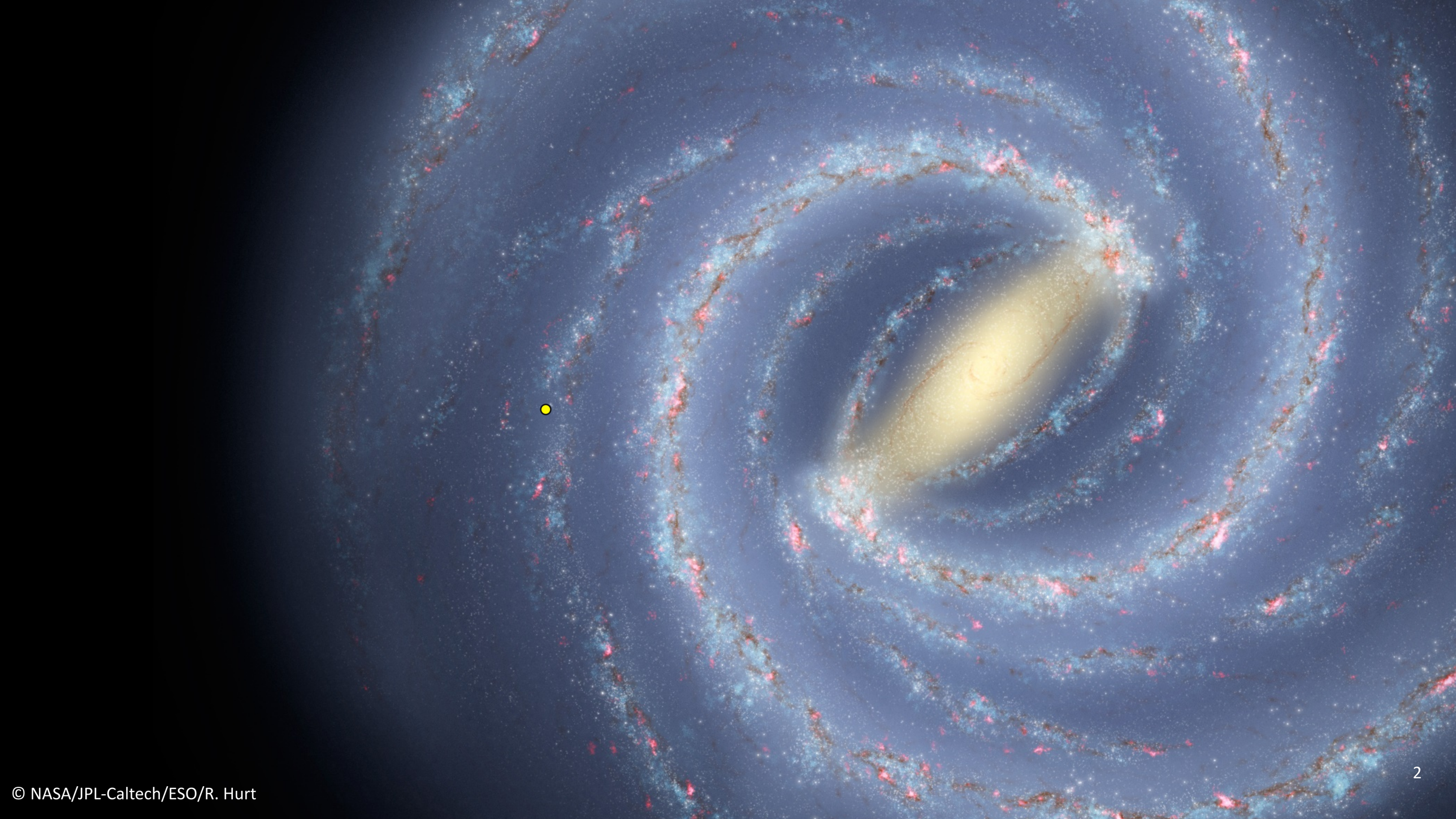
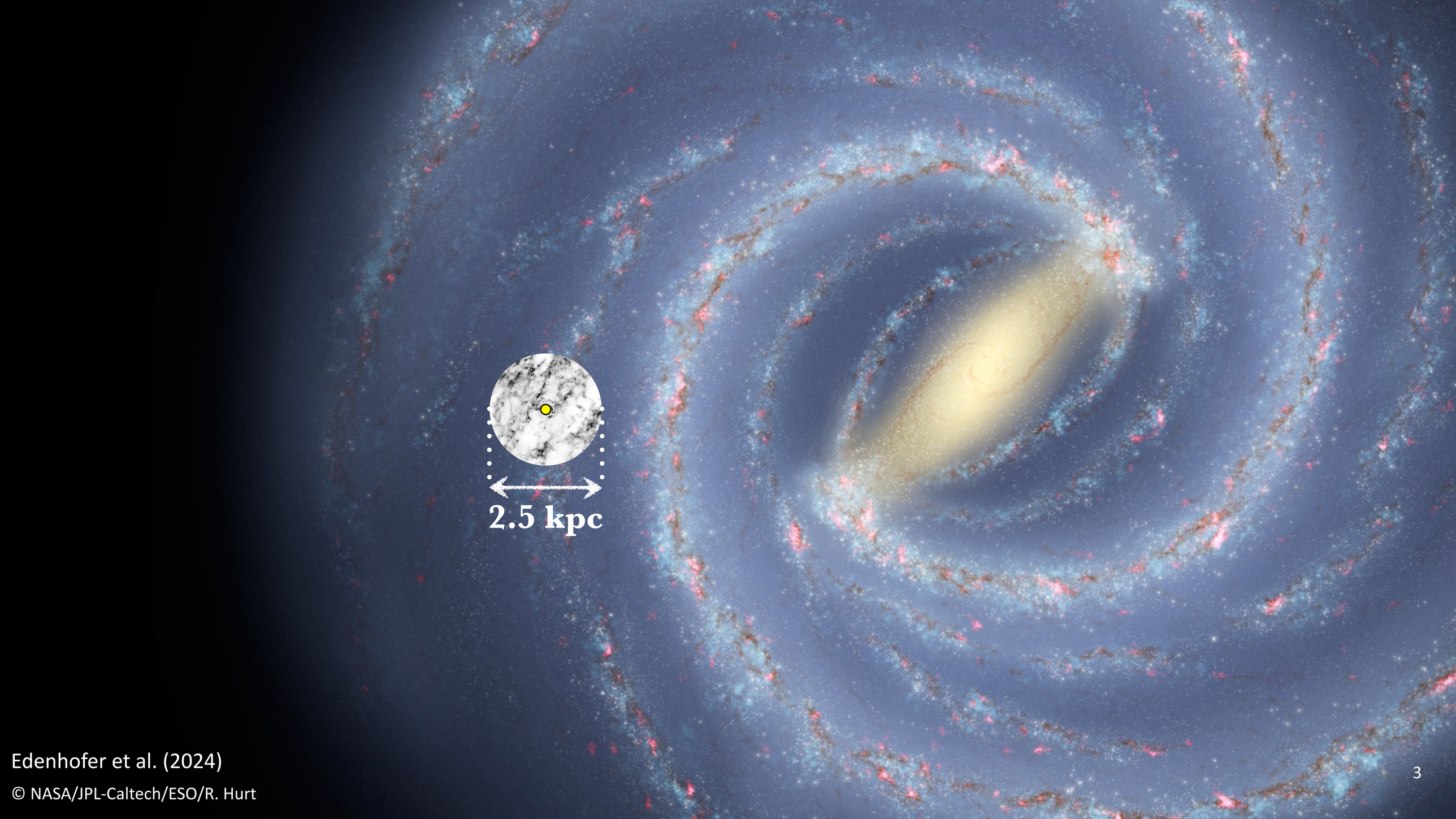
Learning with Gaps

A Domain-Adaptive SBI Framework for Mapping Young Stars from Incomplete, Multi-Survey Data

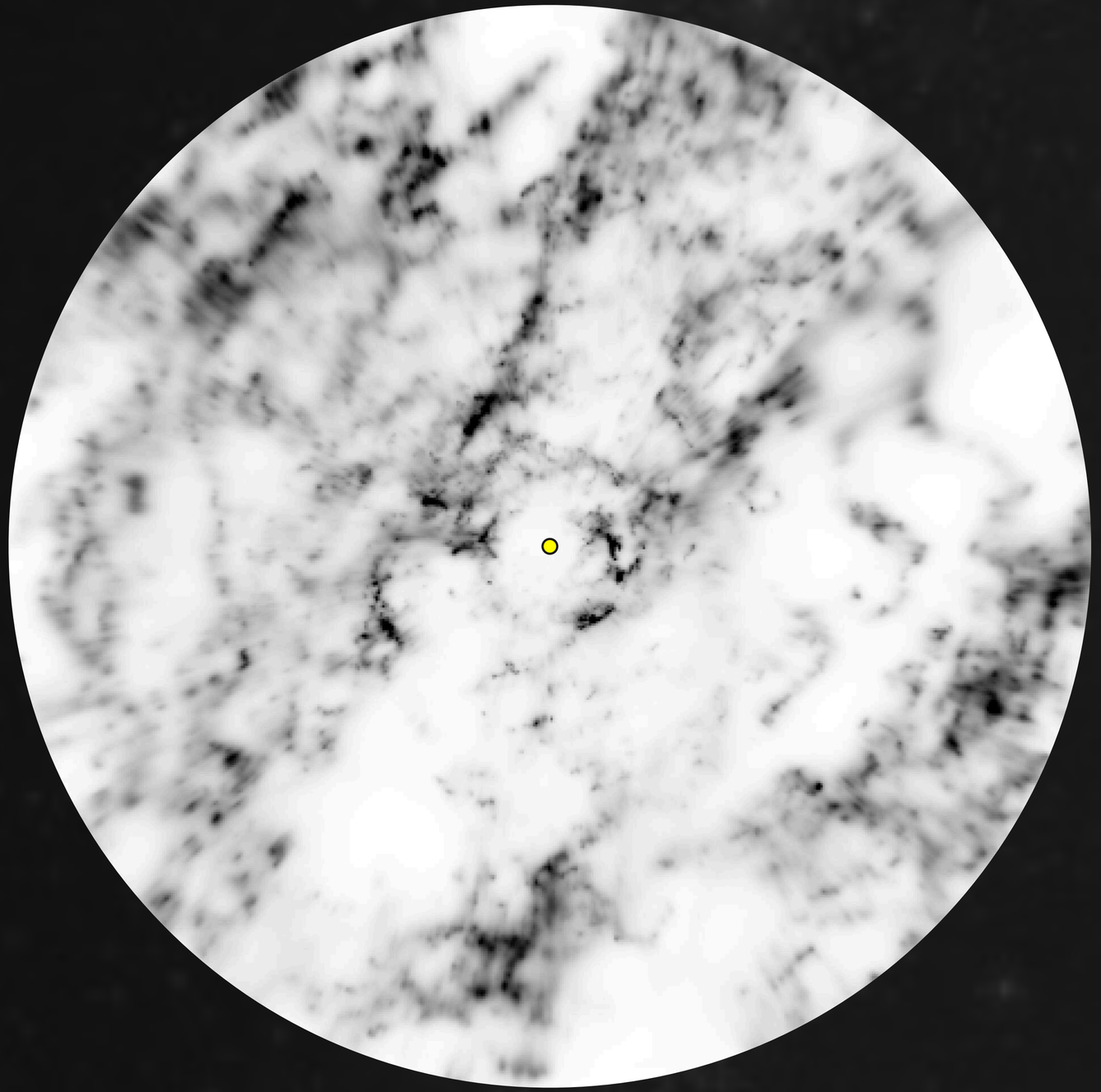
Sebastian Ratzenböck @CfA

In collaboration with
Catherine Zucker (CfA), Joshua Speagle (UToronto), Phillip Cargile (CfA), Philipp Frank (Stanford), Andrew Saydjari (Princeton)

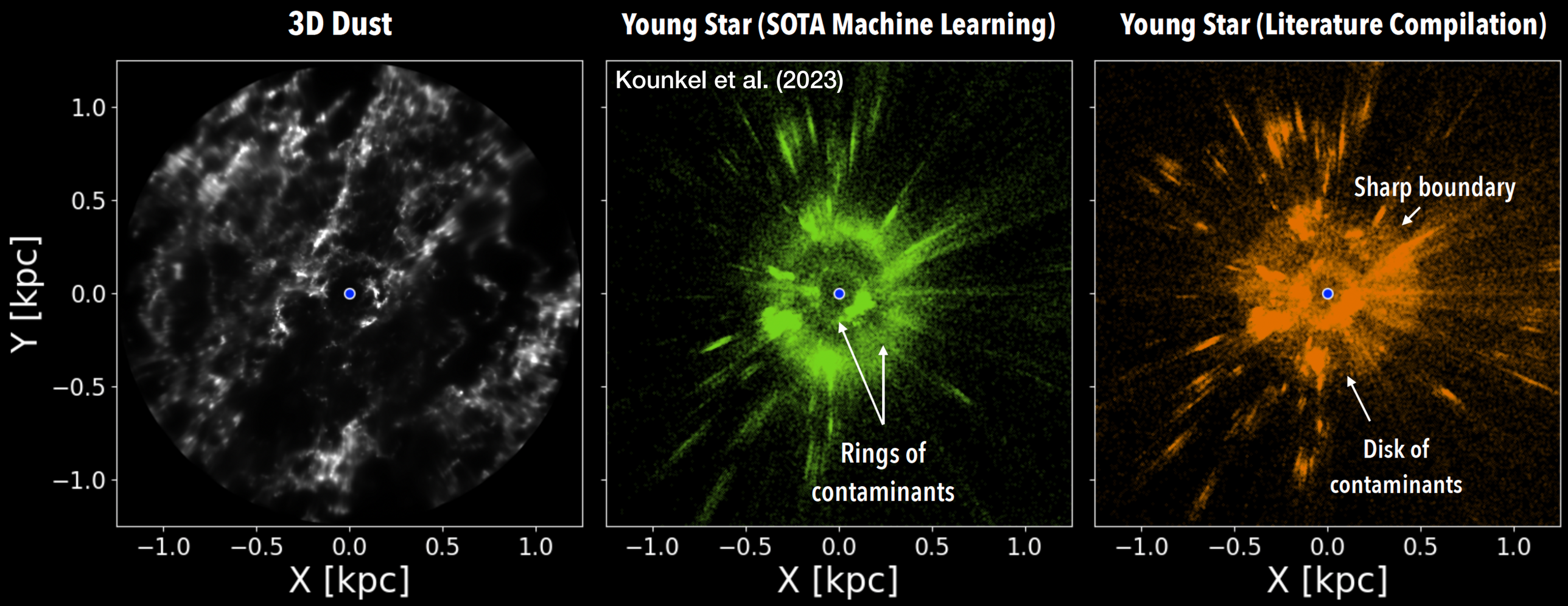




Edenhofer et al. (2024)

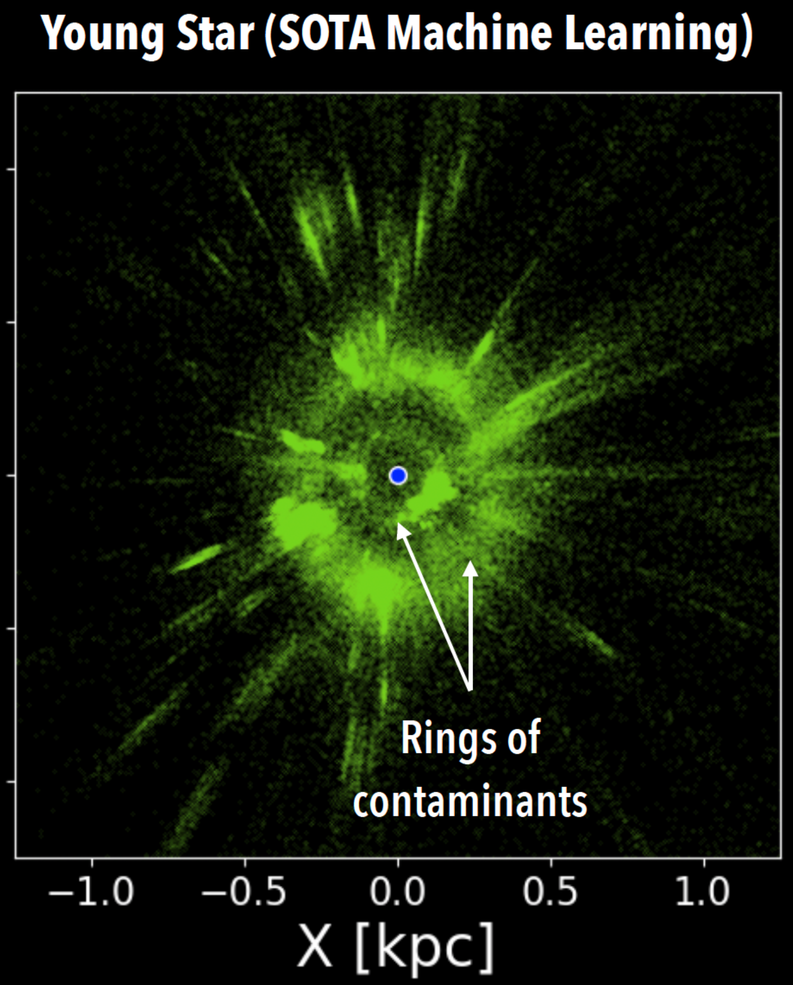


Edenhofer et al. (2024)


3D Dust
Young Star (SOTA Machine Learning)
Young Star (Literature Compilation)
Kounkel et al. (2023)
Rings of contaminants
Sharp boundary
Disk of contaminants
Y [kpc]
X [kpc]
X [kpc]
X [kpc]
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

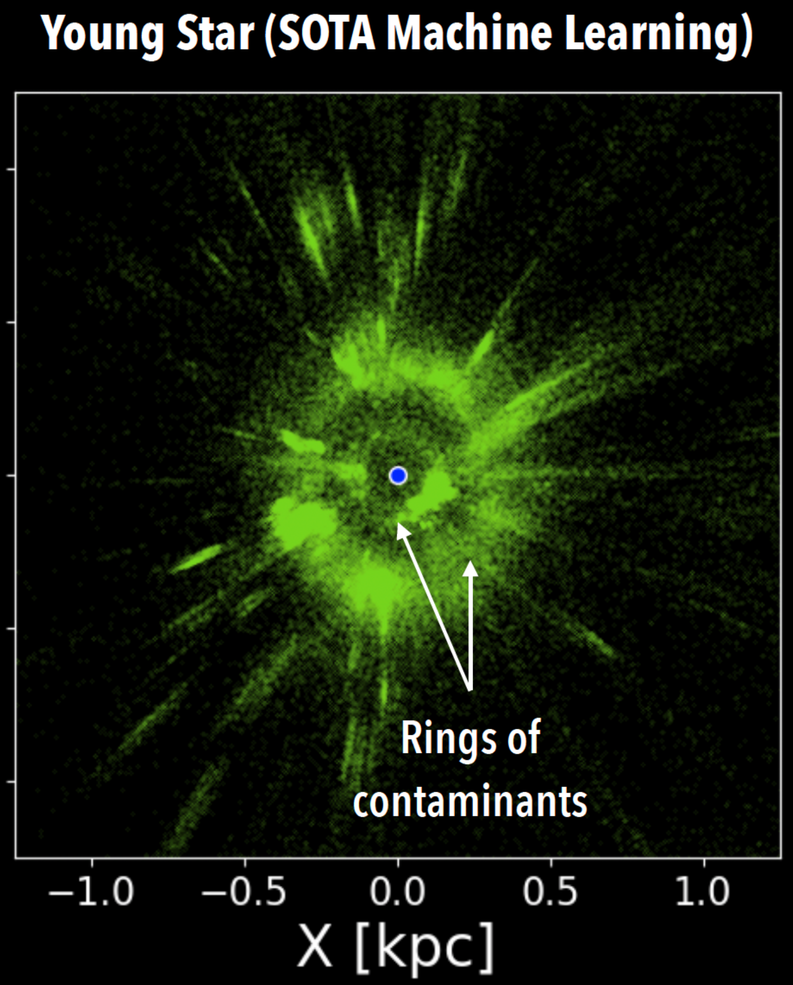
Understanding Galactic baryon cycle

- YSOs connect *cloud* ↔ *stars* ↔ *feedback*



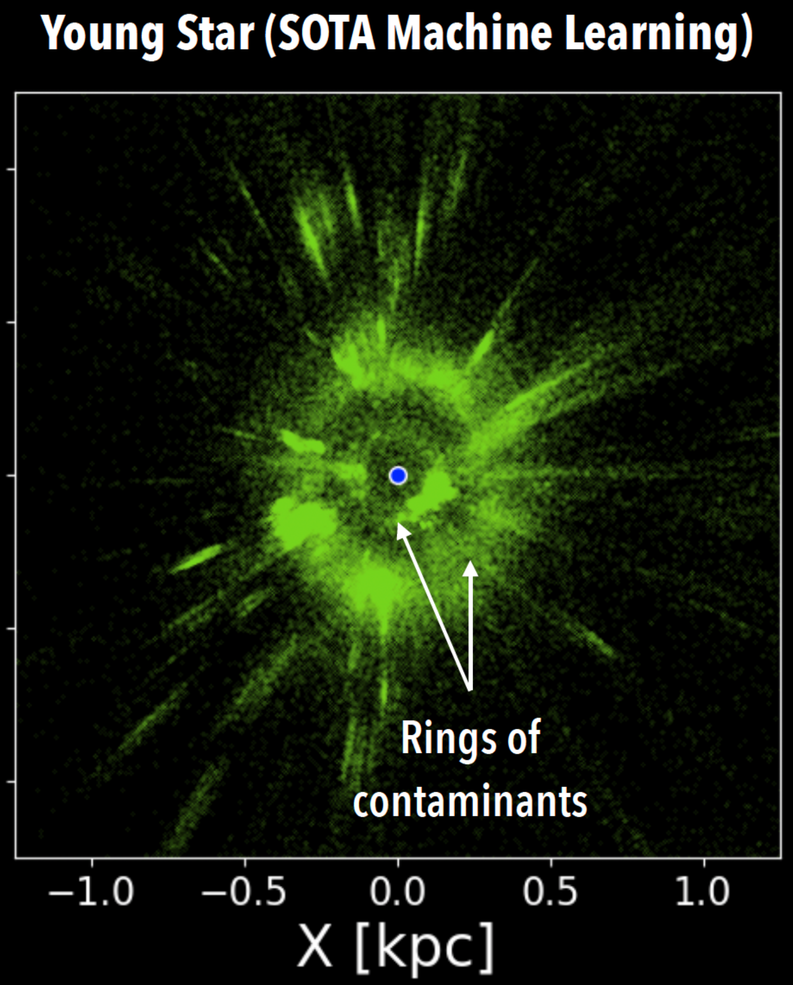
Understanding Galactic baryon cycle

- YSOs connect *cloud* ↔ *stars* ↔ *feedback*
- How does the Milky Way convert gas into stars?



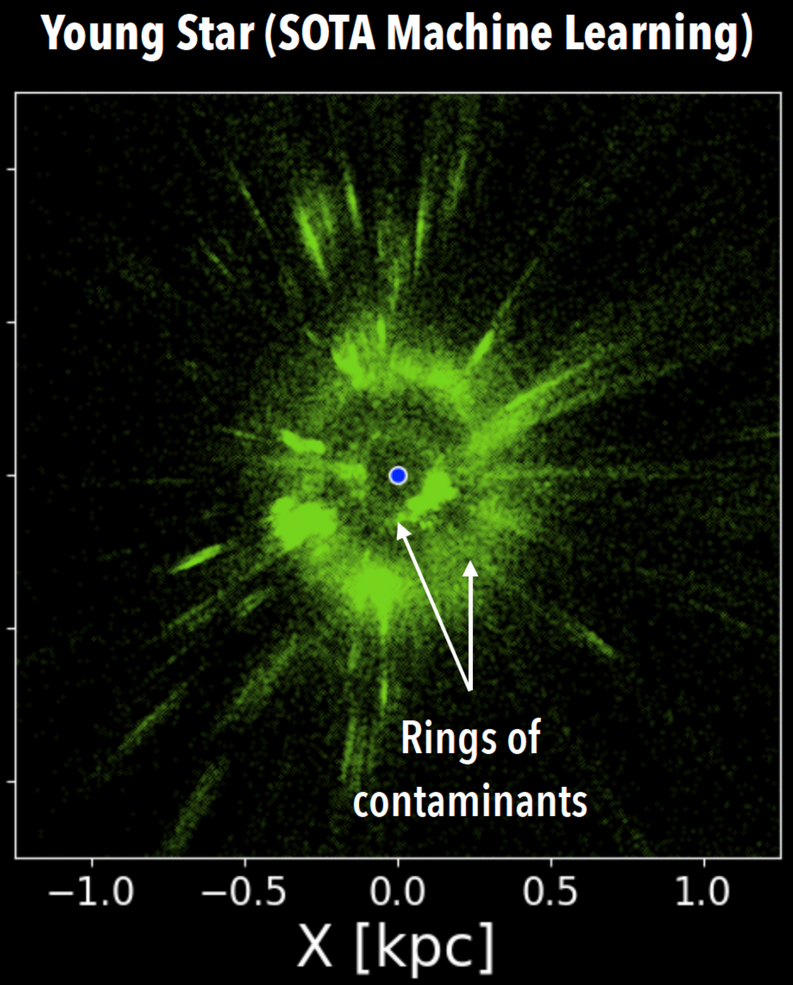
Understanding Galactic baryon cycle

- YSOs connect *cloud* ↔ *stars* ↔ *feedback*
- How does the Milky Way convert gas into stars?
- How do stars leave their birth clouds and shape the ISM?



Understanding Galactic baryon cycle

- YSOs connect *cloud* ↔ *stars* ↔ *feedback*
- How does the Milky Way convert gas into stars?
- How do stars leave their birth clouds and shape the ISM?
- How do supernovae regulate, trigger, or suppress new generations of stars?



Aim to improve YSO catalog

- Employ dedicated YSO models

Aim to improve YSO catalog

- Employ dedicated YSO models
- Data fusion: use as many informative data sets as possible

Aim to improve YSO catalog

- Employ dedicated YSO models
- Data fusion: use as many informative data sets as possible
- Produce well-calibrated posteriors over stellar parameters given spectra & photometric observations

Aim to improve YSO catalog

- Employ dedicated YSO models
- Data fusion: use as many informative data sets as possible
- Produce well-calibrated posteriors over stellar parameters given spectra & photometric observations
- Scale inference to > 1M - 1B stars

Challenges with “1 model does it all” approach

- Fusing surveys is hard due to different
  - resolutions & depths
  - coverage
  - instrument response
  - noise model
  - …

Challenges with “1 model does it all” approach

- Fusing surveys is hard due to different
  - resolutions & depths
  - coverage
  - instrument response
  - noise model
- **Model misspecification** leads to domain shift between simulated and real data

Challenges with “1 model does it all” approach

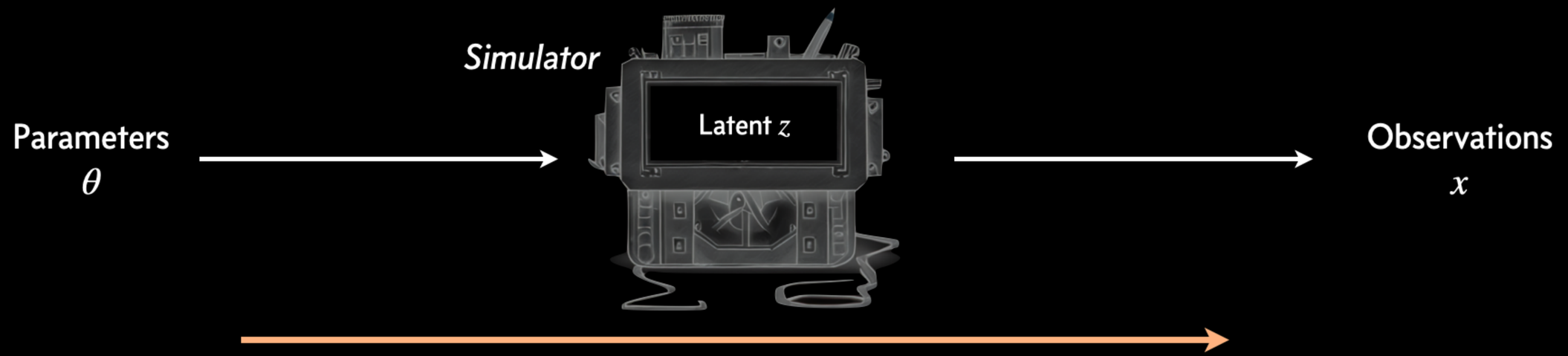
- Fusing surveys is hard due to different
  - resolutions & depths
  - coverage
  - instrument response
  - noise model
- **Model misspecification** leads to domain shift between simulated and real data

→ ***Domain-Adaptive*** *SBI w/* ***incomplete, multi-survey*** *data*

Model implementation

I. SBI model

Simulation based inference setup



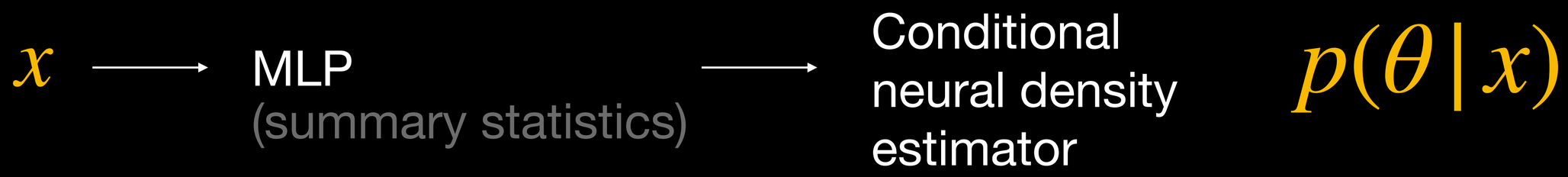
Prediction:

- Mechanistic forward model
- We can generate samples from a simulator $x \sim p(x \mid \theta)$

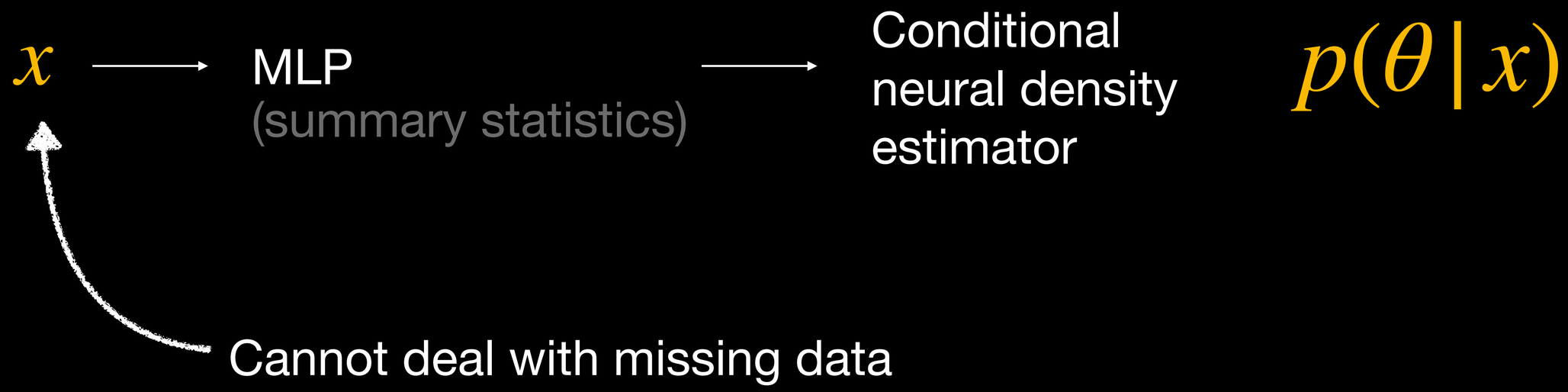
Inference:

- Likelihood $p(x \mid \theta) = \int \mathrm{d}z\, p(x, z \mid \theta)$ is intractable
- *Inference is challenging*

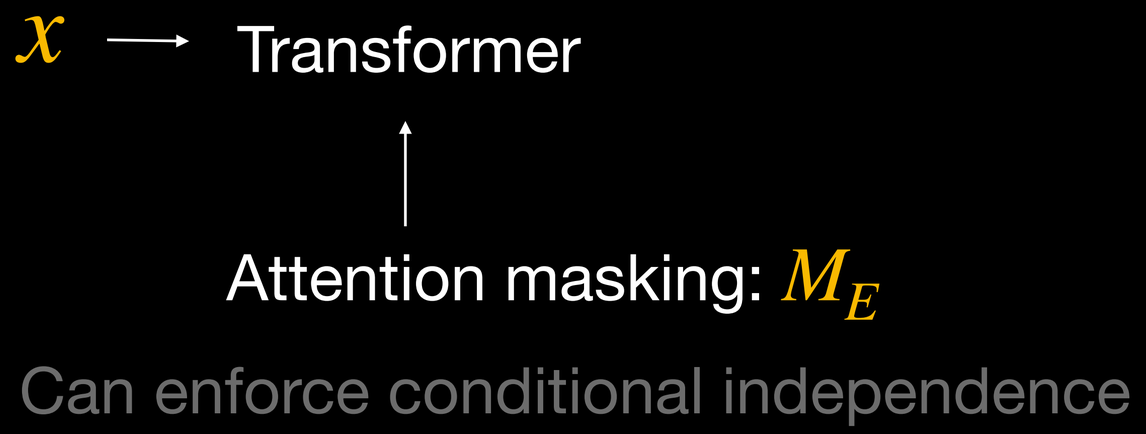
Neural posterior estimation



Neural posterior estimation



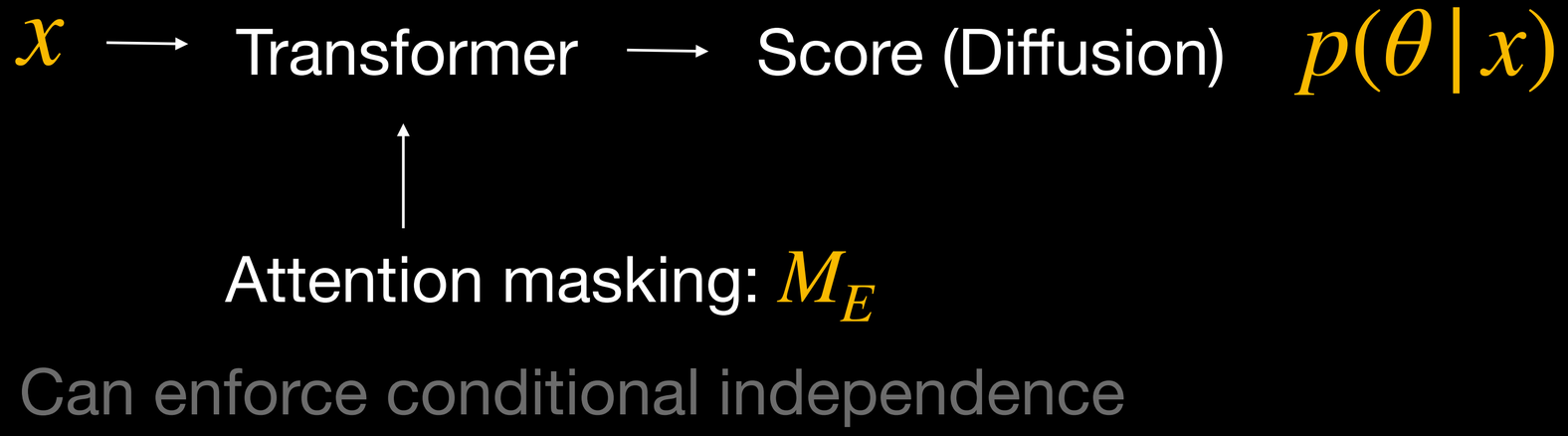
Simformer: learning with incomplete data



Can enforce conditional independence

Gloeckler et al. (2024)

Simformer: learning with incomplete data



Can enforce conditional independence

Gloeckler et al. (2024)

Model implementation

II. Domain adaption

Architecture

$(X, \theta) \longrightarrow$ Transformer $\longrightarrow$ Score (Diffusion)

Input split into *simulated*, *real* & *paired* data

$$\theta$$

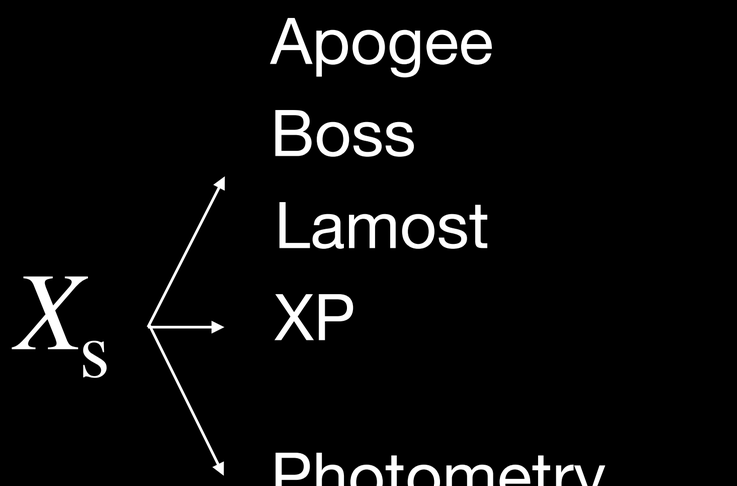
$$X \rightarrow \begin{matrix} X_{\text{sim}} \\ X_{\text{real}} \\ X_{\text{sim-real-pairs}} \end{matrix}$$

→ Transformer → Diffusion

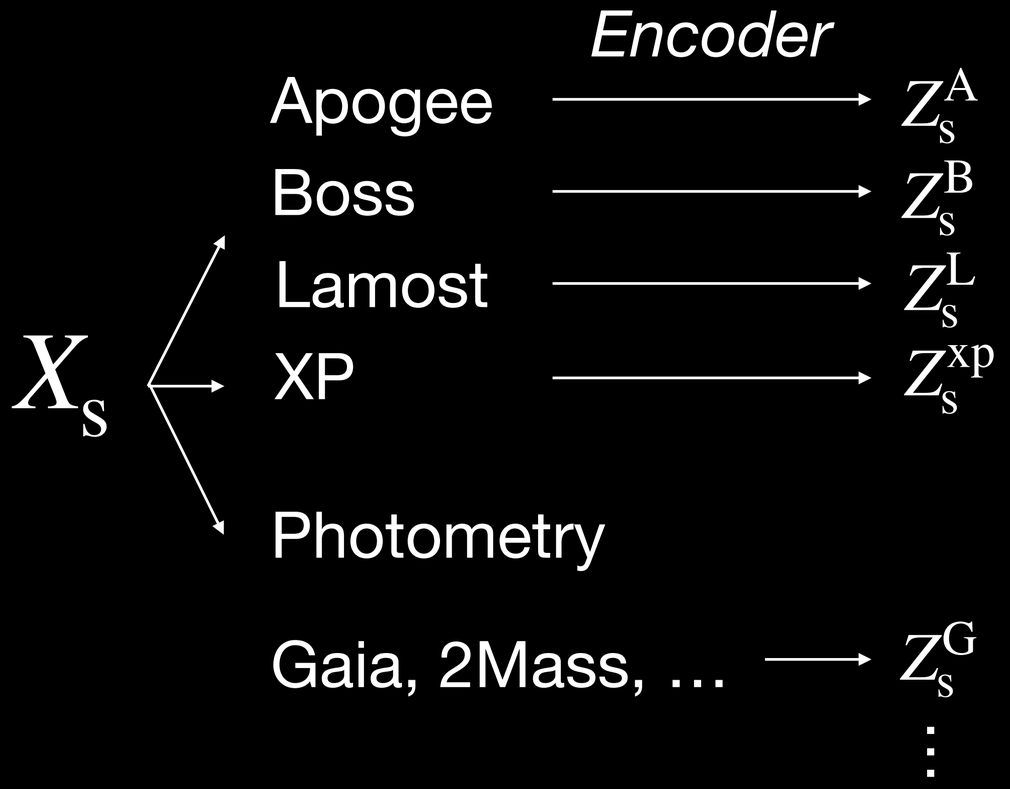
Modality encoders

$$X \rightarrow \begin{matrix} X_{\mathrm{s}} \\ X_{\mathrm{r}} \\ X_{\mathrm{s-r}} \end{matrix}$$

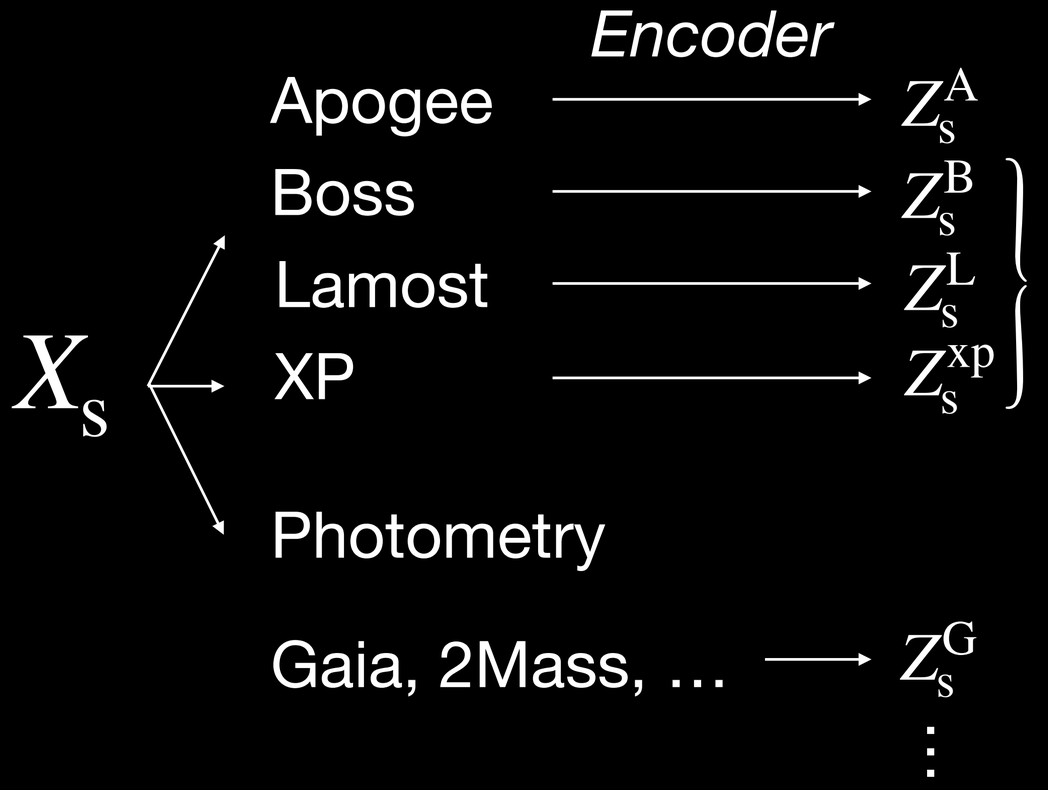
Modality encoders: split into indiv. spectra



Modality encoders: encode



Modality encoders: contrastive loss

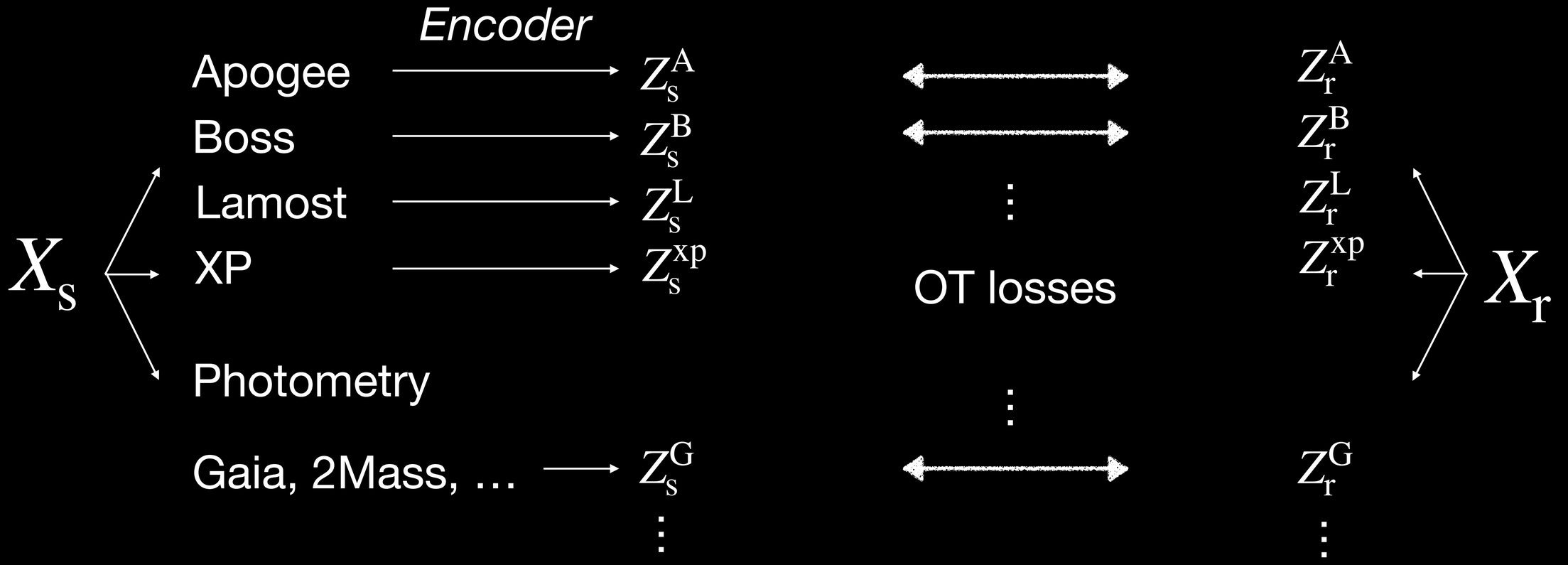


Pool embedding (mean)

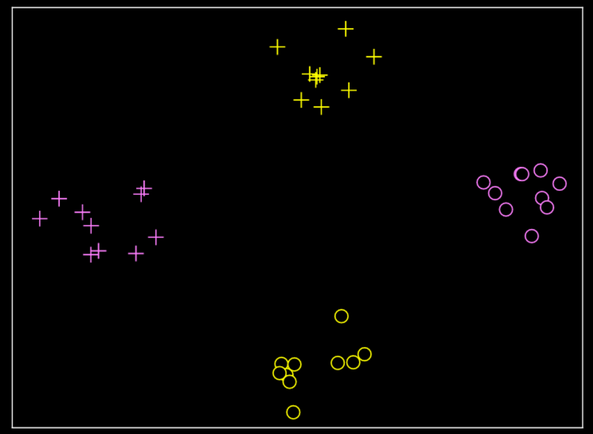
Contrastive loss: $Z^{\{xp,L,B\}}$ & Z^{pool}

Goal: learn shared information across spectra, that's (hopefully) invariant of detector systems

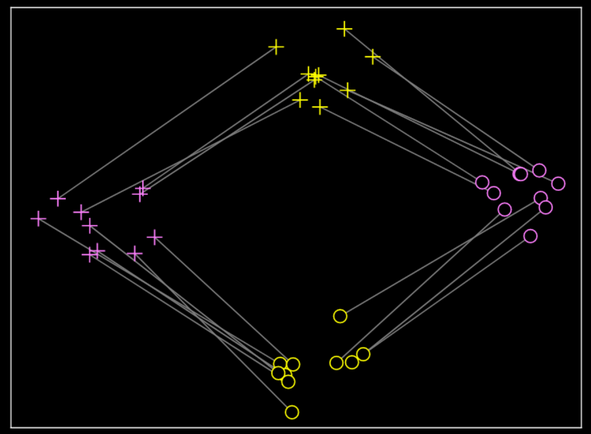
Modality encoders: alignment loss



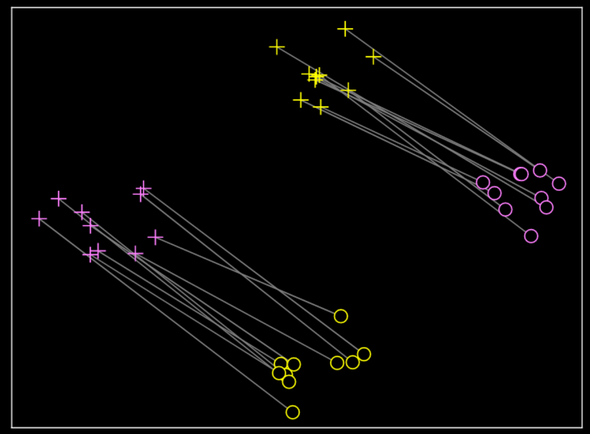
Alignment via optimal transport



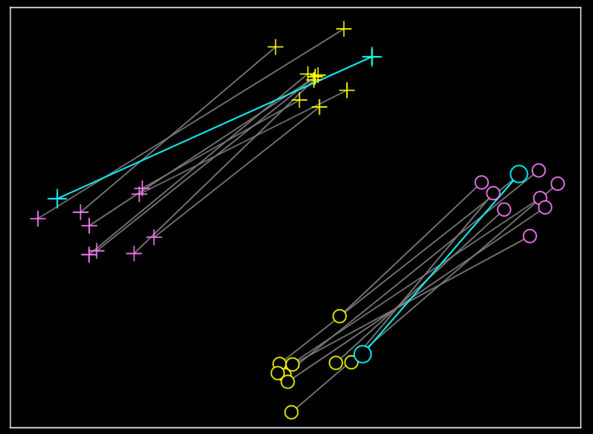
(a) Samples



(b) Exact



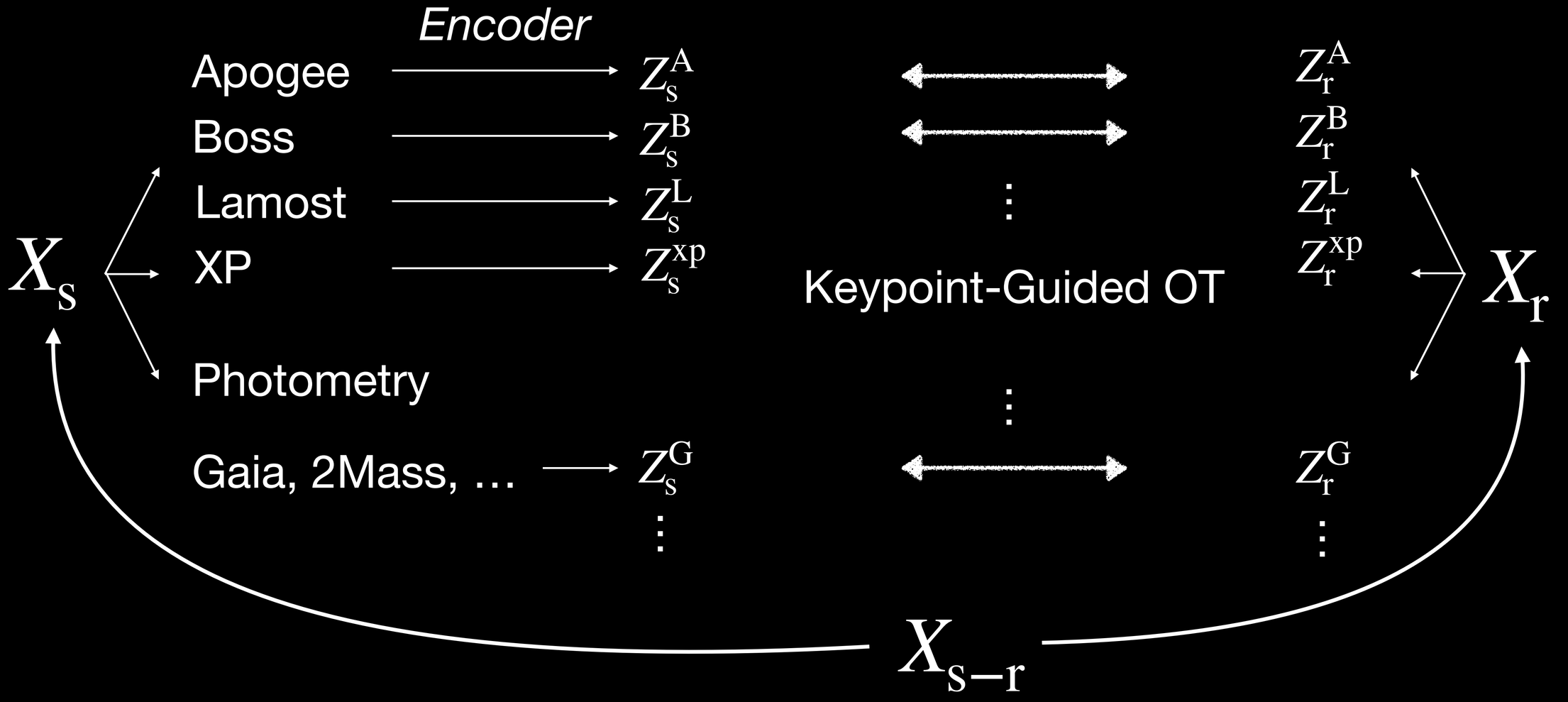
(c) Gromov-Wasserstein



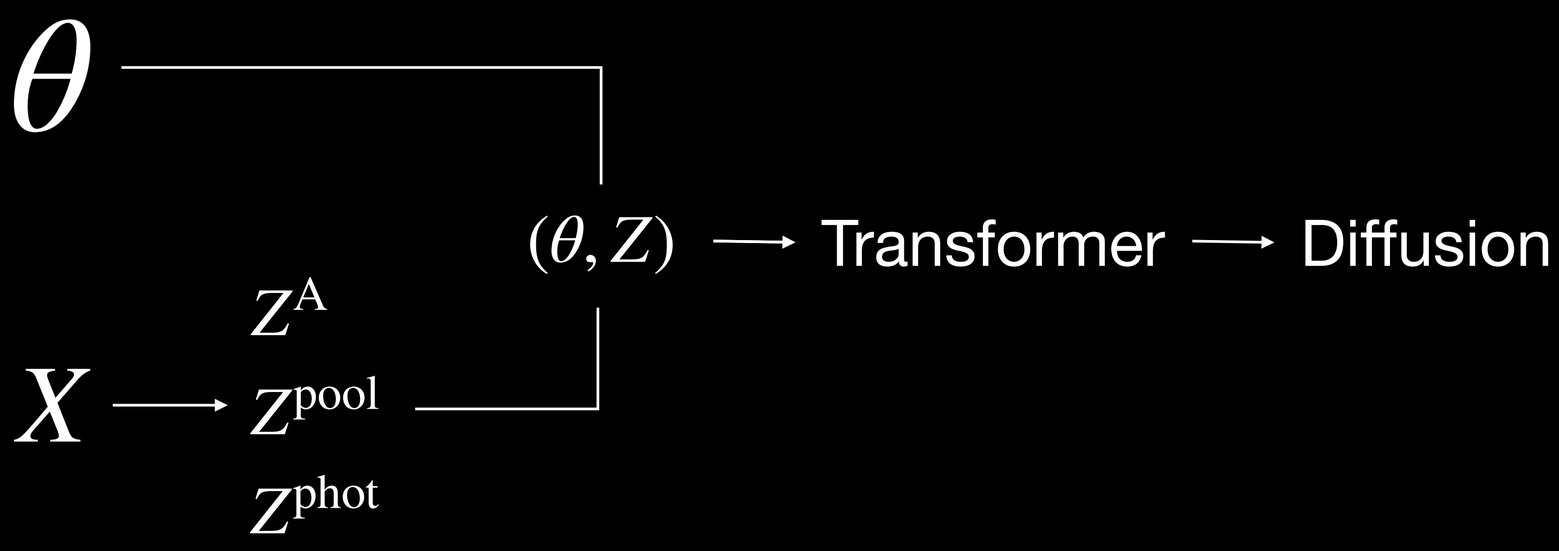
Gu et al. (2022)

Gu et al. (2022)

Additional pair constraint



Final model

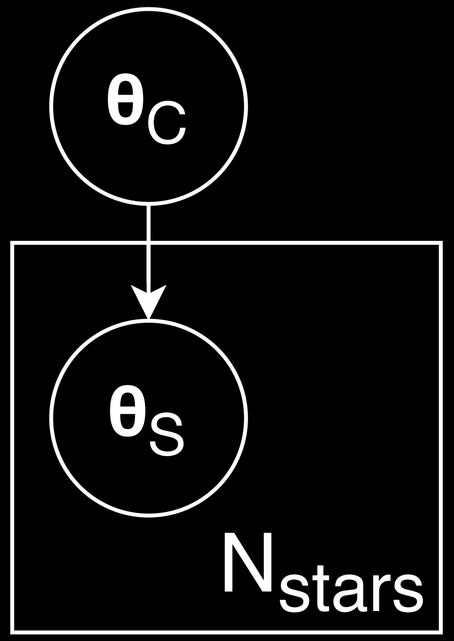


Loss

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) + \mathscr{L}_C + \mathscr{L}_{OT} + \mathscr{L}_{R-S}$$

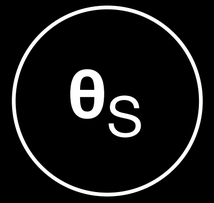
Forward model

Clusters/young stars

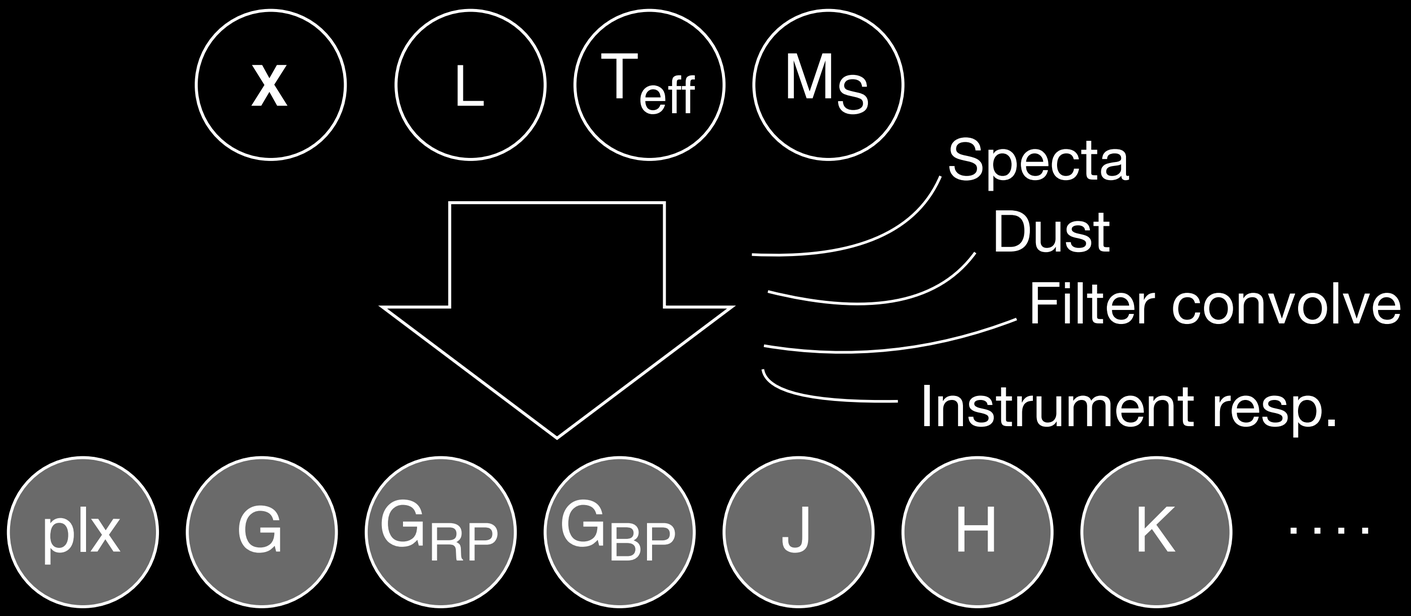


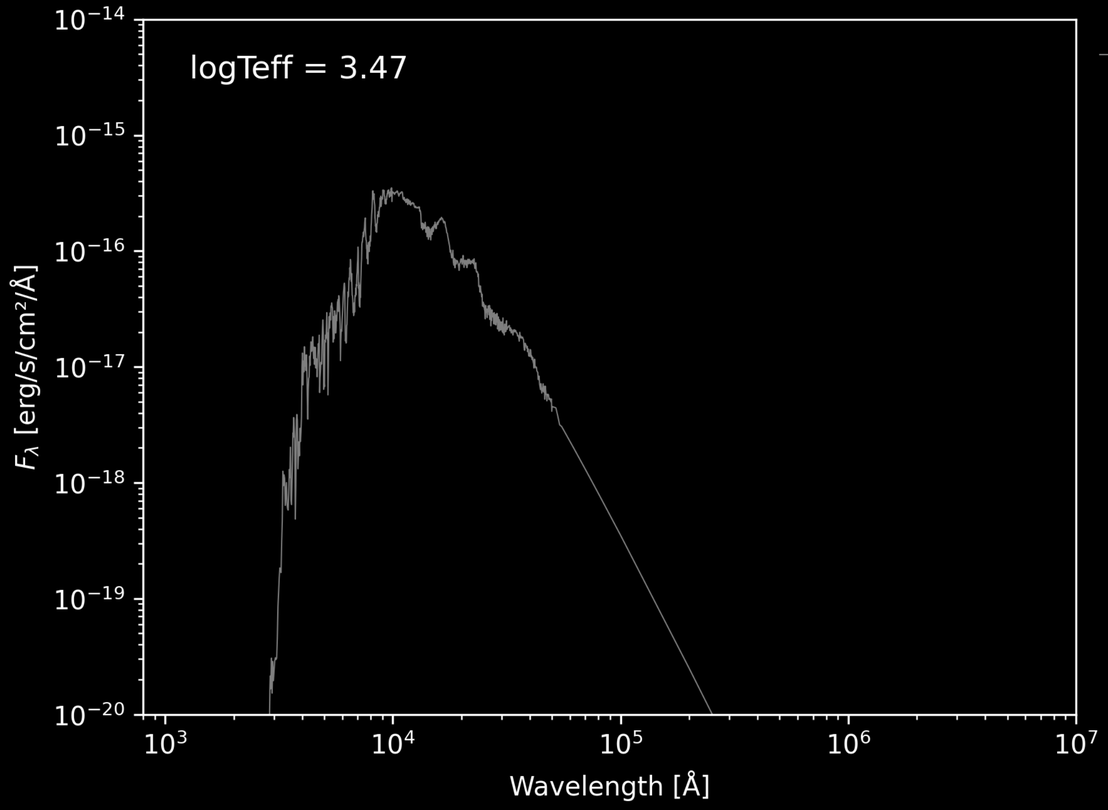
Milky Way model

- Rybizki et al. (2018)
- *galaxia* code
- thin+thick+halo+bulge

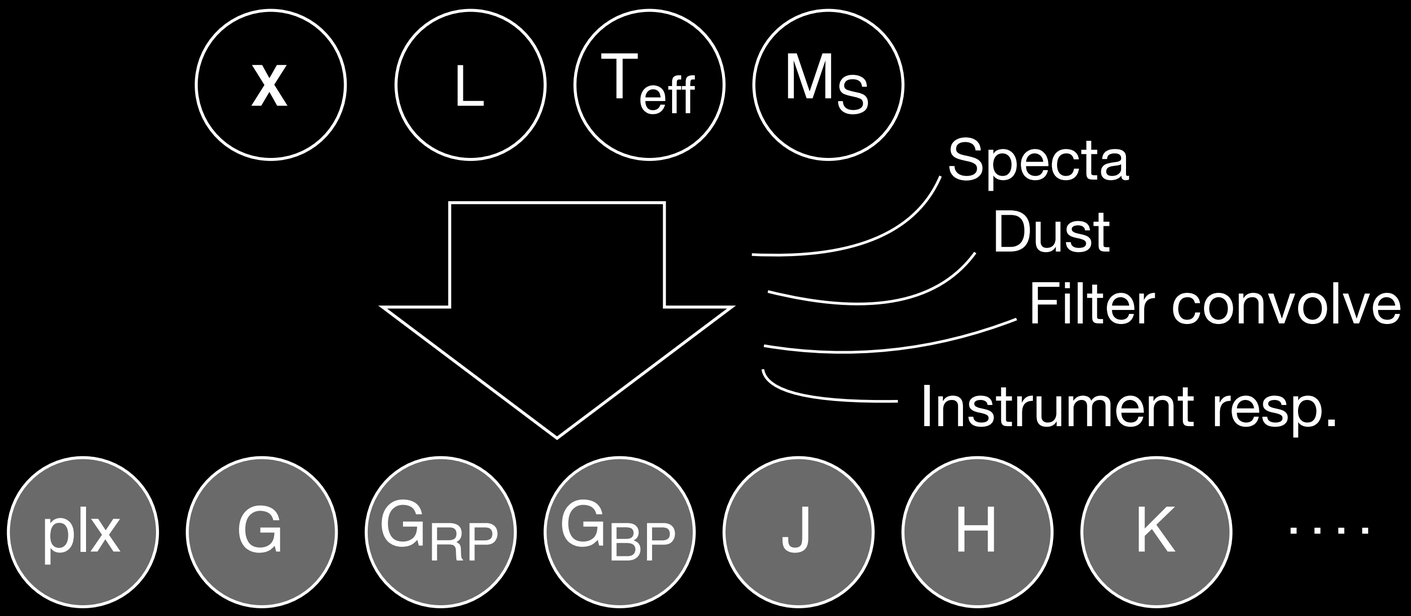


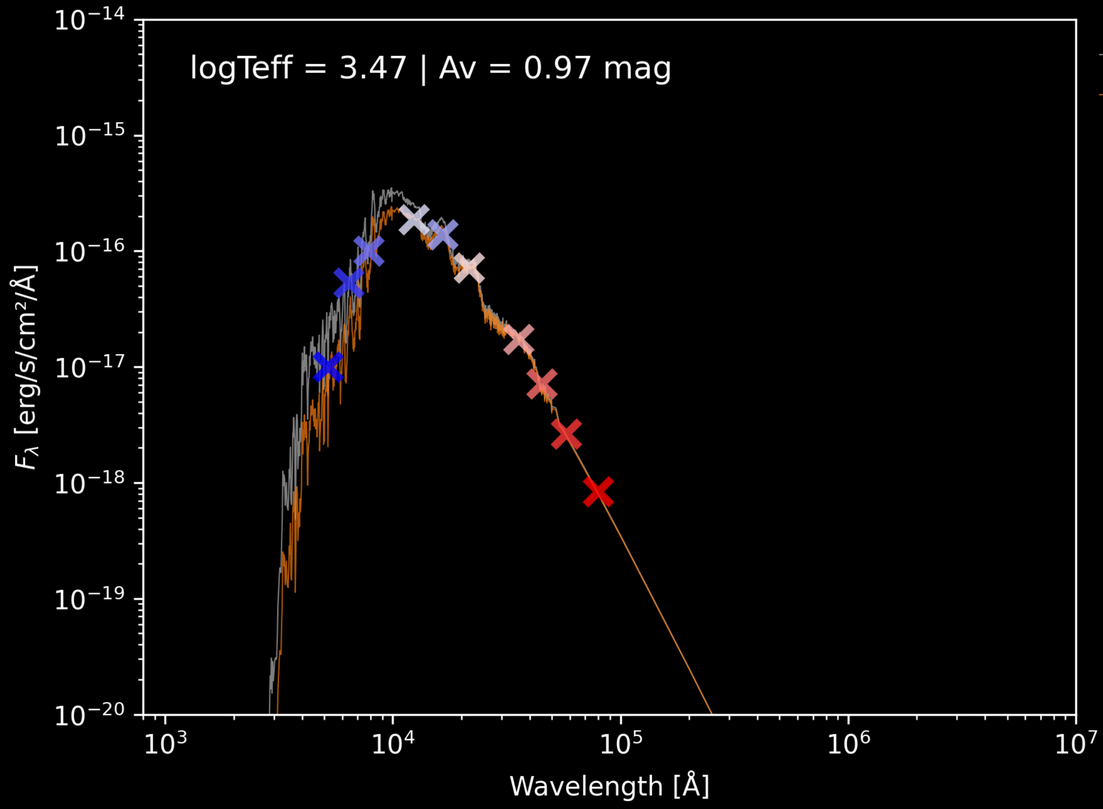
Fwd model


x
L
T_{eff}
M_S
Specta
Dust
Filter convolve
Instrument resp.
plx
G
G_{RP}
G_{BP}
J
H
K
....

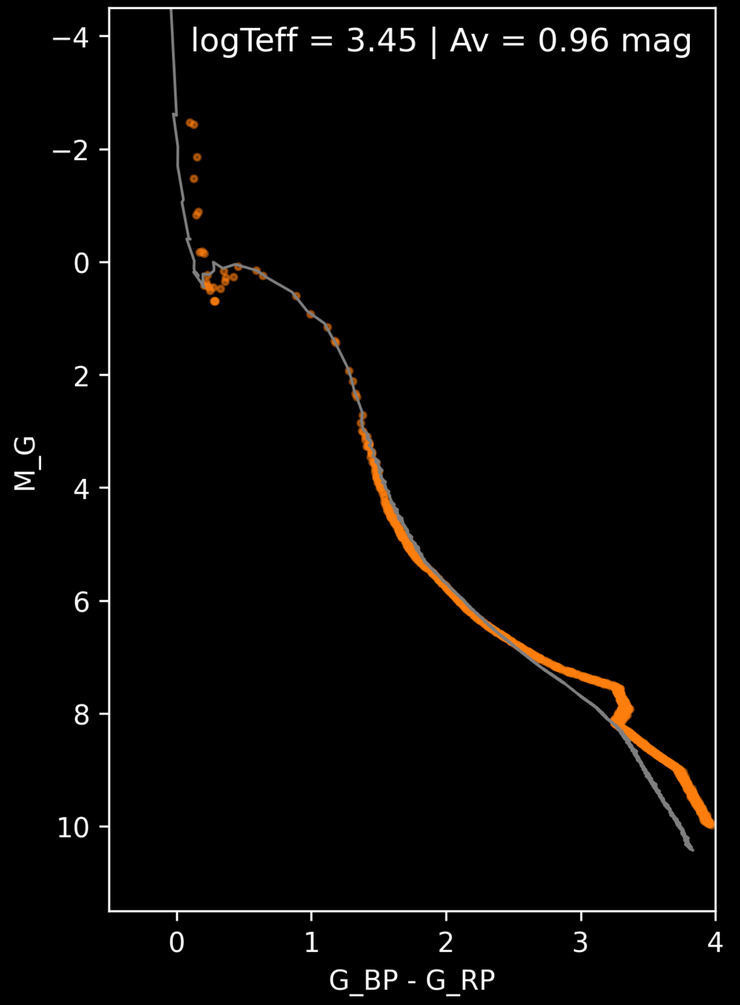

logTeff = 3.47
10^{-14}
10^{-15}
10^{-16}
10^{-17}
10^{-18}
10^{-19}
10^{-20}
F_λ [erg/s/cm²/Å]
10^3
10^4
10^5
10^6
10^7
Wavelength [Å]

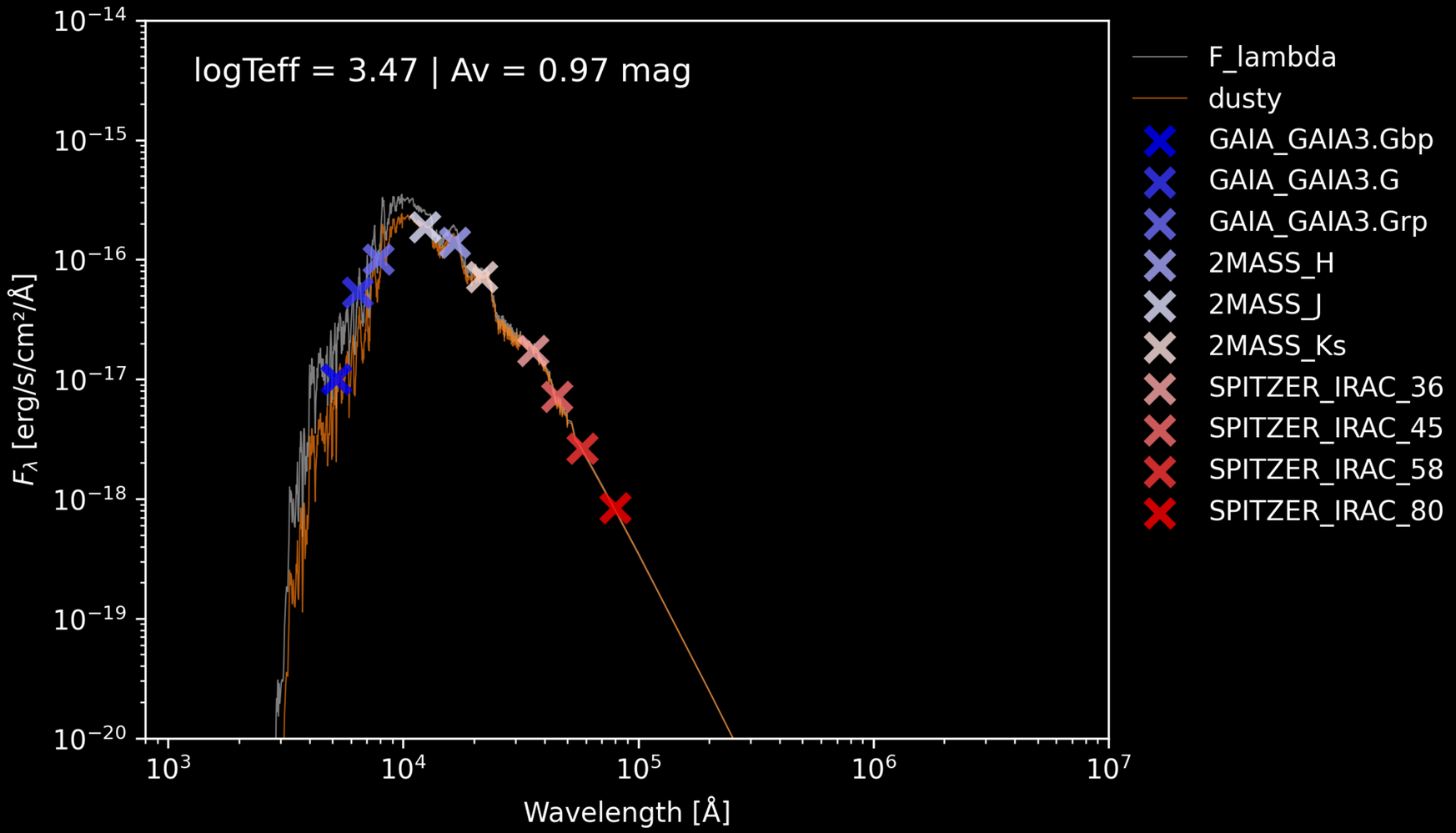
Fwd model


x
L
T_{eff}
M_S
Specta
Dust
Filter convolve
Instrument resp.
plx
G
G_{RP}
G_{BP}
J
H
K
....

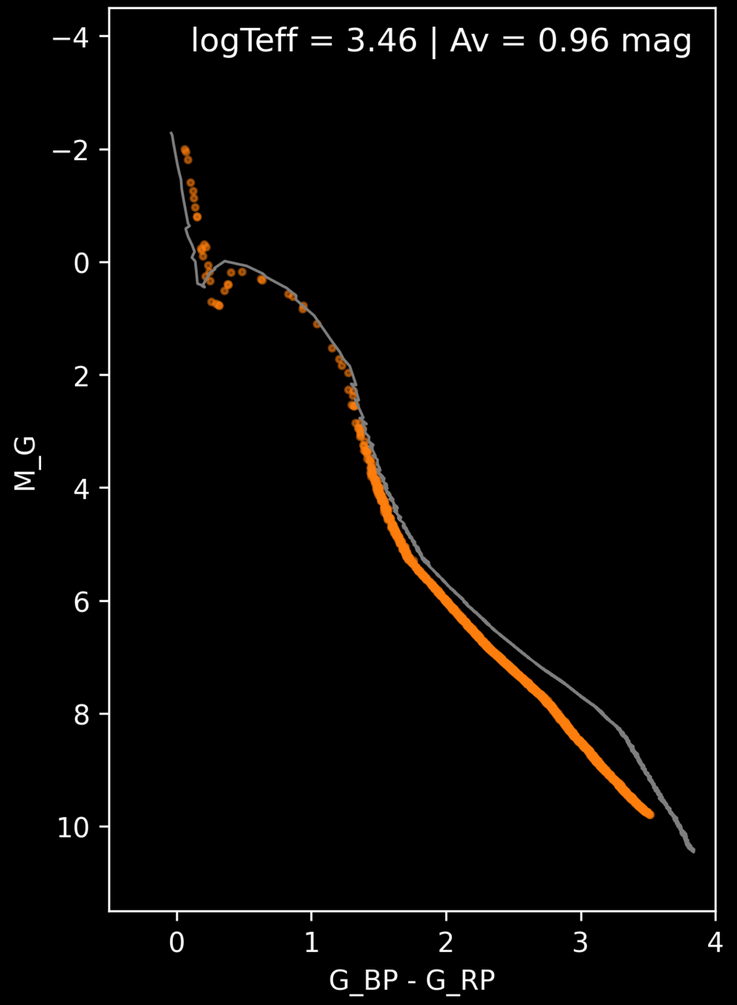

logTeff = 3.47 | Av = 0.97 mag
F_λ [erg/s/cm²/Å]
Wavelength [Å]
10^{-14}
10^{-15}
10^{-16}
10^{-17}
10^{-18}
10^{-19}
10^{-20}
10^3
10^4
10^5
10^6
10^7

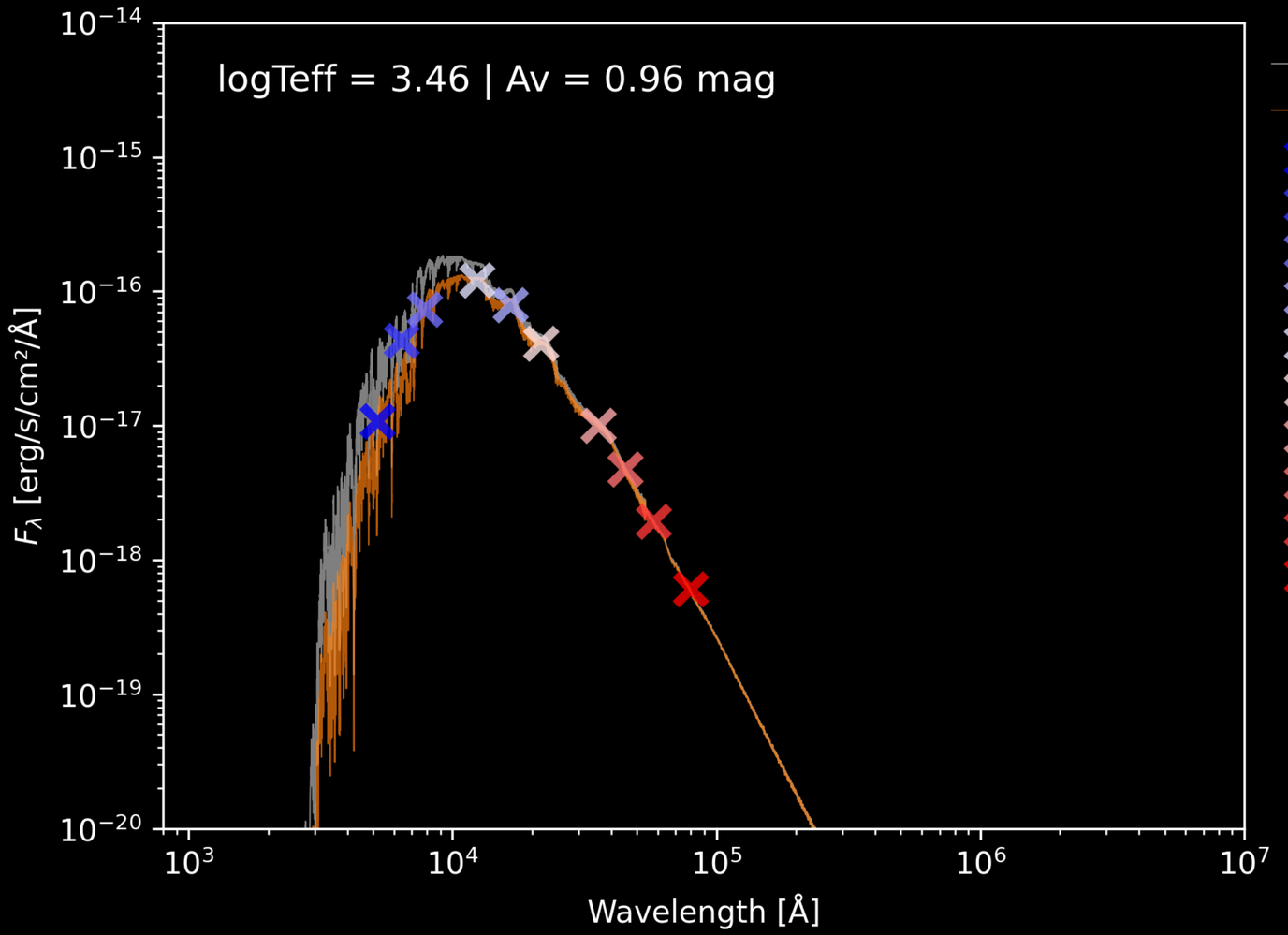
Model differences (BaSeL)



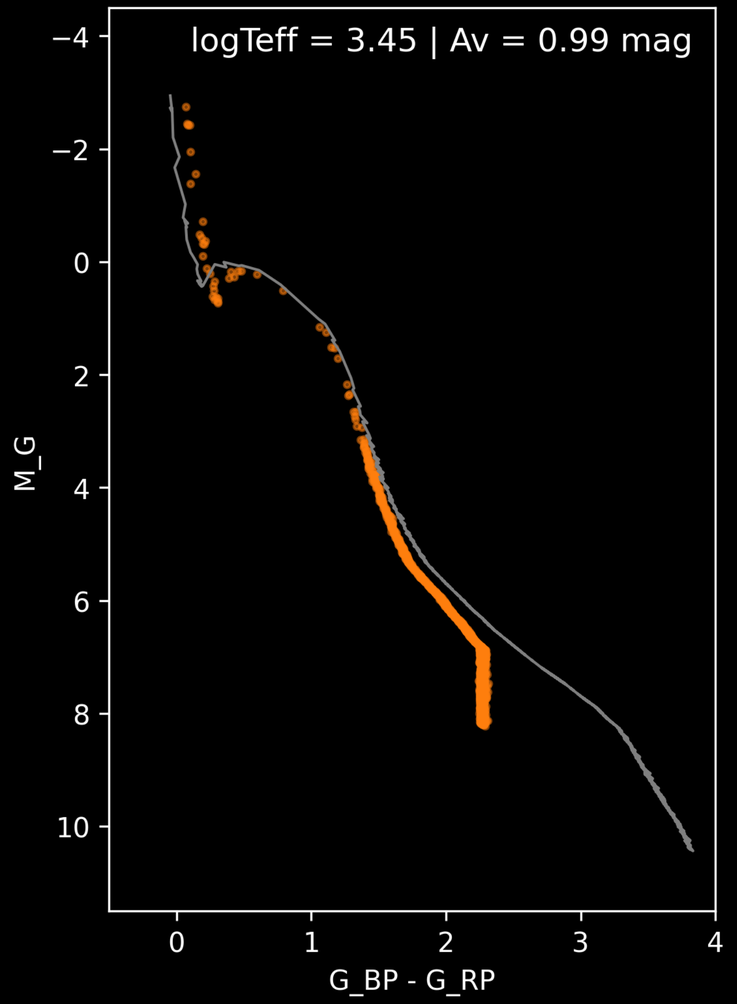


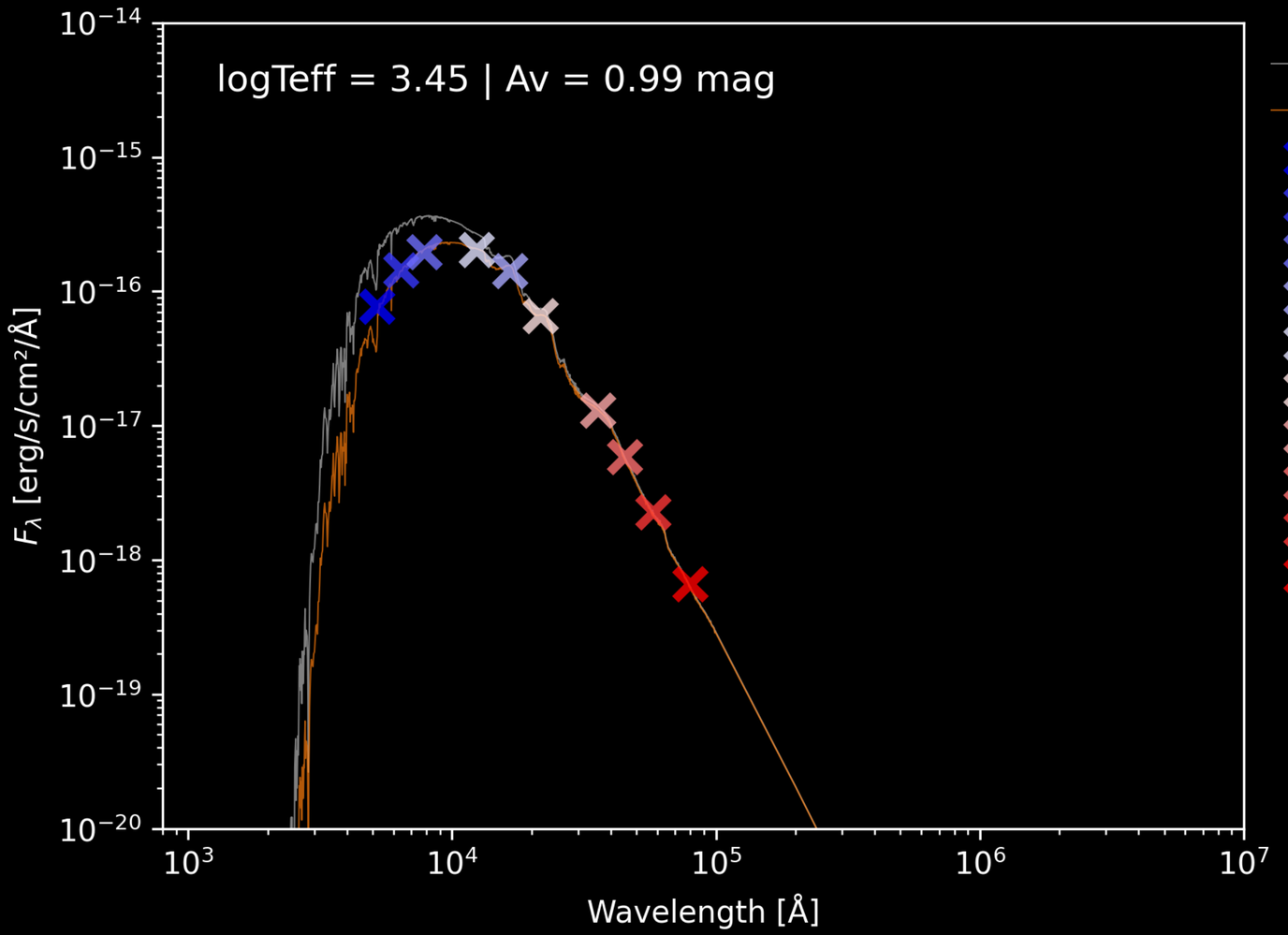
Model differences (BTSettl)





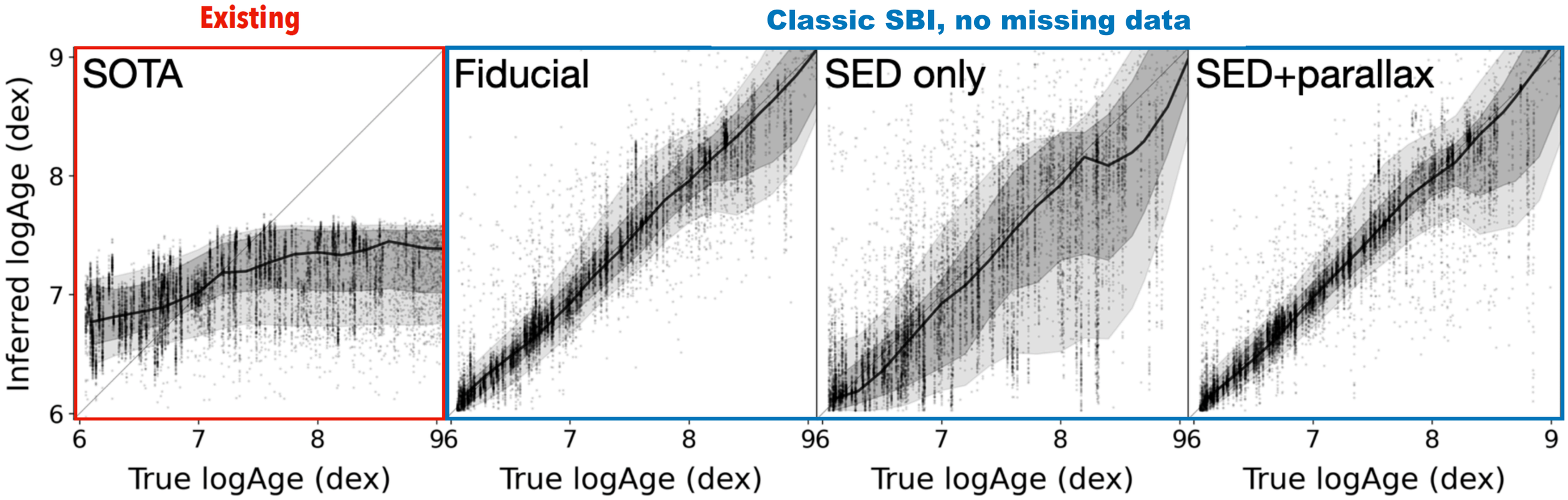
Model differences (Kurucz)



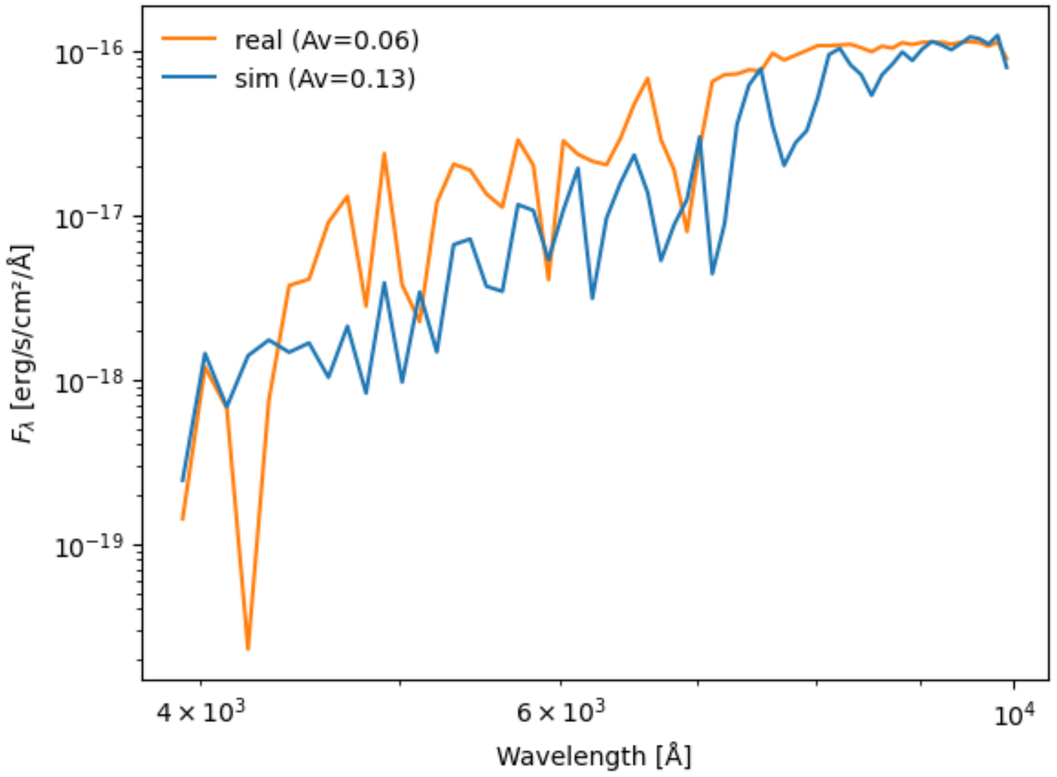


Results

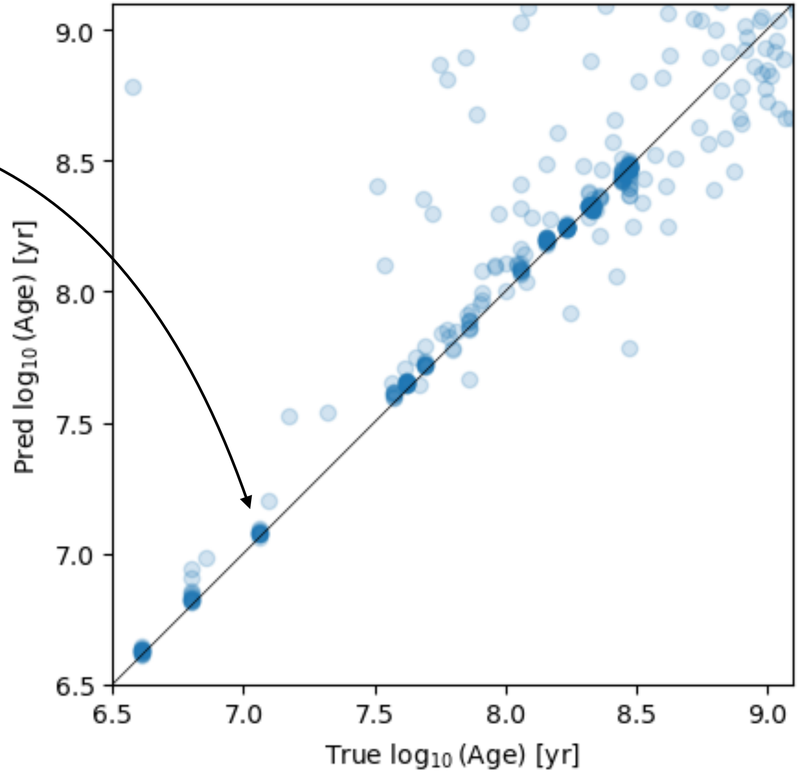
Pilot study: Gaia+2MASS+WISE



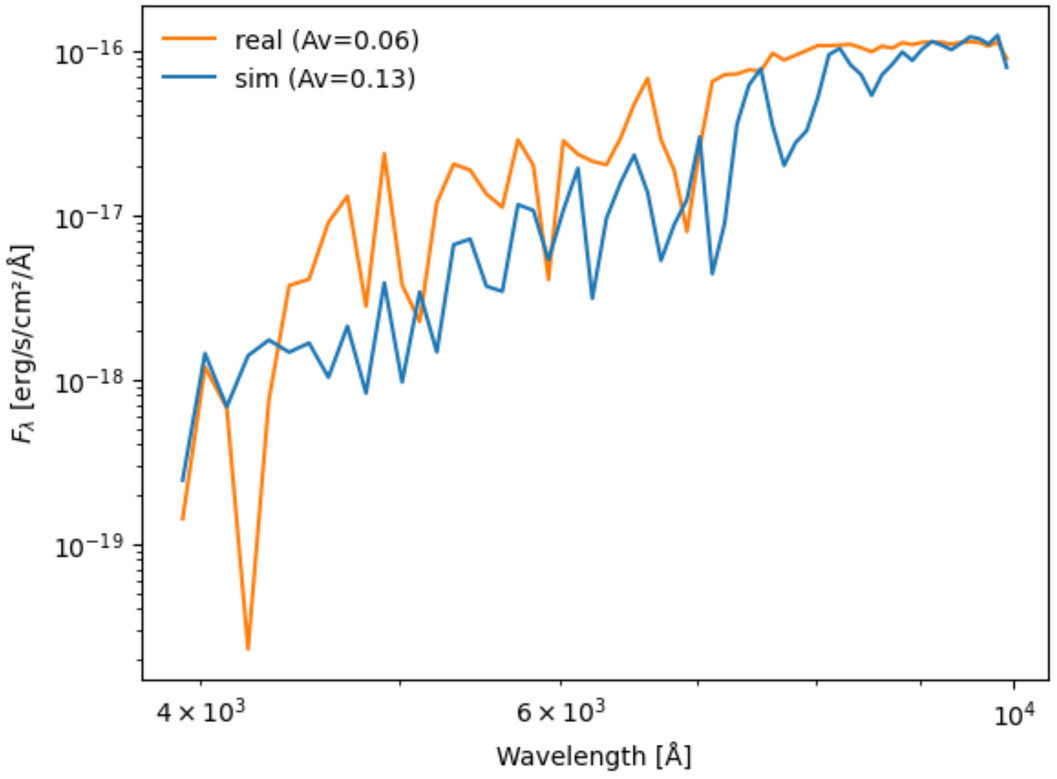
Updated pipeline + 50% missing + XP spectra

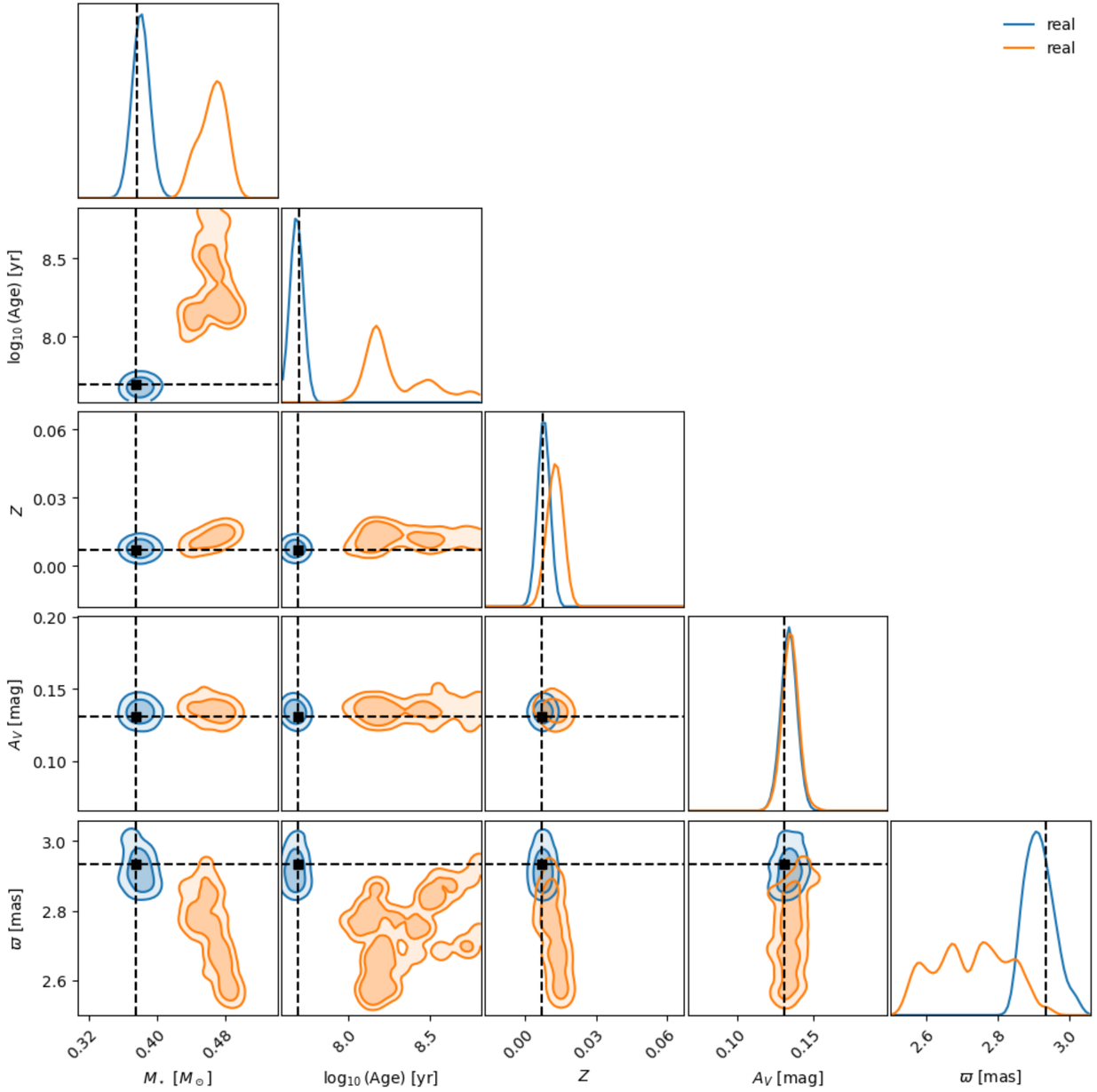


Posterior mean **sim**

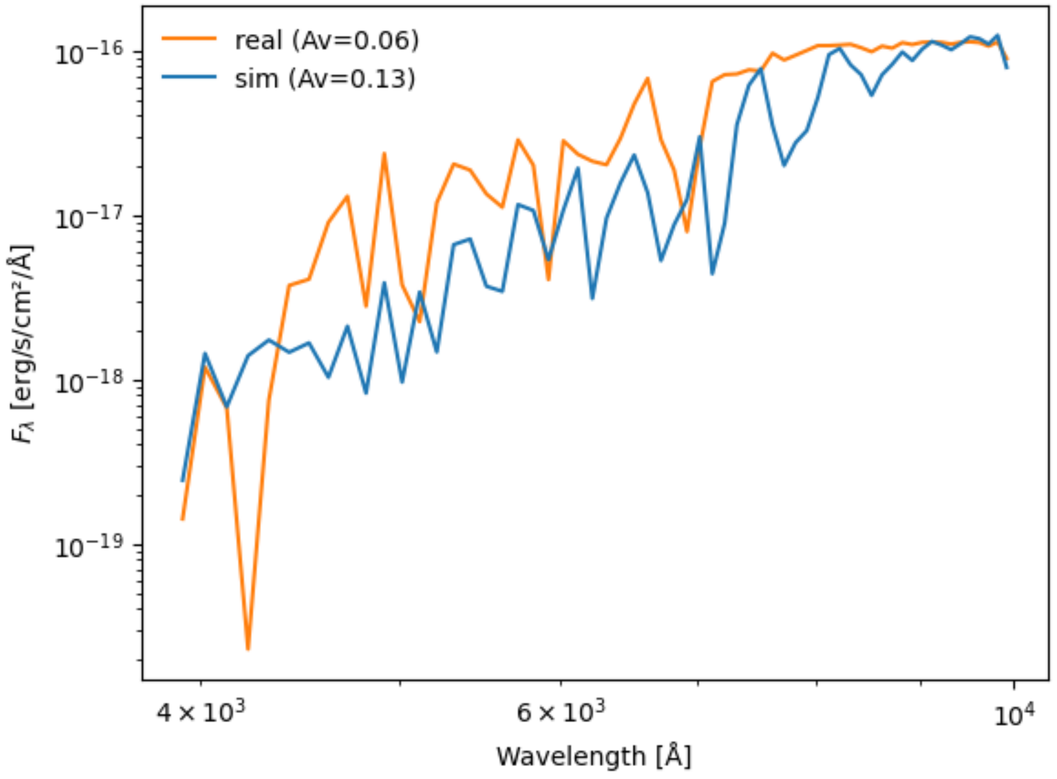


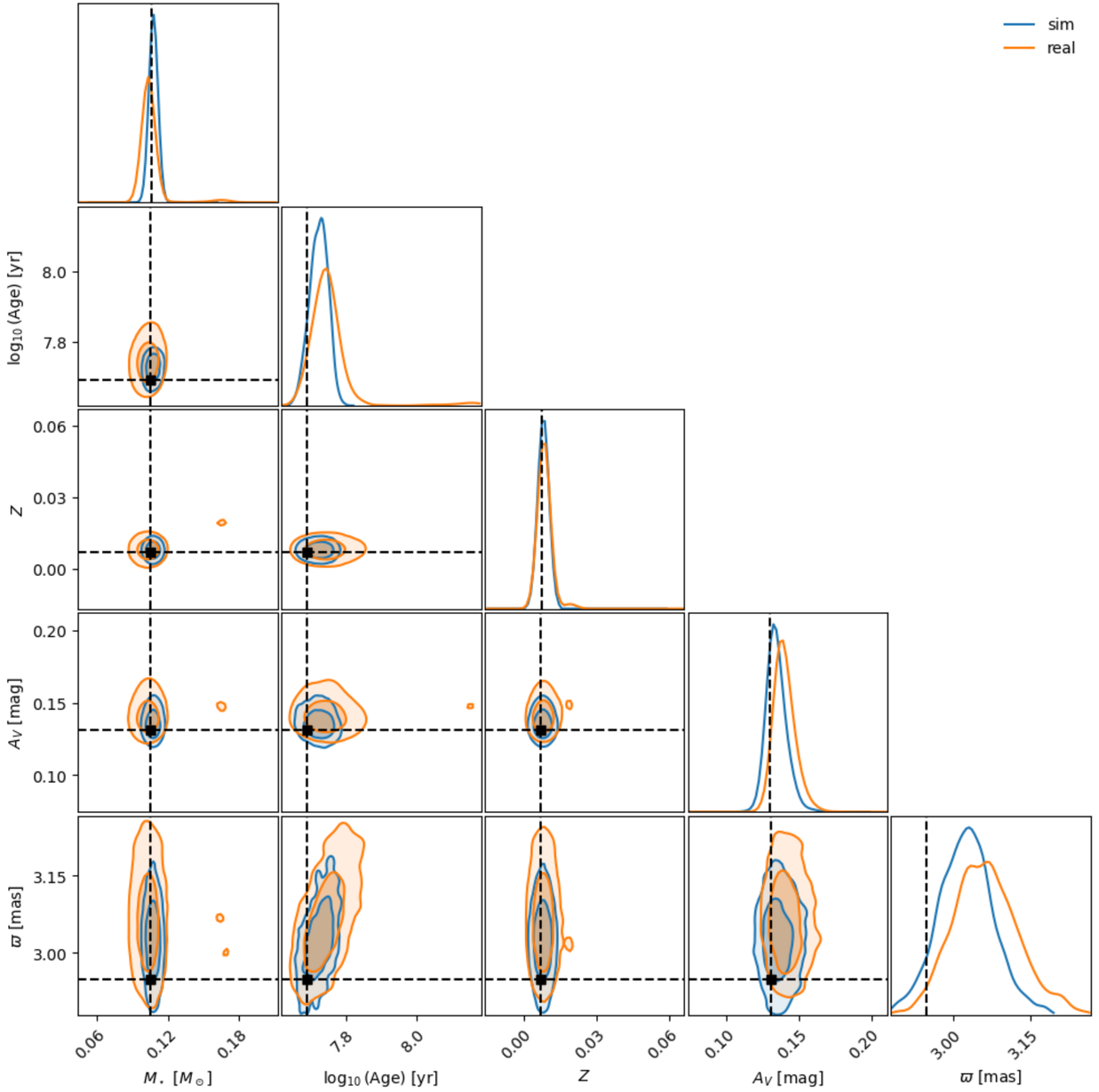
Without DA


real (Av=0.06)
sim (Av=0.13)
10^{-16}
10^{-17}
10^{-18}
10^{-19}
F_λ [erg/s/cm²/Å]
4×10^3
6×10^3
10^4
Wavelength [Å]


real
real
$\log_{10}$(Age) [yr]
8.5
8.0
Z
0.06
0.03
0.00
A_V [mag]
0.20
0.15
0.10
ϖ [mas]
3.0
2.8
2.6
0.32
0.40
0.48
$M_\star$ [$M_\odot$]
8.0
8.5
$\log_{10}$(Age) [yr]
0.00
0.03
0.06
Z
0.10
0.15
A_V [mag]
2.6
2.8
3.0
ϖ [mas]

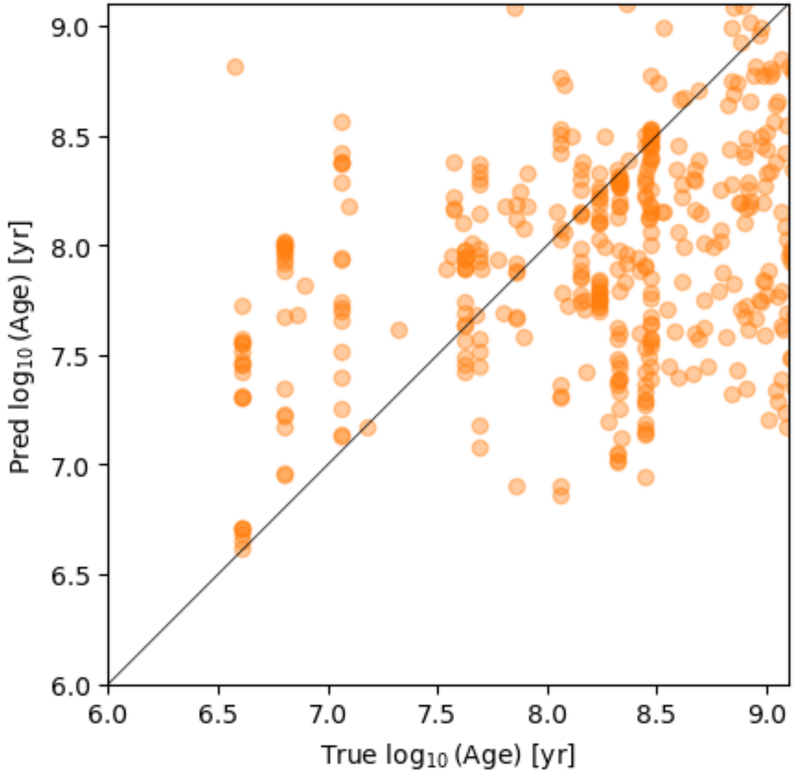
With DA


real (Av=0.06)
sim (Av=0.13)
10^{-16}
10^{-17}
10^{-18}
10^{-19}
F_λ [erg/s/cm²/Å]
4×10^3
6×10^3
10^4
Wavelength [Å]


sim
real
$\log_{10}$(Age) [yr]
8.0
7.8
Z
0.06
0.03
0.00
A_V [mag]
0.20
0.15
0.10
ϖ [mas]
3.15
3.00
0.06
0.12
0.18
$M_\star$ [$M_\odot$]
7.8
8.0
$\log_{10}$(Age) [yr]
0.00
0.03
0.06
Z
0.10
0.15
0.20
A_V [mag]
3.00
3.15
ϖ [mas]

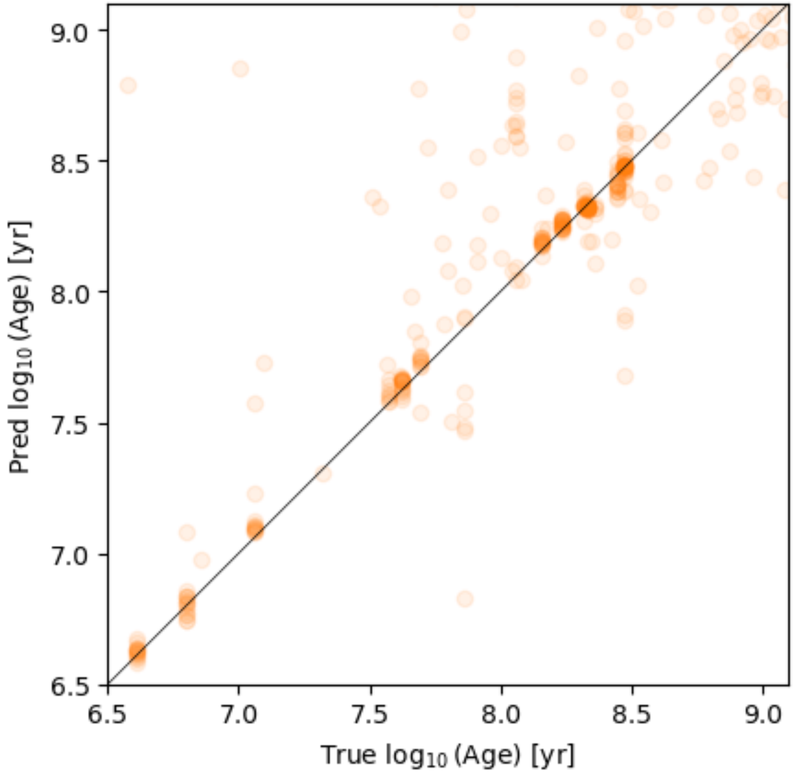
Predictions

Posterior mean **real**



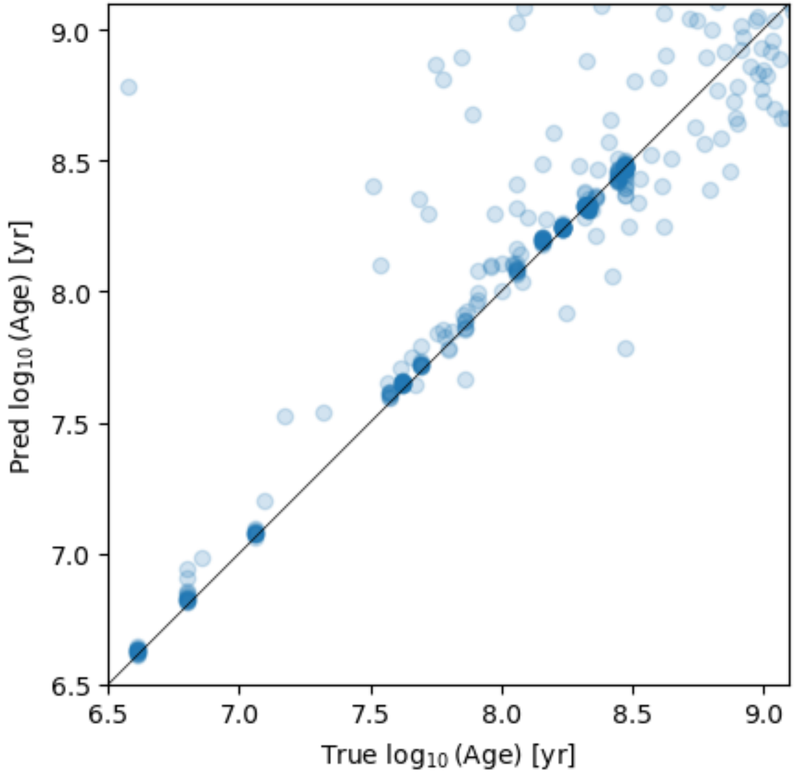
no domain adaption

Posterior mean **real**



with domain adaption

Posterior mean **sim**



Summary

- Combine **diffusion models** + **transformer model** to learn arbitrary **conditionals** and **marginals**
- Add OT + pair + contrastive loss to close domain gap
- Obtain promising results on simulations

Thank you!

Backup

Quick recap

Score based diffusion models

Langevin dynamics

Rossky, Doll & Friedman (1978)

Besag (1994)

Roberts & Tweedie (1996)

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathcal{N}(0, I_d)$$

…MCMC method that samples from $p(\mathbf{x})$

Langevin dynamics

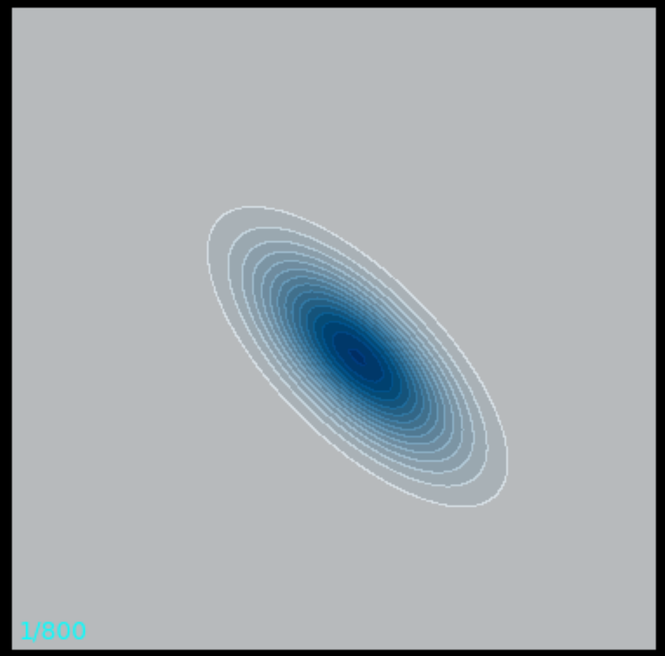
Rossky, Doll & Friedman (1978)

Besag (1994)

Roberts & Tweedie (1996)

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathcal{N}(0, I_d)$$

…MCMC method that samples from $p(\mathbf{x})$



Langevin dynamics

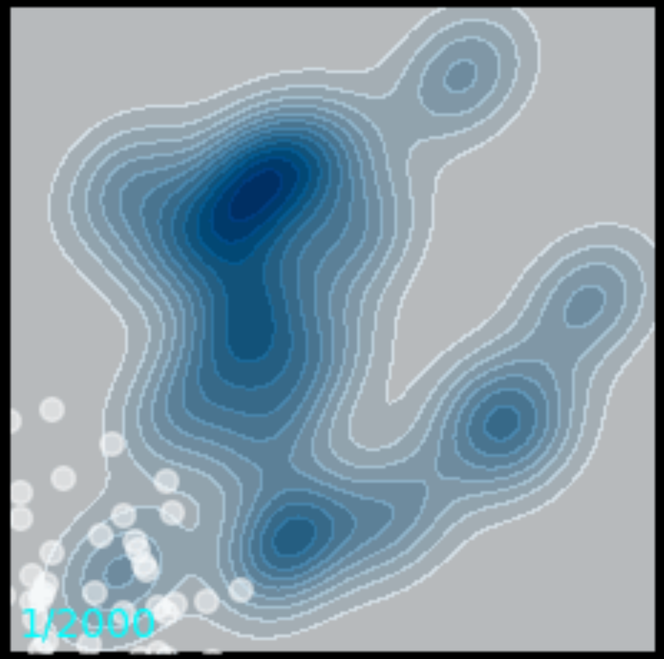
Rossky, Doll & Friedman (1978)

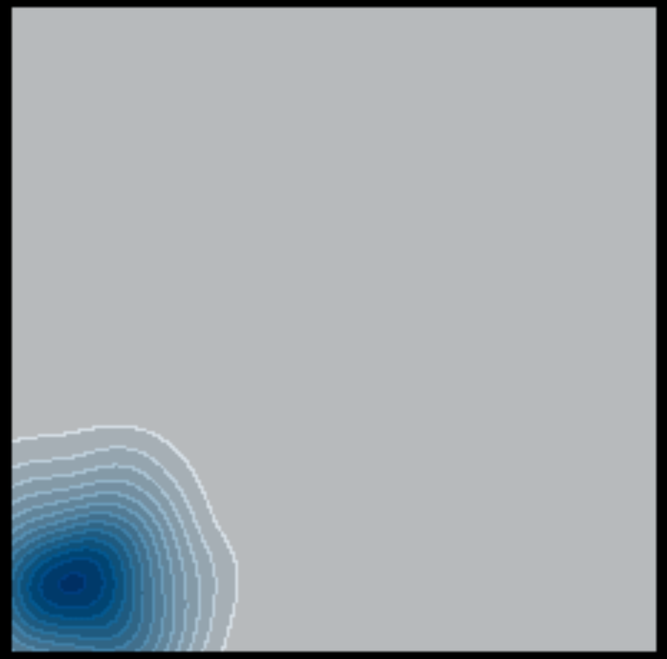
Besag (1994)

Roberts & Tweedie (1996)

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathcal{N}(0, I_d)$$

…MCMC method that samples from $p(\mathbf{x})$





Langevin dynamics requires score

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathcal{N}(0, I_d)$$

…MCMC method that samples from $p(\mathbf{x})$

Would like to learn this

$$\mathcal{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\, s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}) \,||_2^2]$$

Problem 1: don't have access to $p(\mathbf{x})$

Estimating the score

$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\,s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x})\,||_2^2]$$

Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\,s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x})\,||_2^2]$$

Estimating the score

$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\,s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x})\,||_2^2]$$

Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\,s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x})\,||_2^2]$$

$$p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}\,|\,\mathbf{0}, \sigma^2 1_d)$$

Estimating the score

$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\, s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}) \,||_2^2]$$

Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) \,||_2^2]$$

$$p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}\,|\,\mathbf{x}, \sigma^2 1_d)$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}}) \,||_2^2]$$

Estimating the score

$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\, s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}) \,||_2^2]$$

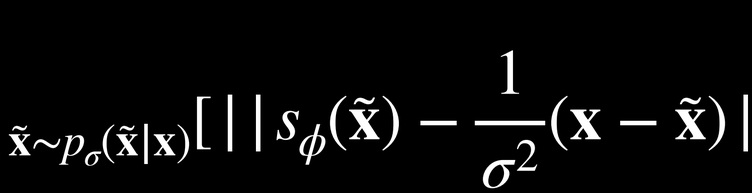
Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) \,||_2^2]$$

$$p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}\,|\,\mathbf{x}, \sigma^2 1_d)$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}}) \,||_2^2]$$



Estimating the score

$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\, s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}) \,||_2^2]$$

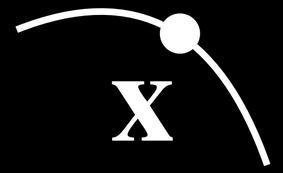
Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) \,||_2^2]$$

$$p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}\,|\,\mathbf{x}, \sigma^2 1_d)$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}}) \,||_2^2]$$



Estimating the score

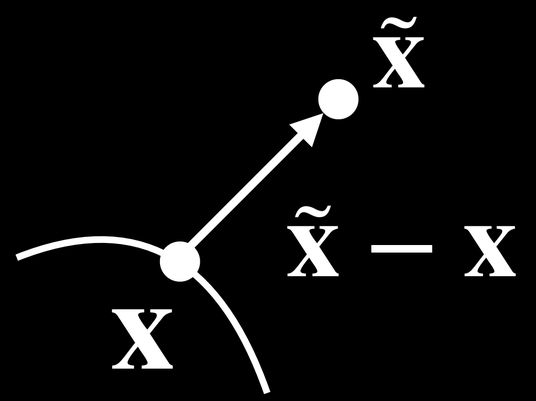
$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\, s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}) \,||_2^2]$$

Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) \,||_2^2]$$

$$p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}\,|\,\mathbf{x}, \sigma^2 1_d)$$



$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}}) \,||_2^2]$$

Estimating the score

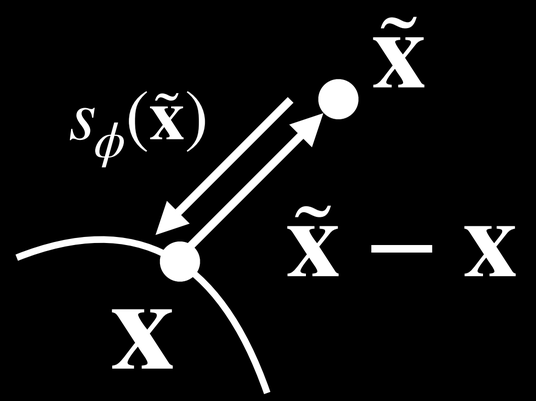
$$\mathscr{L} = \mathbb{E}_{p(\mathbf{x})}[\,||\, s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}) \,||_2^2]$$

Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) "*A Connection Between Score Matching and Denoising Autoencoders*"

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \nabla_{\tilde{x}} \log p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) \,||_2^2]$$

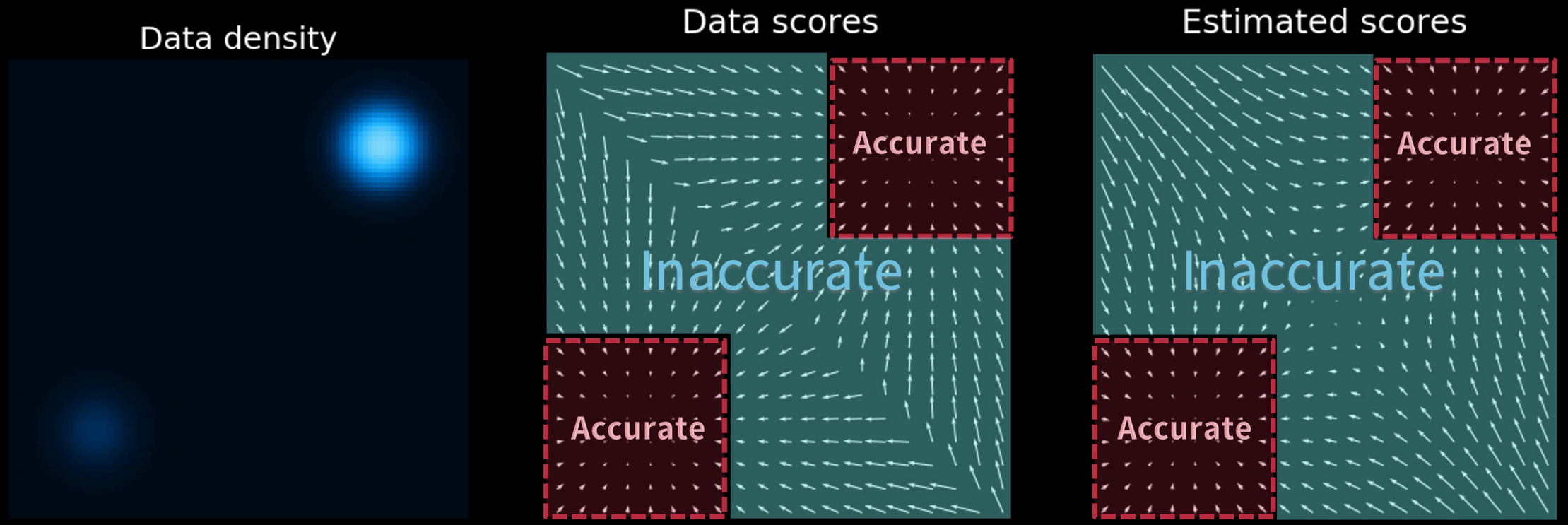
$$p_\sigma(\tilde{\mathbf{x}}\,|\,\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}\,|\,\mathbf{x}, \sigma^2 1_d)$$



$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}[\,||\, s_\phi(\tilde{\mathbf{x}}) - \frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}}) \,||_2^2]$$

Training score models

Problem 2: low coverage of data space → inaccurate score



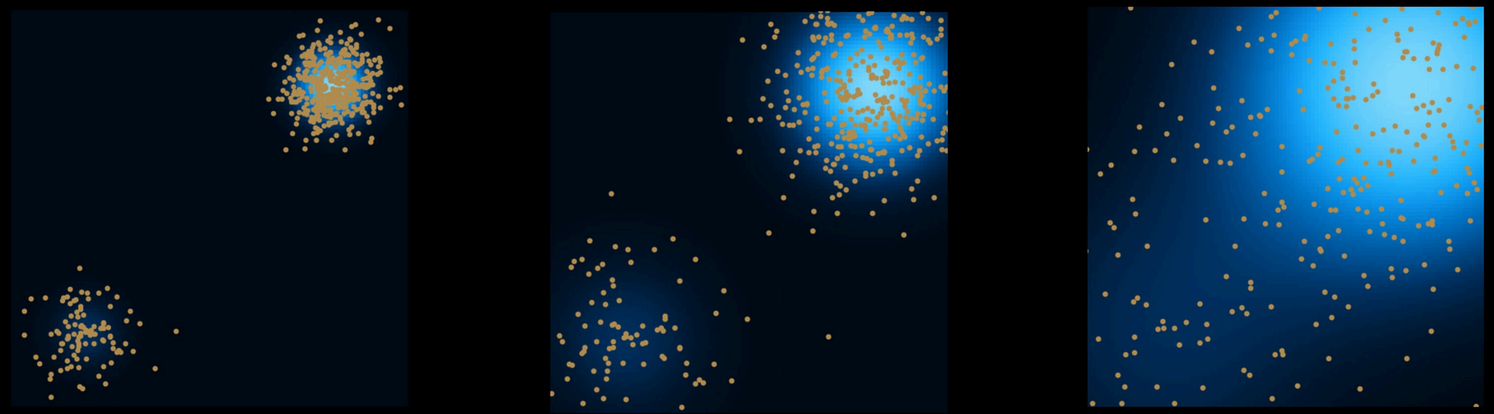
Training score models: multiple noise levels

Fix: add gradual noise

$$p_t(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{x} \mid \mu_t(x), \sigma_t(x))$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_t(\tilde{\mathbf{x}}|\mathbf{x}),t\sim\mathcal{U}(0,1)}[\,||\, s_\phi(\tilde{\mathbf{x}}, t) - \nabla_{\tilde{x}} \log p_t(\tilde{\mathbf{x}} \mid \mathbf{x})\,||_2^2]$$

$\sigma_1 < \sigma_2 < \sigma_3$



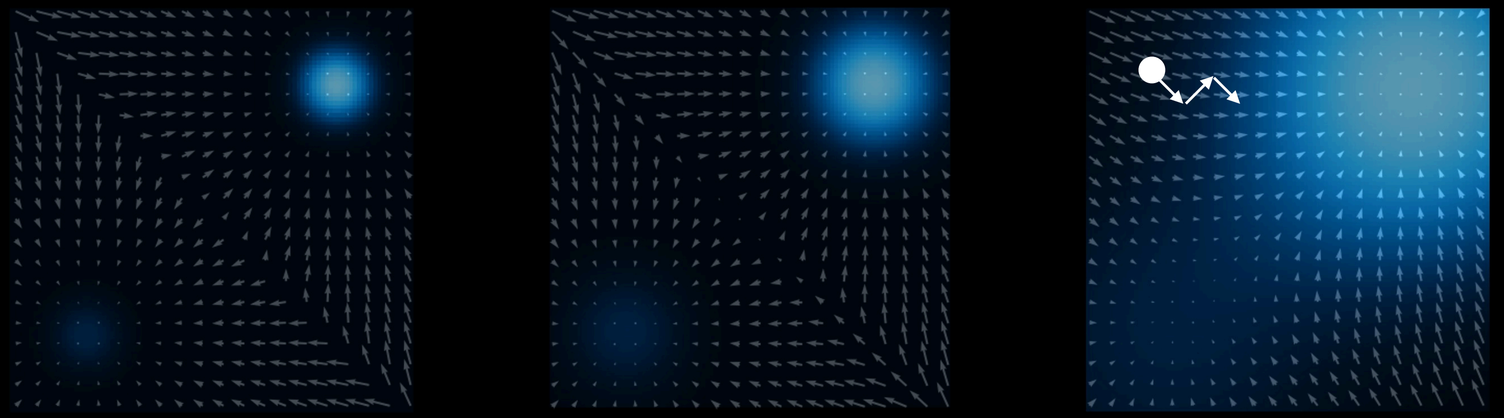
Training score models: multiple noise levels

Fix: add gradual noise

$$p_t(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{x} \mid \mu_t(x), \sigma_t(x))$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_t(\tilde{\mathbf{x}}|\mathbf{x}),t\sim\mathcal{U}(0,1)}[\,||\, s_\phi(\tilde{\mathbf{x}}, t) - \nabla_{\tilde{x}} \log p_t(\tilde{\mathbf{x}} \mid \mathbf{x})\,||_2^2]$$

$\sigma_1 < \sigma_2 < \sigma_3$



Sampling via **annealed** Langevin dynamics

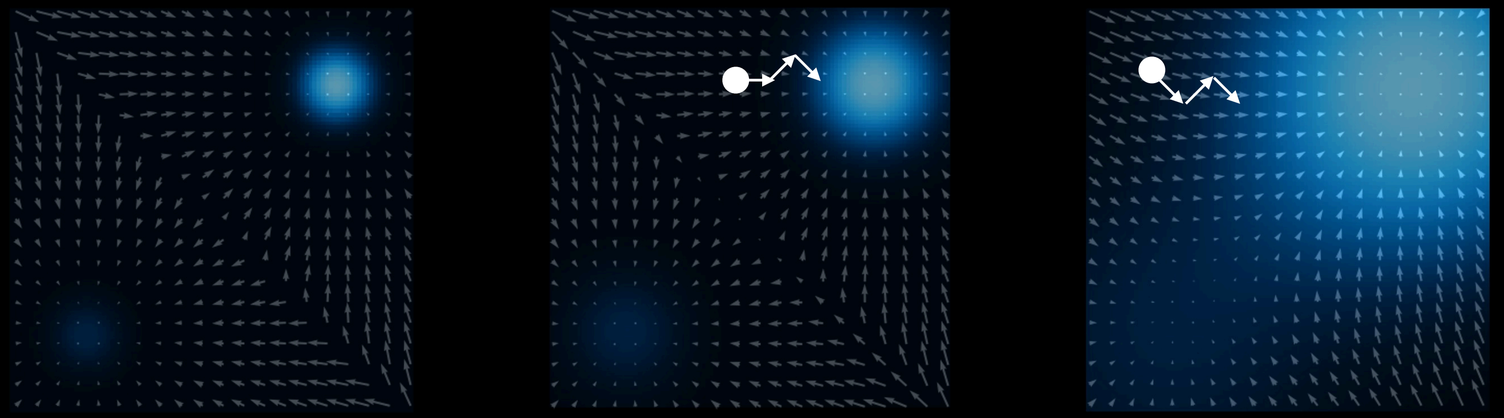
Training score models: multiple noise levels

Fix: add gradual noise

$$p_t(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{x} \mid \mu_t(x), \sigma_t(x))$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_t(\tilde{\mathbf{x}}|\mathbf{x}),t\sim\mathcal{U}(0,1)}[\,||\, s_\phi(\tilde{\mathbf{x}}, t) - \nabla_{\tilde{x}} \log p_t(\tilde{\mathbf{x}} \mid \mathbf{x})\,||_2^2]$$

$\sigma_1 < \sigma_2 < \sigma_3$



Sampling via **annealed** Langevin dynamics

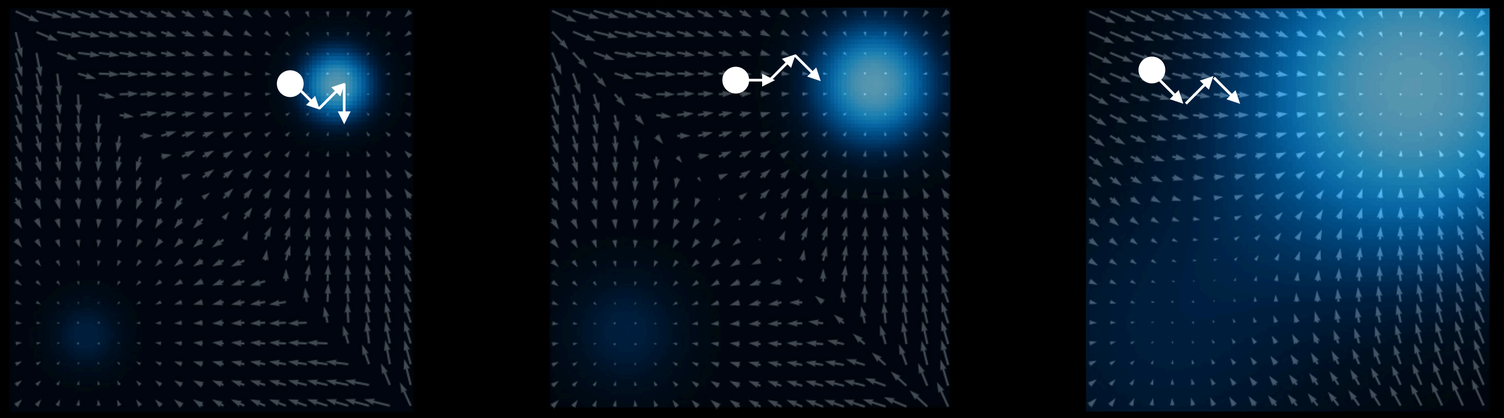
Training score models: multiple noise levels

Fix: add gradual noise

$$p_t(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{x} \mid \mu_t(x), \sigma_t(x))$$

$$\mathscr{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\, \tilde{\mathbf{x}}\sim p_t(\tilde{\mathbf{x}}|\mathbf{x}),t\sim\mathcal{U}(0,1)}[\,||\, s_\phi(\tilde{\mathbf{x}}, t) - \nabla_{\tilde{x}} \log p_t(\tilde{\mathbf{x}} \mid \mathbf{x})\,||_2^2]$$

$\sigma_1 < \sigma_2 < \sigma_3$



Sampling via **annealed** Langevin dynamics

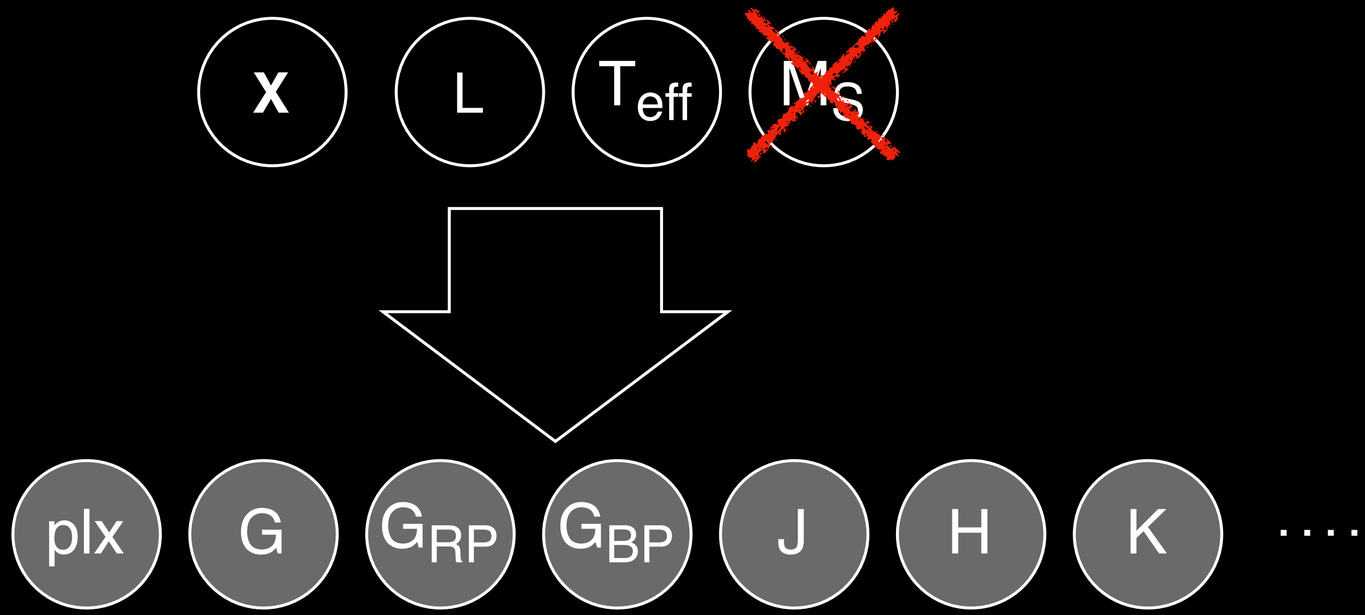
YSO models

Fwd model: YSO (Robitaille17+Richardson+24)

- Mass not input parameter
  - Unlink to evo. tracks

  ➡ logg = 4

  ➡ Small impact for $T_{\rm eff} \in [3,20]$ kK
- Low resolution (R~15)



Fwd model: YSO (Robitaille17+Richardson+24)

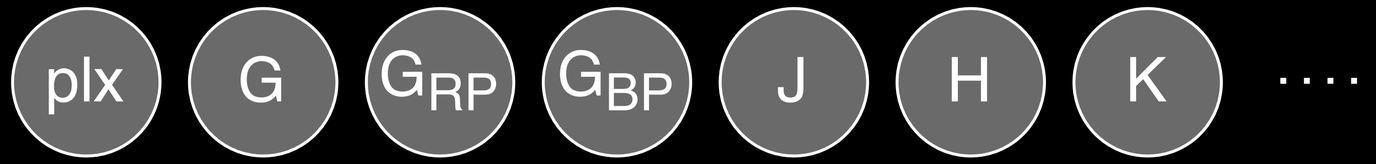
- Mass not input parameter
  - Unlink to evo. tracks

➡ logg = 4

➡ Small impact for $T_{\rm eff} \in [3,20]$ kK

- Low resolution (R~15)
- Introduces many additional parameters

Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	−2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	−2	−1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear



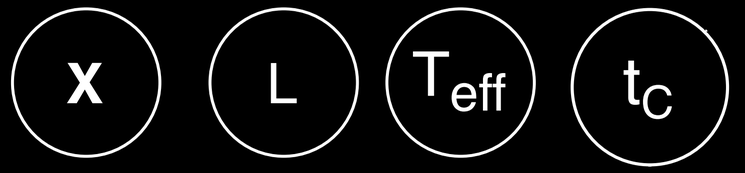
Fwd model: YSO (Robitaille17+Richardson+24)

- Mass not input parameter
  - Unlink to evo. tracks

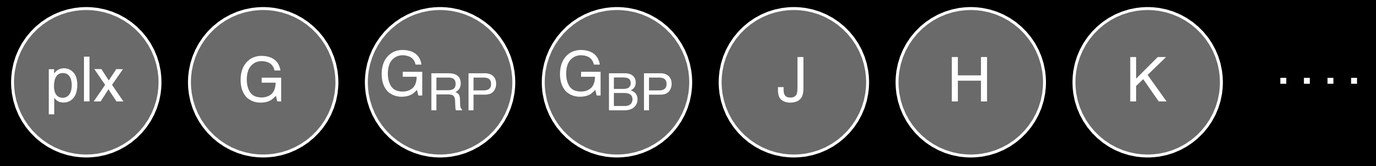
➡ logg = 4

➡ Small impact for $T_{\rm eff} \in [3,20]$ kK

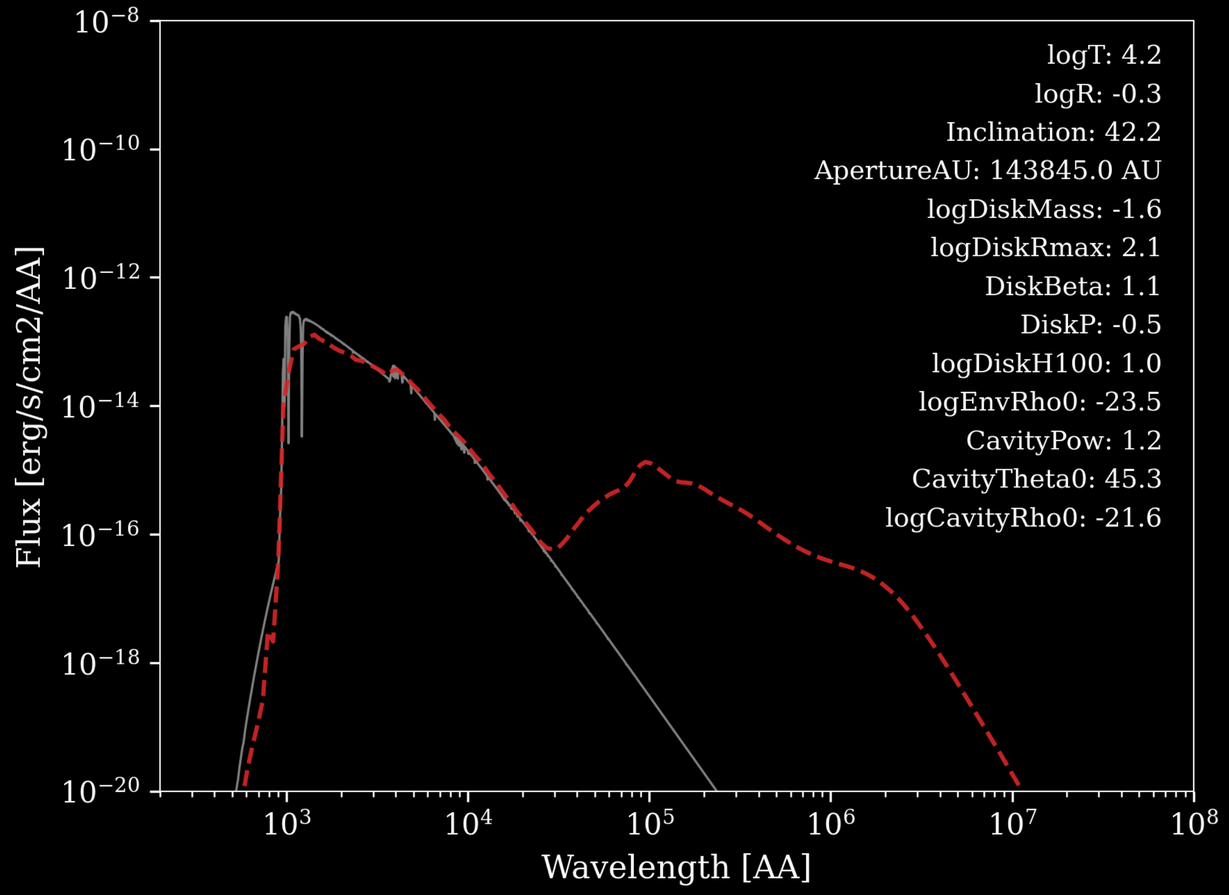
- Low resolution (R~15)
- Introduces many additional parameters



Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	−2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	−2	−1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

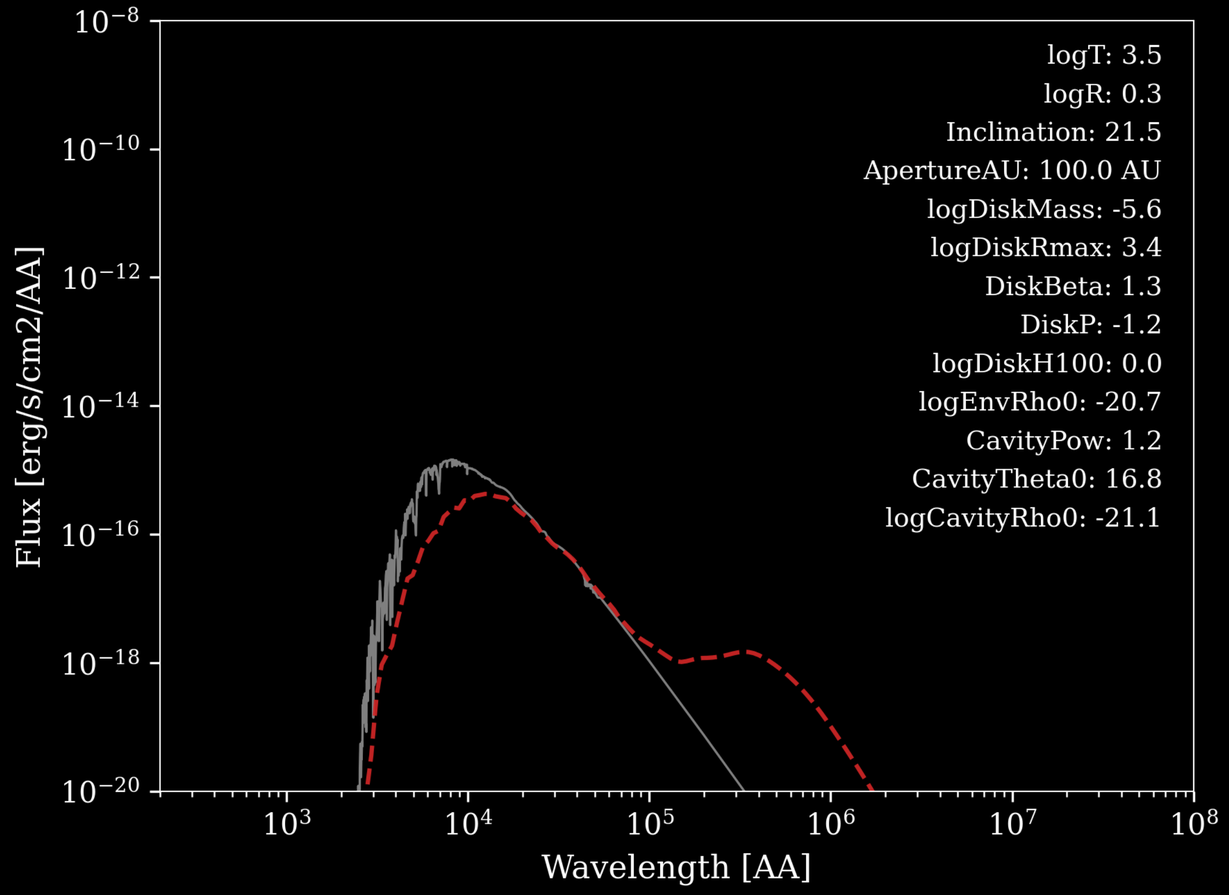


Fwd model: YSO (Robitaille17+Richardson+24)



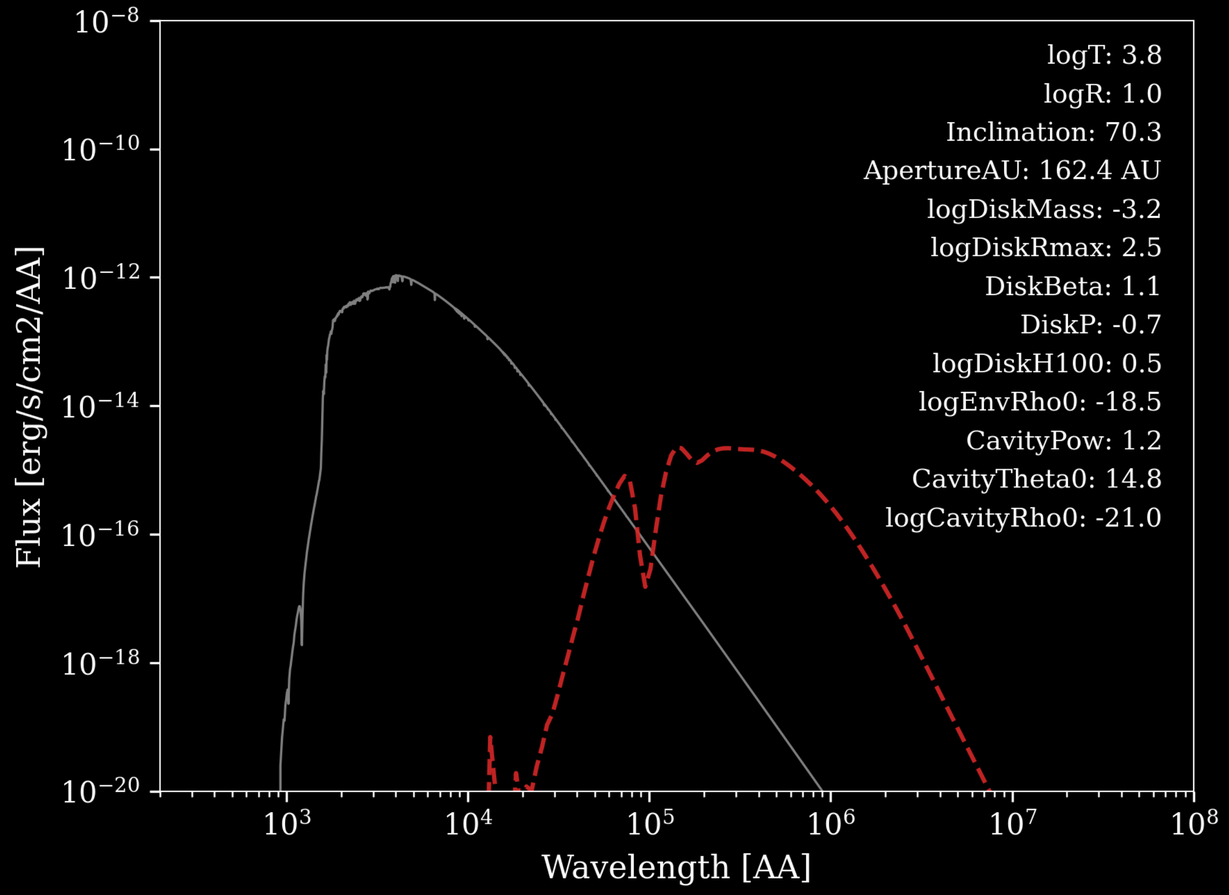
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

Fwd model: YSO (Robitaille17+Richardson+24)



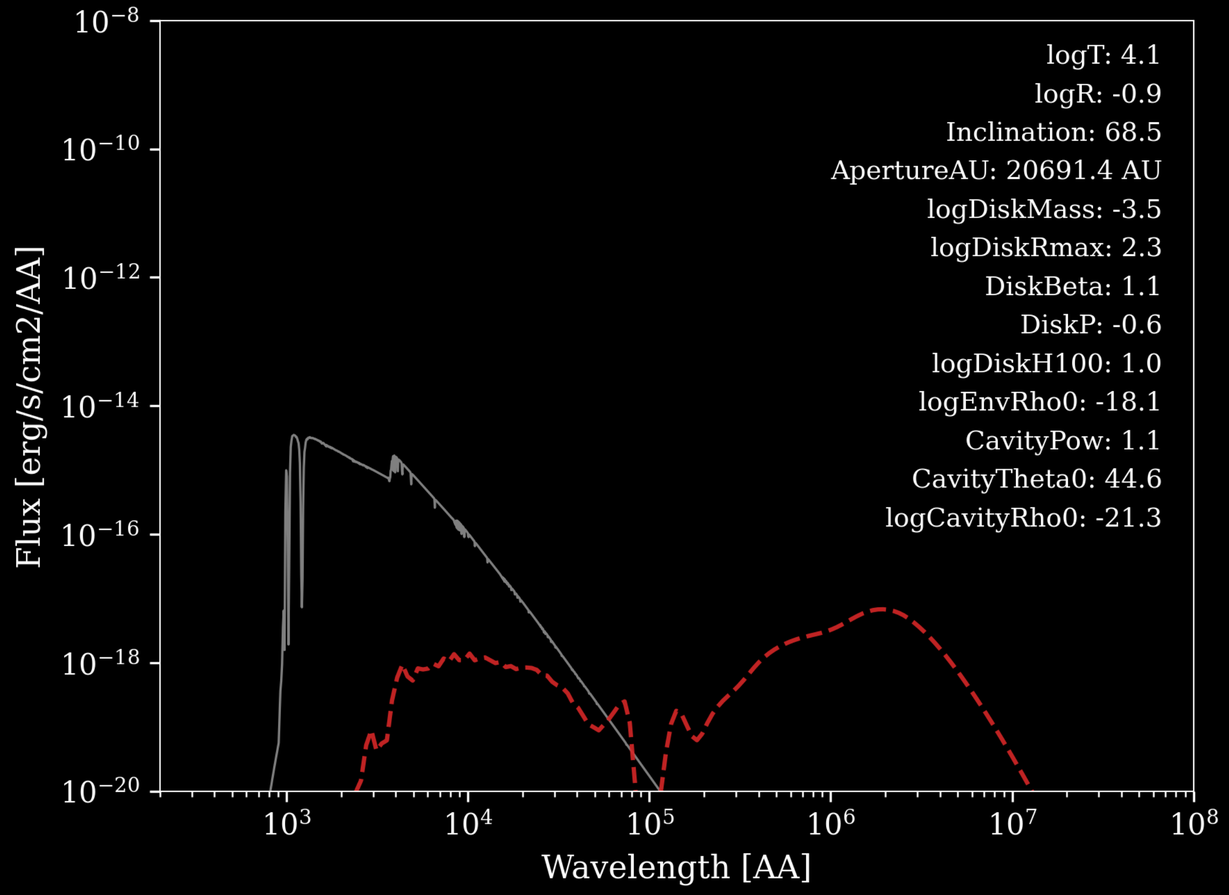
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{\rm 100AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

Fwd model: YSO (Robitaille17+Richardson+24)



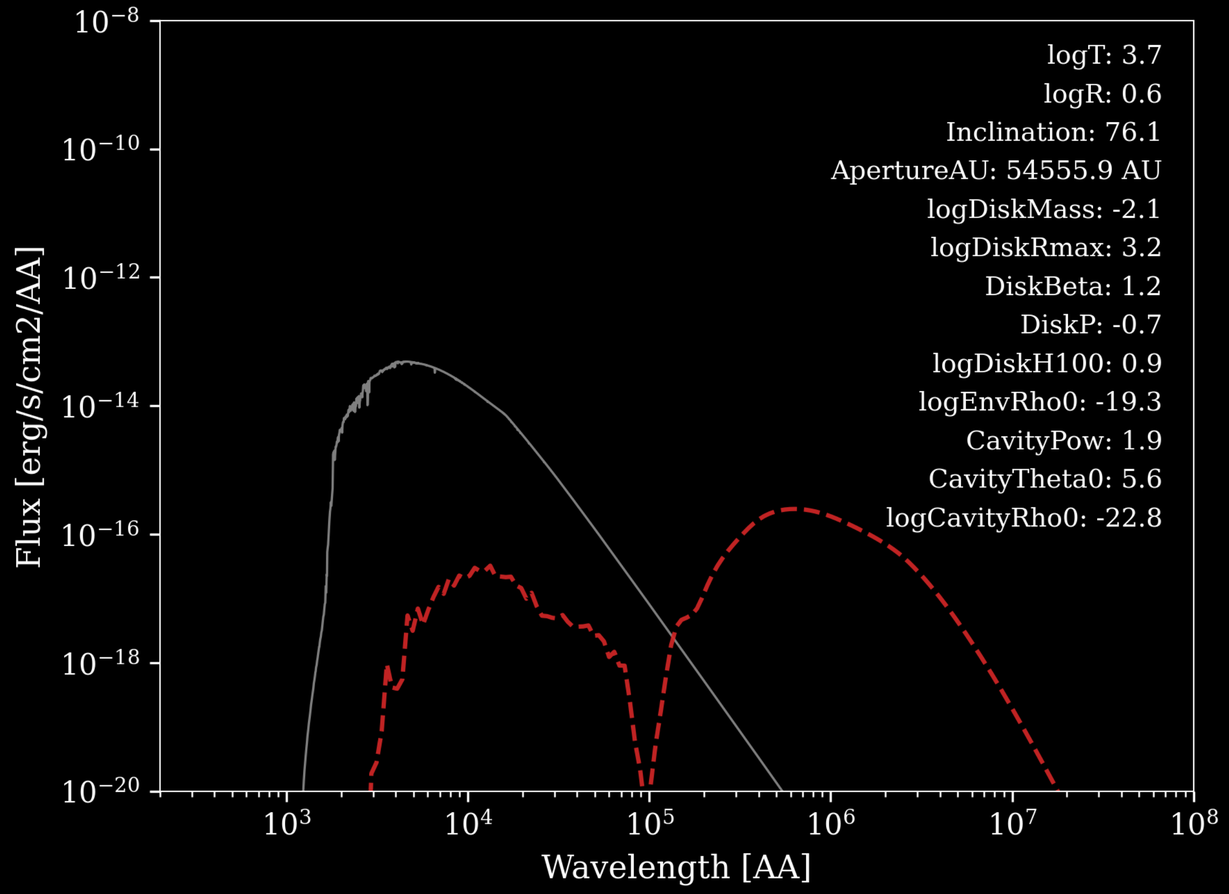
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

Fwd model: YSO (Robitaille17+Richardson+24)



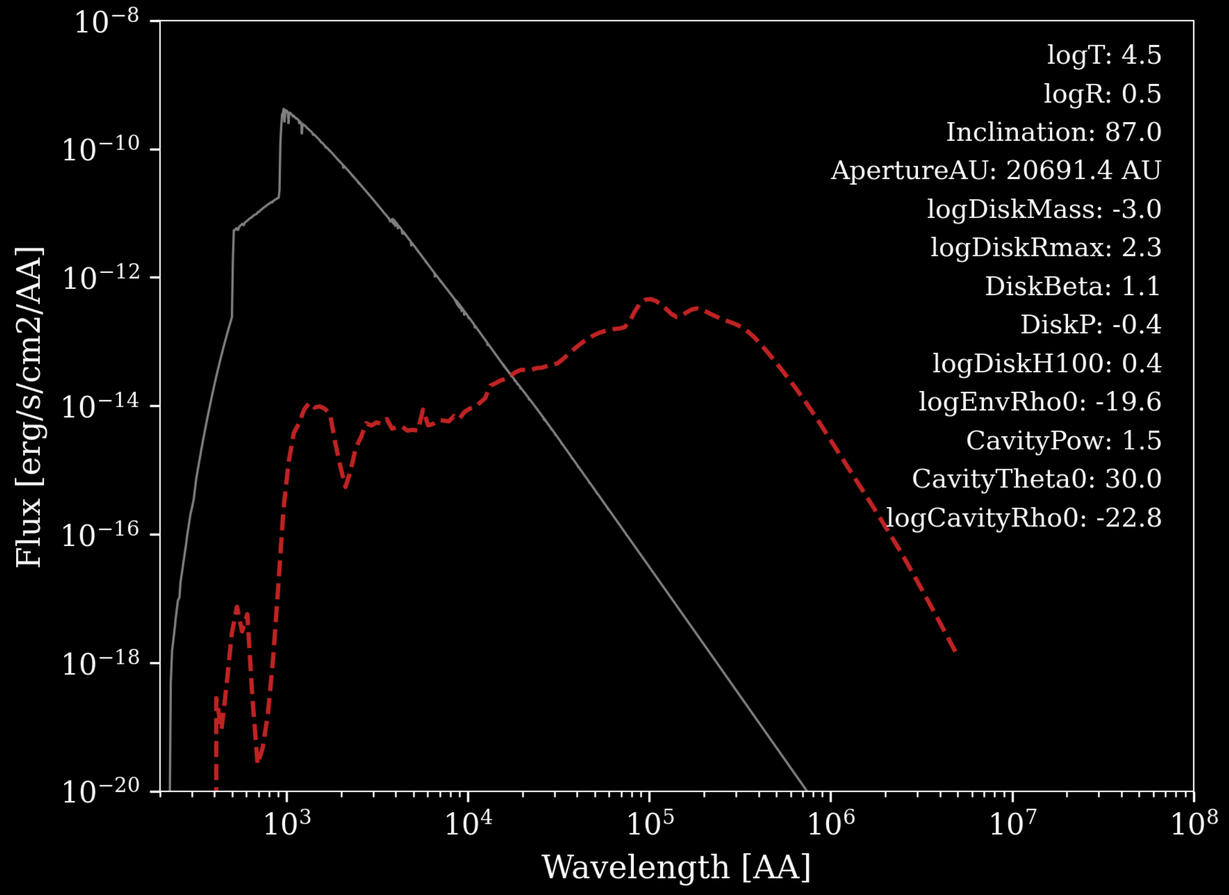
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

Fwd model: YSO (Robitaille17+Richardson+24)



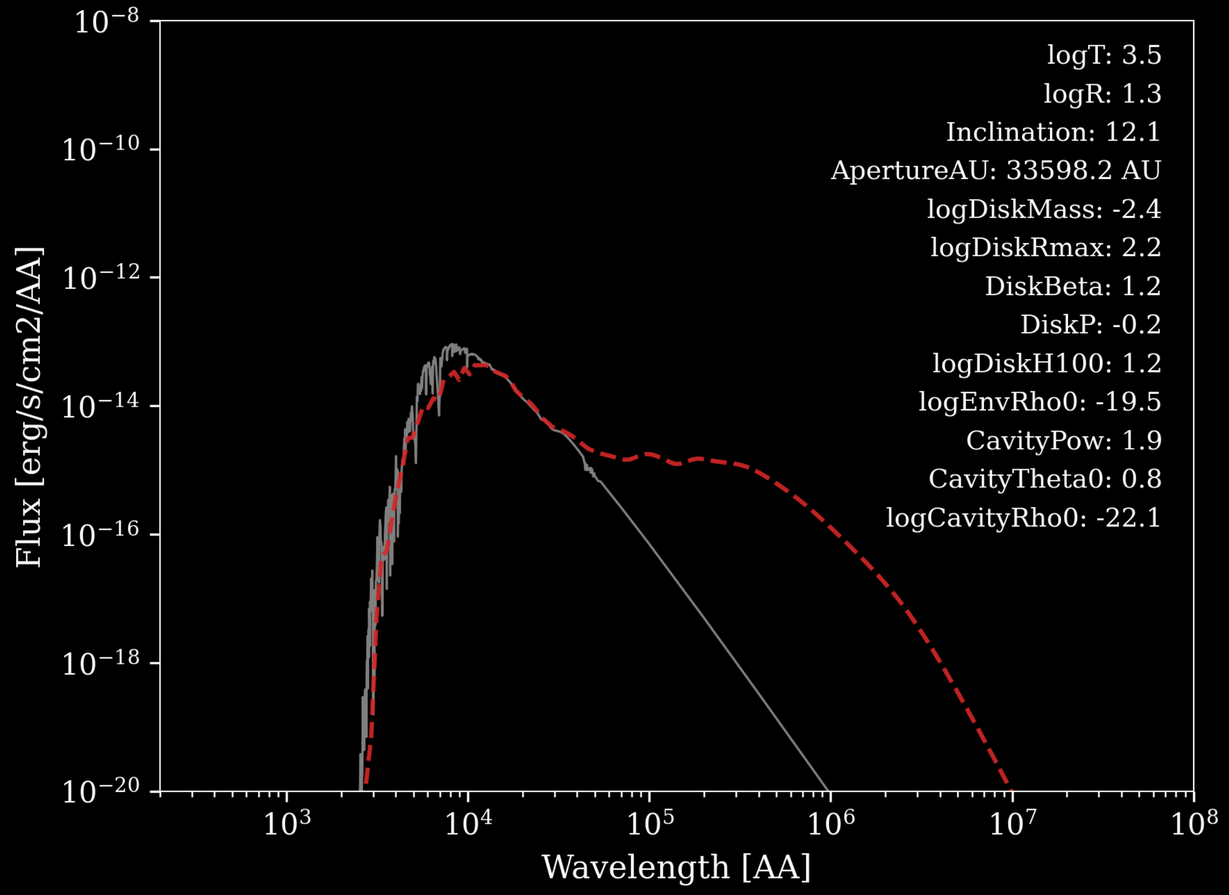
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

Fwd model: YSO (Robitaille17+Richardson+24)



Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{\rm 100AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

Fwd model: YSO (Robitaille17+Richardson+24)



Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_\star$	$0.1\,R_\odot$	$100\,R_\odot$	Log
Stellar temperature	$T_\star$	2000 K	30 000 K	Log
Disk mass [dust]	$M_{\rm disk}$	$10^{-8}\,M_\odot$	$0.1\,M_\odot$	Log
Disk inner radius	$R_{\rm min}^{\rm disk}$	$R_{\rm sub}$	$1000\,R_{\rm sub}$	Log
Disk outer radius	$R_{\rm max}^{\rm disk}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\rm AU}$	1 AU	20 AU	Log
Envelope density [dust]	$\rho_0^{\rm env}$	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	$R_{\rm c}$	50 AU	5000 AU	Log
Cavity density [dust]	$\rho_0^{\rm cav}$	10^{-23} g/cm^3	10^{-20} g/cm^3	Log
Cavity opening angle	θ_0	0°	60°	Linear
Cavity power	c	1	2	Linear

YSO (Robitaille17+Richardson+24)

Model set	Icon	Star	Disk	Envelope	Cavity	Ambient	Inner radius	Variables	Models
`s-s-i`		yes	…	…	…	…	…	2	10 000
`sp-s-i`		yes	passive	…	…	…	$R_{\rm sub}$	7	10 000
`sp-h-i`		yes	passive	…	…	…	variable	8	10 000
`s-smi`		yes	…	…	…	yes	$R_{\rm sub}$	2	10 000
`sp-smi`		yes	passive	…	…	yes	$R_{\rm sub}$	7	10 000
`sp-hmi`		yes	passive	…	…	yes	variable	8	10 000

sp-smi		yes	passive	…	…	yes	R_{sub}	7	10 000
sp-hmi		yes	passive	…	…	yes	variable	8	10 000
s-p-smi		yes	…	power-law	…	yes	R_{sub}	4	10 000
s-p-hmi		yes	…	power-law	…	yes	variable	5	10 000
s-pbsmi		yes	…	power-law	yes	yes	R_{sub}	7	10 000
s-pbhmi		yes	…	power-law	yes	yes	variable	8	10 000
s-u-smi		yes	…	Ulrich	…	yes	R_{sub}	4	10 000
s-u-hmi		yes	…	Ulrich	…	yes	variable	5	10 000
s-ubsmi		yes	…	Ulrich	yes	yes	R_{sub}	7	10 000
s-ubhmi		yes	…	Ulrich	yes	yes	variable	8	10 000

s-pbhmi		yes	…	power-law	yes	yes	variable	8	10 000
s-u-smi		yes	…	Ulrich	…	yes	R_{sub}	4	10 000
s-u-hmi		yes	…	Ulrich	…	yes	variable	5	10 000
s-ubsmi		yes	…	Ulrich	yes	yes	R_{sub}	7	10 000
s-ubhmi		yes	…	Ulrich	yes	yes	variable	8	10 000
spu-smi		yes	passive	Ulrich	…	yes	R_{sub}	8	10 000
spu-hmi		yes	passive	Ulrich	…	yes	variable	9	10 000
spubsmi		yes	passive	Ulrich	yes	yes	R_{sub}	11	40 000
spubhmi		yes	passive	Ulrich	yes	yes	variable	12	80 000

YSO (Robitaille17+Richardson+24)

`sp-h-i`		yes	passive	…	…	…	variable	8	10 000
`s-smi`		yes	…	…	…	yes	$R_{\rm sub}$	2	10 000
`sp-hmi`		yes	passive	…	…	yes	variable	8	10 000
`spubhmi`		yes	passive	Ulrich	yes	yes	variable	12	80 000

Model implementation

II. Learning to condition and marginalize

Conditional properties: 3D example

$\ell(\phi, M_C, \mathbf{x}) = (1 - M_C) \times (s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}))$ $\mathbf{x} = (x, y, z)$ $M_C = (0,0,1)$

Conditional properties: 3D example

$$\ell(\phi, M_C, \mathbf{x}) = (1 - M_C) \times (s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}))$$

$\mathbf{x} = (x, y, z)$ $M_C = (0,0,1)$

$$\log p(x, y, z) = \log p(x, y \mid z) + \log p(z)$$

Conditional properties: 3D example

$$\ell(\phi, M_C, \mathbf{x}) = (1 - M_C) \times (s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}))$$

$\mathbf{x} = (x, y, z)$ $M_C = (0,0,1)$

$$\log p(x, y, z) = \log p(x, y \mid z) + \log p(z)$$

$$\nabla \log p(x, y, z) = \nabla \log p(x, y \mid z) + \nabla \log p(z)$$

Conditional properties: 3D example

$$\ell(\phi, M_C, \mathbf{x}) = (1 - M_C) \times (s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}))$$

$\mathbf{x} = (x, y, z)$ $M_C = (0,0,1)$

$$\log p(x, y, z) = \log p(x, y \mid z) + \log p(z)$$

$$\nabla \log p(x, y, z) = \nabla \log p(x, y \mid z) + \nabla \log p(z)$$

$$\nabla_{x,y} \log p(x, y \mid z) + \nabla_{x,y} \log p(z) = \nabla_{x,y} \log p(x, y \mid z)$$

Conditional properties: 3D example

$\ell(\phi, M_C, \mathbf{x}) = (1 - M_C) \times (s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x}))$ $\quad \mathbf{x} = (x, y, z) \quad M_C = (0,0,1)$

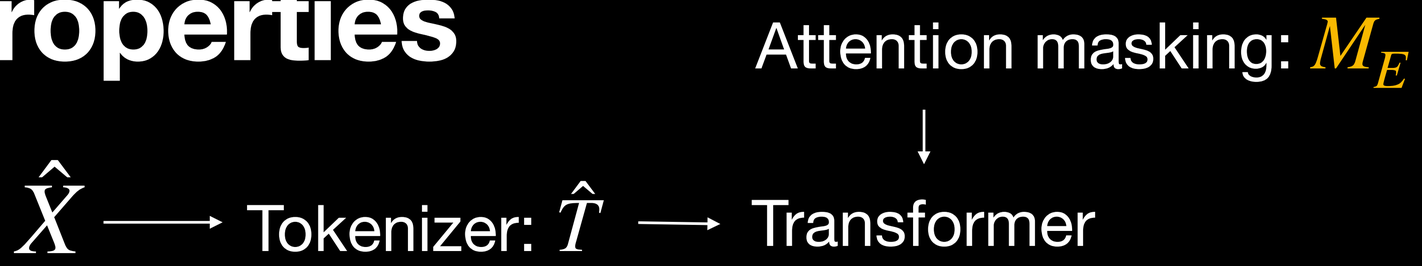
$\log p(x, y, z) = \log p(x, y \mid z) + \log p(z)$

$\nabla \log p(x, y, z) = \nabla \log p(x, y \mid z) + \nabla \log p(z)$

$\nabla_{x,y} \log p(x, y \mid z) + \nabla_{x,y} \log p(z) = \nabla_{x,y} \log p(x, y \mid z)$

$\nabla_z \log p(x, y \mid z) + \nabla_z \log p(z) \leftarrow$ Is set to 0 due to $(1 - M_C)$ in ℓ

Marginalization properties



Marginalization properties

Attention masking: M_E

$\downarrow$

$\hat{X} \longrightarrow$ Tokenizer: $\hat{T} \longrightarrow$ Transformer

$$\hat{T} \in (N, D_{\hat{X}}, E) \rightarrow S \propto QK^T \in (N, D_{\hat{X}}, D_{\hat{X}})$$

$$\text{softmax}(S + M_E)\, V\, P \in (N, D_{\hat{X}})$$

$$M_E = \begin{pmatrix} 0 & 0 & -\infty \\ 0 & 0 & -\infty \\ -\infty & -\infty & 0 \end{pmatrix} \longrightarrow s_\phi^{M_E}(\hat{x}^{M_C}) = \begin{pmatrix} f(x,y) \\ f(x,y) \\ f(z) \end{pmatrix}$$

Marginalization properties

Attention masking: M_E

$\downarrow$

$\hat{X} \longrightarrow$ Tokenizer: $\hat{T} \longrightarrow$ Transformer

$$\hat{T} \in (N, D_{\hat{X}}, E) \rightarrow S \propto QK^T \in (N, D_{\hat{X}}, D_{\hat{X}})$$

$$\text{softmax}(S + M_E)\, VP \in (N, D_{\hat{X}})$$

$$M_E = \begin{pmatrix} 0 & 0 & -\infty \\ 0 & 0 & -\infty \\ -\infty & -\infty & 0 \end{pmatrix} \longrightarrow s_\phi^{M_E}(\hat{x}^{M_C}) = \begin{pmatrix} f(x,y) \\ f(x,y) \\ f(z) \end{pmatrix}$$

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) = (1 - M_C) \cdot \left(s_\phi^{M_E}(\hat{\mathbf{x}}_t^{M_C}, t) - \nabla_{\hat{\mathbf{x}}_t} \log p_t(\hat{\mathbf{x}}_t | \hat{\mathbf{x}}_0) \right) \propto \begin{pmatrix} f(x,y) - \Delta x \\ f(x,y) - \Delta y \\ f(z) - \Delta z \end{pmatrix}$$

$$\propto \hat{x}_0 - \hat{x}_t$$

$$\hat{x}_t \sim p_t(\hat{x}_t \,|\, \hat{x}_0) = \mathcal{N}(\mu_t(\hat{x}_0), \sigma_t(\hat{x}_0))$$

Marginalization properties

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) = (1 - M_C) \cdot \left(s_\phi^{M_E}(\hat{\mathbf{x}}_t^{M_C}, t) - \nabla_{\hat{\mathbf{x}}_t} \log p_t(\hat{\mathbf{x}}_t | \hat{\mathbf{x}}_0) \right) \propto \begin{pmatrix} f(x,y) - \Delta x \\ f(x,y) - \Delta y \\ f(z) - \Delta z \end{pmatrix}$$

$$\propto \hat{x}_0 - \hat{x}_t$$

$$\mathscr{L} \propto (f(x,y) - \Delta x)^2 + (f(x,y) - \Delta y)^2 + (f(z) - \Delta z)^2$$

Marginalization properties

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) = (1 - M_C) \cdot \left(s_\phi^{M_E}(\hat{\mathbf{x}}_t^{M_C}, t) - \nabla_{\hat{\mathbf{x}}_t} \log p_t(\hat{\mathbf{x}}_t | \hat{\mathbf{x}}_0) \right) \propto \begin{pmatrix} f(x,y) - \Delta x \\ f(x,y) - \Delta y \\ f(z) - \Delta z \end{pmatrix}$$

$$\propto \hat{x}_0 - \hat{x}_t$$

$$\mathscr{L} \propto (f(x,y) - \Delta x)^2 + (f(x,y) - \Delta y)^2 + (f(z) - \Delta z)^2 =$$

$$= \nabla \log p(x,y) + \nabla \log p(z) = \nabla \log p(x,y)p(z)$$