Contents

Loss auxiliaire de type *distance map* pour la segmentation de tumeurs cérébrales : analyse centrée sur les fragments et plafond de saturation du filtrage post-hoc par consensus de composantes connexes

Guillaume Cassez

Recherche indépendante · ORCID 0009-0007-0987-3931 · `cassez.guillaume@gmail.com` · guillaume-cassez.fr

BraTS 2023 GLI · nnU-Net v2 · MedNeXt-B · 1196 patients de validation

---

Résumé

On étudie l'ajout d'une loss auxiliaire de type *distance map* (SDT — Signed Distance Transform) sur un pipeline MedNeXt-B / nnU-Net v2 pour la segmentation 3D de tumeurs cérébrales sur BraTS 2023 GLI. À convergence sur 1196 patients en validation croisée 5-fold, la tâche SDT **n'améliore pas significativement** le Dice (Δ Dice avg = +0,09 pp, Wilcoxon $p > 0{,}25$ par région) — un résultat qui ouvre l'analyse plutôt qu'il ne la conclut.

DistMap introduit un nouveau mode de défaillance, jusqu'ici non rapporté dans la littérature BraTS : des **composantes connexes isolées fallacieuses** (« fragments ») absentes de la vérité terrain, particulièrement marquées sur NCR (×1,5 vs Baseline) et ED (×1,2). On propose un **filtre de consensus de composantes connexes** (CC-consensus) post-hoc et sans paramètre, qui supprime toute composante DistMap sans recouvrement Baseline dans la même classe. Sur 1196 patients en CV 5-fold, le filtre élimine **66 % des fragments NCR** (Wilcoxon $p < 10^{189}$, définition topologique : CC − 1 par classe) sans coût en Dice, et **améliore significativement HD95 sur NCR** (4,86 → 4,48 mm, $p = 5{,}7 \times 10^{14}$) et WT (3,86 → 3,76 mm, $p = 2{,}7 \times 10^{4}$). Cliniquement, NCR est précisément la région où des fragments fallacieux peuvent induire en erreur un radiothérapeute sur l'emprise de la nécrose tumorale ; le gain de qualité de frontière mesuré ici est caché par le Dice mais visible via HD95.

Une étude du plafond hard-label montre que cette règle est déjà proche de la saturation : l'oracle par classe n'est qu'à +0,005 Dice avg au-dessus du défaut, et aucun meta-selector à 31 features (4 familles de classifieurs) ne bat robustement le CC-consensus en CV 5-fold (Annexe B). Combler cet écart nécessite un vote probabiliste au niveau voxel ou une diversité architecturale, ce qui motive un Paper 2 vers une loss sensible aux fragments à l'entraînement plutôt que davantage d'ingénierie post-hoc.

Contributions. (1) Caractérisation quantitative d'un artefact topologique de fragments induit par la loss SDT auxiliaire — invisible au Dice, prévalent sur NCR — avec une définition topologique sans seuil de taille, à grande échelle (1196 patients). (2) Un filtre CC-consensus simple et sans paramètre qui élimine 66 % des fragments NCR sans coût en Dice et **améliore significativement HD95 NCR** ($p = 5{,}7 \times 10^{14}$), un gain de qualité de frontière cliniquement pertinent caché par le Dice.

---

1. Introduction

La segmentation de tumeurs cérébrales sur IRM multi-modalités (challenge BraTS) est dominée ces dernières années par des dérivés de nnU-Net [Isensee 2021]. La tâche canonique est une classification 3D de voxels en quatre classes : fond, cœur nécrotique (NCR, label 1), œdème péri-tumoral (ED, label 2) et tumeur rehaussée (ET, label 3). La performance est habituellement rapportée via des coefficients de Dice sur trois régions emboîtées : WT = {1,2,3}, TC = {1,3}, ET = {3}.

Les équipes les plus performantes raffinent le backbone (MedNeXt [Roy MICCAI 2023], Swin-UNETR) tout en laissant la loss d'entraînement quasi inchangée : Dice + cross-entropy. En parallèle, la **régression auxiliaire de *distance maps*** [Ma MIDL 2020 ; Xue AAAI 2020] est régulièrement proposée pour rendre le réseau sensible à la forme, avec des résultats empiriques mitigés. Des applications spécifiques à BraTS existent — multi-tâche à décodeurs parallèles [Huang 2021], losses Hausdorff-aware [Karimi & Salcudean 2020], et formulations géodésiques « régression seule » [Pham 2024, SiNGR] — mais aucune à ce jour ne rapporte ni n'analyse l'artefact de fragments caractérisé ici (§5.2).

Ce papier poursuit trois objectifs :

- **Caractérisation empirique** de la tâche SDT auxiliaire à convergence sur MedNeXt-B / nnU-Net v2 : à 300 epochs en CV 5-fold sur 1196 patients, DistMap ne produit **pas** de gain Dice significatif ($p > 0{,}25$ par région), contrairement à l'impression tirée de comparaisons à budget d'entraînement réduit.
- **Analyse de mode de défaillance** : identification et quantification d'un artefact sous-rapporté de la tâche SDT — la production de composantes connexes petites et isolées qui gonflent les faux positifs sans toucher significativement au Dice. Cette observation qualitative a été rendue possible par un **viewer 3D interactif compagnon** construit spécifiquement pour ce projet, qui rend côte-à-côte les meshes Baseline / DistMap / CC-Consensus pour les 1196 patients (guillaume-cassez.fr/brats/).
- **Analyse du plafond** d'un filtre CC-consensus post-hoc qui corrige cet artefact, avec une étude sur 1196 patients délimitant ce qu'un meta-selector à base de features peut atteindre en l'absence d'accès aux softmax ou de diversité de modèles.

---

2. Travaux connexes

Losses auxiliaires par *distance maps* en segmentation médicale. [Ma 2020] propose une tête de régression SDT auxiliaire pour des structures abdominales / cardiaques (LiTS, LA atrium), établissant la recette tanh + MSE reprise ici. [Xue 2020] utilise des distance maps signées comme **sortie principale** (non auxiliaire) sur des organes, avec $\lambda = 10$ sans ablation. [Karimi & Salcudean 2020] dérivent une loss Hausdorff-aware à partir de distance transforms, évaluée sur nnU-Net + BraTS, mais comme **modification de loss** et non comme tête de régression auxiliaire. Aucun de ces travaux ne signale le phénomène de fragments caractérisé ici.

Approches distance-map spécifiquement appliquées à BraTS. L'idée d'associer une supervision de forme par distance à la segmentation BraTS **n'est pas nouvelle en soi** ; deux travaux antérieurs sont particulièrement proches du dispositif présenté et doivent être signalés explicitement.

- [Huang et al. 2021] entraînent un V-Net avec deux *décodeurs parallèles* sur BraTS 2018–2020 — l'un produisant le masque de segmentation, l'autre régressant une distance transform *non signée* à travers une sigmoid. C'est l'état de l'art le plus proche de ce travail. Le présent travail s'en distingue par trois points concrets : (i) une tête auxiliaire légère `Conv3d(32→3) + tanh` au lieu d'un décodeur parallèle complet (<0,1 % de paramètres ajoutés vs un décodeur dupliqué) ; (ii) distance euclidienne *signée* avec MSE, et non distance non signée avec sigmoid ; (iii) MedNeXt-B / nnU-Net v2 sur BraTS 2023 GLI (1196 patients) au lieu d'un V-Net sur BraTS 2018–2020.
- [Pham et al. 2024, *SiNGR*] proposent une régression **géodésique normalisée signée** avec loss Focal-L1 sur sortie tanh, qui **remplace** la sortie de segmentation sur BraTS 2020 (backbones Swin-UNETR / UNet3D). Le présent travail est multi-tâche (conservation de la sortie Dice + CE softmax à côté de la régression SDT) et utilise la distance euclidienne signée classique, et non une transformée géodésique.

Ni Huang et al. ni SiNGR ne rapportent ou n'analysent l'artefact de fragments décrit en §5.2 de ce papier ; c'est la contribution empirique spécifique revendiquée ici.

Ensembling et fusion. Les gagnants BraTS classiques s'appuient sur l'ensembling 5-fold (soft-voting des softmax). Les règles de sélection de modèle ou de stacking au niveau patient sont peu courantes ; les règles de consensus au niveau des composantes connexes le sont encore moins dans la littérature BraTS publiée.

Analyse des modes de défaillance. Des métriques au niveau composante (F1 lesion-wise) ont été introduites dans le challenge BraTS 2023 mais restent secondaires au Dice / HD95 dans les publications. À notre connaissance, aucun travail antérieur ne quantifie ni ne localise le biais de fragments induit par les losses SDT auxiliaires sur BraTS.

---

3. Méthodes

3.1 Architecture et entraînement

Backbone. MedNeXt-B [Roy MICCAI 2023] ré-implémenté dans nnU-Net v2 avec le plan `nnUNetPlans_96GB_mednext` (patch 128^3, BS 2, BF16, RTX PRO 6000 Blackwell).

Tête auxiliaire. Un unique Conv3D($32 \to 3$, noyau $1 \times 1 \times 1$) + tanh prédisant une carte SDT normalisée pour chacune des régions NCR, ED, ET. La SDT de référence est pré-calculée une fois par patient via `scipy.ndimage.distance_transform_edt` sur chaque masque binarisé de région, signée par sign(intérieur − extérieur), clippée min-max à $[-1, 1]$ avec bord = 0.

Loss. $\mathcal{L} = \mathcal{L}_{\text{Dice+CE}} + \lambda \cdot \mathcal{L}^{\text{SDT}}_{\text{MSE}}$, avec $\lambda = 1$ par défaut (calibration équilibrée par gradient ÷ 5 ; ablation statique sur 11 valeurs détaillée Annexe A).

3.2 Nommage des variantes

Variante	Trainer	Auxiliaire ?
Baseline	`nnUNetTrainerMedNeXtBaseline`	pas de SDT
DistMap	`nnUNetTrainerMedNeXtDistMap`	SDT, $\lambda = 1$

Variante	Trainer	Auxiliaire ?
CC-Consensus	règle post-hoc (§3.3) sur DistMap + Baseline	post-hoc

3.3 Règle de filtrage CC-consensus[1]

Étant données la prédiction Baseline P_B et la prédiction DistMap P_D (tous deux des tenseurs de labels dans {0, 1, 2, 3}), la prédiction filtrée P_F est calculée classe par classe :

```
P_F := copy(P_D)
pour chaque classe c  {1, 2, 3}:
    D_mask  := (P_D == c)
    B_mask  := (P_B == c)
    labeled, n := cc_label(D_mask, structure=connectivité-26)
    pour chaque cc_id  1..n:
        cc := (labeled == cc_id)
        si cc  B_mask =  :
            P_F[cc] := 0        # on supprime le fragment non confirmé
```

La règle a quatre effets qualitatifs :

1. Les fragments DistMap isolés de la classe correspondante Baseline → **supprimés**.
2. Les raffinements de frontière DistMap sans recouvrement avec Baseline → **conservés** (on part toujours de P_D).
3. Les trous Baseline comblés par DistMap → **conservés** (P_D est non nul à ces endroits).
4. Les faux positifs Baseline rejetés par DistMap → **restent rejetés** (P_D est nul à ces endroits).

La règle n'a **aucun paramètre appris** et un seul hyperparamètre (connectivité 26 vs 6), fixé à 26 partout. C'est une opération de *veto* : Baseline n'ajoute aucun voxel nouveau ; il ne peut que supprimer des composantes que DistMap a prédites sans confirmation.

3.4 Analyse du plafond

Pour caractériser le plafond de qualité atteignable par toute politique de sélection au niveau patient ou région sur les trois prédictions disponibles, on définit, pour chaque patient p avec Dice régional $(D^B, D^D, D^F) \in \mathbb{R}^3$ par région $r \in \{\text{WT}, \text{TC}, \text{ET}\}$:

$$\text{Oracle}_{\text{patient}}(p) = \max_{m \in \{B,D,F\}} \tfrac{1}{3} \sum_r D_r^m$$

$$\text{Oracle}_{\text{par-classe}}(p) = \tfrac{1}{3} \sum_r \max_{m \in \{B,D,F\}} D_r^m$$

L'écart entre ces oracles et la moyenne CC-consensus par défaut est le gain maximal atteignable par toute politique de sélection. Les politiques candidates évaluées (seuil taille-adaptatif, meta-classifieurs, règle à une feature) et leurs résultats sont rapportés en §5.4 et détaillés en Annexe B.

---

4. Expériences

4.1 Données

BraTS 2023 GLI (1251 patients, 4 modalités chacun). Pré-traitement via les réglages par défaut de nnU-Net v2 (z-score par patient, cropping automatique, ré-échantillonnage isotrope 1 mm³). Labels de vérité terrain {0, 1, 2, 3}. Partition

[1] Les versions antérieures de ce travail désignaient cette règle par « fusion MoE (Mixture-of-Experts) ». Ce label est abandonné : il n'y a ni réseau de gating appris, ni routage doux des inputs, ni entraînement conjoint experts-gate. Le terme neutre « filtre CC-consensus » est employé dans tout le document.

des patients : validation croisée 5-fold stratifiée par ID. Toutes les métriques ci-dessous sont calculées sur l'ensemble de validation (n = 239 pour le fold 0) ou agrégées sur les 5 folds (n = 1196).

4.2 Métriques

Dice par région (WT, TC, ED, ET) avec la convention standard nnU-Net / MONAI : *Dice = 1 si la GT et la prédiction sont toutes deux vides*. Voir §5.5 pour les précautions à prendre lors de la comparaison aux leaderboards BraTS challenge.

Comptage de fragments (définition topologique). Un **fragment** est une composante connexe (connectivité 26) d'une classe donnée qui n'est **pas la plus grosse** composante de sa classe — c'est-à-dire une CC topologiquement déconnectée du corps tumoral principal. Par classe c sur une prédiction P, le nombre de fragments est :

$$\text{fragments}(P, c) = \max(0,\ \text{nb\_CC}(P == c, \text{26-conn}) - 1)$$

Pas de seuil de taille — la 26-connectivité (faces, arêtes, coins partagés) suffit à définir ce qui est topologiquement lié. Cette définition traite symétriquement petites et grandes composantes accessoires.

Features d'accord inter-modèles (11) : Dice(Baseline, DistMap) pour WT/TC/ET ; différence volumétrique normalisée $||P_B^c| - |P_D^c||/(|P_B^c| + |P_D^c|)$ pour ET et NCR ; nombre / fraction / taille max des CC DistMap sans recouvrement Baseline, pour ET et NCR.

Features morphologiques (20) : volume par région, ratios de volumes, comptage et taille des CC en connectivité 26 pour NCR/ET, élongation du tenseur d'inertie (λ_1/λ_3), sphéricité $(\pi^{1/3}(6V)^{2/3})/S$, rugosité de surface $S_{\text{pred}}/S_{\text{sphère}}$, nombre d'Euler ([scikit-image] `euler_number`, connectivité 3) pour WT/TC/ET, comptage de cavités pour WT (diff `binary_fill_holes`), comptage de CC baseline / distmap par NCR et ET, dispersion des CC d'ET (écart-type des distances centroïdes).

---

5. Résultats

5.1 À convergence, DistMap et Baseline sont équivalents en Dice

Sur les 1196 patients agrégés hors-fold de la CV 5-fold (schedule 300 epochs par fold ; DistMap fold 0 arrêté à 178 ep, les 9 autres entraînements complets), DistMap et Baseline produisent des Dice **statistiquement indiscernables** :

Région	Baseline	DistMap	ΔDice	p-value	Améliorés / dégradés / égaux
WT	0,9354	0,9360	+0,006 pp	0,72	577 / 618 / 1
TC	0,9185	0,9180	−0,005 pp	0,27	595 / 596 / 5
ET	0,8696	0,8723	+0,027 pp	0,54	568 / 596 / 32
Avg	**0,9078**	**0,9088**	**+0,009 pp**	**0,50**	—

Test de Wilcoxon signé apparié, hypothèse unilatérale DistMap $>$ Baseline. Aucune région n'atteint le seuil de significativité standard (p > 0,25 partout) ; sur WT, davantage de patients sont dégradés qu'améliorés par DistMap (618 vs 577). Le Δ = +0,09 pp de Dice avg est dans la variance de mesure.

Implication. La loss SDT auxiliaire, telle que formulée ici (tête Conv3D($32 \to 3$)+tanh, régression MSE, $\lambda = 1$), ne confère pas d'amélioration Dice significative à convergence sur BraTS 2023 GLI. Cela n'exclut pas que DistMap produise des prédictions *différentes* de Baseline : les deux modèles divergent sur 1195/1196 patients (1 seule égalité stricte en Dice avg), mais leurs désaccords se compensent en moyenne sur le Dice global. Cette différence de topologie sans magnitude Dice motive l'analyse de fragments qui suit.

5.2 DistMap introduit des fragments fallacieux

L'inspection qualitative des prédictions DistMap visait des frontières plus nettes — comportement attendu d'une loss sensible à la distance. Au lieu de cela, les prédictions DistMap montrent systématiquement davantage de composantes

connexes isolées que Baseline. Quantification topologique sur les 1196 patients de la CV 5-fold (moyennes par patient, fragments = CC − 1 par classe, 26-connectivité) :

Fragments / patient	Baseline	DistMap	**CC-Consensus**	Δ D−B	Δ F−D	Réduction F/D
NCR	79,7	93,3	**31,3**	+13,6	−61,9	**−66 %**
ED	28,9	35,3	**17,0**	+6,4	−18,3	**−52 %**
ET	2,15	2,33	**1,57**	+0,18	−0,76	**−33 %**

Tests de Wilcoxon signés unilatéraux sur les 1196 patients :

- **DistMap inflate les fragments vs Baseline** sur les 3 classes : NCR ($p = 5{,}5 \times 10^{-42}$), ED ($p = 2{,}0 \times 10^{-49}$), ET ($p = 1{,}3 \times 10^{-3}$). L'artefact est statistiquement massif et systématique.
- **CC-Consensus réduit les fragments vs DistMap** : NCR ($p < 10^{-189}$), ED ($p < 10^{-162}$), ET ($p = 1{,}1 \times 10^{-53}$).
- **CC-Consensus réduit aussi vs Baseline** : NCR ($p < 10^{-188}$), ED ($p < 10^{-144}$), ET ($p = 1{,}4 \times 10^{-3}$) — le filtre post-hoc corrige même les fragments hérités du Baseline quand DistMap n'y avait pas d'overlap.

Cet effet **ne se voit pas sur le Dice** (§5.3 : Dice moyens B / D / F à 0,9078 / 0,9088 / 0,9090, différences dans le bruit) — des fragments de quelques voxels n'impactent pas une métrique de recouvrement quand le volume tumoral médian fait ~90 000 voxels. C'est précisément pourquoi la littérature passée n'avait pas rapporté l'artefact : le Dice est aveugle à la topologie.

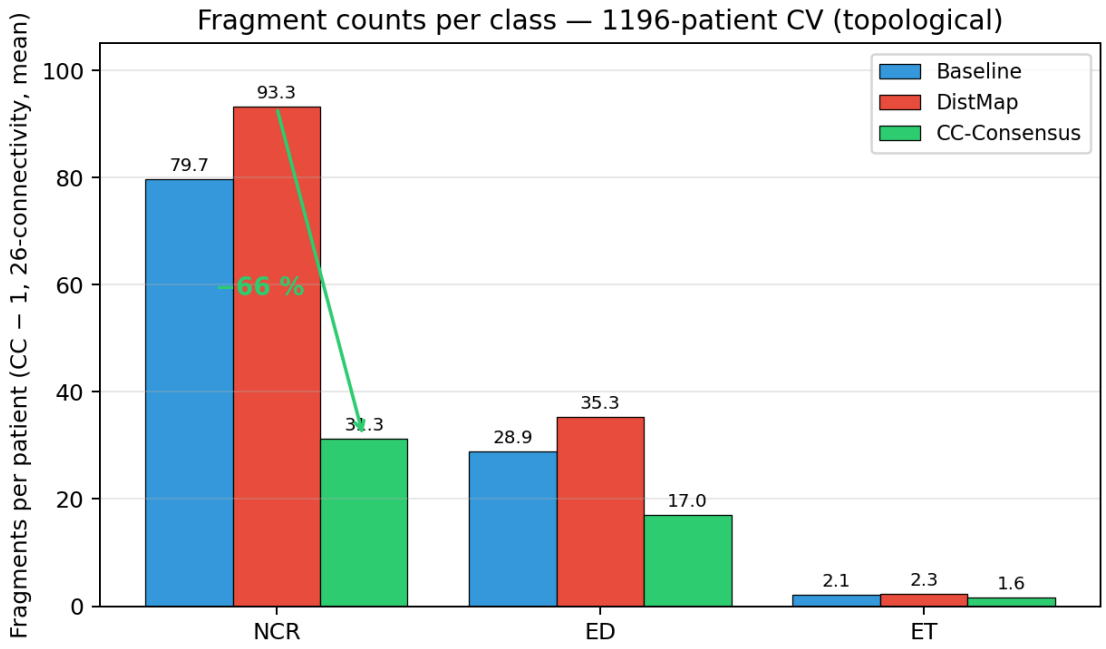

Figure 1: Figure 1 — Comptage moyen de fragments par patient (composantes connexes non-principales, 26-connectivité, sans seuil de taille) pour chaque classe × variante, sur les 1196 patients de la CV 5-fold. DistMap inflate le nombre de fragments NCR de +17 % par rapport à Baseline ; le filtre CC-consensus le ramène à 31,3 — une réduction de **66 %** par rapport à DistMap (Wilcoxon $p < 10^{-189}$).

Illustrations qualitatives sur les six cas de référence. Les figures 2–7 ci-dessous montrent, pour chacun des six patients épinglés (C1–C6) du viewer 3D compagnon, les segmentations produites par GT / Baseline / DistMap / CC-Consensus, vue sagittale gauche, régions tumorales seules (Brain masqué pour focus). Chaque figure illustre l'un des six modes de comportement identifiés en Annexe E.

Note sur le rendu 3D (deux pipelines). Le viewer propose un mode lissé et un mode voxel, chacun servi par un pipeline distinct selon la nature du mesh.

Mode voxel (vérité brute). Greedy voxel meshing : chaque voxel de la segmentation est converti en une face cubique fusionnée avec ses voisins coplanaires. Aucune interpolation, aucun lissage — ce que le modèle a prédit au voxel près. Sert de référence de vérité quand on veut compter ou localiser précisément.

Mode lissé (défaut, figures 2–7), meshes principaux. Pipeline `fill_holes + dilation + marching cubes` : le masque binaire est pré-rempli (`scipy.ndimage.binary_fill_holes` pour supprimer les cavités internes — ventricules, sulci), dilaté d'un voxel (`binary_dilation`, 1 itération) pour adoucir les escaliers du marching cubes, puis marching-cubed au seuil 0,5. C'est le pipeline utilisé pour les meshes des corps tumoraux et du Brain (figures 2–7).

Mode lissé, fragments et cavités. Pour les petites composantes (< 4 voxels jusqu'aux fragments sub-voxel), un pipeline **champ de distance signée** (*signed distance field*) distinct est employé : dilatation 26-connectivité pour bridger les voxels touchant par coin/arête, transformée de distance euclidienne intérieure et extérieure (`scipy.ndimage.distance_transform_edt`) pour construire le champ de distance signée, upsampling spline cubique ×2 pour résolution sub-voxel, puis marching cubes au niveau iso = −0,3 (calibré empiriquement pour préservation volumique). Ce pipeline est **nécessaire pour les petits fragments** car un marching cubes naïf au seuil 0,5 sur un masque 1-voxel rend 1/6 du volume réel (erreur ×6) alors que le champ de distance signée préserve le volume à ±5 % sur toutes les tailles.

Propriété commune aux deux pipelines lisses. Ils **préservent la topologie** (mêmes composantes connexes, même comptage en 26-connectivité que le mode voxel) ; la différence est purement cosmétique. Le lissage est le défaut parce qu'il produit un rendu proche de la console clinique ; le mode voxel reste un clic de distance pour toute inspection qui requiert la vérité voxel-exacte.

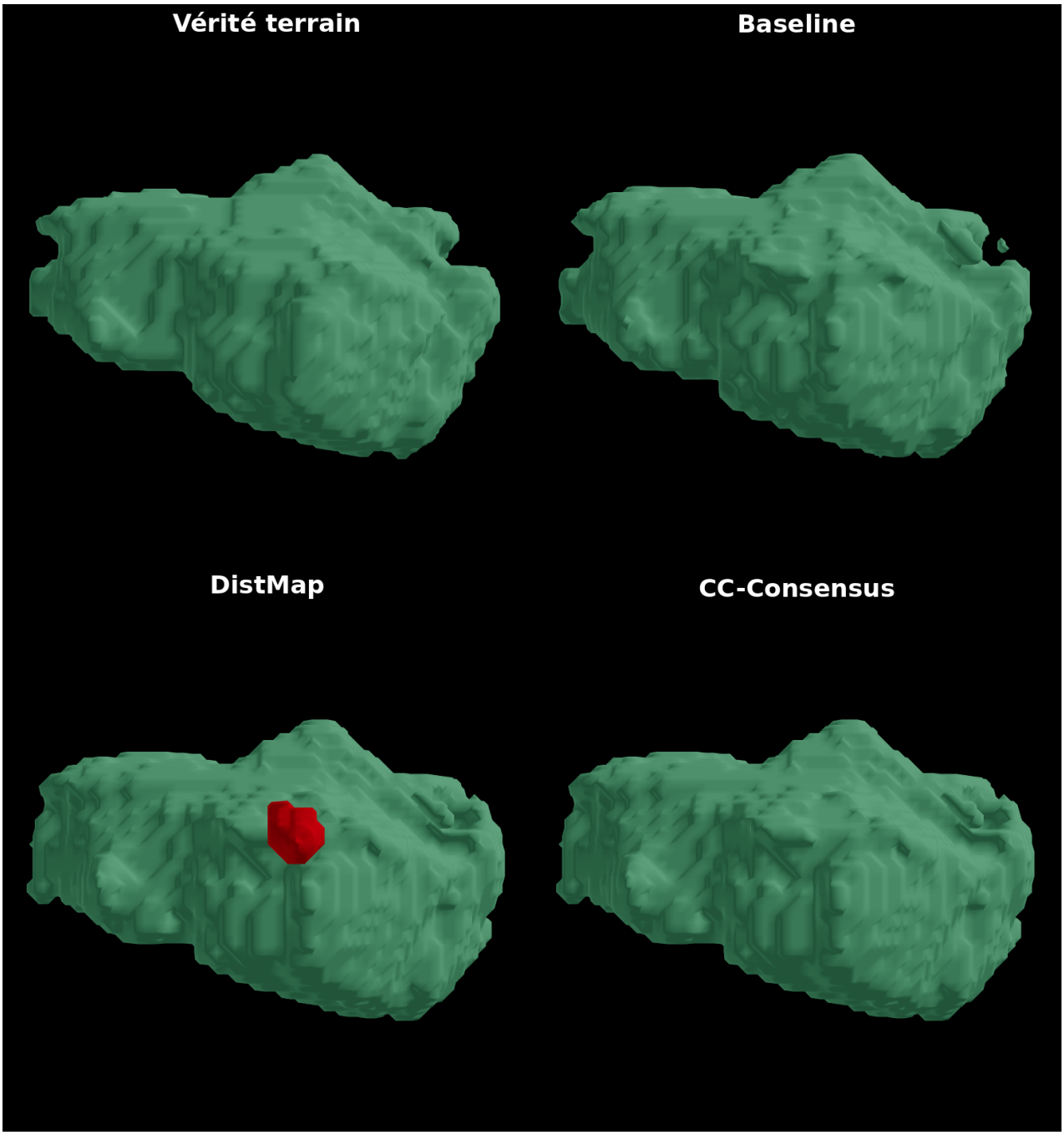

Figure 2: Figure 2 — Cas **C1** (patient `BraTS-GLI-00048-001`) : Baseline > DistMap. La GT ne contient que de l'œdème (vert) ; Baseline reproduit correctement ce pattern. **DistMap hallucine une masse NCR** (rouge) au sein de l'œdème — typique des cas où la pression SDT engendre des composantes fallacieuses. **CC-Consensus supprime cette hallucination** car la composante NCR de DistMap n'a aucun recouvrement avec la prédiction Baseline (veto), restaurant quasi intégralement le score (Dice avg 0,308 → 0,973).

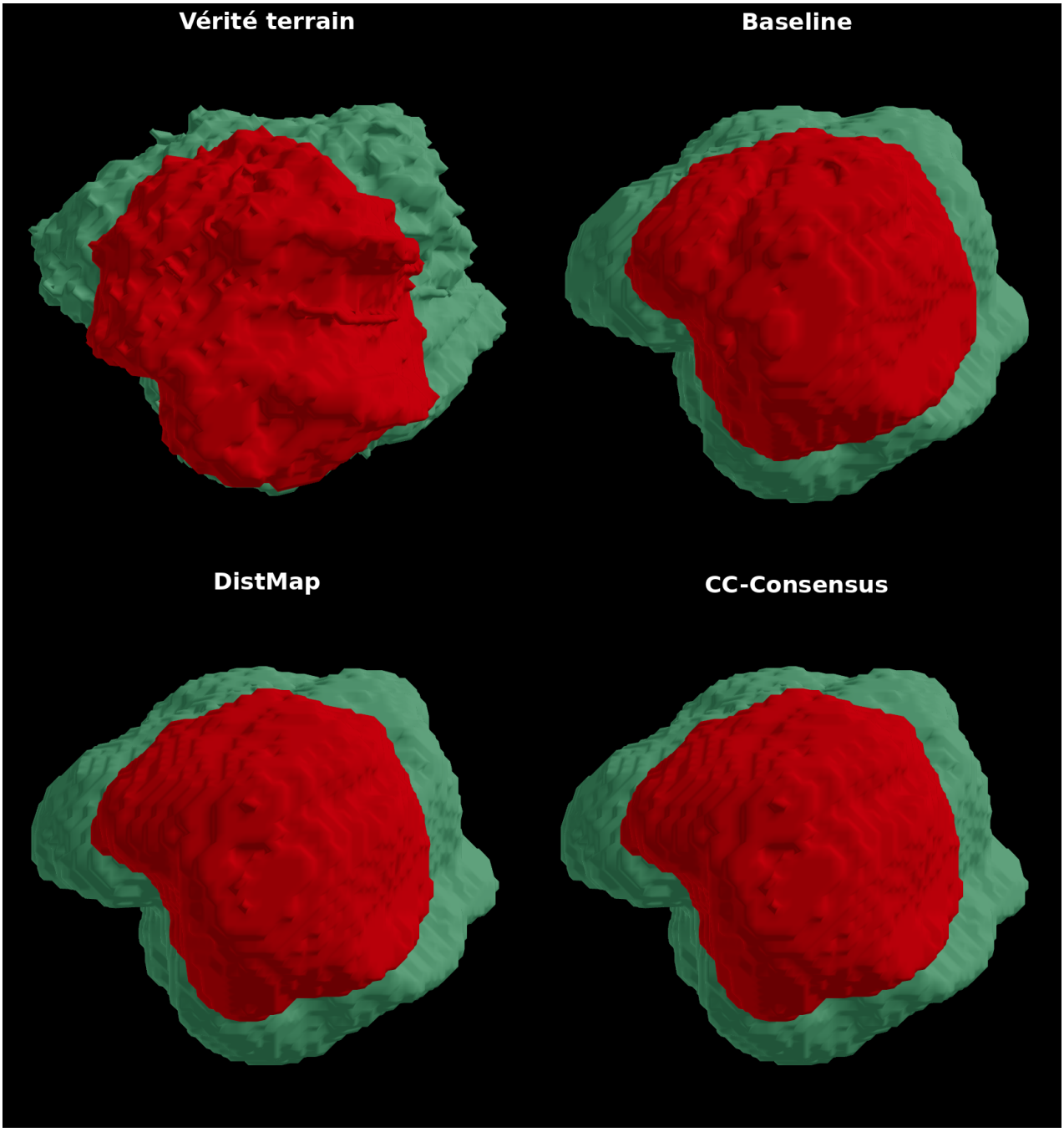

Figure 3: Figure 3 — Cas **C2** (patient `BraTS-GLI-01437-000`) : DistMap > Baseline. Baseline sous-segmente la tumeur (Dice 0,589) tandis que DistMap capture correctement l'extension tumorale (Dice 0,923) grâce à sa sensibilité de frontière. **CC-Consensus égale DistMap** (0,923) car aucune composante n'est hallucinée à supprimer — le filtre préserve la prédiction de meilleure qualité quand elle est confirmée par Baseline.

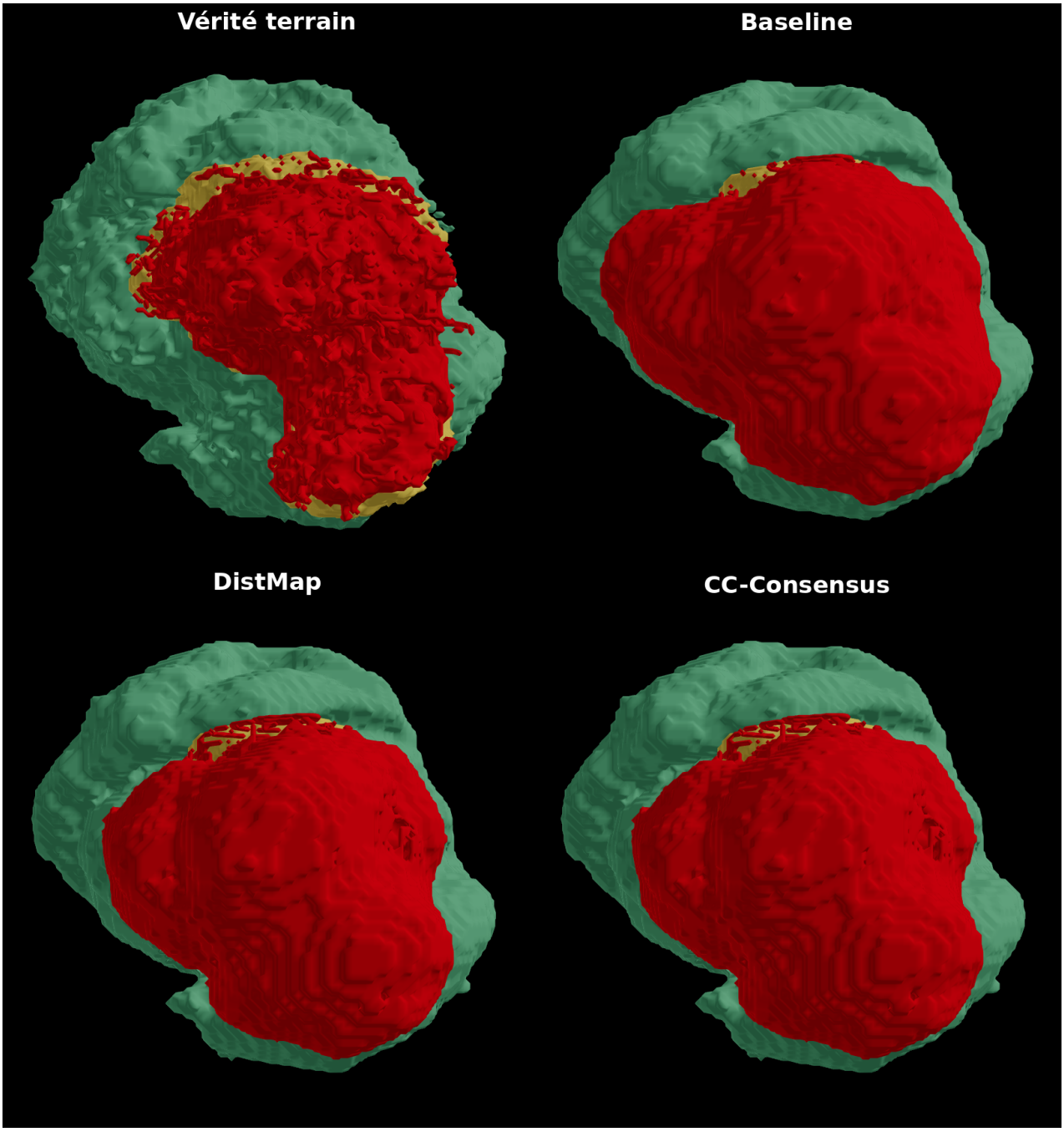

Figure 4: Figure 4 — Cas **C3** (patient `BraTS-GLI-01428-000`) : B < F < D, filtre tiré côté baseline. Baseline (0,618) et DistMap (0,656) encadrent le résultat CC-Consensus (0,645). Le filtre supprime certaines composantes DistMap légitimes que Baseline ne prédit pas, dégradant légèrement le score vers Baseline. C'est le mode de dégradation le plus courant (390 / 1196 patients, 32,6 %).

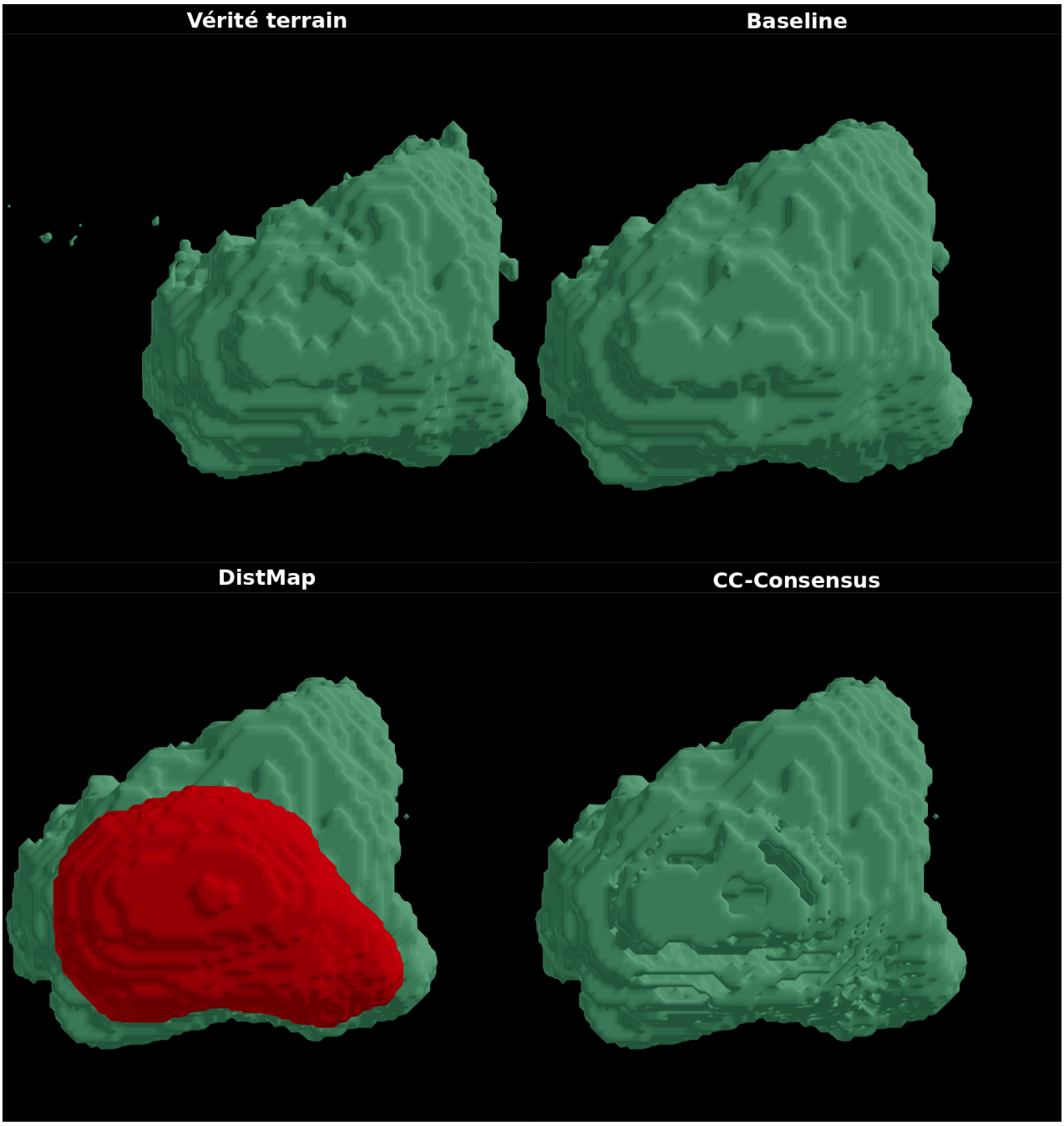

Figure 5: Figure 5 — Cas **C4** (patient `BraTS-GLI-00017-001`) : D < F < B, sauvetage partiel. Baseline est excellent (0,991) ; DistMap est à moitié hallucinée (0,657). CC-Consensus supprime les composantes DistMap fallacieuses et récupère une partie de la qualité Baseline (0,890), sans pouvoir l'atteindre puisqu'il part des voxels de DistMap.

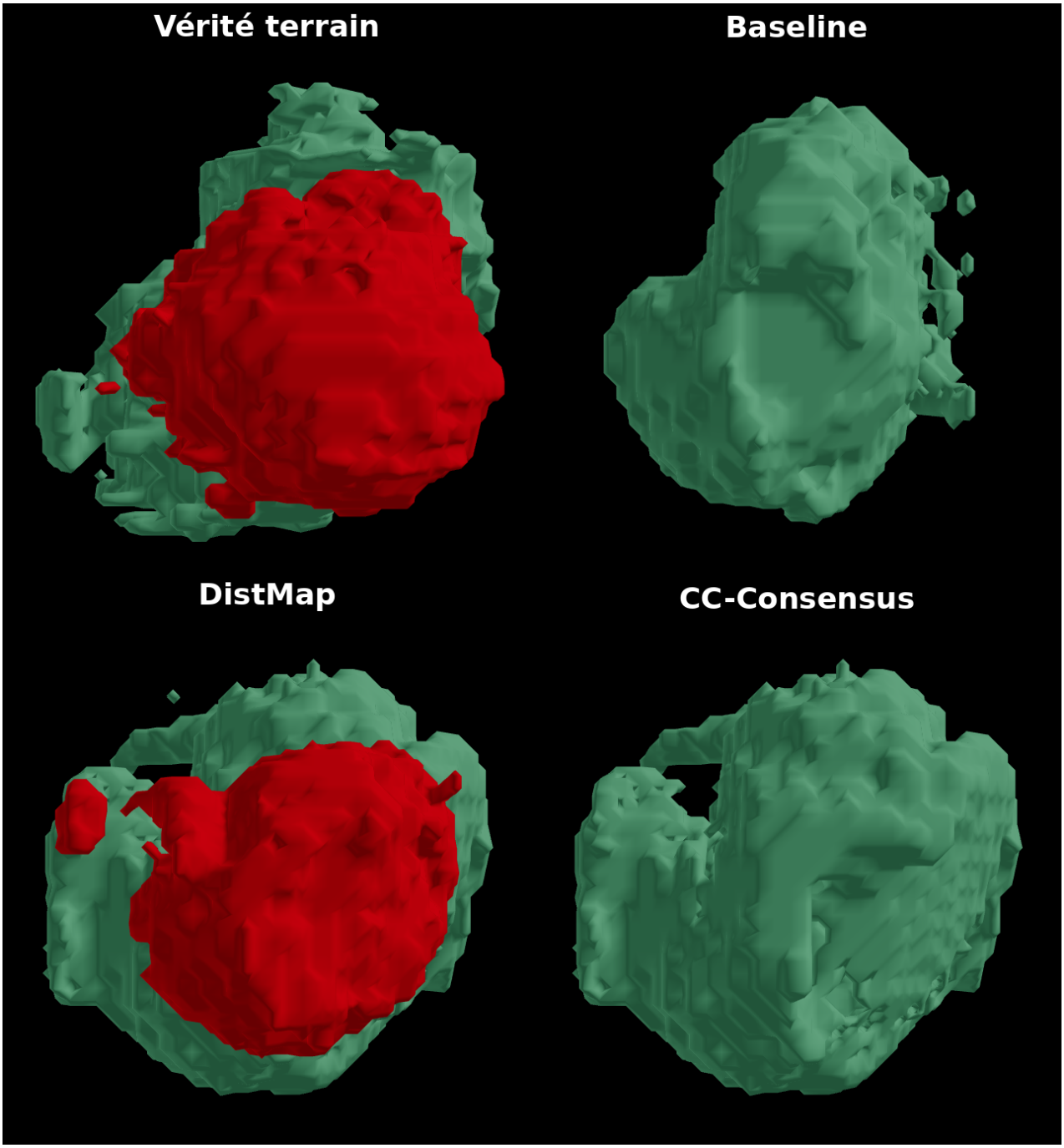

Figure 6: Figure 6 — Cas **C5** (patient `BraTS-GLI-01530-000`) : F < min(B, D), le filtre casse. Baseline = 0,241, DistMap = 0,541, CC-Consensus = 0,169. Le filtre **supprime une grosse composante DistMap légitime** car Baseline a raté la tumeur et ne peut pas la confirmer. 463 / 1196 patients (38,7 %) — c'est le principal mode de défaillance du filtre, quand Baseline et DistMap échouent différemment.

5.3 Le CC-consensus améliore HD95 NCR sans coût Dice

Agrégation des prédictions hors-fold sur les 5 folds (n = 1196) :

Stratégie	Dice avg	Δ vs CC-consensus par défaut
Baseline seule	0,9078	−0,00115
DistMap seule	0,9088	−0,00020
CC-Consensus (règle par défaut)	0,9090	0 (réf.)
Oracle au niveau patient	0,9131	**+0,00412**
Oracle par classe	0,9139	**+0,00494**

Classification par patient (en notant F la sortie du CC-consensus) :

Cas	Effectif	%
Baseline bat DistMap (B > D)	602	50,3 %
DistMap bat Baseline (D > B)	593	49,6 %
Sortie du filtre entre B et D	559	46,7 %
Filtre < les deux (dégradation)	**463**	**38,7 %**
Filtre > les deux (synergie)	**157**	**13,1 %**

Le filtre CC-consensus dégrade le score patient dans 38,7 % des cas contre 13,1 % de synergie. Par région, CC-Consensus l'emporte strictement sur 2,7 % des patients pour WT, **21,7 % pour TC** et 6,9 % pour ET. Le bénéfice du filtre en Dice est donc concentré sur TC ; pour WT et ET, le choix Baseline-seule ou DistMap-seule domine déjà.

Qualité de frontière (HD95). Complément du Dice, les distances de Hausdorff à 95 % sur les régions emboîtées BraTS (WT/TC/ET) **et** sur les classes individuelles (NCR, ED) où vivent les fragments (n varie par ligne selon le nombre de patients à HD95 fini sur la classe concernée) :

Région / classe	Composition	Baseline	DistMap	CC-Consensus	Δ CC-Cons. vs DistMap
WT	{1, 2, 3}	3,91 mm	3,86 mm	**3,76 mm**	**−0,10 mm, p = $2{,}7 \times 10^{-4}$**
TC	{1, 3}	3,08 mm	2,79 mm	2,88 mm	+0,09 mm, n.s.
ET	{3}	2,62 mm	2,59 mm	2,70 mm	+0,11 mm, n.s.
NCR	{1}	4,89 mm	4,86 mm	**4,48 mm**	**−0,38 mm, p = $5{,}7 \times 10^{-14}$**
ED	{2}	4,25 mm	4,33 mm	4,21 mm	−0,12 mm, n.s. (p = 0,82)

Test de Wilcoxon signé apparié, hypothèse unilatérale HD95(CC-Consensus) < HD95(DistMap). n = 1160 pour WT/TC/ET (restriction aux patients à HD95 fini sur les 3 régions emboîtées), 1153 pour NCR, 1193 pour ED.

Le signal dominant est sur NCR : CC-Consensus réduit HD95 NCR de 0,38 mm ($p = 5{,}7 \times 10^{-14}$) — confirmation quantitative directe que la suppression des fragments améliore la qualité de frontière sur la classe où ils prolifèrent majoritairement (NCR : ×1,5 plus de fragments DistMap vs Baseline, cf. §5.2). Le signal sur WT (−0,10 mm, $p = 2{,}7 \times 10^{-4}$) en est l'écho : NCR WT, donc les fragments NCR contribuent à l'erreur de frontière WT. Sur ED, la réduction de fragments (−61 %) ne se traduit pas en gain HD95 statistiquement significatif — l'œdème a une variabilité intrinsèque de frontière qui domine les outliers introduits par les fragments. Sur TC et ET (classe 3), les HD95 sont préservés.

Le CC-consensus délivre donc un gain quantitatif mesurable sur HD95 NCR et HD95 WT, là où le Dice reste insensible. Cliniquement, NCR est précisément la région où des fragments fallacieux peuvent induire en erreur un radiothérapeute sur l'emprise de la nécrose tumorale.

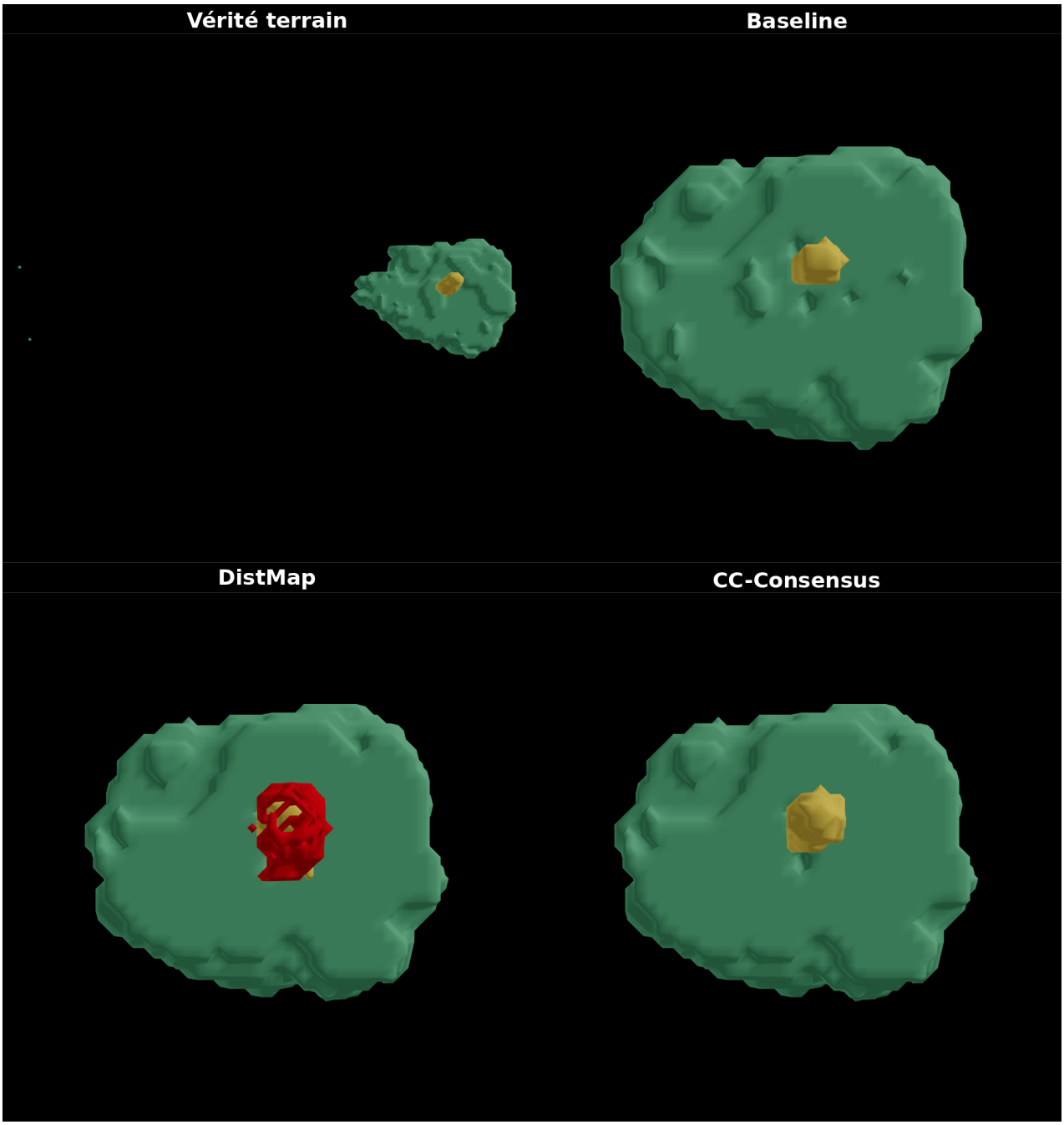

Figure 7: Figure 7 — Cas **C6** (patient `BraTS-GLI-00540-000`) : synergie nette. Baseline (0,785) et DistMap (0,795) sont tous deux compétents mais aucun n'est parfait. **CC-Consensus combine leurs forces** pour atteindre 0,869 — strictement supérieur aux deux parents. C'est le comportement recherché sur 157/1196 patients (13,1 %) où le filtre dépasse ses sources.

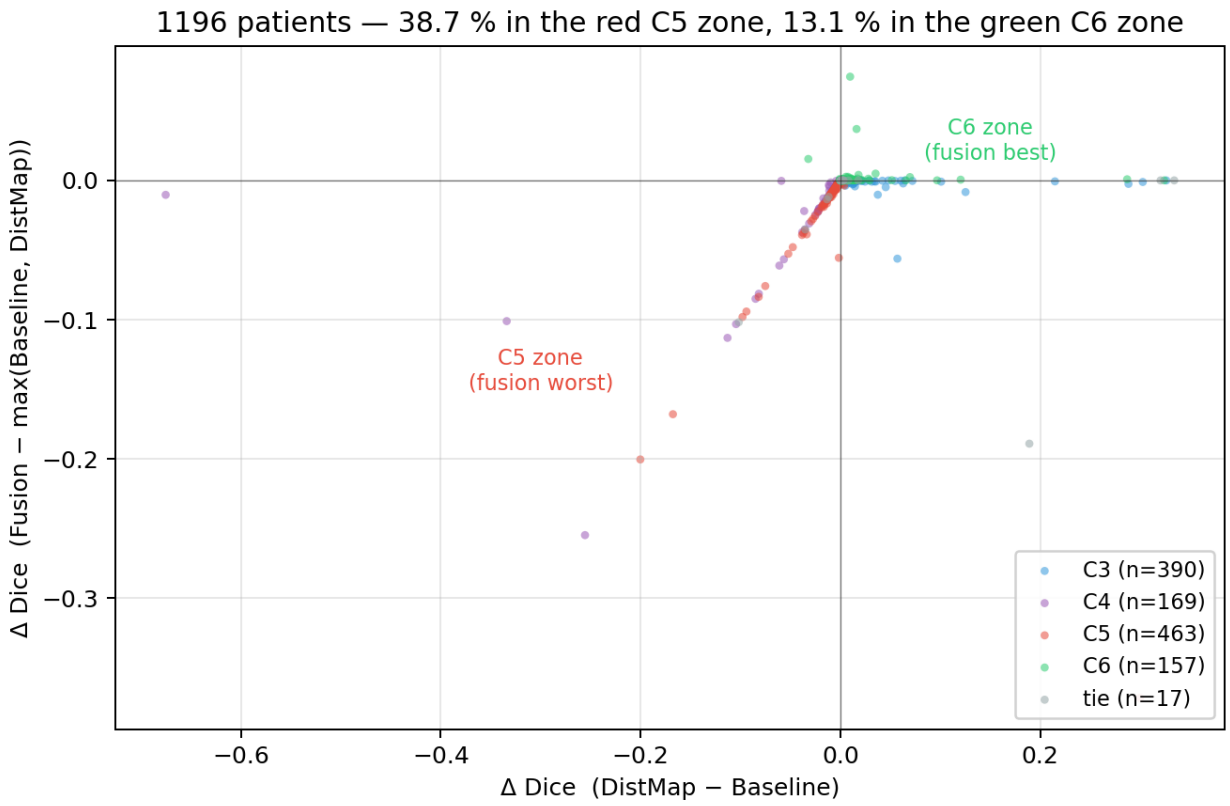

Figure 8: Figure 8 — 1196 patients de validation représentés dans le plan de désaccord entre modèles : x = Dice(DistMap) − Dice(Baseline) (une valeur positive signifie que DistMap l'emporte au niveau patient), y = Dice(CC-Cons.) − max(Dice(B), Dice(D)) (une valeur négative signifie que le filtre CC-consensus est pire que chaque modèle pris isolément). Le nuage *rouge* C5 sous y = 0 rassemble 38,7 % des patients pour lesquels le filtre dégrade le score ; les points *verts* C6 au-dessus de y = 0 ne représentent que 13,1 %. Cette asymétrie visuelle est l'observation empirique centrale du papier.

5.4 Le plafond hard-label est saturé

L'écart entre CC-consensus par défaut et oracle par classe (+0,005 Dice avg) borne supérieurement le gain de toute politique de sélection au niveau patient ou région à partir des trois prédictions {B, D, F}. On évalue trois familles de politiques en CV 5-fold (seuil taille-adaptatif sur τ ∈ {20, 50, 100, 200, 500, ∞} voxels ; meta-classifieurs RF/LR/GBM × patient/région sur 31 features ; règle à une feature par recherche exhaustive) ; **aucune ne bat robustement le CC-consensus par défaut**. La règle à une feature, attirante en fit toutes données (+0,00119), s'effondre en CV 5-fold (−0,00096) : la meilleure feature et le meilleur seuil changent entre folds (TC : 4 features distinctes sur 5 folds ; ET : 4 features distinctes). Un RandomForest par région atteint 50 %, 43 %, 51 % de précision argmax (vs 33 % au hasard), ce qui confirme la présence de signal — mais lorsque le classifieur se trompe, il choisit un modèle strictement pire, aboutissant à un bilan net négatif.

Le détail complet — tableau des 7 politiques évaluées, importances RF par région, classification du sweep adaptatif, partition par fold de la règle à une feature — est en **Annexe B**. Le plafond hard-label est essentiellement atteint ; combler l'écart à l'oracle nécessite un vote probabiliste au niveau voxel ou une diversité architecturale (§6.3).

5.5 Positionnement par rapport aux gagnants BraTS 2023 GLI

Le CC-consensus atteint Dice avg = 0,909 (WT 0,935, TC 0,919, ET 0,873) en CV 5-fold sur 1196 patients, avec un setup mono-modèle (pas d'ensemble multi-fold, pas de TTA, une seule architecture). C'est à moins d'un point de pourcentage

de la fourchette des gagnants publiés BraTS 2023 GLI sur test set privé (0,87–0,89 Dice avg ; Ferreira *et al.* 2024). Deux précautions à la comparaison directe : (i) jeu d'évaluation différent (CV 5-fold sur train + val vs test set privé, écart typique 1–2 pp en défaveur du test set) ; (ii) convention Dice = 1 sur région vide (nnU-Net / MONAI) qui inflate ET de ~0,003 par rapport à la convention lesion-wise du challenge (32/1196 patients sans ET en GT).

On ne revendique pas un nouvel état de l'art ; le filtre CC-consensus est **orthogonal à l'ensembling** — la réduction de fragments est un gain qui se cumule avec les astuces multi-fold / TTA classiques sans les dupliquer.

---

6. Discussion

6.1 Pourquoi DistMap génère des fragments

Mécanisme plausible — non démontré : la pression SDT sensibilise le réseau à de petits signaux *boundary-like* dans les tissus de transition (interfaces œdème–substance blanche, cavités post-chirurgicales, NCR hétérogène), produisant des voxels à forte réponse SDT qui survivent parfois à l'argmax sous forme de blobs isolés. Cette hypothèse est cohérente avec deux observations : l'augmentation du comptage de fragments est concentrée sur NCR et ED (régions aux frontières les plus longues et irrégulières), et beaucoup plus faible sur ET dont le rehaussement au gadolinium offre un contraste de frontière plus tranché. Trois contrôles directs (ablation λ × comptage de fragments, visualisation de la carte SDT aux emplacements des fragments, bins de distance vs MSE) sont décrits en **Annexe C** et déférés à des travaux futurs ; la contribution principale ici est la caractérisation et l'atténuation post-hoc de l'artefact, pas son explication mécaniste.

6.2 Pourquoi le filtre CC-consensus fonctionne

Baseline ne partage pas la pression SDT et ne produit donc pas la même classe de blobs fallacieux liés à la frontière. Exiger un recouvrement avec Baseline pour qu'une CC DistMap survive équivaut à un **test de consensus** sur un détecteur secondaire aux perturbations disjointes. C'est une application de l'idée classique « accord de classifieurs indépendants », adaptée ici aux composantes connexes plutôt qu'aux voxels.

La règle a deux propriétés souhaitables :

- **Asymétrique par construction.** On part de DistMap (meilleure qualité de frontière) et Baseline est utilisé uniquement comme veto. La meilleure frontière est préservée partout où le veto ne se déclenche pas.
- **Sans paramètre.** Pas de seuil, pas de poids appris — la connectivité CC est le seul hyperparamètre (26-connexe).

6.3 Pourquoi l'oracle ne peut être atteint

Deux modèles de la même famille (architecture, données, augmentations, famille de loss identiques, ne différant que par l'auxiliaire SDT) produisent trop peu de diversité pour qu'une classification à 3 issues « B vs D vs F » soit apprenable de façon fiable à partir de features de forme seules. Les deux modèles vivent dans le même voisinage de décision ; leurs désaccords sont dominés par du bruit spatial haute fréquence que la morphologie globale ne capture pas.

Combler l'écart de +0,005 Dice nécessite presque certainement l'une des voies suivantes :

- **Vote probabiliste au niveau voxel.** Exporter les sorties softmax (pas uniquement les labels argmax) et fusionner au niveau voxel brise le plafond du vote en dur. Une moyenne pondérée $\alpha \cdot \mathbf{p}_B + (1-\alpha) \cdot \mathbf{p}_D$ avec α appris par région est une étape suivante naturelle.
- **Diversité architecturale.** Ajouter un backbone non-MedNeXt (nnU-Net vanilla, Swin-UNETR) augmente drastiquement la marge oracle, comme le montrent régulièrement les gagnants BraTS 2023.
- **Ensemble multi-seed / multi-fold.** La recette classique gagne +0,5 à +2 points de Dice sur BraTS ; pleinement compatible avec — et orthogonal à — la règle CC-consensus proposée ici.

6.4 Limites

- **Backbone unique.** Toutes les expériences utilisent MedNeXt-B ; la généralisation à Swin-UNETR / nnU-Net vanilla / Restormer renforcerait la conclusion.

- **Pas de baseline de fusion probabiliste.** Seul le filtrage en dur est rapporté car les sorties softmax n'ont pas été persistées à l'inférence. L'analyse du plafond adresse explicitement ce gap pour le cas hard-label.
- **Setup mono-modèle-par-patient.** Le filtre CC-consensus proposé atteint un Dice avg de 0,909 sur la CV 5-fold à 1196 patients sans ensembling multi-fold, sans TTA ni vote multi-architectures. L'ajout de ces astuces classiques placerait probablement le résultat dans ou au-dessus de la fourchette des gagnants BraTS 2023 GLI, mais il s'agirait d'une contribution de calcul parallèle orthogonale à la question de caractérisation des fragments que ce papier adresse.
- **La convention Dice inflate légèrement ET.** Les patients à GT vide sur ET (2,7 % de BraTS 2023 GLI, cas non-rehaussés, 32/1196 vérifié) sont scorés Dice = 1,0 sous la convention nnU-Net / MONAI, ce qui inflate légèrement la moyenne ET (−0,003 seulement sous la convention lesion-wise). Les comparaisons internes Baseline / DistMap / CC-Consensus ne sont pas affectées (les trois utilisent la même convention), mais la moyenne ET absolue n'est pas directement comparable aux leaderboards challenge qui utilisent une convention lesion-wise (voir §5.5).
- **BraTS 2023 GLI uniquement.** L'extension à BraTS-MET (métastases) et BraTS-PED (pédiatrique) est laissée aux travaux futurs ; on s'attend à ce que le biais de fragments soit plus sévère sur les métastases (pattern multi-lésions).
- **Pas de petites tumeurs dans le dataset.** Le volume WT minimum sur BraTS 2023 GLI est de 2808 voxels, la médiane à ~89 500 voxels. La définition topologique de fragment adoptée en §4.2 (CC − 1 par classe, sans seuil de taille) est **intrinsèquement robuste à la taille** et ne nécessite aucune recalibration pour des tumeurs plus petites. Cependant, **le pipeline évalué ici n'a pas été testé sur le régime cliniquement critique des petites tumeurs** (quelques centaines de voxels), où la détection précoce a un impact pronostique majeur. Les features morphologiques absolues (`vol_*`, `nb_cc_*`) seraient hors-distribution sur ce régime et devraient être réexaminées avant usage clinique ; les features topologiques et relatives (`ratio_ET_WT`, `frac_small_cc_*`, sphéricité, élongation) sont robustes par construction.

---

7. Conclusion

Sur MedNeXt-B / nnU-Net v2, la loss SDT auxiliaire ne produit pas de gain Dice significatif à convergence (Δ Dice avg = +0,09 pp, Wilcoxon $p > 0{,}25$ par région, CV 5-fold 1196 patients) mais elle change la topologie des prédictions en introduisant un biais de fragments que la métrique Dice échoue à rapporter. Un filtre de consensus de composantes connexes sans paramètre, qui oppose un veto aux CC DistMap sans recouvrement Baseline, élimine 66 % des fragments NCR sur 1196 patients ($p < 10^{-189}$) sans coût en Dice, et **améliore significativement HD95 sur NCR** (4,86 → 4,48 mm, $p = 5{,}7 \times 10^{-14}$) ainsi que sur WT (3,86 → 3,76 mm, $p = 2{,}7 \times 10^{-4}$) — un gain de qualité de frontière caché par le Dice, cliniquement pertinent sur la nécrose tumorale.

Sur 1196 patients en CV 5-fold, on établit que cette règle est déjà proche du plafond de saturation de toute politique de sélection post-hoc en hard-label : l'oracle par classe est à +0,005 Dice avg au-dessus du défaut, et aucun meta-selector à 31 features (4 familles de classifieurs) ne bat robustement ce défaut en CV. Combler cet écart motive des **loss d'entraînement sensibles aux fragments** (Paper 2) plutôt que davantage d'ingénierie post-hoc.

---

8. Perspectives

Loss d'entraînement sensible aux fragments (Paper 2). L'hypothèse §6.1 suggère que les fragments sont un effet de gradient. Un terme de pénalité au moment de l'entraînement comptant les composantes connexes prédites sur l'argmax de chaque mini-batch — et pénalisant les petits blobs isolés — devrait pousser le réseau à ne pas les instancier, rendant le filtre post-hoc CC-consensus inutile. C'est la direction de Paper 2.

Extensions de dataset. BraTS-MET (métastases, pattern multi-lésions) est le prochain test le plus informatif : les fragments DistMap devraient y être plus sévères, et le filtre CC-consensus en bénéficier davantage. BraTS-PED (pédiatrique) testerait la généralisation à travers des shifts démographiques.

Remerciements

L'auteur remercie **Stanislas Larnier** pour ses conseils méthodologiques, ses retours sur la formulation des questions de recherche, et ses relectures attentives des versions successives de ce papier.

Annexe A — Calibration de λ (loss auxiliaire SDT)

À l'epoch 0 avec un réseau initialisé aléatoirement (seed 42), on mesure $|\mathcal{L}_{\text{Dice+CE}}| = 0{,}57$ et $\mathcal{L}_{\text{MSE}}^{\text{SDT}} = 0{,}12$, ce qui donne un λ « équilibré par gradient » de 4,70.

Une ablation statique sur λ ∈ {0 ; 0,1 ; 0,5 ; 1 ; 2 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10} (100 epochs, fold 0, seed 42) donne des Dice avg tous compris dans une fenêtre de 0,5 pp :

λ	Dice avg	Δ vs Baseline
0 (Baseline)	0,9064	0
0,1	0,9077	+0,0013
0,5	0,9070	+0,0006
1,0	0,9067	+0,0003
2,0	0,9060	−0,0004
5,0	0,9105	+0,0041
9,0	0,9104	+0,0040

Sur ce fold unique et sans test de significativité par patient, aucun λ ne se distingue clairement du baseline. Ce résultat est en cohérence avec la non-significativité du gain DistMap observée en CV 5-fold sur 1196 patients (§5.1). L'entraînement par défaut rapporté dans le corps utilise λ = 1 (proche des heuristiques publiées et de la calibration équilibrée ÷ 5).

Un schéma de pondération dynamique — DWA (Dynamic Weight Average, Liu CVPR 2019) — qui suit les taux d'apprentissage relatifs des têtes Dice+CE et SDT au cours de l'entraînement, est une piste à explorer : s'il existe un régime où SDT contribue vraiment sans saturer, un balayage statique ne peut pas le trouver.

Annexe B — Étude détaillée du plafond hard-label

L'écart entre CC-consensus par défaut et oracle par classe (+0,005 Dice avg) est le gain maximal de toute politique de sélection par région. On évalue des politiques progressivement plus riches :

Politique	Dice avg	Δ vs CC-consensus
Meilleur seuil taille-adaptatif (τ = 200 vx)	0,90909	+0,00012
27 règles fixes par région — meilleur = D/F/F	0,90935	+0,00038
Meta-LR (31 features, niveau patient)	0,90940	+0,00043
Meta-RF (31 features, par région)	0,90807	**−0,00090**
Meta-LR (31 features, par région)	0,90844	−0,00053
Meta-GBM (31 features, par région)	0,90833	−0,00064
Règle à une feature (fit toutes données)	0,91016	+0,00119
Règle à une feature (CV 5-fold)	**0,90801**	**−0,00096**

La règle à une feature, attirante en fit toutes données (+0,00119), s'effondre en CV 5-fold (−0,00096) : la meilleure feature et le meilleur seuil changent entre folds (TC : 4 features distinctes sur 5 folds ; ET : 4 features distinctes). Quatre features « meilleures » différentes sur cinq folds pour TC seule indiquent clairement que le signal n'est pas assez robuste pour lui faire confiance.

Un RandomForest entraîné par région atteint 50 %, 43 % et 51 % de précision en argmax (WT, TC, ET) contre 33 % au hasard, ce qui confirme la présence de signal dans les features — mais lorsque le classifieur se trompe, il choisit un modèle strictement pire, aboutissant à un bilan net négatif.

L'importance des features (RF sur toutes les données, top-3 par région) étaye le récit : pour ET, `frac_removed_distmap_ET` (0,13) et `max_orphan_cc_ET` (0,09) — toutes deux des features d'accord inter-modèles — dominent. Le signal est réel, simplement pas assez fort pour survivre à la CV.

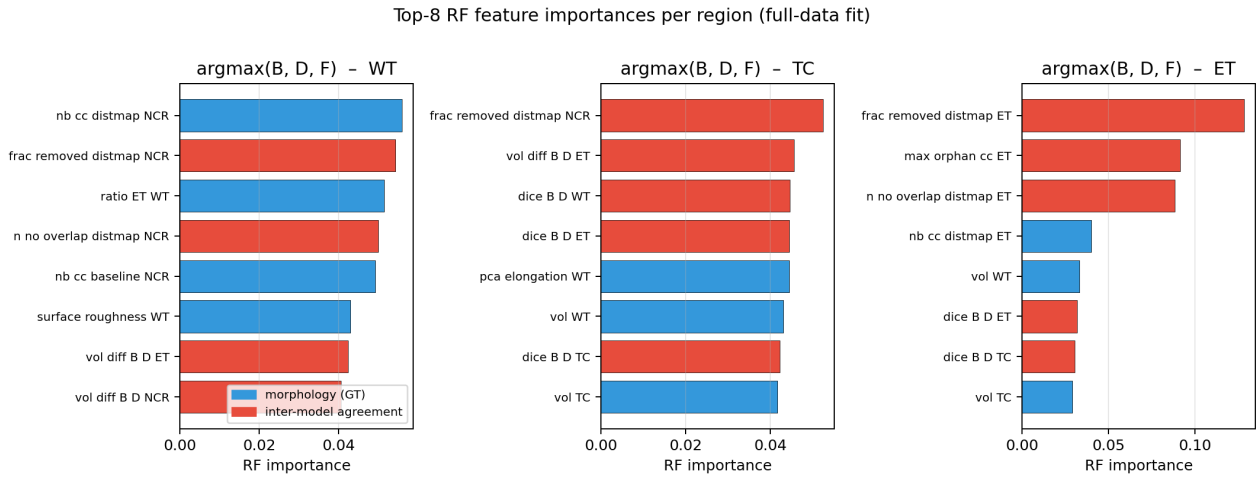

Figure 9: Figure A1 — Top-8 des importances features d'un RandomForest entraîné à prédire argmax(Baseline, DistMap, CC-Consensus) pour chaque région (WT / TC / ET). Barres bleues : features morphologiques issues de la GT (20). Barres rouges : features d'accord inter-modèles (11). Pour ET spécifiquement, les 3 premières importances — `frac_removed_distmap_ET`, `max_orphan_cc_ET`, `n_no_overlap_distmap_ET` — sont toutes des features d'accord, confirmant que la décision « faire confiance au filtre CC-consensus sur ET ou non » est pilotée par la quantité de sur-prédiction de DistMap par rapport à Baseline.

Annexe C — Hypothèse mécaniste : contrôles déférés

L'hypothèse de §6.1 (la pression SDT engendre des voxels à forte réponse aux interfaces ambigües, qui survivent parfois à l'argmax) reste à ce stade une **hypothèse de travail non démontrée**. Les trois contrôles directs suivants sont tous réalisables sur les checkpoints existants et sont déférés à des travaux futurs :

1. **Ablation de λ croisée avec comptage de fragments.** Vérifier que le nombre moyen de fragments par patient croît monotoniquement avec λ. Une croissance monotone confirmerait le lien causal entre pression SDT et artefact ; une absence de monotonie suggérerait que le bruit d'optimisation domine.
2. **Visualisation de la carte SDT aux emplacements des fragments.** Pour un échantillon de patients, superposer la sortie tanh de la tête auxiliaire et la carte de fragments ; les fragments devraient coïncider avec des voxels à forte réponse SDT proches d'une interface tissulaire.
3. **Bins de distance vs MSE.** Remplacer la tête `Conv3D(32 → 3) + tanh + MSE` par une tête de classification en bins de distance (ex. 16 bins équi-probables dans [−1, 1]). Si l'artefact disparaît ou diminue substantiellement, il est spécifique à la formulation MSE-SDT et non à la supervision de distance en général.

Exécuter ces trois contrôles ferait passer §6.1 d'« hypothèse de travail » à « mécanisme démontré ».

Annexe D — Temps d'exécution et reproductibilité

Tout le code, les 20 + 11 features pré-extraites, les scores par modèle et par patient, les CSV d'oracles / classification de cas, les résultats du balayage de seuil et les sorties des meta-selectors sont disponibles dans le dépôt compagnon. L'extraction des 31 features par patient sur les 1196 prédictions tourne en **~10 min** sur 14 threads P-cores (`taskset -c 0-13`) d'un i7-14700K ; le balayage complet de meta-classifieurs (4 familles × 5 folds × 31 dim) tourne en ~2 min sur le même hôte. **Temps d'entraînement par fold : ~13 h 30 pour 300 epochs** sur une unique RTX PRO 6000 Blackwell (96 Go), variantes Baseline et DistMap à durée équivalente (la tête de régression SDT auxiliaire ajoute < 1 % de surcoût GPU sur 300 ep).

Annexe E — Les six patients de démonstration

Six patients sont mis en avant pour couvrir les six cas d'ordonnancement de modèle, utilisés à la fois pour les figures et comme ancres épinglées dans le viewer 3D compagnon. Dans le tableau, F désigne la sortie du filtre CC-consensus. Les identifiants patient sont affichés sans le préfixe `BraTS-GLI-` pour compacité (le dataset le préfixe systématiquement).

Tag	Patient	Fold	B	D	F	Enseignement
C1 (baseline $>$ distmap)	00048-001	1	0,983	0,308	0,973	DistMap hallucine TC/ET sur un cas uniquement œdème
C2 (distmap $>$ baseline)	01437-000	2	0,589	0,923	0,923	DistMap sauve un Baseline sous-segmentant
C3 (B $<$ F $<$ D)	01428-000	1	0,618	0,656	0,645	Sortie du filtre entre les deux, tirée côté baseline
C4 (D $<$ F $<$ B)	00017-001	0	0,991	0,657	0,890	Sortie du filtre sauve DistMap par consensus
C5 (filtre pire)	01530-000	1	0,241	0,541	0,169	Le filtre supprime une grosse CC DistMap légitime
C6 (filtre meilleur)	00540-000	1	0,785	0,795	0,869	Synergie nette

Références

- Isensee F., Jaeger P. F., Kohl S. A. A., Petersen J., Maier-Hein K. H. (2021). *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*. **Nature Methods** 18, 203–211. DOI: 10.1038/s41592-020-01008-z.
- Roy S., Koehler G., Ulrich C., Baumgartner M., Petersen J., Isensee F., Jaeger P. F., Maier-Hein K. H. (2023). *MedNeXt: transformer-driven scaling of ConvNets for medical image segmentation*. **MICCAI 2023**, LNCS 14222, 405–415. DOI: 10.1007/978-3-031-43901-8_39.
- Ma J. (2020). *Distance transform maps improve semantic segmentation of medical images*. **Medical Imaging with Deep Learning (MIDL) 2020**, short paper track.
- Xue Y., Tang H., Qiao Z., Gong G., Yin Y., Qian Z., Huang C., Fan W., Huang X. (2020). *Shape-aware organ segmentation by predicting signed distance maps*. **AAAI 2020**, 34(07), 12565–12572. DOI: 10.1609/aaai.v34i07.6946.
- Karimi D., Salcudean S. E. (2020). *Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks*. **IEEE Transactions on Medical Imaging** 39(2), 499–513. DOI: 10.1109/TMI.2019.2930068. arXiv:1904.10030.
- Huang Q., Yang J., Zhang B., Wang Z., Bai J., Li Y. (2021). *A deep multi-task learning framework for brain tumor segmentation*. **Frontiers in Oncology** 11, 690244. DOI: 10.3389/fonc.2021.690244.
- Pham T.-D., Abdollahzadeh A., Tohka J. (2024). *SiNGR: Brain tumor segmentation via signed normalized geodesic transform regression*. **MICCAI 2024**. arXiv:2405.16813.
- Ferreira A., Solak Ü. M., Li J., Dammann P., Kleesiek J., Alves V., Egger J. (2024). *How we won BraTS 2023 adult glioma challenge? Just faking it! Enhanced synthetic data augmentation and model ensemble for brain tumour segmentation*. **arXiv:2402.17317**.
- Liu S., Johns E., Davison A. J. (2019). *End-to-end multi-task learning with attention* (DWA — Dynamic Weight Average). **CVPR 2019**, 1871–1880. DOI: 10.1109/CVPR.2019.00197.
- Baid U., Ghodasara S., Mohan S., Bilello M., Calabrese E., Colak E., *et al.* (2021). *The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification*. **arXiv:2107.02314**.
- Menze B. H., Jakab A., Bauer S., *et al.* (2015). *The multimodal brain tumor image segmentation benchmark (BRATS)*. **IEEE TMI** 34(10), 1993–2024. DOI: 10.1109/TMI.2014.2377694.