Comparación de funciones de pérdida para predicción de características de galaxias en base a sus halos

Ivan Oyarzún Rojas
ivan.oyarzun@usm.cl
Universidad Técnica Federico Santa María
Santiago, Región Metropolitana, Chile

Benjamin Espinoza
benjamin.espinozahu@usm.cl
Universidad Técnica Federico Santa María
Santiago, Región Metropolitana, Chile

ABSTRACT

Comprender las propiedades del halo y de las galaxias es fundamental para el proceso de estudio de la formación de galaxias para el campo de la cosmología. Al no ser una relación trivial de estudiar, diversas herramientas han sido empleadas para su estudio y desarrollo. El fin de este artículo es centrar su foco en el uso de redes neuronales artificiales para lograr esto, buscando obtener mejoras en una de las soluciones propuestas en la actualidad, que mediante el uso de simulaciones cosmológicas magnetohidrodinámicas Illustris TNG300, predicen propiedades (masa estelar, color, tasa de formación estelar específica y radio) de las galaxias centrales a partir de las propiedades del halo. El medio para lograr mejoras para este modelo es un enfoque basado en cuantiles para su proceso de entrenamiento, a través de otra función de pérdida. De esta manera se espera demostrar un mejor rendimiento respecto a la anterior solución utilizando dos funciones de pérdida propuestas.

KEYWORDS

Galaxias, Halos, Redes Neuronales

1 INTRODUCCIÓN

1.1 Marco Teórico

Un punto clave para la comprensión de las galaxias a profundidad es conocer la relación entre estas y sus halos de materia oscura, lo que da mayor entendimiento al cómo las propiedades y la evolución de las galaxias se ve directamente influenciada por sus halos y sus características. Esto es de gran relevancia para el estudio de la influencia de la materia oscura en las estructuras de masa bariónica.

Siguiendo la bien explicada línea mencionada en el artículo de Santi et al. [3] junto a Rodrigues et al. [6], podemos diferenciar propiedades tanto para las galaxias como para sus halos.
Las propiedades de los halos en que se centrará el estudio constan de las siguientes:

(1) Masa Virial ($M_{vir}[h^{-1}M_{\odot}]$, definido básicamente como la masa encerrada dentro del radio virial R_{vir}. En el trabajo de de Santi et al. [3] se define un corte de $\log_{10}(M_{vir}[h^{-1}M_{\odot}) \geq 10.5$.
(2) Concentración Virial c_{vir}, definido como una relación entre el radio virial R_{vir} y el radio de escala R_s de forma $c_{vir} = R_{vir}/R_s$.
(3) Edad del Halo, desplazamiento al rojo de $z_{1/2}$. Desplazamiento al rojo donde por primera vez la mitad de la masa actual se a acumulado en un solo subhalo.
(4) Spin del Halo, λ_{halo}. Definido citando el trabajo de Bullock et al. [1] como $\lambda_{halo} = |J|\sqrt{2}M_{vir}V_{vir}R_{vir}$, con J como su momento angular del halo y V_{vir} como su velocidad circular en R_{vir}.
(5) Sobredensidad, definido como la densidad de los subhalos dentro de una esfera de radio $R = 3h^{-1}Mpc$.

A partir de esto último, se define la entrada del modelo de redes neuronales a estudiar.

En concordancia de esto y como valores de salida, se definen en de Santi et al. [3] conjuntamente con Rodrigues et al. [6] las propiedades a estudiar de las galaxias como:

- Masa estelar, definido como la masa de entre todas las partículas dentro de cada subhalo. También es definido un corte inferior de $\log_{10}(M_*[h^{-1}M_{\odot}) \geq 8.75$.
- Tamaño de la galaxia, que se parametriza como el radio de media masa estelar.
- Tasa de formación estelar específica (sSFR $[año^{-1}h]$) tasa total por cada partícula de gas contenida en el subhalo por unidad de masa solar.
- Color $g - i$, obtenido de IllustrisTNG mediante la suma de luminosidades de la totalidad de las partículas estelares de cada subhalo.

De las cuales, este estudio se centrará netamente en **Masa estelar** y **Color g-i** para realizar un estudio de mayor detalle y concordancia a las bases este (de Santi et al. [3] y Rodrigues et al. [6]).

1.2 Motivación

El principal motor detrás de esta investigación radica en su potencial para contribuir tanto al campo de la cosmología como al desarrollo de técnicas avanzadas en Deep Learning.
En el ámbito de la cosmología, esta investigación es de gran relevancia ya que aborda la relación entre los halos de materia oscura y las galaxias. Comprender esta relación es crucial para profundizar los estudios de los procesos de formación y evolución de las galaxias dentro de la estructura a gran escala del Universo. Los halos de materia oscura son estructuras fundamentales que influyen en la dinámica y en la distribución de la materia visible, por lo que su estudio aporta a una gran cantidad de materias dentro de la cosmología.
Desde la perspectiva del Deep Learning, esta investigación ofrece una oportunidad para explorar nuevas metodologías y enfoques distintos a los tradicionales. En particular, se destaca el uso de técnicas de entrenamiento basadas en cuantiles y sus funciones de pérdida asociadas. Este método de enfocar problemas de regresión a funciones de pérdidas de carácter cuantilico, se pueden dar inferencias más concretas y precisas sobre las relaciones entre las variables de estudio.

1.3 Resumen del Trabajo Relacionado

Un primer acercamiento a este problema lo da de Santi et al. [3], con la utilización de diferentes técnicas clásicas de *Machine Learning* en conjunto al uso de redes neuronales, utilizando el error cuadrático medio (MSE Loss) como función de pérdida.

En una continuación a este trabajo se cambia el enfoque a utilizar únicamente redes neuronales, en donde se trata el problema como uno de clasificación, por lo que se utiliza la *Cross-Entropy Loss* como función de pérdida, prediciendo probabilidades de las características de salida para posteriormente reinterpretarles con valores numéricos concretos.

Es este último el trabajo que se utilizará como punto de partida, utilizando una arquitectura de red neuronal similar y el mismo tipo de datos, con la intención de lograr mejores resultados o resultados que generen competencia a partir de un enfoque diferente.

Para tratar de abordar este problema, las funciones de pérdida basadas en cuantiles son más apropiadas, ya que se busca abordar el aprendizaje conjunto de las distribuciones de los datos. Un enfoque notable es el de Yan et al. [9], en donde se utiliza la suma de las regresiones de los cuantiles (1) con el objetivo de una mejor aproximación de los cuantiles de las características de salida, buscando a su vez mejorar las predicciones.

También es importante considerar el trabajo de Tagasovska and Lopez-Paz [7], donde proponen la *Simultaneous Quantile Regression*(SQR) (2), un enfoque de aproximación continua de cuantiles, similar a (1), pero con una orientación un poco más moderna, con la característica de que permite aproximación continua de cuantiles, reduciendo errores como el eleatorio y el cruce de cuantiles (reducción de inconsistencias al predecir cuantiles de forma conjunta).

$$L(y, Q^t(x)) = \sum_{k=1}^{K} L_{\tau_k}(y, q_k^t(x)) \tag{1}$$

$$\hat{f} \in \arg\min_f \frac{1}{n} \sum_{i=0}^{n} \mathop{\mathbb{E}}_{\tau \sim U[0,1]} [\ell_\tau(f(x_i, \tau), y_i)] \tag{2}$$

1.4 Contribuciones

Las contribuciones de la solución propuesta a este problema es, en primer lugar, utilizar funciones de pérdida acordes a cómo se está trabajando actualmente, es decir, mediante distribuciones, en vez de reducir el problema a uno de clasificación. En segundo lugar, tratar de mejorar la precisión de las predicciones logradas en la literatura actual, pero particularmente en el trabajo de Rodrigues et al. [6].

1.5 Estructuración del Trabajo

El documento que continúa organizará de la siguiente manera: el trabajo relacionado al campo que es relevante para el estudio se trata en la sección 2. Posteriormente, se define directamente la pregunta de investigación que será respondida en el desarrollo del artículo en la sección 3. Luego se definirán los experimentos realizados y los resultados , para así finalizar con las conclusiones en las secciones 5,6 y 7 respectivamente.

2 TRABAJO RELACIONADO

Illustris TNG goza de una gran cantidad de datos, por lo mismo, llama la atención de múltiples investigadores tanto del área como aledaños a adentrarse en sus simulaciones.

El trabajo de Wu and Jespersen [8] busca dar un enfoque interesante entrenando una message-passing graph neural networks (GNNs) también a partir de datos obtenidos de la simulación TNG-300-1. El modelo de GNN predice masa estelar a partir de las posiciones, cinemática, masas y velocidades circulares máximas del subhalo central y subhalos vecinos. Para el modelo GNN se define el vector de características $V_i = (\boldsymbol{x_i}, \boldsymbol{v_i}, M_{halo,i}, V_{max,i})$ respectivamente a las características antes mencionadas. Los subhalos con una distancia de enlace de al menos $L = 5Mpc$ se conectan con una arista. Se agrega un padding de $2.5Mpc$. Por cada arista $\mathcal{E}_{i,j}$ se calcula la distancia Euclidiana cuadrada $d_{ij} \equiv ||\boldsymbol{x_i} - \boldsymbol{x_j}||$, el producto interno ente los vectores unitarios $\boldsymbol{e_i} \cdot \boldsymbol{e_j}$ y el producto interno ente los vectores unitarios $\boldsymbol{e_i} \cdot \boldsymbol{e_{i-j}}$, donde el vector unitario se calcula como $\boldsymbol{e_i} \equiv (\boldsymbol{x_i} - \bar{\boldsymbol{x}})/||\boldsymbol{x_i} - \bar{\boldsymbol{x}}||$. Junto a esto, se agregan capas de Max Pooling en cada nodo. Además, un conjunto de fully-connected layers (FCL) con 256 latent channels y 128 hidden channels. Se predicen dos campos para cada nodo: la masa estelar en escala logarítmica: $\log(M_{\star,i}/M_\odot)$ y la varianza logarítmica: $\log \Sigma_i$. La función de pérdida consta de dos componentes: el error cuadrático medio para la masa estelar logarítmica: $||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2$ y la diferencia cuadrada entre la varianza predicha y medida: $||\hat{\Sigma} - (\hat{\boldsymbol{y}} - \boldsymbol{y})^2||^2$. Adicionalmente, se utiliza AdamW como optimizador.
Los resultados del modelo se muestran en la tabla 1 del trabajo de Wu and Jespersen [8], donde se da a entender lo bien que funciona **GNN (3*d*)** con un RMSE (raiz del error cuadrático medio) de $0.129dex$, un MAE (promedio de errores absolutos) de $0.125dex$, un NMAD (desviación mediana absoluta normalizada) de $0.102dex$, un coeficiente de correlación de Pearson (medida lineal de dependencia entre variables) ρ de 0.975, un coeficiente de determinación R^2 (calidad del modelo al replicar) de 0.951, un Bias de $0.8 \pm 0.6 \cdot 10^{-3} dex$ y un porcentaje de Outlier (valores atípicos) de $0.68 \pm 0.00\%$. Lo que resalta por encima de los modelos de RF, AM y SHAM entrenados y probados.

Por otro lado, se presenta el trabajo de Chittenden and Tojeiro [2], donde la característica temporal se toma a consideración, trabajo en el cual se definen características temporales, y atemporales de los halos. En este se establece una red semi-recurrente, utilizando una red recurrente para las características temporales, considerando 33 instancias temporales por característica y una red densa para las atemporales, las cuales se combinan en una única red completamente conectada. La arquitectura específica de la red se puede ver en la figura 10 del trabajo de Chittenden y Tojeiro, con una función de activación ELU en cada capa. La envergadura de este trabajo abarca muchas predicciones sobre la evolución de las galaxias, pero una interesante para el actual enfoque es la relación entre la masa del halo y la masa estelar, en donde los autores Chittenden and Tojeiro [2] logran generar predicciones precisas, aunque con tendencia a subestimar los valores con un MAD (desviación mediana absoluta) de $0.079dex$ para galaxias centrales y $0.094dex$ para galaxias satelites.

3 PREGUNTA DE INVESTIGACIÓN

¿Es posible lograr mejoras en los resultados mediante el entrenamiento de una Red Neuronal con una función de pérdida basada en cuantiles para el estudio de la relación halo-galaxia?

Utilizando una red neuronal basada en el trabajo de Rodrigues et al. [6] y los mismos datos, se tiene la intención de realizar el entrenamiento con 2 funciones de pérdidas basadas en cuantiles, donde se trabajará la *Simultaneous Quantile Regression* (SQR) (Tagasovska and Lopez-Paz [7]) y la suma de las regresiones de cuantiles (SRC) (Yan et al. [9]) para competir contra la red neuronal del trabajo de Rodrigues et al. [6], que utiliza la *Cross-Entropy Loss* como función de pérdida.

4 METODOLOGÍA

4.1 Materiales

Todos los datos a utilizar en estre trabajo proviene de la simulación IlustrisTNG300-1, la cuál es definida en base a las propiedades establecidas en el marco teórico.

Se trabajará con información atemporal en un único vector unidimensional de entrada, y un único vector unidimensional de salida. Para mantener consistencia con los trabajos anteriores (de Santi et al. [3] y Rodrigues et al. [6]), se hará el mismo procesamiento de datos, que consiste en definir de manera aleatoria el sSFR del 14% de los datos ya que este es nulo, utilizando una distribución Gaussiana $\mathcal{N}(\mu = -13.5, \sigma = 0.5)$.

Para el entrenamiento, el total se distribuira de tal forma que 70%, 15% y 15% correspondan al conjunto de entrenamiento, validación y prueba respectivamente.

4.2 Métricas de rendimiento

Para cuantificar el rendimiento del modelo, se utilizará el test de Kolmogorov–Smirnov (KS Test) $\Delta = max(|F_1 - F_2|)$, donde menor sea el valor, mejor rendimiento se tiene actualmente.

Para visualizar más a fondo las predicciones logradas, se compararán los Espectros de Potencia (Power Spectrum) obtenidos tanto a partir de la simulacíon TNG-300-1, las predicciones logradas con los modelos basados en funciones de péridas cuantilicas y el modelo nueronal del trabajo de Rodrigues et al. [6]. Para comparar los valores de TNG junto a los valores predichos, se utilizaran las posiciones del centro de sus halos. Los espectros a comparar se centran en $P(M_*, Colorg - i)$, es decir, en las características de masa estelar (M_*) y color ($Colorg - i$). Adicionalmente, se comparará la medida de ajuste χ^2, donde un menor valor indicará un mejor ajuste al modelo original TNG.

4.3 Arquitectura

Se proponen 2 arquitecturas acordes a la loss en cuestión, donde cada una se enfoca en predecir cada característica de manera individual, NN_{smass} y NN_{color}, para también poder calcular su unión NN_{union}, y su forma conjunta NN_{join}. Se utilizará NN_{class} como referencia, el cual viene del trabajo de Rodrigues et al. [6]

4.3.1 Suma de regresiones de cuantiles (SRC). Se propone un arquitectura de regresión, la cual consta de una red feed-forward de 3 capas ocultas, con activaciones no lineales ReLu. Es necesaria la definición de los cuantiles que generará esta arquitectura, donde se utilizan el percentil 0.25, 0.5 y 0.75 para cada una de las características a predecir, con los cuales se calcula la función de pérdida SRC, comparando cada salida del modelo, con su percentil específico, para luego realizar la suma de estos. Es por eso que es necesario una salida por cada percentil utilizado. En el caso de las características unitarias, se generan 3 predicciones por cada una, utilizando los percentiles mencionados anteriormente. Para el caso conjunto, son 6 predicciones, por lo cual se repiten los percentiles utilizados quedando [0.25, 0.5, 0.75, 0.25, 0.5, 0.75] para que la funcion se calcule correctamente, quedando las primeras 3 salidas correspondientes a la masa estelar y las últimas 3 al color $g - i$.

4.3.2 Aproximación continua de cuantiles (SQR). Arquitectura para tratar el problema en forma de regresión con 3 capas ocultas y activaciones no lineales ReLu, donde en la salida de las capas internas se realiza layer normalization para estabilizar el entrenamiento. Además, se utiliza el método definido en el trabajo de Tagasovska and Lopez-Paz [7] para la predicción de cuantiles de forma continua con un método de extensión de dimensionalidad para el aprendizaje de los cuantiles de cada fila del input del modelo. El modelo utiliza la función de pérdida SQR antes mencionada en su versión clásica para salida unidimensional como Masa estelar o Color g-i y en una versión modificada ligeramente para salidas de mayor dimensionalidad como Masa estelar y Color g-i en conjunto, definida como la función de pérdida SQR clásica (2), pero realizando un promedio de sus resultados.

Junto a esto, para este caso fueron entrenados los modelos con el dataset augmentado con SMOGN para mejorar las predicciones de los tipos de galaxias de menor representación en los datos de TNG300-1, donde se realiza el aumento de datos en conjunto con un muestreo estratégico al dividir el conjunto de datos con el fin de que los trazadores sean distribuidos lo más equitativamente posible (lo mencionado respecto a los trazadores y galaxias de menor representación se explicará en detalle en la sección 5).
Además de esto, se define un modelo extra de estimación conjunta con una función de pérdida SQR de alta dimensionalidad modificada para que reciba un castigo adicional basado en la métrica de Wasserstein entre los valores predichos y los valores originales, definida como la función SQR en alta dimensión como la anteriormente mencionada, pero agregando un castigo adicional de la distancia de Wasserstein entre las distribuciones de las predicciones y los valores originales.

5 EXPERIMENTOS

5.1 Métricas numéricas.

Con los datos para realizar las pruebas, se compara el rendimiento de los modelos Se utilizará el test de Kolmogorov-Smirnov como antes se menciona en base a la distribución del conjunto de predicciones de características, pero además se utilizarán otras métricas como el MSE, MAE, RMSE, R^2, MAD y MAPE para la medición de la calidad de las predicciones de forma puntual a la vez.

5.2 Power Spectrum.

Como se mencionó en la sección anterior, la construcción de los Power Spectrum es el motivo por el cual es necesario la separación estratégica de los datos junto al proceso de SMOGN de augmentación de datos.

Para introducir este método de evaluación, es importante comprender qué son los Power Spectrum y cuál es su significado e importancia para los estudios cosmológicos. Según un clásico libro de la cosmología, Dodelson and Schmidt [4] explica los Power Spectrum en el contexto de la cosmología como el medio para estudiar las inhomogeneidades (variaciones o irregularidades) de la distribución de la materia. Matemáticamente, los Power Spectrum se calculan con la transformada de Fourier de una función de densidad de materia o temperatura, tal y como define Dodelson and Schmidt [4] en la ecuación (3).

$$\langle \bar{\delta}(\vec{k})\bar{\delta}(\vec{k'})\rangle = (2\pi)^2 P(k)\delta^3(\vec{k}-\vec{k'}) \tag{3}$$

Siendo $\delta(k)$ la transformada de Fourier de la fluctuación de densidad $\delta(x)$, δ^3 la función delta de Dirac, que restringe que los modos de onda sean iguales $(k - k')$. En base a esto, es posible llegar al cálculo de los Power Spectrum de los datos predichos. Ahora bien, este calculo es efectuado de forma práctica gracias a la misma librería utilizada en el trabajo de Rodrigues et al. [6], **NBODYKIT**, del trabajo realizado por Hand et al. [5].

Posteriormente a esta definición, es importante introducir el concepto de **trazador**, que en el contexto de la cosmología, Dodelson and Schmidt [4] lo define como los objetos o fenómenos utilizados para mapear la distribución de la materia en el Universo. Gracias a estos trazadores, es posible realizar el estudio estadístico en base a características de las galaxias sobre las cuales centrar el foco. Para objeto de esta investigación, se definen 7 trazadores en base a las dos características a predecir de las galaxias (masa estelar y color), donde se dividen en grupos las galaxias según restricciones conjuntas para masa estelar, $(9.5, 10.5], > 10.5 y \leq 9.5$ y para color $> 1.05, (0.80, 1.05] y \leq 0.80$. En base a estos trazadores es la separación estratégica realizada.

Con los datos predichos de cada modelo, se calculan y grafican los Power Spectrum de cada una de las predicciones, además de los valores originales y se comparan sus precisiones. Basado en el trabajo de Rodrigues et al. [6], se grafican los espectros de potencia de cada trazador definido en la **Tabla 1** y en la **Tablas 2** para SRC y SQR respectivamente para cada uno de los modelos. En ambas tablas se muestran también la cantidad de objetos por cada trazador para interpretar mejor los errores de cada caso. En conjunto con esto, se grafican también los valores de los residuos para cada modelo respecto al original, donde estos residuos se definen también de la misma forma que en el trabajo de Rodrigues et al. [6], es decir, como en la fórmula (4) para el cálculo de la varianza y (5) para los residuos definidas en su artículo. Adicionalmente, sus correspondientes valores de χ^2 para cada residuo graficado, tanto en un espacio reducido desde 0.1 hasta 0.4 como en el espacio completo.

$$\frac{\sigma^2_{\alpha,i}}{P^2_{\alpha,i}} = \frac{2}{V\tilde{V}}\left(\frac{1+\bar{n}_\alpha P_{\alpha,i}}{\bar{n}_\alpha P_{\alpha,i}}\right)^2, \tag{4}$$

Trazador	$\log(M_*[h^{-1}\,M_\odot])$	Color g-i	# Objects
$\alpha = 1$	(9.5, 10.5]	> 1.05	1497
$\alpha = 2$	> 10.5	> 1.05	1942
$\alpha = 3$	≤ 9.5	(0.80, 1.05]	1777
$\alpha = 4$	(9.5, 10.5]	(0.80, 1.05]	2300
$\alpha = 5$	> 10.5	(0.80, 1.05]	482
$\alpha = 6$	≤ 9.5	≤ 0.80	11172
$\alpha = 7$	(9.5, 10.5]	≤ 0.80	6908

Table 1: Distribución de objetos para el modelo de suma de regresiones de cuantiles SRC, donde se define cada trazador (α), el rango de masa estelar, el rango de color y la cantidad de objetos por trazador.

Trazador	$\log(M_*[h^{-1}\,M_\odot])$	Color g-i	# Objects
$\alpha = 1$	(9.5, 10.5]	> 1.05	3028
$\alpha = 2$	> 10.5	> 1.05	3924
$\alpha = 3$	≤ 9.5	(0.80, 1.05]	3643
$\alpha = 4$	(9.5, 10.5]	(0.80, 1.05]	4453
$\alpha = 5$	> 10.5	(0.80, 1.05]	944
$\alpha = 6$	≤ 9.5	≤ 0.80	22311
$\alpha = 7$	(9.5, 10.5]	≤ 0.80	13847

Table 2: Distribución de objetos para el modelo de aproximación continua de cuantiles SQR, donde se define cada trazador (α), el rango de masa estelar, el rango de color y la cantidad de objetos por trazador.

donde $\tilde{V} = 4\pi k_i^2 \Delta k/(2\pi)^3$, definiendo en conjunto la ecuación de los residuos (5).

$$\frac{\left(P^{\text{pred}}_{\alpha,i} - P^{\text{TNG300}}_{\alpha,i}\right)^2}{\sigma^2_{\alpha,i}}. \tag{5}$$

Además, para complementar las interpretaciones de los espectros de potencia, se grafican también los errores relativos para cada trazador y para cada modelo en cuestión, para ser lo más coherente con el trabajo de Rodrigues et al. [6] posible. Para el caso de la suma de regresiones de cuantiles, se hizo el cálculo con el percentil 50. Para los modelos SQR se utilizan como tal puesto que no se centra en un cuantil concreto.

6 RESULTADOS

Antes de analizar los resultados, es importante definir que los valores numéricos del KS-Test son obtenidos directamente del trabajo de Rodrigues et al. [6] y dado que en el mismo trabajo no se especifíca concretamente el método de cálculo del KS Test para la alta dimensión, se evalúan en este caso dos formas de cálculo, *flat* donde se realiza un aplanamiento de las salida para 1 dimensión y así compararles y *mean* donde se efectúa el KS Test por columnas individualmente y se promedian los resultados finales. Ambos casos son comparados con el valor obtenido del trabajo de Rodrigues et al. [6]. Además, para las otras métricas numéricas, se evalúa el modelo del mismo trabajo obtenido del GitHub entregado en la sección de

DATA AVAILABILITY del trabajo bajo el mismo conjunto de datos en que se evalúan los otros modelos, donde al ser el mismo conjunto de datos bajo los mismos criterios del trabajo de Rodrigues et al. [6], es aceptable el cálculo y la comparación de los modelos bajo estos criterios.

6.1 Modelos de funcion de pérdida SQR

6.1.1 Métricas numéricas. Se evalúa el resultado de las predicciones de cada modelo, donde NN_{mass} es el modelo construido con la función de pérdida SQR que predice únicamente masa estelar, NN_{color} es el modelo construido con la función de pérdida SQR que predice únicamente color g-i, NN_{joint} es el modelo construido con la función de pérdida SQR en alta dimensión que predice de forma conjunta tanto masa estelar como color g-i, NN_{joint_w} es el modelo construido con la función de pérdida SQR en alta dimensión con castigo adicional Wasserstein que predice de forma conjunta tanto masa estelar como color g-i, NN_{class} es el modelo mencionado en el trabajo de Rodrigues et al. [6].

Los valores del KS-Test presentados en la **Tabla 3** muestra los KS-Test de cada tipo de predicción, donde en gran medida son ambos tipos de modelos considerablemente similares y cercanos, igualandose en ordenes de magnitud en todos los casos exceptuando en las predicciones de unión, donde en ninguno de los dos métodos del cálculo del KS-Test, NN_{class} supera al modelo basado en SQR. Los resultados de la la **Tabla 4**, en primera parte, muestran un a clara tendencia que en predicciones individuales, el modelo de NN_{smass} tiene todas sus métricas con resultados altamente positivos, mientras que NN_{color} tiene bajos errores en general pero muestra un gran MAPE y muy bajo R^2.

Ahora bien, el centro de este análisis recae en los modelos de predicciones conjuntas (masa estelar y color), puesto que es el foco de las predicciones de NN_{class}. Tanto en MSE, MAE, RMSE y MAD, el modelo de NN_{joint_w} obtiene los mejores resultados, indicando un menor error general en sus predicciones, tanto para absolutos, medios y medianos. Para la métrica del R^2, también NN_{joint_w} indica un mejor resultado, pero no logra un valor particularmente alto en ninguno de los tres casos, lo que puede dar indicios de que no es una métrica adecuada para la evaluación bajo este contexto. Finalmente, el MAPE (Error Porcentual Medio Absoluto) toma un valor menor para las predicciones conjuntas clásicas NN_{joint}, demostrando un mejor rendimiento respecto a los otros dos modelos. Para cerrar el análisis, si bien el modelo NN_{joint_w} destaca por sobre los otros dos de predicciones conjuntas (NN_{joint} y NN_{class} de Rodrigues et al. [6]), es notable que los tres modelos tienen excelentes resultados, pero es remarcable lo logrado por los modelos SQR, es decir, competir directamente con el modelo NN_{class} en la calidad de las predicciones, demostrando el nivel de robustez de la función de pérdida SQR.

6.1.2 Gráficos de las predicciones del modelo. Las figuras 1 y 2 muestran gráficamente el cómo se distribuyen tanto masa como color respecto a los valores originales TNG300-1. Para el caso de la distribución de masa, las predicciones individuales se ajustan considerablemente bien a los datos originales, teniendo unos ligeros errores en las masas de menor magnitud al ser un cambio brusco las distribuciones originales en conjunto a otros errores en las de mayor magnitud que posiblemente son outlayers. Para el caso del color, la distribución predicha sigue considerablemente bien la forma de los datos originales, teniendo un error ligeramente mayor respecto a la masa en los valores de menor magnitud, junto a algunos pocos de mayor magnitud. Por su parte, para las distribuciones conjuntas representadas en las figuras 3, 5, 4 y 6, las tres predicciones muestran grandes resultados tanto en masa como en color, con ligeras diferencias que hacen destacar las predicciones de la masa estelar del modelo JOINT Wasserstein (NN_{joint_w}), pero las predicciones del color g-i son similares y de buena precisión, destacando ligeramente las de los modelos JOINT (NN_{joint}) y UNION (NN_{union}).

KS Test		
	SQR	NN_{class}
NN_{smass}	0.0072	**0.002**
NN_{color}	0.0094	**0.004**
$NN_{union}(flat)$	**0.0047**	0.010
$NN_{union}(mean)$	**0.0083**	0.010
$NN_{joint}(flat)$	0.00511	0.005
$NN_{joint}(mean)$	0.0095	**0.005**
$NN_{joint\_w}(flat)$	0.0067	**0.005**
$NN_{joint\_w}(mean)$	0.0108	**0.005**

Table 3: Resultados del KS Test de los modelos basados en SQR vs modelo base de NN_{class} del documento original de predicción conjunta de Rodrigues et al. [6]

Métricas de desempeño					
	NN_{smass}	NN_{color}	NN_{joint}	NN_{joint_w}	NN_{class}
MSE	0.0326	0.0610	0.0474	**0.0463**	0.0539
MAE	0.1398	0.1884	0.1655	**0.1640**	0.1790
RMSE	0.1806	0.2470	0.2178	**0.2152**	0.2323
R^2	0.9158	0.0357	0.4709	**0.4851**	0.4452
MAD	0.1128	0.1489	0.1324	**0.1318**	0.1442
MAPE	1.47	56.64	**27.56**	28.12	28.91

Table 4: Resultados de cada uno de los modelos producidos de SQR, comparados con el modelo base de NN_{class} del documento original de predicción conjunta de Rodrigues et al. [6]

6.1.3 Power spectrum. Primeramente, en la figura 7, los espectros de potencia muestran un ajuste considerablemente acertado a lo largo de la mayoría del gráfico con excepción de los extremos, específicamente en la parte final de la curva, cosa que se muestra en el gráfico de los residuos, los cuales van en subida continua, demostrando que los errores se disparan en la sección final del espectro. Los errores en la sección final, es decir, en valores de k mayores de altas frecuencias, demostrando que para las zonas donde se necesita una precisión mayor por las características de sus estructuras al ser más pequeñas, todos los modelos coinciden en problemas de ajuste al espectro original. Con diferencia, el trazador con mayor error es el 5, pues es el tipo de galaxias con menor representación en los datos de TNG300-1, incluso tras el proceso de aumentado de datos SMOGN, lo contrario pasa para el trazador 6, que es el tipo de galaxias con mayor representación, demostrando el menor error entre todos los trazadores. Además de esto, los

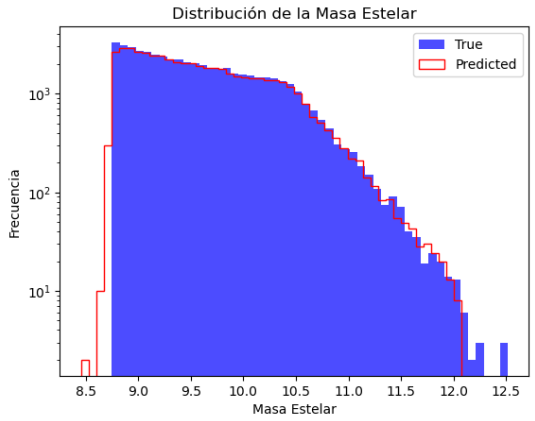

Figure 1: Distribución individual de la masa estelar. Masa estelar verdadera (barras azules solidas) y masa estelar predicha (barras rojas vacías) para la predicción del modelo basado en las funciones de pérdida SQR.

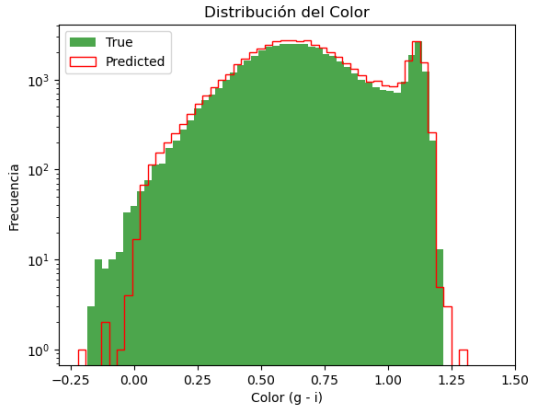

Figure 2: Distribución individual del color de las galaxias. Color g-i verdadero (barras verdes solidas) y color g-i predicho (barras rojas vacías) para la predicción del modelo basado en las funciones de pérdida SQR.

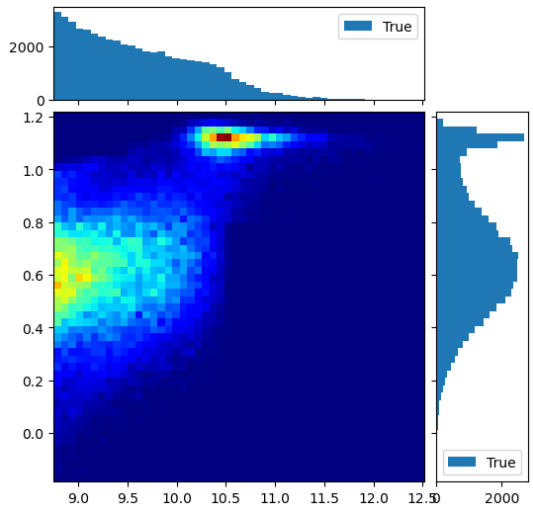

Figure 3: Distribución conjunta real TNG300-1 de color y masa estelar de las galaxias para los datos aumentados SMOGN utilizada para los modelos basados en las funciones de péridad SQR.

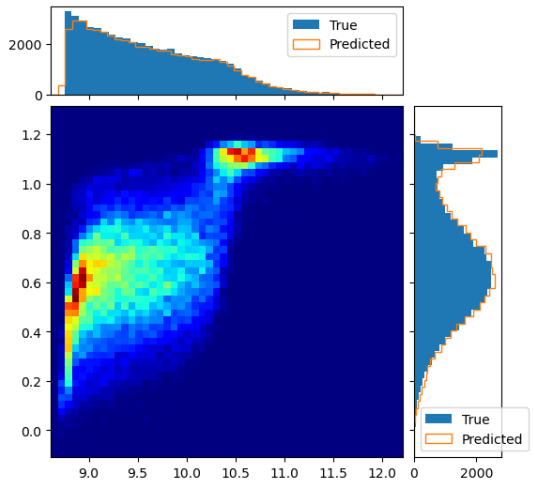

Figure 4: Distribución conjunta predicha para el modelo JOINT de color y masa estelar de las galaxias para los modelos basados en las funciones de péridad SQR.

valores de χ^2 son disparados en todos los casos pues al tener una gran cantidad de datos y por el crecimiento exponencial de estos valores con los pequeños errores, donde la sección final de mayor problema de ajuste proboca la mayor parte del error de los χ^2. Ahora bien, para la figura 8, que consisten en los mismos graficos representados en la figura 7 pero acotados en el mismo rango de los gráficos de los Power Spectrum presentados en el trabajo de Rodrigues et al. [6], tiene la misma tendencia al ser los mismos gráficos, pero los χ^2 tienen un valor más sencillo de interpretar y más representativo del desempeño general del modelo. Para los trazadores 1, 5 y 6, el modelo NN_{joint} obtiene menores errores en sus aproximaciones generales, demostrando un mejor desempeño general en los trazadores 5 y 6, coincidentemente los trazadores de las galaxias de menor y mayor repesentación respectivamente. Por su parte, para los trazadores 2 y 3, el modelo de mejor desempeño es NN_{joint_w}, pero con un mayor inestabilidad en el trazador 5. Para el trazador 4 y 7, el modelo de mejor desempeño es el NN_{union} con un ligero margen. Ahora bien, el modelo NN_{class} no se queda atrás, pues si bien no obtiene el mejor resultado exactamente en ningún trazador específico, obtiene resultados mucho más robustos a niveles generales, teniendo un desempeño digno de destacar en todos los trazadores que compiten con cada uno de los modelos a su manera. Finalmente, en la figura 9, todos los modelos muestran un comportamiento muy similar y tienen errores bastante bajos, exceptuando en el trazador número 5, lo que se explica por la

baja representación de este en el conjunto de datos de este tipo de galaxias.

6.2 Modelos de funcion de pérdida mediante suma de regresiones de cuantiles

6.2.1 Métricas numéricas. Se evaluó el rendimiento de los modelos generados para la masa estelar y el color de forma individual, unida y conjunta, considerando los valores obtenidos vistos en el trabajo de Rodrigues et al. [6]. Los resultados se pueden ver en la **Tabla 5**. Los valores del KS-Test presentados en la **Tabla 6** muestra los KS-Test de cada tipo de predicción, donde para ningun caso el modelo basado en SRC obtiene un KS Test lo suficientemente bueno como para competir con NN_{class}, manteniédose en general a un orden de magnitud de distancia en todos los casos.

Los resultados de la la **Tabla 5**, también el objeto del análisis son los modelos de predicciones conjuntas (masa estelar y color), puesto que es el foco de las predicciones de NN_{class}. Tanto en MSE, RMSE y R^2, el modelo de NN_{class} obtiene los mejores resultados, indicando un menor error general en sus predicciones para absolutos

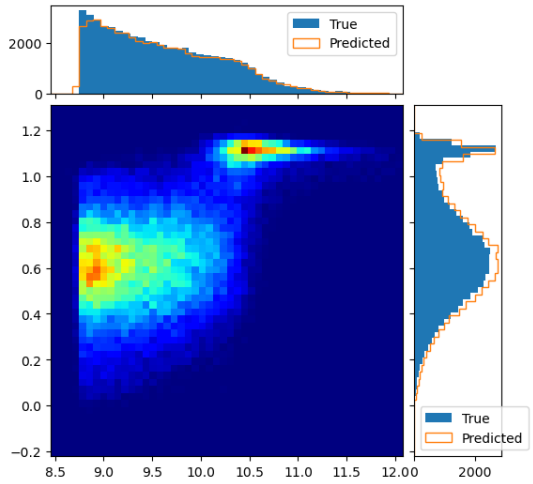

Figure 5: Distribución conjunta predicha para el modelo UNION de color y masa estelar de las galaxias para los modelos basados en las funciones de péridad SQR.

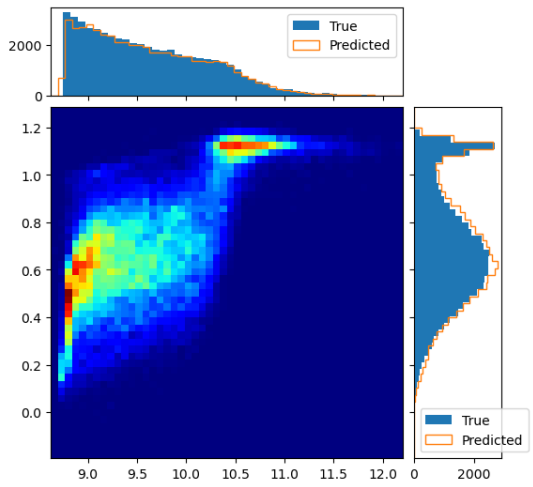

Figure 6: Distribución conjunta predicha para el modelo JOINT Wasserstein de color y masa estelar de las galaxias para los modelos basados en las funciones de péridad SQR.

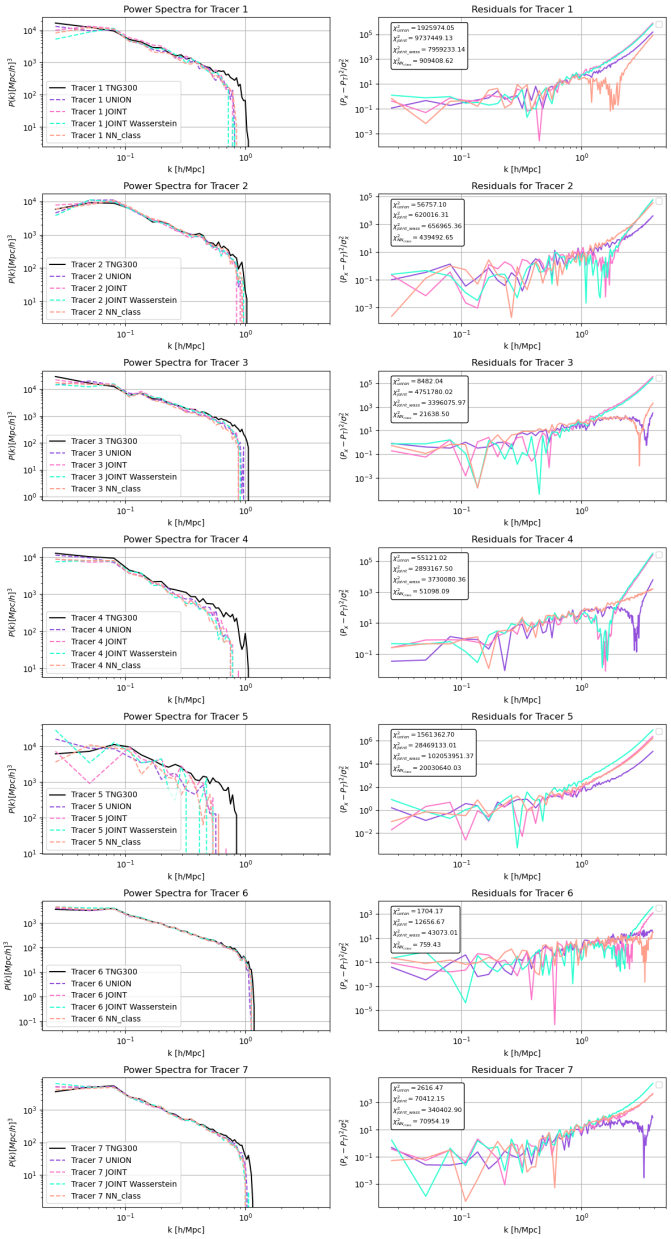

Figure 7: Cálculos de los espectros de potencia para los modelos basados en SQR. Junto a ellos, sus residuos correspondientes respecto a los espectros de potencia construidos a partir de TNG300-1. Sin límite de ejes.

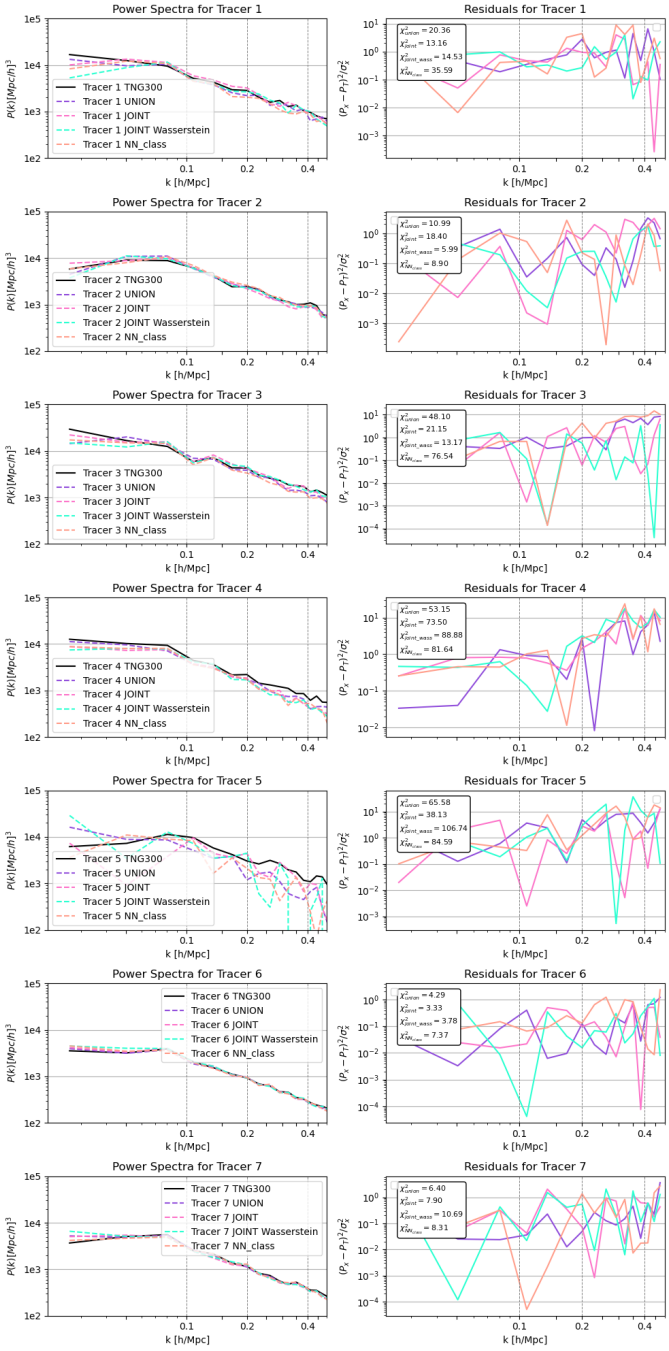

Figure 8: Cálculos de los espectros de potencia para los modelos basados en SQR. Junto a ellos, sus residuos correspondientes respecto a los espectros de potencia construidos a partir de TNG300-1. Limitado en sus ejes.

medios. Para las métricas MAE, MAD y MAPE, NN_{joint} obtiene mejores resultados para las predicciones conjuntas, demostrando un rendimiento competitivo respecto al modelo NN_{class}, con menores errores absolutos medios y medianos.

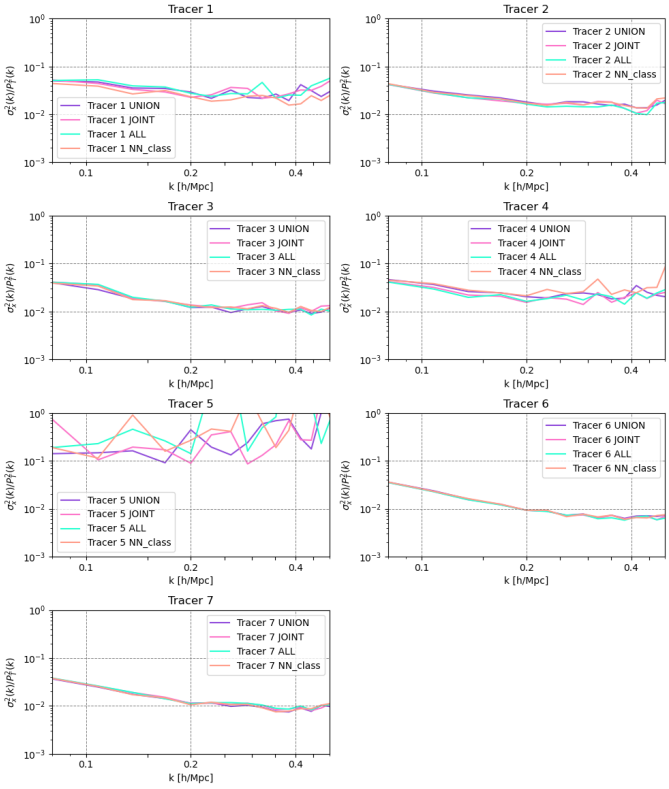

Figure 9: Errores relativos de los espectros de potencia de los modelos basados en SQR.

Métricas de desempeño				
	NN_{mass}	NN_{color}	NN_{joint}	NN_{class}
MSE	0.1417	0.1797	0.3110	**0.0550**
MAE	0.1198	0.1570	**0.1385**	0.1805
RMSE	0.3764	0.4240	0.5576	**0.2345**
R^2	0.9380	0.3426	0.6278	**0.4338**
MAD	0.0967	0.1257	**0.1129**	0.1453
MAPE	1.26	54.29	**30.03**	32.66

Table 5: Resultados de la suma de las regresiones de cuantiles (SRC) vs modelo base de NN_{class} del documento original de predicción conjunta de Rodrigues et al. [6]

Para este caso, los modelos obtienen métricas relativamente similares para cada caso exceptuando para MSE, donde el modelo NN_{class} logra un resultado de un orden de magnitud completo menor respecto a NN_{joint}, demostrando una importante tendencia a mejores resultados en sus predicciones.

6.2.2 Gráficos de las predicciones del modelo. Las figuras 10 y 11 muestran gráficamente el cómo se distribuyen tanto masa como color respecto a los valores originales TNG300-1. Para el caso de la distribución de masa, las predicciones individuales correctamente desde las masas entre 8.5 y 9.0 en adelante, pero con un mayor

KS Test		
	SRC	NN_{class}
NN_{smass}	0.044	**0.002**
NN_{color}	0.206	**0.004**
$NN_{union}(flat)$	0.103	**0.010**
$NN_{union}(mean)$	0.125	**0.010**
$NN_{joint}(flat)$	0.097	**0.005**
$NN_{joint}(mean)$	0.113	**0.005**

Table 6: Resultados del KS Test de la suma de las regresiones de cuantiles (SRC) vs modelo base de NN_{class} del documento original de predicción conjunta de Rodrigues et al. [6]

error de ajuste respecto a los gráficos anteriores SQR en las frecuencias pero sobre todo en los valores de masas menores a 8.5, donde realmente no existen masas de esas magnitudes y el modelo predice una cantidad importante a tener en cuenta en esa zona. Para el caso del color, las frecuencias de magnitudes predichas logran captar la forma de la distribución pero no ajustar correctamente las cantidades, demostrando la principal fuente de error de este modelo con problemas considerables en predicciones del color. Por su parte, para las distribuciones conjuntas representadas en las figuras 12, 14 y 13, las predicciones de masa estelar del modelo NN_{joint} logra resultados aceptables mientras que el modelo de predicciones individuales UNION NN_{union} obtiene resultados peores respecto a NN_{joint} y por supuesto respecto a los originales, pero el principal problema está en las predicciones del color, donde al tener problemas tan grandes en los resultados, perjudica al gráfico conjunto en su completitud, motivo que más adelante demostrará la gran importancia de este problema en las gráficas de los Power Spectrum.

6.2.3 Power spectrum. En las figuras 15, 16 y 17 se muestran los espectros de potencia de los modelos basados en la función de pérdida SRC, sus residuos y errores relativos respectivamente. Existe poco que mencionar en esta sección, pues dado los grandes errores en las predicciones del color de las galaxias de estos modelos, al filtrar los datos por trazadores surgen problemas considerables en la construcción de los espectros, llevando a errores sustanciales.

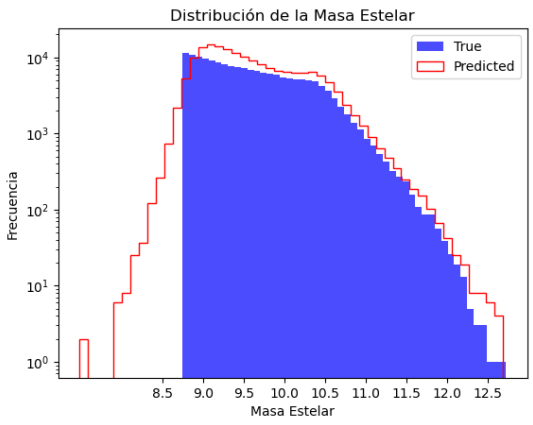

Figure 10: Distribución individual de la masa estelar. Masa estelar verdadera (barras azules solidas) y masa estelar predicha (barras rojas vacías.

7 CONCLUSIONES

En conclusión, los modelos construidos basados en SRC no logran resultados efectivos, donde si bien sus métricas numéricas no son especialmente negativas, en comparación a los modelos $NNclass$ se quedan cortos especialmente en sus predicciones del color, dando problemas considerables que impiden su uso práctico real. Por su parte, los modelos basados en SQR logran resultados prometedores en comparación a NN_{class}, logrando asemejarse considerablemente al modelo objetivo y competirle en rendimiento y calidad de las soluciones. Para las construcciones de los espectros de potencia, el modelo basado en SRC obtiene resultados que dejan basante que desear, pero por su lado el modelo basado en SQR logra resultados importantes y positivos, superando en precisión en algunos

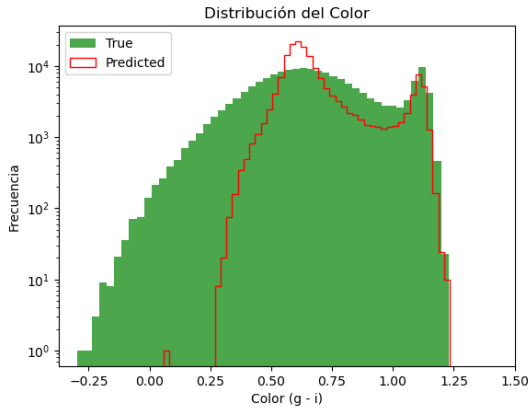

Figure 11: Distribución individual del color de las galaxias. Color g-i verdadero (barras verdes solidas) y color g-i predicho (barras rojas vacías).

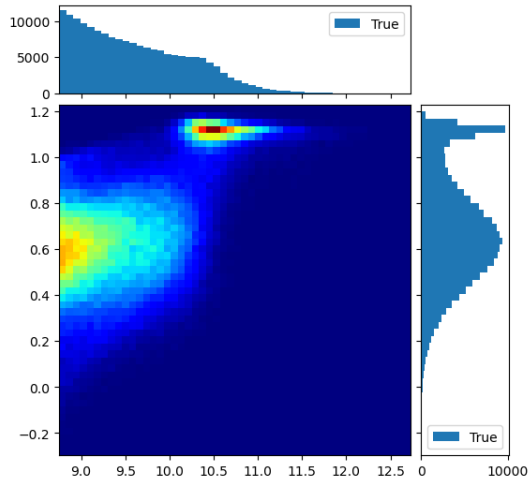

Figure 12: Distribución conjunta real TNG300-1 de color y masa estelar de las galaxias.

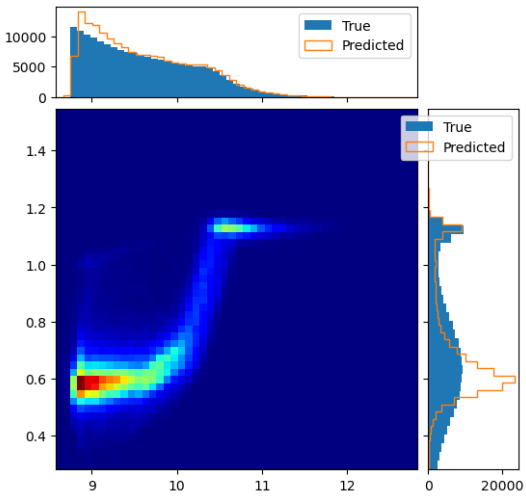

Figure 13: Distribución conjunta predicha para el modelo JOINT de color y masa estelar de las galaxias.

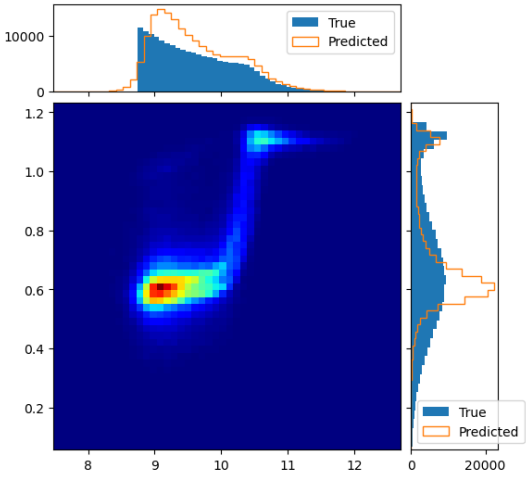

Figure 14: Distribución conjunta por unión predicha para el modelo UNION de color y masa estelar de las galaxias.

trazadores al modelo NN_{class}.
Claro esta que aún existen muchos puntos a mejorar con métodos diferentes a los ya estudiados, como pueden ser modelos enfocados en los espectros de potencia, para lograr un ajuste con menor error, modelos que necesiten menor cantidad de datos para lograr buenos resultados o métodos de tratamiento de datos para reducir los problemas en las galaxias menos representadas en TNG300-1. Respectos a las funciones de pérdidas cuantilicas y en respuesta a la hipótesis, la función de mejores resultados SQR demuestra un gran desempeño y confirma lo expuesto en el trabajo de Tagasovska and Lopez-Paz [7], donde para datos complejos como lo son los de TNG300-1 logra resultados positivos, tanto en su versión original como en su versión implementada de alta demensionalidad,

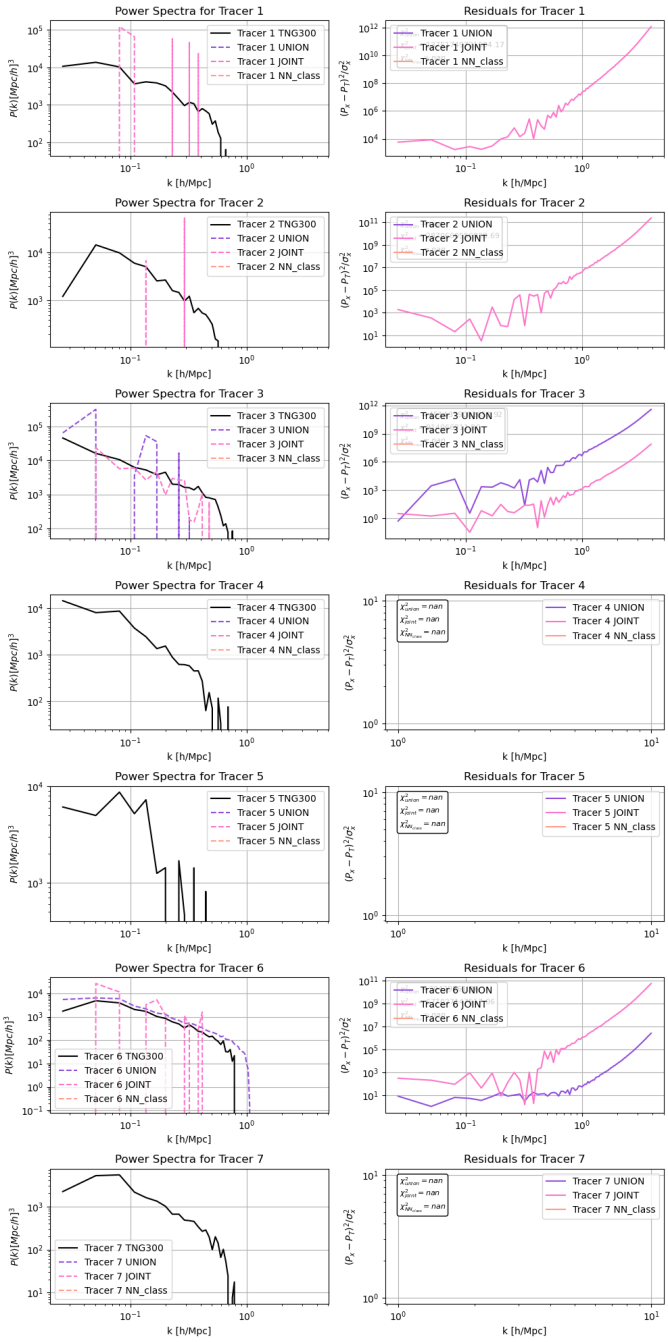

Figure 15: Cálculos de los espectros de potencia para los modelos basados en SCR. Junto a ellos, sus residuos correspondientes respecto a los espectros de potencia construidos a partir de TNG300-1. Sin límite de ejes.

incluyendo su punto extra de reducción en los tiempos de entrenamiento. Además de esto, un punto a destacar de los modelos basados en SQR es que considerando su calidad de predicciones, el tiempo que les conlleva realizarlas en considerablemente menor

al del modelo NN_{class} considerando predicción en conjunto con interpretación numérica de los resultados, es decir, la interpretación de la probabilidad predicha por el modelo NN_{class}.

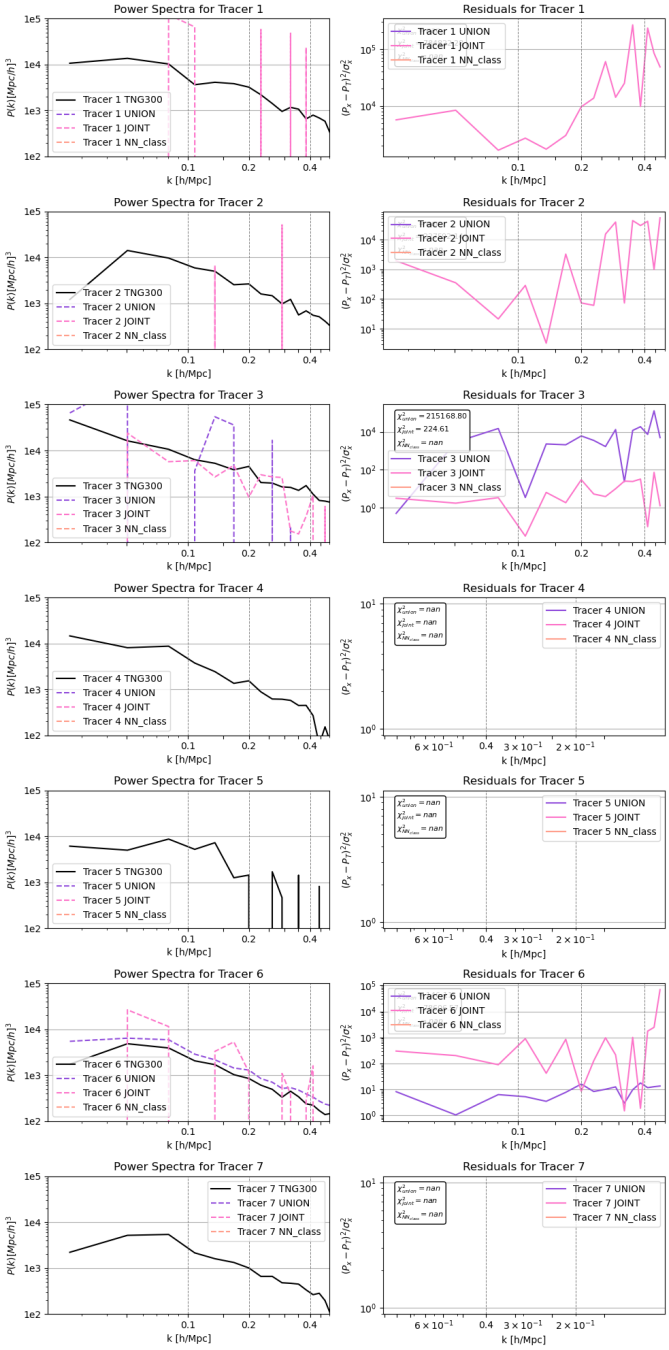

Figure 16: Cálculos de los espectros de potencia para los modelos basados en SCR. Junto a ellos, sus residuos correspondientes respecto a los espectros de potencia construidos a partir de TNG300-1. Limitado en sus ejes.

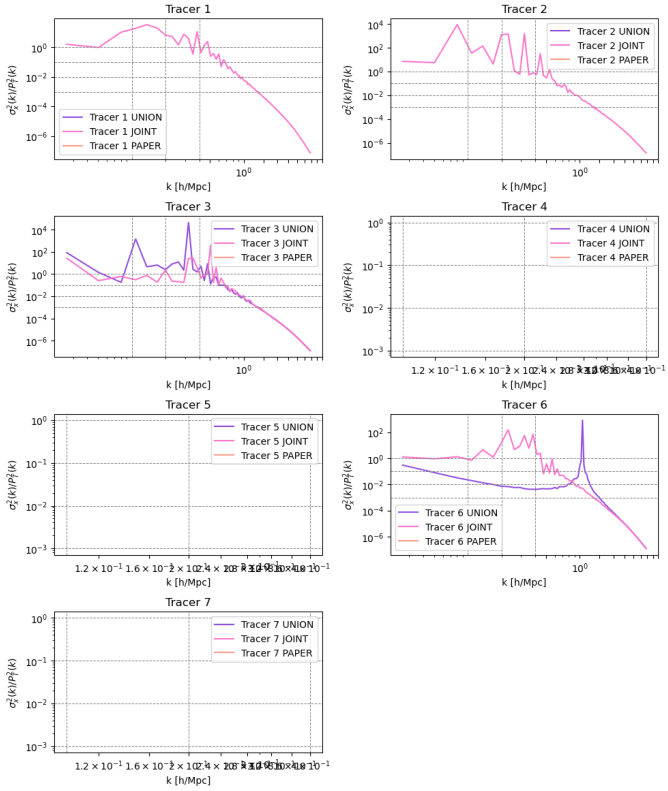

Figure 17: Errores relativos de los espectros de potencia de los modelos basados en SCR.

8 DISPONIBILIDAD DE LOS RESULTADOS

Todos los modelos entrenados en este trabajo se encuentran en el siguiente repositorio de GitHub:
https://github.com/iogurth/QuantilLossHaloGalaxy

9 REFERENCIAS

[1] James S Bullock, Avishai Dekel, Tsafrir S Kolatt, Andrey V Kravtsov, Anatoly A Klypin, Cristiano Porciani, and Joel R Primack. 2001. A universal angular momentum profile for galactic halos. *The Astrophysical Journal* 555, 1 (2001), 240.

[2] Harry George Chittenden and Rita Tojeiro. 2022. Modelling the galaxy–halo connection with semi-recurrent neural networks. *Monthly Notices of the Royal Astronomical Society* 518, 4 (11 2022), 5670–5692. https://doi.org/10.1093/mnras/stac3498 arXiv:https://academic.oup.com/mnras/article-pdf/518/4/5670/47839727/stac3498.pdf

[3] Natalí S M de Santi, Natália V N Rodrigues, Antonio D Montero-Dorta, L Raul Abramo, Beatriz Tucci, and M Celeste Artale. 2022. Mimicking the halo–galaxy connection using machine learning. *Monthly Notices of the Royal Astronomical Society* 514, 2 (May 2022), 2463–2478. https://doi.org/10.1093/mnras/stac1469

[4] Scott Dodelson and Fabian Schmidt. 2020. *Modern cosmology*. Academic press.

[5] Nick Hand, Yu Feng, Florian Beutler, Yin Li, Chirag Modi, Uroš Seljak, and Zachary Slepian. 2018. nbodykit: An open-source, massively parallel toolkit for large-scale structure. *The Astronomical Journal* 156, 4 (2018), 160.

[6] Natália V N Rodrigues, Natalí S M de Santi, Antonio D Montero-Dorta, and L Raul Abramo. 2023. High-fidelity reproduction of central galaxy joint distributions with neural networks. *Monthly Notices of the Royal Astronomical Society* 522, 3 (April 2023), 3236–3247. https://doi.org/10.1093/mnras/stad1186

[7] Natasa Tagasovska and David Lopez-Paz. 2019. Single-Model Uncertainties for Deep Learning. arXiv:1811.00908 [stat.ML]

[8] John F. Wu and Christian Kragh Jespersen. 2023. Learning the galaxy-environment connection with graph neural networks. arXiv:2306.12327 [id='astro-ph.IM' full$_{n}ame$ =' $InstrumentationandMethodsforAstrophysics'is_{a}ctive$ = $Truealt_{n}ame$ = $Nonein_{a}rchive$ =' $astro$ –

$ph' is_general = False description = ' Detector and telescope design, experiment proposals. Laboratory Astrophysics. Methods for data distribution. Statistical methods. Software and data quality and control '.

[9] Xing Yan, Yonghua Su, and Wenxuan Ma. 2023. Ensemble Multi-Quantiles: Adaptively Flexible Distribution Prediction for Uncertainty Quantification. arXiv:2211.14545 [cs.LG]