Contents

La précision paie : une loss de frontière à poids fixe ancre l'extrémité précision de l'axe rappel–précision et gagne comme veto du consensus sous les métriques officielles BraTS-2023 **2**
Résumé . . . 3
1. Introduction . . . 4
2. Travaux connexes . . . 4
3. Méthodes . . . 4
3.1 Données et backbone . . . 4
3.2 Loss de frontière . . . 5
3.3 Sélection du poids et robustesse . . . 5
3.4 Évaluation . . . 5
4. Résultats . . . 5
4.1 Versus Baseline — aucun effet significatif au centre . . . 5
4.2 Versus la tête SDT — un miroir orienté précision (exploratoire, robuste) . . . 6
4.3 Kervadec comme veto de consensus . . . 13
4.4 Robustesse à la graine d'entraînement (3 graines, fold 0) . . . 13
5. Discussion . . . 15
6. Limites . . . 16
7. Conclusion . . . 16
Contributions des auteurs . . . 16
Références . . . 16

La précision paie : une loss de frontière à poids fixe ancre l'extrémité précision de l'axe rappel–précision et gagne comme veto du consensus 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 2.6 (2026-09-22) — provenance mesurée des 55 exclusions, deux statistiques réénoncées à pleine précision. Aucun résultat rapporté ne change.

Provenance. Le §3.1 énonçait « 1251 cas » alors que toute statistique de ce papier est calculée sur le jeu d'étude nnU-Net de **1196** cas ; la formulation laissait lire que 1251 cas avaient été entraînés. Le §3.1 distingue désormais le téléchargement (1251 cas unifiés) du jeu d'étude (1196) et renvoie à `analysis/DATASET_EXCLUSIONS_fr.md`, qui liste les 55 cas écartés un par un — déclencheur, date git de l'entrée déclenchante, et nouvelle mesure de leur intégrité fichier (55 exclus contre 55 témoins inclus : 0 fichier illisible, 0 NaN/Inf, 0 label hors jeu, 0 affine divergent, (240, 240, 155) partout). L'exclusion est une précaution conservative au niveau du patient, enregistrée en mars 2026, en amont de tout entraînement et de toute métrique ; le miroir anglais est livré sous `analysis/DATASET_EXCLUSIONS.md` avec son CSV `analysis/exclusions_1251_to_1196.csv`.

Statistiques réénoncées (§4.4). Le Dice de validation nnU-Net multi-graine de la loss de frontière vaut 0,9060 ± 0,0018 (par graine 0,9062 / 0,9078 / 0,9042), imprimé 0,9061 ± 0,0018 jusqu'ici — soit la moyenne de valeurs par graine déjà arrondies à 4 décimales — et son test t apparié face à Baseline (0,9074 ± 0,0007) donne t = −1,02, p = 0,42, là où le t = 0,96, p = 0,44 imprimé jusqu'ici sortait de ce même test appliqué à des valeurs d'abord arrondies. Les deux lectures restent très loin de la significativité (df = 2) : aucune conclusion ne change. `data/nnunet_validation_dice_multiseed.json` porte désormais les six mesures lues des logs d'entraînement et les deux calculs.

Auto-contenance de l'archive. Tout chemin cité dans le manuscrit résout dans cette archive : le pointeur final `Références :` nommait un chemin inexistant ici (`../references.bib` -> `references.bib`), et les trois légendes de cas épinglés citent désormais `data/patient_level_clean_BKD.json`, l'artefact qui porte les comptes patient qu'elles annoncent.

Version 2.5 (2026-09-21) — contributions des auteurs. Aucun chiffre modifié. Une section **Contributions des auteurs** a été ajoutée après le §7, déclarant les mêmes rôles que dans les Papiers 1 et 3 (G.C. : conception de l'étude, entraînements, évaluations et analyses, rédaction ; S.L. : questions de recherche, conseil méthodologique, relectures des versions successives). Ce bandeau est ajouté maintenant parce que la section était entrée sans lui — le changelog public ne décrivait pas l'artefact.

Version 2.4 (2026-09-19) — titre resserré, règles de mise en page. Aucun chiffre modifié. L'accroche est inchangée ; le sous-titre est raccourci. Le bloc titre démarre désormais une page neuve — jamais coupé entre deux pages — et les Références sont rendues en fin de PDF (citations résolues depuis `references.bib` ; plus aucun marqueur de citation brut dans le texte). Le plan des sections suit déjà le squelette commun des Papiers 1 et 3 (Méthodes dont Données et Évaluation, Résultats, Discussion, Limites autonomes, Conclusion).

Version 2.3 (2026-09-01) — auteurs et titre. Aucun chiffre modifié. D'un commun accord entre les deux auteurs, Guillaume Cassez est premier auteur et Stanislas Larnier deuxième. Le titre mène désormais par ce que la loss rapporte — l'extrémité précision de l'axe et le veto gagnant — le résultat nul de la loss seule est énoncé dans le papier, pas dans son titre.

Version 2.2 (2026-08-22) — passe pré-relecture. Aucun chiffre modifié. Identifiants des auteurs complétés (Stanislas Larnier : IdHAL `stanislas-larnier` ; Guillaume Cassez : ORCID 0009-0007-0987-3931). Les résultats de consensus du résumé précisent désormais leur population : les chiffres K∩B provi-

ennent du protocole multi-graine (3 graines × fold 0, n = 720 poolé), le chiffre du veto D∩K de la CV 5-fold complète (n = 1196). Méthode du programme confirmée le 2026-08-22 : 3 graines et 5 folds pour tous les papiers.

Version 2.1 (2026-08-11) — passe éditoriale. Aucun chiffre modifié. Le résumé mène désormais par le résultat de consensus, seul résultat significatif de l'étude, et le résultat nul de la loss de frontière seule vient après ; l'arbitrage que le papier documente — situer la loss sur l'axe rappel/précision, puis la dépenser dans l'accord — est énoncé d'emblée.

Version 2 (2026-06-14) — note de version. Remplace la v1. Évaluation refaite sous les **métriques officielles BraTS-2023** à convergence (300 epochs) sur la validation croisée 5-fold complète (n = 1196). La loss de frontière est **non significative vs Baseline** sur toute métrique officielle ; son trait distinctif est d'occuper l'extrémité *précision* d'un axe rappel/précision. Ajout v2 : Kervadec est un **veto** efficace dans le filtre CC-consensus de l'article compagnon DistMap (sa haute spécificité préserve le rappel comme veto). Ajout v2 (2026-06-14) : **analyse multi-graine** (3 graines × fold 0) qui confirme la non-significativité vs Baseline et établit deux faits nouveaux — (i) Kervadec resserre fortement la dispersion inter-graine des métriques *lesion-wise*, à l'opposé de DistMap ; (ii) le filtre de consensus K∩B reproduit le gain *lesion-wise* WT du filtre D∩B de Paper 1, montrant que l'amélioration est portée par le mécanisme de consensus, non par la loss auxiliaire.

Brouillon de travail — Article 2 du programme « losses auxiliaires » BraTS. Les affirmations quantitatives CV proviennent de `code/paper_stats.py` (`data/cv_paper_stats.json`) ; les tableaux de `code/make_tables.py` (`table_paper2.{md,tex}`). Les chiffres multi-graine proviennent de `code/evaluate_multiseed_kervadec.py` (`data/multiseed_stats.json`). Toute métrique officielle BraTS-2023 est rapportée ; un critère principal pré-spécifié est séparé des analyses exploratoires ; un résultat non significatif / négatif est rapporté comme tel.

Résumé

Deux modèles qui doivent être d'accord battent chacun des deux pris seul. Sous les métriques officielles lésion-wise de BraTS-2023, un consensus de composantes connexes sans aucun paramètre — ne garder que les composantes tumorales sur lesquelles deux modèles concordent — retire environ la moitié des lésions faussement positives d'un modèle et produit les seules améliorations significatives de cette étude : Dice lésion-wise de la tumeur entière **+3,89 pp** (p = 0,005) et HD95 **−15,66 mm** (p = 0,001) pour le filtre Kervadec∩Baseline (3 graines × fold 0, n = 720 poolé), et HD95 **−9,95 mm** ($p = 2{,}3\times10^{-20}$; CV 5-fold, n = 1196) quand le modèle à loss de frontière vetote le modèle à distance map du Paper 1 compagnon. Ce gain ne coûte aucun entraînement, aucun paramètre appris et aucun hyper-paramètre : c'est l'accord qui travaille, et il se reproduit avec les deux losses auxiliaires (K∩B reproduit le D∩B du Paper 1, +4,28 pp / −17,2 mm), ce qui identifie le mécanisme plutôt que l'ingrédient.

L'ingrédient, lui, est nul, et nous le rapportons en entier. La loss de frontière de Kervadec et al. (Kervadec et al. 2019 ; Kervadec et al. 2021) vise à réduire l'erreur de contour dans la segmentation fortement déséquilibrée. On teste si son ajout (poids fixe $\lambda = 0{,}05$, agissant directement sur les logits, sans tête auxiliaire) à un backbone MedNeXt / nnU-Net performant améliore la segmentation de gliomes de l'adulte, évalué sous le jeu **complet** de métriques officielles BraTS-2023 en validation croisée 5-fold (n = 1196). Face à Baseline, la loss de frontière est **non significative sur les critères principaux pré-spécifiés** (Dice lesion-wise $\Delta = +0{,}000$, Holm p = 0,34 ; HD95 lesion-wise $\Delta = -0{,}58$ mm, Holm p = 0,31), **neutre en Dice** de chevauchement (Δ moyen = +0,000, p = 0,06), **nulle sur le HD95 régional legacy** (Δ moyen = −0,09 mm, p = 0,11) et **neutre en détection** (sensibilité régionale p = 0,26, spécificité p = 0,08). En bref, une pénalité de frontière à poids fixe est statistiquement indiscernable de Baseline sur un backbone qui possède déjà une supervision profonde — un résultat négatif rapporté ici en entier, avec un contrôle de robustesse du choix de poids qui écarte l'objection du mauvais réglage.

Ce à quoi cette loss est bonne, c'est à vetoter. Sa seule propriété distinctive et robuste en solo apparaît **par contraste avec une tête auxiliaire de distance signée (SDT)** : elle est significativement **plus spécifique** (spécificité régionale r = +0,27, $p = 2{,}6\times10^{-16}$) et produit **moins de lésions fallacieuses cœur-tumoral / rehaussantes** (lésions FP TC $\Delta = -0{,}066$, ET $\Delta = -0{,}195$, Holm p = 0,004), au prix du rappel — l'image miroir du biais de rappel de la tête SDT. Cette précision est exactement ce qui en fait un bon corroborateur : comme veto, elle retire des lésions faussement positives

sans presque toucher au rappel (0,006 lésion manquée ajoutée par cas contre 0,009 pour un veto par Baseline). Une analyse **multi-graine** indépendante (3 graines × fold 0) confirme la non-significativité vs Baseline (t-test p = 0,42) et établit que Kervadec **resserre** la dispersion inter-graine des métriques lésion-wise (σ LW Dice 3 à 12× plus faible que Baseline selon la région), à l'opposé de DistMap. La lecture pratique : situer une loss auxiliaire sur l'axe rappel–précision, puis la dépenser dans l'accord plutôt que dans le classement.

1. Introduction

Les losses sensibles à la frontière partent du constat que les losses de chevauchement sous-pondèrent le placement du contour. La loss de frontière (Kervadec et al. 2019 ; Kervadec et al. 2021) intègre une distance de type *level-set* le long du contour prédit et a aidé des backbones faibles. Qu'elle aide un backbone 3D **moderne et performant** (MedNeXt (Roy et al. 2023) dans nnU-Net (Isensee et al. 2021)) déjà doté d'une supervision profonde et d'un objectif Dice + entropie croisée est bien moins clair.

On teste cela honnêtement : on évalue la loss de frontière à poids fixe sous **toutes** les métriques officielles BraTS-2023, on sépare un critère principal pré-spécifié des analyses exploratoires, on rapporte le résultat quel qu'en soit le signe, et on devance l'objection du mauvais réglage par un contrôle de robustesse sur la sélection du poids.

Contributions : 1. Une évaluation à convergence, en 5-fold, d'une loss de frontière à poids fixe sur un backbone gliome 3D performant sous le jeu complet de métriques officielles BraTS-2023. 2. Un **résultat négatif** rapporté de façon transparente : aucune amélioration significative vs Baseline sur la moindre métrique principale ou moyennée par région, avec un contrôle de robustesse écartant un mauvais réglage du poids. 3. Le seul signal robuste — que la loss de frontière est la **contrepartie orientée précision** d'une tête auxiliaire SDT (significativement plus spécifique, moins de lésions fallacieuses TC/ET, rappel plus faible) — situant les deux losses auxiliaires sur un unique axe rappel–précision. 4. Une **analyse multi-graine** (3 graines × fold 0) qui (i) confirme la non-significativité vs Baseline, (ii) montre que la loss de frontière *régularise* la variance lesion-wise inter-graine, et (iii) établit que la règle de consensus de composantes connexes produit son gain lesion-wise WT indépendamment de la loss auxiliaire servant de modèle primaire (K∩B reproduit D∩B).

2. Travaux connexes

Losses à intégrale de frontière (Kervadec et al. 2019 ; Kervadec et al. 2021) vs régression de carte de distance (Ma et al. 2020 ; Xue et al. 2020 ; Dangi et al. 2019) : deux familles d'objectifs sensibles à la forme / à la frontière ; l'article compagnon (Paper 1) étudie la seconde. Backbone et entraînement : nnU-Net (Isensee et al. 2021), MedNeXt (Roy et al. 2023), Dice / GDL (Milletari et al. 2016 ; Sudre et al. 2017). Données et métriques : BraTS (Menze et al. 2015 ; Bakas et al. 2017 ; Baid et al. 2021) ; métriques lesion-wise (Moawad et al. 2023 ; Saluja et al. 2023) ; distance de surface (Nikolov et al. 2021).

3. Méthodes

3.1 Données et backbone

Identiques à l'article compagnon Paper 1. **Données.** Gliome de l'adulte BraTS-2023 (GLI) (Baid et al. 2021) : **1251 cas unifiés téléchargés, 1196 dans le jeu d'étude** — ce jeu d'étude est celui qu'utilisent tous les entraînements et toutes les métriques ci-dessous. Les 55 cas écartés sont listés un par un (déclencheur, date git de l'entrée déclenchante, nouvelle mesure de leur intégrité fichier : 55 exclus contre 55 témoins inclus, 0 fichier illisible, 0 NaN/Inf, 0 label hors jeu, 0 affine divergent) dans `analysis/DATASET_EXCLUSIONS_fr.md`. L'écart est une précaution au niveau du patient, enregistrée en mars 2026, en amont de tout entraînement et de toute métrique ; aucun chiffre publié ne change. **Backbone.** Splits officiels nnU-Net 5-fold, MedNeXt-B via nnU-Net v2 (`3d_fullres`, `nnUNetPlans_96GB_mednext`), Dice + entropie croisée avec supervision profonde, 300 epochs/fold, RTX PRO 6000. Toutes les méthodes partagent les folds (statistiques appariées, n = 1196).

3.2 Loss de frontière

On ajoute le terme de frontière de Kervadec sous la forme `seg_loss + λ · mean(softmax(logits) · SDT_signée)` ($\theta_0 = 30$, $\lambda = 0{,}05$ fixe), agissant directement sur les logits de segmentation, **sans tête auxiliaire**. Implémentation : `code/nnUNetTrainerMedNeXtKervadec.py` ; cache de SDT signée précalculé pour tous les cas.

3.3 Sélection du poids et robustesse

$\lambda \in \{0{,}05 ; 0{,}15 ; 0{,}5 ; 1{,}5 ; 5{,}0\}$ a été balayé sur un calendrier court (50 epochs, fold 0) ; $\lambda = 0{,}05$ maximise le pseudo-Dice EMA. Ce critère étant fondé sur le chevauchement, on re-classe en outre le balayage sur le HD95 plafonné aux sentinelles (médiane + moyenne) : $\lambda = 0{,}05$ est déjà quasi-optimal en HD95 (plus faible HD95 médian WT, le moins de ratés catastrophiques), et un λ plus grand **dégrade** le HD95 (la médiane croît de façon monotone ; $\lambda = 5{,}0$ s'effondre). Le résultat nul par région ci-dessous n'est donc pas un artefact de sélection du poids.

3.4 Évaluation

Protocole de métriques officielles et statistiques identiques à Paper 1 (BraTS-2023-Metrics officiel, paramètres GLI ; Wilcoxon apparié bilatéral ; r rang-bisériel avec + en faveur de Kervadec et son IC bootstrap ; correction de Holm entre régions ; FDR de Benjamini–Hochberg sur la famille exploratoire des 216 tests ; critère principal pré-spécifié = les deux métriques de classement officielles moyennées par région). On évalue Kervadec vs Baseline et Kervadec vs la tête SDT (DistMap). Pipeline CV : `code/eval_lesionwise_kervadec.py` → `code/paper_stats.py`. Pipeline multi-graine (§4.4) : `code/evaluate_multiseed_kervadec.py`.

4. Résultats

4.1 Versus Baseline — aucun effet significatif au centre

- **Critères principaux (nuls) :** Dice lesion-wise $\Delta = +0{,}000$ (Holm p = 0,34) ; HD95 lesion-wise $\Delta = -0{,}58$ mm (r = +0,06, p brut = 0,155, Holm p = 0,31).
- **Chevauchement (Dice-neutre) :** Dice legacy moyen $\Delta = +0{,}000$ (p = 0,063 ; Tableau 1). Un unique effet nominal — Dice legacy cœur-tumoral $\Delta = +0{,}001$ (p = 0,005, Holm 0,015†) — est significatif mais d'amplitude négligeable.
- **Frontière (nul) :** HD95 régional legacy moyen $\Delta = -0{,}09$ mm (r = +0,07, p = 0,109).
- **Détection (neutre) :** sensibilité régionale moyenne $\Delta = -0{,}001$ (p = 0,256) ; spécificité moyenne $\Delta = +0{,}000$ (r = +0,06, p = 0,079). Les comptes de lésions FP/FN vs Baseline sont tous non significatifs (Tableau 2). La loss de frontière ne se sépare pas de Baseline.

Critères principaux pré-spécifiés (Kervadec vs Baseline, métriques officielles de classement moyennées par région, correction de Holm sur la famille de 2 métriques) :

- Dice lésion-wise : $\Delta = +0{,}000$, r = +0,032, p brut = 0,339, p Holm = 0,339 → non significatif
- HD95 lésion-wise : $\Delta = -0{,}582$ mm, r = +0,056, p brut = 0,155, p Holm = 0,311 → non significatif

Tableau 1 — Métriques officielles de classement et legacy (Kervadec vs Baseline)

Δ = Kervadec − Baseline. r = corrélation rang-bisériale appariée (+ favorise Kervadec). * p brut < 0,05 ; † p corrigé de Holm < 0,05 (sur les 3 régions). n = 1196.

Métrique	Région	Kervadec	Baseline	Δ	r	gagnés/perdus	p	p(Holm)
Dice lésion-wise	WT	0.811	0.812	-0.001	+0.03	614/580	0.352	0.658
	TC	0.869	0.870	-0.001	+0.07	643/543	0.039*	0.118
	ET	0.800	0.798	+0.003	+0.03	603/557	0.329	0.658
	moy	0.827	0.826	+0.000	+0.03	635/559	0.339	–
HD95 lésion-wise (mm)	WT	53.87	53.09	+0.78	+0.06	337/289	0.216	0.631
	TC	24.33	24.68	-0.35	-0.01	222/225	0.889	0.889
	ET	42.71	44.89	-2.18	+0.07	200/173	0.210	0.631
	moy	40.31	40.89	-0.58	+0.06	458/388	0.155	–
Dice legacy	WT	0.936	0.935	+0.000	+0.03	619/575	0.397	0.648
	TC	0.918	0.918	+0.001	+0.09	645/542	0.005*	0.015†
	ET	0.870	0.869	+0.000	+0.03	606/557	0.324	0.648
	moy	0.908	0.907	+0.000	+0.06	646/548	0.063	–
HD95 legacy (mm)	WT	5.77	5.79	-0.02	+0.06	291/269	0.252	0.589
	TC	6.02	6.00	+0.01	+0.06	214/199	0.282	0.589
	ET	11.36	11.62	-0.26	+0.08	172/145	0.196	0.589
	moy	7.72	7.81	-0.09	+0.07	420/354	0.109	–

Tableau 2 — Détection de lésions (Kervadec vs Baseline)

Δ = Kervadec − Baseline. r = corrélation rang-bisériale appariée (+ favorise Kervadec). * p brut < 0,05 ; † p corrigé de Holm < 0,05 (sur les 3 régions). n = 1196.

Métrique	Région	Kervadec	Baseline	Δ	r	gagnés/perdus	p	p(Holm)
Lésions FP	WT	0.334	0.337	-0.003	+0.00	112/112	0.945	1.000
	TC	0.122	0.132	-0.010	-0.02	38/46	0.725	1.000
	ET	0.641	0.718	-0.077	+0.14	80/74	0.158	0.474
	moy	0.366	0.396	-0.030	+0.02	172/176	0.777	–
Lésions FN	WT	0.079	0.080	-0.001	+0.11	5/4	0.739	0.739
	TC	0.033	0.037	-0.003	+0.43	5/2	0.206	0.412
	ET	0.038	0.042	-0.004	+0.56	6/2	0.132	0.395
	moy	0.050	0.053	-0.003	+0.37	10/7	0.244	–
Sensibilité	WT	0.929	0.929	-0.000	-0.03	574/618	0.438	0.438
	TC	0.919	0.921	-0.002	+0.04	636/544	0.193	0.386
	ET	0.876	0.877	-0.000	+0.07	623/532	0.051	0.154
	moy	0.908	0.909	-0.001	+0.04	627/567	0.256	–
Spécificité	WT	1.000	1.000	+0.000	+0.07	626/566	0.039*	0.118
	TC	1.000	1.000	+0.000	+0.00	573/607	0.908	0.908
	ET	1.000	1.000	-0.000	-0.04	584/593	0.259	0.517
	moy	1.000	1.000	+0.000	+0.06	616/578	0.079	–

4.2 Versus la tête SDT — un miroir orienté précision (exploratoire, robuste)

La loss de frontière diffère nettement de la tête auxiliaire SDT (Tableaux 3–4, Fig. 1–2) : - **Plus spécifique :** spécificité régionale r = +0,27 (moyenne $\mathbf{p = 2{,}6\times10^{-16}}$; toutes régions Holm†). - **Moins de lésions fallacieuses :** lésions FP TC Δ = −0,066 (p = 0,002, Holm 0,004†), ET Δ = −0,195 (p = 0,001, Holm 0,004†). - **Rappel plus faible / plus de lésions ratées :** sensibilité régionale r = −0,18 (moyenne $p = 1{,}1\times10^{-7}$). - **Meilleur HD95 lesion-wise rehaussant :** Δ = −4,05 mm (p = 0,015, Holm 0,046†) et Dice lesion-wise ET Δ = +0,008 (p = 0,020).

La loss de frontière occupe donc l'extrémité **précision** d'un axe rappel–précision dont l'extrémité rappel est la tête SDT (Paper 1) ; aucune des deux ne bat Baseline sur les métriques de classement officielles.

Critères principaux pré-spécifiés (Kervadec vs DistMap, métriques officielles de classement moyennées par région, correction de Holm sur la famille de 2 métriques) :

- Dice lésion-wise : Δ = +0,003, r = +0,033, p brut = 0,324, p Holm = 0,647 $\rightarrow$ non significatif
- HD95 lésion-wise : Δ = −1,700 mm, r = +0,027, p brut = 0,487, p Holm = 0,647 $\rightarrow$ non significatif

Tableau 3 — Métriques officielles de classement et legacy (Kervadec vs DistMap)

Δ = Kervadec − DistMap. r = corrélation rang-bisériale appariée (+ favorise Kervadec). * p brut < 0,05 ; † p corrigé de Holm < 0,05 (sur les 3 régions). n = 1196.

Métrique	Région	Kervadec	DistMap	Δ	r	gagnés/perdus	p	p(Holm)
Dice lésion-wise	WT	0.811	0.816	-0.005	-0.01	606/589	0.785	0.785
	TC	0.869	0.862	+0.008	+0.05	613/576	0.161	0.323
	ET	0.800	0.792	+0.008	+0.08	618/545	0.020*	0.061
	moy	0.827	0.823	+0.003	+0.03	616/579	0.324	–
HD95 lésion-wise (mm)	WT	53.87	51.11	+2.76	-0.03	328/332	0.486	0.971
	TC	24.33	28.14	-3.81	+0.03	225/232	0.626	0.971
	ET	42.71	46.76	-4.05	+0.14	231/193	0.015*	0.046†
	moy	40.31	42.01	-1.70	+0.03	450/418	0.487	–
Dice legacy	WT	0.936	0.936	-0.000	+0.01	607/588	0.786	0.786
	TC	0.918	0.917	+0.001	+0.05	617/572	0.111	0.332
	ET	0.870	0.871	-0.001	+0.05	612/553	0.158	0.332
	moy	0.908	0.908	-0.000	+0.03	616/579	0.362	–
HD95 legacy (mm)	WT	5.77	5.51	+0.26	-0.01	284/292	0.838	1.000
	TC	6.02	5.89	+0.13	-0.05	206/222	0.384	1.000
	ET	11.36	11.54	-0.18	+0.01	183/174	0.873	1.000
	moy	7.72	7.65	+0.07	-0.02	400/393	0.646	–

Tableau 4 — Détection de lésions (Kervadec vs DistMap)

Δ = Kervadec − DistMap. r = corrélation rang-bisériale appariée (+ favorise Kervadec). * p brut < 0,05 ; † p corrigé de Holm < 0,05 (sur les 3 régions). n = 1196.

Métrique	Région	Kervadec	DistMap	Δ	r	gagnés/perdus	p	p(Holm)
Lésions FP	WT	0.334	0.297	+0.037	-0.10	116/132	0.131	0.131
	TC	0.122	0.188	-0.066	+0.39	61/34	0.002*	0.004†
	ET	0.641	0.836	-0.195	+0.26	110/73	0.001*	0.004†
	moy	0.366	0.440	-0.075	+0.08	206/185	0.123	–
Lésions FN	WT	0.079	0.075	+0.004	-0.34	4/9	0.166	0.472
	TC	0.033	0.030	+0.003	-0.50	2/6	0.157	0.472
	ET	0.038	0.035	+0.003	-0.03	4/7	0.366	0.472
	moy	0.050	0.047	+0.003	-0.30	10/16	0.064	–
Sensibilité	WT	0.929	0.931	-0.002	-0.20	486/706	2.4e-09*	7.3e-09†
	TC	0.919	0.921	-0.002	-0.08	551/631	0.018*	0.018†
	ET	0.876	0.880	-0.004	-0.15	494/659	1.6e-05*	3.2e-05†
	moy	0.908	0.910	-0.002	-0.18	495/700	1.1e-07*	–
Spécificité	WT	1.000	1.000	+0.000	+0.25	705/488	1.2e-13*	3.5e-13†
	TC	1.000	1.000	+0.000	+0.17	654/529	6.6e-07*	6.6e-07†
	ET	1.000	1.000	+0.000	+0.23	684/498	3.2e-12*	6.5e-12†
	moy	1.000	1.000	+0.000	+0.27	732/463	2.6e-16*	–

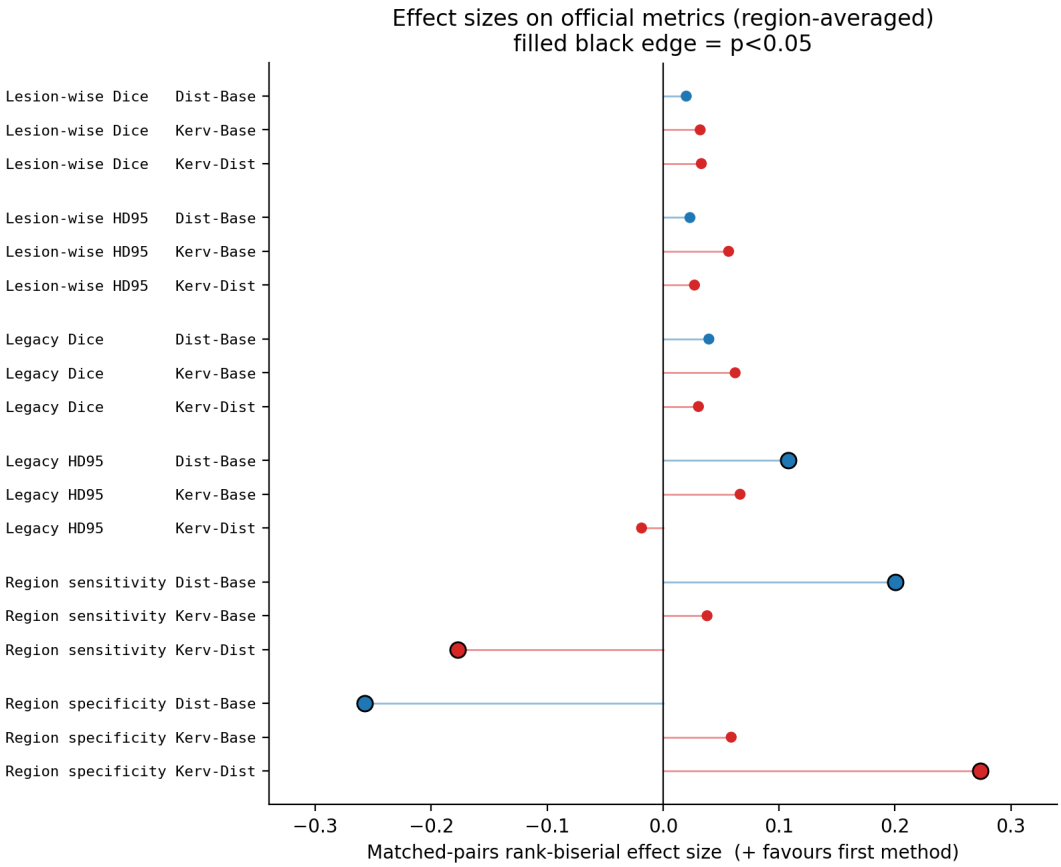

Figure 1: Figure 1 — Tailles d'effet rang-bisériel (moyennées par région) pour chaque comparaison, avec intervalles de confiance : la loss de frontière reste dans le bruit face à Baseline sur les critères principaux, et se distingue de la tête SDT sur la spécificité et les lésions fallacieuses.

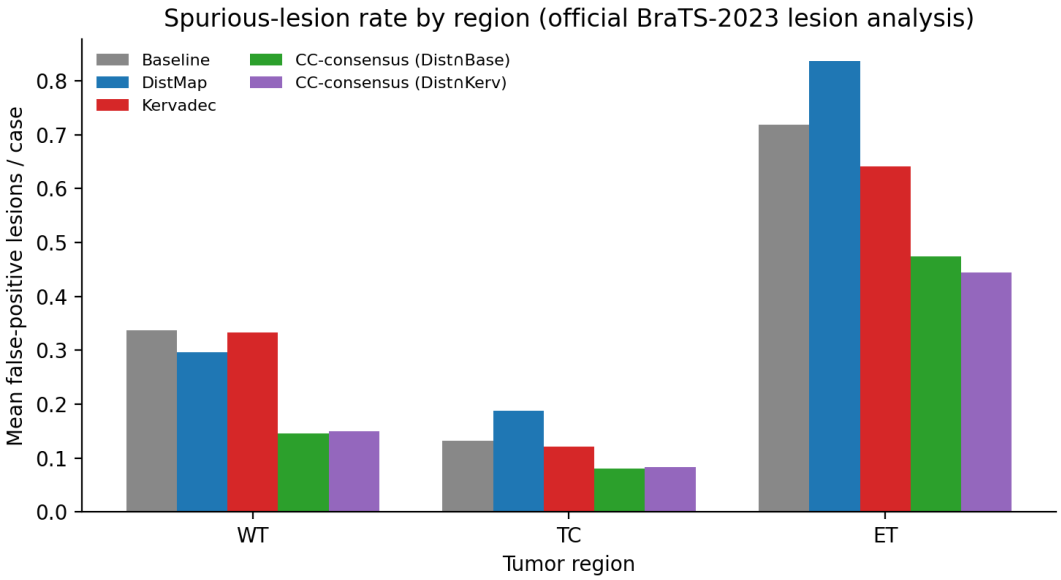

Figure 2: Figure 2 — Taux de lésions fallacieuses par région × méthode : signature du biais de précision de la loss de frontière (moins de lésions FP TC/ET que la tête SDT, au prix du rappel).

Trois patients épinglés du viewer 3D compagnon (https://guillaume-cassez.fr/imagerie-medicale/brats/2023-distance-map/viewer/, régime officiel : fond blanc, nettoyage de composantes du barème BraTS-2023, couleurs de régions du viewer) incarnent ces comptes au niveau patient : un cas où la loss de frontière sauve la segmentation quand les deux autres modèles échouent (Figure 3), un cas où elle échoue et Baseline domine (Figure 4), et un cas où la tête SDT sauve (Figure 5).

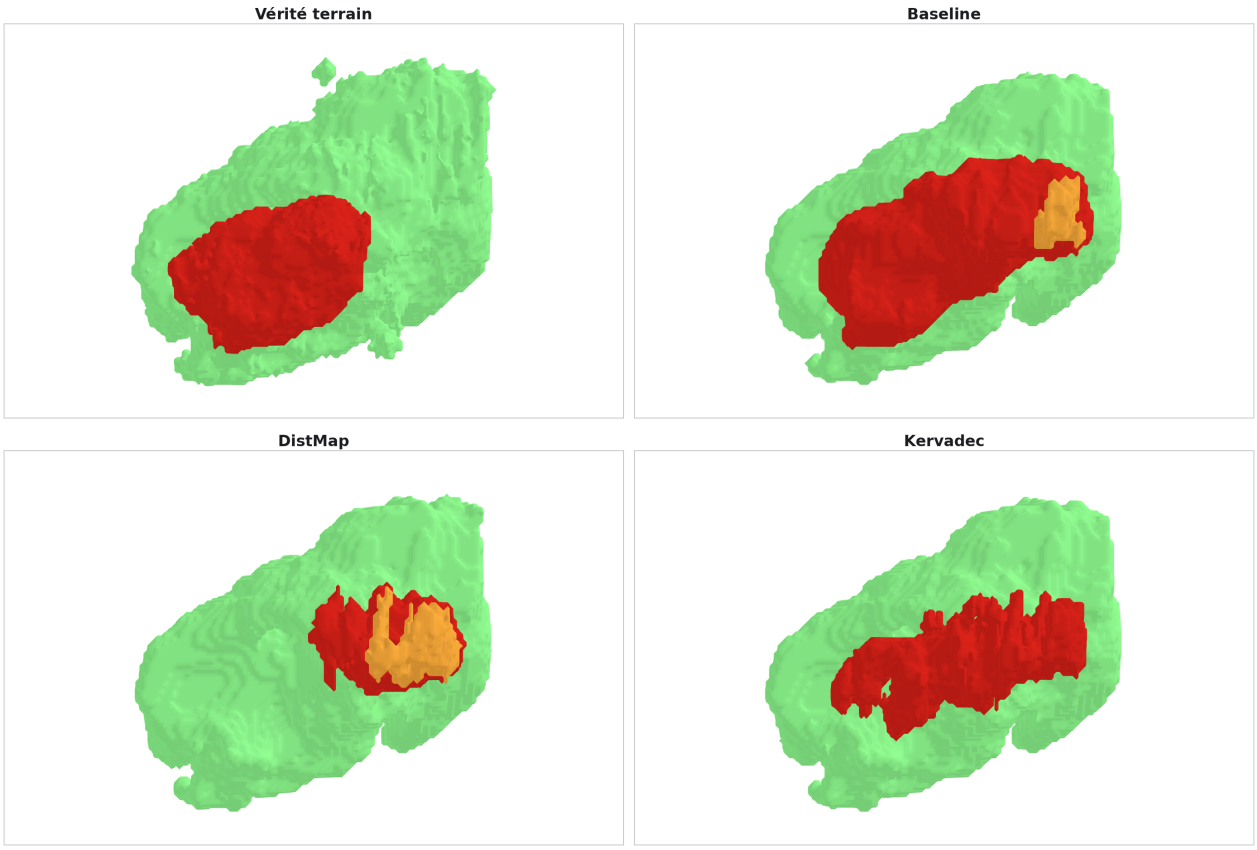

Figure 3: Figure 3 — Cas épinglé P2A (`BraTS-GLI-01486-000`, fold 1) : **Kervadec sauve.** Kervadec (0,822) ≫ Baseline (0,526) > DistMap (0,324) : la loss de frontière capture la tumeur là où les deux autres échouent — 415/1196 patients (34,7 %) voient Kervadec battre à la fois Baseline et DistMap en Dice lésion-wise (régime officiel clean, moyenne des 3 régions). Comptes patient et scores des cas épinglés : `data/patient_level_clean_BKD.json`. Rendu exact du viewer 3D compagnon : fond blanc, filtre officiel, vue sagittale, cerveau masqué.

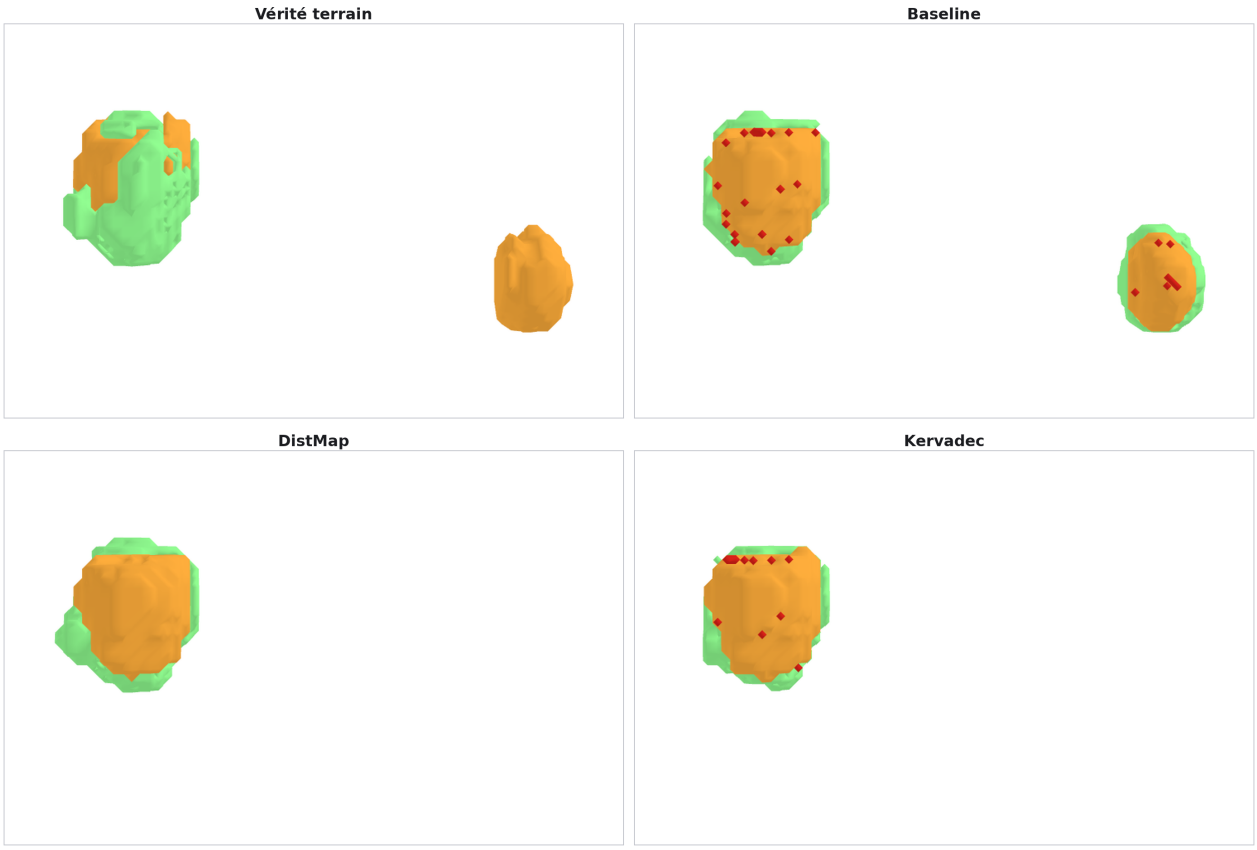

Figure 4: Figure 4 — Cas épinglé P2B (`BraTS-GLI-00732-001`, fold 4) : **Kervadec échoue.** Baseline (0,897) ≫ Kervadec (0,476) ≈ DistMap (0,473) — le miroir du cas précédent : 346 patients (28,9 %) voient Baseline battre les deux losses auxiliaires. Trop stricte sur le contour, la loss de frontière manque ici des portions de tumeur que la CE + Dice régionale capture ; sur ce patient, aucune loss auxiliaire n'aide. Comptes patient et scores des cas épinglés : `data/patient_level_clean_BKD.json`.

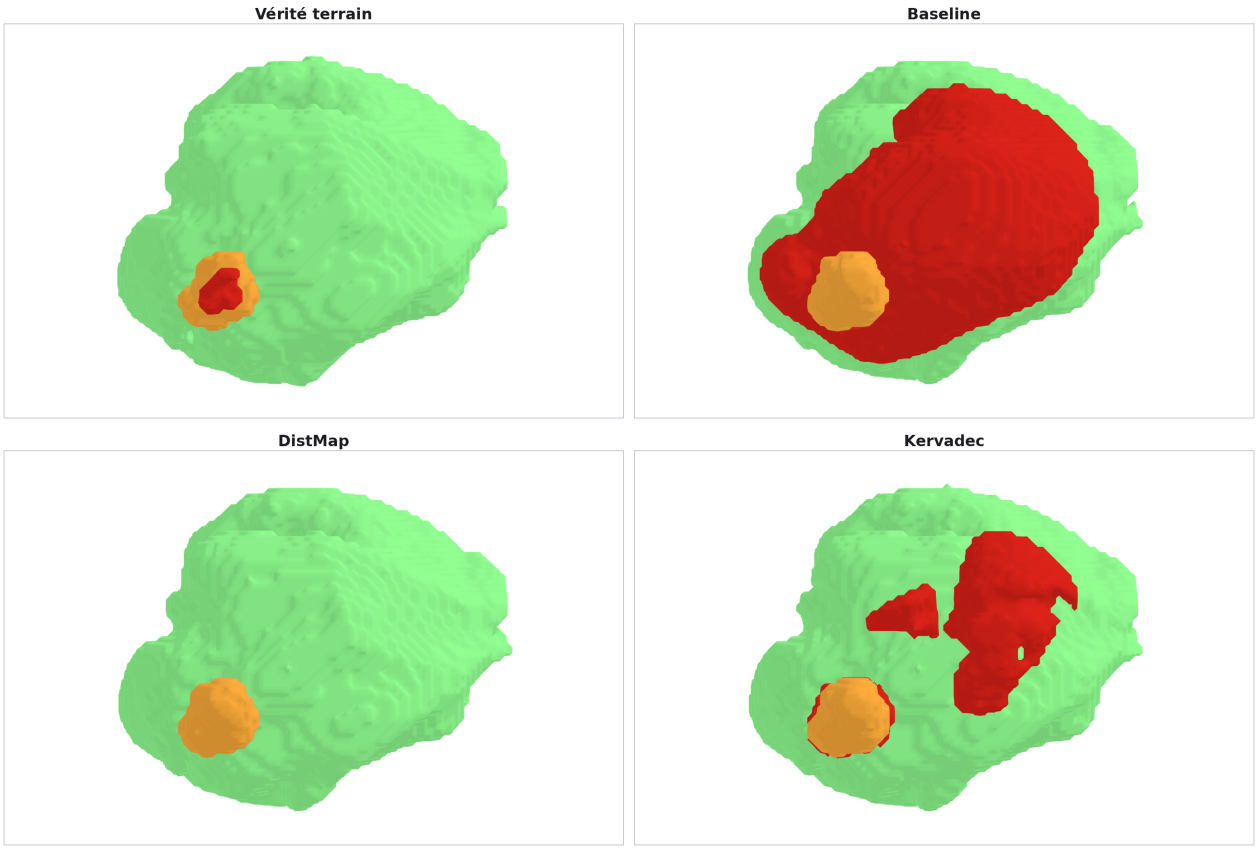

Figure 5: Figure 5 — Cas épinglé P2C (`BraTS-GLI-00525-000`, fold 4) : **DistMap sauve.** DistMap (0,905) > Kervadec (0,698) > Baseline (0,623) — 433 patients (36,2 %) voient DistMap battre les deux autres experts : la régression de carte de distance signée rappelle des structures que la perte régionale lisse. Comptes patient et scores des cas épinglés : `data/patient_level_clean_BKD.json`.

4.3 Kervadec comme veto de consensus

Sa haute spécificité fait de Kervadec un veto efficace dans le filtre CC-consensus étudié dans l'article compagnon DistMap. Filtrer les composantes de DistMap par accord de Kervadec (`CC(D∩K)`) retire ~la moitié des lésions faux positifs de DistMap (0,44 → 0,23/cas) tout en préservant le rappel un peu mieux qu'un veto Baseline (lésions ratées ajoutées 0,006 vs 0,009/cas), et donne une configuration qui bat significativement Baseline sur les métriques de classement officielles (HD95 lesion-wise −9,95 mm, p = $2{,}3\times10^{-20}$). La loss de frontière ne contribue donc pas comme amélioration autonome, mais comme **composant de veto orienté précision** — sa haute spécificité est exactement la propriété requise d'un veto.

4.4 Robustesse à la graine d'entraînement (3 graines, fold 0)

L'évaluation CV (§4.1–4.3) utilise la graine 42 sur les 5 folds. Pour distinguer un signal robuste d'une fluctuation d'initialisation et pour mesurer la variance inter-graine de la loss de frontière, trois entraînements indépendants (graines 1, 2, 3) sont conduits pour Baseline et Kervadec (λ = 0,05) sur le fold 0 (n $\approx$ 240 patients), 300 epochs chacun. Ce protocole reproduit l'analyse multi-graine de l'article compagnon DistMap (Paper 1, §4.6).

Dice de validation fold 0 (nnU-Net, 240 patients) :

Config	Graine 1	Graine 2	Graine 3	Moy. ± σ
Baseline	0,9078	0,9066	0,9078	**0,9074 ± 0,0007**
Kervadec (λ=0,05)	0,9062	0,9078	0,9042	**0,9060 ± 0,0018**
Δ (B − K)	+0,16 pp	−0,12 pp	+0,36 pp	+0,13 pp

Test t apparié (n=3 graines) : t = −1,02, p = 0,42 — non significatif (artefact `data/nnunet_validation_dice_multiseed.json` : les six logs d'entraînement et les deux calculs). Conforme à la CV (§4.1), Kervadec et Baseline sont indiscernables en Dice de chevauchement ; l'écart moyen (+0,13 pp en faveur de Baseline) reste bien inférieur à 1 pp et au seuil de significativité.

Métriques officielles BraTS-2023 et CC-consensus (K∩B) par graine. Évaluation complète sur les 3×240 patients (n $\approx$ 720 paires valides ; ~5 % d'échecs *BraTS-Metrics*) avec les métriques officielles *BraTS-2023-Metrics* (Legacy Dice, LW Dice, HD95). Sur Legacy Dice, Kervadec et Baseline restent statistiquement indiscernables (p > 0,40 Wilcoxon). Le filtre CC-consensus K∩B (composantes de Kervadec recouvrant une prédiction Baseline) améliore significativement **LW Dice WT** (+3,89 pp, p = 0,005) et **LW HD95 WT** (−15,66 mm, p = 0,001).

Tableau A — Baseline vs Kervadec (λ=0,05), métriques officielles, fold 0, 3 graines

Métrique	Région	Baseline (moy.±σ)	Kervadec (moy.±σ)	Δ (K−B)	p t-test	p Wilcoxon
Legacy Dice	WT	0,9361 ± 0,0006	0,9354 ± 0,0014	−0,06 pp	0,66	0,86
Legacy Dice	TC	0,9208 ± 0,0006	0,9211 ± 0,0027	+0,03 pp	0,89	0,94
Legacy Dice	ET	0,8668 ± 0,0029	0,8625 ± 0,0007	−0,43 pp	0,23	0,41
LW Dice	WT	0,8103 ± 0,0080	0,8231 ± 0,0021	+1,29 pp	0,12	0,70
LW Dice	TC	0,8670 ± 0,0067	0,8779 ± 0,0024	+1,10 pp	0,09	0,34
LW Dice	ET	0,8071 ± 0,0109	0,8123 ± 0,0009	+0,52 pp	0,54	0,98
Legacy HD95	WT	6,42 ± 0,28 mm	6,57 ± 0,12 mm	+0,15 mm	0,40	0,89
Legacy HD95	TC	5,95 ± 0,90 mm	6,06 ± 0,63 mm	+0,10 mm	0,92	0,97
Legacy HD95	ET	13,85 ± 1,33 mm	15,02 ± 0,65 mm	+1,17 mm	0,14	0,37
LW HD95	WT	54,5 ± 2,8 mm	49,2 ± 1,4 mm	−5,35 mm	0,058	0,54
LW HD95	TC	26,8 ± 2,7 mm	22,4 ± 0,6 mm	−4,39 mm	0,099	0,34
LW HD95	ET	41,6 ± 4,7 mm	40,3 ± 0,7 mm	−1,32 mm	0,70	0,95

Aucune différence significative Baseline/Kervadec sur les 12 comparaisons (p > 0,05 pour tous les tests). Deux tendances sous-significatives vont dans le sens d'une frontière mieux placée : LW HD95 WT −5,35 mm (p = 0,058) et LW HD95 TC −4,39 mm (p = 0,099) — cohérentes avec le caractère spécifique de la loss, sans atteindre le seuil avec n = 3 graines.

Tableau B — CC-consensus (K∩B) vs Baseline, métriques officielles, fold 0, 3 graines

Métrique	Région	Baseline (moy.±σ)	CC(K∩B) (moy.±σ)	Δ	p t-test	p Wilcoxon
Legacy Dice	WT	0,9361 ± 0,0006	0,9356 ± 0,0014	−0,05 pp	0,74	0,93
Legacy Dice	TC	0,9208 ± 0,0006	0,9215 ± 0,0028	+0,07 pp	0,76	0,68
Legacy Dice	ET	0,8668 ± 0,0029	0,8630 ± 0,0005	−0,38 pp	0,21	0,76
LW Dice	WT	0,8103 ± 0,0080	**0,8492 ± 0,0016**	**+3,89 pp**	**0,026**	**0,005**
LW Dice	TC	0,8670 ± 0,0067	0,8776 ± 0,0019	+1,06 pp	0,11	0,47
LW Dice	ET	0,8071 ± 0,0109	0,8120 ± 0,0005	+0,49 pp	0,58	0,43
Legacy HD95	WT	6,42 ± 0,28 mm	6,57 ± 0,10 mm	+0,15 mm	0,44	0,90
Legacy HD95	TC	5,95 ± 0,90 mm	6,09 ± 0,60 mm	+0,13 mm	0,90	0,37
Legacy HD95	ET	13,85 ± 1,33 mm	15,05 ± 0,63 mm	+1,20 mm	0,14	0,31
LW HD95	WT	54,5 ± 2,8 mm	**38,9 ± 0,8 mm**	**−15,66 mm**	**0,017**	**0,001**
LW HD95	TC	26,8 ± 2,7 mm	22,7 ± 0,7 mm	−4,14 mm	0,13	0,64
LW HD95	ET	41,6 ± 4,7 mm	40,5 ± 0,5 mm	−1,06 mm	0,77	0,77

Deux conclusions :

- **Kervadec régularise la dispersion lesion-wise.** Sur le Dice de chevauchement (pseudo-Dice nnU-Net et Legacy Dice), les trois configurations ont une stabilité comparable, Kervadec se plaçant entre Baseline et DistMap. Mais sur les métriques *lesion-wise* officielles — celles qui pénalisent les fragments —, Kervadec présente une variance inter-graine nettement inférieure à Baseline : σ(LW Dice) passe de 0,0080→0,0021 (WT, ×3,9), 0,0067→0,0024 (TC, ×2,8) et 0,0109→0,0009 (ET, ×12). C'est **l'opposé** de DistMap, plus instable que Baseline sur le même protocole (Paper 1, §4.6 : σ nnU-Net ×4). En réprimant les composantes fallacieuses, la loss de frontière resserre la dispersion lesion-wise : la signature de son biais de précision (haute spécificité, peu de faux positifs, §4.2).
- **Le filtre K∩B reproduit le gain WT du filtre D∩B.** En gardant les composantes connexes de Kervadec recouvrant une prédiction Baseline (K∩B), le filtre de consensus améliore significativement LW Dice WT (+3,89 pp, p = 0,005 Wilcoxon) et LW HD95 WT (−15,66 mm, p = 0,001), sans coût sur Legacy Dice. Ce gain reproduit de près celui du filtre D∩B sur le même protocole (Paper 1, §4.6 : +4,28 pp / −17,2 mm) : la règle de consensus produit l'amélioration lesion-wise WT **indépendamment de la loss auxiliaire servant de modèle primaire**. Le gain est porté par le consensus de composantes connexes, non par la loss. Cela complète §4.3 : Kervadec est utile en consensus aussi bien comme **veto** (D∩K, filtrer DistMap) que comme **modèle primaire** (K∩B, filtré par Baseline).

5. Discussion

Une loss de frontière à poids fixe n'aide pas un backbone performant. Avec supervision profonde et objectif Dice + CE déjà présents, une pénalité de contour ajoutée laisse chevauchement, frontière régionale et détection inchangés vs Baseline. C'est un résultat négatif utile qui borne le bénéfice attendu des pénalités de frontière « bolt-on » pour les backbones gliome 3D de l'état de l'art — rapporté de façon transparente, avec un contrôle de sélection du poids qui exclut le mauvais réglage. L'analyse multi-graine confirme : à convergence sur 3 graines indépendantes, l'écart Baseline/Kervadec est nul (p = 0,42).

Son trait distinctif est la précision, visible seulement par contraste. Face à la tête SDT, la loss de frontière est significativement plus spécifique et produit moins de lésions fallacieuses TC/ET, au prix du rappel. Lues ensemble avec Paper 1, les deux losses auxiliaires sont des images miroir sur un axe rappel–précision : la tête SDT échange spécificité contre rappel, la loss de frontière échange rappel contre spécificité, et aucune n'améliore les métriques de classement officielles vs Baseline. Ce contraste — pas une victoire sur Baseline — est la contribution. La signature multi-graine le confirme : la loss de frontière *régularise* la dispersion lesion-wise là où la tête SDT l'amplifie.

Le mécanisme du gain est le consensus, pas la loss. Que le gain lesion-wise WT apparaisse à l'identique pour D∩B (Paper 1) et K∩B (§4.4), avec deux losses auxiliaires aux biais opposés, établit que le levier est le filtre de consensus de composantes connexes — un post-traitement sans paramètre — et non la loss auxiliaire elle-même. La loss auxiliaire ne sert qu'à fournir un second avis dont l'accord avec Baseline sélectionne les composantes fiables.

6. Limites

La comparaison principale (vs Baseline) est nulle ; on la présente comme telle. Les avantages de spécificité / FP ne tiennent que face à la tête SDT et sont exploratoires (critères principaux nuls), bien que robustes ($p \leq 10^{-7}$, survivant à Holm). Sensibilité / spécificité régionales sont au niveau voxel. L'analyse multi-graine porte sur le seul fold 0 avec n = 3 graines : elle estime la variance d'initialisation, pas la variance inter-fold, et sa puissance pour les tendances sous-significatives (LW HD95 WT/TC) est faible. Un seul poids fixe a été porté en CV complète, justifié par le contrôle de robustesse du balayage.

7. Conclusion

Sur un backbone gliome MedNeXt / nnU-Net performant, une loss de frontière de Kervadec à poids fixe ne donne **aucune amélioration significative vs Baseline** sous quelque métrique officielle BraTS-2023 que ce soit, ce que confirme une analyse multi-graine indépendante (3 graines × fold 0, p = 0,42). Sa seule propriété robuste et distinctive est d'être le **miroir orienté précision** d'une tête auxiliaire de distance signée — plus spécifique, moins de lésions fallacieuses, rappel plus faible — propriété qui en fait aussi un *régularisateur* de la variance lesion-wise et un *veto* de consensus efficace. Le gain mesurable ne vient pas de la loss mais du filtre de consensus de composantes connexes, qui produit la même amélioration lesion-wise WT que la loss serve de veto (D∩K) ou de modèle primaire (K∩B). Rapporté en entier, ce travail délimite quand les losses de frontière aident (pas ici, sur un backbone fort) et ce qu'elles changent réellement (le point de fonctionnement rappel–précision, et la stabilité d'entraînement).

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 : `references.bib` (toutes vérifiées ; livrée dans cette archive).

Références

- 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.
- Bakas, Akbari, Sotiras, Bilello, Rozycki, Kirby et al. (2017). *Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features*. Scientific Data. doi:10.1038/sdata.2017.117.
- Dangi, Linte, Yaniv (2019). *A distance map regularized CNN for cardiac cine MR image segmentation*. Medical Physics. doi:10.1002/mp.13853.
- 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.
- 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.
- 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).
- 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.
- Milletari, Navab, Ahmadi (2016). *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. 2016 Fourth International Conference on 3D Vision (3DV). doi:10.1109/3DV.2016.79.
- 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.

- Nikolov, Blackwell, Zverovitch, Mendes, Livne, De Fauw et al. (2021). *Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study*. Journal of Medical Internet Research. doi:10.2196/26151.
- 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*. .
- Sudre, Li, Vercauteren, Ourselin, Jorge Cardoso (2017). *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS 2017). doi:10.1007/978-3-319-67558-9_28.
- 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.