La représentation préconditionne l'efficience en échantillons à information constante

Un prédicteur structurel de l'écart d'apprentissage entre sérialisations informationnellement équivalentes, sans modèle et sondé causalement

Jérôme Vétillard
VP R&D + Engineering / Chief Product Officer, TweenMe by Qualees

---

Résumé

Deux sérialisations d'une même cohorte longitudinale peuvent porter une information démontrablement identique et pourtant s'apprendre avec des facilités différentes. Nous étudions ce phénomène sous une contrainte dure : les deux représentations sont mutuellement reconstructibles par une bijection déterministe, de sorte qu'aucune ne contient plus d'information que l'autre. À l'aide de générateurs synthétiques *ex nihilo* à vérité terrain contrôlée, nous définissons une *ontologie opérationnelle* comme le choix de l'axe de tête dans le tenseur de données, c'est-à-dire l'axe traité comme l'unité échangeable sur laquelle on constitue les lots. Nous rapportons un faisceau de résultats mutuellement corroborants. (i) Un croisement asymétrique : à information égale et budget égal, la représentation dont l'axe de tête coïncide avec l'unité de prédiction apprend mieux, et cet avantage survit à l'encodeur maintenu fixe, ce qui isole la représentation de l'encodeur et décompose le croisement brut en une part de représentation et une part d'encodeur. (ii) Un contrôle nul confirme que le pipeline ne fabrique aucun écart en l'absence de mécanisme asymétrique. (iii) Le croisement se reproduit sans entraînement, par une sonde des k plus proches voisins : c'est donc une propriété de la métrique induite. (iv) Une quantité structurelle sans modèle, la pureté de voisinage, prédit le signe et l'amplitude de l'écart de performance sur une famille de tâches (Pearson $r = 0{,}996$; bootstrap par grappes et permutation rapportés). (v) Une intervention préservant l'information sur la métrique déplace conjointement la performance et la pureté, à étiquette fixée et information fixée ($r = 0{,}942$), ce qui établit la causalité *au niveau de la métrique*. Le faisceau se reproduit dans quatre mondes synthétiques différant par la modalité d'observation, dont un domaine de graphes appris par passage de messages ; nous disons explicitement que ces mondes partagent une structure latente à deux échelles, de sorte que leur indépendance est de modalité, non de squelette génératif. Nous ajoutons une mesure directe données-vers-performance et une validation sur deux panels longitudinaux réels, où le prédicteur tient ($r = 0{,}98$ et $r = 0{,}996$). Nous montrons honnêtement que la pureté de voisinage est un médiateur suffisant mais non unique : une séparabilité fondée sur les distances prédit tout aussi bien. Nous traçons aussi la frontière propre de la théorie : la pureté prédit les apprenants dominés par le voisinage induit (une machine à noyau, un ensemble d'arbres) et, comme préenregistré pour échouer, ne prédit pas un apprenant linéaire global sur une étiquette de structure XOR ; enfin le prédicteur se dégrade progressivement, non catastrophiquement, lorsque la bijection n'est qu'approchée. Nous formulons le résultat comme une interaction entre représentation, apprenabilité et efficience en échantillons, et nous en déclarons les limites.

1 Introduction

1.1 La représentation comme biais inductif

Apprendre, c'est généraliser à partir de ressemblances. Un modèle qui n'a jamais vu un cas exact doit décider à quel cas connu l'assimiler ; toute généralisation présuppose une notion de proximité. Sans biais, toutes les hypothèses compatibles avec les données restent à égalité (Mitchell, 1980), et aucun apprenant n'est

meilleur qu'un autre en moyenne sur tous les problèmes (Wolpert, 1996; Wolpert and Macready, 1997). Cette tradition parle du biais du *modèle* : la localité d'une convolution, la séquentialité d'une récurrence, les entités et relations d'un réseau de graphes (Battaglia et al., 2018). Un second biais, distinct et plus rarement nommé, est induit par la *représentation*. De Finetti l'a formalisé (de Finetti, 1937) : déclarer des observations échangeables, c'est poser un modèle latent, et choisir l'unité échangeable fixe l'objet dont on suppose la répétition.

Les deux biais se composent plutôt qu'ils ne se substituent. Une représentation préconditionne le biais architectural, au sens de l'optimisation numérique où un préconditionneur ne change pas la solution d'un système mais la facilité de l'atteindre. La chaîne n'est pas Architecture $\rightarrow$ Fonction mais Architecture $\rightarrow$ Représentation $\rightarrow$ Fonction : le même encodeur, entraîné sur deux sérialisations d'un même jeu de données, peut apprendre deux fonctions différentes, parce que la représentation a déjà fixé les proximités à l'intérieur desquelles opère le biais architectural.

1.2 La représentation tensorielle

Concrètement, une ontologie opérationnelle est un choix de disposition tensorielle : quel axe est l'axe échangeable, parcouru en lots i.i.d., et comment s'imbriquent les axes restants. Une cohorte peut être rangée en $[\text{patient}, \text{temps}, \text{variables}]$ ou en $[\text{temps}, \text{patient}, \text{variables}]$: deux tenseurs faits des mêmes octets, réversibles l'un dans l'autre, mais dont l'axe de tête diffère, et avec lui ce que la machine tient pour proche. Un lecteur venu de la représentation des connaissances objectera qu'il s'agit d'une disposition tensorielle, non d'une ontologie au sens des entités, relations et types. L'objection est juste quant au vocabulaire et hors sujet quant au mécanisme : la décision que l'ingénierie des données appelle l'ontologie, le choix de l'unité échangeable, est computationnellement un choix de disposition, et c'est ce choix, non un vocabulaire, dont nous montrons qu'il agit sur l'apprentissage. Dans tout ce qui suit, « représentation » est opérationnalisée comme le choix de l'axe de regroupement, et aucune prétention n'est portée au-delà.

1.3 Ce que la littérature ne teste pas

L'apprentissage de représentations étudie comment *apprendre* des attributs (Bengio et al., 2013), non comment une disposition fixée, à information constante, change l'apprenabilité. La littérature sur l'agrégation et l'inférence écologique, de Simpson (1951) à Robinson (1950), avertit que le niveau d'agrégation change ce qu'on lit, mais pour l'inférence, non pour l'efficience en échantillons. Du côté du modèle, le noyau tangent neuronal (Jacot et al., 2018), le biais spectral (Rahaman et al., 2019) et la double descente (Belkin et al., 2019; Nakkiran et al., 2019) caractérisent ce qu'une architecture donnée trouve facile, à représentation des données fixée. Dans le périmètre de notre revue de ces littératures, nous n'avons pas trouvé de travaux antérieurs isolant la représentation tout en maintenant l'information démontrablement constante par une bijection explicite ; nous l'énonçons comme une recherche bornée, non comme une revendication exhaustive de nouveauté.

1.4 Contribution et périmètre

Nous apportons un prédicteur structurel sans modèle de l'écart d'apprentissage entre sérialisations informationnellement équivalentes, une ablation causale qui le fait passer du corrélationnel au causal au niveau de la métrique, une démonstration d'indépendance sur quatre mondes synthétiques et deux panels réels, et un bornage honnête du médiateur. La démonstration est synthétique et sur panels réels, à vérité terrain contrôlée ou observée ; nous tenons la frontière de bout en bout.

2 Travaux connexes

Apprentissage de représentations et de métriques. L'apprentissage de représentations cherche des attributs qui facilitent les tâches en aval (Bengio et al., 2013) ; l'apprentissage de métrique optimise une distance pour une tâche. Tous deux *apprennent* la représentation ou la métrique à partir des données. Nous fixons au contraire deux représentations a priori, les contraignons à une information identique, et étudions l'écart d'apprenabilité qui en résulte, ce qui isole l'effet du choix de disposition plutôt que celui d'une transformation apprise.

Biais inductif, invariance et géométrie. Les biais inductifs relationnels (Battaglia et al., 2018) et l'apprentissage profond géométrique (Bronstein et al., 2021) formalisent la manière dont la structure architecturale encode des hypothèses. C'est le biais du modèle. Notre objet est le biais complémentaire et moins étudié de la représentation des données, et notre contrôle à encodeur fixé (section 3.4) sépare les deux.

La géométrie de l'apprentissage. Le noyau tangent neuronal (Jacot et al., 2018), le biais spectral (Rahaman et al., 2019), le goulot d'étranglement informationnel (Tishby et al., 1999) et la double descente (Belkin et al., 2019; Nakkiran et al., 2019) expliquent ce qu'une architecture fixée trouve facile. Ils agissent du côté du modèle ; nous montrons que la représentation des données est un levier sur la même métrique effective.

Échangeabilité et agrégation. L'échangeabilité de De Finetti (de Finetti, 1937) lie le choix de l'unité à un modèle latent. La littérature sur Simpson et l'inférence écologique (Simpson, 1951; Robinson, 1950) montre que le niveau d'agrégation change les conclusions inférentielles ; nous transportons l'avertissement de l'inférence vers l'efficience en échantillons.

Similarité et projectibilité. La lecture selon laquelle une représentation est une hypothèse sur les similarités pertinentes a des antécédents chez Edelman (1998), et plus en amont dans la projectibilité de Goodman (1955) et les espèces naturelles de Quine (1969). Notre ajout est que cette similarité, une fois mesurée, prédit et (au niveau de la métrique) cause l'efficience en échantillons.

3 Méthodes

3.1 Générateurs *ex nihilo* à vérité terrain contrôlée

Le générateur principal (monde A) produit une cohorte longitudinale synthétique dotée d'un mécanisme de survie à risques proportionnels en temps discret. Le risque instantané du patient i dépend séparément de deux composantes d'un facteur latent,

$$h_i(t) = h_0(t)\, \exp\!\big(\beta_{\text{bet}}\, L_i + \beta_{\text{wit}}\, w_i(t) + \gamma^\top Z_i\big), \tag{1}$$

où L_i est le niveau structurel du patient (composante inter-individuelle), $w_i(t)$ l'écart intra-individuel, Z_i des covariables statiques, et h_0 un risque de base de Weibull. Des signes opposés de β_{bet} et β_{wit} donnent un régime de type Simpson. Les apprenants ne voient que des biomarqueurs mesurés bruités aux visites ; la vérité terrain n'est conservée que pour la validation. Trois mondes supplémentaires partagent le squelette de test mais diffèrent par le processus génératif : le monde B, un panel continu à innovations de Student-t, niveaux bimodaux, processus intra oscillatoire à retour à la moyenne, et attrition ; le monde C, une matrice bipartite à facteurs latents (utilisateurs par objets) avec une nuisance par objet et une parcimonie native, sans temps ; le monde D, un graphe issu d'un modèle à blocs stochastiques dont le signal des nœuds se scinde en une composante lisse sur le graphe (populationnelle) et une composante idiosyncrasique (locale). Les équations complètes figurent en annexe A. Les quatre mondes diffèrent sur au moins deux axes chacun (mécanisme, bruit, dépendance, distribution des niveaux, observation), mais ils partagent une décomposition latente à deux échelles, que nous discutons comme une limite.

3.2 Équivalence informationnelle et trois sens du mot information

La contrainte est dure : les deux sérialisations d'un jeu de données sont mutuellement reconstructibles par une bijection déterministe, vérifiée par un test d'égalité exacte aller-retour sur l'ensemble des configurations. Cela impose de distinguer trois sens du mot *information*. La bijection garantit une information *de Shannon* identique : les mêmes bits, chacun fonction sans perte de l'autre. Elle ne garantit pas une *accessibilité algorithmique* identique : les mêmes bits peuvent être arbitrairement plus difficiles à exploiter sous une disposition donnée. Et elle ne garantit pas une *efficience statistique* identique : la cible peut être une statistique exhaustive minimale dans une représentation et une fonction profondément intriquée dans l'autre. Notre résultat vit dans l'écart entre le premier sens et les deux autres. Pour le monde D, l'équivalence est plus lâche qu'une transposition stricte (deux fonctions d'un substrat partagé), coût honnête du régime géométrique.

3.3 Tâches et cible paramétrée

Une tâche *individuelle* prédit une issue individuelle à partir de l'historique propre de l'individu (monde A : concordance de survie en *landmark*). Une tâche *transversale* prédit une quantité relative à la population, rendue transversale par un effet de vague, un décalage aléatoire par tranche jouant le rôle d'un effet de lot que seul un modèle voyant la tranche peut retirer. Un curseur w interpole l'unité d'étiquette de l'individuel ($w = 0$) au transversal ($w = 1$) en mélangeant des composantes standardisées de contexte patient et de contexte tranche. Les mondes C et D transposent la dualité en objet-contre-utilisateur et graphe-contre-attribut.

3.4 Encodeurs à budget égalisé et contrôle à encodeur fixé

Chaque ontologie reçoit l'encodeur adapté à sa structure, à budget de paramètres égalisé (GRU causal et Transformer, MLP agrégé personne-période, encodeur d'ensembles). Comme apparier chaque ontologie à un encodeur différent confond représentation et encodeur, nous introduisons un contrôle à encodeur fixé : un même encodeur d'ensembles appliqué de deux façons ne différant que par l'axe de regroupement sur lequel il agrège (par tranche contre par patient). Dans le monde D, l'encodeur fixé est une couche de passage de messages à un saut dont le contexte agrégé est le voisinage du graphe ou un ensemble global aveugle à la géométrie. Les hyperparamètres figurent en annexe B.

3.5 Alignement structurel, sans modèle

Une représentation induit une métrique, donc un graphe de voisinage. La *pureté de voisinage* est, pour une représentation et une étiquette, la fraction moyenne des k plus proches voisins partageant l'étiquette du point ; c'est une propriété du couple (métrique, étiquette), calculée sans aucun modèle entraîné. Comme concurrent, nous calculons une *séparabilité* de type Fisher fondée sur les distances.

3.6 Contrôles de robustesse et de validité

Nous rapportons des estimations multi-graines avec intervalles bootstrap et croisons les indicateurs avec trois apprenants (régression logistique, SVM linéaire, forêt aléatoire) ; la forêt aléatoire, dépourvue de toute notion intégrée de voisinage, prémunit contre une théorie auto-validante. Tous les modèles séquentiels sont strictement causaux (anti-fuite) ; un *landmark* fixe l'horizon de risque ; un générateur nul doit prédire un écart nul. Nous préenregistrons le signe prédit, une bande d'amplitude et un seuil de réfutation avant chaque expérience d'indépendance et de profondeur (annexe C).

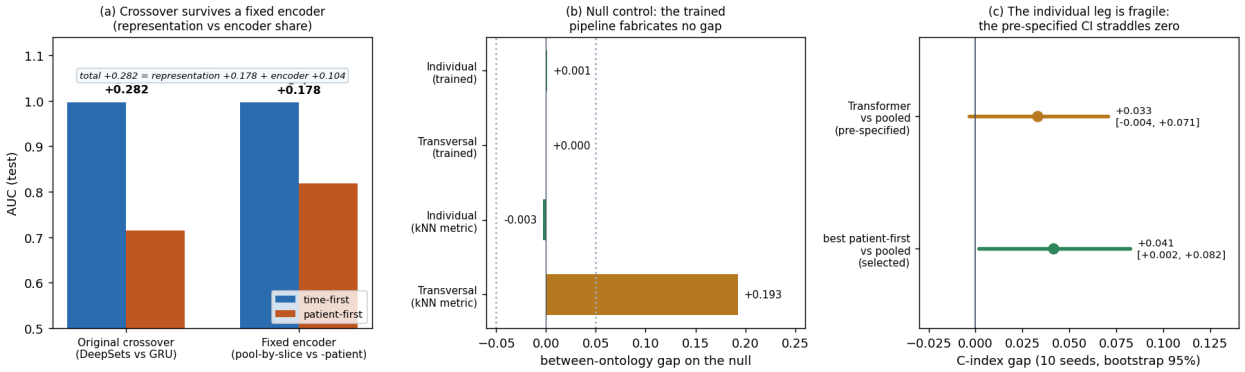

FIGURE 1 – (a) Le croisement transversal survit à un encodeur fixé, décomposant le $0{,}282$ brut en parts de représentation et d'encodeur. (b) Sur un générateur nul, le pipeline entraîné ne produit aucun écart, alors même que la métrique sans modèle continue de favoriser temps-en-tête : dissociation entre métrique et apprenabilité. (c) Le pied individuel est fragile : l'IC bootstrap pré-spécifié enjambe zéro.

3.7 Ablation causale

Pour passer de la prédiction à la cause, nous intervenons sur la métrique à étiquette fixée et information fixée, par un redimensionnement diagonal inversible du bloc discriminant, $\text{rep}_s = [\, s\, z_{\text{aligné}},\ (1/s)\, z_{\text{désaligné}},\ \text{covariables}\,]$. Une application inversible préserve l'information mais change la métrique euclidienne, donc la pureté et l'apprenabilité pour un apprenant sensible à la métrique. Nous mesurons, sans standardisation, la pureté et l'AUC d'une machine à noyau à base radiale. Deux conditions de portée s'appliquent : les redimensionnements forment une sous-famille de l'ensemble des déformations métriques, et l'intervention agit sur des coordonnées construites, de sorte qu'elle établit la causalité le long de métrique $\rightarrow$ voisinage $\rightarrow$ performance, et non de bout en bout depuis la représentation.

4 Résultats

4.1 Le croisement asymétrique et sa décomposition honnête

Sur deux tâches portant sur les mêmes données, à budget égalisé, le classement s'inverse avec la tâche : patient-en-tête domine la tâche individuelle (concordance $0{,}785$ contre $0{,}764$), temps-en-tête domine la tâche transversale (AUC $0{,}997$ contre $0{,}715$). Deux réserves maintiennent l'honnêteté du constat (figure 1). Le croisement est *asymétrique* : l'avantage transversal est net ($\approx 0{,}28$), l'avantage individuel est modeste ($\approx 0{,}02$), et un bootstrap à dix graines de la comparaison individuelle pré-spécifiée donne $+0{,}033$ avec un IC à 95% de $[-0{,}004, 0{,}071]$, qui enjambe zéro. Ensuite, le contrôle à encodeur fixé décompose l'écart transversal : $0{,}282 = \underbrace{0{,}178}_{\text{représentation}} + \underbrace{0{,}104}_{\text{encodeur}}$. Un peu plus du tiers du croisement affiché en titre était un artefact d'encodeur ; la plus grande part est représentationnelle.

4.2 Contrôle nul et dissociation métrique contre apprenabilité

Sur le générateur neutre du point de vue de la représentation, l'écart entraîné est de $+0{,}001$ (individuel) et $+0{,}000$ (transversal), très en dessous du seuil préenregistré de $0{,}05$; retirer la vague fait s'effondrer l'avantage transversal de $+0{,}28$ à zéro. La métrique sans modèle, elle, favorise toujours temps-en-tête de

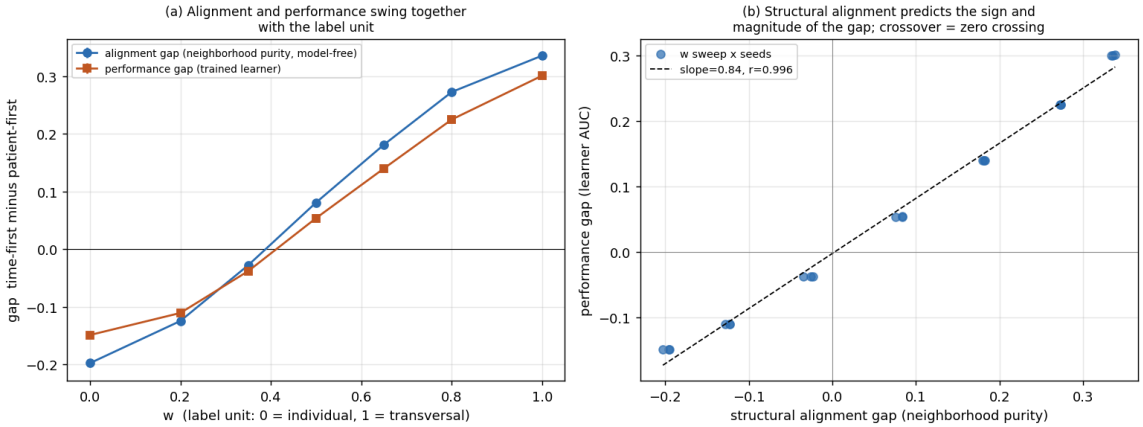

FIGURE 2 – (a) Alignement et performance basculent ensemble avec l'unité d'étiquette et franchissent zéro au même point. (b) L'écart d'alignement structurel prédit le signe et l'amplitude de l'écart de performance ($r = 0{,}996$, pente $0{,}84$).

$+0{,}19$ sur l'étiquette transversale sans vague, alors que les modèles entraînés sont à parité : un cas où un écart d'alignement structurel ne produit pas d'écart de performance parce que le modèle désavantagé reconstruit l'information manquante. Ce n'est pas une fuite, mais la distinction métrique contre apprenabilité prise sur le fait.

4.3 Reproduction géométrique sans entraînement

Une sonde des k plus proches voisins, n'entraînant aucun modèle, reproduit le croisement : sur la tâche transversale, la métrique temps-en-tête atteint une AUC de $0{,}998$ contre $0{,}740$, tandis que sur la tâche individuelle les deux métriques sont à parité, ce qui concorde avec la fragilité de ce pied. Le croisement est donc une propriété de la métrique, antérieure à tout optimiseur.

4.4 Un prédicteur structurel sans modèle

En paramétrant les tâches par w, nous mesurons l'écart d'alignement sans modèle et l'écart de performance d'un apprenant véritable (figure 2). Ils basculent ensemble du négatif au positif et le croisement devient un passage par zéro, avec $r = 0{,}996$ et une pente de $0{,}84$. Parce qu'une corrélation aussi élevée sur un balayage lisse invite à la suspicion, nous en rapportons les statistiques : 21 points répartis sur 7 niveaux de tâche ; bootstrap naïf $[0{,}995, 0{,}998]$; bootstrap par grappes rééchantillonnant des niveaux de tâche entiers $[0{,}994, 0{,}999]$; permutation $p < 0{,}001$. La réserve honnête ne porte pas sur la significativité mais sur l'interprétation, puisque alignement et performance sont tous deux pilotés par la famille de tâches, ce qui est précisément pourquoi l'ablation causale est nécessaire.

4.5 Robustesse

Sur dix graines, quatre indicateurs structurels et trois apprenants, la corrélation reste dans $[0{,}987, 0{,}999]$; la forêt aléatoire, sans aucune notion de voisinage, est elle-même prédite à $0{,}999$ par une mesure de voisinage, ce qui écarte l'artefact d'optimisation et la théorie auto-validante (annexe D).

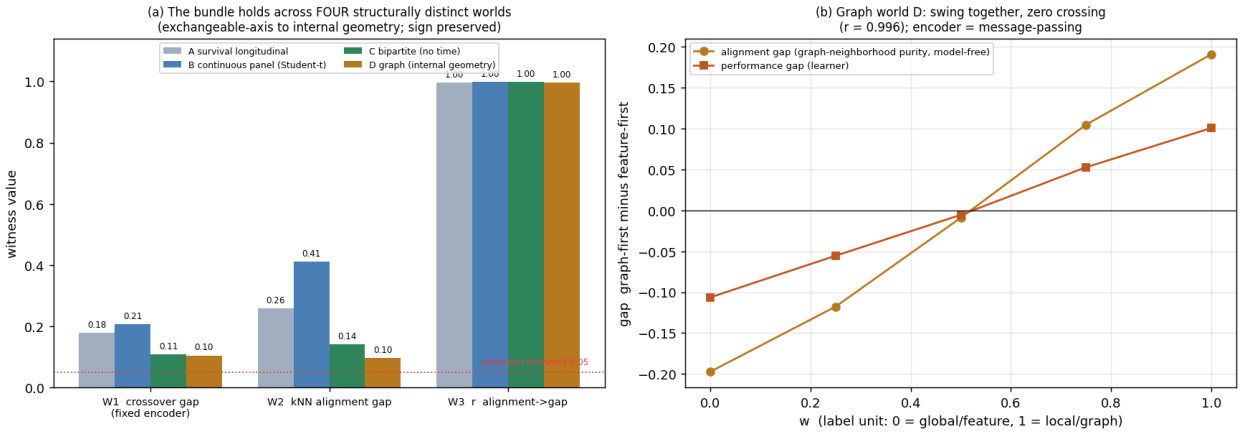

FIGURE 3 – (a) Les trois témoins sur quatre mondes ; le signe est préservé partout, l'amplitude est propre à chaque monde. (b) Le balayage alignement-écart dans le monde de graphes D, appris par un encodeur à passage de messages ($r = 0{,}996$).

TABLE 1 – Trois témoins et statistiques de corrélation sur quatre mondes. T1 : écart de croisement à encodeur fixé. T2 : écart d'alignement kNN. T3 : r de Pearson alignement-écart, avec intervalle bootstrap par grappes à 95% et permutation p.

Monde	T1	T2	T3 (r)	grappes 95%	perm. p
A survie longitudinale	0,178	0,258	0,996	[0,994, 0,999]	$< 0{,}001$
B panel continu (Student-t)	0,207	0,411	0,999	[0,997, 1,000]	$< 0{,}001$
C bipartite (sans temps)	0,108	0,141	0,998	[0,993, 0,999]	$< 0{,}001$
D graphe (géométrie interne)	0,104	0,097	0,996	[0,980, 0,998]	$< 0{,}001$

4.6 Indépendance expérimentale sur quatre mondes

Nous avons répliqué le faisceau réduit (croisement à encodeur fixé, alignement kNN, signe de la relation alignement-écart) sur trois générateurs supplémentaires, préenregistrés avant chaque exécution (figure 3, table 1). L'écart de croisement à encodeur fixé vaut $0{,}178$, $0{,}207$, $0{,}108$, $0{,}104$ de A à D ; l'écart d'alignement $0{,}258$, $0{,}411$, $0{,}141$, $0{,}097$; la corrélation alignement-écart se situe entre $0{,}996$ et $0{,}999$, avec des intervalles bootstrap par grappes au-dessus de $0{,}98$ et une permutation $p < 0{,}001$ dans chaque monde. Le passage du monde C au monde D quitte la modalité du tenseur échangeable pour un graphe doté d'une géométrie interne apprise par passage de messages. Nous revendiquons la généralité du signe, non celle de l'amplitude. Deux limites doivent être énoncées avec le résultat : les quatre mondes partagent un squelette latent à deux échelles, de sorte que leur indépendance est de *modalité*, non statistique, et un générateur brisant cette structure reste à tester ; et ils passent par une seule implémentation, de sorte qu'un artefact de pipeline commun n'est pas exclu.

4.7 Mesure directe données-vers-performance

Nous balayons la taille de la cohorte d'entraînement sous encodeur fixé, sur la tâche transversale (figure 4). La représentation temps-en-tête atteint une AUC $> 0{,}90$ dès $N \leq 200$ et reste proche du plafond ; la représentation patient-en-tête plafonne sous $0{,}85$ et ne referme pas l'écart sur deux ordres de grandeur, sans jamais atteindre $0{,}85$. C'est un énoncé précis et, à un égard, plus fort qu'une différence d'efficience en échantillons. Là où les deux sérialisations peuvent converger, l'effet apparaît comme une différence de

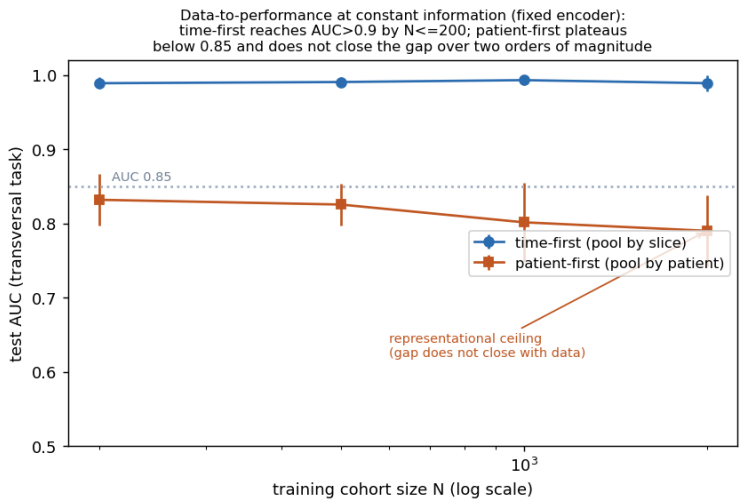

FIGURE 4 – Données-vers-performance à information constante, encodeur fixé, tâche transversale. Temps-en-tête atteint une AUC $> 0{,}9$ dès $N \leq 200$; patient-en-tête plafonne sous $0{,}85$ et ne referme pas l'écart : un plafond représentationnel.

vitesse ; ici, parce que le regroupement désaligné ne peut pas débiaiser la vague, il apparaît comme un plafond représentationnel. Nous gardons le vocabulaire exact et traitons ces deux affirmations comme distinctes plutôt que comme une seule. L'affirmation d'efficience en échantillons, selon laquelle la représentation la mieux alignée atteint un seuil commun avec moins d'échantillons, vaut là où les deux sérialisations convergent. L'affirmation plus forte de plafond représentationnel, selon laquelle la représentation désalignée n'atteint pas du tout la performance de la représentation alignée, vaut là où le débiaisage est impossible. Elles partagent une cause, la représentation fixant ce qui est facile à apprendre, mais elles ne sont pas le même énoncé, et une tâche exhibe l'une ou l'autre selon que le regroupement désaligné peut ou non, en principe, récupérer l'étiquette. Nous avons vérifié cette dépendance directement : sur une tâche bipartite dont l'étiquette est un véritable agrégat de groupe mais ne porte aucune nuisance inaccessible, le regroupement désaligné ne plafonne pas et les deux représentations convergent à $0{,}001$ près ; le plafond est donc spécifique aux tâches où le regroupement désaligné ne peut pas récupérer l'étiquette, et ailleurs l'effet est la plus douce différence de vitesse.

4.8 De la prédiction à la cause, et le mécanisme

En intervenant sur la métrique par redimensionnement inversible à étiquette fixée et information fixée (figure 6a), la pureté et la performance de la machine à noyau s'élèvent ensemble, $r = 0{,}942$, au-dessus du seuil préenregistré de $0{,}7$. Les deux extrêmes sont les deux ontologies : le choix d'ontologie est donc un point sur un continuum de métriques. Le prédicteur est par là causal au niveau de la métrique ; nous redisons que cela établit métrique $\rightarrow$ voisinage $\rightarrow$ performance, et non la chaîne complète depuis la représentation, dont le premier maillon repose sur la construction et sur la reproduction sans modèle exposée plus haut. Le lien est également fondé : sur les données d'ablation, l'erreur de l'apprenant est affine en l'impureté (figure 6b), $R^2 = 0{,}887$, pente positive, comme le prédit un biais de régularité, quoique la pente ($\approx 0{,}21$) reflète notre usage de $1 - \text{AUC}$ plutôt que du taux d'erreur de classification et ne doive pas être surinterprétée.

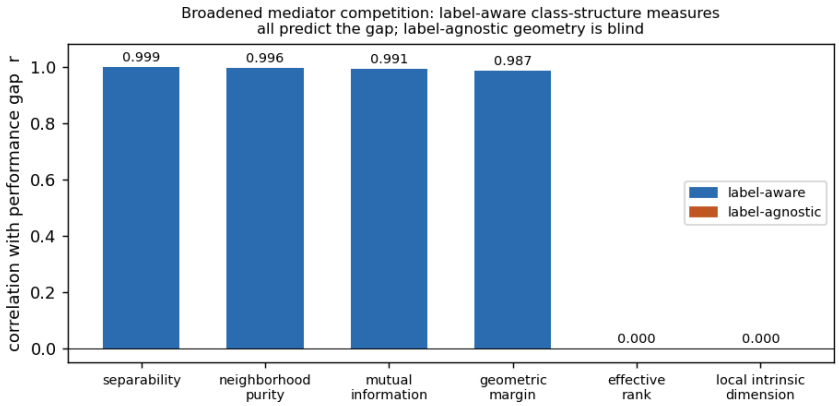

FIGURE 5 – Compétition élargie entre médiateurs sur la famille de tâches. Quatre mesures structurelles sensibles à l'étiquette prédisent toutes l'écart de performance ($r \in [0{,}99, 1{,}00]$) ; deux descripteurs géométriques aveugles à l'étiquette y sont insensibles ($r = 0{,}00$). La quantité active est la structure de classes, sensible à l'étiquette, dans la métrique, dont la pureté de voisinage est un estimateur suffisant.

4.9 Minimalité, honnêtement réfutée

La pureté de voisinage seule explique $R^2 = 0{,}992$ de l'écart de performance, et l'ajout de la séparabilité ne l'améliore que de $0{,}006$: la pureté est une statistique suffisante (figure 6c). Mais elle n'est pas unique. La séparabilité seule atteint $0{,}997$, marginalement mieux, et les corrélations partielles penchent contre la pureté (perf∼sép sachant pureté $= 0{,}86$; perf∼pureté sachant sép $= 0{,}48$). Pureté et séparabilité sont des reflets colinéaires de la structure des classes dans la métrique induite ; les données ne couronnent pas la pureté. Élargir la compétition (figure 5) à quatre mesures structurelles sensibles à l'étiquette (pureté, séparabilité, information mutuelle, marge géométrique) et deux descripteurs géométriques aveugles à l'étiquette (rang effectif, dimension intrinsèque locale) affine le diagnostic : les quatre mesures sensibles à l'étiquette prédisent toutes l'écart (corrélation entre $0{,}99$ et $1{,}00$), tandis que les deux descripteurs aveugles à l'étiquette ne prédisent rien ($r = 0{,}00$), parce qu'ils ne distinguent pas les deux sérialisations, qui sont des recentrages des mêmes points. Le médiateur est donc la structure de classes, sensible à l'étiquette, dans la métrique ; des descripteurs purement géométriques, aveugles à l'étiquette, ne suffisent pas. Nous revendiquons un médiateur causal, structurel, de niveau métrique, dont la pureté est un estimateur suffisant et commode, non la variable fondamentale unique.

4.10 Validation sur panels longitudinaux réels

Nous testons enfin le prédicteur sans modèle sur deux panels réels mesurés, dont les deux sérialisations sont une transposition stricte (donc informationnellement équivalentes) : *dietox* (72 porcs, 12 visites, poids) et *BtheB* (100 patients, score de dépression BDI sur 5 visites, clinique). En balayant l'unité d'étiquette et en mesurant l'écart d'alignement sans modèle contre l'écart de performance de l'apprenant (figure 7), le prédicteur tient : $r = 0{,}983$ sur le panel agricole et $r = 0{,}996$ sur le panel clinique, tous deux avec un passage par zéro. L'effet survit donc au bruit de mesure réel sur des trajectoires véritablement observées. Comme contrôle de réalisme de surface, nous calibrons le monde A sur la cohorte réelle PBC (annexe E) : taux d'événement $0{,}385$, bilirubine médiane $1{,}4$ (IIQ $0{,}8$–$3{,}4$), suivi médian de 1730 jours. Nous restons prudents sur ce que cela valide : les panels réels confirment la moitié aval, corrélationnelle, de la chaîne, à savoir que la pureté de voisinage prédit l'écart d'apprentissage sur trajectoires observées, mais non la moitié causale, puisque l'intervention représentation-vers-métrique de la section 4.8 demeure entièrement synthétique. Une

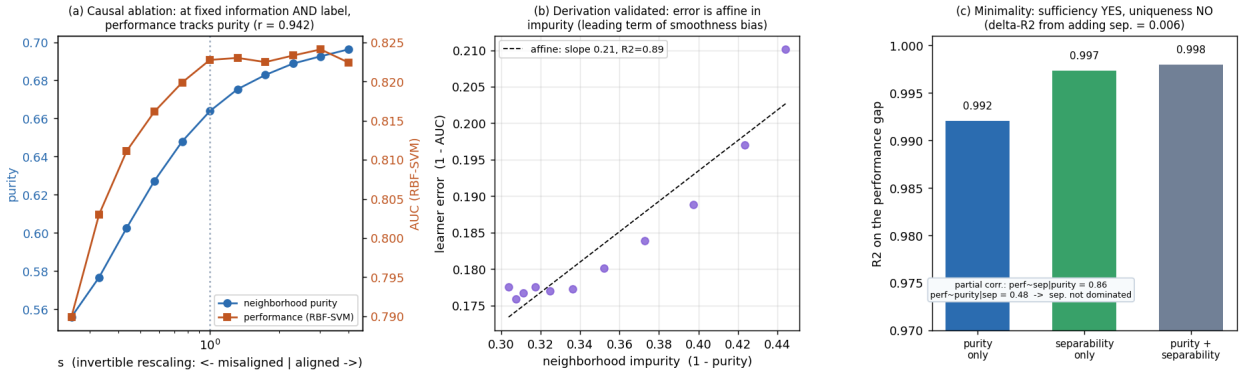

FIGURE 6 – (a) Ablation causale : sous un redimensionnement préservant l'information et fixant l'étiquette, la performance suit la pureté ($r = 0{,}942$). (b) L'erreur est affine en l'impureté ($R^2 = 0{,}887$). (c) Minimalité : la pureté est suffisante mais non unique ; la séparabilité prédit tout aussi bien.

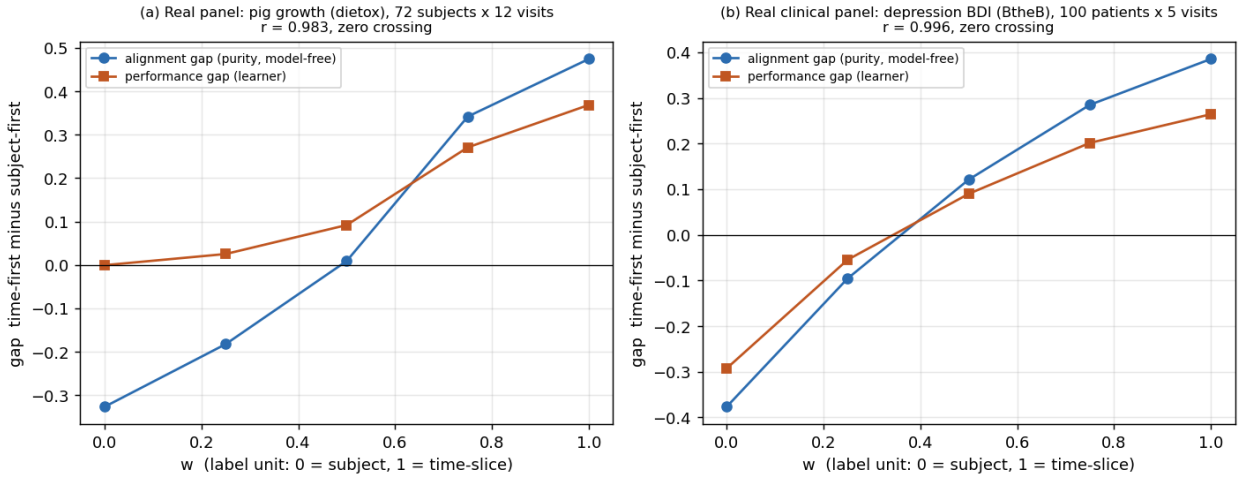

FIGURE 7 – Le prédicteur structurel sans modèle sur deux panels longitudinaux réels. (a) Croissance porcine (dietox), $r = 0{,}983$. (b) Scores cliniques de dépression (BtheB), $r = 0{,}996$. Les deux franchissent zéro.

preuve du lien causal sur données réelles est laissée aux travaux futurs.

4.11 Régime d'échec du prédicteur et robustesse à la quasi-bijection

Un prédicteur mature doit nommer là où il échoue et tolérer l'imperfection ; nous avons préenregistré les deux tests (figure 8). **Régime d'échec.** Sur une tâche balayée du linéairement séparable au structuré en XOR, à représentation fixée, la pureté de voisinage reste modérée ($\approx 0{,}81$ au pôle XOR) et une machine à noyau RBF, biaisée vers la régularité, la suit ($r = 0{,}99$ sur le balayage ; AUC $0{,}99$ au pôle XOR), mais un apprenant linéaire global s'effondre au niveau du hasard ($0{,}50$) malgré cette pureté modérée. Le prédicteur caractérise donc l'apprentissage dominé par le voisinage induit ; un apprenant à biais global sur une étiquette localement pure mais globalement non linéaire lui échappe. Formulé positivement, le prédicteur est valide pour les apprenants dont la généralisation est gouvernée par l'homogénéité locale des étiquettes dans la métrique de la représentation, c'est-à-dire les apprenants biaisés vers la régularité ou à voisinage lipschitzien (plus proches voisins, machines à noyau, et empiriquement les ensembles d'arbres) ; une caractérisation formelle de ce domaine de validité, en termes de classe d'hypothèses admettant une décomposition de voisinage

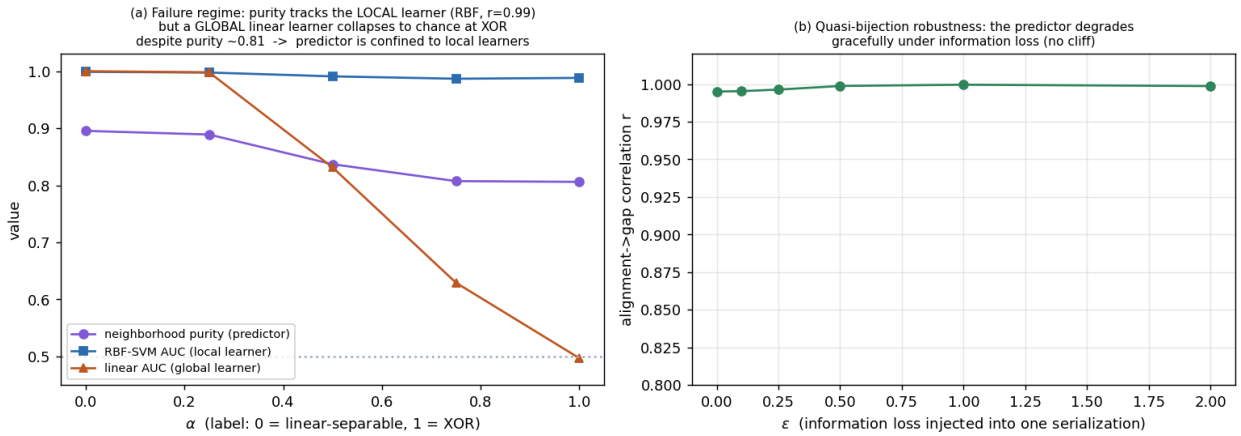

FIGURE 8 – (a) Régime d'échec : la pureté et un apprenant RBF local restent élevés tandis qu'un apprenant linéaire global s'effondre au niveau du hasard au pôle XOR ; le prédicteur est confiné aux apprenants dominés par le voisinage. (b) Robustesse : sous perte d'information injectée (bijection approchée), la corrélation alignement-écart se dégrade progressivement, sans falaise.

lipschitzienne, est laissée aux travaux futurs. Notre préenregistrement supposait que l'échec frapperait l'apprenant à noyau ; l'inverse s'est produit, l'apprenant à noyau local suit la pureté et l'apprenant linéaire global lui échappe, et nous rapportons l'inversion comme le préenregistrement l'exige. **Robustesse à la quasi-bijection.** En injectant une perte d'information dans une des sérialisations, de sorte que les deux représentations ne soient plus qu'approximativement reconstructibles, la corrélation alignement-écart se dégrade progressivement plutôt qu'en falaise : elle reste au-dessus de 0,99 jusqu'à une corruption substantielle, parce que le prédicteur est empirique et mesure la pureté de la représentation effectivement fournie. La bijection exacte est une condition suffisante propre, non une exigence fragile.

5 Discussion

Mécanisme. Une représentation induit une *famille de métriques dépendante du modèle* : une convolution, un Transformer et un réseau à passage de messages n'en voient pas exactement la même, et le contrôle à encodeur fixé tient le modèle constant afin que le contraste de métrique soit imputable à la représentation. Strictement, la métrique opérante est *co-induite* par la représentation et la classe d'hypothèses, et n'est pas une propriété de la représentation seule ; notre revendication est donc conditionnelle : à classe d'hypothèses fixée, la représentation déplace la métrique, ce qu'opérationnalise exactement le contrôle à encodeur fixé. Cette métrique fixe ce qui compte comme proche, ce qui fixe un voisinage dont l'homogénéité en étiquettes gouverne ce qu'un apprenant biaisé vers la régularité trouve facile. L'ablation établit cela comme un mécanisme au niveau de la métrique ; l'analyse de minimalité le borne : la quantité active est la structure de classes dans la métrique, et la pureté en est une mesure suffisante.

La représentation comme hypothèse sur les similarités pertinentes. Nous proposons, à titre de lecture, qu'une représentation est une hypothèse sur les similarités pertinentes (Edelman, 1998; Goodman, 1955; Quine, 1969) ; notre ajout est que cette similarité, une fois mesurée, prédit et (au niveau de la métrique) cause l'efficience en échantillons.

Efficience en échantillons, optimisation et généralisation. Notre protocole sépare la part structurelle du contenu brut, mais non pleinement la facilité d'optimisation, la qualité d'apprentissage et la généralisation ; la mesure directe de la section 4.7 montre que l'effet transversal est un plafond, non une simple vitesse.

Efficience en échantillons contre optimisation. Nous opposons représentation et optimisation par souci de clarté, mais elles interagissent : l'ablation emploie une machine à noyau sans descente de gradient, ce qui isole en partie l'effet de la dynamique d'optimisation ; pour autant, une représentation remodèle aussi le paysage de perte que parcourt un optimiseur du premier ordre, interaction que nous n'étudions pas.

Limites. La démonstration est synthétique, complétée par deux panels réels ; les quatre mondes synthétiques partagent un squelette à deux échelles et une seule implémentation ; l'ablation couvre une sous-famille de changements métriques et reste locale à la métrique ; la minimalité est réfutée, non établie ; la dérivation formelle est validée dans sa forme, non écrite comme une borne. La validité du prédicteur est confinée à l'apprentissage dominé par le voisinage induit (section 4.11), de sorte qu'elle couvre une large part de l'apprentissage automatique moderne mais non les apprenants dont le biais est global ou symbolique. Parce que les deux agrégations induisent des distributions d'entrée et de mini-lots différentes même à encodeur fixé, un confondant distributionnel résiduel ne peut être totalement exclu. La famille de tâches est une interpolation construite d'une unité d'étiquette, dans les mondes synthétiques comme dans les panels réels ; savoir si des tâches véritablement indépendantes et non interpolées reproduisent la même continuité reste à tester. La vision avec sa localité spatiale et le langage avec son ordre séquentiel ne sont pas testés pour eux-mêmes, et un graphe ne capture pas la géométrie invariante par translation d'une grille.

Une heuristique d'ingénierie. Identifier l'unité de décision ; construire l'ontologie dont les objets élémentaires coïncident avec elle ; vérifier qu'elle préserve les dépendances pertinentes ; adapter l'architecture seulement ensuite. L'ontologie n'est pas un gain inconditionnel mais un pari sur l'alignement, et ce pari peut se perdre : le cas le plus instructif est celui où la représentation la plus riche apprend le moins, parce que sa richesse est orthogonale à la question posée.

6 Conclusion

À information constante et démontrablement égale, la représentation fixe les proximités qu'un apprenant utilise, et donc son efficience en échantillons. Nous donnons une quantité structurelle sans modèle qui prédit, en signe et en amplitude, l'écart d'apprentissage entre sérialisations équivalentes ; nous montrons par une intervention préservant l'information qu'elle médie causalement cet écart au niveau de la métrique ; nous montrons que le phénomène se reproduit sur quatre mondes synthétiques et deux panels réels ; et nous montrons, contre notre propre attente initiale, que le médiateur est structurel mais non uniquement la pureté de voisinage. Ce qui est démontré est un prédicteur causal et structurel de l'écart d'apprentissage ; ce qui reste est une dérivation écrite et une démonstration au-delà du régime à deux échelles, tabulaire et de graphes.

Déclaration d'impact élargi (Broader Impact Statement)

Ce travail est méthodologique et repose sur des générateurs synthétiques et des jeux de données de recherche publics et désidentifiés (dietox, BtheB, PBC). Il suggère qu'à information constante, le choix de la représentation des données change matériellement l'apprenabilité, ce qui peut orienter vers une modélisation plus économe en échantillons et moins intensive en calcul dans les domaines longitudinaux et relationnels, y compris en santé. Le même levier, employé sans précaution, peut biaiser les motifs qu'un modèle trouve faciles ; les praticiens devraient traiter le choix d'ontologie comme une décision de modélisation ayant des conséquences sur l'équité et la distribution des erreurs, et non comme une étape neutre de prétraitement. Aucune donnée de sujets humains n'a été collectée ; le jeu de données clinique utilisé est un jeu d'enseignement publié de longue date et désidentifié.

Références

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv :1806.01261*, 2018.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32) :15849–15854, 2019. doi : 10.1073/pnas.1903070116.

Y. Bengio, A. Courville, and P. Vincent. Representation learning : A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8) :1798–1828, 2013. doi : 10.1109/TPAMI.2013.50.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning : Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv :2104.13478*, 2021.

B. de Finetti. La prévision : ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7 (1) :1–68, 1937.

S. Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4) :449–467, 1998.

N. Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA, 1955.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel : Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

T. M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, 1980.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent : Where bigger models and more data hurt. *arXiv preprint arXiv :1912.02292*, 2019.

W. V. Quine. Natural kinds. In *Ontological Relativity and Other Essays*, pages 114–138. Columbia University Press, New York, 1969.

N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, et al. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 5301–5310, 2019.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15 (3) :351–357, 1950.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13(2) :238–241, 1951.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7) : 1341–1390, 1996. doi : 10.1162/neco.1996.8.7.1341.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1) :67–82, 1997. doi : 10.1109/4235.585893.

A Équations des générateurs

Monde A (survie). Effets aléatoires $(L_i, \text{pente}_i) \sim \mathcal{N}(0, \Sigma)$; le processus intra $w_i(t)$ est une tendance linéaire plus un AR(1) gaussien ; marqueur observé = latent + bruit de mesure gaussien ; risque instantané comme en éq. (1) avec un h_0 de Weibull ; temps d'événement par risque cumulé inverse ; censure administrative et aléatoire.

Monde B (panel continu). Niveau L_i tiré d'un mélange gaussien à deux composantes (bimodal) ; processus intra par retour à la moyenne $w_{i,b} = \phi\, w_{i,b-1} + \epsilon$, $\epsilon \sim t_\nu$, plus une sinusoïde propre au patient $A_i \sin(\omega_i t + \varphi_i)$; observé $= L_i + w +$ osc+ bruit de mesure t_4 ; attrition aléatoire des visites. Pas de survie ; le marqueur lui-même porte le signal à deux échelles.

Monde C (bipartite). Facteurs utilisateurs P_u, facteurs objets $Q_i \sim \mathcal{N}(0, I_d)$; valeur $V_{ui} = \langle P_u, Q_i \rangle + a_u$ (niveau objet fixé à zéro afin que la nuisance par objet soit une vague pure) ; observé $O_{ui} = V_{ui} + \delta_i +$ bruit avec un décalage par objet δ_i ; entrées observées avec une densité ρ (parcimonie). Objet-en-tête agrège la colonne (retire δ_i), utilisateur-en-tête la ligne.

Monde D (graphe). Modèle à blocs stochastiques avec probabilités d'arêtes intra et inter ; signal des nœuds $o_i = s_i + g_i +$ bruit avec une part lisse constante par bloc s_i et une part locale i.i.d. g_i ; le contexte graphe-en-tête est la moyenne des voisins d'adjacence (un pas de passage de messages), le contexte attribut-en-tête la moyenne globale. L'étiquette locale est $g_i > 0$; l'étiquette globale est s_i au-dessus de sa médiane.

B Encodeurs et hyperparamètres

Encodeurs séquentiels : GRU (état caché 24–32) et Transformer à une couche avec masquage causal ($d_{\text{model}} = 20$, 2 têtes). MLP agrégé personne-période (état caché ≈ 58). Encodeur d'ensembles : MLP ϕ, ρ avec contexte agrégé par moyenne, état caché 32. Encodeur de graphe : passage de messages par moyenne à un saut, avec la même forme ϕ, ρ, état caché 32. Optimiseur Adam, taux d'apprentissage 10^{-2} ; époques 100–400 selon l'expérience ; budgets de paramètres égalisés au sein de chaque comparaison et journalisés. Le kNN et la pureté utilisent $k = 20$–25. La machine à noyau RBF utilise un $\gamma = 0{,}3$ fixe et aucune standardisation dans l'ablation.

C Synthèse du préenregistrement

Chaque expérience d'indépendance et de profondeur a été préenregistrée avec un horodatage UTC avant exécution, fixant le signe prédit, une bande d'amplitude et un seuil de réfutation. Horodatages : contrôles de rang 0, `2026-08-08T18:44Z` ; monde B, `2026-08-09T13:46Z` ; monde C, `2026-08-09T14:18Z` ; monde D, `2026-08-09T14:45Z` ; profondeur (ablation, minimalité, dérivation), `2026-08-09T15:01Z` ; bordure (régime d'échec, quasi-bijection), `2026-08-09T17:56Z`. Toutes les prédictions et tous les seuils sont publiés dans le paquet de reproductibilité. Nous notons un résultat honnête : le préenregistrement du régime d'échec supposait que l'apprenant à noyau échapperait au prédicteur, alors que c'est l'apprenant linéaire global qui y a échappé ; l'inversion est rapportée en section 4.11.

D Tables de robustesse et de corrélation

La matrice de robustesse croise quatre indicateurs structurels (pureté de voisinage à $k \in \{10, 25, 50\}$ et séparabilité) avec trois apprenants (régression logistique, SVM linéaire, forêt aléatoire) sur dix graines ;

la corrélation alignement-écart reste dans $[0{,}987, 0{,}999]$, la forêt aléatoire étant prédite à $0{,}999$. Minimalité : $R^2_{\text{pureté}} = 0{,}992$, $R^2_{\text{sép}} = 0{,}997$, $R^2_{\text{deux}} = 0{,}998$; corrélations partielles perf∼sép|pureté $= 0{,}864$, perf∼pureté|sép $= 0{,}480$. Compétition élargie (robustesse post hoc, non préenregistrée) : corrélation de chaque écart structurel avec l'écart de performance sur la famille de tâches : séparabilité $0{,}999$, pureté de voisinage $0{,}996$, information mutuelle $0{,}991$, marge géométrique $0{,}987$ (toutes sensibles à l'étiquette), contre rang effectif $0{,}000$ et dimension intrinsèque locale $0{,}000$ (toutes deux aveugles à l'étiquette). Le contrôle de plafond sans vague (section 4.7) et cette compétition élargie sont des analyses de robustesse post hoc et ne faisaient pas partie du préenregistrement.

E Détails de la validation sur données réelles

Les panels sont chargés depuis des jeux de données R publics (dietox, geepack ; BtheB, HSAUR ; pbc, survival). Chaque panel est disposé en matrice sujet-par-visite par report de la dernière observation ; les deux sérialisations sont la matrice et sa transposée (bijection stricte). Le balayage de l'unité d'étiquette mélange des composantes standardisées de contexte tranche et de contexte sujet ; l'écart d'alignement est la pureté de voisinage (sans modèle), l'écart de performance celui d'un apprenant logistique, sur cinq partitions aléatoires par w. Résultats : dietox $r = 0{,}983$; BtheB $r = 0{,}996$; les deux franchissent zéro. Cibles de calibration PBC (cohorte de référence, $n = 418$) : taux d'événement (transplantation ou décès) $0{,}385$, bilirubine médiane $1{,}4$ mg/dL (IIQ $0{,}8$–$3{,}4$), suivi médian 1730 jours, âge médian 51 ans.

F Liste de contrôle de reproductibilité

Le code, les graines, les préenregistrements et un script de reproduction en une commande sont publiés. Tout l'aléa est initialisé par graine ; les figures se régénèrent depuis les fichiers de résultats sauvegardés. Les générateurs synthétiques sont *ex nihilo*, avec paramètres publiés ; les panels réels sont publics. Le calcul est modeste (CPU uniquement ; l'expérience la plus lourde, la courbe d'apprentissage de l'encodeur d'ensembles par sujet, s'exécute en quelques dizaines de minutes sur un seul cœur). Aucune donnée ni aucun modèle propriétaire ne sont utilisés.