



HAL
open science

Intelligence artificielle et recherche de lentilles gravitationnelles dans les grands relevés astronomiques

Quentin Petit

► **To cite this version:**

Quentin Petit. Intelligence artificielle et recherche de lentilles gravitationnelles dans les grands relevés astronomiques. Astrophysique [astro-ph]. Université de Bordeaux, 2023. Français. NNT : 2023BORD0275 . tel-04438431

HAL Id: tel-04438431

<https://theses.hal.science/tel-04438431>

Submitted on 5 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



université
de **BORDEAUX**



THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE
DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

ECOLE DOCTORALE SCIENCES PHYSIQUES ET DE L'INGÉNIEUR
ASTROPHYSIQUE, PLASMAS, NUCLÉAIRE

Par **Quentin PETIT**

Intelligence artificielle et recherche de lentilles gravitationnelles dans les
grands relevés astronomiques

Sous la direction de : **Christine DUCOURANT**
Co-direction : **Eric SLEZAK**

Soutenance prévue le 03 novembre 2023

Membres du jury :

Mme. Roser PELLÓ	Astronome	LAM	Rapporteure / Présidente
Mme. Geneviève SOUCAIL	Astronome	Observatoire Midi-Pyrénées	Rapporteure
M. Dominique SLUSE	Professeur	Université de Liège	Examineur
M. Alexandre BOUCAUD	Ingénieur de recherche	APC	Examineur
Mme. Christine DUCOURANT	Astronome adjoint	LAB	Directrice de thèse
M. Eric SLEZAK	Astronome	Observatoire de la Côte d'Azur	Co-Directeur de thèse
M. Ludovic DELCHAMBRE	Maître de conférence	Université de Liège	Invité

Intelligence artificielle et recherche de lentilles gravitationnelles dans les grands relevés astronomiques

Résumé : L'effet de lentille gravitationnelle est l'une des conséquences les plus spectaculaires et les plus utiles de la relativité générale. C'est une technique précise pour mesurer la distribution de masse des amas de galaxies, étudier leur contenu en matière noire et leur profil de masse, et l'une des rares techniques qui permettent de mesurer la constante de Hubble H_0 avec une précision $<4\%$. Dans cette thèse, nous avons développé une approche basée sur l'apprentissage automatique pour rechercher des quasars lentillés à quatre images (quads) dans le catalogue Gaia DR3. En raison du nombre limité de quads connus à ce jour (environ 70 quads connus au 15/07/2023), il est essentiel de réaliser des simulations réalistes de lentilles gravitationnelles pour entraîner les algorithmes d'apprentissage automatique que j'ai employés.

Tout d'abord, nous détaillons la méthode que j'ai développée pour améliorer les simulations des déflecteurs en utilisant des distributions réalistes de paramètres. Cette méthode analyse la forme, l'alignement et la distribution de masse des halos de matière dans les simulations cosmologiques EAGLE.

J'ai ensuite utilisé les informations statistiques obtenues sur la forme de ces halos de matière noire et de matière baryonique pour réaliser des simulations de lentilles gravitationnelles à l'aide d'un modèle SIE+*shear* en explorant un large éventail de configurations possibles. Ces simulations ont été utilisées pour entraîner l'algorithme de machine learning XGBoost, basé sur le *gradient tree boosting*. En combinant les distributions de paramètres obtenues à partir de simulations EAGLE avec l'algorithme d'apprentissage XGBoost, nous avons obtenu huit modèles entraînés. Nous confrontons alors les résultats de cet apprentissage aux découvertes récentes de quads pour évaluer l'efficacité de cette approche. Nous avons confirmé l'efficacité de ces modèles en évaluant un échantillon de cas faux positifs issus de précédentes campagnes d'observation, montrant une réduction significative des scores des faux positifs. De plus, parmi un échantillon de 20 quads confirmés, 16 ont été correctement identifiés avec un score élevé, renforçant ainsi la fiabilité de notre approche. Les résultats obtenus soulignent l'amélioration significative apportée par notre méthode par rapport aux approches précédentes.

En exploitant les modèles entraînés sur les données d'EOLENS, le catalogue de multiplets issu de la chaîne d'analyse GravLens du Gaia Data Processing and Analysis Consortium (Gaia DPAC), nous avons établi une liste prometteuse de plusieurs dizaines de nouveaux candidats de quasars lentillés à quatre images. Ces candidats seront l'objet de futures campagnes d'observation pour confirmer ou infirmer leur nature.

Mots-clés : Lentilles gravitationnelles, Quasars, Galaxies, Intelligence artificielle, Machine learning

Artificial intelligence for gravitational lenses search in astronomical all-sky surveys

Abstract: The gravitational lensing effect is one of the most spectacular and useful consequences of general relativity. It is a precise technique for measuring the mass distribution of galaxy clusters, studying their dark matter content and mass profile. Moreover, it is one of the few methods allowing for the measurement of the Hubble constant, H_0 , with an accuracy of $<4\%$. In this thesis, we developed a machine learning-based approach to search for quadruply imaged lensed quasars (quads) in the Gaia DR3 catalog. Due to the limited number of known quads to date (approximately 70 known quads as of July 15, 2023), realistic gravitational lensing simulations are crucial to train the machine learning algorithms that I employed.

Firstly, we detail the method I developed to improve deflector simulations by using realistic distributions of parameters. This method analyzes the shape, alignment, and mass distribution of halos in the EAGLE cosmological simulations.

Subsequently, I utilized statistical information obtained from the shapes of these dark and baryonic matter halos to perform gravitational lensing simulations using an SIE+shear model, exploring a wide range of possible configurations. These simulations were used to train the XGBoost machine learning algorithm, based on gradient tree boosting. By combining parameter distributions obtained from EAGLE simulations with the XGBoost learning algorithm, we obtained eight trained models.

Next, we compared the results of this learning approach to recent quad discoveries to assess the effectiveness of our method. We confirmed the efficiency of these models by evaluating a sample of false positive cases from previous observation campaigns, which showed a significant reduction in false positive scores. Additionally, out of a sample of 20 confirmed quads, 16 were correctly identified with high scores, thus reinforcing the reliability of our approach. The obtained results highlight the significant improvement our method brings compared to previous approaches.

By applying the trained models to the EOENS data, the multiplet catalog from the GravLens analysis chain of the Gaia Data Processing and Analysis Consortium (Gaia DPAC), we established a promising list of several dozen new candidate quadruply imaged lensed quasars. These candidates will be the subject of future observation campaigns to confirm or refute their nature.

Keywords: Gravitational Lenses, Quasars, Galaxies, Artificial Intelligence, Machine Learning

Laboratoire d’Astrophysique de Bordeaux (LAB)

UMR 5804. Université de Bordeaux, 351 cours de la Libération CS10004 33405 Talence CEDEX

À Caroline

Remerciements

Je remercie vivement :

Christine Ducourant sans qui tout ce travail n'aurait pas vu le jour. Tes encouragements et ton soutien indéfectible tout au long de cette aventure scientifique ont été essentiels pour faire de cette thèse une réalité. Merci pour ta confiance renouvelée depuis maintenant six ans.

Eric Slezak, dont les précieux conseils scientifiques et techniques ont joué un rôle crucial dans la réussite de cette thèse. Ta présence bienveillante et tes éclairages ont été des atouts précieux qui ont enrichi cette expérience doctorale. Je te suis profondément reconnaissant pour ton accompagnement tout au long de ce parcours. Ensemble, Christine Ducourant et Eric Slezak ont formé une équipe de direction de thèse exemplaire, et je suis honoré d'avoir eu la chance de travailler sous leur guidance éclairée.

Geneviève Soucail et Roser Pelló pour avoir accepté d'être les rapporteurs de ce travail.

Dominique Sluse et Ludovic Delchambre pour leur accueil chaleureux à Liège. Leur précieuse aide et leurs conseils éclairés sur le monde captivant des lentilles gravitationnelles a contribué de manière significative à la qualité et à la rigueur de mes travaux.

Alexandre Boucaud, pour avoir accepté de faire partie de mon jury de thèse et pour avoir partagé une des semaines les plus enrichissantes de ce parcours scientifique.

Arnaud, Didier, Jeremy, Pierre, Nathalie, David et tous les membres de ce laboratoire pour avoir partagé vos connaissances et votre expertise avec générosité.

Monsieur Pascal Bordé, directeur du Laboratoire d'Astrophysique de Bordeaux, pour sa confiance et son accueil chaleureux.

Sur une note plus personnelle, je remercie également Rabah et Pierre pour tous les bons souvenirs qu'on partage et partagera encore ensemble. Je tiens à remercier tout particulièrement Mathieu. Depuis plus de dix ans, tu as été mon compagnon de route à travers d'innombrables aventures. Ce n'est que le début et je suis reconnaissant d'avoir un ami comme toi à mes côtés.

Enfin un grand merci à mes parents, vous m'avez toujours laissé libre de mes choix et soutenu dans mes études. Votre soutien inébranlable a été la clé de mon épanouissement personnel et académique. C'est grâce à votre amour et vos encouragements que j'ai pu accomplir ce parcours avec succès. Votre fierté est mon plus grand accomplissement et je vous adresse mes remerciements les plus sincères pour avoir été les piliers de mon succès et pour m'avoir permis de réaliser mes rêves.

Sommaire

I	Introduction	1
II	La recherche de quasars multi-imaginés par le phénomène de lentille gravitationnelle	7
II.1	État de l’art sur l’identification des lentilles gravitationnelles	7
II.1.1	Historique	7
II.1.1.1	La découverte des premiers arcs	9
II.1.1.2	Première détection d’un quasar multi-imaginé	9
II.1.2	Développement des techniques automatisées pour la recherche de lentilles dans de grands volumes de données	13
II.2	Gaia change la donne : un catalogue de positions et de flux	14
II.3	Le machine learning pour analyser Gaia	15
II.3.1	Le choix de l’apprentissage supervisé	17
II.3.2	Les algorithmes de classification	18
II.3.3	Le gradient tree boosting	21
II.3.3.1	Les arbres de décision	21
II.3.3.2	Les algorithmes à apprentissage d’ensemble	23
II.3.3.3	Le choix de XGBoost	26
III	Profil de masse des galaxies	29
III.1	Les différentes méthodes utilisées pour mesurer la distribution de la masse dans les galaxies	30
III.1.1	La courbe de rotation des galaxies	30
III.1.2	Les simulations hydrodynamiques d’univers	31
III.2	Mesure des halos de la simulation EAGLE	34
III.2.1	Sélection des halos	38
III.2.2	Mesurer les paramètres de forme des halos	39
III.2.2.1	Trouver la position du centre des halos	41
III.2.2.2	Calcul d’une sphère de travail	44
III.2.2.3	Ajustement d’un ellipsoïde	45

III.2.3	Bootstrap pour estimer les erreurs sur la morphologie	46
III.2.4	Courbe de croissance en masse des halos	50
III.2.5	Code et environnement de calcul	50
III.3	Distribution statistique des paramètres des halos d'EAGLE	53
III.3.1	Distribution en masse des halos dans les simulations de EAGLE	53
III.3.2	Morphologie des halos	53
III.3.2.1	La répartition de la masse des halos en 3D	54
III.3.2.2	La répartition de la masse des halos en 2D	58
III.3.3	Distribution de la masse et concentration	60
IV	Simuler des lentilles gravitationnelles réalistes	65
IV.1	Introduction aux simulations de lentilles gravitationnelles	65
IV.1.1	Résoudre l'équation de la lentille	66
IV.1.2	Calcul du rayon d'Einstein	68
IV.1.3	Amplification et cisaillement des images	70
IV.2	Choix d'un modèle de profil de densité	72
IV.3	Simulation des quads	80
IV.3.1	Des quasars comme source d'arrière plan	80
IV.3.2	Calcul des rayons d'Einstein de couples QSO/halos de EAGLE	80
IV.3.3	Tirage des positions des quasars	83
IV.3.4	Ajout d'un cisaillement externe ou <i>shear</i>	84
IV.3.5	Ajout des incertitudes sur la position des images	85
IV.4	Préparation des données pour l'apprentissage supervisé	86
IV.4.1	Alignement et normalisation de toutes les configurations	86
IV.4.2	Calcul des angles entre les images	88
IV.4.3	Sélection des sources qui ne sont pas des quads	91
V	Apprentissage supervisé pour la recherche de quads	93
V.1	Espace des paramètres pour l'entraînement supervisé	93
V.1.1	Entraînement de XGBoost sur plusieurs sous-catalogues	93
V.1.2	Résultat des entraînements	98
V.2	Comparaisons des scores avec des quads connus	102
VI	Application aux données de Gaia	109
VI.1	Identification de candidats lentilles dans le catalogue Gaia GravLens FPR	109
VI.1.1	Préselection des configurations à analyser dans Gaia	110
VI.1.2	Identification d'un nouveau critère de sélection	111
VI.1.3	Recherche de nouvelles lentilles gravitationnelles de quasars	112

VI.2 Liste de candidats	113
VII Bilan et perspectives	119
Annexes	124
Annexe A Article publié	125

Liste des Figures

II.1	Zoom sur la région centrale de l'amas Abell 370	10
II.2	Lentille gravitationnelle Q0957+561 et PG 1115+080 capturée par le télescope Pan-STARRS	11
II.3	Lentille gravitationnelle Q2237+0305, appelée "Croix d'Einstein", observé en 1990 par la Faint Object Camera (FOC) installé sur le télescope spatial Hubble	12
II.4	Nombre de quasars multi-imaginés par phénomène de lentille gravitationnelle au cours du temps	13
II.5	Exemple d'hyperplans séparateurs	19
II.6	Représentation schématique d'un arbre de décision	21
II.7	Illustration tirée du livre "Boosting" de Schapire et Freund	25
III.1	Projection du cube de la simulation EAGLE sur les axes XY au redshift $z = 0,5$	35
III.2	Projection bidimensionnelle du groupe de halos no. 25 de la simulation RefL0025N0376 au redshift $z=0,5$	39
III.3	Rapport de la masse des sous-halos sur la masse totale du halo en fonction de la masse totale des halos pour l'ensemble de la simulation RefL0025N0376	41
III.4	Projection bidimensionnelle du groupe de halos no. 17 de la simulation RefL0025N0376 au redshift $z = 0,5$	42
III.5	Zoom sur le centre du halo principal no. 25 de la simulation RefL0025N0376 au redshift $z = 0,5$ projeté dans les plans (XY) et (YZ)	43
III.6	Figure illustrant le processus de calcul de la sphère de travail pour le halo no. 20	45
III.7	Projections sur les axes principaux (X'Y') et (Y'Z') du groupe de halos no. 11 de la simulation de référence RefL0025N0376	47
III.8	Différence relative ($\Delta a, \Delta b, \Delta c$) des demi-axes récupérés (a,b,c) et différence des angles ($\psi_{sim}, \theta_{sim}, \phi_{sim}$) sur 10 000 halos simulés en fonction du nombre de particules	48
III.9	Différences relatives ($\Delta a, \Delta b, \Delta c$) et différence entre les angles récupérés (ψ, θ, ϕ) sur 10 000 halos simulés en fonction du rapport d'axe (b/a)	49
III.10	Courbes de croissance en masse de quatre halos extraits de la simulation EAGLE RefL0025N0376 à un redshift $z=0,5$	51

III.11	Plan schématique du programme développé pour étudier la forme, l'orientation et la distribution de masse des halos d'EAGLE	52
III.12	Distribution de masse de 681 438 halos EAGLE pour tous les redshifts et pour chaque type de particule	54
III.13	Distribution des rapports d'axes b/a et c/a selon le type de particule	55
III.14	Figure illustrant les rapports d'axes de l'ellipsoïde d'ajustement en fonction de la masse totale des halos	56
III.15	Figure illustrant les rapports d'axes de l'ellipsoïde d'ajustement en fonction du redshift des halos	58
III.16	Répartition des rapports d'axes des halos projetés dans trois directions selon le type de particule	59
III.17	Rapports d'axe \tilde{A} , \tilde{B} et \tilde{C}	61
III.18	Répartition du rapport entre le rayon de demi-masse de la matière noire et les halos de gaz et les halos de gaz et d'étoiles	62
III.19	Répartition des rayons à demi-masse pour chaque type de particules	63
IV.1	Schéma d'un système de lentille gravitationnelle typique	66
IV.2	Position et déformation des sources d'arrière-plan (à $z_s = 2.0$) au sein d'une lentille gravitationnelle (à $z_d = 0.3$)	73
IV.3	Simulation d'une lentille gravitationnelle avec un modèle SIS et une source ponctuelle (quasar)	76
IV.4	Cartographie du <i>cut</i> et de la caustique dans le plan de la source	77
IV.5	Simulations de lentilles gravitationnelles en fonction de l'ellipticité	79
IV.6	Distribution des magnitudes Gaia G en fonction du redshift des quasars du catalogue MILLIQUAS	81
IV.7	Distribution du cisaillement (shear) utilisé dans les simulations de quads	85
IV.8	Représentation des configurations simulées normalisées dans le plan "Norm. x_1 " et "Norm. x_2 "	87
IV.9	Représentation des trois possibilités d'angles centrées sur l'image A, BAC , BAD et DAC pour une configuration de quatre images	88
IV.10	Représentation sans cisaillement et avec cisaillement de la différence des angles $BAD - DAC$ en fonction des angles BAC	89
IV.11	Correspondance entre les zones dans le plan des angles et le plan des coordonnées normalisées	89
IV.12	Correspondance entre les zones des images C et des images D dans le plan "Norm. x_1 " et "Norm. x_2 "	90
IV.13	Correspondance et représentation de la position des images C et D des simulations avec du <i>shear</i> et des groupes d'objets de Gaia dans le plan "Norm. x_1 " et "Norm. x_2 "	92

V.1	Histogrammes de 16 des 41 paramètres les plus porteurs d'information dans le catalogue d'entraînement no.2	96
V.2	Importance des fonctionnalités dans le modèle XGBoost entraîné sur le catalogue no.2 : Shear	101
V.3	Histogramme des scores obtenus précédemment par l'algorithme Extremely Randomized Trees (ERT) par rapport aux scores obtenus dans notre étude avec XGBoost	103
V.4	Représentation des deux configurations qui ont obtenu un score élevé grâce à notre méthode utilisant XGBoost	104
V.5	Position des images ABCD des configurations Gaia EDR3 6126672995504821120 et Gaia EDR3 1321244403970596480 dans le plan normalisé "Norm.x ₁ " et "Norm.x ₂ " . . .	105
V.6	Meilleur score de probabilité P_{geo} en fonction du meilleur score de probabilité P_{dist} pour les lentilles gravitationnelles connues	106
V.7	Images des lentilles gravitationnelles connues qui ont un score P_{dist} et $P_{geo} < 0.4$. . .	107
VI.1	Meilleur score de probabilité géométrique P_{geo} en fonction du meilleur score parmi les probabilités incluant la distance P_{dist} pour les 1 893 multipléts sélectionnés de GravLens	113
VI.2	Images des candidats lentilles dans les régions denses en étoiles	114
VI.3	Exemple de candidats lentilles avec un score visuel A ou A+	117

Liste des tableaux

II.1	Algorithme de <i>boosting</i> AdaBoost	26
III.1	Médiane et écart absolu médian (MAD) des rapports d'axes b/a et c/a des 485 796 halos des simulations RefL0025N0376 et RefL0025N0752	55
III.2	Médiane et écart absolu médian des rapports d'axes projetés b_p/a_p , selon le type de particule	59
V.1	Ensemble des paramètres des catalogues d'entraînements	94
V.2	Liste des catalogues pour l'entraînement de XGBoost	95
V.3	Matrice de confusion du modèle d'entraînement d'XGBoost à partir du catalogue no.2.	99
V.4	Scores du modèle d'entraînement d'XGBoost à partir du catalogue no.2.	100
VI.1	Répartition des multiplets en fonction des cadrants 1, 2, 3 et 4 de la figure VI.1	112
VI.2	Liste des 59 candidats quads prometteurs identifiés dans le catalogue GravLens	115

Chapitre I

Introduction

Les lentilles gravitationnelles sont un phénomène lié à la courbure locale de l'espace-temps causée par la présence d'une masse. En termes simples, lorsque la lumière d'une source lointaine traverse une concentration de masse, la trajectoire de la lumière est déviée, créant ainsi une image déformée et amplifiée de la source. Ce phénomène a été prédit pour la première fois par Albert Einstein en 1915 dans le cadre de sa théorie de la relativité générale et confirmé en 1919 par A.Eddington qui photographie la déviation des rayons lumineux à proximité du Soleil. La première observation directe de lentilles gravitationnelles n'a cependant été faite qu'en 1979 par l'astronome britannique Dennis Walsh et ses collaborateurs du Kitt Peak Observatory. Ces astronomes observent ainsi deux images d'un quasar baptisé Q0957+561A-B. Les deux objets, séparés de six secondes d'arc, sont de magnitude 16,87 et 17,11 dans la bande G et présentent rigoureusement le même spectre, avec un décalage identique vers le rouge de 1,407. Ce sont donc deux images de la même source.

Il existe plusieurs types de lentilles gravitationnelles qui peuvent être classées en fonction de leurs caractéristiques et de leur effet sur la lumière. Les lentilles gravitationnelles fortes sont causées par des masses importantes, appelées déflecteurs, et peuvent entraîner la formation d'images multiples de la source d'arrière-plan. Ces images peuvent prendre la forme d'arcs, de demi-arcs ou même de structures plus complexes. Par exemple, lorsqu'une source est fortement déformée par un déflecteur massif, elle peut apparaître sous la forme d'une croix résultant du dédoublement ou du quadruplement de son image. Les lentilles gravitationnelles faibles sont les cas où la déviation de la lumière est faible par rapport à celle des lentilles gravitationnelles fortes et nécessite une analyse statistique sur de grands échantillons pour être détectée. Elles sont causées par l'accumulation de la matière tout au long de la ligne de visée entre la source lumineuse et l'observateur, à l'opposé d'un objet massif localisé.

Depuis une vingtaine d'années, les lentilles gravitationnelles jouent un rôle crucial dans l'étude de la distribution de la matière noire, une forme de matière invisible qui n'émet aucun rayonnement

électromagnétique (Dalal and Kochanek, 2002; Gilman et al., 2018). Bien que nous ne puissions pas observer directement la matière noire, son existence et sa répartition peuvent être déduites indirectement à partir des effets gravitationnels qu'elle induit, notamment à travers le phénomène de lentille gravitationnelle.

La constante de Hubble-Lemaître, notée H_0 , est une grandeur fondamentale en cosmologie qui mesure la vitesse d'expansion de l'univers aujourd'hui. Les mesures de cette constante, réalisées par différentes méthodes, ont conduit à des résultats difficiles à concilier (Vanderriest et al., 1989; Schechter et al., 1997; Kochanek et al., 2006; Wong et al., 2020). La méthode directe, mise en œuvre par le programme SH0ES, aboutit à une valeur de $H_0 = 74,03 \pm 1,42 \text{ km.s}^{-1}\text{Mpc}^{-1}$ (Riess et al., 2019). L'équipe SH0ES procède à l'observation d'étoiles variables et de supernovae, en mesurant précisément la distance qui les sépare de la Terre. Par la suite, ils déterminent la vitesse d'éloignement de ces objets en utilisant le décalage vers le rouge des raies spectrales (redshift), ce qui leur permet ainsi de déduire la valeur de la constante de Hubble. Une autre méthode indirecte pour mesurer la constante de Hubble utilise les données du satellite Planck. Au lieu de mesurer directement la vitesse d'expansion de l'univers à partir des distances et des vitesses des galaxies, cette méthode se base sur les mesures précises des fluctuations de température dans le fond diffus cosmologique (CMB) recueillies par Planck. Les fluctuations du CMB sont influencées par plusieurs paramètres cosmologiques, dont la constante de Hubble H_0 . En analysant les données du CMB et en effectuant des ajustements statistiques, les équipes travaillant sur Planck ont pu déduire une valeur pour H_0 de $67,4 \pm 0,5 \text{ km.s}^{-1}\text{Mpc}^{-1}$ (Planck Collaboration, 2020).

Une troisième méthode consiste à étudier une lentille gravitationnelle forte d'un objet dont la luminosité varie au cours du temps et à exploiter les caractéristiques des images multiples ainsi formées. Lorsque la source d'arrière-plan présente une variabilité intrinsèque, cette variabilité se manifeste dans chacune des images multiples, mais avec un décalage temporel dû aux différents chemins parcourus par la lumière (appelés *time delays* en anglais). Ce décalage temporel, appelé *temporisation* (Δt), dépend à la fois de la distance de retard $D_{\Delta t}$ (*time-delay distance* en anglais) et de la distribution de masse de la lentille. Plus précisément, on a $\Delta t = D_{\Delta t} \Delta \phi / c$, où $\Delta \phi$ représente la différence de potentiel de Fermat, qui est déterminée par la répartition de masse de la lentille, et c correspond à la vitesse de la lumière. Ainsi, en mesurant le retard à partir des courbes de lumière photométrique des images de quasar et en modélisant la distribution de masse de la lentille, on peut déterminer la distance de retard du système de lentille et utiliser la relation distance-redshift pour contraindre les modèles cosmologiques. Plus précisément, la distance de retard est :

$$D_{\Delta t} \equiv (1 + z_L) \frac{D_L D_S}{D_{LS}} \quad (\text{I.1})$$

où z_L est le redshift de la galaxie défectrice, D_L est la distance au déflecteur, D_S est la distance

angulaire de la source et D_{LS} est la distance angulaire entre le déflecteur et la source.

Les galaxies à noyau actif (AGN pour "Active Galaxy Nucleus" en anglais) ont une luminosité variable, ce qui permet de les utiliser pour estimer la constante de Hubble-Lemaître, H_0 . Un AGN fait référence à une région compacte et très lumineuse au centre d'une galaxie. Les AGN sont alimentés par l'accrétion de matière sur le trou noir supermassif présent dans le noyau galactique. Les AGN peuvent être classés en différents types en fonction de leurs propriétés observées, et l'une des classifications les divise en AGN de type I et de type II.

Les AGN de Type I se caractérisent par la présence de larges raies d'émission dans leur spectre. Ces larges raies d'émission proviennent de nuages de gaz à proximité du trou noir supermassif qui se déplacent à des vitesses élevées. L'élargissement des raies d'émission est le résultat de l'effet Doppler provoqué par le mouvement rapide du gaz. Les AGN de type I présentent généralement d'autres caractéristiques observationnelles telles qu'une émission continue optique et UV forte et variable, de larges lignes d'émission autorisées (comme la série de Balmer de l'hydrogène) et, dans certains cas, un spectre en loi de puissance dans les rayons X. Ces propriétés suggèrent que la région d'émission est directement visible pour l'observateur, ce qui nous permet de voir les larges raies d'émission et l'émission continue provenant du disque d'accrétion autour du trou noir.

Des exemples d'AGN de type I incluent les quasars, les galaxies Seyfert (Seyfert Type 1) et certains blazars. Ces objets sont généralement très lumineux et peuvent émettre d'énormes quantités d'énergie à travers le spectre électromagnétique, éclipsant souvent toute la galaxie dans laquelle ils résident. Il convient de noter que la classification en AGN de type I et de type II est basée sur la présence ou l'absence de larges raies d'émission dans le spectre, qui se produit en raison de différents angles de vision et de l'obscurcissement des régions centrales par le gaz et la poussière environnants dans les AGN de type II.

Les AGN ont été découverts pour la première fois au début du XXe siècle, lorsque des astronomes ont détecté des raies d'émission très intenses dans le spectre de certaines galaxies, telles que NGC 1068 et NGC 4151. Ces galaxies ont ensuite été cataloguées par Karl Seyfert en 1943, qui a démontré que les noyaux non résolus spatialement de ces galaxies avaient des dimensions très petites, de l'ordre de 1 à 100 pc. Le premier quasar, 3C 273, a été découvert en 1962 par Maarten Schmidt (Schmidt, 1963). Il s'agissait d'un objet détecté dans le domaine radio d'apparence stellaire (donc non résolu angulairement) en optique. Son apparence stellaire fut à l'origine du nom donné à ce type d'objet "quasi stellar object - quasar", en abrégé QSO. Ce sont des objets extrêmement lumineux et éloignés avec un redshift généralement compris entre 0,1 et 7,5. Ainsi, l'objet 3C 273 a une magnitude apparente de +12.9 mag pour un redshift de $z=0.158$. Si on calcule sa magnitude absolue à partir de la magnitude apparente et de la distance ainsi estimée, on obtient -26.7 mag, ce qui est environ 10 fois plus brillant que la galaxie la plus brillante jamais observée à l'époque de sa découverte.

Lorsqu'un quasar est dans l'alignement d'une galaxie massive, il peut être observé sous la forme de quatre images déformées autour de la lentille, formant ainsi un quadruplet (également appelé quad). En mesurant les variations de luminosité dans les séries temporelles photométriques des images du quasar, il devient alors possible d'estimer les décalages spectraux attendus en fonction des paramètres de la lentille, y compris la constante de Hubble. Plusieurs études ont déjà utilisé cette méthode pour estimer la constante de Hubble. Par exemple, le programme H0LiCOW (H0 Lenses in COSMOGRAIL's Wellspring) a analysé des quads lentillés et a obtenu une estimation de la constante de Hubble avec une valeur de $H_0 = 73,3 \pm 1,7$ km/s/Mpc (Wong et al., 2020). Ce résultat renforce les résultats obtenus par la collaboration SH0ES. Il convient également de souligner que l'utilisation des sirènes gravitationnelles, qui sont des événements astrophysiques associés aux fusions de trous noirs ou d'étoiles à neutrons, constitue une alternative prometteuse pour la mesure de la constante de Hubble, comme l'illustre la récente mesure réalisée par LIGO Scientific Collaboration and Virgo Collaboration (2021) avec une valeur de $H_0 = 69 \pm 16$ km/s/Mpc.

Cette dernière décennie, la recherche de nouvelles lentilles gravitationnelles s'est tournée vers l'utilisation de techniques de *machine learning* afin de détecter de nouveaux candidats. Cette approche se concentre principalement sur l'analyse des images provenant de vastes relevés astronomiques tels que le SDSS et DESI réalisés à l'aide de télescopes terrestres (Abazajian et al., 2009; DESI Collaboration, 2023). Le nombre de lentilles connues est extrêmement limité, avec environ 70 lentilles composées de plus de deux images identifiées à ce jour (Ducourant et al., 2018; Stern et al., 2021; Lemon et al., 2023). En raison de la rareté de ces phénomènes, il est nécessaire de couvrir une grande partie du ciel pour obtenir un échantillon important de lentilles gravitationnelles. De plus, la plupart des lentilles connues présentent une séparation angulaire relativement importante, de l'ordre de 1 à 3 secondes d'arc entre leurs différentes images. Cependant, il est probable que de nombreuses configurations de lentilles soient plus compactes, avec des séparations angulaires inférieures à 1 seconde d'arc. En raison des limitations imposées par l'atmosphère terrestre, les observations effectuées depuis le sol ne peuvent atteindre cette résolution. Dans ce contexte, les observations spatiales offrent une couverture homogène de l'ensemble du ciel, permettant ainsi de collecter des informations sur un plus grand nombre de régions.

Deux télescopes spatiaux se révèlent particulièrement intéressants pour la recherche de lentilles gravitationnelles : Gaia et Euclid (Gaia Collaboration, 2016; Euclid Collaboration, 2019). Euclid, lancé en juillet 2023, est un satellite dont l'objectif principal est d'étudier l'énergie noire. En amont, nous avons le satellite Gaia. Lancée le 19 décembre 2013, Gaia est une mission scientifique majeure de l'ESA dont l'objectif est de cartographier l'ensemble du ciel en trois dimensions et de créer le catalogue le plus précis jamais réalisé des étoiles de notre galaxie. Grâce à ces mesures, Gaia peut déterminer avec une précision exceptionnelle les positions, les mouvements, les parallaxes et les caractéristiques physiques des étoiles. Le satellite est capable de mesurer les positions des sources

avec une précision de quelques dizaines de microsecondes d'arc. Ainsi, en plus de couvrir l'ensemble de la sphère céleste, Gaia possède une résolution angulaire suffisante pour mesurer avec précision la position des images des lentilles gravitationnelles compactes.

Il est cependant difficile de repérer des configurations de lentilles gravitationnelles parmi près de deux milliards de sources présentes dans le catalogue Gaia (Gaia Collaboration, 2021). Afin de relever ce défi, l'utilisation d'algorithmes de machine learning s'avère essentielle pour apprendre à identifier les configurations de lentilles gravitationnelles de façon automatique. Une étude récente menée par Delchambre et al. (2019) s'est intéressée à cette problématique. Dans cette étude, un algorithme d'apprentissage supervisé, basé sur les arbres extrêmement aléatoires (ERT), a été entraîné à détecter les candidats lentilles dans les données de Gaia. Cependant, les simulations de lentilles gravitationnelles qui ont été réalisées pour entraîner ces ERT présentent des lacunes en raison d'une distribution uniforme des paramètres du modèle $NSIE+shear$ utilisé pour les simulations de lentilles gravitationnelles de quasars.

L'objectif de cette thèse s'inscrit dans la continuité des travaux précédents. Elle vise à améliorer les simulations de lentilles gravitationnelles de quasars en utilisant des distributions de paramètres réalistes pour les galaxies défectrices. Ce manuscrit présente les résultats de mes recherches, organisées en cinq parties distinctes.

Tout d'abord, la première partie se concentre sur une analyse approfondie des différentes approches utilisées dans la découverte et la recherche de lentilles gravitationnelles de quasars. Cette section détaille les méthodes et les techniques mises en œuvre pour détecter ces phénomènes.

Ensuite, la deuxième partie propose une étude détaillée de la distribution de masse des galaxies. À travers cette analyse, nous examinons en détail la forme, l'alignement et la répartition de la masse des halos de matière noire et de matière baryonique dans les simulations cosmologiques EAGLE (Evolution and Assembly of GaLaxies and their Environments). L'objectif est de mesurer les propriétés statistiques de la distribution de masse des galaxies afin de les utiliser pour la simulation des lentilles gravitationnelles réalistes.

Dans la troisième partie, nous nous penchons sur le processus complet de création de simulations réalistes de quasars lentillés à quatre images, en utilisant un modèle $SIE+shear$ pour les simulations de lentilles gravitationnelles de quasars. Cette étape permet de générer un catalogue réaliste de simulations de lentilles gravitationnelles, qui servira à améliorer l'entraînement des algorithmes d'apprentissage automatique.

Dans la quatrième partie, nous explorons l'utilisation de ces simulations réalistes pour améliorer l'entraînement des algorithmes d'apprentissage automatique. Nous mettons en œuvre différentes approches et évaluons les performances des modèles d'apprentissage sur ces données réalistes.

Enfin, dans la dernière partie, nous présentons les résultats de la recherche de nouveaux candidats dans le catalogue de Gaia, en exploitant les connaissances acquises grâce à l'apprentissage des configurations de quasars lentillés par les algorithmes d'apprentissage automatique. Cette section met en évidence les découvertes prometteuses réalisées dans notre recherche de nouveaux candidats.

———— Chapitre II ————

La recherche de quasars multi-imaginés par le phénomène de lentille gravitationnelle

Dans ce premier chapitre, je commencerai par un état de l'art sur l'identification des lentilles gravitationnelles, mettant en lumière les développements clés de cette discipline. Je retracerai l'histoire de cette exploration, depuis ses débuts jusqu'aux avancées récentes. En particulier, je me pencherai sur le développement des techniques automatisées qui ont permis d'élargir le champ des découvertes. J'aborderai également l'impact de la mission Gaia, qui fournit un catalogue précis des positions et des flux, ouvrant ainsi de nouvelles possibilités pour l'analyse des lentilles gravitationnelles. De plus, j'explorerai le rôle crucial de l'apprentissage automatique dans l'analyse des données Gaia, en me concentrant sur le choix de l'apprentissage supervisé et les algorithmes de classification utilisés. Enfin, je présenterai en détail le *gradient tree boosting*, une méthode d'apprentissage puissante qui sera utilisée dans le cadre de cette thèse.

II.1 État de l'art sur l'identification des lentilles gravitationnelles

II.1.1 Historique

La première mention de l'effet de lentille gravitationnelle est attribuée à Isaac Newton dans son ouvrage "Opticks" publié en 1704. Dans cet ouvrage, Newton envisage la possibilité que les corps puissent agir à distance sur la lumière et courber ses rayons. Plus tard, Johann Georg von Soldner reprend cette idée en utilisant la loi newtonienne de la gravité. Partant de l'hypothèse que les rayons lumineux sont constitués de particules de faible masse, il calcule la déviation d'un rayon lumineux émis tangentiellement à la surface de la Terre (Soldner, 1804). À partir de ce calcul, Soldner

détermine l'angle de déviation, noté $\hat{\alpha}$, d'un rayon lumineux provenant d'une étoile lointaine et passant près de la surface du Soleil. Cet angle de déviation est donné par l'expression suivante :

$$\hat{\alpha} = \frac{2GM}{bc^2} \quad (\text{II.1})$$

où G est la constante newtonienne de gravitation, c la vitesse de la lumière, M la masse du Soleil (ou, plus généralement, la masse du déflecteur), et b le paramètre d'impact du rayon lumineux incident, c'est-à-dire la plus petite distance entre le centre du Soleil et le rayon lumineux incident si celui-ci n'avait pas été dévié.

Plus tard, Albert Einstein utilisa en 1911 le principe d'équivalence pour calculer l'angle de déviation et dérive la formule de Soldner. En 1915, il appliqua les équations de la Relativité Générale en champ complet et découvrit que l'angle de déviation était en réalité le double de sa prévision précédente, ce facteur deux étant dû à la courbure de la métrique. La formule d'Einstein pour l'angle de déviation, dérivée de la Relativité Générale, est la suivante :

$$\hat{\alpha} = \frac{4GM}{c^2b} \quad (\text{II.2})$$

La prévision d'Albert Einstein concernant la déviation de la lumière d'une étoile lorsqu'elle passe près de la surface du Soleil, avec un angle de déviation de $1,74''$, ouvrit la possibilité de vérifier expérimentalement cette déviation. Étant donné que la lumière du Soleil rendait impossible l'observation directe d'une étoile à proximité, cette observation devait être effectuée pendant une éclipse solaire totale. En 1919, deux expéditions organisées pour observer l'éclipse solaire totale de Sobral et de Principe ont rapporté avec succès l'observation de la déviation d'une étoile en arrière-plan (Dyson et al., 1920) avec comme valeur celle donnée par le calcul d'Albert Einstein. Cette observation historique a marqué la première confirmation directe de l'effet de lentille gravitationnelle et la validation de la relativité générale comme théorie de la gravitation.

Albert Einstein, en 1936, écrivit un article (Einstein, 1936) dans lequel il calcula l'effet de lentille par une étoile, y compris la position des images, leur séparation et leurs amplifications. Il conclut que la séparation angulaire entre les deux images serait beaucoup trop petite (de l'ordre de la milli-seconde d'arc) pour être résolue, indiquant qu'il y a peu de chances d'observer ce phénomène. En 1937, Fritz Zwicky met en évidence le potentiel des galaxies pour former des images multiples de sources en arrière-plan, avec des angles suffisamment grands pour être observés (Zwicky, 1937). Cette observation a élargi le champ des recherches sur les lentilles gravitationnelles en incluant les galaxies comme des déflecteurs potentiellement puissants.

II.1.1.1 La découverte des premiers arcs

Dans les années 1970, de nombreuses applications des lentilles gravitationnelles ont été proposées, mais elles sont restées des spéculations théoriques sans confirmation observationnelle. La principale raison était le manque d'outils adaptés pour observer les phénomènes prédits. La détection d'événements de lentilles gravitationnelles fortes nécessite un alignement exceptionnel entre l'observateur, la lentille et la source. Ces conditions sont rarement remplies, ce qui nécessite d'observer de vastes portions du ciel.

Lorsque la source d'arrière-plan est plus étendue, telle qu'une galaxie, les différents éléments de son image peuvent se déformer et fusionner, créant ainsi des structures allongées connues sous le nom d'arcs gravitationnels. Les premiers arcs gravitationnels ont été observés par deux équipes indépendantes en 1986 et 1987 (Lynds and Petrosian, 1986; Soucaïl et al., 1987). Ces observations initiales ne permettaient pas de conclure avec certitude qu'il s'agissait d'arcs gravitationnels. Ce n'est qu'après la mesure du décalage spectral de l'arc dans l'amas Abell 370 par Soucaïl et al. (1988) qu'il a pu être démontré que cet arc, avec un décalage spectral de $z = 0,724$, ne faisait pas partie de l'amas de galaxies avec un décalage spectral de $z = 0,37$. Cette mesure a ainsi confirmé qu'il s'agissait du premier exemple d'arc gravitationnel.

Les arcs gravitationnels constituent un outil puissant pour contraindre la distribution de masse des amas de galaxies. Un exemple notable de modélisation de masse d'amas est celui réalisé pour l'amas Abell 370 par Kneib et al. (1993). Leur modèle considère deux profils de masse centrés sur les deux galaxies elliptiques géantes de cet amas. Les contraintes du modèle ont été fixées par l'arc géant (une triple image) et une autre triple image, B2-B3-B4 sur la figure II.1. Celle-ci montre un zoom sur une partie de l'amas Abell 370. Depuis lors, d'autres avancées significatives ont été réalisées dans la modélisation des amas de galaxies, notamment dans le cadre d'une exploration de l'espace des paramètres par une approche bayésienne (Jullo et al., 2007) et du programme observationnel Hubble Frontier Fields (HFF) visant la mise en évidence des premières galaxies (Coe et al., 2015; Lotz et al., 2017; Mahler et al., 2018). Avec notamment des observations sur six amas massifs avec lentilles gravitationnelles, dont l'amas Abell 370, le programme HFF a réussi à sonder les galaxies à des profondeurs atteignant des magnitudes optiques/infrarouges de 29 ABmag.

II.1.1.2 Première détection d'un quasar multi-imagé

C'est grâce à l'étude menée par Walsh et al. (1979) que la première observation d'un quasar lentillé fut réalisée. Les chercheurs observèrent deux sources d'apparence quasi-stellaires (qso), nommées Q0957+561 et illustrées à la figure II.2 (panneau gauche), dans une image beaucoup plus récente que l'observation originale de Walsh. Les deux sources présentent une magnitude de 16,87 et 17,11 dans la bande G et sont séparées par une distance angulaire de 5,7 secondes d'arc. Ces deux sources ont une couleur et un spectre similaires, suggérant qu'il s'agissait de deux images provenant d'une

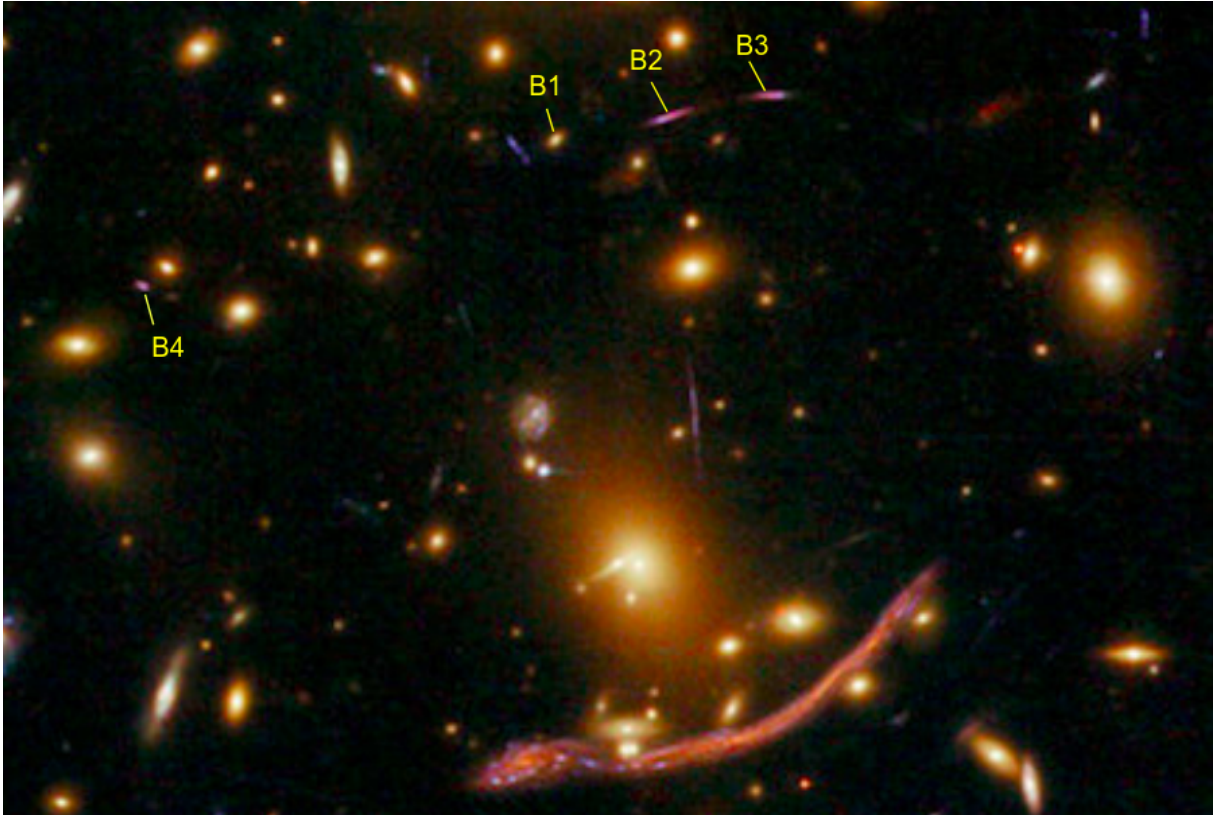


FIGURE II.1 – Zoom sur la région centrale de l’amas Abell 370. B2-B3 sont des images triangulaires typiques prévues dans le cas de déflecteurs multi-composants. B1 ne doit pas être la troisième image attendue, qui est prédite en B4 par leur modèle.

même source, observées à travers une lentille gravitationnelle. Cette découverte marque la première observation d’un quasar lentillé, où une seule source lumineuse apparaît sous la forme de deux images distinctes. Ce phénomène, connu sous le nom d’images multiples, est une manifestation caractéristique des lentilles gravitationnelles.

Un an plus tard, l’équipe dirigée par [Weymann et al. \(1980\)](#) (voir le panneau de droite de la figure [II.2](#)) a découvert le premier "triple quasar" connu sous le nom de PG 1115+080. Ce système se compose de trois images distinctes d’une même source lointaine, révélant ainsi la présence d’une lentille gravitationnelle qui dévie la lumière provenant du quasar. Un modèle théorique développé par [Young et al. \(1981\)](#) suggère que les images observées du système PG 1115+080 sont en réalité constituées d’au moins cinq composantes, formées par un déflecteur étendu telle qu’une galaxie spirale. Initialement considérée comme une configuration de trois images, des recherches approfondies ont révélé que le système PG 1115+080 est en réalité composé d’au moins quatre images, dont deux étaient indiscernables depuis le sol en raison de leur faible séparation angulaire. Cela a rendu leur identification extrêmement difficile avec les télescopes optiques terrestres. La composante la plus

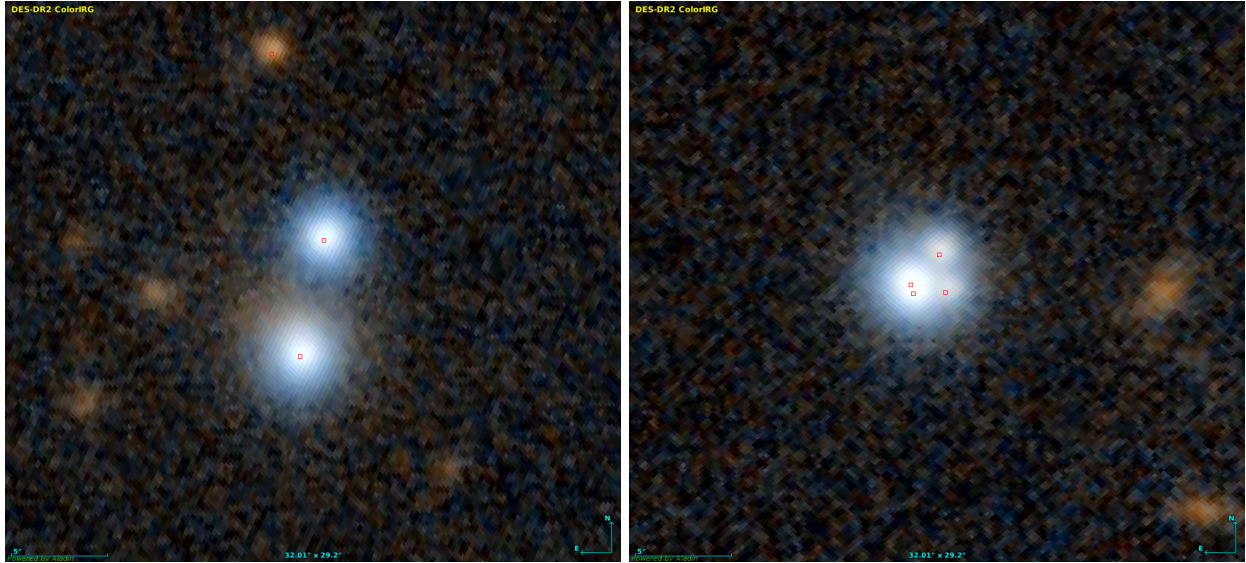


FIGURE II.2 – Dans le premier panneau nous avons une image du système de lentille gravitationnelle Q0957+561 capturée par le télescope Pan-STARRS. Le quasar présent dans cette image présente un décalage spectral de $z_s = 1.41$. Dans le panneau de droite, nous avons une image du système de lentille gravitationnelle PG 1115+080 également obtenue par le télescope Pan-STARRS. Le quasar à l’origine de ces images multiples possède quant à lui un décalage spectral de $z_s = 1.72$. Les carrés rouges sont les positions des sources mesurées par le satellite Gaia.

brillante, désignée sous le nom de Q1115+080A, correspond en réalité à une paire d’images séparées par 0,5 seconde d’arc. Ce système de lentille gravitationnelle a été étudié dans le cadre de la première série d’observations de lentilles gravitationnelles réalisées avec la caméra Wide Field/Planetary (WFPC) du Hubble Space Telescope (HST) (Kristian et al., 1993). Ces observations ont permis de résoudre le système en quatre sources ponctuelles et en un objet étendu rouge, vraisemblablement la galaxie lentille. À l’heure actuelle, les dernières observations réalisées depuis l’espace par le satellite Gaia ont permis de résoudre spatialement les quatre images, identifiables par les carrés rouges sur la figure II.2.

Une méthode couramment utilisée pour détecter les lentilles gravitationnelles consiste à analyser les relevés de décalages spectraux (redshift surveys). Ces relevés sont des catalogues de spectres des galaxies permettant ainsi de détecter des caractéristiques spectrales distinctes qui pourraient indiquer la présence d’une lentille gravitationnelle. Selon des estimations théoriques (Kochanek 1992; Mortlock and Webster, 2000), on estime que le taux de découverte de lentilles gravitationnelles devrait être d’environ un pour chaque 10^4 à 10^5 spectres. C’est grâce à cette approche que la lentille Q2237+0305 a été identifiée dans le relevé de décalages spectraux CfA (Huchra et al., 1985), tout comme la lentille SDSS0903+5028 dans le relevé du Sloan Digital Sky Survey (SDSS) (Johnston et al., 2003). La lentille Q2237+0305, illustrée sur la figure II.3, fait partie des configurations particulières

appelées "croix d'Einstein". Dans cette configuration, la lentille et la source sont presque parfaitement alignées, ce qui dispose les images secondaires en une croix caractéristique.

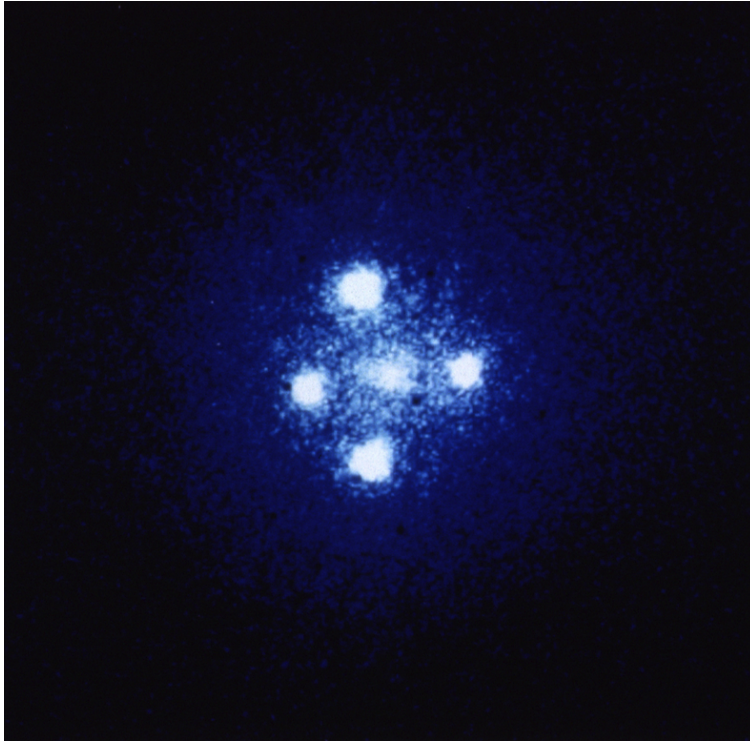


FIGURE II.3 – Lentille gravitationnelle Q2237+0305, appelée "Croix d'Einstein", observé en 1990 par la Faint Object Camera (FOC) installé sur le télescope spatial Hubble. On distingue les quatre images d'un quasar très éloigné ($z_s = 1.695$) vu au travers d'une galaxie proche ($z_l = 0.0394$) (point central). La séparation angulaire entre les images supérieure et inférieure est de 1,6 seconde d'arc.

Avec l'avènement de nouveaux projets d'observations et l'automatisation des techniques d'identification des lentilles, alimentés par des relevés optiques à grande échelle tels que le SDSS (Abazajian et al., 2009) et le Dark Energy Survey (DES) (Dark Energy Survey Collaboration, 2016), le paysage a évolué. Des efforts ciblés, tel que le Sloan Bright Arcs Survey (Diehl et al., 2009; Kubo et al., 2009; Wiesner et al., 2012), ont également contribué de manière significative à la recherche spécifique d'arcs gravitationnels, marquant une avancée dans la méthodologie de découverte. Cependant, c'est particulièrement avec le déploiement du satellite Gaia que la multiplication des découvertes s'est accélérée, comme illustré dans la figure II.4, présentant l'évolution du nombre de quasars multi-imagés au fil du temps (Ducourant et al., 2018; Stern et al., 2021).

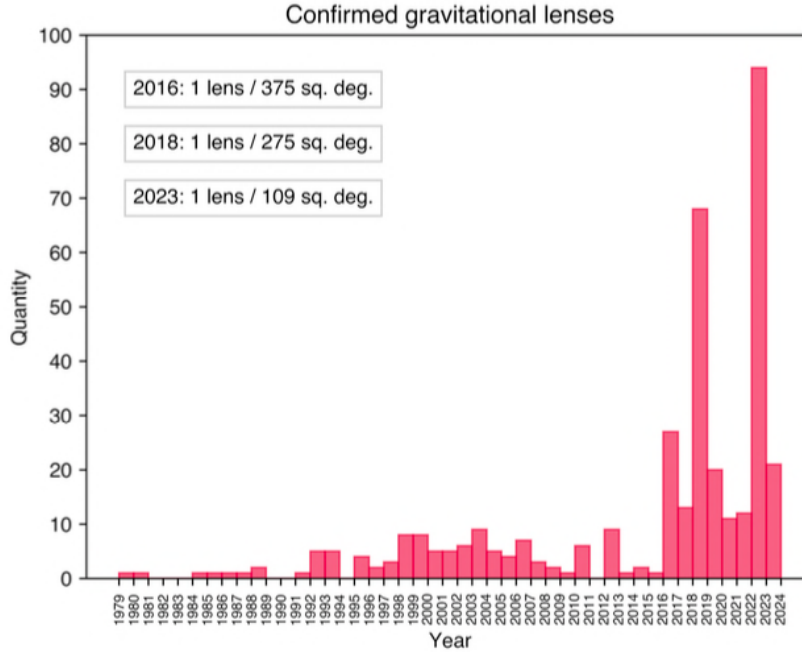


FIGURE II.4 – Nombre de quasars multi-imagés par phénomène de lentille gravitationnelle au cours du temps. La date 2016 correspond à la première publication du catalogue Gaia DR1.

II.1.2 Développement des techniques automatisées pour la recherche de lentilles dans de grands volumes de données

Les premières méthodes automatisées pour la recherche de lentilles gravitationnelles se sont principalement concentrées sur la détection des structures en forme d’arcs et d’anneaux. Par exemple, le Strong Lensing Legacy Survey-ARCS (SARCS) (More et al., 2012), basé sur l’analyse des données du Canada-France-Hawaii Telescope Legacy Survey (CFHTLS), utilise un estimateur local d’allongement au niveau des pixels pour détecter les arcs gravitationnels. Une autre approche morphologique, proposée par Seidel and Bartelmann (2007), se concentre sur la détection de caractéristiques fines telles que les lignes et les courbes dans les images astronomiques. Une méthode alternative développée par Bom et al. (2017) utilise la méthode de filamentation Mediatrix, qui décompose les objets allongés en segments le long de leur ligne de crête d’intensité.

Ces dernières années ont été marquées par de nombreux grands relevés imageant tout ou partie du ciel, tels que le Two Micron All Sky Survey (2MASS), le Catalina Real-Time Survey (CRTS), le Wide-field Infrared Survey Explorer (WISE), le Dark Energy Survey (DES) et le Panoramic Survey Telescope and Rapid Response System (Pan-STARRS). La taille considérable des catalogues associés a motivé l’utilisation de techniques d’analyse plus automatisées encore, notamment l’utilisation de réseaux de neurones artificiels (ANN) et plus spécifiquement, de réseaux de neurones convolutifs (CNN), pour

la détection de nouvelles lentilles gravitationnelles. Les CNNs se révèlent particulièrement pertinents pour traiter la variabilité spatiale présente dans ces données astronomiques. Leur capacité à gérer l'invariance par translation et à s'adapter à des échelles multiples offre une approche puissante pour la détection automatique de lentilles gravitationnelles, qui, en raison de leur diversité et de leur complexité, bénéficient grandement d'une approche pilotée par les données capable d'apprendre et de reconnaître des motifs complexes au sein de vastes ensembles de données labélisées.

Par exemple, l'étude de [Cañameras et al. \(2020\)](#) sur Pan-STARRS 3π utilise un algorithme CNN. Le principe de cette méthode repose sur l'utilisation d'images découpées provenant du relevé Pan-STARRS, sur lesquelles des lentilles gravitationnelles simulées ont été ajoutées. En entraînant le réseau CNN à partir de ces images contenant des lentilles artificielles, il devient capable d'identifier automatiquement dans ces portions d'images les régions associées à de nouveaux candidats de lentilles gravitationnelles.

Cependant, tous les grands relevés mentionnés jusqu'à présent ont été ou sont réalisés depuis le sol, ce qui limite les études à la recherche de configurations présentant une large séparation angulaire, généralement de l'ordre de 1,5 seconde d'arc, entre deux images. Afin d'identifier les lentilles gravitationnelles compactes, avec une séparation angulaire inférieure à 1 seconde d'arc, la solution est de se tourner vers des observations spatiales. Le télescope spatial Gaia, lancé en 2013 par l'Agence spatiale européenne (ESA), offre une résolution spatiale et une sensibilité exceptionnelle (0.18"), permettant ainsi une recherche plus précise et complète des lentilles gravitationnelles de petite séparation angulaire. L'utilisation de ces observations spatiales complète les recherches réalisées au sol.

II.2 Gaia change la donne : un catalogue de positions et de flux

Le satellite d'observation Gaia, lancé en 2013 par l'Agence spatiale européenne (ESA), a fourni le relevé astrométrique depuis l'espace le plus vaste et le plus précis jamais réalisé. Depuis le début de sa mission en 2015, Gaia effectue un relevé complet du ciel, mesurant les positions (α , δ), les parallaxes et les mouvements propres des étoiles, ainsi que les magnitudes G , G_{BP} et G_{RP} de près de 2 milliards de sources stellaires. L'une des capacités les plus impressionnantes de Gaia est sa précision astrométrique exceptionnelle, capable de mesurer les positions relatives des sources ponctuelles avec une précision de quelques microsecondes d'arc ([Gaia Collaboration, 2016](#)).

Gaia a publié plusieurs remises de données (Data Release, DR) contenant des informations sur les étoiles et autres objets célestes observés. La première DR, Gaia DR1, a été publiée en septembre 2016 et comprenait les positions et les magnitudes de plus d'un milliard d'étoiles sur tout le ciel ([Gaia Collaboration, 2016](#)). En avril 2018, la Gaia DR2 a été publiée, fournissant des mesures astrométriques et photométriques plus précises pour environ 1,7 milliard d'étoiles, ainsi que des

paramètres astrométriques et photométriques pour des milliers d’objets du Système Solaire, des étoiles variables, des quasars et d’autres sources (Gaia Collaboration, 2018).

La dernière Data Release, Gaia Data Release 3 (DR3) (Gaia Collaboration, 2023) a été publiée en 2023. La Gaia DR3 contient des mesures astrométriques, photométriques et spectroscopiques améliorées pour un plus grand nombre de sources. Cette version comprend également plusieurs nouveaux catalogues, notamment des catalogues dédiés aux galaxies et aux quasars (Gaia Collaboration, 2023; Ducourant et al., 2023).

Dans le domaine de la recherche de lentilles gravitationnelles, Gaia joue un rôle majeur. La méthode appliquée aux données Gaia pour détecter les lentilles gravitationnelles telle que développée par Delchambre et al. (2019) se base sur des algorithmes d’apprentissage automatique supervisés. Ces algorithmes sont entraînés à reconnaître les signatures caractéristiques des lentilles gravitationnelles dans les données de Gaia, en se basant sur des exemples connus de lentilles gravitationnelles et des configurations d’étoiles aléatoires.

L’avantage de cette approche est qu’elle permet d’identifier de façon complètement automatisée. Les algorithmes apprennent à distinguer les caractéristiques astrométriques et photométriques spécifiques des lentilles gravitationnelles. En identifiant ces signatures, les algorithmes peuvent détecter de nouvelles lentilles gravitationnelles dans les données observées par Gaia.

Dans la suite, je vais me pencher sur l’utilisation du *machine learning* pour analyser les données du satellite Gaia afin d’y détecter les lentilles gravitationnelles. Plus précisément, j’aborderai le choix de l’apprentissage supervisé comme approche principale pour cette tâche et je présenterai les algorithmes de classification utilisés. Ensuite, je me concentrerai sur une méthode spécifique appelée le *gradient tree boosting*. Je détaillerai les concepts des arbres de décision, des algorithmes à apprentissage d’ensemble et enfin, j’explorerai le fonctionnement du *gradient tree boosting*.

II.3 Le machine learning pour analyser Gaia

Toute la problématique de cette thèse porte sur l’entraînement d’un algorithme d’apprentissage automatique supervisé, utilisant la méthode du *gradient tree boosting*, pour la reconnaissance de configurations de lentilles gravitationnelles dans les données du satellite Gaia. Le défi de cette étude réside dans le fait qu’il existe un nombre très limité de vraies lentilles confirmées disponibles pour constituer un ensemble de données d’entraînement étiqueté fiable. Pour pallier cette contrainte, des simulations de lentilles gravitationnelles ont déjà été effectuées (Delchambre et al., 2019), mais elles ont présenté des lacunes en raison d’une distribution uniforme des paramètres du modèle de simulation.

Afin d’améliorer les résultats et d’obtenir un catalogue d’entraînement plus réaliste, j’ai réalisé de

nouvelles simulations de lentilles gravitationnelles en utilisant des distributions statistiques pour les paramètres impliqués basées sur mes travaux concernant la mesure des paramètres de forme des halos de matière noire dans la simulation EAGLE (The EAGLE team, 2017) pris comme déflecteurs pour générer des configurations de lentilles plus réalistes en plaçant une distribution de quasars en arrière-plan. Je détaillerai le processus de réalisation de ces simulations de lentilles gravitationnelles réalistes ainsi que la construction du catalogue d'entraînement pour l'algorithme d'apprentissage dans le chapitre [IV](#).

Le *machine learning* (apprentissage automatique) permet de développer des modèles prédictifs efficaces. Une fois que le modèle a été entraîné, il peut être utilisé pour faire des prédictions sur des données avec lesquelles l'algorithme n'a jamais été entraîné mais similaires à celles-ci. Le type d'algorithme de machine learning qui fonctionne le mieux dépend du problème considéré, de la nature des données et des ressources disponibles. Les modèles de machine learning peuvent généralement être classés dans trois catégories selon le type d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

L'apprentissage supervisé est un type d'apprentissage qui nécessite la connaissance préalable de ce que les valeurs de sortie sur un échantillon type sont. Les données d'entraînement sont étiquetées et les algorithmes supervisés apprennent à partir de ces exemples pour produire une fonction qui prédit les sorties à partir des entrées.

L'apprentissage supervisé suit généralement un processus bien défini pour s'assurer que le modèle est également performant sur des données nouvelles et ainsi assurer une généralisation optimale du modèle. Tout d'abord, le jeu de données est divisé en trois parties. La majorité, soit environ 70 à 80%, constitue les données d'entraînement. Ces données servent à initier le modèle, permettant à celui-ci d'apprendre des relations complexes présentes dans le jeu de données. Ensuite, environ 10 à 15% des données sont réservées à la validation. Cette phase intervient après l'entraînement initial et vise à affiner les hyperparamètres afin de prévenir le surajustement du modèle. Le surajustement, ou surapprentissage, se produit lorsque le modèle s'adapte trop précisément aux données d'entraînement, capturant même le bruit ou les fluctuations aléatoires. Cela se traduit par une performance médiocre sur de nouvelles données, car le modèle a appris des modèles spécifiques aux données d'entraînement qui ne généralisent pas bien. Une fois les hyperparamètres ajustés, le modèle est soumis à une phase de réentraînement en utilisant à la fois les données d'entraînement et de validation. Cette étape combine les données d'entraînement et de validation, favorisant une généralisation plus robuste. Enfin, pour évaluer la performance réelle du modèle, le reste des données, soit environ 10 à 15%, est utilisé comme ensemble de test. Cette phase permet de mesurer l'efficacité du modèle sur des données inconnues, offrant une évaluation réaliste de sa capacité à généraliser au-delà des exemples d'entraînement.

Les applications de l'apprentissage supervisé comprennent la classification, la régression et la

prédiction. L'apprentissage non supervisé, en revanche, est utilisé lorsque l'information utilisée pour entraîner le modèle n'est ni classifiée ni étiquetée. Les algorithmes non supervisés cherchent alors à découvrir des structures et des relations cachées dans les données avec ou sans contraintes extérieures. Les applications de l'apprentissage non supervisé comprennent, par exemple, la détection d'anomalies et la réduction de dimensionnalité. Enfin, l'apprentissage par renforcement est une méthode d'apprentissage automatique dans laquelle un agent apprend à prendre des décisions en interagissant avec son environnement. Dans ce cadre, l'agent reçoit une récompense ou une pénalité en fonction des actions qu'il entreprend, ce qui lui permet d'ajuster son comportement pour maximiser la récompense globale.

II.3.1 Le choix de l'apprentissage supervisé

Le choix d'utiliser l'apprentissage supervisé dépend des objectifs spécifiques de notre problème d'apprentissage et des données dont nous disposons. Examinons d'abord l'approche de l'apprentissage non supervisé. Cette méthode est utilisée lorsque l'objectif est d'explorer les structures sous-jacentes dans les données, sans nécessiter de données étiquetées. Elle peut aider à découvrir des regroupements, des corrélations ou des associations entre les données. Les algorithmes de regroupement (clustering), par exemple, peuvent être utilisés pour détecter des structures cosmologiques telles que des surdensités ou des amas de galaxies en analysant les positions et les propriétés des objets dans le ciel (Euclid Collaboration, 2019). D'autres approches non supervisées permettent également d'analyser et de classer les spectres provenant de vastes relevés spectroscopiques (Garcia-Dias et al., 2018). La classification basée sur les spectres permet de capturer les variations des paramètres atmosphériques des étoiles, tels que les abondances chimiques et la vitesse de rotation, ainsi que d'autres facteurs significatifs. L'algorithme est alors capable de différencier les populations stellaires du bulbe et du halo de la Voie Lactée, tout en retrouvant les étoiles naines, les sous-géantes, les étoiles de branche des géantes rouges et les étoiles de branche des géantes bleues.

De l'autre côté, l'apprentissage supervisé est couramment utilisé pour des tâches de prédiction et de classification, dans lesquelles l'objectif est de prédire une valeur continue ou de classer des données dans des catégories prédéfinies. Contrairement à l'apprentissage non supervisé, l'apprentissage supervisé nécessite des exemples d'entraînement étiquetés, où les données sont préalablement annotées avec par exemple leur classe. Toutefois, dans certaines situations où les données étiquetées sont limitées, des techniques telles que l'augmentation de données, comme jouer sur l'angle de rotation des images et l'apprentissage semi-supervisé peuvent être utilisées pour contourner le nombre limité de données étiquetées. Une autre approche fréquemment adoptée pour pallier le manque d'échantillons étiquetés consiste à générer un catalogue simulé de données étiquetées.

Dans le cadre de cette thèse, la recherche de quasars lentillés à quatre images est limitée par le nombre extrêmement restreint de cas connus (environ 70 cas), ce qui à partir d'eux seuls rend

impossible la constitution d'un catalogue étiqueté suffisamment grand qui couvrirait l'ensemble des configurations possibles. Face à cette limitation, nous avons choisi de nous concentrer sur le développement d'une méthode basée sur l'apprentissage supervisé s'appuyant sur des techniques d'apprentissage automatique d'ensemble pour améliorer la précision des prédictions par rapport à un modèle unique. Cela aide à réduire le surajustement et à fournir des résultats plus robustes. Les prédictions du modèle sont donc plus stables et fiables face à différentes conditions ou variations dans les données. Un modèle robuste est moins susceptible de produire des résultats excessivement influencés par des données spécifiques d'entraînement ou par des variations mineures dans les données d'entrée.

Pour mieux organiser les algorithmes d'apprentissage automatique, il est utile de les classer selon leurs similitudes. Il existe de nombreux modèles différents, chacun adapté à des types de problèmes spécifiques. Ainsi, le choix du modèle le plus approprié pour un problème donné est crucial pour obtenir les meilleurs résultats. En apprentissage supervisé, la notion centrale est celle de la perte d'information, qui mesure l'écart entre les prédictions du modèle et les valeurs réelles. En général, l'apprentissage consiste à minimiser une fonction de perte, qui est une mesure de la qualité de la prédiction du modèle. Cette fonction peut être choisie pour minimiser l'erreur ou maximiser la vraisemblance du modèle. L'apprentissage est donc souvent un processus itératif qui consiste à ajuster les paramètres du modèle pour minimiser la fonction de perte et améliorer la qualité des prédictions. Cette fonction de perte utilisée pour l'optimisation est à ne pas confondre avec la métrique utilisée lors de l'évaluation des performances, qui utilisera d'autres mesures pour étudier la qualité finale du modèle.

II.3.2 Les algorithmes de classification

En apprentissage automatique, la classification vise à identifier la classe (ou l'étiquette) d'un objet en utilisant ses caractéristiques. Les classificateurs linéaires accomplissent cette tâche en prenant une décision de classification basée sur la valeur d'une combinaison linéaire des caractéristiques. Les caractéristiques d'un objet sont souvent présentées sous la forme d'un vecteur appelé vecteur d'observation (ou *feature vector* en anglais).

En classification binaire, l'objectif d'un classificateur linéaire est de séparer un espace de grande dimension en deux parties à l'aide d'un hyperplan ou frontière de décision (voir Figure II.5). L'hyperplan est défini par un vecteur normal \vec{w} et un biais w_0 , et il divise l'espace en deux parties telles que tous les points d'un côté de l'hyperplan sont classés en tant que 1 et les autres en tant que -1. L'hyperplan est également appelé hyperplan séparateur ou séparatrice. Le choix de l'hyperplan est déterminé par les vecteurs de poids appris à partir des données d'apprentissage. Plus précisément, l'hyperplan est perpendiculaire au vecteur de poids et il est situé à une distance $w_0/|\vec{w}|$ de l'origine de l'espace. Les points qui se trouvent à une distance plus grande de l'hyperplan sont plus sûrs dans

leur classification que les points qui se trouvent à proximité de l'hyperplan.

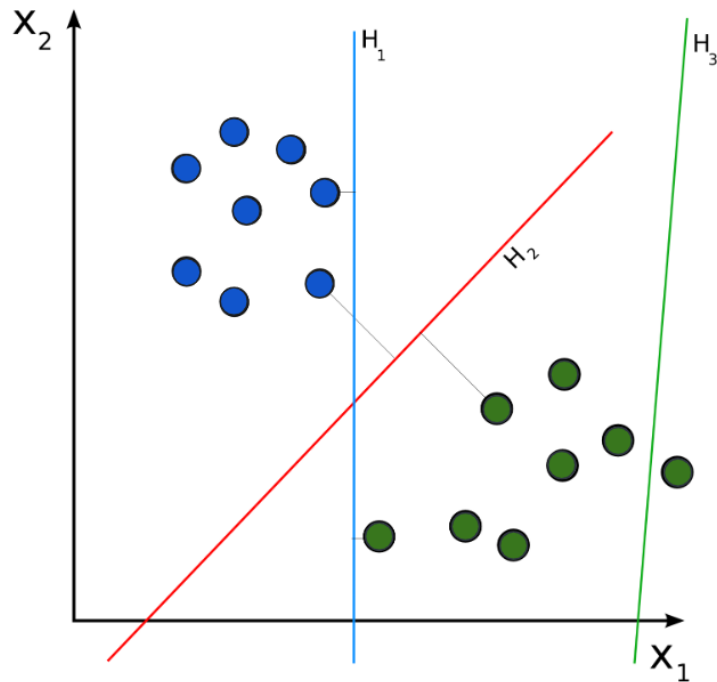


FIGURE II.5 – Exemple d'hyperplans séparateurs. Dans ce cas, les points pleins et les points vides peuvent être correctement classés par n'importe quel nombre de classificateurs linéaires. H_1 (bleu) les classe correctement, tout comme H_2 (rouge). H_2 pourrait être considéré comme "meilleur" dans le sens où il est également le plus éloigné des deux groupes. H_3 (vert) ne parvient pas à classer correctement les points.

Cependant, il existe des problèmes de classification qui ne peuvent pas être résolus efficacement par un simple hyperplan, surtout lorsque les données ne sont pas linéairement séparables dans l'espace d'origine. C'est là que les *Support Vector Machines* (SVM) peuvent intervenir. *Support Vector Machine* (SVM) est un algorithme d'apprentissage automatique supervisé utilisé pour des tâches de classification et de régression. Il est particulièrement efficace pour résoudre des problèmes de classification binaire, mais peut également être étendu pour traiter des classifications multi-classes. Dans le cadre des SVM, l'algorithme prend un ensemble de données d'entrée et les classe dans l'une des deux catégories en construisant un hyperplan optimal dans un espace de caractéristiques de haute dimension. Cet hyperplan est choisi de manière à maximiser la marge entre les deux classes, ce qui permet une meilleure généralisation et une performance améliorée sur des données non vues auparavant.

Le terme "vecteur de support" désigne les points de données les plus proches de la frontière de décision, jouant un rôle crucial dans la définition de l'hyperplan. Ces vecteurs de support sont essentiels pour

déterminer l'hyperplan optimal et classer de nouvelles instances. Dans cette optique, l'utilisation d'un noyau (*kernel*) linéaire permet de transformer linéairement l'espace des caractéristiques d'origine des données en un espace de dimension supérieure, facilitant ainsi la séparation linéaire des données.

Maintenant, tournons-nous vers une autre approche en matière d'apprentissage supervisé : le "gradient tree boosting". Le gradient tree *boosting* repose sur un ensemble d'arbres de décision, où chaque arbre est construit de manière itérative pour corriger les erreurs commises par les arbres précédents. Ce processus d'apprentissage par renforcement permet de créer des modèles plus complexes et plus flexibles, capables de capturer des relations non linéaires entre les caractéristiques et les étiquettes de classification.

Le gradient tree *boosting*, également connu sous le nom de gradient *boosted* decision trees (GBDT), est un modèle d'apprentissage automatique qui combine les idées des arbres de décision et de la descente de gradient par *boosting* pour construire un modèle prédictif fort et précis. Les arbres de décisions sont des modèles d'apprentissage automatique qui utilisent une structure arborescente pour prendre des décisions en évaluant séquentiellement différentes caractéristiques d'un ensemble de données. Le *boosting* est une technique d'ensemble qui consiste à combiner plusieurs modèles faibles pour obtenir un modèle plus fort. Un modèle faible dans ce contexte est un modèle d'apprentissage dont la performance est légèrement meilleure que celle d'un modèle aléatoire. En revanche, un modèle fort résulte de la combinaison pondérée de plusieurs modèles faibles, créant ainsi un modèle global capable de fournir des prédictions plus précises. L'idée de base du gradient tree *boosting* est d'ajouter de manière itérative des arbres de décision au modèle, chaque nouvel arbre étant formé pour corriger les erreurs des arbres précédents. L'algorithme fonctionne en ajustant initialement un modèle simple, tel qu'un arbre de décision, aux données. Les résidus ou les erreurs du modèle initial sont ensuite utilisés comme variable de réponse pour former l'arbre suivant. Ce processus est répété plusieurs fois. La prédiction finale est obtenue en combinant les prédictions de tous les arbres. L'algorithme peut être personnalisé à l'aide de divers hyperparamètres, tels que le taux d'apprentissage, qui contrôle la contribution de chaque nouvel arbre au modèle, et la profondeur des arbres, qui contrôle la complexité du modèle.

La "descente de gradient" fait référence à un algorithme d'optimisation itératif utilisé pour minimiser une fonction de coût en ajustant itérativement les valeurs des paramètres dans la direction opposée au gradient de la fonction de coût. On cherche ainsi à atteindre le minimum de la fonction de coût.

Dans la suite, je vais détailler chacun de ces points pour mieux comprendre comment fonctionne l'algorithme de gradient tree *boosting*.

II.3.3 Le gradient tree boosting

II.3.3.1 Les arbres de décision

Les arbres de décision sont une méthode courante d'apprentissage supervisé non paramétrique utilisée pour la classification et la régression. Ils fonctionnent en créant un modèle qui prédit la catégorie ou la classe pour chaque entrée en apprenant des règles de décision simples déduites des caractéristiques des données. Concrètement, l'arbre est construit en sélectionnant la caractéristique qui divise au mieux les données. Les arbres de décision sont souvent représentés sous forme d'un arbre avec chaque nœud représentant une caractéristique, chaque branche représentant une règle de décision basée sur cette caractéristique, et chaque feuille représentant une prédiction de la classe.

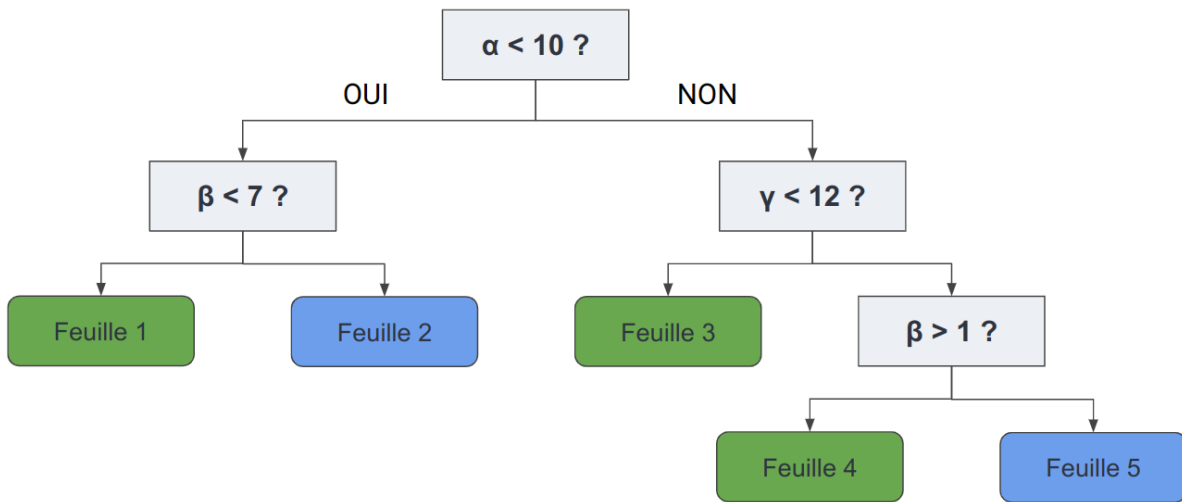


FIGURE II.6 – Représentation d'un arbre de décision. Les nœuds de l'arbre, représentés par des rectangles de couleur gris clair, sont des points de décision, chacun contenant une condition spécifique basée sur une caractéristique des données. Ces conditions guident la division des données en sous-groupes à chaque nœud. Les feuilles de l'arbre, illustrées par des rectangles de couleur aux bords arrondis, marquent les extrémités des branches et représentent la décision finale, souvent exprimé par une étiquette de classe ou une prédiction (la classe bleue ou verte). Les branches, symbolisées par des lignes, connectent les nœuds de l'arbre, illustrant les relations entre les différentes conditions.

Un arbre de décision ou de classification est une structure d'arbre dans laquelle chaque nœud interne (non feuille) est étiqueté. Chaque branche qui part d'un nœud représente une valeur possible de la caractéristique et mène à un autre nœud ou une feuille. Une feuille représente la décision finale, c'est-à-dire l'étiquette de classe. L'arbre est construit en choisissant les caractéristiques à diviser et les conditions de division pour maximiser une mesure de gain d'information. L'algorithme de construction est généralement basé sur une stratégie de division répétitive de l'ensemble de données d'entrée. À chaque nœud, l'algorithme calcule une mesure de qualité de la division et sélectionne la caractéristique et la valeur de seuil qui produisent la meilleure division. La construction de l'arbre se

poursuit jusqu'à ce qu'un critère d'arrêt soit satisfait, tel que la profondeur maximale de l'arbre ou le nombre minimum d'échantillons par feuille. Le résultat est un arbre qui peut être utilisé pour la classification ou la régression selon la nature du problème à résoudre. L'algorithme CART (Classification and Regression Trees) (Breiman, 1984) est souvent utilisé pour construire des arbres de décision. L'algorithme CART est très avantageux par rapport aux autres algorithmes, car il est capable de gérer à la fois les arbres de régression et les arbres de classification en utilisant des métriques spécifiques pour chaque cas. Par exemple, la métrique utilisée pour l'arbre de régression est la MSE (Mean Squared Error), tandis que pour l'arbre de classification, l'impureté de Gini est utilisée comme mesure de qualité de la division.

L'impureté de Gini est une mesure de la probabilité d'un classement incorrect d'une nouvelle instance d'une variable aléatoire en cours de classification. C'est une mesure de la probabilité que deux éléments choisis au hasard dans un ensemble soient étiquetés de manière différente. Pour un ensemble d'éléments avec J classes, la probabilité de choisir un élément avec l'étiquette i est p_i , et la probabilité de choisir un élément d'une autre classe est $1 - p_i$. L'impureté de Gini est calculée comme la somme pondérée des probabilités de chaque classe i multipliées par la probabilité de mal classer un élément dans cette classe, soit $1 - p_i$. En d'autres termes, l'impureté de Gini se calcule comme suit :

$$I_G(p) = \sum_{i=1}^J p_i(1 - p_i) = \sum_{i=1}^J p_i - \sum_{i=1}^J p_i^2 = 1 - \sum_{i=1}^J p_i^2 \quad (\text{II.3})$$

Cette mesure est comprise entre 0 et 1, où une valeur de 0 indique un ensemble homogène contenant des éléments appartenant tous à la même classe, tandis qu'une valeur de 1 indique un ensemble parfaitement équilibré dans lequel les éléments sont répartis uniformément entre toutes les classes possibles.

Les arbres de décision présentent certaines limites, notamment le surapprentissage, l'instabilité et la sensibilité aux données d'entrée. Le surapprentissage peut se produire lorsque l'arbre de décision est trop complexe ou que l'ensemble d'entraînement est trop petit. Dans ces cas, l'arbre de décision peut mémoriser les données d'entraînement plutôt que d'apprendre le modèle sous-jacent, ce qui se traduit par une faible généralisation des caractéristiques des données. L'instabilité et la sensibilité font référence au fait que de petites modifications des données d'entraînement ou des paramètres de l'algorithme peuvent entraîner des changements significatifs dans l'arbre de décision obtenu. L'instabilité peut survenir lorsque l'arbre de décision est construit à l'aide d'un ensemble de données spécifique qui peut ne pas bien représenter la population. Dans ce cas, l'arbre de décision peut surapprendre et être incapable de généraliser à de nouvelles données. La sensibilité, quant à elle, fait référence au fait que de petites modifications des données d'entrée peuvent entraîner la génération d'arbres décisionnels différents. Cela peut rendre les arbres décisionnels moins robustes et moins

fiables, en particulier lorsque les données d'entrée sont bruitées ou contiennent des erreurs. Ces problèmes peuvent être résolus en utilisant des méthodes d'apprentissage d'ensemble qui combinent plusieurs arbres décisionnels construits à partir de différents sous-ensembles de données dans le but de réduire l'impact des arbres individuels.

Je vais maintenant aborder les algorithmes à apprentissage d'ensemble. Plutôt que de se fier à un seul modèle, ces méthodes exploitent la combinaison de multiples modèles d'apprentissage pour obtenir des prédictions plus robustes et précises. Les algorithmes à apprentissage d'ensemble tirent parti de la diversité des modèles et de leur capacité à se corriger mutuellement, ce qui permet d'améliorer significativement les performances de prédiction. Dans la partie qui suivra, j'examinerai le *boosting*.

II.3.3.2 Les algorithmes à apprentissage d'ensemble

L'apprentissage d'ensemble est une technique d'apprentissage automatique qui consiste à combiner plusieurs modèles pour améliorer la précision et la robustesse de la prédiction. L'idée derrière l'apprentissage d'ensemble est que plusieurs modèles faibles peuvent être combinés pour créer un modèle fort. Un modèle faible dans le contexte de l'apprentissage d'ensemble est un modèle dont la performance prédictive est légèrement meilleure que celle d'une prédiction aléatoire. En d'autres termes, un modèle faible a une précision qui dépasse le hasard, mais il peut ne pas être très performant. D'un autre côté, un modèle fort est un modèle qui a une capacité de prédiction significativement meilleure que celle d'un modèle faible. Lorsque plusieurs modèles faibles sont combinés ils peuvent former un modèle fort, capable de fournir des prédictions plus précises et robustes que celles de n'importe quel modèle individuel.

L'apprentissage d'ensemble peut être appliqué à un large éventail d'algorithmes d'apprentissage automatique, notamment les arbres de décision et les réseaux neuronaux. Il existe plusieurs types de méthodes d'apprentissage d'ensemble comme le *bootstrap aggregating (bagging)* (Breiman, 1996) et le *boosting* (Schapire and Robert, 1990)

Le boosting

Le *boosting* a été proposé dans la littérature sur la théorie de l'apprentissage informatique (Schapire and Robert, 1990; Freund, 1995) et a depuis reçu beaucoup d'attention. Le *boosting* fonctionne en exécutant à plusieurs reprises un algorithme d'apprentissage faible donné (*weak learner* en anglais) sur diverses distributions sur les données d'apprentissage, puis en combinant les classificateurs produits par l'apprenant faible en un seul classificateur composite.

L'algorithme AdaBoost (Freund, 1995), premier algorithme de *boosting*, fonctionne par itérations successives d'un apprenant faible. La distribution utilisée à l'itération t est notée D_t , et le poids attribué à l'exemple d'apprentissage i est noté $D_t(i)$. Ce poids peut être considéré comme une mesure de l'importance d'une classification correcte de l'exemple i à l'itération en cours. Initialement, tous

les poids sont fixés de manière égale, mais à chaque itération, les poids des exemples mal classés sont augmentés, permettant ainsi aux exemples difficiles d'obtenir des poids de plus en plus élevés. Cette augmentation progressive des poids pousse l'apprenant faible à se concentrer davantage sur ces exemples difficiles, en cherchant à les classer correctement. Voyons comment mettre en place un tel algorithme dont le "pseudocode" est représenté par l'algorithme de la table [II.1](#).

La tâche de l'apprenant faible consiste à trouver un classificateur faible $h_t : \mathcal{X} \rightarrow -1, +1$ adapté à la distribution D_t . La qualité d'un classificateur faible est évaluée par son erreur pondérée ϵ_t selon la distribution D_t . L'erreur pondérée ϵ_t correspond à la probabilité de h_t de mal classer un exemple aléatoire sélectionné selon D_t . Cette notion de classificateur faible h_t est similaire à celle des hyperplans séparateurs abordée dans la section sur les classificateurs linéaires (voir Section [II.3.2](#)). L'apprenant faible cherche donc à choisir un classificateur faible h_t avec une faible erreur pondérée ϵ_t . Une fois que le classificateur faible h_t est obtenu, AdaBoost attribue un paramètre α_t comme indiqué dans l'algorithme de la table [II.1](#). Intuitivement, α_t représente l'importance accordée à h_t . On observe que $\alpha_t > 0$ lorsque $\epsilon_t < \frac{1}{2}$, et que α_t augmente à mesure que ϵ_t diminue. Ainsi, plus le classificateur de base h_t est précis, plus il est considéré comme important dans le processus de *boosting*.

Après un certain nombre d'itérations, AdaBoost fusionne les classificateurs faibles en un seul classificateur combiné, noté H . Cette fusion est réalisée par un vote pondéré des classificateurs faibles. En d'autres termes, le vote du t -ème classificateur de base h_t est pondéré par le paramètre α_t précédemment déterminé. La formule résultante pour la prédiction de H est la suivante :

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right) \quad (\text{II.4})$$

La figure [II.7](#) illustre le processus itératif d'AdaBoost. Dans cet exemple, nous considérons trois itérations avec $t = 1, 2, 3$. Lors de la première itération, un premier classificateur faible est sélectionné pour $t = 1$. Trois exemples sont mal classés et, par conséquent, leur poids est augmenté pour la deuxième itération à $t = 2$. Cette fois-ci, un deuxième classificateur faible réussit à classer correctement les exemples qui avaient été mal classés lors de l'itération précédente, mais il introduit également de nouvelles erreurs de classification. Ainsi, lors de la troisième itération pour $t = 3$, les exemples mal classés lors de l'itération précédente sont à nouveau pondérés avec des poids plus élevés. Finalement, le classificateur combiné $H(x)$ est calculé en prenant le signe de la somme pondérée des trois hypothèses faibles, c'est-à-dire $H(x) = \alpha_1 h_1 + \alpha_2 h_2 + \alpha_3 h_3$ pour cet exemple.

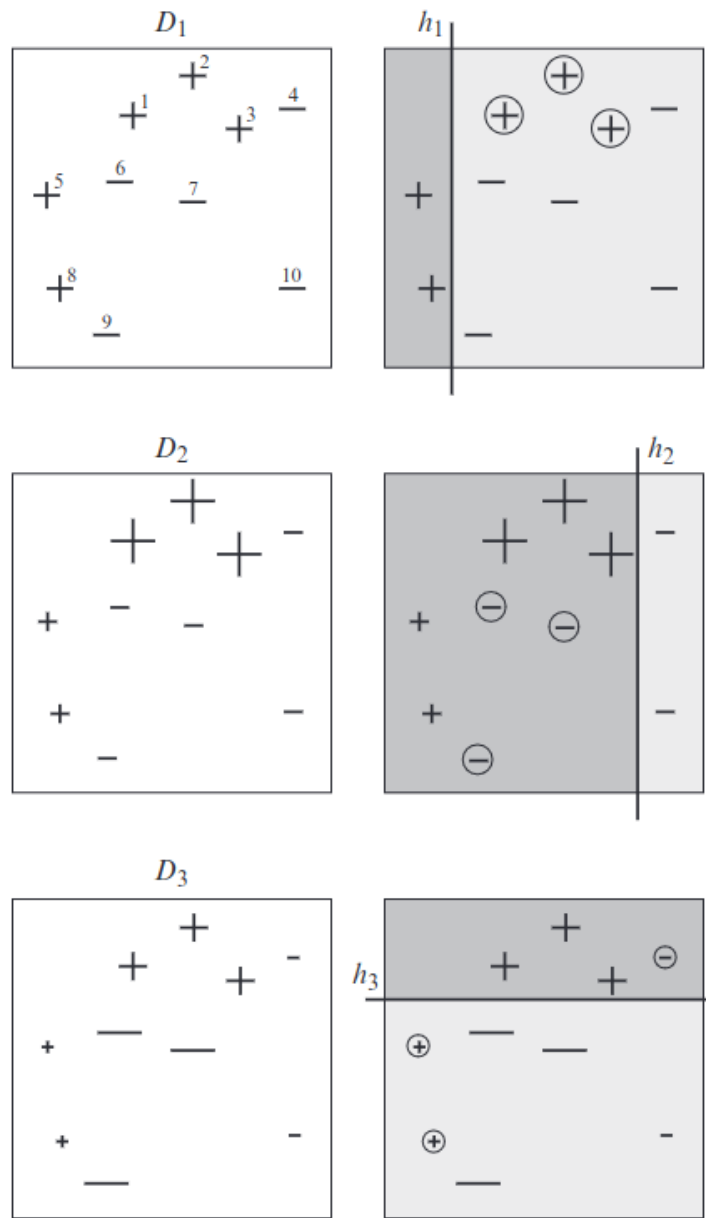


FIGURE II.7 – Illustration tirée du livre "Boosting" de Schapire et Freund. Chaque ligne correspond à une itération, avec $t = 1, 2, 3$. La case de gauche de chaque ligne représente la distribution D_t , où la taille de chaque exemple est mise à l'échelle proportionnellement à son poids sous cette distribution. De plus, chaque case de droite montre l'hypothèse faible h_t , avec une région plus foncée indiquant la prédiction positive. Les exemples mal classés par h_t sont encadrés pour une meilleure visualisation.

TABLE II.1 – **Algorithme 1.1.**
L'algorithme de *boosting* AdaBoost.

Soit $(x_1, y_1), \dots, (x_m, y_m)$ où $x_i \in X, y_i \in \{-1, +1\}$.

Initialiser $D_1(i) = 1/m$ pour $i = 1, \dots, m$.

Pour $t = 1, \dots, T$:

- Entraîner l'apprenant faible en utilisant la distribution D_t .
- Obtenir l'hypothèse faible $h_t : X \rightarrow \{-1, +1\}$.
- Objectif : sélectionner h_t pour minimiser l'erreur pondérée :

$$\epsilon_t \doteq \Pr_{i \sim D_t}[h_t(x_i) \neq y_i].$$

- Choisir $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$.
- Mettre à jour, pour $i = 1, \dots, m$:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

où Z_t est un facteur de normalisation (choisi pour que D_{t+1} soit une distribution).

Sortie de l'hypothèse finale :

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right).$$

II.3.3.3 Le choix de XGBoost

Les méthodes d'apprentissage automatique reposent d'une manière ou d'une autre sur l'optimisation d'une fonction d'objectif ou de perte (*loss function* en anglais). Par exemple, dans sa forme la plus simple de régression linéaire, à partir des exemples d'entraînements $(x_1, y_1), \dots, (x_m, y_m)$, où $x_i \in \mathfrak{R}^n$ et $y_i \in \mathfrak{R}$, on cherche à trouver un vecteur de poids \mathbf{w} tel que $\mathbf{w} \cdot \mathbf{x}_i$ soit une bonne approximation de y_i . Plus précisément, le but est de trouver $\mathbf{w} \in \mathfrak{R}^n$ en minimisant la moyenne (ou somme) des erreurs quadratiques.

$$\mathcal{L}(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m (\mathbf{w} \cdot \mathbf{x}_i - y_i)^2 \tag{II.5}$$

Ici, l'erreur quadratique de chaque exemple $(\mathbf{w} \cdot \mathbf{x}_i - y_i)^2$ est la fonction de perte - dans ce cas, la perte carrée ou quadratique - et le but est de minimiser la moyenne des pertes sur tous les m exemples.

Dans le cadre d'AdaBoost, nous pouvons le comprendre comme une méthode de minimisation basée sur la descente de coordonnées d'une fonction d'optimisation spécifique bien que nous ne détaillions pas ici la méthode de descente de coordonnées. Je me concentre plutôt sur le Gradient Tree *boosting*, qui est l'algorithme d'optimisation que je vais utiliser. Plus précisément, j'examine XGBoost (pour eXtreme Gradient Boosting) (Chen and Guestrin, 2016). L'algorithme combine l'apprentissage d'ensemble avec des arbres de décision pour créer un modèle prédictif robuste. Grâce à sa capacité à capturer des relations complexes entre les variables d'entrée, XGBoost excelle dans la classification de données et est particulièrement adapté aux problèmes de grande dimension.

XGBoost fonctionne en entraînant un ensemble d'arbres de décision successifs. Chaque arbre est ajouté à l'ensemble de manière itérative, en cherchant à améliorer les prédictions du modèle existant. À chaque itération, le modèle tente de prédire les résidus (différence entre les prédictions actuelles et les vraies valeurs) plutôt que les valeurs brutes elles-mêmes. Cela permet de réduire l'erreur résiduelle à chaque étape et d'améliorer progressivement la précision du modèle. Les arbres de décision sont construits de manière à minimiser la fonction de perte et à intégrer la régularisation pour éviter le surajustement.

Le réglage précis des hyperparamètres de XGBoost est un élément essentiel pour obtenir des performances optimales du modèle. Ils déterminent des aspects tels que la profondeur des arbres, le taux d'apprentissage, le nombre d'itérations, la régularisation, et bien d'autres. Dans la partie consacrée à l'apprentissage dans le chapitre [V](#), je détaillerai les différentes stratégies et techniques utilisées pour régler ces hyperparamètres, en optimisant les performances du modèle pour la reconnaissance de configurations de lentilles gravitationnelles. Je discuterai également des méthodes d'évaluation et de validation du modèle afin d'obtenir des résultats fiables et généralisables.

Au terme de ce chapitre, j'ai abordé le défi de la reconnaissance de configurations de lentilles gravitationnelles dans les données de Gaia. J'ai souligné la limitation des ensembles de données d'entraînement disponibles. Afin de surmonter ces contraintes, j'ai développé une nouvelle approche basée sur l'utilisation des halos de la simulation EAGLE comme déflecteurs, combinée à des distributions réalistes de quasars en arrière-plan. Cette approche a été choisie pour améliorer la construction du catalogue d'entraînement et obtenir des simulations de lentilles gravitationnelles plus réalistes. Dans le prochain chapitre, j'aborderai en détail l'approche qui a été retenue pour établir de meilleures distributions a priori de probabilité pour la morphologie des galaxies déflectrices. Je décrirai notamment l'analyse de la distribution de masse au sein des vastes simulations cosmologiques.

Chapitre III

Profil de masse des galaxies

Dans ce chapitre, je me concentre sur les moyens d'améliorer les simulations de lentilles gravitationnelles existantes en obtenant les distributions statistiques réalistes des paramètres de la morphologie des galaxies défectrices. Pour ce faire, j'ai examiné en détail la structure des halos massifs au sein de vastes simulations cosmologiques.

Le chapitre se divise en deux parties. Dans la première partie, je présenterai les différentes méthodes généralement utilisées pour étudier la distribution de la masse dans les galaxies. Parmi ces méthodes, la courbe de rotation des galaxies fournit une quantité observable utilisée pour estimer la distribution de masse. Je discuterai également des simulations cosmologiques, qui permettent d'observer et d'analyser l'évolution de la répartition de la masse à grande échelle. J'analyserai leur pertinence pour notre objectif d'amélioration des simulations de lentilles gravitationnelles, en tant qu'outils complémentaires pour mieux comprendre la distribution de masse et son impact sur les lentilles gravitationnelles.

Dans la seconde partie, je me concentrerai sur la mesure des halos de matière dans la suite de simulations cosmologiques EAGLE ([The EAGLE team, 2017](#)), qui ont été utilisés comme déflecteurs dans mes simulations de lentille. Je décrirai en détail le processus de sélection des halos et de mesure de leurs paramètres de forme, en supposant que ces halos représentent une approximation réaliste de la distribution de matière dans les galaxies. Ces informations sont essentielles pour obtenir des distributions de probabilité plus réalistes et mieux adaptées à la morphologie des galaxies défectrices. Ces nouvelles distributions a priori permettront de simuler des configurations réalistes de lentilles gravitationnelles de quasars. Mon objectif est d'obtenir des résultats plus précis et fiables pour la détection de lentilles gravitationnelles de quasars à partir des données de Gaia.

III.1 Les différentes méthodes utilisées pour mesurer la distribution de la masse dans les galaxies

III.1.1 La courbe de rotation des galaxies

Une méthode courante pour mesurer la répartition de la matière au sein des galaxies consiste à utiliser la dynamique des étoiles dans une galaxie à disque pour déterminer sa courbe de rotation. En effet, la force gravitationnelle exercée par la masse incluse jusqu'à la distance de l'objet étudié régit les vitesses orbitales des étoiles et des nuages de gaz dans la galaxie. En se basant sur l'hypothèse de l'équilibre dynamique, il est possible d'appliquer le théorème du viriel pour estimer la masse en question compte tenu de la vitesse mesurée. Ainsi, en mesurant les vitesses radiales des étoiles ou des nuages de gaz, on peut déterminer les vitesses circulaires à différentes distances du centre galactique. La courbe de rotation traduit la variation de cette vitesse en fonction de la distance radiale des étoiles de la galaxie au centre.

Pour mesurer les vitesses radiales des étoiles ou du gaz dans la galaxie, on mesure généralement les décalages Doppler des raies spectrales émises par les objets en mouvement. Dans le cas spécifique du gaz, la raie à 21 cm de l'hydrogène est souvent utilisée, car elle permet de sonder la répartition pratiquement deux fois plus loin que par les étoiles. En analysant ces vitesses, on peut déduire la distribution de masse au sein de la galaxie et étudier sa structure. C'est grâce à cette analyse que peut être établie la relation entre la vitesse de rotation et la répartition de masse. Selon la 3e loi de Kepler, on s'attendrait à voir rapidement une diminution de la vitesse de rotation en fonction de la distance au centre de la galaxie. Or, il a été observé dès 1932 (Oort, 1932) que les étoiles et le gaz à la périphérie des galaxies se déplacent plus rapidement que prévu sur la base de la masse comptabilisée pour les étoiles et le gaz seuls.

Les observations pionnières de Zwicky (Zwicky, 1937) ont émis l'hypothèse de l'existence d'une composante massive invisible en se basant sur l'étude du déplacement des galaxies au sein d'amas proches. Par la suite, les observations de Rubin et Ford (Rubin and Ford, 1970; Rubin et al., 1980) ont confirmé cette hypothèse au sein même des galaxies en montrant que la vitesse de rotation des étoiles dans les galaxies ne décroissait pas au-delà d'une certaine distance au centre galactique. Au lieu de cela, la vitesse de rotation augmentait rapidement jusqu'à atteindre une valeur d'environ $V \simeq 125$ km/s à une distance d'environ $R \simeq 5$ kpc. Au-delà de cette distance, la courbe de rotation montrait une augmentation plus lente. Ainsi, ces observations ont révélé que la distribution de masse dans les galaxies ne se concentrait pas uniquement au centre, mais s'étendait sur de larges distances. Cette composante invisible est aujourd'hui connue sous le nom de matière noire, souvent abrégée en DM (Dark Matter).

La masse visible, composée d'étoiles et de gaz, n'est pas suffisante pour expliquer la vitesse de rotation

observée, même pour des distances galactocentriques inférieures à celle du Soleil dans la Voie Lactée. En outre, cette masse invisible qui n'émet pas de rayonnement électromagnétique est constituée de particules autres que les baryons, qui eux interagissent avec le rayonnement électromagnétique. La mesure des courbes de rotation montre que la matière noire contribue de manière plus que significative à la masse totale des galaxies. En effet, le rapport masse-luminosité (M/L) moyen est d'environ 5, et il est encore plus élevé pour les galaxies irrégulières. Ainsi, la contribution de la matière noire est dominante et prédominante.

La prédominance de la matière noire dans la masse totale des galaxies en fait un élément essentiel dans le contexte des lentilles gravitationnelles. En effet, la matière noire non seulement creuse le potentiel gravitationnel des galaxies, mais elle s'étend également bien au-delà du rayon optique, ce qui augmente la section efficace des effets gravitationnels. Considérant la relativité générale, nous "observons" la matière noire par les effets qu'elle engendre plutôt que par une observation directe. Il est donc essentiel de tenir compte de cette composante lors de la simulation des lentilles gravitationnelles. Pour obtenir des informations sur la morphologie et la distribution de la matière noire, il est possible de recourir à des simulations numériques, telles que les simulations hydrodynamiques d'univers. Les simulations hydrodynamiques d'univers sont des modèles numériques qui tentent de reproduire le comportement dynamique des fluides, tels que les gaz et les matières noires, à l'échelle cosmologique. Elles intègrent les principes de l'hydrodynamique pour étudier les mouvements des gaz dans des cubes univers en tenant compte de la gravité et d'autres interactions comme les forces de pression. Le terme cube d'univers fait référence à la manière dont ces simulations divisent l'espace en grilles tridimensionnelles, représentées sous forme de cubes, pour faciliter le calcul des interactions entre différentes régions.

Cependant, il convient de noter que la prédiction précise du contenu en matière noire des galaxies nécessite de prendre en compte des échelles à la fois grandes et petites, ainsi que des caractéristiques spécifiques, comme la nature non collisionnelle de la matière noire et la physique complexe des baryons, incluant les rétroactions des AGN (Active Galactic Nucleus). Dans ces simulations, le profil théorique de densité NFW (Navarro-Frenk-White) (Navarro et al., 1996) est souvent utilisé pour décrire les halos de matière noire, mais il est également important de considérer la présence éventuelle d'un cœur dans le profil réel. Ainsi, grâce au développement de modèles de simulations hydrodynamiques et aux avancées des infrastructures de supercalcul, il devient possible de prédire de manière plus approfondie le contenu en matière noire des galaxies et d'estimer leur morphologie avec une meilleure précision.

III.1.2 Les simulations hydrodynamiques d'univers

Du point de vue théorique, le scénario hiérarchique de formation des structures postule que les galaxies se forment et évoluent à l'intérieur de halos de matière noire froide (*Cold Dark Matter*

en anglais). Selon la théorie, la matière noire est supposée se comporter comme un fluide non collisionnel, dont les hétérogénéités s'amplifient avec le temps une fois les particules qui la constituent découplées du rayonnement (Vogelsberger et al., 2012; Vogelsberger et al., 2014). En raison de la nature hautement non linéaire des équations qui régissent l'évolution de ces hétérogénéités, il est difficile de prédire analytiquement leur évolution tardive lorsque le contraste de densité devient supérieur à 180. Le contraste de densité fait référence à la variation de densité d'une région donnée par rapport à la densité moyenne de l'univers. Mathématiquement, il est souvent exprimé comme :

$$\delta = \frac{\rho - \bar{\rho}}{\bar{\rho}} \quad (\text{III.1})$$

où δ représente le contraste de densité, ρ est la densité locale, et $\bar{\rho}$ est la densité moyenne cosmique. Les simulations hydrodynamiques d'univers utilisent ces contrastes pour représenter les fluctuations de densité qui évoluent au fil du temps, conduisant à la formation de structures telles que galaxies, amas de galaxies, etc. De plus, des modèles comme le profil NFW (Navarro-Frenk-White) ont été développés pour décrire les halos de matière noire. Les profils NFW sont caractérisés par une densité qui suit une forme spécifique :

$$\rho(r) = \frac{\rho_0}{\frac{r}{r_s} \left(1 + \frac{r}{r_s}\right)^2}, \quad (\text{III.2})$$

où ρ_0 est une densité de référence et r_s est une échelle radiale caractéristique. Ces modèles sont couramment utilisés dans les simulations hydrodynamiques pour représenter la distribution de matière noire et comprendre l'influence de cette dernière sur la formation des structures cosmiques.

Pour suivre cette évolution, il est possible d'utiliser des simulations numériques, notamment les simulations hydrodynamiques "Dark Matter Only" (DMO) qui se concentrent exclusivement sur le suivi gravitationnel de la matière noire sans tenir compte des autres composantes que sont les étoiles et le gaz. Ces simulations cosmologiques d'envergure permettent de prendre en considération les coalescences successives des halos de matière noire. Les halos de matière noire simulés à l'aide de ces modèles DMO présentent des profils de densité caractérisés par une augmentation de la densité vers leur centre. Ces profils de densité sont bien décrits par des modèles tels que le profil Navarro-Frenk-White (NFW) (Navarro et al., 1996) ou le profil Einasto (Einasto, 1965). Par exemple, le profil NFW correspond à une courbe de rotation de la matière noire qui augmente dans les 15 kpc intérieurs des halos suffisamment massifs ($> 10^{18}, M_{\odot}$) qui sont susceptibles de former des galaxies (Springel et al., 2008; Vogelsberger et al., 2009).

Les avancées récentes dans les modèles de formation des galaxies, combinées à la disponibilité croissante de puissance de calcul, ont permis de dépasser l'approche ne considérant que la matière

noire seule (DMO) en incluant la physique des gaz et en tenant compte des processus astrophysiques complexes dans les simulations de formation des galaxies (Springel and Hernquist, 2003; Springel et al., 2005; Schaye et al., 2010; Crain et al., 2010; Hopkins et al., 2014; Vogelsberger et al., 2014; Schaye et al., 2015; Somerville and Davé, 2015). En plus de résoudre les propriétés hydrodynamiques du gaz, les simulations de formation de galaxies appliquent des modèles de raffinement de maillage. Le raffinement de maillage est une technique numérique qui permet de suivre la dynamique de processus se déroulant à des échelles plus petites que la maille standard de la simulation, en particulier la formation des étoiles dans les cœurs denses des nuages moléculaires. Il est important de noter que la gestion de ces processus de rétroaction, dont les mécanismes, l’ampleur et la durée sont mal connus, reste un défi majeur. Des projets de simulation tels que EAGLE (Schaye et al., 2015), GAEA (Hirschmann et al., 2016), IllustrisTNG (Pillepich et al., 2018; Nelson et al., 2018) et Thesan (Kannan et al., 2022) ont rencontré un franc succès en reproduisant, par exemple, la fonction de luminosité observée des galaxies à travers les époques, ainsi qu’une multitude d’autres propriétés des galaxies liées aux étoiles et au gaz. Cependant, il reste encore des aspects à améliorer et des défis à relever. Les simulations de galaxies actuelles, telles qu’EAGLE, représentent un état de l’art solide et fournissent des prévisions crédibles sur la formation des galaxies, notamment en ce qui concerne la distribution de la matière noire et baryonique. Cependant, des efforts continus sont nécessaires pour améliorer la modélisation des processus de rétroaction, comprendre les mécanismes sous-jacents et résoudre les problèmes restants.

Les propriétés structurelles des galaxies, telles que les courbes de rotation (Scannapieco et al., 2012), ainsi que la répartition de la masse de gaz dans les halos (Pillepich et al., 2018), peuvent différer significativement d’un modèle à un autre. Certaines simulations, comme EAGLE (Schaller et al., 2015), NIHAO (Wang et al., 2015), Auriga (Grand et al., 2017) et Magneticum (Teklu et al., 2018), prédisent des galaxies analogues à la Voie lactée (MW-analogue) avec des courbes de rotation qui sont plates à l’extérieur jusqu’à quelques kpc du centre, en accord qualitatif avec les observations. En revanche, Illustris (Genel et al., 2014; Vogelsberger et al., 2014) prédit des courbes de rotation qui augmentent en dessous d’environ 15 kpc.

Dans cette étude, j’ai utilisé les simulations hydrodynamiques cosmologiques du projet EAGLE (Schaye et al., 2015; Crain et al., 2015) pour examiner la forme, l’alignement et la distribution de masse de la matière noire et de la matière baryonique. EAGLE est bien adapté à cette application car il offre une résolution spatiale et de masse élevée, ce qui permet une analyse détaillée de la forme des galaxies et de leur environnement. De plus, EAGLE couvre une large gamme de volumes cosmologiques, nous permettant d’explorer différentes échelles. Cette diversité de volumes offre également une taille d’échantillon importante, ce qui permet d’obtenir des statistiques significatives pour identifier les tendances globales dans les propriétés étudiées. De plus, EAGLE a été calibré en utilisant des observations comme la fonction de masse stellaire (*galaxy stellar mass function*,

GSMF) (González Delgado et al., 2014), assurant ainsi une correspondance précise avec les données réelles. Cela nous assure que les simulations de lentilles gravitationnelles de quasars que j’ai produits reflètent suffisamment la réalité.

III.2 Mesure des halos de la simulation EAGLE

La suite de simulations hydrodynamiques cosmologiques EAGLE (Schaye et al., 2015; Crain et al., 2015) a eu pour objectif de comprendre la formation et l’évolution des galaxies. Ce qui distingue EAGLE des jeux de simulations précédents, c’est sa capacité à reproduire un large éventail de propriétés observées des galaxies, en particulier à des échelles plus petites, typiquement de l’ordre de quelques kiloparsecs. La simulation EAGLE utilise une résolution spatiale et de masse plus élevée par rapport à de nombreuses simulations précédentes, avec une résolution de quelques centaines de parsecs, et des modèles physiques plus détaillés pour capturer les processus tels que la formation des étoiles, la croissance des trous noirs supermassifs et l’effet du rayonnement ultraviolet et des rayons X sur l’environnement galactique. Les simulations cosmologiques comme EAGLE, qui considèrent à la fois la matière noire et la matière baryonique, permettent de modéliser de manière réaliste la distribution de masse des galaxies.

La simulation EAGLE est réalisée à l’aide d’une version modifiée du code GADGET-3 tree-SPH (Springel et al., 2005), qui permet de gérer à la fois la dynamique de la matière noire et l’évolution des baryons. Le code utilise une méthode de raffinement de maillage basée sur des sous-grilles pour capturer les processus physiques spécifiques. Par exemple, le refroidissement radiatif est considéré en utilisant le modèle décrit par Wiersma et al. (2009), qui tient compte de la composition multi-éléments du gaz interstellaire. La formation d’étoiles est modélisée en suivant l’approche de Schaye and Dalla Vecchia (2008), qui considère les processus tels que la formation des disques galactiques et la régulation de la formation stellaire par la rétroaction des supernovae. L’enrichissement en métaux est également inclus dans la simulation en se basant sur le modèle de Wiersma et al. (2009). De plus, la simulation EAGLE intègre la formation de trous noirs supermassifs, qui est traitée en suivant les travaux de Springel et al. (2005) et Rosas-Guevara et al. (2015), en incluant la croissance des trous noirs par accrétion de matière. La rétroaction des étoiles massives et des AGNs est également traitée, ce qui permet de modéliser son impact sur la formation et l’évolution des galaxies (Dalla Vecchia and Schaye, 2012; Schaye et al., 2015). Dans la simulation, les galaxies sont identifiées en utilisant des critères basés sur les propriétés des particules, telles que la densité, la température et la formation stellaire. La base de données publique EAGLE, qui comprend la base de données des galaxies EAGLE (McAlpine et al., 2016) et les données sur les particules¹ EAGLE (The EAGLE team, 2017), fournit un accès aux résultats de la simulation pour la communauté scientifique.

1. <http://icc.dur.ac.uk/Eagle/>

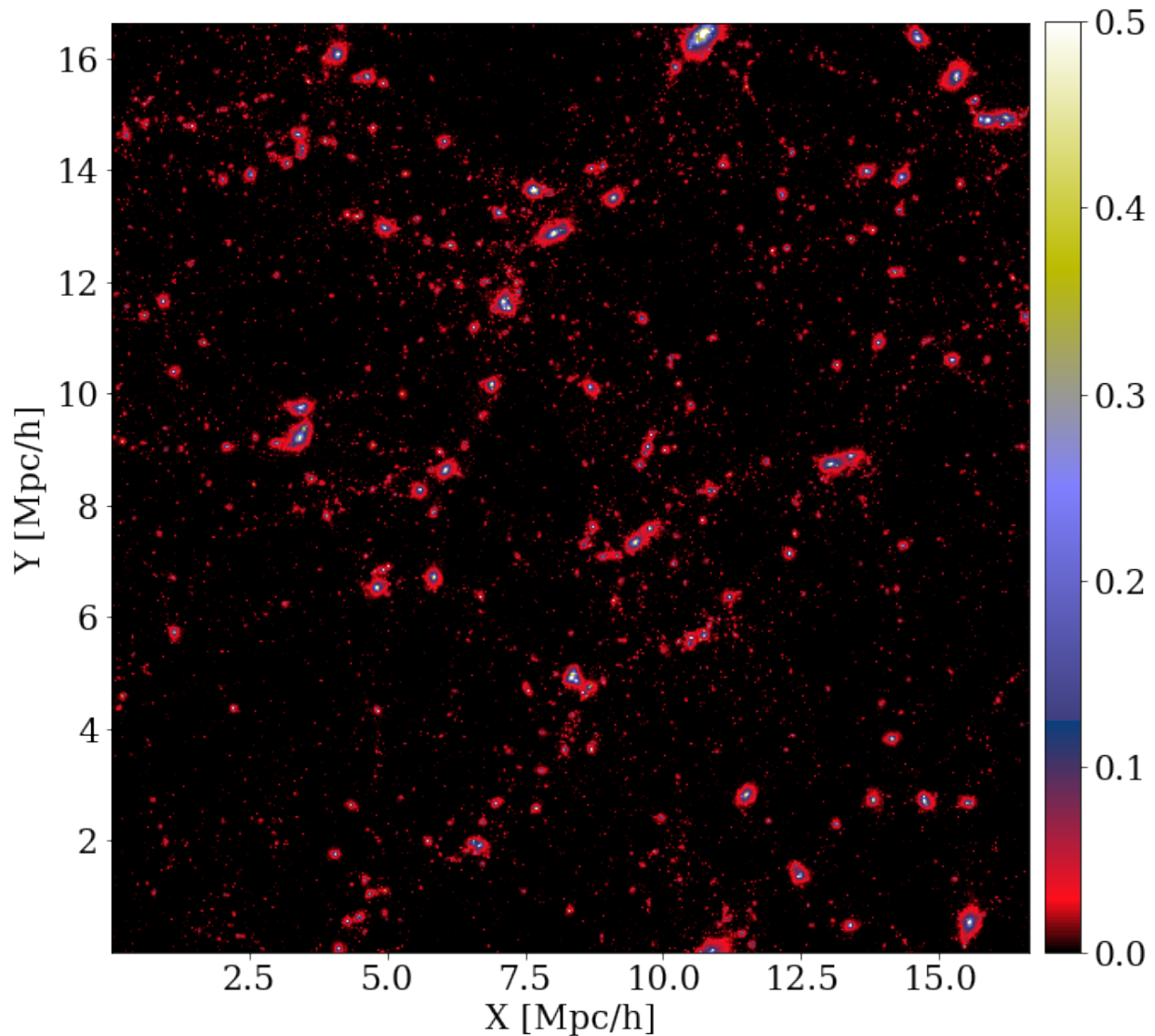


FIGURE III.1 – Projection du cube de la simulation EAGLE sur les axes XY au redshift $z = 0,5$.

Chaque simulation EAGLE comprend une série de 29 instantanés couvrant une large plage d'époques de 0,26 Gyr ($z=15$) à 13,46 Gyr ($z=0$). Les simulations EAGLE ont été soigneusement calibrées pour reproduire la fonction de masse stellaire observée des galaxies à $z = 0$, tout en respectant des contraintes réalistes en termes de taille, de masse, de rayon, de taux de formation d'étoiles, de métallicité et de contenu en gaz. Ainsi, les simulations EAGLE fournissent une résolution suffisante de l'ordre de quelques kiloparsec pour étudier individuellement les caractéristiques des galaxies simulées et pour effectuer des analyses détaillées.

La Figure [III.1](#) illustre la projection sur les axes XY du cube de la simulation EAGLE de référence

RefL0025N0376 à un redshift de 0,5. Cette représentation met en évidence la structure à grande échelle de l'univers simulé, caractérisée par la présence de filaments délimitant des régions vides et à l'intersection desquels se trouvent des nœuds riches en galaxies, les amas de galaxies. Cette distribution est un exemple de la notion de "toile cosmique" évoquée dans les travaux antérieurs (Bond et al., 1996; Faucher-Giguère et al., 2008). Ce motif est principalement déterminé par les fluctuations initiales de densité dans l'univers primordial, qui évoluent de manière non linéaire pour former le réseau filamentaire observé. Dans la simulation EAGLE, qui considère à la fois la matière noire et la matière baryonique, cette structure à grande échelle est reproduite en intégrant les processus physiques détaillés de formation des galaxies.

Dans ce travail, j'ai analysé les simulations de référence EAGLE, RefL0025N0376 et RefL0025N0752. Ce choix est motivé par la nécessité de tester intensivement mon code d'analyse tout en limitant le temps de calcul ce qui est un aspect important compte tenu de la complexité des simulations et des ressources informatiques disponibles. Ces simulations suivent un volume périodique cubique de côté $L = 25$ cMpc. Le volume contient initialement respectivement 376^3 et 752^3 particules de gaz et de matière noire avec des masses initiales respectives de $1.81 \times 10^6 M_\odot$ et $9.7 \times 10^6 M_\odot$. Ces conditions initiales ont été choisies en utilisant les valeurs des paramètres cosmologiques dérivées de la publication des résultats initiaux de la mission Planck sur le fond de rayonnement cosmologique (Planck Collaboration, 2014), à savoir $\Omega_0 = 0.307$, $\Omega_b \simeq 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 \simeq 0.83$, $n_s = 0.9611$, et $h \simeq 0,68$, $Y = 0,248$. Le rapport entre les particules baryoniques et non baryoniques est donc égal à $(\Omega_0 - \Omega_b)/\Omega_b \equiv (m_{DM}/m_b) \simeq 5,4$.

Dans la simulation EAGLE, les étoiles se forment à partir de particules de gaz lorsque la densité locale de gaz dépasse un seuil donné (voir Schaye et al., 2015, pour une présentation détaillée). Étant donné la résolution limitée de la simulation, il est nécessaire d'imposer un seuil de formation d'étoiles au-delà duquel une phase de gaz interstellaire froide doit se former. Cela signifie que, lorsque la densité de gaz dépasse ce seuil, le gaz commence à se condenser et à former des nuages denses propices à la formation d'étoiles. La probabilité qu'une particule de gaz soit convertie en une particule d'étoile dépend de la pression (Schaye, 2004; Schaye and Dalla Vecchia, 2008) et de la métallicité de la particule de gaz. En d'autres termes, la pression exercée par le gaz et sa composition métallique influencent la probabilité de formation d'étoiles.

L'évolution des étoiles dans la simulation est suivie en utilisant des durées de vie dépendantes de leur métallicité, basées sur les travaux de Portinari et al. (1998). Les étoiles perdent progressivement de la masse au cours de leur évolution et cette masse perdue est redistribuée aux particules de gaz voisines. La métallicité des particules étoile est utilisée pour les étiqueter dans la classification de Harvard B, A et G. Les vents de ces étoiles, ainsi que les supernovae à effondrement de cœur, jouent un rôle important dans cette redistribution en injectant de l'énergie et du moment dans le milieu interstellaire. Les rendements nucléosynthétiques décrits par Marigo (2001) sont utilisés pour

déterminer la quantité de masse perdue par les étoiles et les quantités d'éléments chimiques qui sont libérées dans le milieu interstellaire.

Il est également important de noter que les étoiles en tant que population contribuent à un fluide non collisionnel, ce qui signifie que leur dynamique n'est pas gouvernée par des collisions individuelles entre les étoiles elles-mêmes. Cela a des implications importantes pour la formation et l'évolution des structures de l'univers. Ces processus de formation et d'évolution des étoiles, y compris les effets des vents stellaires, du rayonnement et des supernovae, jouent un rôle crucial dans la régulation de la formation des galaxies et dans la redistribution de la masse et des éléments chimiques dans le milieu interstellaire. Ils sont notamment cruciaux pour comprendre la formation des étoiles massives, qui ont un impact significatif sur la dynamique et l'évolution des galaxies.

Dans la pratique, nous évitons de qualifier les structures de la simulation EAGLE en tant que galaxies. Nous préférons utiliser le terme plus général de "halo". Plusieurs études antérieures ont déjà examiné divers aspects de la morphologie des halos dans les simulations EAGLE (Velliscig et al., 2015; Shao et al., 2016; Hill et al., 2021). Notre étude vise à compléter ces travaux en apportant une nouvelle méthode pour déterminer les paramètres de forme et les orientations des halos de matière noire, du gaz et des étoiles. En utilisant des ellipsoïdes pour modéliser directement les formes des composants de matière noire, de gaz et d'étoiles dans un espace tridimensionnel, il est possible de suivre l'évolution de ces formes au fil du temps et de suivre leurs variations en fonction de la masse totale du halo.

De plus, j'ai mesuré les formes projetées sur le plan du ciel (2D) afin de les comparer aux observations, notamment les images de galaxies observées. En évaluant les ellipticités projetées des composants de matière noire, de gaz et d'étoiles des galaxies simulées, on obtient la distribution statistique des ellipticités. Cette distribution statistique est un paramètre essentiel pour la simulation des lentilles gravitationnelles, car elle permet de reproduire les caractéristiques observées des lentilles gravitationnelles et d'effectuer des comparaisons quantitatives avec les observations.

Il est important de distinguer la fréquence des ellipticités, c'est-à-dire la proportion d'objets ayant une certaine ellipticité, de la répartition des ellipticités selon d'autres paramètres, tels que la masse totale du halo. En étudiant la distribution statistique des ellipticités selon différents paramètres, nous pouvons obtenir des informations précieuses sur la formation et l'évolution des galaxies, ainsi que sur les processus physiques qui influencent leur morphologie.

La masse des particules de matière noire dans ces simulations reste constante tout au long de la simulation, avec une valeur de $m_{\text{DM}} = 9.7 \times 10^6 M_{\odot}$. Bien que les particules de matière noire n'interagissent pas directement entre elles, elles s'agrègent sous l'influence mutuelle de leur attraction gravitationnelle, ce qui permet la formation de structures cosmiques. Cependant, en raison de leur caractère non collisionnel, les particules de matière noire ne peuvent former des structures avec une

morphologie très piquée ou spécifique à une physique particulière.

Au cours de la simulation, certaines particules de gaz peuvent accumuler une masse supérieure à leur masse initiale, établie à $m_g = 1.81 \times 10^6 M_\odot$. Cette accumulation de masse est principalement due à la conversion de ces particules de gaz en étoiles. En effet, les particules de gaz ont une forte probabilité de se transformer en étoiles plutôt que d'accumuler davantage de masse. Cette conversion en étoiles est régulée par des processus physiques tels que le refroidissement radiatif, la formation d'étoiles et les retours d'énergie des étoiles massives et des supernovae.

III.2.1 Sélection des halos

Les structures de matière noire sont initialement identifiées à l'aide de l'algorithme "friends-of-friends" (FoF) avec une longueur de liaison fixée à 0,2 fois l'espacement moyen entre les particules (Davis et al., 1985). Les particules autres que la matière noire, telles que celles associées au gaz, aux étoiles et aux trous noirs, sont regroupées avec les particules de matière noire les plus proches.

Les structures gravitationnellement liées au sein des groupes FoF sont identifiées grâce à l'algorithme SUBFIND (Springel et al., 2001; Dolag et al., 2009). SUBFIND fonctionne en calculant une estimation locale de la densité aux positions de toutes les particules du groupe. Il considère toute région localement surdense dans ce champ de densité comme une sous-structure candidate. Pour trouver ces régions, SUBFIND abaisse lentement un seuil de densité globale dans le champ de densité. La plupart du temps, les régions surdenses isolées ne font que croître en taille à mesure que le seuil est abaissé, sauf lorsqu'il y a fusion de deux régions distinctes pour former un domaine commun. Lors de ces fusions, les contours de deux régions distinctes se rejoignent en un point-selle, ce qui modifie la topologie du contour d'isodensité. Ainsi, les régions entourées par un contour d'isodensité traversant un point-selle sont considérées comme des sous-structures candidates, également appelées "sous-halos".

Dans le cadre de cette étude, les sous-halos sont définis comme des satellites du halo principal. Le sous-halo présentant le potentiel gravitationnel le plus grand au sein d'un groupe FoF est désigné comme le "halo principal". Toute particule liée au groupe mais ne faisant pas partie d'un sous-halo est attribuée au halo principal.

Les halos EAGLE ont des formes et des tailles variées et ont souvent des structures complexes. Afin d'analyser toutes les simulations et de modéliser chaque halo quelle que soit sa forme, j'ai développé une procédure d'analyse spécifique. Cette procédure a été appliquée pour chaque type de particule (matière noire, gaz, étoiles) ainsi que pour la somme des composantes baryoniques et pour l'ensemble des particules combinées. J'ai également effectué une analyse de la morphologie des halos projetés sur les plans [xy], [yz] et [xz] afin de dériver la distribution statistique des paramètres morphologiques comme s'ils étaient projetés sur un plan tangent à la sphère céleste. La distribution

de masse projetée est un paramètre clef de la modélisation des lentilles gravitationnelles².

III.2.2 Mesurer les paramètres de forme des halos

Dans la pratique, les halos d'EAGLE présentent une structure complexe et souvent irrégulière sous forme d'un amas composé d'un halo principal et de sous-halos satellites. La figure III.2 présente un exemple de cette structure. J'ai représenté ici l'ensemble des particules de matière noire, de gaz, d'étoiles et de trous noirs du halo no25 de la petite simulation RefL0025N0376 à un redshift $z=0,5$. Pour visualiser les caractéristiques de ce halo, j'ai deux perspectives : une projection sur le plan [XY] de EAGLE (panneau de gauche) et une projection sur le plan [YZ] (panneau de droite). Cette représentation inclut un halo principal et des sous-halos qui forment des agrégats de matière au sein et à la périphérie du halo principal.

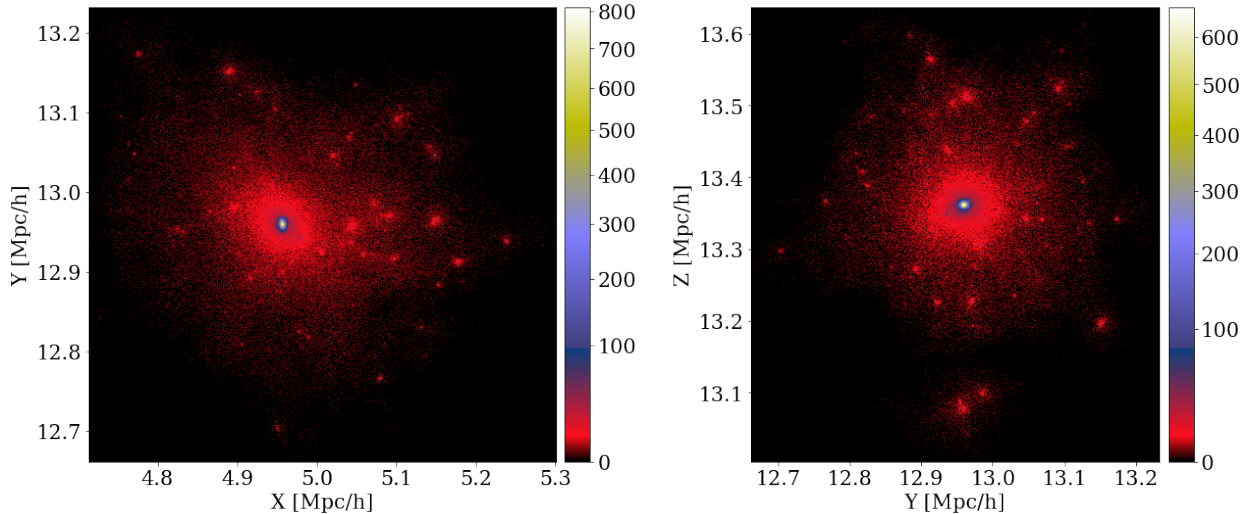


FIGURE III.2 – Projection bidimensionnelle du groupe de halos no. 25 de la simulation RefL0025N0376 au redshift $z=0,5$. Les coordonnées sont dans le système de référence initial de la simulation EAGLE (X,Y,Z). Tous les types de particules sont représentés. Pour rendre les sous-structures visibles, une échelle de couleur logarithmique a été appliquée. Chaque image fait 500x500 pixels et la couleur représente le nombre de particules dans chaque pixel. Notez que les petits points brillants visibles dans l'amas ne doivent pas être confondus avec du bruit - ce sont en fait des sous-halos liés gravitationnellement et correspondant aux noyaux survivants de halos qui sont tombés dans ce halo à un moment antérieur.

Au cours du processus de formation des halos, ceux-ci ne sont pas des entités isolées, mais des systèmes dynamiques en interaction avec leur environnement à plus grande échelle. Ces interactions et les coalescences entre structures qui suivent donnent lieu à la formation de halos pouvant présenter des structures complexes en fonction du rapport de masse entre les objets fusionnant et l'ancienneté

2. Dans le cadre de l'approximation de la lentille mince, détails dans le chapitre IV

de cette fusion. Par exemple, lorsqu'une fusion de halos se produit, les particules des halos et des sous-halos se mélangent et se redistribuent, créant ainsi des structures complexes comprenant souvent des régions de faible densité très étalées qui témoignent de la formation hiérarchique parfois mouvementée des halos.

Le défi que j'ai dû relever réside dans le fait que les sous-structures présentes dans les halos peuvent exercer une influence significative sur la mesure de leur morphologie, même si elles contribuent proportionnellement peu à la masse totale du halo. Cela dépend du choix entre moments géométriques et moments barycentriques pour évaluer la forme générale des halos. Les sous-structures étalées peuvent avoir un impact non négligeable sur les moments géométriques, tandis que les moments barycentriques peuvent être moins sensibles à ces structures étalées. Par conséquent, pour mesurer la forme générale des halos et tenir compte des structures très étalées, la méthode d'analyse utilisée ne doit pas être trop sensible aux structures étalées et doit considérer les caractéristiques globales du halo.

La figure III.3 montre la proportion que représentent les sous-halos dans les halos d'EAGLE en fonction de la masse totale de ces halos. La majorité des halos (89%) ont une masse totale inférieure à $10^{10} M_{\odot}$. Parmi ces halos, 85% d'entre-eux ont une part de sous-halos qui représente moins de 20% de la masse totale du halo. Cette figure montre aussi qu'au sein de la simulation il existe des sous-halos qui représentent une masse importante par rapport à la masse totale. Pour les cas où la masse totale est faible, $M_{tot} < 10^{10} M_{\odot}$, les sous-halos sont constitués essentiellement de particules de matière noire et parfois de gaz. Ils ne contiennent en général pas plus de 50 particules d'étoiles. Nous avons alors des halos principaux accompagnés de sous-halos de matière noire importants qui influenceront sur la mesure de la distribution de masse du halo principal.

Lorsque la masse totale du halo est élevée, $M_{tot} > 10^{10} M_{\odot}$, il arrive parfois qu'un ou plusieurs sous-halos massifs constituent une part significative de la masse totale, atteignant jusqu'à près de 60% de ce total. Dans ces situations, on considère que ces sous-halos, combinés au halo principal, forment un regroupement de halos. La modélisation des lentilles gravitationnelles induites par ces regroupements de halos nécessiterait donc la prise en compte de plusieurs distributions de masses distinctes. Ayant choisi de concentrer notre étude sur les simulations de lentilles gravitationnelles avec une seule distribution de masse, j'ai appliqué un filtre qui exclut ces cas particuliers, représentant moins de 3% de l'ensemble des halos, pour ne retenir que ceux où la masse des sous-halos ne dépasse pas 20% de la masse totale des halos.

La méthode habituelle pour étudier la forme des halos consiste à les analyser à l'intérieur d'une ouverture circulaire fixe (Thob et al., 2019; Hill et al., 2021). Afin d'analyser toutes les simulations et de modéliser chaque halo quelle que soit sa forme, tout en limitant l'influence des structures de faible densité très étalées qui se trouvent principalement en périphérie du halo principal, j'ai développé une nouvelle approche itérative. Cette approche vise à exclure ces structures de grande taille tout en

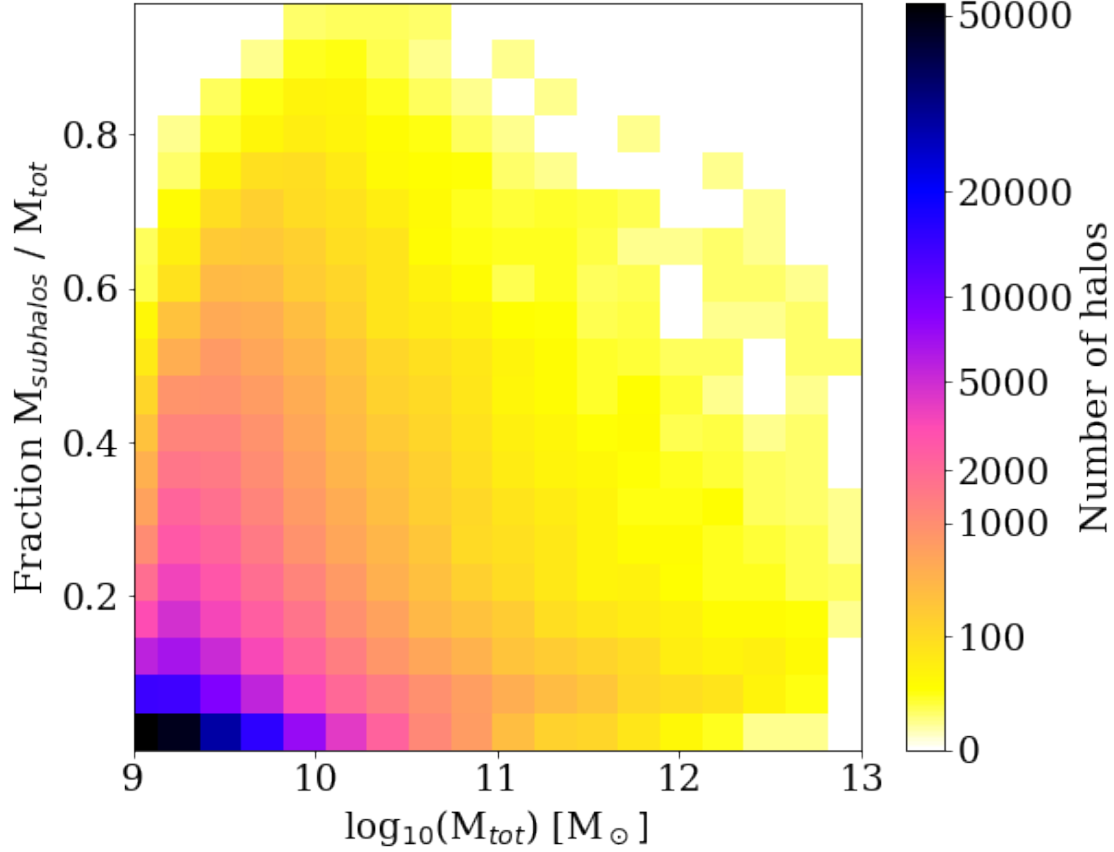


FIGURE III.3 – Rapport de la masse des sous-halos sur la masse totale du halo en fonction de la masse totale des halos pour l’ensemble de la simulation RefL0025N0376. L’échelle de couleur est normalisée par une loi de puissance par un exposant $\gamma = 0.3$.

préservant la majeure partie de la masse des halos. Ma méthode repose sur trois étapes essentielles. Tout d’abord, j’ai identifié avec précision le centre des halos. Ensuite, j’ai déterminé une sphère de travail qui permet d’appliquer mon analyse tout en excluant les structures étendues de faible masse situées aux bords du halo principal. Enfin, l’analyse est itérée plusieurs fois à partir des résultats de chaque itération précédente.

III.2.2.1 Trouver la position du centre des halos

Mon approche commence par une détermination précise du centre du halo principal en tenant compte de toutes les particules constituant le halo, tout en négligeant l’influence des sous-structures périphériques. Dans cette étude, j’ai choisi de définir le centre des halos comme étant l’emplacement du pic de densité de particules le plus élevé. Ce choix est motivé par l’objectif de définir une référence pour la distribution de masse qui sera utilisée ultérieurement lors des simulations de lentilles gravitationnelles. Si j’avais choisi le centre de masse, par exemple, j’aurais étudié la distribution de

masse centrée sur le centre de masse du système halo+sous-halos, ce qui n'aurait pas été utilisable pour les simulations de lentilles simples que je projette de réaliser.

Prenons l'exemple du halo no. 17 pour illustrer. La figure III.4 présente les projections XY et YZ du halo, mettant en évidence à la fois le halo central et plusieurs sous-halos massifs. Si j'appliquais la méthode du centre de masse, j'identifierai l'intersection des droites bleues. Cependant, notre objectif est d'étudier des cas plus simples, où un seul de ces halos agit en tant que lentille gravitationnelle pour un quasar donné. Ainsi, ce qui nous intéresse principalement est la morphologie et la distribution de masse centrées sur le halo principal dont le centre médian est affiché à l'intersection des droites blanches. Cette approche permet de mesurer la distribution de masse, influencée par les sous-halos, tout en maintenant une configuration de lentille gravitationnelle simple.

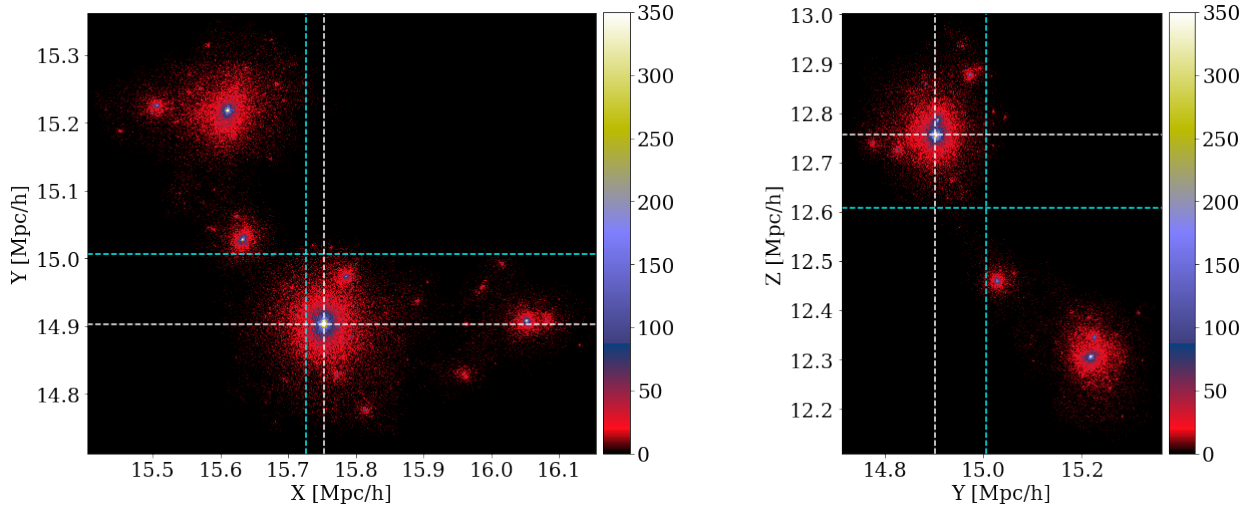


FIGURE III.4 – Projection bidimensionnelle du groupe de halos no. 17 de la simulation RefL0025N0376 au redshift $z = 0,5$. Les coordonnées sont dans le système de référence initial de la simulation EAGLE (X,Y,Z). Chaque image fait 500x500 pixels et la couleur représente le nombre de particules dans chaque pixel; tous les types de particules sont considérés. On distingue le halo principal au centre et plusieurs sous-halos massifs périphériques. L'intersection des droites bleues représente le centre de masse de l'ensemble halo+sous-halos et l'intersection des droites blanches représente le centre du halo principal.

Pour déterminer le centre du halo, j'ai procédé comme suit : tout d'abord, j'ai calculé les coordonnées médianes (x_0, y_0, z_0) de l'ensemble des particules. Ensuite, une nouvelle position médiane (x_0', y_0', z_0') est dérivée des particules englobées dans une sphère centrée sur (x_0, y_0, z_0) et de rayon égal au dixième de la distance de la particule la plus éloignée par rapport à la première estimation du centre médian précédent. Cette deuxième étape est répétée en utilisant les coordonnées du nouveau centre (x_0', y_0', z_0') jusqu'à ce que le critère de convergence soit atteint, c'est-à-dire lorsque la variation entre deux itérations successives est inférieure à 5 kpc. La position médiane finale des

particules restantes est alors considérée comme le centre définitif du halo. Pour vérifier la convergence du processus, un contrôle visuel a été effectué lors du développement de la méthode sur plusieurs dizaines de halos. En pratique, la procédure nécessite en général trois itérations pour atteindre une convergence satisfaisante. Dans la figure III.5, on observe l'évolution de l'estimation du centre du halo entre la première et la dernière itération, mettant en évidence l'importance des itérations pour obtenir une estimation précise du centre.

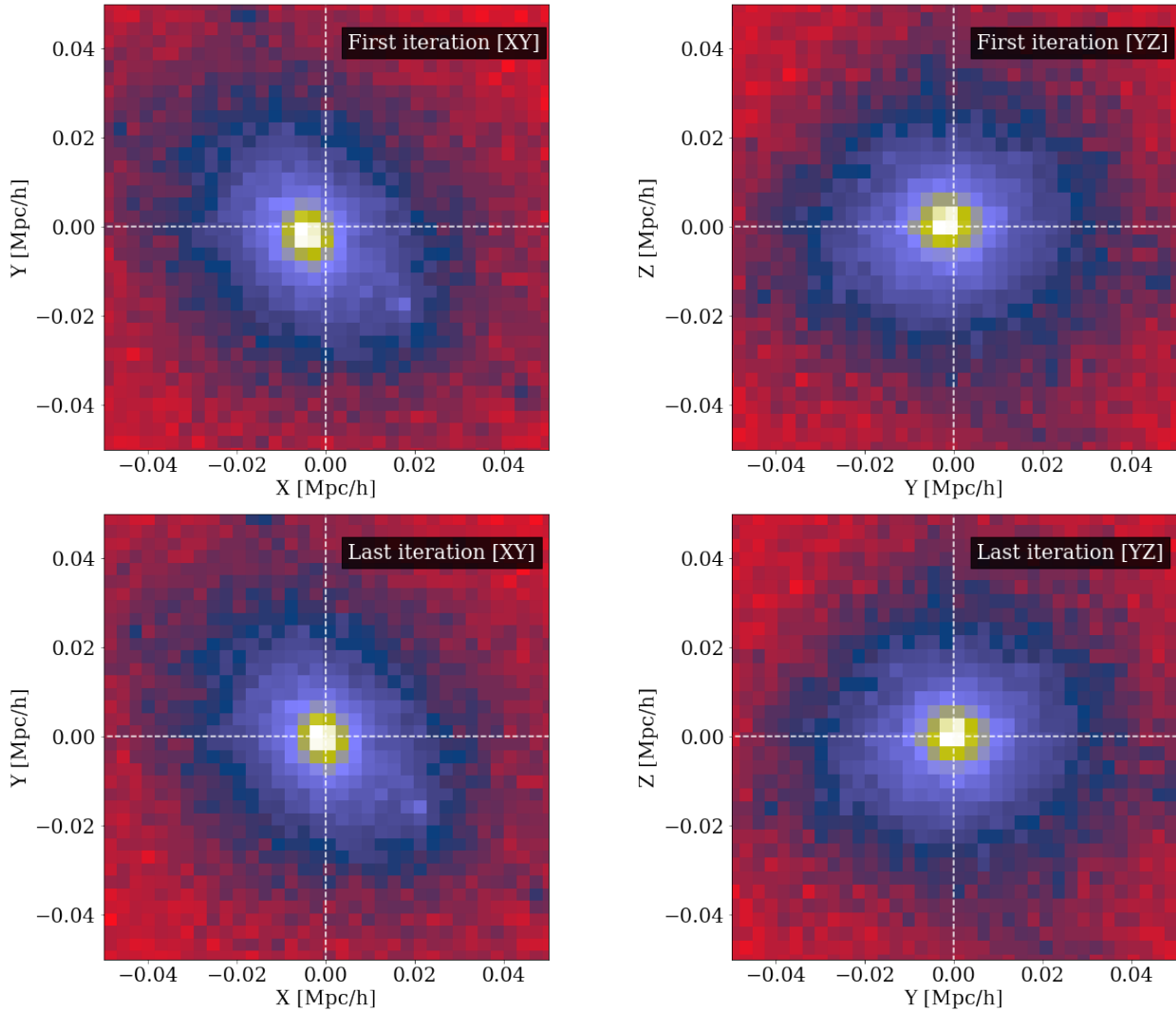


FIGURE III.5 – Zoom sur le centre du halo principal no. 25 de la simulation RefL0025N0376 au redshift $z = 0,5$ projeté dans le plan (XY) à gauche et dans le plan (YZ) à droite. (Haut) Première itération pour identifier le centre du halo. Les coordonnées ont été ramenés à l'origine pour centrer le halo en $(0, 0, 0)$. (Bas) Troisième et dernière itération qui identifie correctement le pic de densité du halo.

III.2.2.2 Calcul d'une sphère de travail

Pour modéliser la morphologie des halos, j'ai utilisé un modèle d'ellipsoïde dont les paramètres libres sont calculés à l'aide d'une Analyse en Composantes Principales (ACP). L'ACP permet de quantifier la forme en accédant aux moments de degré 2 de la distribution. Ainsi, les valeurs propres obtenues grâce à l'ACP reflètent les caractéristiques morphologiques du halo et son orientation.

La détermination du centre des halos au pic de densité permet de réaliser l'étude de la distribution de masse dans une sphère de travail. En choisissant une sphère de travail, nous définissons une région spécifique du halo à étudier, en excluant les structures périphériques. Le fait de choisir une sphère de travail détermine les particules incluses dans l'analyse, ce qui peut influencer la distribution des particules et donc la matrice de covariance de l'ACP. Si la sphère de travail est bien choisie et capture adéquatement la morphologie du halo, les valeurs propres obtenues refléteront les demi-axes de l'ellipsoïde ajusté au halo. Le choix de la taille de la sphère de travail est donc crucial. Si la sphère est trop petite, elle peut ne pas capturer suffisamment de particules pour représenter fidèlement la morphologie globale du halo. D'autre part, si la sphère est trop grande, elle peut inclure des zones que nous ne voulons pas dans notre analyse faussant les estimations.

Je détermine la sphère de travail par son rayon, r_{80cc} , la sphère qui englobe une masse suffisante sans inclure les sous-structures à grande distance du centre et peu massives. Sur la base de tests, j'ai choisi empiriquement de définir r_{80cc} comme le rayon auquel la densité a chuté de 80% par rapport à la densité d'un noyau central. Pour dériver r_{80cc} , il est nécessaire d'établir le profil de densité du halo. Pour cela, je dois d'abord déterminer la taille du noyau central du halo, r_{noyau} , que je définis comme le rayon qui contient 10% de la masse totale du halo.

Pour mesurer r_{noyau} , j'ai calculé une courbe de croissance de la masse cumulée du halo depuis le centre vers la périphérie. Pour construire cette courbe de croissance, j'additionne la masse des particules dans des couronnes successives de rayon égal à 1/50 de l'extension spatiale totale du halo. La figure [III.6](#) montre la courbe de croissance mesurée pour le halo no. 20. Lorsque la masse dérivée ne s'écarte pas de la moyenne des trois dernières mesures de plus de 1σ , correspondant à la déviation standard des trois dernières mesures, le processus s'arrête; on considère que l'asymptote a été atteinte. Les irrégularités sur la pente de la courbe traduisent la présence des agrégats de matières (les sous-halos) dans le halo principal ou à son voisinage direct. À partir de cette courbe, il est facile de déterminer le rayon central du halo, englobant 10% de la masse totale asymptotique, en effectuant une simple interpolation linéaire.

La mesure de r_{noyau} permet d'estimer la densité du cœur du halo et de construire le profil de densité radiale du halo. Le rayon auquel la densité chute de 80% par rapport à la densité du noyau central, r_{80cc} , est ainsi mesuré via une interpolation linéaire. Ce rayon définit la sphère dans laquelle l'analyse de la morphologie sera effectuée.

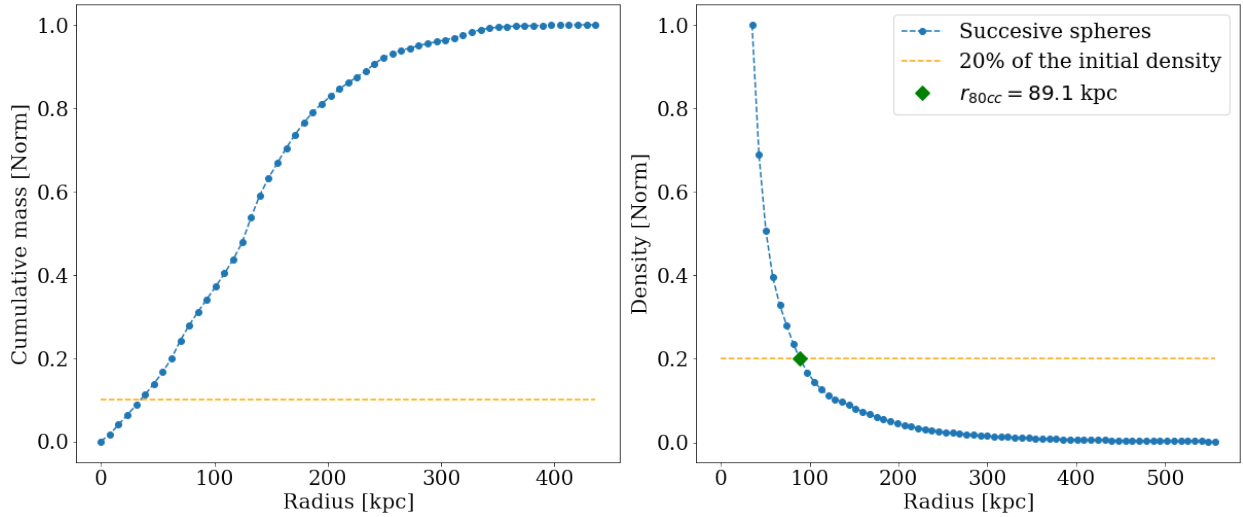


FIGURE III.6 – Figure illustrant le processus de calcul de la sphère de travail pour le halo no. 20. (Gauche) La courbe de croissance de la masse du halo, où nous mesurons le rayon du noyau central, noté r_{noyau} , correspondant à 10% de la masse totale asymptotique. (Droite) Le profil de densité construit à partir du rayon du noyau central. Le rayon de la sphère de travail, noté $r_{80\text{cc}}$, correspond au rayon où la densité a diminué à 20% de sa valeur initiale.

Il convient de mentionner deux cas spécifiques qui nécessitent une attention particulière lors de l’application de la méthode décrite dans cette section. Le premier cas concerne la composante gaz des halos. Dans certains cas, la détermination du profil de densité ne fonctionne pas. En raison d’une valeur élevée de la turbulence du gaz suite aux collisions associées à la formation des structures, il peut arriver que ce gaz ne présente pas un noyau central identifiable. Par conséquent, si la présence d’un noyau central n’est pas identifiable du fait d’une densité qui augmente avec le rayon, l’ensemble du halo est exclu de l’analyse.

Le deuxième cas concerne les petits halos, notamment les halos d’étoiles et ceux avec un faible nombre de particules. Dans de tels cas, j’ai fixé une limite inférieure de 30 kpc pour le rayon de la sphère de travail, $r_{80\text{cc}}$. Cette valeur de 30 kpc est couramment utilisée dans l’étude de la morphologie des halos dans les simulations cosmologiques (Thob et al., 2019; Hill et al., 2021), car elle englobe la majeure partie de la matière baryonique.

III.2.2.3 Ajustement d’un ellipsoïde

Comme mentionné précédemment, la forme choisie pour caractériser les halos d’EAGLE est un ellipsoïde. Les paramètres de cet ellipsoïde sont a , b et c , respectivement le demi-grand, intermédiaire et petit axe. Ces paramètres sont déterminés par une analyse en composantes principales (ACP). A partir de ces valeurs, on définit les rapports d’axes q_i qui sont respectivement b/a , c/a et c/b pour $i \in [1; 3]$. On définit aussi la sphéricité $S = c/a$. Ainsi, $S = 0$ correspond à un disque parfaitement

plat et $S = 1$ correspond à une sphère parfaite.

Une fois que la première estimation des demi-axes est obtenue, j'ai appliqué plusieurs fois l'ACP aux particules sélectionnées à l'intérieur d'un ellipsoïde défini par les demi-axes obtenus lors de l'itération précédente. Cela me permet d'obtenir de nouvelles estimations des demi-axes. Après un nombre limité d'itérations (maximum 10 itérations), je mets fin aux itérations et obtiens les estimations finales des demi-axes.

Dans le cas d'une distribution tridimensionnelle de particules, les axes principaux de la distribution définissent un nouveau repère orthogonal (X', Y', Z') . Ce repère a subi une rotation par rapport au repère EAGLE initial (X, Y, Z) définie par trois angles d'Euler (ψ, θ, ϕ) qui sont donnés par la matrice de rotation \mathbf{P} . Cette matrice permet de convertir la position des particules donnée dans le référentiel de simulation EAGLE (X, Y, Z) vers le nouveau référentiel défini par les axes principaux (X', Y', Z') de la distribution.

La Figure III.7 représente le rayon de la sphère de travail obtenue pour le gaz, les étoiles et l'ensemble des deux composantes du halo no.11. Dans cette représentation, le halo est aligné sur ses axes principaux et centré sur l'origine. On distingue le gaz qui forme des bras. Notre méthode d'analyse nous permet de déterminer avec précision l'axe d'aplatissement du gaz, nous pouvons clairement observer la vue de côté du disque dans la projection YZ. De plus, le rayon de la sphère de travail s'ajuste de manière adéquate en fonction du type de particule, ce qui facilite une étude détaillée des paramètres morphologiques spécifiques à chaque composante.

III.2.3 Bootstrap pour estimer les erreurs sur la morphologie

Il est crucial de connaître l'impact du nombre de particules sur la précision de la récupération des paramètres morphologiques des halos (une distribution trop clairsemée peut conduire à une estimation erronée) ainsi que l'influence de la sphéricité du halo (une distribution se rapprochant d'une sphère peut être difficile à analyser). Afin d'explorer la dépendance vis-à-vis du nombre de particules, j'ai effectué des simulations de 10 000 halos synthétiques contenant de 10 à 2000 particules, en maintenant une sphéricité constante. J'ai réalisé cette étude pour deux valeurs de sphéricité différentes : $c_{sim}/a_{sim} = 0,72$ (halos plutôt sphériques) dans un cas, et $c_{sim}/a_{sim} = 0,25$ (halos aplatis en forme de disque) dans l'autre.

Pour simuler les halos, j'ai tiré aléatoirement les coordonnées des particules à partir d'une distribution normale multivariée avec une moyenne $\vec{\mu} = [\mu_x, \mu_y, \mu_z]$ et une matrice de covariance $\Sigma = diag(a_{sim}^2, b_{sim}^2, c_{sim}^2)$. J'ai ensuite fait tourner toute la distribution des particules de trois angles d'Euler choisis au hasard $(\psi_{sim}, \theta_{sim}$ et $\phi_{sim})$ afin de simuler des halos orientés aléatoirement.

Sur la figure III.8, je présente les erreurs relatives³ ($\Delta a, \Delta b, \Delta c$) et la différence entre les angles

3. $\Delta x = (x - x_{sim})/x_{sim}$

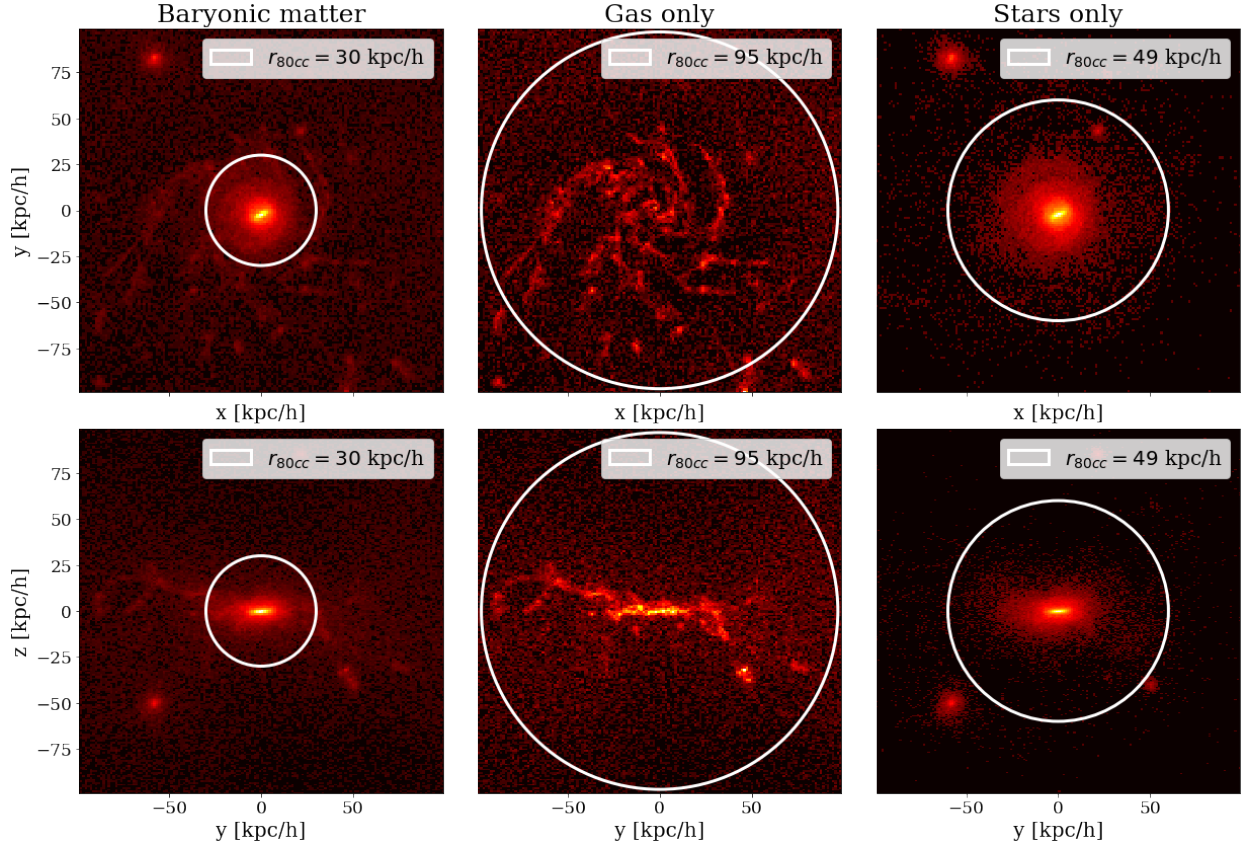


FIGURE III.7 – Projections sur les axes principaux ($X'Y'$) et ($Y'Z'$) du groupe de halos no. 11 de la simulation de référence RefL0025N0376. Dans la première colonne, nous observons la somme des baryons, c'est-à-dire le gaz et les étoiles. La colonne du milieu représente spécifiquement le gaz, tandis que la colonne de droite est dédiée aux étoiles. Chaque projection est accompagnée de la représentation du rayon de la sphère de travail correspondant à chaque type de particule.

estimés (ψ, θ, ϕ) et leur vraie valeur ($\psi_{sim}, \theta_{sim}, \phi_{sim}$) en fonction du nombre de particules et pour deux sphéricités. Les mesures des demi-axes sont généralement bons à un niveau meilleur que 10%, et généralement à un niveau meilleur que 5% pour les halos de plus de 100 particules, quelle que soit la sphéricité. Il y a une dégradation visible de la mesure des demi-axes lorsque le nombre de particules est inférieur à 100. Le demi-axe 'c' a alors tendance à être légèrement sous-estimé de plus de 5% pour les halos de moins de 100 particules. L'orientation des halos est globalement bien récupérée à mieux que 10 degrés dans tous les cas. La sphéricité des halos influence la qualité de l'orientation récupérée comme prévu : pour les halos quasi-sphériques avec plus de 100 particules, les erreurs sont de 8 à 10 degrés ou moins, tandis que pour les halos aplatis, elles sont d'environ 2 à 3 degrés.

Pour explorer cette fois-ci la dépendance de la qualité de l'ajustement à la sphéricité des halos, j'ai

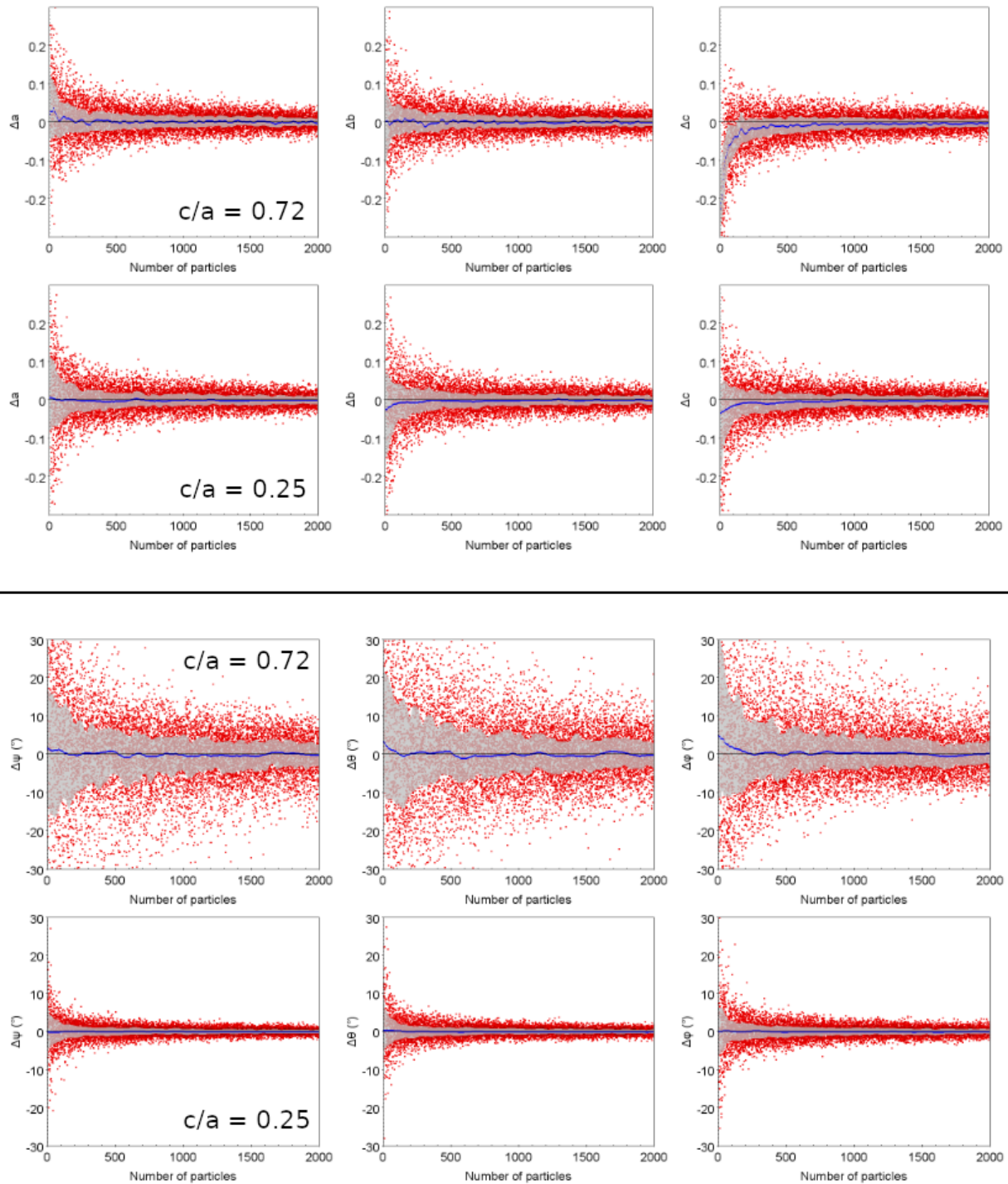


FIGURE III.8 – Analyse de 10 000 halos simulés. Différence relative (Δa , Δb , Δc) des demi-axes récupérés (a,b,c) (panneau supérieur) et différence des angles (ψ_{sim} , θ_{sim} , ϕ_{sim}) par rapport à leur vraie valeur (panneau inférieur) en fonction du nombre de particules pour des halos quasi-sphériques ($c_{sim}/a_{sim}=0,72$) et des halos aplatis ($c_{sim}/a_{sim}=0,25$). Les lignes bleues indiquent la médiane des distributions et la zone grise montre les quartiles.

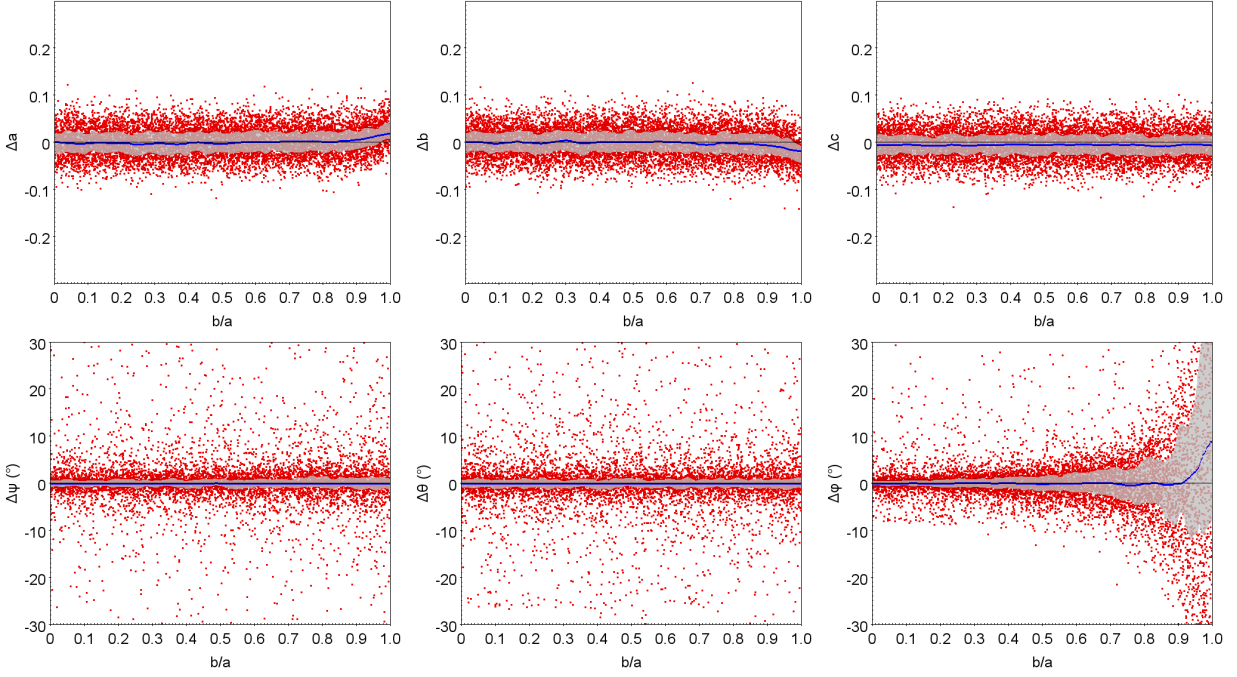


FIGURE III.9 – Analyse de 10 000 halos simulés de 500 particules chacun. Ligne supérieure : différences relatives (Δa , Δb , Δc) dans les demi-axes récupérés (a,b,c) en fonction du rapport d’axe (b/a). Ligne inférieure : Différence entre les angles récupérés (ψ , θ , ϕ) et leurs vraies valeurs simulées exprimées en deg en fonction du rapport d’axe (b/a). Les lignes bleues indiquent la médiane des distributions et la zone grise montre les quartiles.

simulé 10 000 nouveaux halos synthétiques de 500 particules avec des rapports d’axe choisis au hasard tels que $(b_{sim}/a_{sim}) < 1$ et $(c_{sim}/b_{sim}) < 1$. J’ai ainsi obtenu des estimations de leurs demi-axes (a, b, c) et de leurs angles d’Euler (ψ , θ , ϕ). La figure [III.9](#) permet d’observer que les demi-axes (a,b,c) sont bien récupérés. L’orientation des halos est également bien mesurée avec des erreurs relatives non influencées par la sphéricité du halo, sauf pour ϕ (la troisième rotation). La distribution des quartiles de $\Delta\phi$ indique une dégradation sévère (erreurs jusqu’à 10 deg et plus) pour les sphéricités supérieures à 0,9 avec une surestimation systématique de ϕ .

L’estimation de la valeur des demi-axes principaux est bonne lorsque le nombre de particules incluses dans les halos est supérieur à 100. Les demi-axes sont généralement récupérés à mieux que 5%. Les angles d’Euler des halos sont bien récupérés à mieux que 10 degrés pour les halos avec un nombre de particules supérieur à 100, mais leur récupération dépend de la sphéricité du halo, se dégradant pour les halos quasi-sphériques avec $c/a > 0,9$.

III.2.4 Courbe de croissance en masse des halos

Une fois que la morphologie du halo est établie, il devient possible de construire la courbe de croissance en masse qui décrit la distribution de masse du halo à partir de couronnes respectant cette fois la forme de l'ellipsoïde ajusté à la distribution des particules. Le calcul de la courbe de croissance s'arrête lorsque celle-ci atteint une asymptote, c'est-à-dire lorsque la masse dérivée ne s'écarte pas de la moyenne des trois dernières mesures de plus de 1σ . La Figure III.10 présente plusieurs courbes de croissance en masse pour les halos 18, 22, 36 et 39. Le rayon fait référence au demi-grand axe a exprimé en kpc/h. Les couleurs sont associées aux différents types de particules. Il est intéressant de noter que les particules correspondant aux étoiles se trouvent principalement concentrées dans le cœur des halos. Ce comportement sera discuté plus en détail dans la section suivante. On peut observer que la pente des courbes de croissance change parfois de manière abrupte. Ces changements sont la signature de la présence de sous-halos massifs à proximité ou dans le halo principal, sur lequel cette étude est centrée.

Dans les études de classification des galaxies, la concentration de lumière (Conselice, 2003) est largement utilisée pour caractériser la distribution de la luminosité des galaxies et permettre la distinction entre les galaxies de type précoce, dominées par leur bulbe, et les galaxies de type tardif. En me basant sur cette définition de la concentration de lumière, je peux calculer la concentration de masse selon le même principe. La concentration de masse peut être estimée en calculant le rapport entre le rayon à 80% et le rayon à 20% de la masse totale du halo étudié (eq. III.3). Cette mesure de concentration donne une indication de la manière dont la masse est répartie dans le halo. Une valeur plus élevée de concentration indique une concentration accrue de la matière dans le noyau du halo.

$$C = 5 \times \log \left(\frac{r_{80}}{r_{20}} \right), \quad (\text{III.3})$$

III.2.5 Code et environnement de calcul

J'ai élaboré une méthodologie complète afin d'étudier en détail la forme, l'orientation et la répartition de la masse au sein des halos de la simulation cosmologique EAGLE. Toutes les étapes que j'ai décrites précédemment sont appliquées de manière systématique à chacune des composantes individuelles des halos. Cette approche permet de comparer les morphologies et les alignements spécifiques de chaque composante entre elles. De plus, une variante de cette méthode a été développée pour les projections en deux dimensions. Cela permet d'analyser les caractéristiques morphologiques et les alignements non seulement dans l'espace tridimensionnel, mais également dans les projections en deux dimensions. Ainsi, nous obtenons une compréhension exhaustive de la structure interne des halos dans toutes ses dimensions.

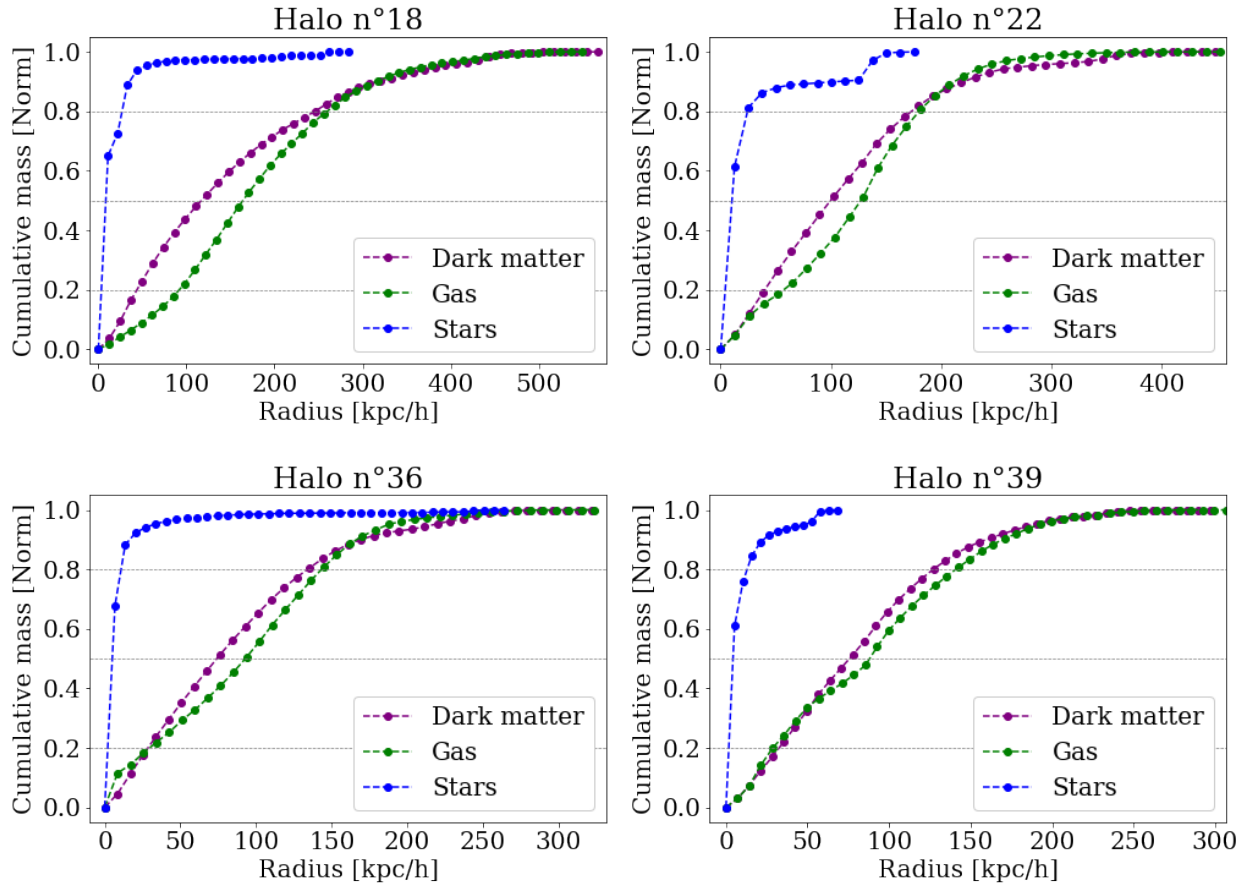


FIGURE III.10 – Courbes de croissance en masse de quatre halos extraits de la simulation EAGLE RefL0025N0376 à un redshift $z=0,5$. Le rayon désigne le demi-axe le plus grand a exprimé en kpc/h. Les couleurs violet, vert et bleu représentent respectivement les particules de matière noire, de gaz et d'étoiles constituant chaque halo.

La Figure [III.11](#) illustre le schéma du programme que j'ai développé au cours de ma thèse, offrant ainsi une vue d'ensemble de son fonctionnement. Après avoir extrait les données de la simulation EAGLE, je génère un sous-catalogue dédié à chaque instantané de la simulation. Ce sous-catalogue est préparé spécifiquement pour être compatible avec MOGPY (Morphology Of Galaxies in Python⁴), un package Python que j'ai développé. MOGPY met à disposition l'ensemble des outils et des méthodes que j'ai mis au point pour estimer la forme, l'orientation et la distribution de masse des halos de EAGLE. Il convient de noter que ce package peut également être adapté aisément à d'autres simulations cosmologiques telles que IllustrisTNG (Vogelsberger et al., 2014).

L'application de MOGPY a présenté une difficulté majeure liée à la gestion des ressources de calcul. Les deux catalogues de la simulation EAGLE que j'ai analysés occupent respectivement 224 Go

4. <https://github.com/Ethhj/MOGPy>

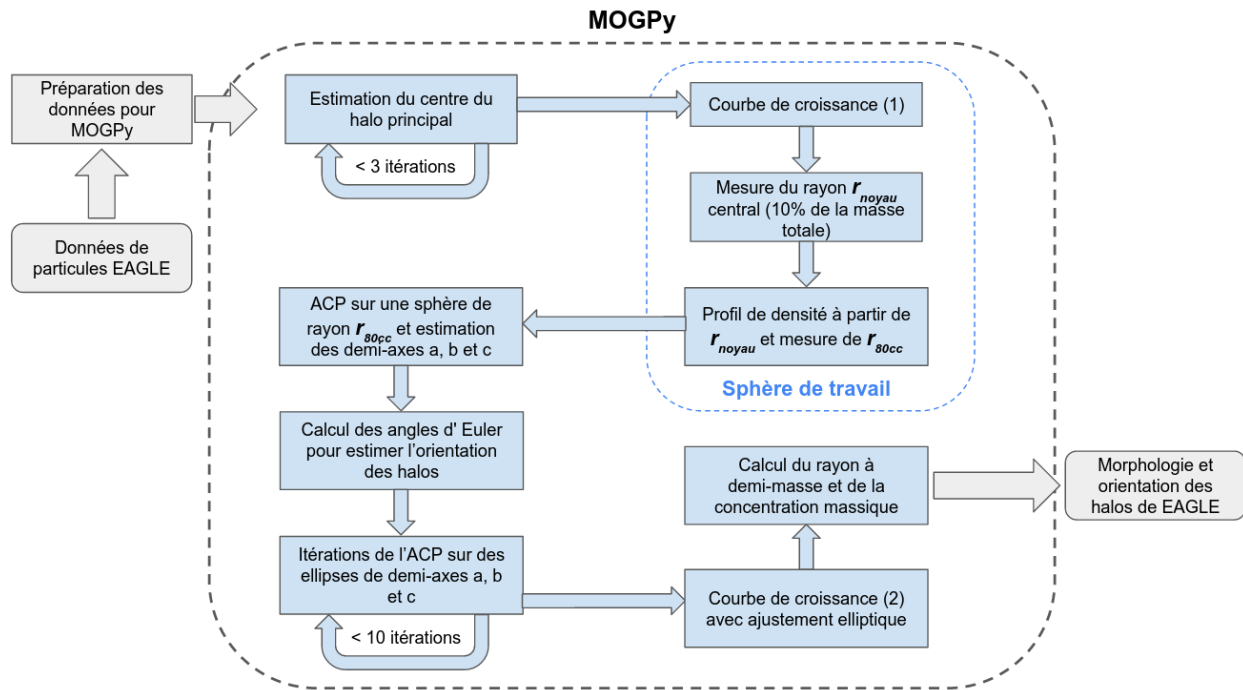


FIGURE III.11 – Plan schématique du programme que j’ai développé pour étudier la forme, l’orientation et la distribution de masse des halos d’EAGLE. Les données en entrée et en sortie sont tabulaires.

et 1,6 To d’espace, avec chaque instantané occupant entre 8 et 58 Go. La taille des fichiers a donc nécessité une gestion rigoureuse lors du chargement des données en mémoire. Pour surmonter ces obstacles, les analyses ont principalement été exécutées sur des serveurs de calcul dédiés. Un travail considérable a été réalisé pour permettre le lancement simultané de plusieurs analyses sur ces serveurs afin de réduire le temps de calcul. L’architecture du code a été spécifiquement conçue pour traiter individuellement chaque composante et chaque halo, offrant ainsi la possibilité de choisir l’ordre d’exécution et d’effectuer une parallélisation partielle ou totale. Malgré ces développements, les catalogues de plus grande taille n’ont pas pu être entièrement analysés pendant la phase de développement, en raison de contraintes temporelles.

L’application de la méthodologie décrite dans cette partie m’a permis de mesurer et d’analyser en détail la morphologie, l’orientation et la distribution de masse des halos d’EAGLE. Dans la section suivante, je présente les résultats obtenus pour les simulations de basse et haute résolution, RefL0025N0376 et RefL0025N0752, de la simulation cosmologique EAGLE.

III.3 Distribution statistique des paramètres des halos d'EAGLE

Grâce au développement et l'application de la méthodologie décrite dans la section précédente, j'ai pu produire une publication scientifique [Petit et al. \(2023\)](#) qui porte sur la simulation RefL0025N0376. Depuis, j'ai complété ces mesures en terminant l'analyse de la simulation RefL0025N0752 pour obtenir un total de 340 719 halos dont les résultats seront présentés dans cette section.

J'ai restreint toutes les analyses aux halos contenant au moins 100 particules afin d'obtenir des estimations fiables des paramètres, comme expliqué dans la section [III.2.3](#). J'ai examiné deux scénarios pour chacun des halos étudiés. Dans le premier scénario, j'ai mesuré les caractéristiques en incluant à la fois le halo principal et les sous-halos associés. Dans le second scénario, je me suis concentré uniquement sur le halo principal. Cette approche m'a permis d'analyser un ensemble de 681 438 cas avec (i) 485 796 halos de matière noire, (ii) 151 894 halos de gaz, et (iii) 17 808 halos d'étoiles. Parmi les 151 894 halos de gaz, 134 548 ne contiennent pas un minimum de 100 étoiles, tandis que 17 346 halos en contiennent plus de 100. Je distinguerai ces deux types de halos de gaz respectivement comme "w/o stars" (sans étoiles) et "w/ stars" (avec étoiles) pour les halos de gaz. Pour commencer, examinons la distribution de ces halos en termes de masse.

III.3.1 Distribution en masse des halos dans les simulations de EAGLE

La figure [III.12](#) présente l'histogramme des masses des 681 438 halos étudiés. J'ai représenté la masse totale ainsi que la masse de chaque composant. Il convient de noter qu'il existe peu de trous noirs massifs dans chaque halo, et leur présence n'affecte donc pas directement la forme générale des halos. Par conséquent, ils sont considérés uniquement dans l'analyse de la masse totale des halos. La masse totale des halos est clairement dominée par la matière noire, qui représente en moyenne 94,2% de la masse totale des halos. Le gaz et les étoiles constituent respectivement 5,7% et 0,1% de la masse totale, répartie de manière significativement moindre.

Du côté des halos de gaz, la masse des halos qui contiennent moins de 100 étoiles est plus faible que les halos qui en contiennent plus de 100. Les étoiles, pour se former, ont besoin que des particules de gaz dépassent un seuil de densité donné. Les halos qui n'atteignent pas cette densité ne forment donc que très peu ou pas d'étoiles.

III.3.2 Morphologie des halos

Mon étude s'est concentrée sur l'analyse des rapports d'axes, b/a et c/a , qui décrivent la morphologie des halos à travers l'ajustement d'un ellipsoïde aux données. J'ai examiné ces rapports d'axes dans un contexte général, en prenant en compte l'ensemble des halos dans les simulations EAGLE. Cette approche permet d'obtenir une vision d'ensemble de la forme des halos dans les simulations et de comprendre leur distribution et leur variabilité. Ensuite, je me suis intéressé plus particulièrement à

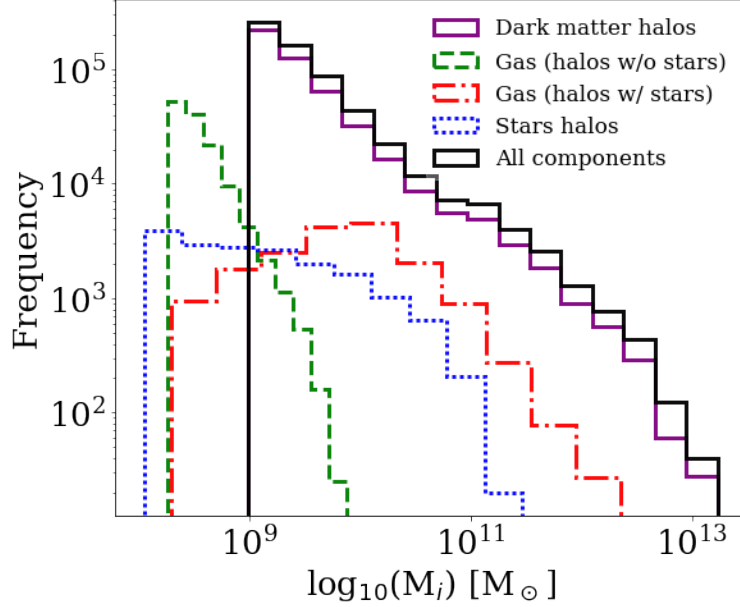


FIGURE III.12 – Distribution de masse de 681 438 halos EAGLE pour tous les redshifts et pour chaque type de particule ($i \in [\text{DM}, \text{Gaz}, \text{Étoiles}, \text{Total}]$). Sur l’axe des ordonnées, nous avons le nombre de halos, tandis que sur l’axe des abscisses, la masse de chaque halo est indiquée pour chaque type de particule. La ligne continue noire représente la masse totale des halos en prenant en compte la masse de toutes les particules, y compris les trous noirs présents dans la simulation. La ligne continue mauve correspond à la composante de matière noire. Les lignes rouges et vertes correspondent respectivement aux halos de gaz contenant suffisamment de particules d’étoiles et aux halos de gaz sans une quantité suffisante de particules d’étoiles. Enfin, la ligne bleue représente les halos d’étoiles.

l’étude de la morphologie des halos projetés sur le plan du ciel telle qu’elle est observée par les grands relevés. Pour ce faire, j’ai étudié les halos projetés dans les plans XY, XZ et YZ. Cette approche réaliste est essentielle pour la modélisation des galaxies déflectrices dans les lentilles gravitationnelles fortes⁵, où la connaissance précise de l’ellipticité de la galaxie responsable de l’effet de lentille est cruciale.

III.3.2.1 La répartition de la masse des halos en 3D

J’ai commencé par étudier le cas en trois dimensions des données. La Figure III.13 présente l’histogramme des rapports d’axes b/a (panneau de gauche) et c/a (panneau de droite) des halos analysés. La Table III.1 résume les caractéristiques médianes des distributions des rapports d’axes selon le type de particule. Nous observons tout d’abord que l’ensemble des halos de la simulation EAGLE est bien décrit par une médiane b/a d’environ 0,8. En ce qui concerne la sphéricité c/a ,

5. Seulement pour la modélisation du déflecteur dans le cadre de l’approximation de la lentille mince, plus de détails dans le chapitre IV

nous constatons qu'elle est davantage centrée autour de la médiane c/a d'environ 0,5, mais nous remarquons également que les halos de gaz contenant un nombre important d'étoiles (halos w/ stars) présentent un aplatissement beaucoup plus prononcé que les autres composantes, avec une médiane de sphéricité c/a de 0,3. Ces halos de gaz se caractérisent par une forme ellipsoïdale plus aplatie. Nous avons mentionné précédemment que ces halos de gaz ont une masse plus élevée que les autres halos de gaz. Cela suggère que les halos de gaz qui se développent suffisamment, que ce soit par accrétion ou par fusion avec d'autres halos, tendent à gagner en aplatissement au fur et à mesure de leur croissance.

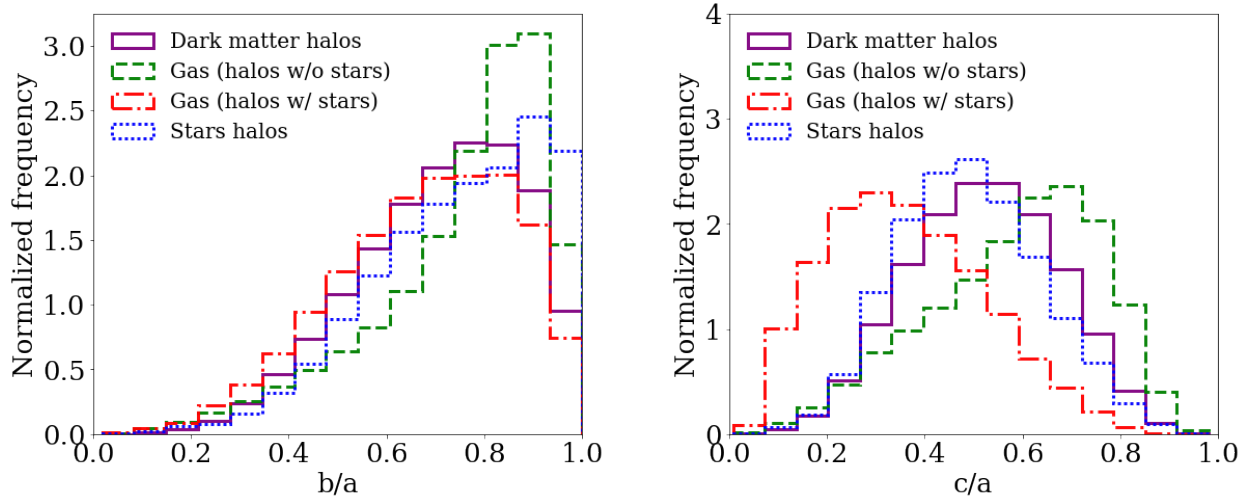


FIGURE III.13 – Distribution des rapports d'axes b/a (panneau de gauche) et c/a (panneau du droite) selon le type de particule. Les halos DM (485 796) sont représentés en rose. Les halos de gaz sont divisés en deux groupes : les halos de gaz sans étoile (134 548) en vert et les halos de gaz avec étoiles (17 346) en rouge. Les halos d'étoiles (17 808) sont représentés en bleu.

TABLE III.1 – Analyse des 485 796 halos des simulations RefL0025N0376 et RefL0025N0752. La médiane et l'écart absolu médian (MAD) des rapports d'axes b/a et c/a sont fournis selon le type de particule contenue dans les halos : DM (485 796 halos), gaz (151 894 halos w/o étoiles, 17 346 w/ étoiles) et étoiles (17 808 halos).

	DM	Gas (w/ stars)	Gas (w/o stars)	Stars
b/a	0.73 ± 0.18	0.69 ± 0.15	0.80 ± 0.15	0.77 ± 0.19
c/a	0.52 ± 0.16	0.35 ± 0.17	0.61 ± 0.18	0.48 ± 0.15

Néanmoins, cette représentation de la distribution des rapports d'axes ne permet pas d'examiner la répartition de ces rapports en fonction de la masse des halos. Dans la Figure III.14, j'ai tracé la relation entre les rapports d'axes b/a et c/a en fonction de la masse totale des halos. J'ai divisé la figure en deux niveaux : en haut, j'ai présenté les halos de matière noire et d'étoiles, tandis qu'en

bas, j'ai affiché les halos de gaz avec et sans une quantité suffisante d'étoiles. Sur chaque panneau, j'ai représenté les rapports d'axes correspondant à l'ensemble des particules.

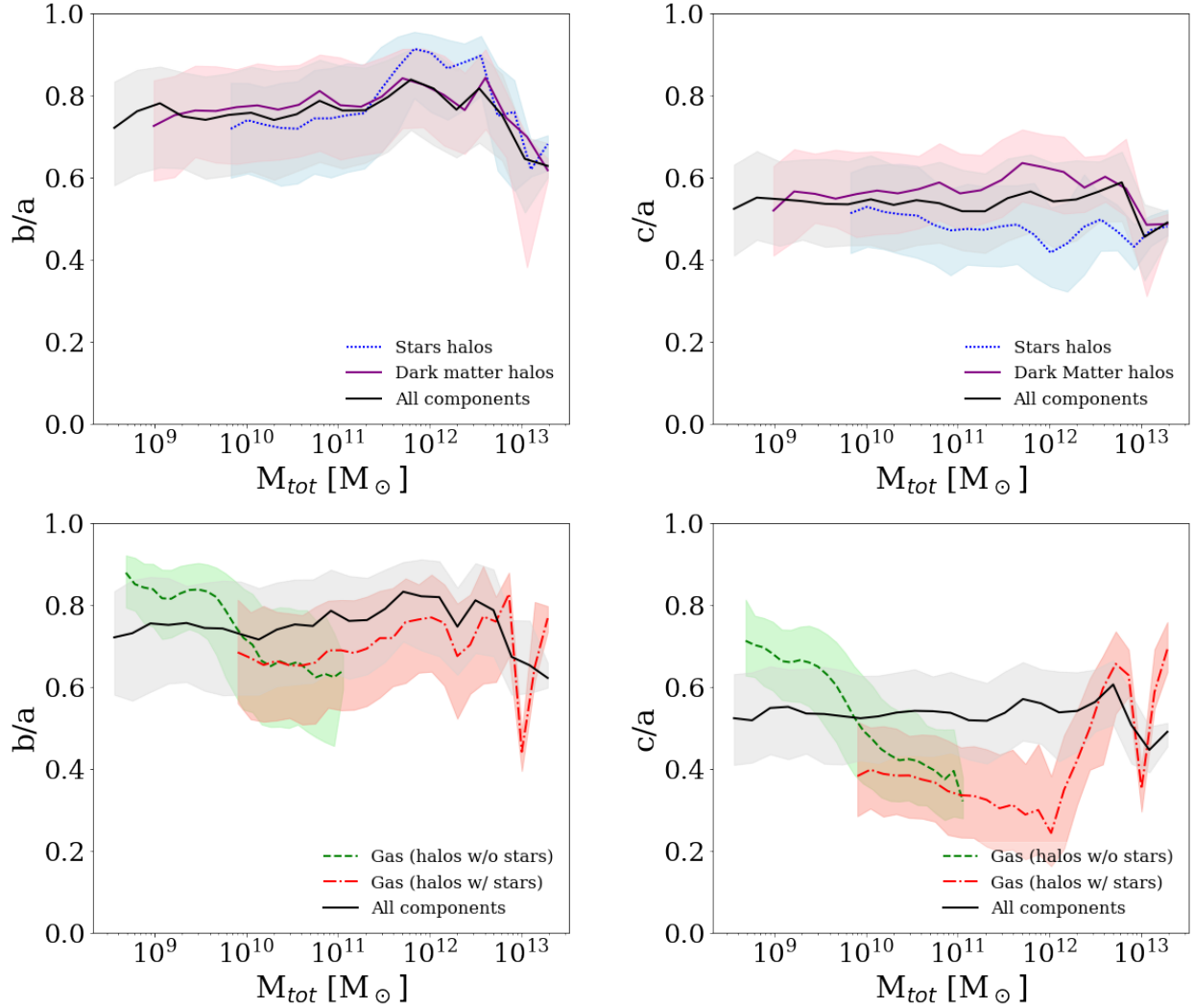


FIGURE III.14 – Figure illustrant les rapports d'axes de l'ellipsoïde d'ajustement en fonction de la masse totale des halos. Les lignes indiquent les valeurs médianes tandis que l'ombrage indique les 25e - 75e quartiles. (Haut) Les halos de matières noire sont représentés en mauve trait plein et les étoiles sont représentées en bleu avec une ligne en petits points. (Bas) Les halos de gaz sont représentés selon qu'ils accompagnent un halo de 100 particules d'étoiles (en rouge) ou non (en vert). Le trait plein en noir représente les rapports d'axe de tous les composants ensemble.

Cette figure révèle plusieurs points importants. Tout d'abord, il est clairement visible que la matière noire dicte la morphologie générale des halos, étant donné qu'elle représente la majeure partie de leur masse. De plus, nous pouvons remarquer que la répartition des étoiles est légèrement plus aplatie que celle des halos de matière noire qui les contiennent. Cependant, cet aplatissement est moins prononcé

que celui du gaz. La dépendance de l’aplatissement du gaz en fonction de la masse totale des halos est clairement visible à travers le rapport c/a . Nous constatons que ce rapport d’axes diminue de $c/a = 0,8$ à $c/a = 0,3$ pour les halos ayant une masse d’environ $10^{12} M_{\odot}$. La transition entre les halos ayant moins de 100 particules d’étoiles et ceux en ayant davantage se situe entre une masse de $10^{10} M_{\odot}$ et $10^{11} M_{\odot}$. Enfin, nous observons une brusque augmentation du rapport d’axes c/a pour les halos très massifs de gaz ($M > 10^{13} M_{\odot}$). Nous émettons l’hypothèse que la formation de ces halos extrêmement massifs est le résultat de fusions de halos, et que lors de ces fusions, l’aplatissement des halos de gaz est perturbé, expliquant ainsi cette augmentation soudaine du rapport d’axes c/a .

La figure III.15 illustre l’évolution des rapports d’axes des halos en fonction du redshift. Nous observons une tendance générale des rapports d’axes à devenir de plus en plus sphériques à mesure que le redshift diminue, indépendamment de la masse et du type de particule. Ces résultats sont en accord avec les études menées par Allgood et al. (2006) et Tenneti et al. (2014), où des tendances similaires ont été observées.

Plus précisément, nos résultats confirment ceux de Allgood et al. (2006) et Tenneti et al. (2014) pour les halos de matière noire, montrant que les rapports d’axes moyens, tels que b/a et c/a , augmentent à mesure que nous nous approchons du temps présent ($z = 0$). Cette évolution n’est pas influencée par la physique baryonique, car elle est également observée dans les simulations ne considérant que la matière noire (Bryan et al., 2013; Tenneti et al., 2014). Les halos se forment par l’accrétion de matière le long des filaments de la toile cosmique. Ce processus entraîne une anisotropie initiale, avec une virialisation progressive qui conduit finalement à un ellipsoïde pour ces particules non collisionnelles. Au fur et à mesure que le temps s’écoule, ces halos tendent à se rapprocher d’une forme plus sphérique.

Concernant la matière stellaire, nous observons une corrélation générale entre les rapports d’axes moyens et ceux de la matière noire. Les particules d’étoiles, tout comme les particules de matière noire, se comportent comme des particules non collisionnelles. Elles évoluent dans le potentiel gravitationnel défini par le halo dominé par la matière noire. Ainsi, les rapports d’axes des halos de matière stellaire suivent généralement les tendances observées pour les halos de matière noire.

Cependant, une exception notable se manifeste dans l’évolution des rapports d’axes des halos de gaz massifs ($M_{\odot} > 10^{11}$). Dans ce cas, nous observons une tendance inverse pour le rapport c/a . Ces halos ont tendance à s’aplatir de plus en plus avec le temps, ce qui se traduit par des rapports d’axes de plus en plus petits. Cette observation suggère que les processus spécifiques liés au gaz, tels que la pression et les interactions hydrodynamiques, peuvent avoir un impact significatif sur la forme des halos de gaz massifs. Ces effets peuvent modifier l’évolution des rapports d’axes par rapport aux halos de matière noire et de matière stellaire qui vont avoir tendance à rester plus ou moins sphérique (Schaller, 2015; Cataldi et al., 2021).

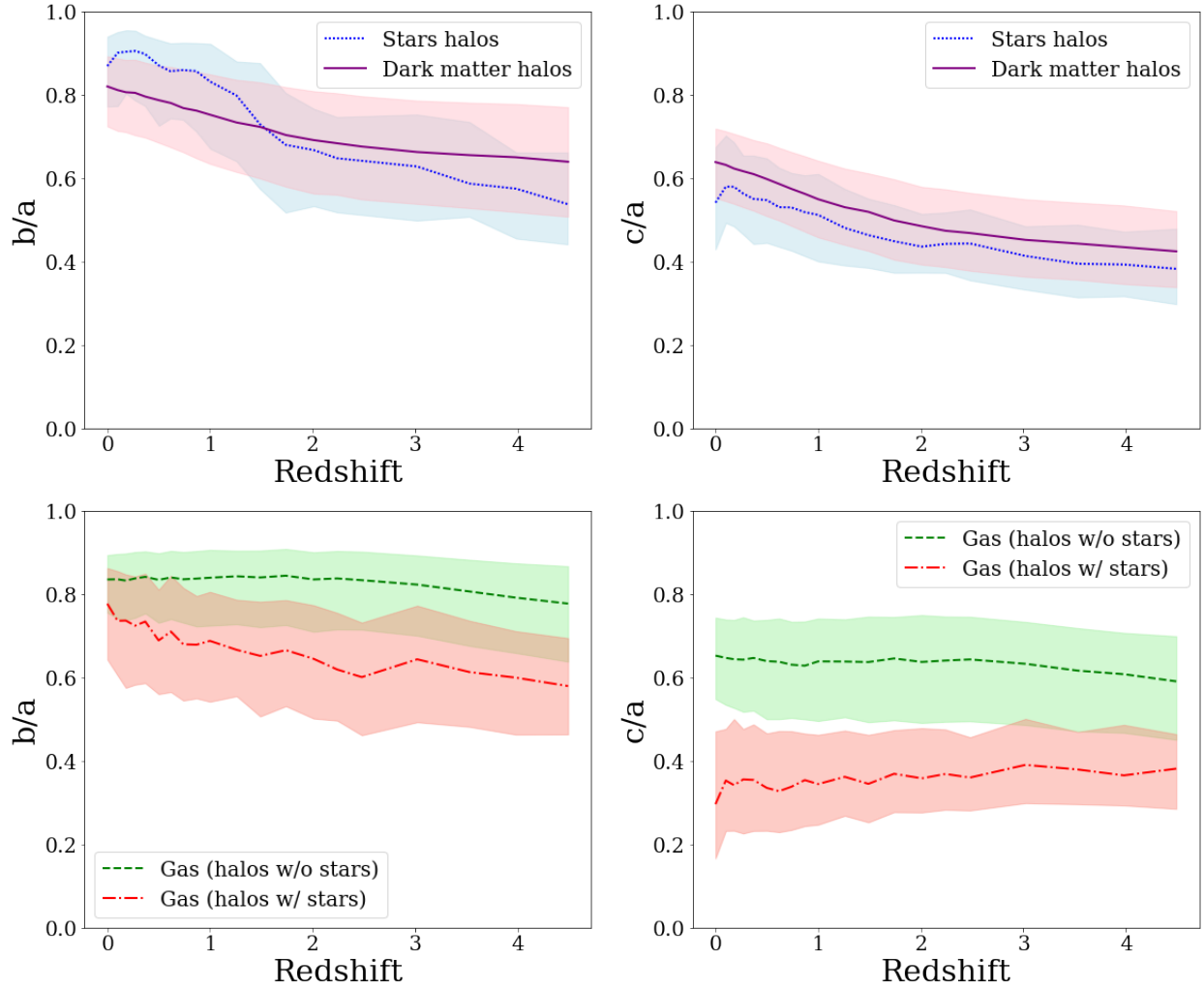


FIGURE III.15 – Figure illustrant les rapports d’axes de l’ellipsoïde d’ajustement en fonction du redshift des halos. Les lignes indiquent les valeurs médianes tandis que l’ombrage indique les 25e - 75e quartiles. (Haut) Le rapport d’axe b/a . (Bas) Le rapport d’axe c/a . Les halos de matières noire sont représentés en mauve trait plein et les étoiles sont représentées en bleu avec une ligne en petits points. Les halos de gaz sont représentés selon qu’ils accompagnent un halo de 100 particules d’étoiles (en rouge) ou non (en vert).

III.3.2.2 La répartition de la masse des halos en 2D

Le travail présenté dans la section précédente étudie la forme tridimensionnelle et la distribution de masse des halos. Au lieu de cela, les analyses courantes de la morphologie des galaxies à l’aide de relevés d’imagerie décrivent généralement des paramètres caractérisant les propriétés des halos projetés sur le plan du ciel. Dans cette section, j’analyse les propriétés des halos EAGLE lorsqu’ils sont projetés sur le plan du ciel, permettant une comparaison avec les relevés d’imagerie au sol

et spatiaux. Sous l'hypothèse raisonnable que les halos sont uniformément orientés, j'ai utilisé les coordonnées des particules dans les plans XY, YZ et XZ comme projections des halos dans trois directions. La figure III.16 présente le rapport d'axe (b_p/a_p) des halos projetés à partir des simulations étudiées de EAGLE. Le tableau III.2 résume les caractéristiques médianes des distributions projetées des rapports d'axes selon le type de particule.

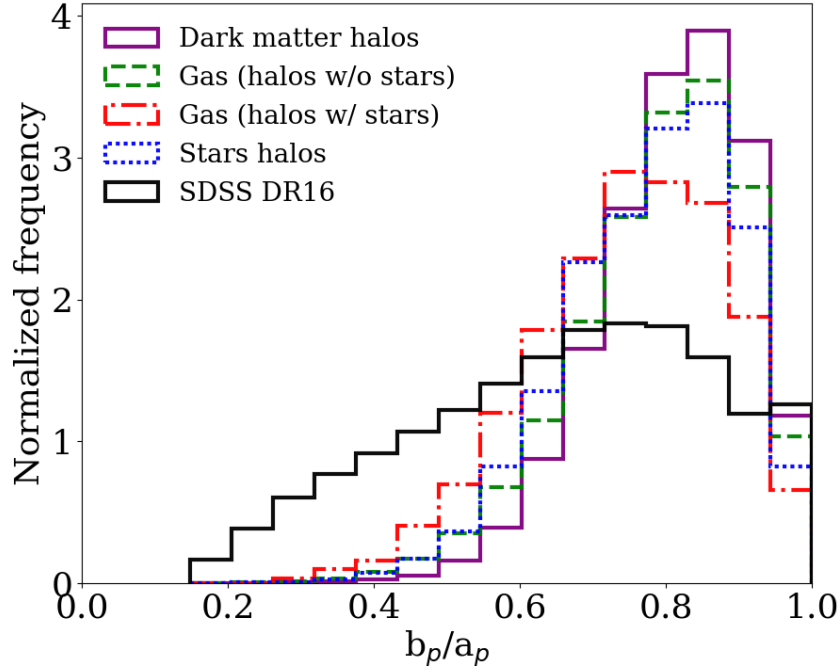


FIGURE III.16 – Répartition des rapports d'axes des halos projetés dans trois directions selon le type de particule : DM (1 439 574 halos), gaz (403 644 halos de gaz sans étoiles, 34 224 avec un nombre minimal de 100 étoiles), étoiles (35 610 halos).

TABLE III.2 – Médiane et écart absolu médian des rapports d'axes projetés b_p/a_p , selon le type de particule : DM (485 796 halos), gaz (151 894 halos w/o étoiles, 17 346 w/ étoiles) et étoiles (17 808 halos).

	DM	Gas	Gas	Stars
		(w/ stars)	(w/o stars)	
b_p/a_p	0.82 ± 0.10	0.75 ± 0.14	0.80 ± 0.12	0.78 ± 0.13

La distribution globale des rapports d'axes projetés de DM a une valeur médiane de $(b_p/a_p)_{DM} = 0,82 \pm 0,10$ correspondant à des halos quasi-sphériques. Le gaz et les étoiles ont des propriétés similaires avec des profils projetés un peu plus allongés avec des valeurs de rapport d'axe de $(b_p/a_p)_{gas} = 0.75 \pm 0.14$ pour les halos de gaz qui contiennent plus de 100 particules d'étoiles et

$(b_p/a_p)_{gas} = 0.80 \pm 0.12$ pour les halos de gaz qui contiennent moins de 100 particules d'étoiles. Les halos d'étoiles ont un rapport d'axe médian de $(b_p/a_p)_{stars} = 0,78 \pm 0,13$.

La valeur médiane des rapports d'axes des étoiles peut être comparée à la morphologie dérivée d'études d'imagerie telles que le SDSS DR16 (Ahumada et al., 2020) ou dérivée des observations spatiales de Gaia (Ducourant et al., 2022; Gaia Collaboration et al., 2022). J'ai analysé la distribution des rapports d'axes de 390 612 galaxies sélectionnées du SDSS DR16 pour lesquelles les paramètres d'un profil de Vaucouleurs sont fournis. La valeur médiane de leurs rapports d'axes est $(b/a)_{SDSS} = 0.65 \pm 0.21$. Les estimations SDSS décrivent le contenu stellaire des galaxies avec une forme plus allongée que celle observée dans la simulation EAGLE. Le même échantillon de galaxies, mais observé par Gaia avec un profil de Vaucouleurs, a les valeurs de rapport d'axe suivantes : $(b/a)_{Gaia} \sim 0.73 \pm 0.14$. En fonction de l'observation au sol ou spatiale, les valeurs médianes d'un rapport d'axe sont quelque peu différentes, mais elles sont en bon accord entre elles et avec la valeur EAGLE, à 1σ près.

L'un des résultats les plus significatifs de cette étude est la contribution majeure que j'apporte à l'amélioration des simulations de lentilles gravitationnelles. En effet, grâce à la distribution des rapports d'axes projetés que j'ai établie, nous sommes en mesure de générer des simulations de lentilles gravitationnelles présentant des paramètres plus réalistes que ce qui a été réalisé dans des études antérieures sur la recherche automatique de lentilles gravitationnelles dans de vastes relevés (Sonnenfeld et al., 2018; Avestruz et al., 2019; Delchambre et al., 2019).

Dans ces précédentes études, les paramètres utilisés pour modéliser la distribution de masse des galaxies déflectrices étaient sélectionnés de manière aléatoire, en supposant des distributions uniformes a priori pour chacun d'entre eux. En revanche, notre approche repose sur une distribution des rapports d'axes qui permet de mieux reproduire la réalité observée. Cette avancée constitue une amélioration majeure car elle permet de capturer plus fidèlement la diversité des configurations de lentilles gravitationnelles présentes dans l'univers.

III.3.3 Distribution de la masse et concentration

A cette étape, je me suis intéressé à la répartition des particules dans les halos pour étudier la distribution de masse, ce qui donne un aperçu de leur structure interne. Dans la figure III.17 je présente la distribution des rapports des demi-axes $\tilde{A}_{gas} = a_{DM}/a_{gas}$ et $\tilde{A}_{stars} = a_{DM}/a_{stars}$ ajustés des halos selon le type de particules. Le même principe est appliqué pour calculer \tilde{B} et \tilde{C} . Nous rappelons au lecteur que la longueur des demi-axes est obtenue à partir de la racine carrée des valeurs propres de l'ACP. Ainsi, la mesure des axes principaux reflète la taille caractéristique dans chaque direction des halos. De cette façon, le rapport \tilde{A} peut être utilisé pour estimer l'extension de la distribution. Si la valeur du demi-axe est petite, les halos ont une petite extension, tandis que plus la valeur est grande, plus le halo est étendu. Cela signifie que $\tilde{A} > 1$ indique que les halos de gaz ou d'étoiles sont moins étendus que les halos de DM selon cet axe principal, ce qui est le

cas pour les halos d'étoiles. Il est important de rappeler ici que ce rapport fournit seulement une mesure de l'étendue des halos et ne fournit pas une mesure de la concentration. Comme mentionné précédemment, les halos de gaz sont divisés en deux groupes : les halos de gaz appauvris en étoiles et les halos de gaz contenant un nombre suffisant d'étoiles. Lorsque nous considérons toutes les particules ensemble, les distributions résultantes sont similaires à celles de la matière noire, ce qui explique pourquoi le cas de toutes les particules n'est pas ajouté aux figures.

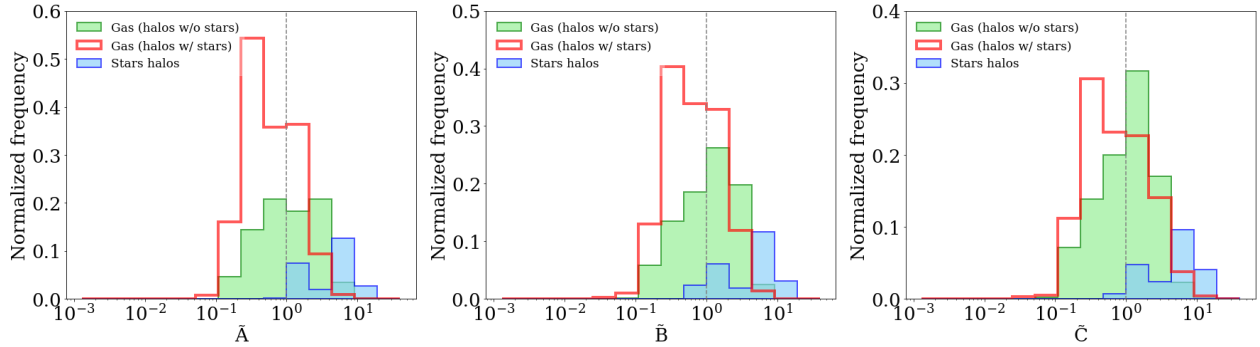


FIGURE III.17 – Rapports d'axe \tilde{A} , \tilde{B} et \tilde{C} . Les valeurs inférieures à 1 correspondent à un demi-axe de gaz (ou d'étoiles) plus grand que leur équivalent de matière noire. Le panneau du milieu correspond à la distribution \tilde{B} et le panneau de droite correspond à la distribution \tilde{C} . Les couleurs sont respectivement le vert et le rouge pour les deux cas distincts du gaz $\tilde{A} = a_{DM}/a_{gaz}$ et le bleu pour $\tilde{A} = a_{DM}/a_{stars}$. Les lignes constantes $\tilde{A}=1$, $\tilde{B}=1$ et $\tilde{C}=1$ sont également tracées en tant que références.

La dispersion des étoiles est également > 1 pour les rapports \tilde{B} et \tilde{C} ce qui montre bien que les étoiles se forment dans les régions de gaz suffisamment denses donc dans le cœur des halos. Cependant, nous constatons, pour un certain nombre de halos de gaz, que leurs demi-axes a , b et c sont plus grands que ceux de leur équivalents en matière noire. Cela signifie que ces halos avec des rapports < 1 présentent une dispersion plus importante le long des axes principaux de l'ellipsoïde ajusté de la matière noire.

Pour aller plus loin dans l'analyse de la dispersion des halos, intéressons-nous au rayon effectif à demie masse, R_{50} . Ce rayon correspond à la taille de demi-grand axe pour un ellipsoïde qui englobe 50% de la masse totale du halo considéré. La figure [III.18](#) montre la distribution du rapport entre le rayon de demi-masse de la matière noire et les halos de gaz (panneau de gauche) et les halos de gaz et d'étoiles (panneau de droite), respectivement.

La distribution des rapports de rayons effectifs en fonction du type de particules révèle que les particules d'étoiles sont principalement concentrées dans le cœur des halos. Aucun halo stellaire ne présente un rayon de demi-masse effectif plus grand que celui de son homologue de gaz correspondant. En revanche, les simulations EAGLE mettent en évidence la présence de halos de gaz plus diffus que leurs homologues de matière noire. Cependant, cela n'est pas observé pour les étoiles. En effet, les

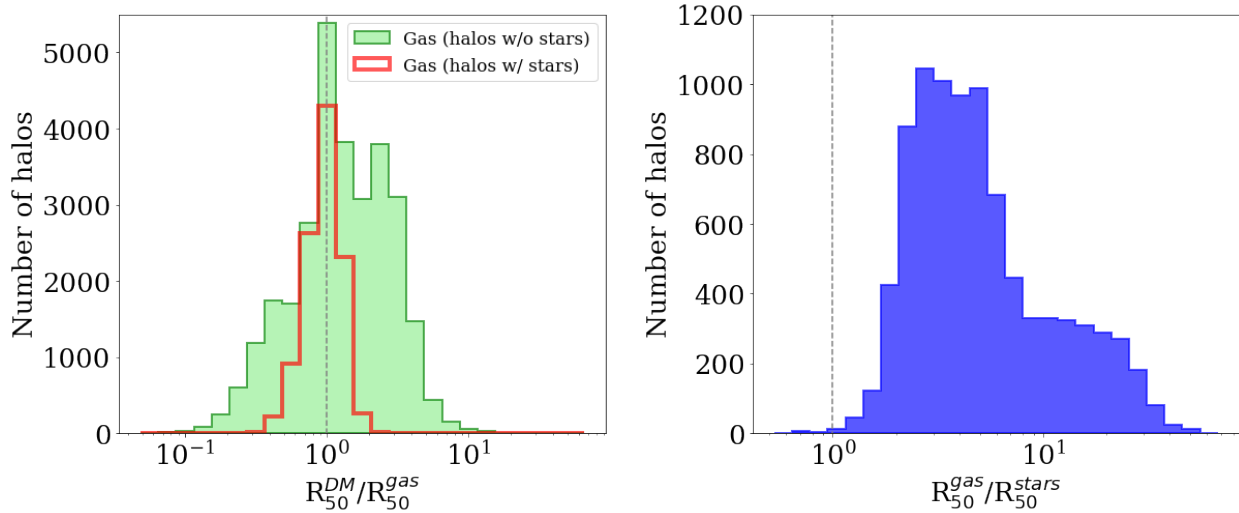


FIGURE III.18 – Répartition du rapport entre le rayon de demi-masse de la matière noire et les halos de gaz (panneau de gauche) et les halos de gaz et d’étoiles (panneau de droite). Les halos sont composés d’un nombre minimum de 200 particules. Le jeu de couleurs est similaire à celui des figures précédentes. Les lignes constantes $R_{50}^{DM}/R_{50}^{gas}=1$ et $R_{50}^{gas}/R_{50}^{stars}=1$ sont également tracées en tant que références.

étoiles se forment principalement au cœur des halos de gaz, comme le confirme la Figure III.18, où aucun halo d’étoiles ne présente un rayon effectif à demi-masse plus grand que celui du halo de gaz qui l’abrite.

Comme nous le détaillerons dans le chapitre suivant, l’un des paramètres essentiels utilisés pour modéliser la distribution de masse des simulations de lentilles gravitationnelles est la dispersion de vitesse, notée σ_v . Cette grandeur est estimée initialement en utilisant la relation de Faber-Jackson, qui s’exprime par $\sigma_v = \sqrt{\frac{GM_{tot}}{R_{\frac{1}{2}}}}$. Dans cette approche, nous nous intéressons à la distribution des rayons à demi-masse des halos présents dans les simulations EAGLE que nous avons étudiés. La figure III.19 illustre cette distribution des rayons à demi-masse pour l’ensemble des particules considérées dans notre étude. Nous observons que les rayons à demi-masse des halos de gaz qui contiennent au moins 100 particules d’étoiles (en rouge sur la figure) sont plus grand que tous les autres composants avec une médiane de $r_{50} = 102 \pm 56$ kpc. Cela signifie que ces halos de gaz sont plus diffus. Cette représentation graphique constitue une étape cruciale pour comprendre la structure et les caractéristiques des halos dans notre échantillon.

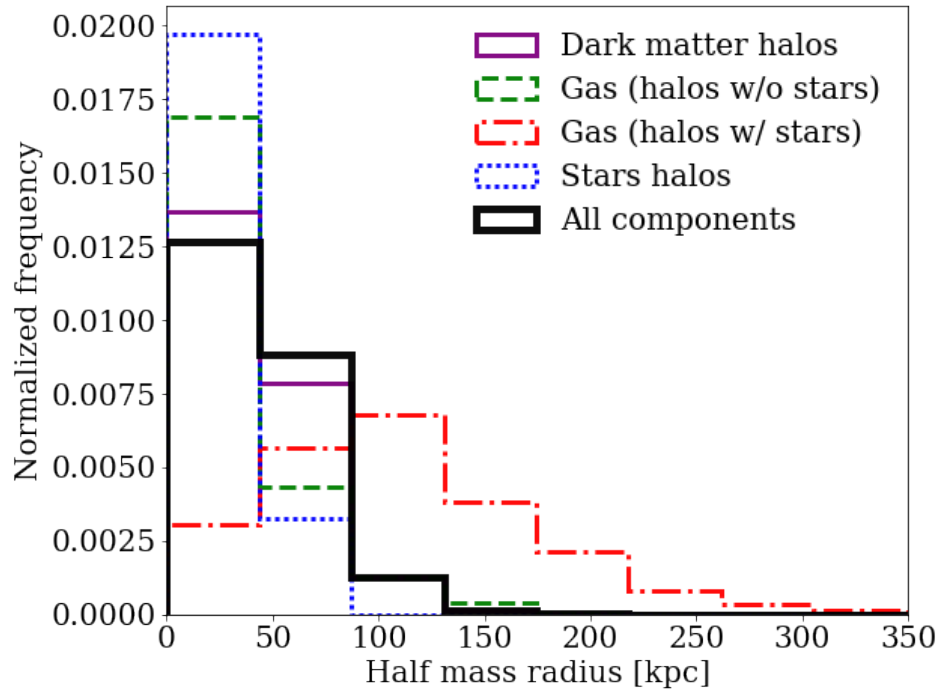


FIGURE III.19 – Répartition des rayons à demi-masse pour chaque type de particules. Les halos de matière noire sont représentés en mauve trait plein et les étoiles sont représentées en bleu avec une ligne en petits points. Les halos de gaz sont représentés selon qu'ils accompagnent un halo de 100 particules d'étoiles (en rouge) ou non (en vert).

Chapitre IV

Simuler des lentilles gravitationnelles réalistes

Ce chapitre a pour objectif de présenter la méthode utilisée pour simuler un catalogue réaliste de quasars lentillés à quatre images (quads) à partir des distributions de masse mesurées dans les simulations cosmologiques EAGLE. Je commencerai par un aperçu des bases du formalisme des lentilles gravitationnelles. Ensuite, j'examinerai l'impact du choix d'un modèle SIS (*Singular Isothermal Sphere*) ou SIE (*Singular Isothermal Ellipsoid*) pour rendre compte du profil de densité de la distribution de masse sur le nombre d'images et sur leurs agencements. Dans la troisième partie, j'aborderai le catalogue de quasars que j'ai utilisé comme sources d'arrière-plan dans les simulations. Par la suite, je détaillerai toutes les étapes que j'ai mises en place pour réaliser la simulation des quads. Enfin, j'analyserai les simulations qui constitueront l'échantillon d'apprentissage de l'algorithme de classification.

IV.1 Introduction aux simulations de lentilles gravitationnelles

La lumière se propage le long des géodésiques nulles de la métrique espace-temps. Cela signifie que la lumière suit la trajectoire de plus court chemin déterminée par la géométrie locale de l'espace-temps. Cependant, dans la plupart des situations astrophysiques en champ faible, une description approximative beaucoup plus simple de la propagation de la lumière peut être utilisée. Cette approximation est connue sous le nom de théorie des lentilles gravitationnelles minces, ou approximation de la lentille mince. Dans cette section, je vais résumer les équations de base pour la description de la déviation de la lumière dans ce cas spécifique.

IV.1.1 Résoudre l'équation de la lentille

L'équation de la lentille est l'expression qui relie l'angle de déviation $\hat{\alpha}$ à la position angulaire de l'image amplifiée observée par rapport à la position angulaire de la source initiale dans le plan du ciel. Pour obtenir cette équation, nous devons prendre en compte la géométrie du système ainsi que la distribution de masse des lentilles. Prenons comme exemple la figure [IV.1](#) qui illustre la déviation de la lumière émise par une source S lorsqu'elle passe à travers une lentille de très faible épaisseur.

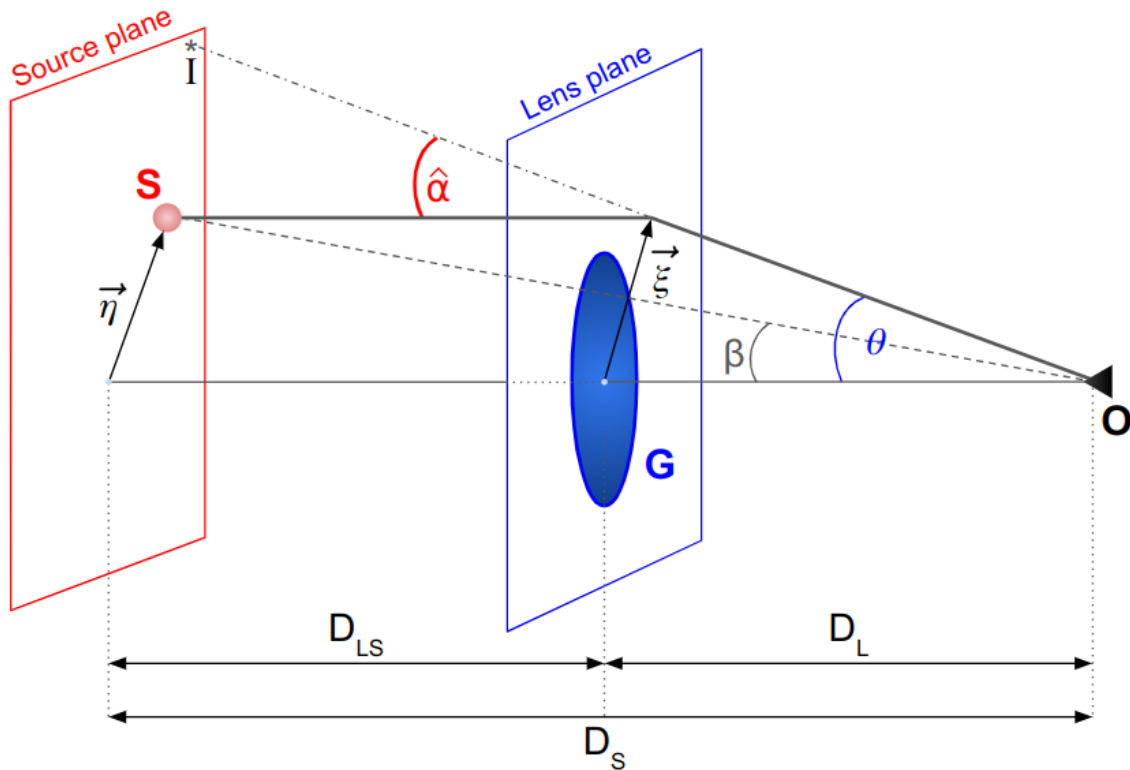


FIGURE IV.1 – Schéma d'un système de lentille gravitationnelle typique dans le cadre de l'approximation de la lentille mince.

Dans ce cas, toute la déviation de la lumière se produit à proximité de la lentille. Étant donné que cette région est extrêmement petite au regard des distances entre l'observateur et la lentille ainsi qu'entre la lentille et la source, nous pouvons utiliser le principe de l'approximation de la lentille mince. Selon cette approximation, la lentille est considérée comme un plan mince de masse perpendiculaire à l'axe optique, qui correspond à la ligne de visée reliant l'observateur au centre de la lentille. En utilisant cette approximation, nous pouvons définir deux plans distincts.

Le premier plan est le plan de la source S , situé à une distance D_S de l'observateur O et contenant

la source. La position angulaire de la source par rapport à l'axe optique est donnée par le vecteur η , tandis que β représente la position angulaire de la source dans le plan du ciel par rapport à l'axe optique. Dans l'approximation des petits angles, nous avons la relation suivante :

$$\eta = D_S \beta \quad (\text{IV.1})$$

Le deuxième plan est le plan de la lentille G , perpendiculaire à la ligne de visée. Ce plan est considéré comme infiniment mince et contient une répartition surfacique de la masse du déflecteur $\Sigma(\xi)$. Une image secondaire de la source est identifiée par ses coordonnées ξ dans le plan de la lentille G et sa position angulaire θ sur le ciel, liées par la relation suivante :

$$\xi = D_L \theta \quad (\text{IV.2})$$

Nous allons maintenant établir l'équation de la lentille, qui permet de relier la position angulaire de l'image secondaire observée à sa position angulaire réelle dans le ciel. Cette relation dépend de l'angle de déviation $\hat{\alpha}(\xi)$ associé à la densité de surface $\Sigma(\xi)$. Si les vecteurs $\vec{\theta}$, $\vec{\beta}$ et $\hat{\alpha}$ sont petits, la position réelle de la source et sa position observée sur le ciel sont liées par une relation très simple, qui peut être obtenue facilement à partir de la géométrie de la figure [IV.1](#). Ainsi, nous obtenons l'équation de la lentille :

$$\vec{\theta} D_S = \vec{\beta} D_S + \hat{\alpha} D_{LS} \quad (\text{IV.3})$$

Nous définissons également l'angle de déviation réduit :

$$\vec{\alpha} \equiv \frac{D_{LS}}{D_S} \hat{\alpha}(\vec{\theta}) \quad (\text{IV.4})$$

En utilisant l'équation [IV.3](#) nous pouvons exprimer $\vec{\beta}$ en fonction de $\vec{\theta}$ et $\vec{\alpha}(\vec{\theta})$:

$$\vec{\beta} = \vec{\theta} - \vec{\alpha}(\vec{\theta}) \quad (\text{IV.5})$$

Afin de simplifier les équations ultérieures, nous introduisons un facteur d'échelle ξ_0 sur le plan de la lentille, correspondant à l'échelle de longueur $\eta_0 = \xi_0 D_S / D_L$ sur le plan source. Ensuite, nous définissons les vecteurs sans dimension :

$$\vec{x} \equiv \frac{\vec{\xi}}{\xi_0}; \vec{y} \equiv \frac{\vec{\eta}}{\eta_0} \quad (\text{IV.6})$$

et l'angle de déviation mis à l'échelle :

$$\vec{\alpha}(\vec{x}) = \frac{D_L D_{LS}}{D_S \xi_0} \hat{\alpha}(\xi_0 \vec{x}) \quad (\text{IV.7})$$

En utilisant l'équation [IV.3](#) nous obtenons l'équation de la lentille dédimensionnée :

$$\vec{y} = \vec{x} - \vec{\alpha}(\vec{x}) \quad (\text{IV.8})$$

Selon cette équation, si un photon est émis à partir de la position angulaire \vec{y} , il sera observé par l'observateur à la position \vec{x} à condition que l'angle de déviation à la position \vec{x} satisfasse l'équation [IV.8](#). Il est possible que, pour une valeur donnée de \vec{y} , l'angle α vérifie l'équation [IV.8](#) pour plusieurs positions \vec{x} . Dans ce cas, l'équation de la lentille admet plusieurs solutions, ce qui signifie qu'une source située à la position \vec{y} sera observée sous la forme d'images multiples dans le plan de la lentille.

Lors de l'observation d'une lentille gravitationnelle, la position vraie de la source demeure inconnue, tandis que nous pouvons mesurer la position de ses images secondaires. Néanmoins, nous pouvons déduire la position vraie de la source en moyennant un modèle de la distribution de masse de la lentille, c'est-à-dire en résolvant l'équation de la lentille pour $\vec{\beta}$. L'équation de la lentille est linéaire en $\vec{\beta}$, ce qui signifie qu'il existe une solution unique pour chaque image. Théoriquement, si nous sommes en mesure d'identifier plusieurs images provenant de la même source et que le modèle de masse de la lentille est correct, nous obtenons la même solution de l'équation de la lentille pour toutes les images.

IV.1.2 Calcul du rayon d'Einstein

Dans le cadre de l'approximation de la lentille mince, nous supposons que toute la masse de la lentille est concentrée dans son plan. Ainsi, le plan de la lentille est caractérisé par sa densité surfacique de masse $\Sigma(\vec{\xi})$. Cette densité est définie comme l'intégrale de la densité volumique de masse $\rho(\vec{\xi}, z)$ le long de l'axe perpendiculaire au plan de la lentille :

$$\Sigma(\vec{\xi}) = \int \rho(\vec{\xi}, z) dz \quad (\text{IV.9})$$

Ici, $\vec{\xi}$ est un vecteur bidimensionnel représentant les coordonnées dans le plan de la lentille. En d'autres termes, $\vec{\xi}$ spécifie la position d'un point dans ce plan. Nous définissons maintenant la densité surfacique de masse dédimensionnée ou convergence, κ , comme le ratio entre la densité surfacique de masse et la densité surfacique de masse critique Σ_{CR} , qui est fonction des distances.

$$\kappa(\vec{\theta}) \equiv \frac{\Sigma(\vec{\theta})}{\Sigma_{cr}} \text{ with } \Sigma_{cr} = \frac{c^2}{4\pi G} \frac{D_S}{D_L D_{LS}} \quad (\text{IV.10})$$

Considérons une lentille de masse ponctuelle. En choisissant la position de la lentille comme le centre du repère de référence, c'est-à-dire en prenant en compte les angles β et θ à partir de la position de la lentille, l'angle de déviation peut être exprimé comme suit :

$$\hat{\alpha}(\vec{\xi}) = \frac{4GM}{c^2} \frac{\vec{\xi}}{|\vec{\xi}|^2} = \frac{4GM}{c^2 D_L} \frac{\vec{\theta}}{|\vec{\theta}|^2} = \hat{\alpha}(\vec{\theta}) \quad (\text{IV.11})$$

Dans cette expression, nous avons utilisé la relation géométrique reliant la longueur physique ξ , l'angle θ et la distance D_L , à savoir $\xi = D_L \theta$. Dans le livre [Meneghetti \(2022\)](#), il est indiqué que le vecteur $\hat{\alpha}$ pointe en dehors de la lentille et que nous pouvons donc omettre le signe vectoriel dans plusieurs des équations suivantes. Ainsi, nous obtenons :

$$\hat{\alpha} = \frac{4GM}{c^2 D_L \theta} \quad (\text{IV.12})$$

En reprenant l'équation de la lentille [IV.3](#), nous pouvons écrire :

$$\beta = \theta - \frac{4GM}{c^2 D_L \theta} \frac{D_{LS}}{D_S} \quad (\text{IV.13})$$

Cette équation est plus couramment exprimée de manière concise en introduisant le rayon d'Einstein :

$$\theta_E \equiv \sqrt{\frac{4GM}{c^2} \frac{D_{LS}}{D_L D_S}} \quad (\text{IV.14})$$

Nous obtenons alors :

$$\beta = \theta - \frac{\theta_E^2}{\theta} \quad (\text{IV.15})$$

En divisant par θ_E et en posant $y = \beta/\theta_E$ et $x = \theta/\theta_E$, nous pouvons exprimer l'équation de la lentille exprimé en unité du rayon d'Einstein :

$$y = x - \frac{1}{x} \quad (\text{IV.16})$$

En multipliant par x on peut écrire :

$$x^2 - xy - 1 = 0 \quad (\text{IV.17})$$

qui a donc deux solutions :

$$x_{\pm} = \frac{1}{2} \left[y \pm \sqrt{y^2 + 4} \right] \quad (\text{IV.18})$$

Il est important de préciser que dans notre notation, les coordonnées des images dans le plan de la lentille sont données par $x = [x_1, x_2]$, tandis que la position de la source dans le plan source est représentée par $y = [y_1, y_2]$. Par conséquent, l'équation précédente indique qu'une source dont la lumière est déviée par une lentille de masse ponctuelle aura toujours deux images. Ces deux images seront situées de part et d'autre de la lentille, et l'une des images se trouvera toujours à l'intérieur d'un cercle de rayon $x = 1$ et l'autre à l'extérieur. Ce cercle correspond à la solution d'un alignement parfait où $y = 0$, ce qui conduit aux solutions $x_{\pm} = \pm 1$. Ainsi, les images s'étaleront le long d'un anneau de rayon θ_E , un aspect connu sous le nom d'anneau d'Einstein.

IV.1.3 Amplification et cisaillement des images

L'une des conséquences principales des lentilles gravitationnelles est la distorsion des images provenant des sources en arrière-plan. Cette distorsion devient visible lorsque la source est étendue, par exemple lorsqu'il s'agit de galaxies d'arrière-plan se présentant en arcs allongés. La distorsion se produit en raison de la déviation différentielle des faisceaux lumineux. Considérons un point sur le plan de la lentille situé à la position θ_0 , et examinons ce qui se passe pour un deuxième point proche, $\theta = \theta_0 + d\theta$. L'angle de déviation peut alors s'approximer par :

$$\vec{\alpha} \simeq \vec{\alpha}_0 + \frac{d\vec{\alpha}}{d\vec{\theta}} d\vec{\theta} \quad (\text{IV.19})$$

En utilisant l'équation de la lentille (Eq. IV.3), les points $\vec{\theta}_0$ et $\vec{\theta}$ sont projetés sur les points $\vec{\beta}_0$ et $\vec{\beta} = \vec{\beta}_0 + d\vec{\beta}$ sur le plan source. Par conséquent, le vecteur $(\vec{\beta} - \vec{\beta}_0)$ est donné par :

$$(\vec{\beta} - \vec{\beta}_0) = \left(I - \frac{d\vec{\alpha}}{d\vec{\theta}} \right) (\vec{\theta} - \vec{\theta}_0) \quad (\text{IV.20})$$

Ainsi, la distorsion des images peut être décrite par une matrice jacobienne :

$$A \equiv \frac{\partial \vec{\beta}}{\partial \vec{\theta}} = \left(\delta_{ij} - \frac{\partial \alpha_i(\vec{\theta})}{\partial \theta_j} \right) \quad (\text{IV.21})$$

où θ_i représente la i -ème composante de $\vec{\theta}$ dans le plan de la lentille.

Cependant, il convient de noter que notre étude se concentre spécifiquement sur les lentilles gravitationnelles de quasars. Dans ce contexte, plutôt que d'observer une distorsion des images, nous aurons une amplification de la brillance des quasars. Le théorème de Liouville (Etherington, 1933) stipule que, en l'absence d'émission et d'absorption de photons, le flux lumineux par unité de surface d'une source de lumière reste constant au cours de sa propagation, même si la lumière est déviée. Cela signifie que la luminosité reste la même, même si elle est déformée par la lentille gravitationnelle. Cependant, en raison de la déviation de la lumière, l'angle solide sous lequel la source est observée peut changer. L'angle solide mesure la taille apparente de la source depuis l'observateur. Si cet angle solide diminue, cela signifie que la source apparaît plus brillante, car la même quantité de lumière est concentrée sur une zone plus petite. C'est ce qu'on appelle l'amplification ou la *magnification*.

Étant donné l'équation IV.21, l'amplification (*magnification*) est donnée par l'inverse du déterminant de la matrice jacobienne. Pour cette raison, la matrice $M = A^{-1}$ est appelée tenseur d'amplification. Ici, le tenseur d'amplification décrit à quel point l'intensité d'une image est amplifiée ou diminuée dans différentes directions. Nous définissons donc :

$$\mu \equiv \det M = \frac{1}{\det A} = \frac{1}{(1 - \kappa)^2 - \gamma^2} \quad (\text{IV.22})$$

Dans cette équation, κ et γ sont des paramètres qui décrivent les propriétés de la lentille gravitationnelle. κ représente la convergence, qui est liée à la concentration de masse totale de la lentille, et γ représente la composante tangentielle du cisaillement interne (*shear* en anglais), qui est liée à l'asymétrie de la distribution de la masse. Dans le cas des lentilles à symétrie axiale, les vecteurs propres du tenseur de cisaillement sont orientés tangentiellement et radialement par rapport aux contours d'isodensité surfacique de la lentille. Cela signifie que l'amplification varie selon ces deux directions. On définit donc deux quantités qui sont les solutions $\mu = \frac{1}{\lambda}$, où μ est l'amplification et λ est une valeur propre du tenseur d'amplification.

$$\mu_t = \frac{1}{\lambda_t} = \frac{1}{1 - \kappa - \gamma} \quad (\text{IV.23})$$

$$\mu_r = \frac{1}{\lambda_r} = \frac{1}{1 - \kappa + \gamma} \quad (\text{IV.24})$$

Ces quantités sont souvent appelées les facteurs d'amplification tangentiel et radial, respectivement. Elles indiquent l'amplification le long des directions tangentielles et radiales par rapport aux contours d'isodensité de la lentille.

L'amplification des images est une caractéristique des lentilles gravitationnelles, où différentes images peuvent être amplifiées à des degrés variés. Dans certaines situations particulières, l'amplification théorique devient infinie. Cela se produit lorsque les valeurs propres du tenseur d'amplification, λ_t et λ_r , sont égales à zéro. Lorsque $\lambda_t = 0$, on parle de la ligne critique tangentielle. Lorsque $\lambda_r = 0$, on parle de la ligne critique radiale. Ces deux conditions définissent deux courbes distinctes dans le plan de la lentille. La présence de ces lignes critiques tangentielle et radiale permet de reconnaître les lentilles gravitationnelles, car elles donnent lieu à des configurations où les images sont amplifiées à des degrés différents. L'une des manifestations les plus connues est l'anneau d'Einstein, qui est formé lorsque la ligne critique tangentielle se referme sur elle-même, créant ainsi un anneau de rayon particulier où l'amplification est infinie.

Lorsque la ligne critique tangentielle est projetée dans le plan de la source par l'équation de la lentille, elle correspond à la ligne caustique au point $\beta = 0$ ($y = 0$). Pour une lentille à symétrie axiale, la ligne caustique et la ligne critique tangentielle projetée dans le plan de la source se superposent.

Les sources d'arrière-plan dans les systèmes de lentilles gravitationnelles peuvent être des sources étendues telles que des galaxies, ou des sources ponctuelles, notamment des quasars. Ces sources d'arrière-plan subissent une amplification et une déformation résultant de l'effet de cisaillement propre à la lentille. Par conséquent, si la source d'arrière-plan est une galaxie, par exemple, elle sera déformée et se manifestera sous la forme d'un ou de plusieurs arcs étendus. La figure [IV.2](#) illustre cette déformation observée sur une source d'arrière-plan de forme circulaire à travers le modèle SIE. J'ai réalisé ces figures avec python en reconstruisant les images d'une source circulaire étendue (panneau de gauche) et d'une source ponctuelle (panneau de droite). On peut clairement y observer les différentes images de la source étendue se présentant sous la forme d'arcs. Il est intéressant de noter ici que les images des sources étendues peuvent se joindre et se mélanger alors que des images aux mêmes positions de sources ponctuelles (panneau de droite) ne le seraient pas.

IV.2 Choix d'un modèle de profil de densité

La description la plus simple pour modéliser le déflecteur est un déflecteur (appelé lentille par la suite) possédant une symétrie axiale ou circulaire. Pour ce type de lentille, le potentiel de la lentille (qui est une mesure de la distribution de masse de la lentille) est constant le long de cercles centrés sur le centre de la lentille. Cette propriété de symétrie permet de choisir l'origine du référentiel au centre de la lentille, ce qui facilite les calculs et les équations associées.

L'angle de déviation pour une lentille à symétrie axiale s'écrit :

$$\alpha(\theta) = \frac{D_{LS}}{D_S} \frac{4GM(\theta)}{c^2 D_L \theta} \quad (\text{IV.25})$$

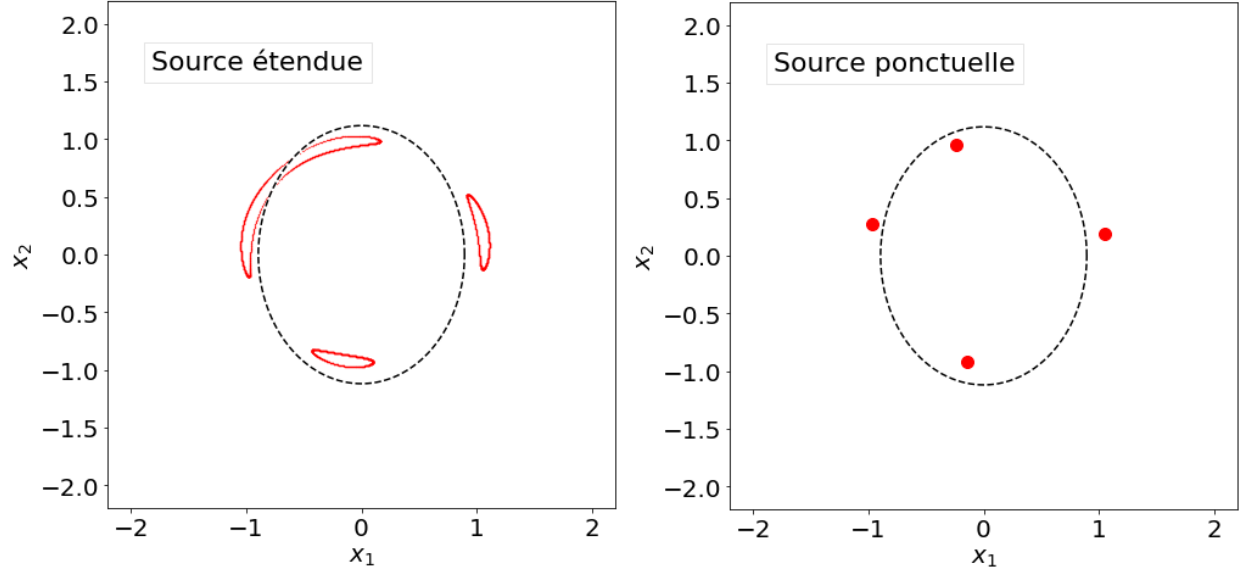


FIGURE IV.2 – Position et déformation des sources d’arrière-plan (au redshift $z_s = 2.0$) au sein d’une lentille gravitationnelle (au redshift $z_d = 0.3$) avec un rapport d’axe de $b/a = 0.7$. La ligne critique est représentée par une ligne en pointillés.

Cette formule est identique à celle de l’angle de déviation pour une lentille de masse ponctuelle, la masse M est remplacée par la masse $M(\theta)$ contenue dans le cercle de rayon θ . Cette analogie avec la loi de Newton, qui indique que tout se passe comme si la masse était concentrée en un seul point, justifie l’identité des formules utilisées dans le cas des lentilles gravitationnelles. Cela signifie que, grâce à la symétrie axiale de la lentille, les propriétés de la lentille sont uniquement déterminées par le profil de masse $M(\theta)$, ou alternativement par le profil de la densité de surface $\Sigma(\theta)$.

De manière similaire à la section précédente, nous pouvons exprimer l’angle de déviation sous une forme adimensionnelle en utilisant une échelle linéaire arbitraire, notée ξ_0 , correspondant à $\theta_0 = \xi_0/D_L$. Ainsi, l’angle de déviation peut être écrit comme suit :

$$\alpha(x) = \frac{D_L D_{LS}}{\xi_0 D_S} \hat{\alpha}(\xi_0 x) \quad (\text{IV.26})$$

$$= \frac{M(\xi_0 x)}{\pi \xi_0^2 \Sigma_{cr}} \frac{1}{x} = \frac{m(x)}{x} \quad (\text{IV.27})$$

Ici, nous avons introduit la fonction de masse :

$$m(x) \equiv \frac{M(\xi_0 x)}{\pi \xi_0^2 \Sigma_{cr}} \quad (\text{IV.28})$$

Ainsi, l'équation de la lentille peut être exprimée comme :

$$y = x - \frac{m(x)}{x} \quad (\text{IV.29})$$

Lorsque nous utilisons la notation adimensionnelle, un choix naturel pour ξ_0 est $\xi_0 = D_L \theta_E$. Dans ce cas, la fonction de masse correspond à la masse $M(\theta)$ en unités de masse $M(\theta_E)$:

$$m(x) = \frac{M(\theta)}{M(\theta_E)} \quad (\text{IV.30})$$

Les formules que nous venons de dériver sont valables pour toute lentille à symétrie axiale. En raison de sa simplicité, l'un des modèles les plus largement utilisés pour les lentilles à symétrie axiale est la sphère isotherme singulière pour *Singular Isothermal Sphere* en anglais (SIS) (Kormann et al., 1994). Ce modèle repose sur l'idée que la matière de la lentille se comporte comme un gaz parfait confiné par un potentiel gravitationnel à symétrie sphérique.

Dans ce modèle, on suppose que le gaz constituant la lentille est en équilibre thermique et hydrostatique, ce qui permet de dériver un profil de densité satisfaisant ces conditions. Ce profil de densité tridimensionnel est donné par :

$$\rho(r) = \frac{\sigma_v^2}{2\pi G r^2} \quad (\text{IV.31})$$

Où σ_v est la dispersion de vitesse et $r = \sqrt{\xi^2 + z^2}$ est la distance au centre de la sphère. Pour obtenir la densité surfacique projetée sur le plan de la lentille, on intègre la densité tridimensionnelle le long de l'axe de visée z . Cela donne :

$$\Sigma(\xi) = 2 \int_0^\infty \rho(\xi, z) dz \quad (\text{IV.32})$$

$$= \frac{\sigma_v^2}{2G\xi} \quad (\text{IV.33})$$

Ce profil de densité a une singularité à $\xi = 0$, où la densité est infinie. En posant $\xi_0 = D_L \theta_0$, nous obtenons :

$$\theta_0 = 4\pi \left(\frac{\sigma_v}{c} \right)^2 \frac{D_{LS}}{D_S} \quad (\text{IV.34})$$

Ainsi, pour le profil SIS, la fonction de masse est simplement $m(x) = x$, et l'angle de déviation

s'écrit :

$$\alpha(x) = \frac{x}{|x|} \tag{IV.35}$$

Étant donné que $|x| = 1$ lorsque $y = 0$, on constate que θ_0 correspond effectivement au rayon d'Einstein pour le modèle SIS. Deux solutions distinctes de l'équation de la lentille existent, correspondant aux positions $x_- = y - 1$ et $x_+ = y + 1$, situées de part et d'autre du centre de la lentille. La séparation angulaire entre ces deux images est toujours de $\Delta(\theta) = 2\theta_E$. Cependant, si $y > 1$, l'équation de la lentille possède une unique solution, à savoir $x_+ = y + 1$. Ainsi, le cercle de rayon $y = 1$ constitue la frontière qui sépare les régions du plan source présentant des images de multiplicités différentes. Cette ligne est couramment appelée le *cut*. Dans le cadre du modèle SIS, le rayon d'Einstein et le *cut* partagent la même valeur. La figure [IV.3](#) illustre l'exemple d'une lentille gravitationnelle simulée à partir d'un modèle SIS. Lorsque la source ponctuelle est parfaitement alignée sur la ligne de visée, la solution théorique de l'équation de la lentille prédit en effet la formation de quatre images. Cependant, en réalité, on observe souvent un anneau plutôt que quatre images distinctes.

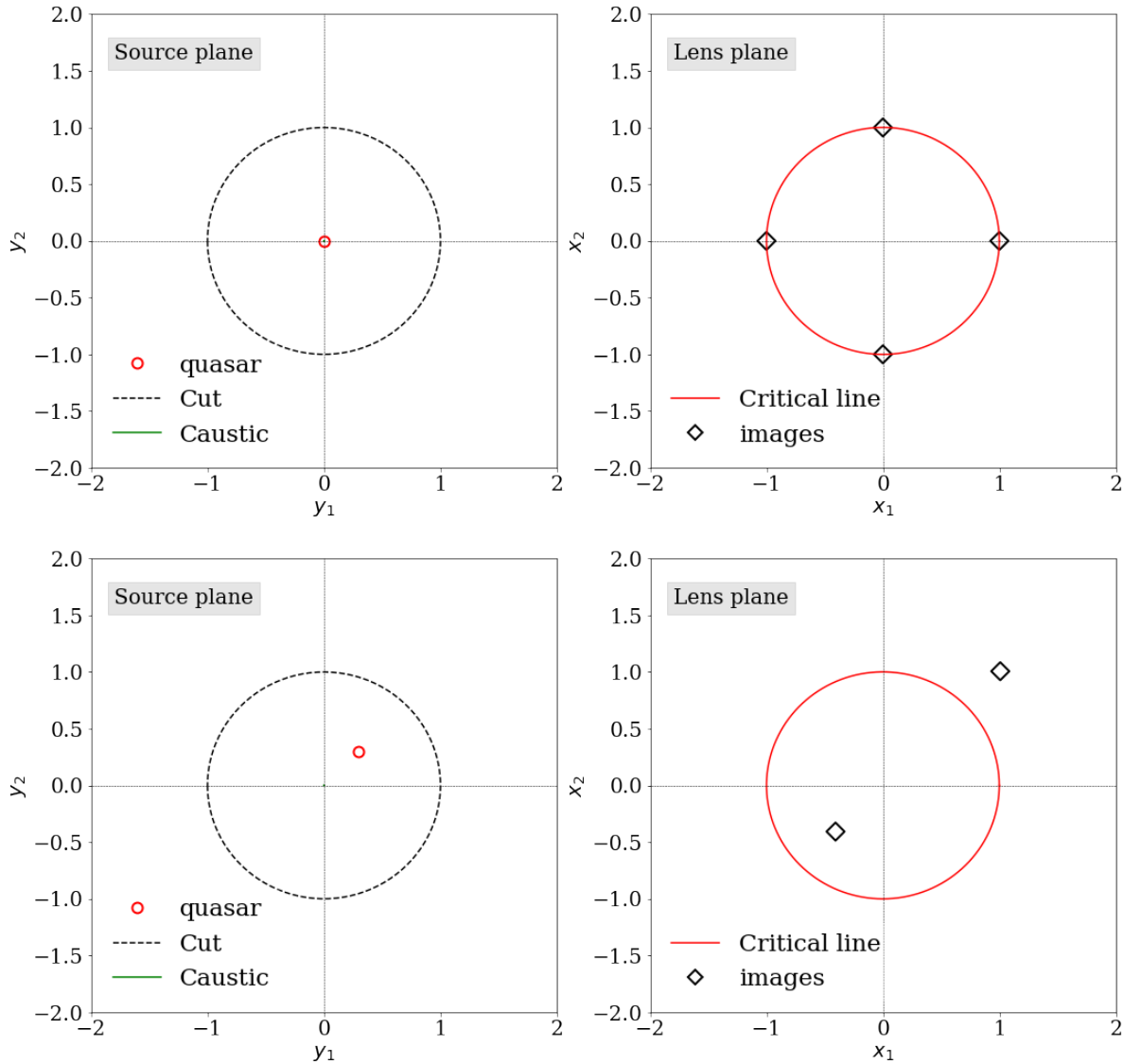


FIGURE IV.3 – Simulation d’une lentille gravitationnelle avec un modèle SIS et une source ponctuelle (quasar) représenté par le cercle rouge dans le plan source (panneau de gauche). Les images correspondantes sont représentées par des losanges dans le plan de la lentille (panneau de droite). La ligne critique tangentielle est indiquée en rouge, tandis que le *cut* est représenté par une ligne en pointillé noire dans le plan source.

Considérons maintenant comment l’ellipticité affecte les propriétés de la lentille. L’ajout d’une ellipticité supprime la symétrie axiale de la lentille, donnant ainsi le nom de SIE, pour *Singular Isothermal Ellipsoid*, au modèle. Lorsque la lentille présente une ellipticité, la géométrie de la ligne caustique tangentielle et de la ligne critique tangentielle est modifiée par rapport à celle d’une lentille à symétrie axiale. Dans le cas d’une lentille elliptique, la ligne caustique tangentielle n’est plus un

point, mais plutôt une structure plus complexe ressemblant à un astroïde. Cette caustique peut avoir deux ou quatre cuspidés (points de rebroussement) en fonction de l'ellipticité et de la forme spécifique de la lentille. Les cuspidés représentent des régions de concentration d'amplification et de déformation plus élevées.

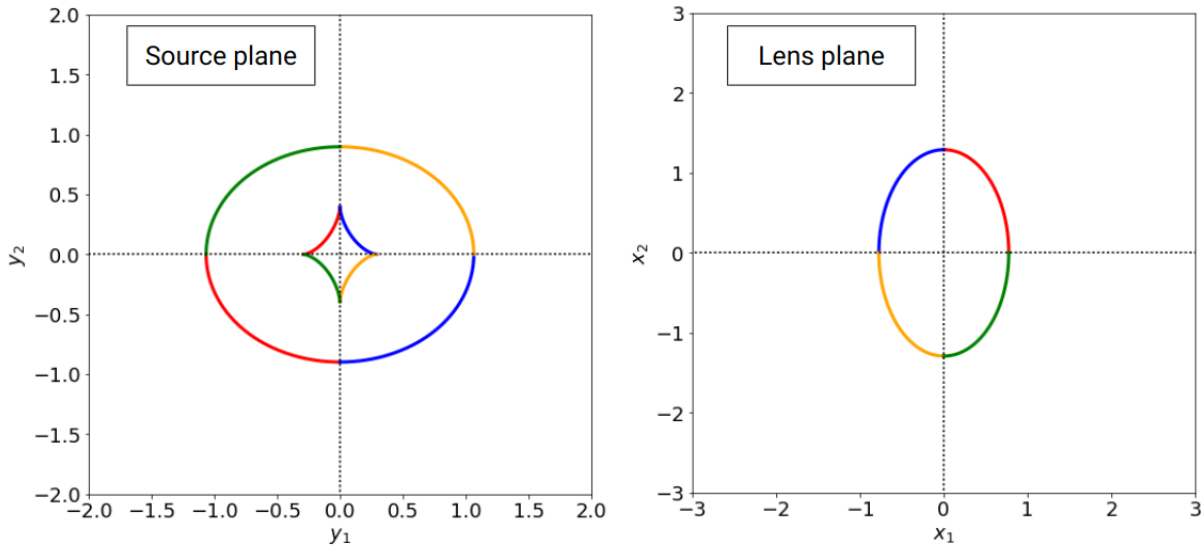


FIGURE IV.4 – Cartographie du *cut* et de la caustique dans le plan de la source (panneau de gauche). Cartographie des points de la ligne critique dans le plan de la lentille (panneau de droite). Différentes couleurs correspondent à différents quadrants. Figure issue du livre *Introduction to Gravitational Lensing* de [Meneghetti \(2022\)](#).

On introduit l'ellipticité en remplaçant la distance ξ par $\sqrt{\xi_1^2 + f^2 \xi_2^2}$, où ξ_1 et ξ_2 sont les coordonnées sur le plan de la lentille et $f = b/a$ est le rapport d'axe de l'ellipse.

Avec cette modification, la densité surfacique projetée sur le plan de la lentille devient :

$$\Sigma(\vec{\xi}) = \frac{\sigma_v^2}{2G} \frac{\sqrt{f}}{\sqrt{\xi_1^2 + f^2 \xi_2^2}} \quad (\text{IV.36})$$

Ce profil de densité surfacique est constant le long des ellipses de petit-axe ξ et de grand-axe ξ/f , avec l'axe majeur aligné selon l'axe ξ_2 .

Il est utile de voir comment des morceaux de la ligne critique tangentielle sont projetés sur le plan source. Dans le panneau de gauche de la figure [IV.4](#) issu du livre de Massimo Meneghetti ([Meneghetti, 2022](#)), la ligne critique est divisée en quatre parties correspondant aux quadrants dans le plan de la lentille. Chaque partie est associée à une couleur différente pour indiquer les

points critiques avec des angles polaires spécifiques dans chaque plage $[0, \pi/2)$, $[\pi/2, \pi)$, $[\pi, 3/2\pi)$ et $[3/2\pi, 2\pi)$. Dans le panneau de droite, les points correspondants sur la caustique sont affichés en utilisant les mêmes couleurs. La couleur de chaque point sur la caustique correspond à la couleur de la partie correspondante de la ligne critique. La règle générale observée est que la cartographie des points critiques suit une règle de gauche à droite. Cela signifie que les points critiques dans le premier et le quatrième quadrant du plan de la lentille (angles polaires ϕ appartenant à $[0, \pi/2)$ et $[3/2\pi, 2\pi)$) sont cartographiés sur le deuxième et le troisième quadrant du plan source, et vice versa. Cependant, la cartographie des points sur le *cut* (ligne caustique) suit une règle diagonale. Par exemple, la portion du *cut* située dans le premier quadrant du plan source correspond aux angles polaires ϕ dans la plage $[\pi, 3/2\pi)$. Cette règle diagonale indique que les points sur le *cut* sont projetés de manière spécifique en fonction de leur position sur la caustique.

La multiplicité des images produites par le modèle SIE dépend des formes de la caustique et du *cut* et de la position de la source par rapport à ces deux courbes. Pour des petites ellipticités, donc f grand, les deux cas possibles sont :

- La source se trouve dans le *cut*, ce qui formera deux images ;
- La source se trouve dans le *cut* et dans la caustique ce qui formera quatre images de la source d'arrière plan.

En revanche, pour de grandes ellipticités, correspondant à de petites valeurs de f , la caustique tangentielle s'étend à l'extérieur du *cut* selon l'axe y_2 .

J'ai représenté sur la figure [IV.5](#) plusieurs lentilles avec chacune une ellipticité différente, passant ainsi d'une configuration avec quatre images à une configuration avec seulement trois images. Les cas avec seulement trois images ne seront pas étudiés dans le cadre de cette thèse. Cependant, l'exploration de ces configurations avec trois images constitue une perspective d'amélioration intéressante pour les travaux futurs.

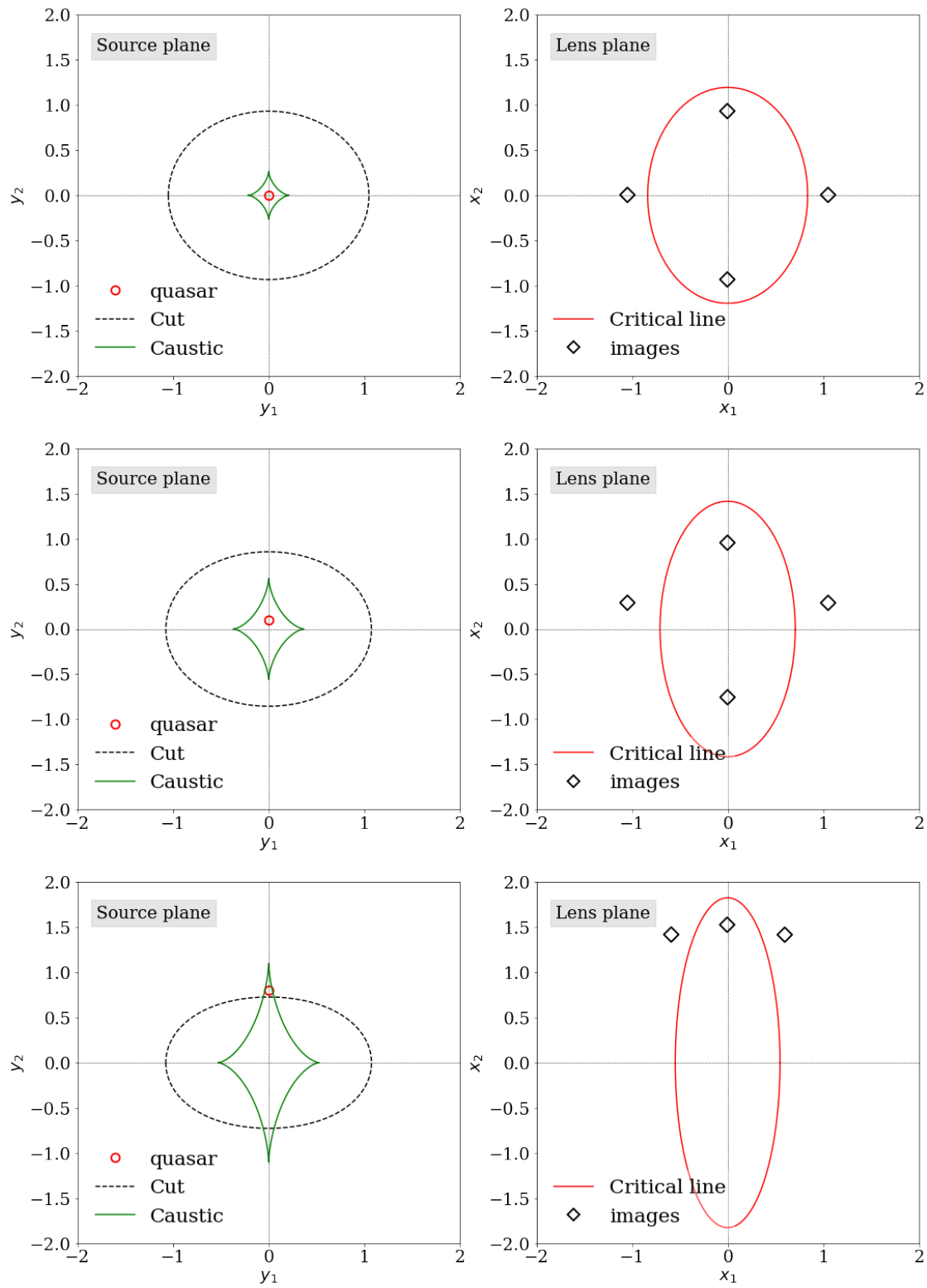


FIGURE IV.5 – Simulations de lentilles gravitationnelles en fonction de l'ellipticité.

Au cours de cette section, j’ai précisé que le modèle de lentille gravitationnelle SIE offre la possibilité de simuler des configurations avec des images multiples. Il est l’un des modèles les plus réalistes pour reproduire fidèlement ce qui est observé. C’est donc le modèle que j’ai sélectionné pour la génération de lentilles gravitationnelles réalistes. Ce modèle repose sur la représentation de la distribution de masse du déflecteur sous la forme d’ellipses d’iso-densité. Comme nous l’avons vu, la modélisation de ces lentilles nécessite la prise en compte de plusieurs paramètres, notamment l’ellipticité du profil de masse, notée $f = b/a$, la dispersion de vitesse (σ_v) et le redshift de la galaxie et de la source d’arrière plan. La distribution de chacun de ces paramètres a été déterminée à partir des simulations cosmologiques EAGLE (Petit et al., 2023). Dans la section suivante, j’aborderai la méthodologie développée pour réaliser les simulations des lentilles gravitationnelles.

IV.3 Simulation des quads

IV.3.1 Des quasars comme source d’arrière plan

Pour réaliser les simulations, nous avons choisi d’utiliser le Million Quasars (MILLIQUAS) Catalog, Version 7.10 (15 April 2023) (Flesch, 2021) qui recense plus de 845 286 QSO et AGN de type I. Ce catalogue est une source de données complète, mise à jour jusqu’au 5 février 2023, qui contient également 553 966 candidats QSO, identifiés grâce à l’association radio/rayons X, ainsi que des blazars et des objets de type II, pour un total de 1 461 834 objets. Près de 73% de ces objets ont une astrométrie Gaia-EDR3 ou Pan-STARRS.

Nous croisons (*cross-match*) des données avec Gaia DR3 afin d’obtenir une liste de 544 981 quasars pour lesquels nous disposons d’une mesure de redshift. Dans le cadre de cette mise en correspondance, nous récupérons également le flux et la magnitude Gaia dans la bande G. La figure IV.6 illustre la distribution des magnitudes Gaia des quasars en fonction du redshift. La distribution des quasars a un pic autour d’un décalage $z = 1.5$, et des quasars sont détectés jusqu’à $z = 4$. Par ailleurs, la médiane des magnitudes en bande G est de 20.

IV.3.2 Calcul des rayons d’Einstein de couples QSO/halos de EAGLE

Cette partie se concentre sur le calcul du rayon d’Einstein et son influence sur la taille des configurations observables dans les lentilles gravitationnelles. Le rayon d’Einstein est déterminé par la masse des déflecteurs, qui correspond dans notre étude aux halos observés dans EAGLE, ainsi que par les distances relatives entre la lentille et la source. Je vais donc examiner comment ces paramètres influent sur la taille des configurations que nous pouvons détecter. Le rayon d’Einstein que nous avons défini pour le modèle SIS s’écrit :

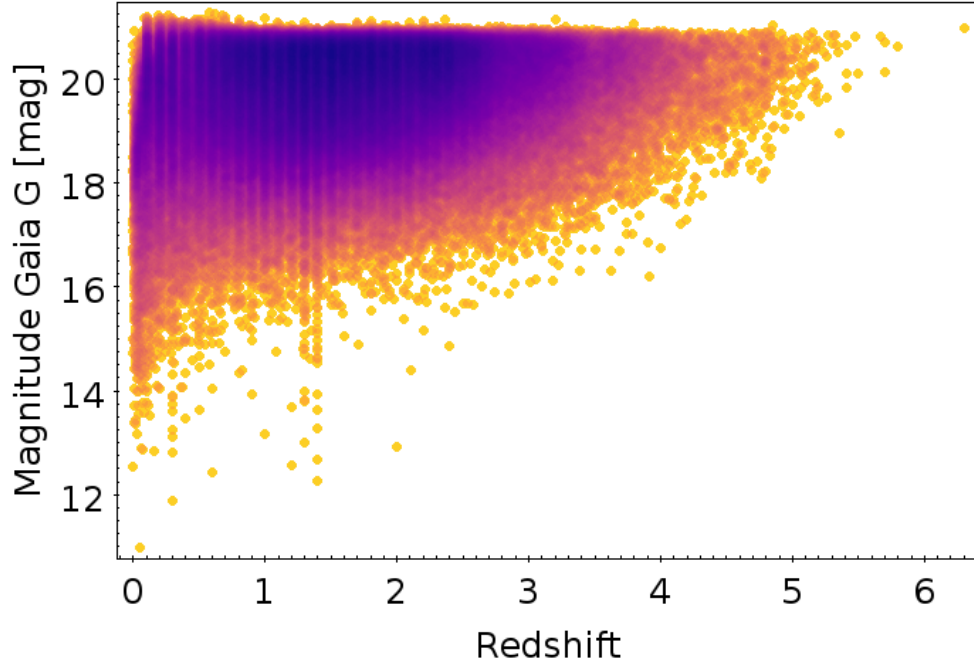


FIGURE IV.6 – Distribution des magnitudes Gaia G en fonction du redshift des quasars du catalogue MILLIQUAS (Flesch, 2021).

$$\theta_E = 4\pi \left(\frac{\sigma_v}{c} \right)^2 \frac{D_{LS}}{D_S} \quad (\text{IV.37})$$

où D_{LS} est la distance entre le déflecteur et la source et D_S la distance de la source. σ_v est la dispersion de vitesse des déflecteurs et est obtenue à partir de la relation de Faber-Jackson (Faber and Jackson, 1976). La relation de Faber-Jackson est une relation empirique observée pour les galaxies de type elliptique ou les bulbes des galaxies spirales, excluant les disques. Elle indique que la dispersion de vitesse est corrélée à la luminosité de la galaxie selon une loi de puissance, généralement exprimée comme :

$$\sigma_v \propto L^\alpha \quad (\text{IV.38})$$

où α est un coefficient déterminé empiriquement. Il est important de noter que cette relation peut également être étendue aux halos de matière noire.

En supposant une valeur constante pour le rapport masse-luminosité (M/L), où la masse est proportionnelle à la luminosité avec un coefficient d'environ 5, on peut réécrire l'équation de dispersion de vitesse en utilisant la masse totale M_{tot} :

$$\sigma_v \propto \sqrt{\frac{L}{R_{\frac{1}{2}}}} = \sqrt{\frac{G(M/L)L}{R_{\frac{1}{2}}}} = \sqrt{\frac{GM_{tot}}{R_{\frac{1}{2}}}} \quad (\text{IV.39})$$

M_{tot} est la masse totale de la galaxie déflectrice et $R_{\frac{1}{2}}$ est le rayon de demi-masse. Cette équation est une simplification de l'équation plus générale de la vitesse de dispersion basée sur le théorème du viriel dans un système en équilibre. Il est important de noter que cette équation est une approximation. Une méthode plus précise consisterait à utiliser directement le sigma projeté dans le plan du ciel pour les simulations EAGLE et calculer un rayon d'Einstein sur la base de celui-ci.

Afin de produire une variété réaliste d'images multiples, j'ai constitué une population réaliste de quasars et de galaxies déflectrices. Étant donné le grand nombre de combinaisons possibles entre les 340 719 halos, issus des simulations EAGLE avec des mesures de distribution de masse et de redshift, et les 544 981 quasars, il n'est pas pertinent de les calculer toutes. Même en ne considérant que les cas où le quasar est situé plus loin que le déflecteur, cela représente plusieurs dizaines de milliards de combinaisons potentielles. Afin d'obtenir une distribution réaliste des rayons d'Einstein (θ_E), j'ai opté pour une approche où je sélectionne aléatoirement 500 quasars pour chaque halo de EAGLE. Ces 500 quasars sont choisis aléatoirement de manière à respecter la distribution en redshift et en magnitude de l'échantillon initial. Cette méthode nous permet de calculer 170 359 500 rayons θ_E qui correspondent à la distribution des quasars et des galaxies déflectrices.

Mon objectif principal est de détecter les images multiples dans les données Gaia, ce qui nécessite une résolution angulaire d'au moins 300 millisecondes d'arc (mas). Pour satisfaire cette condition, les rayons d'Einstein peuvent être filtrés pour ne conserver que ceux qui peuvent donner des configurations avec une séparation d'au moins 300 mas. Les résultats de l'étude morphologique des halos des simulations EAGLE révèlent que ces halos présentent une ellipticité projetée relativement sphérique sur le plan du ciel, avec un rapport d'axes (b_p/a_p) d'environ 0,8. En utilisant cette information, on peut faire l'approximation simplifiée que les images observées des lentilles gravitationnelles seront séparées d'environ $2\theta_E$, où θ_E est le rayon d'Einstein. Par conséquent, nous pouvons sélectionner uniquement les rayons $\theta_E > 150$ mas.

A partir de l'ensemble de la distribution des rayons d'Einstein, je regarde la masse minimale pour ce rayon minimum. J'ai ainsi obtenu une masse minimale $M_{\min} = 4.39 \times 10^{10} M_{\odot}$ nécessaire pour former des configurations de lentilles gravitationnelles avec une séparation minimale de 300 mas. J'ai identifié ainsi un total de 12 255 halos provenant des simulations de l'univers EAGLE (RefL0025N0376 et RefL0025N0752), ce qui se traduit par 36 765 configurations distinctes une fois projetées dans trois directions différentes. Ces halos sont répartis sur 13 redshifts, s'étendant de $z = 0.1$ à $z = 1.74$.

Dans la section ci-dessous, j'explorerai la manière de simuler les différentes configurations à quatre images à partir de cette distribution de rayons d'Einstein que je viens de calculer.

IV.3.3 Tirage des positions des quasars

À partir des millions de rayons d'Einstein que nous avons calculés, mon objectif est de simuler toutes les configurations possibles d'alignement entre les halos d'EAGLE, agissant en tant que défecteurs, et les quasars en tant que sources d'arrière-plan. Par convention, je considère le centre des halos comme l'origine des simulations, ce qui signifie que l'alignement du système halo/quasar dépend uniquement de la position du quasar. Comme expliqué dans le chapitre consacré au formalisme des lentilles, afin d'obtenir des configurations à quatre images avec un modèle SIE, la source en arrière-plan doit être située à l'intérieur de la caustique tangentielle.

Pour résoudre l'équation de la lentille pour le modèle SIE, j'ai utilisé la méthode proposée par [Kormann et al. \(1994\)](#) et développée par Massimo Meneghetti dans son livre [\[Meneghetti, 2022\]](#). Le principe de base de cette méthode est la détermination d'une fonction $F(\phi)$.

$$y_1 = x_1 - \alpha_1(x) \tag{IV.40}$$

$$y_2 = x_2 - \alpha_2(x) \tag{IV.41}$$

En multipliant l'équation [IV.40](#) par $\cos \phi$ et l'équation [IV.41](#) par $\sin \phi$, nous obtenons les expressions suivantes :

$$y_1 \cos \phi = x_1 \cos \phi - \alpha_1(x) \cos \phi = x \cos^2 \phi - \alpha(x, \phi) \cos^2 \phi \tag{IV.42}$$

$$y_2 \sin \phi = x_2 \sin \phi - \alpha_2(x) \sin \phi = x \sin^2 \phi - \alpha(x, \phi) \sin^2 \phi \tag{IV.43}$$

Il est important de rappeler que $\alpha(x, \phi) = \psi(\phi)$ pour la lentille SIE. En combinant les équations [IV.42](#) et [IV.43](#), nous obtenons une expression permettant de calculer la distance depuis le centre de la lentille en fonction de l'angle ϕ :

$$x(\phi) = y_1 \cos \phi + y_2 \sin \phi + \psi(\phi) \tag{IV.44}$$

En réinsérant cette expression dans l'équation de la lentille, nous aboutissons à la fonction $F(\phi)$. Cette équation décrit mathématiquement la relation entre les coordonnées de la source et les images pour une lentille SIE.

1. Introduction to gravitational lensing / with python examples

$$F(\phi) = \left[y_1 + \frac{\sqrt{f}}{f'} \operatorname{arcsinh} \left(\frac{f'}{f} \cos \phi \right) \right] \sin \phi - \left[y_2 \frac{\sqrt{f}}{f'} \operatorname{arcsin} (f' \sin \phi) \right] \cos \phi \quad (\text{IV.45})$$

où $[y_1, y_2]$ est la position de la source, f le rapport d'axe, $f' = \sqrt{1 - f^2}$ et ϕ est l'angle polaire. Le problème se réduit donc à trouver les zéros de $F(\phi)$. Les solutions ne peuvent pas être trouvées analytiquement : un algorithme de recherche de racine tel que la méthode de Brent est employé (Brent 1972).

IV.3.4 Ajout d'un cisaillement externe ou *shear*

Dans le domaine des simulations de lentilles gravitationnelles fortes, il est essentiel de prendre en considération une composante appelée le shear externe. Contrairement au shear interne, qui émane de la distribution de masse dans le plan de la lentille, le shear externe provient de sources de masse situées le long de la ligne de visée. Cette distorsion supplémentaire joue un rôle significatif dans la déformation gravitationnelle observée. Intégrer le shear externe, ajouté à posteriori dans les simulations, est indispensable pour obtenir une représentation plus complète et précise des effets gravitationnels sur la lentille.

Le *shear* utilisé dans cette étude est basé sur la distribution estimée par [Holder and Schechter \(2003\)](#), qui a été obtenue à partir des simulations publiques du projet *Semi-Analytic Galaxy Formation - GIF* ([Kauffmann et al., 1999](#)). Ces simulations fournissent la distribution de matière noire ainsi que des estimations des positions, des vitesses, des luminosités et des couleurs des galaxies. La Figure [IV.7](#) présente la distribution du cisaillement qui a été adoptée pour cette étude.

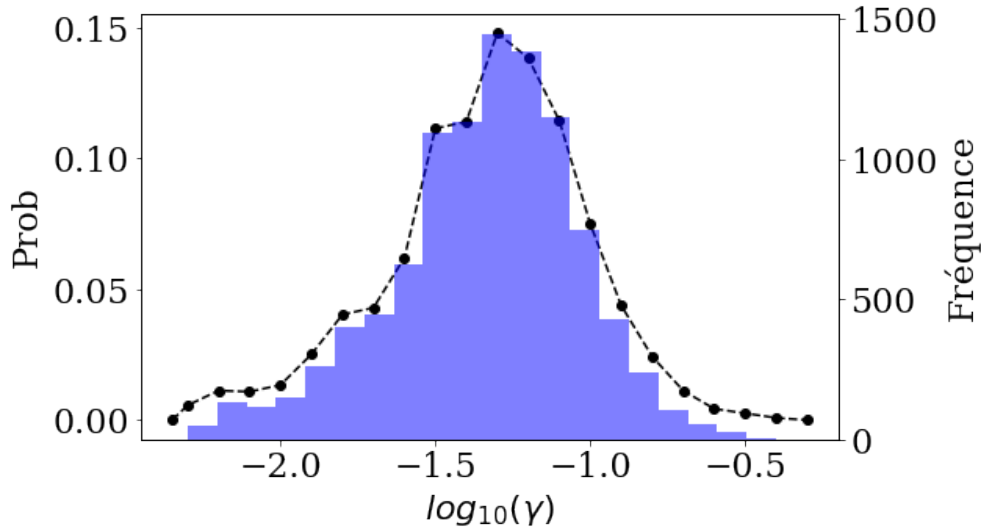


FIGURE IV.7 – Distribution du cisaillement (shear) provenant de l’étude de [Holder and Schechter \(2003\)](#) (représentée par la courbe en pointillés) et tirage aléatoire de $N = 10\,000$ valeurs de shear respectant cette distribution (représenté par l’histogramme bleu).

IV.3.5 Ajout des incertitudes sur la position des images

Les observations des lentilles gravitationnelles dans la réalité sont soumises à des incertitudes et des erreurs de mesure inhérentes aux instruments utilisés ainsi qu’aux propriétés des objets étudiés tel que l’effet de microlentille qui va introduire un déplacement par rapport à l’estimation initiale. Dans le contexte de lentillage gravitationnel fort, l’effet de microlentille fait référence aux variations ou perturbations à petite échelle causées par des étoiles individuelles ou des objets compacts au sein de la galaxie déflectrice. Le potentiel gravitationnel de la galaxie de premier plan entraîne une déviation de la lumière provenant de la source d’arrière-plan le long de plusieurs chemins, chacun correspondant à une image distincte de la source. Lorsque la lumière traverse des régions contenant des étoiles individuelles ou des objets compacts, le champ gravitationnel de ces objets induit de légères déviations dans la trajectoire des rayons lumineux. Ces événements de microlentille sont temporaires et entraînent des fluctuations faibles mais notables de la luminosité observée des images. La durée et l’amplitude de l’effet de microlentille dépendent de divers facteurs, tels que la masse et la distribution des objets de microlentille intermédiaire, le mouvement relatif entre la source, la lentille et l’observateur, ainsi que la taille de la source qui est soumise au lentillage. Une conséquence notable de cet effet de microlentille est la possibilité de modifier l’ordre d’amplification des images. Afin de tenir compte des sources d’incertitudes, nous bruitons les positions des images du quasar avec des valeurs typiques des erreurs de Gaia pour ces gammes de magnitudes. Étant donné la corrélation entre l’incertitude de mesure et la luminosité de la source, nous recueillons les magnitudes

des sources pour créer un ensemble de dix distributions, chacune centrée sur une magnitude donnée et couvrant l'ensemble des magnitudes simulées dans nos lentilles gravitationnelles. Cette approche nous permet de simuler des configurations plus fidèles à la réalité, en prenant en compte les erreurs de mesure susceptibles d'affecter les observations réelles.

IV.4 Préparation des données pour l'apprentissage supervisé

La recherche de lentilles gravitationnelles en utilisant l'apprentissage automatique supervisé est un défi complexe, car les données dont nous disposons sont limitées aux coordonnées et flux lumineux des sources détectées, sans accès à des images directes. Pour surmonter cette limitation, il est essentiel de construire un espace de représentation adéquat où le classement des objets sera facilité. Dans cet espace, les objets d'intérêt, tels que les configurations de lentilles gravitationnelles, peuvent être concentrés dans un ou plusieurs sous-espaces de plus petite dimensionnalité, tandis que les autres objets se situent à l'extérieur de ces sous-espaces. Par exemple, il peut exister des feuillettes distincts dans cet espace qui regroupent les configurations de lentilles gravitationnelles, facilitant ainsi le classement par l'algorithme XGBoost. En construisant préalablement des quantités discriminantes dans cet espace, si elles sont connues, nous pouvons améliorer l'efficacité et la robustesse de l'algorithme de classement.

IV.4.1 Alignement et normalisation de toutes les configurations

J'ai commencé par calculer le rapport des flux entre chaque paire d'images, ce qui permet de classer les images en fonction de leur amplification respective ou de leur flux relatif. Les images sont ensuite nommées de A à D, en ordre décroissant d'amplification. Ensuite, j'ai calculé toutes les distances possibles entre les différentes images de la lentille, ce qui donne six distances au total : \overline{AB} , \overline{AC} , \overline{AD} , \overline{BC} , \overline{BD} et \overline{CD} . Pendant ce calcul, j'ai extrait les valeurs des distances minimale (MinDist) et maximale (MaxDist).

J'ai simulé les configurations en tirant aléatoirement la position du quasar dans la caustique, ce qui conduit à des configurations orientées dans différentes directions et de tailles différentes. Afin de les ramener à un référentiel commun, j'ai procédé de la manière suivante. Tout d'abord, j'ai centré toutes les configurations sur l'image la plus brillante (A). J'ai fait cela car la galaxie déflectrice n'est pas toujours visible dans les observations réelles et c'est toujours la position de l'image la plus brillante qui est choisie comme l'origine du repère. Ensuite, pour choisir la direction du repère, j'ai orienté toutes les configurations de telle sorte que la seconde image la plus brillante (B) soit toujours alignée selon l'axe des ordonnées. Enfin, j'ai normalisé toutes les distances en divisant toute les coordonnées par la distance maximale (MaxDist). Examinons ce que cela donne pour l'ensemble des configurations que j'ai simulées. La figure [IV.8](#) présente la répartition des 4 images des configurations simulées, réorientées et normalisées. J'ai séparé les simulations de lentilles en deux catégories : la

première où je n'ai pas inclus de *shear* (15 085 configurations) et la seconde où j'ai ajouté un *shear* externe aléatoire mentionné dans la partie [IV.3.4](#) (44 339 configurations). Les nouvelles coordonnées de chaque image ont été renommées "Norm. x_1 " et "Norm. x_2 ".

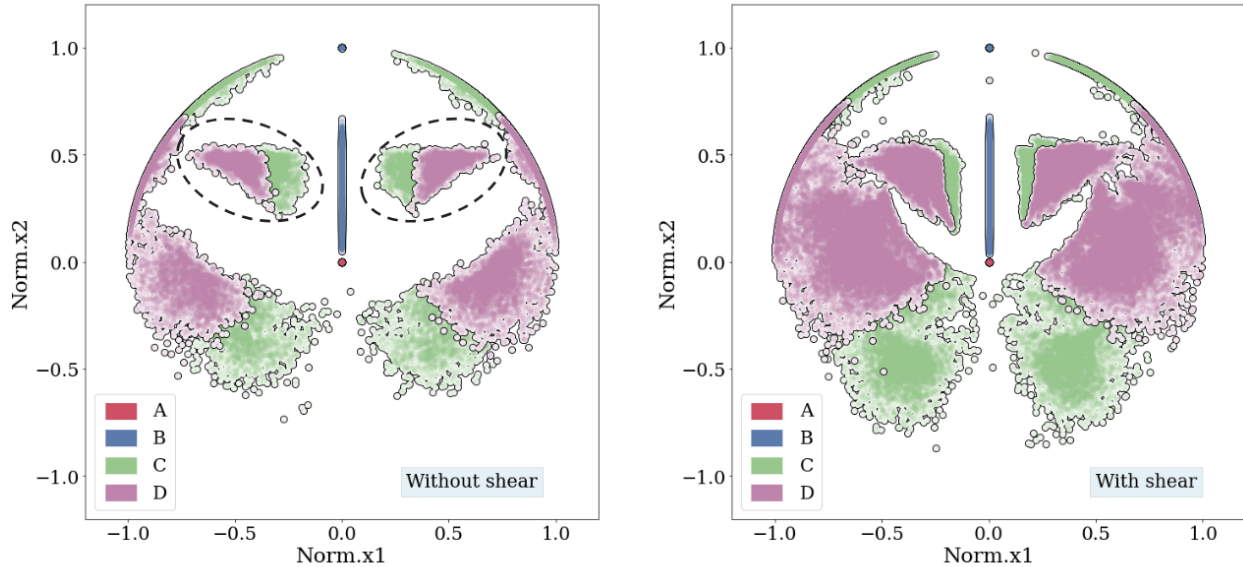


FIGURE IV.8 – Représentation des configurations simulées normalisées dans le plan "Norm. x_1 " et "Norm. x_2 ".

Sur cette figure, les images A, B, C et D sont respectivement associées aux régions en rouge, bleu, vert et rose. La première observation notable est que les points C et D tombent dans des zones bien séparées. En examinant la figure sans *shear* dans le panneau de gauche, nous distinguons deux zones centrales que j'ai entourée d'une ellipse de part et d'autre de l'axe des ordonnées. En sélectionnant manuellement les configurations pour lesquelles les images C et D tombent dans ces ellipses, j'ai remarqué que toutes les images B de ces configurations se concentrent au point bleu en haut de la figure, avec les coordonnées (0, 1). Il existe également deux autres zones particulières dans cette figure. Il s'agit des arcs présents dans la moitié supérieure pour les images C et D, ainsi que des extensions situées en bas pour les images C et D également. Nous verrons dans la prochaine section une méthode indirecte pour sélectionner toutes les configurations de ces zones et ainsi pouvoir les regrouper en trois catégories distinctes. Ces trois catégories sont liées aux angles que forment les images les unes avec les autres.

On constate également que le *shear* externe étend la distribution des coordonnées normalisées, couvrant ainsi une plus grande étendue spatiale.

IV.4.2 Calcul des angles entre les images

De manière similaire au calcul des distances, j'ai procédé au calcul de l'ensemble des angles pour chaque trio d'images. Ainsi, j'ai obtenu douze angles distincts : ABC , BAD , ADC , BCD , ABD , CBD , BAC , DAC , BDA , BDC , BCD , ACB , DCA . Comme les coordonnées des images sont interdépendantes, il est possible d'identifier les angles autorisés dans les configurations réalistes de lentilles gravitationnelles à quatre images.

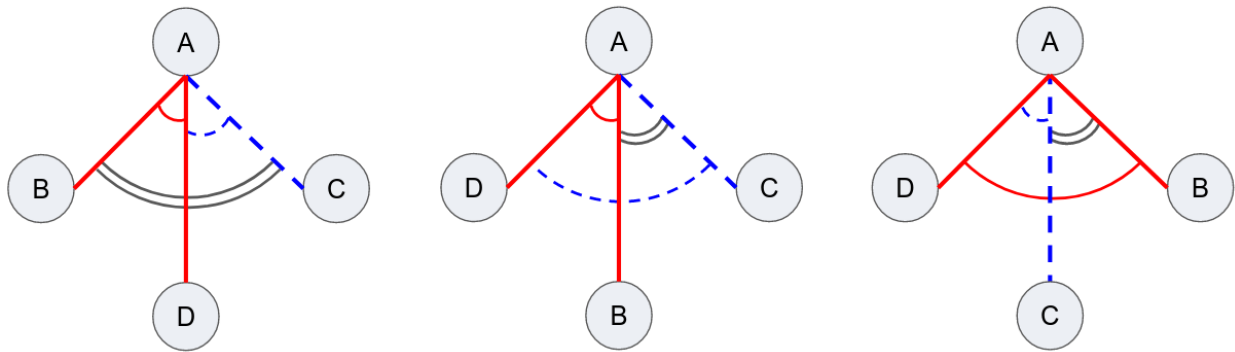


FIGURE IV.9 – Représentation des trois possibilités d'angles centrées sur l'image A, BAC , BAD et DAC pour une configuration de quatre images en fonction de la position relative de chacun.

Pour une lentille gravitationnelle, il existe trois configurations possibles en ce qui concerne l'ordre de luminosité des images, en les regardant dans le sens trigonométrique ou dans le sens horaire de la configuration. Ces trois possibilités sont représentées sur la figure [IV.9](#), où les angles BAD , DAC et BAC sont mis en évidence en fonction de la position de chaque image. La configuration que j'ai déjà identifiée dans la section précédente, où B est opposé à A et où C et D sont ordonnées de part et d'autre de l'axe \overline{AB} , correspond à la configuration centrale. Intuitivement, nous pouvons supposer que les deux autres zones de la figure [IV.8](#) correspondent aux deux autres cas. Pour vérifier cette hypothèse, j'ai examiné la distribution de ces trois angles BAD , DAC et BAC les uns par rapport aux autres.

La figure [IV.10](#) présente la distribution de l'angle BAC en fonction de la différence $BAD - DAC$. Le panneau de gauche montre la distribution sans cisaillement, tandis que le panneau de droite montre la distribution avec cisaillement. Nous pouvons distinguer trois zones clairement séparées, bien que la zone supérieure en rouge soit plus étendue dans le cas de la distribution avec cisaillement. Je vais maintenant examiner à quoi correspondent ces trois groupes dans le plan Norm. x_1 - Norm. x_2 .

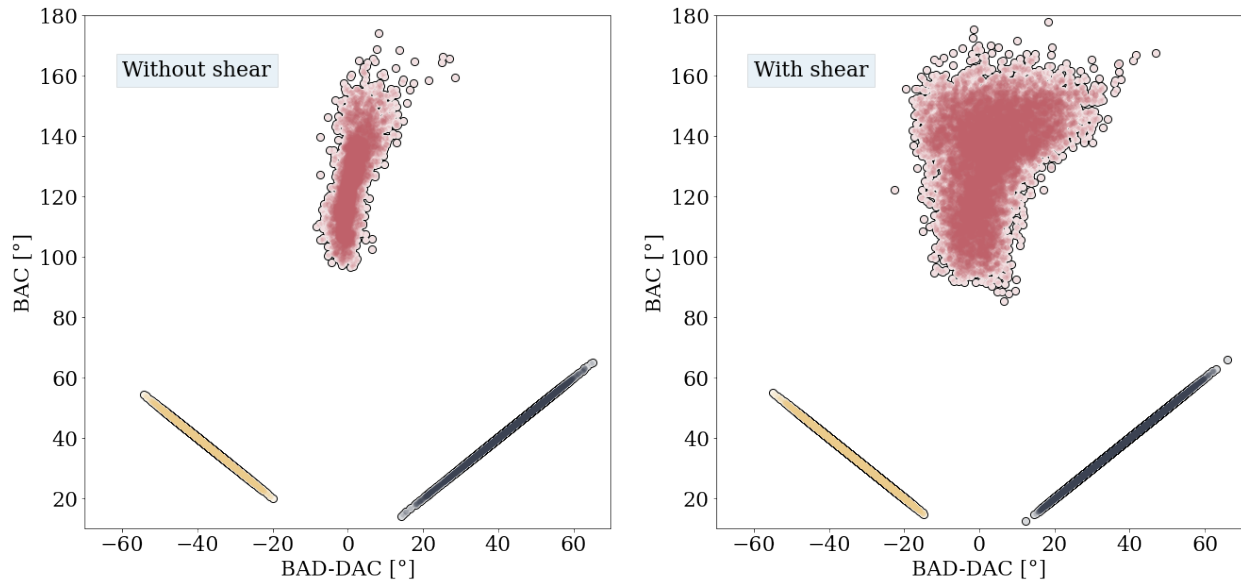


FIGURE IV.10 – Représentation sans cisaillement (panneau de gauche) et avec cisaillement (panneau de droite) de la différence des angles $BAD - DAC$ en fonction des angles BAC pour l'ensemble des configurations à quatre images simulées. Les couleurs indiquent trois zones distincts.

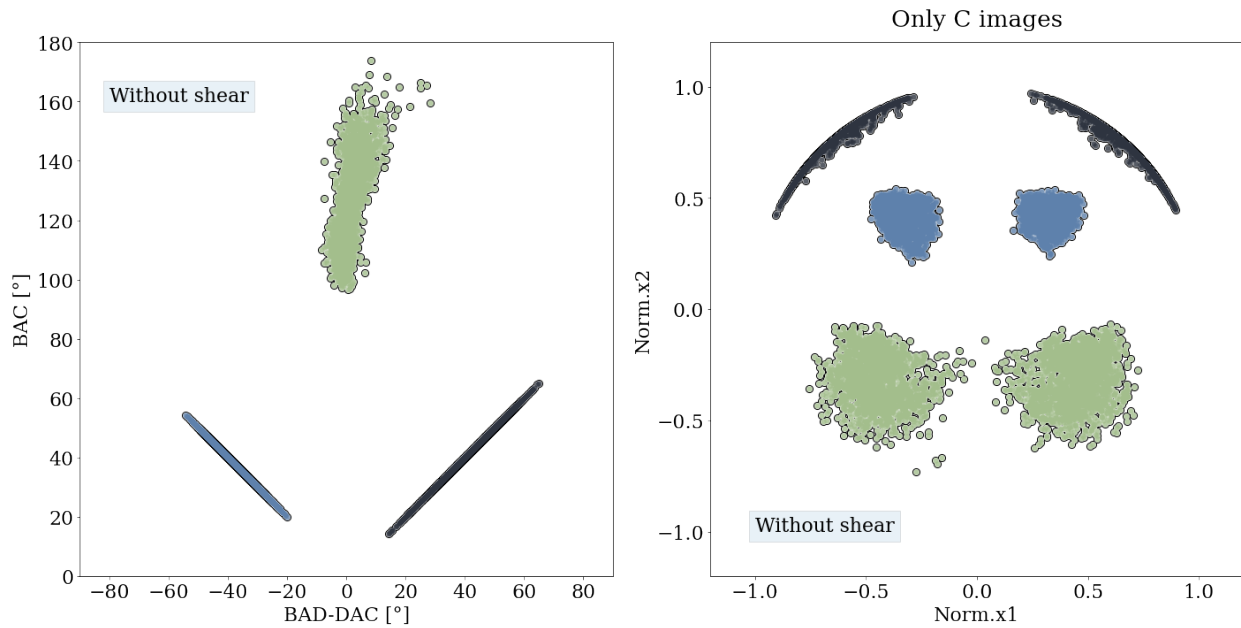


FIGURE IV.11 – Correspondance entre les zones dans le plan des angles (panneau de gauche) et le plan des coordonnées normalisées (panneau de droite). Dans le plan des coordonnées normalisées, seules les images C ont été représentées. Les couleurs sont associées à chacune des zones du plan des angles.

La figure [IV.11](#) présente les trois zones distinctes identifiées précédemment dans le panneau de gauche, ainsi que les trois zones correspondantes dans le plan Norm. x_1 - Norm. x_2 dans le panneau de droite. Pour plus de clarté, j'ai représenté uniquement les coordonnées des images C des configurations de lentilles gravitationnelles. Nous pouvons observer trois zones distinctes qui correspondent aux zones identifiées précédemment. La zone bleue représente toutes les configurations situées au centre de la figure des angles [IV.9](#). Ensuite, la zone noire correspond aux configurations en forme d'arc où le point C est à l'opposé de A, c'est-à-dire que le point C se trouve régulièrement à la distance maximale, bien que ce ne soit pas toujours le cas. Enfin, la dernière configuration concerne les cas où le point D est à l'opposé de A. Cette représentation met en évidence que le point C et, de la même manière, le point D ne peuvent pas être situés n'importe où dans le plan Norm. x_1 - Norm. x_2 .

Examinons maintenant les contraintes associées à ces différentes zones. La figure [IV.12](#) présente les zones couvertes par les images C (panneau de gauche) et les images D (panneau de droite). Ces configurations sont issues des simulations avec cisaillement, d'où la dispersion plus étendue des points. Chaque zone est numérotée pour faciliter la lecture de la figure. L'interprétation de cette figure est la suivante : si une image C se trouve dans la zone 1, alors l'image D correspondante doit également se trouver dans la zone 1. Ainsi, une contrainte forte est imposée sur la position relative des images C et D.

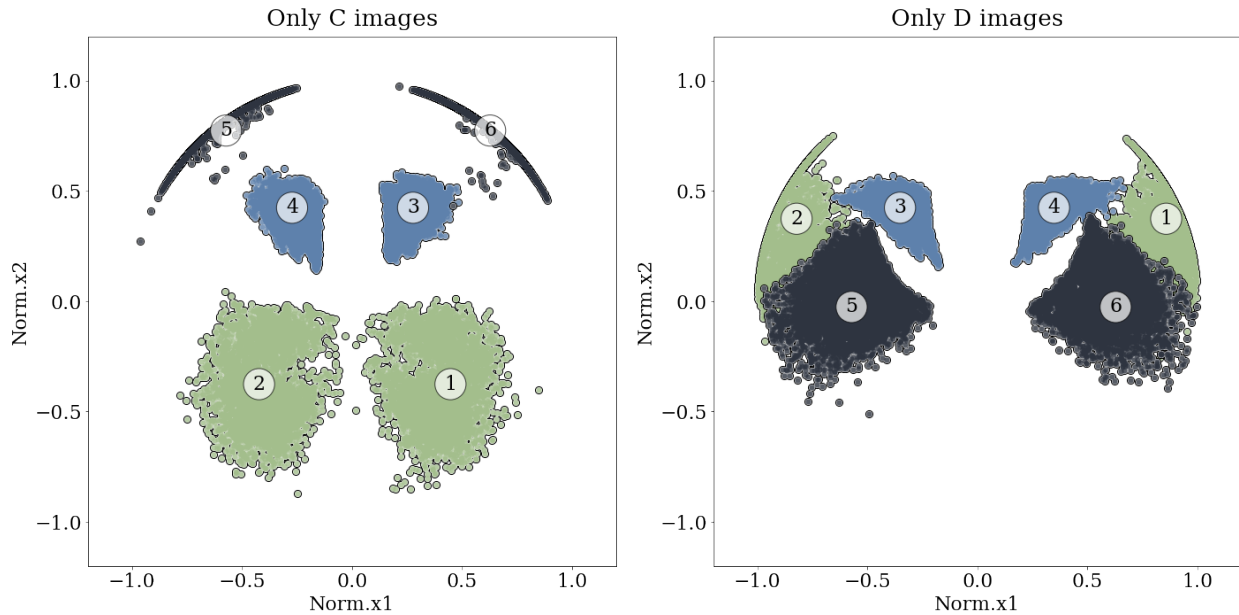


FIGURE IV.12 – Correspondance entre les zones des images C (panneau de gauche) et des images D (panneau de droite) dans le plan "Norm. x_1 " et "Norm. x_2 ". Chaque couleur est associée aux différents angles identifiés dans la figure [IV.9](#). Les chiffres représentent l'association des images C et D de chaque configuration.

Cette visualisation me permet de définir des zones de forte contraintes sur des configurations de

candidats. J'aborderai les contraintes de ces zones dans le chapitre [VI](#).

IV.4.3 Sélection des sources qui ne sont pas des quads

Pour construire un catalogue d'entraînement à partir des quads simulés, il est important d'inclure une deuxième classe qui comprend des configurations différentes de celles des quads. Ces configurations peuvent être générées en sélectionnant de manière aléatoire des ensembles de quatre points ou en utilisant des ensembles de quatre étoiles observées par Gaia, pour lesquelles nous avons des mesures de flux pour chaque image. Dans un souci de réalisme, j'ai opté pour l'extraction des configurations à partir des données de Gaia, données par [Delchambre et al. \(2019\)](#). Le critère de sélection principal repose sur la taille de la configuration. Étant donné que les quads simulés présentent une séparation maximale entre deux images d'environ 6 secondes d'arc, seules les configurations d'étoiles plus petites ont été choisies, car nous recherchons des configurations compactes.

Cette sélection de sources a été initialement effectuée dans le but de rechercher des candidats dans les données de Gaia DR2 (Data Release 2). Toutefois, dans le cas présent, mon objectif est d'utiliser ce catalogue de sources non pas pour identifier des candidats, mais pour construire un catalogue d'entraînement ne contenant que des configurations *a priori* d'étoiles.

Les critères initiaux de sélection qui ont été utilisés sont les suivants : (i) les images de lentilles doivent posséder des parallaxes, $\bar{\omega}$, et des mouvements propres, $(\mu_{\alpha^*}, \mu_{\delta})$, où $\mu_{\alpha^*} = \mu_{\alpha} \cos \delta$, négligeables puisque les images des quasars lentillés n'ont pas de mouvements propres. (ii) Il doit y avoir au moins quatre images afin de fournir un nombre suffisant de contraintes pour l'identification des candidats de lentilles. (iii) La séparation angulaire maximale entre n'importe quelle paire d'images doit être inférieure à 6 arcsec. Et enfin, (iv) la différence absolue de magnitude G entre les composantes doit être inférieure à 4 mag. J'ai introduit un critère supplémentaire basé sur les résultats obtenus dans l'étude de [Delchambre et al. \(2019\)](#). J'ai décidé de ne sélectionner que les groupes d'objets ayant obtenu les scores d'être une lentille les plus bas. En sélectionnant les groupes avec les scores de prédiction les plus faibles, j'ai privilégié des configurations qui sont moins susceptibles de correspondre à des lentilles gravitationnelles, ce qui renforce la robustesse de mon catalogue d'entraînement.

J'ai appliqué ensuite la même méthode de normalisation utilisée pour les simulations de lentilles afin de calculer les paramètres supplémentaires. Pour ce faire, j'ai établi l'ordre des images en fonction de leur luminosité, en utilisant les valeurs de magnitude dans la bande G mesurées par Gaia. Ensuite, j'ai calculé les nouvelles coordonnées normalisées pour chaque image en respectant les conditions énoncées précédemment. Cette approche de normalisation permet d'obtenir des paramètres identiques et comparables pour toutes les configurations d'étoiles, facilitant ainsi l'apprentissage automatique.

La figure [IV.13](#) met en évidence la distinction entre les simulations de lentilles gravitationnelles avec cisaillement (panneau de gauche) et les groupes d'étoiles provenant de Gaia (panneau de droite). Il

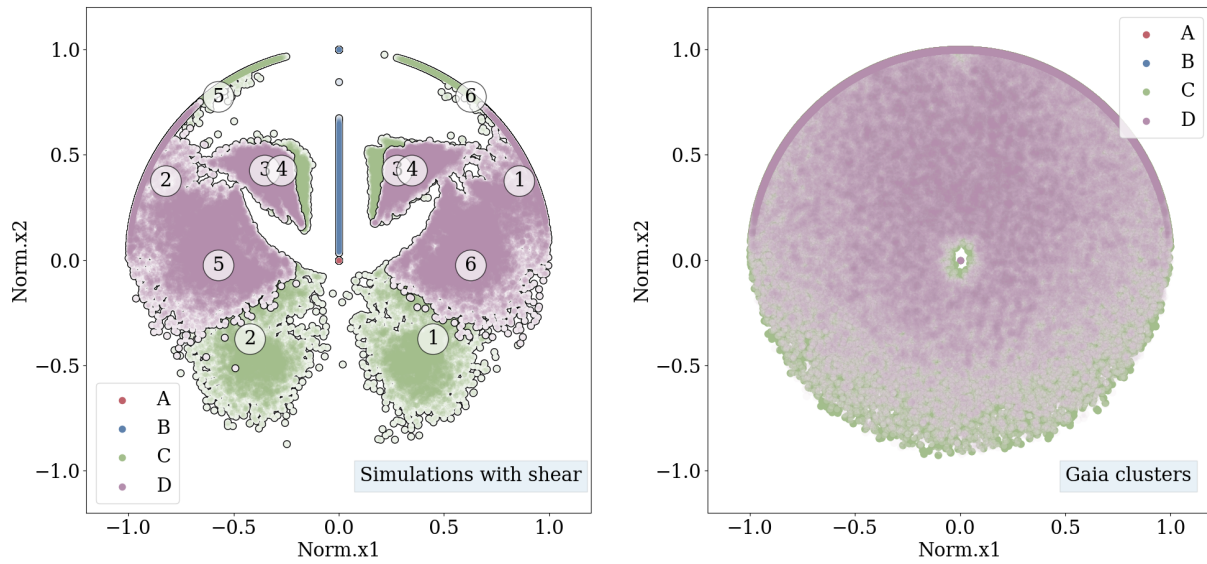


FIGURE IV.13 – Correspondance entre les zones des images C et des images D des simulations avec du *shear* dans le plan normalisé. Les chiffres représentent l’association des images C et D de chaque configuration. Panneau de droite : représentation de la position des images C et D des groupes d’objets de Gaia . Les images C sont situées sous les images D au premier plan.

est observé que les groupes d’étoiles de Gaia ne présentent aucun ordre spécifique entre les images. Cela indique que les images ABCD de ces groupes d’étoiles peuvent être disposées dans n’importe quel ordre et n’importe quel angle, contrairement aux simulations de lentilles gravitationnelles. C’est donc bien une contrainte forte pour la recherche de lentilles gravitationnelles.

Dans le prochain chapitre, j’aborderai l’apprentissage de l’algorithme XGBoost à partir des données présentées dans ce chapitre en constituant plusieurs catalogues d’entraînement.

Chapitre V

Apprentissage supervisé pour la recherche de quads

Dans le cadre de ce chapitre, nous procédons à l'entraînement de l'algorithme de machine learning XGBoost sur le catalogue d'entraînement que j'ai construit dans le chapitre précédent. Tout d'abord, je m'attarderai sur l'espace des paramètres utilisé et j'expliquerai la méthodologie de sélection que nous avons adoptée. Ensuite, je procéderai à l'entraînement de l'algorithme et j'évaluerai ses performances sur des quads confirmés et non utilisés pour l'entraînement, afin d'évaluer sa capacité de discernement entre les lentilles gravitationnelles et les groupes d'étoiles.

V.1 Espace des paramètres pour l'entraînement supervisé

Au cours du chapitre précédent, j'ai effectué des simulations pour obtenir un catalogue réaliste de quads. J'ai également récupéré un catalogue de groupes d'étoiles à partir des données de Gaia DR2, permettant ainsi de construire un catalogue d'entraînement réaliste combinant les deux types d'objets. La table [V.1](#) récapitule l'ensemble des paramètres disponibles pour l'entraînement de l'algorithme de machine learning XGBoost ([Chen and Guestrin, 2016](#)). Nous avons donc un total de 41 paramètres pour l'entraînement.

V.1.1 Entraînement de XGBoost sur plusieurs sous-catalogues

Dans le but d'évaluer les performances de l'algorithme XGBoost pour l'apprentissage des configurations de lentilles gravitationnelles, j'ai constitué plusieurs catalogues d'entraînement pour ce qui concerne les simulations de lentilles en considérant ou pas un cisaillement. Cette dichotomie permettra d'analyser les différences de performance de l'algorithme d'apprentissage dans ces deux cas. Une autre motivation pour créer plusieurs catalogues d'entraînement a été la suivante : les

TABLE V.1 – Ensemble des paramètres des catalogues d’entraînements

Paramètres	Description
NormX_ima_i (A-D)	Coordonnées X normalisées et orientées des images A-D dans le plan "Norm. x_1 " et "Norm. x_2 "
NormY_ima_i (A-D)	Coordonnées Y normalisées et orientées des images A-D dans le plan "Norm. x_1 " et "Norm. x_2 "
d1 - d6	Ensemble des six distances possibles entre les quatre images
Nd1 - Nd6	Ensemble des six distances normalisées par la distance maximale entre les quatre images
MaxDist	Distance maximale entre les quatre images
MinDist	Distance minimale entre les quatre images
Nmu_1 - Nmu_4	Rapports de luminosité entre les quatre images normalisées par l’image la plus lumineuse
Geo_flux_ratio_1, ..., Geo_flux_ratio_3	Rapports de luminosité des images en fonction de l’ordre géométrique
ABC, ..., DAC	Ensemble des 12 angles possibles entre les quatre images

simulations sont basées sur les halos provenant de simulations, ce qui signifie que la quantité de lentilles simulées dans chaque gamme de masse influence le contenu du catalogue d’entraînement. En particulier, les halos massifs avec une masse supérieure à $10^{12} M_{\odot}$ sont relativement rares dans la simulation que nous avons analysée. Lorsque des tirages aléatoires sont effectués pour créer le catalogue d’entraînement, il est donc fortement vraisemblable que très peu de halos de cette masse et au-delà soient sélectionnés. Afin de surmonter cette limitation, des contraintes ont été introduites lors de certains tirages pour inclure un nombre significatif de halos très massifs. Ces catalogues spécifiques sont appelés ci-après "BigHalos". La table [V.2](#) récapitule les différents catalogues de simulations de lentilles utilisés pour l’entraînement.

La figure [V.1](#) présente la distribution en nombre de 16 des 41 paramètres les plus porteurs d’information dans le catalogue no.2 prenant en compte un cisaillement. L’analyse de ces distributions permet de faire plusieurs observations. Tout d’abord, nous constatons une différence significative entre les distances caractérisant les configurations de lentilles gravitationnelles et celles des configurations d’étoiles. Cette différence est particulièrement marquée au niveau de la distance maximale, où les deux distributions se chevauchent très peu. Cela met en évidence que les configurations de lentilles gravitationnelles sont beaucoup plus compactes en moyenne que les configurations d’étoiles à quatre images. Afin d’éviter tout rejet systématique des configurations étendues de plus de 4 secondes d’arc, il est important de ne pas se fier uniquement à ce paramètre pour l’apprentissage de XGBoost.

Pour prendre en compte cette différence marquée sur les distances, en plus des quatre catalogues

TABLE V.2 – Liste des catalogues pour l’entraînement de XGBoost.

Paramètres	Description
Catalogue no.1 : NoShear	Catalogue sans cisaillement et sans halos massif supplémentaire
Catalogue no.2 : Shear	Catalogue avec cisaillement et sans halo massif supplémentaire
Catalogue no.3 : NoShearBigHalos	Catalogue sans cisaillement et avec des halos massifs supplémentaires
Catalogue no.4 : ShearBigHalos	Catalogue avec cisaillement et halos massifs supplémentaires

d’entraînement, j’ai effectué deux entraînements. Le premier entraînement est réalisé en utilisant uniquement les critères morphologiques, tels que les positions normalisées ($NormX_ima_i$ et $NormY_ima_i$) pour chaque image (A-D) et les angles. Le second entraînement inclut les critères morphologiques ainsi que tous les critères de taille, tels que les distances d1-d6, Nd1-Nd6 et les distances MaxDist et MinDist. Cette approche en deux étapes permet d’évaluer d’abord la géométrie du phénomène en établissant un premier score, puis d’observer comment ce score évolue en considérant également les caractéristiques de taille du phénomène. Ainsi, en sortie de l’algorithme XGBoost, il est possible d’obtenir huit scores de probabilité : P_{Gi} , avec i variant de 1 à 4, correspondant respectivement aux quatre catalogues (Table [V.2](#)) pour l’entraînement géométrique, ainsi que P_{Di} , avec i variant de 1 à 4, pour l’entraînement incluant les distances.

Une autre remarque importante à souligner à partir de la figure [V.1](#) concerne la distribution des angles entre les différentes images. Comme je l’ai montré dans le chapitre précédent, les configurations de lentilles gravitationnelles à quatre images occupent des positions spécifiques dans l’espace des paramètres. J’ai identifié six zones distinctes dans lesquelles les images C et D sont situées. Cette tendance à occuper un sous-espace particulier est aussi constatée dans la distribution des angles entre les images. En effet, nous pouvons observer (figure [V.1](#)) que les angles présentent des répartitions bien définies, ce qui suggère qu’ils portent une information discriminante pour l’apprentissage de XGBoost.

Examinons maintenant le choix des hyperparamètres de XGBoost. Pour rappel, les hyperparamètres sont des paramètres ajustables avant l’entraînement d’un modèle d’apprentissage qui influencent le comportement de l’algorithme et la performance du modèle résultant. Ils ne sont pas appris à partir des données, mais doivent être définis avant le processus d’entraînement. Parmi ces paramètres, nous avons :

- (i) Le nombre d’arbres de décision à entraîner dans l’algorithme XGBoost, appelé "n_estimators".

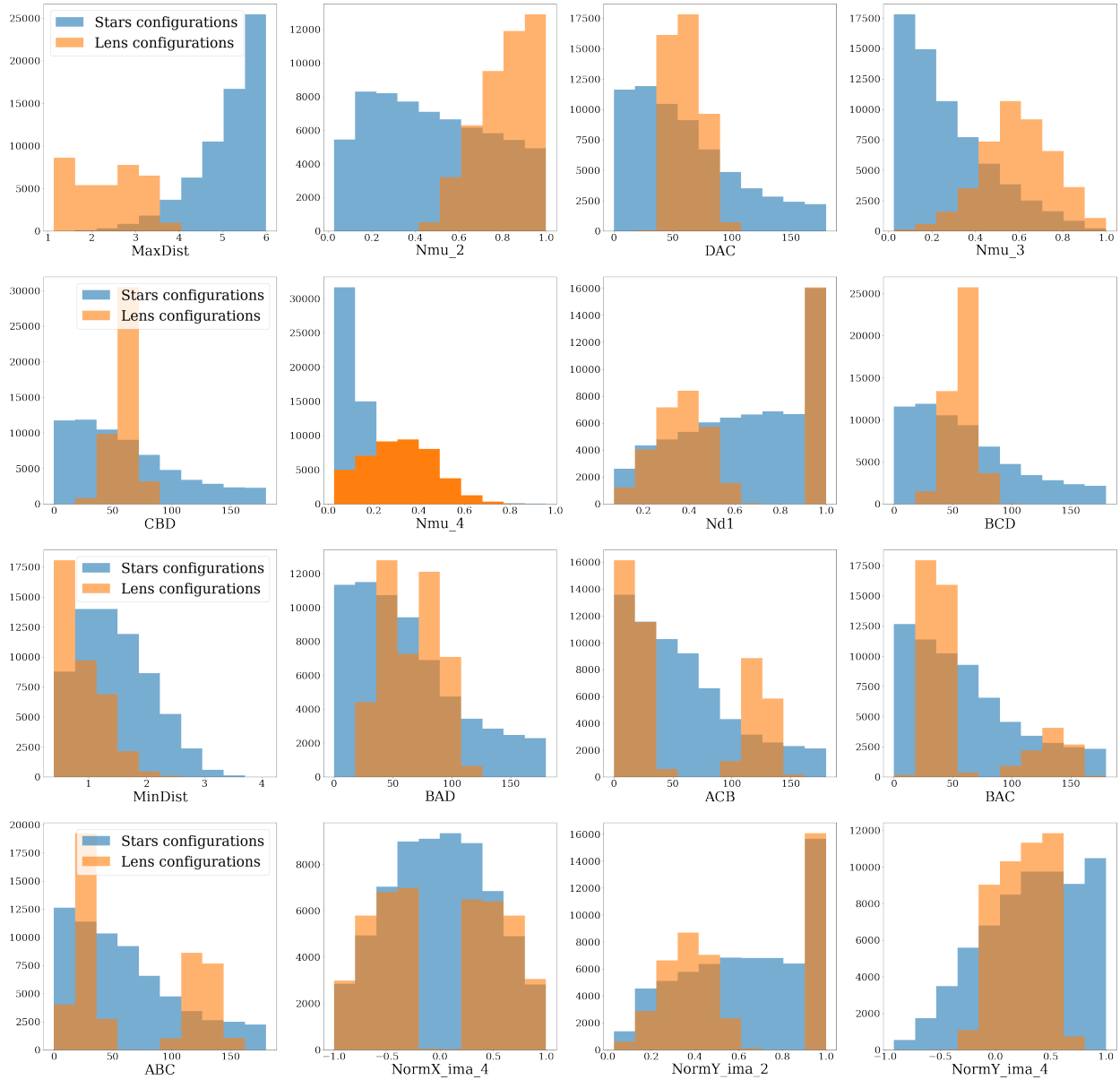


FIGURE V.1 – Histogrammes de 16 des 41 paramètres les plus porteurs d'information dans le catalogue d'entraînement no.2 prenant en compte un cisaillement selon le critère Feature Importance décrit dans la section [V.1.2](#)

XGBoost construit des arbres de décision ultérieurs pour corriger les erreurs commises par les arbres précédents. Chaque arbre suivant se concentre sur la prédiction des résidus de l'arbre précédent.

(ii) La profondeur maximale des arbres de décision, appelée "max_depth". Une valeur plus élevée permet au modèle d'apprendre des relations plus complexes dans les données, mais peut également conduire à un surajustement (*overfitting*).

(iii) Le taux d'apprentissage, également appelé "learning_rate", qui détermine la vitesse à laquelle l'algorithme apprend à partir des données. Le taux d'apprentissage détermine le poids de chaque arbre à la prédiction finale. Il met à l'échelle les prédictions de chaque arbre en les multipliant par le taux d'apprentissage. Un taux d'apprentissage inférieur à 1 réduit la contribution de chaque arbre, tandis qu'un taux d'apprentissage supérieur à 1 amplifie la contribution. Un taux d'apprentissage faible introduit un effet de réduction, rendant l'algorithme plus conservateur. Cela oblige chaque nouvel arbre à accorder plus d'attention aux résidus, réduisant ainsi l'impact d'un seul arbre. Cela peut aider à éviter le surajustement et à améliorer la généralisation du modèle.

Par défaut, j'ai sélectionné les hyperparamètres suivants pour l'entraînement : un nombre d'arbres de 10 avec une profondeur de 5. Le taux d'apprentissage est fixé à 0.3.

La première étape de l'apprentissage est l'étape de prétraitement des données ou de préparation des données. Lors de cette étape, on sépare le jeu de données en deux parties distinctes : les variables indépendantes (X), qui représentent les caractéristiques ou les variables prédictives, et la variable dépendante (y), qui représente la variable cible ou la classe que l'on cherche à prédire. Dans notre cas, j'ai nommé la variable dépendante "TargetGL" qui vaut 1 si l'objet est de la classe groupe d'étoiles, et -1 si l'objet est de la classe lentille gravitationnelle de quasar.

La prochaine étape consiste à diviser les données en un ensemble d'entraînement, un ensemble de validation et un ensemble de test comme mentionné dans la section [III.3](#). Dans notre cas, j'ai choisi que 70% des données seront utilisées pour l'ensemble d'entraînement, 15% des données seront utilisées pour l'ensemble de validation et les 15% restants seront utilisés pour l'ensemble de test.

Enfin, avant de lancer l'entraînement à proprement parler, j'ai utilisé la matrice de données DMatrix. Cette matrice est optimisée pour un stockage efficace des données et un accès rapide lors de l'entraînement du modèle. Elle utilise des structures de données compactes et des formats compressés pour économiser de l'espace mémoire. De plus, elle exploite des optimisations internes dans XGBoost, telles que le partitionnement des données et le calcul parallèle, ce qui accélère considérablement le temps de calcul.

Voici le bloc de code python permettant de créer la matrice de données et de réaliser l'entraînement :

```
1 import xgboost as xgb
2
3 # Charger les donnees et creer la matrice optimisee pour XGBoost
4 data = xgb.DMatrix(X, label=y)
5
6 # Choix des hyperparametres
7 params = {'max_depth': 5, 'eta': 0.3, 'objective': 'binary:logistic'}
8 num_boost_round = 10 # Nombre d'arbres (iterations)
9
```

```
10 # Entraînement du modele
11 model = xgb.train(params, data, num_boost_round=num_boost_round)
```

V.1.2 Résultat des entraînements

Pour suivre la performance de XGBoost pendant l’entraînement et ajuster les hyperparamètres, on utilise l’ensemble de validation. L’ensemble de validation a été défini précédemment comme étant un échantillon aléatoire représentant 15% des données.

Analysons les performances de XGBoost sur le catalogue no.2 comprenant le cisaillement (*shear*) et excluant les grands halos supplémentaires. C’est le catalogue le plus réaliste dans le sens où il respecte les distributions de paramètres des simulations EAGLE et la distribution de quasars sans ajout de halos artificiels. J’ai considéré ici l’entraînement avec l’ensemble des 41 paramètres. Je commence par entraîner un modèle XGBoost initial en utilisant les données d’entraînement avec les hyperparamètres par défaut : un nombre maximum d’arbres de 10, une profondeur maximale de 3, et un taux d’apprentissage de 0,1. Ensuite, j’évalue les performances de ce modèle initial en utilisant les données de validation, obtenant ainsi une première estimation des performances.

L’ajustement des hyperparamètres intervient à ce stade. Je peux le faire manuellement ou en utilisant des techniques automatisées comme la recherche par grille (*GridSearchCrossValidation*). À cette étape, je me limite à utiliser les données d’entraînement pour réentraîner le modèle en utilisant le nouveau jeu d’hyperparamètres sélectionné. La méthode *GridSearchCrossValidation* est une technique automatisée d’ajustement d’hyperparamètres. Elle explore différentes combinaisons prédéfinies d’hyperparamètres pour un modèle, évalue chaque combinaison à l’aide de la validation croisée (*cross-validation*) et sélectionne celle qui offre les meilleures performances. Par exemple, l’utilisation de la méthode de la recherche par grille avec l’entraînement du catalogue no.2, on obtient les hyperparamètres suivants : un nombre maximum d’arbres de 500, une profondeur maximale de 7, et un taux d’apprentissage de 0,1.

Après avoir réglé les hyperparamètres, je passe à l’entraînement du modèle final. Cette fois, j’utilise à la fois les données d’entraînement et de validation pour garantir une meilleure généralisation du modèle. Pour évaluer la performance de l’entraînement final, j’utilise la matrice de confusion. Cette matrice est une représentation tabulaire qui compare les prédictions d’un modèle aux véritables valeurs de la classe. La matrice de confusion est obtenue en comparant les prédictions du modèle avec les classes réelles des données de test. Elle se divise en quatre sections :

- Vrais Positifs (VP) : Les cas où le modèle prédit correctement une classe positive.
- Vrais Négatifs (VN) : Les cas où le modèle prédit correctement une classe négative.
- Faux Positifs (FP) : Les cas où le modèle prédit à tort une classe positive (fausse alarme).

— Faux Négatifs (FN) : Les cas où le modèle prédit à tort une classe négative (omission).

La matrice de confusion offre une vue détaillée des performances d'un modèle de classification en identifiant les erreurs de prédiction et en quantifiant sa capacité à distinguer les différentes classes. Elle constitue un outil essentiel pour évaluer la qualité d'un modèle et ajuster ses paramètres en conséquence.

	Classe réelle positive	Classe réelle négative
Prédiction positive	65 658 (VP)	35 (FP)
Prédiction négative	19 (FN)	44 320 (VN)

TABLE V.3 – Matrice de confusion du modèle d'entraînement d'XGBoost à partir du catalogue no.2.

La table [V.3](#) présente la matrice de confusion de l'apprentissage de XGBoost à partir du catalogue d'entraînement no.2 : Shear. La classe positive correspond aux groupes d'étoiles et la classe négative correspond aux lentille gravitationnelle de quasar. Pour évaluer la performance d'un modèle de classification, on utilise souvent différentes métriques. Voici quatre métriques couramment utilisées, chacune apportant une perspective unique sur la qualité des prédictions du modèle :

1. **Précision :**

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

La précision mesure la proportion de prédictions positives du modèle qui sont effectivement correctes. Elle est particulièrement utile lorsque la minimisation des faux positifs est cruciale.

2. **Rappel (Sensibilité) :**

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Le rappel indique la capacité du modèle à détecter la totalité des instances de la classe positive. Il est important dans les situations où la non-détection de cas positifs peut avoir des conséquences graves.

3. **Exactitude :**

$$\text{Exactitude} = \frac{\text{VP} + \text{VN}}{\text{Total des Échantillons}}$$

L'exactitude mesure globalement la proportion de prédictions correctes du modèle, qu'elles soient positives ou négatives. C'est une métrique générale de la performance du modèle.

4. **Indice de Matthews (MCC) :**

$$\text{MCC} = \frac{(\text{VP} \times \text{VN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{VP} + \text{FP})(\text{VP} + \text{FN})(\text{VN} + \text{FP})(\text{VN} + \text{FN})}}$$

L'indice de Matthews est une mesure de la qualité globale des prédictions du modèle, prenant en compte tous les éléments de la matrice de confusion. Il varie de -1 (prédictions parfaitement opposées) à +1 (prédictions parfaites), avec 0 indiquant des performances aléatoires.

Métrique	Score
Précision	0.9992
Rappel (Sensibilité)	0.9996
Exactitude	0.9995
Indice de Matthews (MCC)	0.9990

TABLE V.4 – Scores du modèle d'entraînement d'XGBoost à partir du catalogue no.2.

En examinant les résultats obtenus (table [V.4](#)), on constate de très bonnes performances du modèle de classification. La précision, mesurant la proportion de prédictions positives correctes parmi toutes les prédictions positives, est élevée à 0.9992. De même, le rappel, indiquant la capacité du modèle à détecter la totalité des instances de la classe positive, atteint 0.9996, soulignant une très bonne sensibilité. L'exactitude globale du modèle, qui prend en compte à la fois les prédictions positives et négatives, est également très bonne à 0.9995. Enfin, l'indice de Matthews (MCC), mesurant la qualité globale des prédictions en prenant en compte tous les éléments de la matrice de confusion, est proche de 0.9990, ce qui confirme la robustesse du modèle.

L'exactitude de ces entraînements peut être comparée aux 0,98 obtenus par le meilleur classificateur de lentille du "*Strong Gravitational Lens Finding Challenge*" ([Lanusse et al., 2018](#)) et aux 0,99503 de l'entraînement des ERT sur des configurations à quatre images ABCD de l'étude de [Delchambre et al. \(2019\)](#). Le "*Strong Gravitational Lens Finding Challenge*" est une compétition dont l'objectif a été de développer des modèles de *machine learning* capables de détecter avec précision les lentilles gravitationnelles fortes dans des images astronomiques provenant de télescopes (SDSS et KiDS) et d'observatoires spatiaux (Euclid). Les participants ont eut accès à des ensembles de données étiquetés, comprenant des images où des lentilles gravitationnelles étaient présentes. Pour atteindre les objectifs, des méthodes automatisées ont été développées, impliquant des approches telles que l'inspection visuelle, la détection d'arcs et d'anneaux, l'utilisation de machines à vecteurs de support (SVM) et de réseaux neuronaux convolutionnels (CNN). Les participants ont été chargés de classer 100 000 objets candidats comme étant des lentilles gravitationnelles ou non, dans le but de perfectionner les méthodes automatisées pour repérer ces lentilles dans de grands ensembles de données.

Nos résultats suggèrent que notre modèle est très performant dans la classification des données, avec une capacité à minimiser les erreurs de prédiction. Cependant, il est toujours important d'évaluer d'autres métriques.

Pour cela, nous commençons par examiner ce que l'on appelle l'importance des fonctionnalités

(feature importance). Cette mesure permet de déterminer quelles fonctionnalités ont le plus d'impact sur les prédictions et peuvent donc être considérées comme les plus informatives pour le modèle. L'importance des fonctionnalités est calculée par XGBoost en utilisant différentes méthodes, telles que la fréquence d'utilisation d'une fonctionnalité lors de la construction des arbres de décision et l'amélioration moyenne du score de division obtenue grâce à cette fonctionnalité. La figure [V.2](#) présente l'importance des fonctionnalités (F score) pour cet entraînement.

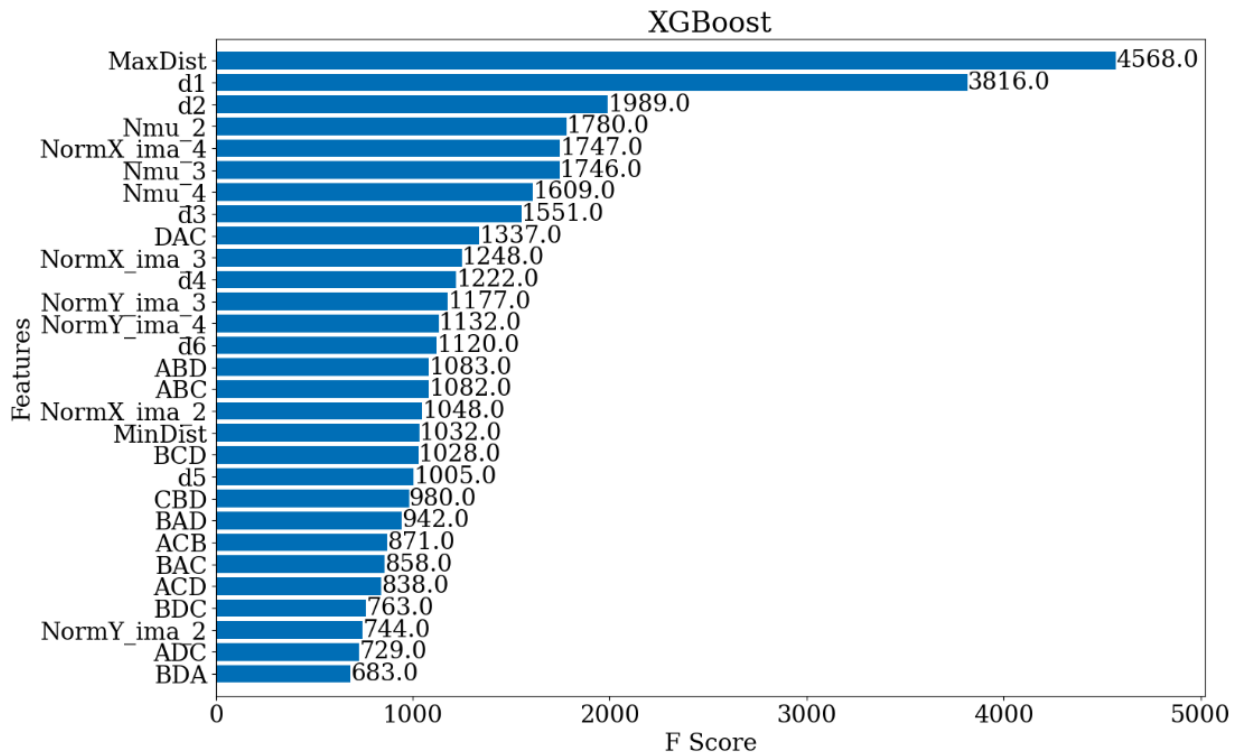


FIGURE V.2 – La figure présente l'importance des fonctionnalités dans le modèle XGBoost entraîné sur le catalogue no.2 : Shear. L'importance des fonctionnalités mesure leur contribution relative aux prédictions du modèle. Les fonctionnalités sont ordonnées de haut en bas en fonction de leur importance décroissante, avec des barres plus longues indiquant une plus grande importance.

L'analyse de l'importance des fonctionnalités révèle que le paramètre le plus crucial utilisé par XGBoost pour classer les configurations est la distance maximale entre les images, appelée MaxDist. Cette observation était attendue, car nous avons précédemment remarqué que les configurations d'étoiles ont en moyenne une séparation maximale plus grande que les configurations de lentilles gravitationnelles provenant de distributions de paramètres réalistes. Ainsi, il est essentiel de réaliser des apprentissages supplémentaires sans prendre en compte l'information sur les distances entre les images si l'on souhaite éviter que XGBoost ne rejette automatiquement toutes les configurations étendues. Parmi les autres paramètres importants, on retrouve les rapports de luminosité (Nmu) et les angles, ce qui est cohérent avec les histogrammes observés dans le catalogue d'entraînement,

comme illustré à la figure [V.1](#). Cette constatation confirme l'importance de ces caractéristiques pour la classification précise des configurations de lentilles gravitationnelles.

Suite à l'évaluation des performances de notre modèle, nous allons désormais aborder une étape supplémentaire de validation. Dans la section suivante, nous utiliserons un ensemble de données constitué de lentilles gravitationnelles connues. L'objectif sera de tester la capacité de notre modèle à identifier et à classer ces lentilles gravitationnelles dans un contexte réel.

V.2 Comparaisons des scores avec des quads connus

Après avoir effectué l'entraînement de XGBoost sur les quatre catalogues en utilisant deux approches distinctes, à savoir l'entraînement géométrique et l'entraînement en tenant compte des distances, je vais maintenant examiner en détail les performances de prédiction de XGBoost sur les quads déjà connus. Dans un premier temps, je vais d'abord évaluer les performances de XGBoost en ce qui concerne le taux de faux positifs. En effet, une partie essentielle de cette thèse vise à réduire de manière significative ce taux, c'est-à-dire à minimiser le nombre de candidats quads qui, en réalité, ne sont pas quatre images du même quasar. À cette fin, j'ai identifié 28 cas de faux positifs à partir de campagnes d'observations précédentes menées par le groupe GraL sur divers télescopes entre 2018 et 2023 tels que le ESO/NTT (New Technology Telescope) avec l'instrument EFOSC2, le Keck (au W. M. Keck Observatory) avec l'instrument LRIS, le Gemini South avec l'instrument GMOS-S, et le SOAR avec l'instrument GTHS, entre autres. Le groupe de travail Gaia GraL (Gaia Gravitational Lenses) s'est constitué en 2014 et a pour objectif d'identifier les lentilles gravitationnelles dans les données de Gaia, de valider les candidats par des observations au sol et de modéliser les lentilles (paramètres des quasars et des défecteurs). A ce jour, le groupe GraL a confirmé 18 nouveaux quads dont 12 ont été publiés ([Stern et al., 2021](#)) et 66 nouveaux doublets.

Les configurations sélectionnées se composent toutes de quatre images, pour lesquelles nous avons récupéré les coordonnées Gaia EDR3 ainsi que la magnitude Gaia G. Des observations spectroscopiques directes de ces images ont permis de confirmer que 27 des 28 candidats ne correspondaient pas à des quasars lentillés à quatre images, malgré les scores élevés obtenus par les méthodes d'apprentissage utilisées précédemment.

La figure [V.3](#) présente le meilleur des scores de prédiction P obtenus sur les quatre catalogues et les deux approches. Ces résultats sont comparés aux scores obtenus précédemment par un entraînement supervisé utilisant l'algorithme de Extremely Randomized Trees (ERT) de [Delchambre et al. \(2019\)](#). Cette représentation met en évidence les améliorations apportées par la nouvelle méthodologie en termes de taux de faux positifs. En effet, j'ai considérablement réduit la probabilité de fausses détections pour ces 27 configurations. Cela permet d'envisager une meilleure identification et rejet des faux positifs à l'avenir, en utilisant les entraînements avec XGBoost et les simulations réalistes

de lentilles gravitationnelles.

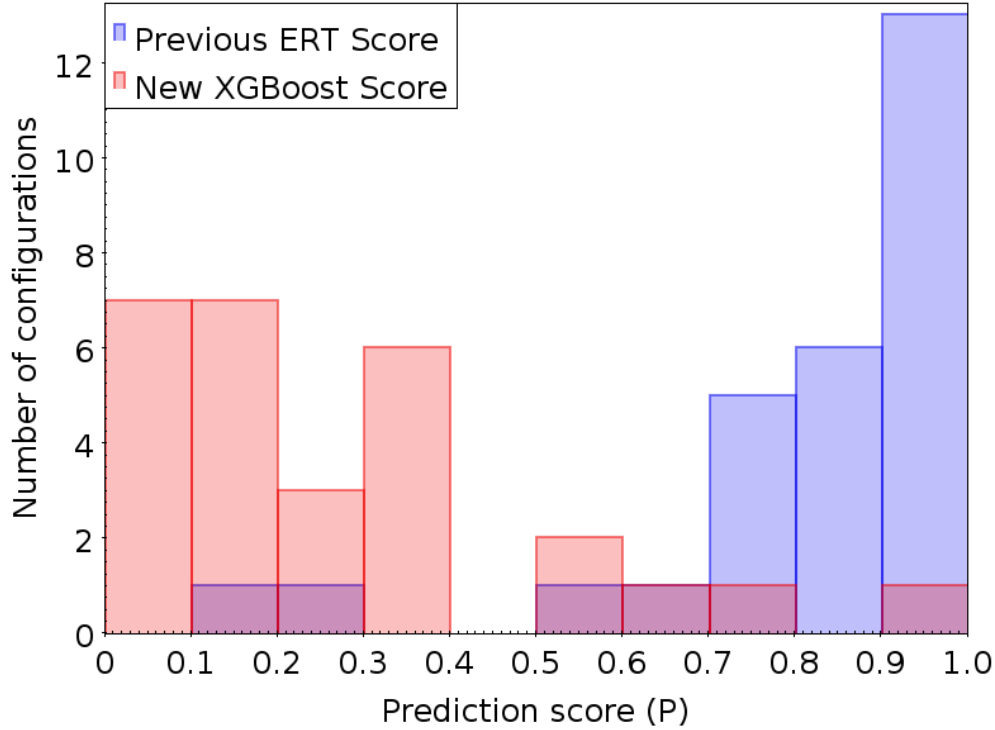


FIGURE V.3 – Histogramme des scores obtenus précédemment par l’algorithme Extremely Randomized Trees (ERT) en bleu, par rapport aux scores obtenus dans notre étude avec XGBoost en rouge, pour un ensemble de 28 configurations identifiées comme des faux positifs. Cette représentation met en évidence une nette amélioration, avec des scores significativement plus faibles pour les faux positifs lors de l’entraînement le plus récent avec XGBoost.

Cet histogramme constitue un résultat très encourageant. Chacune de ces configurations a été préalablement sélectionnée en se basant sur des critères similaires à ceux évoqués dans le chapitre [IV](#). En effet, les images de la configuration doivent avoir des parallaxes et des mouvements propres négligeables, étant donné que les images des quasars lentillés ne présentent pas de mouvement propre. De plus, les images de chaque configuration doivent présenter des couleurs similaires. Ainsi, toutes ces configurations faussement positives ont satisfait à ces critères, malgré le fait de ne pas être des quads authentiques. Ceci renforce l’idée que l’apprentissage automatique est une technique efficace pour détecter de nouveaux quads.

Examinons en détail les quelques cas qui présentent néanmoins un score élevé avec la nouvelle méthode utilisant XGBoost. Parmi les 28 cas examinés dans cette étude, il existe un cas présentant une probabilité moyenne élevée de $P_{\text{xgb}} = 0.96$ représenté dans la figure [V.4](#) (panneau de gauche). Cette configuration, qui contient la source A la plus brillante avec un SourceID Gaia EDR3 6126672995504821120, a été observée en 2019 au moyen de l’instrument GMOS-S du télescope

Gemini South. Malgré les efforts déployés pour son analyse (Stern et al., 2023), aucune conclusion définitive n’a pu être établie à son sujet. Lors de l’évaluation précédente effectuée par ERT, cette configuration avait obtenu un score de $P_{\text{ert}} = 0.88$. Cependant, grâce à la récente contribution du score de prédiction généré par XGBoost, cette configuration se révèle être un bon candidat pour une nouvelle observation visant à confirmer ou à infirmer sa nature.

La deuxième configuration qui obtient un score de prédiction supérieur à $P_{\text{xgb}} > 0.7$ correspond à l’image la plus lumineuse associée au SourceID Gaia EDR3 1321244403970596480, avec un score $P_{\text{xgb}} = 0.78$. Cette configuration est illustrée dans le panneau de droite de la figure V.4 et représente un cas limite où notre méthodologie identifie une lentille gravitationnelle en termes de taille et de configuration géométrique. Cependant, après avoir observé les spectres (private communication of Djorgovski et al.), il semble probable que cette configuration soit en réalité une fusion de galaxies. Afin de visualiser le fait que ces configurations ressemblent à des configurations de lentilles gravitationnelles, nous les avons représentées sur la figure des zones que nous avons discutée dans le chapitre précédent. Nous remarquons que leur géométrie, en particulier l’ordre des images ABCD, correspond aux distributions caractéristiques des simulations de lentilles gravitationnelles à quatre images. Nous constatons que l’image C se situe dans la zone 1 (ou 6), tandis que l’image D correspondante se trouve également dans la zone 1 (ou 6) correspondante. Ainsi, ces deux configurations font partie des cas difficiles à classer.

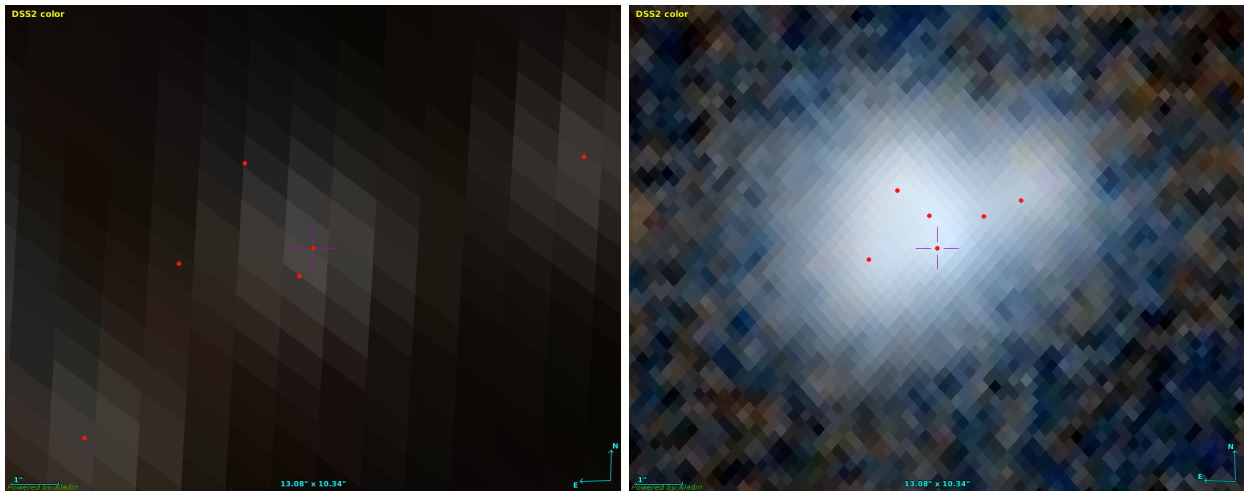


FIGURE V.4 – Représentation des deux configurations qui ont obtenu un score élevé grâce à notre méthode utilisant XGBoost. Le panneau de gauche représente la configuration avec la source la plus brillante Gaia EDR3 6126672995504821120 correspondant à une configuration non concluante pour le moment. Le panneau de droite représente une des configurations limites qui présente une forte ressemblance avec une lentille gravitationnelle sans en être une.

Penchons-nous maintenant sur les configurations de quads qui ont été observées et confirmées. Mon objectif est de vérifier si la méthodologie que j’ai développée permet de détecter correctement des

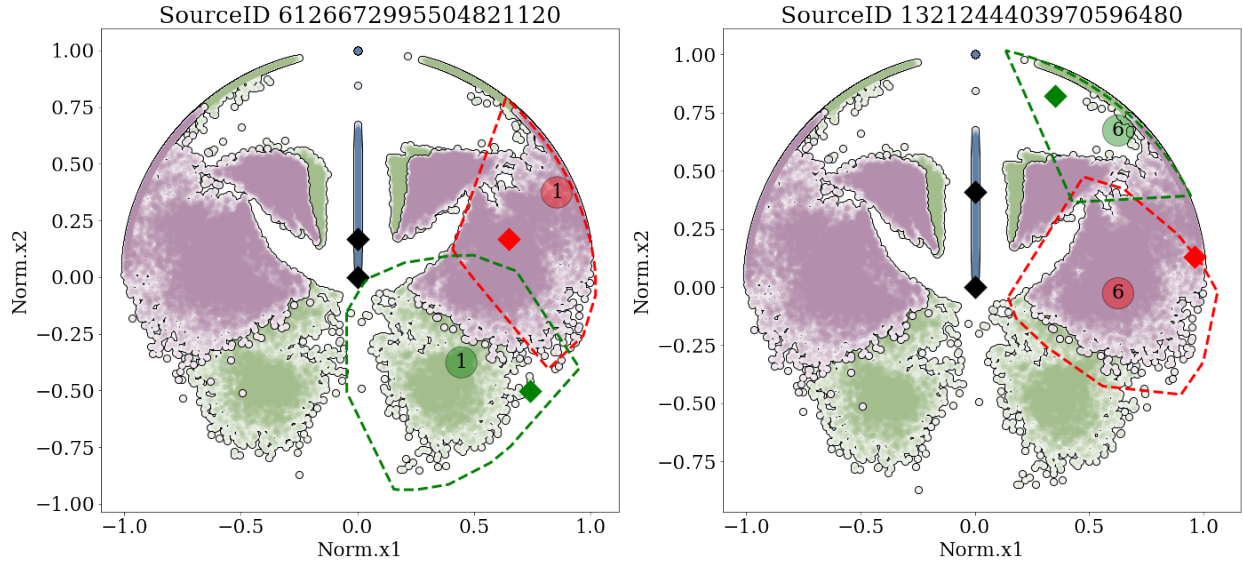


FIGURE V.5 – Position des images ABCD des configurations Gaia EDR3 6126672995504821120 et Gaia EDR3 1321244403970596480 dans le plan normalisé "Norm.x₁" et "Norm.x₂". Les zones dans lesquelles tombent ces images sont identifiées en vert pour les images C et en rouge pour les images D.

quads déjà connus. À cette fin, j'ai sélectionné 20 quads validés, pour lesquels chaque image possède une mesure dans le catalogue de Gaia DR3. La figure [V.6](#) présente le meilleur score de probabilité estimé par XGBoost entre les quatre entraînements géométriques, P_{geo} , en fonction du meilleur score de probabilité entre les quatre entraînements incluant la distance, P_{dist} .

Nous observons quatre cadrants distincts pour les lentilles connues délimités par les lignes sur la figure, que je vais étudier en détail afin de mieux comprendre les résultats obtenus par l'apprentissage de XGBoost.

Le premier cadrant correspond aux configurations qui sont correctement identifiées avec des scores $P_{dist} > 0,8$ et $P_{geo} > 0,8$. Dans l'échantillon, cela représente 12 configurations sur 20, soit 60% des lentilles. XGBoost parvient à bien reconnaître ces configurations. Elles se caractérisent par leur compacité, avec une séparation maximale moyenne de $MaxDist = 2,41$ secondes d'arc et une séparation minimale moyenne de $MinDist = 0,91$ secondes d'arc. Étant donné que ces configurations sont plutôt compactes, elles obtiennent toutes un score $P_{dist} > 0,8$, correspondant aux tailles caractéristiques des lentilles simulées.

Le deuxième cadrant à regarder concerne le cadrant 3, où l'on trouve six configurations qui ne sont pas correctement identifiées par XGBoost, que ce soit avec le score P_{dist} ou le score P_{geo} . Ces six configurations sont illustrées dans la figure [V.7](#). Plusieurs raisons expliquent pourquoi ces configurations ne sont pas bien reconnues.

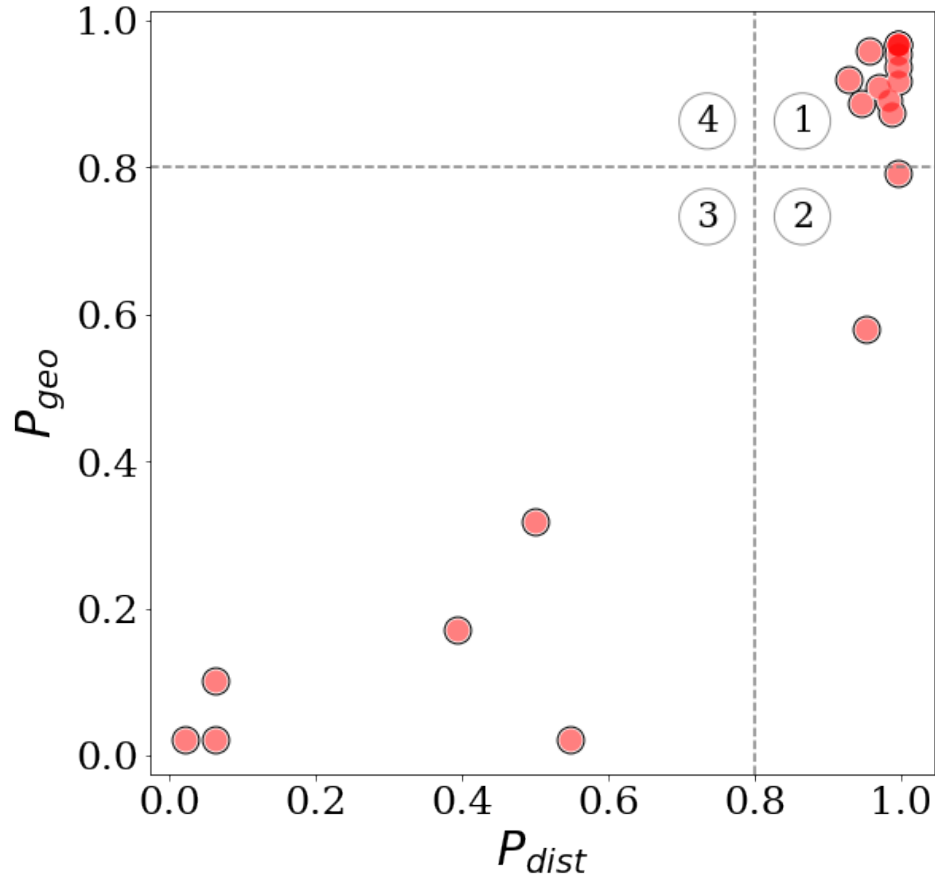


FIGURE V.6 – Meilleur score de probabilité P_{geo} en fonction du meilleur score de probabilité P_{dist} pour les lentilles gravitationnelles connues.

La première raison concerne les configurations de la ligne du haut de la figure. Ces trois configurations ne sont pas compactes et présentent une séparation angulaire maximale respective de 10,07, 5,26 et 3,68 secondes d’arc. De plus, elles sont très allongées, ce qui diffère des configurations les plus courantes obtenues à l’aide du modèle de simulation de lentilles gravitationnelles SIE+*shear*. Dans certains cas, il peut s’agir d’une lentille dont le déflecteur est un amas de galaxies.

En ce qui concerne les deux cas en bas à gauche de la figure, ces configurations présentent une des quatre images décalée par rapport aux configurations habituelles. Ce décalage pourrait être dû à des effets de microlentilles supplémentaires. Ces quelques cas mettent en évidence les limites de la méthodologie de cette thèse qui ne prend pas en compte toute la diversité des configurations. Une perspective d’amélioration consisterait à prendre en compte des effets supplémentaires comme plusieurs profils de masse ou en ne considérant que trois des quatre images.

En conclusion de cette analyse des configurations de quads déjà connus, nous pouvons considérer le

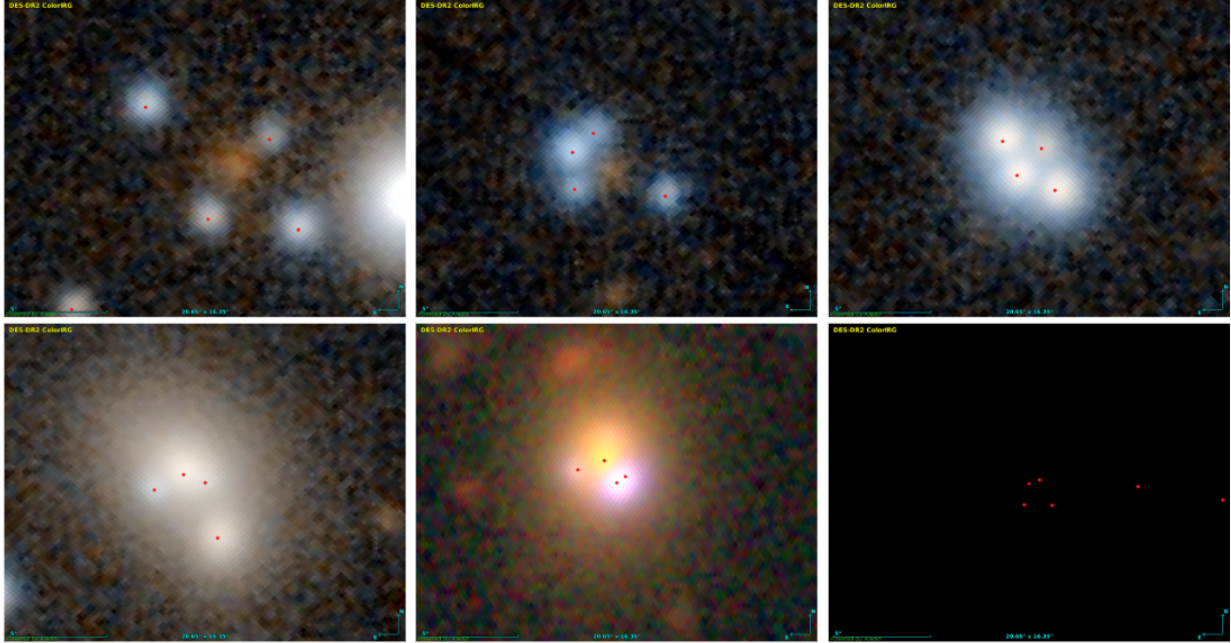


FIGURE V.7 – Images des lentilles gravitationnelles connues qui ont un score P_{dist} et $P_{geo} < 0.4$. Toutes les images sont à la même échelle.

cadrant 1 pour l'identification de candidats quads dans les données de Gaia. Le quadrant 1 permet d'obtenir un indice de confiance élevé. Les cadrants 2 et 4 peuvent être pris en compte, mais avec une confiance moindre, et ils doivent être utilisés avec précaution et comparés à d'autres critères de sélection afin de déterminer si ce sont des configurations probables. Enfin, le quadrant 3 est le quadrant avec la confiance la plus basse. En revanche, cela ne signifie pas qu'il faut rejeter l'ensemble des configurations de ce quadrant puisqu'un score, même faible de l'ordre de $P > 0,6$, peut indiquer un bon candidat.

Chapitre VI

Application aux données de Gaia

Dans les chapitres précédents, j’ai présenté la méthode de simulation de lentilles gravitationnelles de quasar à quatre images à partir d’une distribution réaliste des paramètres. J’ai utilisé ces simulations pour construire plusieurs catalogues d’entraînement pour l’algorithme de *machine learning* XGBoost. Les résultats de XGBoost sur ces catalogues d’entraînement ont été évalués sur des configurations de quads connus et d’anciens candidats invalidés.

Dans ce chapitre, je mettrai en œuvre la méthodologie que j’ai développée tout au long de cette thèse afin d’identifier de nouveaux candidats dans le catalogue Gaia GravLens FPR (Gaia collaboration & Krone-Martins et al. 2023, soumis à A&A). Il s’agit d’un catalogue de sources extraites des observations de Gaia dans 3 millions de champs de 6 secondes d’arc de rayon centrés sur un candidat quasar déjà connu. Dans un premier temps, j’appliquerai des critères de présélection pour identifier les configurations de sources dans le catalogue Gaia GravLens FPR qui sont les plus susceptibles de correspondre à des quads. Ensuite, j’utiliserai les modèles entraînés de XGBoost pour calculer les scores de prédiction pour chacune des configurations. Enfin, je procéderai à une inspection visuelle des candidats les plus prometteurs, ce qui permettra d’établir une liste de candidats en vue d’observations futures.

VI.1 Identification de candidats lentilles dans le catalogue Gaia GravLens FPR

Les publications actuelles des données Gaia demeurent incomplètes pour les séparations angulaires les plus petites (Arenou et al., 2017; Fabricius et al., 2021; Torra et al., 2021). Par conséquent, certaines images de lentilles gravitationnelles connues n’ont pas été publiées dans Gaia DR3 (Ducourant et al., 2018). Cette situation a ralenti l’identification de nouvelles lentilles, car la plupart des lentilles gravitationnelles de quasars encore non découvertes sont caractérisées par de faibles séparations angulaires. Pour remédier à cela, le Gaia Data Processing and Analysis Consortium (Gaia DPAC) a

mis au point une chaîne de traitement spécifique visant à analyser l'environnement des quasars et à produire un catalogue de sources plus complet aux faibles séparations. Ce travail fait l'objet d'une publication "Focused Product Release" (FPR). L'objectif de cette FPR est de détecter les quasars lentillés en analysant les détections Gaia à proximité de ces quasars, afin de générer une liste de sources secondaires qui complètent les publications actuelles des données Gaia.

La chaîne d'analyse mise en place dans le cadre de cette FPR est appelée GravLens. Ce pipeline utilise un algorithme spécifique appelé DBSCAN (Density-Based Spatial Clustering of Applications with Noise) pour détecter les mesures de Gaia autour des quasars à moins de 6 secondes d'arc. DBSCAN est un algorithme de clustering non supervisé largement utilisé dans les domaines de l'apprentissage automatique et de l'analyse de données. Son principe repose sur la construction d'un graphe de voisinage reliant des points (représentant ici des détections individuelles en ascension droite et en déclinaison) dont la distance est inférieure à 100 mas. Ainsi, DBSCAN regroupe automatiquement les détections individuelles de Gaia, permettant ainsi d'identifier de nouvelles sources qui n'avaient pas été publiées dans les catalogues Gaia précédents. Ce nouveau catalogue de multiplets, regroupant les quasars et les nouvelles sources identifiées autour d'eux, est ensuite utilisé pour rechercher de nouveaux candidats de lentilles gravitationnelles. Des informations plus détaillées sont fournies dans l'article [Gaia Collaboration et al. \(2023\)](#).

VI.1.1 Préselection des configurations à analyser dans Gaia

Pour la recherche de nouvelles lentilles gravitationnelles de quasars, j'ai donc utilisé le catalogue de multiplets issu du FPR Gaia, nommé *GravLens*. La liste initiale des quasars comprend 3 760 480 sources présentes dans Gaia DR3. Ce catalogue est enrichi de 103 000 nouvelles sources identifiées par GravLens qui ne sont pas présentes dans Gaia DR3. Dans GravLens, 87% des quasars ne présentent aucun voisin dans un rayon de 6 secondes d'arc, et seulement 4% de ces multiplets sont composés de plus de deux composants, soit environ 159 000 configurations. Pour la recherche spécifique de quasars lentillés à quatre images, je me suis concentré uniquement sur un sous-échantillon ne comprenant que les multiplets comportant exactement quatre sources identifiées. Ce sous-échantillon représente un total de 17 610 multiplets.

L'objectif final de cette thèse est de constituer une liste de candidats aussi exhaustive que possible afin de maximiser nos chances de détecter de nouvelles lentilles gravitationnelles. À cette fin, je n'ai procédé à aucun filtrage ou nettoyage des données basé sur des critères tels que les mouvements propres, les parallaxes, les redshifts ou les couleurs. En revanche, j'ai appliqué une contrainte importante concernant la position des images. Par exemple, dans les configurations de lentilles gravitationnelle de quasars à quatre images, il n'existe aucune configuration où une des images est contenu dans un triangle formé par les trois autres images.

Pour détecter ces configurations, j'ai utilisé un algorithme qui calcule l'aire du triangle formé par

trois des images de la configuration. Ensuite, j'ai évalué l'aire des trois sous-triangles formés par le quatrième point à tester et les trois autres points du triangle initial. Si l'aire du triangle complet est égale à la somme des aires des sous-triangles, cela indique que le point à tester se trouve à l'intérieur du triangle complet. Parmi les 17 610 multipléts, j'ai constaté que 3 504 d'entre eux ont une image située à l'intérieur d'un triangle formé par les trois autres. Cette contrainte permet donc d'éliminer 20% des multipléts qui ne satisfont pas cette condition.

Ces multipléts à quatre composantes où l'une des composantes se trouve à l'intérieur d'un triangle formé par les trois autres peut correspondre à des configurations dont l'une des images est la galaxie déflectrice qui est détectée. Les entraînements que j'ai réalisé ne prennent pas en compte ces configurations et sont donc rejetées de mon analyse. En revanche, la prise en compte de ces configurations, ainsi que des configurations avec plus de quatre sources, représente une amélioration potentielle de la méthodologie que j'ai développée. Cependant, cette extension n'a pas été explorée dans le cadre de cette thèse.

VI.1.2 Identification d'un nouveau critère de sélection

L'étude des configurations de lentilles gravitationnelles à quatre images a permis d'identifier des zones bien distinctes sur la représentation normalisées $\text{Norm}.x_1$ - $\text{Norm}.x_2$ de la figure [IV.12](#). Ces zones peuvent servir à établir un nouveau critère supplémentaire qui n'a pas encore été exploré pour affiner la recherche de lentilles gravitationnelles de quasars. Pour ce faire, j'ai développé une procédure qui effectue des tests sur chaque configuration afin de vérifier si l'image C et l'image D se trouvent dans des zones complémentaires.

Pour cela, j'ai d'abord appliqué un algorithme dit de la marche de Jarvis sur chaque zone identifiée. Cet algorithme permet de calculer l'enveloppe convexe d'un ensemble de points. Le principe est le suivant : je commence par sélectionner le point ayant la plus petite coordonnée y (et la plus petite coordonnée x en cas d'égalité) comme point de départ de l'enveloppe convexe. Ce point sera appelé le point pivot. Puis, je recherche le point suivant parmi tous les autres points pour trouver celui qui forme l'angle le plus petit avec le segment reliant le point pivot au point courant. Cela implique généralement le calcul de l'angle de chaque point par rapport au point pivot. Une fois que le point suivant a été identifié, il devient le nouveau point pivot, et le processus est répété jusqu'à ce que le point pivot initial soit atteint. Cela garantit que l'enveloppe convexe est fermée. J'ai appliqué cet algorithme à l'ensemble des six zones de la figure [IV.12](#).

Afin de prendre en compte un plus grand nombre de cas que ceux présents dans les simulations de quads, j'ai agrandi de manière arbitraire de 10% la taille de ces zones. Pour cela, j'ai calculé le centroïde de chaque contour et j'ai appliqué un facteur d'échelle. Cela agrandit ou réduit la distance entre chaque point et le centroïde selon le facteur d'échelle, tout en maintenant la même relation de position relative entre les points.

Enfin, j'ai appliqué un algorithme qui commence par vérifier si l'image C se situe dans l'une des six zones définies pour les images C dans les simulations de lentilles. Ensuite, l'algorithme examine si l'image D se trouve également dans la zone correspondante pour les images D. Cette condition de correspondance entre les zones constitue une contrainte importante sur les configurations, car seules 1 908 des 14 106 configurations de multiplets à quatre images (soit 13%) satisfont à cette condition. Ce test, que j'ai appelé "test in_contour", attribue la valeur 0 à la configuration si les images C et D ne se trouvent pas dans des zones correspondantes, et la valeur 1 si les images C et D se situent dans des zones complémentaires.

VI.1.3 Recherche de nouvelles lentilles gravitationnelles de quasars

J'ai appliqué les huit modèles entraînés avec XGBoost au chapitre précédent aux données GravLens, ce qui donne un total de huit scores. Quatre scores sont basés uniquement sur les paramètres géométriques ($P_{G1}, P_{G2}, P_{G3}, P_{G4}$), tandis que les quatre autres scores incluent également l'information sur les distances ($P_{D1}, P_{D2}, P_{D3}, P_{D4}$). En plus de ces probabilités individuelles, j'ai calculé les moyennes suivantes : μ_G correspondant à la moyenne des probabilités géométriques, μ_D correspondant à la moyenne des probabilités avec les distances, et μ_{all} correspondant à la moyenne de toutes les probabilités, y compris le score du test "in_contour". Pour évaluer les écarts entre les différentes probabilités, j'ai calculé l'écart maximum parmi les huit scores. Cette valeur permet de repérer rapidement les configurations qui ont un score élevé ($P > 0,9$) pour un modèle d'apprentissage et un score bas ($P < 0,1$) pour un autre modèle. Ainsi, je suis en mesure de caractériser les multiplets de GravLens en utilisant un total de 13 critères. C'est l'analyse de ces 13 critères qui permet de sélectionner les meilleurs candidats parmi les multiplets.

La figure VI.1 présente les résultats des meilleures probabilités P_{geo} et P_{dist} parmi les quatre scores géométriques et les quatre scores incluant la distance, respectivement. Cette visualisation est similaire à celle utilisée dans le chapitre précédent (voir figure V.6). Seules les configurations qui satisfont le critère "in_contour" sont représentées ici, soit un total de 1 893 multiplets. La répartition des multiplets en fonction des cadrants 1, 2, 3 et 4 est récapitulée dans le tableau VI.1.

TABLE VI.1 – Répartition des multiplets en fonction des cadrants 1, 2, 3 et 4 de la figure VI.1.

	Cadrant 1	Cadrant 2	Cadrant 3	Cadrant 4
in contour	67	19	1 708	99
in and out contour	82	46	13 806	154

En sélectionnant les cadrants 1, 2 et 4 qui satisfont au critère "in_contour", j'ai identifié un total de 185 candidats potentiels, ce qui est un nombre important et très encourageant. Afin de classer ces candidats, je vais m'appuyer sur l'ensemble des critères que j'ai établis. J'ai également réalisé une évaluation visuelle pour confirmer le classement final des candidats et éviter toute configuration

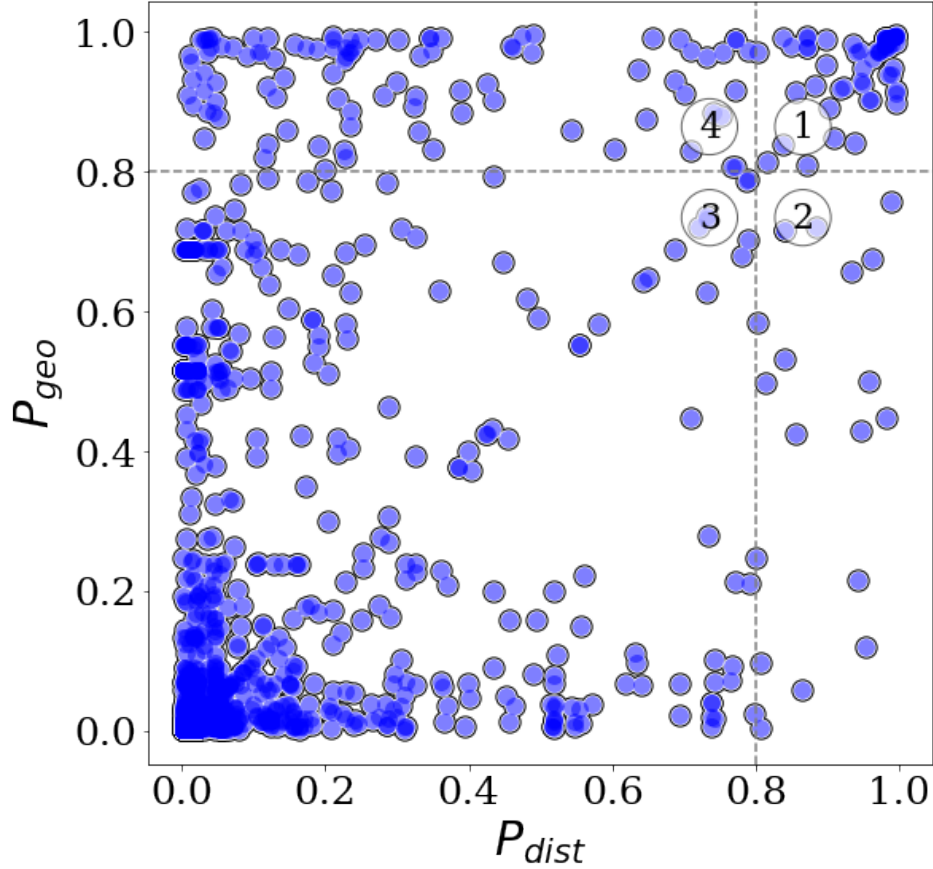


FIGURE VI.1 – Meilleur score de probabilité géométrique P_{geo} en fonction du meilleur score parmi les probabilités incluant la distance P_{dist} pour les 1 893 multiplets sélectionnés de GravLens.

probable mais dont l'environnement, par exemple, indiquerait peu de chances qu'il s'agisse d'une lentille.

VI.2 Liste de candidats

Dans cette section, je procède à l'analyse de la liste des 67 candidats les plus prometteurs obtenus à partir de l'apprentissage de XGBoost. Les 13 critères que j'ai établi pour chaque configuration sont utilisés pour classer les candidats. Une première inspection visuelle a permis d'éliminer 8 candidats supplémentaires puisque une ou plusieurs composantes sont en fait associés à des étoiles.

Je présente dans la table [VI.2](#) la liste des 59 candidats restants les plus prometteurs. Les paramètres suivants sont inclus dans cette liste : "MaxDist", qui correspond à la distance maximale entre les images, et "MinDist", qui correspond à la distance minimale entre les images. Les magnitudes minimale et maximale des images mesurées dans la bande G de Gaia sont notées Mag_A et Mag_D

respectivement. De plus, le score moyen de toutes les probabilités, noté μ_{all} , est également fourni pour chaque candidat.

Ces candidats ont été classés visuellement de A à C, où A correspond aux candidats les plus prometteurs. La notation C est attribuée aux candidats présents dans des régions denses en étoiles, comme indiqué dans la figure VI.2. La figure VI.3 présente six des meilleurs candidats ayant reçu un score visuel A. Ces représentations affichent les images provenant du DESI Legacy Imaging Surveys (Collaboration, 2019) ou de Pan-STARRS1 (Chambers et al., 2016), associées à chaque candidat lentille.

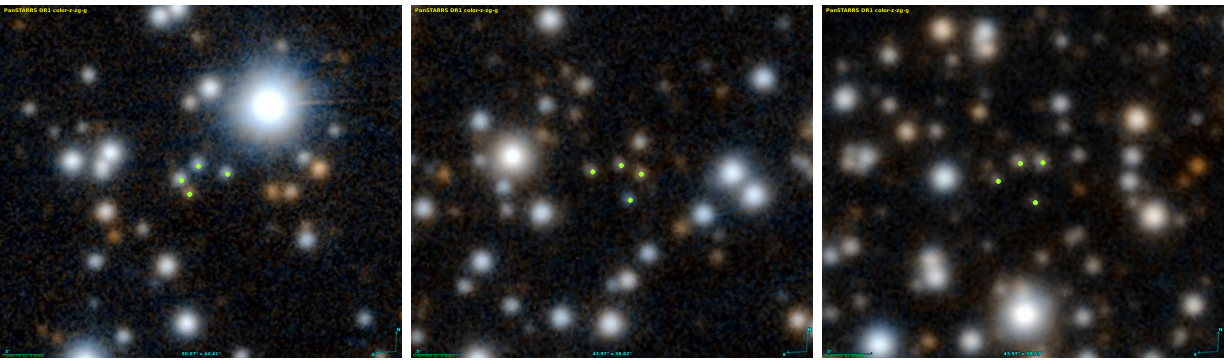


FIGURE VI.2 – Images des candidats lentilles dans les régions denses en étoiles. Les points verts correspondent aux coordonnées de chaque image de la configuration dans le catalogue GravLens.

Les résultats obtenus concernent les configurations du quadrant 1 de la figure VI.1. Cependant, nous avons de bonnes raisons de penser que les candidats du quadrant 4 sont également des candidats compacts prometteurs. De plus, il est très probable que de nombreux candidats des quadrants 2 et 3, ayant un score supérieur à $P > 0,6$, puissent également être des lentilles gravitationnelles de quasars. Ces résultats doivent donc être comparés avec d'autres scores et paramètres pour déterminer si ces candidats sont suffisamment robustes pour être sélectionnés pour les campagnes d'observations destinées à confirmer ou infirmer leur nature de lentilles gravitationnelles de quasars. Ces futures observations seront cruciales pour valider définitivement ces candidats.

TABLE VI.2 – Liste des 59 candidats quads prometteurs identifiés dans le catalogue GravLens.

#	MaxDist (")	MinDist (")	Mag_A	Mag_D	μ_{all}	Visuel
EOLENS 933	7,61	3,11	19,39	20,29	0,56	A
EOLENS 1093	5,92	0,93	19,70	20,56	0,55	C
EOLENS 1337	5,85	2,04	19,46	20,56	0,47	C
EOLENS 1372	5,83	1,58	20,11	20,62	0,71	B
EOLENS 1475	6,63	2,65	20,30	20,73	0,66	B
EOLENS 1918	5,32	2,40	20,31	20,64	0,45	C
EOLENS 1971	5,21	0,18	19,63	20,74	0,74	B
EOLENS 2950	5,29	2,46	19,69	20,63	0,70	C
EOLENS 3054	5,90	3,42	19,93	20,37	0,61	B
EOLENS 3294	3,42	1,35	19,01	20,16	0,80	A
EOLENS 3351	7,01	1,67	19,13	20,37	0,50	B
EOLENS 3678	6,65	0,80	19,44	19,96	0,72	A
EOLENS 3713	7,12	3,22	19,20	20,27	0,62	C
EOLENS 4090	6,50	2,39	19,21	20,54	0,55	B
EOLENS 4192	6,72	0,78	18,95	20,62	0,48	C
EOLENS 4296	7,99	4,00	20,06	20,73	0,75	B
EOLENS 4626	6,59	3,19	20,13	20,60	0,75	A
EOLENS 4715	6,81	3,02	17,77	19,77	0,75	C
EOLENS 4728	6,63	3,01	19,54	20,55	0,66	A
EOLENS 4767	5,96	1,73	18,87	20,44	0,77	A
EOLENS 5387	6,86	0,50	18,39	20,19	0,55	C
EOLENS 5685	6,73	3,29	19,73	20,42	0,60	B
EOLENS 5936	5,19	1,36	18,40	20,33	0,71	A
EOLENS 6234	5,90	0,50	18,47	20,06	0,77	C
EOLENS 6517	8,24	4,43	19,34	20,53	0,74	C
EOLENS 6604	7,01	2,94	19,65	20,67	0,72	A
EOLENS 6814	5,95	2,42	20,39	20,67	0,66	A
EOLENS 6894	5,68	2,21	19,60	20,48	0,68	A
EOLENS 7076	8,27	4,43	20,31	20,64	0,71	B
EOLENS 7215	7,16	2,63	18,75	19,89	0,59	-
EOLENS 7704	5,96	2,34	19,86	20,65	0,76	-

#	MaxDist (")	MinDist (")	Mag_A	Mag_D	μ_{all}	Visuel
EOLENS 7948	4,62	2,47	19,76	20,81	0,45	C
EOLENS 7995	7,70	3,25	19,94	20,45	0,75	B
EOLENS 8184	6,17	3,03	19,24	20,24	0,57	-
EOLENS 8483	5,44	2,41	19,30	20,44	0,60	-
EOLENS 8539	7,08	3,60	19,37	20,76	0,69	-
EOLENS 8691	7,08	1,12	19,53	20,56	0,55	-
EOLENS 9043	6,97	3,41	19,14	20,06	0,39	-
EOLENS 9123	4,68	1,71	19,20	20,49	0,53	C
EOLENS 9815	5,37	1,61	20,01	20,69	0,76	B
EOLENS 10098	6,98	0,84	18,46	20,61	0,52	C
EOLENS 10349	7,13	3,28	20,32	20,39	0,77	-
EOLENS 10416	5,17	2,77	19,40	20,37	0,71	-
EOLENS 10606	5,63	2,53	20,04	20,40	0,72	B
EOLENS 11104	4,19	1,55	19,89	20,46	0,76	-
EOLENS 11377	6,82	2,67	20,44	20,66	0,77	-
EOLENS 11476	5,31	2,85	19,56	20,43	0,66	C
EOLENS 11482	3,47	1,43	19,85	20,64	0,84	B
EOLENS 11997	8,50	3,90	19,82	20,76	0,62	B
EOLENS 12763	6,57	1,64	17,95	20,66	0,52	-
EOLENS 12915	6,34	0,96	18,61	20,77	0,72	-
EOLENS 13030	7,35	2,13	20,09	20,66	0,65	-
EOLENS 13311	5,81	0,77	19,44	20,66	0,76	-
EOLENS 13822	7,15	1,28	18,99	20,09	0,56	-
EOLENS 14158	3,12	0,99	19,90	20,59	0,72	-
EOLENS 14522	5,55	0,61	19,16	20,49	0,69	-
EOLENS 15612	6,53	0,76	19,50	20,12	0,77	-
EOLENS 15736	6,10	1,42	19,26	20,46	0,68	-
EOLENS 17157	0,52	0,31	19,49	20,37	0,59	A

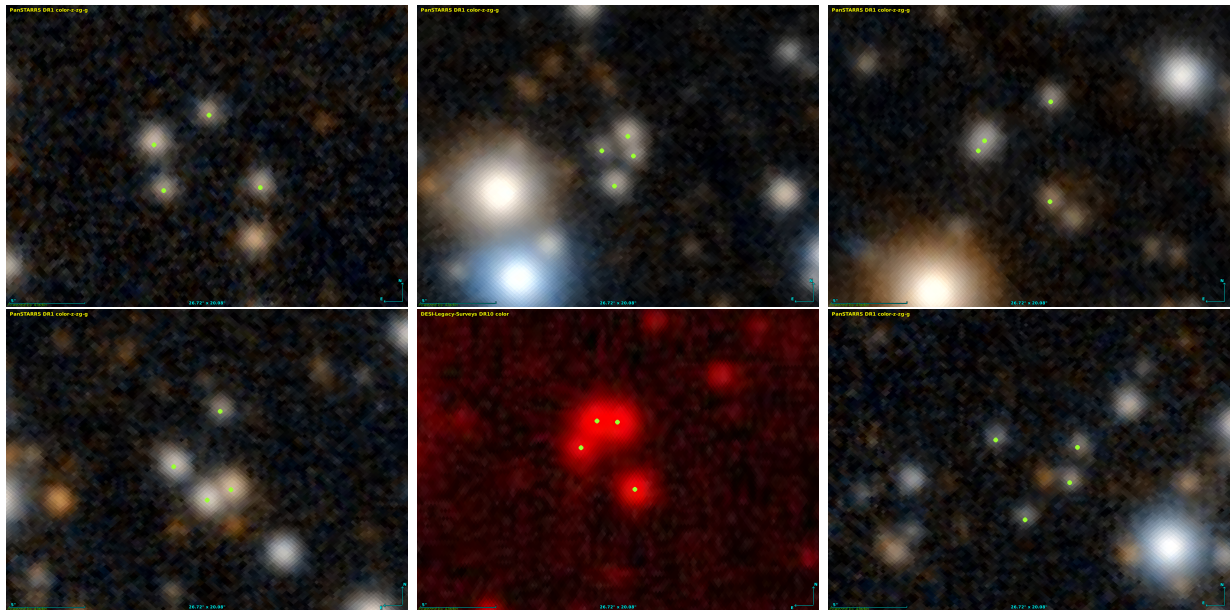


FIGURE VI.3 – Exemple six candidats lentilles avec un score visuel A ou A+. Les points verts correspondent aux coordonnées de chaque image de la configuration dans le catalogue GravLens. De gauche à droite et de haut en bas on a les configurations GravLens #933, #3294, #3678, #4767, #5936 et #6814.

Bilan et perspectives

Les résultats présentés dans ce manuscrit mettent en évidence les opportunités de découvrir de nouvelles lentilles gravitationnelles de quasars à quatre images (quads) dans les données de Gaia. Au fil des différentes sections, j'ai décrit en détail la méthode que j'ai développée pour améliorer les simulations des quads en utilisant des distributions réalistes de paramètres. J'ai comparé les résultats de l'apprentissage de l'algorithme XGBoost, basé sur le *gradient tree boosting*, avec les découvertes récentes de quads pour démontrer l'efficacité de cette approche. Dans cette dernière section, je propose de faire le bilan des contributions de cette méthode pour améliorer la recherche de lentilles gravitationnelles à l'aveugle grâce à l'apprentissage automatique. A ce titre, je soulignerai également les améliorations potentielles à apporter à différents aspects de la méthode afin de renforcer la découverte de nouvelles lentilles gravitationnelles.

Au cours de cette thèse, j'ai développé une méthode complète d'analyse de la morphologie, de l'alignement et de la distribution de masse des halos des simulations cosmologiques EAGLE. Cette méthode est une étape essentielle, car elle permet d'obtenir une mesure précise de la forme et de la distribution de masse des halos qui agissent comme des défecteurs dans les simulations de lentilles gravitationnelles. L'utilisation des halos de matière noire et de matière baryonique dans les simulations EAGLE garantit que les défecteurs possèdent une distribution réaliste des paramètres de forme et de masse. La méthode d'analyse développée pour EAGLE est conçue pour fonctionner sur l'ensemble des halos de matière noire et de matière baryonique, quels que soit leur taille ou leur forme dans les simulations. Pour adapter la mesure des paramètres morphologiques à chaque halo, j'ai développé une méthode de calcul d'une sphère de travail (« *working sphere* »). Cette sphère de travail correspond au rayon auquel la densité par rapport au centre du halo a diminué de 80%. Ce choix permet d'exclure les grandes structures de faible densité à la périphérie des halos. J'ai ensuite utilisé l'analyse en composante principale (ACP) comme outil pour ajuster un ellipsoïde en trois dimensions sur les halos des simulations. Enfin, j'ai construit pour l'ensemble des halos une courbe de croissance en masse permettant ainsi de mesurer les rayons correspondant à 20%, 50% et 80% de la masse des halos. Les résultats de cette étude ont fait l'objet d'une publication scientifique (Petit

et al., 2023) Ce travail a également abouti à la publication d'un package Python appelé MOGPY, regroupant toutes les fonctions développées dans le cadre de l'analyse de EAGLE.

Une fois ces résultats obtenus, j'ai réalisé des simulations réalistes de lentilles gravitationnelles de quasars. Pour cela, j'ai utilisé les halos provenant de la simulation EAGLE comme déflecteurs, en exploitant les informations sur leur forme et leur masse. Pour les sources d'arrière-plan, j'ai utilisé une liste de quasars issus du Million Quasars (MILLIQUAS) Catalog, Version 7.10 (15 April 2023) (Flesch, 2021) qui recense plus de 845 286 QSO et AGN de type I. En plus des simulations de lentilles gravitationnelles de quasars à quatre images, j'ai mis en évidence plusieurs espaces de représentation où le classement des lentilles est facilité. Ces espaces comprennent l'espace des configurations alignées et normalisées (Norm.x1 et Norm.x2) ainsi que les angles entre les différentes images. Ces sous-espaces présentent des structures distinctes qui regroupent les configurations de lentilles gravitationnelles. Ainsi, le calcul de ces paramètres facilite le classement grâce à l'algorithme XGBoost.

J'ai utilisé ces simulations de lentilles gravitationnelles de quasars pour créer quatre catalogues d'entraînement. Chaque catalogue diffère par la présence ou l'absence de *shear* externe ainsi que la présence ou l'absence de halos massifs supplémentaires. Cette approche comparative a permis d'obtenir plusieurs scores de prédiction, un pour chaque catalogue. En complément de ces quatre catalogues, j'ai également réalisé deux types d'entraînement par catalogue. Le premier type d'entraînement prend en compte les paramètres géométriques tels que les angles et la position des images. Le second type d'entraînement inclut non seulement les paramètres géométriques, mais aussi les paramètres de distance. Cette approche à deux entraînements permet d'obtenir deux scores distincts, l'un se concentrant sur la morphologie des configurations uniquement, l'autre tenant compte de la taille des configurations. Au total, j'ai ainsi obtenu huit modèles entraînés avec XGBoost.

En combinant les nouvelles distributions de paramètres issus des simulations de EAGLE avec l'algorithme d'apprentissage XGBoost, j'ai apporté des améliorations significatives dans la détection de lentilles gravitationnelles de quasar à quatre images (quads) à partir des données de Gaia. Les résultats obtenus démontrent une nette amélioration en comparaison de l'estimation précédente de Delchambre et al. (2019), notamment en ce qui concerne les faux positifs. Lors de l'évaluation d'un échantillon de 28 cas faux positifs issus de précédentes campagnes d'observations menées par le groupe GraL, les modèles entraînés avec XGBoost ont considérablement réduit les scores des faux positifs. De plus, parmi un échantillon de 20 quads confirmés, 16 ont été correctement identifiés avec un score supérieur à 0,8, ce qui confirme l'efficacité des modèles développés.

En exploitant les modèles entraînés sur les données d'EOLENS, le catalogue de multiplets issu de la chaîne d'analyse GravLens du Gaia Data Processing and Analysis Consortium (Gaia DPAC), j'ai établi une liste très prometteuse de plusieurs dizaines de candidats de quasars lentillés à quatre images. Cette liste de candidats jouera un rôle clé lors des futures campagnes d'observation, visant

à confirmer ou infirmer la nature de ces configurations. La prochaine campagne d'observation est prévue à l'Observatoire de la Silla sur ESO/NTT (New Technology Telescope) en août 2023, utilisant l'instrument EFOSC2.

Améliorations possible de la méthode

Les simulations réalistes de lentilles gravitationnelles de quasars que j'ai réalisées présentent certaines limites. Elles ne capturent pas toute la diversité des configurations possibles. En premier lieu il faut améliorer l'échantillon d'apprentissage en améliorant les simulations elle-mêmes. Tout d'abord, il faut accroître la complexité du déflecteur en considérant également des distributions de masse moins axi-symétriques et en prenant en compte les déflecteurs secondaires associés à l'environnement proche des quasars. Pour cela, il est possible d'utiliser plusieurs profils de masses pour modéliser le déflecteur. Cette modélisation modifie les positions relatives et la luminosité des images du quasar. Il s'agit également d'inclure l'effet de microlentilles car cela impacte également le phénomène, allant jusqu'à inverser les rapports de luminosité, ce qui met en défaut notre apprentissage actuel.

Ensuite on peut améliorer la sélection des candidats en étendant l'analyse de XGBOOST aux multiplets n'ayant pas exactement quatre images. Considérons d'abord le cas de trois sources dans le multiplet. En effet dans beaucoup de cas, Gaia FPR ne parvient à détecter que trois des images des lentilles. Pour que XGBOOST parvienne à identifier dans les triplets les configurations de lentilles, il faudrait produire des simulations de lentilles et retirer aléatoirement une des images. Généralement c'est l'image la moins brillante qui n'est pas détectées mais parfois les effets de micro-lentilles vont inverser l'ordre d'amplification des images. Il s'agirait donc d'inclure dans notre échantillon d'apprentissage toutes les combinaisons de trois images des simulations de quads. On remarquera également que dans le Gaia FRP les triplet sont extrêmement nombreux et constituent donc un réservoir de lentilles potentielles très important, bien plus que les quadruplets. De la même façon, une autre perspective d'amélioration consiste à étendre l'analyse de XGBOOST aux multiplets de plus de quatre sources dans le champs de vue. En effet, dans certains cas de lentilles connues, les images du quasar se trouvent dans des multiplets à 5, 6 voir 7 sources. Il s'agirait dans ce cas d'analyser toutes les combinaisons de trois sources ou de quatre sources dans ces multiplets et donc d'estimer la probabilité de chacune des configurations.

Le taux de faux positifs est un problème majeur qui se pose lors de la confirmation spectroscopique au sol des candidats. Le temps de télescope est difficile à obtenir aussi il est important que la liste des meilleurs candidats soit aussi pure que possible. Pour réduire le taux de faux positifs on doit inclure des informations photométriques supplémentaires (couleurs Gaia, WISE, DES) dans les paramètres de classification pour mieux détecter le caractère extra-galactique des sources associées aux lentilles. Les quasars sont généralement assez bleus et les diverses images du quasar ont des couleurs assez similaires. Ce double filtre permettrait d'éliminer des multiplets non probables généralement constitués d'étoiles. Pour gagner en robustesse dans l'étape de classification, il s'agirait

en parallèle d'identifier et retirer les paramètres les moins discriminants. Il faudrait également adapter l'algorithme afin qu'il puisse traiter les cas de données manquantes (les sources faibles n'ayant pas toujours d'information photométrique). Un second filtre intéressant à installer serait de tester le nombre de détections de Gaia pour chacune des sources. En effet, nous travaillons sur le catalogue Gaia FRP qui a recherché dans les données Gaia des sources non-publiées dans la Gaia DR3. Certaines des sources que nous avons utilisées se trouvent à la limite de la détection du satellite et bénéficient de très peu de mesures impliquant que leur astrométrie est très peu précise. Dans ce cas on pourrait imaginer d'exclure des multiplets, avant analyse, les sources avec peu de détections. Il faudrait également tenir compte de la latitude galactique des sources. En effet on ne s'attend pas à avoir beaucoup de lentille dans des zones proches du plan Galactique. Ce filtre devrait lui aussi permettre de réduire le taux de faux positifs afin d'obtenir une liste de candidats.

En prenant en compte ces perspectives d'amélioration, nous pourrions mieux caractériser et exploiter les configurations de lentilles gravitationnelles de quasars, renforçant ainsi la précision et la généralisation de notre modèle.

Annexes

Article publié

Shape, alignment, and mass distribution of baryonic and dark-matter halos in one EAGLE simulation

Q. Petit, C. Ducourant, E. Slezak, D. Sluse, and L. Delchambre. Shape, alignment, and mass distribution of baryonic and dark-matter halos in one EAGLE simulation. , 669 :A132, January 2023. doi : 10.1051/0004-6361/202244920. 19 pages. <https://arxiv.org/pdf/2212.08880.pdf>

Abstract :

Accurate knowledge of the morphology of halos and its evolution are key constraints on the galaxy formation model as well as a determinant parameter of the strong-lensing phenomenon. Large-scale cosmic simulations are a tailored tool used to obtain statistics on the shape and mass distributions of these halos according to redshift. Using the cosmological hydrodynamic simulation, the Evolution and Assembly of GaLaxies and their Environments (EAGLE), we aim to provide a comprehensive analysis of the evolution of the morphology of galaxy halos and of their mass distributions with a focus on the snapshot at redshift $z = 0.5$. We developed an iterative strategy involving a principal component analysis (PCA) to investigate the properties of the EAGLE halos and the differences in alignment between the various components. The semi-axes and orientation of the halos are estimated taking into account sub-halos in the simulation. The mass distributions of the dark-matter (DM), gas, and star halos are characterised by a half-mass radius, a concentration parameter and (projected) axis ratios. We present statistics of the shape parameters of 336 540 halos from the EAGLE RefL0025N0376 simulation and describe their evolution from redshift $z = 15$ to $z = 0$. We measured the three-dimensional shape parameters for the DM, the gas, and the star components as well as for all particles. We also measured these parameters for two-dimensional projected distributions. At $z = 0.5$, the minor axis of gas aligns with the minor axis of DM for massive halos ($M > 10^{12} M_{\odot}$), but this alignment is poorer for less massive halos. The DM halos axis ratios b/a and c/a have median values of 0.82 ± 0.11 and 0.64 ± 0.12 , respectively. The gas in halos that also contain stars has

a more flattened shape, with $b/a = 0.70 \pm 0.19$ and $c/a = 0.38 \pm 0.20$. The sphericity of gas in halos w/ and w/o stars appears to be negatively correlated to the total mass, while the sphericity of DM is insensitive to it. The measured projected axis ratios, b_p/a_p , of star halos at $z = 0.5$ have a median value of 0.80 ± 0.07 , which is in good agreement with ground-based and space-based measurements within 1σ . For DM halos, we measure a value of 0.85 ± 0.06 . The evolution of the concentration as a function of redshift is fairly homogeneous for the various components, except for the starless gas halos, which appear much more concentrated for $z > 0.7$.

Shape, alignment, and mass distribution of baryonic and dark-matter halos in one EAGLE simulation

Q. Petit¹, C. Ducourant¹, E. Slezak², D. Sluse³, and L. Delchambre³

¹ Laboratoire d'Astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, Allée Geoffroy Saint-Hilaire, 33615 Pessac, France
e-mail: quentin.petit.1@u-bordeaux.fr

² Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Boulevard de l'Observatoire, CS 34229, 06304 Nice, France

³ Institut d'Astrophysique et de Géophysique, Université de Liège, 19c, Allée du 6 Août, 4000 Liège, Belgium
e-mail: quentin.petit.1@u-bordeaux.fr

Received 8 September 2022 / Accepted 5 December 2022

ABSTRACT

Context. Accurate knowledge of the morphology of halos and its evolution are key constraints on the galaxy formation model as well as a determinant parameter of the strong-lensing phenomenon. Large-scale cosmic simulations are a tailored tool used to obtain statistics on the shape and mass distributions of these halos according to redshift.

Aims. Using the cosmological hydrodynamic simulation, the Evolution and Assembly of GaLaxies and their Environments (EAGLE), we aim to provide a comprehensive analysis of the evolution of the morphology of galaxy halos and of their mass distributions with a focus on the snapshot at redshift $z = 0.5$.

Methods. We developed an iterative strategy involving a principal component analysis (PCA) to investigate the properties of the EAGLE halos and the differences in alignment between the various components. The semi-axes and orientation of the halos are estimated taking into account sub-halos in the simulation. The mass distributions of the dark-matter (DM), gas, and star halos are characterised by a half-mass radius, a concentration parameter and (projected) axis ratios.

Results. We present statistics of the shape parameters of 336 540 halos from the EAGLE RefL0025N0376 simulation and describe their evolution from redshift $z = 15$ to $z = 0$. We measured the three-dimensional shape parameters for the DM, the gas, and the star components as well as for all particles. We also measured these parameters for two-dimensional projected distributions. At $z = 0.5$, the minor axis of gas aligns with the minor axis of DM for massive halos ($M > 10^{12} M_{\odot}$), but this alignment is poorer for less massive halos. The DM halos axis ratios b/a and c/a have median values of 0.82 ± 0.11 and 0.64 ± 0.12 , respectively. The gas in halos that also contain stars has a more flattened shape, with $b/a = 0.70 \pm 0.19$ and $c/a = 0.38 \pm 0.20$. The sphericity of gas in halos w/ and w/o stars appears to be negatively correlated to the total mass, while the sphericity of DM is insensitive to it. The measured projected axis ratios, b_p/a_p , of star halos at $z = 0.5$ have a median value of 0.80 ± 0.07 , which is in good agreement with ground-based and space-based measurements within 1σ . For DM halos, we measure a value of 0.85 ± 0.06 . The evolution of the concentration as a function of redshift is fairly homogeneous for the various components, except for the starless gas halos, which appear much more concentrated for $z > 0.7$.

Key words. methods: numerical – Galaxy: halo – dark matter – gravitational lensing: strong

1. Introduction

The widely accepted Λ -cold dark matter (Λ -CDM) model postulates that a highly non-uniform system of voids, sheets, filaments, and halos – commonly known as the cosmic web – emerges as a result of the gravitational growth of tiny heterogeneities in the early Universe's matter distribution (Bond et al. 1996; Faucher-Giguère et al. 2008). The evolution over time of this large-scale cosmic web is regulated by the dark-energy (DE) and dark-matter (DM) components, since they are today by far the main contributors to the Universe's total mass-energy content. Regarding the visible galaxies, which gather most of the baryonic content, a key assumption of galaxy formation models is that these galaxies develop within the DM halos that permeate the cosmic web (White & Rees 1978; Blumenthal et al. 1984). However, baryonic structures do not always directly trace the form and shape of their DM halos due to the wide variety of dissipative physical processes that are involved in the evolution of

galaxies. Through the redistribution of angular momentum, the geometry of the DM halo influences certain properties of galaxies, such as the development and evolution of bars (Collier et al. 2018, 2019; Kumar et al. 2022).

There are other observational influences on large-scale mass distribution such as weak and strong lensing. With the emergence of weak gravitational lensing as a precision probe of cosmology, galaxy alignments have taken on an added importance because they can mimic cosmic shear, such as the effect of gravitational lensing by a large-scale structure on observed galaxy shapes (Kirk et al. 2015). The shape parameters of the mass distribution of foreground galaxies can be measured from the two-dimensional shear field derived from background galaxies (Natarajan & Refregier 2000). Indeed, the distribution of satellite galaxies preferentially aligns with the major axis of the central galaxy, with a clear dependence on both halo mass and galaxy colours (Wang et al. 2008). Another important tool to probe the total matter distribution of galaxies is strong

gravitational lensing¹. In that context, quadruply imaged quasars (quads) are particularly interesting because of the constraints they put on the radial and azimuthal mass distribution within the deflectors. In multiple imaged lens systems, the observed configuration of images is completely dependent on the projected mass distribution of the deflecting galaxy. Lenses are sought in many large and deep sky surveys such as the Sloan Digital Sky Survey (Tegmark et al. 2004), the Dark Energy Survey (Dark Energy Survey Collaboration 2016) and the *Gaia* data (Krone-Martins et al. 2018; Delchambre et al. 2019; Stern et al. 2021) with machine-learning algorithms. Known lenses are too rare to be used for the training of these algorithms. One needs a set of realistic simulations for that with in particular a realistic distribution of projected ellipticities of the deflecting galaxies. Recent studies show that using a non-singular isothermal ellipsoid projected mass distribution with the addition of an external shear as a lens model helps reproduce the distribution of observed configurations (Witt 1996; Woldesenbet & Williams 2015; Gomer & Williams 2021). Ongoing searches (such as those performed by *Gaia* GraL) allow one to establish samples with well understood selection functions, and pave the way to upcoming facilities such as *Euclid*, which will increase the number of known lenses by a few orders of magnitude.

Observational evidence for non-spherical DM halos (Barnabè et al. 2012; van Uitert et al. 2012) requires an account of baryonic physics on galaxy formation. Running cosmological numerical simulations is a way to improve our understanding of the building-up of galaxies by studying how simulated galaxies manifest themselves in a variety of morphologies ranging from thin to bulge-dominated discs and to ellipsoids of various sizes and masses. To simulate galaxies with statistical characteristic properties similar to the ones measured in real catalogues, it is crucial to combine the evolution of large-scale DM halos with the study of baryon physics at the smallest scales. Hence, highly resolved simulations large enough to include the large-scale structure while capturing the morphology and underlying physics of galaxies are required. Progress has been made in this regard by the current generation of hydrodynamical simulations such as the Magneticum Pathfinder (Remus et al. 2015), the MassiveBlack-II (Khandai et al. 2015), the EAGLE project (Schaye et al. 2014), and more recently the IllustrisTNG simulation (Pillepich et al. 2019). With these numerical simulations, it is possible to make detailed predictions of the cosmological distribution and evolution of the galaxies that can be compared with a wide range of observations, including future large-scale structure surveys, such as those that will be produced by the *Vera Rubin* Observatory (LSST Science Collaboration 2009), the *Euclid* spacecraft (Laureijs et al. 2012), and the *Nancy Grace Roman* Space Telescope (Spergel et al. 2015).

The EAGLE suite of simulations is particularly well suited to this application because these simulations were explicitly calibrated to reproduce the $z = 0$ observed galaxy stellar mass function with realistic constraints on sizes. In this study, we used the hydrodynamic simulation with a volume of side $L = 25$ cMpc, with $N = 376^3$ initial DM particles (Schaye et al. 2014). This simulation has the smallest volume but has the same resolution as larger volume simulations. Several groups have analysed various aspects of the morphologies of the galaxies in the EAGLE simulations (Velliscig et al. 2015; Shao et al. 2016; Hill et al.

2021). In this paper we intend to complement these works on the morphology of DM, gas, and star halos by bringing in a new method to determine the shape parameters of the halos and their orientations. We also analyse in details the mass distribution of the various components of the halos.

We measure directly the shapes of the DM, gas and stars components of halos and subhalos (modeled as ellipsoids in three dimensional space). We examine how shapes evolve with time and as a function of halo total mass. We also measure the projected (2D) shapes for comparison with observations. By measuring projected ellipticities of the DM, gas and stars components of simulated galaxies, we can estimate the distribution of ellipticities. The shape of the halos together with statistics on the ellipticities will allow us to generate simulations of gravitational lenses that will be used to train an automatic detection tools for these objects.

This paper is structured as follows. In Sect. 2, we briefly describe the EAGLE simulation with a focus on the content of the simulation at redshift $z = 0.5$, which is the typical redshift of known gravitational lens deflectors. In Sect. 3, we present our methods to measure the shapes of the DM, gas, and star components of halos and sub-halos. In Sect. 4, we validate our method on simulations and we study the impact of the number of particles and the axis ratios of the halos on the measured parameters. In Sect. 5, we provide a comprehensive description of the shapes and the internal alignment of gas with DM and stars as a function of halo total mass at redshift $z = 0.5$. We also analyse the properties of the projected halos on the sky. In Sect. 6, we examine the dependence of the DM, gas, and star halos on the redshift and the halos' total masses. In Sect. 7, we discuss and summarise our findings.

2. Data

2.1. The EAGLE simulations

The Evolution and Assembly of GaLaxies and their Environments (EAGLE) project (Schaye et al. 2014) is a suite of cosmological, hydrodynamic simulations aimed at studying the evolution of DM and gas in the Universe and the formation of stars and black holes. These simulations are computed using a modified version of the GADGET-3 tree-SPH code from (Springel et al. 2005). They include sub-grid descriptions for radiative cooling (Wiersma et al. 2009a), star formation (Schaye & Dalla Vecchia 2008), multi-element metal enrichment (Wiersma et al. 2009b), black hole formation (Springel et al. 2005; Rosas-Guevara et al. 2015), as well as feedback from massive stars and active galactic nuclei (AGNs; Dalla Vecchia & Schaye 2012; Schaye et al. 2014). The EAGLE public database, which includes the EAGLE galaxy database (McAlpine et al. 2016) and EAGLE particle data (The EAGLE Team 2017), can be accessed and queried via their website².

Each simulation includes 29 snapshots at different redshifts between $z = 0$ and $z = 15$ allowing us to study the evolution of the morphology of galaxies along time. Our primary analysis focuses on the EAGLE simulation with the smallest cosmic volume, RefL0025N0376. This choice is motivated by the need to extensively test our analysis code while limiting the calculation time. This simulation follows a cubic periodic volume of side $L = 25$ cMpc, with $N = 376^3$ collision-less DM particles with an individual mass of $M_{\text{DM}} = 9.7 \times 10^6 M_{\odot}$ and a starting equal number of baryonic (gas) particles, each with an initial

¹ Strong gravitational lensing depicts the formation of multiple images of a background source whose light rays are deflected due to the presence of a large mass concentration like a massive galaxy or a galaxy cluster in the line of sight between the observer and the source.

² <http://icc.dur.ac.uk/Eagle/>

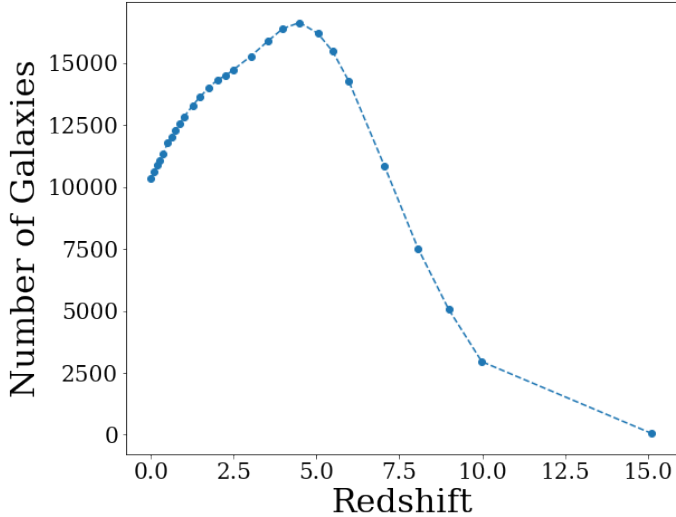


Fig. 1. Number of halos studied in this paper as a function of redshift in the RefL0025N0376 EAGLE simulation.

mass of $m_g = 1.81 \times 10^6 M_\odot$. The initial conditions were created using values of the cosmological parameters derived from the initial Planck data release (Planck Collaboration XVI 2014), namely $\Omega_0 = 0.307$, $\Omega_b \approx 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 \approx 0.83$, $n_s = 0.9611$, and $h \approx 0.68$, $Y = 0.248$. The ratio of the baryonic-to-non-baryonic particles is therefore equal to $(\Omega_0 - \Omega_b)/\Omega_b \equiv (m_{\text{DM}}/m_b) \approx 5.4$.

2.2. Star formation and its energy feedback

In the EAGLE simulation, stars are formed from gas particles when the gas density exceeds a given threshold (see Schaye et al. 2014, for a detailed presentation). The limited resolution of the simulation makes it necessary to impose such a threshold for star formation for a density above which a cold interstellar gas phase is expected to form. The probability that a gas particle is converted into a collision-less star particle depends on the gas particle’s pressure (Schaye 2004; Schaye & Dalla Vecchia 2008) and metallicity. The evolution of stars is tracked using the metallicity-dependent lifetimes of Portinari et al. (1998). The mass lost by stars is distributed among the neighbouring particles through winds from B-, A-, and G-type stars, winds from massive stars, and core-collapse supernovae using the nucleosynthetic yields from Marigo (2001). Stars can also inject energy and moment into the interstellar medium (ISM) through stellar winds, radiation, and supernovae. These processes are particularly important for massive stars.

2.3. Black holes and feedback from AGN

Supermassive black holes (BHs) are settled in halos with a total mass greater than $10^{10} M_\odot$ and which do not already contain a black hole (Springel et al. 2005; Booth & Schaye 2009; Schaye et al. 2014). These halos are identified by a friends-of-friends (FoF) algorithm at epochs corresponding to a logarithmic sampling of the expansion factor. The gas particle whose density is the highest is converted into a collision-less BH particle with sub-grid BH mass $m_{\text{BH}} = 10^5 M_\odot h^{-1}$. BHs grow by accretion of gas. This accretion is dependent on the mass of the BH, the local density and temperature, the velocity of the BH relative to the ambient gas, and the angular momentum of the gas with respect

to the BH. BHs are merged with other BHs if they are separated by a distance that is smaller than both the smoothing kernel of the BH, h_{BH} , and three gravitational softening lengths. A limit on the allowed relative velocity prevents BHs from merging during the initial stages of halo mergers.

The feedback associated with AGNs is implemented in a manner analogous to energy feedback from star formation. This feedback is tuned so that the simulation reproduces the present-day galaxy stellar mass function and realistic sizes for disc galaxies (Crain et al. 2015), and it verifies the close relationship between the stellar mass of galaxies and the mass of their central BHs.

2.4. Halo identification and sample

As mentioned, DM structures are initially identified using the FoF algorithm with a linking length of 0.2 times the mean inter-particle spacing (Davis et al. 1985). Particles other than the DM ones (gas, stars, and black holes) are assigned to the same group as their nearest DM neighbours. Halos are considered as gravitationally self-bound structures within the FoF groups using the SUBFIND algorithm (Springel et al. 2001; Dolag et al. 2009). Unlike FoF, SUBFIND considers all types of particle and identifies self-bound sub-units within a bound structure that are called sub-halos. The local minima in the gravitational potential field are the centres of sub-halo candidates. The sub-halo with a minimum value of the gravitational potential within a FoF group is defined as the ‘main halo’. Any particle bound to the group but not assigned to any other sub-halos is assigned to the main halo. In this study, sub-halos are defined as satellite halos to the main halo.

Our study is based on the evolution of the smallest EAGLE simulation, RefL0025N0376, spanning a stellar mass range from 10^9 to $10^{11} M_\odot$ over a redshift range of $0 \leq z \leq 15$. Figure 1 presents the number of halos from the small simulation according to the snapshot redshifts. The number of detected halos first increases from $z = 15$ to $z = 5$ as their difference in local density is getting larger due to their gravitational build-up. Then, at redshift $z \approx 5$, the hierarchical formation of large and massive halos becomes dominant, leading to a decrease in the total number of individual structures.

2.5. Overall properties of the halos at $z = 0.5$

In this section, we present the halo content of the RefL0025N0376 simulation, and for that we chose the redshift $z = 0.5$, which is the typical redshift of known gravitational lens deflectors. We analyse the mass distribution and look at the distribution of the different types of particles in the halos and sub-halos.

The mass distribution of the DM particles remains constant in the simulation ($m_{\text{DM}} = 9.7 \times 10^6 M_\odot$). DM particles are non-colliding and therefore only interact gravitationally. Several gas particles are more massive than their initial mass ($m_g = 1.81 \times 10^6 M_\odot$) but this situation remains rare due to the high probability of these particles being converted into star particles. As already mentioned, the initial mass ratio between DM and gas particles is exactly the ratio of the density of these two components in the cosmological model used by the EAGLE simulation.

In Fig. 2, we present the fraction of the sub-halo mass to the total mass of the halo as a function of the total mass of halos. In this way, we can detect the presence of massive sub-halos in comparison to their main parent halo. Most halos have a mass between 10^9 and $10^{11} M_\odot$. For the less massive halos, a large

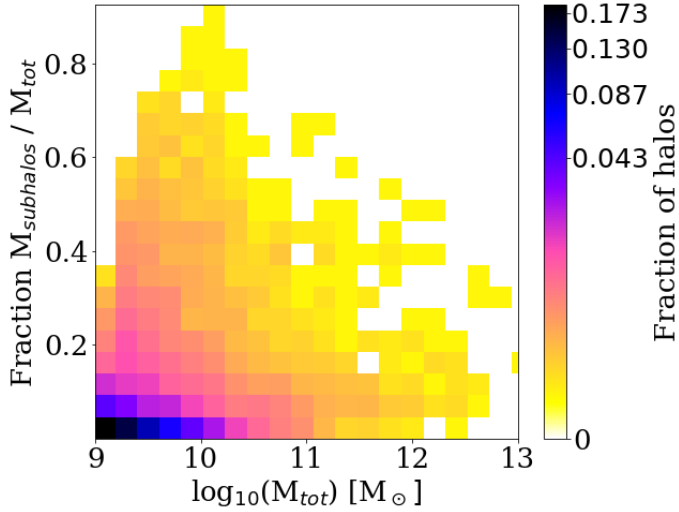


Fig. 2. Fraction of the sub-halo mass to the total mass of the halo as a function of the total mass of halos. Colours represent the fraction of halos at redshift $z = 0.5$.

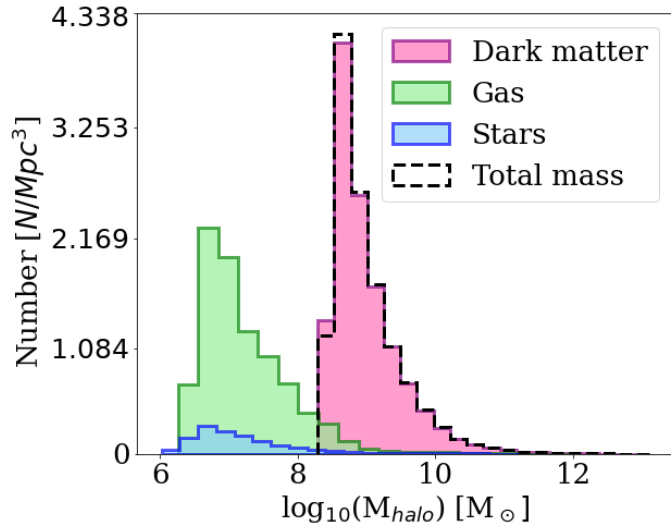


Fig. 3. Mass distribution of 42 766 EAGLE halos at redshift $z = 0.5$ for each type of particle. The y-axis shows the volume densities for a box size of 16.94 Mpc per side. The black dotted line represents the total mass of the halos by summing the mass of all the particles. We notice that the DM particles dominate the halos masses by several orders of magnitude.

proportion of the mass can be accounted for by the sub-halos. This is an issue when analysing the morphology of such halos. To take it into account, any halo whose sub-halos contribute to more than 10% of its total mass are not considered in the results. We found that 88% of all halos identified at redshift $z = 0.5$ have less than 10% of their total mass in sub-halos.

Figure 3 presents the distribution of the mass for the 42 766 halos considered. The DM mass contribution is several orders of magnitude larger than those of the two other types of particles. As expected, the mass of the halos is largely dominated by DM.

Figure 4 presents the number of particles in these halos for each type of particle. There are very few massive black holes in each halo, so they do not directly influence the overall shape. Therefore, they will only be accounted for in the total mass analysis of the halos. On average, 50% of the halos have a small num-

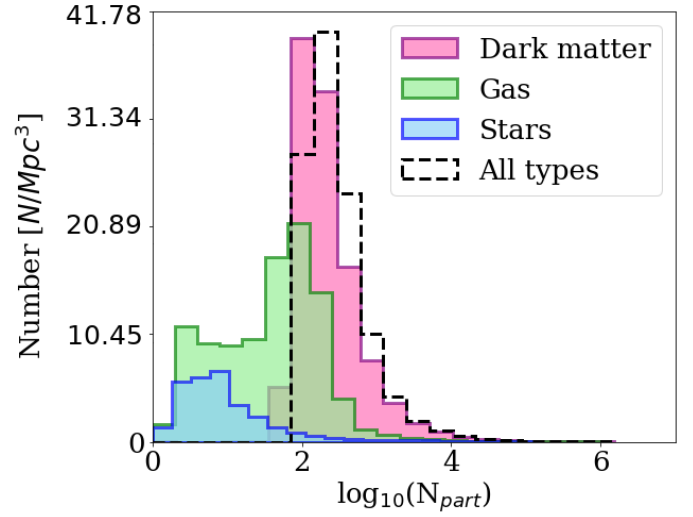


Fig. 4. Distribution of number of particles of each type in halos in RefL0025N0376 simulation at redshift $z = 0.5$. The y-axis shows the volume densities for a box size of 16.94 Mpc per side.

ber of DM particles (50–300 particles), but the number of baryonic particles is even smaller (only 7% have more than 50 gas particles and 2% more than 50 stars). As a result, we observe many dark halos containing very few or no baryonic particles (90% of halos have no stars).

3. Shape analysis

The EAGLE halos have various shapes and sizes and often have rich structures. In order to analyse all the simulations and model each halo whatever its shape is, we developed a specific analysis procedure that is detailed hereafter. It was applied five times, respectively, for each type of particle (DM, gas, stars) as well as for the sum of the baryonic components and all particles together.

We also performed a complementary shape analysis of the projected halos in the $[xy]$, $[yz]$, and $[xz]$ plans to derive the statistical distribution of the shape parameters of once projected on a plane that is tangential to the celestial sphere. The projected mass distribution being a central parameter of the modelling of gravitational lenses.

An accurate measurement of the major and minor axes of an EAGLE halo is not straightforward, because generally the distribution of particles is not homogeneous and presents peripheral sub-structure, as illustrated in Fig. 5, that impacts the measurements. Recent studies of the EAGLE simulations focus on their star and gas content (Thob et al. 2019; Hill et al. 2021) and only study the central part of the halos. The present work intends to assess the overall morphology at a larger scale accounting for most of the matter including DM, which is the gravitational dominant component. A specific strategy was, however, developed to get rid of halo sub-structures.

3.1. Elimination of peripheral sub-structures for shape analysis

3.1.1. Determination of the centre of the halo

We started by determining the centre of the main halo containing all the particles without taking into account the influence of peripheral sub-structures. First, we considered all particles and

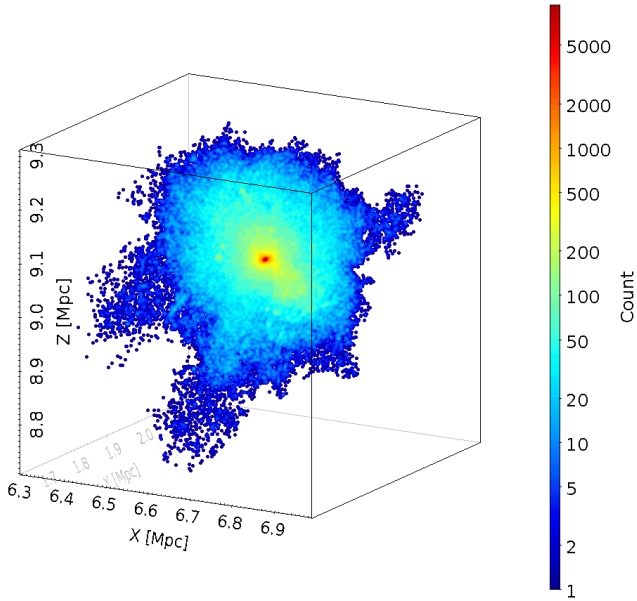


Fig. 5. Three dimensional plot of main halo #20 from the RefL0025N0376 simulation at $z = 0.5$ with a large-scale sub-structure in the down part of size ~ 150 kpc. Coordinates are in the initial EAGLE simulation reference system (X, Y, Z). All types of particles are represented.

found their median position (x_0, y_0, z_0) . Then, a new median position (x_0', y_0', z_0') was derived from the particles encompassed in a sphere centred on (x_0, y_0, z_0) and with a radius equal to one tenth of the maximal distance between any particle and the centre, giving priority to the densest part of the halo. This second step is re-run but using the new centre (x_0', y_0', z_0') until convergence is reached and the final median position of the remaining particles defines the final centre of the halo.

A visual check is used to monitor convergence. This centre corresponds to the location of the highest density peak of particles. In practice, three iterations are often required before convergence.

3.1.2. Working sphere

In order to confidently measure the shape properties of a halo, we need to get rid of the influence of peripheral sub-structures. For this, we aim to determine a working spherical volume characterised by its radius, r_{80cc} , that encompasses a sufficient amount of mass while not including the perturbing sub-structures. r_{80cc} is empirically chosen as the radius at which the density drops by 80% from the central core density. The computed sphere then delimits the volume comprising the particles we chose to consider.

To derive r_{80cc} , it is necessary to establish the density profile of the halo. For this, we first need to determine the size of the spherical core of the halo, r_{core} , that contains 10% of the total mass and within which the core density will be measured.

To do so, we produce a cumulative mass growth curve of the halo (as shown in Fig. 6) by summing the mass of particles in successive spherical shells with width equals to 1/50 of the total spatial extension of the halo. When the derived mass does not change from the mean of the last three measurements by more than 1σ , the process stops. The asymptotic value of the mass growth curve corresponds to the total mass of the halo. From this

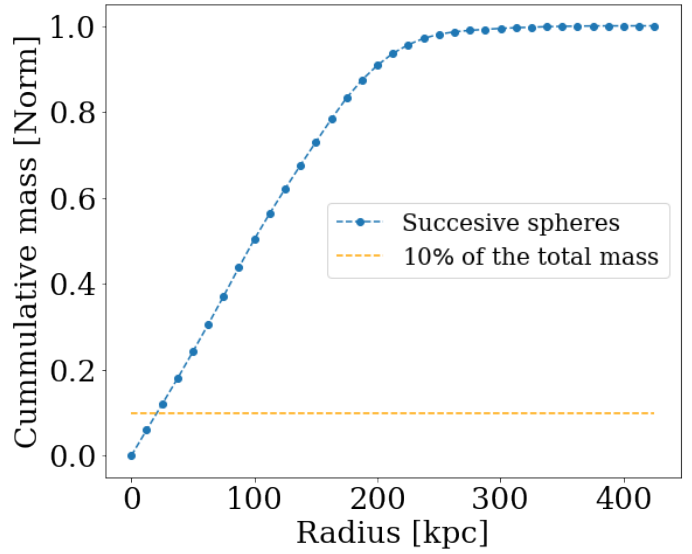


Fig. 6. Cumulative mass growth curve of the main halo #20, obtained from mass measurement in successive concentric spherical shells with width of 1/50 of the total spatial extension. The intersection of the yellow line with the growth curve indicates the core radius r_{core} encompassing 10% of the total mass. Here $r_{core} = 20.92$.

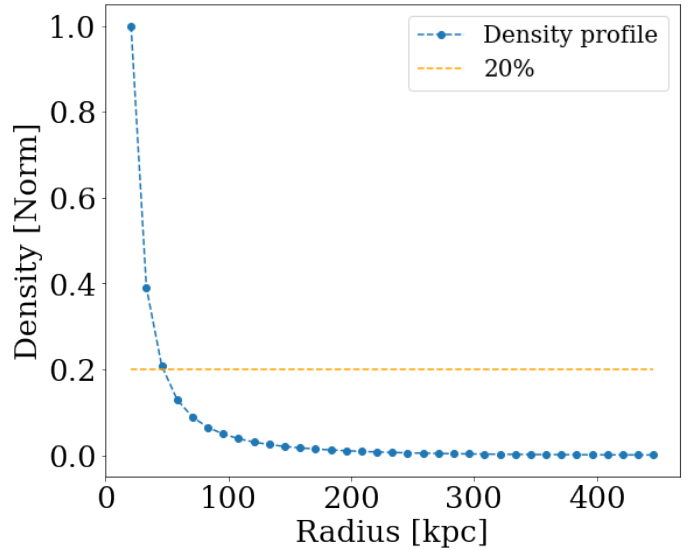


Fig. 7. Density profile of main halo #20, obtained from density estimation in successive concentric spherical shells with width of 1/50 of the total spatial extension. The intersection of the yellow line with the density curve indicates the radius r_{80cc} , at which point the density drops by 80% from the central core density. Here $r_{core} = 20.92$ kpc and $r_{80cc} = 47.25$ kpc.

curve, we determine the core radius of the halo, encompassing 10% of the total (asymptotic) mass.

It is then possible to establish the radial density profile of the halo by estimating the density in the core, r_{core} , and in successive spherical shells with a radius step of 1/50 of the total spatial extension of the halo. We present in Fig. 7 the resulting density profile.

The radius at which the density drops by 80% from the central core density, r_{80cc} , is derived from this figure via a linear interpolation. This radius defines the sphere in which the following shape analysis is performed.

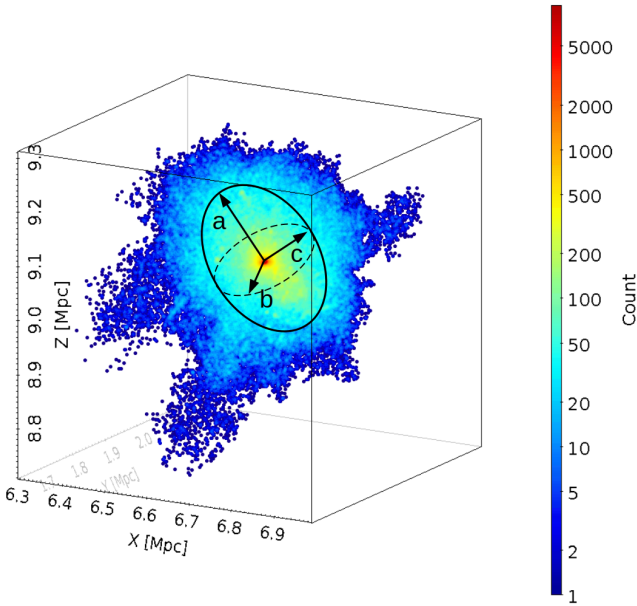


Fig. 8. Three dimensional plot of main halo #20 from EAGLE ($z = 0.5$) with materialisation of the final ellipsoid describing the halo, obtained by an iterative PCA.

3.2. Semi-axis and shape parameter estimation

The semi-axes (a , b , and c , from largest to smallest) of the halo are estimated by applying a principal components analysis (PCA) to the spatial coordinates of the particles contained in the working sphere. The PCA determines the principal axes of the distribution that best explain the (normal) variance in the coordinate space. The length of each principal axis is given by the square root of the eigenvalues of the covariance matrix of the particles coordinates. From a practical point of view, the semi-axes measure the standard deviations of the underlying three dimensional particle distribution.

Once the first estimation of the semi-axes is obtained, the PCA is re-applied iteratively to the selection of particles encompassed in the ellipsoid defined by the semi-axes derived at previous iteration and a new estimation of the semi-axes is produced. After a few iterations (<10) convergence is reached and the final semi-axis estimates (a, b, c) are obtained (see Fig. 8). By convention, $a > b > c$.

From the final estimates of the semi-axes of the fitted ellipsoid, we define the axis ratios b/a , c/a , and c/b . The ratio c/a is also called sphericity ($c/a = 1$ corresponds to a sphere, whereas $c/a = 0$ corresponds to a flat disc if $b \neq 0$).

We have derived in this way, for each halo, estimates of the semi-axes of the ellipsoids corresponding to the distribution of each type of particles (stars, gas, DM) as well as to the distribution of baryonic and all particles: $(a, b, c)_{\text{star}}$, $(a, b, c)_{\text{gas}}$, $(a, b, c)_{\text{DM}}$, $(a, b, c)_{\text{baryonic}}$, $(a, b, c)_{\text{all}}$.

3.3. Orientation of halos

In the case of a three-dimensional distribution of particles, the principal axes of the distribution define a new orthogonal reference system (X', Y', Z'). This reference system is rotated with respect to the initial EAGLE reference system (X, Y, Z) by three Euler angles (ψ, θ, ϕ) that are given in the rotation matrix. This matrix allows us to convert the position of the particles given in the EAGLE simulation reference system (X, Y, Z) to the new

reference system defined by the principal axes (X', Y', Z') of the distribution.

We derived estimates of the three Euler angle this way for each halo, for each type of particle, as well as for the entirety as $(\psi, \theta, \phi)_{\text{stars}}$, $(\psi, \theta, \phi)_{\text{gas}}$, $(\psi, \theta, \phi)_{\text{DM}}$, $(\psi, \theta, \phi)_{\text{all}}$.

3.4. Mass growth curve and mass concentration

Once the shape of a halo is known, it is possible to establish the final mass growth curve characterising its mass distribution. This curve is computed by adding the mass of the particles contained in successive concentric ellipsoid shells with a fixed width, as illustrated in Fig. 9 (left panel). The calculation is stopped when an asymptote is reached. The resulting curve is smoothed by a moving average with a three-point window. Figure 9 (right panel) presents the final mass growth curve of the main halo #20 from EAGLE ($M = 1.18 \times 10^{12} M_{\odot}$, $z = 0.5$, $N_{\text{part}} = 1.62 \times 10^5$ particles).

In galaxy classification studies, light concentration is often used to characterise the light distribution of galaxies in view of discriminating between early-types (bulge dominated) and late-type objects. A similar quantity, the mass concentration, can be estimated as (Conselice 2003):

$$C = 5 \times \log \left(\frac{r_{80}}{r_{20}} \right), \quad (1)$$

where r_{20} and r_{80} are the radii containing 20% and 80% of the total mass, respectively. These radii are measured from the mass growth curve as the intersection of the curve with the horizontal lines corresponding to 20% and 80% of the total mass.

We derived a mass growth curve and a mass concentration this way for each halo, for each type of particle, as well as for the whole halo.

4. Validation of the algorithm

4.1. Simulation of halos

We validated our method by applying it to 10 000 simulated halos with shape parameters ($a_{\text{sim}}, b_{\text{sim}}, c_{\text{sim}}$) randomly selected such that $a > b > c$ and $a < 1$. To simulate halos, we draw the particles coordinates from a multivariate normal distribution with mean $\mu = [\mu_x, \mu_y, \mu_z]$ and covariance matrix $\Sigma = \text{diag}(a_{\text{sim}}^2, b_{\text{sim}}^2, c_{\text{sim}}^2)$. We then rotated the whole distribution of particles by three randomly selected Euler angles ($\psi_{\text{sim}}, \theta_{\text{sim}}$ and ϕ_{sim}) in order to simulate randomly orientated halos. The resulting simulated halo has particle coordinates expressed in the original (X, Y, Z) reference system. We do not consider the possible presence of substructures in the halos since our algorithm is built to minimise the influence of these substructures before applying the morphological analysis.

4.2. Estimation of the halo's parameters

Two parameters may influence the correct recovery of the input parameters defining the shape and orientation of the simulated halos: the number of particles in the halo (too sparse distribution may be poorly recovered) and the sphericity of the halo (distribution close to a sphere may be harder to retrieve).

4.2.1. Dependence on the number of particles

To explore the dependence on the number of particles, we simulated 10 000 synthetic halos containing 10–2000 uniformly

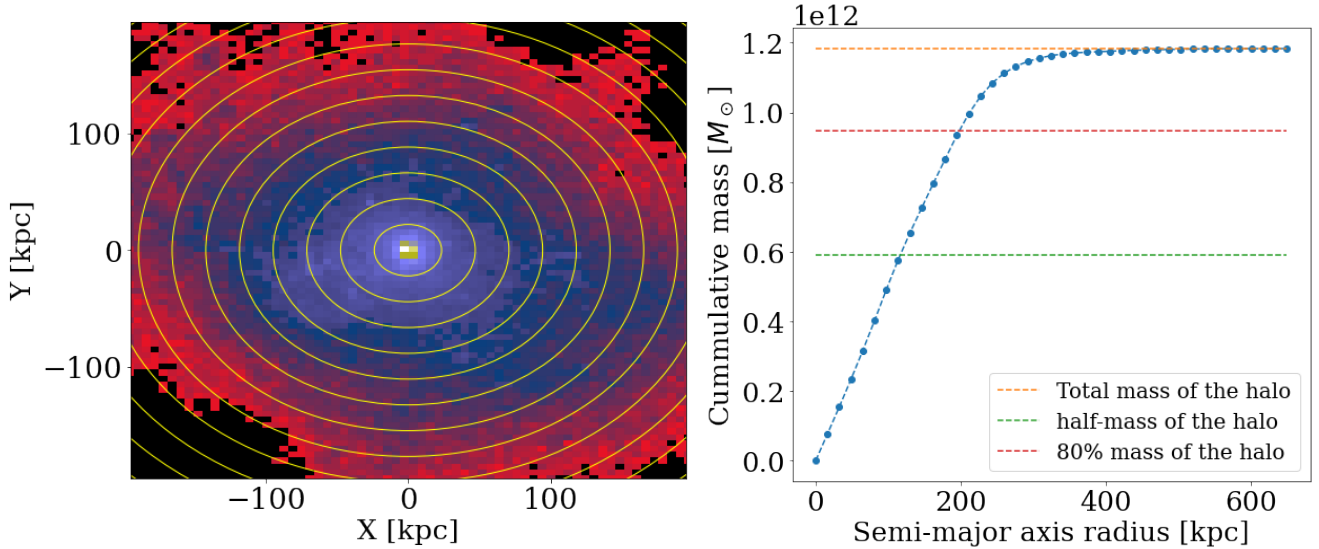


Fig. 9. Calculation of the mass growth curve. *Left panel:* projection along (X, Y) of main halo #20 of EAGLE redshift $z = 0.5$. Yellow ellipses represent successive concentric ellipses. Here, only one in five ellipses is displayed. *Right panel:* mass of the particles contained in each successive ellipse is summed to construct the cumulative halo mass curve. The green dashed line is the asymptotic value of the mass growth curve, which corresponds to the normalised total mass of the halo. The red and blue dashed lines correspond to the radii containing 20% and 80% of the total mass, respectively.

drawn particles, as observed in EAGLE halos (see Fig. 4) and with a fixed sphericity. We did that for two different sphericities: $c_{\text{sim}}/a_{\text{sim}} = 0.72$ (rather spherical halos) in one case and $c_{\text{sim}}/a_{\text{sim}} = 0.25$ (flattened-disc halos) in the other. We then applied our algorithm to these synthetic halos and derived estimates of the semi-axes (a, b, c) and the Euler angles (ψ, θ, ϕ) .

In Appendix A, we present the diagrams presenting the relative errors³ $(\Delta a, \Delta b, \Delta c)$ and the difference between the estimated angles (ψ, θ, ϕ) and their true value $(\psi_{\text{sim}}, \theta_{\text{sim}}, \phi_{\text{sim}})$ as a function of the number of particles and the sphericity.

The semi-axes are generally well recovered to a level better than 10%, and generally to one better than 5% for halos with more than 100 particles, whatever the sphericity is. There is a visible degradation of the quality of the semi-axes when the number of particles is smaller than 100. The semi-axis ‘ c ’ then tends to be slightly underestimated by less than 5% for halos with fewer than 100 particles. We also notice that the orientation of the halos is globally well recovered to better than 10 deg in all cases. The sphericity of the halos is influencing the quality of the recovered orientation as expected: for quasi-spherical halos with more than 100 particles, the errors are of 8–10 deg or fewer, while for flattened halos they are around 2–3 deg.

4.2.2. Dependence on the sphericity

To explore the dependence on the sphericity of the halos, we have simulated 10 000 synthetic halos of 500 particles with randomly selected axis ratios such that $(b_{\text{sim}}/a_{\text{sim}}) < 1$ and $(c_{\text{sim}}/b_{\text{sim}}) < 1$. We analysed these halos with our algorithm and derived estimates of their semi-axes (a, b, c) and their Euler angles (ψ, θ, ϕ) .

The relative errors $(\Delta a, \Delta b, \Delta c)$ on the estimated semi-axes and the difference between the estimated angles (ψ, θ, ϕ) and their true value as a function of their sphericity is presented in Appendix A. We observe that the semi-axes (a, b, c) are well recovered but a slight systematic effect is observed for halos with

high sphericity ($c/a > 0.9$): major-axis a is overestimated by 0.2%, semi-axis b is underestimated by 0.3%, while c does not seem to be influenced by the sphericity of the halos. The orientation of the halos is also well recovered with relative errors not influenced by the sphericity of the halo, except for ϕ (the third rotation). The distribution of quartiles of $\Delta\phi$ indicates a severe degradation (errors up to 10 deg and more) for sphericities larger than 0.9 with a systematic over-estimation of ϕ .

We conclude that the estimate of the value of the principal semi-axes are very well recovered when the number of particles included in the halos are larger than 100. The semi-axes are generally recovered at a rate better than 5%. The Euler angles of the halos are well recovered to better than 10 deg for halos with a number of particles larger than 100, but their recovery is dependent on the sphericity of the halo, degrading for quasi-spherical halos with $c/a > 0.9$.

5. Application to the EAGLE snapshot at redshift $z = 0.5$

We first focus on a comprehensive description of the results obtained for the redshift $z = 0.5$ in the RefL0025N0376 simulation, because one motivation of our work consists of studying strong lensing. We limit our analysis to halos containing at least 100 particles of each type in order to ensure the reliability of the estimated parameters, as seen in Sect. 4.2.1. We derive shape parameters describing (i) the DM distribution for 17 480 halos, (ii) the gas distribution for 2122 halos, and (iii) the star distribution for 682 halos. From the 2122 halos with a sufficient number of gas particles to be considered as a halo, 1440 do not contain the minimum of 100 stars we required, while 682 do (hereafter gas halos ‘w/ stars’, gas halos ‘w/o stars’).

5.1. Orientation of the halos

We begin with an examination of the differences in the alignment of the gas, stars and DM halos as a function of the total mass. We

³ $\Delta x = (x - x_{\text{sim}})/x_{\text{sim}}$.

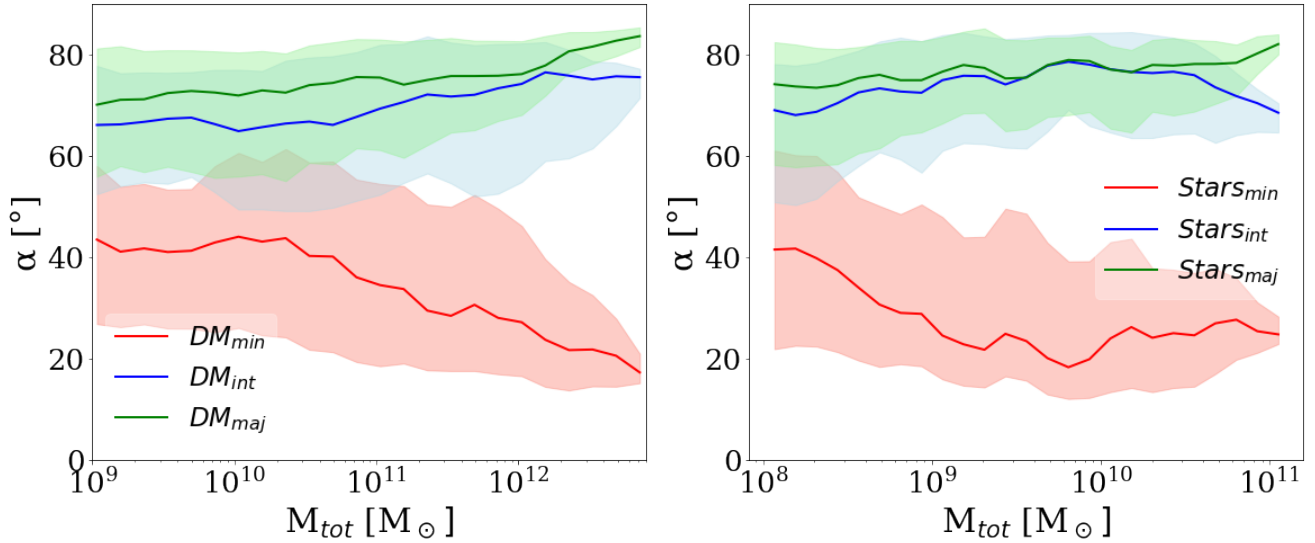


Fig. 10. Difference of alignment angle, α , between the minor axis of gas halos and all semi-axis of the DM (left panel) and star halos (right panel). The solid lines indicate the median values of α , whilst the shading denotes the 25th–75th quartiles. The values are shown as a function of the total mass (M_{tot}). The red, blue, and green curves correspond to the difference of alignment of the gas minor axis with respect to the minor, intermediate, and major axes, respectively.

quantify the differences of alignment of the various components via the angle α , between the semi-axes of the ellipsoid describing the distributions via the scalar product of two major (intermediate or minor) axes (α in case of projected halos). We note that it is more typical in the literature to use the major axis. In contrast, it is the minor axis that is the most natural choice when focusing on discs. The quantity α only varies between 0 and 90 deg where $\theta = 0$ deg indicates perfect alignment and $\theta = 90$ deg indicates orthogonality.

Figure 10 shows the differences of alignment between the minor axes of the gas distribution and all semi-axis of the DM and star halos. In each panel, the difference of alignment is shown as a function of the total mass, M_{tot} . In general, the alignment between the gas and the DM is relatively strong between minor axes, with the median alignment typically being $\alpha \simeq 40$ deg at $M_{\text{tot}} = 10^{10} M_{\odot}$ and declining to $\alpha \simeq 30$ deg at $M_{\text{tot}} = 10^{11} M_{\odot}$ down to $\alpha \simeq 20$ deg at $M_{\text{tot}} > 10^{12} M_{\odot}$. It is important to note that the differences of alignment with the less massive DM halos is imprecise since they have a quasi-spherical shape in most cases (see Sect. 5.3).

The alignment of the gas and the stars is only slightly dependent on the mass of the halos. The respective minor axes of gas and star halos have a median alignment of at least 40 deg and decrease and stabilise at around 20 deg for all masses. This behaviour will have to be confirmed with a more massive sample of halos.

Bett et al. (2010) quantified the difference of alignment angle between the stellar and total matter distribution in a sample of disc galaxies selected from a hydrodynamic simulation. They found that half of these galaxies have a difference of alignment larger than 45 deg. Using a sample of the EAGLE and cosmo-OWLS simulations, Velliscig et al. (2015) reported that the difference of alignment of massive gas halo is as large as 50 deg and dropped to 30 deg for the halo mass range $13 < \log_{10}(M_{200}) < 14$. Despite the lack of data for massive halos beyond $10^{12} h^{-1} M_{\odot}$ in the small EAGLE simulation, both studies are in broad agreement with our findings for similar halo masses. A second paper will follow presenting the results of larger simulations including halos of larger masses.

5.2. Semi-axes

In Fig. 11, we present the distribution of the ratio $\tilde{A}_{\text{gas}} = a_{\text{DM}}/a_{\text{gas}}$ and $\tilde{A}_{\text{stars}} = a_{\text{DM}}/a_{\text{stars}}$ of the fitted semi-axes of halos according to the type of particles. As mentioned before, gas halos are split in two groups, gas halos depleted from stars (1440) and gas halos with a sufficient number of stars (682). When considering all particles together, the resulting distributions are similar to those of DM, and this is why the all-particles case is not added to the figures. Since we see the same behaviour on all three axes (a, b, c), we only present the semi-axis a , but the results also apply for the b and c axes.

We remind the reader that the length of the semi-axes are obtained from the square root of the eigenvalues of the PCA. Thus, the measurement of the principal axes reflects the characteristic size in each direction of the halos. In this way, the \tilde{A} ratio can be used to estimate the extension of the distribution. If the semi-axis value is small, halos have small extension, whereas the larger the value, the more widespread the halo. This means that $\tilde{A} > 1$ denotes that gas or star halos are less extended than DM halos, which is the case for star halos in Fig. 11. For gas halos, we observe two cases: some of the gas halos w/o stars (26%) and w/ stars (42%) have larger semi-axes than the DM one. This behaviour is observed both for gas halos w/o star particles and for halos with at least 100 star particles, but it is more significant for halos with more than 100 star particles.

We then look at the ratio between the length of the gas and DM semi-axes in detail. Figure 11 also presents the distribution of the ratios $\tilde{B}_{\text{gas}} = b_{\text{DM}}/b_{\text{gas}}$ (middle panel) and $\tilde{C}_{\text{gas}} = c_{\text{DM}}/c_{\text{gas}}$ (right panel) for the 2 122 halos of gas and for each semi-axis (a, b, c). For \tilde{B} and \tilde{C} , we only consider halos with $\tilde{A}_{\text{gas}} < 1$ and $\tilde{A}_{\text{stars}}, \tilde{B}_{\text{gas}} < 1$, respectively. We see that we have a significant number of gas halos that have major axis ratios smaller than 1. Looking at the ratios \tilde{B} and \tilde{C} , we see that the gas halos with a ratio \tilde{A} smaller than unity can have a ratio \tilde{B} greater than 1, meaning that the gas particles are less extended than their DM counterpart along this axis. However, this does not represent the majority of halos. A significant proportion of halos (71%) that have a ratio on $\tilde{A} < 1$ also have a ratio on $\tilde{B} < 1$ and $\tilde{C} < 1$.

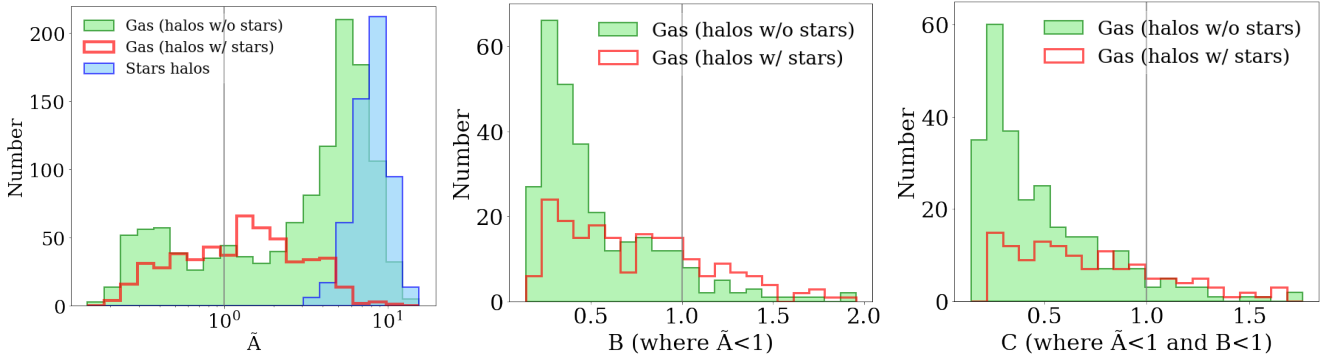


Fig. 11. Axis ratios \tilde{A} , \tilde{B} and \tilde{C} . Each panel corresponds to a sub-distribution. *Left panel:* $\tilde{A}_{\text{gas}} = a_{\text{DM}}/a_{\text{gas}}$ (green and red bins) and $\tilde{A}_{\text{stars}} = a_{\text{DM}}/a_{\text{stars}}$ (blue bins) axis ratios. Values below 1 correspond to a semi-axis of gas (stars) larger than their DM equivalent. *Middle panel:* \tilde{B} distribution with only the $\tilde{A} < 1$ halos. *Bottom panel:* \tilde{C} distribution with only halos with $\tilde{A} < 1$ and $\tilde{B} < 1$.

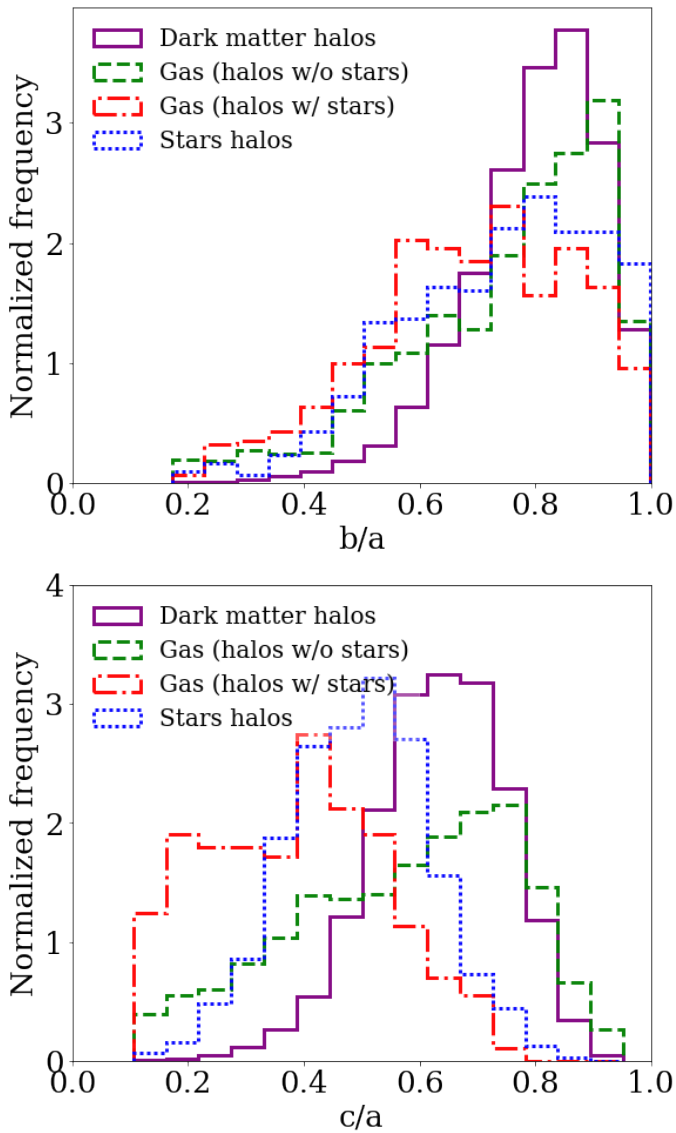


Fig. 12. Distribution of b/a (top panel) and c/a (bottom panel) axis ratios according to the type of particle contained in the RefL0025N0376 simulation at $z = 0.5$. DM halos (17 480) are represented in pink. Gas halos are split in two groups: gas halos w/o stars (1440) in green and gas halos w/ stars (682) in red. Halos of stars (682) are represented in blue.

Table 1. Analysis of the 17 480 halos in the RefL0025N0376 simulation at $z = 0.5$.

DM	Gas (w/ stars)	Gas (w/o stars)	Stars	
b/a	0.82 ± 0.11	0.70 ± 0.19	0.80 ± 0.16	0.76 ± 0.19
c/a	0.64 ± 0.12	0.38 ± 0.20	0.60 ± 0.22	0.50 ± 0.12

Notes. Median and median-absolute deviations of axis ratios b/a and c/a are provided according to the type of particle contained in the halos: DM (17 480 halos), gas (1440 halos w/o stars, 682 w/ stars), and stars (682 halos).

5.3. Shape of halos

In Fig. 12, we present the distribution of the axis ratios b/a and c/a of the 17 480 halos from the RefL0025N0376 simulation according to the types of particles they contain (DM, gas, or stars). Table 1 summarises the median characteristics of the distributions of axis ratios according to the particle types.

The EAGLE halos at $z = 0.5$ are described by a median axis ratio: $b/a \sim 0.8$. The sphericity peaks around $c/a \sim 0.6$, but we notice that the gas component presents an over-representation at small values of sphericity c/a corresponding to flatter ellipsoids of gas. These flatter halos coincidentally correspond to halos with $N_{\text{stars}} > 100$. DM halos have weak ellipticity with strong sphericity, as $b/a \sim 0.8$ and $c/a \sim 0.7$. Because the EAGLE simulation's prescription for star formation uses a Jeans-mass-limiting pressure floor, it is difficult for gas to cool into thin discs before creating stars, resulting in an artificial thickening of the stellar component and perhaps explaining why there are very few star halos with $c/a < 0.3$. Despite the fact that we used a different approach based on a PCA, the median values of the DM, gas, and star sphericity are largely consistent with those retrieved by Velliscig et al. (2015), Shao et al. (2016), and Hill et al. (2021) for halos in the same mass range. Similarly, the distribution of b/a and c/a axis ratios are consistent with those recovered by Tenneti et al. (2014) in the MassiveBlack-II and by Pillepich et al. (2019) in the IllustrisTNG simulations.

We now investigate the dependence of the sphericity on the mass of the halos, with a particular focus on gas halos. Figure 13 presents the distribution of the sphericity, c/a , for DM, gas (split in w/ stars and w/o stars), and star halos. We split star halos into two groups: low-mass halos ($M_{\text{tot}} < 10^{11} M_{\odot}$) and high-mass

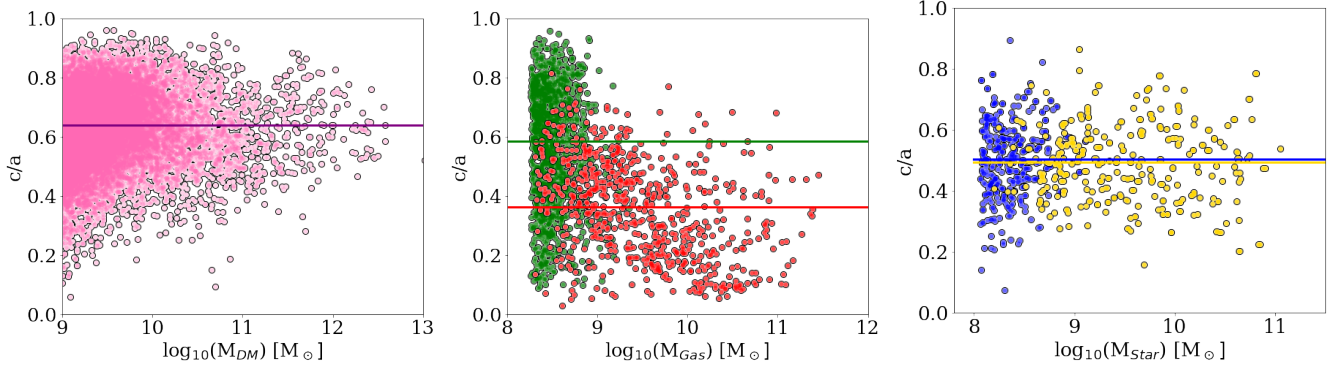


Fig. 13. Analysis of 17 480 halos at $z = 0.5$ of RefL0025N0376 EAGLE simulation. The distribution of the sphericity, c/a , is shown as a function of the total mass of the halos for DM (left panel), gas (middle panel), and stars (right panel). Gas halos are split in two groups: gas halos w/o stars (1440) are in green and gas halos w/ stars (682) are in red. Star halos are split into two groups of approximately equal size: low-mass halos ($M_{\text{tot}} < 10^{11} M_{\odot}$), depicted by blue dots, and high-mass halos ($M_{\text{tot}} > 10^{11} M_{\odot}$), depicted by yellow dots. Colour lines represent the median.

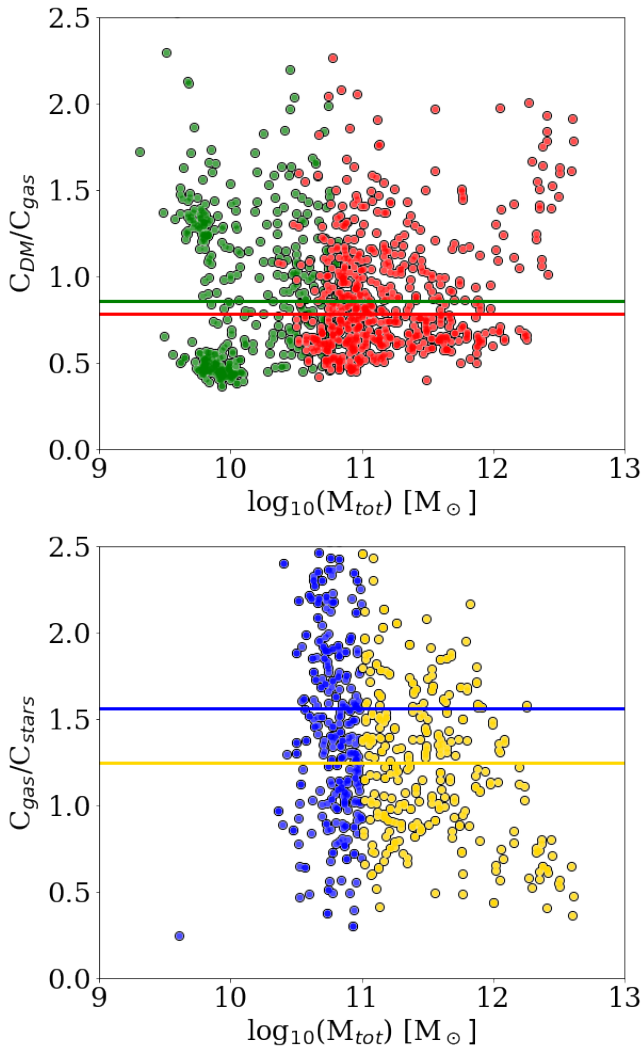


Fig. 14. DM-to-gas (top panel) and gas-to-star halo (bottom panel) concentration ratios as a function of their total mass for 1030 halos in the RefL0025N0376 simulation. Halos are composed of a minimum of 200 particles. Gas halos are split in two groups: gas halos w/o stars (419) in green and gas halos w/ stars (611) in red. Star halos are split into two groups of approximately equal size: low-mass halos ($M_{\text{tot}} < 10^{11} M_{\odot}$), depicted by blue dots, and high-mass halos ($M_{\text{tot}} > 10^{11} M_{\odot}$), depicted by yellow dots. Colour lines represent the median.

halos ($M_{\text{tot}} > 10^{11} M_{\odot}$). This threshold was chosen so as to have 50% of the star halos in each group.

There is a clear dependency on the sphericity of the gas halos with the mass of the halo: the less massive halos tend to have a more spherical distribution of gas with median value of $c/a \sim 0.6$, whereas for high-mass halos ($M_{\text{tot}} > 10^{11} M_{\odot}$), the sphericity has a median value of $c/a \sim 0.4$. Star halos do not follow the same behaviour since we do not observe any influence of mass on the flattening. The lack of massive star halos does not allow us to draw a definitive conclusion. The DM sphericity is totally insensitive to the mass of the halos.

5.4. Mass distribution of halos

In addition to the shape and orientation parameters, it is also interesting to investigate the mass distribution in the halos, which gives insight into their internal structure. For this we used the concentration parameter C from Eq. (1), which relies on the mass growth curve. Figure 14 presents the ratio of the concentration parameters between DM, gas, and star halos as a function of their total mass. In order to avoid poorly constrained concentration estimates, we remove all halos with fewer than 200 particles.

We observe in the top panel of Fig. 14 a concentration ratio $C_{\text{DM}}/C_{\text{gas}}$ slightly smaller than 1 for halos with and without a sufficient number of stars. This shows that the gas of these halos is slightly more concentrated than the DM.

Regarding the concentration of star particles compared to gas particles, we observe that most of the star halos (69%) are less concentrated than gas halos at redshift $z = 0.5$. Looking at the median of the concentration ratios for the two different mass ranges, we see that the more massive stellar halos (in yellow) are more concentrated than the less massive ones (in blue). It is important to remember here that the concentration does not provide a measure of the extent of the halos. Therefore, although the gas may be more concentrated than the stars, the latter are effectively formed in the core of the gas halos as shown in Fig. 15. In this figure, we show the distribution of the ratio between the half-mass radius of DM and gas halos (top panel) and gas and star halos (bottom panel), respectively.

The distribution of the ratio of this effective radii according to the type of particles shows that the star particles are well concentrated in the core of the halos. There is no stellar halo that has an effective half-mass radius larger than its corresponding gas halo. The EAGLE simulation therefore contains gas halos that are more diffuse than their DM counterparts.

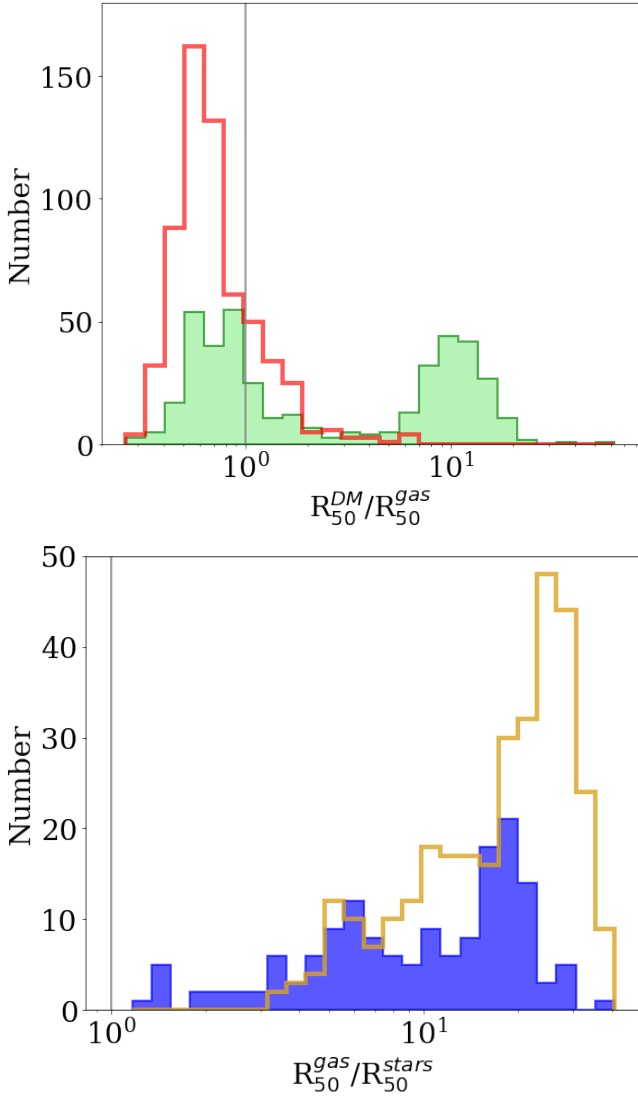


Fig. 15. Distribution of the ratio between the half-mass radius of DM and gas halos (*top panel*) and gas and star halos (*bottom panel*). Halos are composed of a minimum number of 200 particles. Colour scheme is similar to that of Figs. 13 and 14. The constant lines $R_{50}^{\text{DM}}/R_{50}^{\text{gas}} = 1$ and $R_{50}^{\text{gas}}/R_{50}^{\text{stars}} = 1$ are also plotted as references.

5.5. Properties of the projected halos

The work presented in previous sections investigates the three dimensional shape and mass distribution of halos from the RefL0025N0376 ($z = 0.5$) simulation. Instead, common analyses of galaxy morphology using imaging surveys usually derive parameters characterising the properties of projected halos onto the sky plane. Here, we analyse the properties of the EAGLE halos when projected onto the plane of the sky, allowing a comparison with ground-based and space-based imaging surveys and giving a measurement of the gravitational lensing effect that gives, in turn, constraints on the matter distribution along the line of sight. For commodity and under the reasonable assumption that halos are uniformly oriented, we use the coordinates of the particles in the XY , YZ , and XZ planes as the projections of the halos in three directions. Figure 16 presents the axis ratio (b_p/a_p) of the projected halos from the RefL0025N0376 ($z = 0.5$) EAGLE simulation. Table 2 summarises the median character-

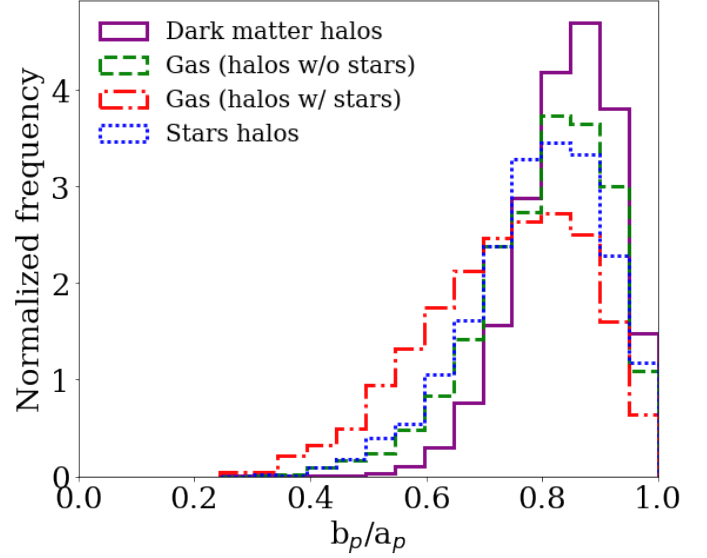


Fig. 16. Unit-area distribution of projected axis ratios of 61 398 projections from the RefL0025N0376 simulation at $z = 0.5$ according to the particle type: DM (61 398 halos), gas (4362 halos w/o stars, 2058 w/ stars), stars (2172 halos).

Table 2. Median and median-absolute deviations of the projected axes ratios b_p/a_p , according to the type of particle: DM (61 398 halos), gas (4362 halos w/o stars, 2058 w/ stars), and stars (2172 halos).

DM	Gas (w/ stars)	Gas (w/o stars)	Stars	
b_p/a_p	0.85 ± 0.06	0.75 ± 0.10	0.82 ± 0.08	0.80 ± 0.07

istics of the projected distributions of axis ratios according to the particle type.

The overall distribution of the axis ratios from DM has a median value of $(b_p/a_p)_{\text{DM}} = 0.85 \pm 0.06$ corresponding to quasi-spherical halos. Gas and stars have similar properties with a little more elongated projected profiles with axis ratio values of $(b_p/a_p)_{\text{gas}} = 0.75 \pm 0.10$ for gas halos with more than 100 star particles and $(b_p/a_p)_{\text{gas}} = 0.82 \pm 0.08$ for gas halos with fewer than 100 star particles. Star halos have a median axis ratio of $(b_p/a_p)_{\text{stars}} = 0.80 \pm 0.07$. The three normalised distributions agree at a 1σ level.

The median value of the axis ratios for stars can be compared to the morphology derived in imaging surveys such as the SDSS DR16 (Ahumada et al. 2020) or derived from the space-based *Gaia* observations (Ducourant et al. 2022; Gaia Collaboration 2022). We analysed the distribution of the axis ratios of 390 612 selected galaxies from the SDSS DR16 for which the parameters of a de Vaucouleurs profile are provided. The mean value of their axis ratios is $(b/a)_{\text{SDSS}} = 0.65 \pm 0.21$. The SDSS estimates describe stellar content of galaxies with more elongated shape than what is observed in the EAGLE simulation at $z = 0.5$. The same sample of galaxies, but observed by *Gaia* with a de Vaucouleurs profile, has the following axis ratio values: $(b/a)_{\text{Gaia}} \sim 0.73 \pm 0.14$. Depending on the survey ground-based or space-based the mean values of an axis ratio are somewhat different, but they are in good agreement one with the other and with the EAGLE value, within 1σ .

Figure 17 presents the ratio of concentration of the projected DM and gas halos (top panel) and gas and star halos (bottom panel) as a function of the total mass.

In it, we see a $C_{\text{DM}}/C_{\text{gas}}$ concentration ratio slightly smaller than 1 between DM and gas. This shows that the gas is slightly more concentrated than the DM. This result is consistent with the observations from Sect. 5.4. Star halos have a lower concentration than gas. This behaviour is the opposite of what is observed for the three dimensional halos.

6. Dependence on the redshift

The RefL0025N0376 simulation contains 29 snapshots with different redshifts varying from $z = 15$ down to $z = 0$. The analysis of the number of halos, as well as their morphology and mass distribution, at the different redshifts allows an insight into the cosmic evolution. In the rest of this study, we chose to represent the evolution of the shape of the halos as a function of the look-back (LB) time. This change of scale between redshift and LB time allows us to have a linear time scale in the EAGLE snapshots and to focus more on the near Universe. We used the same cosmological parameters as the EAGLE simulation for the scale conversion, as provided in Sect. 2.

6.1. Frequencies

The number of halos is not constant with redshift and the presence of stars is linked with the stage of evolution of the halos. In a Λ -CDM universe, high-mass DM halos are formed more recently than their low-mass counterparts (Qu et al. 2017). This is confirmed in Fig. 18, which presents the number of halos at various redshifts and according to the particle type for the RefL0025N0376 simulation. The number of low-mass DM halos increases with time. The number peaks at $LB = 12.13$ Gyr ($z = 3.53$) and then decreases. The rise illustrates that the simulation becomes more and more contrasted (the FoF algorithm detects more and more halos) as structures emerge. From $LB = 10.15$ Gyr ($z = 1.74$), we mainly observe halo mergers in hierarchical mass growth.

We observe a similar behaviour for gas halos at $LB = 12.76$ Gyr ($z = 5.04$), where the number of low-mass halos peaks and then decreases. Low-mass gas halos that formed at high redshifts do not exist in the present day and merged into more massive halos (Qu et al. 2017).

6.2. Morphology

Figure 19 presents the evolution of the half mass radii of the star halos from the RefL0025N0376 simulation as a function of the LB time for two mass ranges.

One first observes that halos more massive than $10^{10} M_{\odot}$ appear from $LB = 13.17$ Gyr ($z = 7.05$) and more massive than $10^{11} M_{\odot}$ from $LB = 12.88$ Gyr ($z = 5.49$) corresponding to the respective time scales necessary for significant fusions of halos producing sufficient stars. The size of the star halos grows with time whatever their mass range is. The growth rate is similar. The growth of halo size results from the merging along time of halos to produce more massive and hence larger halos and by the formation of new stars.

In order to investigate the mass dependence on the distribution of the axis ratio, we plot in Figs. 20 and 21, the median of the axis ratios (b/a) and (c/a) for the mass bins $10^9 \leq M_{\text{tot}} < 10^{10} M_{\odot}$, $10^{10} \leq M_{\text{tot}} < 10^{11} M_{\odot}$, and $10^{11} \leq M_{\text{tot}} < 10^{12} M_{\odot}$.

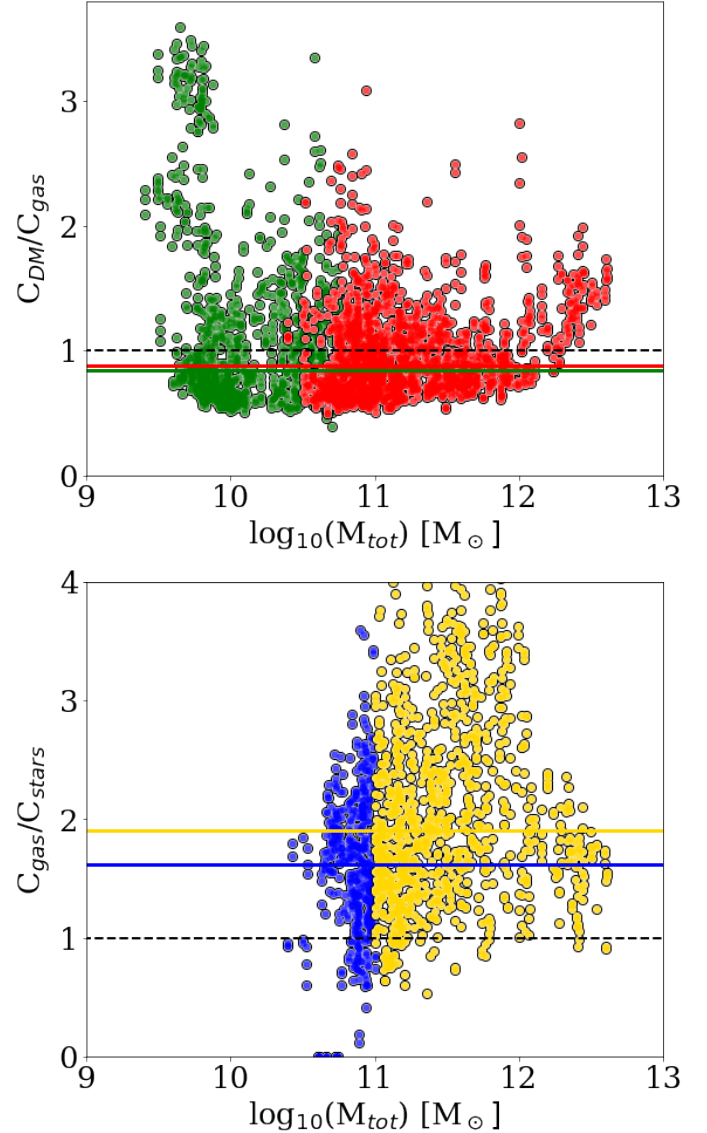


Fig. 17. Concentration ratio of the projected halos at $z = 0.5$ from the RefL0025N0376 simulation. The halos have a minimum of 200 particles. The ratios $C_{\text{DM}}/C_{\text{gas}}$ (top panel) and $C_{\text{gas}}/C_{\text{stars}}$ (bottom panel) are expressed as a function of the total mass of the halo. Gas halos are split in two groups, gas halos w/o stars (1290) in green and gas halos w/ stars (1875) in red. Star halos are split into two groups of approximately equal size: low mass halos ($M_{\text{tot}} < 10^{11} M_{\odot}$), depicted by blue dots, and high mass halos ($M_{\text{tot}} > 10^{11} M_{\odot}$), depicted by yellow dots. Colour lines represent the median. The constant dashed lines $C_{\text{DM}}/C_{\text{gas}} = 1$ and $R_{\text{gas}}/R_{\text{stars}} = 1$ are also plotted as references.

The shape of the halos when all particles are considered is dominated by the DM component if all types of particles are considered. In fact, the shape of the DM component is nearly identical to the total mass distribution, so we did not consider the case of all particles. The global tendency of the axis ratios is to become more and more spherical with time, whatever the mass and the particle type are. These plots can be compared against Allgood et al. (2006) and Tenneti et al. (2014). Our results agree with theirs for DM halos in the sense that the average axis ratios, b/a and c/a , increase as we get closer to the present time ($LB = 0$). This effect is not due to baryon physics since it is also found in DM-only simulations (Bryan et al. 2013; Tenneti et al. 2014). The halos are built up by accreting matter from the cosmic web

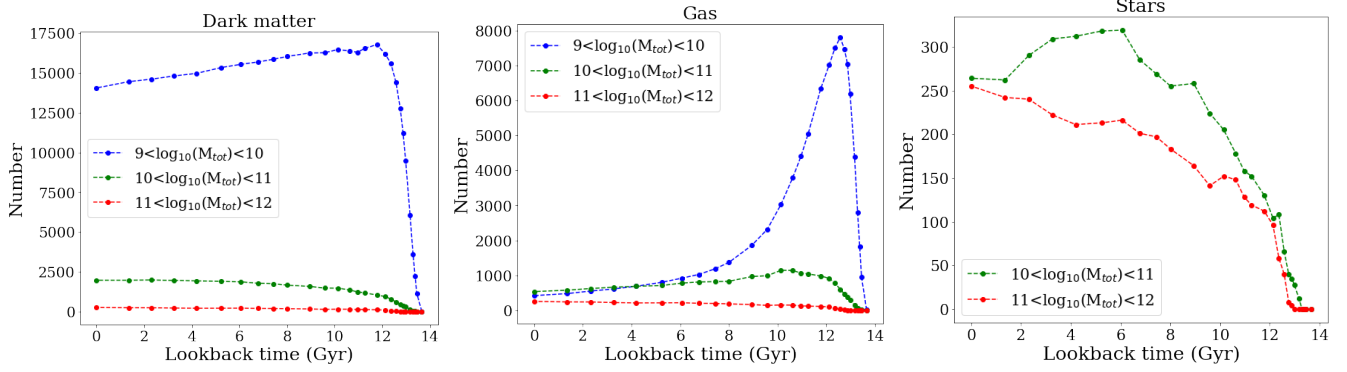


Fig. 18. Variation of number of halos with at least 100 particles from the RefL0025N0376 simulation as a function of redshift. Halos are grouped according to the type of their constituent particles and split in various ranges of solar masses.

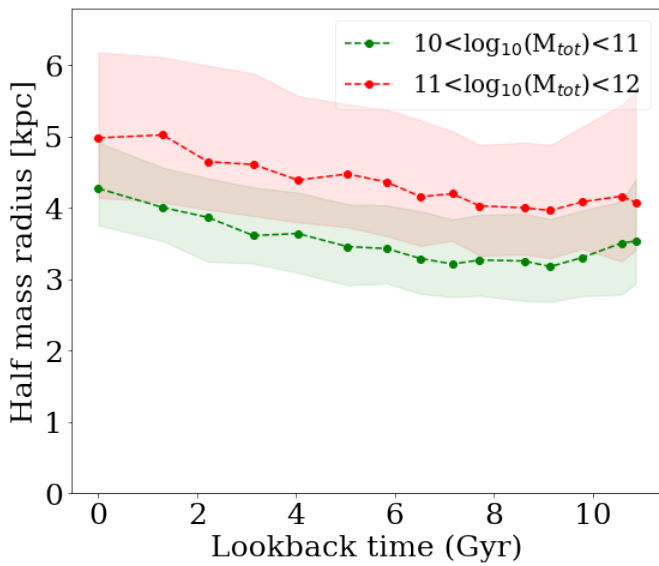


Fig. 19. Median of the half mass major-axis radius of the star halos as a function of the LB time for various solar mass ranges. The solid lines indicate the median values, whilst the shading denotes the 25th–75th quartiles.

filaments, resulting in an initial anisotropy, with a progressive virialisation that leads to a final ellipsoid for these non-colliding particles that approach a spherical halo.

For the stellar matter, we see that the average axis ratios generally follow the DM ratios trends. Star particles are, as DM particles, non-colliding. They behave as massless particles in the gravitational potential defined by the DM-dominated halo. The only exception concerns the very massive gas halos ($M_{\odot} > 10^{11}$), for which an inverse tendency for (c/a) is observed. These halos' discs tend to become more and more flattened with time. For gas, low-mass halos ($M < 10^{10} M_{\odot}$) have larger axis ratios than more massive halos. This is correlated with the absence of stars in the low-mass halos.

Compared to the previous b/a ratio, the c/a axis ratio is slightly smaller than b/a . This results in rather flattened halos. The c/a axis ratio of the gas shows a strong dependence on the mass. The greater the mass, the more the axis ratio tends towards small values of c/a . This shows that gas halos tend to flatten out as they gain mass. The minimum mass observed for a halo of stars to form ($N_{\text{stars}} > 100$) is $M \sim 10^{11} M_{\odot}$, revealing that a necessary mass level is required to form enough stars.

6.3. Concentration as function of redshift

As in Sect. 5.5, we remind the reader that we did not consider halos with fewer than 200 particles for concentration measurements. Figure 22 presents the evolution of the median of the concentration as a function of the LB time.

The concentration of DM and stars increases continuously during the simulation. The DM concentration increases more strongly at the beginning of the simulation ($10 < T_{\text{lookback}} < 13$ Gyr) than at more recent times ($0 < T_{\text{lookback}} < 8$ Gyr). The DM concentration increases as halos form and as these new halos accumulate mass by accreting particles into their gravitational pool. Then, the gas particles follow the gravitational potential and fall in the core of the DM halos. This is observed through a significant increase of the gas concentration in the first gigayear of the simulation up to an LB time of ~ 8 Gyr, where the concentration peaks at around 4.5 for halos that have already formed some stars. This reflects the fact that high-concentration gas halos slowly lose their concentration over time. This can have two possible origins. The first is that the concentrated gas in the heart of the halos is consumed to fuel the formation of new stars. Then, the gas is dispersed by these new stars formed in the heart of the halos. Gas halos that do not form any star particles are small halos ($M_{\text{tot}} < 10^9 M_{\odot}$) and have a very low decreasing concentration as no star consumes the gas.

6.4. Projected axis ratio

In this section, we repeat the analysis performed for the three dimensional halos in Sects. 6.1 and 6.2 for the two dimensional projected halos. Figure 23 presents the projected axis ratio as function of the redshift for each type of particle and by mass range.

We observe that the halos are more spherical on average for low values of the LB time. DM has a constantly increasing distribution. The greater the mass, the greater the axis ratio of the halo. The general evolution of stars is identical to that of DM. For star particles, there is an inversion in the evolution of the most massive halos. For the most massive halos, $M_{\text{tot}} > 10^{11} M_{\odot}$, the axis ratio of the stars tends to be lower. For gas halos of $M > 10^{11} M_{\odot}$, the projected axis ratio has a slight tendency to decrease.

6.5. Projected concentration

Figure 24 presents the evolution of the median of the projected concentrations as a function of the LB time. The behaviour of the DM particles is identical to that already observed on the

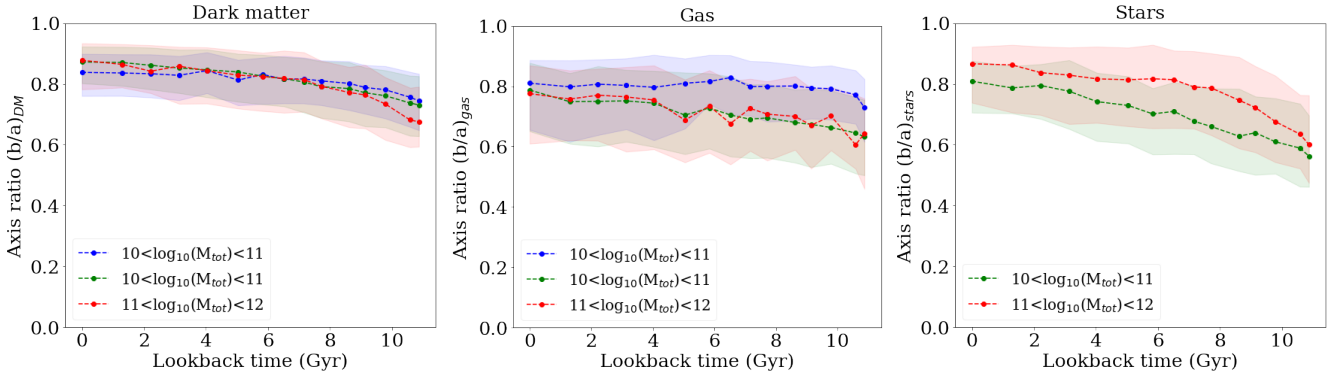


Fig. 20. Dependence of the median of axis ratios (b/a) on the LB time for halos in the RefL0025N0376 simulation according to various solar mass ranges and particle types.

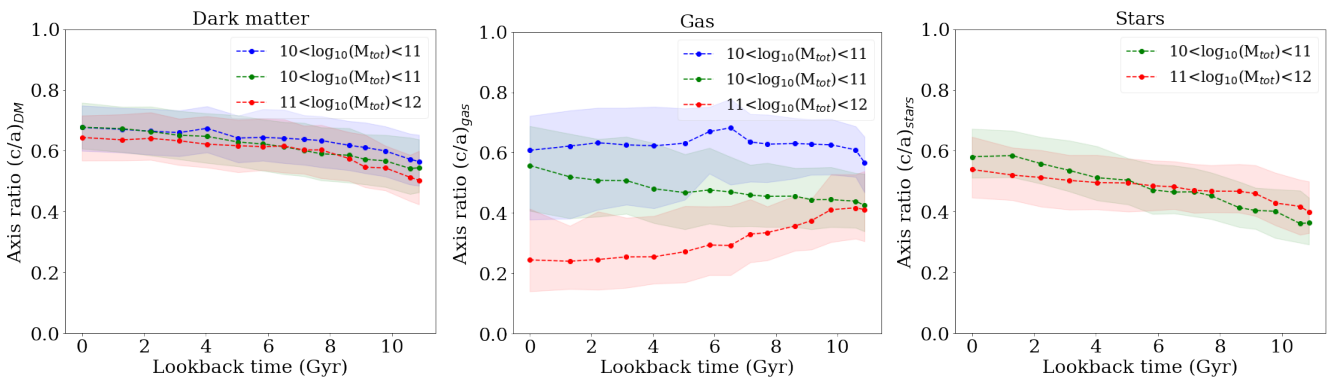


Fig. 21. Dependence of the median of axis ratios (c/a) (sphericity) on the LB time for halos in the RefL0025N0376 simulation according to various solar mass ranges and particle types.

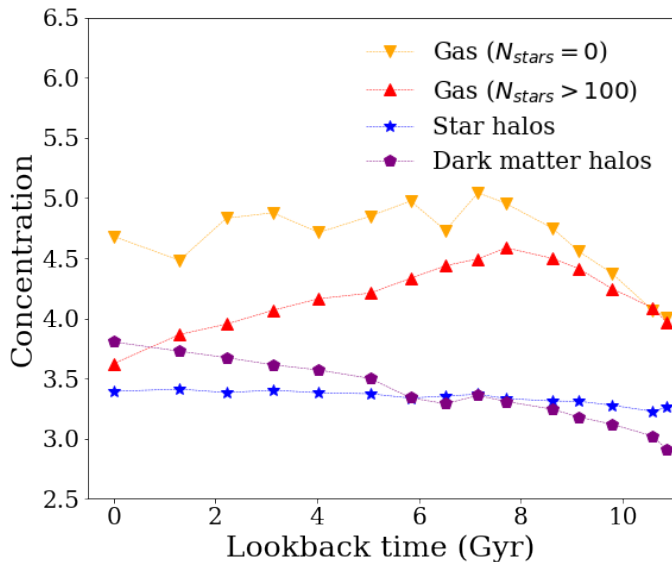


Fig. 22. Concentration as a function of LB time for RefL0025N0376 simulation. The solid lines indicate the median values. Gas halos are split in two groups, gas halos without any stars in yellow and gas halos with more than 100 star particles in red.

concentration measured along the main axes. The concentration of the DM halos increases continuously from three to four. The gas particles have two different behaviours that follow the previously observed trends. The gas halos containing stars see

their projected concentration increase at the beginning of the simulation until they reach a peak around 8–9 Gyr. From this time, the projected concentration of the gas particles decreases continuously from a maximum of 4.5–3.2 at present time. This behaviour indicates the presence of star formation in the gas halos. Stars form in the centre of the halos by consuming some of the gas and then diffusing it as they go. At the beginning of the simulation, the stars, which form mainly in the heart of halos where the gas concentration is sufficient, have a concentration close to that of the gas from which they originate. Then, as more stars form, the projected concentration decreases as these particles diffuse into the halos.

7. Conclusion

The EAGLE project represents a judicious test bed to obtain statistics on the mass distribution of halos, as the suite of simulations were calibrated to reproduce the galaxy stellar mass function and the size-mass relation of late-type galaxies. In this study, we used the small volume RefL0025N0376 simulation to directly measure the shapes and alignments of DM, gas, and star halos. One motivation is to obtain precise knowledge of the projected mass distribution of the deflectors in strong lensing. The shape parameters have been calculated for the 28 available redshift snapshots of the simulation with a special focus at $z = 0.5$. The method we developed analyses the three dimensional particle distribution within an iteratively determined radius excluding large-scale structures, which usually perturb the processing. This radius is set so that $\sim 80\%$ of the mass of the halos is considered.

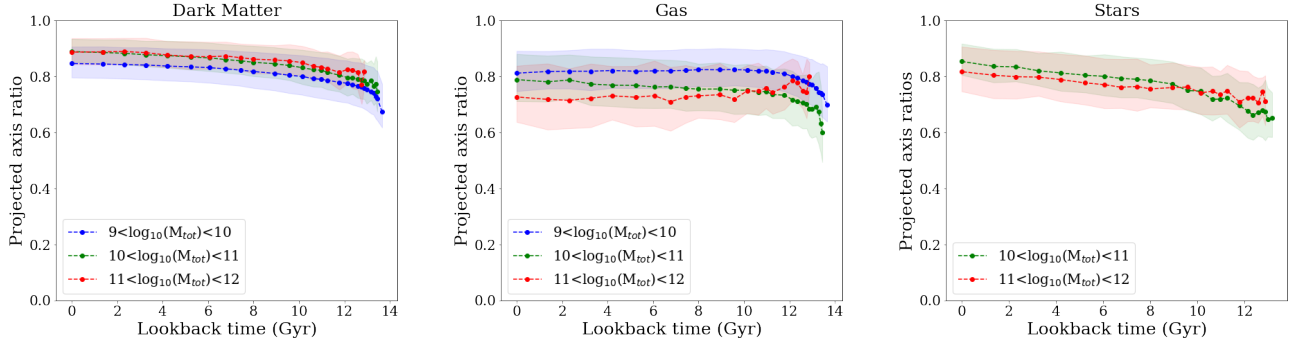


Fig. 23. Dependence of median of projected axis ratios (b_p/a_p) on the LB time for halos in RefL0025N0376 simulation according to various solar mass ranges and particle types.

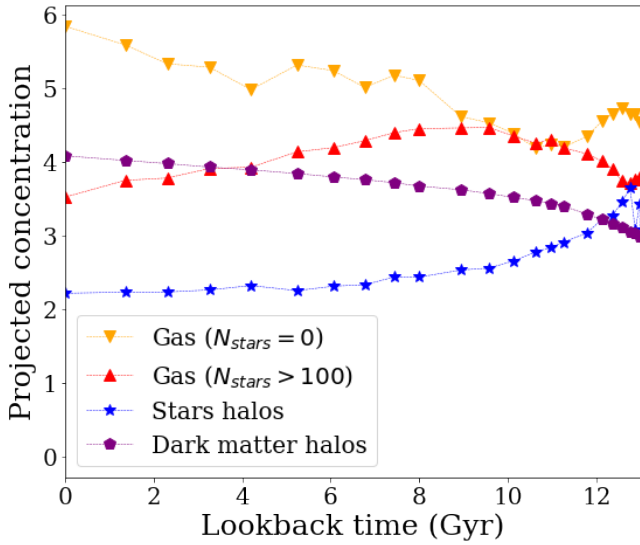


Fig. 24. Projected concentration as a function of the LB time for the RefL0025N0376 simulation. The lines indicate the median values. Gas halos are split in two groups, gas halos without any stars in yellow and gas halos with more than 100 star particles in red.

At $z = 0.5$, the minor axis of the gas halos preferentially aligns with the minor axis of the DM halos. The mean difference in alignment between the minor axes of the gas and of the DM distributions varies from ~ 40 deg to ~ 20 deg for masses of $10^{10} < M_{\text{tot}} < 10^{12} M_{\odot}$. The differences in alignment between the gas and stellar components are larger than those between DM and gas halos. These results are in good agreement with previous studies (Velliscig et al. 2015; Thob et al. 2019; Hill et al. 2021). Since the number of massive halos is small in the simulation we studied, further analyses on the larger EAGLE simulations have already been started.

Concerning the axis ratio of the halos, the b/a for all particles have a median value of 0.81 ± 0.07 , and the sphericity c/a peaks at a median value of 0.64 ± 0.08 . The distribution of the axis ratios of the various components behave similarly, except for gas halos that already form stars and are much more flattened. These halos are more massive than the gas halos that did not form stars.

We also measured the distribution of the projected axis ratio b_p/a_p when the halos are projected on 2D sky maps. Whatever the component under consideration is, the distribution of b_p/a_p has a median value of 0.86 ± 0.05 at $z = 0.5$. As observed for the three dimensional case, this axis ratio has smaller b_p/a_p values for gas halos w/ stars.

We find that gas halos that form stars have larger concentration values than gas halos not forming stars. When comparing star halos to their gas halo counterparts, we notice that most of the star halos (69%) are less concentrated and that this concentration depends on the total mass of the halos.

The value of the concentration parameter of DM and star halos increases continuously from redshift $z = 15$ to $z = 0$. Gas halos that do not form any star appear to be light halos ($M_{\text{tot}} < 10^9 M_{\odot}$). Their concentration increases continuously with time as no star consumes the gas. We find that the total matter distribution in halos is more spherical at higher halo masses and redshifts. The same qualitative trends hold for the star and gas distributions of these halos.

The halos show different rates of variation for each shape parameter depending on the component considered. Our analysis confirms recent observations on the shape and formation of halos. Each component evolves in a particular way with respect to the others, depending on whether the matter is collisional or not. The evolution of gas and star particles are related since stars are derived from the gas. These components are evolving rapidly, forming a diversity of structures, as observed in the current large surveys. On the other hand, DM halos evolve slowly and tend to retain traces of the initial conditions, resulting in more spherical shapes.

Acknowledgements. French CNRS INSU PNgam SNOGaia CS-OASU This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 787886).

References

- Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, *ApJS*, 249, 3
 Allgood, B., Flores, R. A., Primack, J. R., et al. 2006, *MNRAS*, 367, 1781
 Barnabè, M., Dutton, A. A., Marshall, P. J., et al. 2012, *MNRAS*, 423, 1073
 Bett, P., Eke, V., Frenk, C. S., Jenkins, A., & Okamoto, T. 2010, *MNRAS*, 404, 1137
 Blumenthal, G. R., Faber, S. M., Primack, J. R., & Rees, M. J. 1984, *Nature*, 311, 517
 Bond, J. R., Kofman, L., & Pogosyan, D. 1996, *Nature*, 380, 603
 Booth, C. M., & Schaye, J. 2009, *MNRAS*, 398, 53
 Bryan, S. E., Kay, S. T., Duffy, A. R., et al. 2013, *MNRAS*, 429, 3316
 Collier, A., Shlosman, I., & Heller, C. 2018, *MNRAS*, 476, 1331
 Collier, A., Shlosman, I., & Heller, C. 2019, *MNRAS*, 488, 5788
 Conselice, C. J. 2003, *ApJS*, 147, 1
 Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, *MNRAS*, 450, 1937
 Dalla Vecchia, C., & Schaye, J. 2012, *MNRAS*, 426, 140
 Dark Energy Survey Collaboration (Abbott, T., et al.) 2016, *MNRAS*, 460, 1270
 Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, *ApJ*, 292, 371
 Delchambre, L., Krone-Martins, A., Wertz, O., et al. 2019, *A&A*, 622, A165
 Dolag, K., Borgani, S., Murante, G., & Springel, V. 2009, *MNRAS*, 399, 497

- Ducourant, C., Krone-Martins, A., Galluccio, L., et al. 2022, ArXiv e-prints [arXiv:2206.14491]
- Faucher-Giguère, C.-A., Lidz, A., & Hernquist, L. 2008, *Science*, 319, 52
- Gaia Collaboration (Bailer-Jones, C. A. L., et al.) 2022, A&A, in press, <https://doi.org/10.1051/0004-6361/202243232>
- Gomer, M. R., & Williams, L. L. R. 2021, *MNRAS*, 504, 1340
- Hill, A. D., Crain, R. A., Kwan, J., & McCarthy, I. G. 2021, *MNRAS*, 505, 65
- Khandai, N., Di Matteo, T., Croft, R., et al. 2015, *MNRAS*, 450, 1349
- Kirk, D., Brown, M. L., Hoekstra, H., et al. 2015, *Space Sci. Rev.*, 193, 139
- Krone-Martins, A., Delchambre, L., Wertz, O., et al. 2018, *A&A*, 616, L11
- Kumar, A., Das, M., & Kataria, S. K. 2022, *MNRAS*, 509, 1262
- Laureijs, R., Gondoin, P., Duvet, L., et al. 2012, *Proc. SPIE*, 8442, 329
- LSST Science Collaboration (Abell, P. A., et al.) 2009, ArXiv e-prints [arXiv:0912.0201]
- Marigo, P. 2001, *A&A*, 370, 194
- McAlpine, S., Helly, J., Schaller, M., et al. 2016, *Astron. Comput.*, 15, 72
- Natarajan, P., & Refregier, A. 2000, *ApJ*, 538, L113
- Pillepich, A., Nelson, D., Springel, V., et al. 2019, *MNRAS*, 490, 3196
- Planck Collaboration XVI 2014, *A&A*, 571, A16
- Portinari, L., Chiosi, C., & Bressan, A. 1998, *A&A*, 334, 505
- Qu, Y., Helly, J. C., Bower, R. G., et al. 2017, *MNRAS*, 464, 1659
- Remus, R.-S., Dolag, K., Bachmann, L. K., et al. 2015, *IAU Symp.*, 309, 145
- Rosas-Guevara, Y. M., Bower, R. G., Schaye, J., et al. 2015, *MNRAS*, 454, 1038
- Schaye, J. 2004, *ApJ*, 609, 667
- Schaye, J., & Dalla Vecchia, C. 2008, *MNRAS*, 383, 1210
- Schaye, J., Crain, R. A., Bower, R. G., et al. 2014, *MNRAS*, 446, 521
- Shao, S., Cautun, M., Frenk, C. S., et al. 2016, *MNRAS*, 460, 3772
- Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints [arXiv:1503.03757]
- Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, *MNRAS*, 328, 726
- Springel, V., Di Matteo, T., & Hernquist, L. 2005, *MNRAS*, 361, 776
- Stern, D., Djorgovski, S. G., Krone-Martins, A., et al. 2021, *ApJ*, 921, 42
- Tegmark, M., Blanton, M. R., Strauss, M. A., et al. 2004, *ApJ*, 606, 702
- Tenneti, A., Mandelbaum, R., Di Matteo, T., Feng, Y., & Khandai, N. 2014, *MNRAS*, 441, 470
- The EAGLE Team 2017, ArXiv e-prints [arXiv:1706.09899]
- Thob, A. C. R., Crain, R. A., McCarthy, I. G., et al. 2019, *MNRAS*, 485, 972
- van Uitert, E., Hoekstra, H., Schrabback, T., et al. 2012, *A&A*, 545, A71
- Velliscig, M., Cacciato, M., Schaye, J., et al. 2015, *MNRAS*, 453, 721
- Wang, Y., Yang, X., Mo, H. J., et al. 2008, *MNRAS*, 385, 1511
- White, S. D. M., & Rees, M. J. 1978, *MNRAS*, 183, 341
- Wiersma, R. P. C., Schaye, J., & Smith, B. D. 2009a, *MNRAS*, 393, 99
- Wiersma, R. P. C., Schaye, J., Theuns, T., Dalla Vecchia, C., & Tornatore, L. 2009b, *MNRAS*, 399, 574
- Witt, H. J. 1996, *ApJ*, 472, L1
- Woldesenbet, A. G., & Williams, L. L. R. 2015, *MNRAS*, 454, 862

Appendix A: Influence of the number of particles on shape characterisation

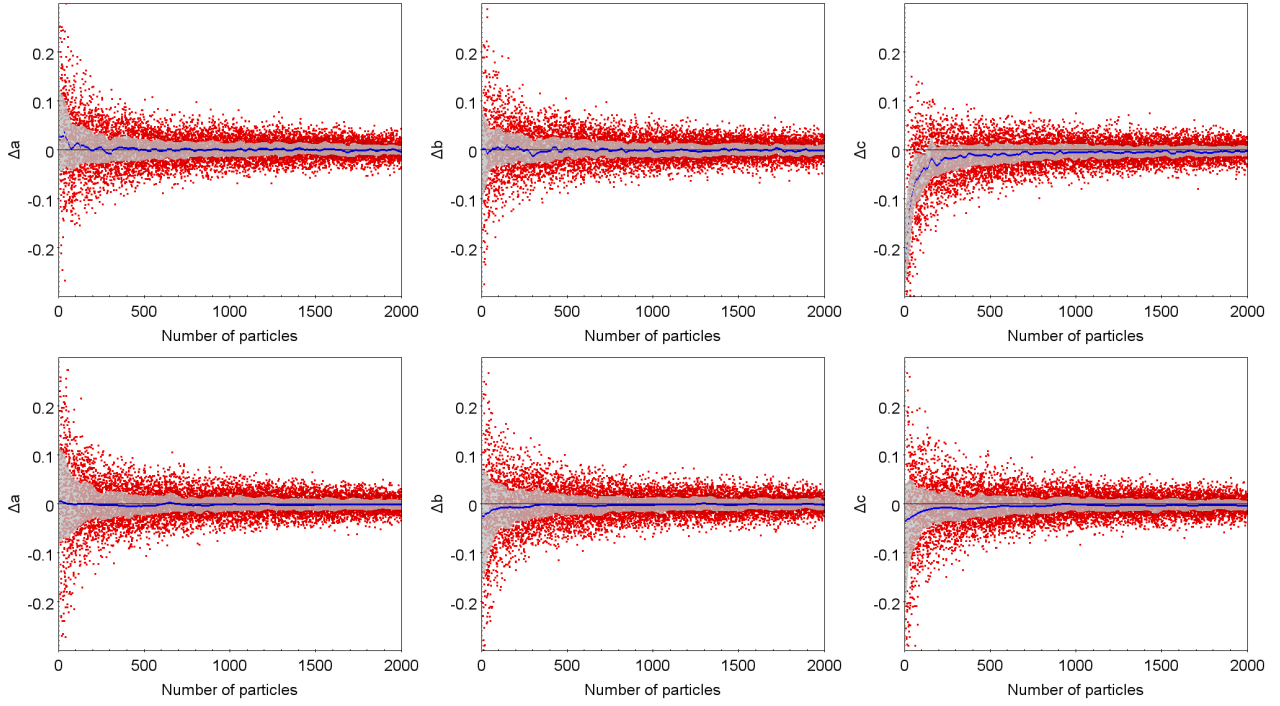


Fig. A.1. Analysis of 10 000 simulated halos. Relative difference (Δa , Δb , Δc) in the recovered semi-axes (a,b,c) as a function of the number of particles of (upper panel) quasi-spherical halos with $c_{sim}/a_{sim}=0.72$ and (lower panel) flattened halos with $c_{sim}/a_{sim}=0.25$. Blue lines indicate the median of the distributions and the grey area shows the quartiles.

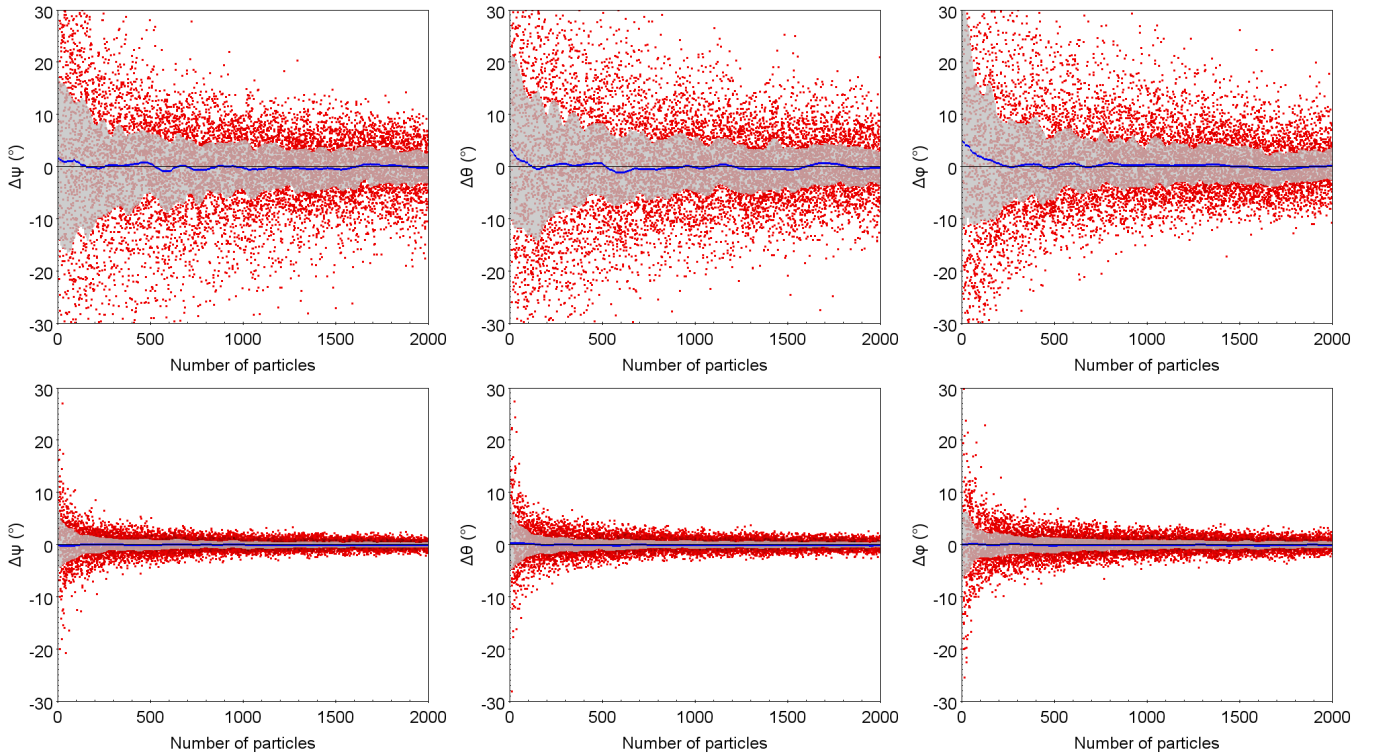


Fig. A.2. Analysis of 10 000 simulated halos. Difference between the recovered angles (ψ , θ , ϕ) and their true simulated values expressed in deg as a function of the number of particles. Blue lines indicate the median of the distributions and the grey area shows the quartiles. Upper panel: Quasi-spherical halos with $c_{sim}/a_{sim}=0.72$. Lower panel: Flattened halos with $c_{sim}/a_{sim}=0.25$.

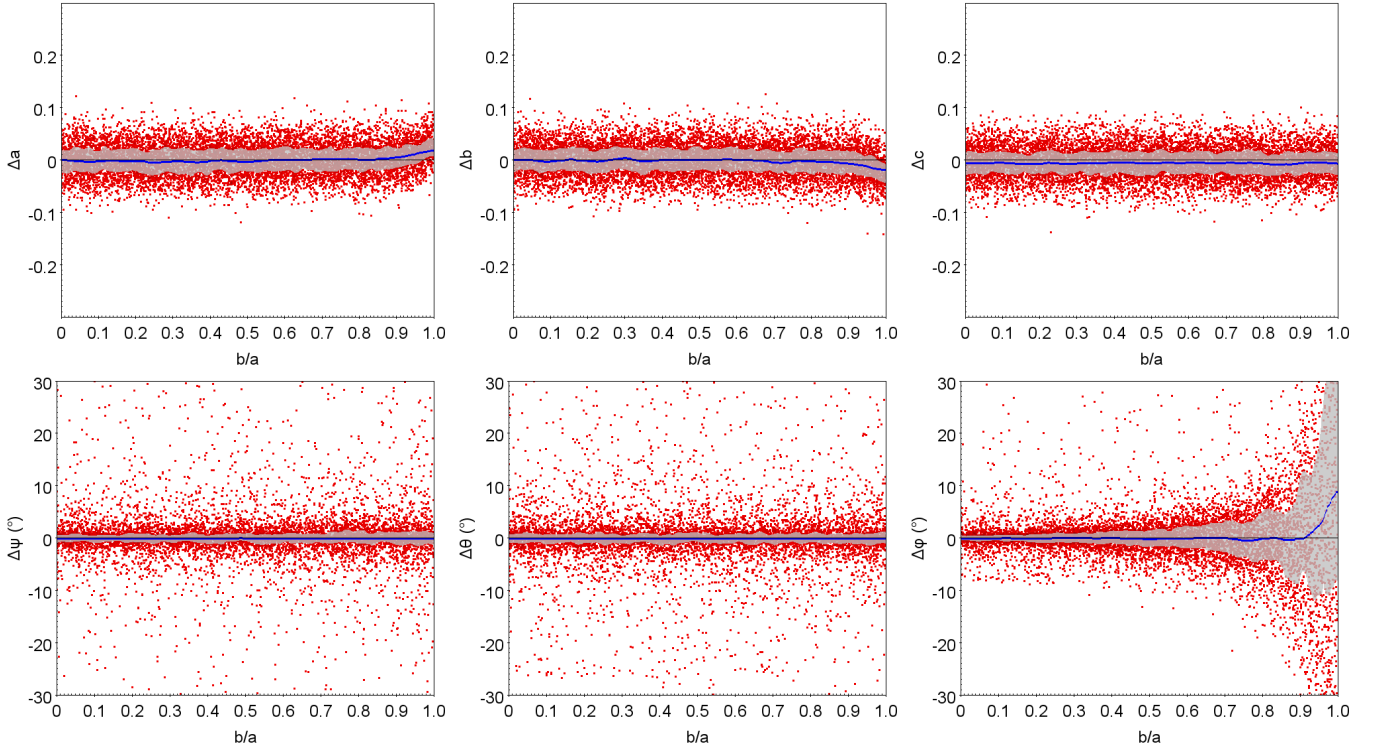


Fig. A.3. Analysis of 10 000 simulated halos with 500 particles. Upper panel: Relative differences (Δa , Δb , Δc) in the recovered semi-axes (a, b, c) as a function of the sphericity (c/a). Lower panel: Difference between the recovered angles (ψ , θ , ϕ) and their true simulated values expressed in deg as a function of the sphericity (c/a). Blue lines indicate the median of the distributions and the grey area shows the quartiles.

Bibliographie

- [1] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, and Marcel A. Agüeros et al. The seventh data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 182(2) :543, may 2009. doi : 10.1088/0067-0049/182/2/543. URL <https://dx.doi.org/10.1088/0067-0049/182/2/543>.
- [2] Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, Santiago Avila, Vladimir Avila-Reese, Carles Badenes, Christophe Balland, Kat Barger, Jorge K. Barrera-Ballesteros, Sarbani Basu, Julian Bautista, Rachael L. Beaton, Timothy C. Beers, B. Izamar T. Benavides, Chad F. Bender, Mariangela Bernardi, Matthew Bershad, Florian Beutler, Christian Moni Bidin, Jonathan Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael R. Blanton, Médéric Boquien, Jura Borissova, Jo Bovy, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Kevin Bundy, Martin Bureau, Adam Burgasser, Etienne Burtin, Mariana Cano-Díaz, Raffaella Capasso, Michele Cappellari, Ricardo Carrera, Solène Chabanier, William Chaplin, Michael Chapman, Brian Cherinka, Cristina Chiappini, Peter Doohyun Choi, S. Drew Chojnowski, Haeun Chung, Nicolas Clerc, Damien Coffey, Julia M. Comerford, Johan Comparat, Luiz da Costa, Marie-Claude Cousinou, Kevin Covey, Jeffrey D. Crane, Katia Cunha, Gabriele da Silva Ilha, Yu Sophia Dai, Sanna B. Damsted, Jeremy Darling, Jr. Davidson, James W., Roger Davies, Kyle Dawson, Nikhil De, Axel de la Macorra, Nathan De Lee, Anna Bárbara de Andrade Queiroz, Alice Deconto Machado, Sylvain de la Torre, Flavia Dell’Agli, Hélion du Mas des Bourboux, Aleksandar M. Diamond-Stanic, Sean Dillon, John Donor, Niv Drory, Chris Duckworth, Tom Dwelly, Garrett Ebelke, Sarah Eftekharzadeh, Arthur Davis Eigenbrot, Yvonne P. Elsworth, Mike Eracleous, Ghazaleh Erfanianfar, Stephanie Escoffier, Xiaohui Fan, Emily Farr, José G. Fernández-Trincado, Diane Feuillet, Alexis Finoguenov, Patricia Fofie, Amelia Fraser-McKelvie, Peter M. Frinchaboy, Sebastien Fromenteau, Hai Fu, Lluís Galbany, Rafael A. Garcia, D. A. García-Hernández, Luis Alberto Garma Oehmichen, Junqiang Ge, Marcio Antonio Geimba Maia, Doug Geisler, Joseph Gelfand, Julian Goddy,

Violeta Gonzalez-Perez, Kathleen Grabowski, Paul Green, Catherine J. Grier, Hong Guo, Julien Guy, Paul Harding, Sten Hasselquist, Adam James Hawken, Christian R. Hayes, Fred Hearty, S. Hekker, David W. Hogg, Jon A. Holtzman, Danny Horta, Jiamin Hou, Bau-Ching Hsieh, Daniel Huber, Jason A. S. Hunt, J. Ider Chitham, Julie Imig, Mariana Jaber, Camilo Eduardo Jimenez Angel, Jennifer A. Johnson, Amy M. Jones, Henrik Jönsson, Eric Jullo, Yerim Kim, Karen Kinemuchi, IV Kirkpatrick, Charles C., George W. Kite, Mark Klaene, Jean-Paul Kneib, Juna A. Kollmeier, Hui Kong, Marina Kounkel, Dhanesh Krishnarao, Ivan Lacerna, Ting-Wen Lan, Richard R. Lane, David R. Law, Jean-Marc Le Goff, Henry W. Leung, Hannah Lewis, Cheng Li, Jianhui Lian, Lihwai Lin, Dan Long, Penélope Longa-Peña, Britt Lundgren, Brad W. Lyke, J. Ted Mackereth, Chelsea L. MacLeod, Steven R. Majewski, Arturo Manchado, Claudia Maraston, Paul Martini, Thomas Masseron, Karen L. Masters, Savita Mathur, Richard M. McDermid, Andrea Merloni, Michael Merrifield, Szabolcs Mészáros, Andrea Miglio, Dante Minniti, Rebecca Minsley, Takamitsu Miyaji, Faizan Gohar Mohammad, Benoit Mosser, Eva-Maria Mueller, Demitri Muna, Andrea Muñoz-Gutiérrez, Adam D. Myers, Seshadri Nadathur, Preethi Nair, Kirpal Nandra, Janaina Correa do Nascimento, Rebecca Jean Nevin, Jeffrey A. Newman, David L. Nidever, Christian Nitschelm, Pasquier Noterdaeme, Julia E. O'Connell, Matthew D. Olmstead, Daniel Oravetz, Audrey Oravetz, Yeisson Osorio, Zachary J. Pace, Nelson Padilla, Nathalie Palanque-Delabrouille, Pedro A. Palicio, Hsi-An Pan, Kaike Pan, James Parker, Romain Paviot, Sebastien Peirani, Karla Peña Ramírez, Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Ignasi Pérez-Ràfols, Patrick Petitjean, Matthew M. Pieri, Marc Pinsonneault, Vijith Jacob Poovelil, Joshua Tyler Povick, Abhishek Prakash, Adrian M. Price-Whelan, M. Jordan Raddick, Anand Raichoor, Amy Ray, Sandro Barboza Rembold, Mehdi Rezaie, Rogemar A. Riffel, Rogério Riffel, Hans-Walter Rix, Annie C. Robin, A. Roman-Lopes, Carlos Román-Zúñiga, Benjamin Rose, Ashley J. Ross, Graziano Rossi, Kate Rowlands, Kate H. R. Rubin, Mara Salvato, Ariel G. Sánchez, Laura Sánchez-Menguiano, José R. Sánchez-Gallego, Conor Sayres, Adam Schaefer, Ricardo P. Schiavon, Jaderson S. Schimoia, Edward Schlafly, David Schlegel, Donald P. Schneider, Mathias Schultheis, Axel Schwope, Hee-Jong Seo, Aldo Serenelli, Arman Shafieloo, Shoaib Jamal Shamsi, Zhengyi Shao, Shiyin Shen, Matthew Shetrone, Raphael Shirley, Víctor Silva Aguirre, Joshua D. Simon, M. F. Skrutskie, Anže Slosar, Rebecca Smethurst, Jennifer Sobeck, Bernardo Cervantes Sodi, Diogo Souto, David V. Stark, Keivan G. Stassun, Matthias Steinmetz, Dennis Stello, Julianna Stermer, Thaisa Storchi-Bergmann, Alina Streblyanska, Guy S. Stringfellow, Amelia Stutz, Genaro Suárez, Jing Sun, Manuchehr Taghizadeh-Popp, Michael S. Talbot, Jamie Tayar, Aniruddha R. Thakar, Riley Theriault, Daniel Thomas, Zak C. Thomas, Jeremy Tinker, Rita Tojeiro, Hector Hernandez Toledo, Christy A. Tremonti, Nicholas W. Troup, Sarah Tuttle, Eduardo Unda-Sanzana, Marica Valentini, Jaime Vargas-González, Mariana Vargas-Magaña, Jose Antonio Vázquez-Mata, M. Vivek, David Wake, Yuting Wang, Benjamin Alan Weaver,

- Anne-Marie Weijmans, Vivienne Wild, John C. Wilson, Robert F. Wilson, Nathan Wolthuis, W. M. Wood-Vasey, Renbin Yan, Meng Yang, Christophe Yèche, Olga Zamora, Pauline Zarrouk, Gail Zasowski, Kai Zhang, Cheng Zhao, Gongbo Zhao, Zheng Zheng, Zheng Zheng, Guangtun Zhu, and Hu Zou. The 16th Data Release of the Sloan Digital Sky Surveys : First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. , 249(1) :3, July 2020. doi : 10.3847/1538-4365/ab929e.
- [3] Brandon Allgood, Ricardo A. Flores, Joel R. Primack, Andrey V. Kravtsov, Risa H. Wechsler, Andreas Faltenbacher, and James S. Bullock. The shape of dark matter haloes : dependence on mass, redshift, radius and formation. , 367(4) :1781–1796, April 2006. doi : 10.1111/j.1365-2966.2006.10094.x.
- [4] F. Arenou, X. Luri, C. Babusiaux, C. Fabricius, A. Helmi, A. C. Robin, A. Vallenari, S. Blanco-Cuaresma, T. Cantat-Gaudin, K. Findeisen, C. Reylé, L. Ruiz-Dern, R. Sordo, C. Turon, N. A. Walton, I. C. Shih, E. Antiche, C. Barache, M. Barros, M. Breddels, J. M. Carrasco, G. Costigan, S. Diakité, L. Eyer, F. Figueras, L. Galluccio, J. Heu, C. Jordi, A. Krone-Martins, R. Lallement, S. Lambert, N. Leclerc, P. M. Marrese, A. Moitinho, R. Mor, M. Romero-Gómez, P. Sartoretti, S. Soria, C. Soubiran, J. Souchay, J. Veljanoski, H. Ziaepour, G. Giuffrida, E. Pancino, and A. Bragaglia. Gaia Data Release 1. Catalogue validation. , 599 :A50, March 2017. doi : 10.1051/0004-6361/201629895.
- [5] Camille Avestruz, Nan Li, Hanjue Zhu, Matthew Lightman, Thomas E. Collett, and Wentao Luo. Automated Lensing Learner : Automated Strong Lensing Identification with a Computer Vision Technique. , 877(1) :58, May 2019. doi : 10.3847/1538-4357/ab16d9.
- [6] C. R. Bom, M. Makler, M. P. Albuquerque, and C. H. Brandt. A neural network gravitational arc finder based on the Mediatrix filamentation method. , 597 :A135, January 2017. doi : 10.1051/0004-6361/201629159.
- [7] J. Richard Bond, Lev Kofman, and Dmitry Pogosyan. How filaments of galaxies are woven into the cosmic web. , 380(6575) :603–606, April 1996. doi : 10.1038/380603a0.
- [8] L. Breiman. Classification and Regression Trees (1st ed.). 1984. doi : 10.1201/9781315139470.
- [9] L. Breiman. Bagging predictors. 1996. doi : 10.1007/BF00058655.
- [10] S. E. Bryan, S. T. Kay, A. R. Duffy, J. Schaye, C. Dalla Vecchia, and C. M. Booth. The impact of baryons on the spins and shapes of dark matter haloes. , 429(4) :3316–3329, March 2013. doi : 10.1093/mnras/sts587.
- [11] R. Cañameras, S. Schuldt, S. H. Suyu, S. Taubenberger, T. Meinhardt, L. Leal-Taixé, C. Lemon, K. Rojas, and E. Savary. HOLISMOKES. II. Identifying galaxy-scale strong gravitational

- lenses in Pan-STARRS using convolutional neural networks. , 644 :A163, December 2020. doi : 10.1051/0004-6361/202038219.
- [12] P. Cataldi, S. E. Pedrosa, P. B. Tissera, and M. C. Artale. Baryons shaping dark matter haloes. , 501(4) :5679–5691, March 2021. doi : 10.1093/mnras/staa3988.
- [13] K. C. Chambers, E. A. Magnier, N. Metcalfe, H. A. Flewelling, M. E. Huber, C. Z. Waters, L. Denneau, P. W. Draper, D. Farrow, D. P. Finkbeiner, C. Holmberg, J. Koppenhoefer, P. A. Price, A. Rest, R. P. Saglia, E. F. Schlafly, S. J. Smartt, W. Sweeney, R. J. Wainscoat, W. S. Burgett, S. Chastel, T. Grav, J. N. Heasley, K. W. Hodapp, R. Jedicke, N. Kaiser, R. P. Kudritzki, G. A. Luppino, R. H. Lupton, D. G. Monet, J. S. Morgan, P. M. Onaka, B. Shiao, C. W. Stubbs, J. L. Tonry, R. White, E. Bañados, E. F. Bell, R. Bender, E. J. Bernard, M. Boegner, F. Boffi, M. T. Botticella, A. Calamida, S. Casertano, W. P. Chen, X. Chen, S. Cole, N. Deacon, C. Frenk, A. Fitzsimmons, S. Gezari, V. Gibbs, C. Goessl, T. Goggia, R. Gourgue, B. Goldman, P. Grant, E. K. Grebel, N. C. Hambly, G. Hasinger, A. F. Heavens, T. M. Heckman, R. Henderson, T. Henning, M. Holman, U. Hopp, W. H. Ip, S. Isani, M. Jackson, C. D. Keyes, A. M. Koekemoer, R. Kotak, D. Le, D. Liska, K. S. Long, J. R. Lucey, M. Liu, N. F. Martin, G. Masci, B. McLean, E. Mindel, P. Misra, E. Morganson, D. N. A. Murphy, A. Obaika, G. Narayan, M. A. Nieto-Santisteban, P. Norberg, J. A. Peacock, E. A. Pier, M. Postman, N. Primak, C. Rae, A. Rai, A. Riess, A. Riffeser, H. W. Rix, S. Röser, R. Russel, L. Rutz, E. Schilbach, A. S. B. Schultz, D. Scolnic, L. Strolger, A. Szalay, S. Seitz, E. Small, K. W. Smith, D. R. Soderblom, P. Taylor, R. Thomson, A. N. Taylor, A. R. Thakar, J. Thiel, D. Thilker, D. Unger, Y. Urata, J. Valenti, J. Wagner, T. Walder, F. Walter, S. P. Watters, S. Werner, W. M. Wood-Vasey, and R. Wyse. The Pan-STARRS1 Surveys. [arXiv e-prints](#), art. arXiv :1612.05560, December 2016. doi : 10.48550/arXiv.1612.05560.
- [14] Tianqi Chen and Carlos Guestrin. XGBoost : A Scalable Tree Boosting System. [arXiv e-prints](#), art. arXiv :1603.02754, March 2016. doi : 10.48550/arXiv.1603.02754.
- [15] Dan Coe, Larry Bradley, and Adi Zitrin. Frontier Fields : High-redshift Predictions and Early Results. , 800(2) :84, February 2015. doi : 10.1088/0004-637X/800/2/84.
- [16] DESI Collaboration. Overview of the DESI Legacy Imaging Surveys. , 157(5) :168, May 2019. doi : 10.3847/1538-3881/ab089d.
- [17] Christopher J. Conselice. The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. , 147(1) :1–28, July 2003. doi : 10.1086/375001.
- [18] Robert A. Crain, Ian G. McCarthy, Carlos S. Frenk, Tom Theuns, and Joop Schaye. X-ray coronae in simulations of disc galaxy formation. , 407(3) :1403–1422, September 2010. doi : 10.1111/j.1365-2966.2010.16985.x.

- [19] Robert A. Crain, Joop Schaye, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, Ian G. McCarthy, John C. Helly, Adrian Jenkins, Yetli M. Rosas-Guevara, Simon D. M. White, and James W. Trayford. The EAGLE simulations of galaxy formation : calibration of subgrid physics and model variations. , 450 (2) :1937–1961, June 2015. doi : 10.1093/mnras/stv725.
- [20] N. Dalal and C. S. Kochanek. Direct Detection of Cold Dark Matter Substructure. , 572(1) : 25–33, June 2002. doi : 10.1086/340303.
- [21] Claudio Dalla Vecchia and Joop Schaye. Simulating galactic outflows with thermal supernova feedback. , 426(1) :140–158, October 2012. doi : 10.1111/j.1365-2966.2012.21704.x.
- [22] Dark Energy Survey Collaboration. The Dark Energy Survey : more than dark energy - an overview. , 460(2) :1270–1299, August 2016. doi : 10.1093/mnras/stw641.
- [23] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large-scale structure in a universe dominated by cold dark matter. , 292 :371–394, May 1985. doi : 10.1086/163168.
- [24] L. Delchambre, A. Krone-Martins, O. Wertz, C. Ducourant, L. Galluccio, J. Klüter, F. Mignard, R. Teixeira, S. G. Djorgovski, D. Stern, M. J. Graham, J. Surdej, U. Bastian, J. Wambsganss, J. F. Le Campion, and E. Slezak. Gaia GraL : Gaia DR2 Gravitational Lens Systems. III. A systematic blind search for new lensed systems. , 622 :A165, February 2019. doi : 10.1051/0004-6361/201833802.
- [25] DESI Collaboration. The Early Data Release of the Dark Energy Spectroscopic Instrument. [arXiv e-prints](#), art. arXiv :2306.06308, June 2023. doi : 10.48550/arXiv.2306.06308.
- [26] H. Thomas Diehl, Sahar S. Allam, James Annis, Elizabeth J. Buckley-Geer, Joshua A. Frieman, Donna Kubik, Jeffrey M. Kubo, Huan Lin, Douglas Tucker, and Anderson West. The Sloan Bright Arcs Survey : Four Strongly Lensed Galaxies with Redshift > 2 . , 707(1) :686–692, December 2009. doi : 10.1088/0004-637X/707/1/686.
- [27] K. Dolag, S. Borgani, G. Murante, and V. Springel. Substructures in hydrodynamical cluster simulations. , 399(2) :497–514, October 2009. doi : 10.1111/j.1365-2966.2009.15034.x.
- [28] C. Ducourant, O. Wertz, A. Krone-Martins, R. Teixeira, J. F. Le Campion, L. Galluccio, J. Klüter, L. Delchambre, J. Surdej, F. Mignard, J. Wambsganss, U. Bastian, M. J. Graham, S. G. Djorgovski, and E. Slezak. Gaia GraL : Gaia DR2 gravitational lens systems. II. The known multiply imaged quasars. , 618 :A56, October 2018. doi : 10.1051/0004-6361/201833480.
- [29] C. Ducourant, A. Krone-Martins, L. Galluccio, R. Teixeira, J. F. Le Campion, E. Slezak, R. de Souza, P. Gavras, F. Mignard, J. Guiraud, W. Roux, S. Managau, D. Semeux, A. Blazere,

- A. Helmer, and D. Pourbaix. Gaia Data Release 3 : Surface brightness profiles of galaxies and host galaxies of quasars. [arXiv e-prints](#), art. arXiv :2206.14491, June 2022.
- [30] C. Ducourant, A. Krone-Martins, L. Galluccio, R. Teixeira, J. F. Le Campion, E. Slezak, R. de Souza, P. Gavras, F. Mignard, J. Guiraud, W. Roux, S. Managau, D. Semeux, A. Blazere, A. Helmer, and D. Pourbaix. Gaia Data Release 3. Surface brightness profiles of galaxies and host galaxies of quasars. , 674 :A11, June 2023. doi : 10.1051/0004-6361/202243798.
- [31] F. W. Dyson, A. S. Eddington, and C. Davidson. A Determination of the Deflection of Light by the Sun’s Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919. [Philosophical Transactions of the Royal Society of London Series A](#), 220 :291–333, January 1920. doi : 10.1098/rsta.1920.0009.
- [32] J. Einasto. On the Construction of a Composite Model for the Galaxy and on the Determination of the System of Galactic Parameters. [Trudy Astrofizicheskogo Instituta Alma-Ata](#), 5 :87–100, January 1965.
- [33] Albert Einstein. Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field. [Science](#), 84(2188) :506–507, December 1936. doi : 10.1126/science.84.2188.506.
- [34] I. M. H. Etherington. On the Definition of Distance in General Relativity. [Philosophical Magazine](#), 15(18) :761, January 1933.
- [35] Euclid Collaboration. Euclid preparation. III. Galaxy cluster detection in the wide photometric survey, performance and algorithm selection. , 627 :A23, July 2019. doi : 10.1051/0004-6361/201935088.
- [36] S. M. Faber and R. E. Jackson. Velocity dispersions and mass-to-light ratios for elliptical galaxies. , 204 :668–683, March 1976. doi : 10.1086/154215.
- [37] C. Fabricius, X. Luri, F. Arenou, C. Babusiaux, A. Helmi, T. Muraveva, C. Reylé, F. Spoto, A. Vallenari, T. Antoja, E. Balbinot, C. Barache, N. Bauchet, A. Bragaglia, D. Busonero, T. Cantat-Gaudin, J. M. Carrasco, S. Diakité, M. Fabrizio, F. Figueras, A. Garcia-Gutierrez, A. Garofalo, C. Jordi, P. Kervella, S. Khanna, N. Leclerc, E. Licata, S. Lambert, P. M. Marrese, A. Masip, P. Ramos, N. Robichon, A. C. Robin, M. Romero-Gómez, S. Rubele, and M. Weiler. Gaia Early Data Release 3. Catalogue validation. , 649 :A5, May 2021. doi : 10.1051/0004-6361/202039834.
- [38] Claude-André Faucher-Giguère, Adam Lidz, and Lars Hernquist. Numerical Simulations Unravel the Cosmic Web. [Science](#), 319(5859) :52, January 2008. doi : 10.1126/science.1151476.

- [39] Eric Wim Flesch. The Million Quasars (Milliquas) v7.2 Catalogue, now with VLASS associations. The inclusion of SDSS-DR16Q quasars is detailed. [arXiv e-prints](#), art. arXiv :2105.12985, May 2021. doi : 10.48550/arXiv.2105.12985.
- [40] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2) :256–285, 1995. ISSN 0890-5401. doi : <https://doi.org/10.1006/inco.1995.1136>. URL <https://www.sciencedirect.com/science/article/pii/S0890540185711364>.
- [41] Gaia Collaboration. The Gaia mission. , 595 :A1, November 2016. doi : 10.1051/0004-6361/201629272.
- [42] Gaia Collaboration. Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties. , 595 :A2, November 2016. doi : 10.1051/0004-6361/201629512.
- [43] Gaia Collaboration. Gaia Data Release 2. Summary of the contents and survey properties. , 616 :A1, August 2018. doi : 10.1051/0004-6361/201833051.
- [44] Gaia Collaboration. Gaia Early Data Release 3. Summary of the contents and survey properties (Corrigendum). , 650 :C3, June 2021. doi : 10.1051/0004-6361/202039657e.
- [45] Gaia Collaboration. Gaia Data Release 3. The extragalactic content. , 674 :A41, June 2023. doi : 10.1051/0004-6361/202243232.
- [46] Gaia Collaboration. Gaia Data Release 3. Summary of the content and survey properties. , 674 :A1, June 2023. doi : 10.1051/0004-6361/202243940.
- [47] Gaia Collaboration, C. A. L. Bailer-Jones, D. Teyssier, L. Delchambre, C. Ducourant, D. Garabato, D. Hatzidimitriou, S. A. Klioner, L. Rimoldini, I. Bellas-Velidis, R. Carballo, M. I. Carnerero, C. Diener, M. Fouesneau, L. Galluccio, P. Gavras, A. Krone-Martins, C. M. Raiteri, R. Teixeira, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, F. Arenou, C. Babusiaux, M. Biermann, O. L. Creevey, D. W. Evans, L. Eyer, R. Guerra, A. Hutton, C. Jordi, U. L. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, C. Soubiran, P. Tanga, N. A. Walton, U. Bastian, R. Drimmel, F. Jansen, D. Katz, M. G. Lattanzi, F. van Leeuwen, J. Bakker, C. Cacciari, J. Castañeda, F. De Angeli, C. Fabricius, Y. Frémat, A. Guerrier, U. Heiter, E. Masana, R. Messineo, N. Mowlavi, C. Nicolas, K. Nienartowicz, F. Pailler, P. Panuzzo, F. Riclet, W. Roux, G. M. Seabroke, R. Sordo, F. Thévenin, G. Gracia-Abril, J. Portell, M. Altmann, R. Andrae, M. Audard, K. Benson, J. Berthier, R. Blomme, P. W. Burgess, D. Busonero, G. Busso, H. Cánovas, B. Carry, A. Cellino, N. Cheek, G. Clementini, Y. Damerdji, M. Davidson, P. de Teodoro, M. Nuñez Campos, A. Dell’Oro, P. Esquej, J. Fernández-Hernández, E. Fraile, P. García-Lario, E. Gosset, R. Haigron, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, J. Hernández,

D. Hestroffer, S. T. Hodgkin, B. Holl, K. Janßen, G. Jevardat de Fombelle, S. Jordan, A. C. Lanzafame, W. Löffler, O. Marchal, P. M. Marrese, A. Moitinho, K. Muinonen, P. Osborne, E. Pancino, T. Pauwels, A. Recio-Blanco, C. Reylé, M. Riello, T. Roegiers, J. Rybizki, L. M. Sarro, C. Siopis, M. Smith, A. Sozzetti, E. Utrilla, M. van Leeuwen, U. Abbas, P. Abraham, A. Abreu Aramburu, C. Aerts, J. J. Aguado, M. Ajaj, F. Aldea-Montero, G. Altavilla, M. A. Álvarez, J. Alves, R. I. Anderson, E. Anglada Varela, T. Antoja, D. Baines, S. G. Baker, L. Balaguer-Núñez, E. Balbinot, Z. Balog, C. Barache, D. Barbato, M. Barros, M. A. Barstow, S. Bartolomé, J. L. Bassilana, N. Bauchet, U. Becciani, M. Bellazzini, A. Berihuete, M. Bernet, S. Bertone, L. Bianchi, A. Binnenfeld, S. Blanco-Cuaresma, T. Boch, A. Bombrun, D. Bossini, S. Bouquillon, A. Bragaglia, L. Bramante, E. Breedt, A. Bressan, N. Brouillet, E. Brugaletta, B. Bucciarelli, A. Burlacu, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, T. Cantat-Gaudin, T. Carlucci, J. M. Carrasco, L. Casamiquela, M. Castellani, A. Castro-Ginard, L. Chaoul, P. Charlot, L. Chemin, V. Chiaramida, A. Chiavassa, N. Chornay, G. Comoretto, G. Contursi, W. J. Cooper, T. Cornez, S. Cowell, F. Crifo, M. Cropper, M. Crosta, C. Crowley, C. Dafonte, A. Dapergolas, P. David, P. de Laverny, F. De Luise, R. De March, J. De Ridder, R. de Souza, A. de Torres, E. F. del Peloso, E. del Pozo, M. Delbo, A. Delgado, J. B. Delisle, C. Demouchy, T. E. Dharmawardena, S. Diakite, E. Distefano, C. Dolding, H. Enke, C. Fabre, M. Fabrizio, S. Faigler, G. Fedorets, P. Fernique, F. Figueras, Y. Fournier, C. Fournon, F. Fragkoudi, M. Gai, A. Garcia-Gutierrez, M. Garcia-Reinaldos, M. García-Torres, A. Garofalo, A. Gavel, E. Gerlach, R. Geyer, P. Giacobbe, G. Gilmore, S. Girona, G. Giuffrida, R. Gomel, A. Gomez, J. González-Núñez, I. González-Santamaría, J. J. González-Vidal, M. Granvik, P. Guillout, J. Guiraud, R. Gutiérrez-Sánchez, L. P. Guy, M. Hauser, M. Haywood, A. Helmer, A. Helmi, M. H. Sarmiento, S. L. Hidalgo, N. Hładczuk, D. Hobbs, G. Holland, H. E. Huckle, K. Jardine, G. Jasiewicz, A. Jean-Antoine Piccolo, Ó. Jiménez-Arranz, J. Juaristi Campillo, F. Julbe, L. Karbevská, P. Kervella, S. Khanna, M. Kontizas, G. Kordopatis, A. J. Korn, Á Kóspál, Z. Kostrzewa-Rutkowska, K. Kruszyńska, M. Kun, P. Laizeau, S. Lambert, A. F. Lanza, Y. Lasne, J. F. Le Champion, Y. Lebreton, T. Lebzelter, S. Leccia, N. Leclerc, I. Lecoeur-Taïbi, S. Liao, E. L. Licata, H. E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, A. Lorca, C. Loup, P. Madrero Pardo, A. Magdaleno Romeo, S. Managau, R. G. Mann, M. Manteiga, J. M. Marchant, M. Marconi, J. Marcos, M. M. S. Marcos Santos, D. Marín Pina, S. Marinoni, F. Marocco, D. J. Marshall, L. Martin Polo, J. M. Martín-Fleitas, G. Marton, N. Mary, A. Masip, D. Massari, A. Mastrobuono-Battisti, T. Mazeh, P. J. McMillan, S. Messina, D. Michalik, N. R. Millar, A. Mints, D. Molina, R. Molinaro, L. Molnár, G. Monari, M. Monguió, P. Montegriffo, A. Montero, R. Mor, A. Mora, R. Morbidelli, T. Morel, D. Morris, T. Muraveva, C. P. Murphy, I. Musella, Z. Nagy, L. Noval, F. Ocaña, A. Ogden, C. Ordenovic, J. O. Osinde, C. Pagani, I. Pagano, L. Palaversa, P. A. Palicio, L. Pallas-Quintela, A. Panahi, S. Payne-Wardenaar, X. Peñalosa Esteller, A. Penttilä,

B. Pichon, A. M. Piersimoni, F. X. Pineau, E. Plachy, G. Plum, E. Poggio, A. Prša, L. Pulone, E. Racero, S. Ragaini, M. Rainer, P. Ramos, M. Ramos-Lerate, P. Re Fiorentin, S. Regibo, P. J. Richards, C. Rios Diaz, V. Ripepi, A. Riva, H. W. Rix, G. Rixon, N. Robichon, A. C. Robin, C. Robin, M. Roelens, H. R. O. Rogues, L. Rohrbasser, M. Romero-Gómez, N. Rowell, F. Royer, D. Ruz Mieres, K. A. Rybicki, G. Sadowski, A. Sáez Núñez, A. Sagristà Sellés, J. Sahlmann, E. Salguero, N. Samaras, V. Sanchez Gimenez, N. Sanna, R. Santoveña, M. Sarasso, M. S. Schultheis, E. Sciacca, M. Segol, J. C. Segovia, D. Ségransan, D. Semeux, S. Shahaf, H. I. Siddiqui, A. Siebert, L. Siltala, A. Silvelo, E. Slezak, I. Slezak, R. L. Smart, O. N. Snaith, E. Solano, F. Solitro, D. Souami, J. Souchay, A. Spagna, L. Spina, F. Spoto, I. A. Steele, H. Steidelmüller, C. A. Stephenson, M. Süveges, J. Surdej, L. Szabados, E. Szegedi-Elek, F. Taris, M. B. Taylor, L. Tolomei, N. Tonello, F. Torra, J. Torra, G. Torralba Elipe, M. Trabucchi, A. T. Tsounis, C. Turon, A. Ulla, N. Unger, M. V. Vaillant, E. van Dillen, W. van Reeve, O. Vanel, A. Vecchiato, Y. Viala, D. Vicente, S. Voutsinas, M. Weiler, T. Wevers, Ł. Wyrzykowski, A. Yoldas, P. Yvard, H. Zhao, J. Zorec, S. Zucker, and T. Zwitter. Gaia Data Release 3 : The extragalactic content. [arXiv e-prints](#), art. arXiv :2206.05681, June 2022.

- [48] Gaia Collaboration, A. Krone-Martins, C. Ducourant, L. Galluccio, L. Delchambre, I. Oreshina-Slezak, R. Teixeira, J. Braine, J. F. Le Campion, F. Mignard, W. Roux, A. Blazere, L. Pegoraro, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, F. Arenou, C. Babusiaux, A. Barbier, M. Biermann, O. L. Creevey, D. W. Evans, L. Eyer, R. Guerra, A. Hutton, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, S. Randich, P. Sartoretti, R. Smiljanic, P. Tanga, N. A. Walton, C. A. L. Bailer-Jones, U. Bastian, M. Cropper, R. Drimmel, D. Katz, C. Soubiran, F. van Leeuwen, M. Audard, J. Bakker, R. Blomme, J. Castaneda, F. De Angeli, C. Fabricius, M. Fouesneau, Y. Fremat, A. Guerrier, E. Masana, R. Messineo, C. Nicolas, K. Nienartowicz, F. Pailler, P. Panuzzo, F. Riclet, G. M. Seabroke, R. Sordo, F. Thevenin, G. Gracia-Abril, J. Portell, D. Teyssier, M. Altmann, K. Benson, J. Berthier, P. W. Burgess, D. Busonero, G. Busso, H. Canovas, B. Carry, N. Cheek, G. Clementini, Y. Damerdji, M. Davidson, P. de Teodoro, A. Dell’Oro, E. Fraile Garcia, D. Garabato, P. Garcia-Lario, N. Garralda Torres, P. Gavras, R. Haigron, N. C. Hambly, D. L. Harrison, D. Hatzidimitriou, J. Hernandez, S. T. Hodgkin, B. Holl, S. Jamal, S. Jordan, A. C. Lanzafame, W. Loffler, A. Lorca, O. Marchal, P. M. Marrese, A. Moitinho, K. Muinonen, M. Nunez Campos, P. Osborne, E. Pancino, T. Pauwels, A. Recio-Blanco, M. Riello, L. Rimoldini, A. C. Robin, T. Roegiers, L. M. Sarro, M. Schultheis, C. Siopis, M. Smith, A. Sozzetti, E. Utrilla, M. van Leeuwen, K. Weingrill, U. Abbas, P. Abraham, A. Abreu Aramburu, C. Aerts, G. Altavilla, M. A. Alvarez, J. Alves, R. I. Anderson, T. Antoja, D. Baines, S. G. Baker, Z. Balog, C. Barache, D. Barbato, M. Barros, M. A. Barstow, S. Bartolome, D. Bashi, N. Bauchet, N. Baudeau, U. Becciani, L. R. Bedin, I. Bellas-Velidis, M. Bellazzini, W. Beordo, A. Berihuete, M. Bernet, C. Bertolotto, S. Bertone, L. Bianchi, A. Binnenfeld, T. Boch, A. Bombrun, S. Bouquillon, A. Bragaglia, L. Bramante,

E. Breedt, A. Bressan, N. Brouillet, E. Brugaletta, B. Bucciarelli, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, S. Cannizzo, R. Carballo, T. Carlucci, M. I. Carnerero, J. M. Carrasco, J. Carretero, S. Carton, L. Casamiquela, M. Castellani, A. Castro-Ginard, V. Cesare, P. Charlot, L. Chemin, V. Chiaramida, A. Chiavassa, N. Chornay, R. Collins, G. Contursi, W. J. Cooper, T. Cornez, M. Crosta, C. Crowley, C. Dafonte, P. de Laverny, F. De Luise, R. De March, R. de Souza, A. de Torres, E. F. del Peloso, M. Delbo, A. Delgado, T. E. Dharmawardena, S. Diakite, C. Diener, E. Distefano, C. Dolding, K. Dsilva, J. Duran, H. Enke, P. Esquej, C. Fabre, M. Fabrizio, S. Faigler, M. Fatovic, G. Fedorets, J. Fernandez-Hernandez, P. Fernique, F. Figueras, Y. Fournier, C. Fouron, M. Gai, M. Galinier, A. Garcia-Gutierrez, M. Garcia-Torres, A. Garofalo, E. Gerlach, R. Geyer, P. Giacobbe, G. Gilmore, S. Girona, G. Giuffrida, R. Gomel, A. Gomez, J. Gonzalez-Nunez, I. Gonzalez-Santamaria, E. Gosset, M. Granvik, V. Gregori Barrera, R. Gutierrez-Sanchez, M. Haywood, A. Helmer, A. Helmi, K. Henares, S. L. Hidalgo, T. Hilger, D. Hobbs, C. Hottier, H. E. Huckle, M. Jablonska, F. Jansen, O. Jimenez-Arranz, J. Juaristi Campillo, S. Khanna, G. Kordopatis, A. Kospal, Z. Kostrzewa-Rutkowska, M. Kun, S. Lambert, A. F. Lanza, Y. Lebreton, T. Lebzelter, S. Leccia, I. Lecoecur-Taibi, G. Lecoutre, S. Liao, L. Liberato, E. Licata, H. E. P. Lindstrom, T. A. Lister, E. Livanou, A. Lobel, C. Loup, L. Mahy, R. G. Mann, M. Manteiga, J. M. Marchant, M. Marconi, D. Marin Pina, S. Marinoni, D. J. Marshall, J. Martin Lozano, J. M. Martin-Fleitas, G. Marton, N. Mary, A. Masip, D. Massari, A. Mastrobuono-Battisti, T. Mazeh, P. J. McMillan, J. Meichsner, S. Messina, D. Michalik, N. R. Millar, A. Mints, D. Molina, R. Molinaro, L. Molnar, G. Monari, M. Monguio, P. Montegriffo, A. Montero, R. Mor, A. Mora, R. Morbidelli, T. Morel, D. Morris, N. Mowlavi, D. Munoz, T. Muraveva, C. P. Murphy, I. Musella, Z. Nagy, S. Nieto, L. Noval, A. Ogden, C. Ordenovic, C. Pagani, I. Pagano, L. Palaversa, P. A. Palicio, L. Pallas-Quintela, A. Panahi, C. Panem, S. Payne-Wardenaar, A. Penttila, P. Pesciullesi, A. M. Piersimoni, M. Pinamonti, F. X. Pineau, E. Plachy, G. Plum, E. Poggio, D. Pourbaix, A. Prsa, L. Pulone, E. Racero, M. Rainer, C. M. Raiteri, P. Ramos, M. Ramos-Lerate, M. Ratajczak, P. Re Fiorentin, S. Regibo, C. Reyle, V. Ripepi, A. Riva, H. W. Rix, G. Rixon, N. Robichon, C. Robin, M. Romero-Gomez, N. Rowell, F. Royer, D. Ruz Mieres, K. A. Rybicki, G. Sadowski, A. Saez Nunez, A. Sagrista Selles, J. Sahlmann, V. Sanchez Gimenez, N. Sanna, R. Santovena, M. Sarasso, C. Sarrate Riera, E. Sciacca, J. C. Segovia, D. Segransan, S. Shahaf, A. Siebert, L. Siltala, E. Slezak, R. L. Smart, O. N. Snaith, E. Solano, F. Solitro, D. Souami, J. Souchay, L. Spina, E. Spitoni, F. Spoto, L. A. Squillante, I. A. Steele, H. Steidelmuller, J. Surdej, L. Szabados, F. Taris, M. B. Taylor, K. Tisanic, L. Tolomei, F. Torra, G. Torralba Elipe, M. Trabucchi, M. Tsantaki, A. Ulla, N. Unger, O. Vanel, A. Vecchiato, D. Vicente, S. Voutsinas, M. Weiler, L. Wyrzykowski, H. Zhao, J. Zorec, T. Zwitter, L. Balaguer-Nunez, N. Leclerc, S. Morgenthaler, G. Robert, and S. Zucker. Gaia Focused Product Release : A catalogue of sources around quasars to

- search for strongly lensed quasars. [arXiv e-prints](#), art. arXiv :2310.06295, October 2023. doi : 10.48550/arXiv.2310.06295.
- [49] Rafael Garcia-Dias, Carlos Allende Prieto, Jorge Sánchez Almeida, and Ignacio Ordovás-Pascual. Machine learning in APOGEE. Unsupervised spectral classification with K-means. , 612 :A98, May 2018. doi : 10.1051/0004-6361/201732134.
- [50] Shy Genel, Mark Vogelsberger, Volker Springel, Debora Sijacki, Dylan Nelson, Greg Snyder, Vicente Rodriguez-Gomez, Paul Torrey, and Lars Hernquist. Introducing the Illustris project : the evolution of galaxy populations across cosmic time. , 445(1) :175–200, November 2014. doi : 10.1093/mnras/stu1654.
- [51] Daniel Gilman, Simon Birrer, Tommaso Treu, Charles R. Keeton, and Anna Nierenberg. Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses. , 481(1) :819–834, November 2018. doi : 10.1093/mnras/sty2261.
- [52] R. M. González Delgado, R. Cid Fernandes, R. García-Benito, E. Pérez, A. L. de Amorim, C. Cortijo-Ferrero, E. A. D. Lacerda, R. López Fernández, S. F. Sánchez, N. Vale Asari, J. Alves, J. Bland-Hawthorn, L. Galbany, A. Gallazzi, B. Husemann, S. Bekeraite, B. Jungwiert, A. R. López-Sánchez, A. de Lorenzo-Cáceres, R. A. Marino, D. Mast, M. Mollá, A. del Olmo, P. Sánchez-Blázquez, G. van de Ven, J. M. Vílchez, C. J. Walcher, L. Wisotzki, B. Ziegler, and CALIFA Collaboration. Insights on the Stellar Mass-Metallicity Relation from the CALIFA Survey. , 791(1) :L16, August 2014. doi : 10.1088/2041-8205/791/1/L16.
- [53] Robert J. J. Grand, Facundo A. Gómez, Federico Marinacci, Rüdiger Pakmor, Volker Springel, David J. R. Campbell, Carlos S. Frenk, Adrian Jenkins, and Simon D. M. White. The Auriga Project : the properties and formation mechanisms of disc galaxies across cosmic time. , 467 (1) :179–207, May 2017. doi : 10.1093/mnras/stx071.
- [54] Alexander D. Hill, Robert A. Crain, Juliana Kwan, and Ian G. McCarthy. The morphology of star-forming gas and its alignment with galaxies and dark matter haloes in the EAGLE simulations. , 505(1) :65–87, July 2021. doi : 10.1093/mnras/stab1272.
- [55] Michaela Hirschmann, Gabriella De Lucia, and Fabio Fontanot. Galaxy assembly, stellar feedback and metal enrichment : the view from the GAEA model. , 461(2) :1760–1785, September 2016. doi : 10.1093/mnras/stw1318.
- [56] Gilbert P. Holder and Paul L. Schechter. External Shear in Quadruply Imaged Lens Systems. , 589(2) :688–692, June 2003. doi : 10.1086/374688.
- [57] Philip F. Hopkins, Dušan Kereš, José Oñorbe, Claude-André Faucher-Giguère, Eliot Quataert, Norman Murray, and James S. Bullock. Galaxies on FIRE (Feedback In Realistic Environ-

- ments) : stellar feedback explains cosmologically inefficient star formation. , 445(1) :581–603, November 2014. doi : 10.1093/mnras/stu1738.
- [58] J. Huchra, M. Gorenstein, S. Kent, I. Shapiro, G. Smith, E. Horine, and R. Perley. 2237+0305 : a new and unusual gravitational lens. , 90 :691–696, May 1985. doi : 10.1086/113777.
- [59] David E. Johnston, Gordon T. Richards, Joshua A. Frieman, Charles R. Keeton, Michael A. Strauss, Gillian R. Knapp, Robert H. Becker, Richard L. White, Eric T. Johnson, Zhaoming Ma, Mark SubbaRao, Neta A. Bahcall, Mariangela Bernardi, Jon Brinkmann, Daniel J. Eisenstein, Masataka Fukugita, Patrick B. Hall, Naohisa Inada, Bartosz Pindor, David J. Schlegel, Ryan Scranton, Erin S. Sheldon, Donald P. Schneider, Alexander S. Szalay, and Donald G. York. SDSS J090334.92+502819.2 : A New Gravitational Lens. , 126(5) :2281–2290, November 2003. doi : 10.1086/379001.
- [60] E. Jullo, J. P. Kneib, M. Limousin, Á. Elíasdóttir, P. J. Marshall, and T. Verdugo. A Bayesian approach to strong lensing modelling of galaxy clusters. *New Journal of Physics*, 9(12) :447, December 2007. doi : 10.1088/1367-2630/9/12/447.
- [61] R. Kannan, E. Garaldi, A. Smith, R. Pakmor, V. Springel, M. Vogelsberger, and L. Hernquist. Introducing the THESAN project : radiation-magnetohydrodynamic simulations of the epoch of reionization. , 511(3) :4005–4030, April 2022. doi : 10.1093/mnras/stab3710.
- [62] Guinevere Kauffmann, Jorg M. Colberg, Antonaldo Diaferio, and Simon D. M. White. Clustering of galaxies in a hierarchical universe - I. Methods and results at $z=0$. , 303(1) :188–206, February 1999. doi : 10.1046/j.1365-8711.1999.02202.x.
- [63] J. P. Kneib, Y. Mellier, B. Fort, and G. Mathez. The distribution of dark matter in distant cluster-lenses : modelling modelling A 370. , 273 :367, June 1993.
- [64] C. S. Kochanek, N. D. Morgan, E. E. Falco, B. A. McLeod, J. N. Winn, J. Dembicky, and B. Ketzeback. The Time Delays of Gravitational Lens HE 0435-1223 : An Early-Type Galaxy with a Rising Rotation Curve. , 640(1) :47–61, March 2006. doi : 10.1086/499766.
- [65] Christopher S. Kochanek. Gravitational Lenses in Redshift Surveys. , 397 :381, October 1992. doi : 10.1086/171794.
- [66] R. Kormann, P. Schneider, and M. Bartelmann. Isothermal elliptical gravitational lens models. , 284 :285–299, April 1994.
- [67] J. Kristian, E. J. Groth, E. J. Shaya, D. P. Schneider, J. A. Holtzman, W. A. Baum, B. Campbell, A. Code, D. G. Currie, G. E. Danielson, S. P. Ewald, J. J. Hester, R. M. Light, C. R. Lynds, Jr. O’Neil, E. J., P. K. Seidemann, and J. A. Westphal. Imaging of the Gravitational Lens

- System PG 1115+080 With the Hubble Space Telescope. , 106 :1330, October 1993. doi : 10.1086/116729.
- [68] Jeffrey M. Kubo, Sahar S. Allam, James Annis, Elizabeth J. Buckley-Geer, H. Thomas Diehl, Donna Kubik, Huan Lin, and Douglas Tucker. The Sloan Bright Arcs Survey : Six Strongly Lensed Galaxies at $z = 0.4-1.4$. , 696(1) :L61–L65, May 2009. doi : 10.1088/0004-637X/696/1/L61.
- [69] François Lanusse, Quanbin Ma, Nan Li, Thomas E. Collett, Chun-Liang Li, Siamak Ravanbakhsh, Rachel Mandelbaum, and Barnabás Póczos. CMU DeepLens : deep learning for automatic image-based galaxy-galaxy strong lens finding. , 473(3) :3895–3906, January 2018. doi : 10.1093/mnras/stx1665.
- [70] C. Lemon, T. Anguita, M. W. Auger-Williams, F. Courbin, A. Galan, R. McMahon, F. Neira, M. Oguri, P. Schechter, A. Shajib, T. Treu, A. Agnello, and C. Spiniello. Gravitationally lensed quasars in Gaia - IV. 150 new lenses, quasar pairs, and projected quasars. , 520(3) : 3305–3328, April 2023. doi : 10.1093/mnras/stac3721.
- [71] LIGO Scientific Collaboration and Virgo Collaboration. A Gravitational-wave Measurement of the Hubble Constant Following the Second Observing Run of Advanced LIGO and Virgo. , 909(2) :218, March 2021. doi : 10.3847/1538-4357/abdeb7.
- [72] J. M. Lotz, A. Koekemoer, D. Coe, N. Grogin, P. Capak, J. Mack, J. Anderson, R. Avila, E. A. Barker, D. Borncamp, G. Brammer, M. Durbin, H. Gunning, B. Hilbert, H. Jenkner, H. Khandrika, Z. Levay, R. A. Lucas, J. MacKenty, S. Ogaz, B. Porterfield, N. Reid, M. Robberto, P. Royle, L. J. Smith, L. J. Storrie-Lombardi, B. Sunnquist, J. Surace, D. C. Taylor, R. Williams, J. Bullock, M. Dickinson, S. Finkelstein, P. Natarajan, J. Richard, B. Robertson, J. Tumlinson, A. Zitrin, K. Flanagan, K. Sembach, B. T. Soifer, and M. Mountain. The Frontier Fields : Survey Design and Initial Results. , 837(1) :97, March 2017. doi : 10.3847/1538-4357/837/1/97.
- [73] R. Lynds and V. Petrosian. Giant Luminous Arcs in Galaxy Clusters. In Bulletin of the American Astronomical Society, volume 18, page 1014, September 1986.
- [74] G. Mahler, J. Richard, B. Clément, D. Lagattuta, K. Schmidt, V. Patrício, G. Soucail, R. Bacon, R. Pello, R. Bouwens, M. Maseda, J. Martinez, M. Carollo, H. Inami, F. Leclercq, and L. Wisotzki. Strong-lensing analysis of A2744 with MUSE and Hubble Frontier Fields images. , 473(1) :663–692, January 2018. doi : 10.1093/mnras/stx1971.
- [75] P. Marigo. Chemical yields from low- and intermediate-mass stars : Model predictions and basic observational constraints. , 370 :194–217, April 2001. doi : 10.1051/0004-6361:20000247.

- [76] S. McAlpine, J.C. Helly, M. Schaller, J.W. Trayford, Y. Qu, M. Furlong, R.G. Bower, R.A. Crain, J. Schaye, T. Theuns, and et al. The eagle simulations of galaxy formation : Public release of halo and galaxy catalogues. *Astronomy and Computing*, 15 :72–89, Apr 2016. ISSN 2213-1337. doi : 10.1016/j.ascom.2016.02.004. URL <http://dx.doi.org/10.1016/j.ascom.2016.02.004>.
- [77] Massimo Meneghetti. *Introduction to Gravitational Lensing : With Python Examples*. 2022.
- [78] A. More, R. Cabanac, S. More, C. Alard, M. Limousin, J. P. Kneib, R. Gavazzi, and V. Motta. The CFHTLS-Strong Lensing Legacy Survey (SL2S) : Investigating the Group-scale Lenses with the SARCS Sample. , 749(1) :38, April 2012. doi : 10.1088/0004-637X/749/1/38.
- [79] Daniel J. Mortlock and Rachel L. Webster. Using galaxy redshift surveys to detect gravitationally lensed quasars. , 319(3) :879–892, December 2000. doi : 10.1046/j.1365-8711.2000.03893.x.
- [80] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The Structure of Cold Dark Matter Halos. , 462 :563, May 1996. doi : 10.1086/177173.
- [81] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Rüdiger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations : the galaxy colour bimodality. , 475(1) :624–647, March 2018. doi : 10.1093/mnras/stx3040.
- [82] J. H. Oort. The force exerted by the stellar system in the direction perpendicular to the galactic plane and some related problems. , 6 :249, August 1932.
- [83] Q. Petit, C. Ducourant, E. Slezak, D. Sluse, and L. Delchambre. Shape, alignment, and mass distribution of baryonic and dark-matter halos in one EAGLE simulation. , 669 :A132, January 2023. doi : 10.1051/0004-6361/202244920.
- [84] Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Rüdiger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, and Federico Marinacci. Simulating galaxy formation with the IllustrisTNG model. , 473(3) :4077–4106, January 2018. doi : 10.1093/mnras/stx2656.
- [85] Planck Collaboration. Planck 2013 results. XVI. Cosmological parameters. , 571 :A16, November 2014. doi : 10.1051/0004-6361/201321591.
- [86] Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. , 641 :A6, September 2020. doi : 10.1051/0004-6361/201833910.
- [87] L. Portinari, C. Chiosi, and A. Bressan. Galactic chemical enrichment with new metallicity dependent stellar yields. , 334 :505–539, June 1998.

- [88] Adam G. Riess, Stefano Casertano, Wenlong Yuan, Lucas M. Macri, and Dan Scolnic. Large Magellanic Cloud Cepheid Standards Provide a 1% Foundation for the Determination of the Hubble Constant and Stronger Evidence for Physics beyond Λ CDM. , 876(1) :85, May 2019. doi : 10.3847/1538-4357/ab1422.
- [89] Y. M. Rosas-Guevara, R. G. Bower, J. Schaye, M. Furlong, C. S. Frenk, C. M. Booth, R. A. Crain, C. Dalla Vecchia, M. Schaller, and T. Theuns. The impact of angular momentum on black hole accretion rates in simulations of galaxy formation. , 454(1) :1038–1057, November 2015. doi : 10.1093/mnras/stv2056.
- [90] V. C. Rubin, Jr. Ford, W. K., and N. Thonnard. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc). , 238 :471–487, June 1980. doi : 10.1086/158003.
- [91] Vera C. Rubin and Jr. Ford, W. Kent. Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. , 159 :379, February 1970. doi : 10.1086/150317.
- [92] C. Scannapieco, M. Wadepuhl, O. H. Parry, J. F. Navarro, A. Jenkins, V. Springel, R. Teyssier, E. Carlson, H. M. P. Couchman, R. A. Crain, C. Dalla Vecchia, C. S. Frenk, C. Kobayashi, P. Monaco, G. Murante, T. Okamoto, T. Quinn, J. Schaye, G. S. Stinson, T. Theuns, J. Wadsley, S. D. M. White, and R. Woods. The Aquila comparison project : the effects of feedback and numerical methods on simulations of galaxy formation. , 423(2) :1726–1749, June 2012. doi : 10.1111/j.1365-2966.2012.20993.x.
- [93] Matthieu Schaller. Effects of baryons on the dark matter distribution in cosmological hydrodynamical simulations. PhD thesis, Durham University, UK, September 2015.
- [94] Matthieu Schaller, Carlos S. Frenk, Richard G. Bower, Tom Theuns, Adrian Jenkins, Joop Schaye, Robert A. Crain, Michelle Furlong, Claudio Dalla Vecchia, and I. G. McCarthy. Baryon effects on the internal structure of Λ CDM haloes in the EAGLE simulations. , 451(2) : 1247–1267, August 2015. doi : 10.1093/mnras/stv1067.
- [95] Schapire and E. Robert. The Strength of Weak Learnability. 1990. doi : 10.1007/bf00116037.
- [96] Joop Schaye. Star Formation Thresholds and Galaxy Edges : Why and Where. , 609(2) : 667–682, July 2004. doi : 10.1086/421232.
- [97] Joop Schaye and Claudio Dalla Vecchia. On the relation between the Schmidt and Kennicutt-Schmidt star formation laws and its implications for numerical simulations. , 383(3) :1210–1222, January 2008. doi : 10.1111/j.1365-2966.2007.12639.x.

- [98] Joop Schaye and Claudio Dalla Vecchia. On the relation between the Schmidt and Kennicutt-Schmidt star formation laws and its implications for numerical simulations. , 383(3) :1210–1222, January 2008. doi : 10.1111/j.1365-2966.2007.12639.x.
- [99] Joop Schaye, Claudio Dalla Vecchia, C. M. Booth, Robert P. C. Wiersma, Tom Theuns, Marcel R. Haas, Serena Bertone, Alan R. Duffy, I. G. McCarthy, and Freeke van de Voort. The physics driving the cosmic star formation history. , 402(3) :1536–1560, March 2010. doi : 10.1111/j.1365-2966.2009.16029.x.
- [100] Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The EAGLE project : simulating the evolution and assembly of galaxies and their environments. , 446(1) :521–554, January 2015. doi : 10.1093/mnras/stu2058.
- [101] Paul L. Schechter, Charles D. Bailyn, Robert Barr, Richard Barvainis, Christopher M. Becker, Gary M. Bernstein, John P. Blakeslee, Schelte J. Bus, Alan Dressler, Emilio E. Falco, Robert A. Fesen, Phillippe Fischer, Karl Gebhardt, Dianne Harmer, Jacqueline N. Hewitt, Jens Hjorth, Todd Hurt, Andreas O. Jaunsen, Mario Mateo, Doerte Mehlert, Douglas O. Richstone, Linda S. Sparke, John R. Thorstensen, John L. Tonry, Gary Wegner, Daryl W. Willmarth, and Guy Worthey. The Quadruple Gravitational Lens PG 1115+080 : Time Delays and Models. , 475 (2) :L85–L88, February 1997. doi : 10.1086/310478.
- [102] M. Schmidt. 3C 273 : A Star-Like Object with Large Red-Shift. , 197(4872) :1040, March 1963. doi : 10.1038/1971040a0.
- [103] G. Seidel and M. Bartelmann. Arcfinder : an algorithm for the automatic detection of gravitational arcs. , 472(1) :341–352, September 2007. doi : 10.1051/0004-6361:20066097.
- [104] Shi Shao, Marius Cautun, Carlos S. Frenk, Liang Gao, Robert A. Crain, Matthieu Schaller, Joop Schaye, and Tom Theuns. Alignments between galaxies, satellite systems and haloes. , 460(4) :3772–3783, August 2016. doi : 10.1093/mnras/stw1247.
- [105] Rachel S. Somerville and Romeel Davé. Physical Models of Galaxy Formation in a Cosmological Framework. , 53 :51–113, August 2015. doi : 10.1146/annurev-astro-082812-140951.
- [106] Alessandro Sonnenfeld, James H. H. Chan, Yiping Shu, Anupreeta More, Masamune Oguri, Sherry H. Suyu, Kenneth C. Wong, Chien-Hsiu Lee, Jean Coupon, Atsunori Yonehara, Adam S. Bolton, Anton T. Jaelani, Masayuki Tanaka, Satoshi Miyazaki, and Yutaka Komiyama. Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). I. Automatic search for galaxy-scale strong lenses. , 70 :S29, January 2018. doi : 10.1093/pasj/psx062.

- [107] G. Soucail, B. Fort, Y. Mellier, and J. P. Picat. A blue ring-like structure in the center of the A 370 cluster of galaxies. , 172 :L14–L16, January 1987.
- [108] G. Soucail, Y. Mellier, B. Fort, G. Mathez, and M. Cailloux. The giant arc in A 370 : spectroscopic evidence for gravitational lensing from a source at $Z=0.724$. , 191 :L19–L21, February 1988.
- [109] V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk, and S. D. M. White. The Aquarius Project : the subhaloes of galactic haloes. , 391(4) : 1685–1711, December 2008. doi : 10.1111/j.1365-2966.2008.14066.x.
- [110] Volker Springel and Lars Hernquist. The history of star formation in a Λ cold dark matter universe. , 339(2) :312–334, February 2003. doi : 10.1046/j.1365-8711.2003.06207.x.
- [111] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. Populating a cluster of galaxies - I. Results at $[formmu]z=0$. , 328(3) :726–750, December 2001. doi : 10.1046/j.1365-8711.2001.04912.x.
- [112] Volker Springel, Tiziana Di Matteo, and Lars Hernquist. Modelling feedback from stars and black holes in galaxy mergers. , 361(3) :776–794, August 2005. doi : 10.1111/j.1365-2966.2005.09238.x.
- [113] D. Stern, S. G. Djorgovski, A. Krone-Martins, D. Sluse, L. Delchambre, C. Ducourant, R. Teixeira, J. Surdej, C. Boehm, J. den Brok, D. Dobie, A. Drake, L. Galluccio, M. J. Graham, P. Jalan, J. Klüter, J. F. Le Campion, A. Mahabal, F. Mignard, T. Murphy, A. Nierenberg, Jr. Scarano, S., J. Simon, E. Slezak, C. Spindola-Duarte, and J. Wambsganss. Gaia GraL : Gaia DR2 Gravitational Lens Systems. VI. Spectroscopic Confirmation and Modeling of Quadruply Imaged Lensed Quasars. , 921(1) :42, November 2021. doi : 10.3847/1538-4357/ac0f04.
- [114] D. Stern, S. G. Djorgovski, A. Krone-Martins, D. Sluse, L. Delchambre, C. Ducourant, R. Teixeira, J. Surdej, C. Boehm, J. den Brok, D. Dobie, A. Drake, L. Galluccio, M. J. Graham, P. Jalan, J. Kluter, J. F. Le Campion, A. Mahabal, F. Mignard, T. Murphy, A. Nierenberg, Jr. Scarano, S., J. Simon, E. Slezak, C. Spindola-Duarte, and J. Wambsganss. VizieR Online Data Catalog : Gaia GraL. VI. Quadruply imaged lensed QSOs (Stern+, 2021). VizieR Online Data Catalog, art. J/ApJ/921/42, February 2023.
- [115] Adelheid F. Teklu, Rhea-Silvia Remus, Klaus Dolag, Alexander Arth, Andreas Burkert, Aura Obreja, and Felix Schulze. Declining Rotation Curves at $z = 2$ in Λ CDM Galaxy Formation Simulations. , 854(2) :L28, February 2018. doi : 10.3847/2041-8213/aaeb4.

- [116] Ananth Tenneti, Rachel Mandelbaum, Tiziana Di Matteo, Yu Feng, and Nishikanta Khandai. Galaxy shapes and intrinsic alignments in the MassiveBlack-II simulation. , 441(1) :470–485, June 2014. doi : 10.1093/mnras/stu586.
- [117] The EAGLE team. The EAGLE simulations of galaxy formation : Public release of particle data. arXiv e-prints, art. arXiv :1706.09899, June 2017.
- [118] Adrien C. R. Thob, Robert A. Crain, Ian G. McCarthy, Matthieu Schaller, Claudia D. P. Lagos, Joop Schaye, Geert Jan J. Talens, Philip A. James, Tom Theuns, and Richard G. Bower. The relationship between the morphology and kinematics of galaxies and its dependence on dark matter halo structure in EAGLE. , 485(1) :972–987, May 2019. doi : 10.1093/mnras/stz448.
- [119] F. Torra, J. Castañeda, C. Fabricius, L. Lindegren, M. Clotet, J. J. González-Vidal, S. Bartolomé, U. Bastian, M. Bernet, M. Biermann, N. Garralda, A. Gúrpide, U. Lammers, J. Portell, and J. Torra. Gaia Early Data Release 3. Building the Gaia DR3 source list - Cross-match of Gaia observations. , 649 :A10, May 2021. doi : 10.1051/0004-6361/202039637.
- [120] C. Vanderriest, J. Schneider, G. Herpe, M. Chevretton, M. Moles, and G. Wlerick. The value of the time delay $\Delta T (A,B)$ for the 'double' quasar 0957+561 from optical photometric monitoring. , 215 :1–13, May 1989.
- [121] Marco Velliscig, Marcello Cacciato, Joop Schaye, Robert A. Crain, Richard G. Bower, Marcel P. van Daalen, Claudio Dalla Vecchia, Carlos S. Frenk, Michelle Furlong, I. G. McCarthy, Matthieu Schaller, and Tom Theuns. The alignment and shape of dark matter, stellar, and hot gas distributions in the EAGLE and cosmo-OWLS simulations. , 453(1) :721–738, October 2015. doi : 10.1093/mnras/stv1690.
- [122] M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, S. Bird, D. Nelson, and L. Hernquist. Properties of galaxies reproduced by a hydrodynamic simulation. , 509(7499) :177–182, May 2014. doi : 10.1038/nature13316.
- [123] Mark Vogelsberger, Amina Helmi, Volker Springel, Simon D. M. White, Jie Wang, Carlos S. Frenk, Adrian Jenkins, Aaron Ludlow, and Julio F. Navarro. Phase-space structure in the local dark matter distribution and its signature in direct detection experiments. , 395(2) : 797–811, May 2009. doi : 10.1111/j.1365-2966.2009.14630.x.
- [124] Mark Vogelsberger, Jesus Zavala, and Abraham Loeb. Subhaloes in self-interacting galactic dark matter haloes. Monthly Notices of the Royal Astronomical Society, 423(4) :3740–3752, 07 2012. ISSN 0035-8711. doi : 10.1111/j.1365-2966.2012.21182.x. URL <https://doi.org/10.1111/j.1365-2966.2012.21182.x>.

- [125] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Greg Snyder, Dylan Nelson, and Lars Hernquist. Introducing the Illustris Project : simulating the coevolution of dark and visible matter in the Universe. , 444(2) :1518–1547, October 2014. doi : 10.1093/mnras/stu1536.
- [126] D. Walsh, R. F. Carswell, and R. J. Weymann. 0957+561 A, B : twin quasistellar objects or gravitational lens ? , 279 :381–384, May 1979. doi : 10.1038/279381a0.
- [127] Liang Wang, Aaron A. Dutton, Gregory S. Stinson, Andrea V. Macciò, Camilla Penzo, Xi Kang, Ben W. Keller, and James Wadsley. NIHAO project - I. Reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations. , 454(1) :83–94, November 2015. doi : 10.1093/mnras/stv1937.
- [128] Ray J. Weymann, David Latham, J. Roger P. Angel, Richard F. Green, James W. Liebert, David A. Turnshek, Diane E. Turnshek, and J. Anthony Tyson. The triple QSO PG1115 + 08 : another probable gravitational lens. , 285(5767) :641–643, June 1980. doi : 10.1038/285641a0.
- [129] Robert P. C. Wiersma, Joop Schaye, and Britton D. Smith. The effect of photoionization on the cooling rates of enriched, astrophysical plasmas. , 393(1) :99–107, February 2009. doi : 10.1111/j.1365-2966.2008.14191.x.
- [130] Robert P. C. Wiersma, Joop Schaye, Tom Theuns, Claudio Dalla Vecchia, and Luca Tornatore. Chemical enrichment in cosmological, smoothed particle hydrodynamics simulations. , 399(2) : 574–600, October 2009. doi : 10.1111/j.1365-2966.2009.15331.x.
- [131] Matthew P. Wiesner, Huan Lin, Sahar S. Allam, James Annis, Elizabeth J. Buckley-Geer, H. Thomas Diehl, Donna Kubik, Jeffrey M. Kubo, and Douglas Tucker. The Sloan Bright Arcs Survey : Ten Strong Gravitational Lensing Clusters and Evidence of Overconcentration. , 761(1) :1, December 2012. doi : 10.1088/0004-637X/761/1/1.
- [132] Kenneth C. Wong, Sherry H. Suyu, Geoff C. F. Chen, Cristian E. Rusu, Martin Millon, Dominique Sluse, Vivien Bonvin, Christopher D. Fassnacht, Stefan Taubenberger, Matthew W. Auger, Simon Birrer, James H. H. Chan, Frederic Courbin, Stefan Hilbert, Olga Tihhonova, Tommaso Treu, Adriano Agnello, Xuheng Ding, Inh Jee, Eiichiro Komatsu, Anowar J. Shajib, Alessandro Sonnenfeld, Roger D. Blandford, Léon V. E. Koopmans, Philip J. Marshall, and Georges Meylan. H0LiCOW - XIII. A 2.4 per cent measurement of H_0 from lensed quasars : 5.3 σ tension between early- and late-Universe probes. , 498(1) :1420–1439, October 2020. doi : 10.1093/mnras/stz3094.
- [133] Peter Young, Robert S Deverill, James E Gunn, James A Westphal, and Jerome Kristian. The triple quasar q1115+ 080a, b, c-a quintuple gravitational lens image. *Astrophysical Journal*, 244(3) :723–735, 1981.

- [134] F. Zwicky. On the probability of detecting nebulae which act as gravitational lenses. Phys. Rev., 51 :679–679, Apr 1937. doi : 10.1103/PhysRev.51.679. URL <https://link.aps.org/doi/10.1103/PhysRev.51.679>.
- [135] F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. , 86 :217, October 1937. doi : 10.1086/143864.