



S-PLUS

Southern Photometric  
Local Universe Survey

# Reconstruction of the Large-Scale Structure of the Universe

## Using photometric redshifts and machine learning techniques

Erik V., Laerte Sodré Jr.

Instituto de Astronomia, Geofísica e Ciências Atmosféricas

21 de fevereiro de 2025



# Visão geral

---

# Visão geral

- **Introdução**
  - ▶ Contexto cosmológico
  - ▶ Estrutura em larga escala
  - ▶ Redshifts espectroscópicos e fotométricos
- **Objetivo**
- **Dados**
  - ▶ Fotométricos e espectroscópicos
  - ▶ Pré-processamento
  - ▶ Pesos por objeto
- **Metodologia**
  - ▶ Redes neurais
  - ▶ Redes neurais Bayesianas
  - ▶ Redes de mistura de densidades
- **Resultados**
  - ▶ Redshifts fotométricos
  - ▶ Funções de densidade de probabilidade
  - ▶ Estrutura em larga escala
- **Conclusões e perspectivas**

# Introdução

---

# O contexto cosmológico

**Estudar a formação e evolução da estrutura em larga escala (LSS) é fundamental para entender o Universo em que vivemos. A partir dela podemos ter insights sobre:**

- Matéria escura
- Energia escura
- Evolução cósmica
- Formação e evolução de galáxias
- Física a nível fundamental

# Estrutura em larga escala

A teia cósmica, ou a LSS, é formada por diferentes componentes. Cada componente apresenta características e possui papéis diferentes na evolução do Universo como um todo.

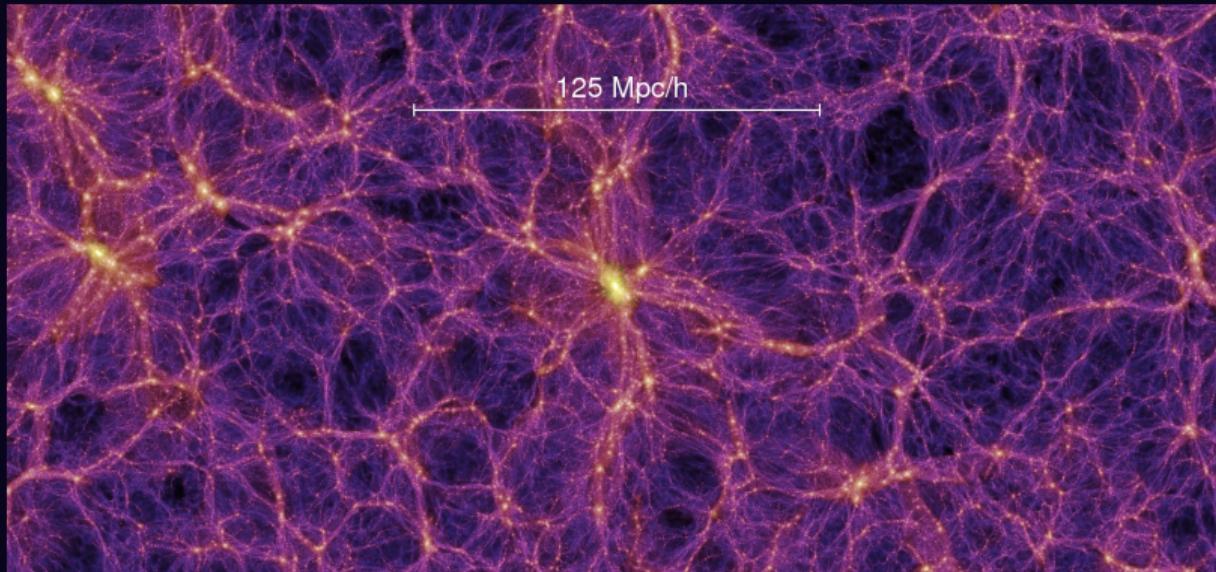


Figura 1: Adaptado de <https://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium/>.

# Estrutura em larga escala

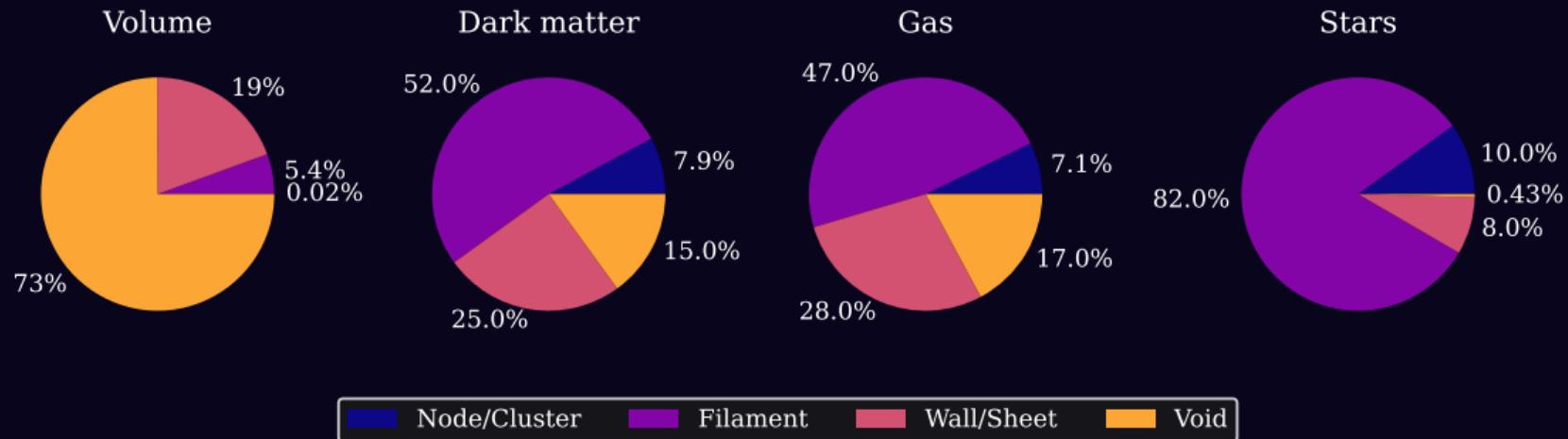


Figura 2: Adaptado de: Ganeshaiyah Veena et al. (2019).

# Estrutura em larga escala



O problema é que estas estruturas são tridimensionais, mas nossas observações são projetadas (bidimensionais)

# Redshifts espectroscópicos e suas limitações

Uma forma de determinar a distância de objetos celestes vêm da lei de Hubble Lemaître (válida para objetos próximos):

$$v_{res} = c \cdot z = H_0 \cdot D$$

Velocidade de recessão [km/s]      Redshift      Distância [Mpc]

Velocidade da luz [km/s]      Cte. Hubble [km/s/Mpc]

A observação de um espectro com alto sinal ruído demanda bastante tempo de observação, ainda mais para objetos fracos

É necessário encontrar uma alternativa

# A alternativa: redshifts fotométricos

Existem uma série de projetos que se baseiam em fotometria para produzir ciência.

- Pode ser considerada uma "aproximação" do espectro
- É obtida de forma muito mais rápida
- Pode alcançar magnitudes mais fracas
- É capaz de gerar dados para muitos objetos e grandes áreas

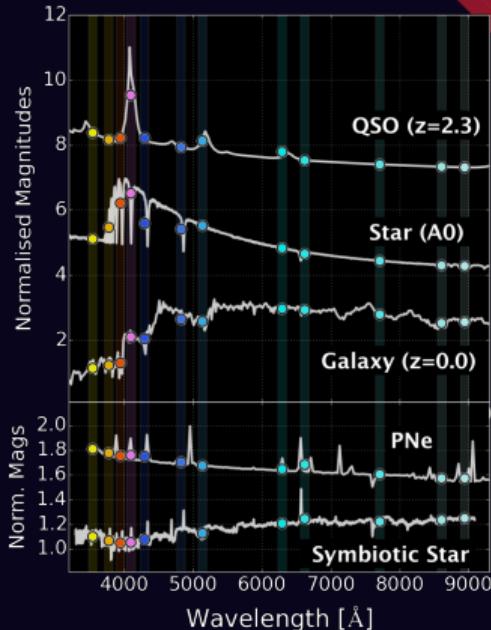


Figura 3: Adaptado de: Mendes de Oliveira et al. (2019).

# A alternativa: redshifts fotométricos

Redshifts fotométricos são estimados por três tipos de algoritmos: aprendizado de máquina, ajuste de templates e códigos híbridos.

## Aprendizado de máquina (ML)

- ✓ Se baseiam no uso de uma amostra de treinamento
- ✓ Flexível em relação a modelos (RFs, KNNs, SVMs)
- ✓ Podem ser mais precisos e rápidos que modelos de ajuste de template
- ✗ Não fornecem uma classificação do objeto (exceto caso seja configurado para isso)
- ✗ Sujeito à vieses devido aos dados

## Ajuste de templates (TF)

- ✓ Fazem uma comparação entre a fotometria de um objeto e uma biblioteca de templates
- ✓ Fornecem uma classificação do objeto junto ao  $z_{\text{phot}}$
- ✓ Maior capacidade de extração
- ✗ Menos precisos, mais lentos
- ✗ Sujeito à vieses devido à escolha dos templates

# O objetivo

---

# O objetivo

## Redshifts fotométricos

- Determinação de redshifts fotométricos de alta precisão
- Funções de densidade de probabilidade bem calibradas
- Galáxias até  $z = 0.8$  e magnitude 21 na banda r

## Estrutura em larga escala

- Utilizar dados fotométricos para a reconstrução da LSS
- Obter um mapeamento similar ao visto usando  $z_{\text{spec}}$

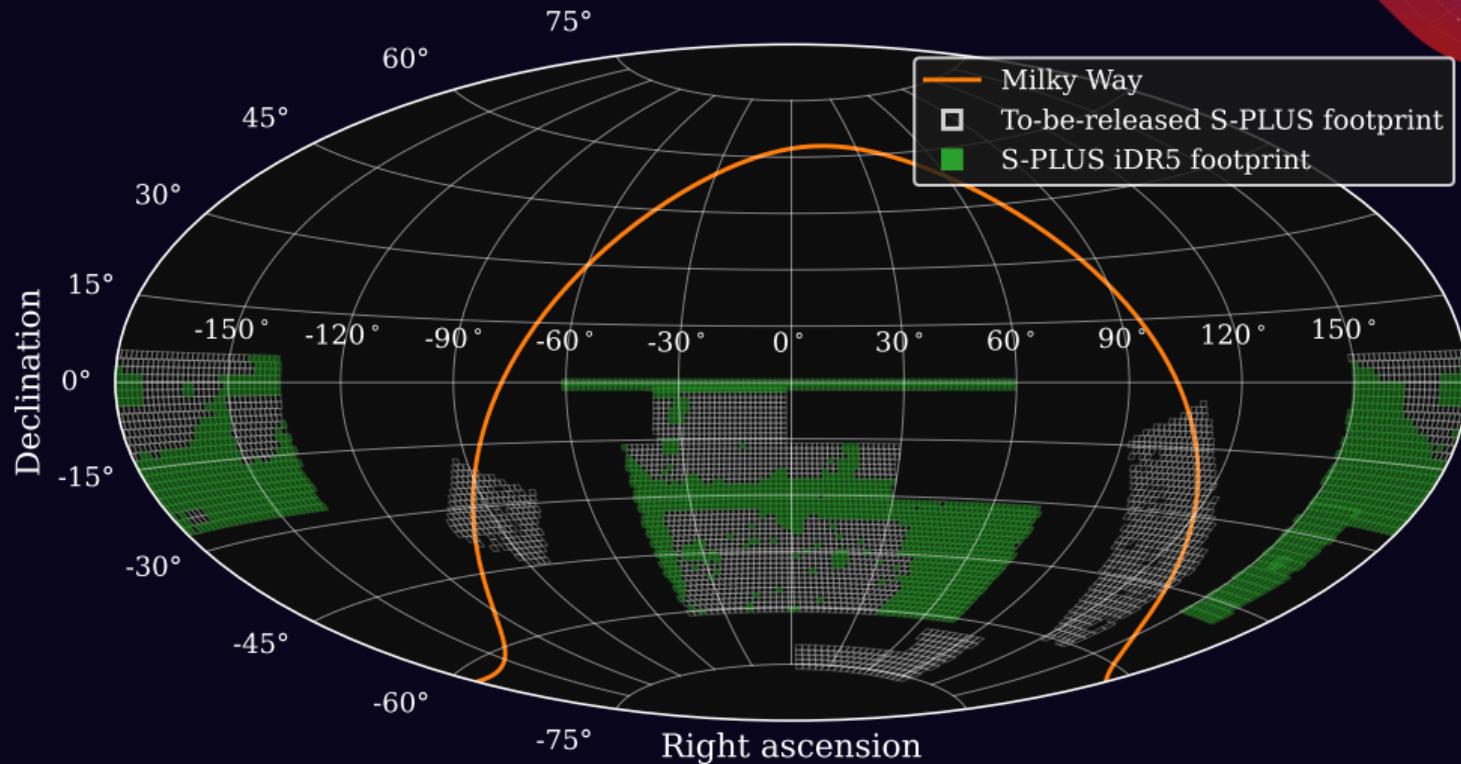
Expandir o conjunto de dados que podemos usar para estudos em diferentes áreas da astronomia

# Dados

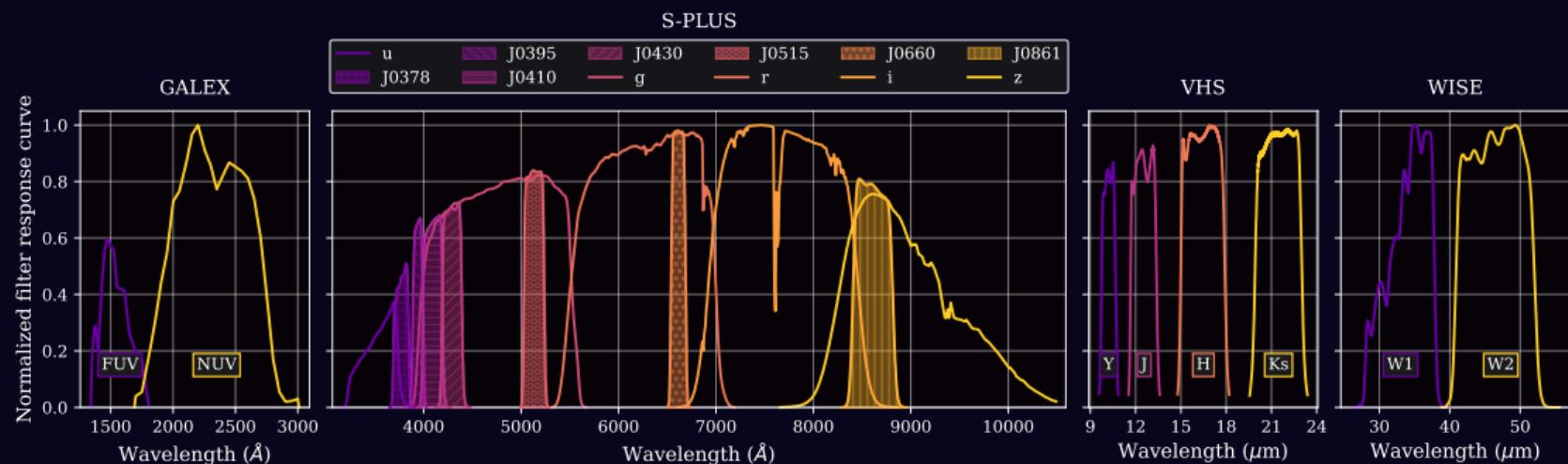
---



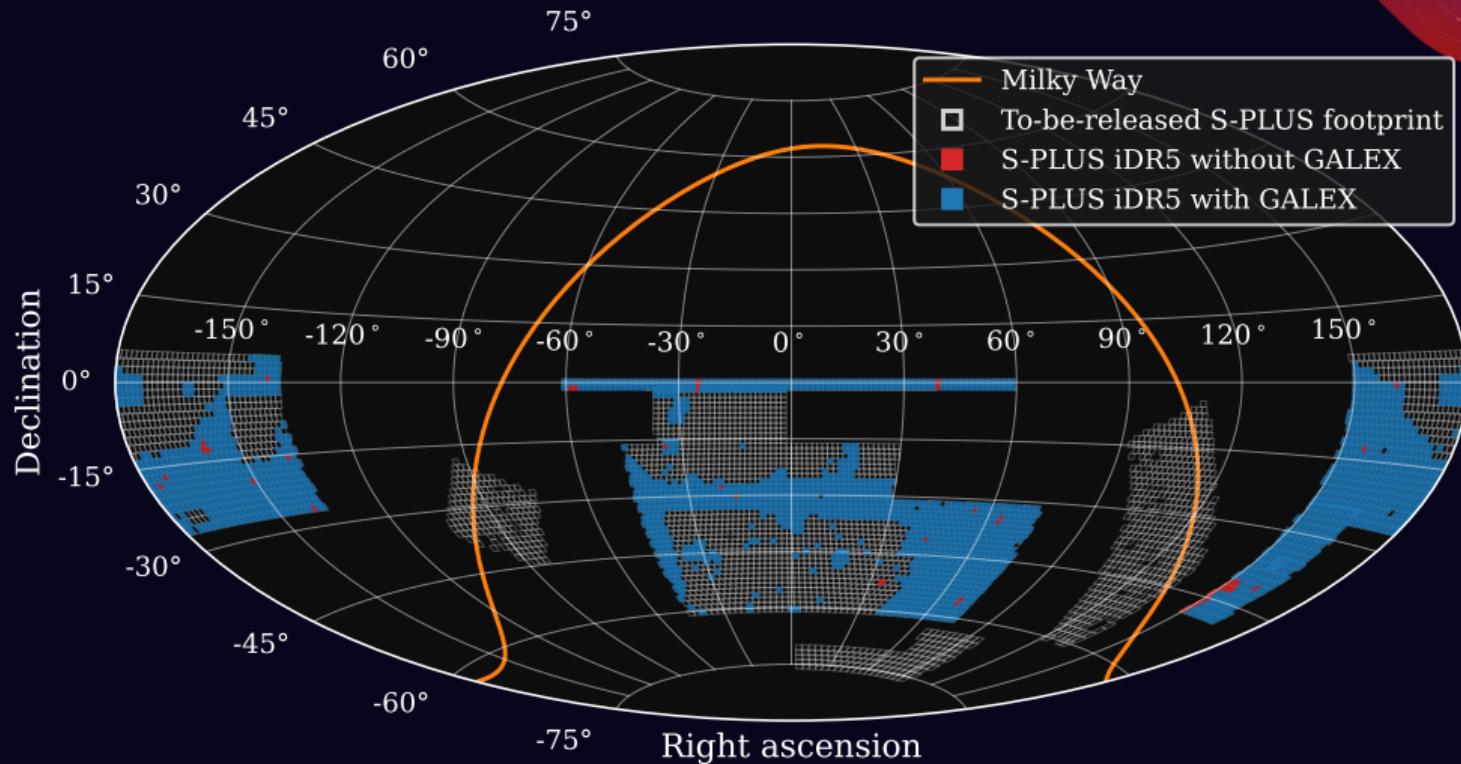
# Dados fotométricos



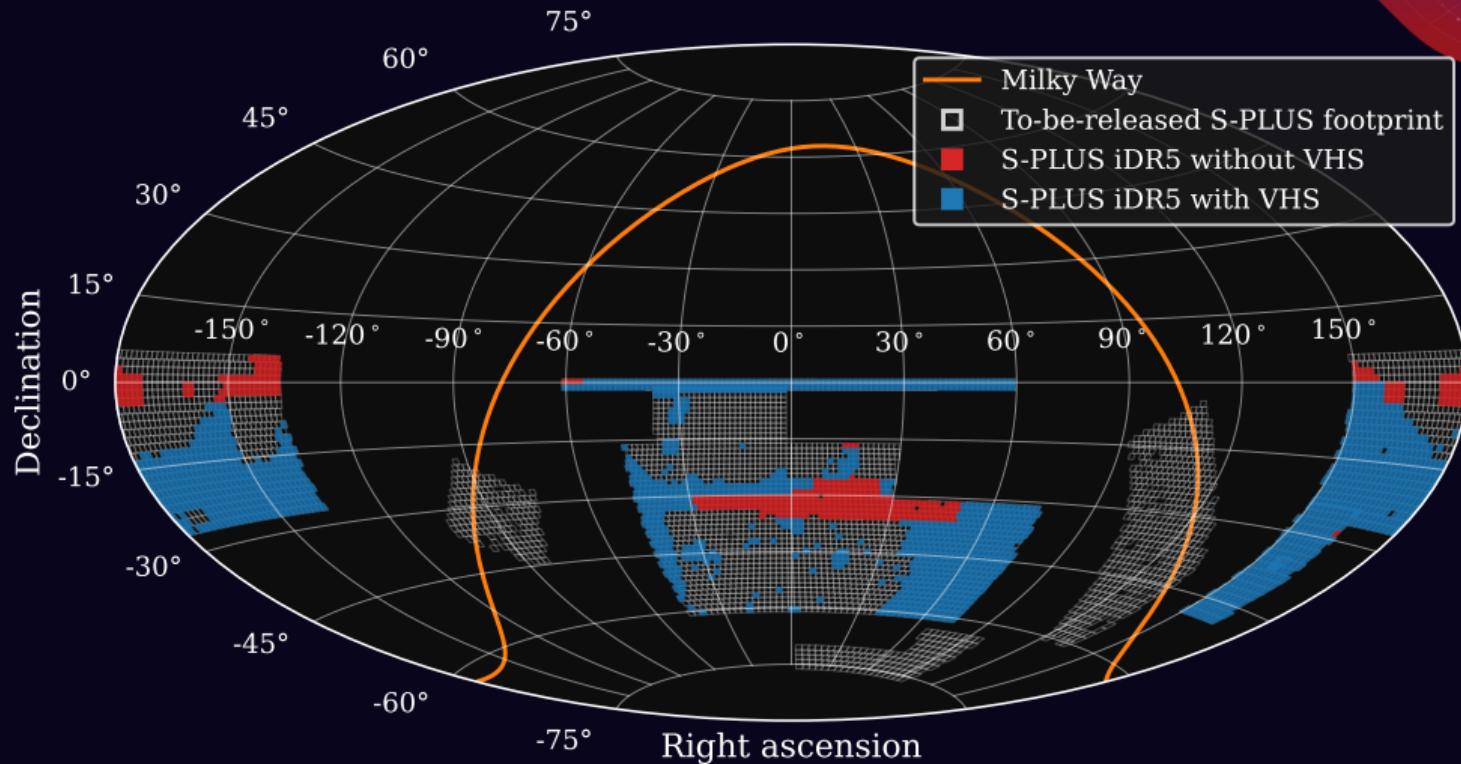
# Dados fotométricos



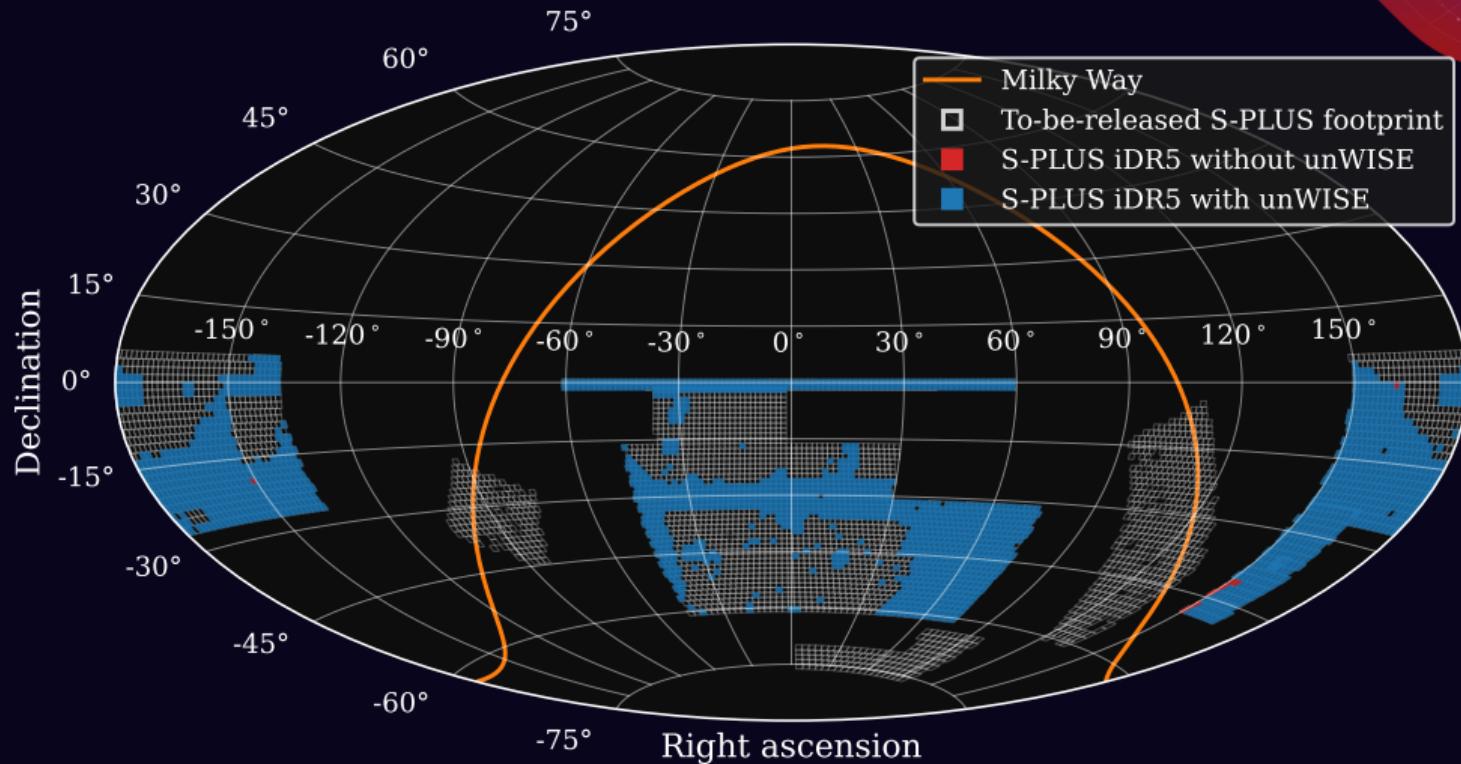
# Dados fotométricos



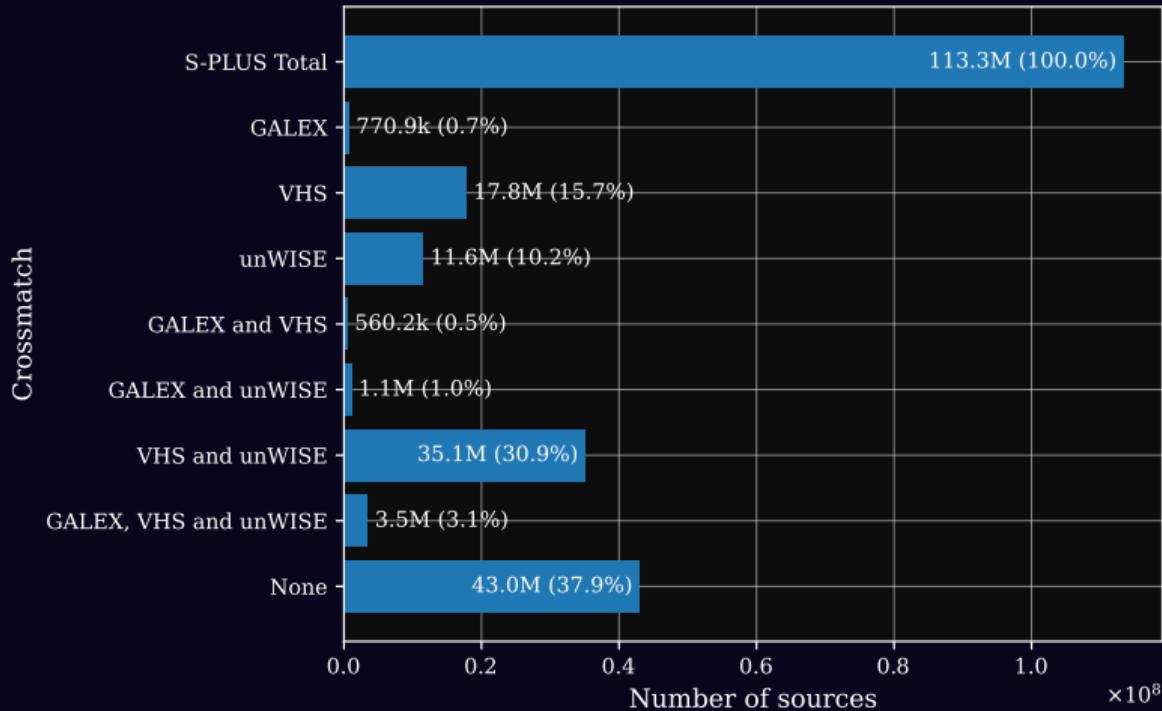
# Dados fotométricos



# Dados fotométricos



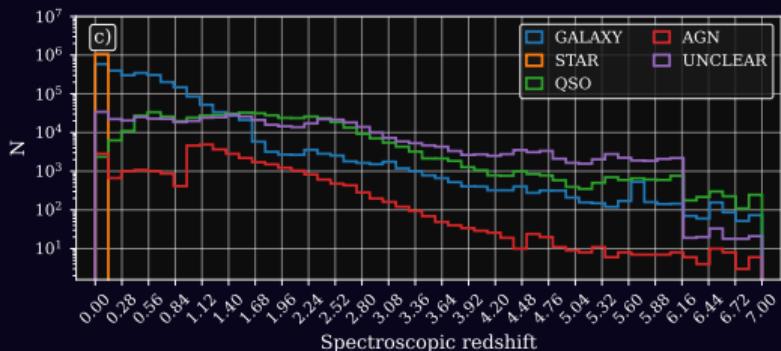
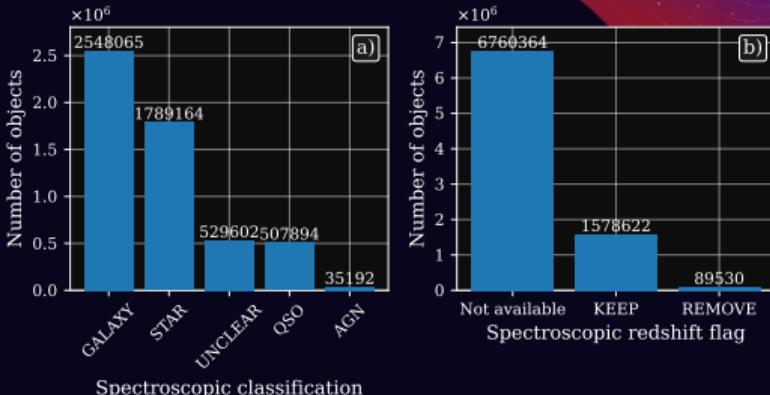
# Dados fotométricos



# Dados espectroscópicos ([https://github.com/ErikVini/specz\\_compilation](https://github.com/ErikVini/specz_compilation))

## Informações do compilado

- Total de catálogos: 5097
- Catálogos usados: 1872
- Total de objetos: 8 437 460
- Pré era GPT!

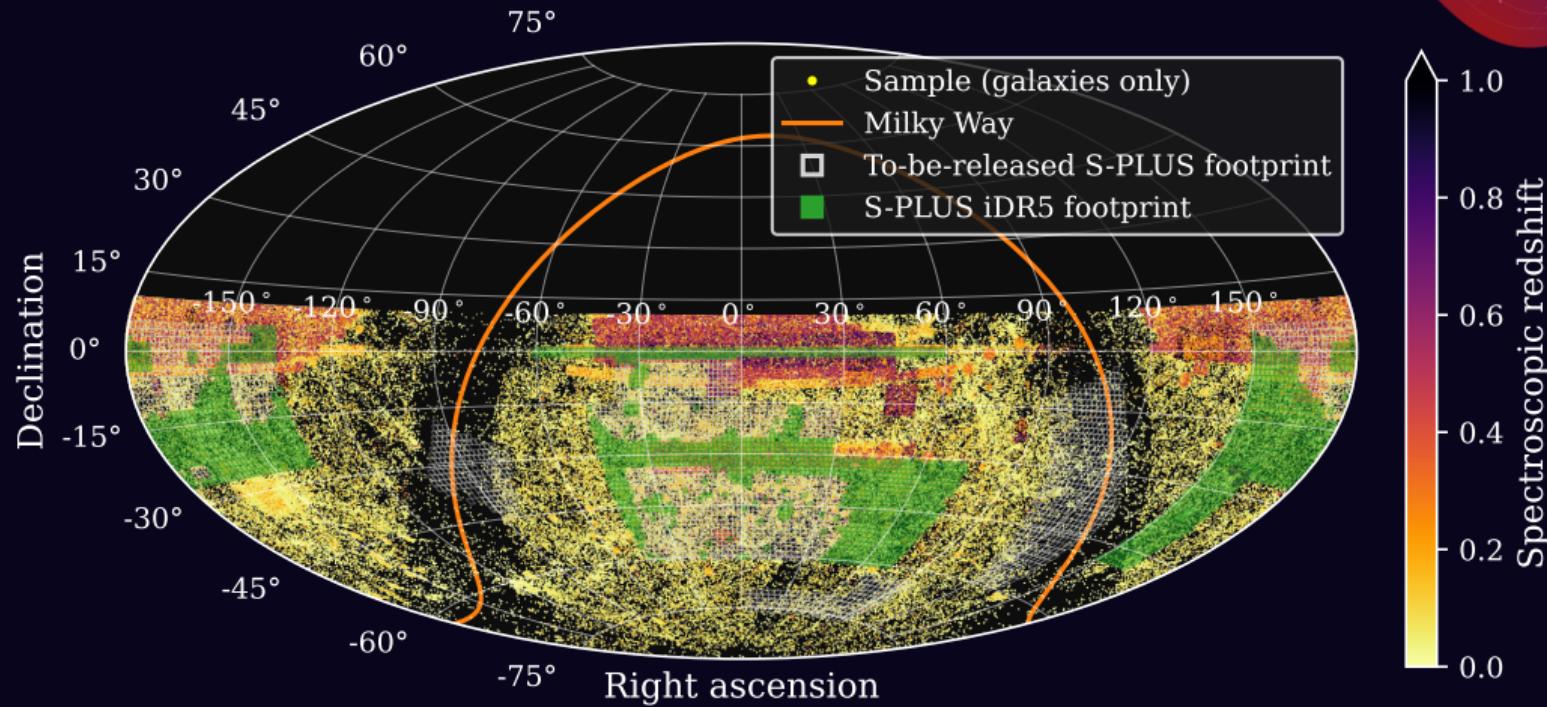


# Dados espectroscópicos

	ID	RA	DEC	z	e_z	f_z	class_spec	original_class_spec	source
0	24745328	270.803750	-3.020833	0.012098	3.335641e-08	KEEP(SPEC)	GALAXY	G	EXTERNAL_NED_2016MNRAS.457.2366S
1	24317251	195.157852	-3.501740	0.004563	5.028434e-08	KEEP(SPEC)	GALAXY	G	EXTERNAL_NED_2016SDSSD.C..0000:
2	23438252	201.650400	8.384936	0.553766	6.033121e-08	KEEP(0)	GALAXY	GALAXY	EXTERNAL_SDSSDR18_BOSS
3	23050127	211.814490	-2.025296	0.681416	7.737748e-08	KEEP(0)	GALAXY	GALAXY	EXTERNAL_SDSSDR18_BOSS
4	23525245	159.850450	4.714067	0.071260	8.633752e-08	KEEP(0)	GALAXY	GALAXY	EXTERNAL_SDSSDR18_SDSS
...	...	...	...	...	...	...	...	...	...
8437455	19110249	342.175542	-44.535939	1.035000		NaN	GRAVLENS (SIMBAD)	NaN	VizieR_J/MNRAS/492/503/tableb5
8437456	19110251	342.192442	-44.525069	2.976000		NaN	GRAVLENS (SIMBAD)	NaN	VizieR_J/MNRAS/492/503/tableb5
8437457	4872418	136.022080	-0.558610	2.589000		NaN	GRAVLENS (SIMBAD)	NaN	VizieR_J/A+A/678/A27/tableb1
8437458	4872414	136.221670	2.338330	1.891000		NaN	GRAVLENS (SIMBAD)	NaN	VizieR_J/A+A/678/A27/tableb1
8437459	4872407	136.557080	-1.012220	2.225100		NaN	GRAVLENS? (SIMBAD)	NaN	VizieR_J/A+A/678/A27/tableb1

8437460 rows × 9 columns

# Dados espectroscópicos



# Criando o catálogo para treinamento

## Crossmatches espectroscópicos

Busca radial (RA, DEC) com raio de 2" em torno de cada objeto do SPLUS

## Crossmatches fotométricos

GALEX (II/335/galex\_ais):

- Busca radial com raio de 2"

VHS (II/367/vhs\_dr5):

- Busca radial com raio de 1"
- Conversão de magnitudes Vega para AB

unWISE (II/363/unwise):

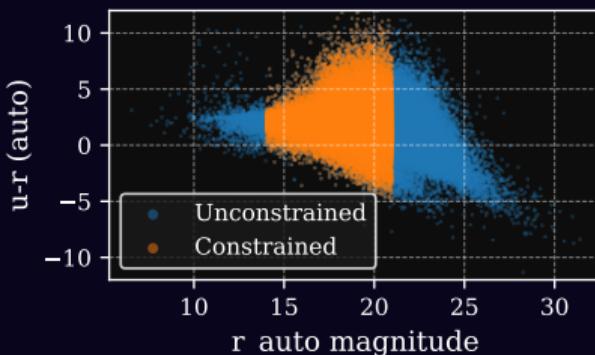
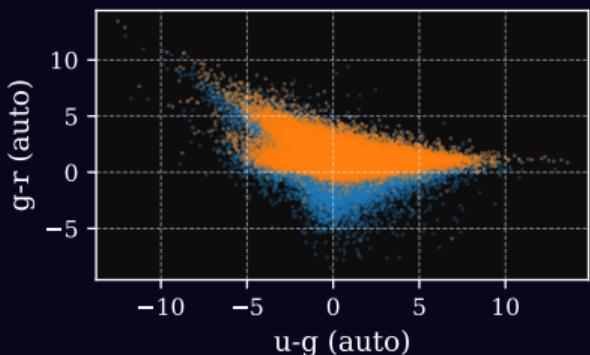
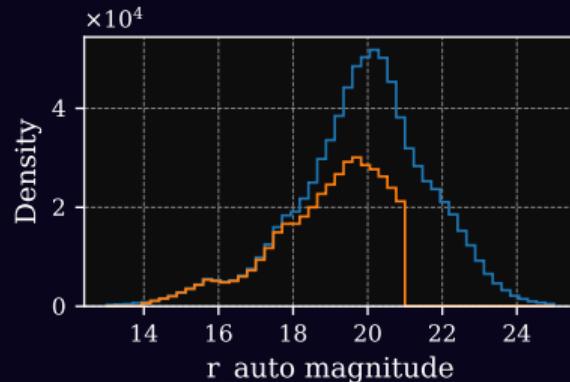
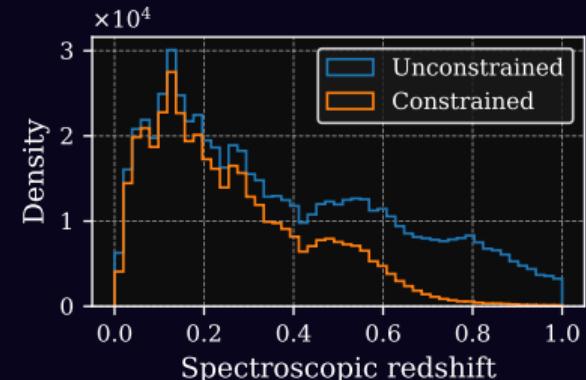
- Busca radial com raio de 1"
- Cálculo das magnitudes e erros
- Conversão de magnitudes Vega para AB

# Pré-processamento

Variable	Constraints
r_auto	[14, 21]
nDet_PStotal	$\geq 1$
SEX_FLAGS_DET	[0, 3]
z	[0.002, 0.8]
e_z	$\leq 0.002$
f_z	not REMOVE
class_spec	GALAXY, SUPERNOVAE or AGN
Separation	$\leq 1''$

- Galáxias
- Nem muito fracas, nem muito brilhantes
- Com boa fotometria
- Com redshifts de boa qualidade e entre 0.002 e 0.8

# Pré-processamento



# Pesos por objeto

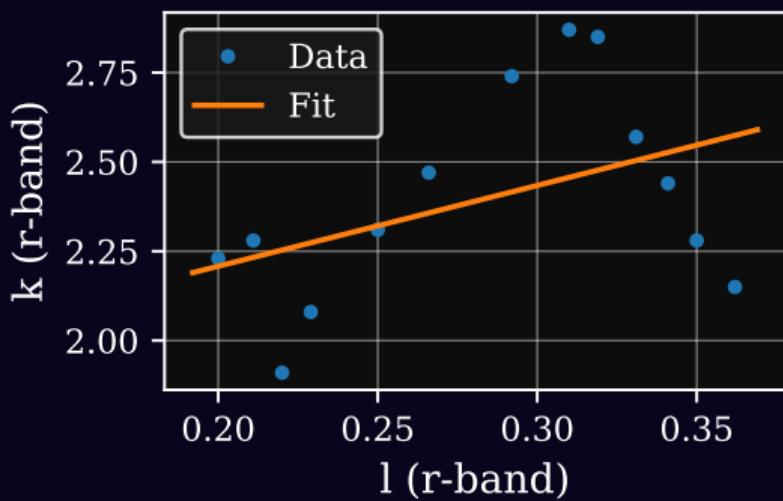
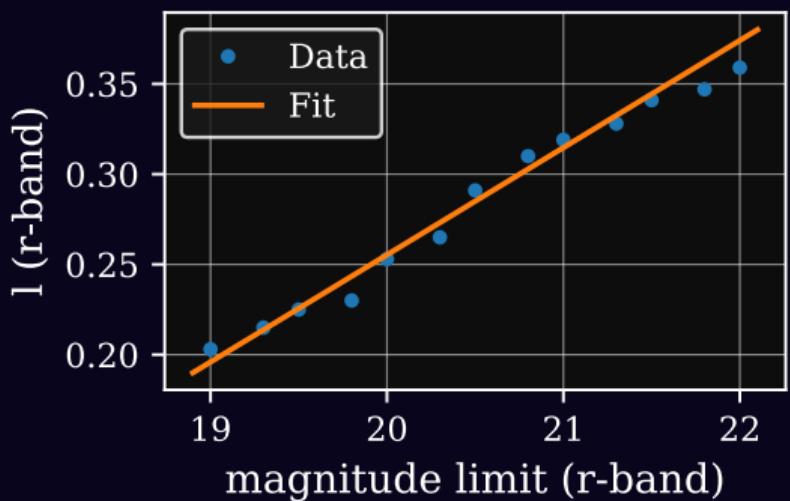
A distribuição de spec-zs da amostra de treino é diferente da distribuição esperada no universo.

$$P(y|\mathcal{D}) = \int P(y, \theta|\mathcal{D})d\theta = \int P(y|\theta, \mathcal{D})P(\theta|D)d\theta$$

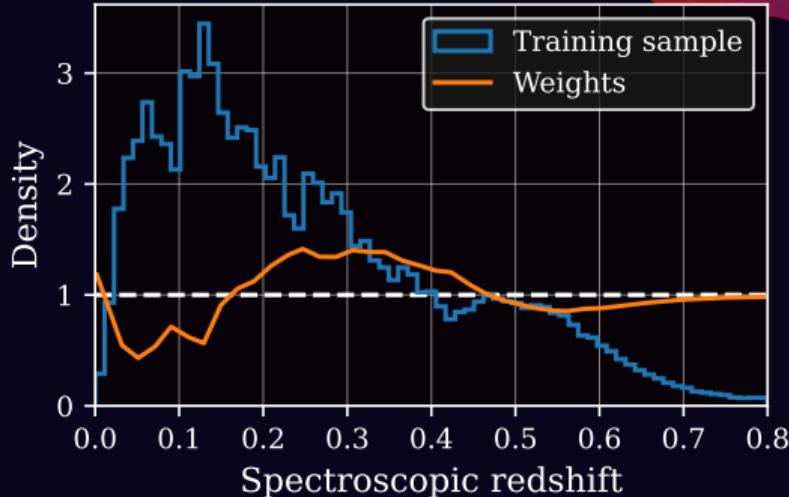
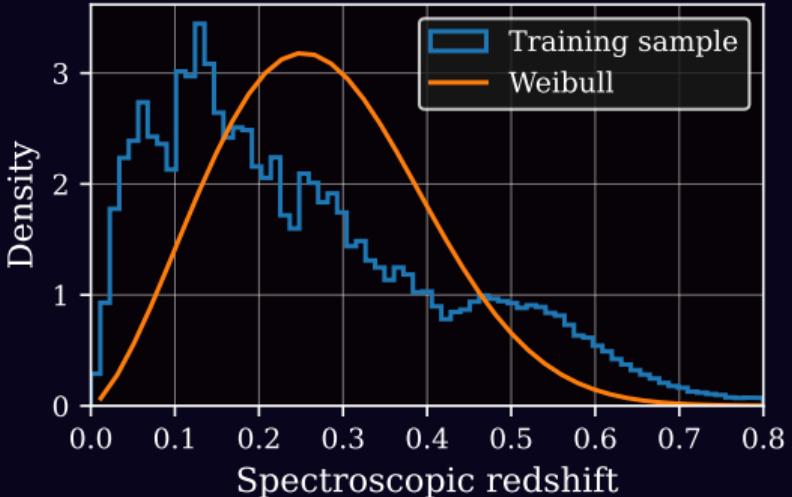
Os resultados para  $y$  dependem dos dados  $\mathcal{D}$ , então o conjunto de treinamento pode introduzir viéses.

# Pesos por objeto

Assumimos que a distribuição de  $z_{\text{spec}}$  é conhecida para surveys limitados por fluxo, e modelamos esta distribuição como função da magnitude usando a amostra do COSMOS2020 (Weaver et al., 2022) e uma função Weibull de dois parâmetros  $l$  e  $k$ .



# Pesos por objeto

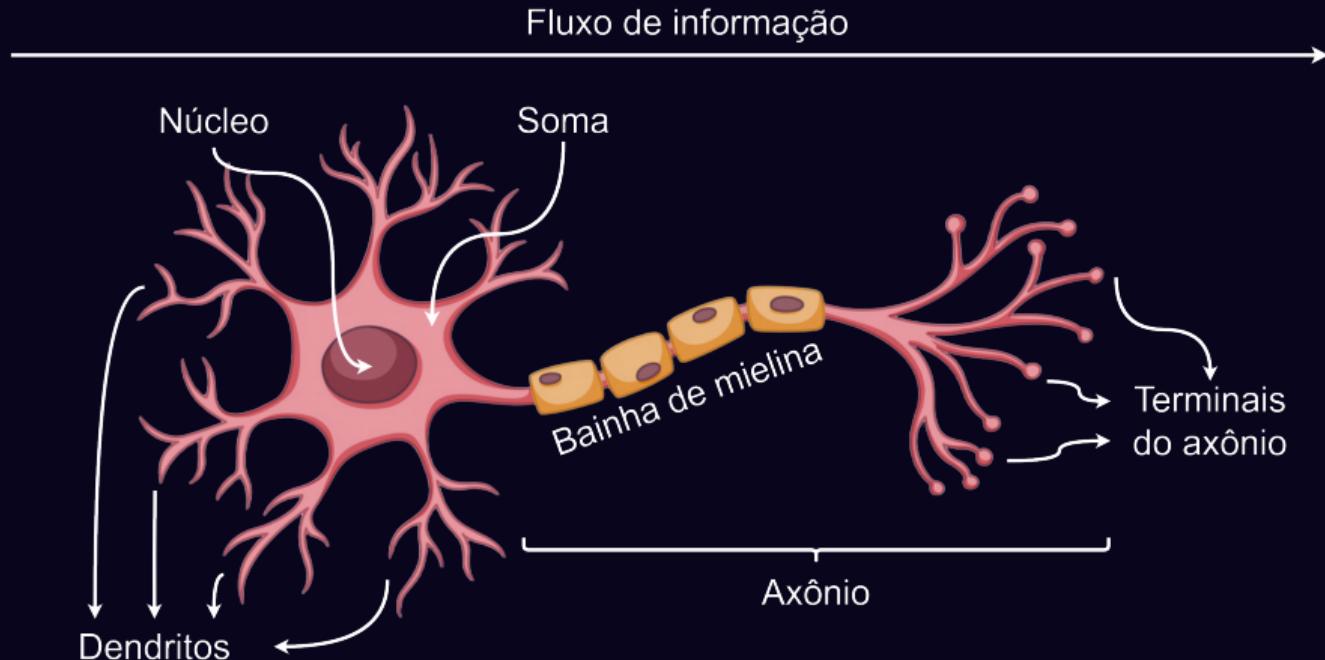


Com essa abordagem, cada objeto contribui de forma diferente no treinamento do modelo

# Metodología

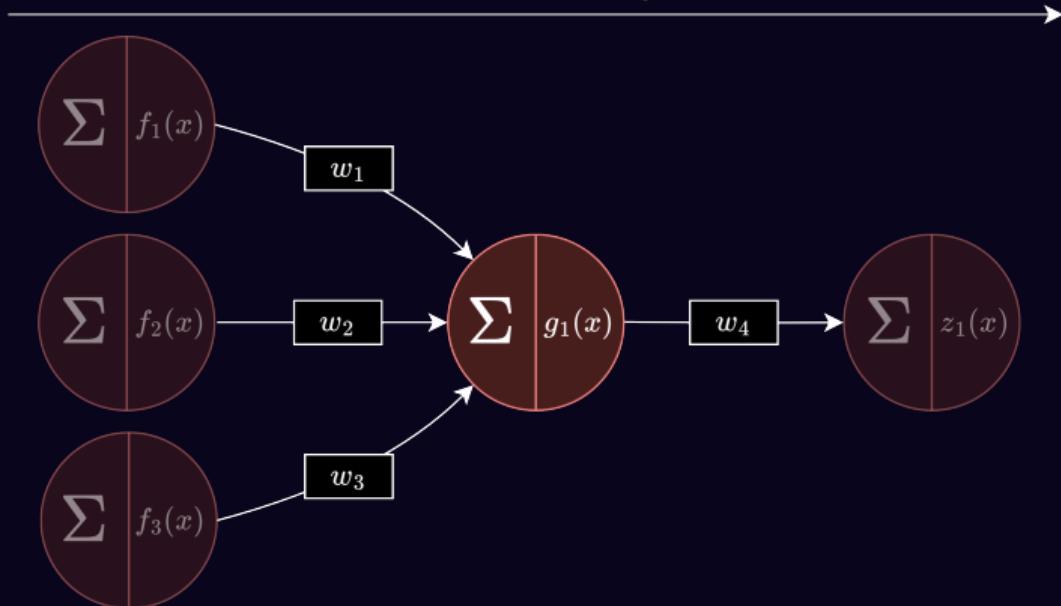
---

# Redes neurais

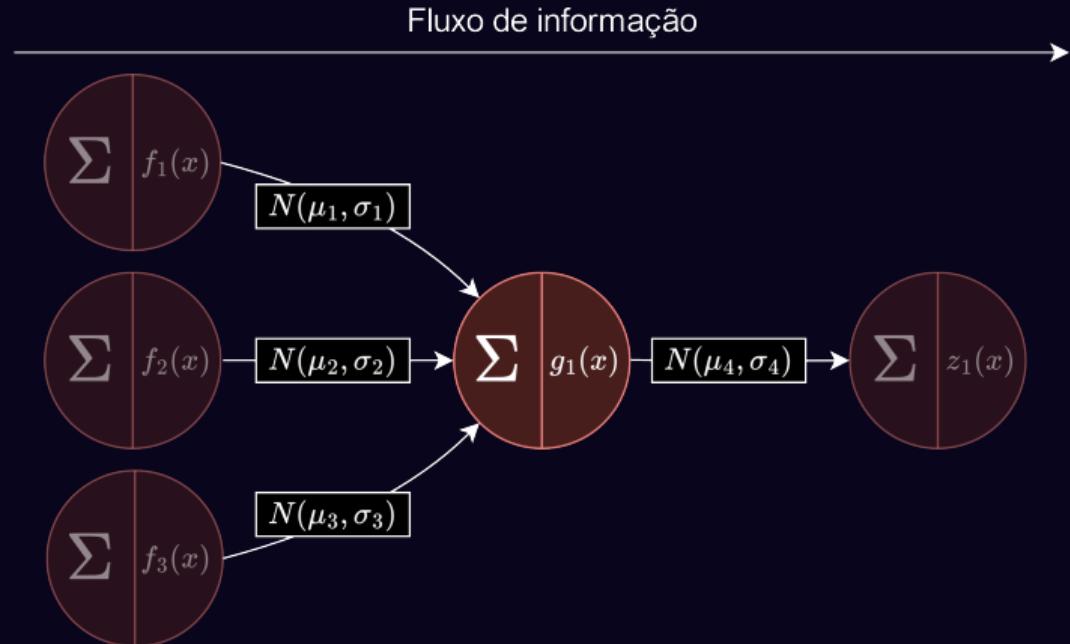


# Redes neurais

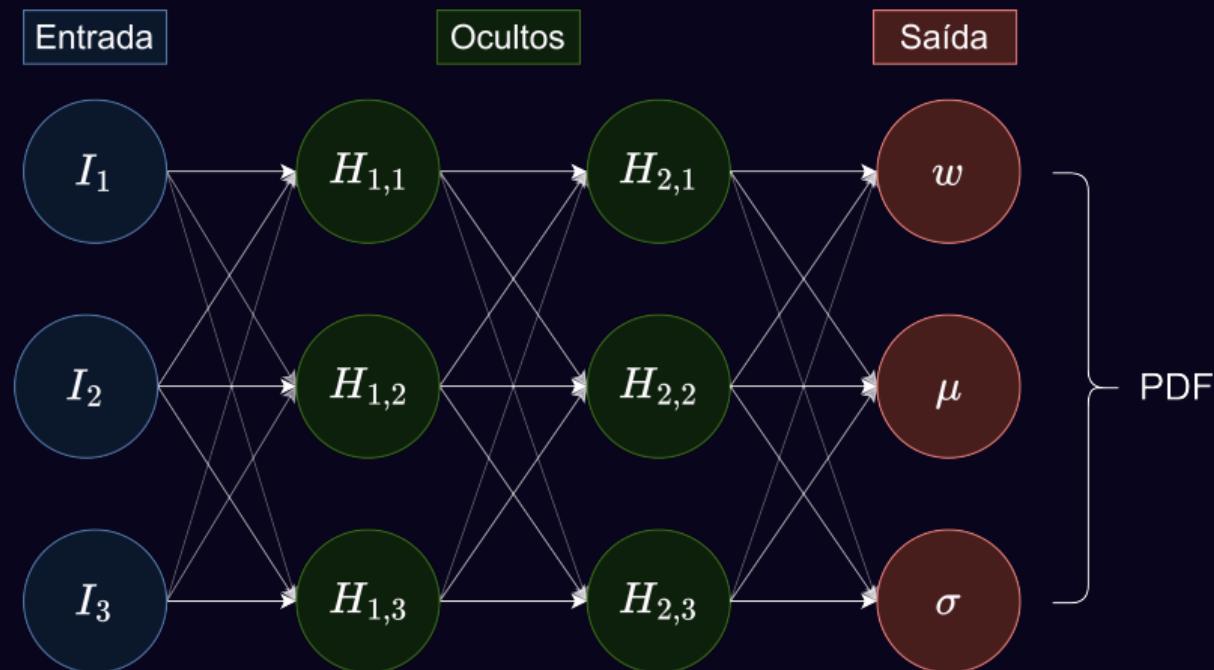
Fluxo de informação



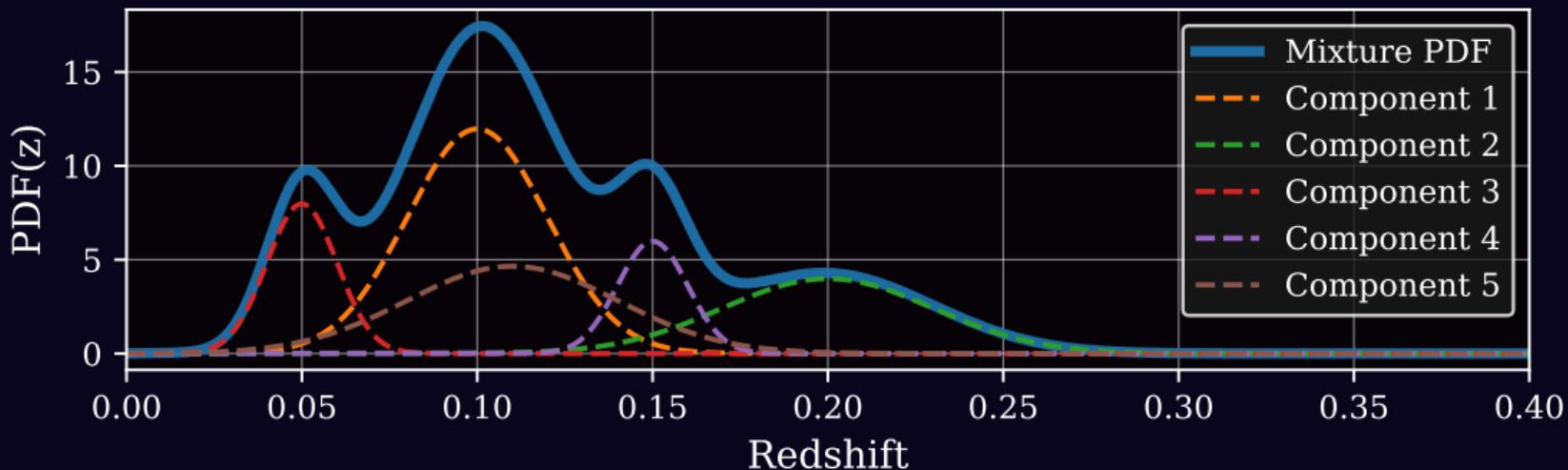
# Redes neurais Bayesianas



# Redes de mistura de densidades



# Redes de mistura de densidades



# Redes de mistura de densidades Bayesiana

Usamos uma combinação da abordagem Bayesiana com a de mistura de densidades, criando uma rede cujos pesos são distribuições e que, como saída, fornece  $N$  componentes de uma mistura de densidades.

## Bayesian Neural Network

- Modelagem de incertezas
  - ▶ Aleatória (dos dados)
  - ▶ Epistêmica (do modelo)

## Mixture Density Network

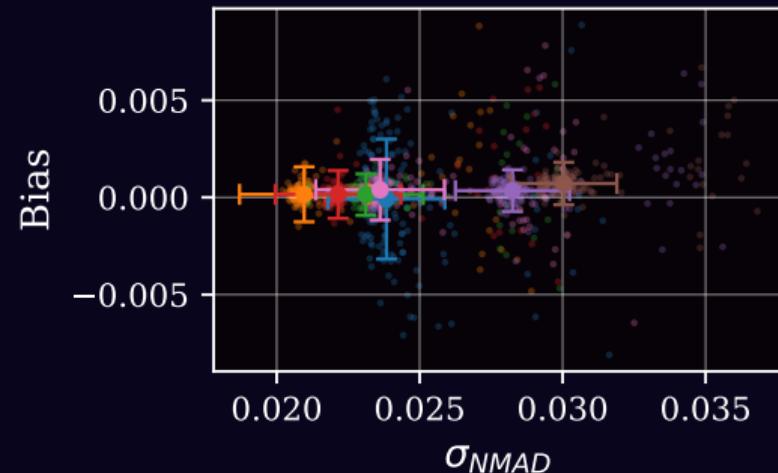
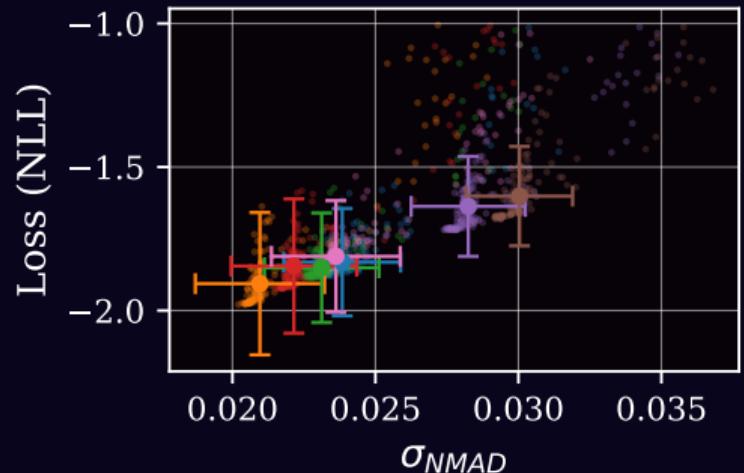
- Gera PDFs bem calibradas diretamente, inteiramente descritas por poucos parâmetros
- É eficiente em termos de armazenamento

# Definindo a arquitetura

---



# Definindo a arquitetura da rede com o Optuna



res	PStotal	aper_3	aper_6	auto	petro	iso
-----	---------	--------	--------	------	-------	-----

# Definindo a arquitetura da rede com o Optuna

Variável	Parâmetros para amostrar	Parâmetros escolhidos
Camadas	3 a 6	4
Neurônios	64 a 128	64
Ativação	Todas as "LU"	gelu
Otimizador	Variações de Adam	AdaBelief
Learning rate	0.001 a 0.03	0.0296
Weight decay	$1 \times 10^{-6}$ a $1 \times 10^{-4}$	$7.5 \times 10^{-6}$
Viés das camadas	True ou False	True
Atenção por feature	True ou False	True

**Em todos os casos, os valores de entrada foram magnitudes, cores e informações morfológicas**

# Resultados

---



# Redshifts fotométricos: estimativas de ponto único

Desvio médio absoluto normalizado ( $\sigma_{\text{NMAD}}$ , Brammer et al., 2008)

$$\sigma_{\text{NMAD}} = 1.48 \times \text{mediana} \left( \left| \frac{\delta z - \text{mediana}(\delta z)}{1 + z_{\text{spec}}} \right| \right)$$

Viés ( $\mu$ )

$$\mu = \text{mediana} (\delta z)$$

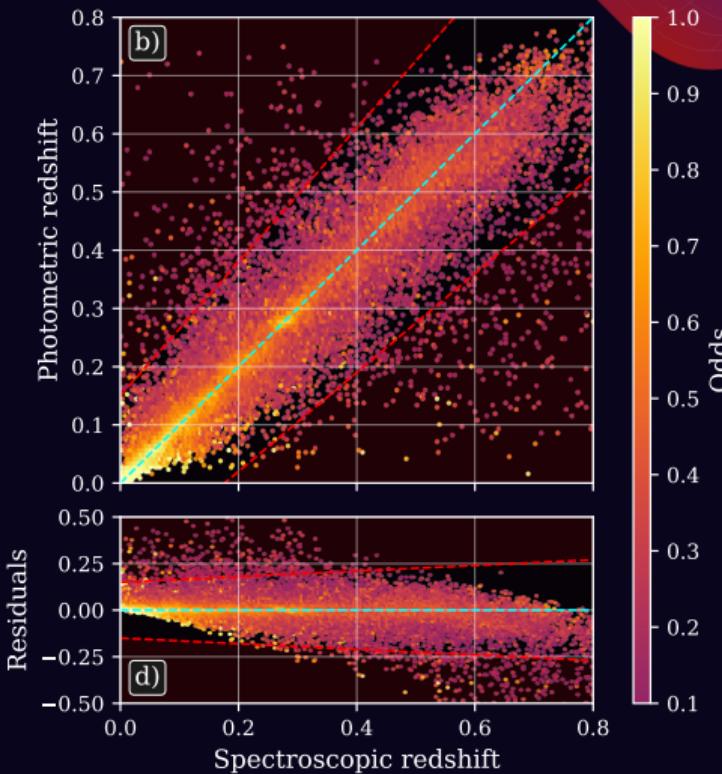
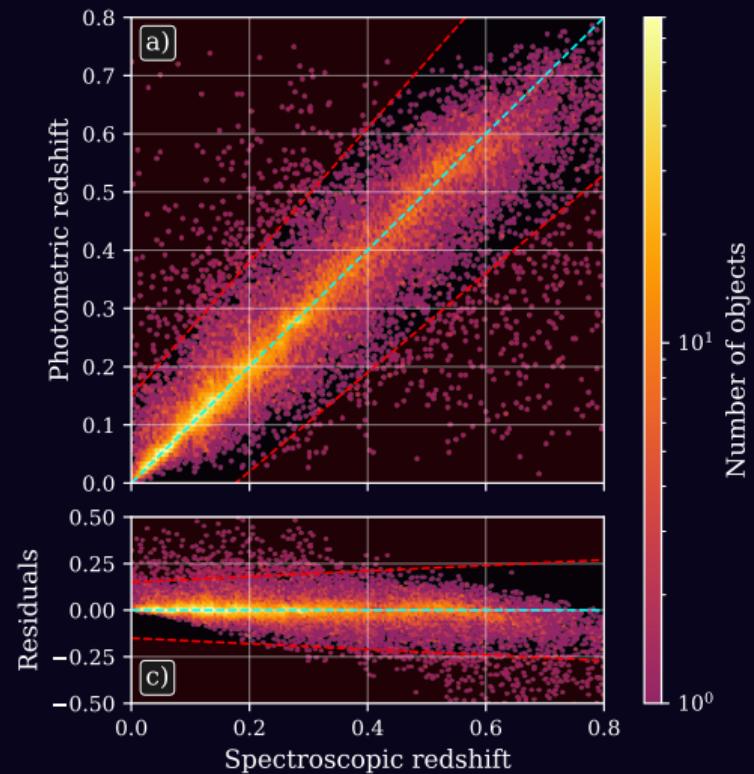
$$\mu_{\text{norm}} = \text{mediana} \left( \frac{\delta z}{1 + z_{\text{spec}}} \right)$$

Fração de outliers ( $\eta$ , Ilbert et al., 2006; Dahlen et al., 2013)

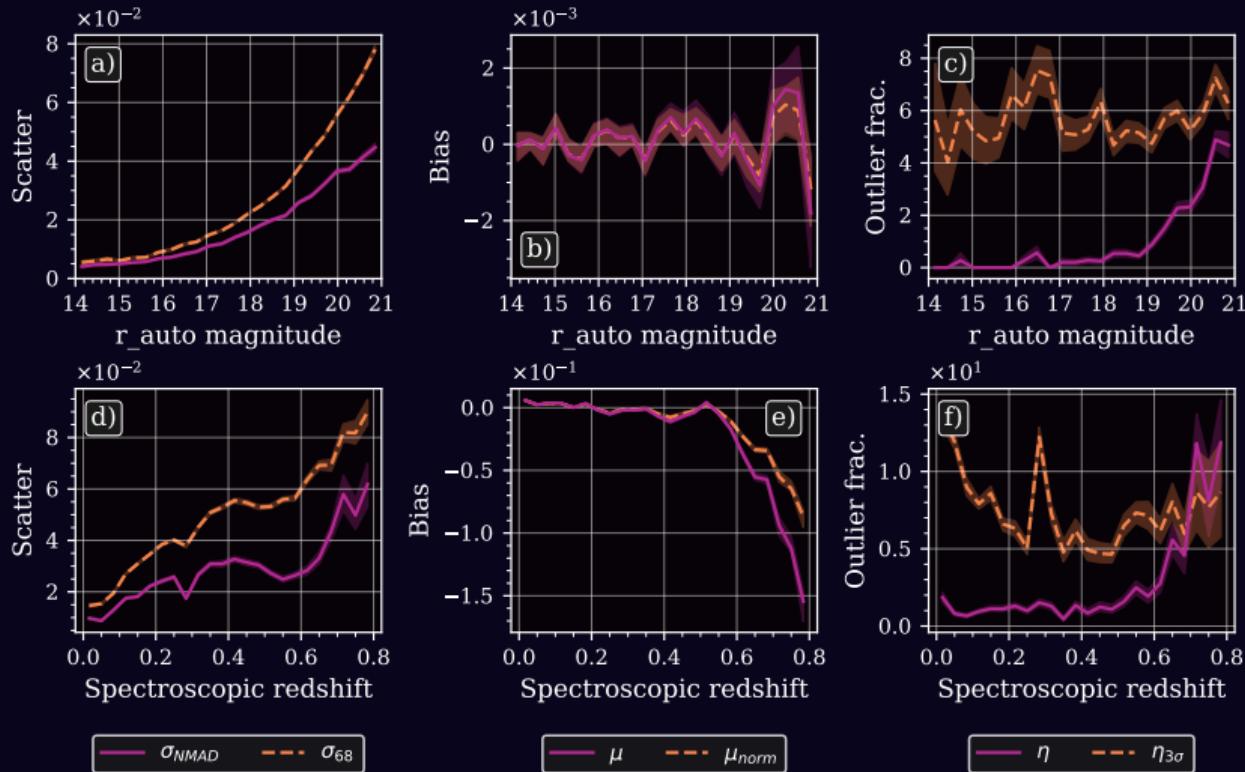
$$\eta = \frac{|\delta z|}{1 + z_{\text{spec}}} > 0.15$$

$$\eta_{N\sigma} = \frac{|\delta z|}{1 + z_{\text{spec}}} > N \cdot \sigma_{\text{NMAD}}$$

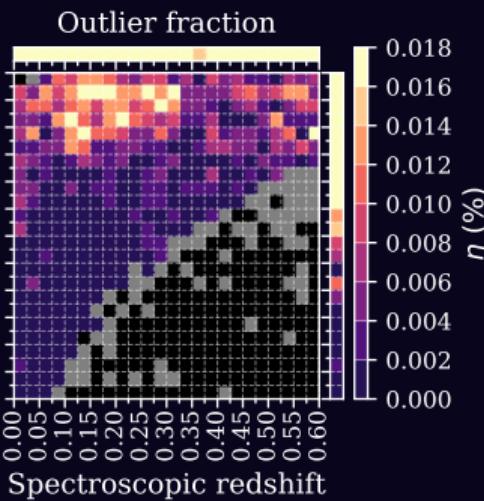
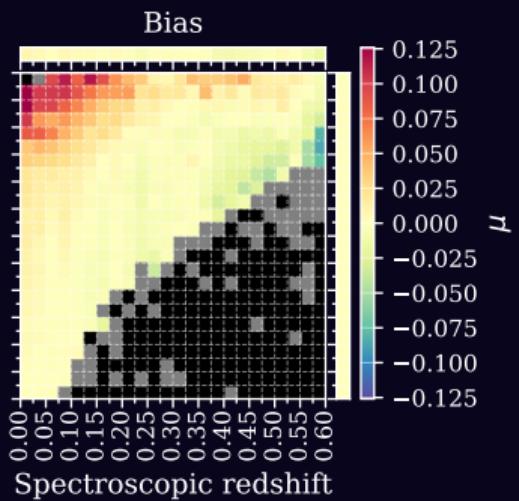
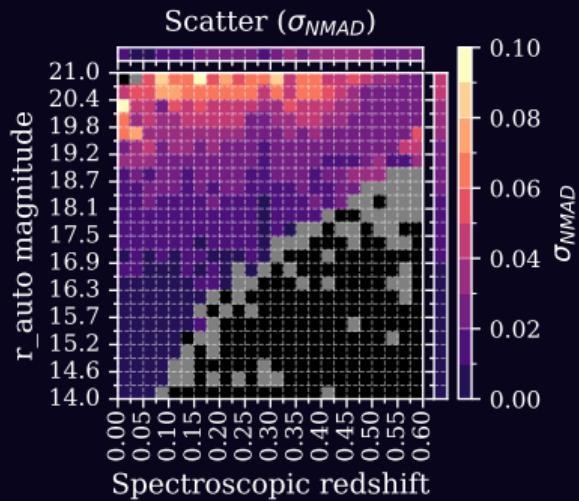
# Redshifts fotométricos: estimativas de ponto único



# Redshifts fotométricos: estimativas de ponto único

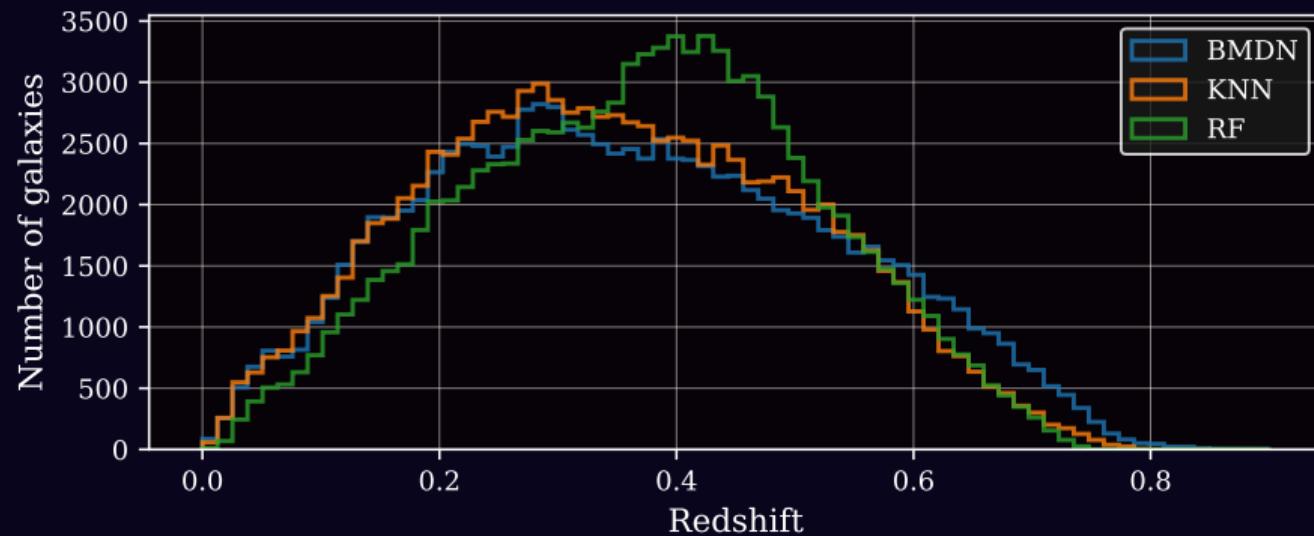


# Redshifts fotométricos: estimativas de ponto único



# Redshifts fotométricos: estimativas de ponto único

Para verificar se a distribuição de photo-zs está como é esperado, comparamos os resultados que obtemos com os resultados de dois outros modelos, um KNN e uma RF, para objetos na Stripe-82.



# Redshifts fotométricos: funções de densidade de probabilidade

## Odds (Benitez, 2000)

$$\text{odds}_i = \int_{z_{\text{peak}, i} - \Delta z}^{z_{\text{peak}, i} + \Delta z} \text{PDF}_i(z) dz,$$

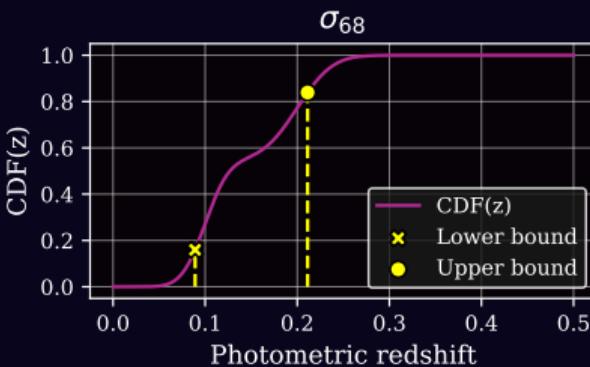
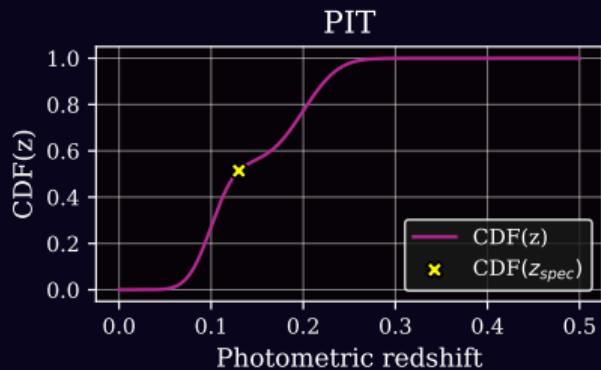
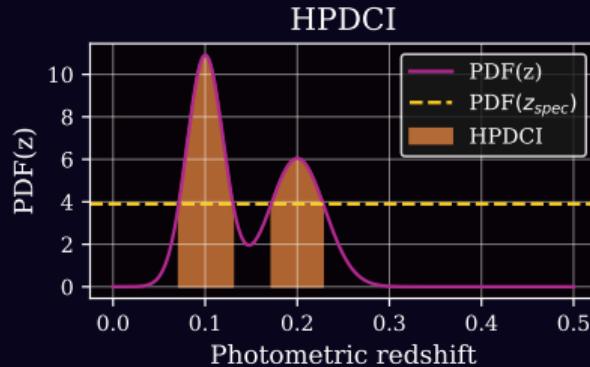
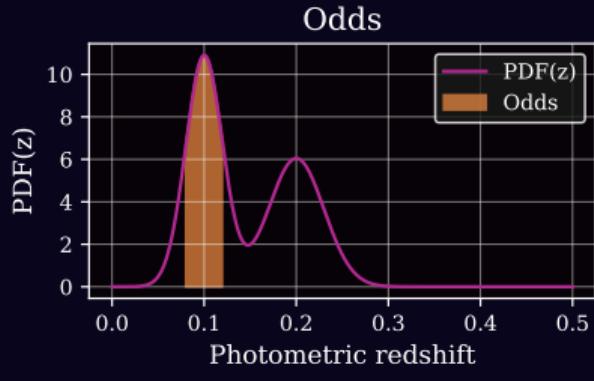
## Probability Integral Transform (Polsterer et al., 2016)

$$\text{PIT}_i = \int_0^{z_{\text{spec}}} \text{PDF}_i(z) dz = \text{CDF}_i(z_{\text{spec}}).$$

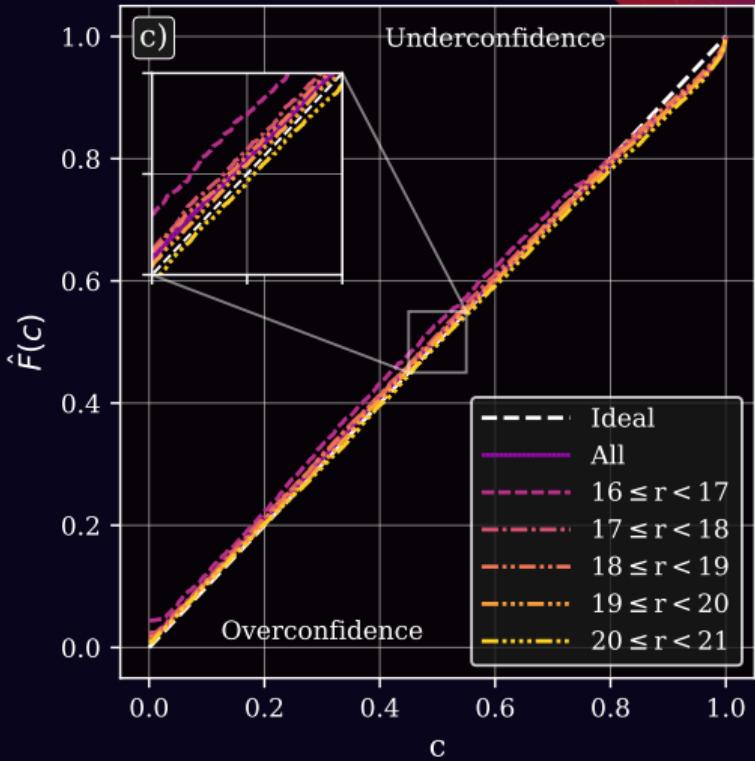
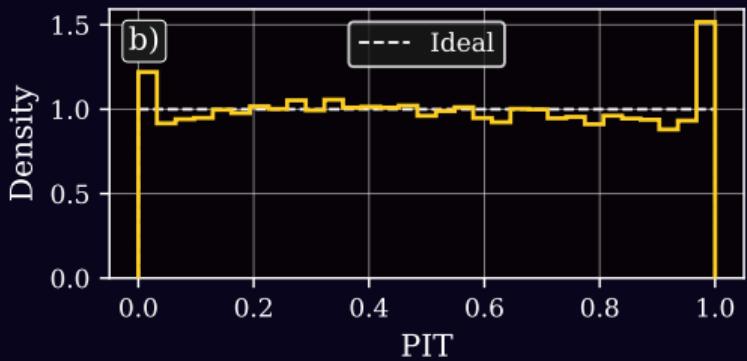
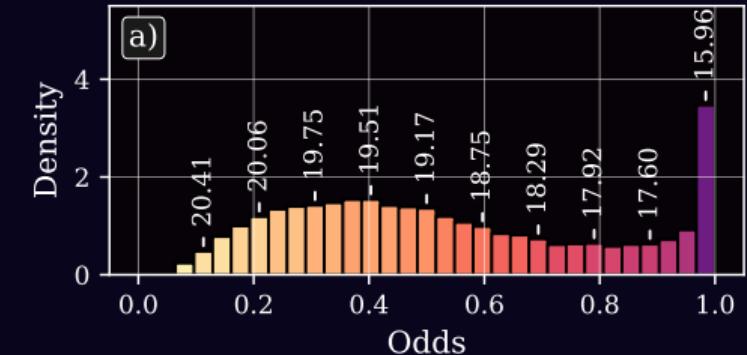
## Highest Probability Density Credible Interval (Wittman et al., 2016)

$\sigma_{68}$  e valor máximo

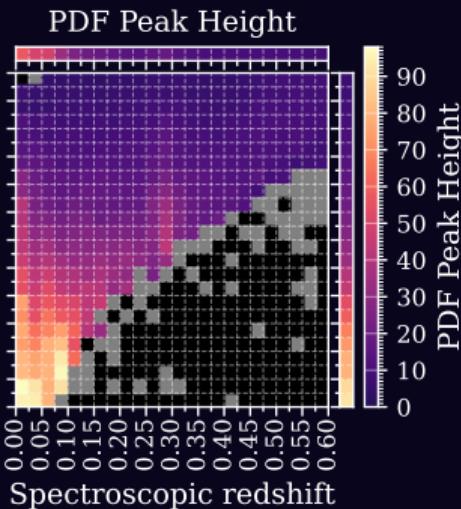
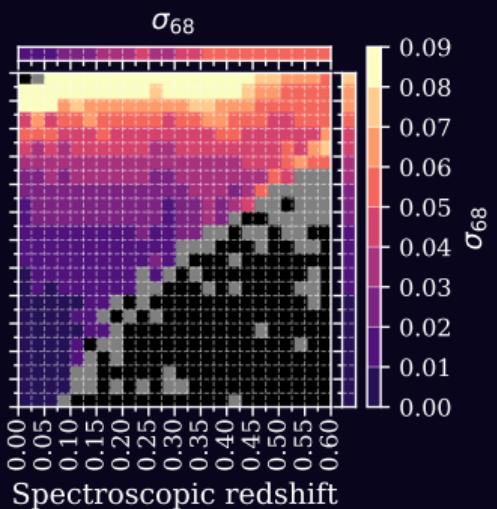
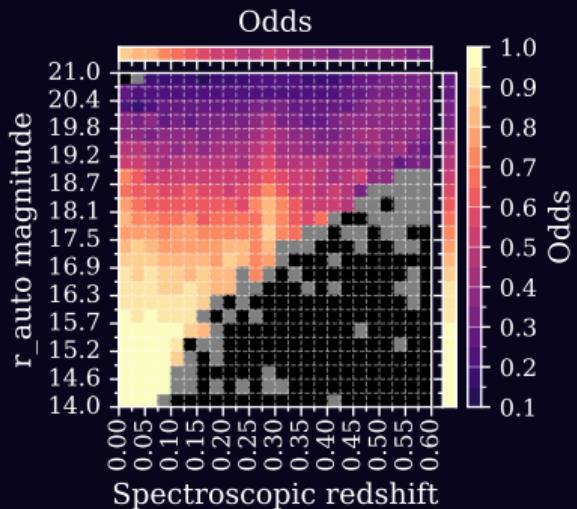
# Redshifts fotométricos: funções de densidade de probabilidade



# Redshifts fotométricos: funções de densidade de probabilidade



# Redshifts fotométricos: funções de densidade de probabilidade

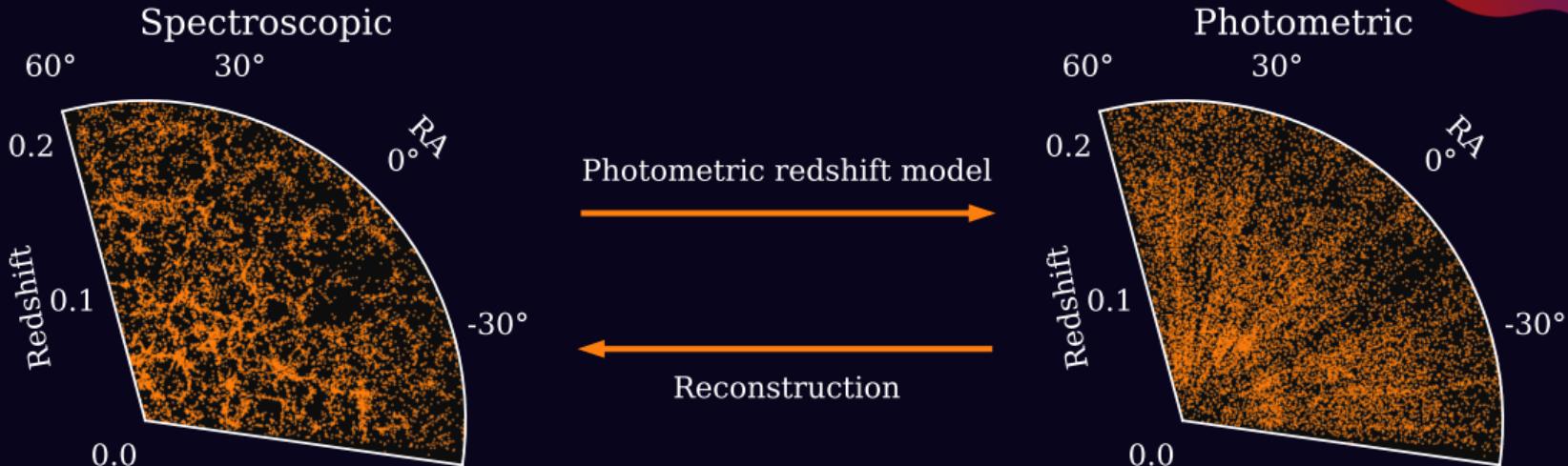


# Estrutura em larga escala

---



# Estrutura em larga escala



Interpretamos que o  $z_{\text{phot}}$  é igual a  $z_{\text{spec}} + \epsilon$

# Estrutura em larga escala

**Autoencoders e U-Nets**

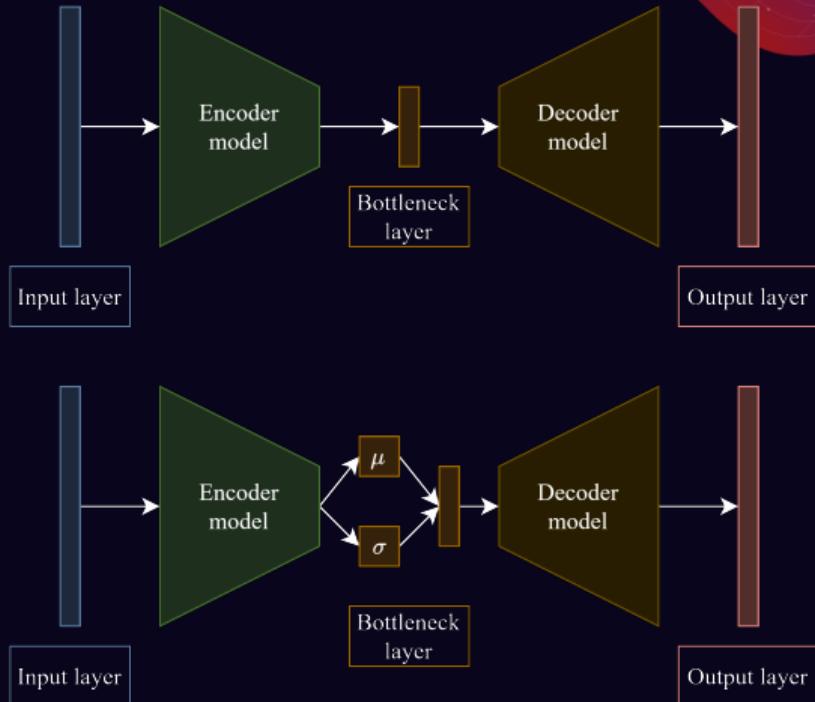
**Denoising Diffusion Probabilistic Models (DDPMs)**

**Graph Neural Networks (GNNs)**

# Autoencoders

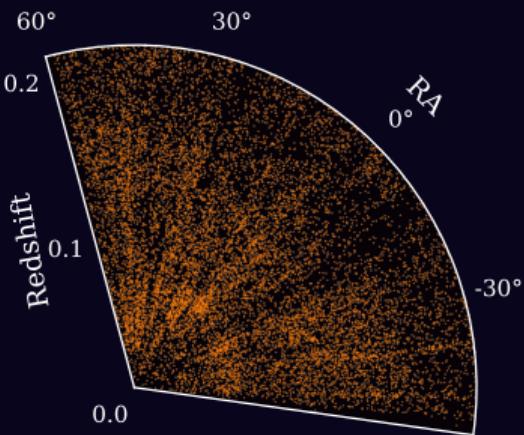
(Kramer, 1991; Kingma e Welling, 2013; Ronnerberger et al., 2015)

- São caracterizadas pela existência de um gargalo
- Treinamento simultâneo de duas redes
- Tem como objetivo reproduzir o input com menos informação
- Remove ruído pois ele não é fundamental na reconstrução do input

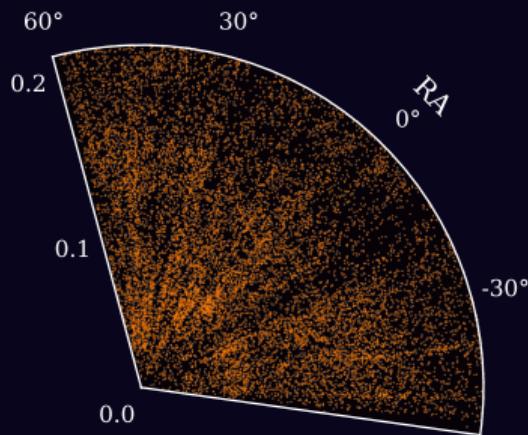


# Autoencoders

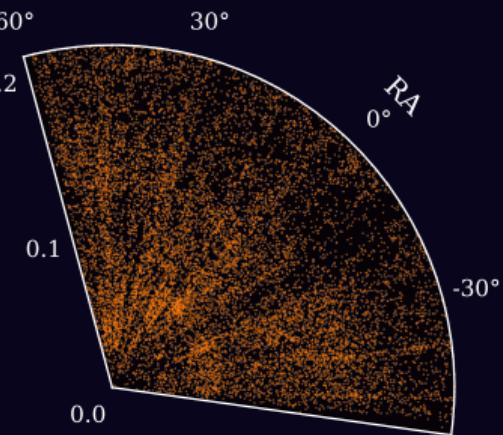
Photometric



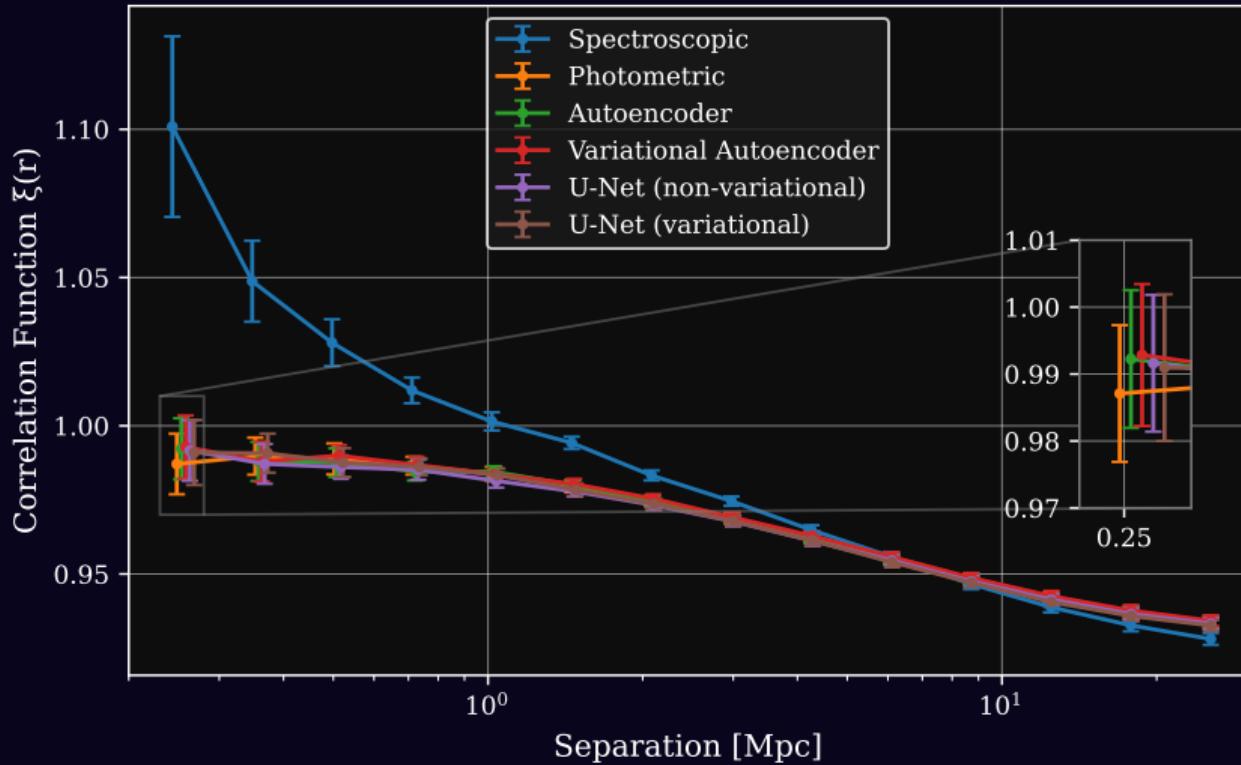
Autoencoder



Variational autoencoder



# Autoencoders



# Denoising Diffusion Probabilistic Models (Ho et al., 2020)



Figura 4: Adaptado de <https://cvpr2022-tutorial-diffusion-models.github.io/>.

É capaz de modelar as incertezas sem suposições simples (por ex. erros são Gaussianos), e pode aprender correlações nas incertezas do photo-z

# Denoising Diffusion Probabilistic Models

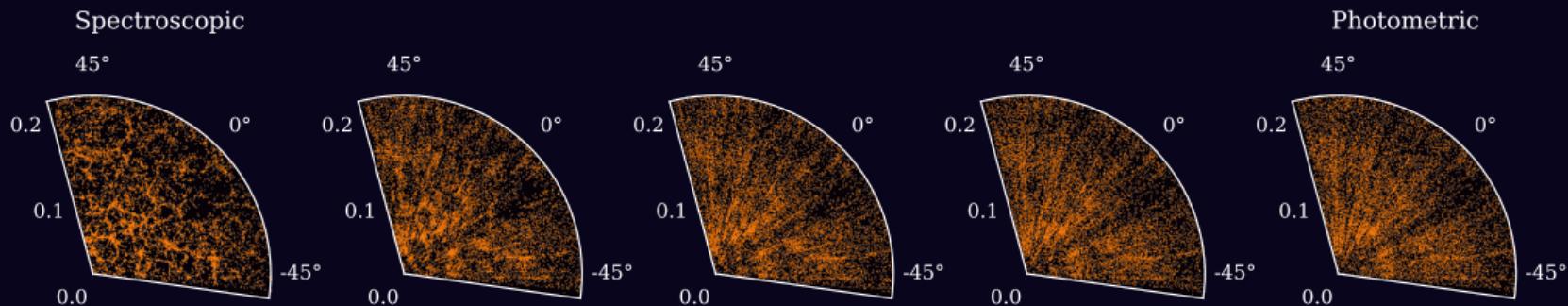
## Tabular

- ✓ Lida com a estrutura natural dos dados
- ✓ Computacionalmente mais leve
- ✗ Sem possibilidade de aprender relações espaciais
- ✗ Não usa arquiteturas comuns a esse problema

## Imagens

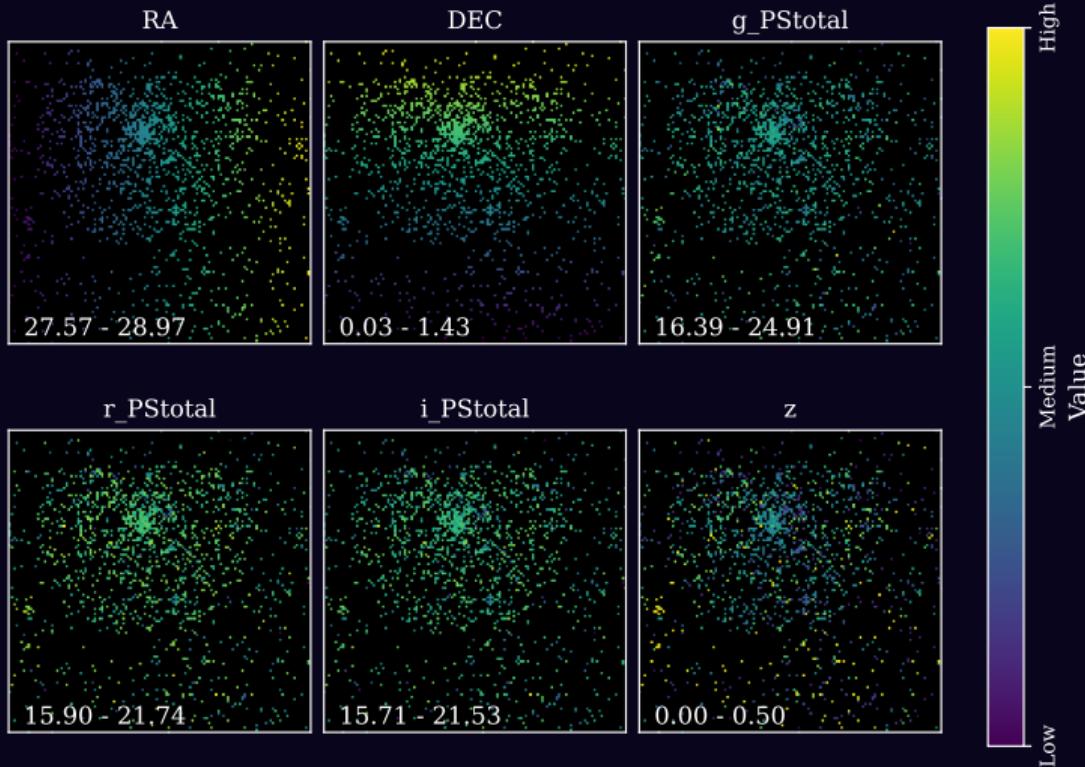
- ✓ Usa arquiteturas conhecidas (U-Net)
- ✓ Aprenderia relações espaciais
- ✓ Entendimento mais simples
- ✗ Computacionalmente mais pesado
- ✗ Modifica a forma natural dos dados

# Denoising Diffusion Probabilistic Models

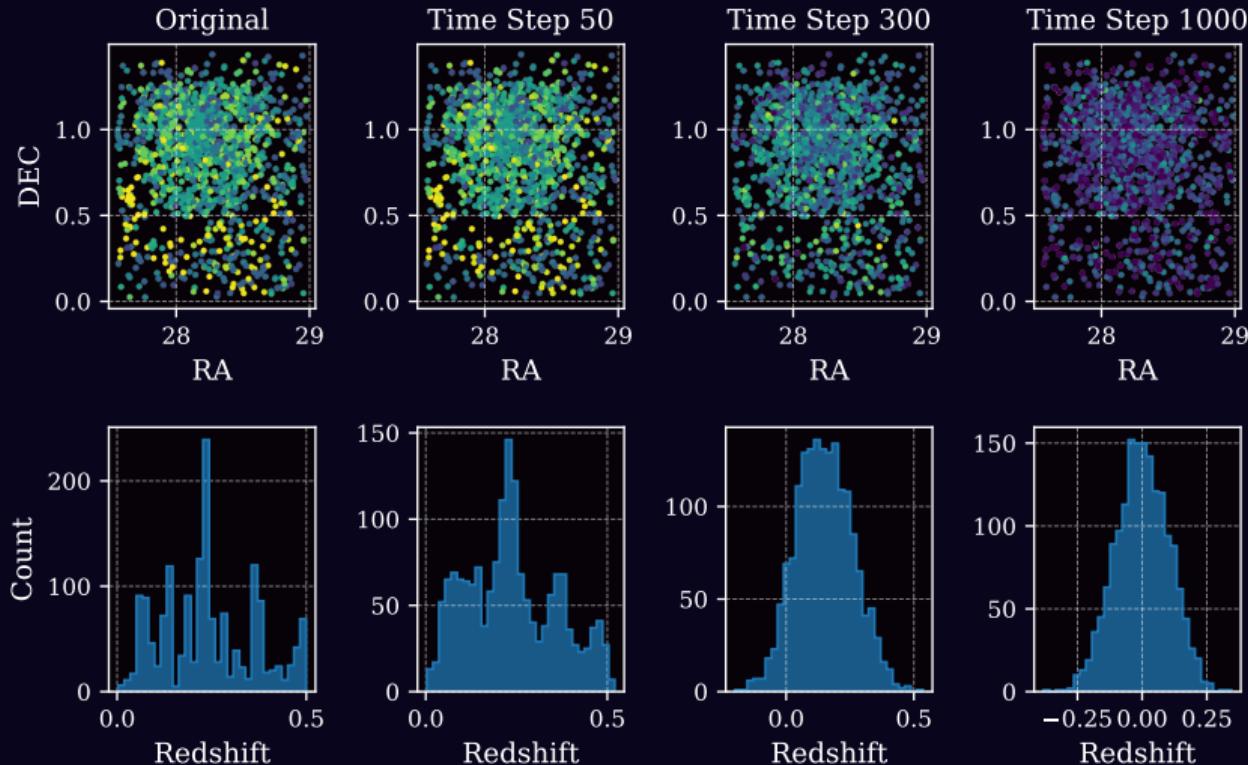


- O ruído é modelado de forma que partimos de  $z_{\text{spec}}$  e chegamos em  $z_{\text{phot}}$
- O modelo é condicionado na fotometria, dispensando a necessidade de saber o timestep  $t$

# Denoising Diffusion Probabilistic Models



# Denoising Diffusion Probabilistic Models



# Graph Neural Networks

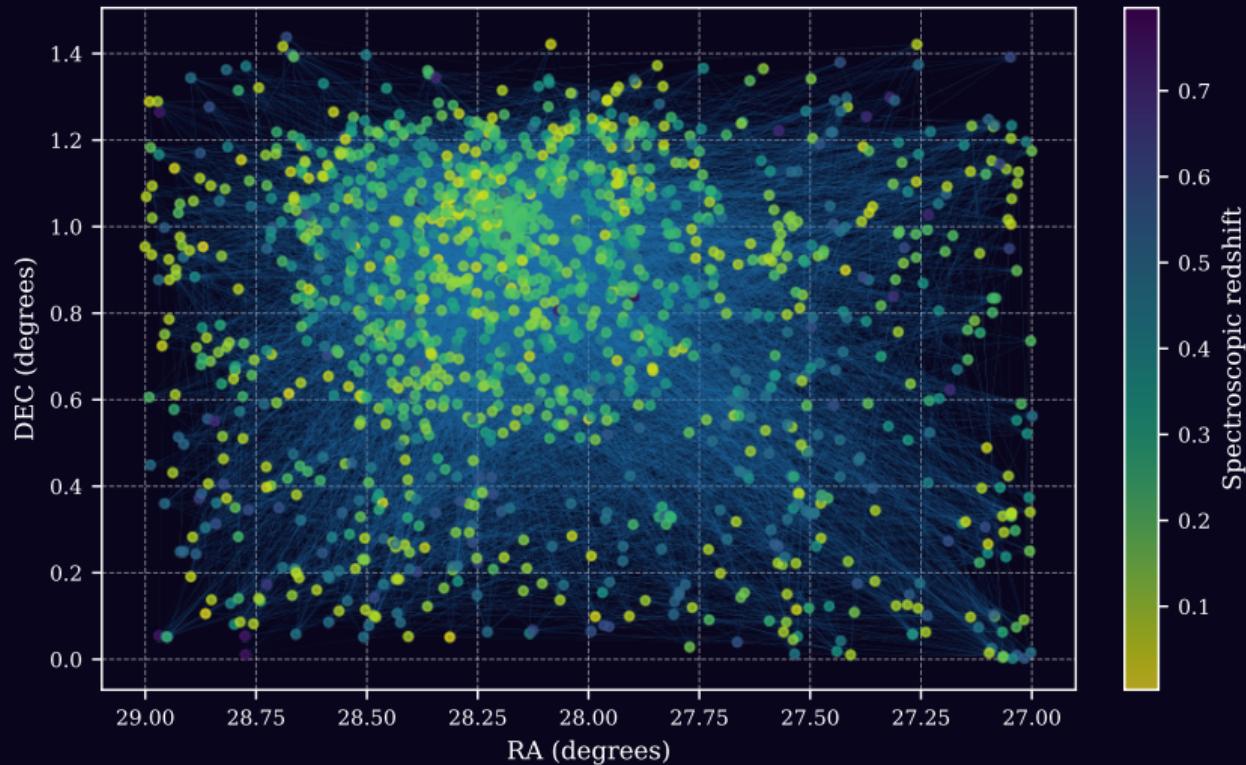
(Gori et al., 2005; Scarselli et al., 2009)

- Sistemas de recomendação
- Detecção de fraudes
- Descoberta de medicamentos
- Identificação de estruturas de proteínas
- Otimização de cadeias de produção

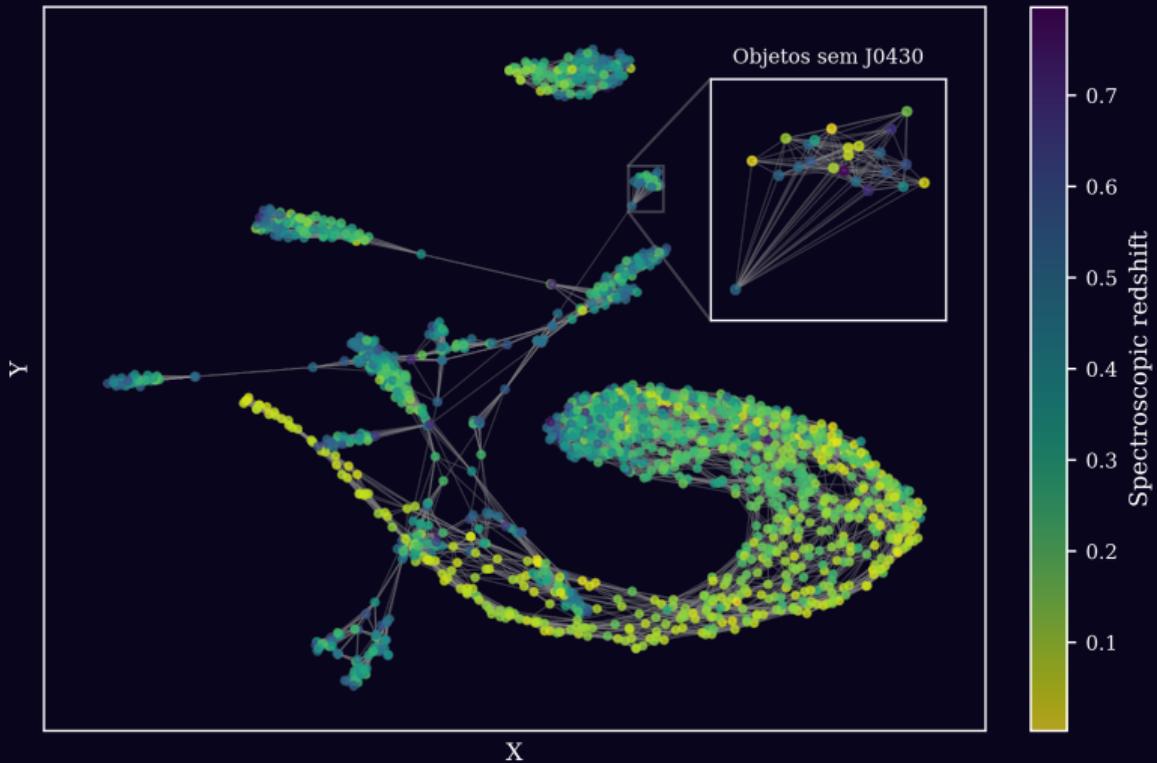


**Figura 5:** Adaptado de <https://graphsandnetworks.com/the-cora-dataset/>

# Graph Neural Networks

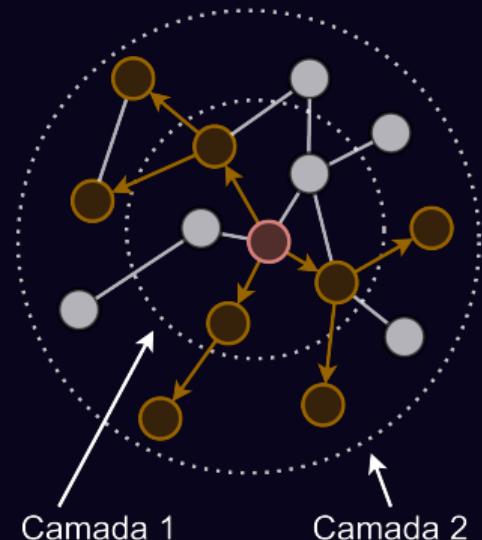


# Graph Neural Networks

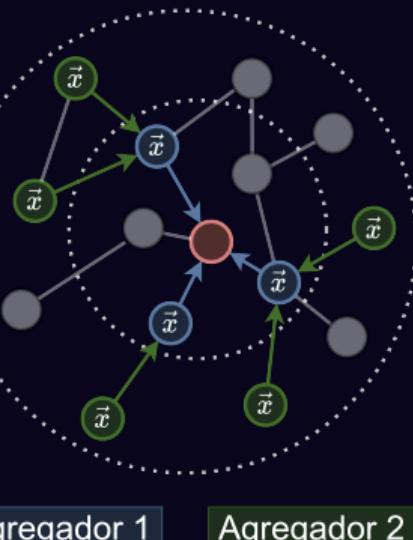


# Graph Neural Networks

1. Amostrar vizinhança



2. Agregar informação  
dos vizinhos



Agregador 2

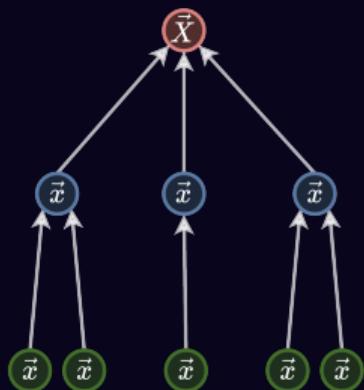


Figura 6: Adaptado de <https://snap.stanford.edu/graphsage/>.

# Conclusões

---

# Conclusões

## Redshifts fotométricos

- O nosso modelo faz estimativas pontuais precisas e acuradas
- Fornece funções de densidade de probabilidade bem calibradas
- Geramos um catálogo com essas informações para toda a colaboração

## Redshifts espectroscópicos

- Criamos o maior compilado de redshifts espectroscópicos do Hemisfério Sul

## Estrutura em larga escala

- A precisão dos  $z_{\text{phot}}$ s serve como ponto de partida para a etapa de recuperação da LSS
- Identificamos possíveis caminhos de progresso (DDPMs, GNNs)
- Este trabalho ainda está em andamento

# Perspectivas futuras

---



# Perspectivas futuras

## Redshifts fotométricos

- Refatoração do código de forma a simplificá-lo e torná-lo mais eficiente
- Desenvolvimento de um código aberto no qual qualquer usuário tem acesso aos modelos e pode fazer estimativas por conta
- Aprimoração dos modelos
  - ▶ Uso de distribuição de magnitudes como input para o treino do modelo
  - ▶ Implementação de uma etapa de template-fitting no processo de treinamento
  - ▶ Arquitetura no estilo Mixture-of-Experts/Transformer

# Perspectivas futuras

## Compilado espectroscópico

- Criar uma nova versão do código que faz o compilado de  $z_{\text{spec}}$ , que seja eficiente e fácil de compreender, para divulgação à comunidade.

## Estrutura em larga escala

- Continuidade nas pesquisas relacionadas à reconstrução da LSS explorando DDPMs e GNNs
- Obtenção da estrutura da LSS reconstruída
- Utilizar este resultado como ponto de partida para outras pesquisas



# Obrigado!