

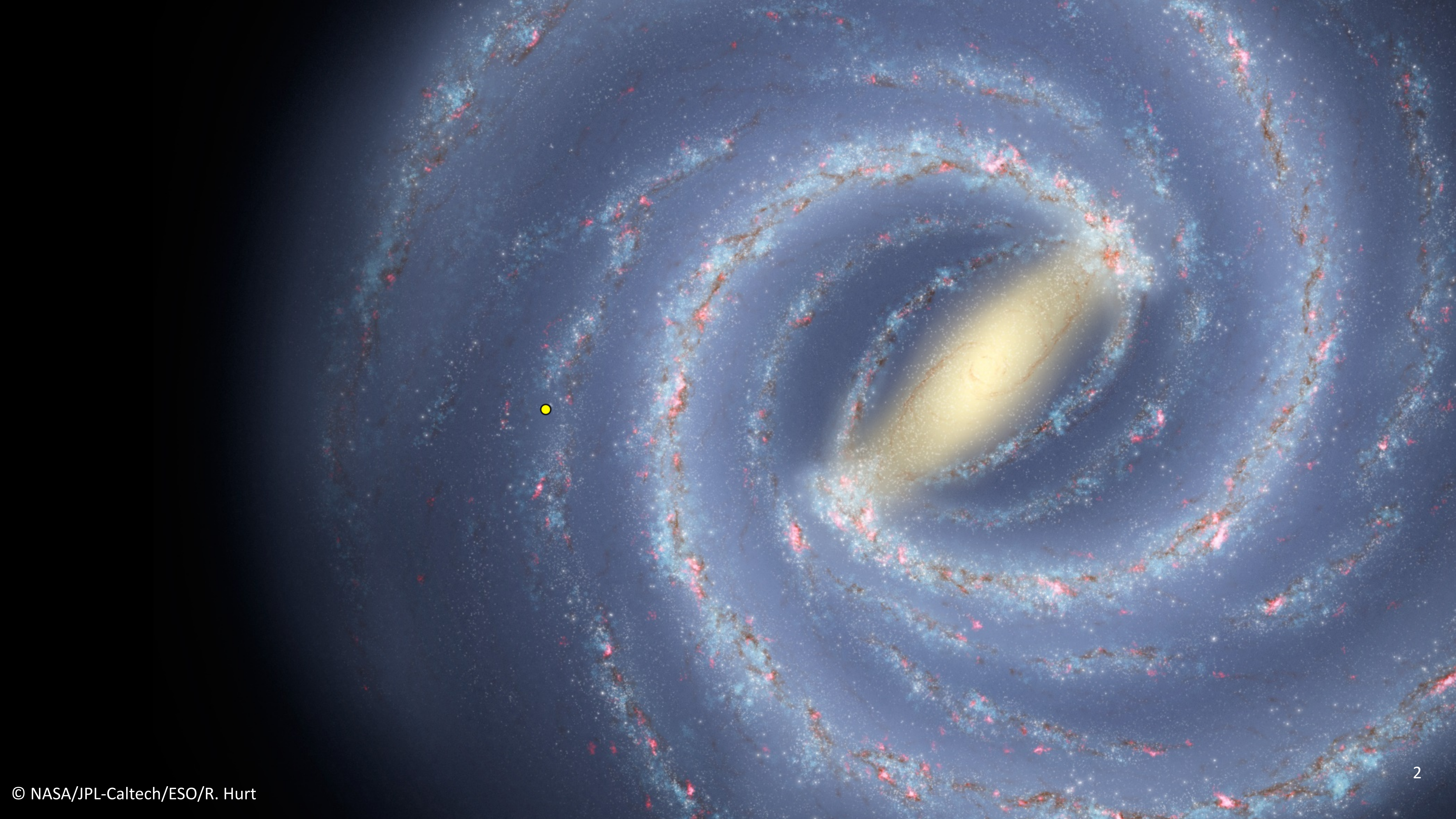
Learning with Gaps

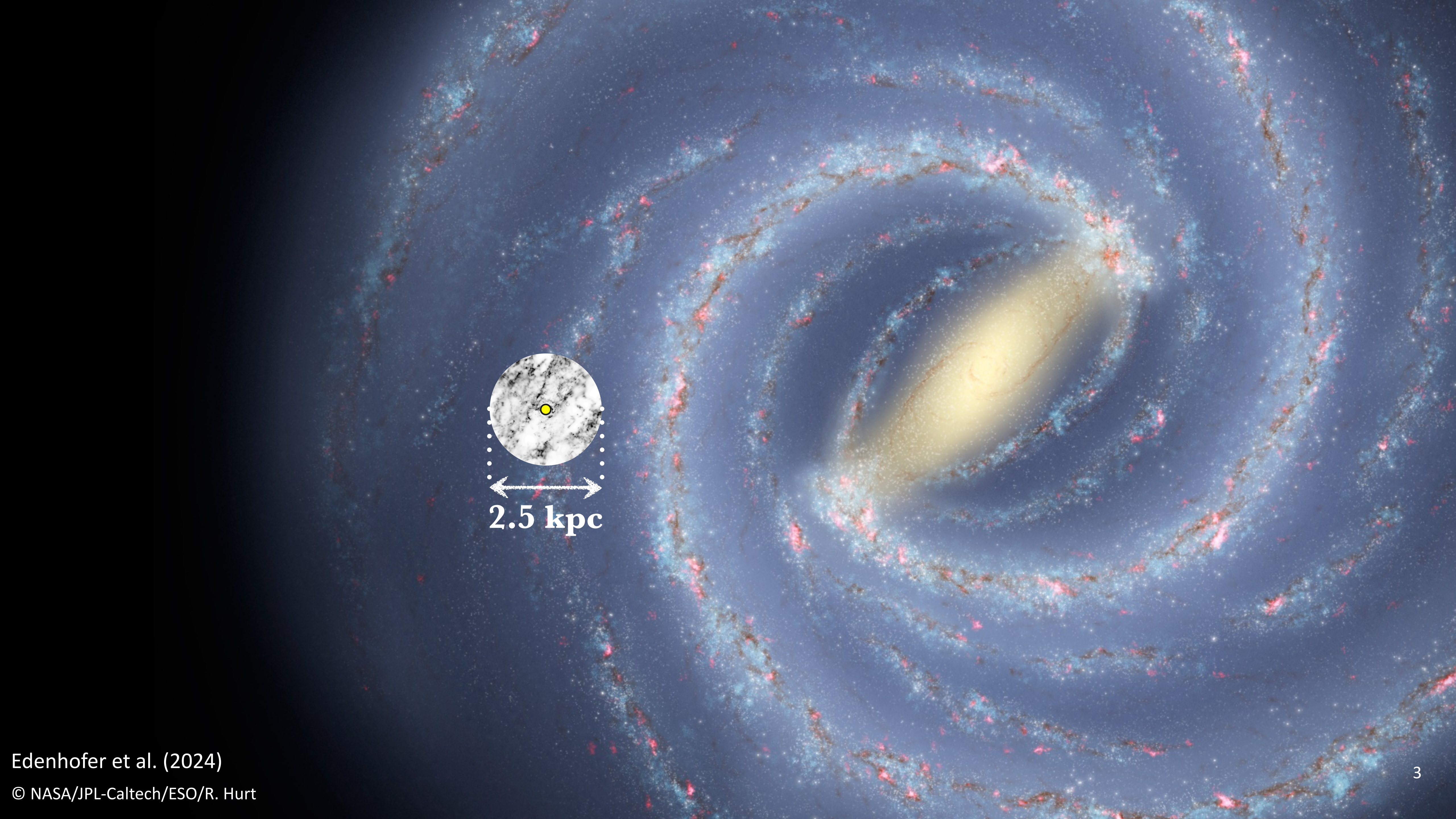
A Domain-Adaptive SBI Framework for Mapping Young Stars from Incomplete, Multi-Survey Data

Sebastian Ratzenböck

In collaboration with

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

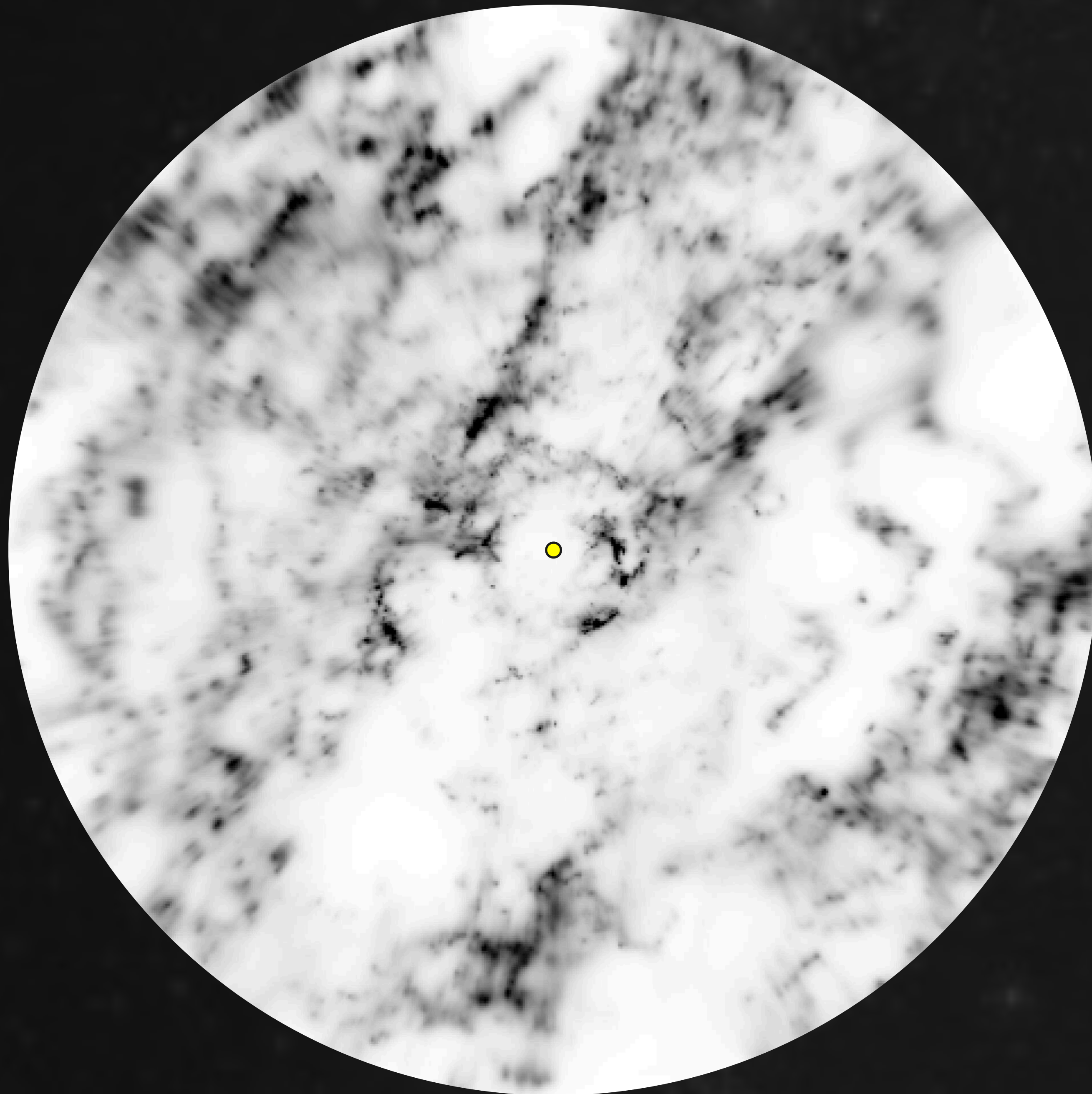
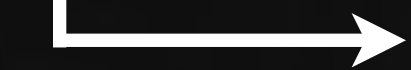




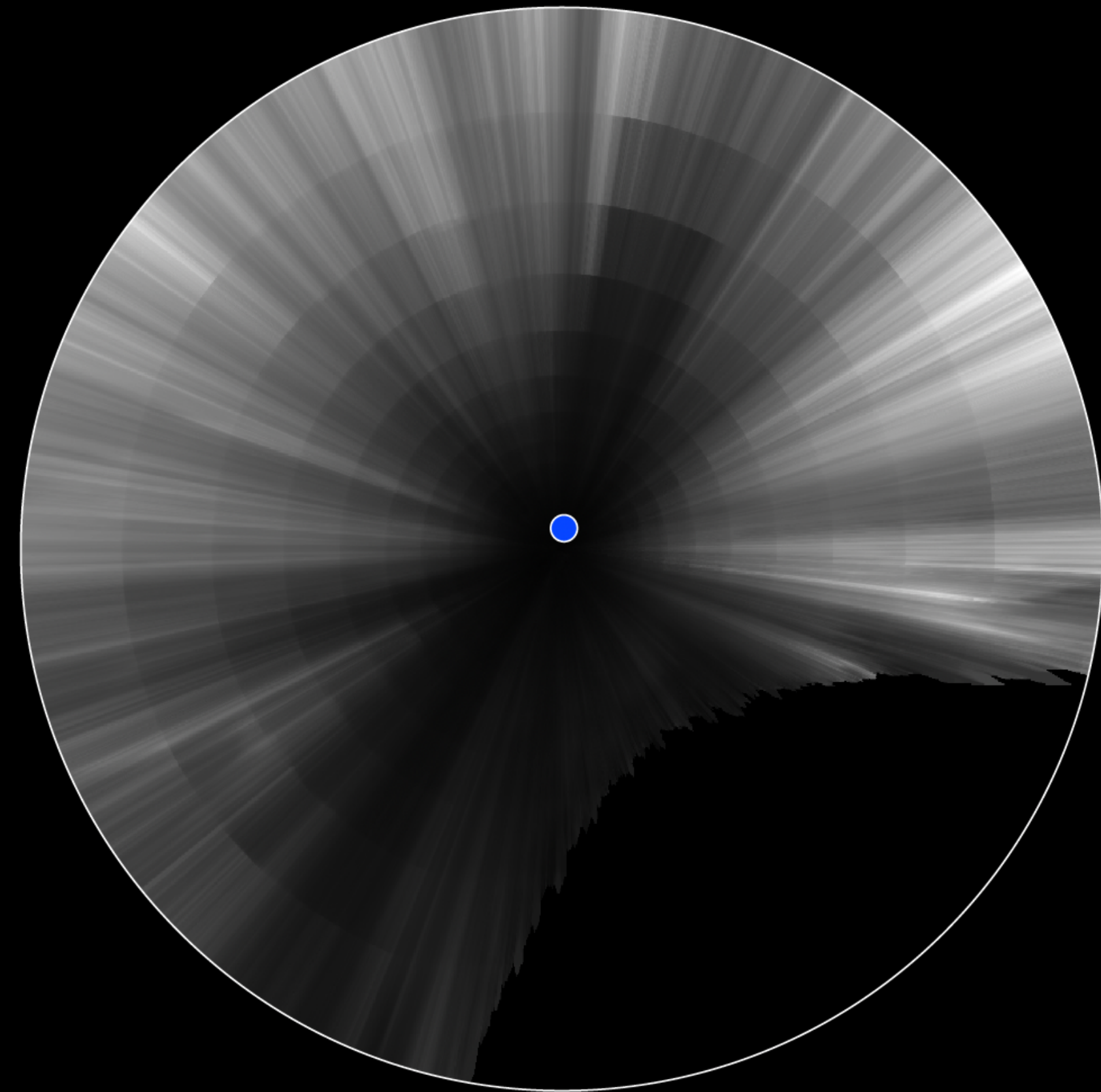
Galactic rotation



Galactic center

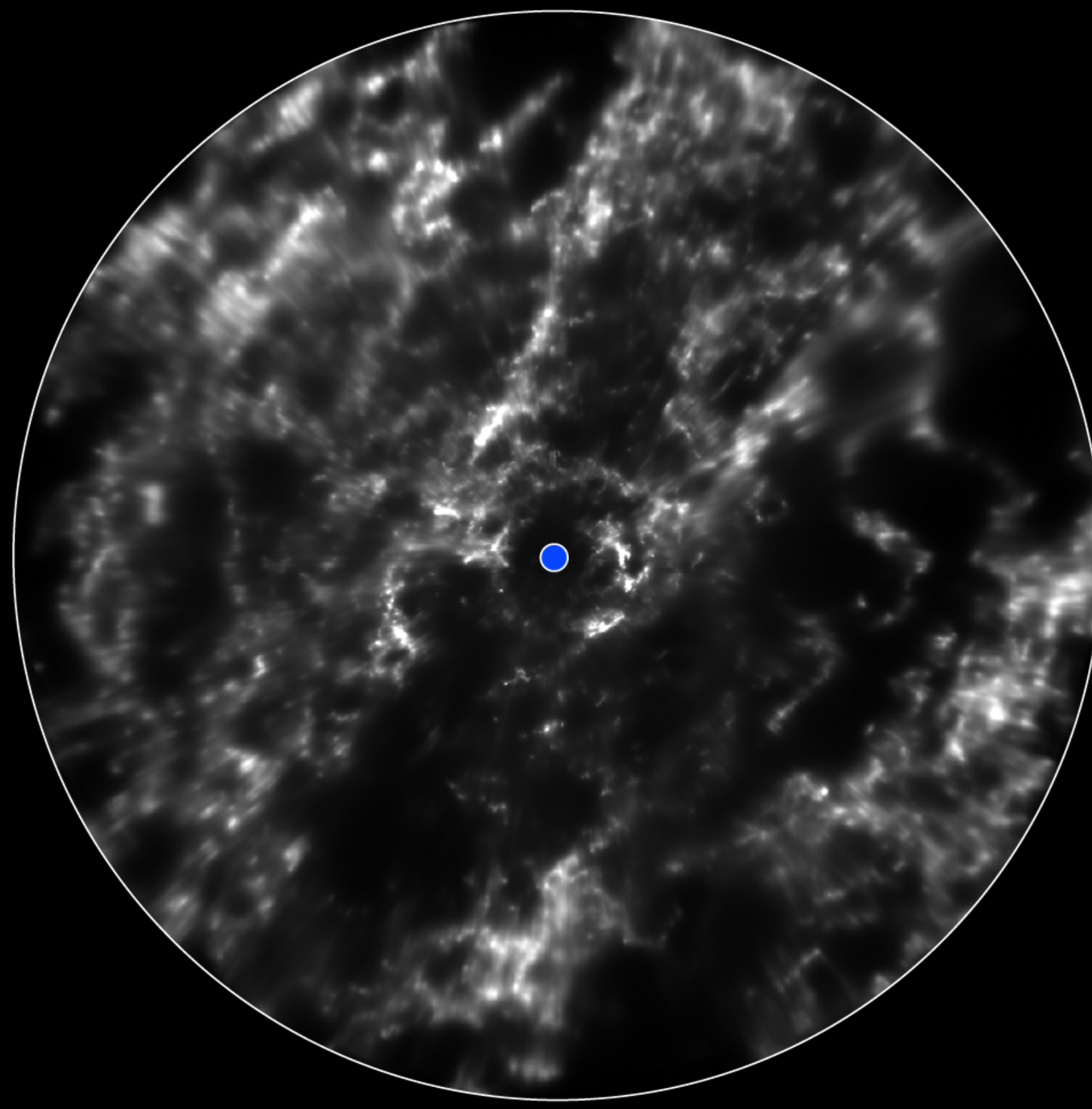


Solar Neighborhood **BEFORE**



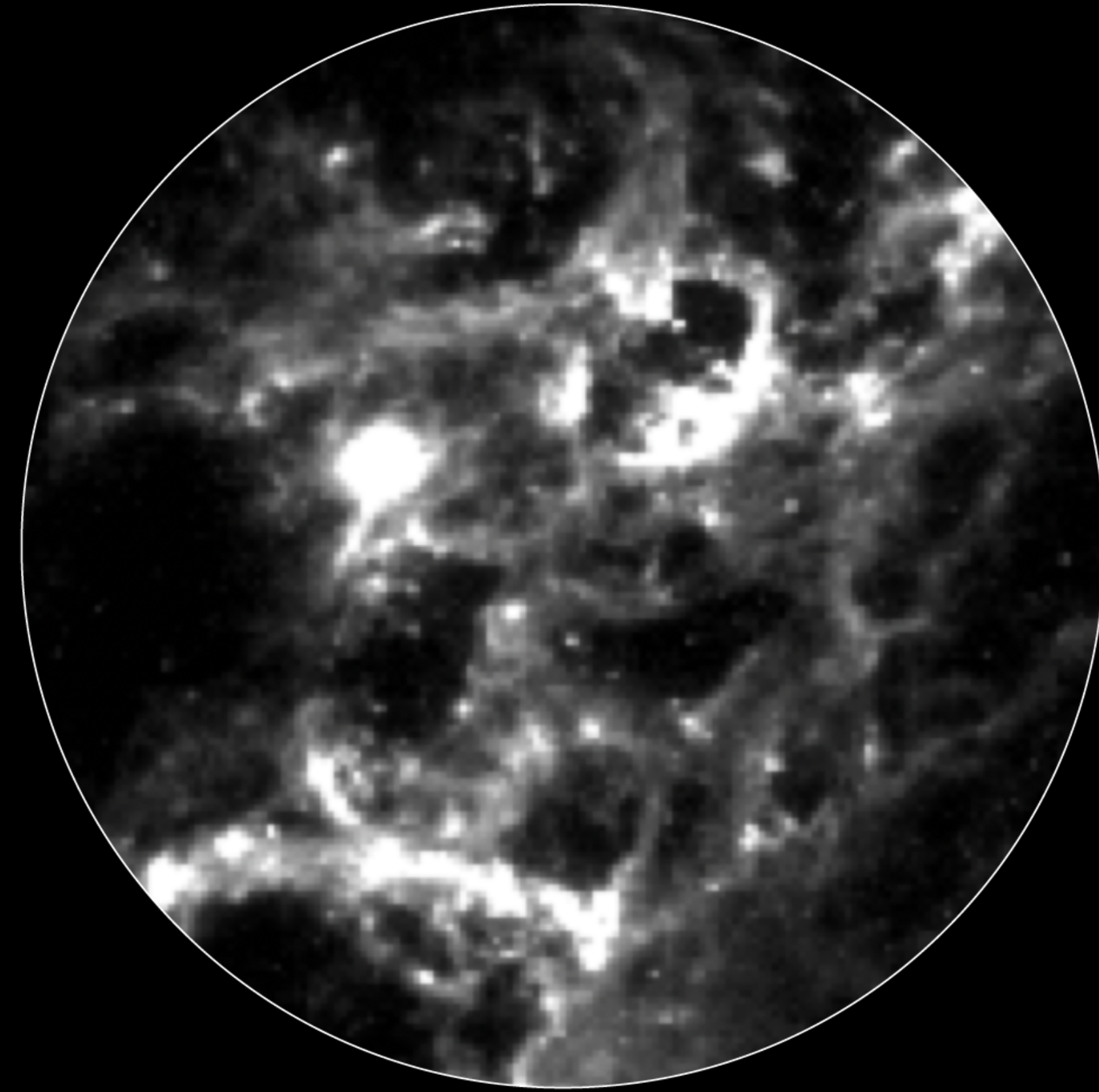
Green et al. 2018

Solar Neighborhood **NOW**



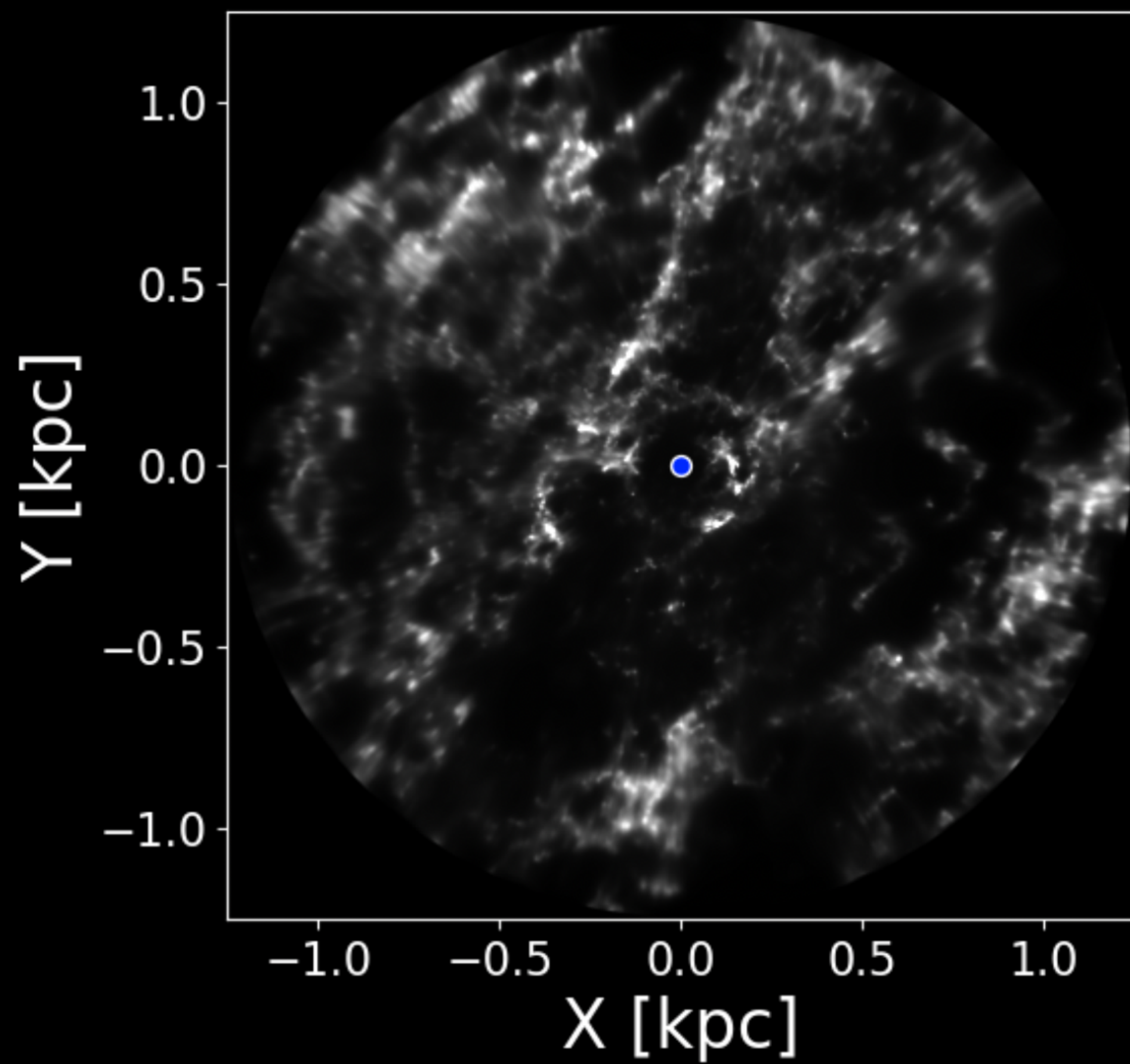
Edenhofer et al. (2024)

External Galaxy Zoom-in **NOW**

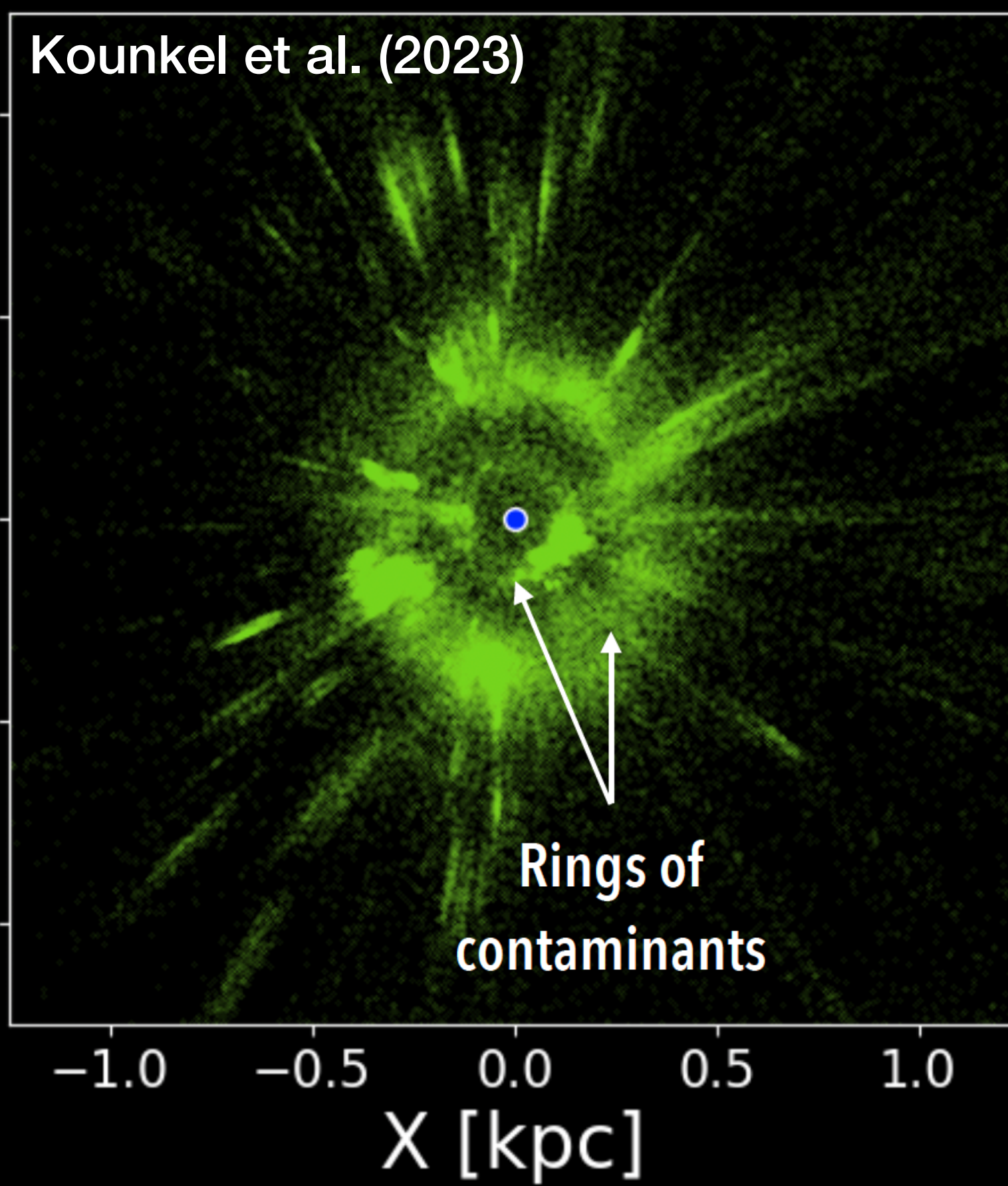


JWST $7\mu\text{m}$ zoom in

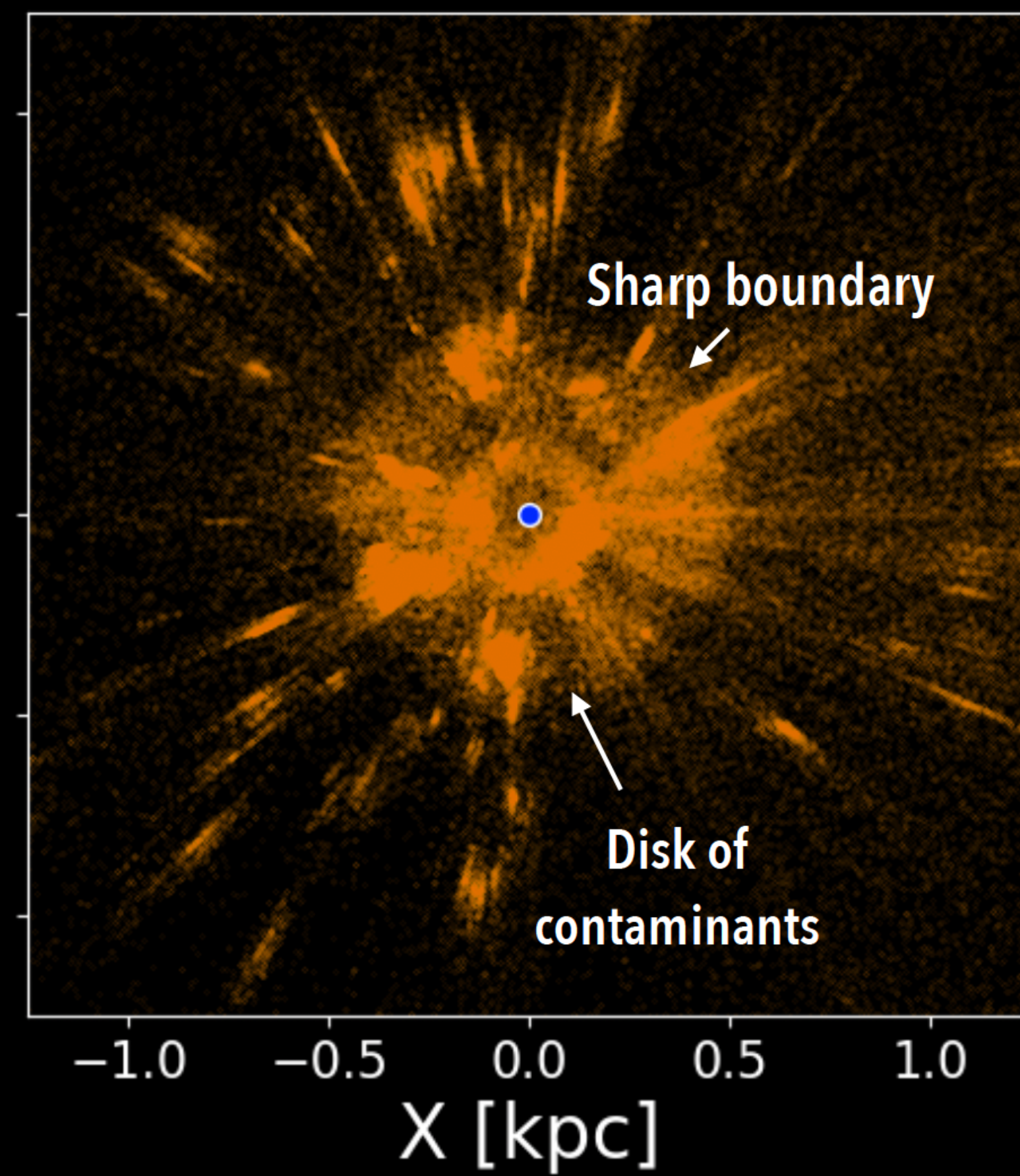
3D Dust



Young Star (SOTA Machine Learning)



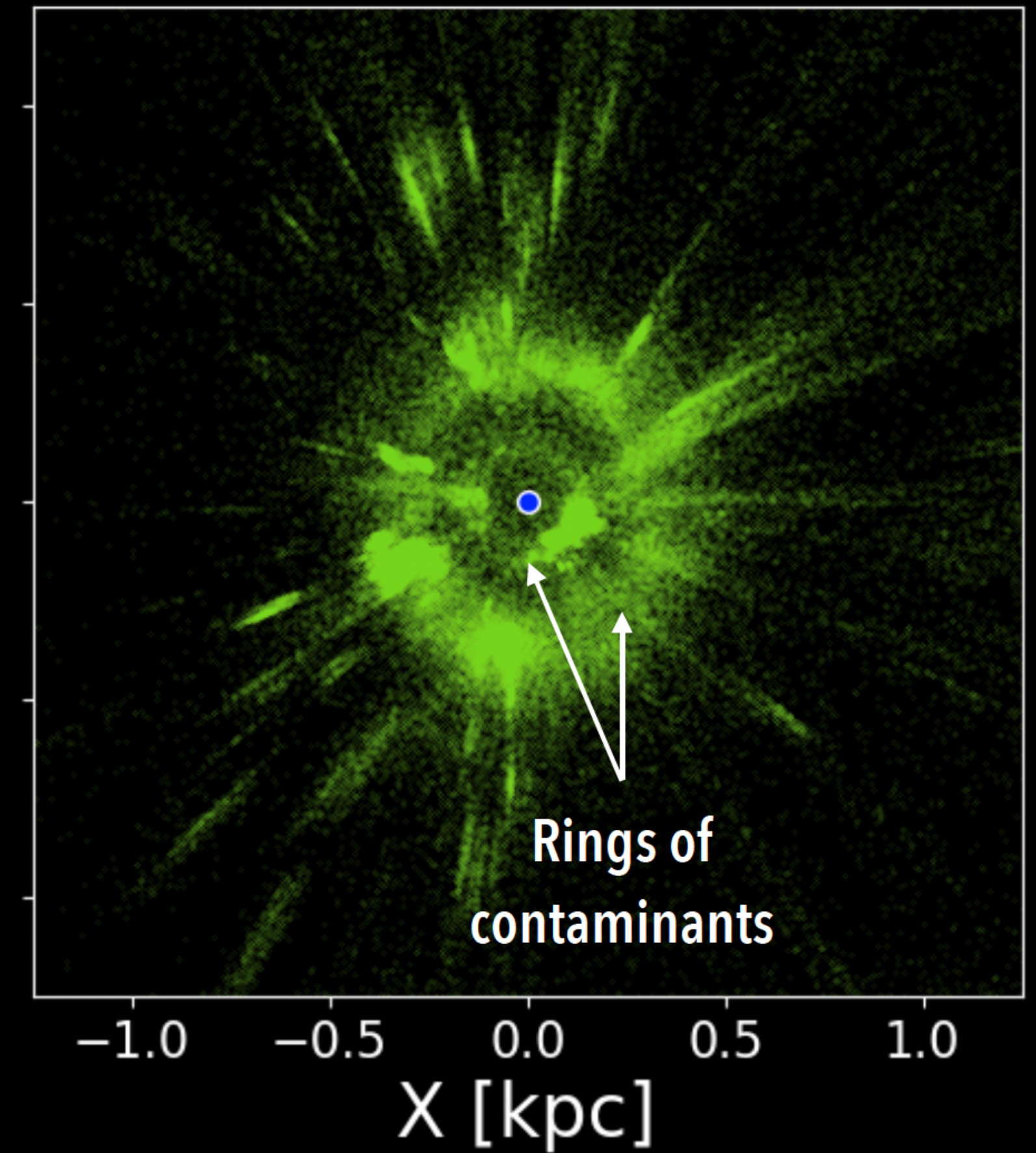
Young Star (Literature Compilation)



Understanding Galactic baryon cycle

- YSOs connect *cloud* $\leftrightarrow$ *stars* $\leftrightarrow$ *feedback*

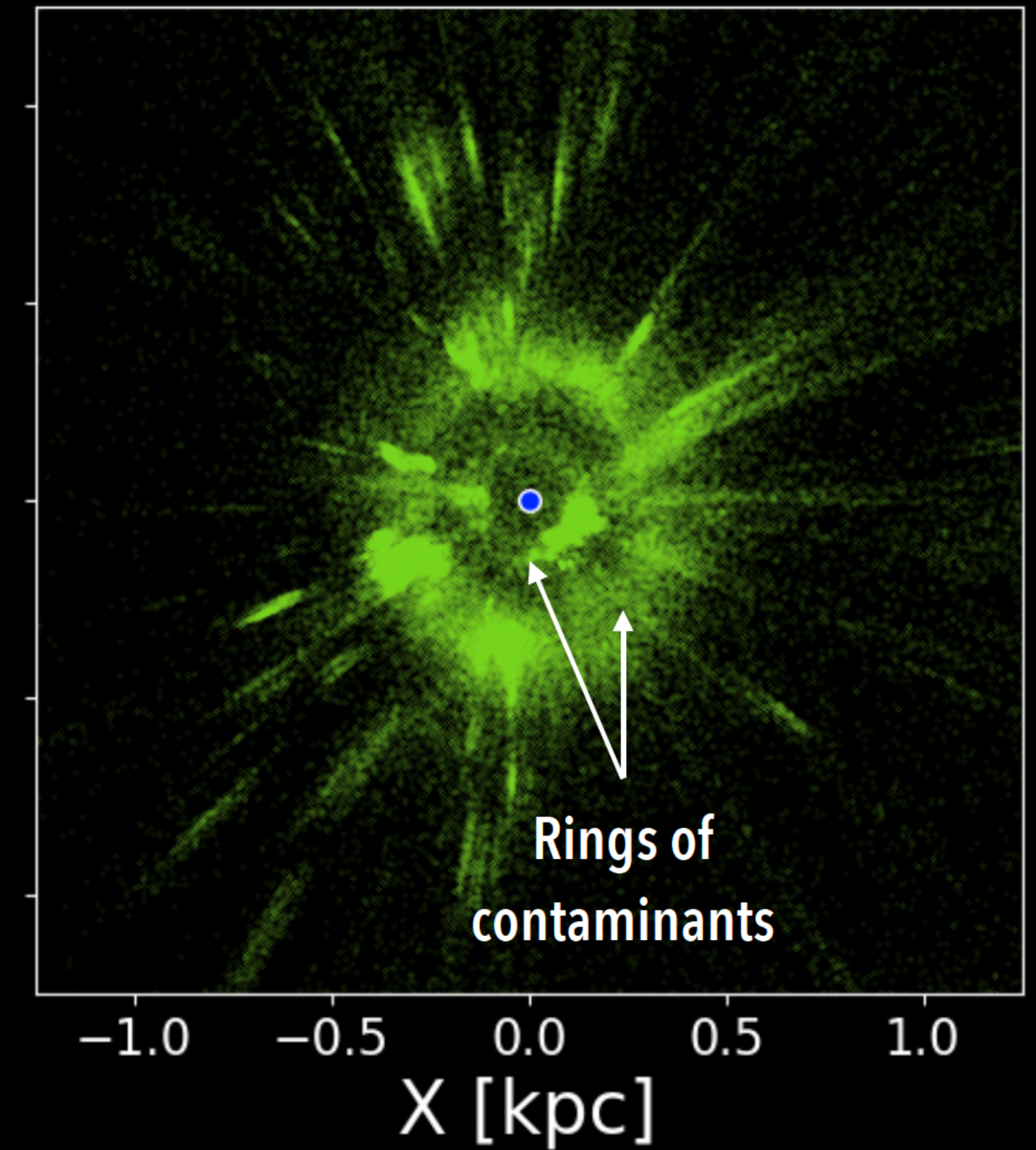
Young Star (SOTA Machine Learning)



Understanding Galactic baryon cycle

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

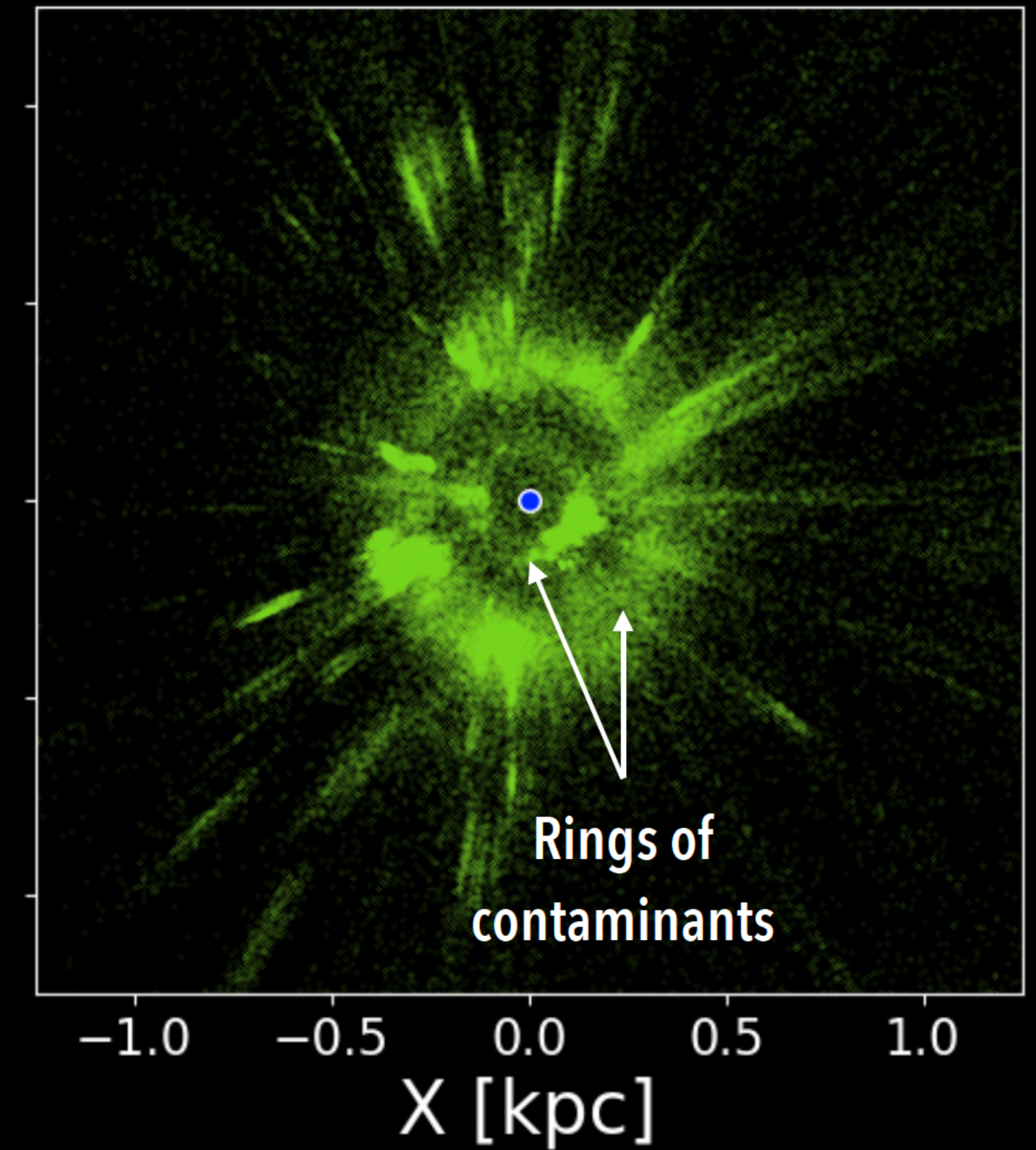
Young Star (SOTA Machine Learning)



Understanding Galactic baryon cycle

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

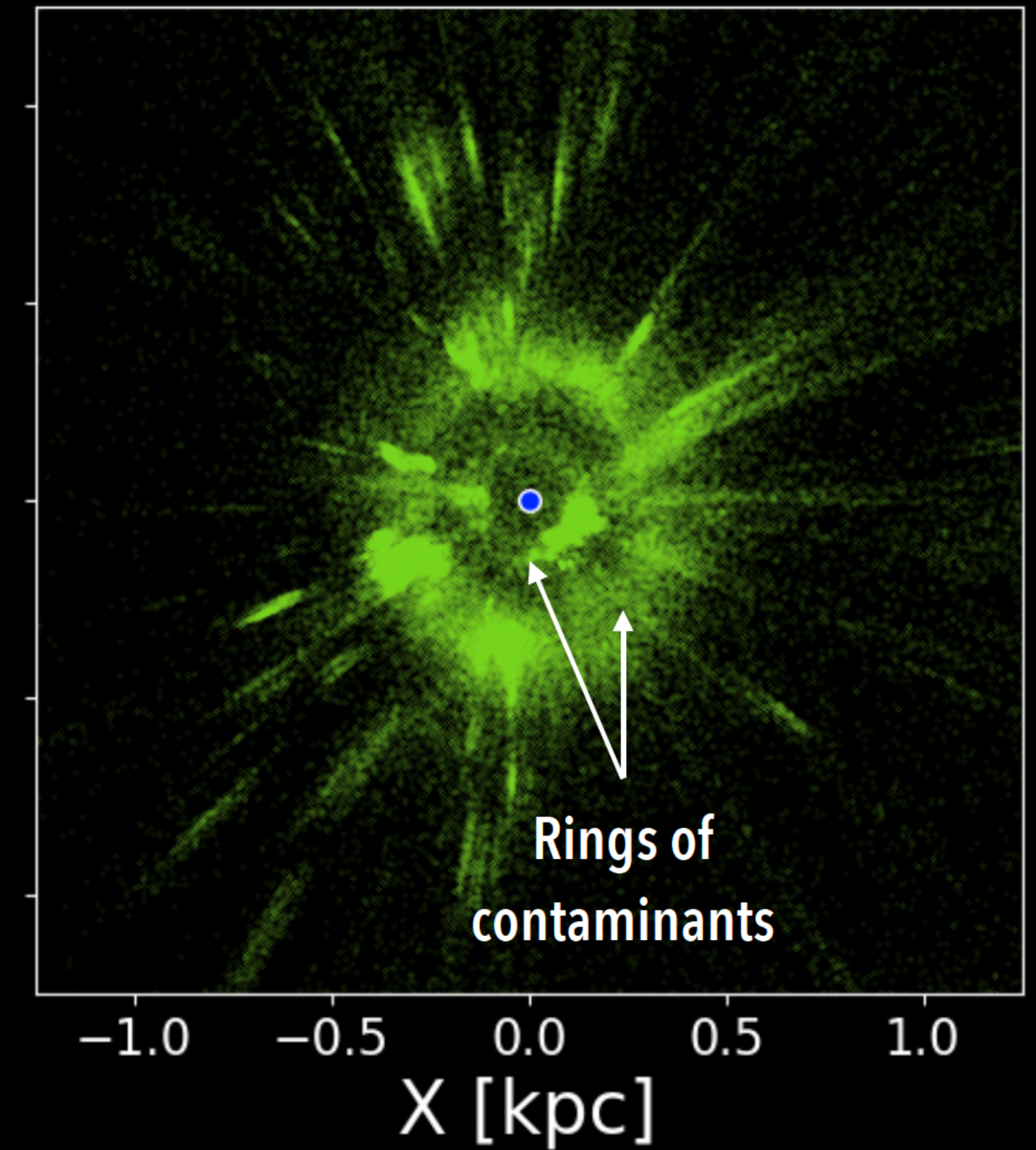
Young Star (SOTA Machine Learning)



Understanding Galactic baryon cycle

- YSOs connect *cloud* $\leftrightarrow$ *stars* $\leftrightarrow$ *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?

Young Star (SOTA Machine Learning)



Untapped potential of current methods

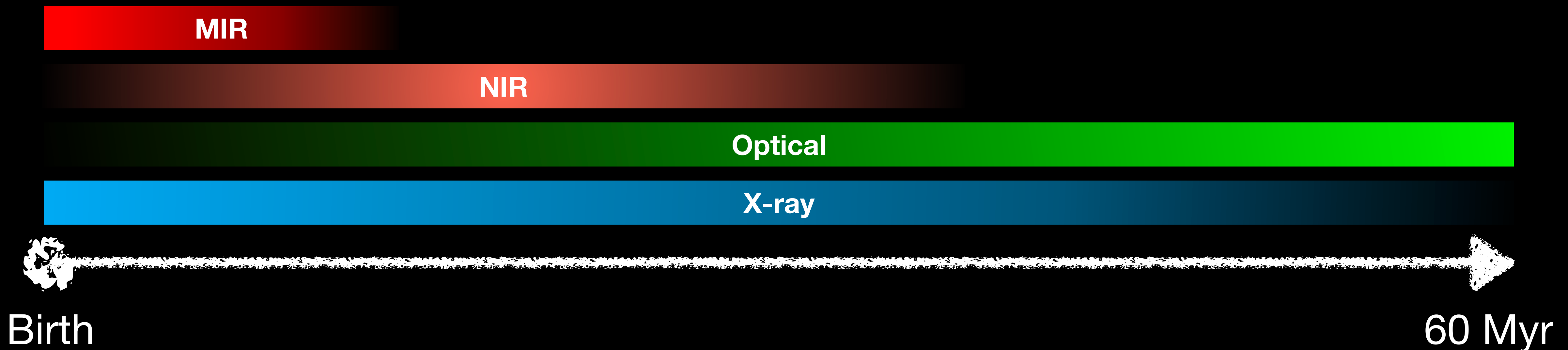
- Dedicated YSO models not employed

Untapped potential of current methods

- Dedicated YSO models not employed
- Simple binary NN classifier and age regressor trained on mock data — no probabilistic approach

Untapped potential of current methods

- Dedicated YSO models not employed
- Simple binary NN classifier and age regressor trained on mock data — no probabilistic approach
- Limited SED coverage with Gaia+2MASS+WISE photometry



What do we want

- Produce well-calibrated stellar parameter predictions

What do we want

- Produce well-calibrated stellar parameter predictions
- Data fusion: use as many informative data sets as possible

What do we want

- Produce well-calibrated stellar parameter predictions
- Data fusion: use as many informative data sets as possible
- Posteriors over stellar parameters given spectra & photometric observations of individual stars

What do we want

- Produce well-calibrated stellar parameter predictions
- Data fusion: use as many informative data sets as possible
- Posteriors over stellar parameters given spectra & photometric observations of individual stars
- Scale inference to $> 1\text{M} - 1\text{B}$ 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

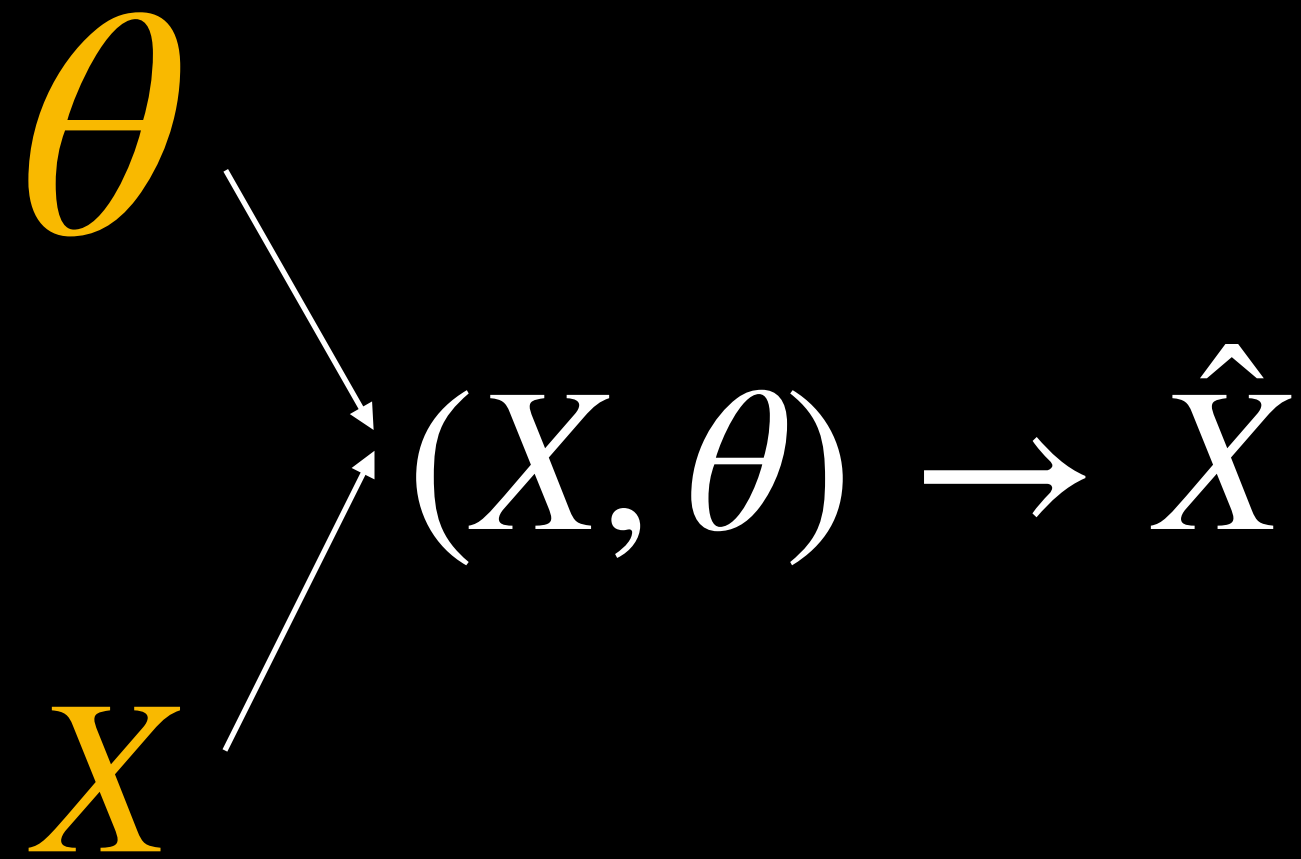
- Fusing surveys is hard due to different
 - resolutions & depths
 - coverage
 - instrument response
 - noise model
- Model misspecification leads to domain shift between simulated and real data

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

Model implementation

I. SBI model

Simformer: learning with incomplete data



Shape $\hat{X} \in (N, d_x + D_\theta = D_{\hat{X}})$

Inputs: θ, X

Simformer: learning with incomplete data

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

Value emb

ID emb

Conditional state emb

Shape $\hat{X} \in (N, D_{\hat{X}}) \rightarrow \hat{T} \in (N, D_{\hat{X}}, E_V + E_{\text{id}} + E_C)$

Inputs: $\theta, X, M_C \in \{0,1\}^{D_x}$

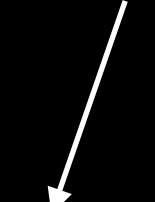
Simformer: learning with incomplete data

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

Shape

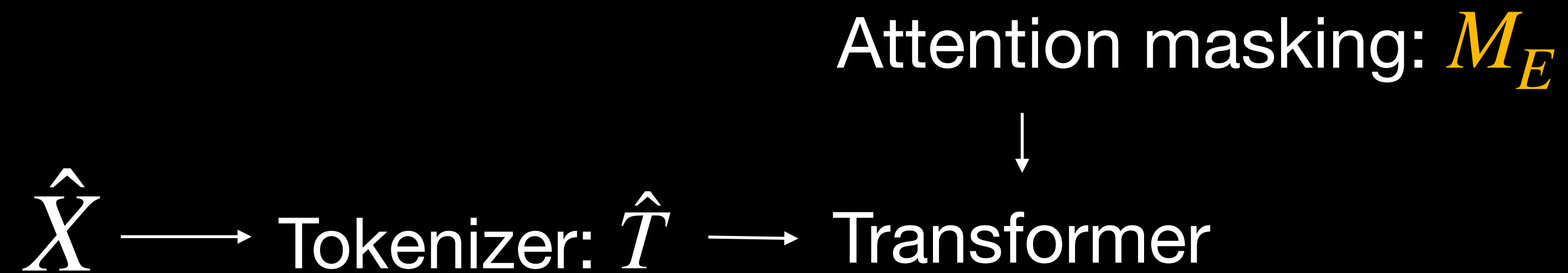
$$\hat{T} \in (N, D_{\hat{X}}, E) \rightarrow S \propto QK^T \in (N, D_{\hat{X}}, D_{\hat{X}})$$

$Q = \hat{T}W_Q \in (N, D_{\hat{X}}, d_v)$



Inputs: $\theta, X, M_C \in \{0,1\}^{D_x}$
 $\phi : \quad W_Q, W_K, W_V \in (E, d_v)$

Simformer: learning with incomplete data

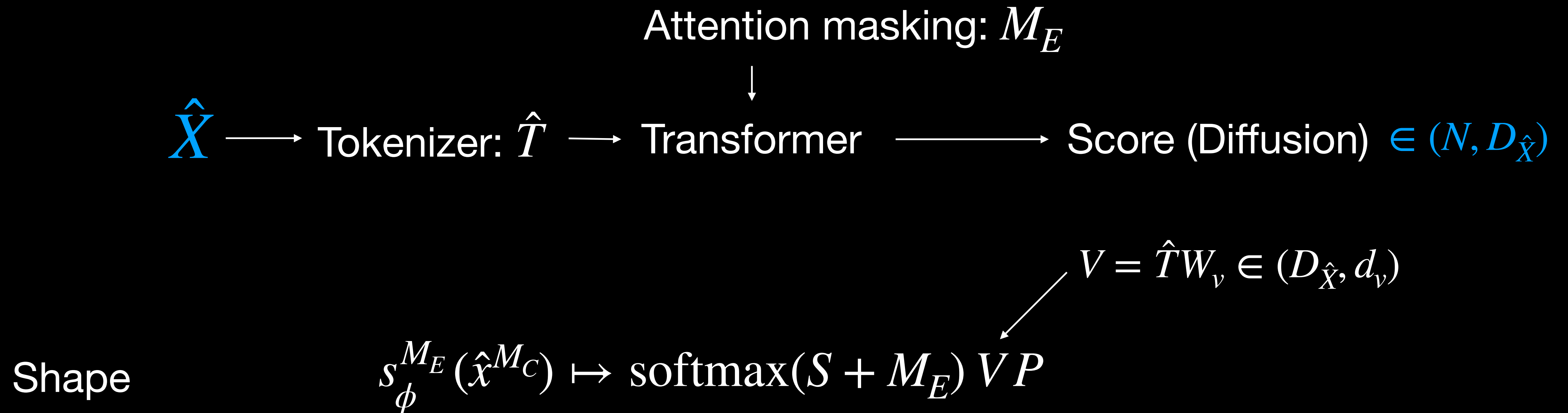


Shape

$$S \rightarrow S + M_E$$

Inputs: $\theta, X, M_C, M_E \in (N, D_{\hat{X}}, D_{\hat{X}})$
 $\phi : \quad W_Q, W_K, W_V$

Simformer: learning with incomplete data



Quick recap

Score based diffusion models

Langevin dynamics

Rossy, 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 $p(\mathbf{x})$

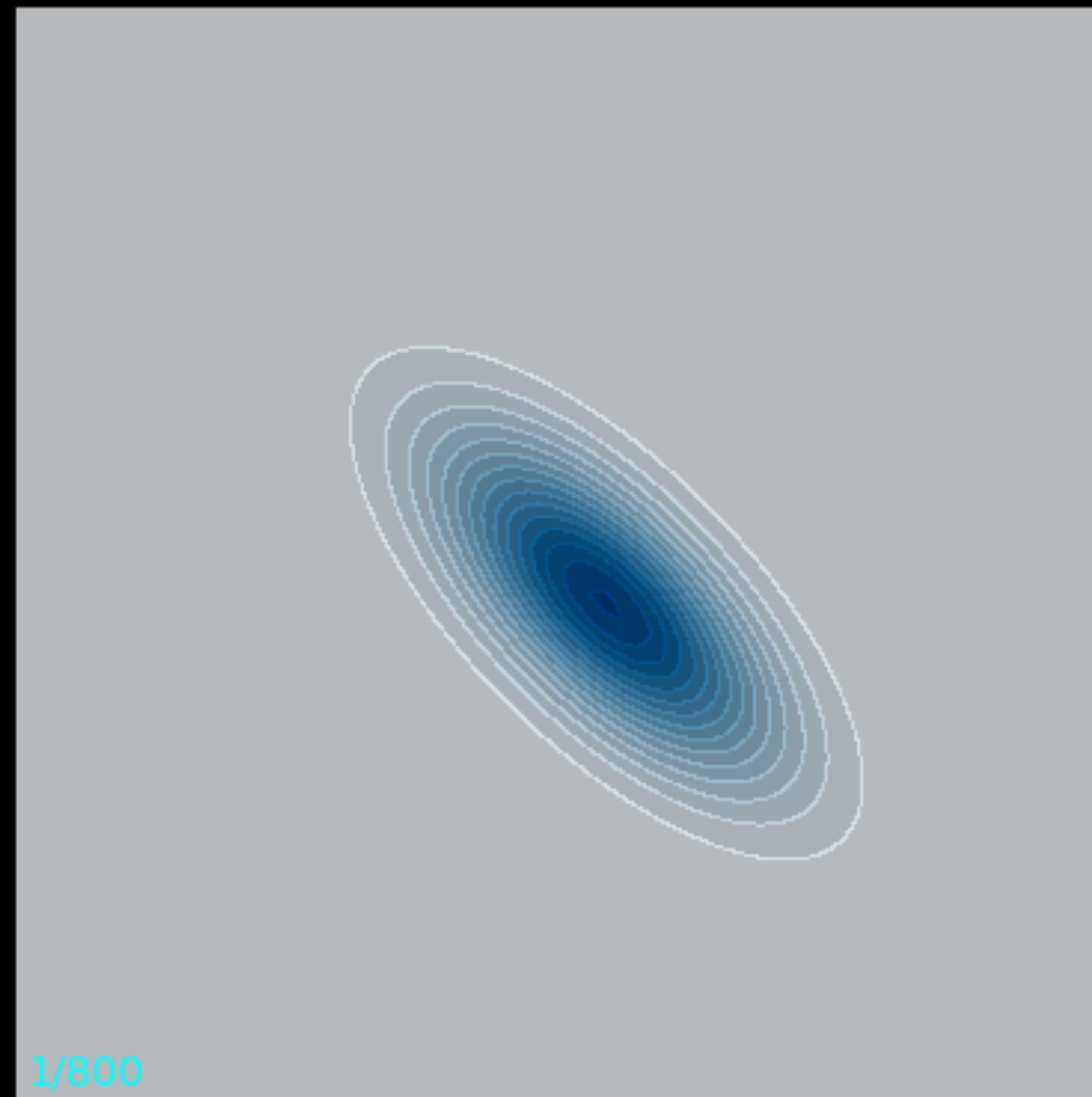
Langevin dynamics

Rossy, Doll & Friedman (1978)

Besag (1994)

Roberts & Tweedie (1996)

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon} \mathcal{N}(0, I_d) \quad \dots \text{MCMC method that samples from } p(\mathbf{x})$$



Langevin dynamics

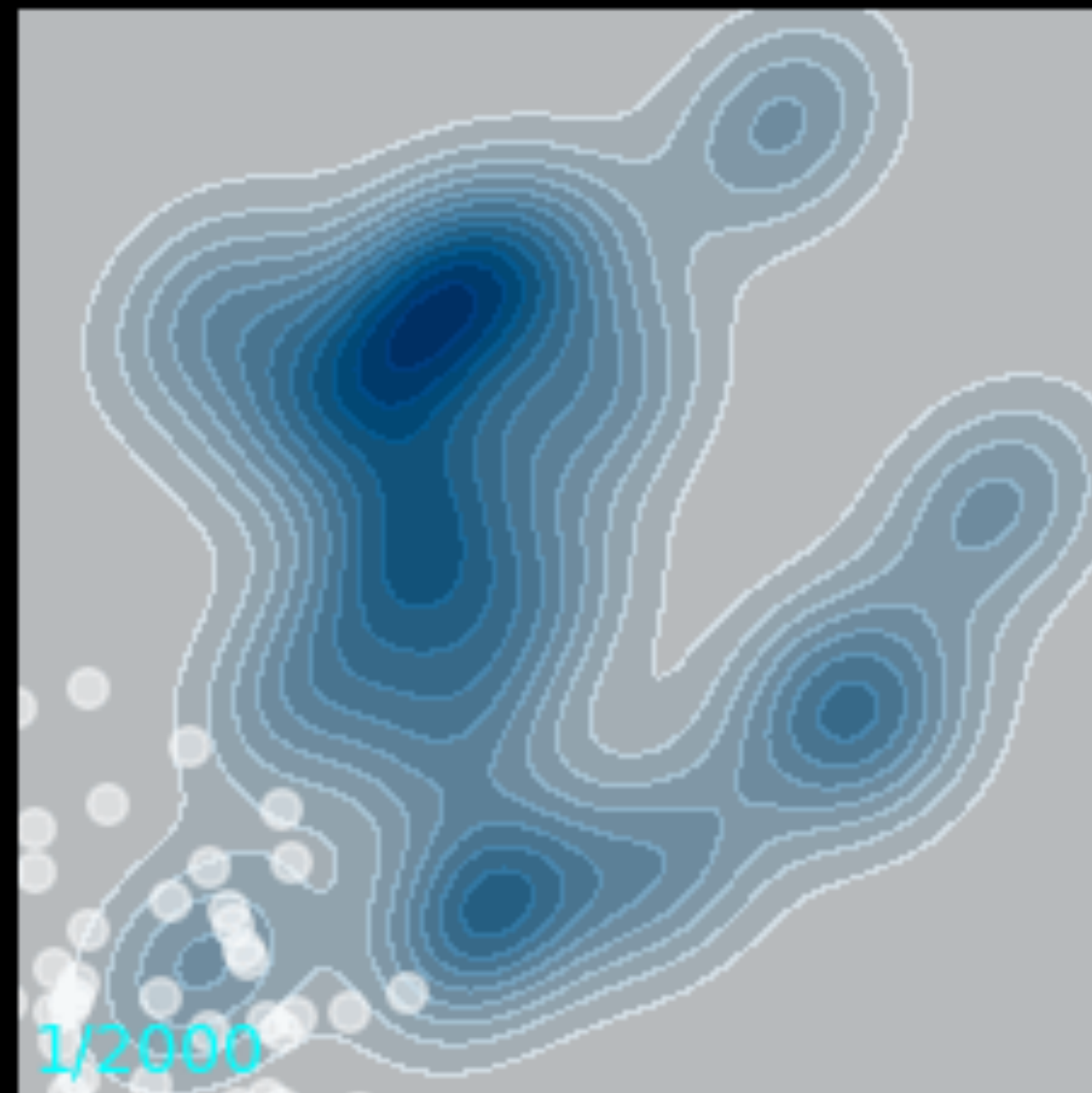
Rossy, Doll & Friedman (1978)

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

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon} \mathcal{N}(0, I_d) \quad \dots \text{MCMC method that samples from } p(\mathbf{x})$$

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

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

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

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

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

Estimating the score

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

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

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

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

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

Estimating the score

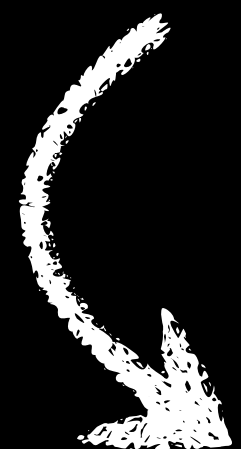
$$\mathcal{L} = \mathbb{E}_{p(\mathbf{x})}[\|s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x})\|_2^2]$$

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

Vincent (2010) “A Connection Between Score Matching and Denoising Autoencoders”

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

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



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

Estimating the score

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

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

Vincent (2010) “A Connection Between Score Matching and Denoising Autoencoders”

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

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



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

$\dot{\mathbf{x}}$

Estimating the score

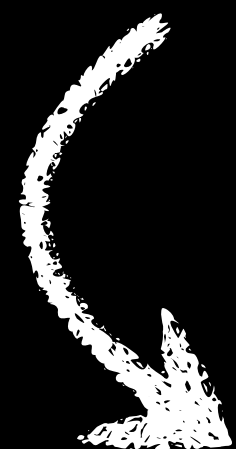
$$\mathcal{L} = \mathbb{E}_{p(\mathbf{x})}[\|s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x})\|_2^2]$$

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

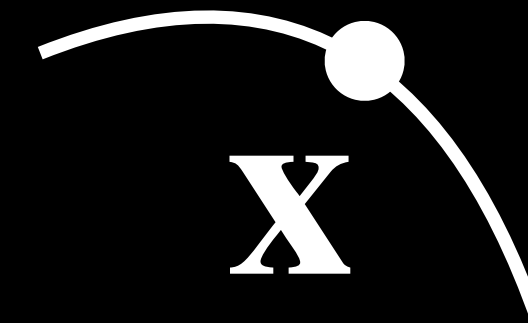
Vincent (2010) “A Connection Between Score Matching and Denoising Autoencoders”

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

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



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



Estimating the score

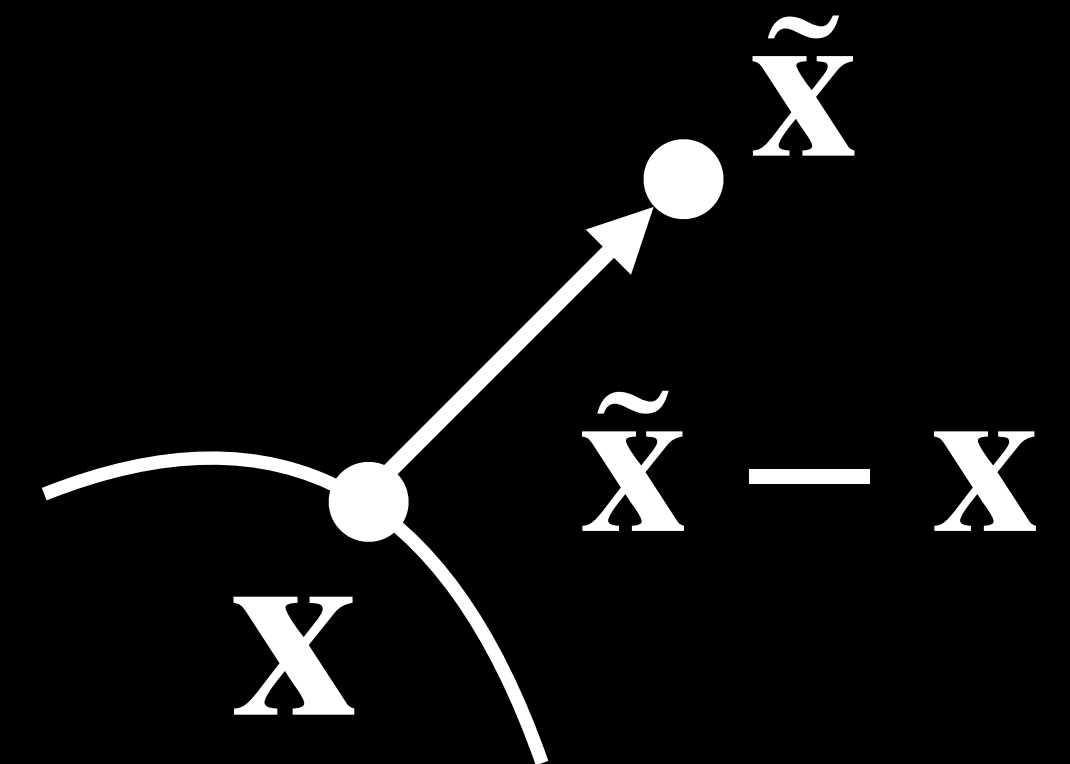
$$\mathcal{L} = \mathbb{E}_{p(\mathbf{x})}[\|s_\phi(\mathbf{x}) - \nabla \log p(\mathbf{x})\|_2^2]$$

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

Vincent (2010) “A Connection Between Score Matching and Denoising Autoencoders”

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

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




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

Estimating the score

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

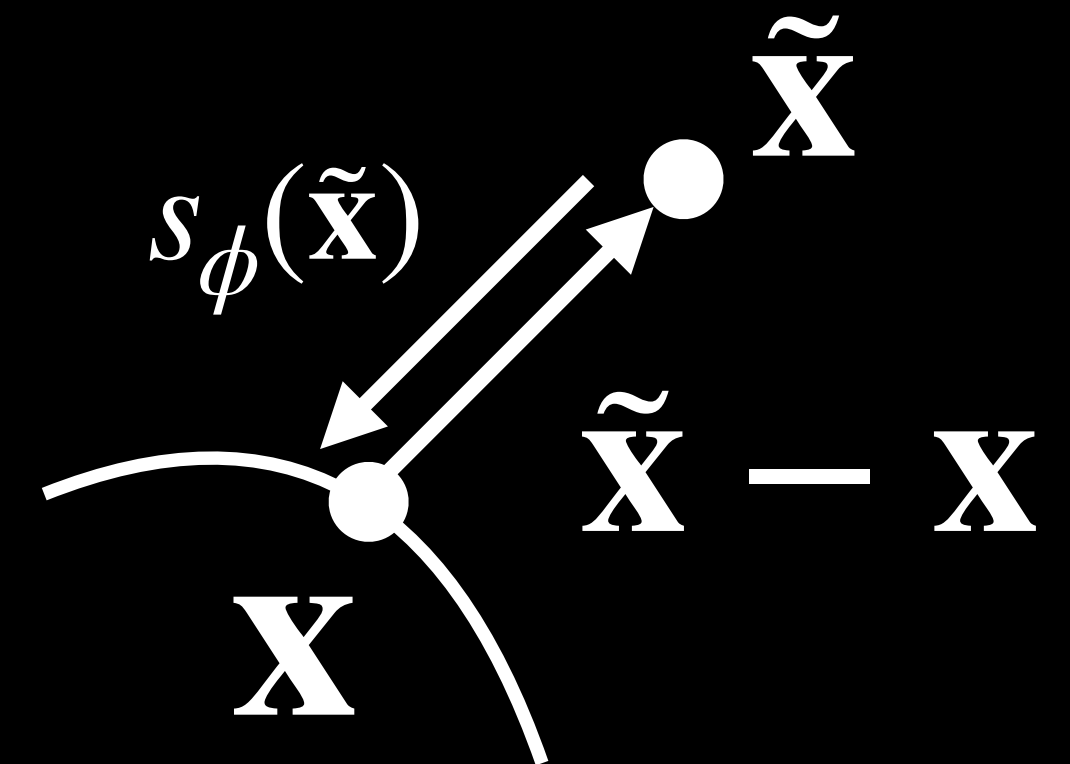
Problem 1: don't have access to $p(\mathbf{x})$

Vincent (2010) “A Connection Between Score Matching and Denoising Autoencoders”

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

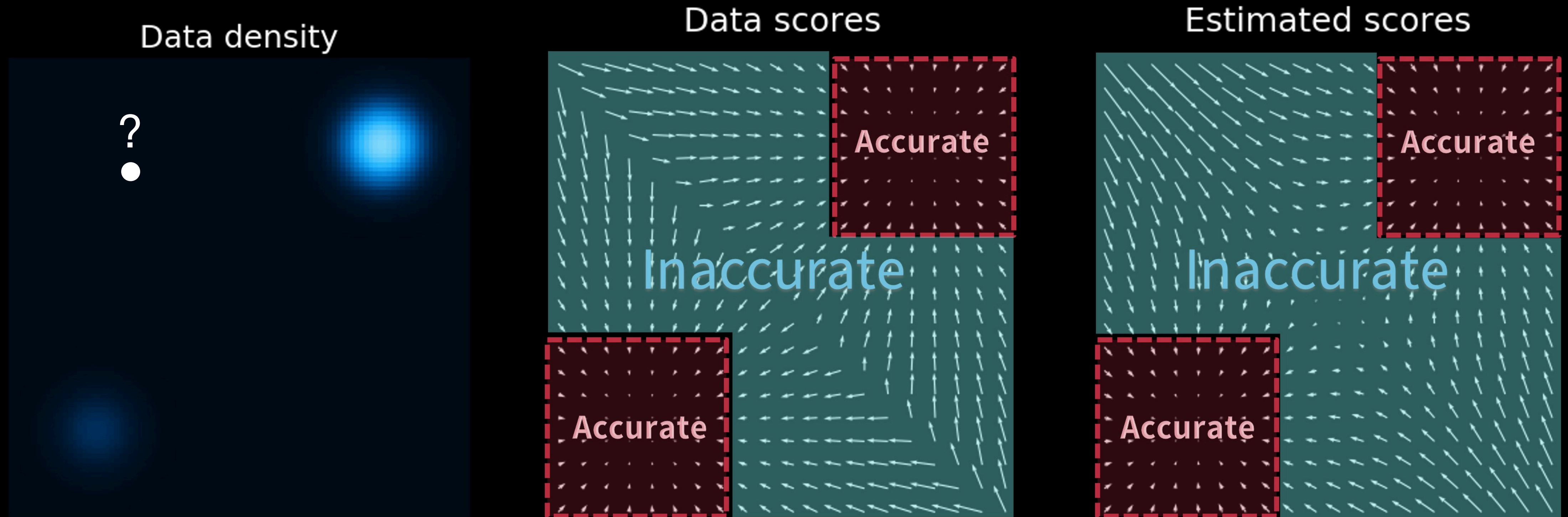
$$p_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x}, \sigma^2 \mathbf{1}_d)$$


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



Training score models

Problem 2: low coverage of data space $\rightarrow$ inaccurate score

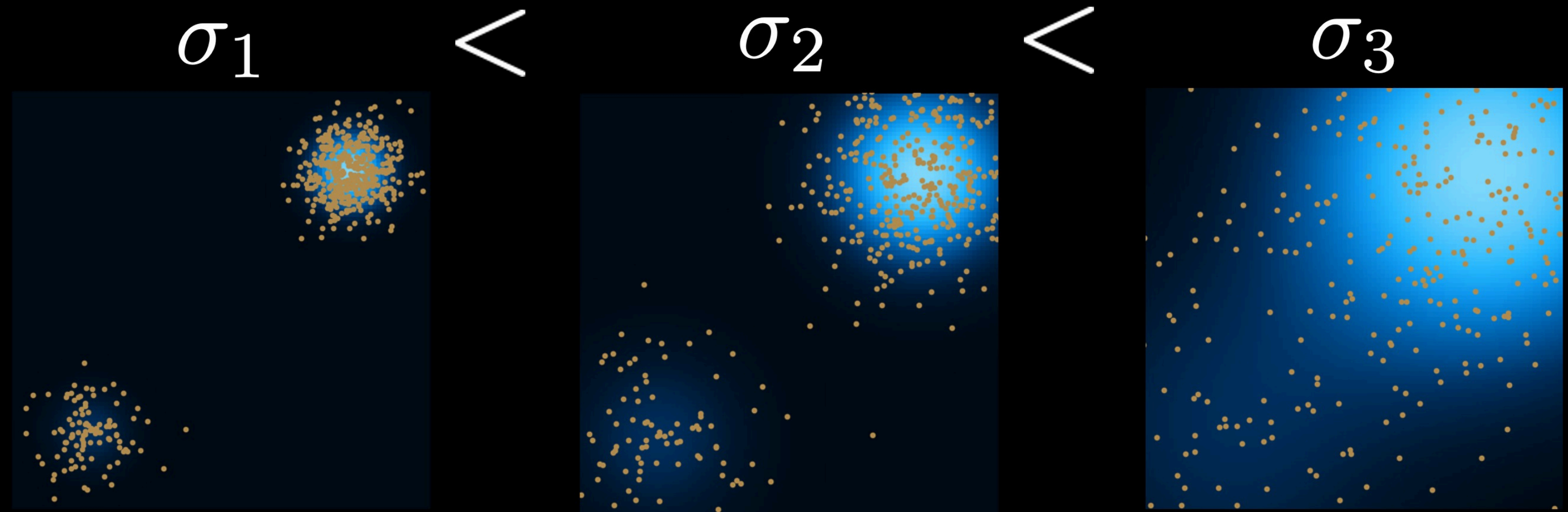


Training score models: multiple noise levels

Fix: add gradual noise

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

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

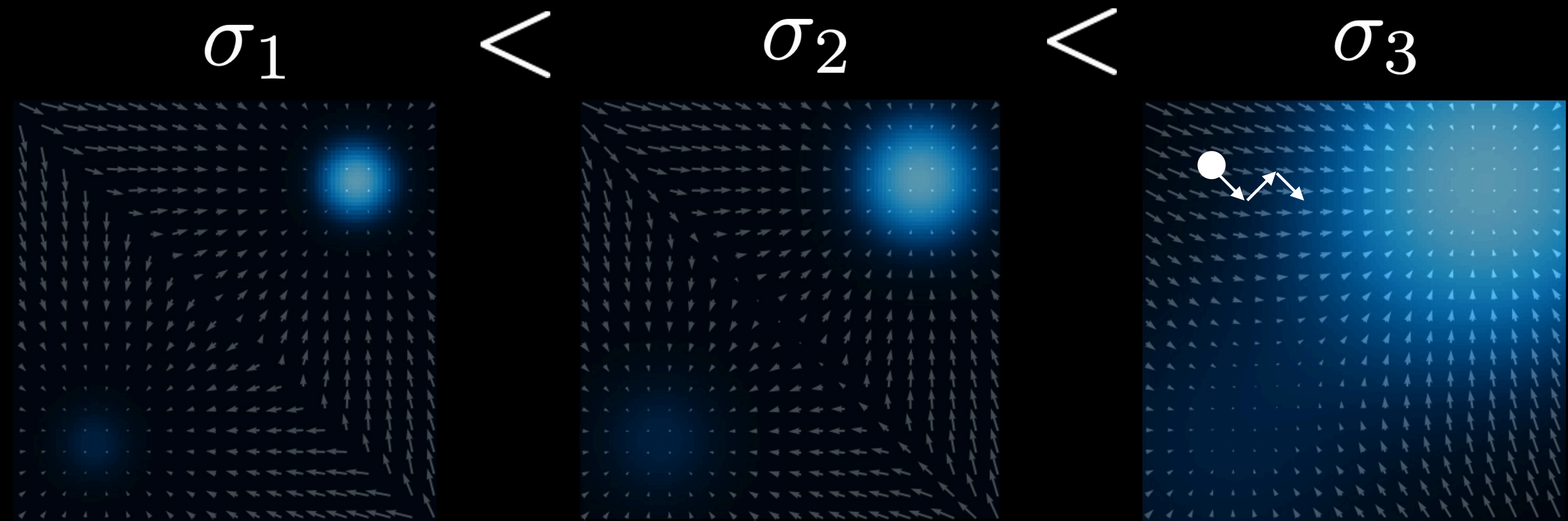


Training score models: multiple noise levels

Fix: add gradual noise

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

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



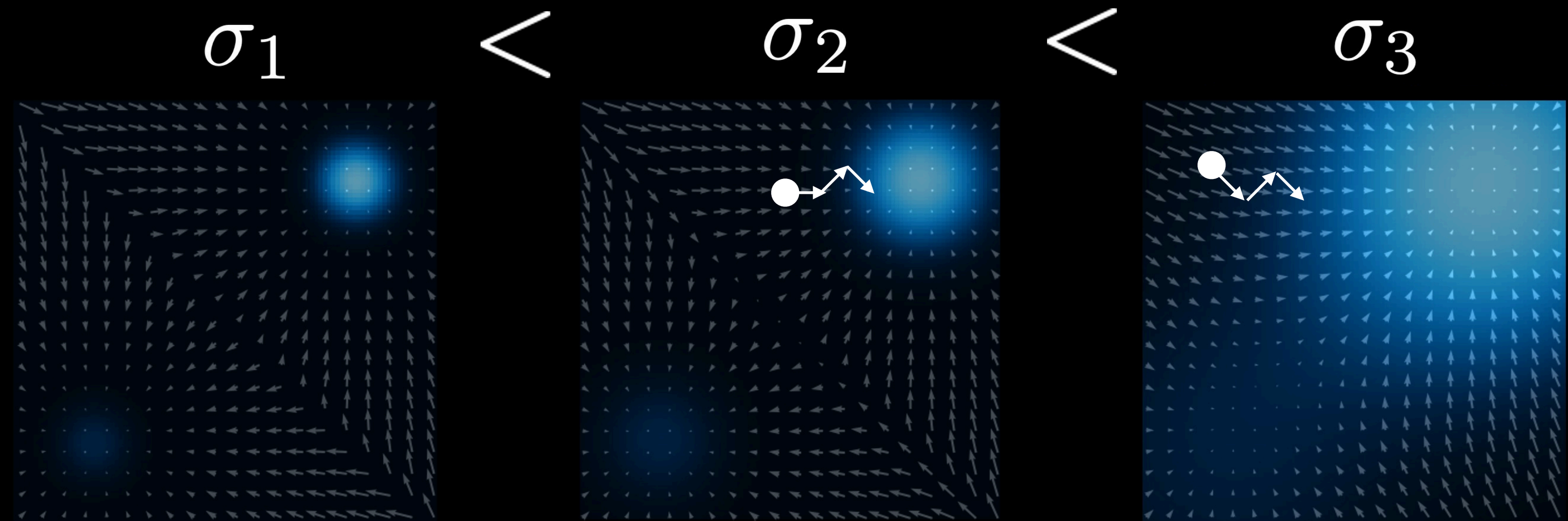
Sampling via **annealed** Langevin dynamics

Training score models: multiple noise levels

Fix: add gradual noise

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

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



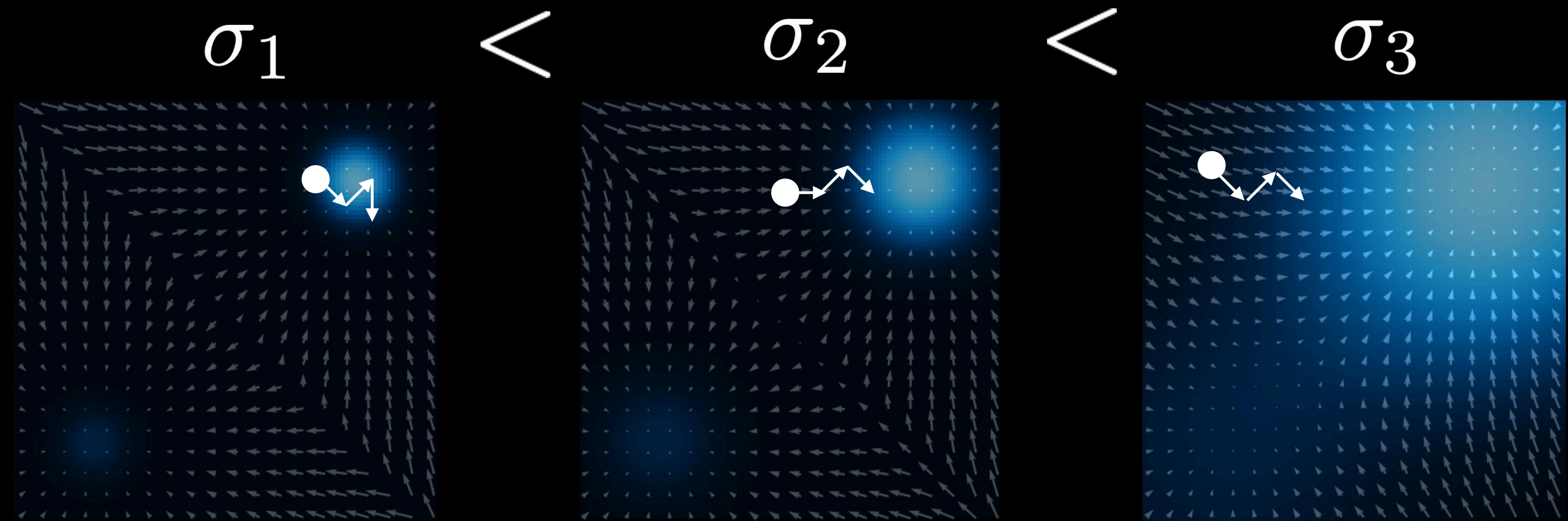
Sampling via **annealed** Langevin dynamics

Training score models: multiple noise levels

Fix: add gradual noise

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

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

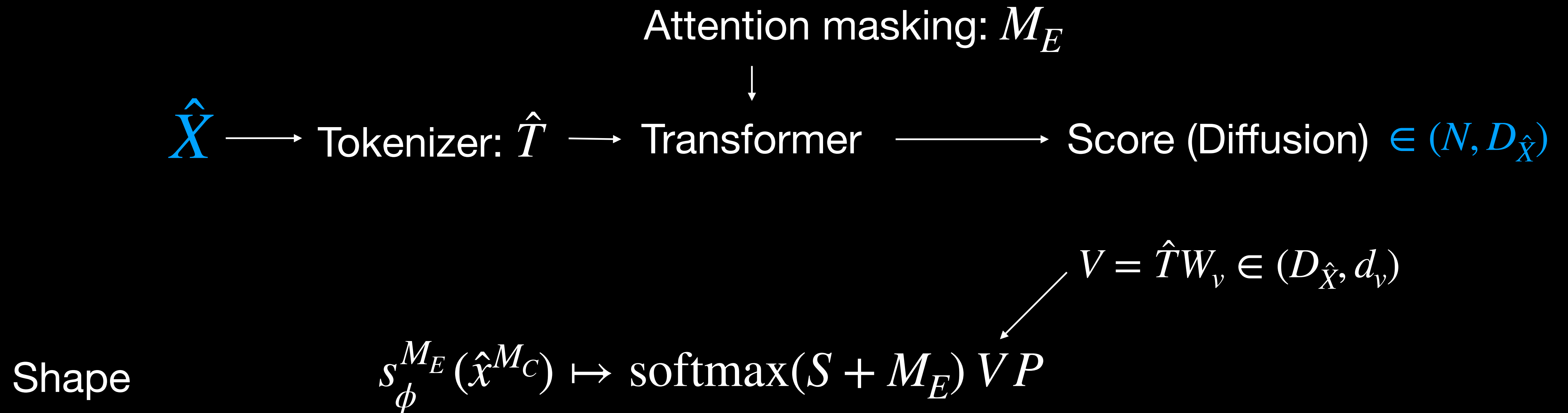


Sampling via **annealed** Langevin dynamics

Model implementation

II. Learning to condition and marginalize

Simformer: learning with incomplete data



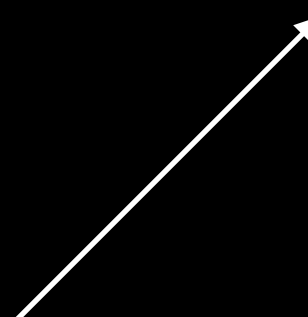
Training

Given Score model: $s_{\phi}^{M_E}(\hat{x}^{M_C})$

Training

Given Score model: $s_{\phi}^{M_E}(\hat{x}^{M_C})$

1. Add noise acc. to noise schedule $\hat{x}_t^{M_C} = (1 - M_C)\hat{x}_t + M_C\hat{x}_0$

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


Training

Given Score model: $s_{\phi}^{M_E}(\hat{x}^{M_C})$

1. Add noise acc. to noise schedule $\hat{x}_t^{M_C} = (1 - M_C)\hat{x}_t + M_C\hat{x}_0$

2. Compute denoising loss

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) = (1 - M_C) \cdot \left(s_{\phi}^{M_E}(\hat{\mathbf{x}}_t^{M_C}, t) - \nabla_{\hat{\mathbf{x}}_t} \log p_t(\hat{\mathbf{x}}_t | \hat{\mathbf{x}}_0) \right)$$

Training

Given Score model: $s_{\phi}^{M_E}(\hat{x}^{M_C})$

1. Add noise acc. to noise schedule $\hat{x}_t^{M_C} = (1 - M_C)\hat{x}_t + M_C\hat{x}_0$

2. Compute denoising loss

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) = (1 - M_C) \cdot \left(s_{\phi}^{M_E}(\hat{\mathbf{x}}_t^{M_C}) - \nabla_{\hat{\mathbf{x}}_t} \log p_t(\hat{\mathbf{x}}_t | \hat{\mathbf{x}}_0) \right)$$

→ With M_C score approximates any conditional

Training

Given Score model: $s_{\phi}^{M_E}(\hat{x}^{M_C})$

1. Add noise acc. to noise schedule $\hat{x}_t^{M_C} = (1 - M_C)\hat{x}_t + M_C\hat{x}_0$

2. Compute denoising loss

$$\ell(\phi, M_C, t, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t) = (1 - M_C) \cdot \left(s_{\phi}^{M_E}(\hat{\mathbf{x}}_t^{M_C}, t) - \nabla_{\hat{\mathbf{x}}_t} \log p_t(\hat{\mathbf{x}}_t | \hat{\mathbf{x}}_0) \right)$$

→ With M_C score approximates any conditional

→ With M_E score approximates any marginal

Model implementation

II. Learning to condition and marginalize

Conditional properties: 3D example

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

Conditional properties: 3D example

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

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

Conditional properties: 3D example

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

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

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

Conditional properties: 3D example

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

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

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

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

Conditional properties: 3D example

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

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

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

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

$$\nabla_z \log p(x, y | z) + \nabla_z \log p(z)$$

Conditional properties: 3D example

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

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

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

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

$$\nabla_z \log p(x, y | z) + \nabla_z \log p(z) \leftarrow \text{Is set to 0 due to } (1 - M_C) \text{ in } \ell$$

Marginalization properties

Attention masking: M_E

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

Marginalization properties

Attention masking: M_E



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

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

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

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

Marginalization properties

Attention masking: M_E



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

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

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

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

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

Marginalization properties

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

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

Marginalization properties

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

$$\begin{aligned} \mathcal{L} &\propto (f(x, y) - \Delta x)^2 + (f(x, y) - \Delta y)^2 + (f(z) - \Delta z)^2 = \\ &= \nabla \log p(x, y) + \nabla \log p(z) = \nabla \log p(x, y)p(z) \end{aligned}$$

Model implementation

III. Domain adaption

Architecture



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

θ

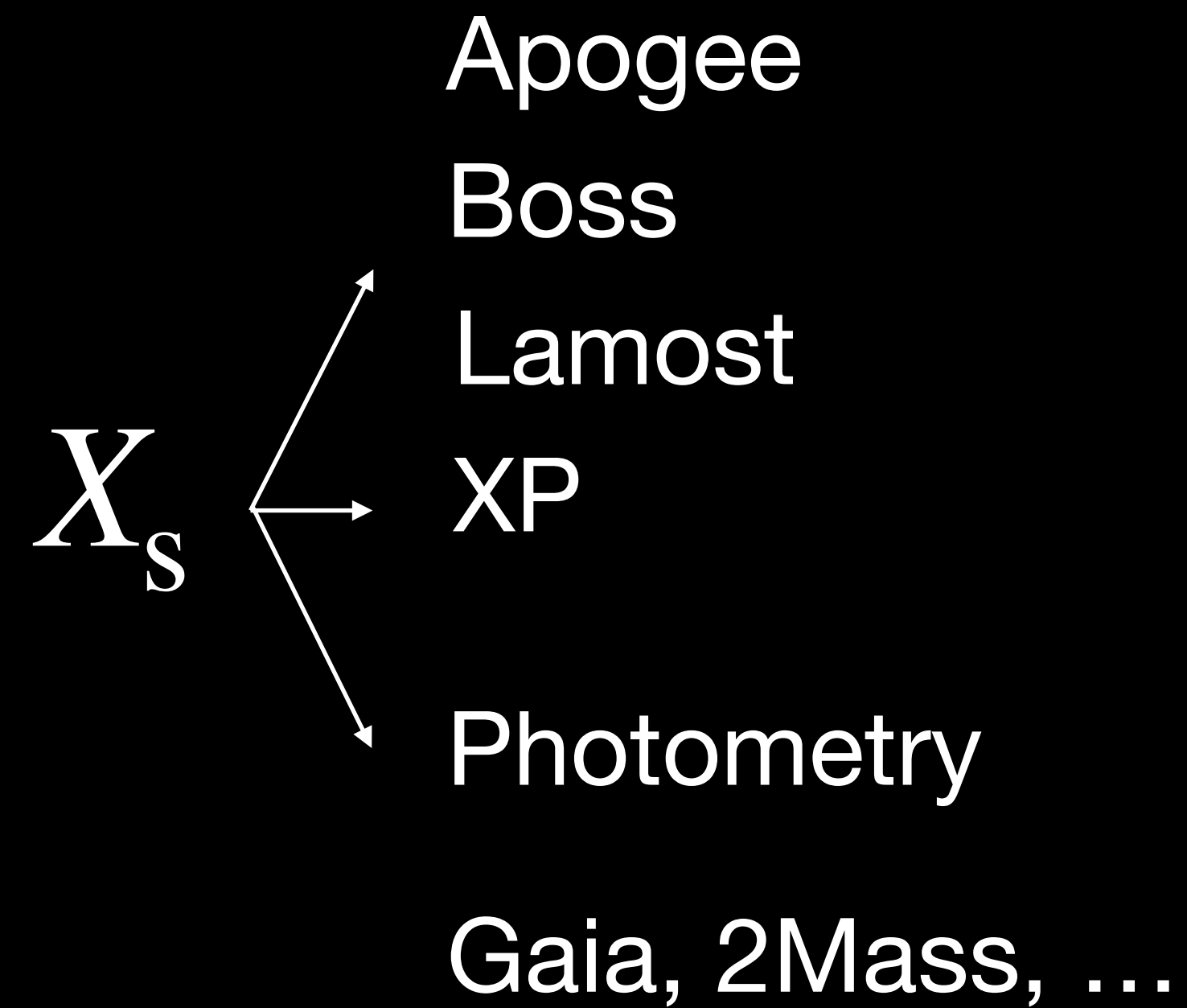
$X \rightarrow \begin{matrix} X_{\text{sim}} \\ X_{\text{real}} \\ X_{\text{sim-real-pairs}} \end{matrix}$

$\rightarrow$ Transformer $\rightarrow$ Diffusion

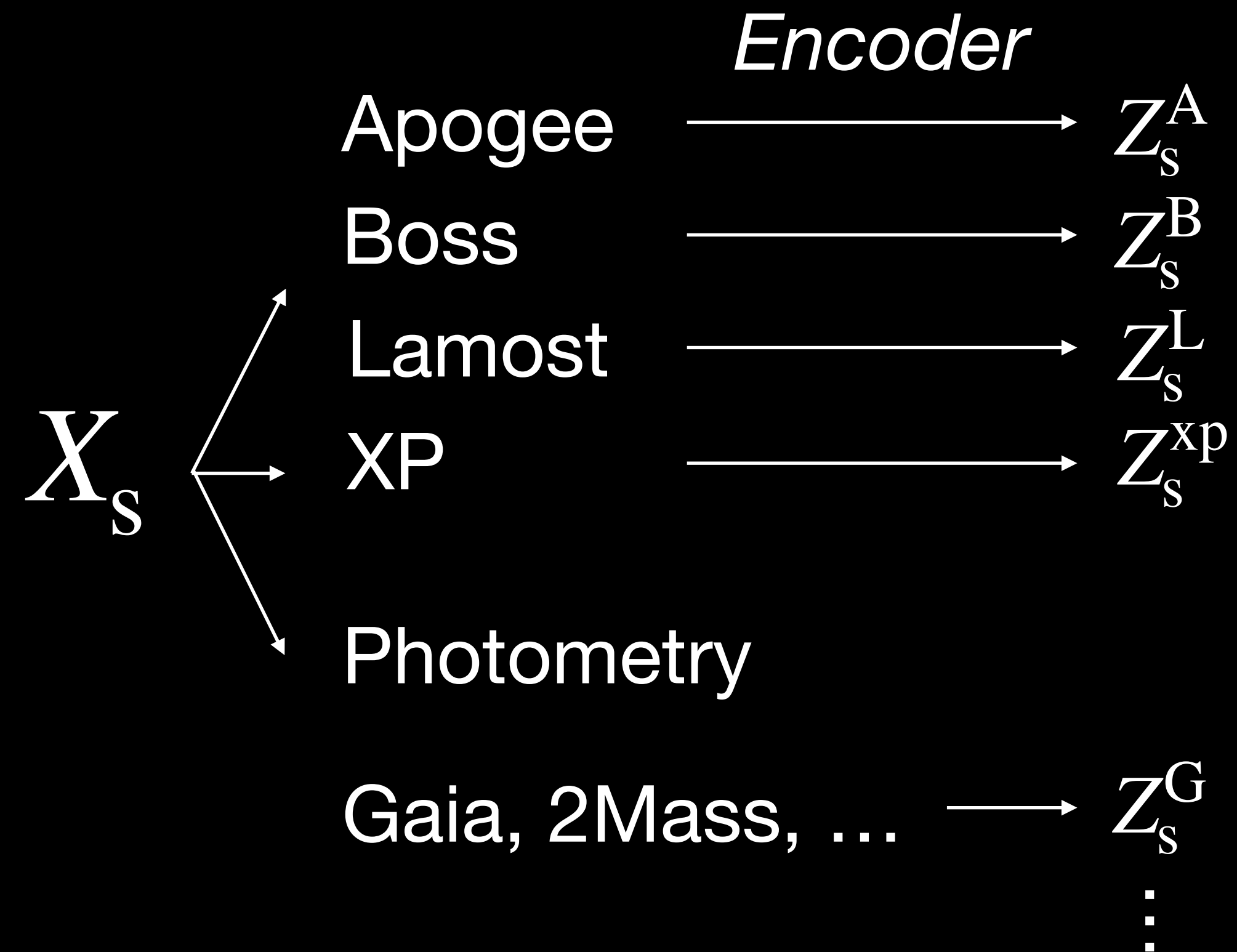
Modality encoders

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

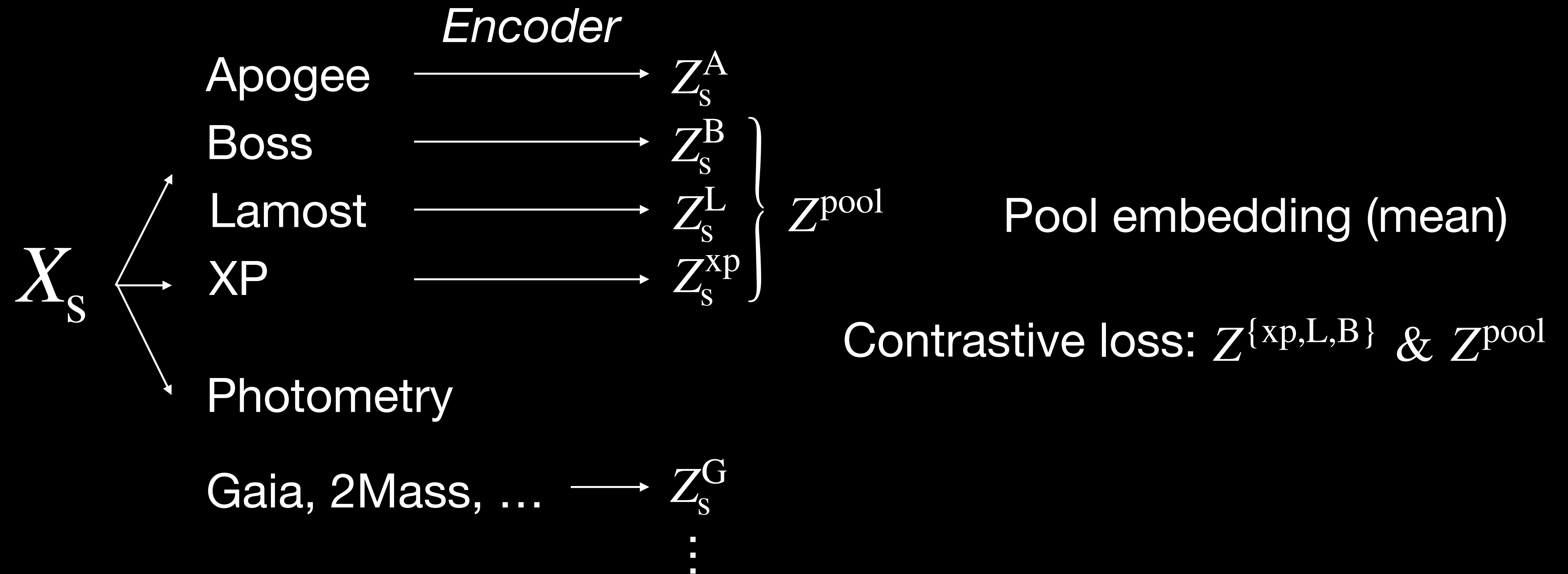
Modality encoders: split into indiv. spectra



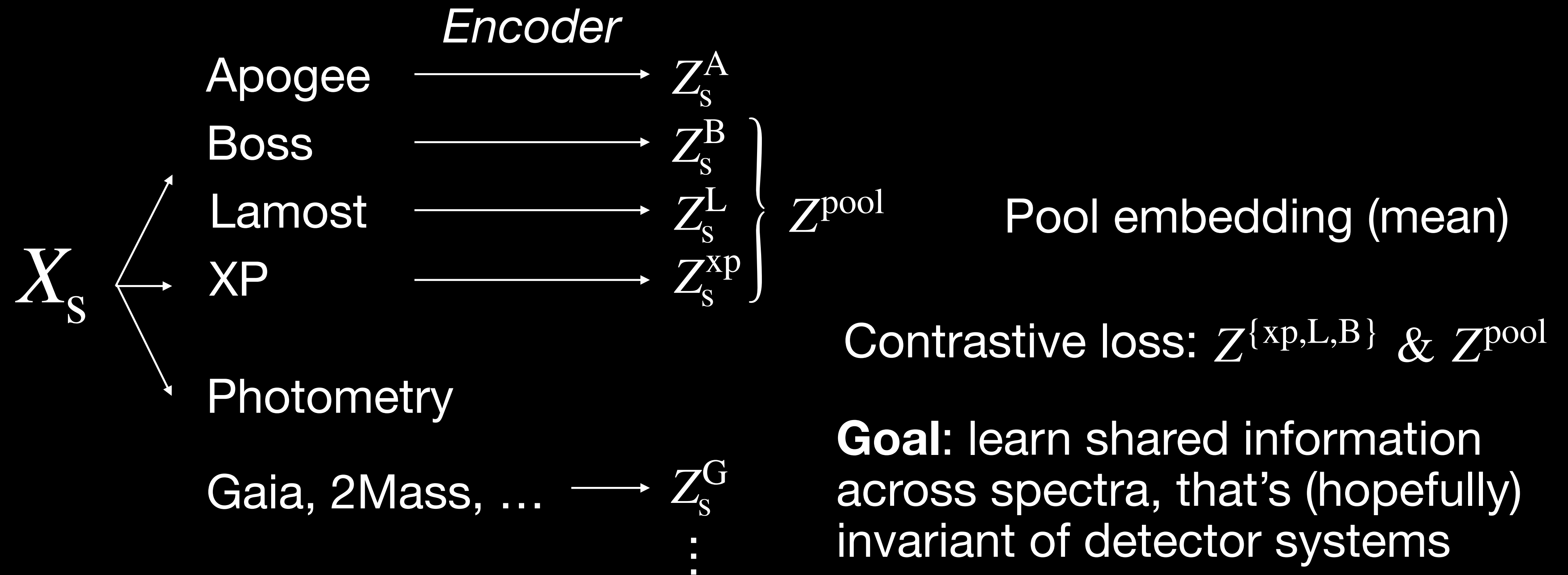
Modality encoders: encode



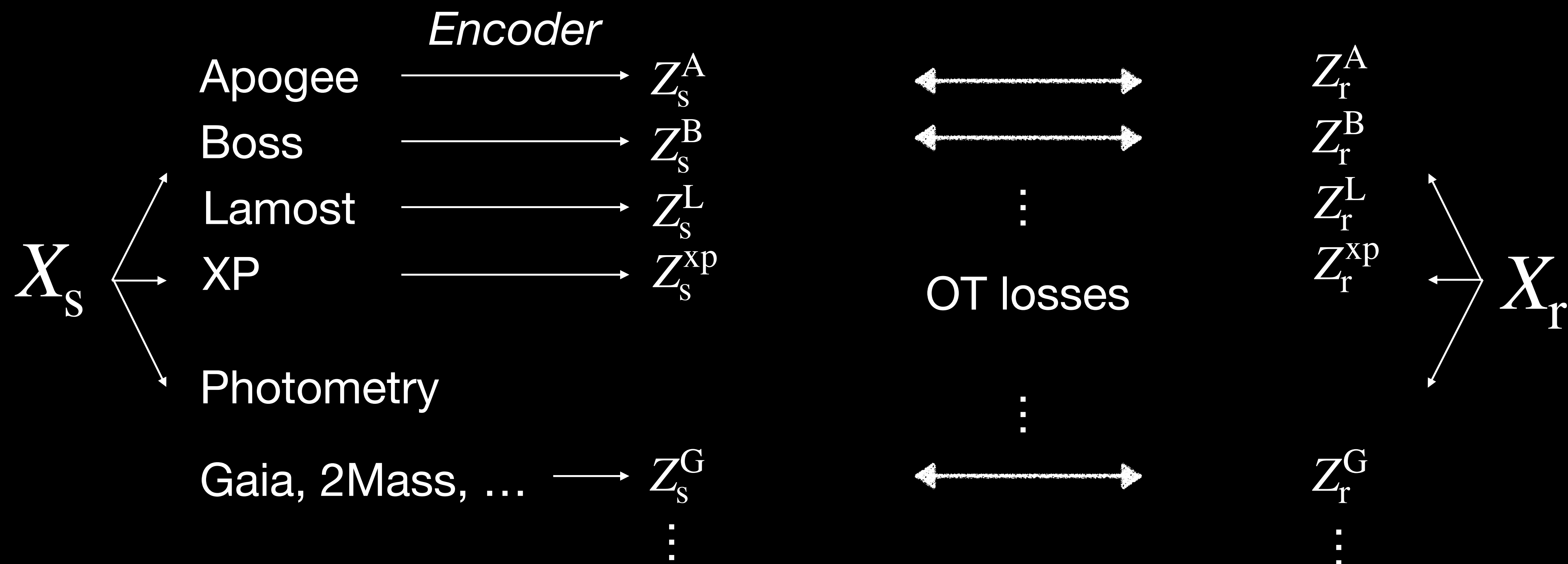
Modality encoders: contrastive loss



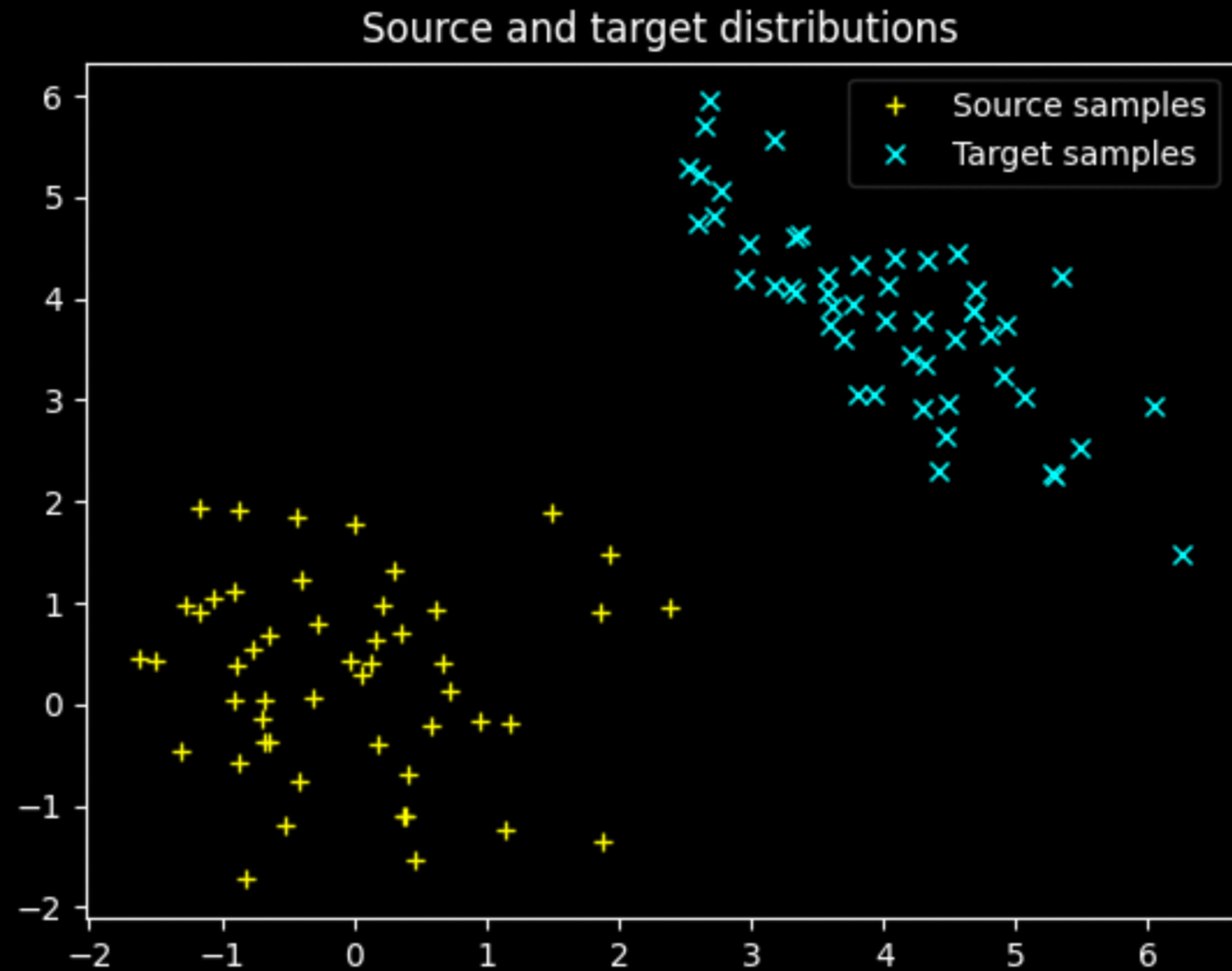
Modality encoders: contrastive loss



Modality encoders: alignment loss

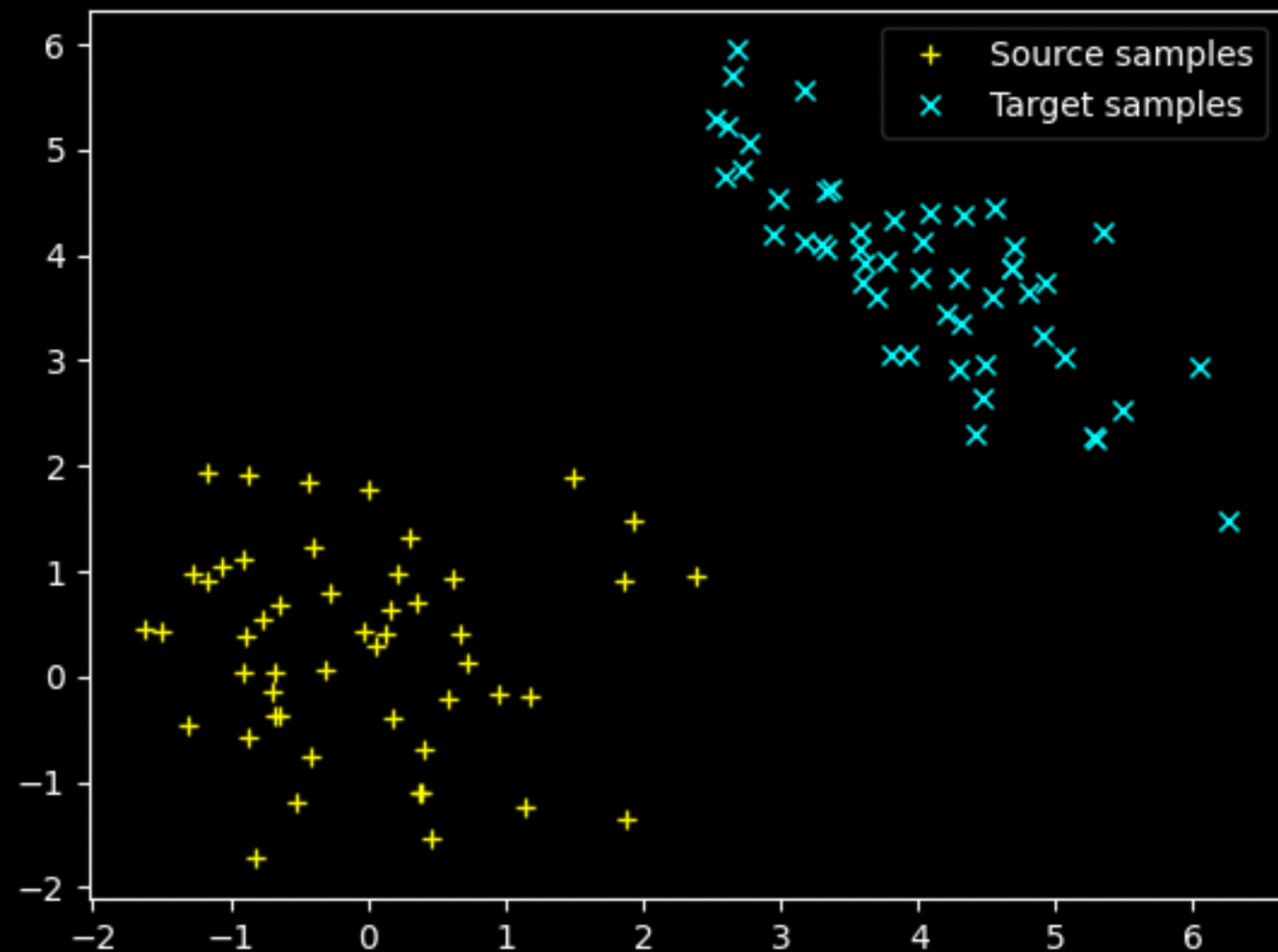


Alignment via optimal transport

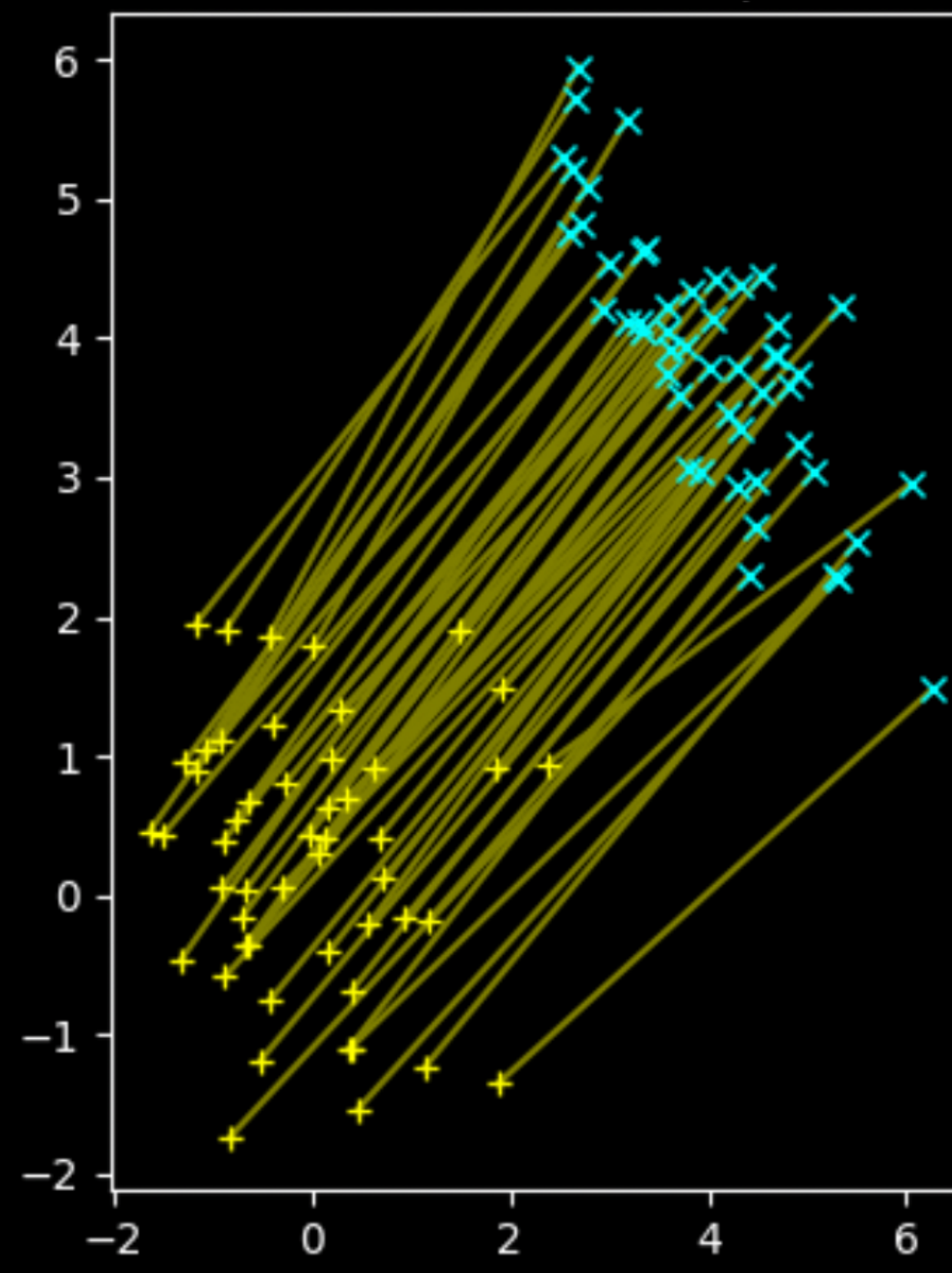


Alignment via optimal transport

Source and target distributions

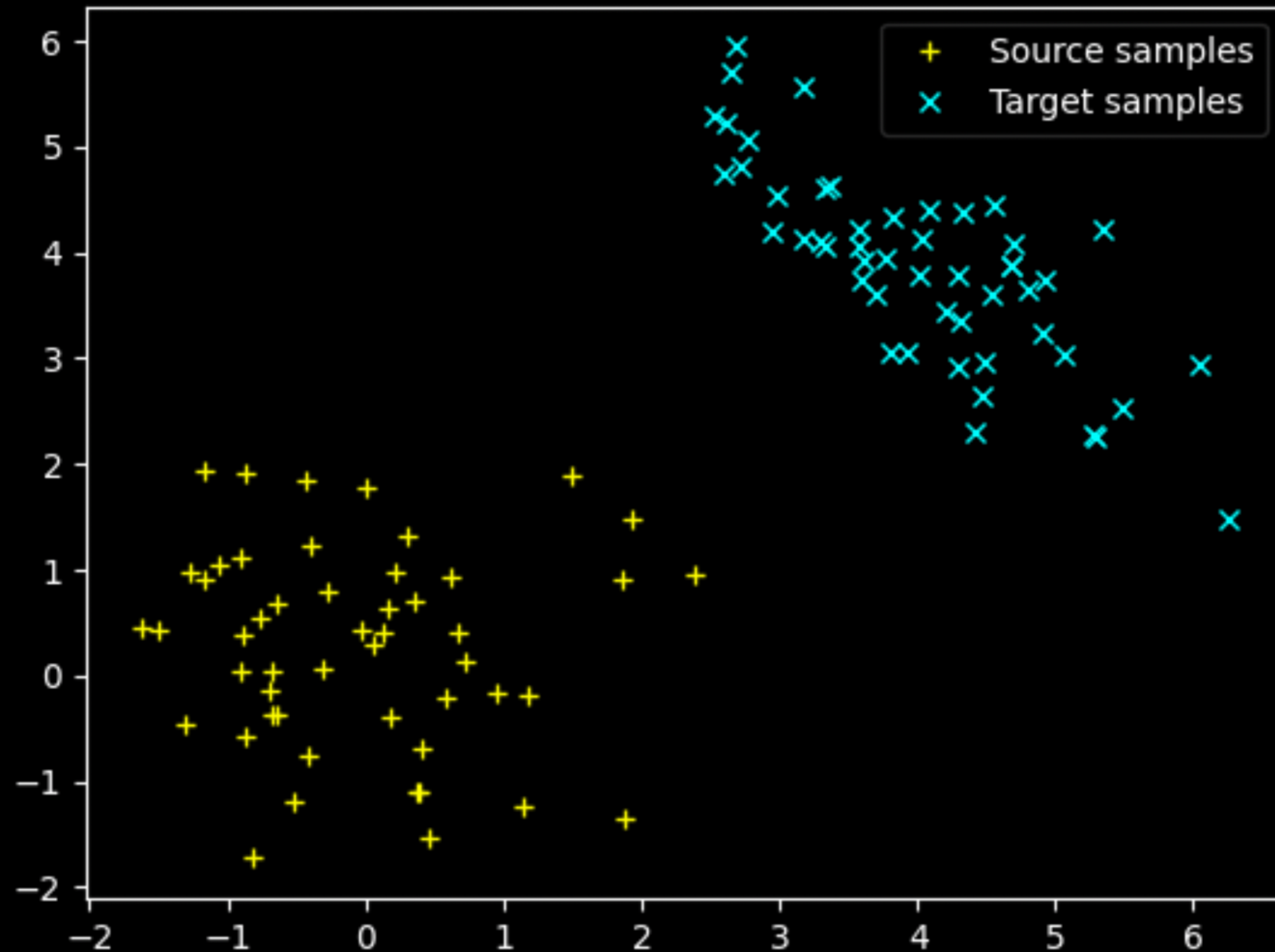


Exact solution

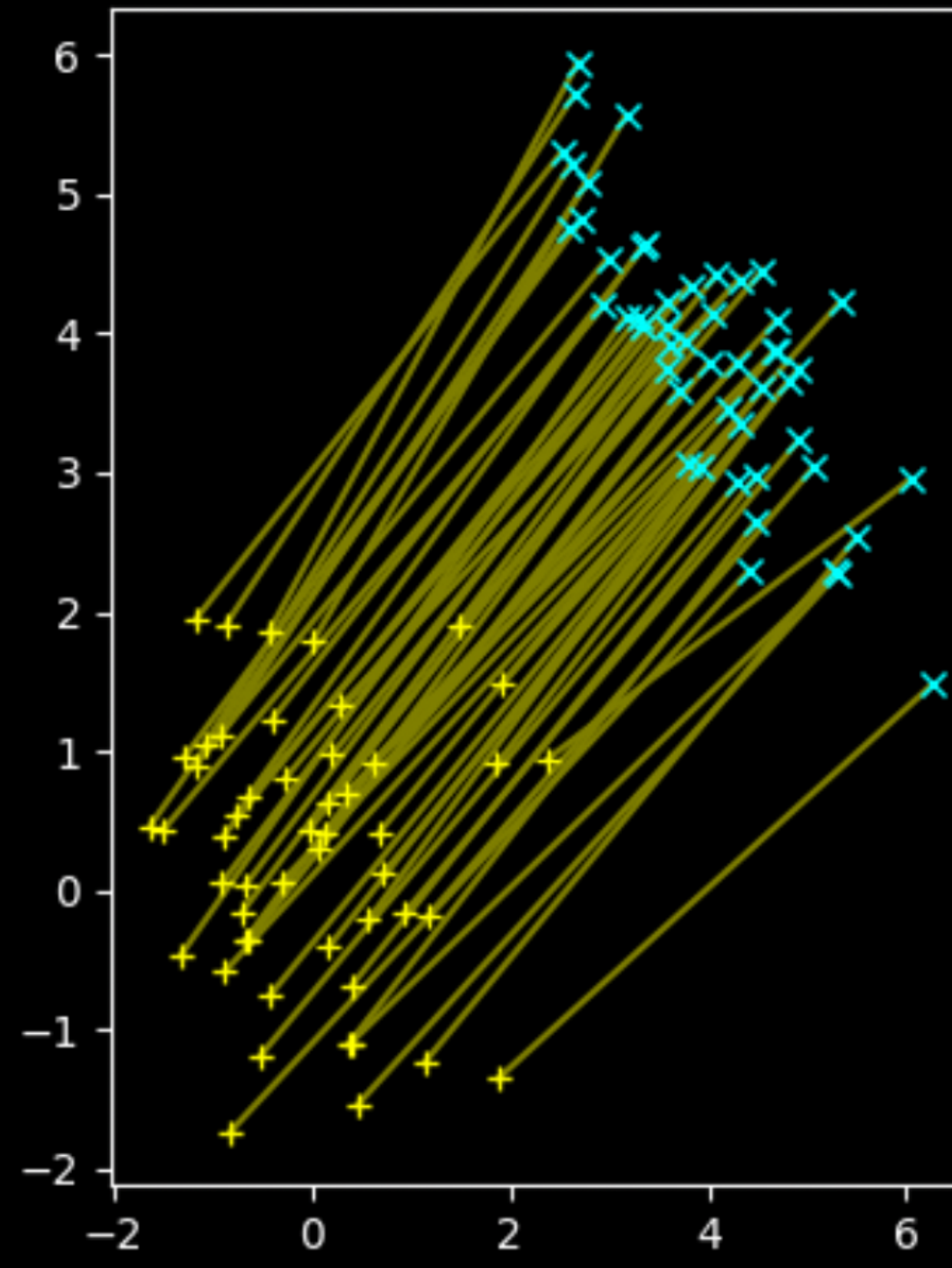


Alignment via optimal transport

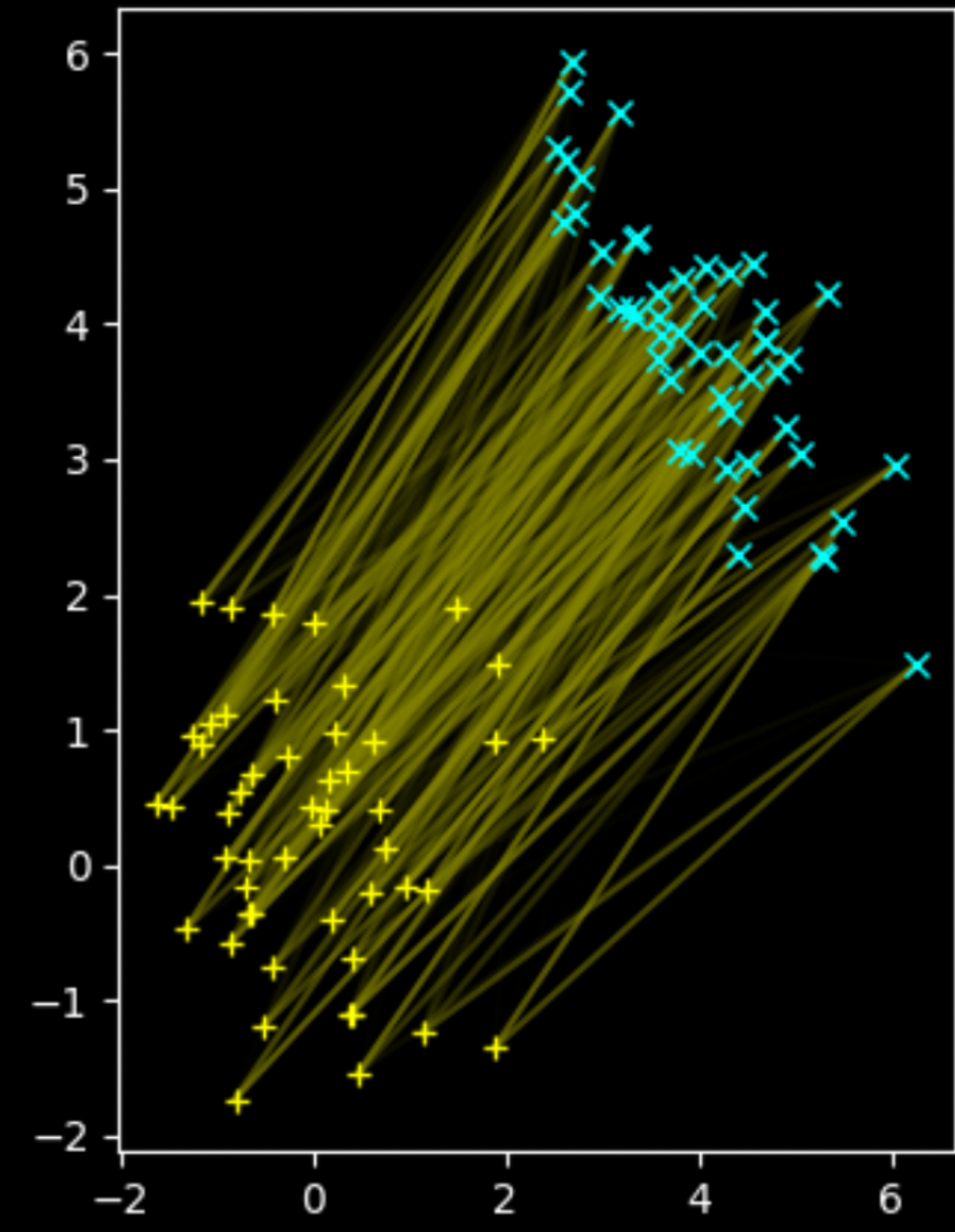
Source and target distributions



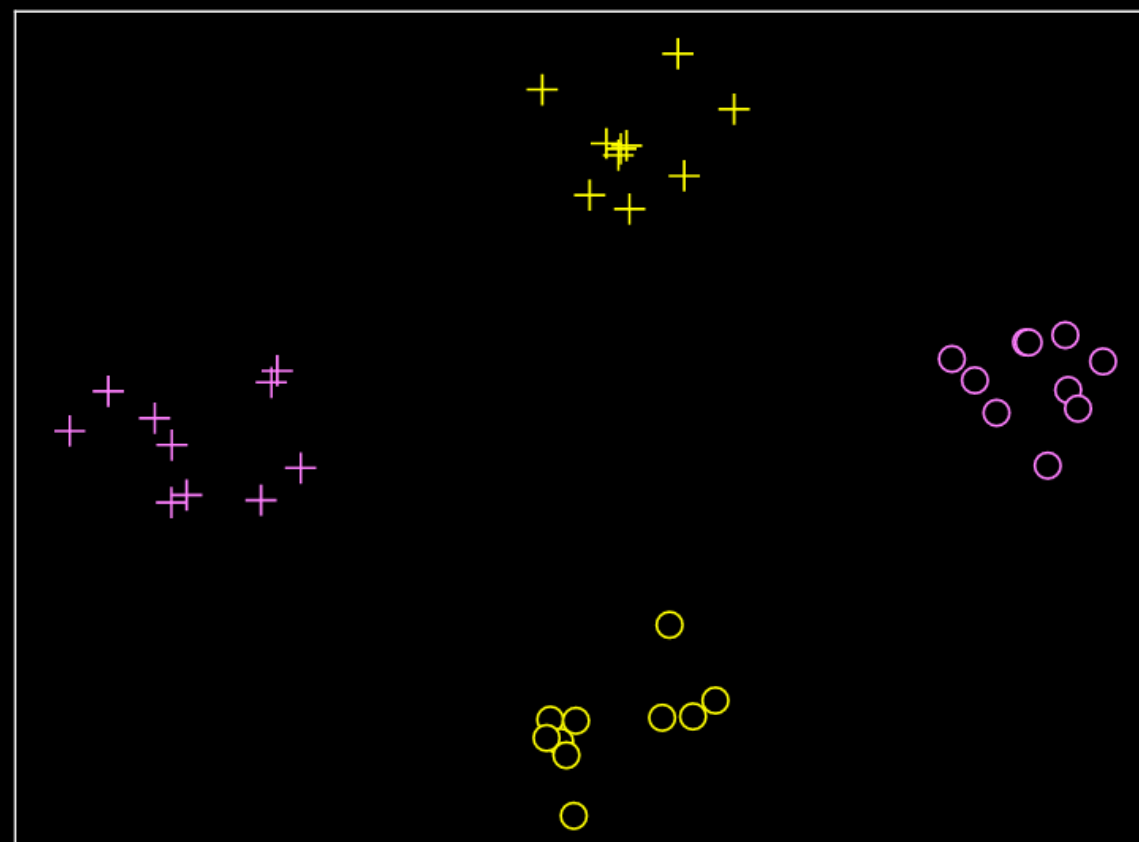
Exact solution



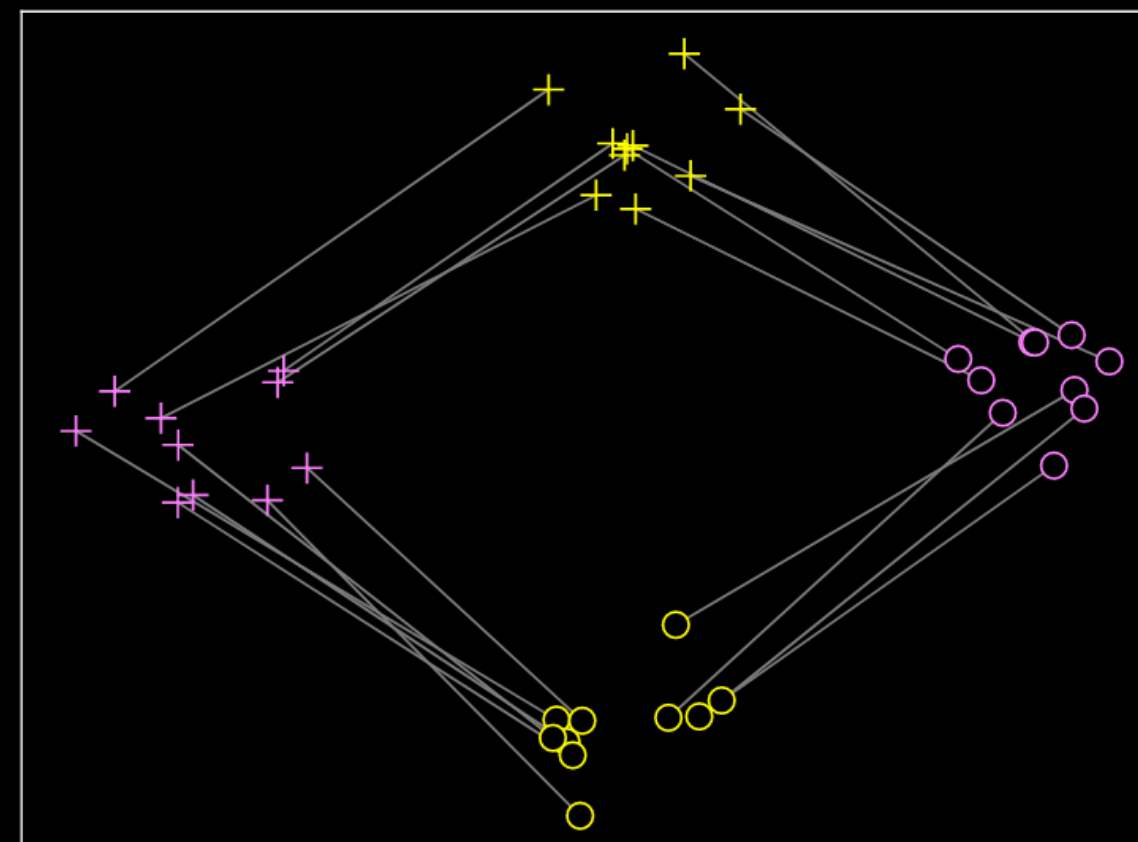
Weak solution



Alignment via optimal transport

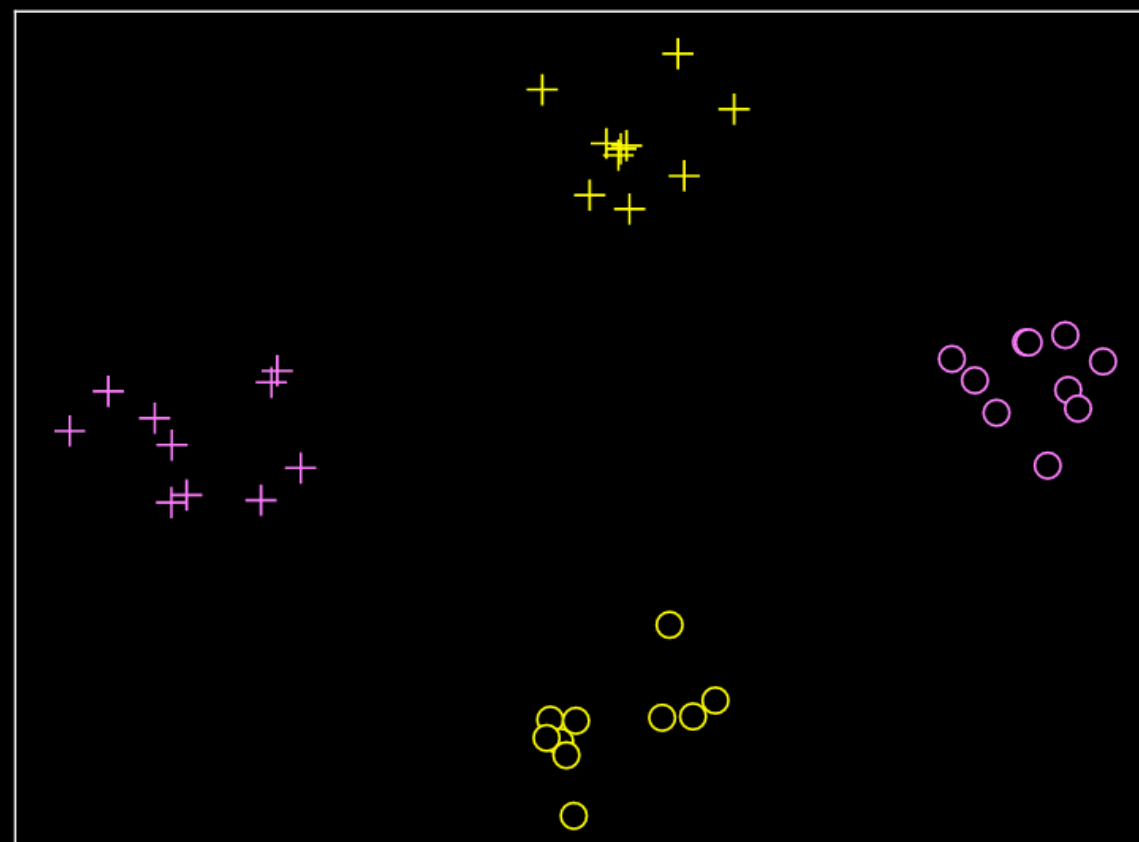


(a) Samples

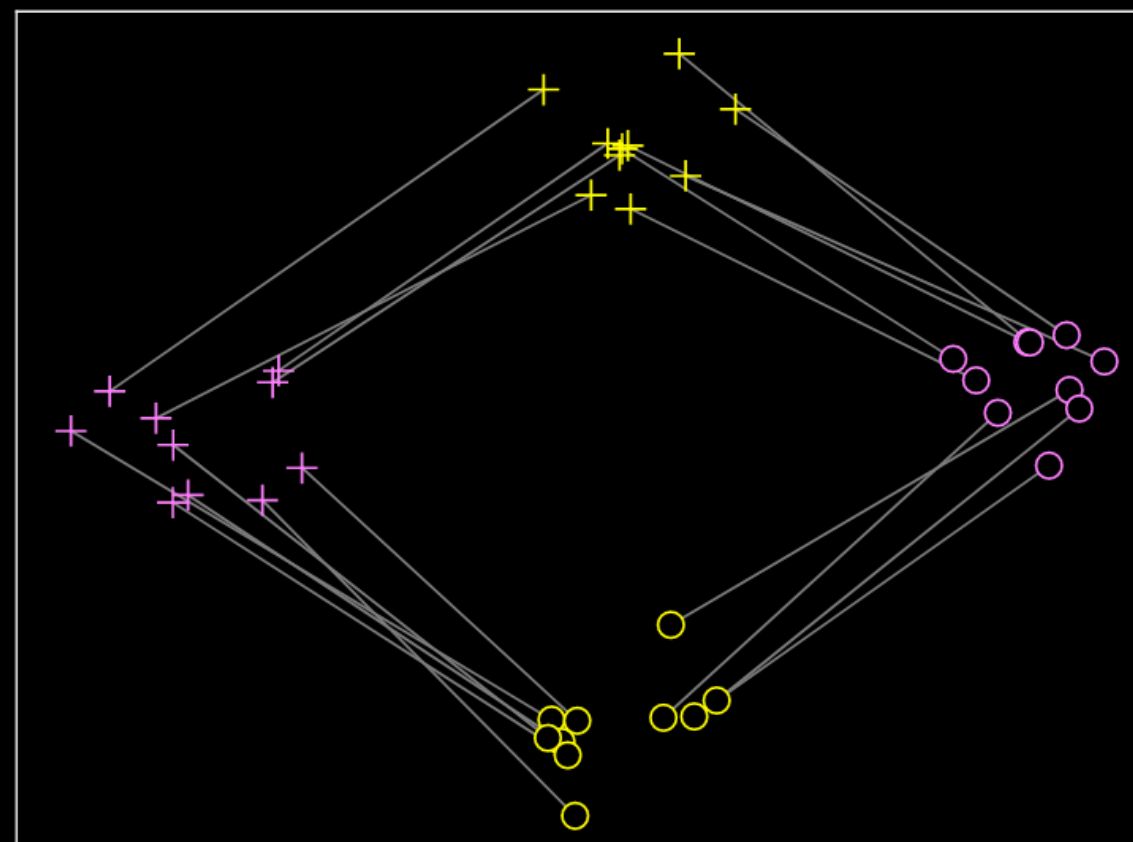


(b) Exact

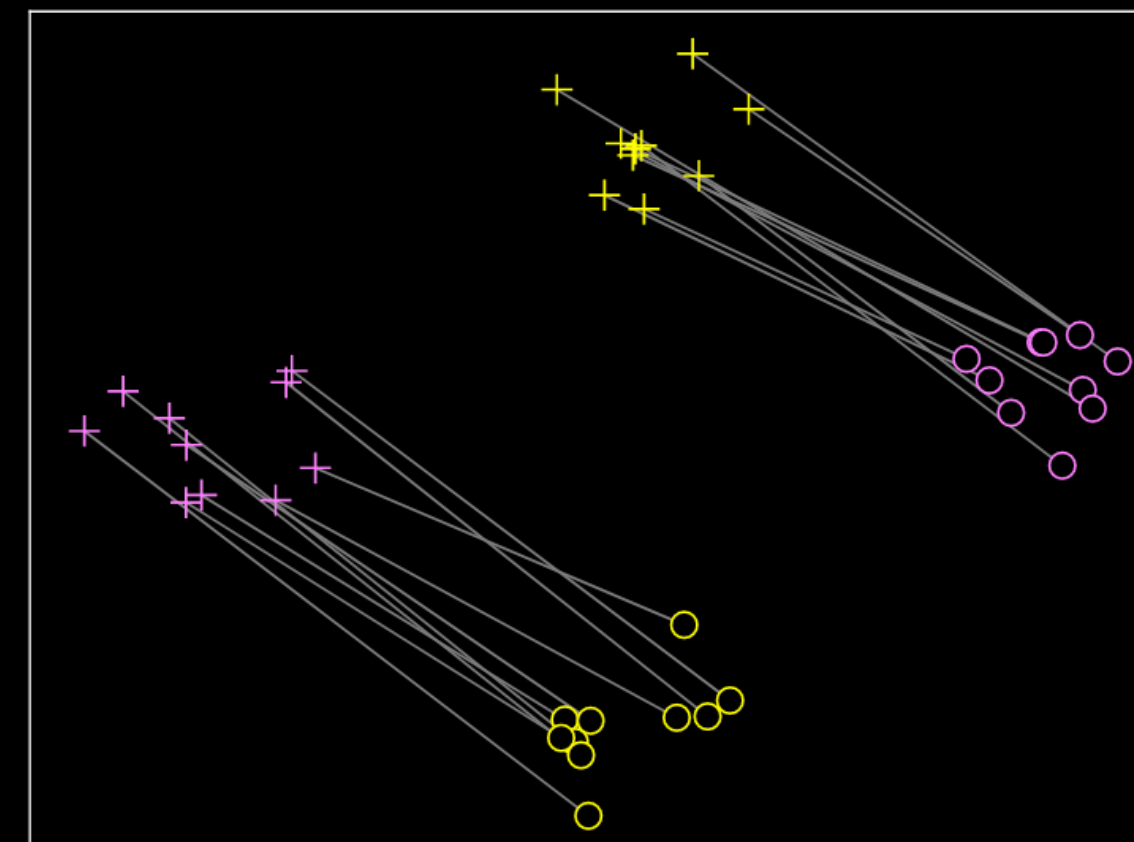
Alignment via optimal transport



(a) Samples

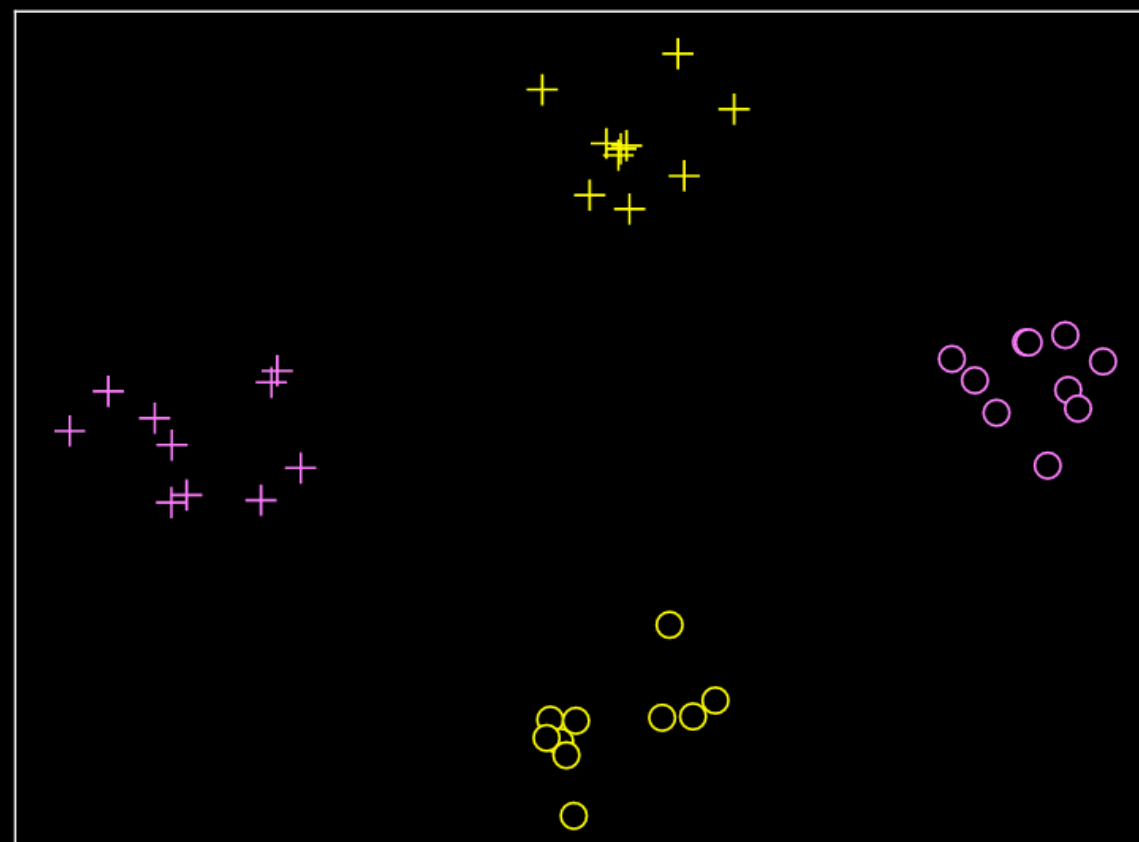


(b) Exact

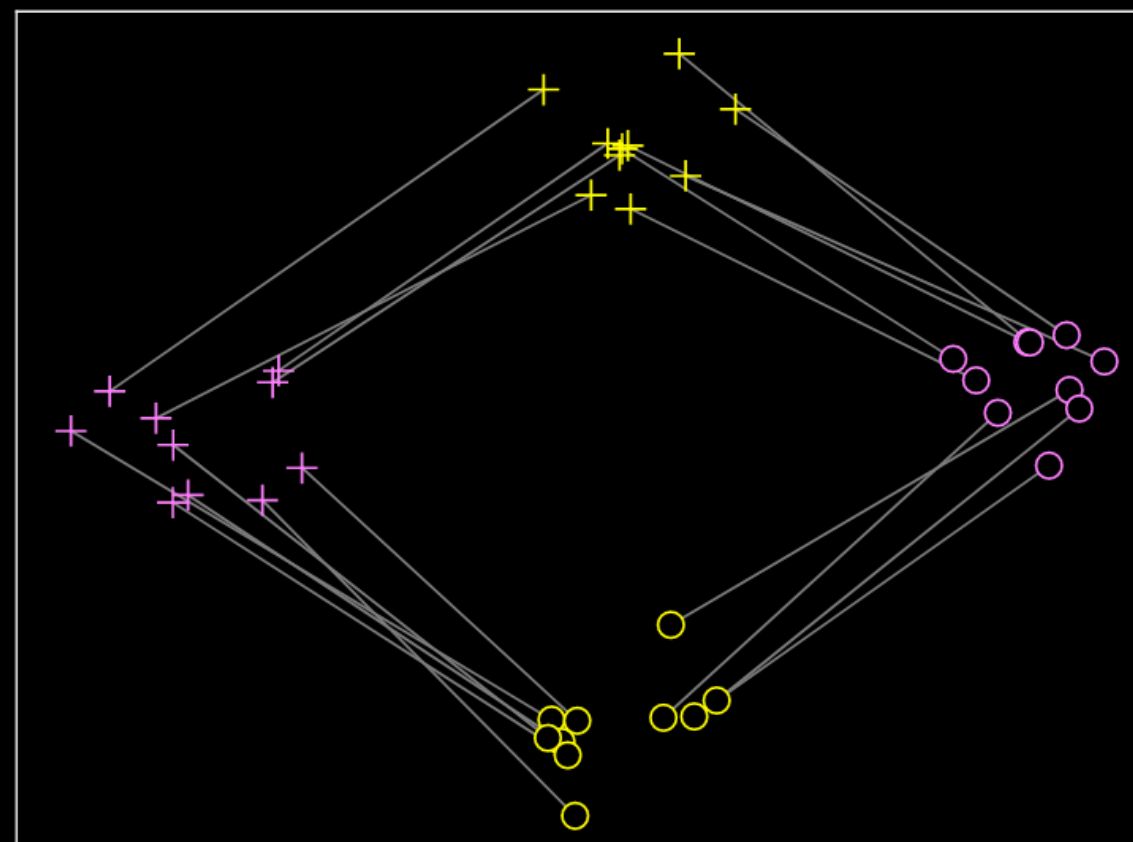


(c) Gromov-Wasserstein

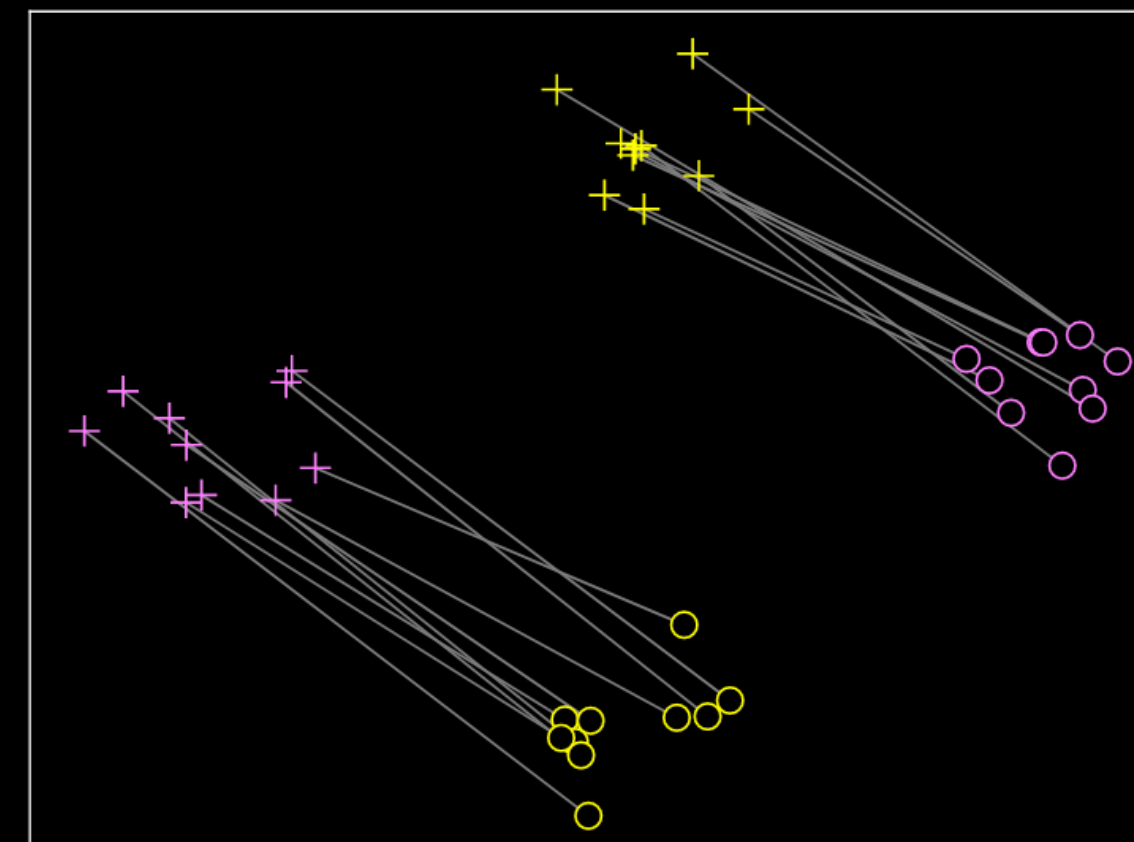
Alignment via optimal transport & key points



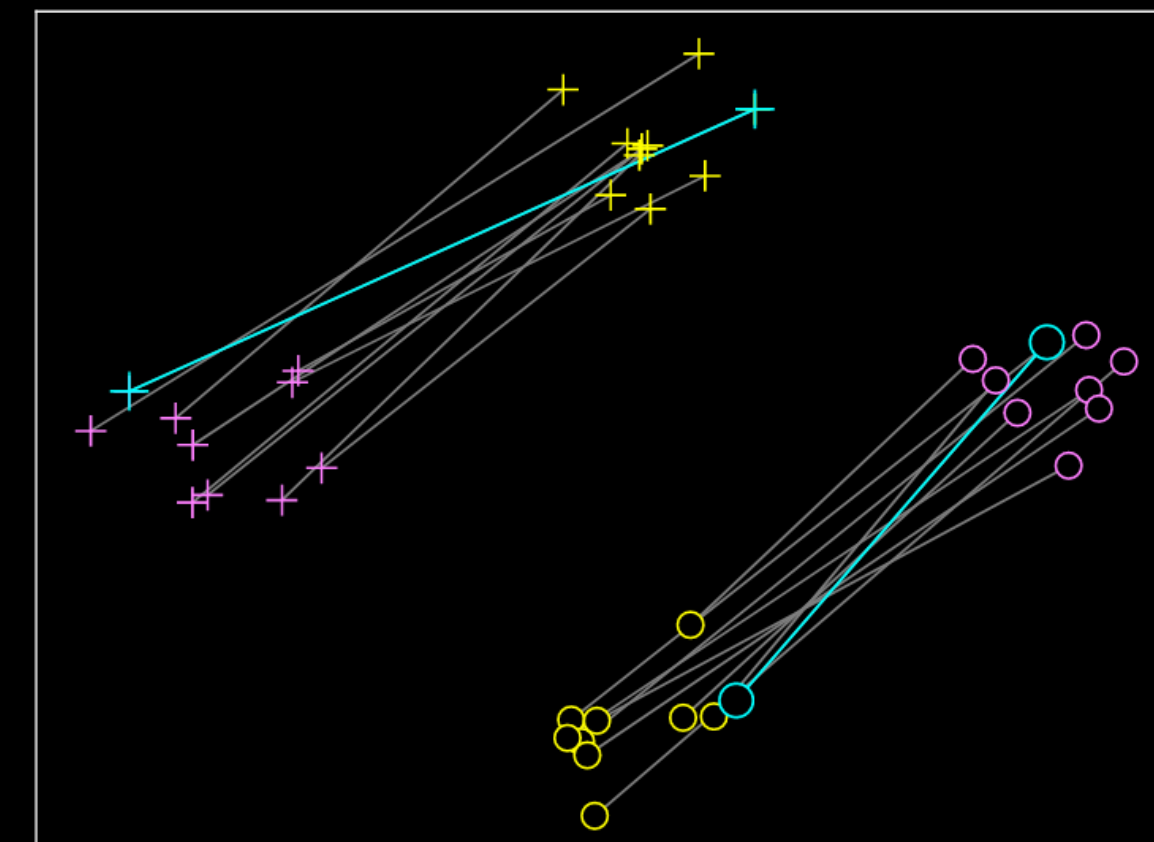
(a) Samples



(b) Exact

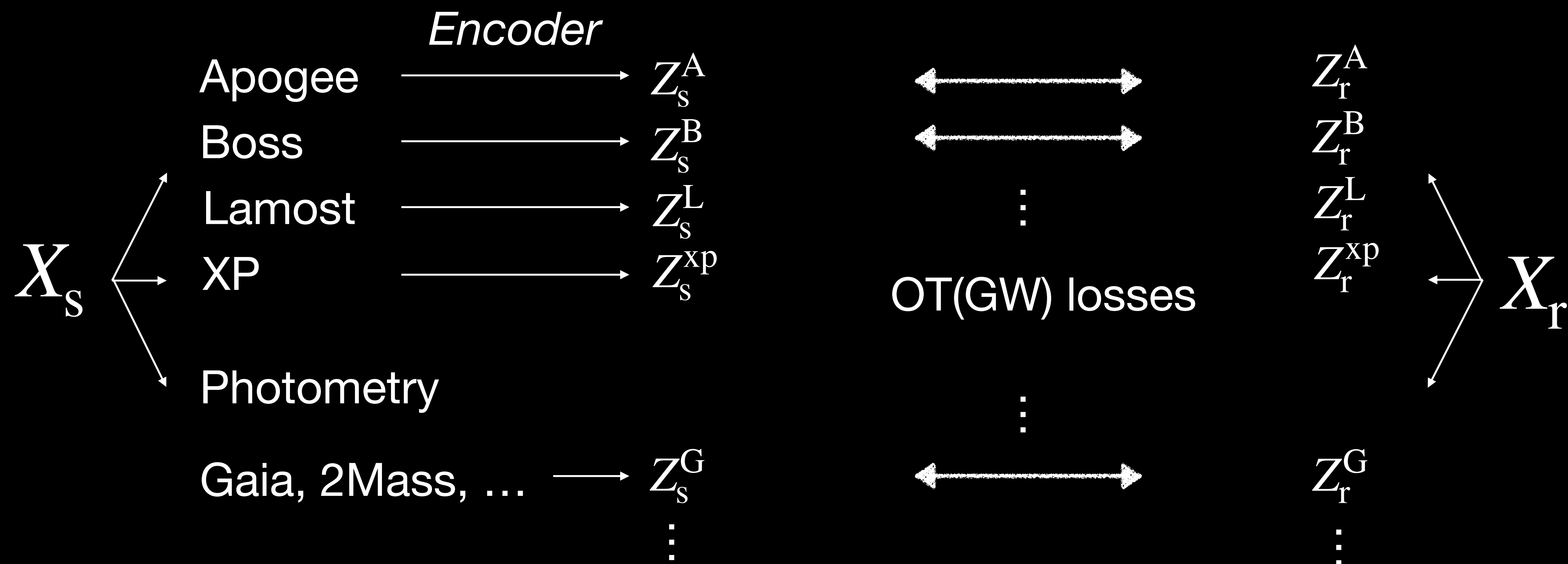


(c) Gromov-Wasserstein

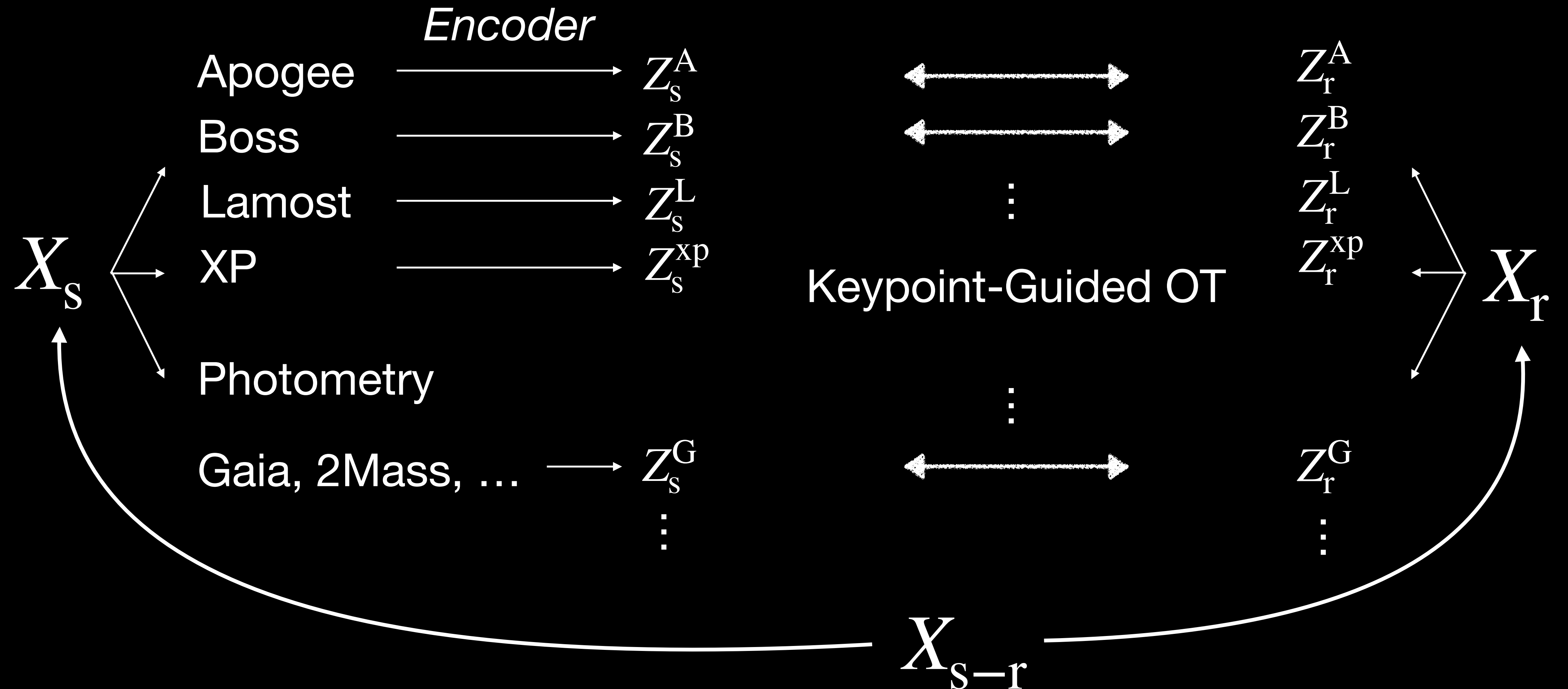


Gu et al. (2022)

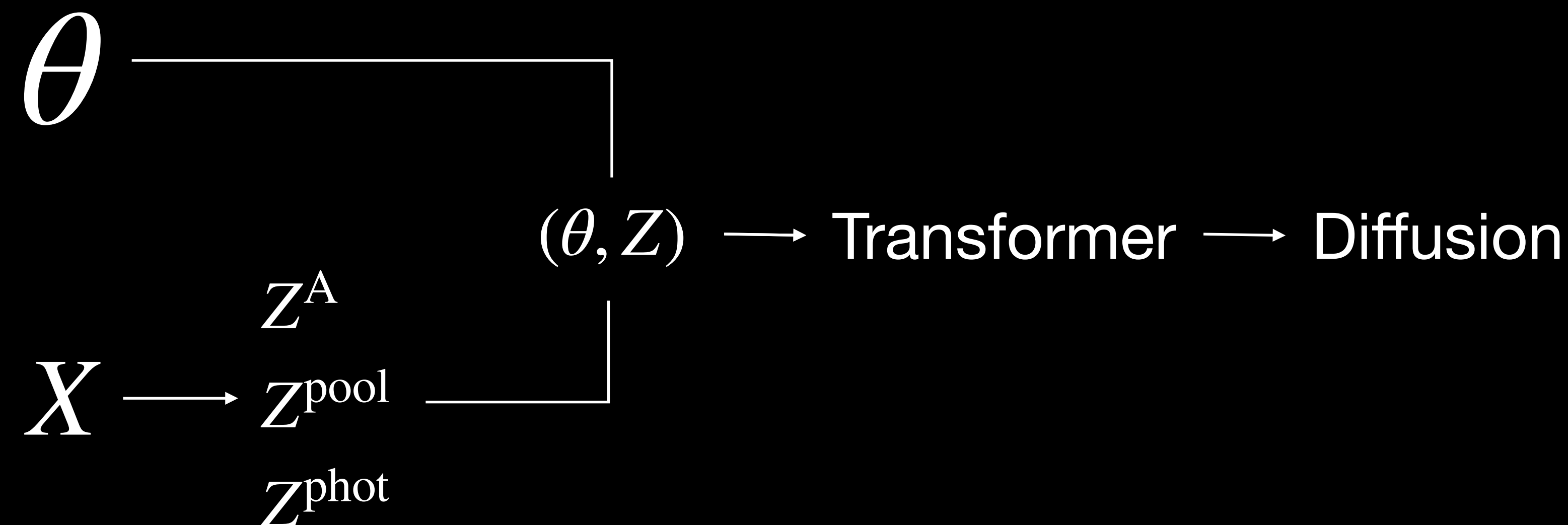
Modality encoders: alignment loss



Additional pair constraint



Final model

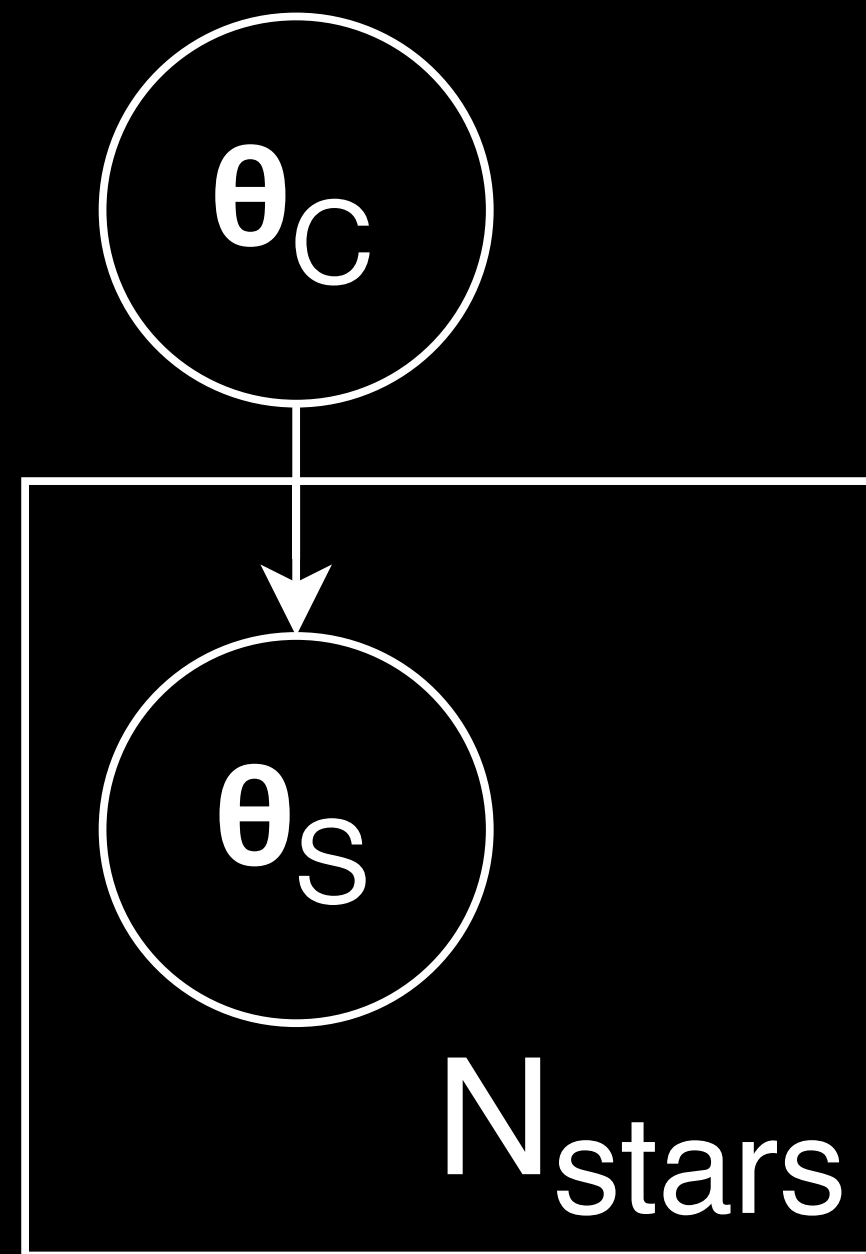


Loss

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

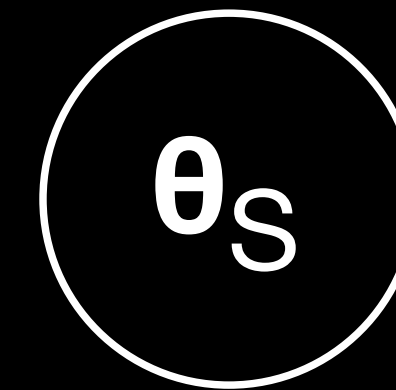
Forward model

Clusters/young stars

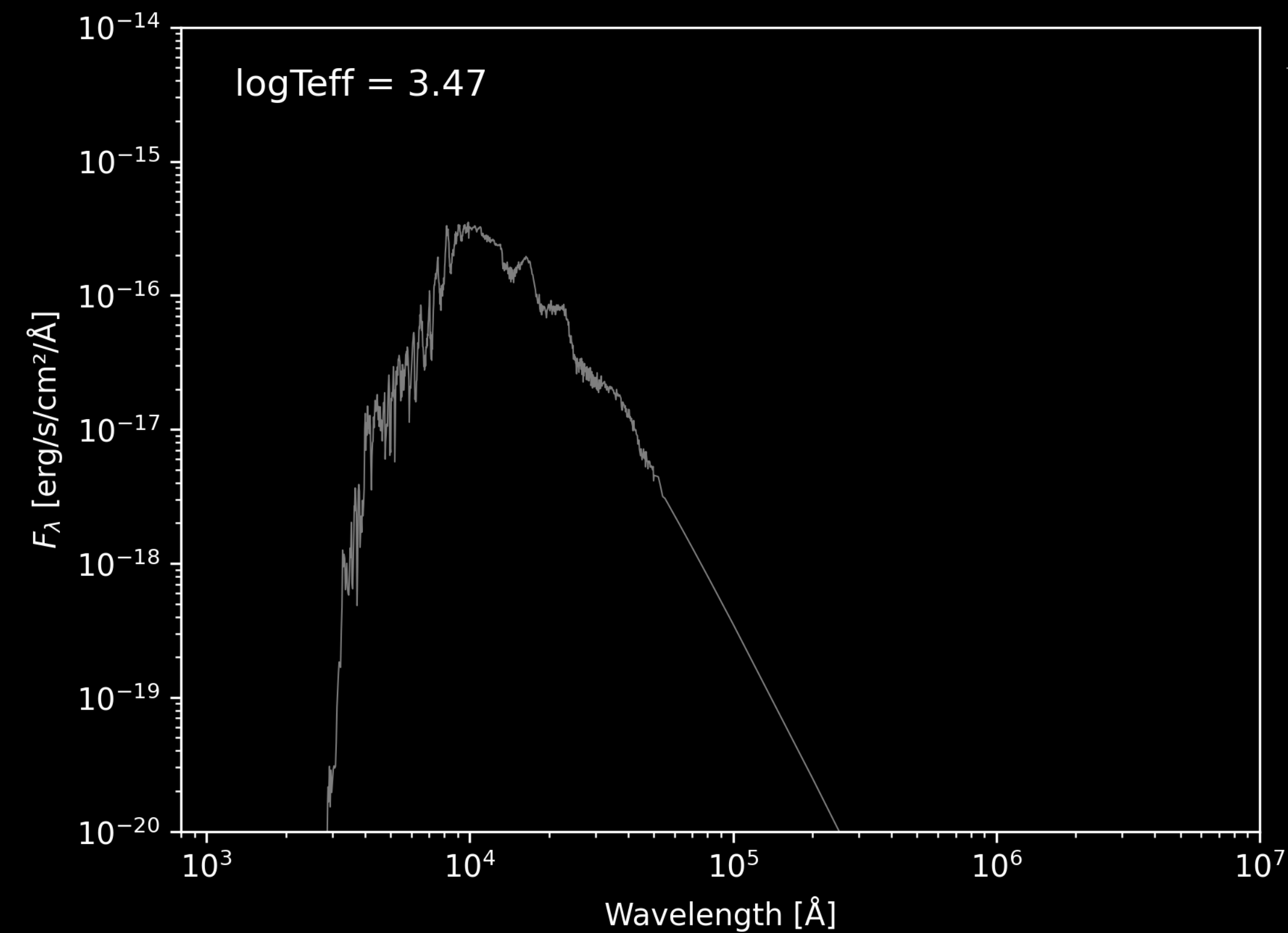
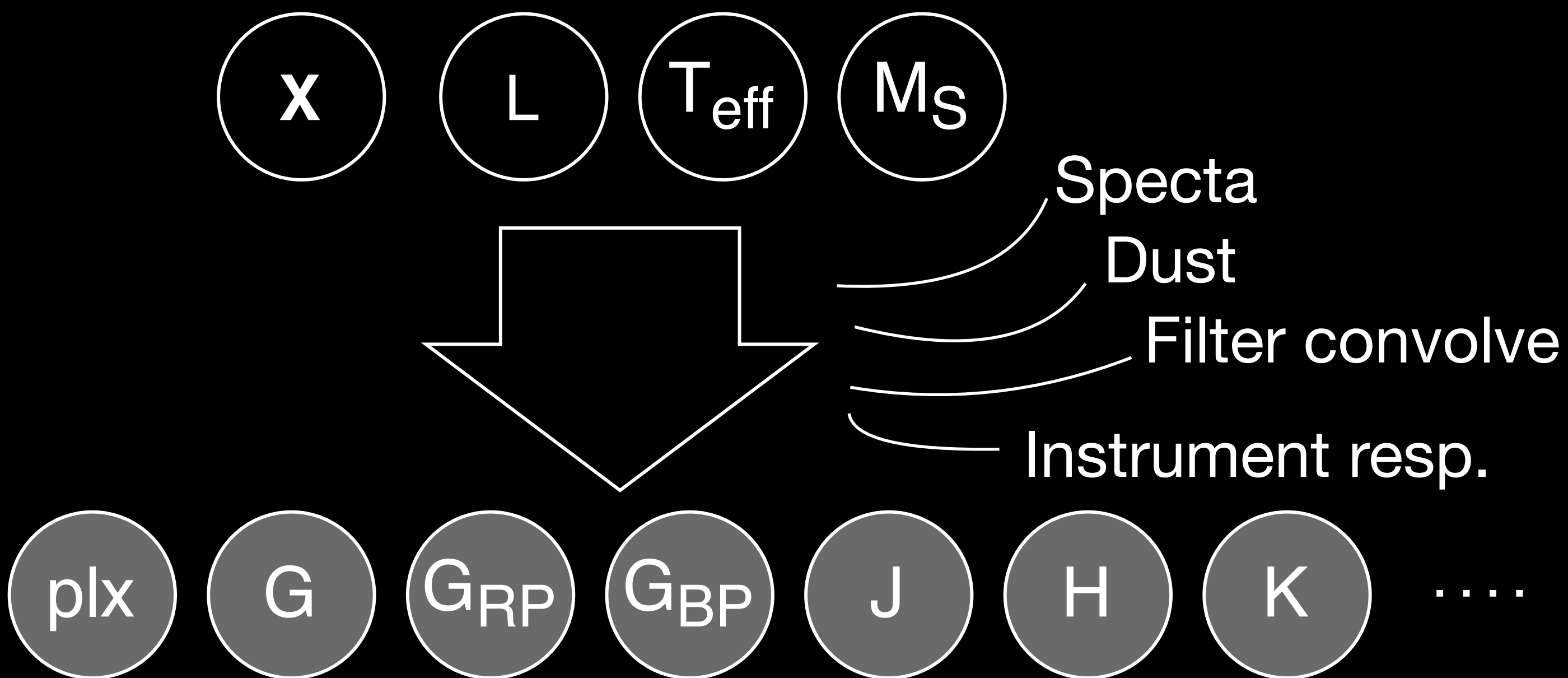


Milky Way model

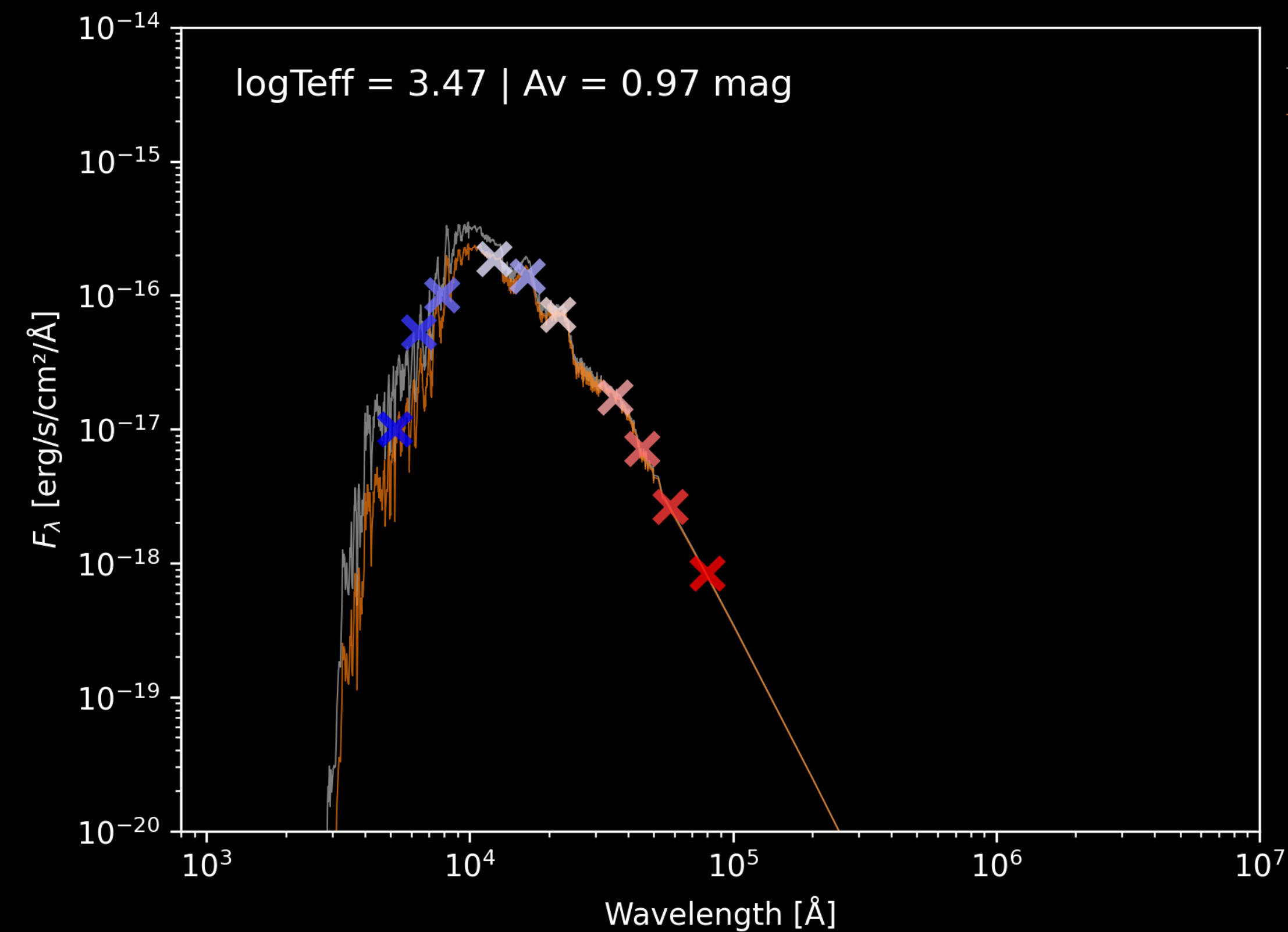
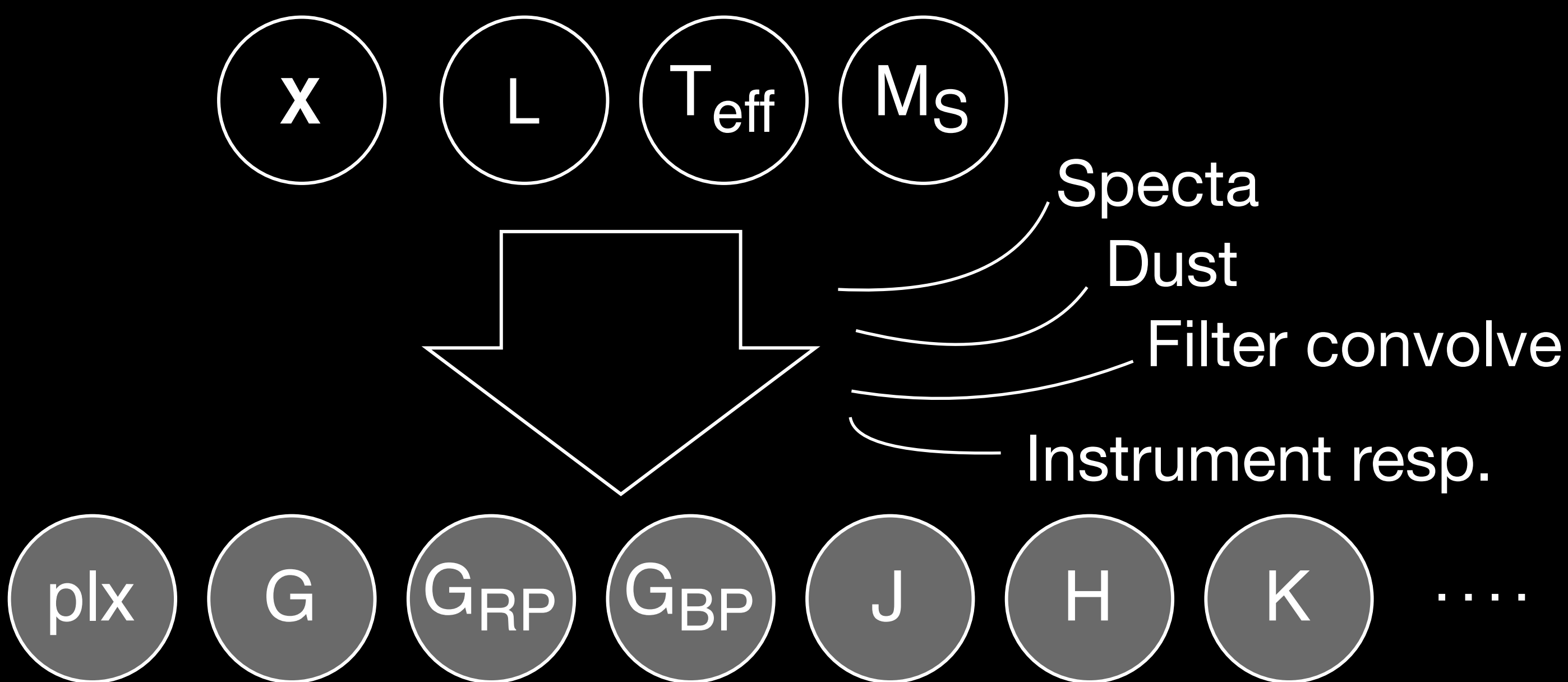
- *galaxia* code
- thin+thick+halo+bulge



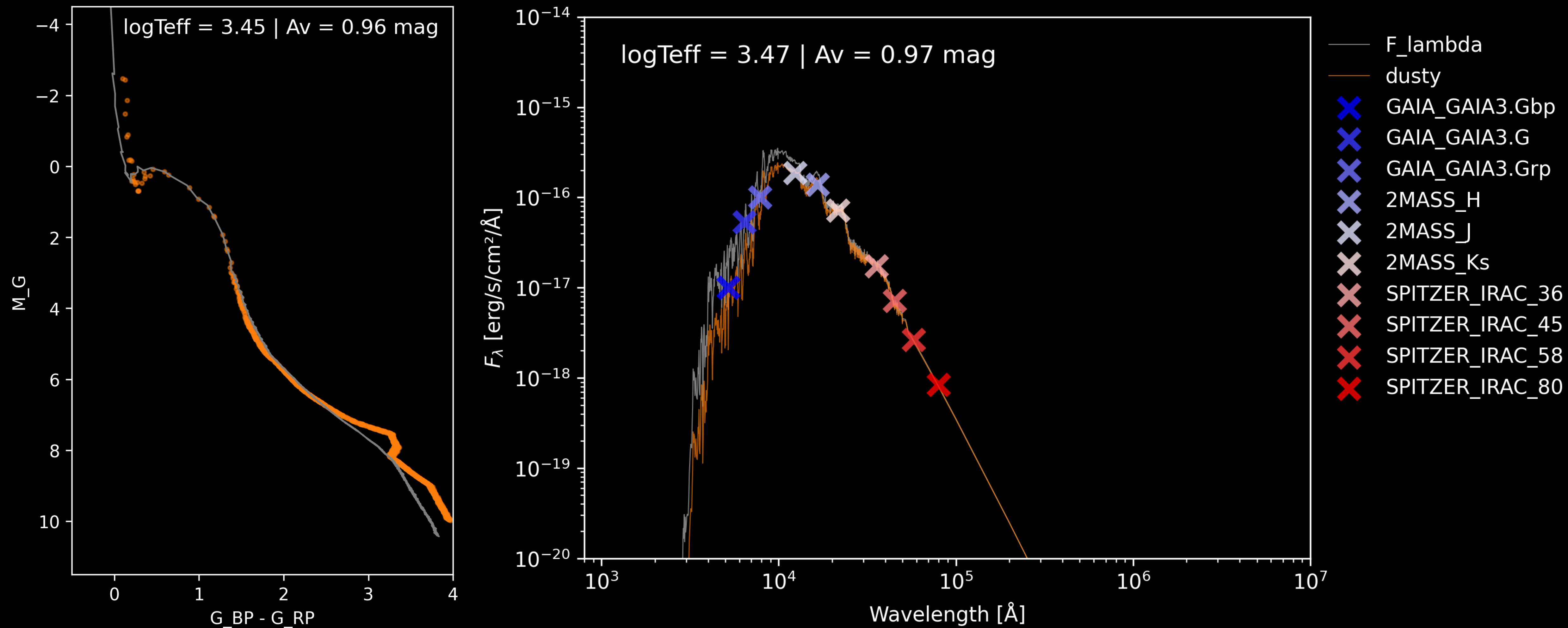
Fwd model



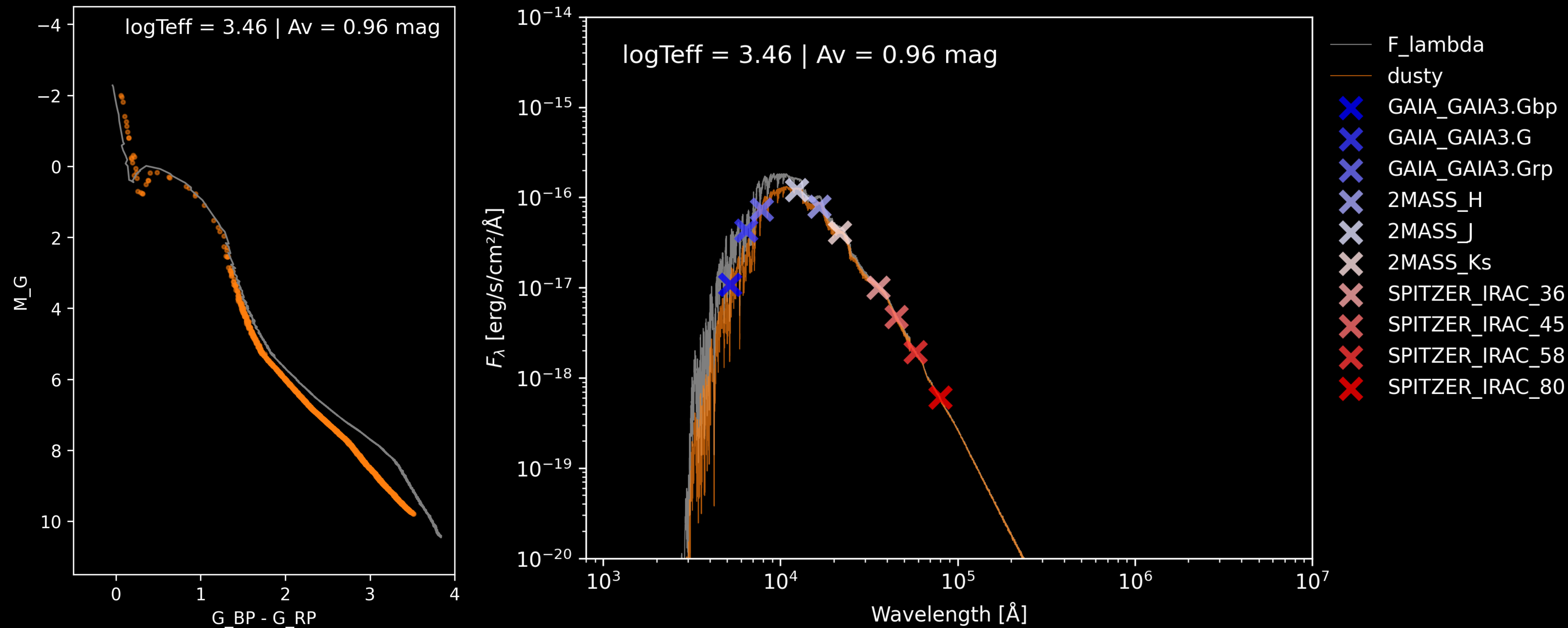
Fwd model



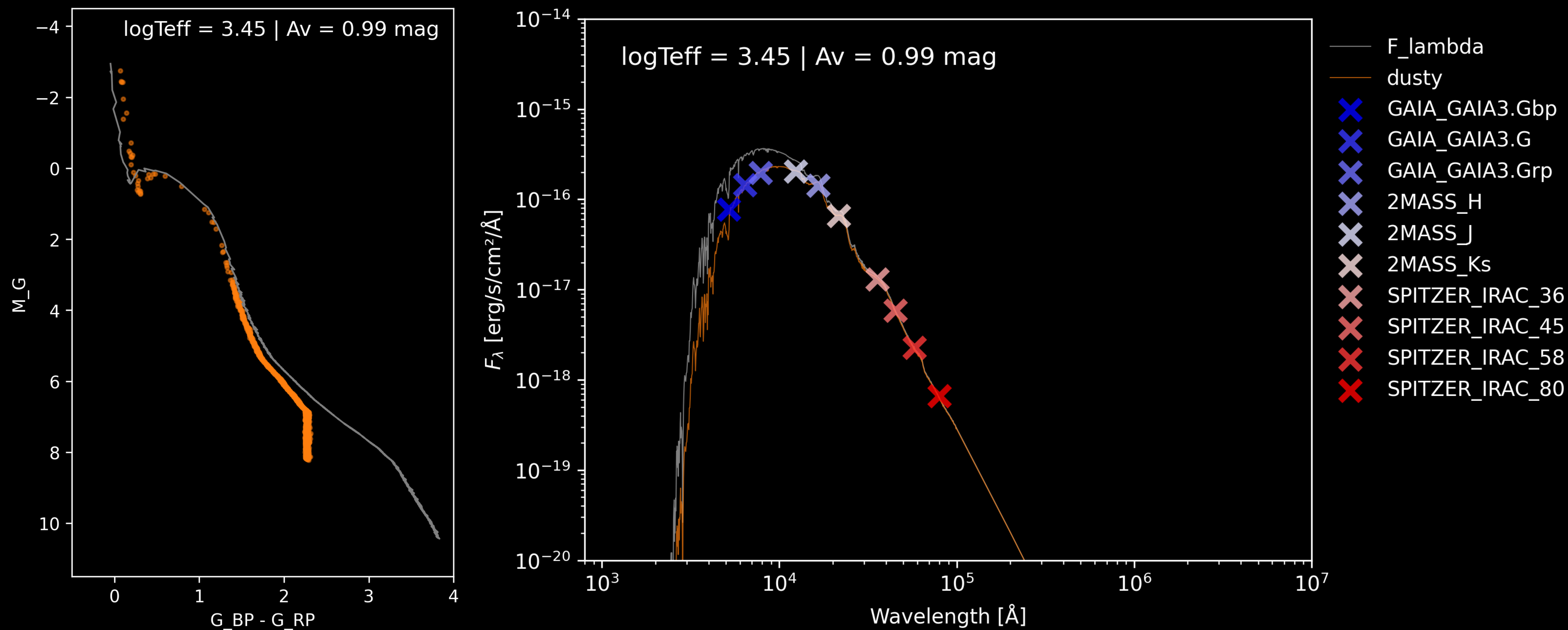
Model differences (BaSeL)



Model differences (BTSettl)



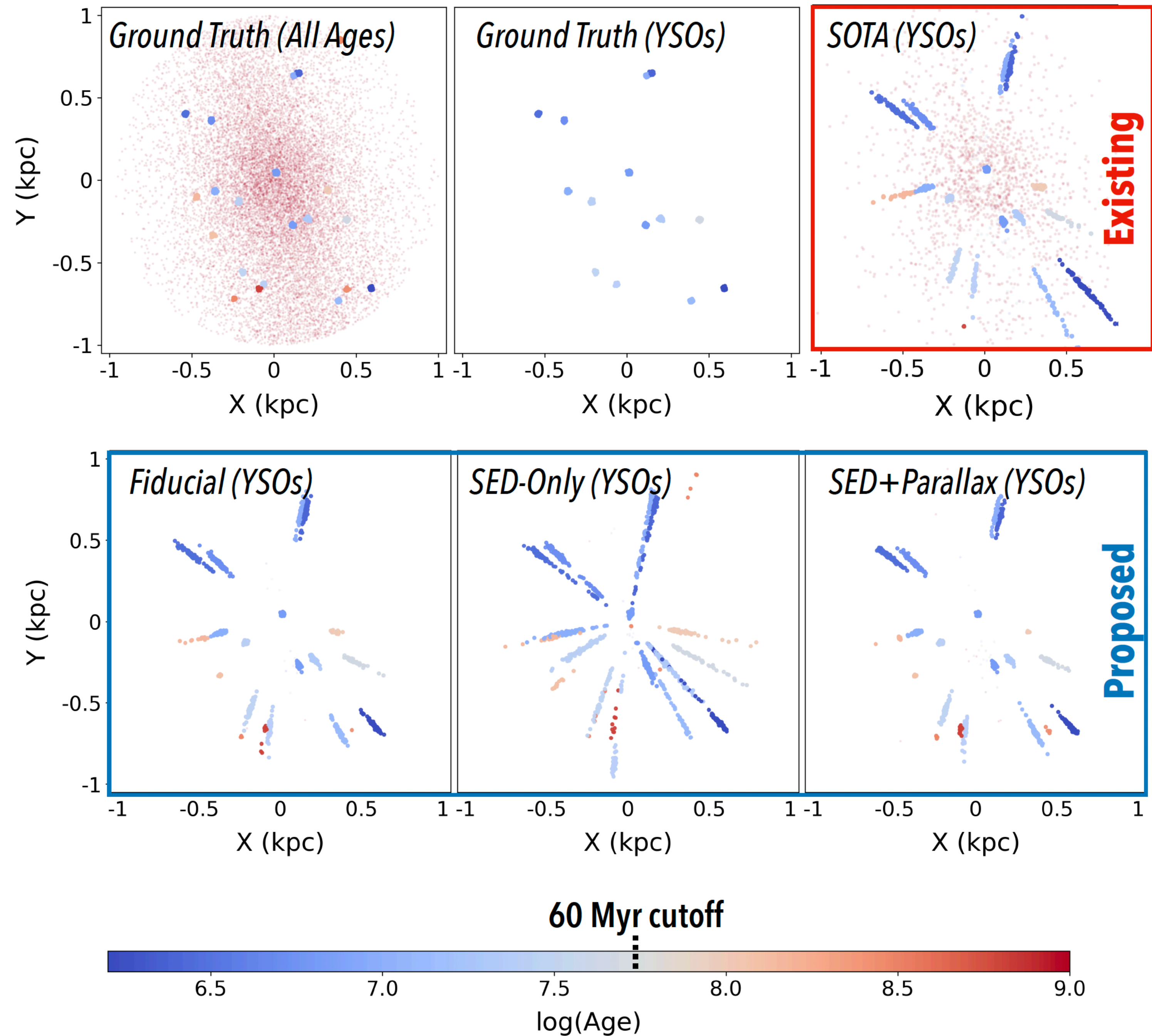
Model differences (Kurucz)



Results

Pilot study

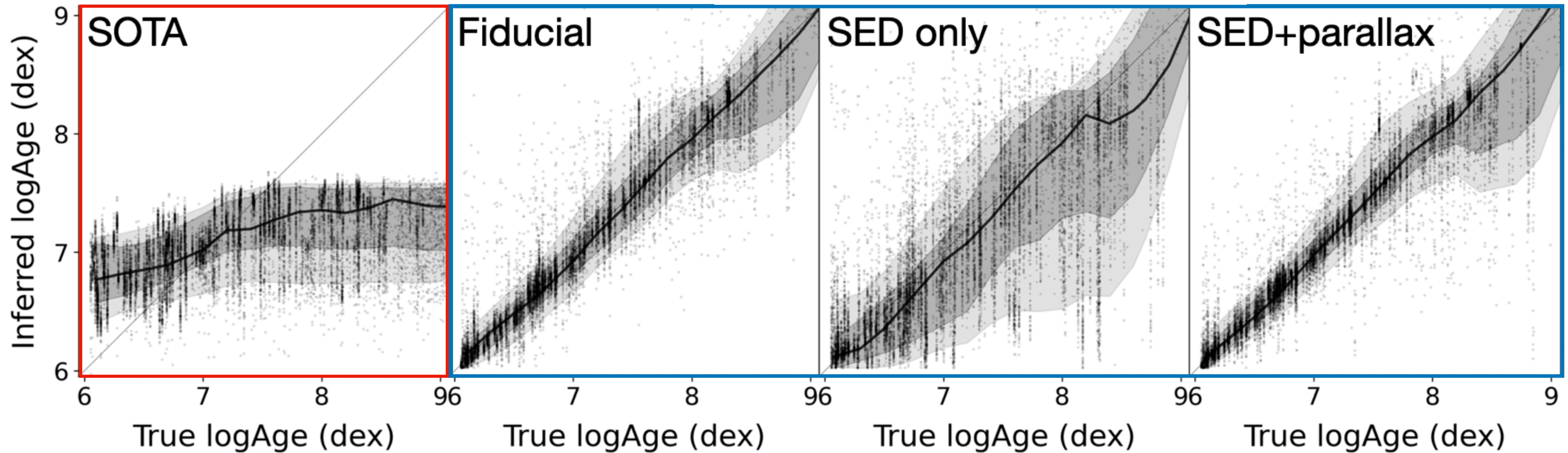
Train NPE on PARSEC
Gaia+2MASS+Wise
photometry + plx + A_v



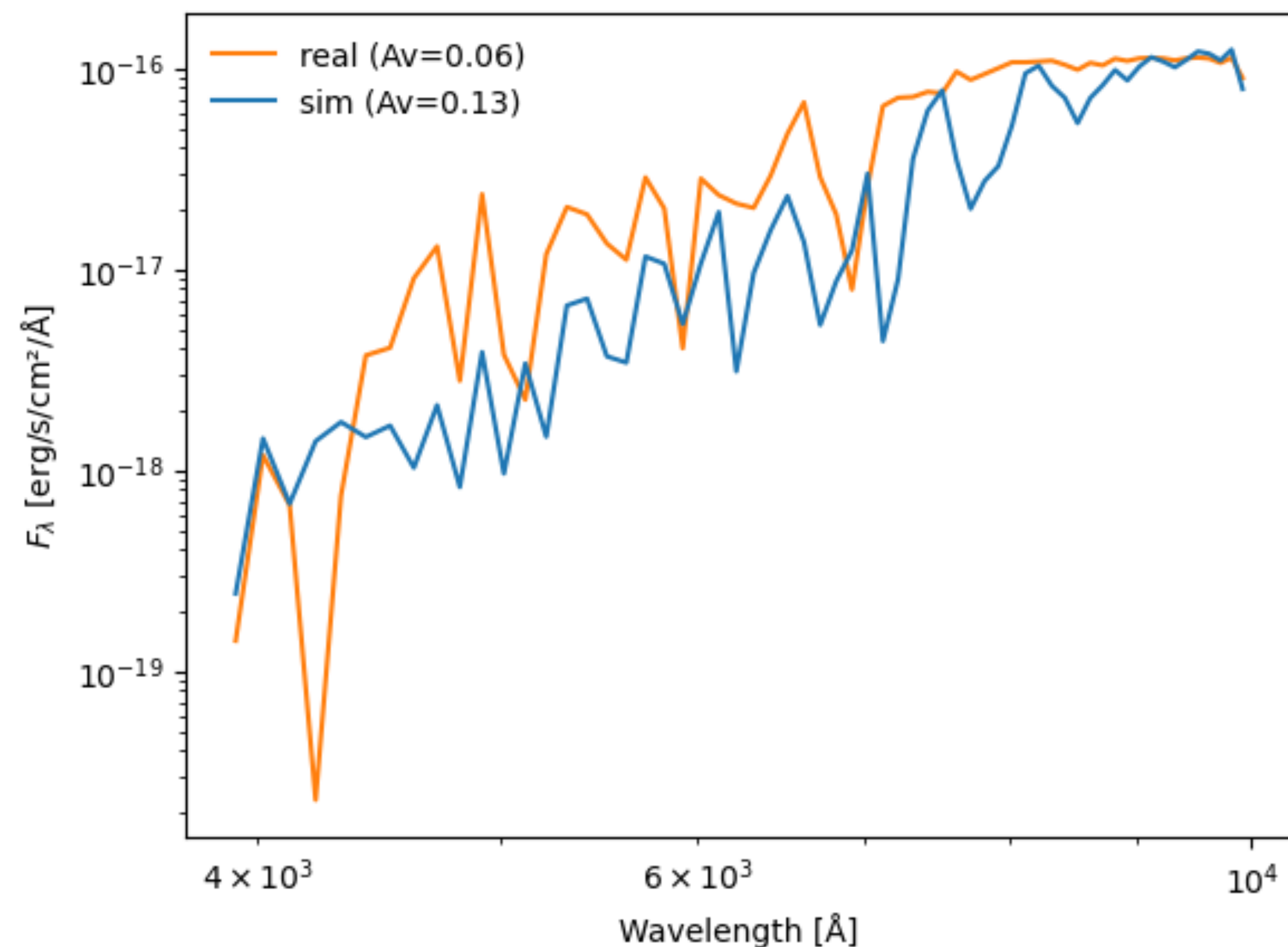
Pilot study: NPE

Existing

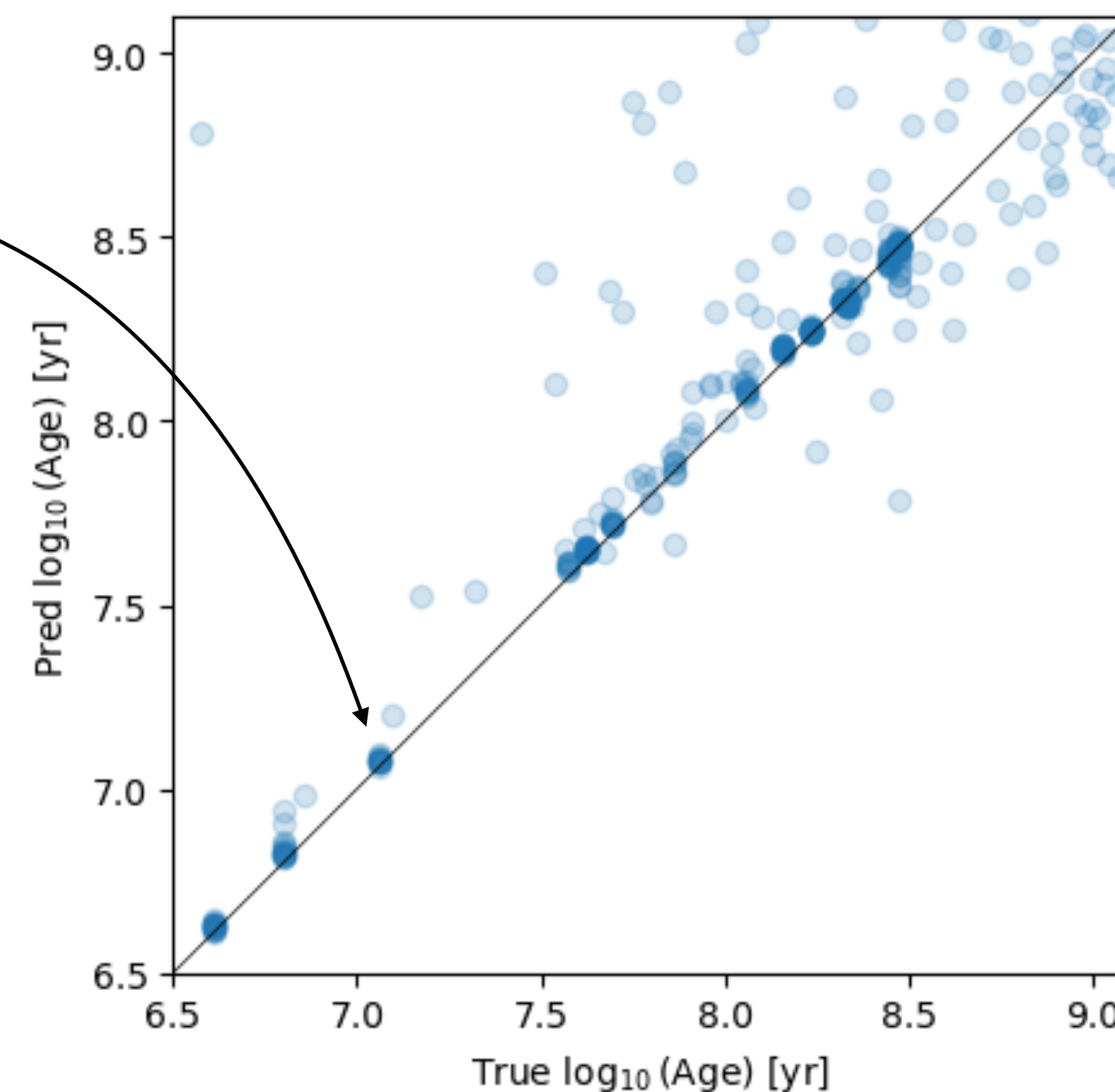
Classic SBI, no missing data



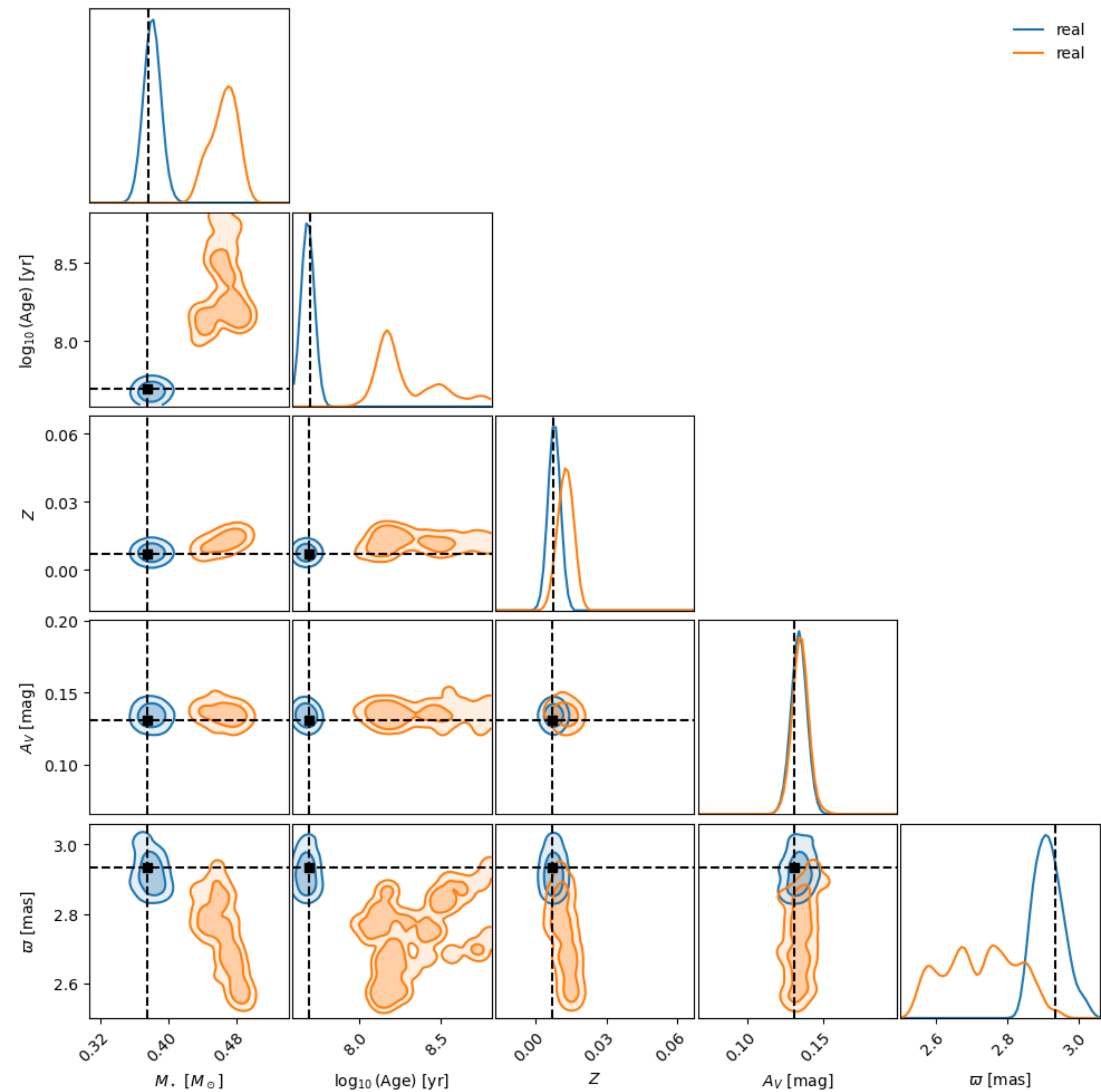
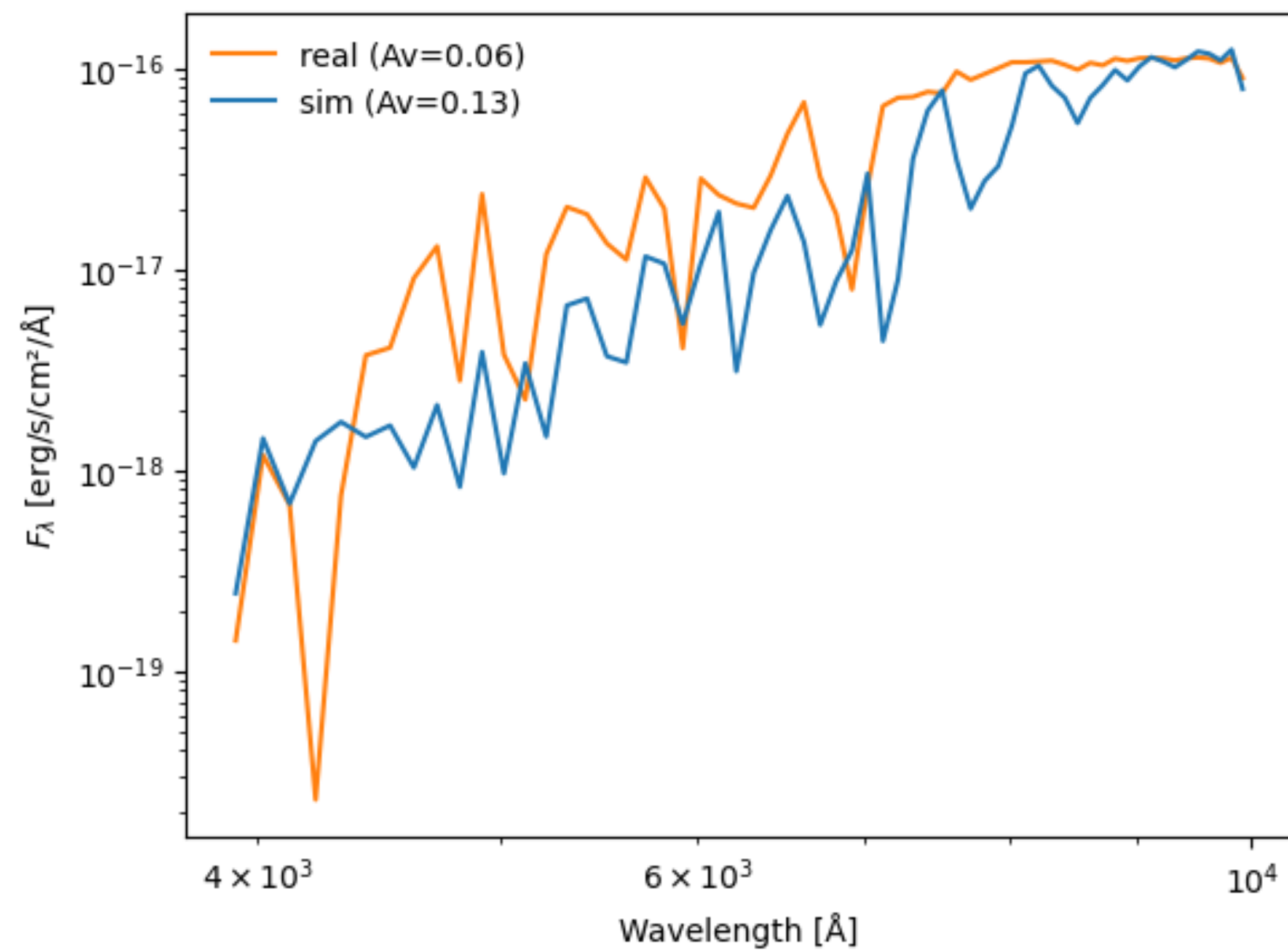
Updated pipeline + 50% missing + XP spectra



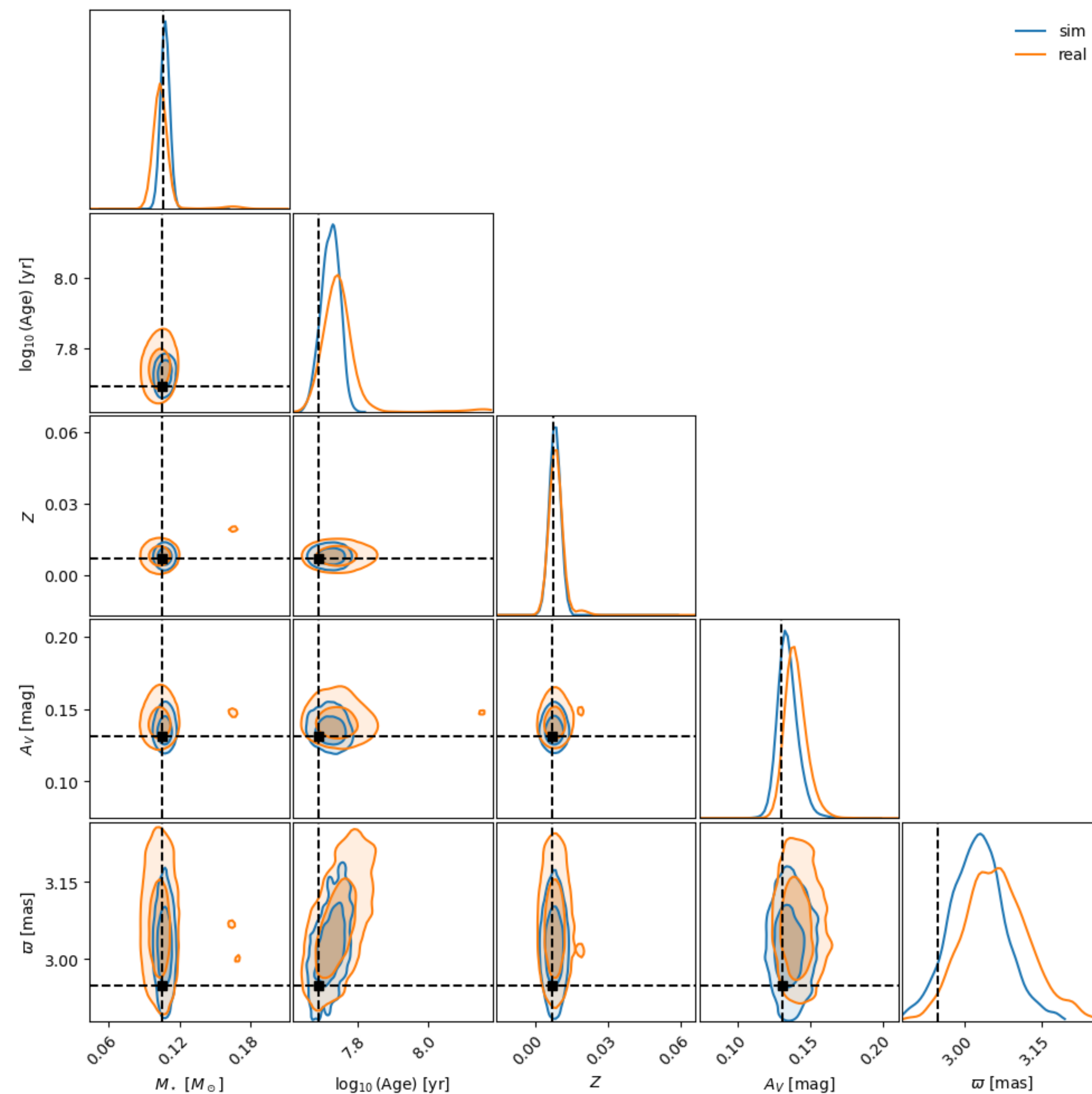
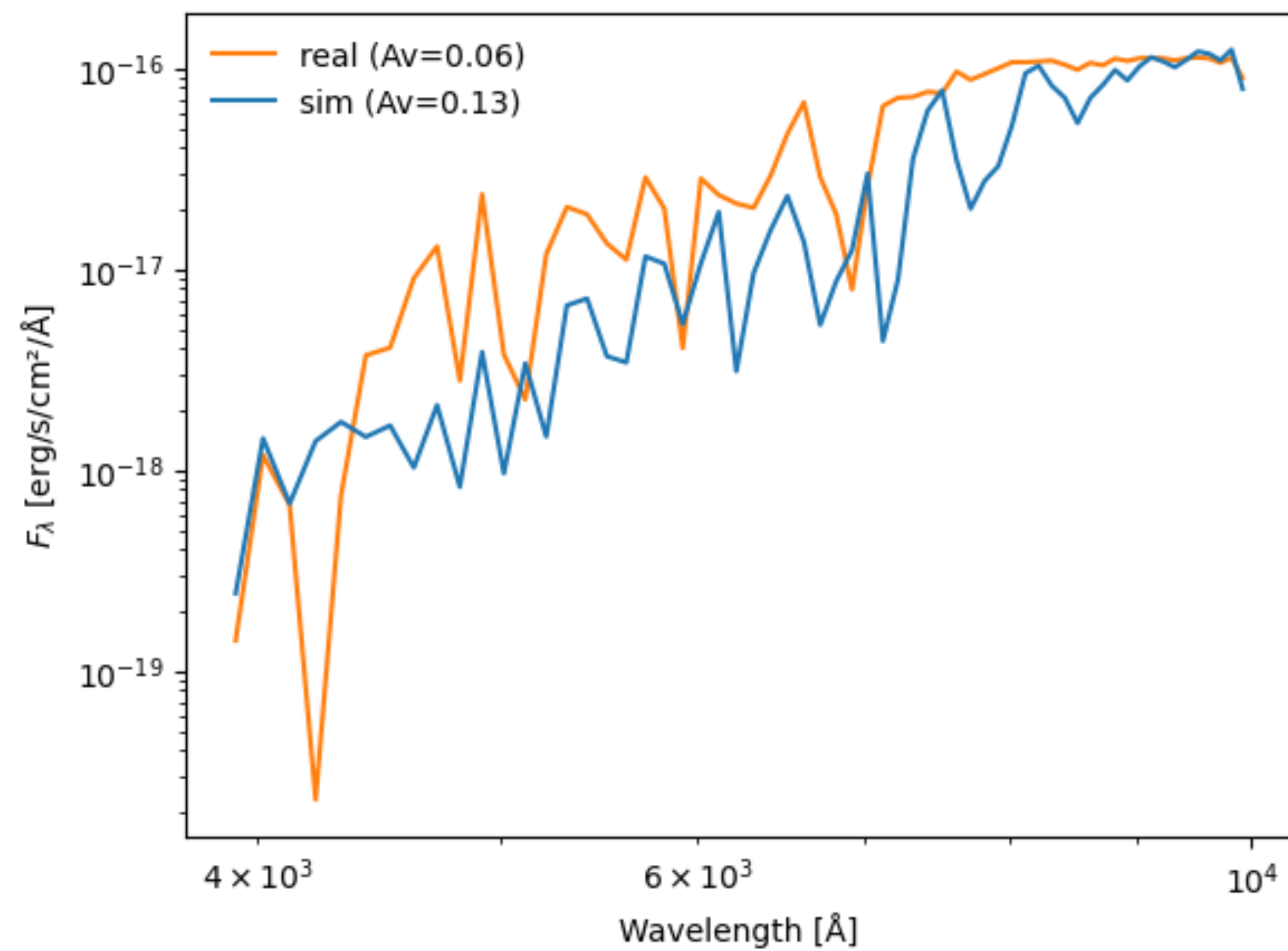
Posterior mean **sim**



Without DA

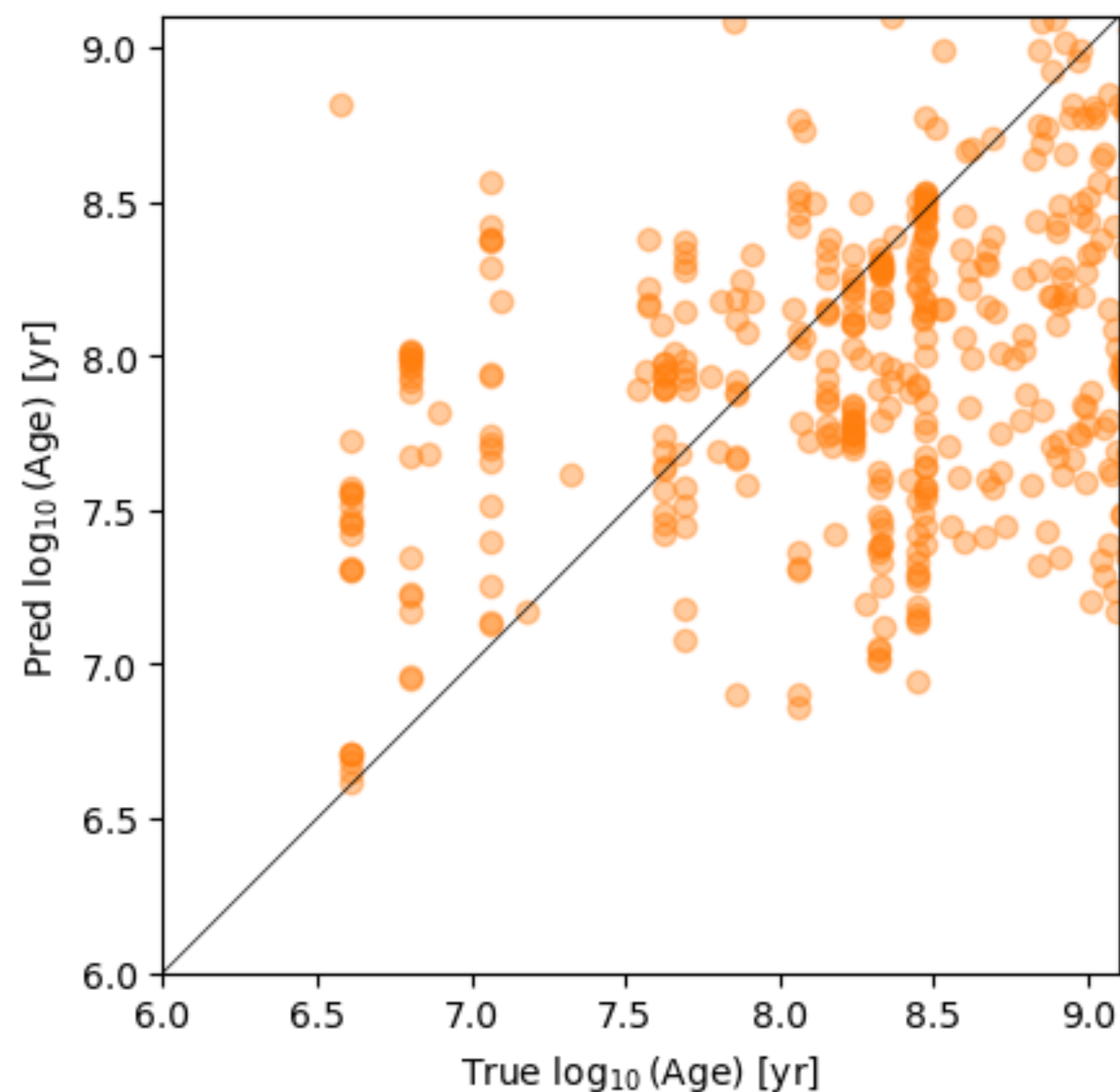


With DA



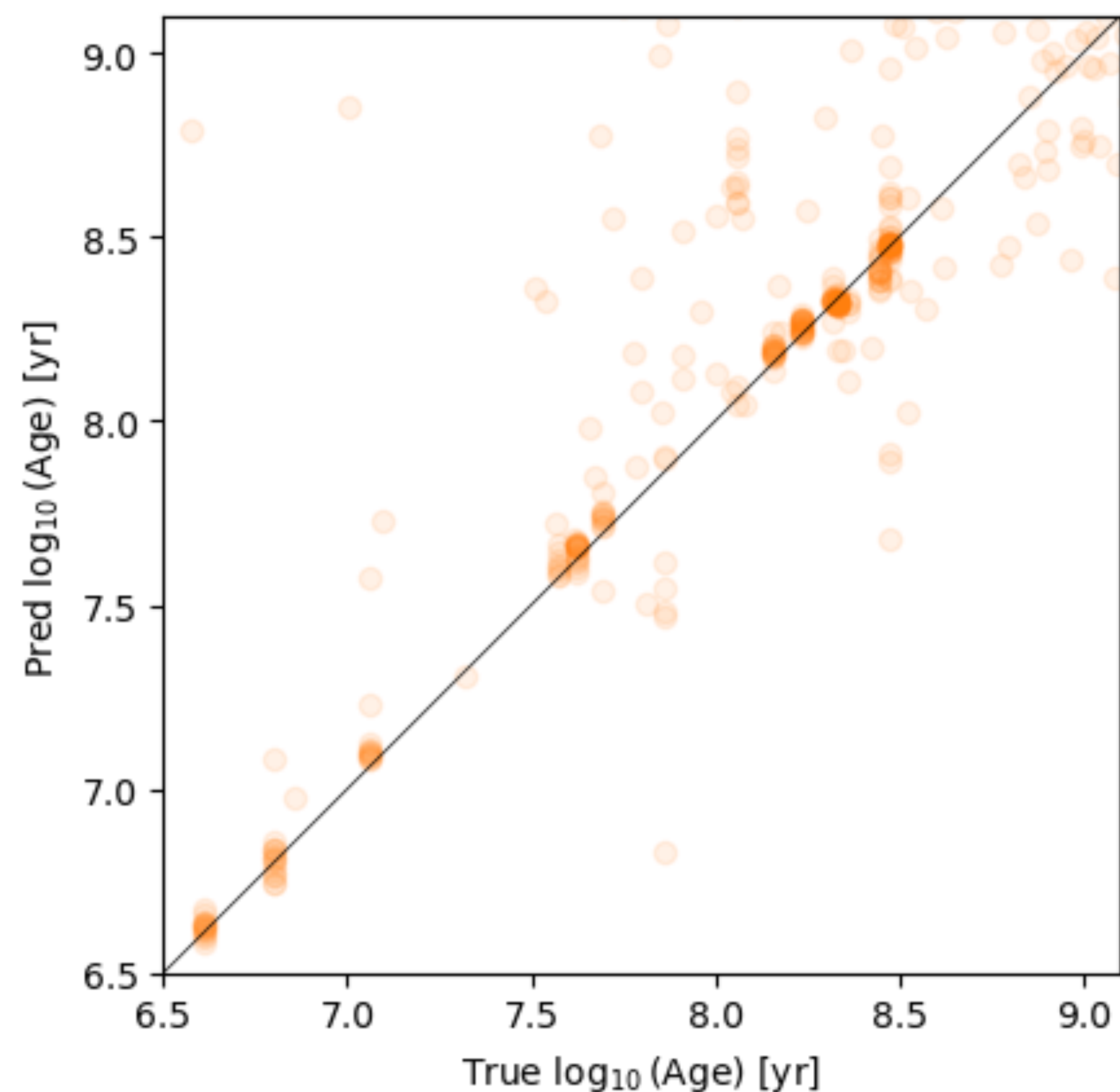
Predictions

Posterior mean **real**



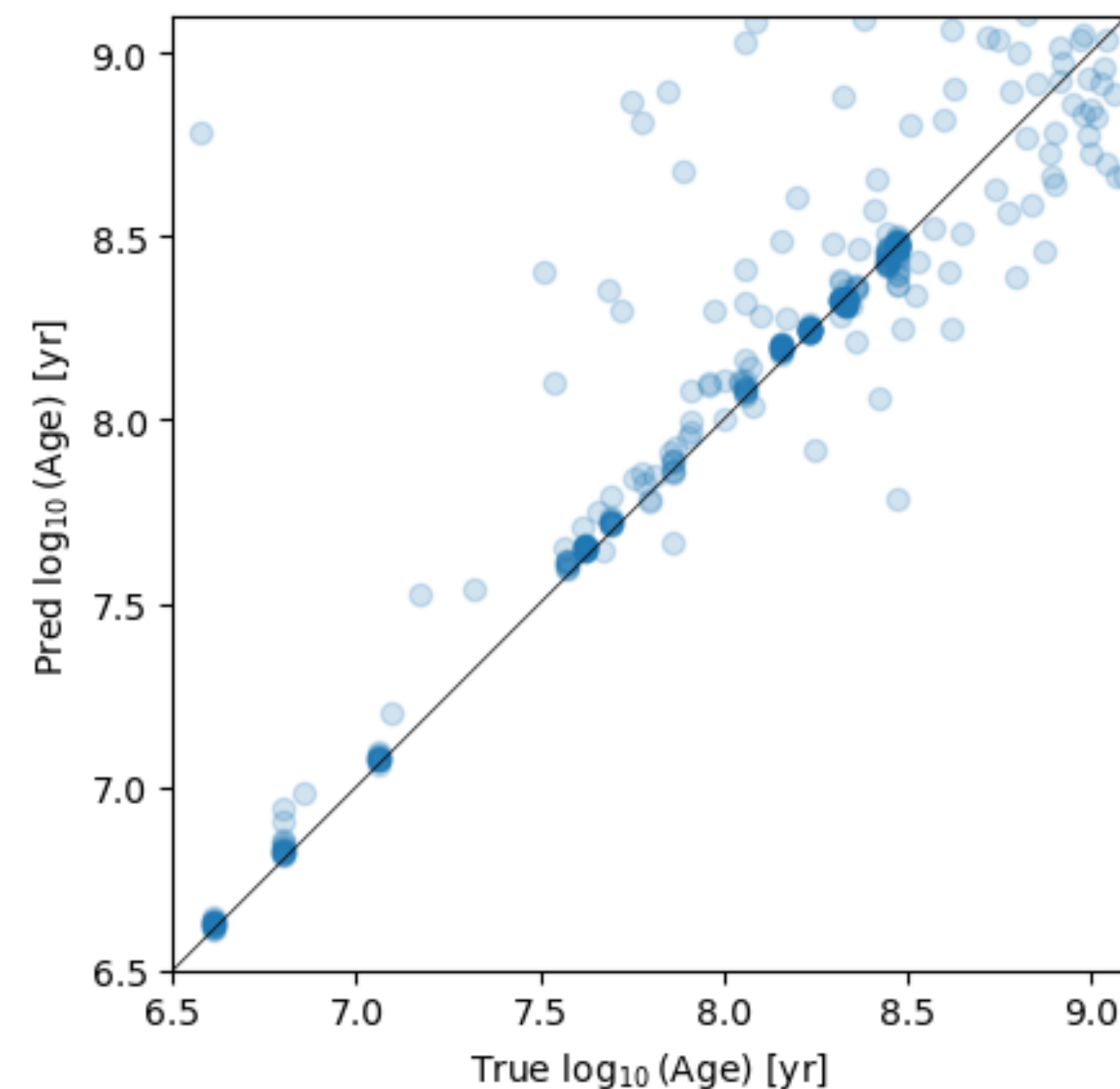
no domain adaption

Posterior mean **real**



with domain adaption

Posterior mean **sim**



Summary

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

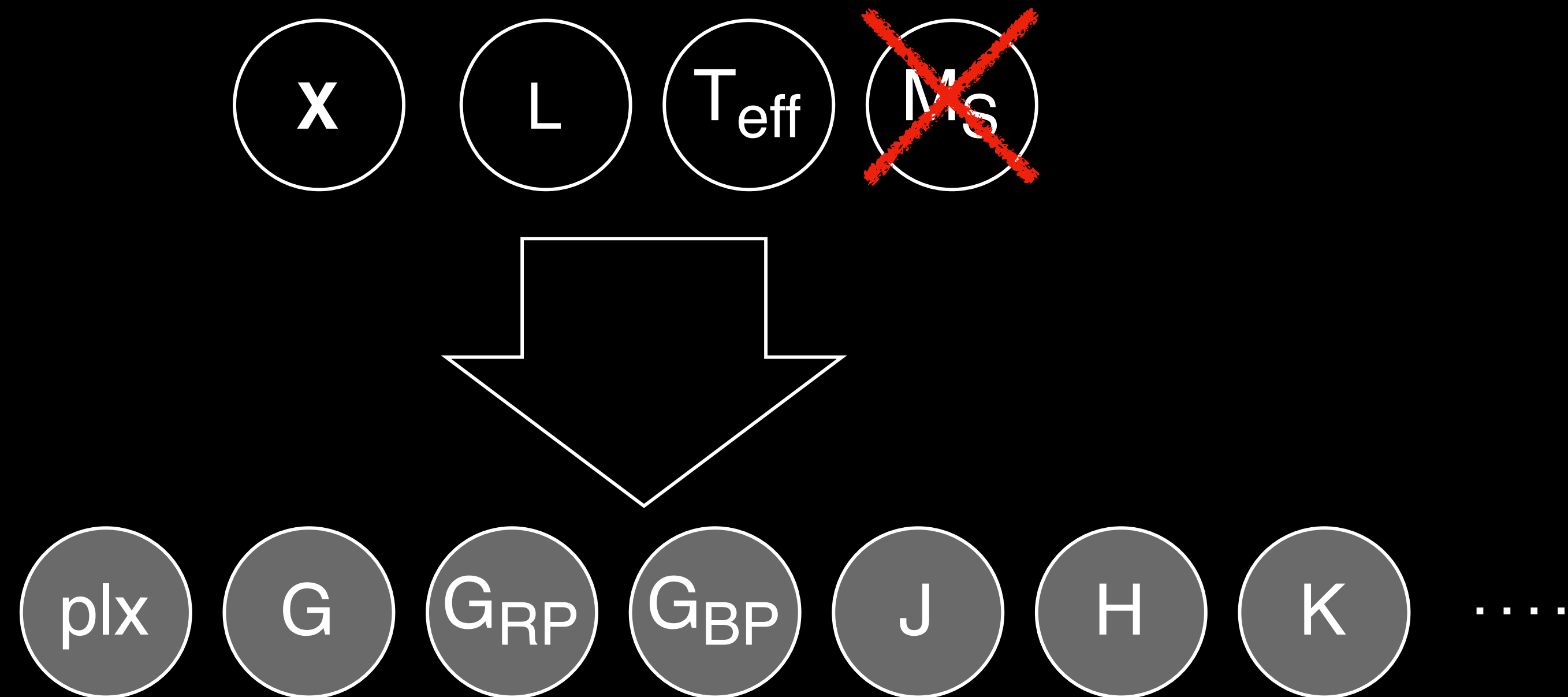
Thank you!

Backup

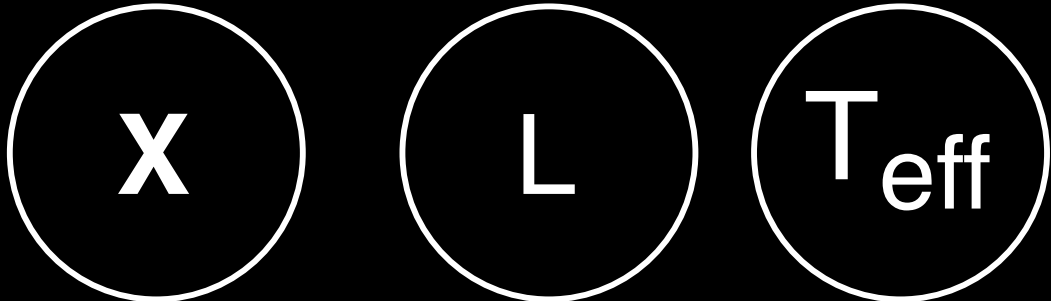
YSO models

Fwd model: YSO (Robitaille17+Richardson+24)

- Mass not input parameter
 - Unlink to evo. tracks
 - ➔ $\log g = 4$
 - ➔ Small impact for $T_{\text{eff}} \in [3, 20] \text{ kK}$
- Low resolution ($R \sim 15$)

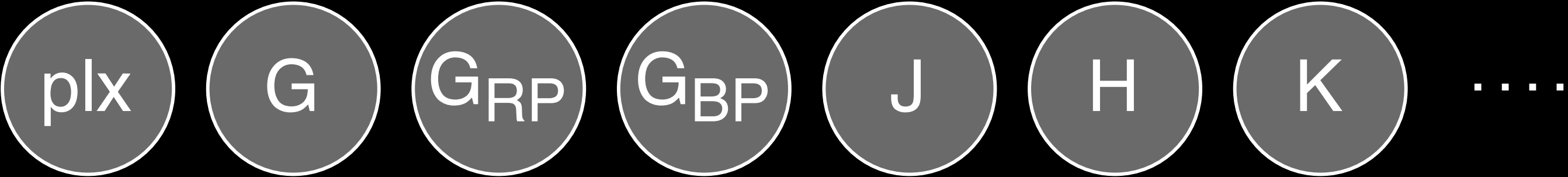


Fwd model: YSO (Robitaille17+Richardson+24)

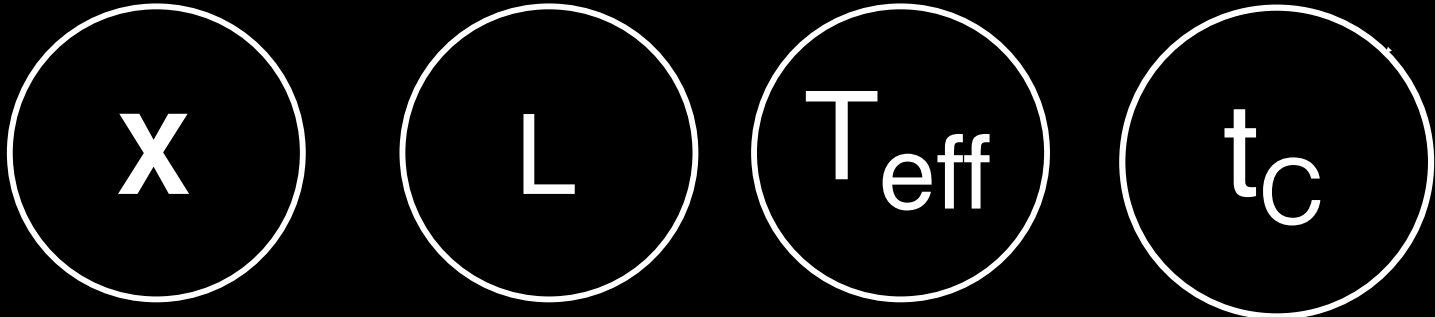


- Mass not input parameter
 - Unlink to evo. tracks
 - ➔ $\log g = 4$
 - ➔ Small impact for $T_{\text{eff}} \in [3, 20] \text{ kK}$
- Low resolution ($R \sim 15$)
- Introduces many additional parameters

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_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_c	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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

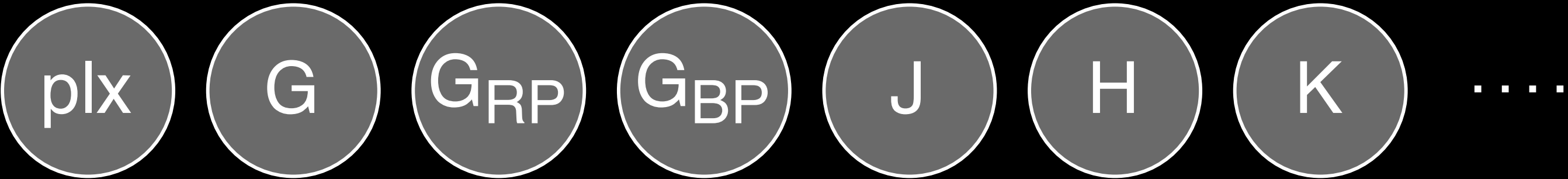


Fwd model: YSO (Robitaille17+Richardson+24)

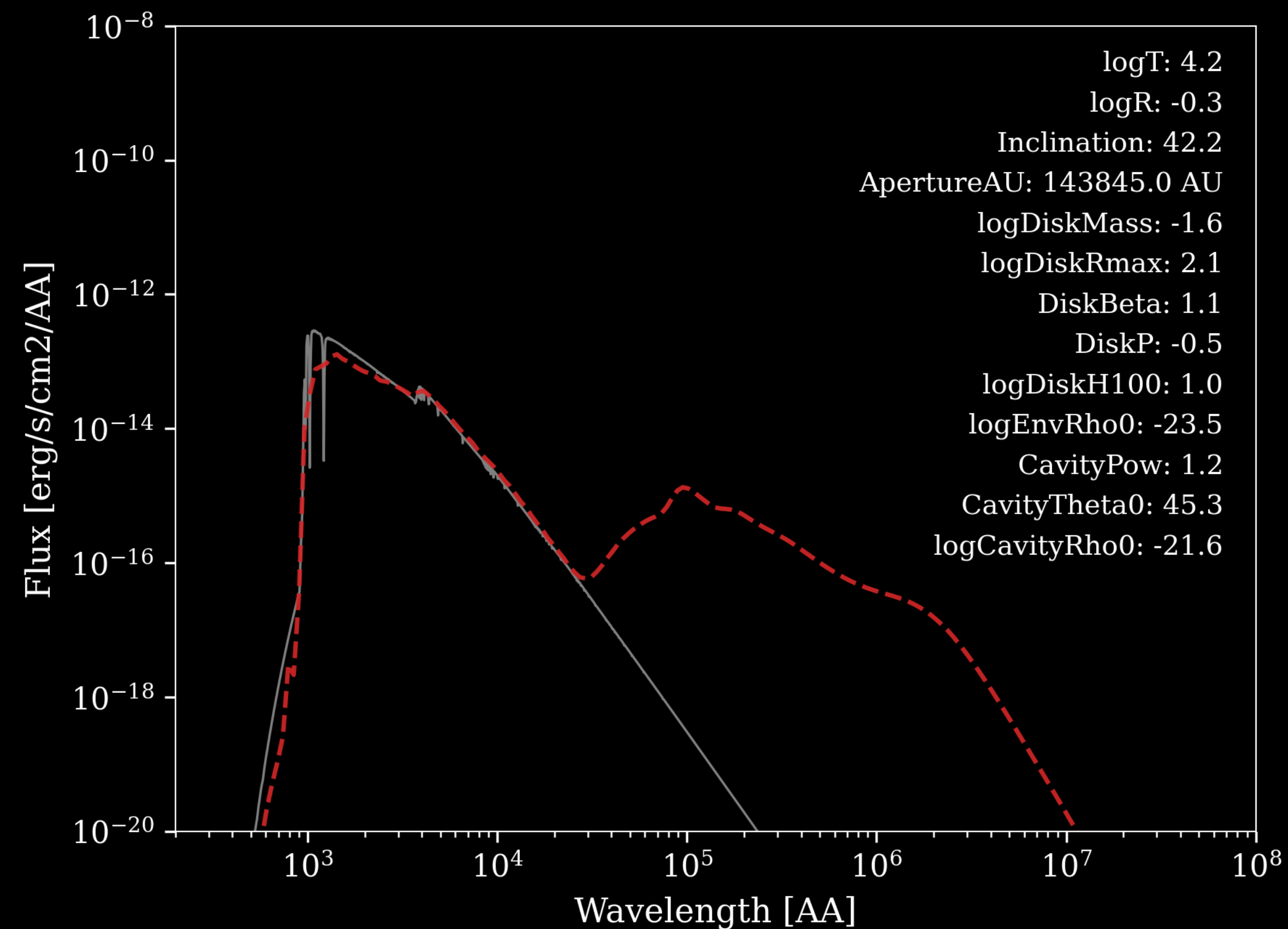


- Mass not input parameter
 - Unlink to evo. tracks
 - ➔ $\log g = 4$
 - ➔ Small impact for $T_{\text{eff}} \in [3,20]$ kK
- Low resolution (R~15)
- Introduces many additional parameters

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_{disk}	$10^{-8} M_\odot$	$0.1 M_\odot$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_c	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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

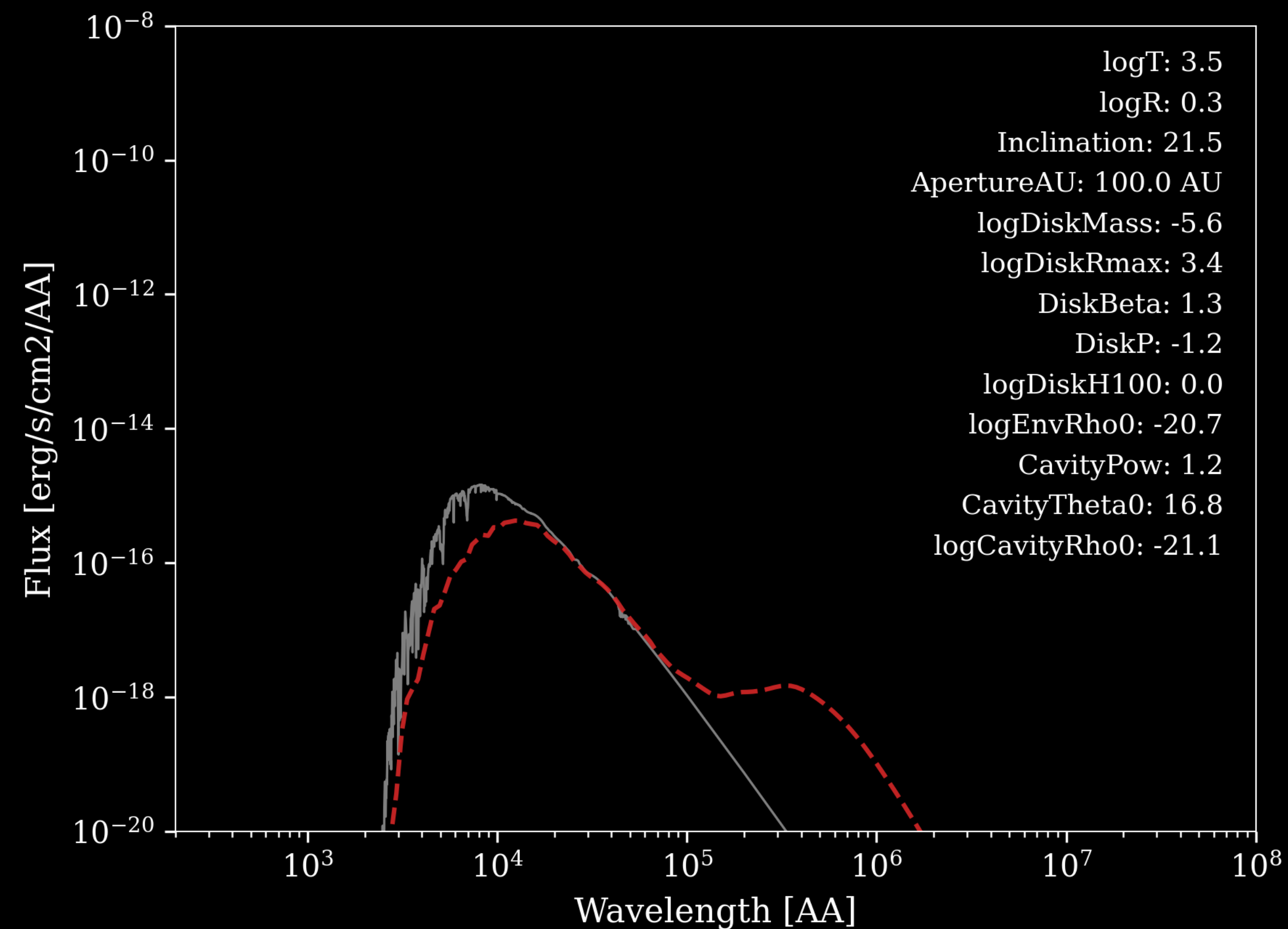


Fwd model: YSO (Robitaille17+Richardson+24)



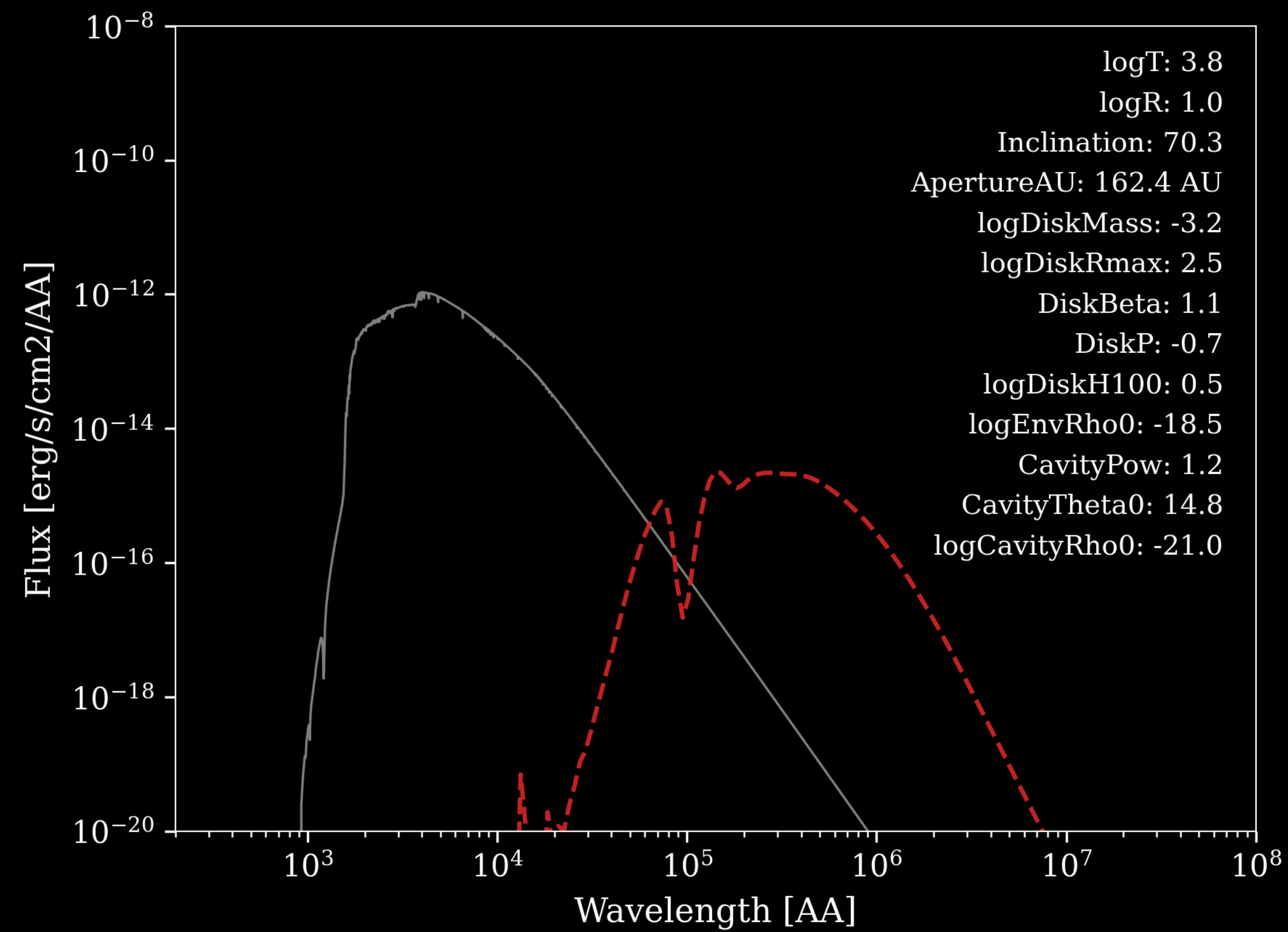
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_{\star}$	$0.1 R_{\odot}$	$100 R_{\odot}$	Log
Stellar temperature	$T_{\star}$	2000 K	30 000 K	Log
Disk mass [dust]	M_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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)



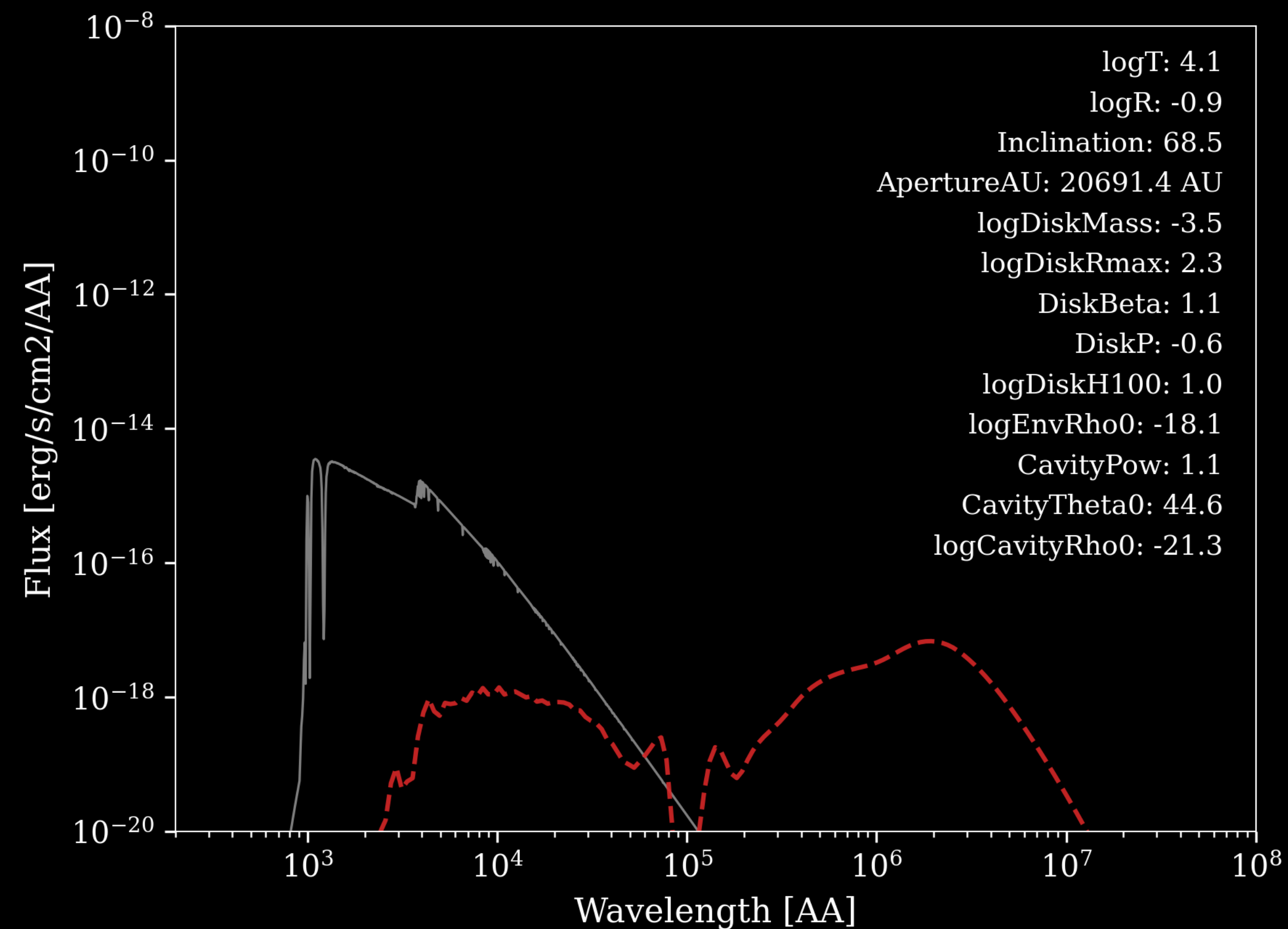
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_{\star}$	$0.1 R_{\odot}$	$100 R_{\odot}$	Log
Stellar temperature	$T_{\star}$	2000 K	30 000 K	Log
Disk mass [dust]	M_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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)



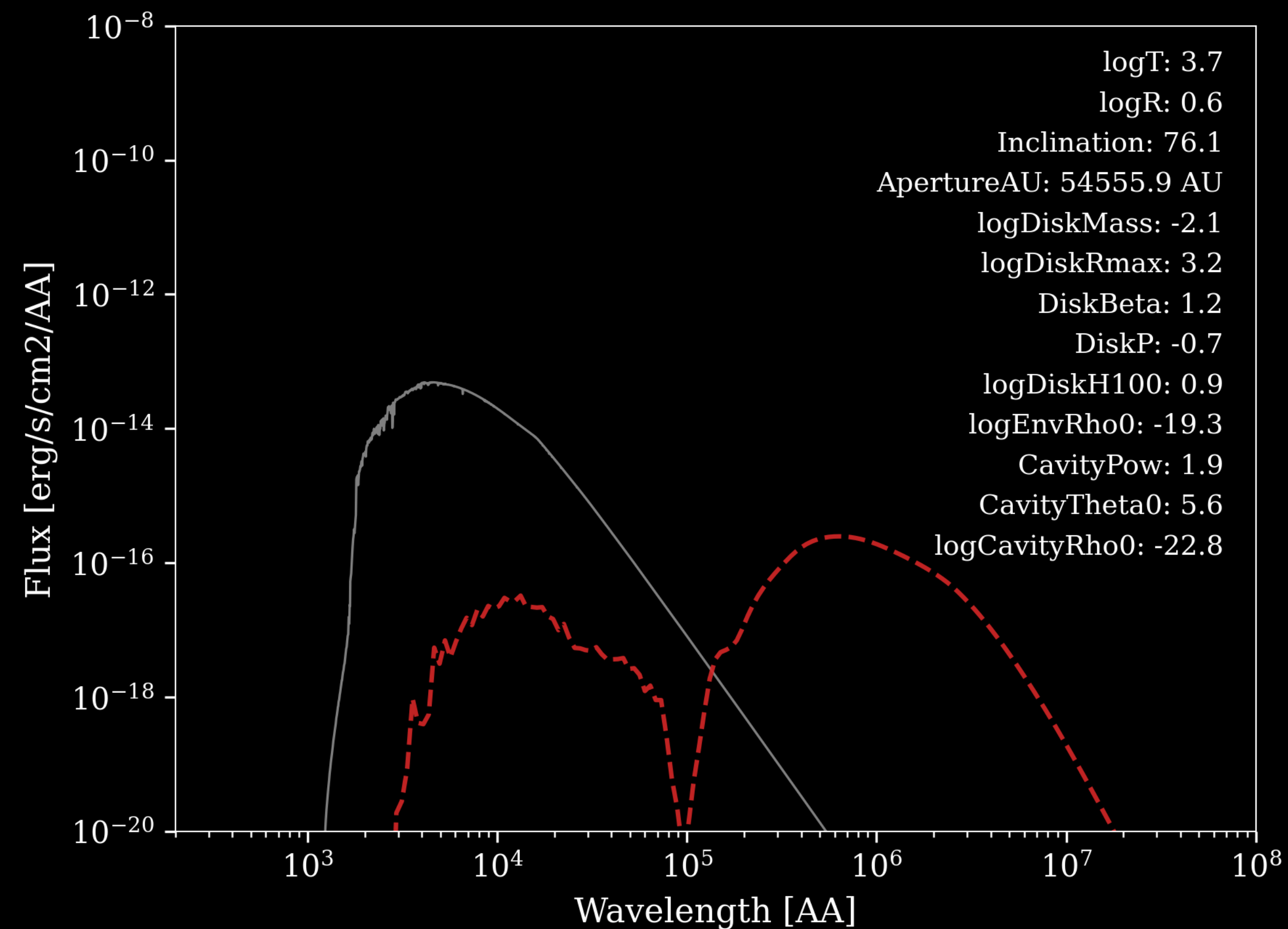
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_{\star}$	$0.1 R_{\odot}$	$100 R_{\odot}$	Log
Stellar temperature	$T_{\star}$	2000 K	30 000 K	Log
Disk mass [dust]	M_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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)



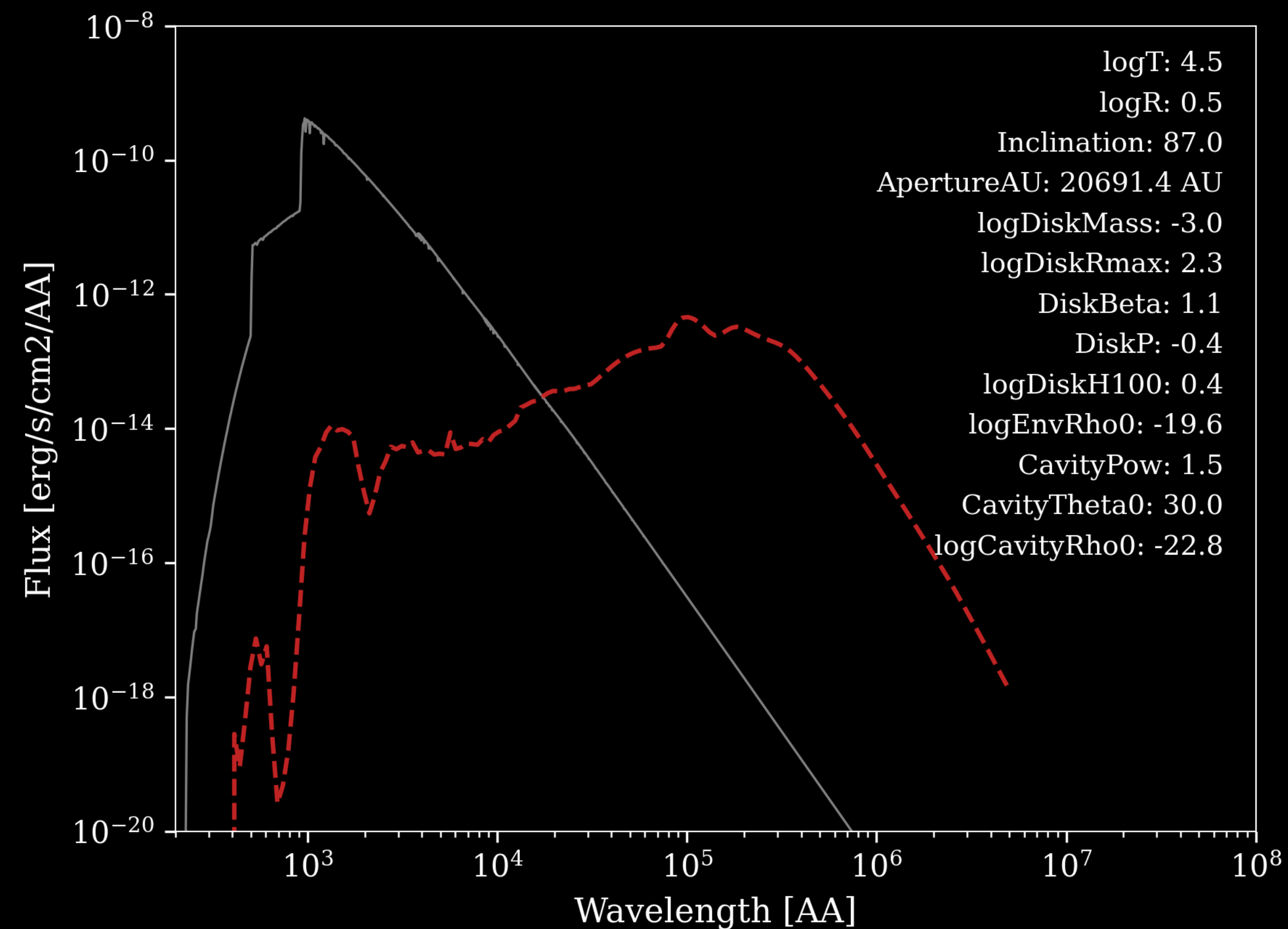
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_{\star}$	$0.1 R_{\odot}$	$100 R_{\odot}$	Log
Stellar temperature	$T_{\star}$	2000 K	30 000 K	Log
Disk mass [dust]	M_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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)



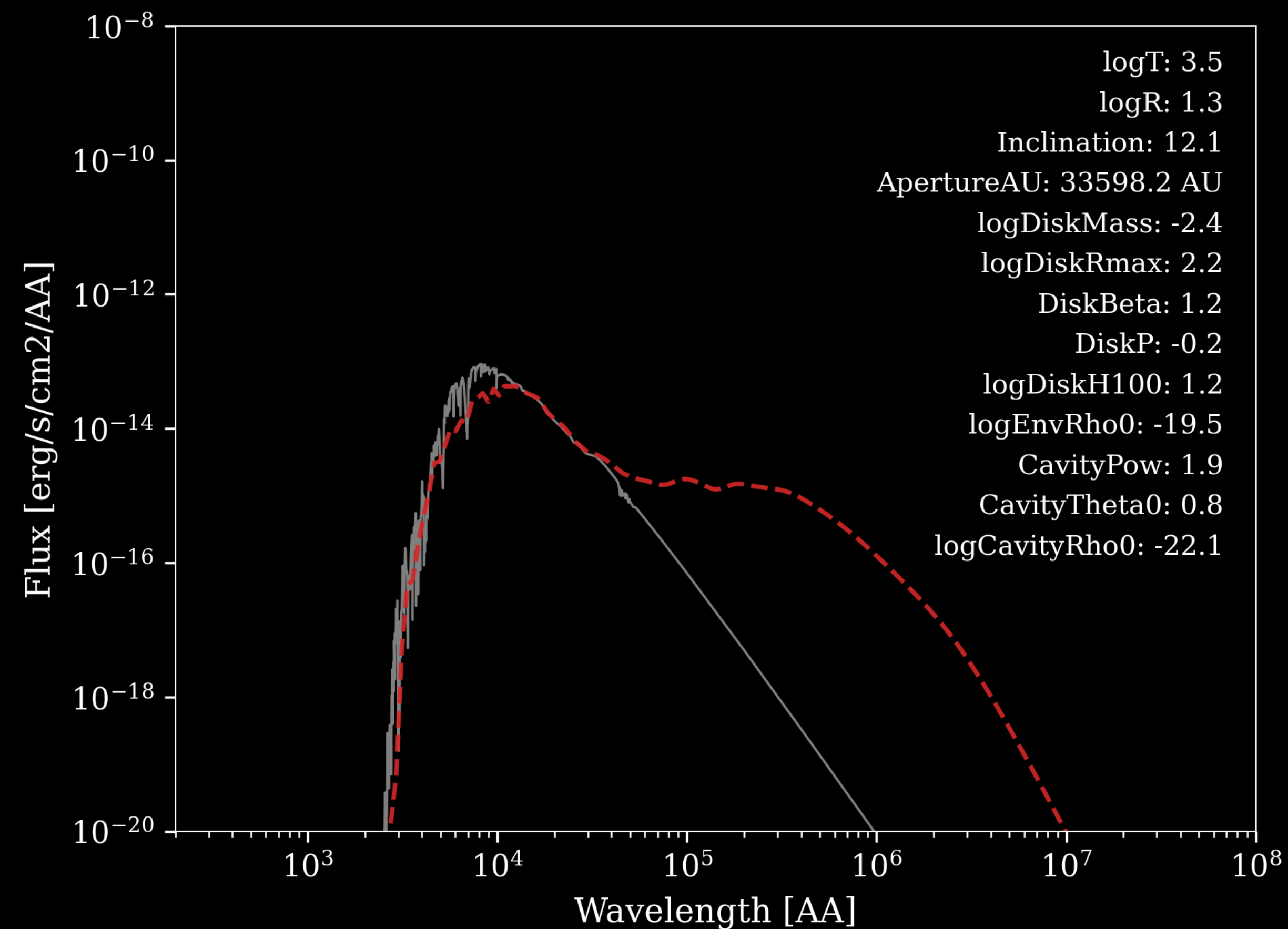
Parameter	Symbol	Minimum	Maximum	Sampling
Stellar radius	$R_{\star}$	$0.1 R_{\odot}$	$100 R_{\odot}$	Log
Stellar temperature	$T_{\star}$	2000 K	30 000 K	Log
Disk mass [dust]	M_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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

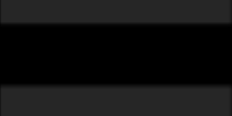
Fwd model: YSO (Robitaille17+Richardson+24)



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_{disk}	$10^{-8} M_{\odot}$	$0.1 M_{\odot}$	Log
Disk inner radius	$R_{\text{min}}^{\text{disk}}$	R_{sub}	$1000 R_{\text{sub}}$	Log
Disk outer radius	$R_{\text{max}}^{\text{disk}}$	50 AU	5000 AU	Log
Disk flaring power	β	1	1.3	Linear
Disk surface density power	p	-2	0	Linear
Disk scaleheight	$h_{100\text{AU}}$	1 AU	20 AU	Log
Envelope density [dust]	ρ_0^{env}	10^{-24} g/cm^3	10^{-16} g/cm^3	Log
Envelope density power	γ	-2	-1	Linear
Envelope centrifugal radius	R_{c}	50 AU	5000 AU	Log
Cavity density [dust]	ρ_0^{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_{sub}	7	10 000
sp-h-i		yes	passive	...	...	...	variable	8	10 000
s-smi		yes	...	...	...	yes	R_{sub}	2	10 000
sp-smi		yes	passive	...	...	yes	R_{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_{sub}	2	10 000
sp-hmi		yes	passive	...	...	yes	variable	8	10 000
spubhmi		yes	passive	Ulrich	yes	yes	variable	12	80 000