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)

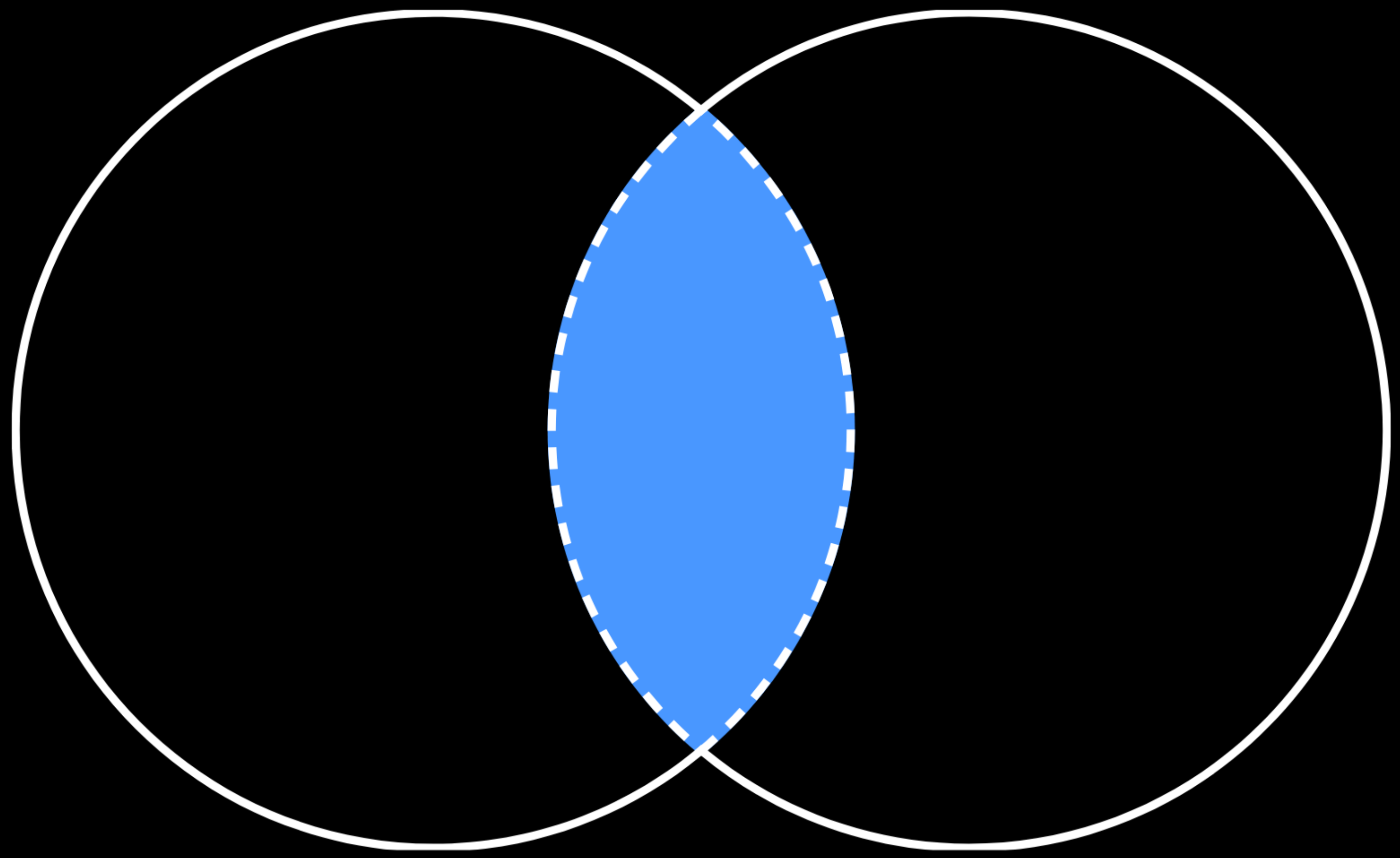
YSOs: Critical link to understanding Galactic baryon cycle

YSOs: Critical link to understanding Galactic baryon cycle

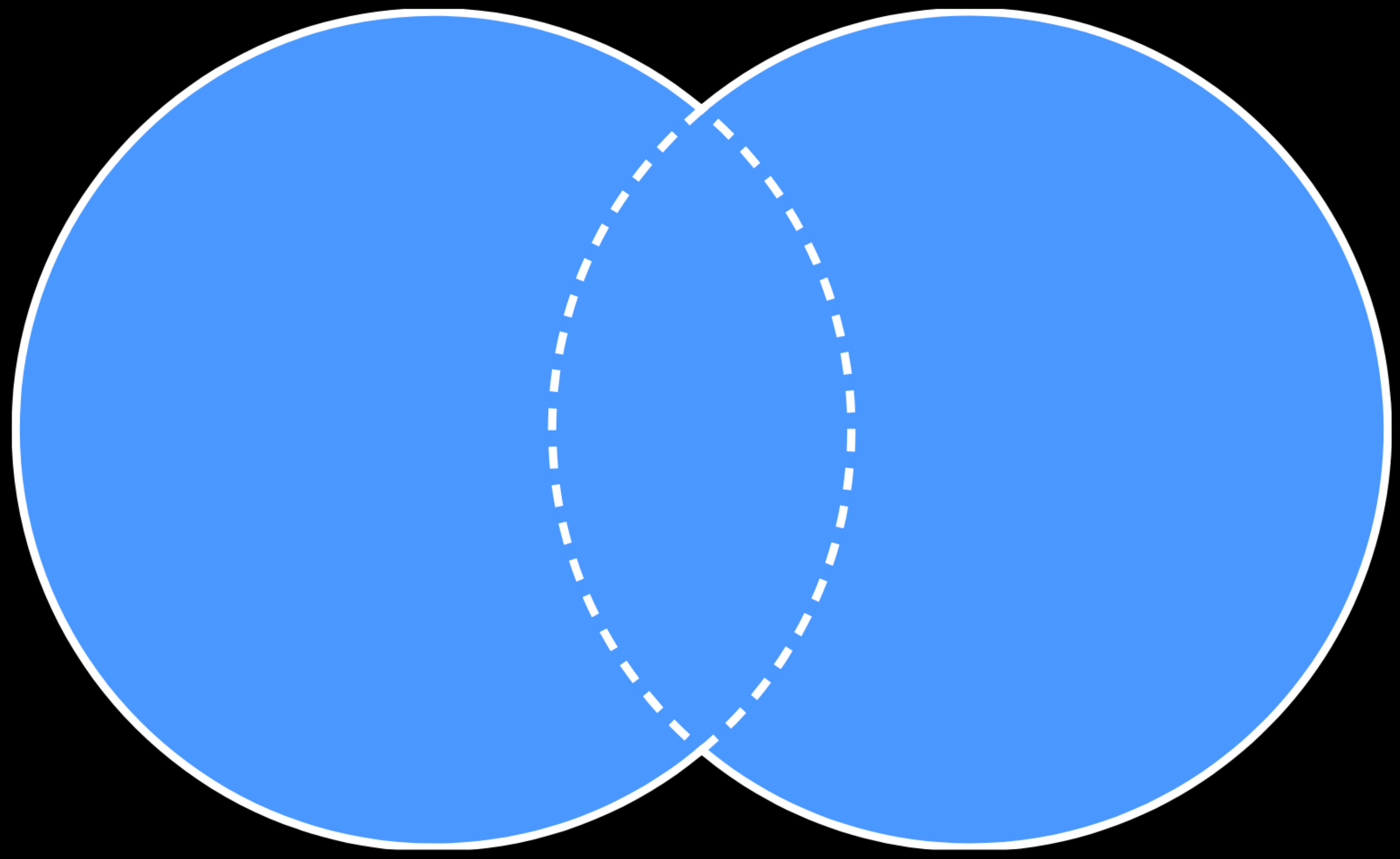
— aim to improve this

Aim to improve YSO catalog

- Data fusion: use as many informative data sets as possible



Often works focus
on intersection



Aim to be truely multi-survey

Aim to improve YSO catalog

- Data fusion: use as many informative data sets as possible

★

Gaia
2MASS
WISE
LAMOST

★

WISE
Spitzer
APOGEE

Aim to improve YSO catalog

- 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

- 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

Typical ML regression

$\vec{x} \longrightarrow$ MLP $\longrightarrow \vec{\theta}$

MLP… Series of ***learnable*** affine transformations of $\vec{x}$

followed by pointwise non-linear map: $f_\phi(\vec{x}) = \hat{\theta}$

ϕ…learable paramerters

Typical ML regression

$\vec{x} \longrightarrow$ MLP $\longrightarrow \vec{\theta}$

MLP… Series of ***learnable*** affine transformations of $\vec{x}$

followed by pointwise non-linear map: $f_\phi(\vec{x}) = \hat{\theta}$

Trained by minizing $||\vec{\theta} - \hat{\theta}||_2$

Typical ML regression

$\vec{x} \longrightarrow \text{MLP} \longrightarrow \vec{\theta}$

Would like to have $p(\vec{\theta} \mid \vec{x})$

Typical ML regression

$\vec{x} \longrightarrow \text{MLP} \longrightarrow \vec{\theta}$

Would like to have $p(\vec{\theta} \mid \vec{x})$

However: — $p(\vec{x} \mid \vec{\theta})$ might not be tractable

— $p(\vec{\theta} \mid \vec{x})$ might not scale to millions - billions of "runs"

Typical ML regression

$\vec{x} \longrightarrow \text{MLP} \longrightarrow \vec{\theta}$

Would like to have $p(\vec{\theta} \mid \vec{x})$

However: — $p(\vec{x} \mid \vec{\theta})$ might not be tractable

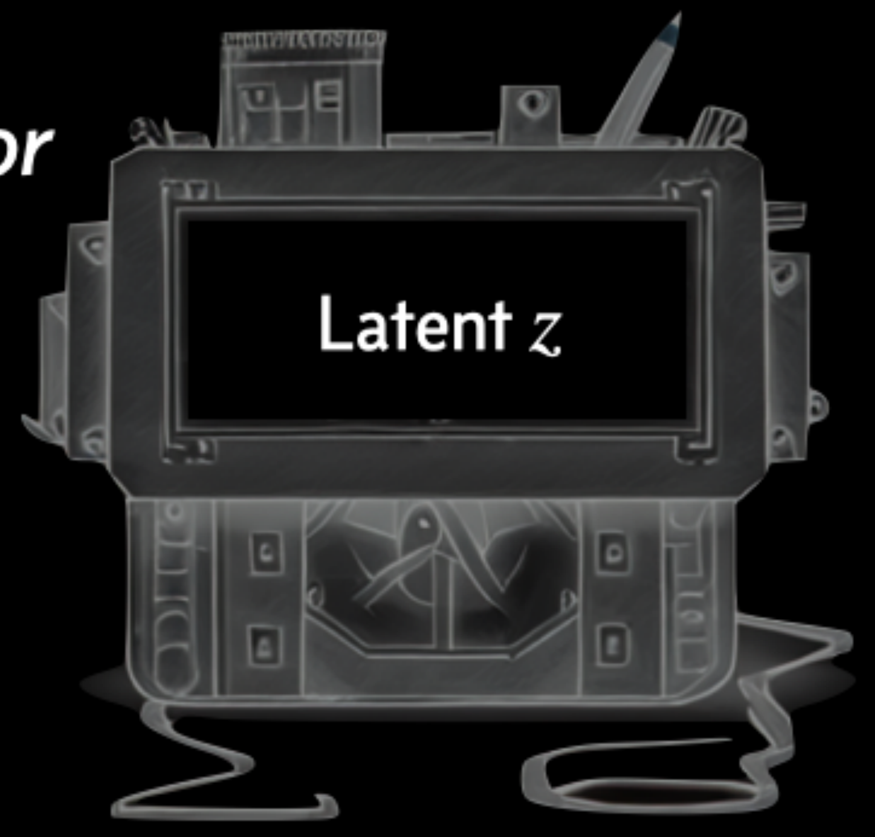
— $p(\vec{\theta} \mid \vec{x})$ might not scale to millions - billions of “runs”

BUT: if we have access to a simulator, we can approximate $p(\vec{\theta} \mid \vec{x})$

Simulation based inference (SBI) setup

Simulator

Parameters
θ



Observations
x

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

$\vec{x} \longrightarrow \text{MLP} \longrightarrow \vec{\theta}$

ML regression

Neural posterior estimation

$\vec{x}$ ⟶ MLP ⟶ $\vec{\theta}$ ML regression

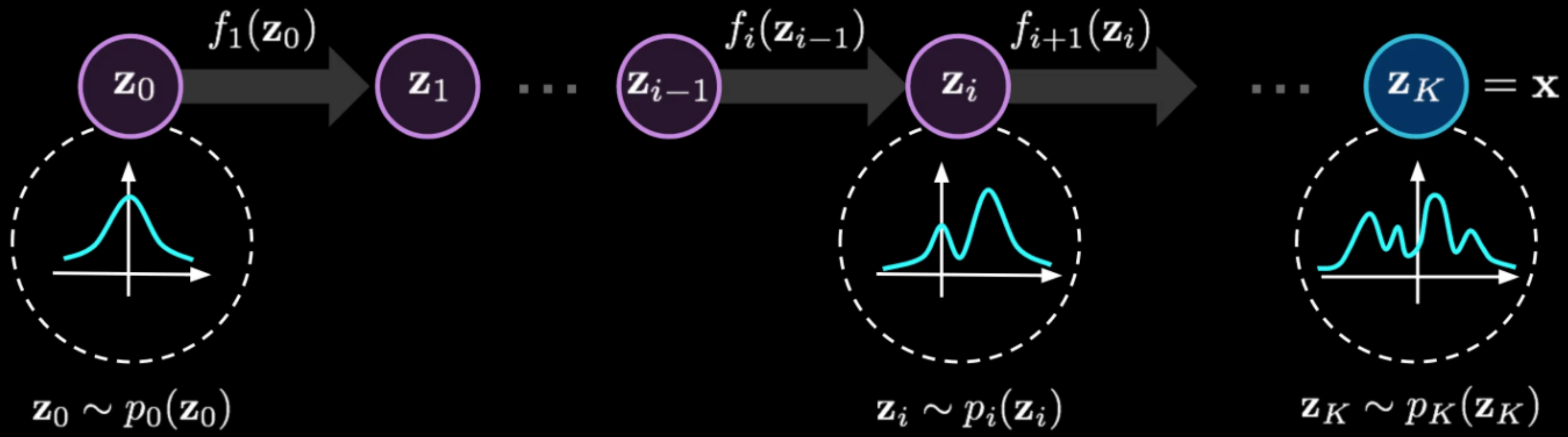
$\vec{x}$ ⟶ MLP $f_\phi(\vec{x}) = s$ ⟶ Conditional neural density estimator $p(\vec{\theta} \mid s) \sim p(\vec{\theta} \mid \vec{x})$

(summary statistics)

Papamakarios & Murray (2016) & Lueckmann et al. (2017)

Normalizing flows

$\vec{x}$ ⟶ MLP $f_\phi(\vec{x}) = s$ (summary statistics) ⟶ Conditional neural density estimator $p(\vec{\theta} \mid s) \sim p(\vec{\theta} \mid \vec{x})$



Parameterized, invertible maps f_i that transform Gaussian into target distribution

Training objective: ***maximum likelihood***

Neural posterior estimation

$\vec{x}$ ⟶ MLP $f_\phi(\vec{x}) = s$ ⟶ Conditional neural density estimator $p(\vec{\theta} \mid s) \sim p(\vec{\theta} \mid \vec{x})$

(summary statistics)

Cannot deal with missing data

Transformer: learning with incomplete data

$\vec{x} \longrightarrow$ Transformer

$\uparrow$

Attention masking: M_E

Can enforce conditional independence

— Effectively marginalize over missing values

Gloeckler et al. (2024)

Transformer: learning with incomplete data

$\vec{x}$ → Transformer → Conditional density estimator (flow matching) $p(\vec{\theta} \mid \vec{x})$

↑

Attention masking: M_E

Transformer: learning with incomplete data

$\vec{x} \longrightarrow$ Transformer $\longrightarrow$ Conditional density estimator (flow matching) $p(\vec{\theta} \mid \vec{x})$

$\uparrow$

Attention masking: M_E

Gaia
2MASS
WISE
LAMOST

WISE
Spitzer
APOGEE

Model implementation

II. Dealing with model misspecification

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

$$x \longrightarrow \begin{array}{l} x_{\mathrm{sim}} \\ x_{\mathrm{real}} \\ x_{\mathrm{sim-real-pairs}} \end{array}$$

Modality encoders: split into indiv. spectra

x_{sim}

Apogee

Boss

Lamost

XP

Photometry

Gaia, 2Mass, …

Modality encoders: encode

MLP encoder

x_{sim}

Apogee $\longrightarrow$ $z_{\mathrm{s}}^{\mathrm{A}} = f_{\phi}(x_{\mathrm{s}}^{\mathrm{A}})$ $\hat{=}$ summary statistics

Boss $\longrightarrow$ $z_{\mathrm{s}}^{\mathrm{B}}$

Lamost $\longrightarrow$ $z_{\mathrm{s}}^{\mathrm{L}}$

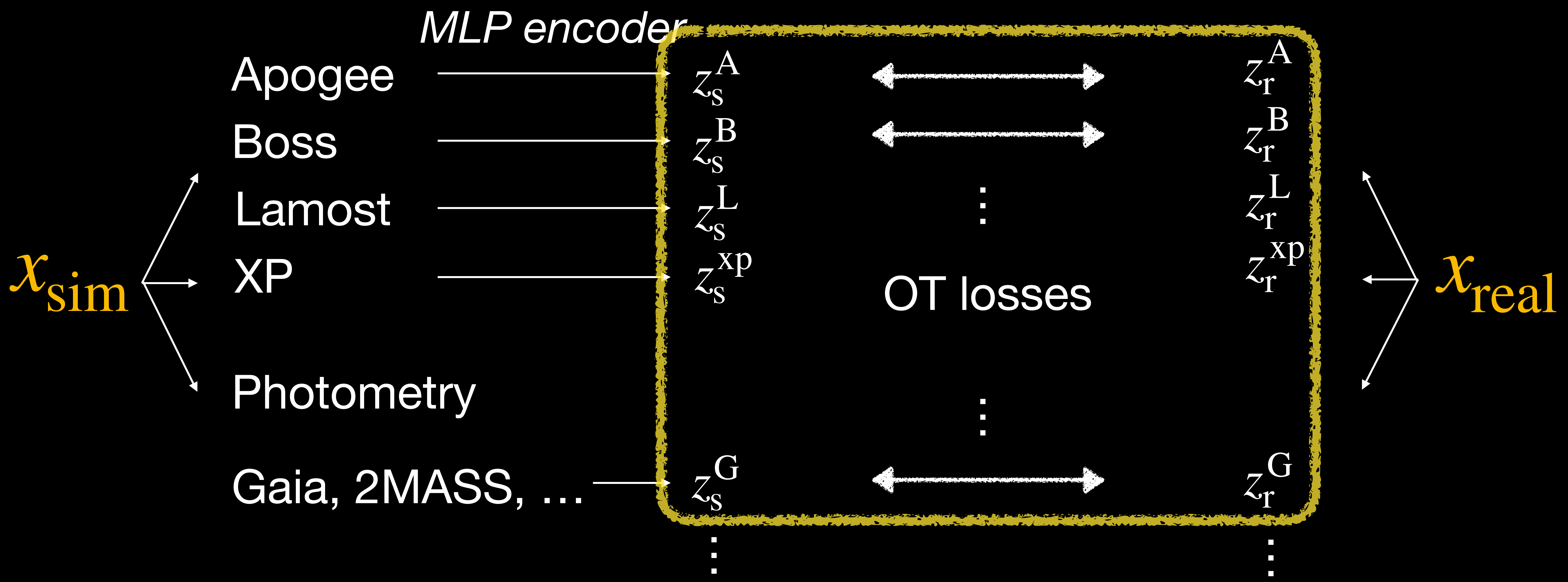
XP $\longrightarrow$ $z_{\mathrm{s}}^{\mathrm{xp}}$

Photometry

Gaia, 2MASS, … $\longrightarrow$ $z_{\mathrm{s}}^{\mathrm{G}}$

$\vdots$ $\vdots$

Modality encoders: alignment loss



Goal: force latent representations of simulated and real data to “look the same”

Sim-real alignment via optimal transport

(a) Samples (b) Exact (c) Gromov-Wasserstein Gu et al. (2022)

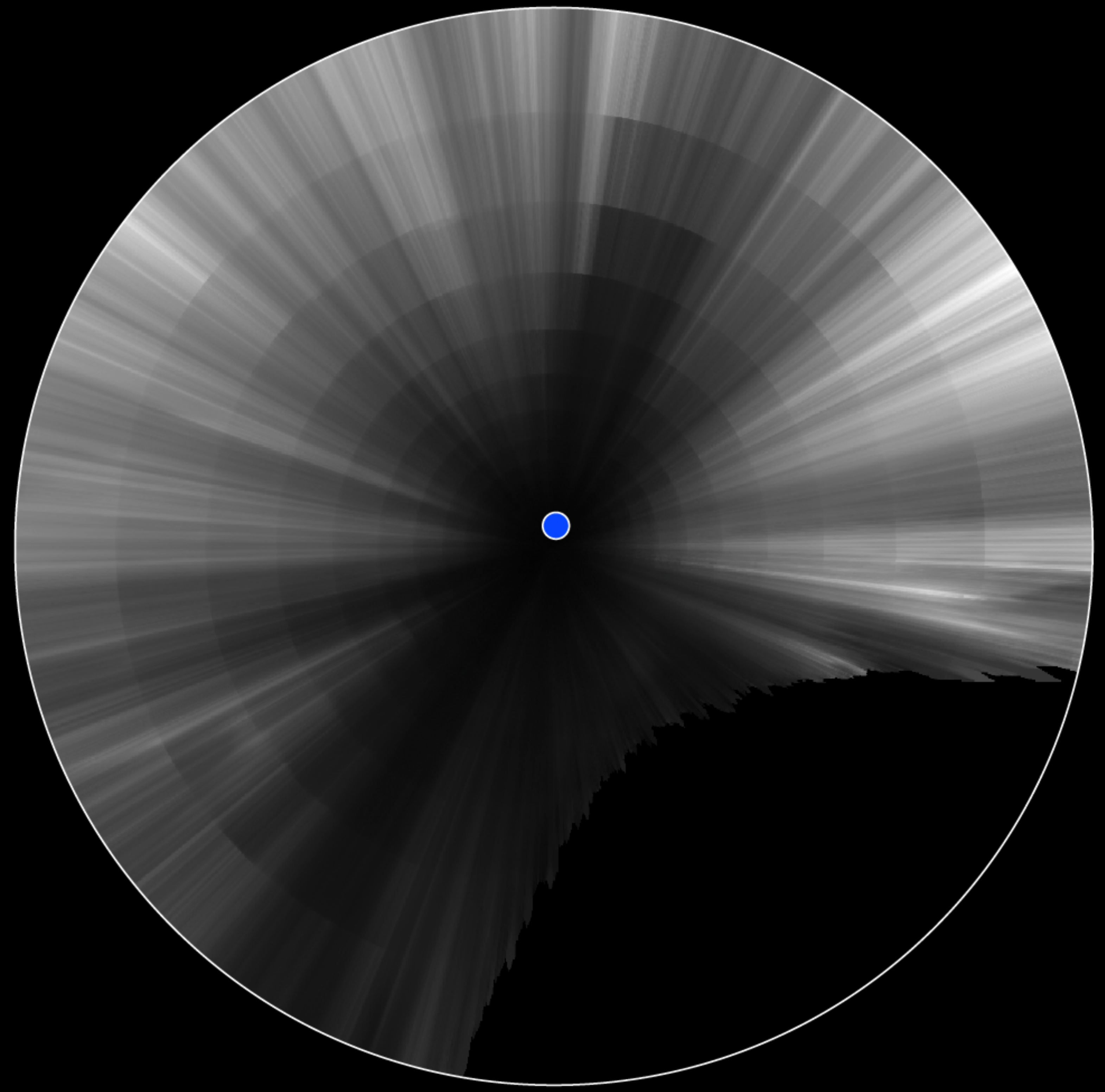
Goal: force latent representations of simulated and real data to “look the same”

Gu et al. (2022)

Let’s test this

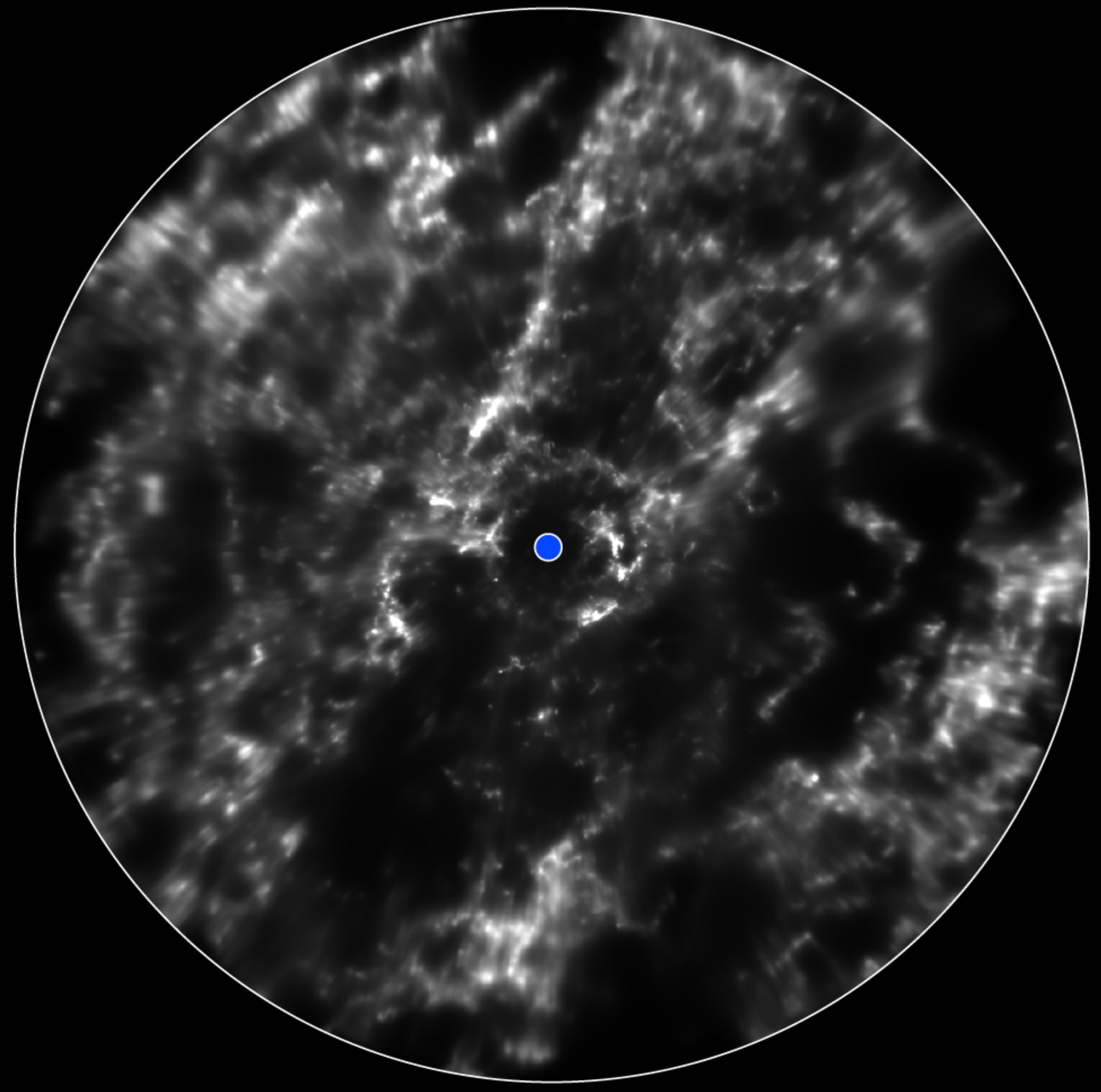
Galaxy A … "sim"

- Dust according to Bayestar2019 (Green+2018)



Galaxy B … "real"

- Dust according to Edenhofer+2024



Galaxy A … “sim”

Synthetic spectra:

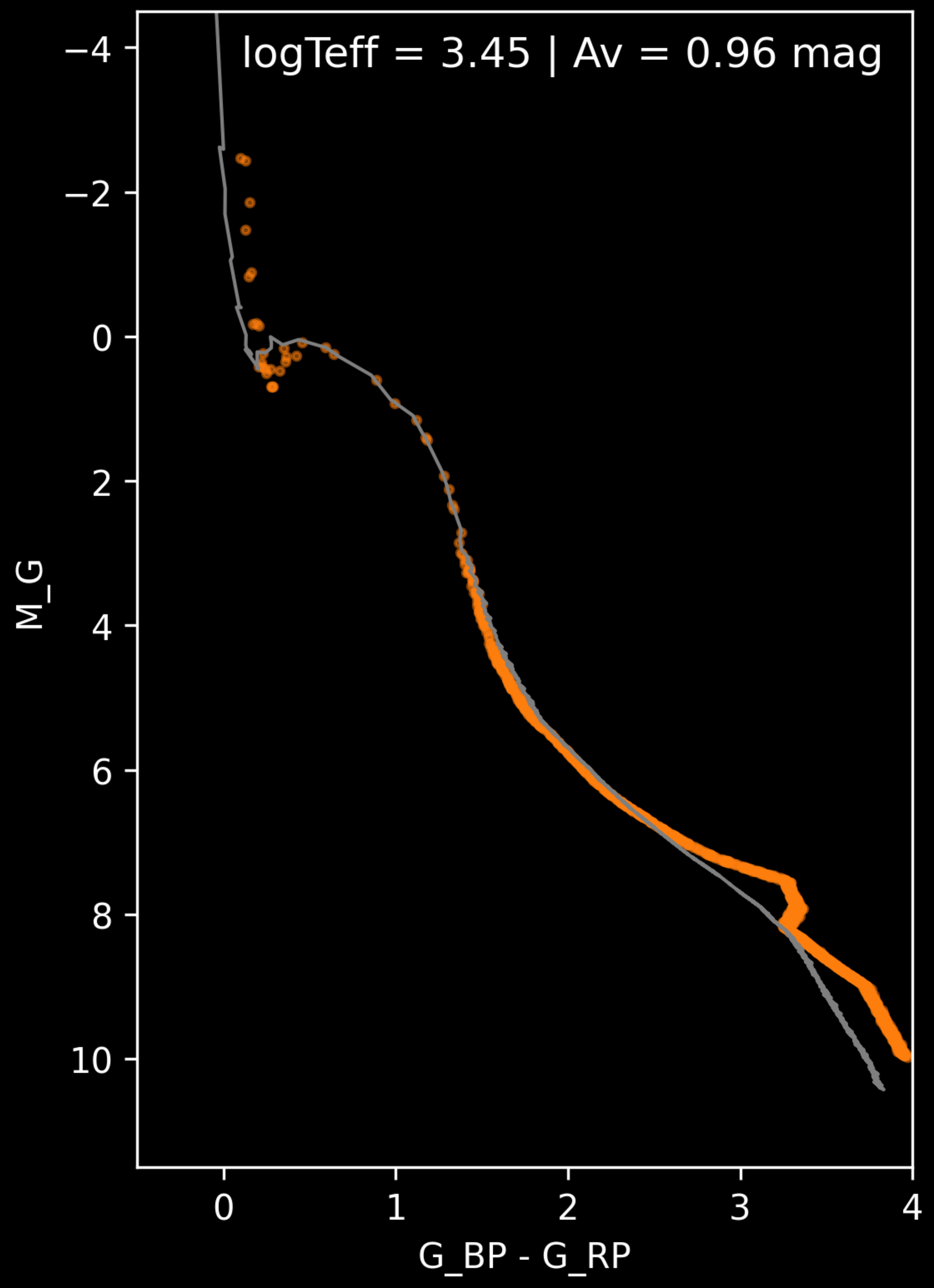
- BaSeL 2.2, ~ Atlas 9 empirically recalibrated (Leujeune+1998)

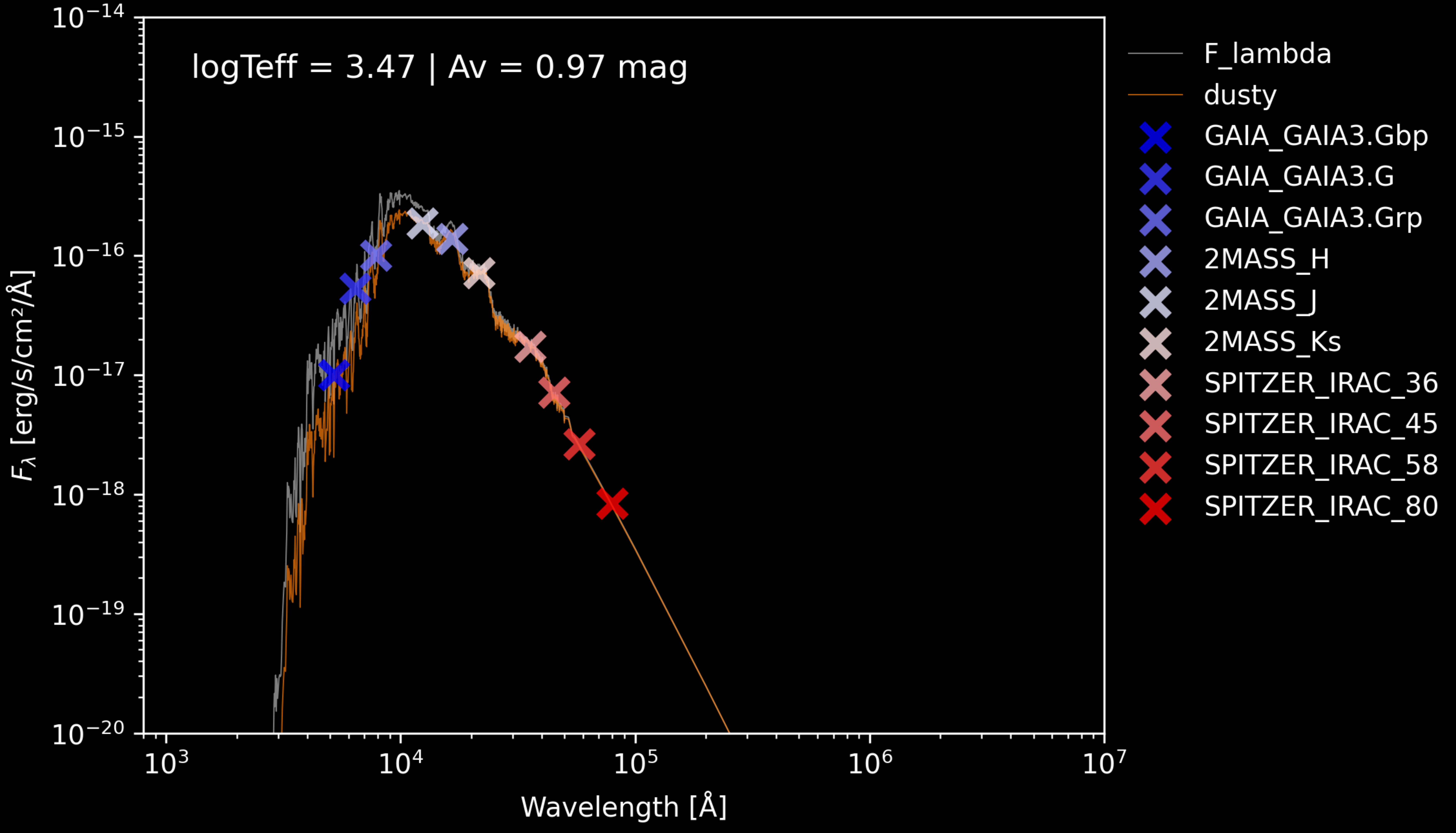
Galaxy B … “real”

Synthetic spectra:

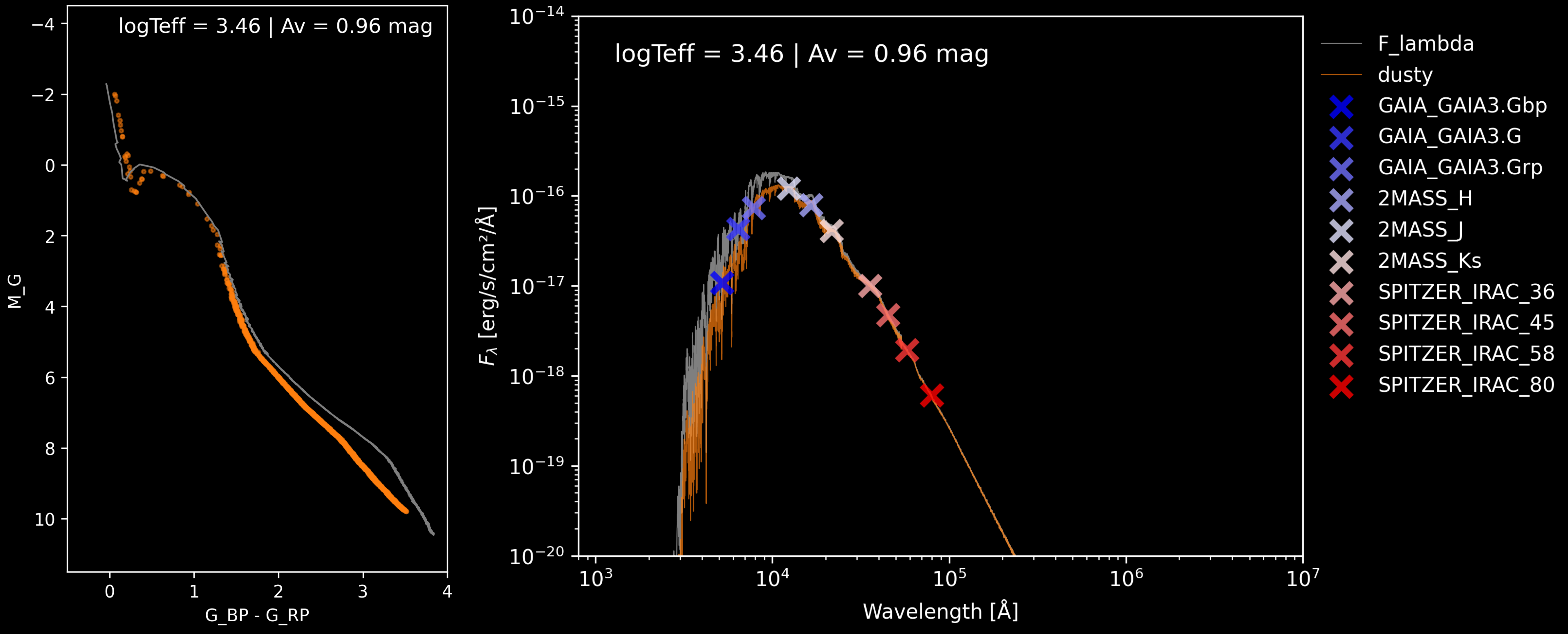
- BT-Settl Library (Allard, Hauschildt and Schweitzer 2000)

Model differences (BaSeL)



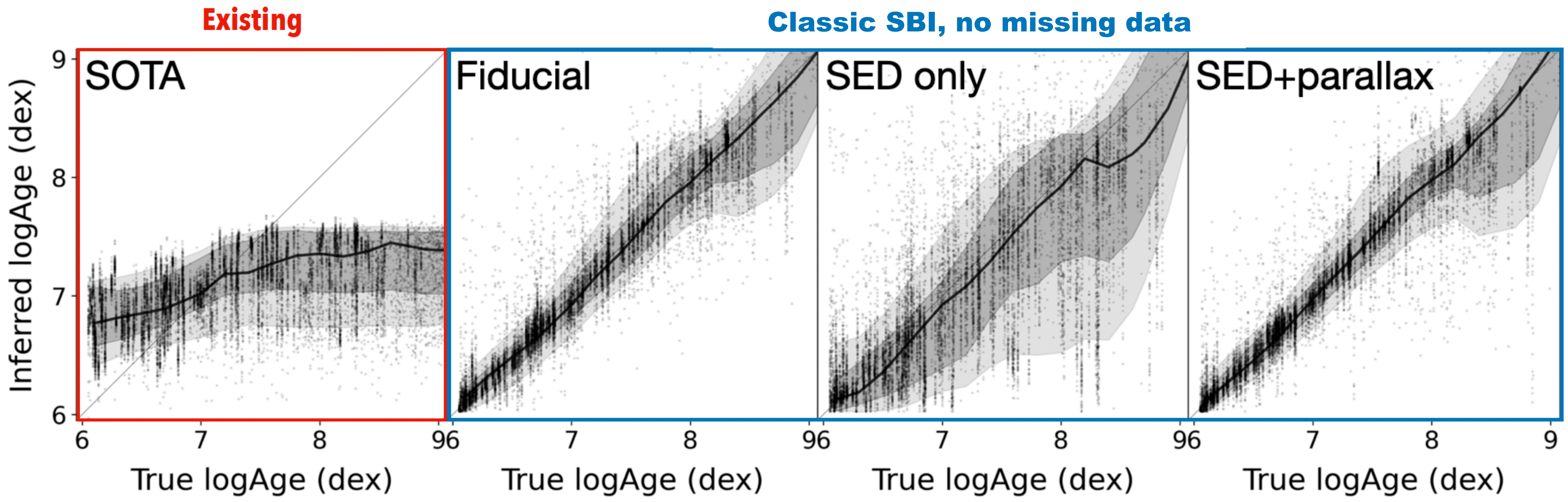


Model differences (BTSettl)

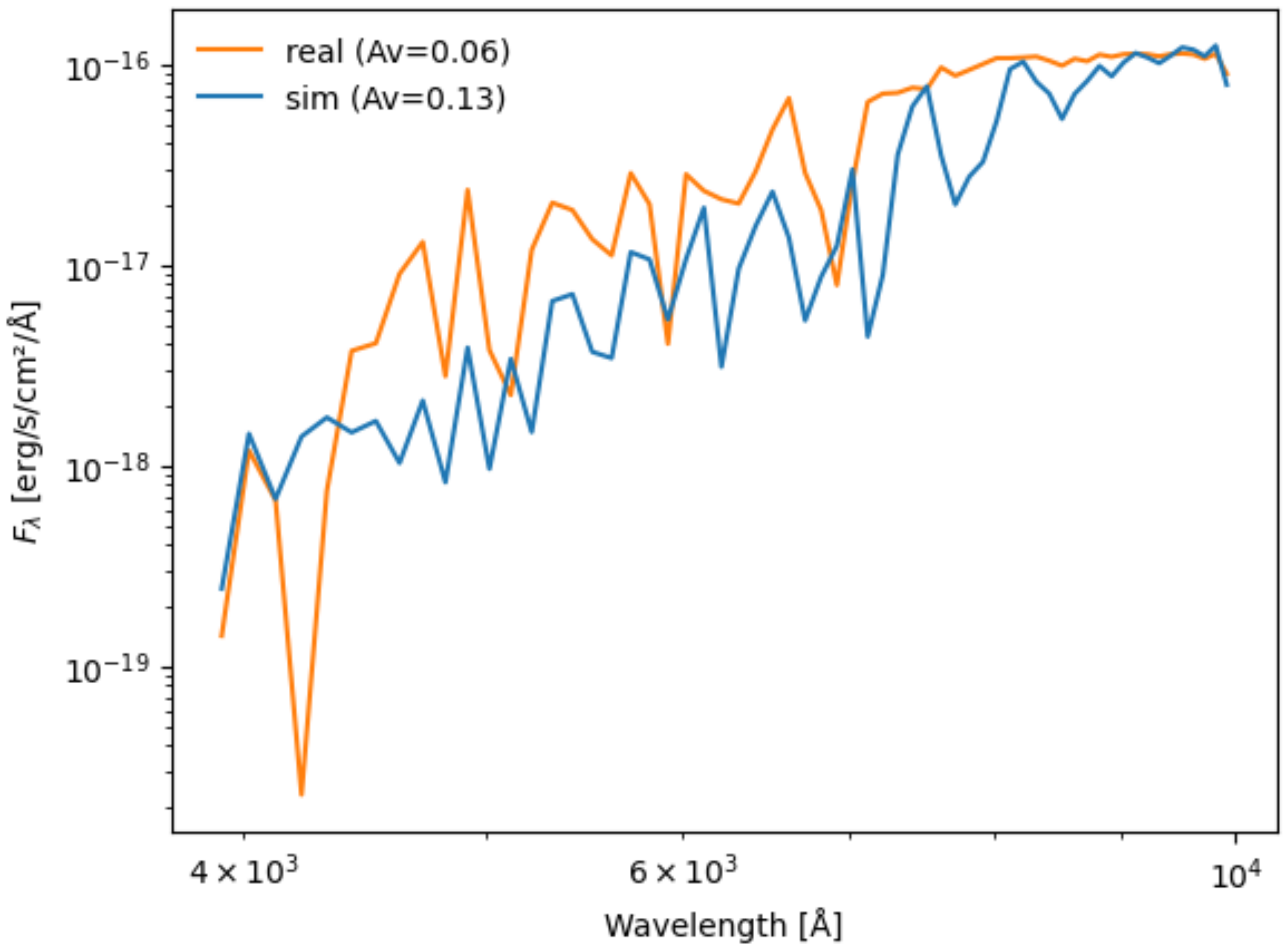


Results

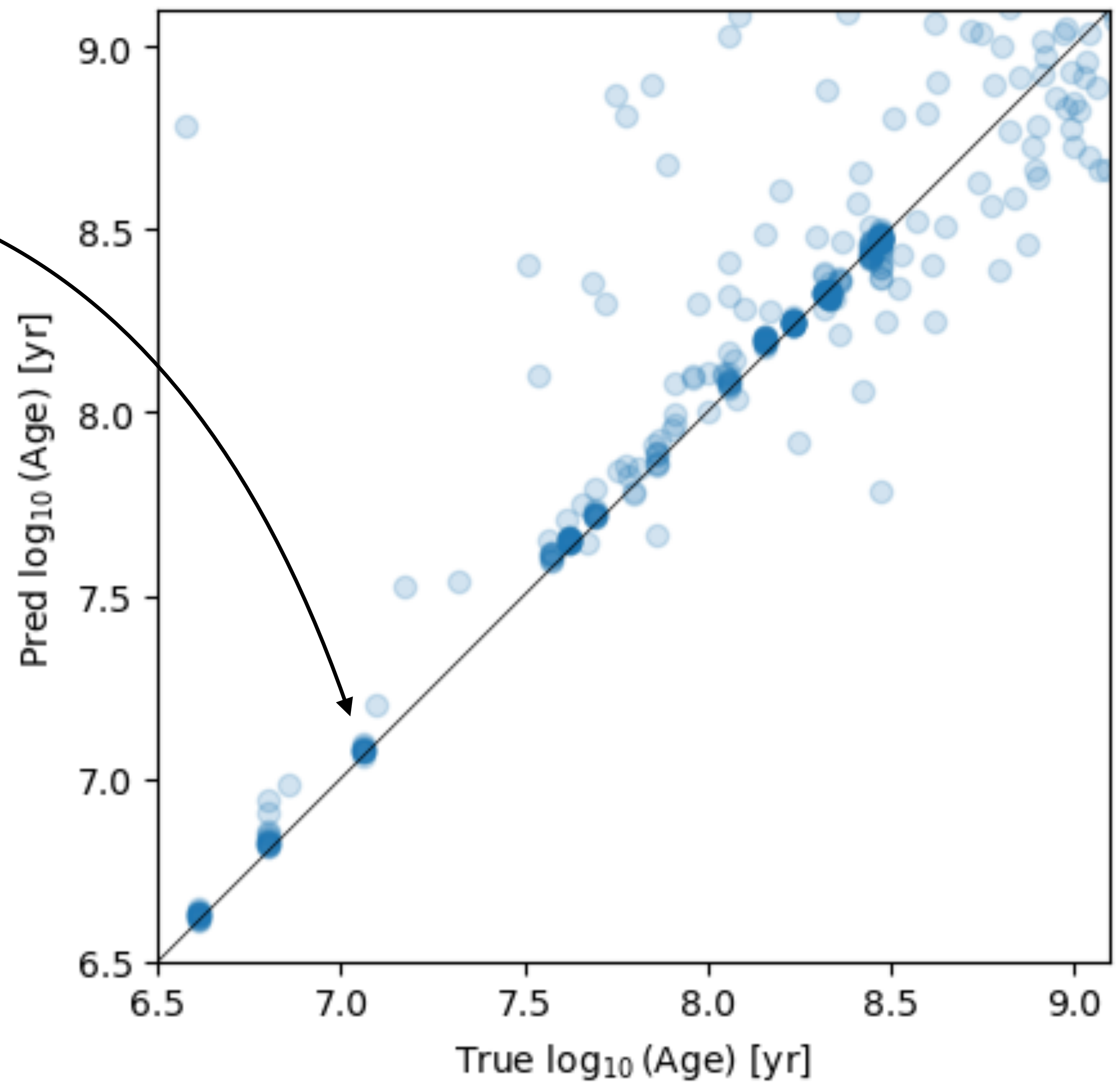
Pilot study: Gaia+2MASS+WISE



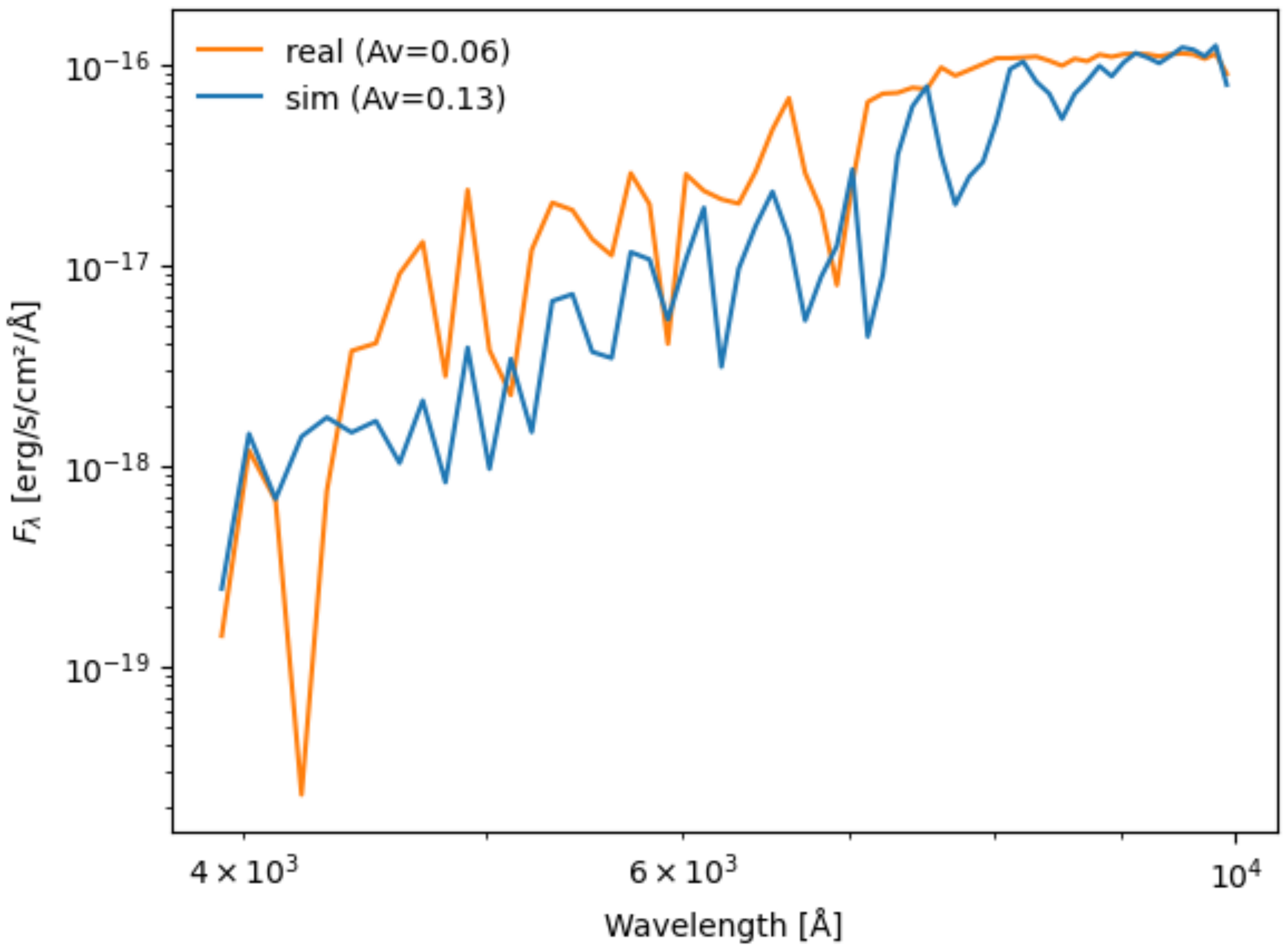
Updated pipeline + XP spectra



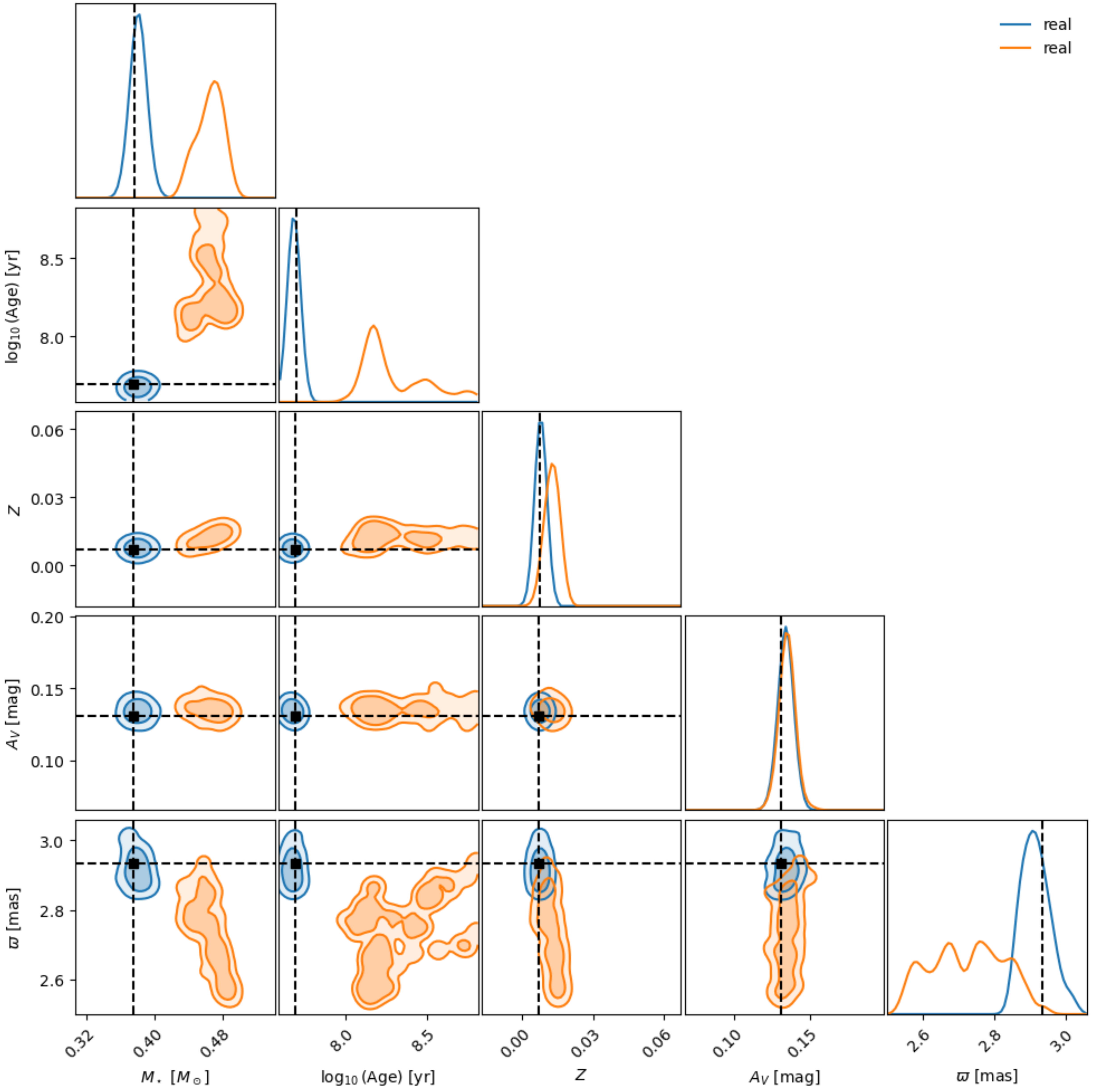
Posterior mean **sim**



Without DA

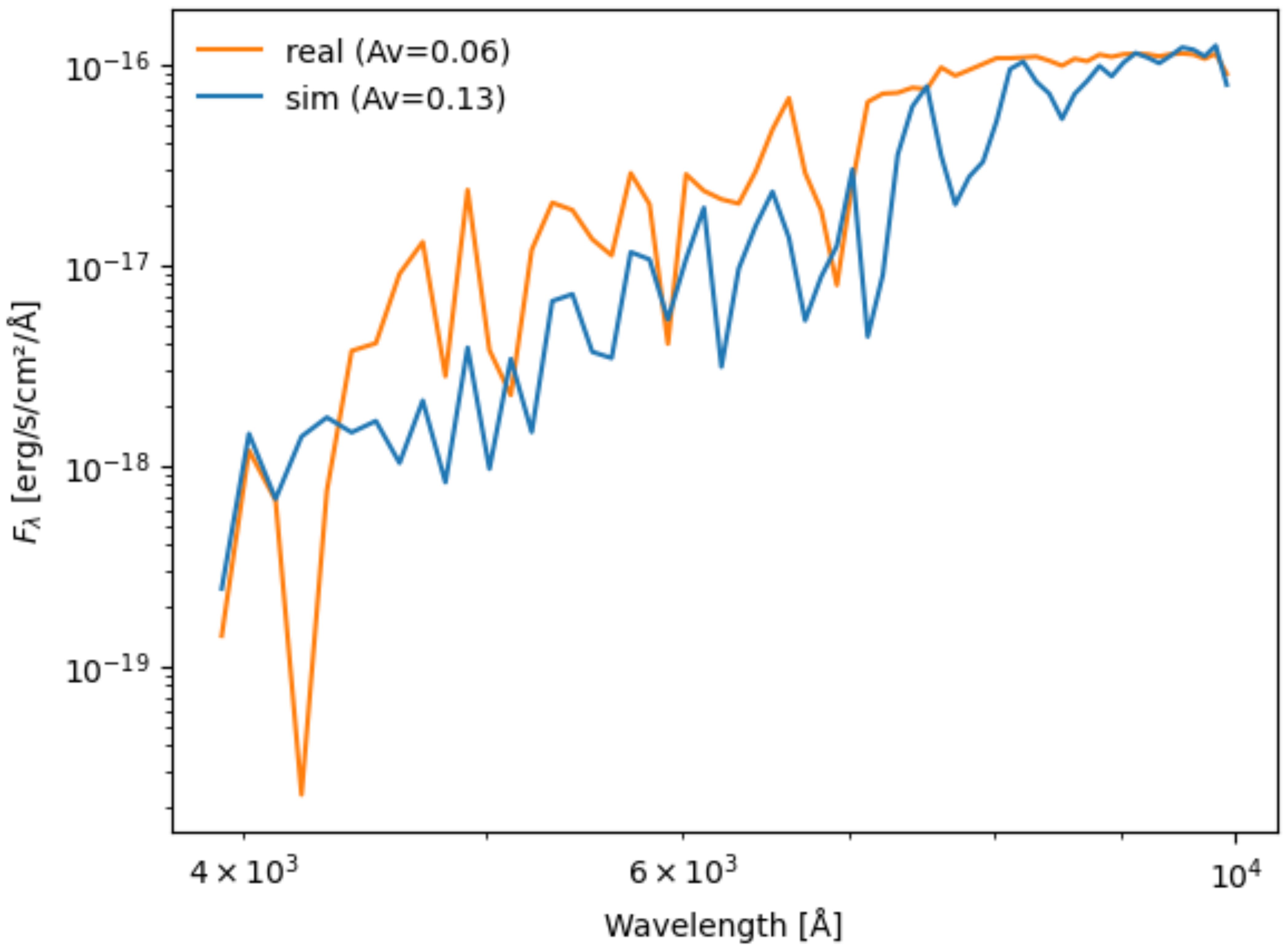


real (Av=0.06)
sim (Av=0.13)
10^{-16}
10^{-17}
10^{-18}
10^{-19}
F_λ [erg/s/cm²/Å]
4×10^3
6×10^3
10^4
Wavelength [Å]

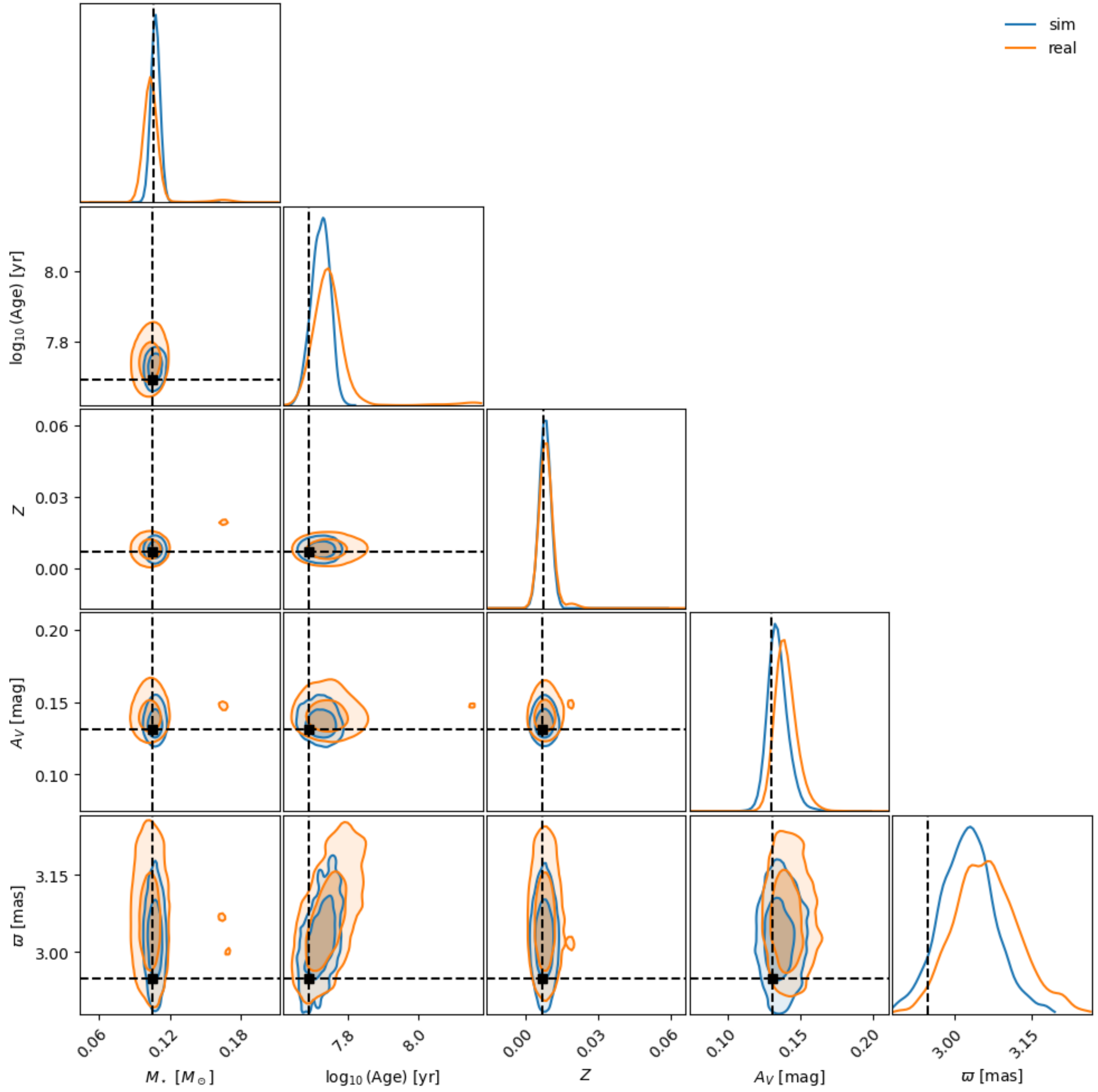


real
real
$\log_{10}$(Age) [yr]
Z
A_V [mag]
ϖ [mas]
$M_\star$ [$M_\odot$]

With DA



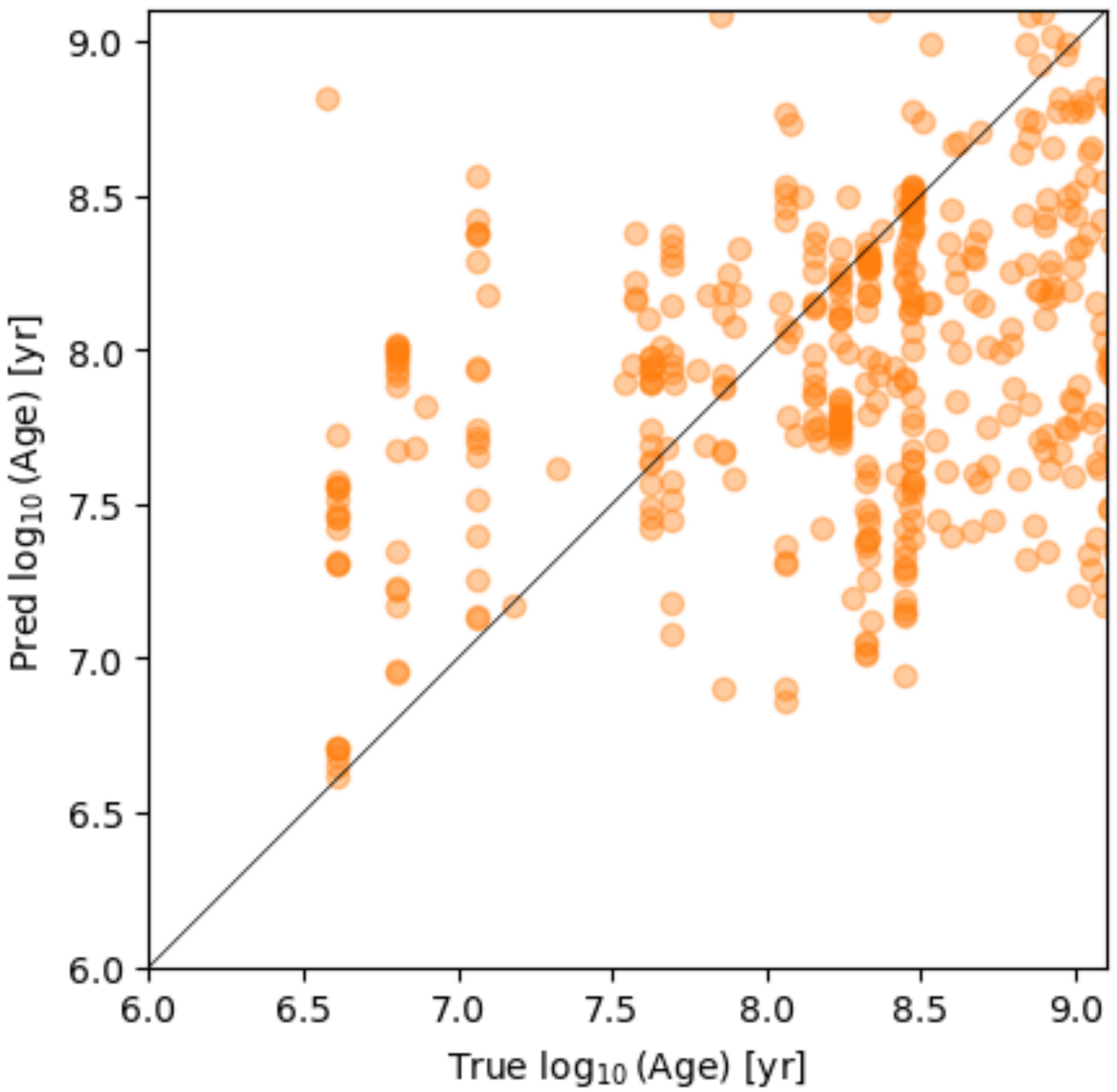
real (Av=0.06)
sim (Av=0.13)
10^{-16}
10^{-17}
10^{-18}
10^{-19}
F_λ [erg/s/cm²/Å]
4×10^3
6×10^3
10^4
Wavelength [Å]



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

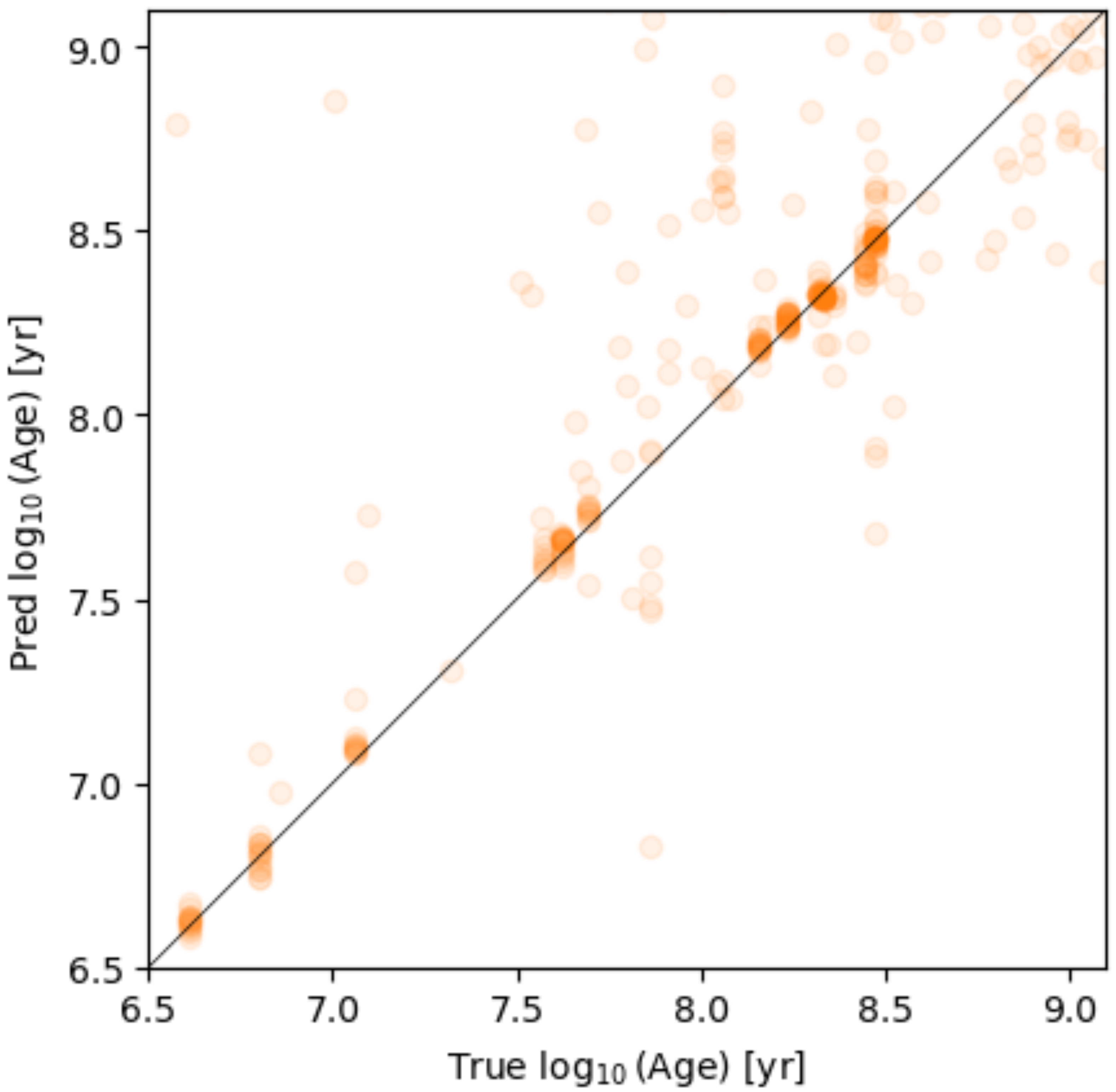
Predictions

Posterior mean “**real**”



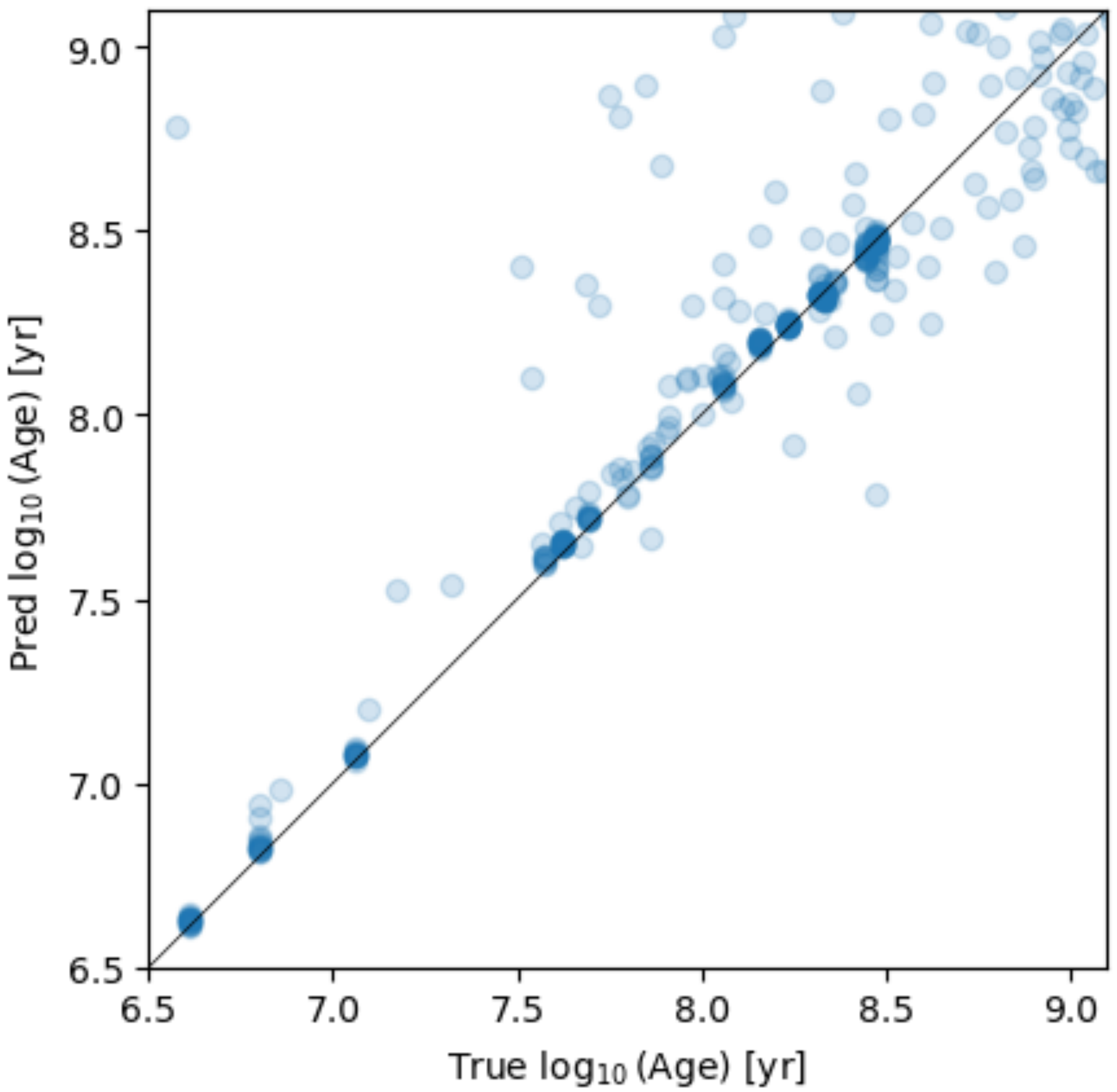
no domain adaption

Posterior mean “**real**”



with domain adaption

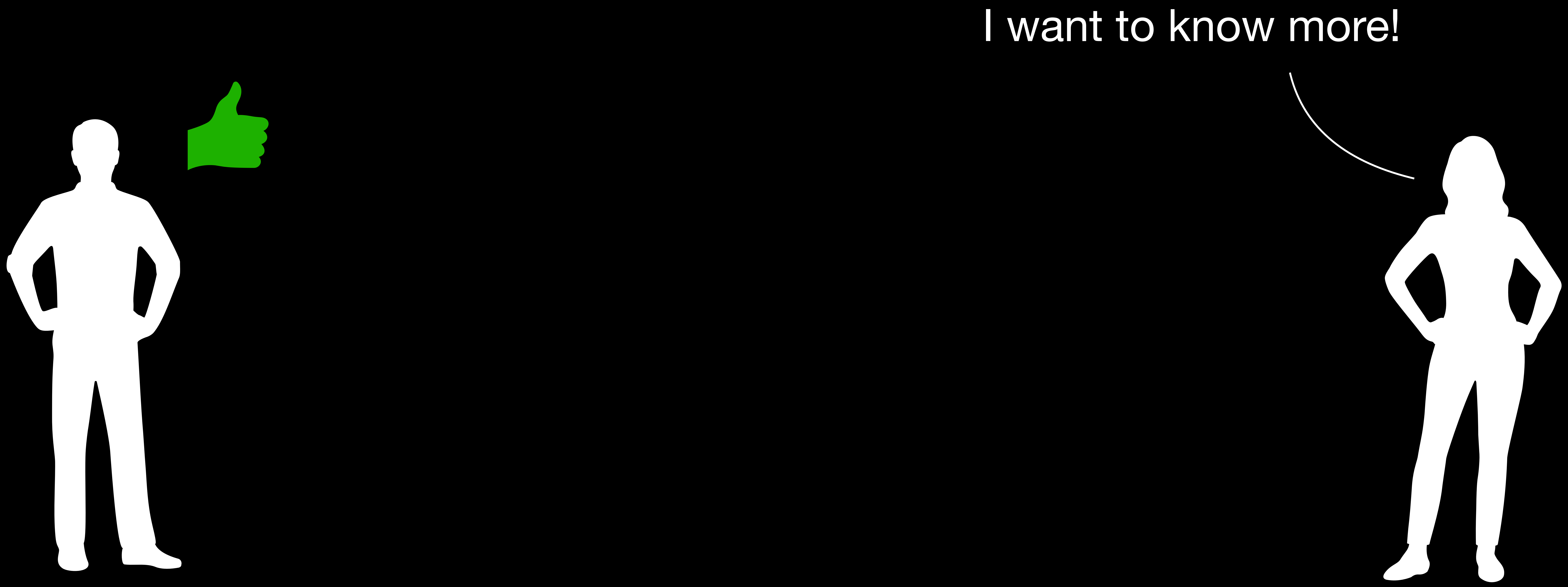
Posterior mean **sim**



Summary

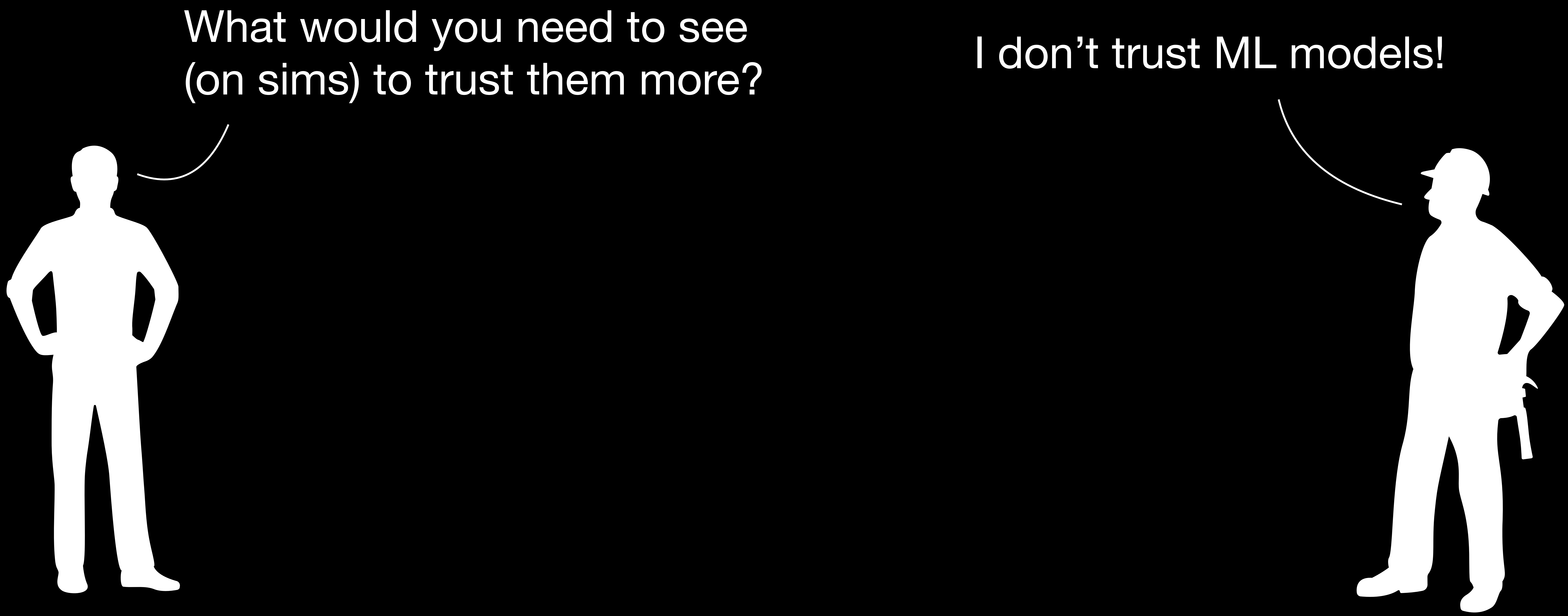
- Combine **flow matching models** + **transformer model** to learn arbitrary **conditionals** and **marginals**
- Add OT + pair loss to close domain gap
- Obtain promising results on simulations

Come find me



Come find me

Come find me


What would you need to see (on sims) to trust them more?
I don’t trust ML models!

Thank you!

Backup