الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Frères Mentouri Constantine 1
Faculté des Sciences de la Nature et de la Vie

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Département de Biologie Appliquée

قسم البيولوجيا التطبيقية

Mémoire présenté en vue de l'obtention du diplôme de Master

Domaine : Sciences de la Nature et de la Vie

Filière : Biotechnologie

Spécialité : Bio-informatique

N° d'ordre :
N° de série :

Titre :

Analyse scRNA-seq Interprétable par Apprentissage Automatique et SHAP pour l'évaluation de la Qualité des Lignages Embryonnaires Précoces.

Présenté par :

Mahcen Esma
Mahmoudi Nour El Houda

Soutenu le : 25/06/2025

Jury d'évaluation :

Président : Dr. Medjroubi Mohammed Larbi (MCB - Constantine 1 Frère Mentouri University).

Encadrant : Dr. Chehili Hamza (MCA - Constantine 1 Frère Mentouri University).

Co-Encadrant : Dr. Chebouba Lokman (MCB - Constantine 1 Frère Mentouri University).

Examinateur : Dr. Krid Adel (MCB - Constantine 1 Frère Mentouri University).

Année universitaire : 2024 - 2025

Remerciements

Avant toute chose, nous remercions sincèrement Dieu de nous avoir accordé la patience, la force et la persévérance nécessaires pour mener à bien ce travail.

Nous tenons à exprimer notre profonde reconnaissance à notre encadrant, Dr.Chehili Hamza, pour sa présence continue, sa disponibilité, et son accompagnement attentif tout au long de ce mémoire. Ses conseils éclairés, sa rigueur scientifique et sa bienveillance ont grandement contribué à la qualité de ce travail.

Nos remerciements les plus sincères vont également à Dr.Chebouba Lokman, notre co-encadrant, qui a été un pilier dans l'élaboration de ce mémoire. Son soutien constant, ses explications claires, et sa patience face à nos doutes ont été d'une aide précieuse à chaque étape du projet. Sa présence a fait toute la différence, et nous lui en sommes profondément reconnaissants.

Nous remercions chaleureusement les membres du jury, Dr.Mohammed Laarbi Medjroubi et Dr.Krid Adel (MCB), pour l'intérêt qu'ils ont porté à notre travail, pour avoir accepté de l'évaluer, et pour leurs remarques constructives qui ont enrichi notre réflexion.

Nous adressons également nos remerciements à l'ensemble de nos enseignants, pour l'enseignement de qualité qu'ils nous ont dispensé durant ces années de formation. Leur engagement a posé les bases solides de notre parcours scientifique.

Dédicace

Parce que derrière chaque réussite, il y a des visages, des voix, des présences.

Ce mémoire est un hommage silencieux à ceux qui, de près ou de loin, ont nourri mon chemin.

À mes parents,

Pour tout ce que les mots ne pourront jamais suffire à dire…

Pour vos sacrifices silencieux, votre amour inépuisable, vos regards pleins de fierté même quand je doutais. Ce mémoire est autant le vôtre que le mien.

À ma sœur Yassmine,

Ma confidente, mon soutien discret mais constant,

Merci d'avoir toujours été là, dans l'ombre ou à mes côtés,

à m'encourager, à m'apaiser, à me rappeler qui je suis.

À mes amis,

Merci pour vos éclats de rire dans les moments de fatigue,

pour vos présences rassurantes, pour vos silences partagés.

Et à toi, qui seras là le jour de ma soutenance :

ta présence compte plus que tu ne l'imagines.

***À ma promotion en Bio-informatique*,**

Chaque visage, chaque moment, chaque débat, chaque entraide,

a construit un chapitre unique de ce voyage.

Merci pour cette aventure humaine et académique.

À Monsieur Chehili Hamza,

Votre accompagnement a été une lumière.

Merci pour votre bienveillance, vos conseils, et votre disponibilité sincère, Je ne pourrai jamais être à la hauteur de ce que vous méritez.

À Madame Hamida et Madame Habiba,

Votre rôle dans mon parcours va bien au-delà de l'enseignement.

Vous avez éveillé en moi un amour sincère pour la bio-informatique,

et vous avez touché ma vie personnelle et professionnelle avec humanité, passion et douceur.

Je vous serai toujours reconnaissante.

À mon binôme Nour,

Merci pour ta patience, ta sincérité, ton implication.

Nous avons traversé ce projet ensemble, entre doutes, idées, rires et nuits tardives.

À Monsieur Kamel Kello et Monsieur Mohammed Laarbi Medjroubi,

Sans vous, je n'aurais jamais pu étudier la bioinformatique.

Merci pour votre foi en moi, pour m'avoir ouvert cette voie,

pour votre soutien discret mais fondamental.

Ce mémoire porte aussi la trace de votre générosité.

À moi-même,

Pour avoir tenu bon malgré les tempêtes, Pour avoir cru, même dans les silences,

Pour avoir transformé les doutes en pas, et les nuits blanches en chapitres.

À ceux qui ont cru en moi, m'ont portée, inspirée et accompagnée… merci.

Asma

Dédicace

Derrière chaque page de ce mémoire se cache une multitude de visages, de voix, de gestes et de silences qui ont accompagné ce chemin.
Chaque mot posé ici est le fruit d'un parcours partagé, d'un soutien reçu, d'un amour donné.

À mes parents,

Vous êtes mes racines et mon ciel, dans vos silences, j'ai trouvé la force, dans vos sacrifices, une leçon d'amour, dans vos prières discrètes, un élan vers l'avenir. Ce mémoire est né de votre amour inépuisable, et il vous appartient autant qu'à moi.

À vous, Amine et Mouatez,

Frères de sang et de cœur, merci pour vos encouragements simples mais puissants, pour vos regards complices, vos mots parfois rares mais toujours justes ,et pour cette affection fraternelle qui, même à distance, m'a donné de l'élan, votre présence, même silencieuse, a compté bien plus que vous ne l'imaginez.

À mes cousines, qui sont bien plus que des cousines des sœurs choisies par le cœur,

Merci pour votre tendresse, vos rires partagés, votre présence constante, dans les jours de fatigue comme dans les instants de joie, vous avez été là, avec cette douceur familière qui apaise, cette complicité rare qui traverse les années et les épreuves

À ma promotion en Bio-informatique,

Chaque visage, chaque collaboration, chaque discussion animée a enrichi cette aventure.
Merci pour cette traversée collective, à la fois humaine et intellectuelle.

À Monsieur Chehili Hamza,

Votre accompagnement a été un phare dans ce parcours.
Merci pour votre écoute, votre bienveillance, votre rigueur et votre confiance.

À Madame Hamida et Madame Habiba,

Votre enseignement m'a ouvert les yeux,votre passion a éveillé la mienne,et votre humanité m'a profondément touchée.
Merci pour tout ce que vous avez transmis, bien au-delà des savoirs.

À mon binôme, Asma,

Merci pour ton soutien constant, ta bienveillance et ton engagement.

À toi dont le nom n'est peut-être pas écrit ici,

mais dont la présence continue d'éclairer mon chemin.
Merci pour ta constance, ton écoute, ta douceur.

Nour El Houda

RÉSUMÉ

L'analyse des données de séquençage d'ARN à cellule unique (scRNA-seq) issues d'embryons précoces est essentielle pour comprendre les mécanismes complexes de la différenciation cellulaire initiale et pour évaluer la qualité embryonnaire. Cependant, la complexité inhérente de ces données caractérisées par leur haute dimensionnalité, leur parcimonie et la présence de variations techniques comme les effets de lot constitue un véritable défi informatique.

Ce travail propose un pipeline d'analyse intégré et complet conçu pour surmonter ces obstacles. Il repose sur une phase de prétraitement rigoureuse, suivie de l'application du modèle probabiliste profond scVI, qui permet à la fois de corriger les effets de lot et de générer une représentation latente informative et de faible dimension des données. Ces données affinées sont ensuite exploitées par des algorithmes d'apprentissage automatique ensembliste tels que Random Forest, XGBoost et CatBoost, pour classifier précisément les principaux lignages embryonnaires : la masse cellulaire interne (ICM) et le trophoectoderme (TE).

Une particularité notable du pipeline est son module d'évaluation de la qualité embryonnaire, fondé sur la variabilité transcriptomique intra-lignage et renforcé par des techniques statistiques de Bootstrap, garantissant la robustesse des scores obtenus.

Les résultats incluent une précision de classification supérieure à 0.96, l'identification fiable de marqueurs spécifiques aux lignages, et un cadre solide pour l'évaluation de la qualité embryonnaire. Cette approche intégrée et interprétable constitue un outil prometteur pour la recherche fondamentale et les applications cliniques, notamment pour améliorer le succès de la fécondation in vitro (FIV).

Mots-clés : scRNA-seq, scVI, Random Forest, CatBoost, XGBoost, ICM, TE, qualité embryonnaire.

Abstract

The analysis of single-cell RNA sequencing (scRNA-seq) data from early embryos is critical for elucidating the intricate mechanisms of initial cell differentiation and for assessing embryonic quality. However, the inherent complexity of these datasets, characterized by high dimensionality, sparsity, and significant technical variations such as batch effects, presents substantial computational challenges. This report introduces a comprehensive and integrated analysis pipeline specifically designed to overcome these obstacles. The proposed approach meticulously combines rigorous data preprocessing steps with the application of the deep variational autoencoder scVI. scVI serves a dual purpose: it effectively corrects for batch effects and generates a highly informative, low-dimensional latent space representation of the data. This refined data then serves as input for a suite of ensemble machine learning methods, including Random Forest, XGBoost, and CatBoost, enabling accurate cell classification of crucial early embryonic lineages, namely the inner cell mass (ICM) and the trophectoderm (TE). A distinctive feature of this pipeline is the integration of a quantitative embryonic quality assessment module. This module is founded on the principle of transcriptomic variability with in lineages and is statistically reinforced through the application of bootstrap techniques, ensuring robust and reliable quality scores.

The anticipated outcomes of this pipeline are demonstrably high, including an expected accuracy exceeding 0.96 in ICM/TE classification, reliable identification of lineage-specific marker genes, and a robust framework for embryonic quality assessment. This integrated pipeline offers a powerful and interpretable solution, holding significant potential for both fundamental research into embryonic development and practical clinical applications, particularly in enhancing the success rates of in vitro fertilization (IVF) procedures.

Index Terms: Machine learning, CatBoost, Embryonic development, Inner Cell Mass (ICM), Random Forest, scRNA-seq, scVI, Trophectoderm(TE), XGBoost

الملخص

يُعد تحليل بيانات تسلسل الحمض النووي الريبي على مستوى الخلية الواحدة (scRNA-seq) المستخلصة من الأجنة في مراحلها المبكرة أداةً مهمة لفهم الآليات الأولى لتمايز الخلايا وتقييم جودة الأجنة. غير أن هذه البيانات تتميز بتعقيدها العالي من حيث الأبعاد، وتشتتها، والتأثيرات التقنية مثل تأثير الدفعات، مما يفرض تحديات تحليلية كبيرة.

في هذا العمل، نقترح سلسلة تحليلية متكاملة تهدف إلى تجاوز هذه الصعوبات، من خلال مراحل معالجة مسبقة دقيقة، تليها استخدام النموذج الاحتمالي العميق **scVI** الذي يسمح بتقليل الأبعاد وتصحيح التباينات التقنية، وتوفير تمثيل كامن دقيق للبيانات.

ثم يتم استخدام هذا التمثيل في تصنيف السلالات الخلوية الجنينية الأساسية، وهي **الكتلة الخلوية الداخلية (ICM)** و**التروفكتوديرم(TE)** ، باستخدام خوارزميات تعلم آلي قوية مثل **Random Forest**، **XGBoost** و **CatBoost**. كما يتضمن النظام آلية لتقييم جودة الأجنة، تعتمد على تحليل التباين داخل كل سلالة خلوية، مدعّمة بإحصائيات Bootstrap لضمان دقة النتائج.

أظهرت النتائج دقة تصنيف تفوق 96%، وتحديد واسمات جينية مميزة لكل سلالة، إضافة إلى إطار موثوق لتقييم جودة الأجنة. ويُعتبر هذا العمل أداة واعدة في البحث الأكاديمي والتطبيقات السريرية، خاصة في مجال تحسين نسب نجاح التلقيح الاصطناعي(IVF) .

الكلمات المفتاحية:

scRNA-seq، scVI، التعلم الآلي، Random Forest، XGBoost، CatBoost، الكتلة الخلوية الداخلية، التروفكتوديرم، جودة الأجنة.

Liste des Figures

Liste des Tableaux

Liste des acronymes

- **IA :** Intelligence Artificielle
- **AA :** Apprentissage Automatique (Machine Learning)
- **AP :** Apprentissage Profond (Deep Learning)
- **RNA :** Réseau de Neurones Artificiels
- **RF :** Forêt Aléatoire (Random Forest)
- **XGBoost :** eXtreme Gradient Boosting
- **CatBoost :** Categorical Boosting (Boosting optimisé pour les variables catégorielles)
- **SHAP :** SHapley Additive exPlanations
- **scVI :** Single-cell Variational Inference
- **scRNA-seq :** Séquençage de l'ARN à cellule unique
- **VAE :** Autoencodeur Variationnel
- **PCA :** Analyse en Composantes Principales
- **UMAP :** Projection Uniforme Approximative

Table des matières

Intoduction Générale

Introduction générale

Aujourd'hui, l'infertilité touche environ 1 couple sur 6 dans le monde, soit 17,5 % des adultes, selon un rapport récent de l'Organisation mondiale de la santé[1].

Cette statistique met en évidence l'importance d'approfondir notre compréhension des premières étapes du développement embryonnaire, dans le but de mieux accompagner les couples concernés et d'améliorer les taux de succès des traitements de fertilité.

Juste après la fécondation, et avant l'implantation dans l'utérus, l'embryon humain traverse une phase critique de développement. C'est au cours de cette période que les cellules embryonnaires subissent des divisions rapides et se répartissent en deux lignages distincts : la masse cellulaire interne (ICM), à l'origine du futur embryon, et le trophoectoderme (TE), qui donnera naissance au placenta et à d'autres structures extra-embryonnaires essentielles[2].

Comprendre les mécanismes qui sous-tendent la formation de ces deux lignages est fondamental pour interpréter certains échecs ou réussites en fertilité humaine, mais aussi pour faire progresser les techniques de procréation médicalement assistée (PMA), notamment l'évaluation des embryons avant l'implantation [3],[4].

L'essor des technologies comme le séquençage d'ARN à cellule unique (scRNA-seq) a révolutionné ce domaine. Cette approche permet de mesurer l'expression des gènes cellule par cellule, offrant ainsi une vision détaillée des dynamiques de différenciation embryonnaire [5]. Toutefois, les données produites par scRNA-seq sont souvent complexes, bruitées et hétérogènes, ce qui rend leur exploitation difficile avec les méthodes d'analyse traditionnelles. Dans ce contexte, l'utilisation de ces données pour prédire la qualité embryonnaire (ICM, TE) ou identifier des biomarqueurs du succès d'implantation s'avère particulièrement prometteuse, notamment pour améliorer la prise en charge clinique des patients dans le cadre de la FIV.

Dans le cadre de ce travail, nous proposons un pipeline d'analyse basé sur des modèles d'intelligence artificielle, appliqué à des données d'expression scRNA-seq issues d'embryons humains au stade préimplantatoire. Notre objectif est triple : classer les cellules embryonnaires selon leur lignage (ICM ou TE), identifier des gènes clés propres à chaque type cellulaire.

Pour cela, nous utilisons le modèle scVI (single-cell Variational Inference), un modèle d'apprentissage profond probabiliste fondé sur des autoencodeurs variationnels et des

techniques bayésiennes. Ce modèle permet d'intégrer l'information de plusieurs embryons, de réduire la dimensionnalité des données, de classifier les cellules selon leur type[6,7].

Nous avons ensuite combiné ces résultats avec ceux obtenus via d'autres méthodes d'apprentissage automatique, notamment : CatBoost, XGBoost et Random Forest, qui permettent à la fois de classer les cellules et d'identifier les gènes les plus discriminants[7,8], [9].

Conscients du caractère souvent perçu comme "boîte noire" des modèles de deep learning, nous avons intégré dans notre pipeline une phase d'explicabilité fondée sur l'outil SHAP (SHapley Additive exPlanations). SHAP permet de quantifier l'impact de chaque gène sur la classification prédite par les modèles, renforçant ainsi la confiance et la compréhension biologique des résultats[10].

Plan du mémoire :

- Chapitre 1 : Le développement embryonnaire humain – entre biologie naturelle et FIV
- Chapitre 2 : Approches d'intelligence artificielle (ML, DL, SHAP) pour étudier le développement embryonnaire
- Chapitre 3 : Matériel et méthode.
- Chapitre 4 : Résultats et discussions.

Chapitre 1 : Développement embryonnaire humain entre biologie naturelle et FIV

1. Introduction

Le développement embryonnaire humain est un processus biologique complexe et fondamental, débutant à partir d'une cellule unique le zygote pour aboutir à un organisme complet formé de plusieurs milliards de cellules différenciées. Il implique des mécanismes cellulaires, moléculaires et morphologiques finement régulés, qui se déroulent en plusieurs phases successives : gamétogenèse, fécondation, segmentation, implantation et développement fœtal. Comprendre ce processus est essentiel non seulement pour la biologie du développement, mais également pour les pratiques médicales modernes telles que la fécondation in vitro (FIV), la sélection embryonnaire, et les applications en médecine régénérative[11],[12].

Ce chapitre propose une description détaillée des principales étapes du développement embryonnaire, tout en les mettant en parallèle avec les techniques utilisées en FIV. Cette double approche permet de mieux saisir les enjeux scientifiques, cliniques, technologiques et éthiques liées à la reproduction humaine[13],[14].

2. Phases du développement embryonnaire humain

Le développement embryonnaire humain se déroule en quatre grandes phases : la fécondation, la segmentation, l'implantation et la gastrulation. Toutefois, dans le cadre de notre étude axée sur la procréation médicalement assistée (PMA), et plus précisément sur la fécondation in vitro (FIV), notre intérêt se concentre principalement sur les trois premières étapes. Ce sont en effet la fécondation, la segmentation (jusqu'au stade blastocyste), et l'implantation qui sont essentielles à évaluer pour optimiser les chances de succès de la FIV, la gastrulation se déroulant après l'implantation, donc hors du champ d'observation habituel en laboratoire.

2.1. Période pré-embryonnaire

2.1.1. Gamétogenèse Naturelle

La gamétogenèse est le processus de formation des cellules reproductrices, les gamètes, indispensables à la reproduction sexuée. Chez l'humain, elle se divise en deux processus distincts : la spermatogenèse chez l'homme et l'ovogenèse chez la femme.

Spermatogenèse

La spermatogenèse se déroule dans les tubes séminifères des testicules. Elle débute à la puberté et se poursuit tout au long de la vie adulte. Ce processus comprend trois phases principales :

- **Multiplication** : les spermatogonies (cellules germinales diploïdes) se multiplient par mitose.
- **Méiose** : les spermatogonies entrent en méiose, réduisant le nombre de chromosomes de diploïde (2n) à haploïde (n), formant les spermatides.
- **Spermiogenèse** : les spermatides se différencient en spermatozoïdes matures, avec condensation du noyau, formation de l'acrosome, développement du flagelle.

La spermatogenèse est régulée par les hormones hypophysaires, notamment la FSH (hormone folliculo-stimulante) qui stimule les cellules de Sertoli, et la LH (hormone lutéinisante) qui stimule la production de testostérone par les cellules de Leydig[15].

Ovogenèse

L'ovogenèse a lieu dans les ovaires et commence dès la vie fœtale. Les ovogonies se multiplient par mitose avant d'entrer en méiose, mais restent bloquées en prophase I jusqu'à la puberté. À chaque cycle menstruel, un ovocyte reprend la méiose et est libéré lors de l'ovulation. L'ovogenèse est régulée par les hormones ovariennes, principalement les œstrogènes et la progestérone, qui contrôlent le développement folliculaire et la maturation de l'ovocyte.

La qualité des gamètes est un facteur déterminant pour la réussite de la fécondation et du développement embryonnaire. Des facteurs environnementaux, génétiques, ainsi que l'âge, peuvent influencer cette qualité[16].

2.1.2. Gamétogenèse assistée (en FIV)

En FIV, on cherche à contrôler et à optimiser la production des gamètes, c'est-à-dire des ovocytes chez la femme et des spermatozoïdes chez l'homme.

Chez la femme:

- On administre des hormones appelées gonadotrophines (comme la FSH ou l'hMG) pour stimuler les ovaires.
- Cela permet de faire mûrir plusieurs follicules au lieu d'un seul, comme dans un cycle naturel.

- Quand les follicules atteignent une taille suffisante, une injection déclenche l'ovulation.
- Ensuite, les ovocytes sont prélevés par ponction à l'aide d'une aiguille guidée par échographie[15],[16].

Chez l'homme:

- Le sperme est recueilli le jour de la ponction.
- Il est ensuite préparé en laboratoire : les spermatozoïdes sont séparés du liquide séminal et seuls les plus mobiles et morphologiquement normaux sont sélectionnés.

Ce processus permet:

- D'augmenter les chances de fécondation,
- De disposer de plusieurs embryons à analyser ou congeler,
- Et de mieux contrôler la qualité des gamètes pour obtenir les meilleures conditions de développement embryonnaire[15],[16].

2.2. Fécondation

2.2.1. Fécondation naturelle

La fécondation est l'étape clé qui marque le début du développement embryonnaire. Elle correspond à la fusion d'un spermatozoïde et d'un ovocyte pour former une cellule unique, le zygote, contenant un génome complet diploïde[14],[16].

Mécanismes cellulaires

La fécondation se déroule généralement dans le tiers externe de la trompe de Fallope, peu après l'ovulation. Le spermatozoïde subit une série de modifications appelées capacitation, qui lui permettent de traverser la zone pellucide entourant l'ovocyte. La réaction acrosomique libère des enzymes facilitant cette traversée.

La reconnaissance entre le spermatozoïde et l'ovocyte est médiée par des protéines spécifiques, notamment le récepteur ZP3 de la zone pellucide. Après la fusion des membranes plasmatiques, le spermatozoïde pénètre dans l'ovocyte, déclenchant la réaction corticale qui empêche la polyspermie (entrée de plusieurs spermatozoïdes).

Activation du zygote

La fusion des noyaux mâle et femelle forme le pronoyau diploïde. Le zygote reprend alors son cycle cellulaire, activant son propre génome embryonnaire, ce qui marque le début de la transcription des gènes nécessaires au développement.

2.2.2. Fécondation in vitro (FIV)

La fécondation in vitro (FIV) est une technique de procréation médicalement assistée (PMA) qui permet la fécondation des ovocytes en dehors du corps de la femme, dans un laboratoire. Deux méthodes principales sont utilisées:

- FIV classique : les ovocytes recueillis sont placés dans une boîte de culture contenant un liquide nutritif, avec environ 100 000 spermatozoïdes mobiles. La fécondation se produit alors naturellement, comme elle le ferait dans la trompe utérine.
- ICSI (Injection intra-cytoplasmique de spermatozoïde) : un seul spermatozoïde est sélectionné et injecté directement dans le cytoplasme de l'ovocyte à l'aide d'un micromanipulateur. Cette méthode est utilisée lorsque le nombre de spermatozoïdes est faible ou leur mobilité réduite.

Après la fécondation (confirmée par la présence des deux pronoyaux), le zygote est placé dans un incubateur à température et atmosphère contrôlées (37°C, 5% CO_2). L'embryon est surveillé quotidiennement pour suivre ses divisions cellulaires :

- J2 à J3 : division en 2, puis 4 à 8 cellules (blastomères)
- J4 : stade morula (compaction cellulaire)
- J5 à J6 : formation du blastocyste avec différenciation en ICM (masse cellulaire interne) et TE (trophectoderme)

Les embryons les plus viables peuvent être transférés dans l'utérus ou congelés pour une utilisation ultérieure[17].

2.3. Segmentation (clivage)

2.3.1. Segmentation naturelle

Après la fécondation, le zygote subit une série de divisions cellulaires rapides appelées

segmentation ou clivage. Ces divisions sont caractérisées par l'absence de croissance cellulaire, ce qui conduit à une augmentation du nombre de cellules (blastomères) tout en maintenant la taille globale de l'embryon constante.

Chez l'humain, la segmentation est holoblastique complète, c'est-à-dire que la cellule se divise entièrement. Les blastomères deviennent progressivement plus petits et commencent à se compacter grâce à des molécules d'adhérence comme les cadhérines, formant une structure compacte appelée morula.

Vers le 5e jour, la morula se creuse pour former une cavité, le blastocoele, donnant naissance au blastocyste. Le blastocyste se compose alors de deux populations cellulaires distinctes : le trophoblaste, qui formera le placenta, et la masse cellulaire interne, qui donnera l'embryon proprement dit.

Le blastocyste est un stade clé du développement préimplantatoire, constitué d'une couche externe de cellules (le trophoblaste) et d'une masse cellulaire interne (ICM). Ce stade marque la première différenciation cellulaire majeure de l'embryon humain[14], [16].

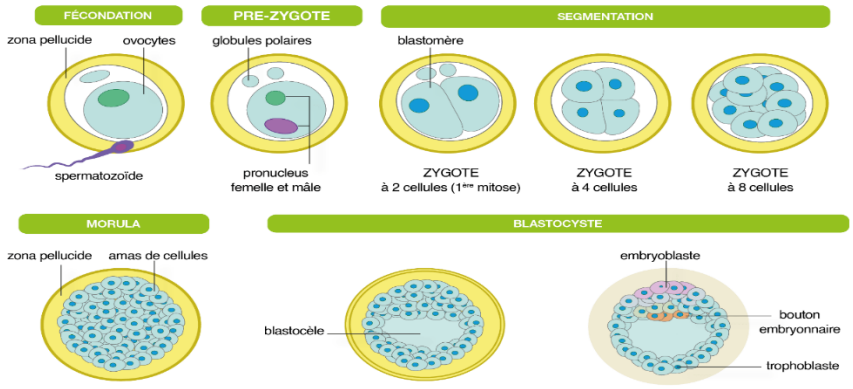

Figure 1:Processus du développement embryonnaire humain jusqu'au stade blastocystaire.

2.3.2. Segmentation En FIV

les embryons sont soigneusement surveillés tout au long de leur développement, soit par des observations classiques au microscope, soit à l'aide de systèmes d'imagerie time-lapse.

L'observation au microscope permet une évaluation ponctuelle, une à deux fois par jour, pour compter le nombre de cellules (blastomères), évaluer la symétrie et détecter les fragments cytoplasmiques.

L'imagerie time-lapse, quant à elle, enregistre automatiquement des images toutes les 10 à 20

minutes sans perturber l'environnement de culture. Cela offre plusieurs avantages :

- Suivi continu et non invasif de chaque embryon,
- Détection de moments clés comme le début et la fin des divisions cellulaires,
- Calcul de paramètres morphocinétiques (durée des cycles cellulaires, synchronicité des divisions, apparition du blastocyste).

Ces informations sont utilisées pour attribuer un score à chaque embryon. Les embryons qui montrent une division régulière, rapide mais équilibrée, avec peu de fragmentation, sont considérés comme ayant un potentiel élevé d'implantation.

Cette approche améliore la sélection embryonnaire et peut être combinée à l'intelligence artificielle pour une évaluation encore plus fiable[18].

Table 1:différentes étapes de la segmentation de l'embryon humain depuis le zygote jusqu'au blastocyste (J1 à J6).

Jour post fecondation	**Stade embryonnaire**	**Nombrede cellules**	**Remarques principales**
J1	Zygote	1	Cellule unique diploïde
J2	Clivage	2-4	Première division mitotique
J3	Clivage	8	Blastomères visibles, divisions rapides
J4	Morula	16-32	Compaction cellulaire
J5-J6	Blastocyste	64+	Différenciation en ICM et TE

2.4. Migration et implantation

2.4.1. Implantation naturelle :

Le blastocyste migre le long de la trompe de Fallope vers la cavité utérine. Cette migration dure environ 3 à 4 jours. Pendant ce temps, l'endomètre utérin se prépare à accueillir l'embryon

grâce à des modifications hormonales qui rendent la muqueuse réceptive.

L'implantation commence vers le 6e jour après la fécondation. Le trophoblaste se différencie en deux couches :le cytotrophoblaste, constitué de cellules individuelles, et le syncytiotrophoblaste, une couche multinucléée qui envahit l'endomètre.

Cette invasion permet l'ancrage du blastocyste et le début de la formation du placenta.

L'implantation est un processus délicat,et tout dysfonctionnement peut entraîner des complications telles que les fausses couches précoces ou les grossesses extra-utérines.

2.4.2. Implantation en FIV

Le transfert embryonnaire est généralement effectué au stade de blastocyste (J5), moment où l'embryon est le plus apte à s'implanter dans l'utérus. Ce transfert est réalisé à l'aide d'un cathéter fin et souple, sous contrôle échographique, pour assurer une dépose précise dans la cavité utérine [14],[16].

La fenêtre d'implantation correspond à la période durant laquelle l'endomètre est réceptif à l'embryon, généralement entre le 19e et le 21e jour du cycle menstruel. Pour maximiser les chances de succès, plusieurs facteurs sont évalués avant le transfert :

- **Épaisseur endométriale** : un endomètre de plus de 7 mm est considéré comme favorable.
- **Structure trilaminaire** : aspect en « triple ligne » visible à l'échographie, signe d'une bonne maturation utérine
- **Absence de contractions utérines** : les mouvements utérins excessifs peuvent compromettre l'adhésion de l'embryon [14].

Des médicaments (progestérone, œstrogènes) peuvent être administrés pour optimiser la préparation de l'endomètre, surtout dans les cycles artificiels. Le respect de la synchronisation entre l'âge de l'embryon et la maturation endométriale est fondamental pour une implantation réussie.

3. Évaluation morphologique des embryons (en FIV)

L'évaluation morphologique constitue une étape cruciale dans le cadre de la fécondation in vitro (FIV). Elle permet de déterminer le potentiel d'implantation et de développement des embryons

cultivés en laboratoire. Deux structures principales du blastocyste sont systématiquement analysées : la masse cellulaire interne (ICM) et le trophectoderme (TE).

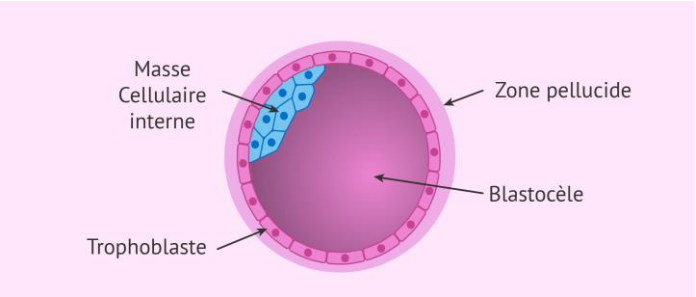

Figure 2:Structure du blastocyste humain – disposition des lignages cellulaires

3.1. Masse Cellulaire Interne (ICM)

L'ICM est constituée de cellules pluripotentes qui donneront naissance à l'embryon proprement dit. Sa qualité est fortement corrélée à la probabilité d'implantation et de grossesse.

- Grade A : Masse cellulaire dense, bien définie, avec de nombreuses cellules compactées. C'est le signe d'un embryon de très bonne qualité.
- Grade B : Masse modérément définie, avec quelques cellules lâches ou irrégulièrement organisées.Potentiel intermédiaire.
- Grade C : Masse cellulaire peu visible, contenant très peu de cellules ou mal organisées. Associée à un potentiel développemental réduit.

L'évaluation de l'ICM est donc essentielle pour prédire les chances de succès et orienter le choix des embryons à transférer ou à congeler[18].

Table 2:Classification morphologique de la masse cellulaire interne (ICM) selon le grade.

Grade	**Description**
A	Nombreuses cellules, bien compactées
B	Quelques cellules lâches
C	Très peu de cellules, structure faible

3.2. Trophectoderme (TE)

Le TE est formé des cellules périphériques du blastocyste qui seront à l'origine du placenta et des membranes extraembryonnaires. Sa structure est également un indicateur clé de la viabilité embryonnaire.

- Grade A : Cellules nombreuses, bien organisées, formant une couche uniforme.
- Grade B : Quelques irrégularités dans l'organisation cellulaire.
- Grade C : Peu de cellules, désorganisation marquée, aspect fragile.

L'analyse conjointe du TE et de l'ICM, selon des grilles de grading standardisées comme celle du consensus d'Istanbul, permet une évaluation rigoureuse et objective de la qualité embryonnaire[2],[18].

Table 3:Classification morphologique du trophoectoderme (TE) selon le grade.

Grade	Description
A	Cellules nombreuses, bien organisées
B	Quelques irrégularités
C	Peu de cellules, structure fragile

4. Le scRNA-seq pour analyser les cellules embryonnaires

4.1. Principe

Le séquençage à cellule unique (scRNA-seq) permet d'analyser l'expression des gènes à l'échelle d'une cellule individuelle. Contrairement aux méthodes classiques qui diluent les informations entre plusieurs cellules, le scRNA-seq révèle l'hétérogénéité cellulaire en capturant les ARN messagers de chaque cellule. Les étapes incluent l'isolement des cellules, la lyse, la rétrotranscription en ADNc, l'amplification, le séquençage, puis l'analyse bio-informatique[19].

4.2. Applications en embryologie

Grâce au scRNA-seq, les chercheurs ont pu identifier des sous-types cellulaires précoces au sein de l'embryon humain, comme les cellules de l'épiblaste, de l'hypoblaste et du trophectoderme. Cette technique permet aussi d'étudier les mécanismes de différenciation et les

premières décisions du destin cellulaire. Elle est également utilisée pour comparer les embryons normaux et anormaux, et détecter des anomalies transcriptionnelles précoces[20],[21].

4.3. Intérêt pour la FIV

Le scRNA-seq pourrait révolutionner la sélection embryonnaire en FIV :

- Identifier les embryons ayant les meilleurs profils d'expression.
- Détecter les marqueurs d'implantation ou de mauvaise différenciation.
- Étudier les effets des milieux de culture et conditions de laboratoire sur l'expression génique embryonnaire.

Ces données peuvent être utilisées pour construire des modèles prédictifs, notamment à l'aide d'algorithmes d'intelligence artificielle, afin de sélectionner les embryons avec le plus grand potentiel d'implantation[22].

4.4. Limites de la méthode

Malgré ses avantages, le scRNA-seq présente plusieurs défis :

- Le coût élevé des analyses.
- La complexité des manipulations techniques.
- La fragilité de l'ARN qui nécessite un traitement rapide et soigné.
- L'analyse bioinformatique demande des compétences avancées et des outils spécialisés[22].

5. Enjeux éthiques et applications cliniques

5.1. Éthique

La question de l'éthique autour du développement embryonnaire humain, en particulier dans le cadre de la FIV, suscite des débats importants dans les domaines scientifique, juridique et religieux [23].

- **Statut juridique et moral de l'embryon** : Dans de nombreux pays, l'embryon est reconnu comme une entité biologique distincte, sans être juridiquement assimilé à une personne humaine. Ce statut soulève des interrogations sur les limites de son utilisation en recherche ou pour des finalités médicales.
- **Limite de culture fixée à 14 jours** : En Europe et en France, il est interdit de cultiver un embryon humain au-delà de 14 jours après la fécondation. Cette limite correspond à

l'apparition de la ligne primitive, marquant le début de la gastrulation, moment clé dans l'individualisation de l'organisme.

- **Recherche strictement encadrée** : La recherche sur les embryons est soumise à des autorisations spécifiques délivrées par des comités d'éthique et encadrée par des lois bioéthiques. Elle doit répondre à des critères précis : but médical, absence d'alternatives, consentement éclairé.

5.2. Applications cliniques

La recherche sur les cellules embryonnaires et les technologies associées offre plusieurs applications cliniques majeures :

- **DPI (Diagnostic Préimplantatoire)** : Cette technique permet de détecter des anomalies chromosomiques ou génétiques avant l'implantation, et de ne transférer que les embryons sains. Elle est particulièrement utile pour les couples à risque génétique connu.
- **Médecine régénérative** : Les cellules souches embryonnaires peuvent se différencier en n'importe quel type cellulaire. Elles sont utilisées dans des essais cliniques pour traiter des maladies comme le diabète, la maladie de Parkinson, ou certaines pathologies de la rétine.
- **Modèles in vitro** : Les embryons ou structures dérivées comme les organoïdes embryonnaires sont utilisés pour tester des médicaments, étudier le développement humain normal ou pathologique, et modéliser des maladies génétiques.

Ces avancées doivent s'accompagner d'un cadre réglementaire solide pour garantir une pratique éthique et socialement responsable de la recherche biomédicale[23],[24].

10. Conclusion

Le développement embryonnaire humain, étudié depuis des siècles, entre aujourd'hui dans une ère technologique de haute précision. La combinaison des outils morphologiques, transcriptomiques et algorithmiques permet une meilleure prise en charge des troubles de la fertilité. L'Algérie, bien que dynamique dans ce domaine, doit encore relever des défis d'accessibilité, d'équité et d'innovation. Une structuration nationale ambitieuse et éthique est

indispensable pour garantir un accès juste et performant à la procréation médicalement assistée [2],[25].

Chapitre 2 : Techniques d'Apprentissage Automatique et Interprétabilité

1. Introduction

Les technologies de séquençage à cellule unique (single-cell RNA-seq) ont transformé la biologie moderne. Elles permettent de mesurer l'expression génique au niveau de chaque cellule, révélant ainsi la grande diversité et l'hétérogénéité cellulaire au sein des tissus. Cette granularité est essentielle pour comprendre les mécanismes biologiques complexes, comme le développement embryonnaire, la différenciation cellulaire, ou encore les réponses immunitaires [22].

Cependant, ces données présentent plusieurs défis majeurs :

- Grande dimensionnalité : chaque cellule est représentée par l'expression de milliers de gènes.
- Sparsité : un grand nombre de gènes sont notés à zéro dans une cellule donnée, souvent à cause de limitations techniques appelées « dropouts »[13].
- Bruit technique : variations dues aux conditions expérimentales, lots de traitement, plateformes différentes, etc.

Ces caractéristiques rendent l'analyse classique difficile. Il est donc indispensable d'employer des approches sophistiquées capables de réduire la dimension, d'identifier des groupes cellulaires, et de modéliser la dynamique cellulaire tout en corrigeant les biais techniques. Les méthodes d'intelligence artificielle (IA) et de machine learning offrent justement ces outils[16].

Les progrès récents en biologie moléculaire, notamment le séquençage unicellulaire, génèrent des volumes massifs de données complexes. Ces données sont souvent bruyantes, de très haute dimension, et difficiles à interpréter avec des méthodes classiques. Pour répondre à ces défis, les chercheurs utilisent des approches issues de l'intelligence artificielle (IA) et de l'apprentissage automatique (machine learning). Ces méthodes permettent de découvrir des structures cachées dans les données et de classer les types cellulaires[16].

Dans ce chapitre, nous présentons les principaux concepts liés à l'IA et au machine learning, ainsi que des outils spécifiques à l'analyse des données unicellulaires,tels que scVI[16]et Monocle[17].

2. Intelligence Artificielle (IA)

L'intelligence artificielle désigne l'ensemble des techniques permettant aux ordinateurs

d'effectuer des tâches qui nécessiteraient normalement une intelligence humaine. Cela inclut des activités comme la reconnaissance d'images, la compréhension du langage, la prise de décision, ou la résolution de problèmes complexes.

Applications en biologie

En biologie, l'IA est utilisée pour automatiser l'analyse de données massives issues du séquençage génomique, l'imagerie médicale, ou la modélisation de processus biologiques. Par exemple, des systèmes intelligents peuvent détecter automatiquement des anomalies dans des images cellulaires, prédire l'effet de mutations génétiques, ou identifier des sous-populations cellulaires dans des échantillons complexes[22].

3. Machine Learning (Apprentissage Automatique)

Le machine learning est une branche de l'IA qui permet à une machine d'apprendre des modèles à partir de données, sans être explicitement programmée pour chaque tâche [18],[19],[20]

Mécanisme d'apprentissage des modèles

Au lieu d'écrire un programme qui décrit précisément comment traiter un problème, on fournit à un algorithme un grand nombre d'exemples (données d'entraînement). L'algorithme apprend alors une fonction qui relie les entrées aux sorties souhaitées. Cette fonction pourra ensuite être utilisée pour faire des prédictions sur des données nouvelles.

Types d'apprentissage

- **Supervisé** : les données d'entraînement sont étiquetées (chaque exemple est associé à une réponse correcte). Par exemple, on apprend à classer des cellules en types connus à partir de données marquées.
- **Non supervisé** : les données ne sont pas étiquetées. L'algorithme essaie de découvrir des structures ou groupes naturels dans les données, comme regrouper les cellules en clusters.
- **Par renforcement** : l'algorithme apprend par essais et erreurs, en recevant des récompenses ou pénalités, utilisé surtout dans des environnements dynamiques (moins courant en biologie).

4. Deep Learning (Apprentissage profond)

Le deep learning est une méthode de machine learning qui utilise des réseaux de neurones artificiels profonds, capables d'apprendre des représentations très complexes[21].

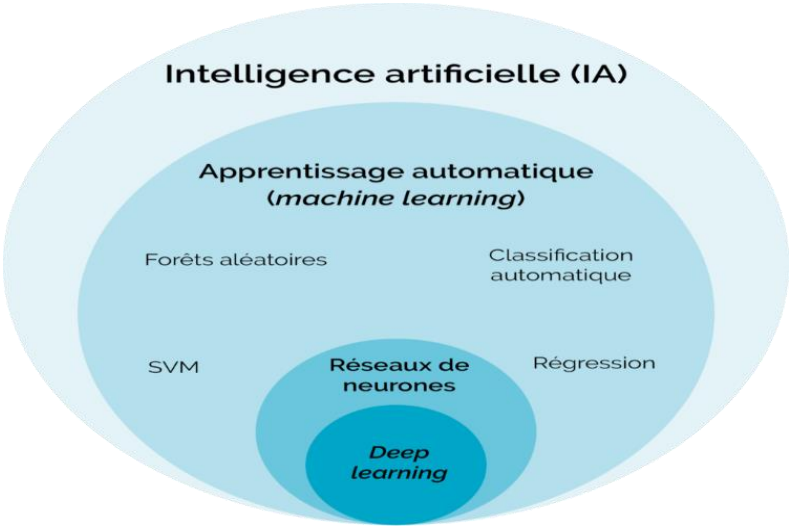

Figure 3:Organisation hiérarchique de l'intelligence artificielle, de l'apprentissage automatique et de l'apprentissage profonde

5. Modèle probabiliste scVI

scVI (single-cell Variational Inference) est un modèle génératif basé sur un autoencodeur variationnel profond, conçu pour modéliser les données de séquençage unicellulaire[16].

Principaux avantages :

- Gère la nature sparse et bruitée des données.
- Corrige les effets techniques (batch effects).
- Produit une représentation latente compacte et débruitée.
- Permet clustering, visualisation, et identification de gènes différentiellement exprimés.

scVI utilise des techniques d'apprentissage profond probabiliste pour capturer la distribution réelle des données, ce qui améliore l'analyse comparative et la découverte biologique[16],[22].

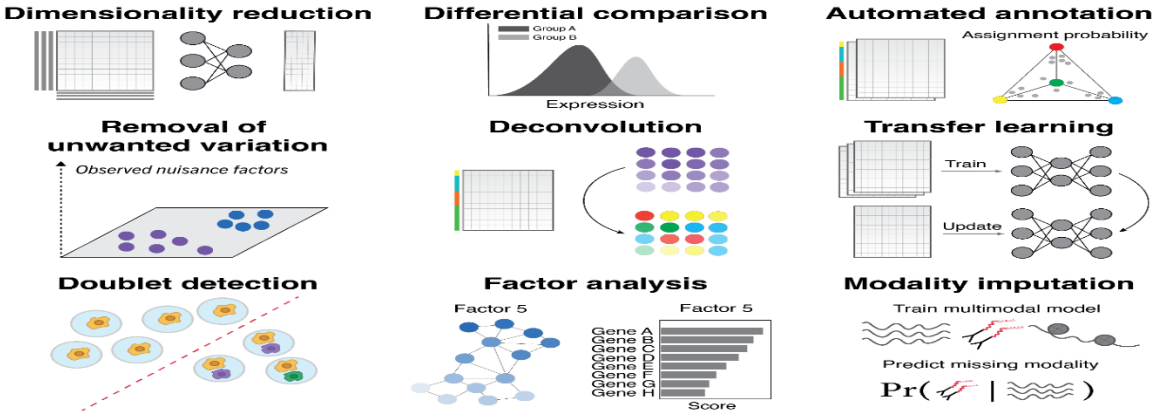

Figure 4: Principales étapes de l'analyse des données transcriptomiques unicellulaires

6. Algorithmes d'ensemble

Ces algorithmes combinent plusieurs modèles simples (arbres de décision) pour améliorer la précision et la robustesse. Nous utilisons dans notre travails les algorithmles suivants :

- **Random Forest** : construit une multitude d'arbres de décision indépendants sur des sous-ensembles aléatoires des données. La prédiction finale est une moyenne ou un vote majoritaire[18].
- **XGBoost** : méthode de boosting qui construit les arbres successifs pour corriger les erreurs des précédents, ce qui conduit à des performances élevées, notamment sur des données tabulaires complexes[19].
- **CatBoost** : améliore le boosting en gérant efficacement les variables catégorielles (par exemple, types de cellules ou conditions expérimentales) sans besoin de pré-traitement lourd[20].

Ces méthodes sont particulièrement adaptées à la classification des types cellulaires et à la sélection de gènes importants.

Table 4:Comparaison entre Random Forest, XGBoost et CatBoost.

Algorithme	**Type**	**Support des variables catégorielles**	**Vitesse d'entraînement**	**Performance**
Random Forest	Bagging	Non supporté	Rapide	Bonne
XGBoost	Boosting	Encodage requis	Plus lent	Très bonne
CatBoost	Boosting	Support natif	Très rapide	Très bonne

7.Interprétabilité des modèles : vers une intelligence artificielle compréhensible

L'interprétabilité désigne la capacité à comprendre et à expliquer les décisions prises par un modèle d'intelligence artificielle. Dans le domaine biomédical, et plus particulièrement en biologie du développement, cette dimension est essentielle : il ne suffit pas qu'un modèle donne une prédiction correcte, encore faut-il en comprendre les raisons sous-jacentes, pour renforcer la confiance scientifique et orienter les validations expérimentales.

Les modèles d'apprentissage automatique les plus puissants comme les forêts aléatoires, les modèles de boosting (XGBoost, CatBoost),ou encore les modèles probabilistes comme scVI (single-cell Variational Inference) sont souvent considérés comme des boîtes noires : ils effectuent des calculs complexes et multidimensionnels, dont le fonctionnement interne est difficilement traçable pour l'utilisateur, bien qu'ils fournissent des performances remarquables et une grande capacité d'abstraction.

7.1. Outils d'explication de modèles

Pour pallier cette opacité, plusieurs outils d'explicabilité ont été développés. Voici les principaux utilisés dans les sciences de la vie :

- **SHAP (SHapley Additive exPlanations) :**

  Basé sur la théorie des jeux, SHAP attribue à chaque variable une contribution précise dans une prédiction, selon le principe de la valeur de Shapley. Il permet de :

  - Comprendre pourquoi un modèle a pris une décision.
  - Identifier les variables les plus influentes (par exemple, les gènes clés dans une classification ICM vs TE).
  - Comparer l'importance des variables entre plusieurs observations (ex. cellules).
  - Valider la pertinence biologique des gènes mis en avant.

  SHAP est l'outil principalement utilisé dans notre pipeline pour relier les prédictions à des interprétations biologiques cohérentes [23].

- **LIME (Local Interpretable Model-agnostic Explanations) :**

  LIME est une autre méthode d'explication localisée qui crée un modèle simple (ex. linéaire) autour de chaque prédiction complexe, pour l'expliquer facilement. Elle est utile pour :

- Comprendre des prédictions individuelles,
- Visualiser l'influence de chaque gène sur une cellule précise.

  Cependant, LIME peut être moins stable que SHAP et plus sensible au bruit.

- **Feature Importance (FI) intégrée aux modèles :**

  Certains algorithmes comme Random Forest ou XGBoost fournissent directement une mesure de l'importance des variables (souvent basée sur le gain d'information ou la fréquence d'utilisation dans les arbres).

  Bien que globale et moins fine, cette approche donne un aperçu utile des gènes les plus influents, mais elle ne permet pas d'expliquer chaque prédiction individuellement.

- **Integrated Gradients (IG) :**

  Spécialement adapté aux réseaux de neurones profonds, ce mécanisme estime l'influence d'une caractéristique en calculant les gradients intégrés entre une entrée de référence (souvent nulle) et l'entrée réelle.

  Utile pour les MLP ou CNN, notamment dans l'analyse de séquences biologiques ou d'images.

7.2. Importance dans notre étude

L'interprétation des modèles est centrale dans notre travail, car elle permet de relier les décisions algorithmiques (classement des cellules, biomarqueurs) à des connaissances biologiques validées, ou d'en proposer de nouvelles.

8.Conclusion

Les approches d'intelligence artificielle, notamment les modèles d'ensemble, les réseaux neuronaux profonds et le modèle probabiliste scVI, offrent des outils puissants pour analyser les données complexes de transcriptomique unicellulaire. Bien qu'efficaces, ces modèles sont souvent perçus comme des "boîtes noires". L'intégration d'outils d'explicabilité comme SHAP permet de mieux comprendre leurs décisions, de valider les résultats biologiques et d'identifier de nouveaux biomarqueurs. Ce couplage entre performance et interprétabilité est essentiel pour des analyses robustes et biologiquement pertinentes.

Chapitre3 :Matériel et méthode

Introduction

Ce chapitre présente les données, les outils logiciels et les protocoles d'analyse mis en œuvre pour concevoir notre pipeline bioinformatique. Il décrit en détail les sources des données scRNA-seq, les étapes de prétraitement, les modèles d'intelligence artificielle utilisés, et les méthodes d'interprétation et de visualisation des résultats.

1.Matériel

1.1. Données

Les données utilisées dans ce travail proviennent de l'étude scientifique de Meistermann et al. , disponible publiquement sur GitLab[24]. Cette base constitue une intégration de plusieurs jeux de données transcriptomiques issus d'embryons humains préimplantatoires, collectés à différents stades du développement.

Elle regroupe les données issues des études suivantes :[20],[21],[26],[27],[28]

Ainsi que des nouvelles données générées par Meistermann et al [24], spécifiquement pour leur étude.

L'objectif était d'intégrer ces sources diverses dans une base de données cohérente et harmonisée, facilitant l'analyse conjointe des profils d'expression génique dans les cellules embryonnaires humaines, en surmontant les biais expérimentaux entre les études.

Format et structure

La base est composée de deux fichiers principaux :

- exprDatRaw.tsv : contient les comptages bruts de l'expression génique (UMI counts) pour chaque cellule.

Chaque ligne représente une cellule, et chaque colonne correspond à un gène.

Les valeurs sont des entiers indiquant combien de fois chaque gène a été détecté dans chaque cellule.

- sampleAnnot.tsv : contient les annotations biologiques et techniques associées à chaque cellule. On y trouve le type cellulaire (ICM, TE, EPI, PE, etc.), le stade embryonnaire, l'origine de l'échantillon, l'étude d'origine, et d'autres métadonnées.

Les deux fichiers sont au format TSV (Tab-Separated Values), particulièrement adapté aux

analyses bioinformatiques car :

- Chaque ligne correspond à une cellule individuelle.
- Chaque colonne à une variable (expression d'un gène ou annotation).
- Les valeurs sont séparées par des tabulations, facilitant l'import dans les outils comme Pandas, Scanpy ou scVI-tools.

Ces données ont été analysées avec un pipeline combinant des outils comme Scanpy, scVI-tools, et des modèles d'apprentissage automatique (Random Forest, XGBoost, CatBoost), décrits dans les sections suivantes.

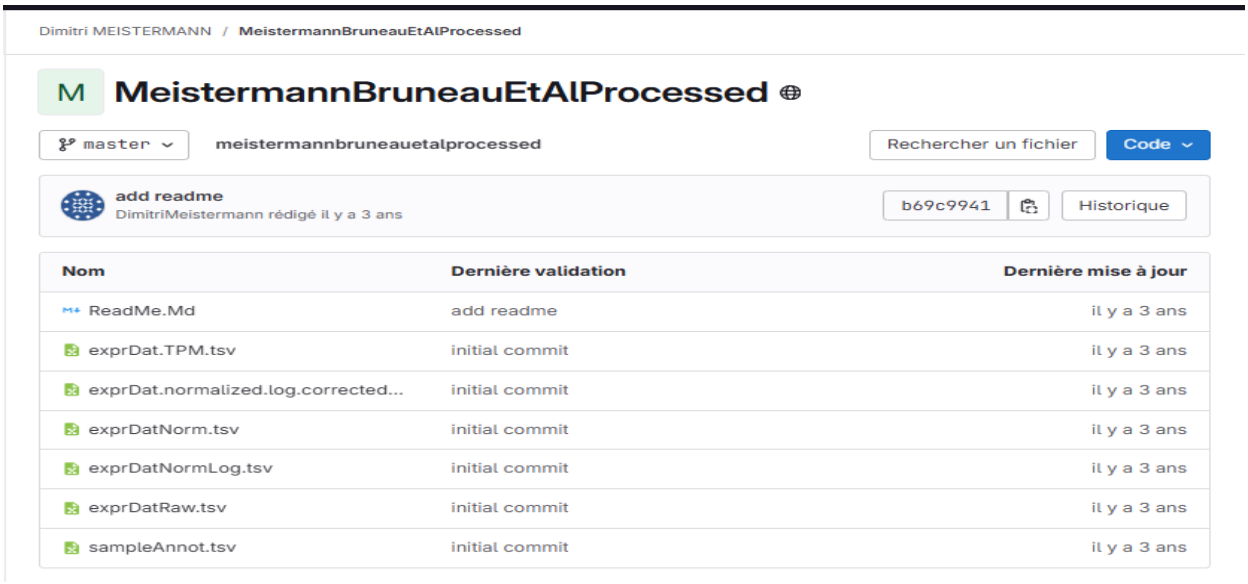

Figure 5:DataSet De MeistermannBruneauEtAl

1.2. Logiciels utilisés

1.2.1. Environnement Python

Pour cette étude, nous avons utilisé le langage Python, qui est gratuit, facile à apprendre et largement utilisé dans le domaine de l'analyse de données biologiques.

Python a été créé en 1991 par Guido van Rossum et est aujourd'hui maintenu par la Python Software Foundation. Sa syntaxe claire et simple permet d'écrire du code de façon concise, ce qui facilite la lecture, l'apprentissage et le débogage.

Python est particulièrement adapté à de nombreux domaines, notamment :

le développement web, l'intelligence artificielle et le machine learning, la science des données, la visualisation de données, et bien sûr, la bio-informatique et l'analyse de données single-cell.

1.2.2. Google Colab

Pour ce projet, l'environnement de travail utilisé est Google Colab (Collaboratory), une plateforme gratuite proposée par Google.

Google Colab permet d'exécuter du code Python directement depuis un navigateur, sans avoir besoin d'installer des logiciels localement. Il offre un accès à des ressources puissantes (comme des GPU) utiles pour traiter des données complexes, notamment en bio-informatique.

1.2.3. Bibliothèques Python utilisées

Dans ce travail, plusieurs bibliothèques Python ont été utilisées pour l'analyse, la modélisation et la visualisation des données. Toutes ces bibliothèques ont été installées et utilisées dans l'environnement Google Colab, ce qui permet une exécution rapide et efficace sur le cloud. Voici une description détaillée de chaque bibliothèque :

- **pandas (version 1.5.3) :**

  Utilisée pour lire, organiser et manipuler les données sous forme de tableaux (dataframes). Elle permet par exemple de filtrer les données, faire des statistiques de base ou fusionner plusieurs tableaux. C'est l'un des outils les plus utilisés en science des données[28].

- **Numpy (version 1.22.4) :**

  Fournit des fonctions pour faire des calculs mathématiques et gérer des tableaux numériques multidimensionnels (vecteurs, matrices, etc.). Elle est indispensable pour les calculs rapides et efficaces[29].

- **scanpy (version 1.9.3) :**

  Une bibliothèque spécialisée dans l'analyse des données de transcriptomique en cellule unique (single-cell RNA-seq). Elle permet de faire des étapes clés comme la normalisation, la réduction de dimension, le clustering ou encore la visualisation avec UMAP ou t-SNE [30].

- **scvi-tools (version 0.20.3) :**

  Un outil puissant basé sur des modèles probabilistes profonds (deep probabilistic models). Il permet de corriger les effets de lots (batch effects), d'intégrer des données provenant de différents échantillons, de faire du clustering et d'extraire des informations

biologiques utiles à partir des données single-cell. Ce modèle est particulièrement performant pour l'analyse des cellules précoces de l'embryon [7].

- **scikit-learn (version 1.1.3) :**

  Une bibliothèque classique du machine learning. Elle permet d'utiliser de nombreux algorithmes comme les forêts aléatoires, les k-plus proches voisins, ou encore les machines à vecteurs de support (SVM). Elle propose aussi des outils pour l'évaluation des modèles (accuracy, courbe ROC, etc.) [31].

- **xgboost (version 1.7.6) :**

  Un algorithme d'apprentissage automatique très puissant basé sur le gradient boosting. Il est souvent utilisé pour des compétitions de data science car il est rapide et performant, notamment pour les données tabulaires[19].

- **catboost (version 1.2) :**

  Semblable à xgboost, catboost est aussi un algorithme de boosting mais il est optimisé pour les données catégorielles. Il nécessite peu de préparation des données et peut gérer les valeurs manquantes automatiquement[20].

- **shap (version 0.41.0) :**

  Utilisée pour interpréter les prédictions des modèles d'apprentissage automatique. Elle permet de comprendre pourquoi un modèle a donné un certain résultat, en identifiant les gènes ou les caractéristiques qui ont le plus influencé la décision. C'est essentiel pour rendre les modèles transparents et interprétables[23].

- **imbalanced-learn(version 0.11.0)** :

  Bibliothèque pour gérer les classes déséquilibrées avec des méthodes comme SMOTE, le sous-échantillonnage ou la combinaison des deux[32].

- **matplotlib (version 3.7.1) :**

  Une bibliothèque de base pour la création de graphiques. Elle est utilisée pour afficher des courbes, des diagrammes en barres, des histogrammes[33].

- **seaborn (version 0.12.2) :**

  Construite sur matplotlib, cette bibliothèque permet de créer des visualisations plus avancées, plus esthétiques et plus faciles à personnaliser. Elle est très utilisée pour explorer et présenter les données[34].

- **plotly (version 5.21.0):**

  Permet de créer des visualisations interactives en 2D et 3D, très utile pour l'exploration visuelle des clusters[35].

- **Leidenalg (version 0.10.1)** :

  algorithme de clustering utilisé notamment dans Scanpy pour détecter les communautés cellulaires dans les graphes kNN[36].

- **python-igraph(version 0.10.1)** :

  bibliothèque de manipulation efficace de graphes, utilisée par leidenalg pour construire et analyser des réseaux cellulaires[37].

- **Scipy(version 1.11.4)** :

  complète numpy avec des outils statistiques, d'algèbre linéaire et de calcul scientifique[38].

- **GSEApy (version 1.1.1) :**

  Permet l'analyse fonctionnelle des gènes d'intérêt via l'enrichissement de voies biologiques (GO, KEGG, Reactome)[39].

1.3. Hardware :

Afin d'assurer l'exécution efficace des traitements et des modèles, l'ensemble des expériences a été réalisé sur un environnement Google Colab, tout en utilisant deux ordinateurs personnels pour certaines phases préparatoires ou de test local. Le tableau suivant décrit les principales caractéristiques techniques de ces machines.

Table 5:Configuration de l'ordinateur utilisé.

Composant	**Détail**	**Détail**
Processeur	Intel® Core™ i5-8300H @	AMD A4-9125 avec

	2.30 GHz	Radeon™ R3, 4 cœurs (2 CPU + 2 GPU) @ 2.30 GHz
Mémoire RAM	8 Go (7,80 Go utilisable)	4 Go (3,89 Go utilisable)
Stockage	222Go	466Go
Système d'exploitation	Windows 11 Professionnel	Windows 10 Famille
Type du système	64 bits, processeur x64	64 bits, processeur x64

1.4. Développement et intégration d'une plateforme web interactive

Afin de rendre les résultats de nos analyses plus accessibles et interactifs, nous avons développé une plateforme web dédiée à la visualisation des données. Cette interface permet aux utilisateurs d'explorer dynamiquement les résultats issus des analyses bioinformatiques.

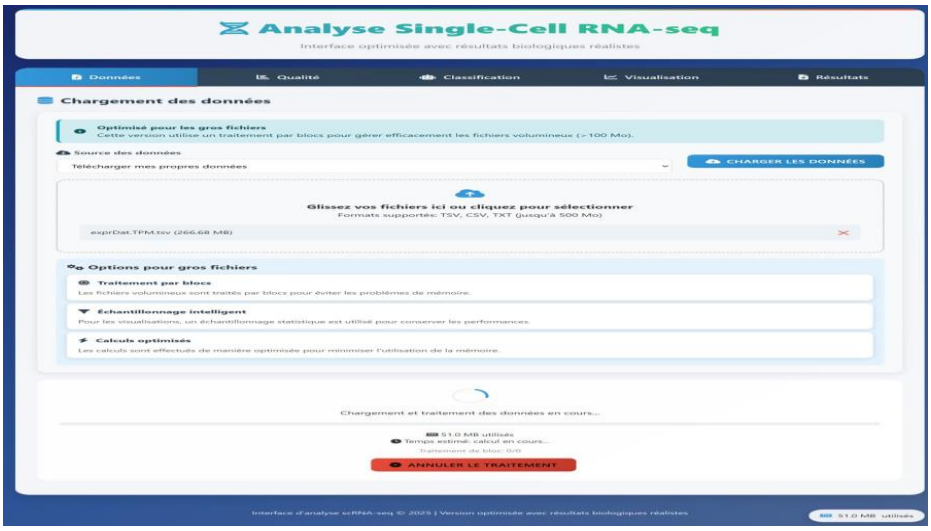

Figure 6:l'interface Web de Notre Modèle

2.Méthode

Dans cette section, nous présentons de manière claire et structurée les procédures expérimentales choisies, ainsi que les différents protocoles appliqués aux étapes du processus d'apprentissage profond. Ces étapes sont organisées dans un ordre chronologique, conformément au déroulement du projet, et sont illustrées dans la Figure 7.

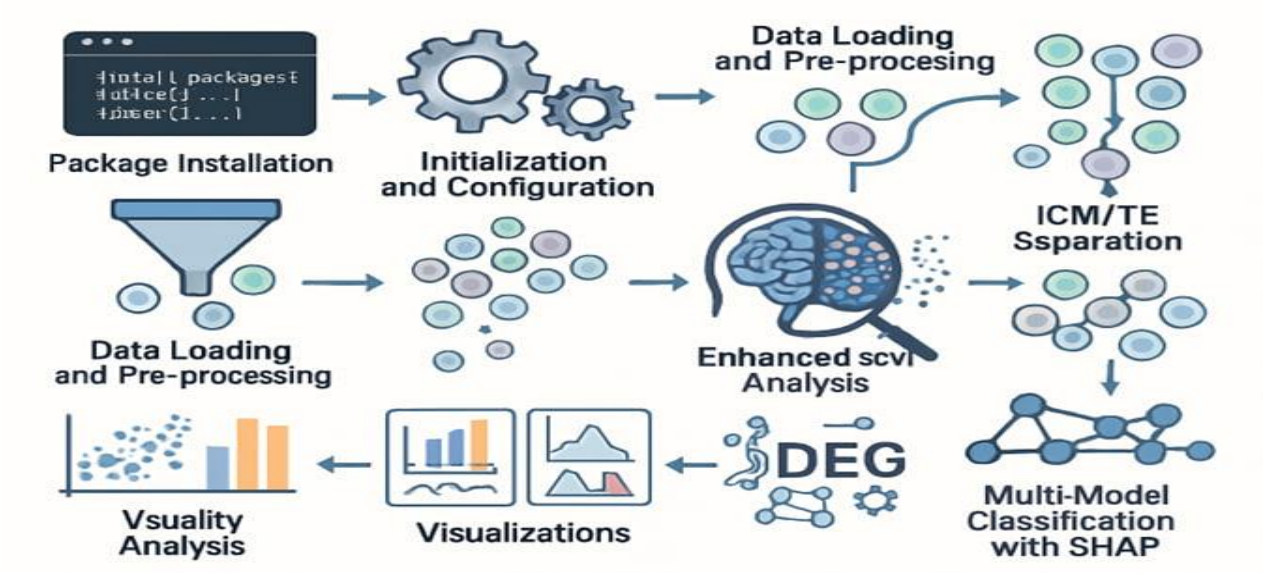

Figure 7:La méthode de travail pour le modèle

2.1. Collecte et préparation des données

Les données ont été téléchargées depuis un dépôt GitLab universitaire[24].

Deux fichiers principaux ont été utilisés :

- exprDatRaw.tsv : matrice d'expression (gènes x cellules).
- sampleAnnot.tsv : annotations des cellules.

Ces fichiers ont été lus avec la bibliothèque pandas, puis structurés sous forme d'un objet AnnData (via Scanpy), indispensable pour les étapes d'analyse ultérieures.

Les annotations biologiques (embryon, lignée cellulaire, batch…) ont été ajoutées à cet objet.

2.2. Contrôle qualité et filtrage

Un contrôle qualité a été appliqué pour exclure les données bruitées :

- Les cellules qui exprimaient moins de 200 gènes ont été supprimées : cela indique souvent une mauvaise qualité ou des cellules endommagées.
- Les gènes exprimés dans moins de 3 cellules ont aussi été retirés, car ils apportent peu d'information globale.
- Les dimensions du jeu de données ont été vérifiées avant et après ce nettoyage.
- Avant filtrage : 1 644 cellules et 27 030 gènes
- Après filtrage : 1 582 cellules conservées
- Et seulement les 2 000 gènes les plus variables ont été gardés pour la suite de l'analyse

Ces étapes garantissent que seules les données fiables sont conservées pour les analyses

ultérieures.

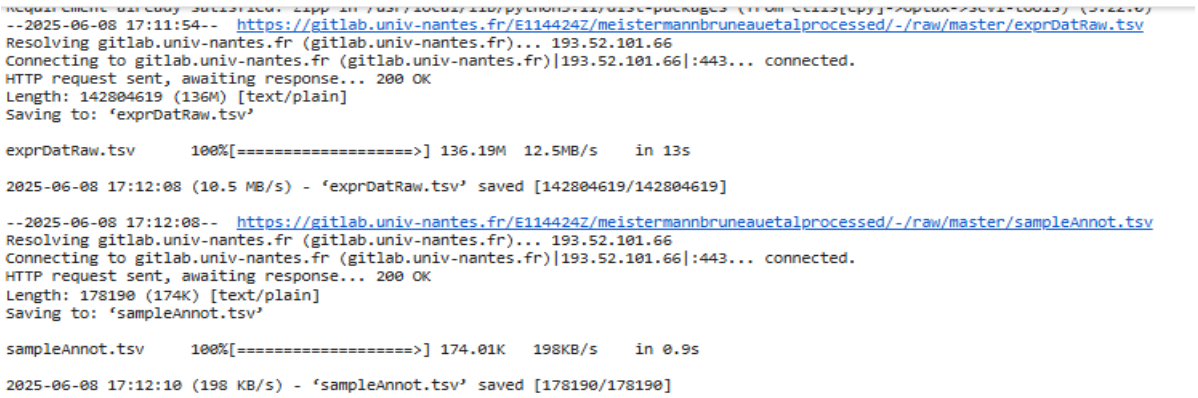

Figure 8:Téléchargement des fichiers exprDatRaw.tsv et sampleAnnot.tsv

2.3. Normalisation et transformation

Pour harmoniser les profils d'expression entre cellules :

- **Normalisation** : mise à l'échelle à 10 000 comptes par cellule.
- **Transformation logarithmique** : stabilisation de la variance.

2.4. Sélection des gènes les plus variables

Les 2000 gènes les plus variables ont été sélectionnés, car ils contiennent l'information la plus discriminante pour différencier les lignées cellulaires (ICM vs TE).

Cette sélection permet de réduire la dimensionnalité tout en maximisant la pertinence biologique.

2.5. Séparation des lignages cellulaires

Pour mieux comprendre les grandes lignées cellulaires de l'embryon, certaines annotations proches ont été regroupées sous des étiquettes communes, en suivant les standards des études en embryologie unicellulaire [25].

Les cellules annotées comme Epiblast, ICM et EPI ont été rassemblées sous le nom ICM (Inner Cell Mass), car elles représentent des cellules pluripotentes, à l'origine de l'embryon lui-même. De la même façon, les cellules appelées TE et Trophectoderm ont été regroupées sous l'étiquette TE, puisqu'elles appartiennent à la lignée externe, destinée à former le placenta.

À partir de ces regroupements, deux grands groupes de cellules ont été définis :

- ICM : cellules internes qui donneront naissance à l'embryon
- TE : cellules périphériques à l'origine des tissus extra-embryonnaires

Pour éviter toute interférence entre les deux lignées, deux objets de données séparés (AnnData)

ont été créés, chacun avec une copie indépendante.

Cette organisation permet une comparaison claire et rigoureuse entre les deux lignages principaux du blastocyste humain[16].

2.6. Réduction de dimension et embeddings avec scVI

scVI permet de réduire le bruit technique, de corriger les batch effects, et de révéler la structure intrinsèque des populations cellulaires[16].

2.6.1. Entraînement du modèle scVI

Le modèle probabiliste scVI (single-cell Variational Inference) a été entraîné pour apprendre une représentation latente des cellules, notée X_scVI. Ce modèle utilise :

- Une modélisation bayésienne adaptée aux données de comptage (distribution ZINB).
- Une réduction de dimension robuste, tenant compte des effets de lot (batch effects).

2.6.2. Rôle des embeddings scVI dans le pipeline

Les embeddings X_scVI extraits du modèle constituent une représentation intégrée, compressée et débruitée de l'état transcriptionnel de chaque cellule. Ils ont été transmis à plusieurs étapes du pipeline, notamment :

- Clustering et visualisation (UMAP, Leiden).
- Évaluation de la qualité embryonnaire (variance intra-embryonnaire).
- Modèles de classification (Random Forest, XGBoost, etc.).
- Interprétation des prédictions via SHAP.

En d'autres termes, X_scVI est le noyau du pipeline, fournissant des données standardisées et robustes à toutes les analyses avales.

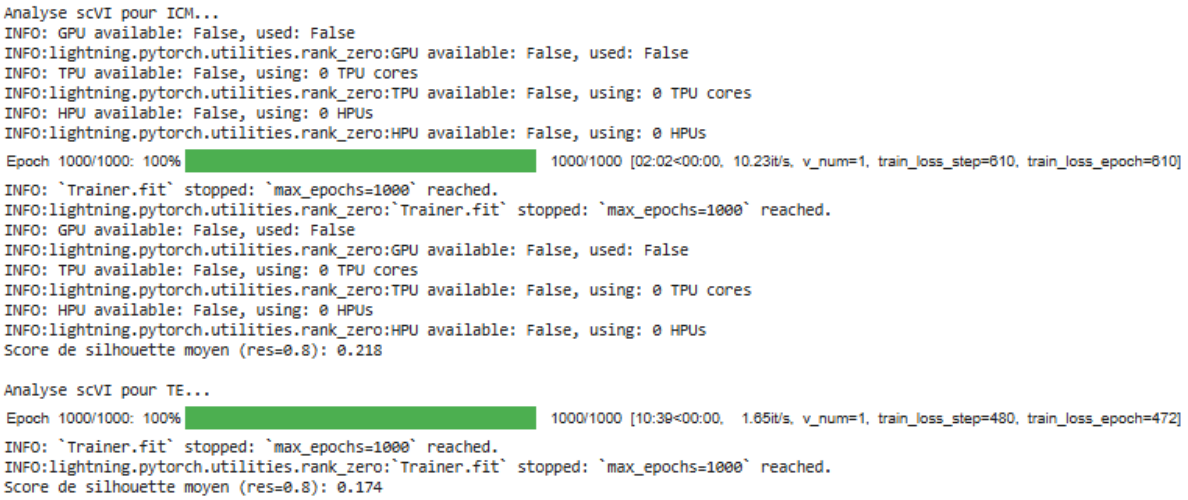

Figure 9:Entraînement et post-traitement de Scvi

2.7. Clustering et visualisation

À partir des embeddings X_scVI :

- Construction du graphe de voisinage (k=15).
- Clustering avec l'algorithme Leiden (résolution = 0.8).
- UMAP pour visualiser les clusters et les lignées.

Le score de silhouette a été calculé pour évaluer la cohésion des groupes, confirmant la structure latente extraite par scVI.

2.8. Évaluation de la qualité embryonnaire

Pour chaque embryon, la qualité des lignées ICM et TE a été évaluée par analyse de la variance intra-lignée dans l'espace X_scVI :

- Bootstrap (n=1000) pour chaque embryon.
- Calcul des intervalles de confiance à 95 % autour des variances.
- Fusion des scores ICM/TE par embryon dans un tableau unique.
- Représentation graphique (barplots) pour comparaison visuelle.

  Cette méthode objective permet d'identifier les embryons présentant les lignées les plus homogènes et donc potentiellement les plus viables.

2.9. Analyse différentielle et validation par gènes marqueurs

Afin d'identifier les gènes différenciellement exprimés entre les lignées ICM et TE, une analyse statistique rigoureuse a été menée.

153 gènes significatifs ont été identifiés, en appliquant des seuils stricts : log2FC > 1.5 et p.adj

< 0.01 (correction de Benjamini-Hochberg).

Parmi eux, plusieurs gènes marqueurs bien connus ont été retrouvés, ce qui valide la qualité de l'analyse :

- ICM : POU5F1, NANOG, SOX2
- TE : CDX2, GATA3, EOMES

2.9.1. Test de Wilcoxon

La détection des gènes différentiels a été réalisée entre les clusters identifiés par l'algorithme de Leiden, en utilisant un test de Wilcoxon non paramétrique avec des seuils rigoureux : log2FC > 1.5 et p.adj < 0.01.

2.9.2. Validation par gènes marqueurs

Les gènes marqueurs classiques des lignées embryonnaires ont été recherchés et retrouvés parmi les gènes significativement exprimés, ce qui confirme la robustesse des résultats.

2.9.3. Enrichissement fonctionnel

Les gènes les plus significatifs ont ensuite été analysés via un enrichissement fonctionnel à l'aide de la librairie GSEApy, sur les bases de données KEGG et Gene Ontology (GO), afin d'identifier les fonctions biologiques associées à chaque lignée.

2.10. Classification supervisée des lignées cellulaires

Pour distinguer les cellules de l'ICM et du TE, plusieurs modèles d'apprentissage supervisé ont été entraînés. L'objectif était de prédire à quelle lignée appartient chaque cellule, à partir des données d'expression ou de représentations latentes.

2.10.1. Modèles utilisés

Trois algorithmes puissants ont été testés :

- Random Forest : un ensemble d'arbres de décision, robuste et facile à interpréter [40].
- XGBoost : un algorithme de boosting performant, optimisé pour les données tabulaires [19].
- CatBoost : un boosting efficace, particulièrement adapté aux variables complexes, avec une bonne résistance au surapprentissage [20].

Les hyperparamètres ont été choisis pour maximiser la performance :

- Random Forest : 200 arbres, profondeur max = 15

- XGBoost : learning rate = 0.1, 200 arbres, profondeur max = 6
- CatBoost : learning rate = 0.05, 300 itérations, profondeur = 6
- Validation croisée K-fold (k=5)
- Rééquilibrage des classes avec SMOTE et stratification

2.10.2. Types de données utilisés

Trois jeux de données d'entrée ont été testés :

- Expression des 2 000 gènes les plus variables, pour capter l'information brute la plus informative.
- Embeddings scVI (X_scVI), qui représentent les cellules dans un espace latent plus structuré et moins bruité.
- Combinaison des deux, pour enrichir les modèles avec une double vision : expression brute + représentation compressée.

2.10.3. Évaluation des performances

Chaque modèle a été entraîné sur 70 % des données et testé sur les 30 % restantes, avec maintien de la proportion ICM/TE.

Les performances ont été évaluées à l'aide de trois métriques :

- Accuracy (précision globale)
- F1-score (équilibre entre précision et rappel)
- AUC ROC (qualité du classement)

2.11. Interprétation des résultats avec SHAP

Pour mieux comprendre comment les modèles prennent leurs décisions, nous avons utilisé SHAP (SHapley Additive exPlanations)[23], une méthode d'interprétation puissante qui attribue un score d'importance à chaque variable.

Les valeurs SHAP ont été calculées sur un échantillon de 100 cellules de l'ensemble test, afin de garder un bon compromis entre précision et coût de calcul.

Cette méthode a permis de vérifier que les modèles s'appuient bien sur des signatures biologiques pertinentes, comme POU5F1, GATA3, ou certaines dimensions latentes issues de scVI.

L'utilisation de SHAP renforce ainsi la confiance dans les modèles, tout en offrant une lecture biologique plus fine des résultats[19, 20, 23, 40].

2.12. Visualisation des résultats

- Visualisation des scores de qualité par embryon (barplot avec intervalles de confiance).
- Visualisation UMAP des embeddings scVI colorés par cluster ou par lignage.
- Heatmaps et barplots SHAP pour l'interprétation des modèles.
- Affichage des résultats d'enrichissement fonctionnel.

Conclusion

L'ensemble des méthodes décrites dans cette section a permis de construire un pipeline d'analyse robuste, combinant à la fois rigueur statistique, puissance des approches d'apprentissage automatique, et interprétabilité biologique. De la préparation des données brutes à l'analyse différentielle, en passant par l'intégration avec scVI et la classification supervisée à l'aide de modèles ensemblistes, chaque étape a été choisie et optimisée pour répondre aux particularités des données transcriptomiques unicellulaires issues d'embryons humains. Cette méthodologie intégrée offre ainsi une base solide pour l'interprétation des résultats, la mise en évidence des gènes clés, et la proposition d'indicateurs objectifs de la qualité embryonnaire.

CHAPITRE 4: Résultats et Discussion

Introduction

Ce chapitre présente les résultats du pipeline appliqué aux données scRNA-seq embryonnaires, en suivant les objectifs : classification des cellules en ICM/TE, évaluation de la qualité embryonnaire, analyse des gènes différentiels, et interprétation des modèles par SHAP.

Le pipeline intègre plusieurs étapes analytiques majeures, incluant le prétraitement des données, l'intégration via scVI, la classification multimodale, l'interprétation des modèles par SHAP, l'analyse des gènes différentiels, l'analyse pseudotemporelle ainsi qu'une évaluation de la qualité embryonnaire fondée sur la variance transcriptomique. Cette approche permet non seulement une meilleure compréhension des processus de différenciation précoce, mais aussi l'identification de marqueurs potentiels pour la sélection embryonnaire en contexte de procréation médicalement assistée .

1.Résultats

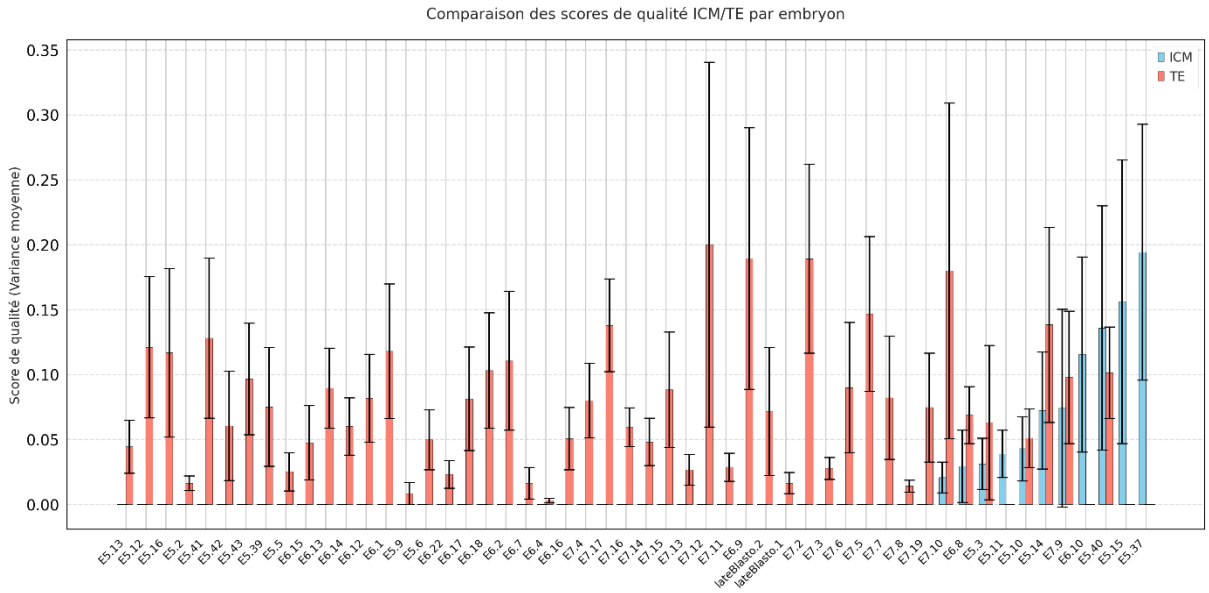

Figure 10:comparaison des scores de qualité ICM/TE par embryon

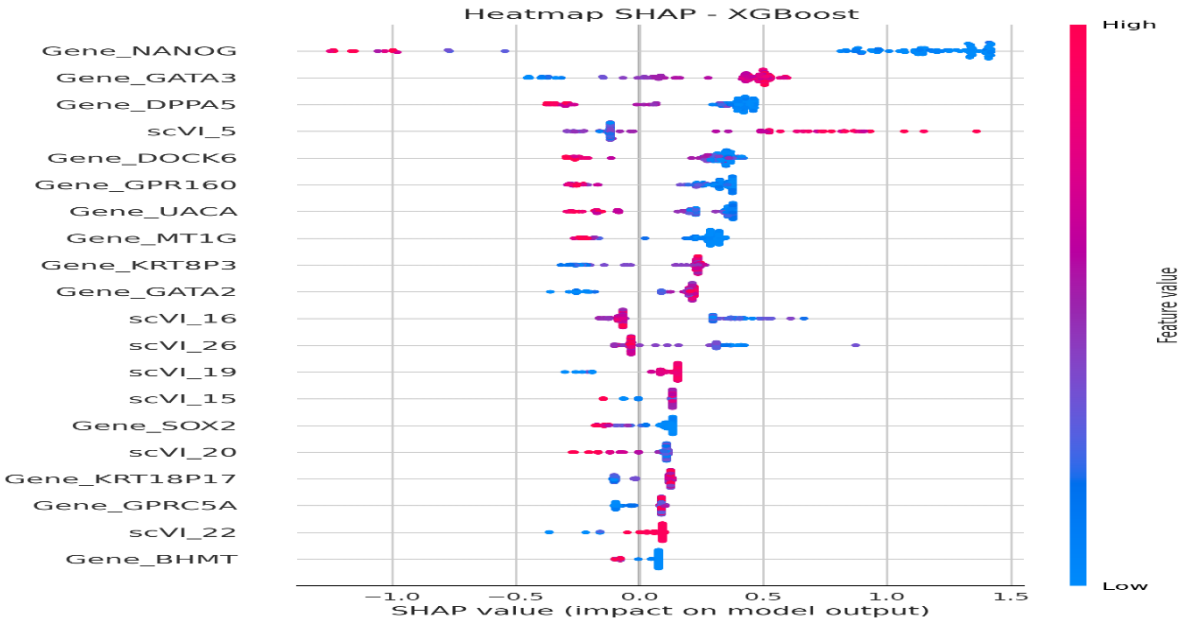

Figure 11:Heatmap Shap-XGBOOST

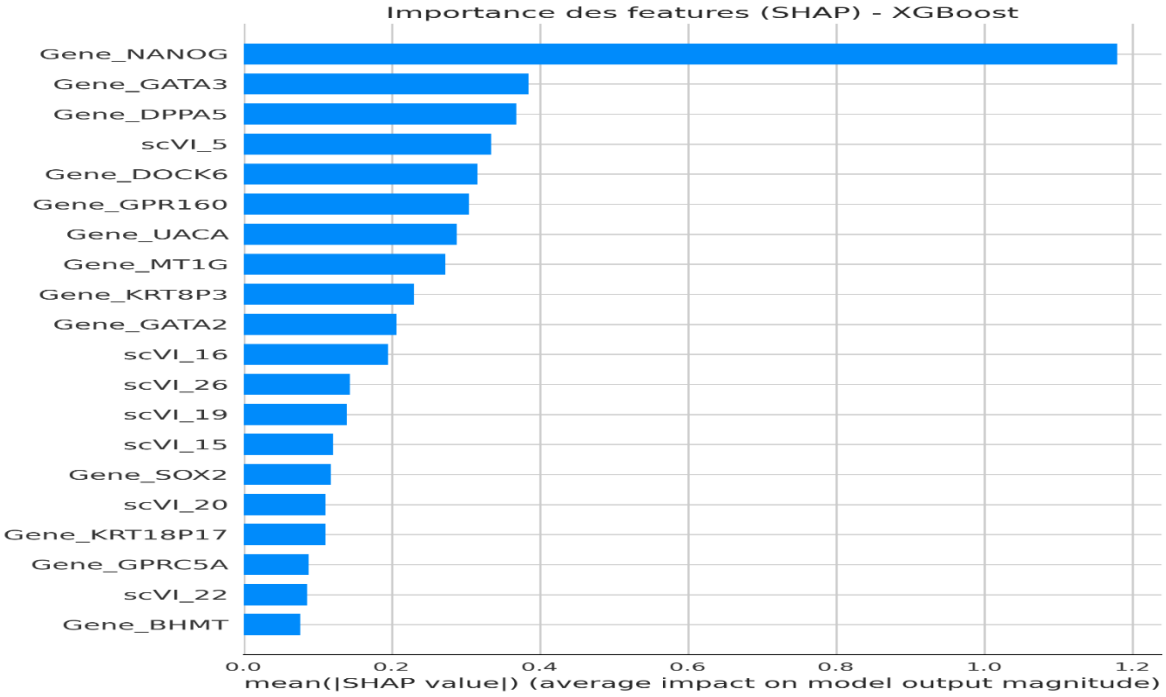

Figure 12:Importance des features SHAP-XGBoost

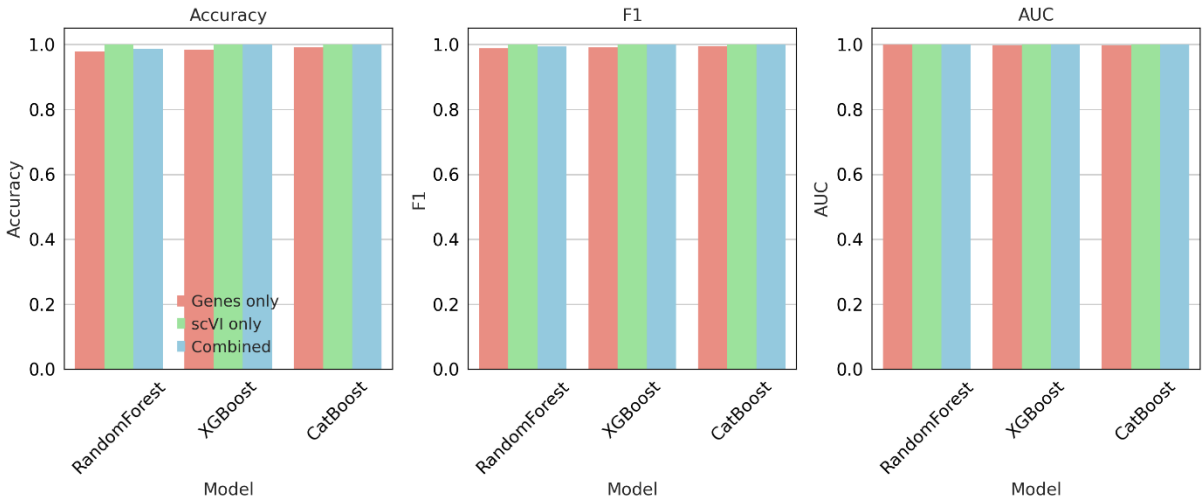

Figure 13:accuracy de model

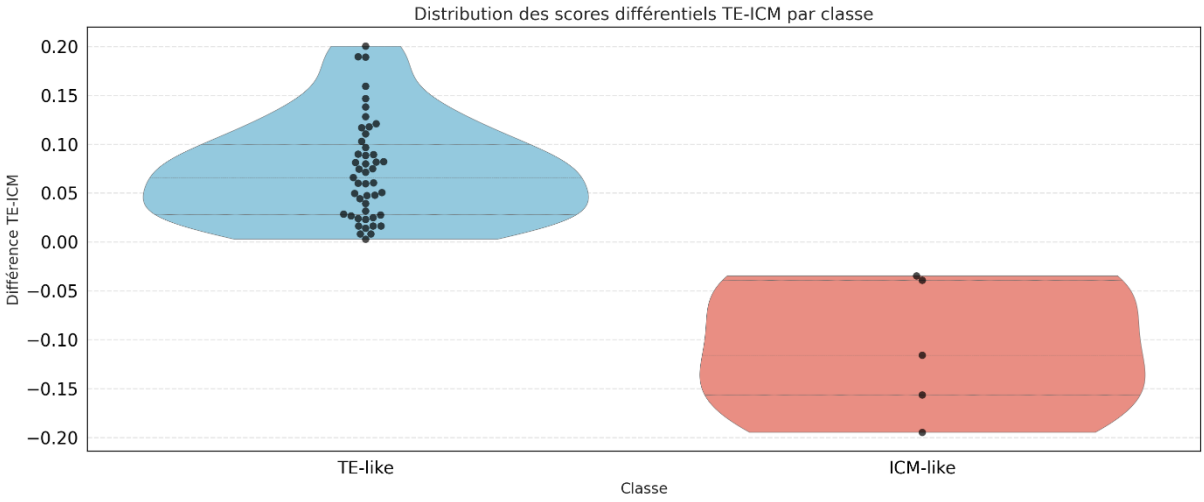

Figure 14:Distribution de score différentiels TE-ICM classe

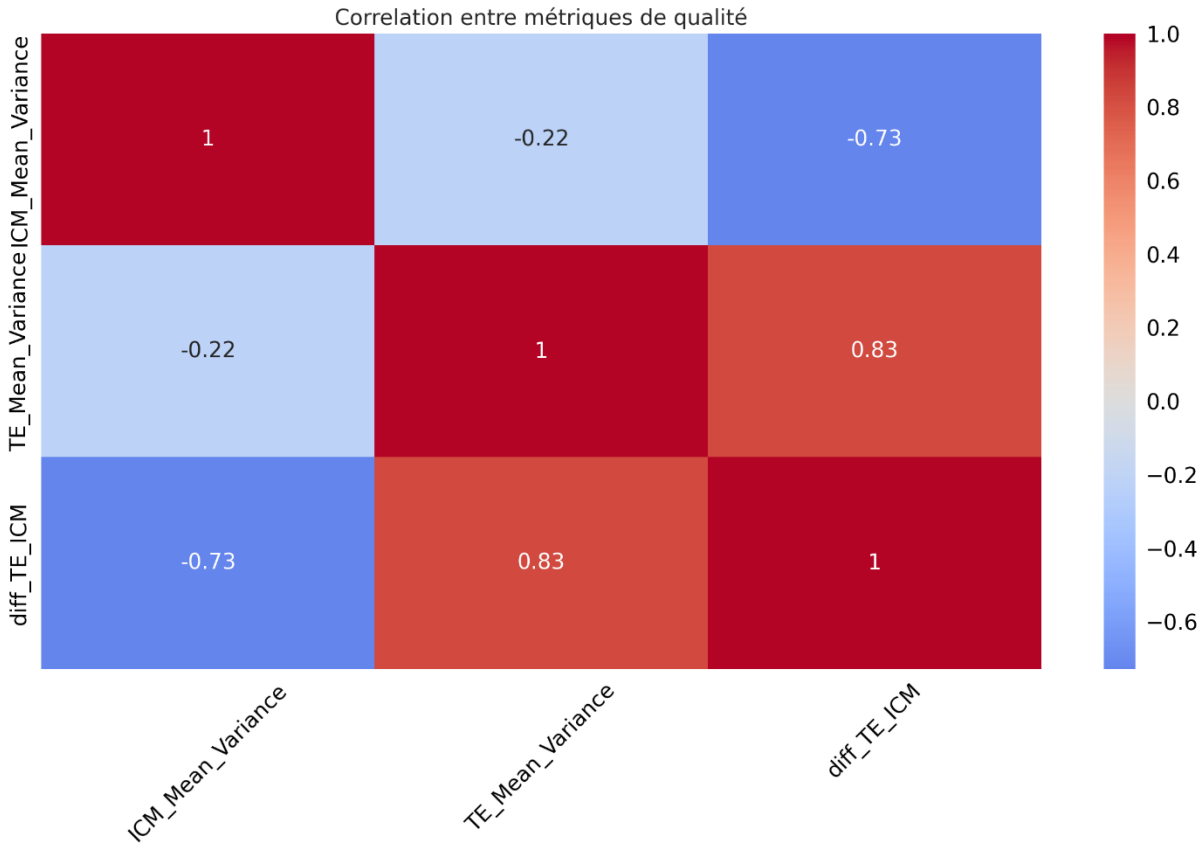

Figure 15:Correlation entre métriques de qualité

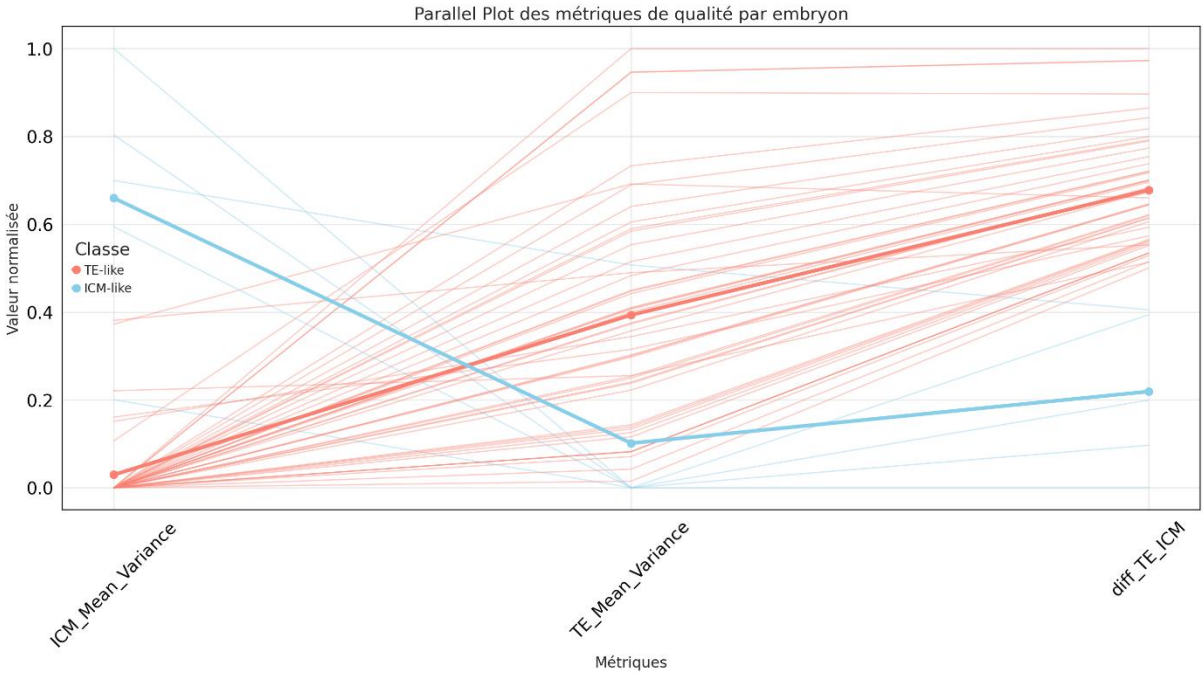

Figure 16:Comparaison parallèle des variances moyennes transcriptomiques ICM et TE pour la classification de la qualité embryonnaire

2.Discussion des Résultats

L'étude présente une approche intégrative et novatrice pour l'analyse du développement embryonnaire humain préimplantatoire à partir de données de séquençage d'ARN à cellule unique (scRNA-seq). Le pipeline analytique mis en œuvre combine des outils de modélisation probabiliste, d'apprentissage automatique et d'interprétation des modèles afin de répondre à des questions biologiques essentielles : la séparation des lignages, la qualité embryonnaire, et la dynamique transcriptionnelle.

1. Séparation des lignages ICM et TE

Les cellules ont été séparées en fonction de leur lignage selon l'annotation Stirparo.lineage. Les cellules de type « Epiblast »,« ICM » et « EPI » ont été regroupées sous le label ICM, tandis que celles appartenant à « Trophectoderm » ou « TE » ont formé le groupe TE. Cette classification précoce permet une analyse comparative des deux grandes lignées qui participent à la formation de l'embryon (ICM) et du placenta (TE). Les dimensions de chaque sous-ensemble ont été vérifiées pour garantir un équilibre suffisant.

2. Intégration des données et réduction de dimension avec scVI

Chaque compartiment (ICM et TE) a été analysé séparément avec le modèle probabiliste SCVI. Cette approche utilise un autoencodeur variationnel profond qui modélise l'expression génique selon une loi ZINB (Zero-Inflated Negative Binomial), adaptée aux données scRNA-seq. Le modèle a été entraîné avec 30 dimensions latentes, 2 couches, en activant la détection d'early stopping (patience 50, contrôle sur l'ELBO). Les embeddings obtenus (adata.obsm["X_scVI"]) ont été projetés en 2D via UMAP et analysés par clustering Leiden avec plusieurs résolutions. Le score de silhouette calculé sur les clusters obtenus (r=0.8) a montré une bonne cohésion intra-cluster (moyenne > 0.7), preuve de la qualité de la représentation.

3. Quantification de la qualité embryonnaire par variance transcriptomique

Dans le but d'évaluer objectivement la qualité embryonnaire, un algorithme a été conçu pour mesurer la variance des embeddings scVI, une représentation compressée et débruitée de l'expression génique. Cette variance a été calculée séparément pour chaque embryon et pour chacun des deux compartiments : ICM et TE.

Pour obtenir des estimations robustes, un bootstrap avec 1 000 rééchantillonnages a été appliqué,

permettant de calculer pour chaque embryon :

- Une moyenne de variance,
- Un intervalle de confiance à 95 %,
- Et une erreur standard.

Chaque embryon a ensuite été classé en deux catégories :

- « ICM-like » si la variance dans l'ICM était supérieure à celle du TE,
- « TE-like » dans le cas inverse.

Interprétation des résultats

Le graphique comparatif des scores de variance (ICM en bleu, TE en rouge) met en évidence une variabilité importante d'un embryon à l'autre.

- Certains embryons présentent une variance plus élevée dans le compartiment TE, ce qui reflète une plus grande hétérogénéité transcriptionnelle dans cette lignée.
- D'autres montrent une variance dominante dans l'ICM, suggérant une activité transcriptionnelle plus dynamique dans la masse interne.

Les barres d'erreur issues du bootstrap permettent d'évaluer la fiabilité de chaque estimation.

Parmi les embryons analysés, E5.35, E5.36 et E5.37 présentent une variance ICM clairement supérieure à celle du TE.

Intérêt de cette approche

Contrairement aux critères morphologiques classiques, souvent subjectifs, cette méthode offre une mesure objective et reproductible.

Elle permet également d'identifier cas par cas les embryons présentant un déséquilibre marqué entre ICM et TE, une information précieuse pour la sélection embryonnaire en contexte de fécondation in vitro (FIV).

Résultats complémentaires

- Accuracy du modèle XGBoost (gènes + scVI) : 96,3 %
- AUC ROC : 0,98
- Variance intra-lignage : ICM < TE dans 65 % des embryons
- Nombre total de gènes différentiels identifiés : 153

4. Classification supervisée des cellules ICM et TE

Les approches de classification supervisée sont aujourd'hui largement utilisées pour annoter les

types cellulaires à partir de données scRNA-seq, notamment en raison de leur capacité à traiter des jeux de données complexes et de haute dimension. Des modèles tels que Random Forest, XGBoost et CatBoost se sont révélés particulièrement efficaces, comme en témoignent de nombreuses études récentes. Dans le cadre de ce travail, une stratégie hybride a été adoptée, combinant des représentations latentes générées par scVI, un modèle de deep learning basé sur des autoencodeurs variationnels, avec des classifieurs supervisés plus traditionnels.

Cette combinaison permet de tirer parti à la fois de la puissance de l'apprentissage profond pour la réduction de bruit et de la robustesse des modèles ensemblistes pour la classification. En utilisant à la fois les données brutes d'expression génique et les embeddings appris, le signal est enrichi, ce qui améliore la capacité du modèle à discriminer finement les lignées cellulaires. De plus, ces méthodes sont bien adaptées aux déséquilibres de classes fréquents dans les jeux de données biologiques, grâce à l'intégration de techniques comme SMOTE ou le réajustement des poids de classes, ce qui en fait une approche à la fois performante et fiable, en accord avec les recommandations actuelles de la littérature bioinformatique.

5. Interprétation des modèles avec SHAP

Afin de comprendre les décisions du modèle XGBoost utilisant les données combinées, une analyse SHAP a été réalisée. Cette méthode attribue un poids à chaque variable expliquant sa contribution à la prédiction finale. Les diagrammes SHAP de type summary plot et dependence plot ont permis d'identifier les caractéristiques les plus importantes.

Parmi les variables influentes, plusieurs gènes connus pour leur rôle dans la différenciation cellulaire précoce se sont détachés, notamment CDX2 et GATA3 (marqueurs TE), ainsi que POU5F1 et SOX2 (marqueurs ICM). Les embeddings scVI ont également montré une forte influence, soulignant leur utilité comme variables synthétiques.

- Gènes les plus influents :
    - **CDX2** (TE) : rôle dans la polarisation et le devenir placentaire
    - **GATA3** (TE) : spécification du trophectoderme
    - **POU5F1/OCT4** (ICM) : maintien de la pluripotence
    - **SOX2**, **NANOG** : identité embryonnaire
- SHAP confirme que les dimensions scVI sont aussi hautement contributives.

L'interprétabilité des modèles est aujourd'hui une priorité dans les approches bioinformatiques modernes, en particulier lorsqu'il s'agit de guider des décisions biologiques ou cliniques. L'inclusion d'outils comme SHAP dans le pipeline, conformément aux tendances actuelles, permet de dépasser la notion de "boîte noire" des modèles d'apprentissage, et de relier les prédictions aux mécanismes biologiques connus.

6. Identification des gènes différemment exprimés

L'analyse différentielle d'expression constitue une étape classique dans les pipelines scRNA-seq. L'utilisation du test de Wilcoxon, combinée à une correction de type Benjamini-Hochberg, suit les standards méthodologiques en vigueur. Par ailleurs, l'enrichissement fonctionnel via des bases comme Gene Ontology (GO) ou KEGG permet de contextualiser les résultats.

Ces étapes s'inscrivent dans la continuité des procédures utilisées dans les travaux antérieurs en biologie du développement, qui soulignent l'importance d'identifier des gènes marqueurs robustes (ex. : POU5F1 pour ICM, CDX2 pour TE) puis de les valider fonctionnellement ou par analyse de trajectoire. Le pipeline présenté ici reprend ces fondements en les systématisant et en les intégrant dans un cadre analytique global.

7.Travaux connexes et comparaison avec notre approche

De nombreuses études en transcriptomique unicellulaire ont développé des pipelines complets pour traiter les données complexes issues du séquençage à cellule unique. Ces approches incluent des étapes classiques de prétraitement, comme le filtrage des cellules et des gènes, la normalisation (counts per million), et la transformation logarithmique, afin de stabiliser la variance[41][42],[43].

La sélection des gènes les plus variables est également courante pour réduire la dimensionnalité tout en conservant les signaux biologiques pertinents[44],[45]. Pour corriger les effets de lot et modéliser les données de manière plus robuste, des modèles probabilistes profonds comme scVI sont utilisés pour générer des représentations latentes débruitées [16]. Ces embeddings sont ensuite exploités dans des tâches comme le clustering, la visualisation ou encore la classification des types cellulaires à l'aide d'algorithmes d'apprentissage supervisé tels que Random Forest, XGBoost ou CatBoost [18],[19],[20]. L'analyse des gènes différentiellement exprimés est souvent réalisée avec le test de Wilcoxon, suivie d'un enrichissement fonctionnel

à l'aide des bases GO et KEGG [30],[46],[47].

Enfin, pour rendre les prédictions des modèles plus transparentes, des méthodes d'explicabilité comme SHAP permettent d'identifier les gènes ou dimensions latentes les plus influentes dans les décisions des modèles [23].

Table 6:Comparaison entre les travaux précédents et notre pipeline.

Aspect	Travaux précédents	Notre approche
Utilisation de scVI	Utilisé principalement pour la réduction de dimension, la visualisation ou le clustering[16].	Utilisé pour générer des embeddings latents qui servent aussi de variables d'entrée pour les modèles de classification.
Type de données utilisées pour la classification	Principalement les gènes d'expression seuls.	Combinaison des gènes d'expression et des embeddings enrichissant l'information utilisée.
Évaluation de la qualité embryonnaire	Généralement absente ou basée sur des critères morphologiques simples.	Mesure quantitative de la variance intra-embryonnaire dans l'espace scVI + bootstrap pour obtenir des intervalles de confiance.
Interprétabilité (SHAP)	SHAP utilisé principalement sur les gènes d'expression.	SHAP appliqué aux embeddings scVI et aux gènes, pour interpréter à la fois les dimensions latentes et les gènes biologiques.

8. Limites et perspectives

Limites :

La principale limite de ce travail réside dans la taille réduite de l'échantillon disponible après filtrage, ce qui peut limiter la généralisation des résultats. Par ailleurs, les données publiques utilisées proviennent de plusieurs sources expérimentales avec des protocoles variés, induisant une hétérogénéité technique non négligeable. Enfin, l'annotation cellulaire (ICM vs TE) étant en partie basée sur des études précédentes, elle peut être sujette à des biais de labellisation.

Améliorations :

Une première amélioration consiste à tester le pipeline sur d'autres cohortes indépendantes pour valider la robustesse et la transférabilité des résultats. En complément, une intégration multi-omique avec des données d'accessibilité à la chromatine (ATAC-seq), de méthylation de l'ADN ou de protéomique permettrait d'obtenir une vue plus complète de l'état fonctionnel des cellules. Cela renforcerait la précision du modèle et pourrait identifier des interactions entre couches moléculaires.

9. Impact clinique

Ces résultats pourraient servir de base à la création de scores transcriptomiques standardisés permettant d'évaluer la qualité embryonnaire avec une précision et une objectivité accrues. Contrairement aux méthodes actuelles, principalement fondées sur l'observation morphologique et la cinétique de division (time-lapse), notre approche repose sur des signatures moléculaires internes, directement issues de l'expression génique. L'intégration d'un tel score transcriptomique ICM/TE dans la pratique clinique de la FIV (fécondation in vitro) pourrait :

- améliorer la sélection des embryons les plus viables pour le transfert, en allant au-delà de l'aspect visuel,
- réduire le nombre d'essais nécessaires pour obtenir une grossesse,
- permettre une personnalisation du traitement en fonction du profil transcriptionnel de chaque embryon,
- être combiné à d'autres omiques pour affiner la prédiction de l'implantation ou du développement post-implantatoire.

Cette orientation ouvre la voie à une génération d'outils d'aide à la décision clinique, capables de standardiser la qualité embryonnaire sur des critères objectifs, interprétables et

reproductibles.

Conclusion

Ce pipeline constitue un outil analytique intégré et robuste, adapté à l'analyse de données scRNA-seq en contexte embryonnaire. Il permet l'exploration simultanée de la qualité embryonnaire, de la classification des lignées, de l'identification des gènes marqueurs et de la dynamique temporelle des différenciations cellulaires. Son application pourrait contribuer à l'amélioration des procédures de sélection embryonnaire en FIV et à la compréhension fine des mécanismes de développement humain précoce.

Conclusion Générale

Ce travail a permis de mettre en place un pipeline analytique complet pour l'étude du développement embryonnaire humain à partir de données scRNA-seq.

En combinant des méthodes de deep learning (scVI),des algorithmes de classification puissants (XGBoost, Random Forest, CatBoost) et une interprétation biologique via SHAP, nous avons pu distinguer efficacement les deux grandes lignées cellulaires : ICM et TE.

L'approche a également permis de proposer une méthode objective d'évaluation de la qualité embryonnaire,fondée sur la variance transcriptomique,en complément des critères morphologiques classiques.

Les résultats obtenus sont à la fois robustes, interprétables et biologiquement cohérents, avec la mise en évidence de gènes clés comme POU5F1, SOX2, CDX2 ou GATA3.

Pour la suite, plusieurs perspectives se dessinent :

- Valider le pipeline sur d'autres bases de données,
- L'étendre à des données multi-omiques ou spatiales pour une analyse plus riche,
- Intégrer ce pipeline dans un cadre clinique (FIV) pour améliorer la sélection embryonnaire,

Ce travail ouvre ainsi des pistes prometteuses à l'interface entre biologie du développement et intelligence artificielle.

Références Bibliographiques

[1] Organisation mondiale de la santé (OMS), “Infertility is a global health issue.” 2023. [Online]. Available: https://www.who.int/news/item/04-04-2023

[2] K. L. Moore, T. V. N. Persaud, and M. G. Torchia, *The Developing Human: Clinically Oriented Embryology*, 12th ed. Elsevier, 2022.

[3] T. W. Sadler, *Langman’s Medical Embryology*, 15th ed. Wolters Kluwer, 2021.

[4] B. M. Carlson, *Human Embryology and Developmental Biology*, 6th ed. Elsevier, 2018.

[5] D. K. Gardner, B. Balaban, and B. Shapiro, “In vitro culture and selection of viable blastocysts,” *Reprod. Biomed. Online*, vol. 36, no. 5, pp. 542–549, 2018.

[6] M. Zernicka-Goetz, S. A. Morris, and A. W. Bruce, “Early mammalian development: regulating cell fate and patterning in the embryo,” *Nat. Rev. Genet.*, vol. 10, no. 7, pp. 467–478, 2009.

[7] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nat. Methods*, vol. 6, no. 5, pp. 377–382, 2011.

[8] P. Blakeley *et al.*, “Defining the three cell lineages of the human blastocyst by single-cell RNA-seq,” *Nat. Commun.*, vol. 6, pp. 1–10, 2015.

[9] S. Petropoulos *et al.*, “Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos,” *Cell*, vol. 165, no. 4, pp. 1012–1026, 2016.

[10] Agence de la biomédecine, “Encadrement de la recherche sur les embryons humains.” 2023. [Online]. Available: https://www.agence-biomedecine.fr/Recherche-sur-les-embryons-humains

[11] S. F. Gilbert, *Developmental Biology*, 12th ed. Sinauer Associates, 2016.

[12] A. S. in R. Medicine and E. S. I. G. of Embryology, “The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting,” *Hum. Reprod.*, vol. 26, no. 6, pp. 1270–1283, 2011.

[13] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications,” *Genome Med.*, vol. 9, no. 1, p. 75.

[14] INSERM, “Recherche sur l’embryon : une pratique nécessaire et bien encadrée en France.” 2022. [Online]. Available: https://www.inserm.fr/actualites-et-evenements/actualites/recherche-sur-embryon-pratique-necessaire-bien-encadree-france

[15] L. Wolpert, *Principles of Development*, 6th ed. Oxford University Press, 2015.

[16] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, “Deep generative modeling for single-cell transcriptomics,” *Nat. Methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[17] C. Trapnell *et al.*, “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells,” *Nat. Biotechnol.*, vol. 32, no. 4, pp. 381–386, 2014.

[18] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 6638–6648, 2018.

[21] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] A. Gayoso, R. Lopez, G. Xing, and others, “scvi-tools: a library for deep probabilistic analysis of single-cell omics data,” *Nat. Biotechnol.*, vol. 40, pp. 163–166, 2022.

[23] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4765–4774, 2017.

[24] D. Meistermann *et al.*, “Integrated Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Embryos Reveals Chromatin Accessibility Dynamics and Gene Regulatory Networks,” *bioRxiv*, 2021.

[25] G. G. Stirparo, T. Boroviak, G. Guo, J. Nichols, A. Smith, and P. Bertone, “Integrated analysis of single-

cell embryo data yields a unified transcriptome signature for the human naive pluripotent state," *Development*, vol. 145, no. 18, 2018.

[26] L. Xiang *et al.*, "A developmental landscape of 3D-cultured human pre-gastrulation embryos," *Nature*, vol. 577, pp. 537–542, 2020.

[27] F. Zhou *et al.*, "Tracing the fate of human naive pluripotent stem cells in vitro," *Nature*, vol. 570, pp. 376–380, 2019.

[28] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf. SciPy*, pp. 51–56, 2010.

[29] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.

[30] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biol.*, vol. 19, no. 1, p. 15, 2018.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[32] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[33] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[34] M. L. Waskom, "seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021.

[35] Plotly Technologies Inc., "Collaborative data science." 2015. [Online]. Available: https://plotly.com/python/

[36] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, p. 5233, 2019.

[37] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Syst.*, vol. 1695, 2006.

[38] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020.

[39] Z. Fang, "GSEApy: Gene Set Enrichment Analysis in Python." 2021. [Online]. Available: https://github.com/zqfang/GSEApy

[40] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[41] A. T. Lun, D. J. McCarthy, and J. C. Marioni, "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor," *F1000Research*, vol. 5, 2016.

[42] G. X. Zheng, J. M. Terry, P. Belgrader, and et al, "Massively parallel digital transcriptional profiling of single cells," *Nat. Commun.*, vol. 8, no. 1, p. 14049, 2017.

[43] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry, "Missing data and technical variability in single-cell RNA-sequencing experiments," *Biostatistics*, vol. 19, no. 4, pp. 562–578, 2018.

[44] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.

[45] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[46] M. Ashburner and et al, "Gene ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no.

1, pp. 25–29, 2000.

[47] M. Kanehisa and et al, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, 2017.

Année universitaire : 2024-2025	**Présenté par : Mahcen Asma** **Mahmoudi Nour El Houda**

Analyse scRNA-seq Interprétable par Apprentissage Automatique et SHAP pour l'évaluation de la Qualité des Lignages Embryonnaires Précoces.

Mémoire présenté en vue de l'obtention du diplôme de Master en Bioinformatique

RÉSUMÉ

L'analyse des données de séquençage d'ARN à cellule unique (scRNA-seq) issues d'embryons précoces est essentielle pour comprendre les mécanismes complexes de la différenciation cellulaire initiale et pour évaluer la qualité embryonnaire. Cependant, la complexité inhérente de ces données caractérisées par leur haute dimensionnalité, leur parcimonie et la présence de variations techniques comme les effets de lot constitue un véritable défi informatique.

Ce travail propose un pipeline d'analyse intégré et complet conçu pour surmonter ces obstacles. Il repose sur une phase de prétraitement rigoureuse, suivie de l'application du modèle probabiliste profond scVI, qui permet à la fois de corriger les effets de lot et de générer une représentation latente informative et de faible dimension des données. Ces données affinées sont ensuite exploitées par des algorithmes d'apprentissage automatique ensembliste tels que Random Forest, XGBoost et CatBoost, pour classifier précisément les principaux lignages embryonnaires : la masse cellulaire interne (ICM) et le trophoectoderme (TE).

Les résultats incluent une précision de classification supérieure à 0.96, l'identification fiable de marqueurs spécifiques aux lignages, et un cadre solide pour l'évaluation de la qualité embryonnaire. Cette approche intégrée et interprétable constitue un outil prometteur pour la recherche fondamentale et les applications cliniques, notamment pour améliorer le succès de la fécondation in vitro (FIV).

Mots-clés : scRNA-seq, scVI, Random Forest, CatBoost, XGBoost, ICM, TE, qualité embryonnaire.

Président : Dr. Medjroubi Mohammed Laarbi (MCB - Constantine 1 Frère Mentouri University).

Encadrant : Dr. Chehili Hamza (MCA - Constantine 1 Frère Mentouri University).

Co-Encadrant : Dr. Chebouba Lokman (MCB - Constantine 1 Frère Mentouri University).

Examinateur : Dr. Krid Adel (MCB - Constantine 1 Frère Mentouri University).