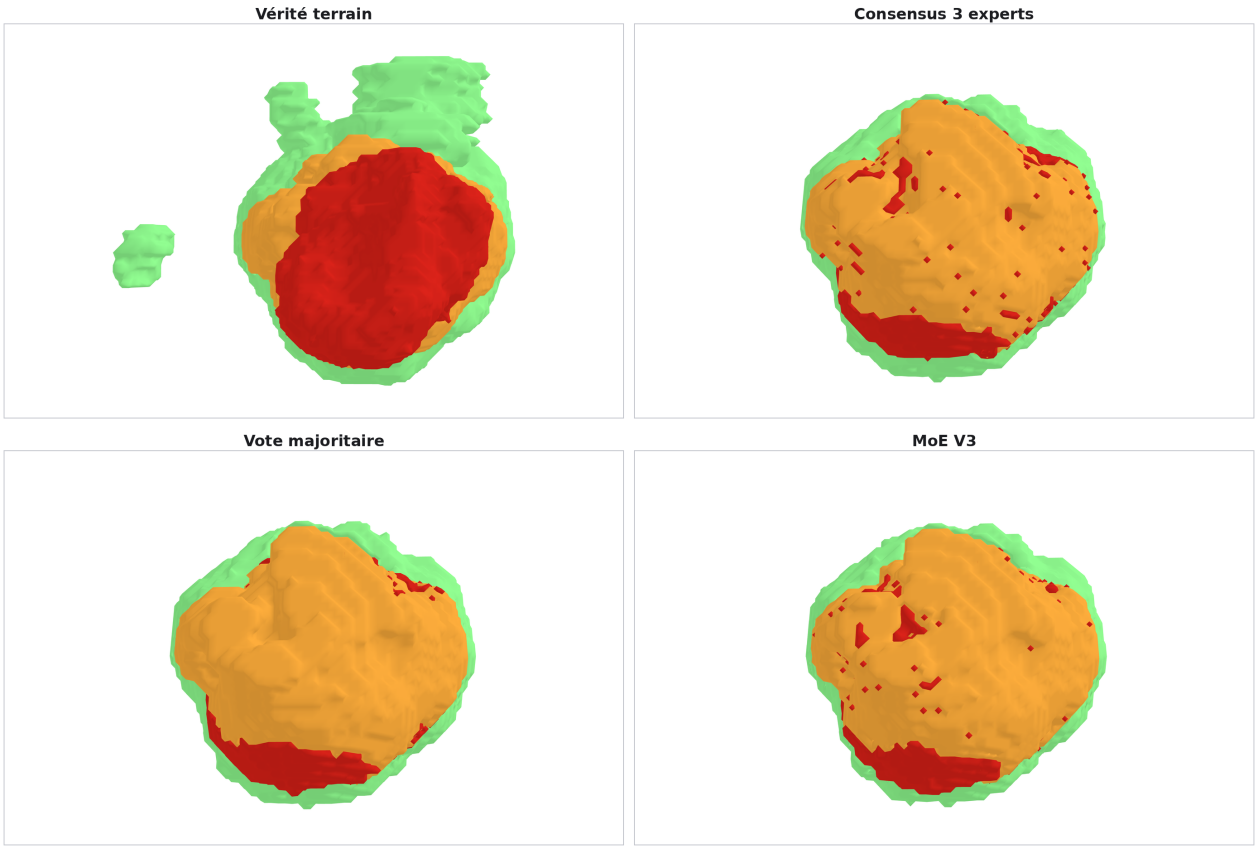

Figure 7: Figure 7 — Cas épinglé P3A (`BraTS-GLI-00442-000`, fold 0) : **la porte devant, de moins que son propre bruit de graine.** MoE V3 patch-wise à 0,7332 de Dice lésion-wise officiel, devant le vote majoritaire (0,7327), le consensus déterministe à trois experts (0,7319) et DistMap (0,7282) sur ce patient, le baseline étant à 0,7309. Son avance sur le vote est plus petite que sa propre étendue de graine ici, de 0,7332 à 0,7347 (graines 42, 1337, 2024). Sur les 42 bras scorés sur ce patient, la porte se classe 11.

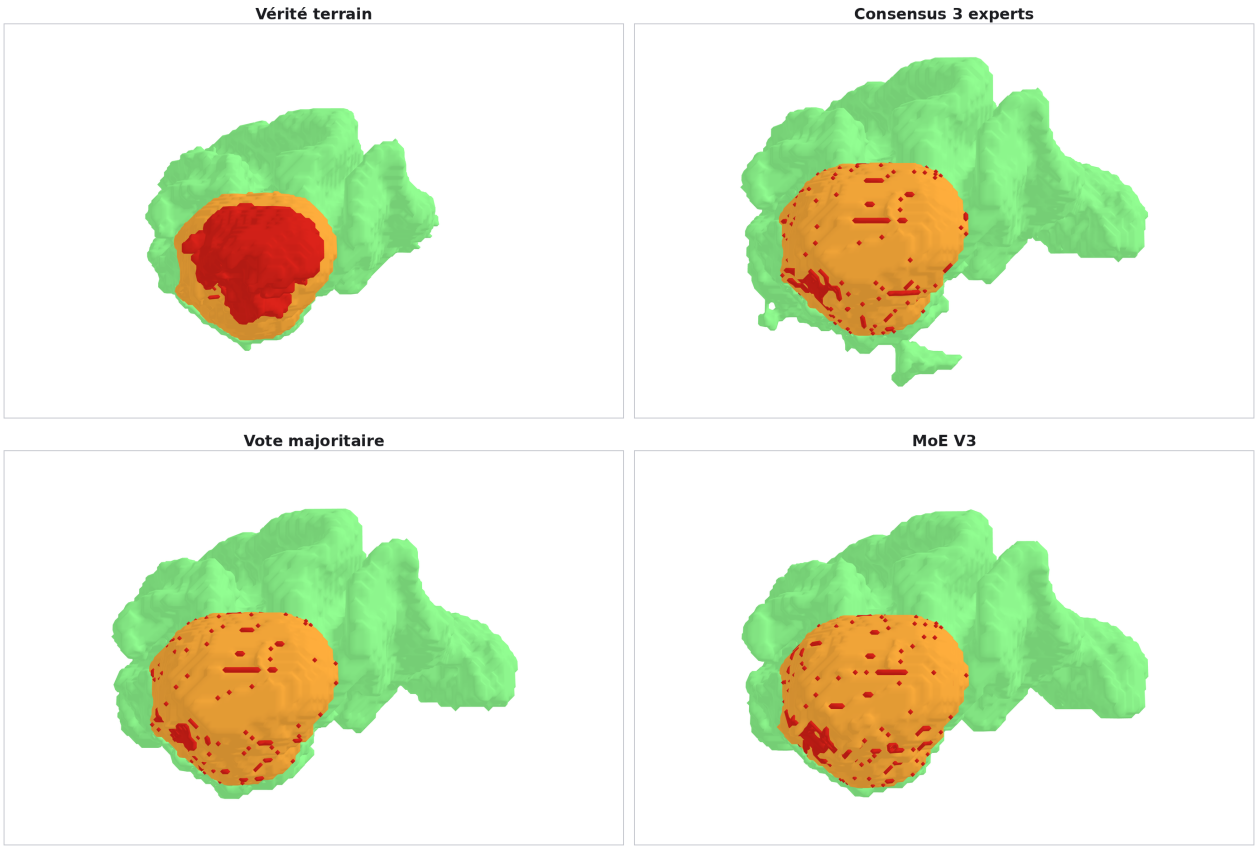

Figure 8: Figure 8 — Cas épinglé P3B (`BraTS-GLI-01004-000`, fold 0) : **la porte tient son niveau.** MoE V3 patch-wise à 0,9631, 0,0011 derrière le meilleur expert seul sur ce patient (DistMap, 0,9642), devant le vote majoritaire (0,9583) et le consensus déterministe à trois experts (0,9554) ; ses trois graines tombent entre 0,9629 et 0,9643. Sur les 42 bras scorés sur ce patient, la porte se classe 13 : sur un patient qu'un expert seul résout déjà, la porte n'ajoute ni ne retire rien.

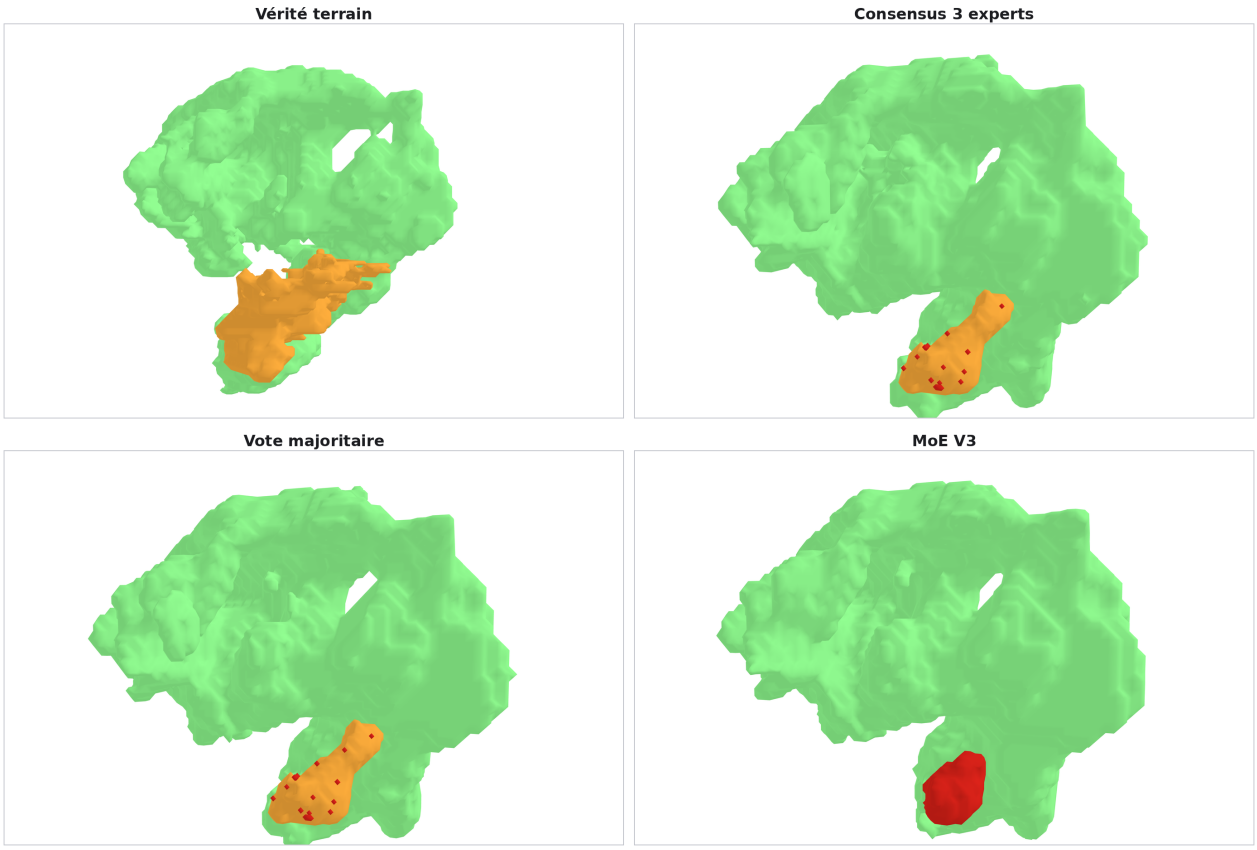

Figure 9: Figure 9 — Cas épinglé P3C (`BraTS-GLI-00621-000`, fold 0) : **la porte casse.** MoE V3 patch-wise s'effondre à 0,3174, 40 sur les 42 bras scorés sur ce patient — sous le baseline (0,5254), sous chaque expert seul, sous le consensus déterministe à trois experts (0,4794) et sous le vote majoritaire (0,4548), qui la bat ici de 0,1373. Après le nettoyage de composantes du barème officiel, le Dice de tumeur rehaussée de la porte vaut 0,0000 (sentinelle HD95) là où le consensus déterministe garde encore 0,3420 ; les trois graines s'effondrent toutes (0,3174, 0,2495, 0,2501) : ce n'est pas un accident de graine mais le visage individuel du verdict de population (§4.2, §4.5).

4.9 Les trois régions ne coûtent pas le même prix, et la moyenne le sait déjà

Les trois régions BraTS sont **emboîtées** : ET ⊂ TC ⊂ WT. Moyenner leurs trois Dice n'est donc pas les pondérer également par tissu. Un voxel de tumeur rehaussée est noté trois fois, un voxel de nécrose deux fois, un voxel d'œdème une seule, et les volumes diffèrent d'un facteur cinq (médianes du fold 0 : WT 89 400, TC 30 702, ET 18 630 voxels ; l'œdème seul fait **62 % du volume de WT**). En linéarisant le Dice autour d'une prédiction parfaite — sa pente vaut −1/(2V), la même pour un faux positif et un faux négatif au premier ordre — le coût d'un voxel mal segmenté sur la moyenne des trois vaut $(1/3)\cdot\Sigma_{R \supset T}\ 1/(2\cdot V_R)$, soit **×7,8 dans la tumeur rehaussée, ×3,7 dans la nécrose, ×1 dans l'œdème** (figure 10a). La métrique hiérarchise déjà les tissus ; elle ne le dit simplement pas. Cette dérivation vaut pour le Dice voxel-wise et au voisinage d'une bonne prédiction, pas pour un cas raté.

Cette pondération implicite n'est pas un artefact à corriger, car elle pointe dans le même sens que la littérature clinique. L'étendue de résection dans le glioblastome se classe sur le volume résiduel de tumeur **rehaussée**, et la survie globale médiane descend de 24 mois en classe 1 à 19, 15 puis 10 mois pour les classes 2 à 4 (Karschnia et al. 2023). Le cœur est ce sur quoi le chirurgien est jugé.

L'affirmation symétrique — l'œdème ne compterait pas — ne survit pas au contact de la même littérature, et nous en énonçons la limite. Le label 2 de BraTS est le **SNFH**, qui fond en une seule classe l'œdème vasogénique et l'infiltration tumorale non rehaussante, et cette infiltration est une cible chirurgicale : la résection maximale de la tumeur non rehaussée est associée à une survie plus longue chez les patients jeunes et IDH sauvage (Molinaro et al. 2020). C'est aussi une cible

radiothérapeutique. La recommandation ESTRO-EANO élargit le GTV de **15 mm** pour construire le CTV (Niyazi et al. 2023), et dans une comparaison de patrons de récidive, inclure l'anomalie T2-FLAIR dans le GTV a supprimé toute récidive marginale (0 sur 48) là où le volume EORTC et le volume 60 Gy du RTOG en laissaient chacun 27 % (Yilmaz et al. 2024). « Le cœur compte davantage » est défendable ; « l'œdème ne compte pas » ne l'est pas.

Le classement de l'étude en dépend-il ? Presque pas. Classer les bras sur le cœur (TC et ET) au lieu de la moyenne des trois donne un τ de Kendall de **0,94** avec le classement publié, et `cc_K_B` reste premier sur la moyenne, sur le cœur et sur TC. Classer sur **WT seul** donne τ = 0,08 (p = 0,60) et renvoie ce même bras au **rang 8** — mais c'est là un énoncé sur WT, où rien ne sépare aucun bras, plutôt qu'un classement concurrent. La décomposition dit pourquoi : `cc_K_B` gagne +0,0106 sur TC et +0,0061 sur ET contre +0,0002 sur WT (figure 10d), et le seul bras qui perde contre le baseline, `cc_B_and_DK` à −0,00067, ne perd **que** sur WT. Le gain de ce papier est un gain sur le cœur.

Ce que la lecture par région change vraiment, c'est le compte des échecs, exactement comme en section 4.8 : le baseline échoue sur **20 patients sur le cœur** contre 9 sur la moyenne des trois, les échecs se répartissant entre ET (25), TC (21) et WT (19) — et les sept sentinelles à 374 mm sont toutes sur ET. Nous ne repondérons donc pas la métrique : le challenge a fait son choix, le classement y est insensible, et un poids fabriqué à la main compterait le cœur deux fois. Nous rapportons, à côté de la moyenne, le taux d'échec sur le cœur et sur la pire région.

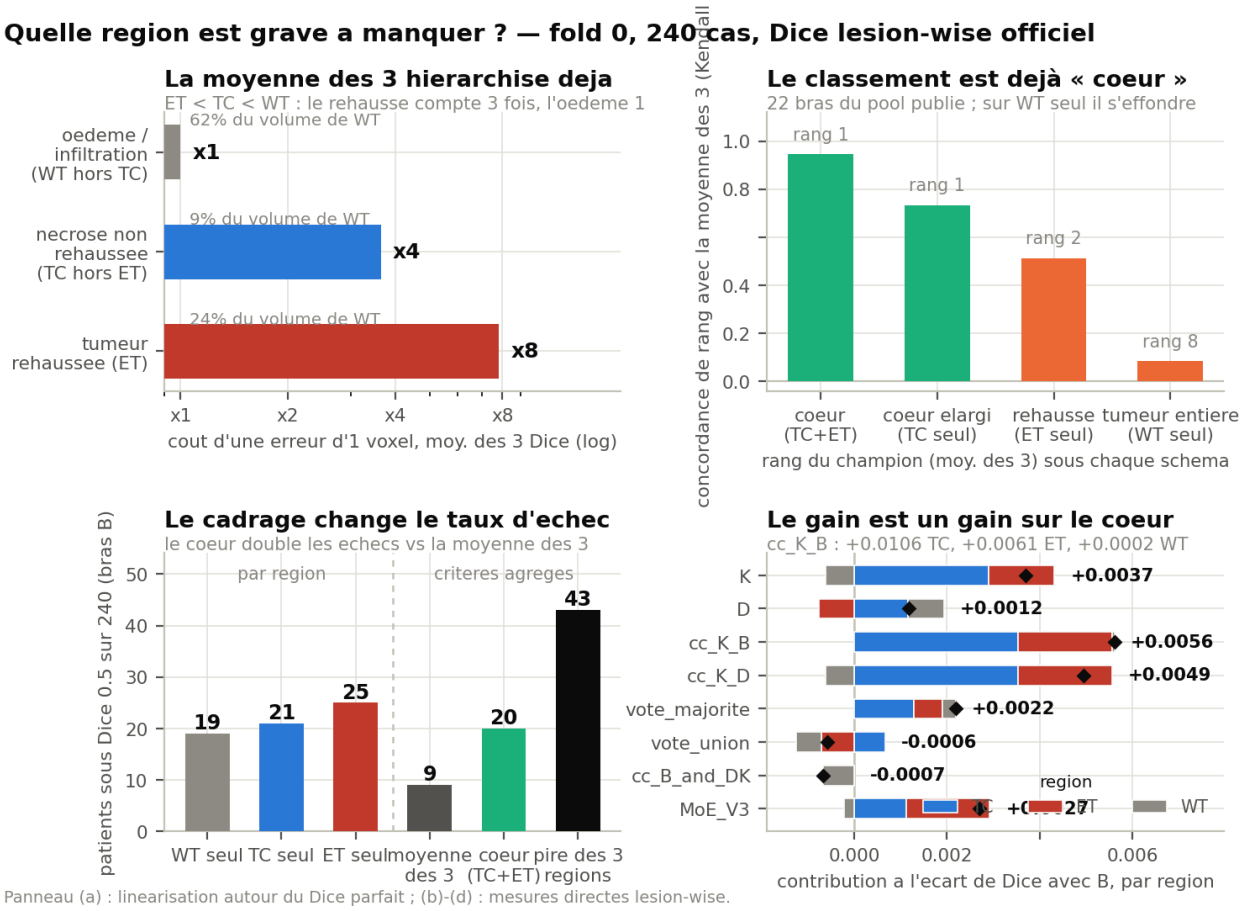

Figure 10: Figure 10 — Les trois régions emboîtées ne coûtent pas le même prix : le coût d'un voxel mal segmenté par tissu, en échelle log (a), le classement lu sur le cœur puis sur WT seul (b), les patients comptés en échec selon l'agrégation (c), et quelle région porte l'écart de chaque bras au baseline, en barres empilées signées (d). Le panneau (a) est une linéarisation autour du Dice voxel-wise parfait ; les trois autres sont des mesures directes en lésion-wise.

5. Discussion

Où mettre l'argent. Une équipe qui dispose d'un backbone solide et d'un budget fixe peut mettre la semaine de GPU suivante dans un mécanisme de combinaison appris, ou dans rien du tout et combiner ce qu'elle a déjà. Sur ce jeu de données, sous la métrique que le challenge note réellement, la réponse dépend du mécanisme appris dont on parle. Une porte patch-wise bâtie sur huit blocs aléatoires derrière un seul baseline rend −0,0007 sur cinq folds et se fait battre par l'opérateur gratuit ; la même architecture portant quatre spécialistes de loss initialisés depuis leurs propres checkpoints rend +0,0057 et bat les vingt-quatre opérateurs gratuits sur une marge moyenne appariée, le meilleur d'entre eux de 0,0044. La recommandation n'est donc pas « le consensus est puissant » — la section 4.7 plafonne à +0,0051 ce qu'il aurait pu être — ni « le routage est puissant » : c'est que ce qu'une porte ajoute ici est *l'initialisation*, et qu'une équipe qui ne peut pas se payer quatre entraînements préalables doit quand même commencer par le veto.

Deux propriétés au-delà du score renforcent cela, et aucune n'est un p. Un opérateur de consensus est **déterministe** : on le relance et le nombre ne bouge pas, là où la porte doit être réentraînée et atterrit n'importe où dans une bande de 0,0013 (0,0034 pour la version à huit blocs aléatoires qu'elle a remplacée). Il est aussi **auditable** : la règle tient en cinq lignes, n'a aucun paramètre appris, et son comportement sur un cas nouveau se prédit en la lisant — ce qui compte dans tout contexte où une modification d'un dispositif doit être justifiée et pas seulement comparée. Le dernier contre-argument de la porte est l'économie de déploiement : une passe avant contre deux ou trois. Sur une station mono-GPU, cet arbitrage est réel, et il figure au tableau 3 ; dans un service de production avec plus d'un GPU, il s'évapore pour

l'essentiel, parce que les passes du consensus sont indépendantes et se parallélisent (section 4.2). Cette étude ne trouve rien du côté précision qui vienne compenser le coût d'entraînement — y compris sur le terrain composite où l'on aurait pu attendre que la porte se rattrape, puisqu'elle n'améliore pas les deux endpoints conjointement plus qu'un baseline réentraîné ne le fait (section 4.3).

Une note sur les scores composites, pour qui serait tenté d'en construire un. Les deux endpoints officiels ne sont pas en tension sur cette tâche : ils bougent ensemble à l'intérieur d'un patient pour tous les bras mesurés ici, si bien qu'une combinaison pondérée des deux n'apporte presque rien que le couple ne portait déjà, et que c'est la pondération qui ferait l'argument. Là où un composite *change* la réponse, il le fait pour la mauvaise raison — le rank score de BraTS et la moyenne se contredisent sur trois bras, toujours parce que le rang compte les victoires et ignore leur taille. La recommandation pratique est de rapporter les deux endpoints séparément, d'ajouter un décompte conjoint par patient quand la question est vraiment « améliore- t-il les deux », et de traiter tout nombre composite unique comme un résumé à vérifier plutôt que comme un critère à optimiser.

Où est passée la place du mécanisme. Un veto par composante connexe est un filtre de taille et de soutien ; le post-traitement officiel est un filtre de taille dont les seuils sont un ordre de grandeur au-dessus des composantes en jeu. Appliquer les deux change un ou deux patients sur 240 — d'où les +0,0019 que le veto apporte dans une marge de +0,0056, et d'où le fait que les votes symétriques, qui peuvent synthétiser des formes qu'aucun expert n'a proposées, soient la sous-famille qui déplace encore la métrique. La condition préalable pour qu'un ensemble paie est que ses membres soient plus en désaccord qu'en erreur (Theisen et al. 2023) ; ici le désaccord vaut 8,3 % de l'erreur, les scores par patient corrèlent de 0,93 à 0,97, et 22 des 28 cas les plus durs sont durs pour les trois experts à la fois. Trois modèles qui partagent un jeu de données, un backbone, un calendrier et une graine, et ne diffèrent que par un terme auxiliaire déjà montré nul, ne sont pas un comité divers — ce sont un modèle mesuré trois fois. Que ce comité-là ait tout de même produit le meilleur bras de l'étude plaide pour le mécanisme, pas contre lui : un comité réellement divers aurait davantage de matière. Les ensembles publiés qui gagnent sur BraTS combinent des *architectures* (Kamnitsas et al. 2018 ; Ferreira et al. 2024), et les MoE médicaux qui rapportent des gains indexent leurs experts sur une hétérogénéité externe, comme des modalités manquantes (Liu et al. 2024). Une porte ne peut exploiter qu'une partition qui existe ; sur une cohorte mono-tâche et mono-protocole, il n'y en a peut-être aucune à trouver.

Un avertissement de protocole qui coupe des deux côtés. Les papiers qui rapportent des gains de consensus ou d'ensemble sur des prédictions *brutes* n'ont pas nécessairement tort : ils mesurent une quantité que le protocole du challenge supprime. Les mêmes douze vetos valent jusqu'à +0,0375 de Dice avant nettoyage et +0,0019 après. Lequel de ces deux nombres on publie est une décision de protocole, pas un fait empirique, et cela change la conclusion d'un facteur vingt. Nous rapportons partout le nombre d'après nettoyage, y compris là où il dégonfle notre propre meilleur bras.

Comment lire un tableau lésion-wise. Les votes symétriques exposent une propriété dont hérite toute comparaison BraTS-2023 : avec une pénalité de 374 mm par lésion manquée ou fausse, une poignée de cas gouverne la moyenne tandis que le test de rangs suit la majorité des patients. Deux de nos trois votes ont une moyenne et un test de rangs de signes opposés ; celui que nous recommandons, la majorité, est celui où ils s'accordent. Nous ne voyons pas de manière honnête de classer des bras sous cette métrique sans rapporter les deux, plus le nombre de patients améliorés et dégradés.

Le bras témoin. Entraîner le même modèle deux fois avec la même graine a déplacé le Dice officiel de 0,0018 et produit un p brut de 0,032. Tout protocole qui aurait déclaré une méthode significative à cette taille d'effet aurait aussi déclaré significatif le non-déterminisme du GPU. Ce n'est pas une bizarrerie locale : en faisant tourner nnU-Net sur 50 graines, Åkesson et al. ont trouvé la meilleure graine surpassant significativement jusqu'à 76 % des autres et ont conclu que la significativité est un indicateur faible d'une vraie différence entre algorithmes d'apprentissage (Åkesson et al. 2024 ; Picard 2021), et un audit récent estime que plus de la moitié des papiers de segmentation en imagerie médicale portent un risque non négligeable de fausse déclaration de supériorité (Christodoulou et al. 2025). Rapporter un bras témoin qui ne teste rien est la défense la moins chère disponible, et nous la recommandons en routine (Bouthillier et al. 2021 ; Maier-Hein et al. 2024). C'est aussi ce qui rend crédible la partie positive de ce papier : le vote majoritaire franchit une barre que le témoin ne franchit pas.

La porte, et ce qu'elle ne soutient pas. La porte change l'initialisation, pas le mécanisme, et trois des chiffres sur lesquels cette discussion s'appuie sont les siens : sa dispersion sur trois graines (0,0013 de Dice) vaut 0,96 fois celle de son propre témoin, si bien que son effet n'est pas à l'intérieur de son propre bruit ; sa moyenne sur cinq folds disjoints vaut +0,0054 avec un signe constant (+0,0061 sur les folds 1 à 4 seuls, quatre favorables sur quatre) ; et la marge en Dice

du consensus sur elle n'est plus établie par le test, alors que la marge en HD95 tient à chaque graine. Ce que la V3 ne fait *pas*, en revanche, c'est soutenir la lecture selon laquelle la porte aurait appris à router. Mesurée à la dernière époque de chacun de ses sept journaux d'entraînement, la préférence de la porte reste proche de l'indifférence — l'expert le plus choisi prend 0,507 à 0,583 des patchs là où 0,25 serait une absence totale de préférence sur quatre experts, et l'entropie de routage sans bruit reste entre 0,892 et 0,995 là où 1,0 est maximal — aucun expert ne meurt dans aucun run, et le degré de décision ne suit pas le gain officiel : le fold le plus décidé (H = 0,892, fold3) gagne le plus (+0,0092) tandis que la partition la plus piquée (part_max = 0,583, fold4) gagne le moins (+0,0025). Le gain est mesuré et constant ; l'histoire de la spécialisation n'est pas ce que ces mesures soutiennent. Investir dans l'accord reste le chemin le moins cher vers un effet comparable, et la porte reste 300 époques en plus de quatre experts déjà entraînés là où un veto ne coûte rien au-delà d'eux.

6. Limites

Les bras de consensus ne sont mesurés que sur le fold 0. C'est une limite d'ordonnancement de calcul et rien de plus — les prédictions des experts existent pour les cinq folds et la mesure est du travail CPU ; la section 4.7 en fait malgré tout une vraie limite, puisque le fold 0 classe démontrablement mal les experts. Tout ce qui est affirmé ici sur les opérateurs de consensus est donc affirmé sur un seul fold, et la section 4.6 chiffre la part d'une marge du fold 0 qui peut s'évaporer sur les autres : des +0,0056 qui mettent un consensus en tête du tableau 2, les +0,0037 attribuables à l'expert primaire sont la part à risque, et l'avantage en face-à-face sur la porte (section 4.2) repose sur le même fold que les graines du fold 0 de la porte — comparaison à conditions égales, mais sur un fold.

Les bras de consensus ne sont par ailleurs mesurés qu'à **une seule graine**, là où les trois experts en ont quatre. L'écart-type inter-graines qui sert partout à normaliser les effets — 0,00186 en Dice lésion-wise et 0,547 mm en HD95 sous le nettoyage officiel — est calculé sur les trois experts et sur eux seuls, et le JSON déclare explicitement cette assiette. Il est donc *emprunté* dès qu'on l'applique à un consensus, exactement au sens que la section 4.5 reproche à la porte. L'emprunt n'est pas neutre, car le bruit de graine dépend du bras : le nettoyage officiel par composantes l'absorbe d'un facteur 10 à 40 sur les bras résiduels et de 1,2 seulement sur une tête libre, et un veto — qui retranche des composantes — pourrait faire l'un comme l'autre. La conséquence est concrète : le z(min) = +3,02 qui place *K veto B* en tête du tableau 4 est un rapport dont le dénominateur a été mesuré sur un autre bras.

Le mélange d'experts est mesuré sur trois graines du fold 0 et sur les quatre folds restants ; le détail est en section 4.5 et aux tableaux 6 et 7. Ce qui compte pour les limites est la grandeur de comparaison : la dispersion du bras lui-même — 0,0013 de Dice entre les trois graines du fold 0, 0,0067 entre les quatre folds restants — et non sa distance à un run de référence unique. L'effet dépasse sa dispersion de graine ; aucun fold pris isolément ne dépasse sa propre différence minimale détectable. La campagne a réentraîné son propre témoin, si bien que sa dispersion ne doit jamais être divisée par le bruit d'un autre bras.

Aucun fold n'a la puissance d'établir son propre effet. La différence minimale détectable sur 239 à 240 patients reste au-dessus de chaque delta par fold de la porte V3 (+0,0027 pour une MDE de 0,0054 au fold 0 ; +0,0082 contre 0,0098 ; +0,0045 contre 0,0108 ; +0,0092 contre 0,0093 ; +0,0025 contre 0,0097). Ce que la cross-validation achète est un signe constant sur cinq populations disjointes (moyenne +0,0054), pas un test significatif dans l'une d'elles.

Le routage de la porte V3 n'est pas décisif, donc son gain n'est pas attribuable à la spécialisation. À la dernière époque de chacun des sept runs — mesurée par `scripts/paper3_v3_routage.py` depuis les journaux d'entraînement eux-mêmes, et non lue d'un résumé (`tables/v3_routage.json`, 7/7 runs à l'époque 299) — l'expert le plus choisi prend 0,507 à 0,583 des patchs, là où 0,25 est une absence totale de préférence sur quatre experts ; l'entropie de routage sans bruit reste entre 0,892 et 0,995, là où 1,0 est maximal ; aucun expert ne meurt (0/4 dans chaque run) ; et aucun lien lisible n'apparaît entre le degré de décision d'un run et ce que ce run gagne. Une seconde asymétrie a sa place ici : le quatrième expert de la porte (le bras blob G) est le pire des quatre pris seul — −0,0038 Dice sur les cinq folds, négatif sur les cinq, 28e des vingt-neuf — si bien que la porte héberge un spécialiste dont la contribution isolée est négative, et aucune ablation expert par expert de la porte n'a été entraînée dans cette étude. Ce que le +0,0057 de la porte doit à chacun de ses quatre experts n'est donc pas mesuré ; ce qui l'est, c'est que son routage ne choisit pas entre eux, et que le même expert utilisé comme veto sur le baseline est le meilleur des douze vetoes à deux experts (+0,0011).

Les prédictions du fold 0 pour B et D proviennent de runs de re-prédiction conservés sous des dossiers à suffixe, et non des sorties de validation d'origine du fold 0, qui ont été perdues ; la provenance du fold 0 n'est donc pas uniforme avec

celle des folds 1 à 4. C'est documenté dans le registre des experts et n'affecte que le fold 0.

STAPLE (Warfield et al. 2004) ne figure pas parmi les opérateurs comparés, faute d'implémentation vérifiée dans ce dépôt ; un combinateur appris au-dessus d'experts gelés est délibérément exclu (section 3.4). Le nettoyage par taille utilise la 6-connexité alors que le scorer et les opérateurs de consensus utilisent la 26-connexité — incohérence documentée, conservée parce que la corriger casserait la comparabilité avec les papers 1 et 2. Enfin, tous les résultats portent sur la cohorte gliome adulte d'une seule édition du challenge, évaluée sur des folds held-out internes et non sur le leaderboard officiel.

7. Conclusion

Vingt-neuf bras, un jeu de données, un jeu de métriques, et un témoin qui ne teste rien. Sous les métriques officielles lésion-wise de BraTS-2023, le bras qui marque le plus est un **mélange d'experts patch-wise appris** dont les quatre experts sont initialisés depuis quatre spécialistes de loss déjà entraînés : +0,0057 de Dice sur son propre baseline en moyenne sur cinq folds disjoints, positif sur cinq d'entre eux, au-dessus du plus petit effet d'intérêt pré-déclaré sur quatre. Les vingt-quatre opérateurs de consensus sans entraînement bâtis à partir des mêmes experts sont tous derrière lui sur une marge moyenne appariée, le meilleur à +0,0013 et sur cinq folds sur cinq ; aucun n'atteint cet effet d'intérêt. La version antérieure de la même couche, bâtie sur huit blocs aléatoires derrière un seul baseline, se situe sous l'opérateur gratuit à −0,0007 — ce n'est donc pas l'architecture qui gagne, c'est l'initialisation.

C'est le résultat négatif dont ce papier porte le titre : **la porte ne choisit pas.** Sur les sept runs la part du plus gros expert va de 0,507 à 0,583 là où 0,25 est l'absence totale de préférence sur 4 experts, l'entropie patch normalisée va de 0,892 à 0,995 là où 1,0 est le maximum, et aucun expert ne meurt nulle part. Un routage qui ne décide pas ne peut pas être le mécanisme du gain, et ce papier ne l'affirme pas ; ce qu'il affirme, c'est qu'héberger quatre spécialistes dans un seul réseau et les faire démarrer depuis leurs propres checkpoints vaut +0,0044 de mieux que tout ce que l'on peut faire gratuitement.

Les limites de cette conclusion sont mesurées et non affirmées, et c'est ce qui la rend utilisable. Aucun fold pris isolément ne franchit sa propre différence minimale détectable, si bien que l'affirmation repose sur un signe constant sur cinq populations disjointes et non sur un test significatif fold par fold ; la référence du fold 0 dérive de −0,0032 entre campagnes de scoring, raison pour laquelle les vingt-neuf bras sont comparés par un seul chemin de code sur un seul baseline par fold ; l'expert à blob loss est le pire des quatre pris seul (−0,0038, aucun des cinq folds positif) et porte quand même le meilleur des douze vetoes à deux experts (+0,0011, deuxième des vingt-quatre opérateurs) ; et le prix est de 300 époques par-dessus quatre entraînements complets, là où un veto ne coûte rien et ses passes se parallélisent sur plusieurs GPU. Les effets comparés ici sont du même ordre que la distance entre un modèle et lui-même — et nous avons mesuré les deux distances au lieu de les supposer.

La recommandation pratique suit, à coût nul : avant de bâtir un comité, mesurer si ses membres sont plus en désaccord qu'en erreur, rapporter un bras témoin qui ne teste rien, et essayer d'abord l'opérateur gratuit. Sur ce jeu de données ces trois vérifications ne pointent plus dans le même sens — et c'est cela, le résultat : l'opérateur gratuit a un plafond, la porte qui le bat ne route pas, et la marge qu'elle capture est celle que porte l'initialisation propre des experts.

Contributions des auteurs

Guillaume Cassez (auteur principal) : conception de l'étude, entraînement des modèles, évaluations et analyses, rédaction du manuscrit. **Stanislas Larnier** : formulation des questions de recherche, conseils méthodologiques, relectures attentives des versions successives du papier. Les deux auteurs ont approuvé la version finale et l'ordre des auteurs.

Références

- 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.