Classification de processus de points sphériques : comparaison entre statistiques spatiales et Graph Neural Networks

Alexandre MARTIN[1] Amel AIT MOUFFOK[2] Eirini FYTILI[2] Jules IMLER[3] Arnaud TÊTE[3] Sylvie BORTOLI[3] Xavier COUMOUL[3] Stephan CLAVEL[2] Xavier DESCOMBES[1]

[1]Université Côte d'Azur, INRIA, CNRS, INSERM

[2]Inserm, Institut de Pharmacologie Moléculaire et Cellulaire (IPMC)

[3]Université Paris Cité, INSERM, Health & Functional Exposomics - HealthFex

Résumé – Cet article compare les méthodes statistiques traditionnelles et les Graph Neural Networks (GNN) pour la classification d'organoïdes de prostate. Nous modélisons les deux phénotypes principaux des organoïdes de prostate - cystiques et choux-fleurs - par des distributions de points sphériques uniformes et agrégées. L'approche statistique utilise les fonctions K, F et G de Ripley, tandis que l'approche GNN s'appuie sur la tessellation de Voronoi pour construire le graphe. Les performances sont évaluées face à deux types de bruit : gaussien et poivre et sel. Nos résultats démontrent que les méthodes statistiques sont plus efficaces pour cette tâche spécifique, mais sont limitées par leur restriction au cas sphérique. Cette étude offre des perspectives pour l'analyse morphologique des organoïdes de prostate afin de comprendre leur développement et leur réponse aux traitements.

Abstract – This paper compares traditional statistical methods and Graph Neural Networks (GNNs) for classifying prostate organoids. We model the two main phenotypes of prostate organoids - cystic and cauliflower-like - using uniform and clustered spherical point distributions. The statistical approach leverages Ripley's K, F, and G functions, while the GNN approach uses Voronoi tessellation for graph construction. Performance is evaluated against two types of noise: Gaussian and salt-and-pepper. Our results demonstrate that statistical methods are more efficient for this specific task, but are limited by their restriction to the spherical case. This study offers perspectives for the morphological analysis of prostate organoids to understand their development and response to treatments.

1 Introduction

Les organoïdes constituent un modèle in vitro précieux pour étudier le développement tissulaire, la pathogenèse et les réponses aux thérapeutiques [5]. Ces structures tridimensionnelles auto-organisées reproduisent partiellement la complexité des organes in vivo, offrant un compromis entre les cultures cellulaires conventionnelles et les modèles animaux. Les organoïdes de prostate, en particulier, présentent des phénotypes diversifiés, notamment les structures cystiques (sphériques avec une cavité centrale) et les structures en "choux-fleurs" (excroissances irrégulières et amas cellulaires). Ces différences morphologiques reflètent des comportements biologiques distincts.

La classification automatisée de ces phénotypes est un défi important pour l'analyse à haut débit et la personnalisation thérapeutique. Les approches traditionnelles reposent sur l'extraction de descripteurs statistiques, tels que la fonction K de Ripley [6] et les fonctions de distribution des distances F et G [3], pour quantifier les relations de voisinage entre cellules et alimenter des classifieurs supervisés.

Parallèlement, les Graph Neural Networks (GNN) [4] ont émergé comme une approche prometteuse pour analyser des données structurées en graphes. En modélisant les distributions cellulaires comme des graphes, les GNN apprennent directement à partir de la topologie de la distribution sans nécessiter l'extraction manuelle de caractéristiques [9].

Cette étude compare ces deux approches pour la classification des organoïdes de prostate, en modélisant les phénotypes cystique et choux-fleurs par des distributions de points sphériques uniformes et agrégées respectivement. L'objectif est d'évaluer leur capacité à discriminer ces morphologies en présence de bruit gaussien et de bruit poivre et sel, tout en explorant leurs limitations et complémentarités.

2 Méthodes

2.1 Modélisation en processus sphériques

Dans notre approche de modélisation des organoïdes, nous avons adopté une étape de projection sphérique essentielle à expliciter. Bien que les organoïdes de type chou-fleur ne présentent pas naturellement une morphologie sphérique, nous normalisons les coordonnées puis projetons les centres des noyaux cellulaires sur une sphère afin d'établir un cadre de référence commun pour l'analyse comparative des phénotypes, comme illustré en Figure 1. Cette transformation géométrique nous permet d'appliquer des processus sphériques pour générer des distributions cellulaires synthétiques qui préservent les relations spatiales importantes entre cellules. L'avantage principal de cette projection est qu'elle normalise les distances intercellulaires, facilitant ainsi la comparaison directe des motifs d'organisation cellulaire entre différents types d'organoïdes, indépendamment de leurs morphologies globales. Ce choix méthodologique constitue une hypothèse de travail fondamentale pour notre analyse, permettant de nous concentrer sur la distribution spatiale des cellules plutôt que sur la forme externe de l'organoïde.

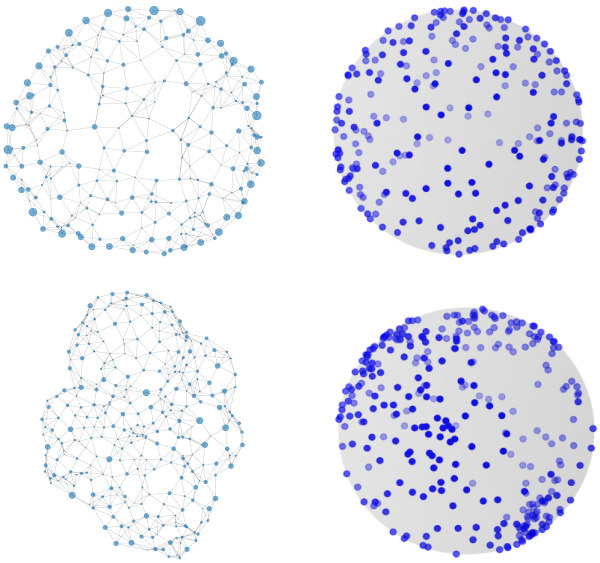

FIGURE 1 : Graphe de l'organoïde (à gauche) et sa projection sur la sphère associée (à droite) pour un organoïde cystique (en haut) et choux-fleur (en bas)

2.2 Génération de données synthétiques

Pour comparer les approches classiques et les GNN dans l'analyse des distributions cellulaires, nous générons deux classes de processus ponctuels sphériques en 3D correspondant aux phénotypes d'organoïdes de prostate observés : uniformes (organoïdes cystiques) et agrégés de type Matérn (organoïdes choux-fleurs).

Les processus uniformes modélisent les organoïdes cystiques, où les cellules sont réparties aléatoirement à la périphérie d'une cavité centrale, suivant :

$$f(x, y, z) = \frac{3}{4\pi}, \quad \text{avec } x^2 + y^2 + z^2 \leq 1 \tag{1}$$

Nous normalisons toutes les coordonnées par le rayon de la sphère englobante ($R = 1$) pour simplifier les analyses.

Les processus de Matérn modélisent les organoïdes choux-fleurs, caractérisés par des amas cellulaires irréguliers. Ils sont générés par un processus parent-enfant où les centres d'agrégats (parents) sont distribués uniformément sur la sphère. Puis, autour de chaque centre, des points (enfants) sont distribués selon une loi gaussienne tronquée de paramètre σ. Nous fixons le nombre de clusters à 10 dans notre étude.

Les deux types de distribution sont visibles en Figure 2.

Pour simuler les imperfections d'acquisition, nous appliquons deux types de bruit aux distributions :

- **Bruit gaussien** : Un bruit gaussien $\mathcal{N}(0, \sigma_g^2)$ est appliqué aux coordonnées de chaque point, avec σ_g variant de 0 à 0,8 par pas de 0,1.
- **Bruit poivre et sel** : Ce bruit modifie la structure même de la distribution en ajoutant et supprimant aléatoirement des points :
  - Ajout de points uniformément répartis sur la sphère (composante « sel »)
  - Suppression aléatoire de points existants (composante « poivre »)

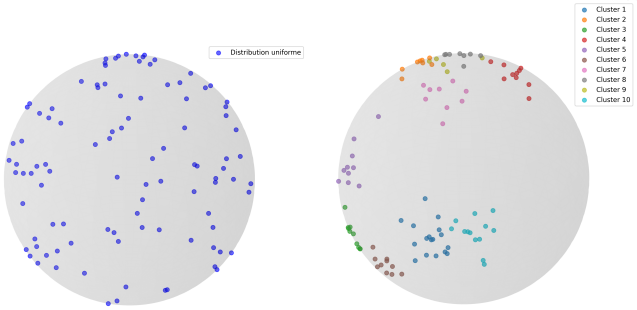

FIGURE 2 : Distribution uniforme (à gauche) et à clusters (à droite)

L'intensité du bruit poivre et sel σ_{ps} varie de 0 à 0,4 par pas de 0,05, où σ_{ps} représente la proportion de points affectés (ajoutés ou supprimés).

Il est à noter que nous contraignons les points à rester sur la sphère après bruitage, pour rester dans le domaine d'application des statistiques spatiales.

2.3 Approche par statistiques spatiales

L'approche classique repose sur l'extraction de trois descripteurs statistiques fondamentaux à partir des distributions de points [1] :

- La fonction K de Ripley, qui caractérise la distribution des distances entre paires de points, estimée par :

$$\hat{K}(r) = \frac{V}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathbf{1}(d_{ij} \leq r) w_{ij} \tag{2}$$

  où V est le volume de la sphère, n le nombre de points, d_{ij} la distance entre les points i et j, et w_{ij} un facteur de correction de bord.

- La fonction F (empty space function), qui mesure la distance d'un point aléatoire au plus proche point du processus, estimée empiriquement sur un ensemble de points tests uniformément distribués.

- La fonction G (nearest neighbor distance function), qui complète les deux fonctions précédentes en caractérisant les distances aux plus proches voisins.

Pour chaque distribution, nous calculons ces trois fonctions sur 20 rayons équidistants, générant ainsi un vecteur de caractéristiques de dimension 60. Ces descripteurs alimentent un classifieur Random Forest (RF) avec 100 arbres et une profondeur maximale de 10.

2.4 Approche par Graph Neural Networks

Notre approche par GNN modélise la distribution cellulaire comme un graphe où :

- Chaque nœud représente un point (cellule) avec ses coordonnées 3D et son volume issu de la tesselation de Voronoi comme attributs

- Les arêtes sont déterminées par la tesselation de Voronoi [2] : deux points sont connectés si leurs cellules de Voronoi partagent une face

Cette construction basée sur Voronoi permet de capturer naturellement les relations de voisinage entre cellules, indépendamment de la distance absolue entre elles.

L'architecture de notre GNN, inspirée du modèle de [7], comprend :

1. Une couche d'entrée projetant les coordonnées 3D dans un espace latent de dimension 64
2. L couches de Graph Attention Network (GATConv), où L varie de 2 à 8
3. Des connexions résiduelles entre chaque couche pour faciliter l'apprentissage
4. Une normalisation par lots (batch normalization) après chaque convolution
5. Une couche de global mean pooling [8]
6. Deux couches fully-connected (128 et 2 neurones) avec activation ReLU et dropout (0,2)

Chaque couche GATConv utilise 4 têtes d'attention, ce qui permet au réseau de capturer différents types de relations entre les nœuds. Les connexions résiduelles pondérées (facteur 0,2) aident à stabiliser l'apprentissage et à combattre le problème d'overfitting.

Pour chaque configuration de L et de niveau/type de bruit, nous réalisons une validation croisée à 5 plis sur 2000 échantillons (1000 par classe) comprenant chacun 100 points, avec optimisation par Adam (learning rate = 0,0005, réduit par un facteur 0,5 lorsque l'accuracy stagne) sur 100 époques maximum avec early stopping.

3 Résultats

3.1 Impact du bruit gaussien

Les résultats obtenus sur la Figure 3 pour la classification d'images et de graphes d'organoïdes à l'aide de différentes architectures de Deep Learning (CNN, GNN) montrent une performance remarquable sous divers niveaux de bruit gaussien. Pour un bruit faible, les modèles statistiques et les GNN de différentes profondeurs (2 à 8) atteignent une précision de 1.0, indiquant une classification parfaite.

À mesure que le bruit augmente, la précision commence à diminuer. Il semble que les modèles statistiques aient une tolérance plus importante au bruit. En général, les modèles GNN de profondeur plus élevée (5 à 8) semblent mieux résister au bruit que les modèles de profondeur plus faible (2 à 4). Cependant, il est important de noter que pour des niveaux de bruit plus élevés, les modèles à grande profondeur (7 et 8) montrent des signes d'overfitting, ce qui peut expliquer la baisse de performance observée.

Ces résultats suggèrent qu'il existe un optimum de la profondeur du GNN. Avec une profondeur trop basse, le réseau ne sera pas capable d'approcher de manière suffisante la fonction de classification optimale. Dans l'autre extrême, avec une profondeur trop importante, le réseau aura une tendance à l'overfitting. Dans notre étude, une profondeur de 5 à 6 couches semble optimale.

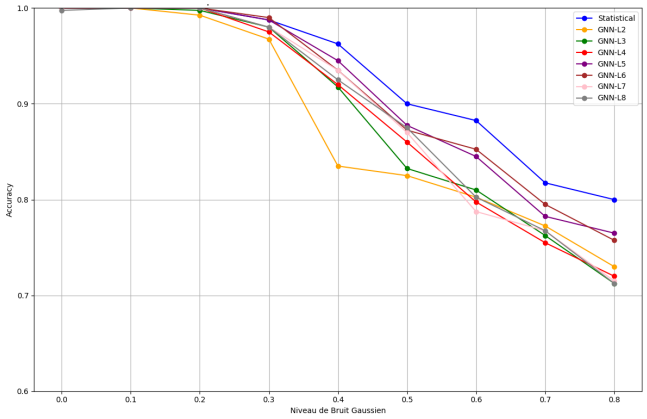

FIGURE 3 : Impact du bruit gaussien et de la profondeur du GNN sur la performance des modèles.

3.2 Impact du bruit poivre et sel

Nous obtenons des résultats similaires pour le bruit poivre et sel, comme illustré dans la Figure 4. Pour des niveaux de bruit faibles, les modèles statistiques et les GNN de différentes profondeurs (2 à 8) atteignent une précision de 1.0. Cependant, à mesure que le bruit augmente, la précision diminue. Les modèles statistiques montrent une meilleure tolérance au bruit, et les GNN de profondeur plus élevée (5 à 8) résistent mieux au bruit que ceux de profondeur plus faible (2 à 4). Pour des niveaux de bruit élevés, les modèles de grande profondeur (7 et 8) montrent des signes d'overfitting. Une profondeur de 5 à 6 couches semble également être optimale pour les GNN dans ce cas-ci.

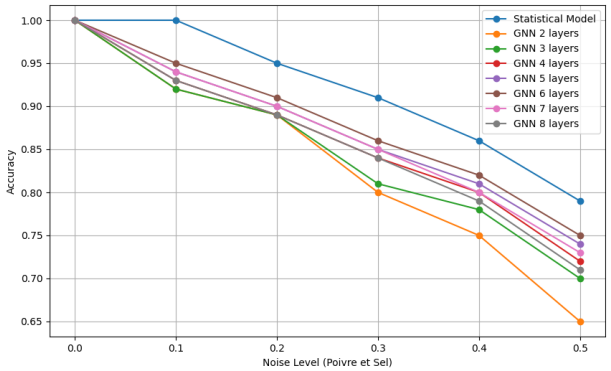

FIGURE 4 : Impact du bruit poivre et sel et de la profondeur du GNN sur la performance des modèles.

3.3 Limitations des approches

Malgré leur efficacité, les méthodes statistiques présentent une limitation majeure : leur forte dépendance aux hypothèses géométriques sous-jacentes. Pour vérifier cette restriction, nous avons testé les deux approches sur des distributions non sphériques (ellipsoïdales avec des ratios d'aspect de 2 :1 à 5 :1).

Dans ce cas, l'accuracy de l'approche statistique chute plus rapidement que celle du GNN. Cette inversion des performances souligne la dépendance des statistiques spatiales aux

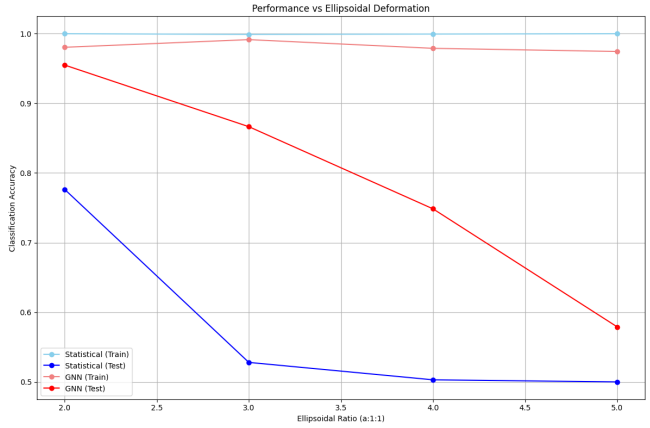

FIGURE 5 : Impact de la différence de géométrie (sphérique vs ellipsoïdale) sur la performance des modèles.

hypothèses géométriques sous-jacentes, là où les GNN démontrent une plus grande flexibilité topologique, et ainsi une meilleure généralisation.

La figure 5 illustre cette sensibilité à la géométrie en comparant les performances des deux types de modèles sur des données ellipsoïdales, ceux-ci ayant été entraînés sur des données sphériques.

4 Conclusion

Cette étude comparative entre méthodes statistiques traditionnelles et Graph Neural Networks pour la classification d'organoïdes de prostate a démontré plusieurs résultats significatifs :

1. Pour les distributions sphériques modélisant les phénotypes cystiques et choux-fleurs, les descripteurs statistiques (fonctions K, F et G de Ripley) surpassent les GNN en termes d'accuracy, avec un écart qui se maintient face aux deux types de bruit (gaussien et poivre et sel).

2. Les deux approches sont plus sensibles au bruit poivre et sel qu'au bruit gaussien, le premier altérant la structure même des distributions spatiales.

3. Pour les GNN, il existe une profondeur optimale (5-6) qui offre le meilleur compromis entre capacité d'apprentissage et risque d'overfitting, indépendamment du type de bruit.

4. Les méthodes statistiques présentent une limitation fondamentale : leur forte dépendance aux hypothèses géométriques. Pour la généralisation aux distributions non-sphériques, les GNN deviennent significativement plus performants.

Ces résultats suggèrent des stratégies complémentaires pour l'analyse morphologique d'organoïdes de prostate : privilégier les méthodes statistiques lorsque la géométrie des structures est régulière et connue a priori, et adopter les GNN pour les circonstances requérant une plus grande flexibilité géométrique.

Les perspectives futures incluent le développement d'approches hybrides combinant la puissance des descripteurs statistiques et la flexibilité topologique des GNN, ainsi que l'application de ces méthodes sur des organoïdes prostatiques réels où l'hétérogénéité cellulaire et les variations de densité ajoutent des niveaux de complexité supplémentaires.

La principale limitation de cette étude réside dans l'utilisation de données synthétiques avec des distributions idéalisées. La modélisation des organoïdes de type choux-fleur par des clusters peut ainsi s'avérer simpliste (il pourrait être intéressant de mixer distribution uniforme et à clusters dans ce cas). Les travaux futurs s'orienteront vers l'extension de ces méthodes à des morphologies plus complexes et des données réelles issues de microscopie 3D, ainsi qu'à l'étude d'autres types d'architectures et de modélisation (architectures adaptées aux nuages de points, ou encore aux graphes géométriques).

Remerciements

Ce projet de recherche a été mené grâce au financement de l'ANR Morpheus (263702).

Références

[1] Adrian BADDELEY et Rolf TURNER : Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3):283–322, 2000.

[2] Manuel CAROLI, Pedro CASTRO, Sébastien LORIOT, Olivier ROUILLER, Monique TEILLAUD et Camille WORMSER : Robust and efficient delaunay triangulations of points on or close to a sphere, 05 2010.

[3] Janine ILLIAN, Antti PENTTINEN, Helga STOYAN et Dietrich STOYAN : Statistical analysis and modelling of spatial point patterns. *John Wiley & Sons*, 2008.

[4] Thomas N KIPF et Max WELLING : Semi-supervised classification with graph convolutional networks. *In International Conference on Learning Representations (ICLR)*, 2016.

[5] Francesco PAMPALONI, Emmanuel G REYNAUD et Ernst HK STELZER : The third dimension bridges the gap between cell culture and live tissue. *Nature reviews Molecular cell biology*, 8(10):839–845, 2007.

[6] Scott ROBESON, Ao LI et Chunfeng HUANG : Point-pattern analysis on the sphere. *Spatial Statistics*, 10, 11 2014.

[7] Petar VELIČKOVIĆ, Guillem CUCURULL, Arantxa CASANOVA, Adriana ROMERO, Pietro LIO et Yoshua BENGIO : Graph attention networks. *International Conference on Learning Representations*, 2018.

[8] Zhitao YING, Jiaxuan YOU, Christopher MORRIS, Xiang REN, Will HAMILTON et Jure LESKOVEC : Hierarchical graph representation learning with differentiable pooling. *In Advances in neural information processing systems*, pages 4800–4810, 2018.

[9] Jie ZHOU, Ganqu CUI, Zhengyan SHENG, Zhengyan YANG, Zhiyuan SU, Cheng LI, Liang-Yu LIU et Maosong SUN : Graph neural networks : A review of methods and applications. *AI Open*, 1:57–81, 2020.