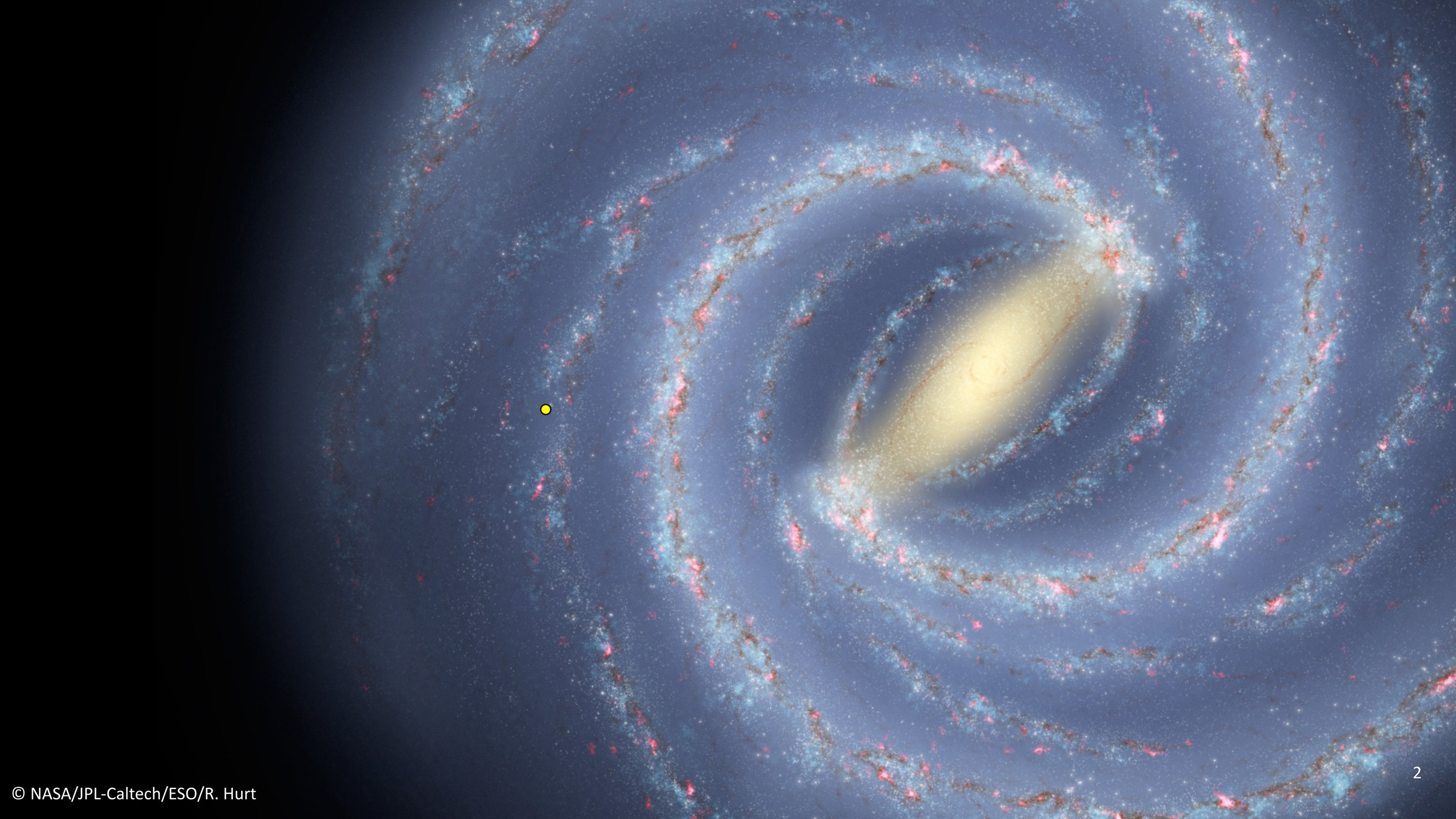
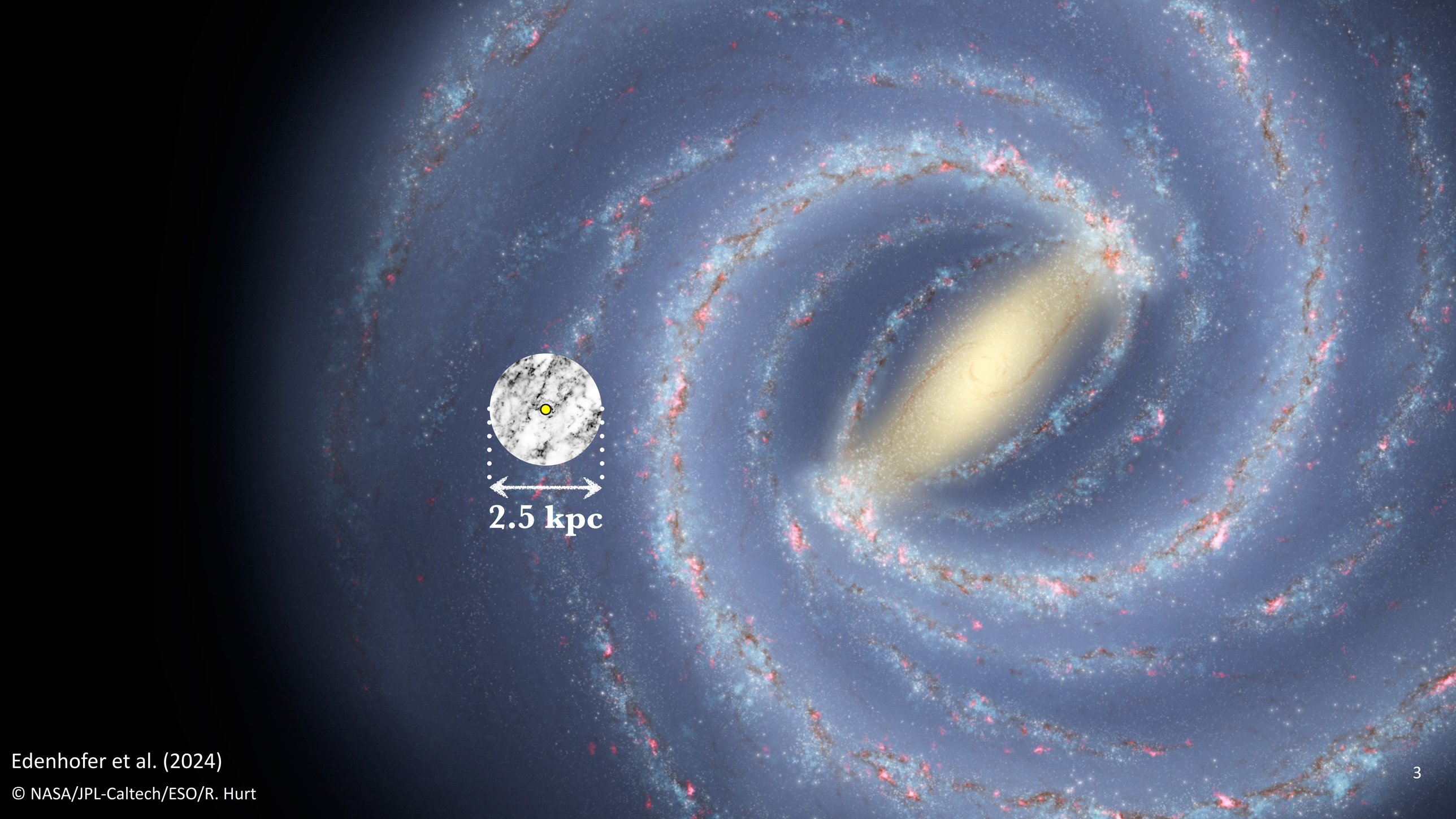
Towards a Complete Census of Young Stars in the local Milky Way

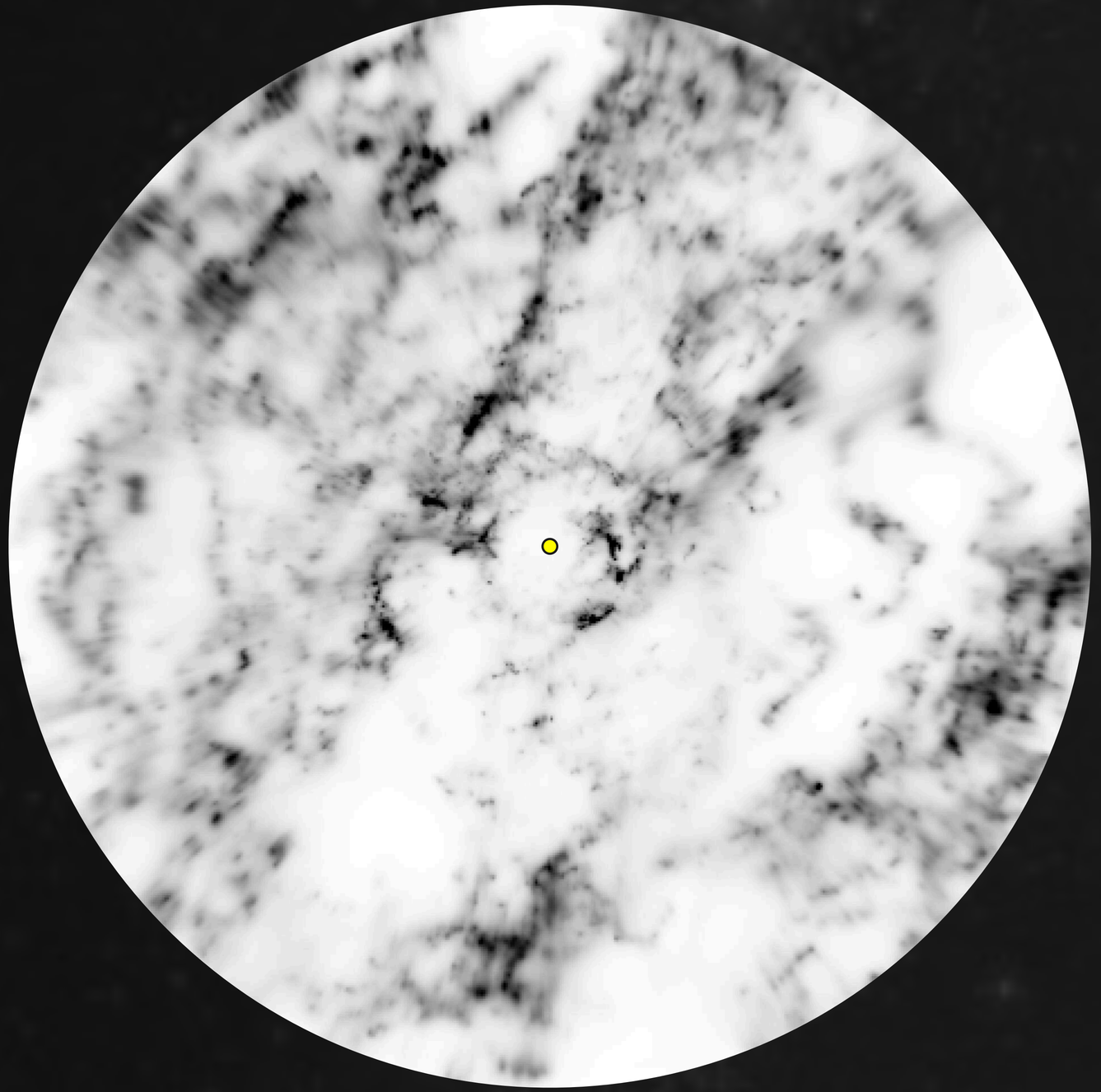
A Self-Consistent Simulation-Based Inference Framework for Incomplete Multi-Survey Data

Sebastian Ratzenböck @CfA
In collaboration with
Catherine Zucker, Joshua Speagle, Phillip Cargile, Philipp Frank, Andrew Saydjari

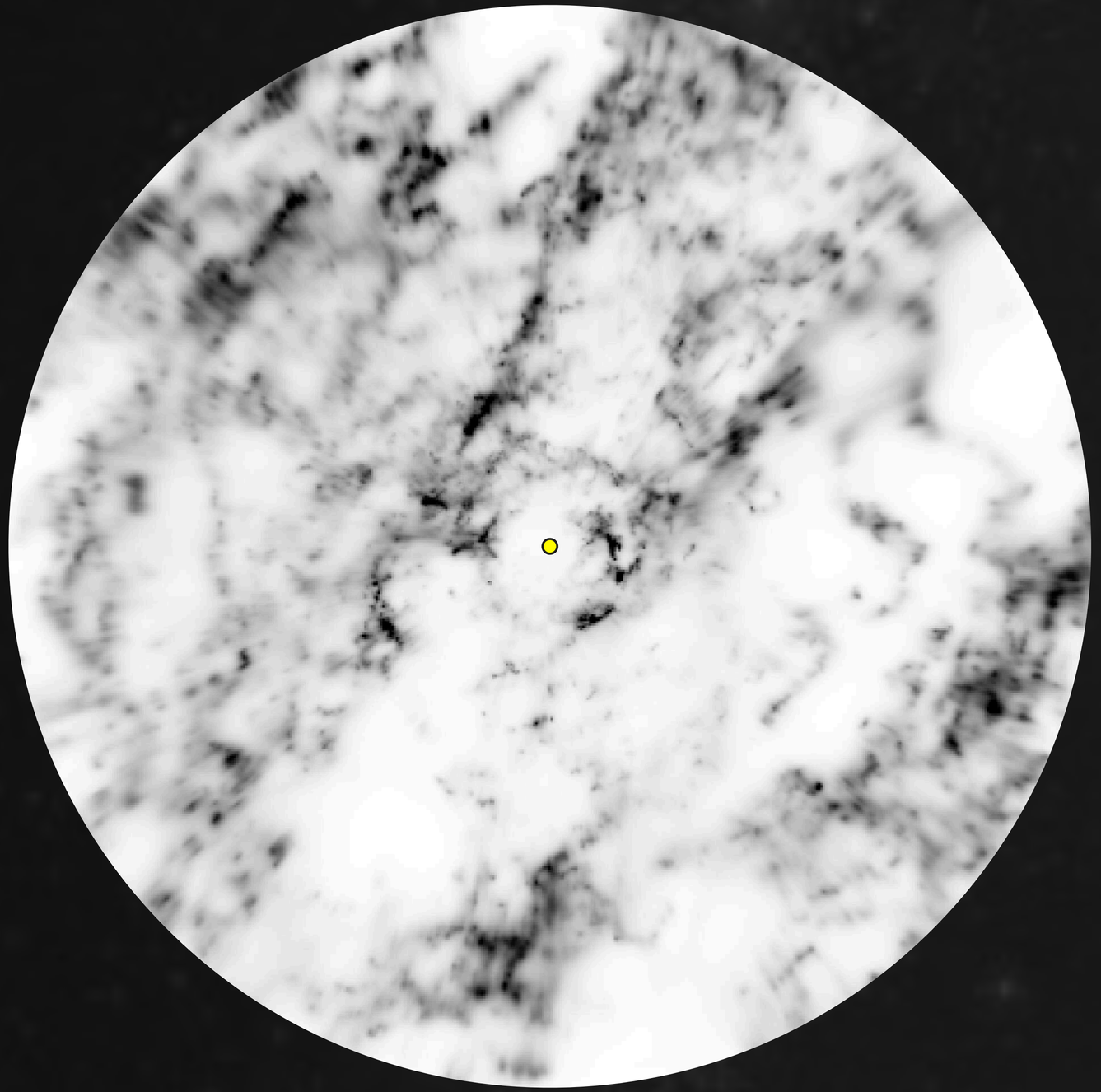




Edenhofer et al. (2024)



Edenhofer et al. (2024)



Galactic rotation

Galactic center

Where are the YSOs?

Edenhofer et al. (2024)

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

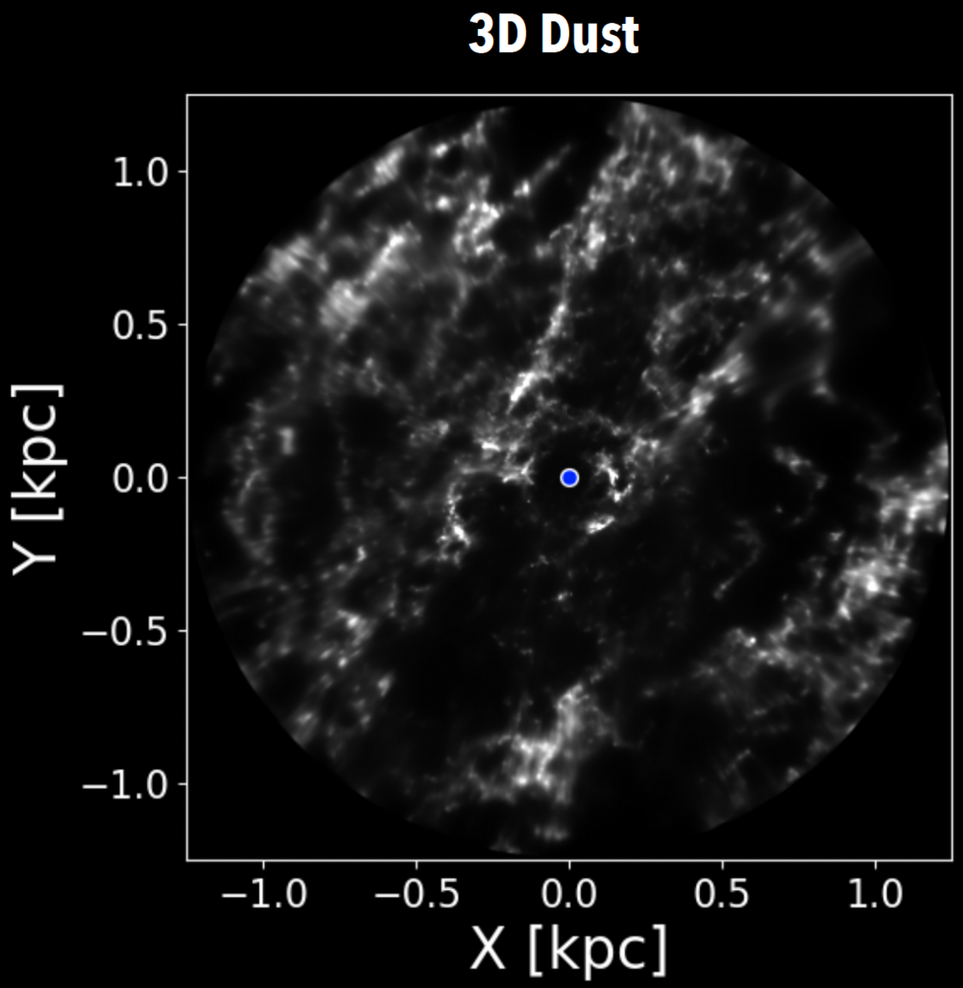
- YSOs connect *cloud* ↔ *stars* ↔ *feedback*
- How does the Milky Way convert gas into stars?
- How do stars leave their birth clouds and shape the ISM?
- How do supernovae regulate, trigger, or suppress new generations of stars?

Understanding Galactic baryon cycle

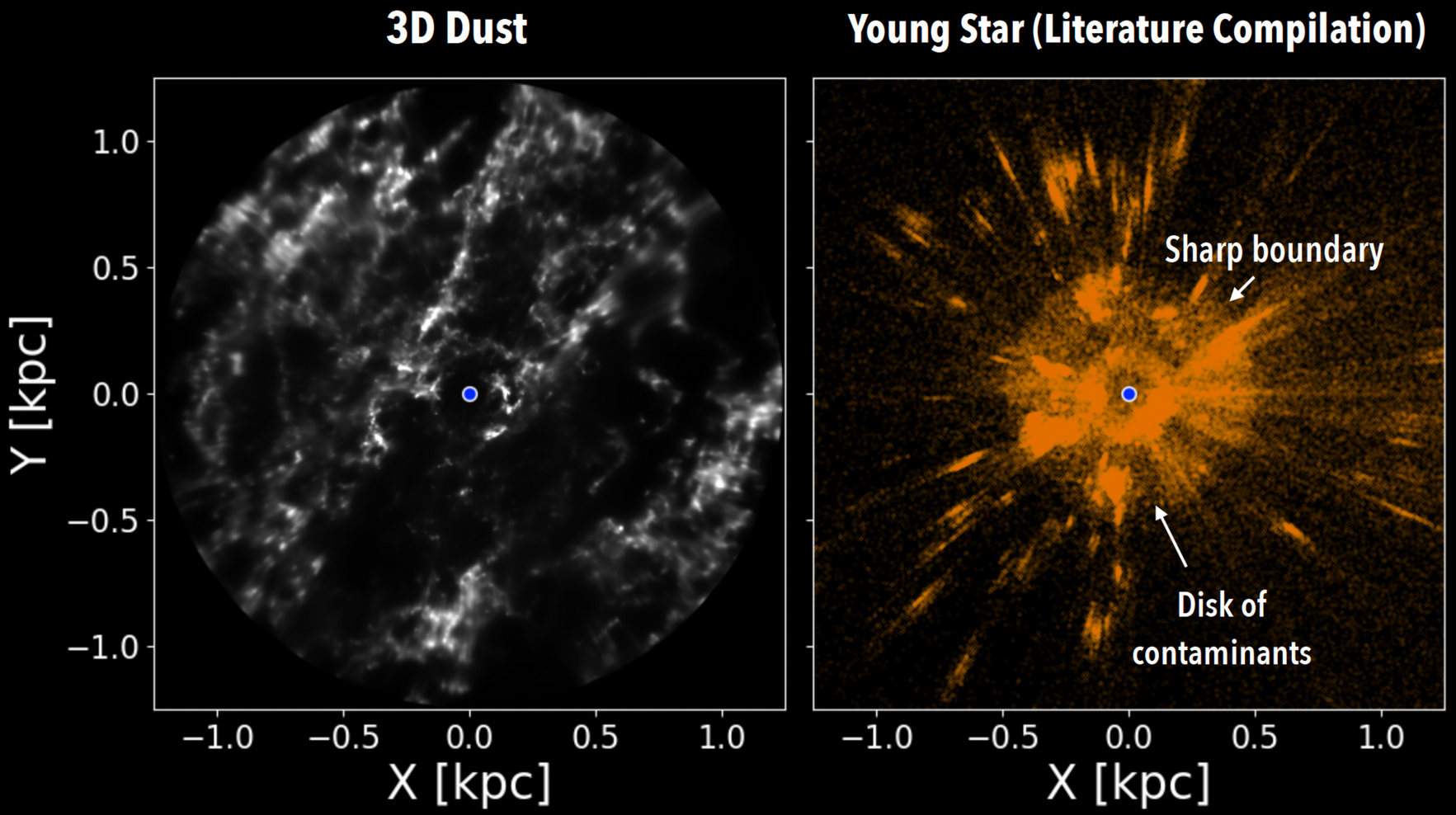
- YSOs connect *cloud* ↔ *stars* ↔ *feedback*
- How does the Milky Way convert gas into stars?
- How do stars leave their birth clouds and shape the ISM?
- How do supernovae regulate, trigger, or suppress new generations of stars?

Goal: study YSOs as function of age in relation to dust

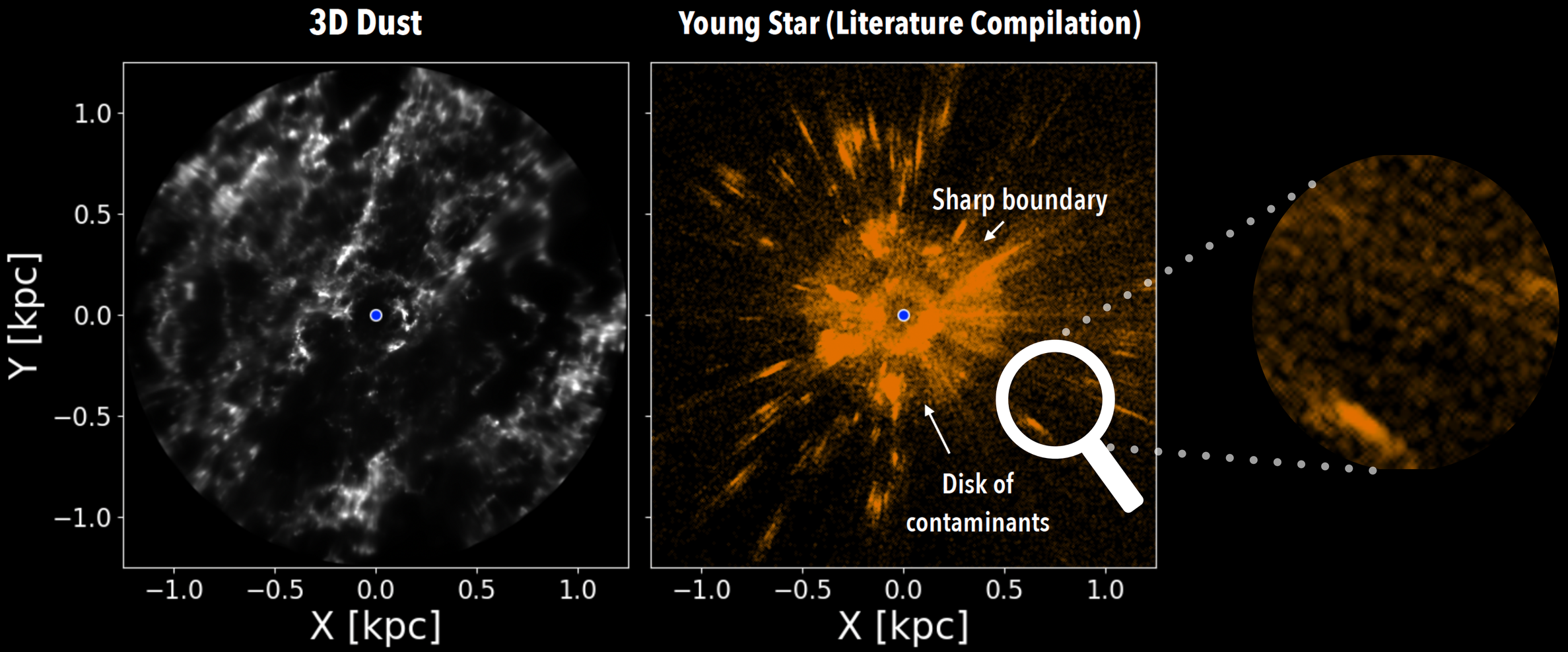
YSOs in the literature



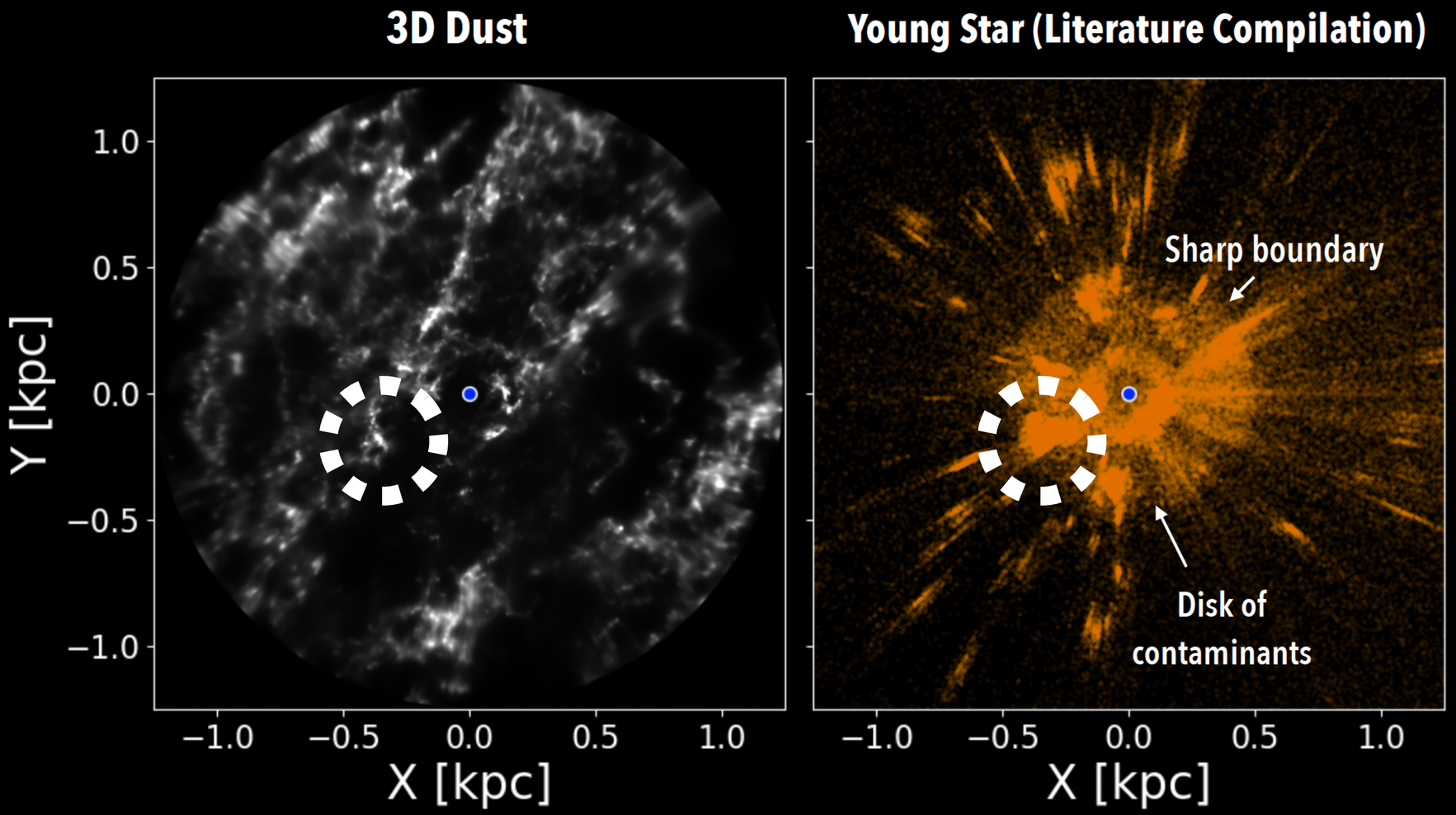
YSOs in the literature



YSOs in the literature



Sample ≙ union of targeted searches



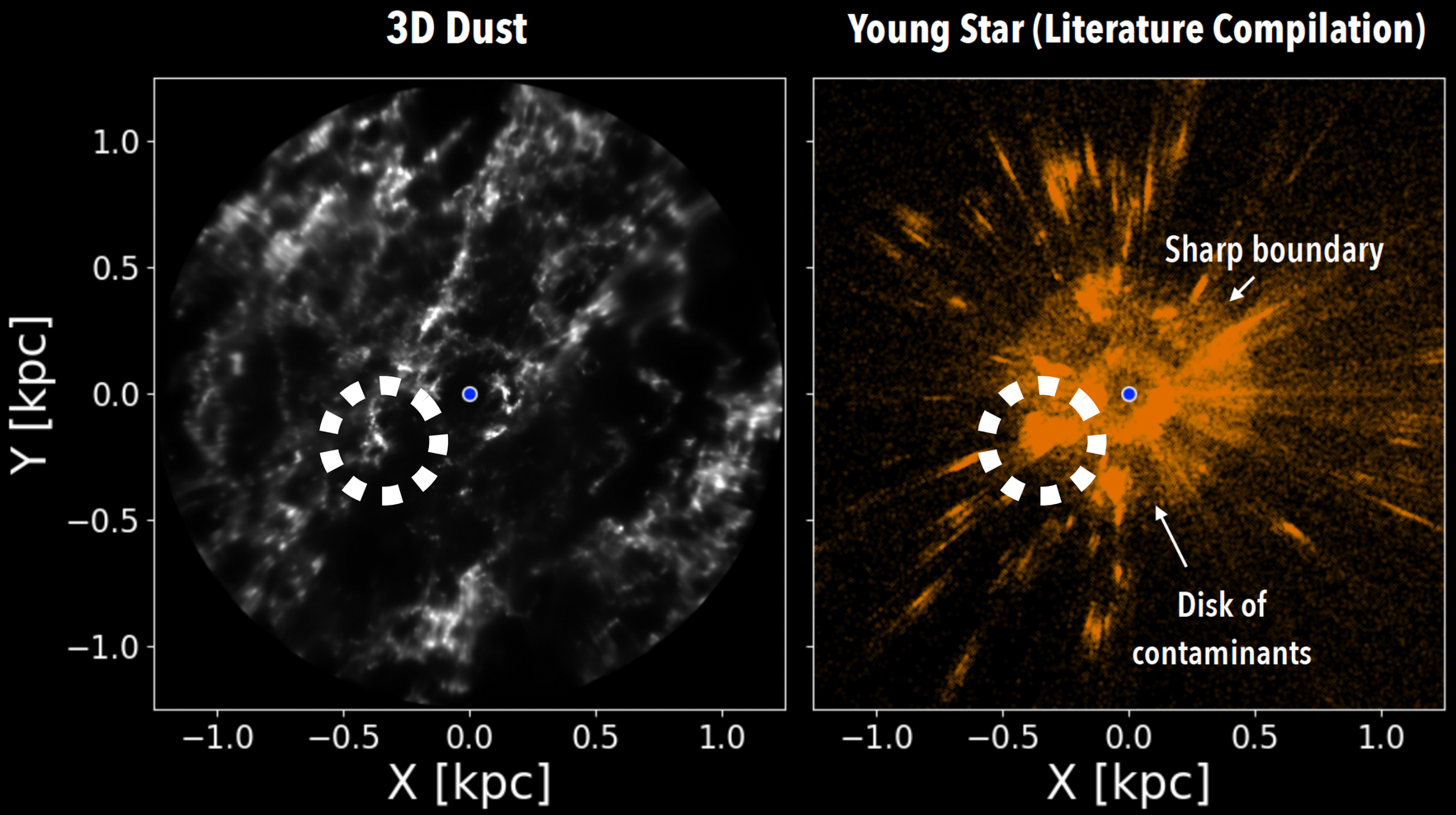
YSOs in Orion SF region

217 papers on YSOs identified in Orion

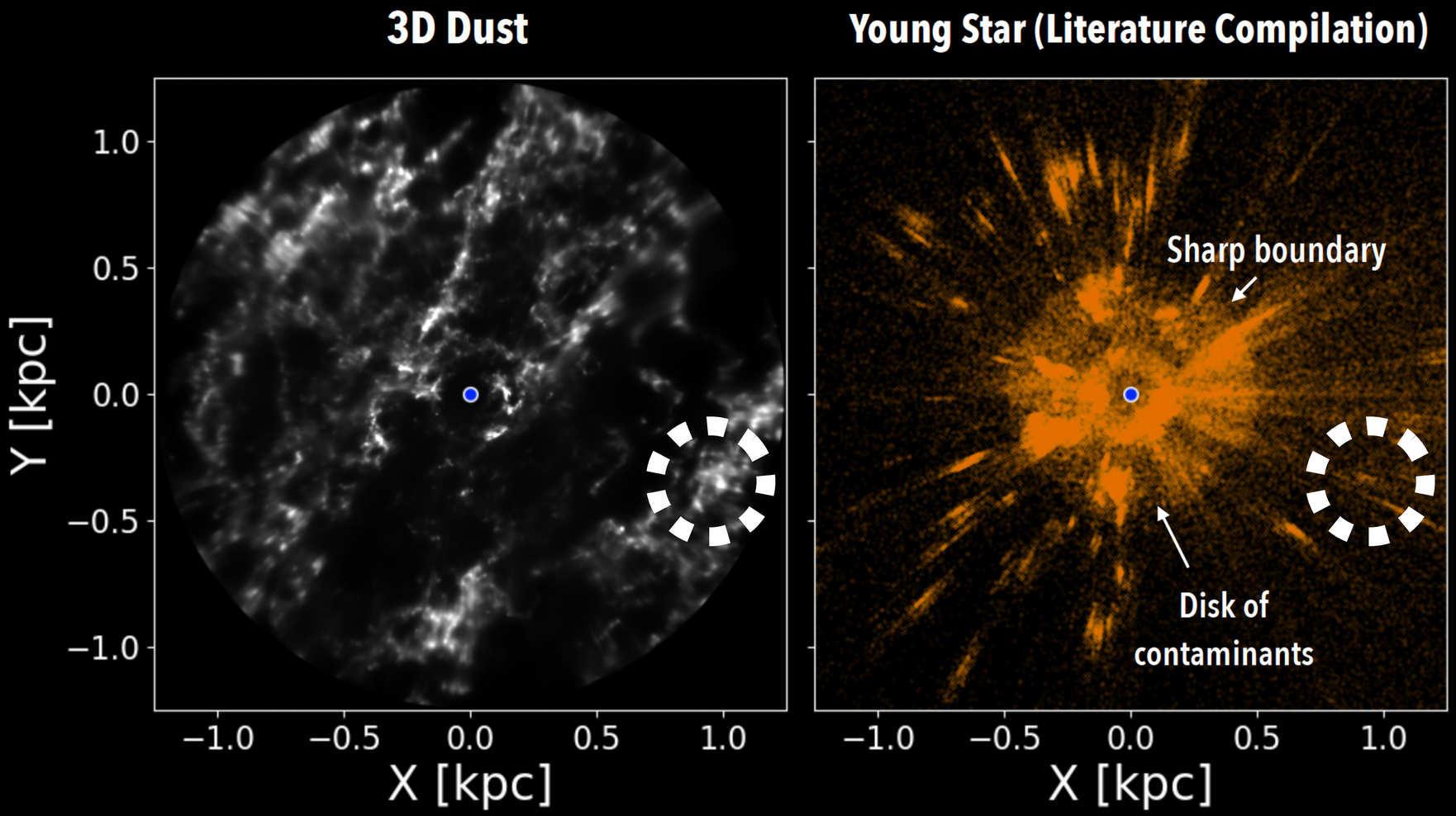
Reference	Data type
Struve (1945)	bin
Popper & Plavec (1976)	bin
Hesser et al. (1976)	bin
Marschall & Mathieu (1988)	bin
Bossi et al. (1989)	bin
Mathieu et al. (1991)	bin
Abt et al. (1991)	bin
Popper (1993)	bin
Gagne & Caillault (1994)	X-ray
Kozlova et al. (1995)	SpT, EW_{acc}, Hα
Gies et al. (1996)	bin
Hillenbrand (1997)	$T_{eff/bol}$, SED
Bolton et al. (1998)	bin
Hillenbrand et al. (1998)	EW_{acc} disc, SED,
Stassun et al. (1999)	Li, RV, Hα, SED
Lloyd & Stickland (1999)	bin
Alcalá et al. (2000)	X-ray, Li, rotation, RV, SpT, Hα, bin
Covino et al. (2000)	bin
Herbst et al. (2000)	rotation, EW_{acc}
Rebull et al. (2000)	SpT, SED
Covino et al. (2001)	bin
Rebull (2001)	rotation, SpT, SED
Dolan & Mathieu (2001)	Li, RV, Hα, bin, SED
Carpenter et al. (2001)	rotation, SED
Rhode et al. (2001)	rotation, bin
Palla & Stahler (2001)	bin
Béjar et al. (2001)	SpT, SED
Kharchenko (2001)	SpT, bin
Herbst et al. (2002)	rotation
Manset & Bastien (2002)	bin
Feigelson et al. (2002)	X-ray
Maheswar et al. (2003)	Li, SpT, Hα, SED
Flaccomio et al. (2003)	X-ray
Skinner et al. (2003)	X-ray
Alcalá et al. (2004)	Li, RV, $T_{eff/bol}$, SpT, Hα, bin
Scholz & Eislöffel (2004)	rotation $T_{eff/bol}$, SpT, logg, EW_{acc}, Hα, SED
Ramírez et al. (2004)	X-ray
Sherry et al. (2004)	SED
Barrado y Navascués et al. (2004)	SpT, Hα, SED
Simon et al. (2004)	X-ray
Ali & Noriega-Crespo (2004)	disc, SED
Stassun et al. (2004)	Li, bin
Baines et al. (2004)	bin
Scholz & Eislöffel (2005)	rotation, SED
Sicilia-Aguilar et al. (2005)	Li, rotation, RV, SpT, Hα, bin
Briceño et al. (2005)	Li, $T_{eff/bol}$, SpT, Hα
Robberto et al. (2005)	bin, SED
Smith et al. (2005)	disc, SED
Getman et al. (2005b)	X-ray
Getman et al. (2005a)	Source list
Kenyon et al. (2005)	Li, RV, $EW_{gravity}$, bin
Burningham et al. (2005)	RV, $EW_{gravity}$, SED
Franciosini et al. (2006)	X-ray, SpT
Köhler et al. (2006)	bin
Herbig & Griffin (2006)	bin
Stassun et al. (2006)	bin
Caballero (2007)	RV, SpT, SED
Barrado Y Navascués et al. (2007)	bin
Reipurth et al. (2007)	SpT, bin
Lee & Chen (2007)	SpT, Hα
Briceño et al. (2007)	Li, RV, Hα
Hernández et al. (2007b)	disc, SED
Barrado y Navascués et al. (2007)	disc, SED
Hernández et al. (2007a)	disc, SED
Irwin et al. (2007)	bin
Caballero (2008)	SED
Stempels et al. (2008)	bin
Caballero & Solano (2008)	SpT, bin, SED
Sacco et al. (2008)	Li, rotation, RV, SpT, EW_{acc}, Hα, bin
López-Santiago & Caballero (2008)	X-ray
Bayo et al. (2008)	$T_{eff/bol}$
Cargile et al. (2008)	bin
Fűrész et al. (2008)	RV, Hα
Luhman et al. (2008b)	SED
Maxted et al. (2008)	RV, $EW_{gravity}$, bin, SED,
Rodríguez-Ledesma et al. (2009)	rotation, SED,
Fang et al. (2009)	Li, SpT, EW_{acc}, Hα, disc, SED,
Biazzo et al. (2009)	rotation, RV, SpT, Hα,
Frasca et al. (2009)	rotation, $T_{eff/bol}$, SpT, SED,
Mace et al. (2009)	bin,
Baldovin-Saavedra et al. (2009)	bin,
Mohanty et al. (2009)	bin,
Tobin et al. (2009)	RV, bin, SED,
Mookerjea & Sandell (2009)	disc, SED,
Hernández et al. (2009)	disc, SED,
Da Rio et al. (2009)	SpT, Hα, SED,
Alecian et al. (2009)	bin,
Parihar et al. (2009)	rotation, Hα, SED,
Connelley & Greene (2010)	SpT, EW_{acc}, veiling, bin, disc, SED,
Grinin et al. (2010)	bin,
Da Rio et al. (2010)	$T_{eff/bol}$, SpT,
Cody & Hillenbrand (2010)	rotation, SED,
Leone et al. (2010)	bin,
Rigliaco et al. (2011)	$T_{eff/bol}$, Hα, disc, SED,
Barrado et al. (2011)	X-ray, disc, SED,
Franciosini & Sacco (2011)	X-ray, SpT, disc,
Bayo et al. (2011)	Li, $T_{eff/bol}$, SpT, $EW_{gravity}$,
Ingleby et al. (2011)	X-ray, SpT, Hα,
van Eyken et al. (2011)	bin,
Morales-Calderón et al. (2011)	rotation,
Béjar et al. (2011)	disc, SED,
Daemgen et al. (2012)	bin,
Caballero et al. (2012)	Li, SpT, EW_{acc}, Hα, disc,
Bayo et al. (2012)	rotation, SpT, EW_{acc}, Hα, disc,
Rigliaco et al. (2012)	SpT, EW_{acc}, Hα,
Kryukova et al. (2012)	disc, SED,
Megeath et al. (2012)	disc, SED,
Gómez Maqueo Chew et al. (2012)	bin,
Da Rio et al. (2012)	$T_{eff/bol}$,
Hsu et al. (2012)	Li, SpT, EW_{acc}, Hα, SED,
Morales-Calderón et al. (2012)	rotation, SpT, bin,
Peña Ramírez et al. (2012)	SED,
Manara et al. (2012)	Hα, disc,
Correia et al. (2013)	bin,
Hillenbrand et al. (2013)	SpT, bin,
Hsu et al. (2013)	Li, RV, SpT, disc, SED,
Stutz et al. (2013)	$T_{eff/bol}$, bin, disc, SED,
Windemuth et al. (2013)	bin,
Pillitteri et al. (2013)	X-ray,
Kim et al. (2013)	X-ray, $T_{eff/bol}$, SpT, bin, disc,
Fang et al. (2013)	Li, SpT, EW_{acc}, Hα, disc, SED,
Szegedi-Elek et al. (2013)	Hα, SED,
Kuhn et al. (2013)	SED,
Povich et al. (2013)	Source list,
Broos et al. (2013)	X-ray, disc,
López-García et al. (2013)	X-ray,
Sergison et al. (2013)	Li, RV, veiling, Hα,
Bouy et al. (2014)	SED,
Pettersson et al. (2014)	Hα, bin, SED,
Fang et al. (2014)	bin,
Sanchez et al. (2014)	$T_{eff/bol}$, SED,
Ingraham et al. (2014)	$T_{eff/bol}$, SpT,
Kuhn et al. (2014)	SpT, bin,
Hernández et al. (2014)	Li, RV, SpT, $EW_{gravity}$, Hα, bin, disc, SED,
Theissen & West (2014)	RV, $T_{eff/bol}$, SpT, $EW_{gravity}$, EW_{acc}, Hα,
Downes et al. (2014)	SpT, Hα,
Spezzi et al. (2015)	disc, SED,
Koenig et al. (2015)	SpT, EW_{acc}, Hα, disc, SED,
Rice et al. (2015)	rotation, bin, SED,
Downes et al. (2015)	$T_{eff/bol}$, SpT, Hα, disc, SED,
Meingast et al. (2016)	SED,
Hasenberger et al. (2016)	SpT, disc,
Megeath et al. (2016)	disc, SED,
Karim et al. (2016)	rotation,
Da Rio et al. (2016)	rotation, $T_{eff/bol}$, logg, SED,
Kounkel et al. (2016a)	Li, RV, $T_{eff/bol}$, bin,
Kounkel et al. (2016b)	bin, disc,
Lewis & Lada (2016)	bin, disc,
Furlan et al. (2016)	$T_{eff/bol}$, disc, SED,
Kim et al. (2016)	Li, $T_{eff/bol}$, SpT, Hα, bin, disc,
Messina et al. (2016)	bin,
Pillitteri et al. (2017)	X-ray,
Fang et al. (2017)	Li, SpT, Hα, disc,
Suárez et al. (2017)	Li, $T_{eff/bol}$, SpT, Hα, disc,
Kounkel et al. (2017a)	Li, SpT, Hα,
Kounkel et al. (2017b)	bin,
Simon & Toraskar (2017)	bin,
Kounkel et al. (2017c)	Li, RV, $T_{eff/bol}$, Hα, disc,
Jaehnig et al. (2017)	bin,
Getman et al. (2017)	X-ray, SED,
Messina et al. (2017)	SED,
Fernandez et al. (2017)	bin,
GRAVITY Collaboration et al. (2018)	bin,
Vioque et al. (2018)	$T_{eff/bol}$, Hα, bin, disc, SED,
Zari et al. (2018)	RV,
Kounkel et al. (2018)	RV, $T_{eff/bol}$, logg,
Caballero et al. (2018)	bin,
Grant et al. (2018)	SpT, disc, SED,
Yao et al. (2018)	$T_{eff/bol}$, logg, EW_{acc}, disc, SED,
Cottle et al. (2018)	disc, SED,
Davies et al. (2018)	bin,
Getman et al. (2018)	bin,
Getman et al. (2018)	disc,
Jayasinghe et al. (2018)	rotation,
Duchêne et al. (2018)	bin,
Großschedl et al. (2019)	disc, SED,
Jerabkova et al. (2019)	bin,
Caballero et al. (2019)	Li, SpT, EW_{acc}, Hα,
Briceño et al. (2019)	Li, $T_{eff/bol}$, SpT, $EW_{gravity}$, Hα, SED,
Kim et al. (2019)	bin, SED,
Kounkel et al. (2019)	RV, $T_{eff/bol}$, logg, veiling, bin, disc,
Mairs et al. (2019)	bin,
McBride & Kounkel (2019)	RV,
Tobin et al. (2019)	bin,
De Furio et al. (2019)	bin,
Suárez et al. (2019)	SED,
Bouma et al. (2020)	bin,
Tokovinin et al. (2020)	bin,
Kounkel et al. (2020)	Source list,
Manzo-Martínez et al. (2020)	SpT, Hα,
Robberto et al. (2020)	$T_{eff/bol}$,
Strampelli et al. (2020)	bin,
Fischer et al. (2020)	SED,
Jackson et al. (2020)	RV, $T_{eff/bol}$, logg, SED,
Lavail et al. (2020)	bin,
Frasca et al. (2021)	Source list,
Pinzón et al. (2021)	rotation, RV, $T_{eff/bol}$, SpT, bin, disc,
Fang et al. (2021)	SpT, veiling, bin,
Habel et al. (2021)	disc,
Laos et al. (2021)	EW_{acc},
Serna et al. (2021)	rotation, $T_{eff/bol}$, bin, SED,
Arun et al. (2021)	bin,
Kos et al. (2021)	Li, rotation, RV, $T_{eff/bol}$, logg,
Proffitt et al. (2021)	bin,
Franciosini et al. (2022)	RV, $T_{eff/bol}$,
Flaischlen et al. (2022)	$T_{eff/bol}$,
Thanathibodee et al. (2022)	$T_{eff/bol}$, SpT, EW_{acc}, Hα, bin,
Kounkel et al. (2022)	rotation,
Pittman et al. (2022)	rotation,
Cao et al. (2022)	$T_{eff/bol}$, disc,
De Furio et al. (2022b)	bin,
Theissen et al. (2022)	rotation, RV, $T_{eff/bol}$, veiling, bin,
De Furio et al. (2022a)	bin,
Hernández et al. (2023)	Li, SpT, Hα,
Federman et al. (2023)	$T_{eff/bol}$, disc,
Damian et al. (2023)	$T_{eff/bol}$, SpT, SED,
Smith et al. (2023)	rotation, $T_{eff/bol}$, disc,

Roquette+2025

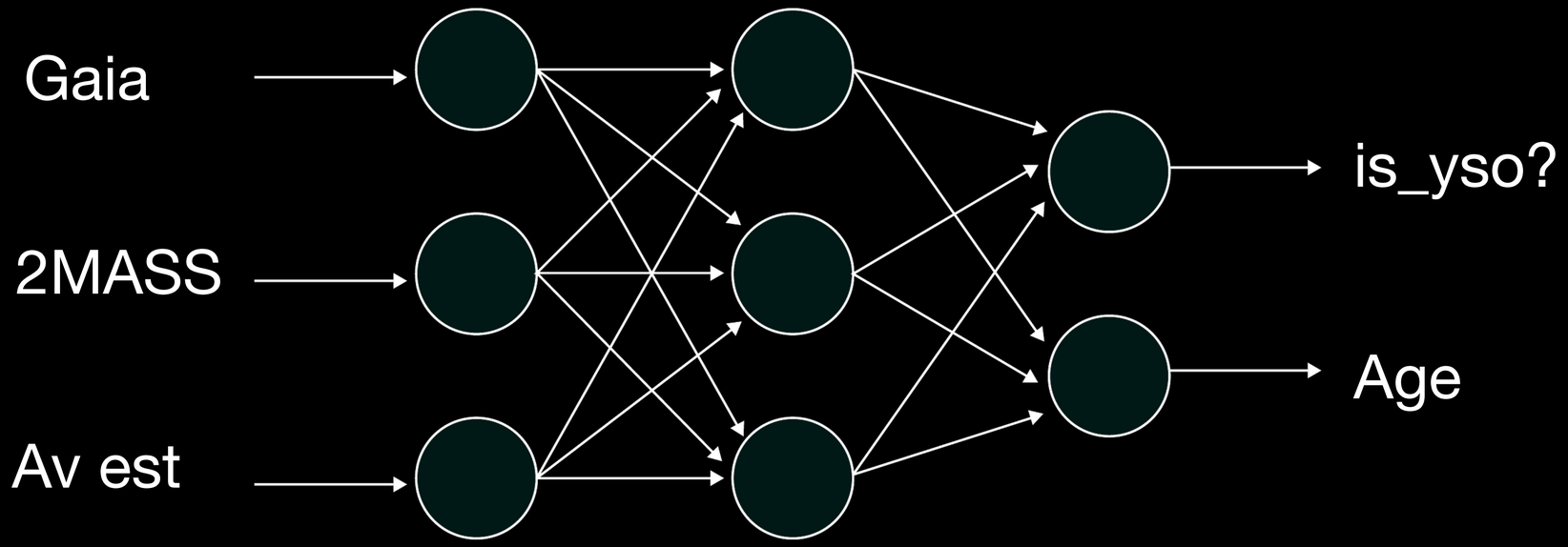
Inhomogenious selection function



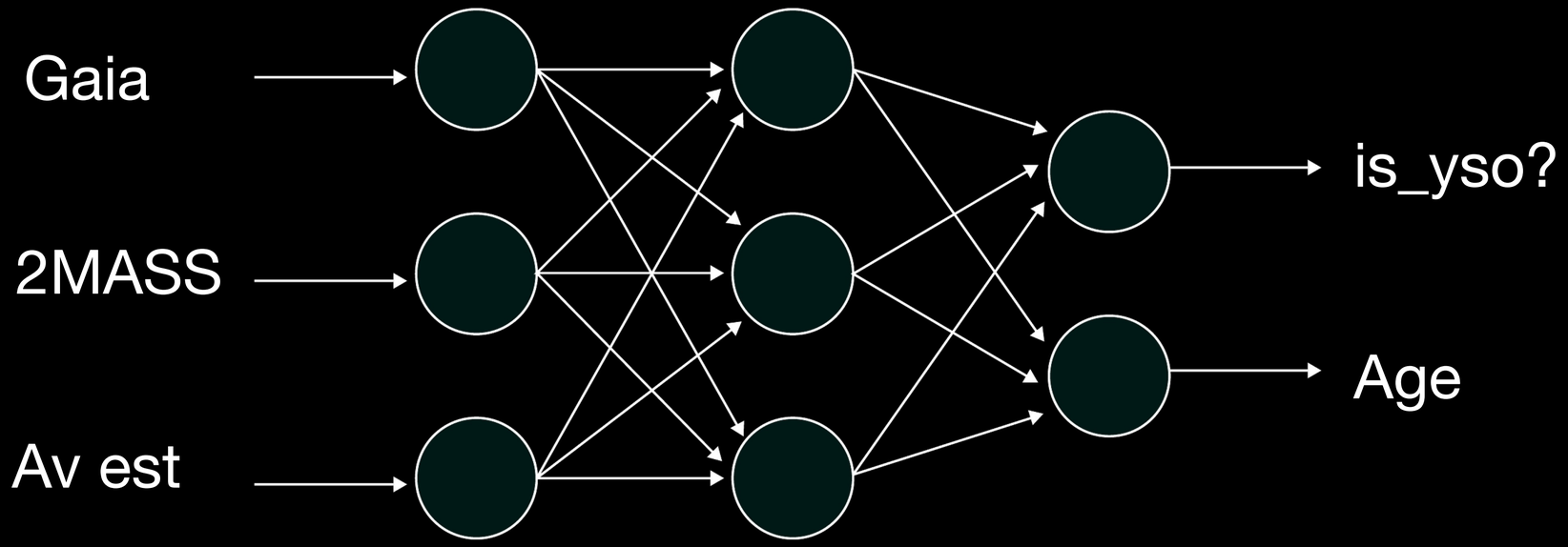
Inhomogenious selection function

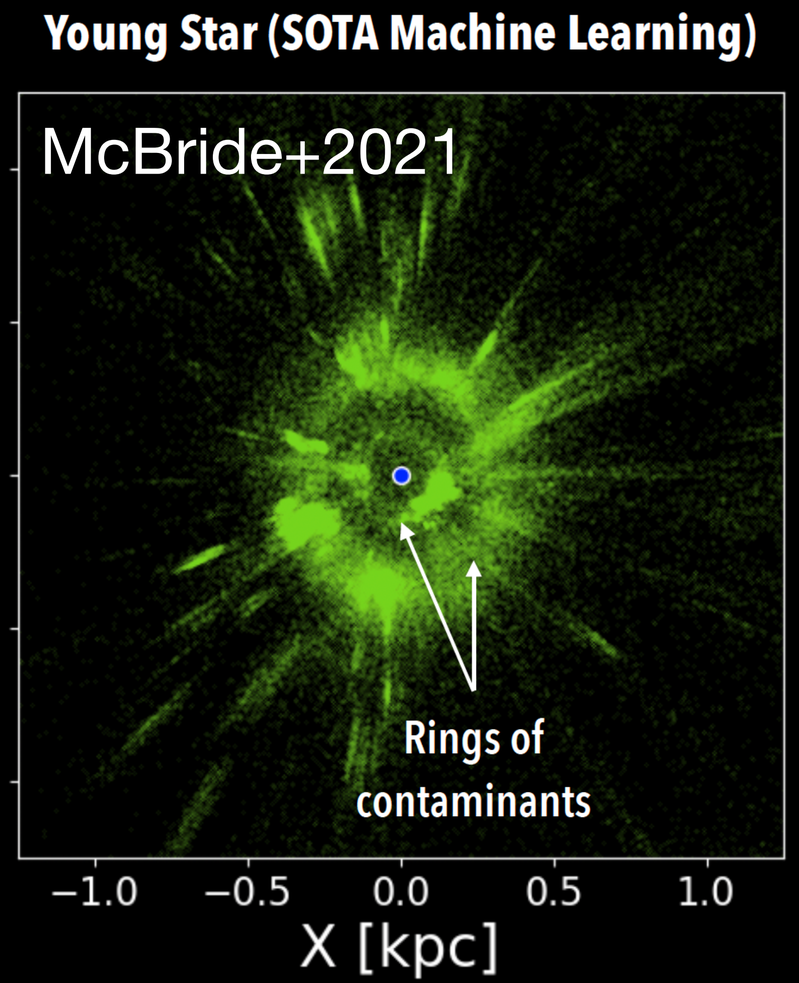


Can ML help?

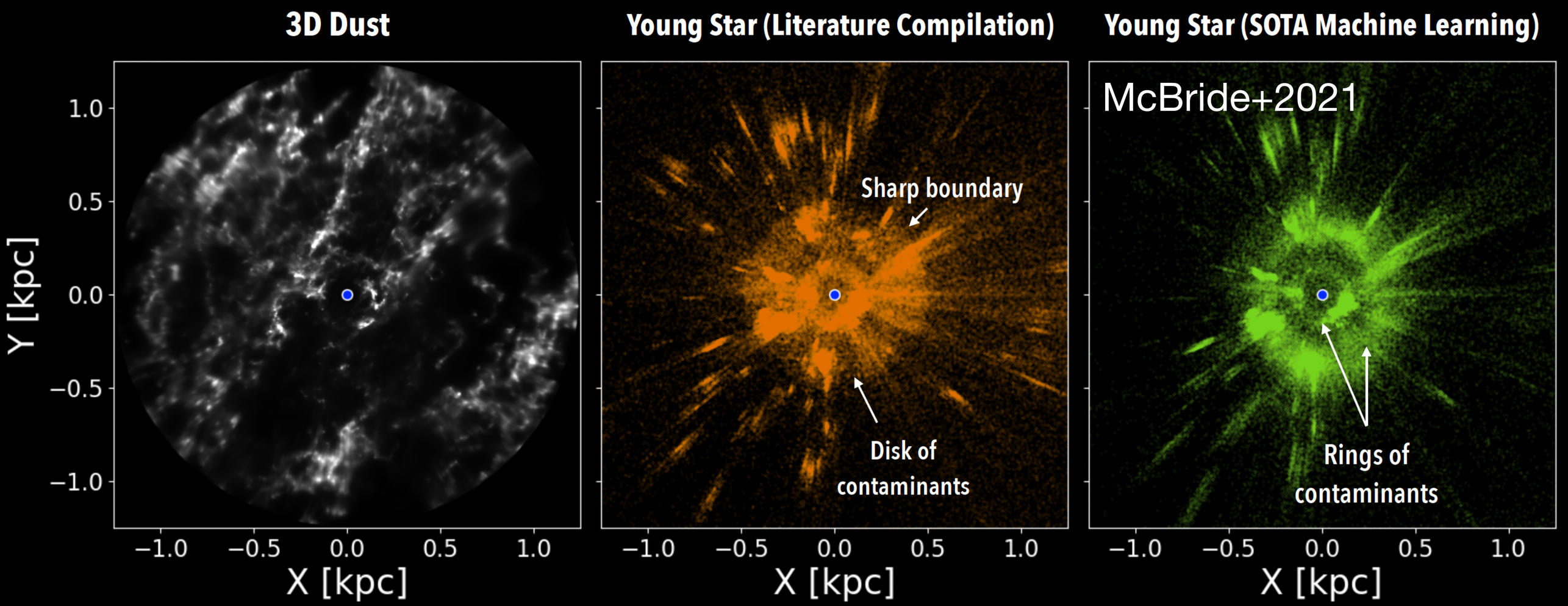


Can ML help?


Gaia
2MASS
Av est
is_yso?
Age


Young Star (SOTA Machine Learning)
McBride+2021
Rings of
contaminants
−1.0
−0.5
0.0
0.5
1.0
X [kpc]

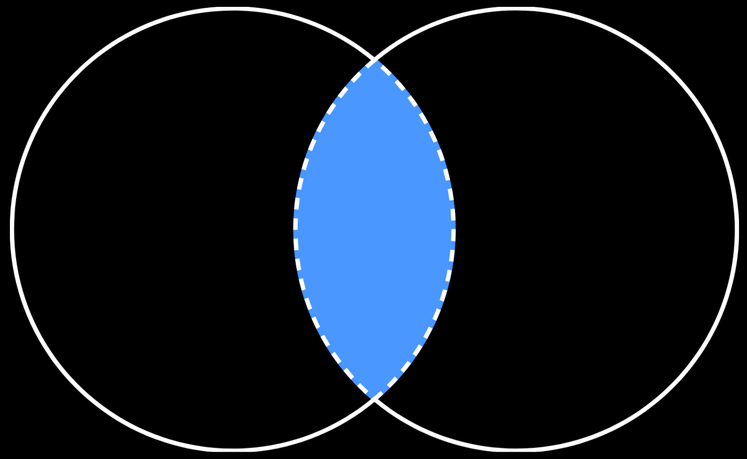
Can ML help?


3D Dust
Young Star (Literature Compilation)
Young Star (SOTA Machine Learning)
McBride+2021
Sharp boundary
Disk of contaminants
Rings of contaminants
Y [kpc]
X [kpc]
X [kpc]
X [kpc]
1.0
0.5
0.0
−0.5
−1.0

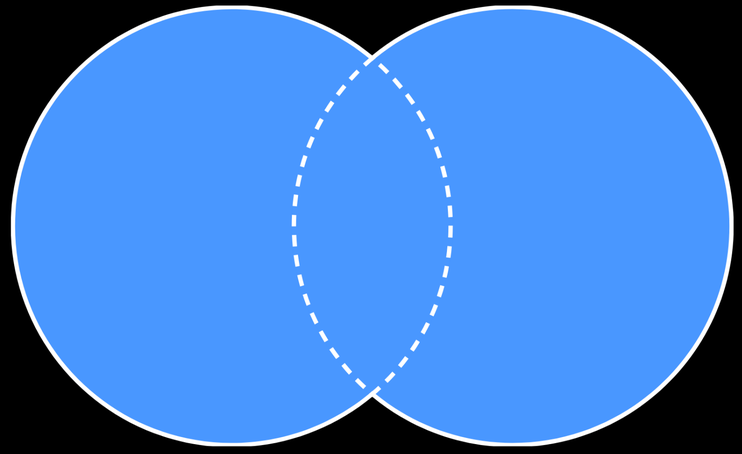
Ways to improve ML YSO catalog

Ways to improve ML YSO catalog

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

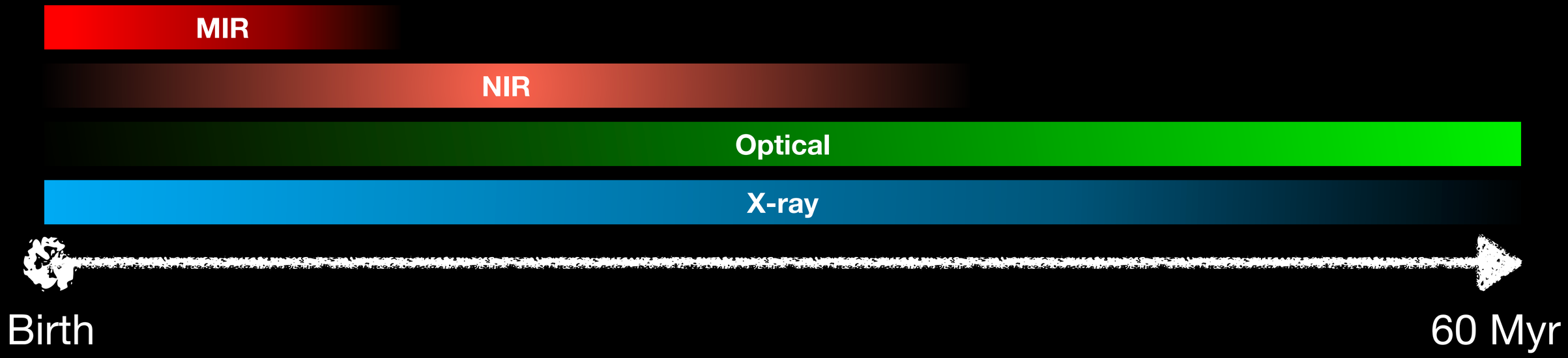


Often works focus on intersection



Aim to be truely multi-survey

Untapped potential of current methods



Many different YSO tracers data souces

- **IR-excess** from disk
- **Lithium Depletion**: Li I (λ6708Å) line
- **Gravity tracers** (weaker *logg* comp. to MS star)
  - Either med to high-R spectra or EW of absorption lines from *logg* (e.g., *Na I, K I, TiO, CaH 3*)
- **Emission lines** from accretion process
  - **Hα line** - indicated also by increased chromospheric activity, or Balmer (Opt), Paschen or Brackett (IR), He I, Ca II near-IR triplet
- **Optical and IR veiling**
  - Excess flux from high-temperature material in inner disc regions
  - dilutes photospheric spectral lines & enhances spectral continuum
- **Variability**
  - Seen e.g. in larger mean flux uncertainties in multi-epoch photometry
- **Increased X-ray emission** & stellar rotation

Ways to improve ML YSO catalog

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

Gaia
2MASS
WISE
LAMOST

WISE
Spitzer
APOGEE

Ways to improve ML YSO catalog

- Data fusion: use as many informative data sets as possible
- Provide well-calibrated posteriors over stellar parameters given spectra & photometric observations

Ways to improve ML YSO catalog

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

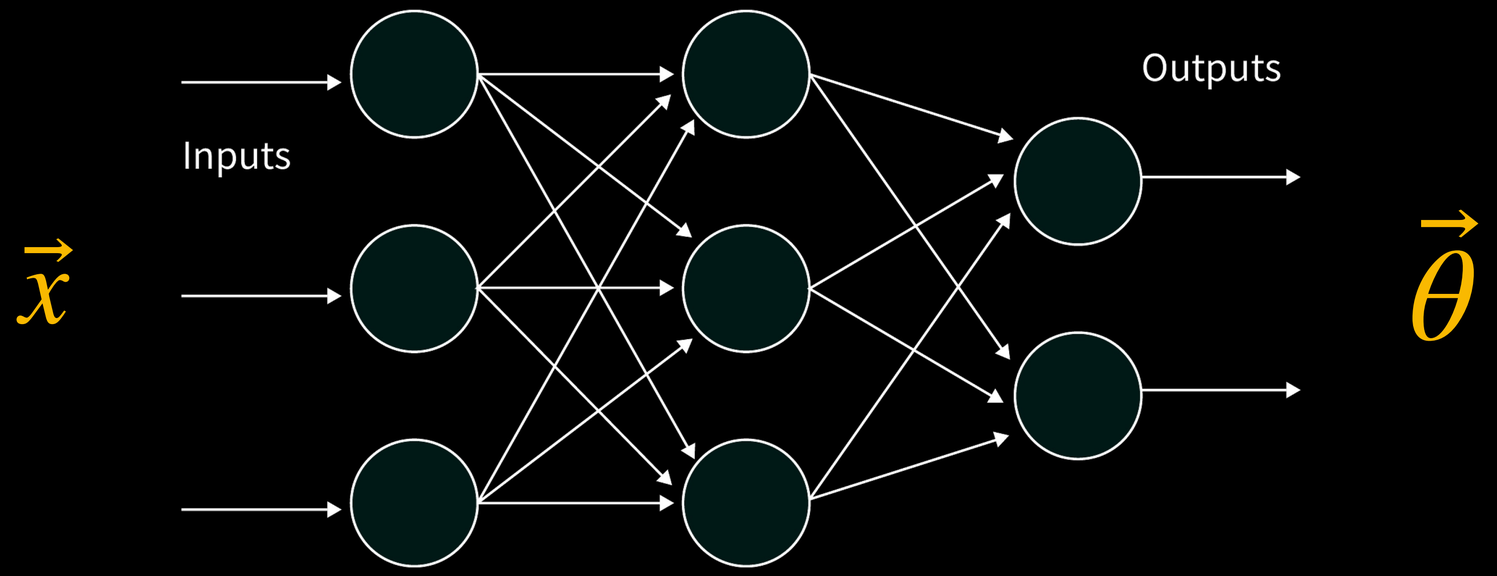
→ *Simulation-based inference for **incomplete**, **multi-survey** data*

Model implementation

I. SBI model

Typical ML regression

Typical ML regression

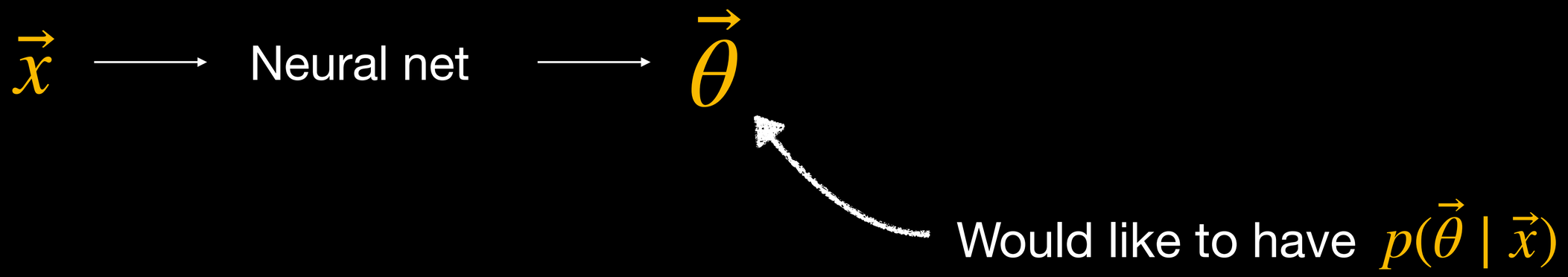


Series of ***learnable*** affine transformations of $\vec{x}$

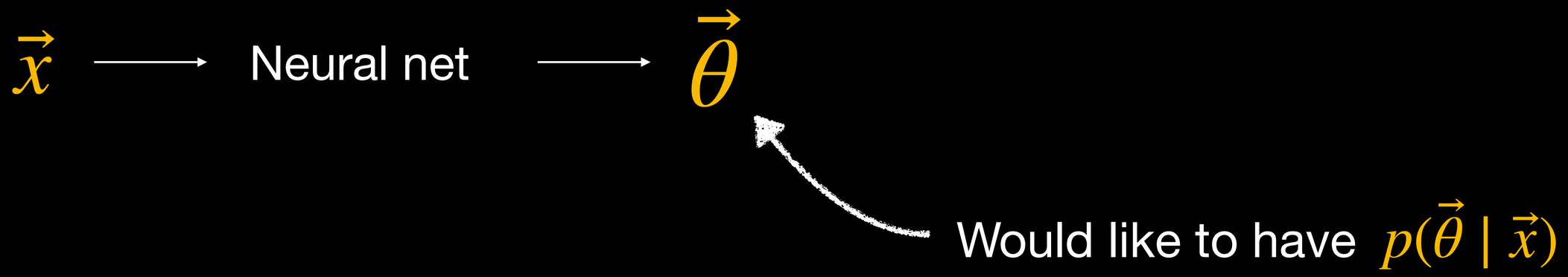
followed by pointwise non-linear map: $f_\phi(\vec{x}) = \hat{\theta}$

ϕ…learnable parameters

Typical ML regression



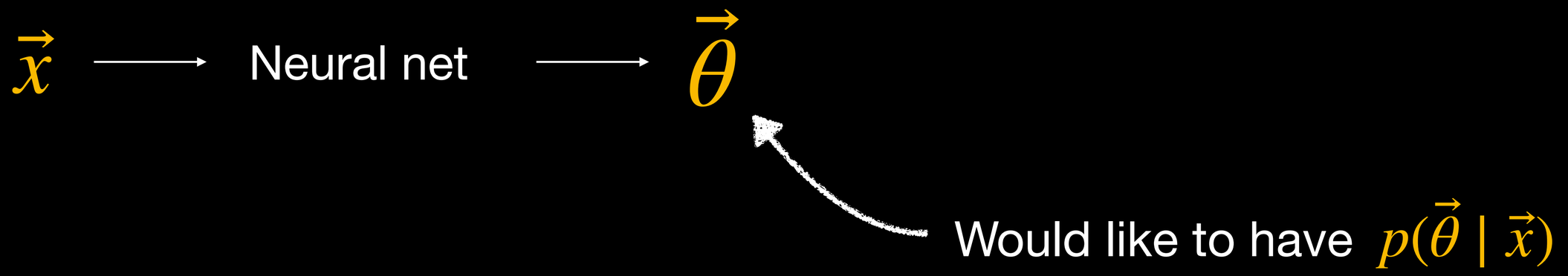
Typical ML regression



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

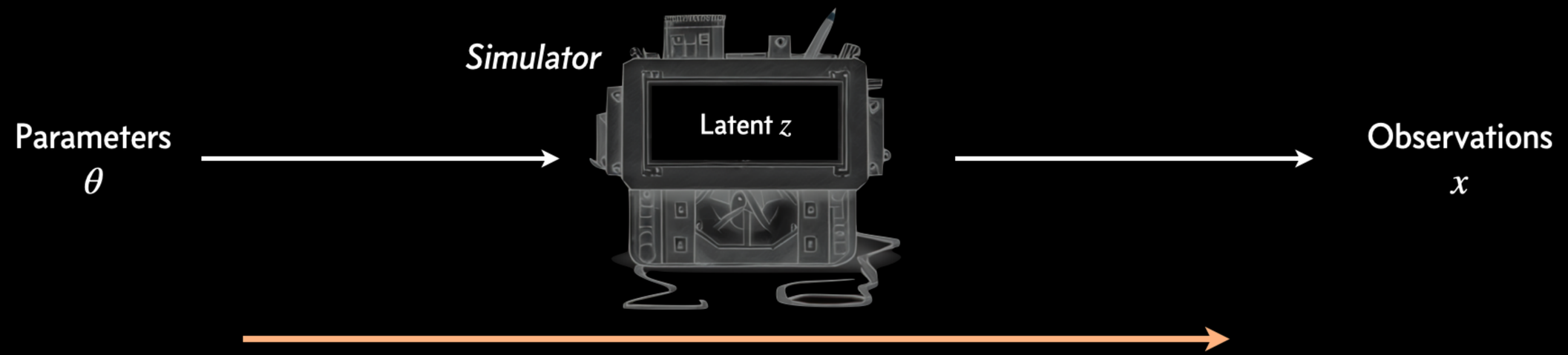


However: — $p(\vec{x} \mid \vec{\theta})$ might not be tractable

— $p(\vec{\theta} \mid \vec{x})$ 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



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

ML regression

Neural posterior estimation

$\vec{x}$ ⟶ Neural net (NN) ⟶ $\vec{\theta}$

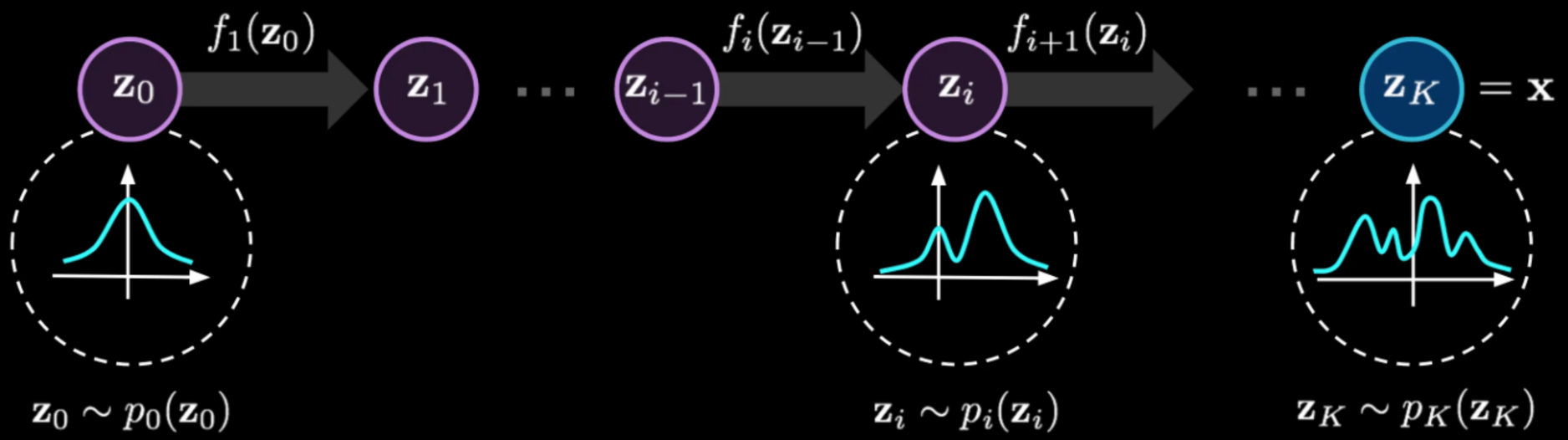
ML regression

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

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

Normalizing flows

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

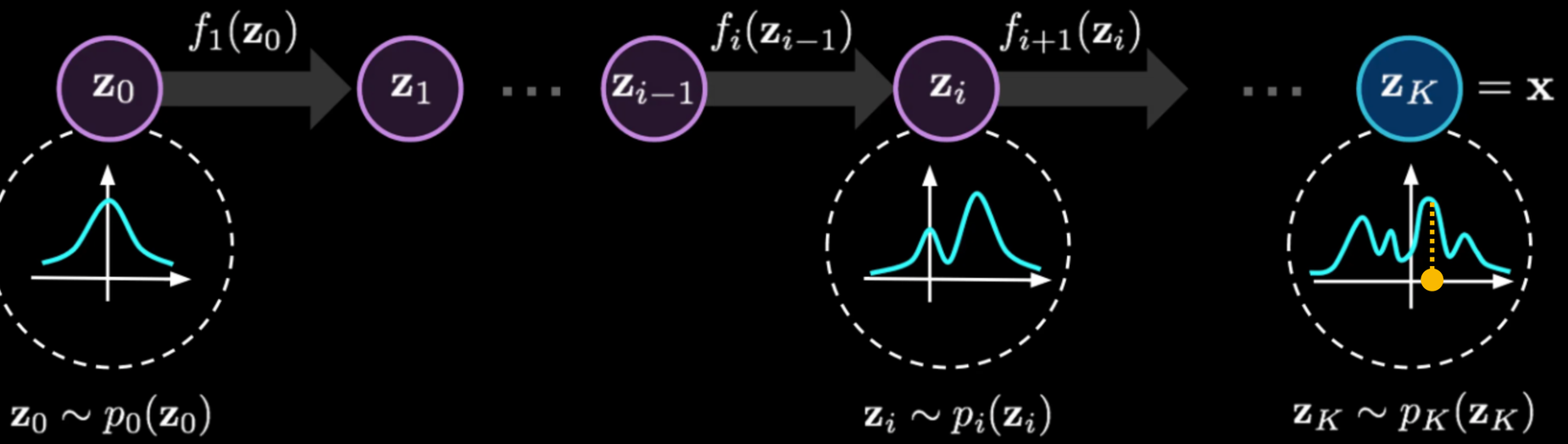


Parameterized, invertible maps f_i that transform Gaussian into target distribution

Training objective: ***maximum likelihood***

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

$\vec{x}$ ⟶ Neural net (NN) ⟶ $\vec{\theta}$

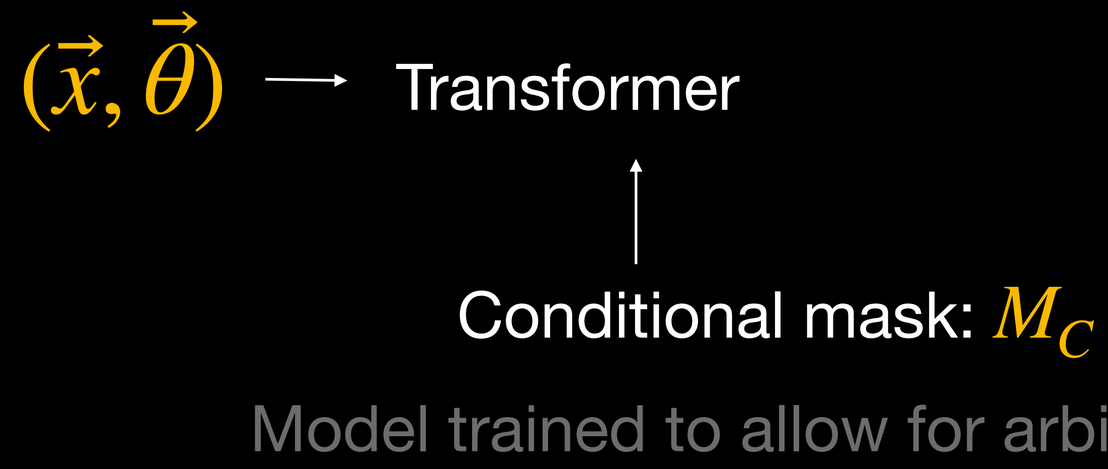
ML regression

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

Cannot handle missing data

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

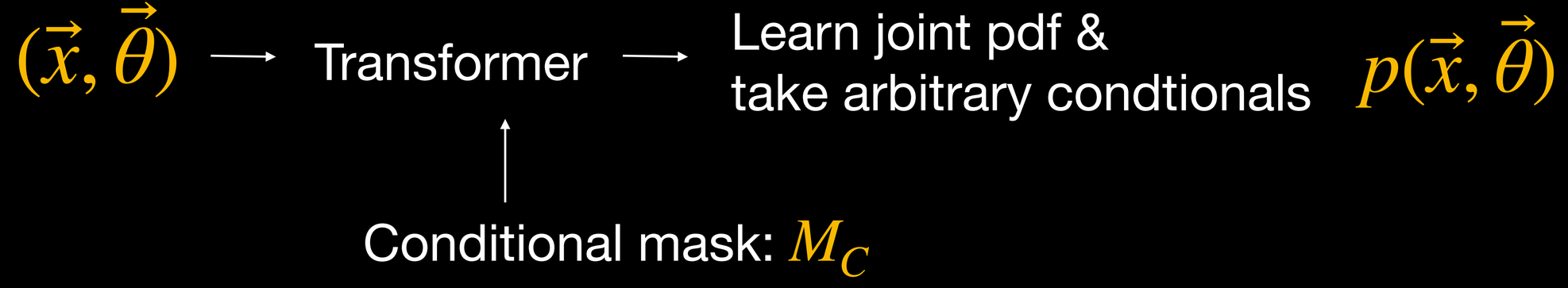
Transformer: learning with incomplete data



Model trained to allow for arbitrary conditioning

Gloeckler et al. (2024)

Transformer: learning with incomplete data

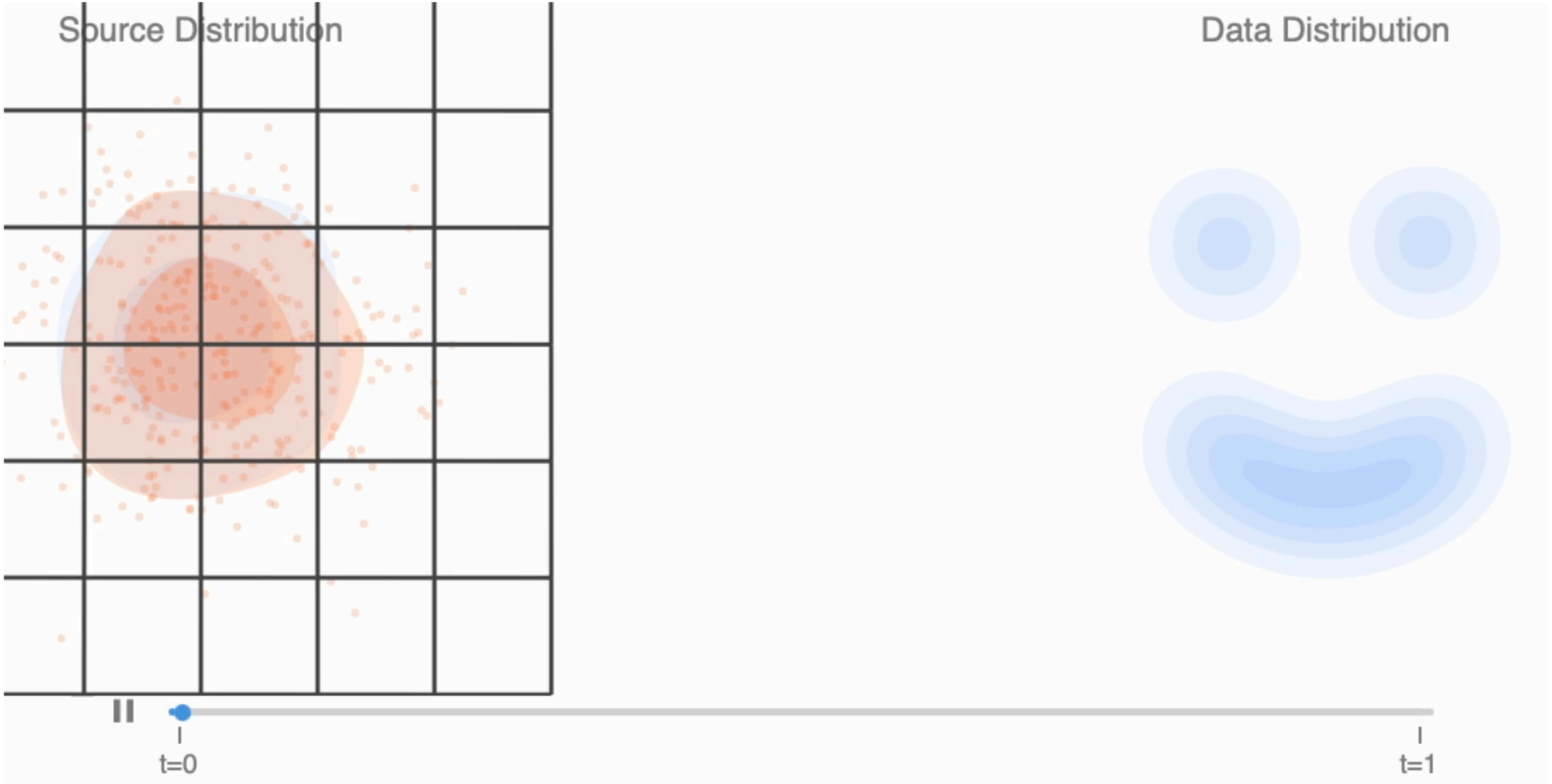


Model trained to allow for arbitrary conditioning

— condition joint only on actually observed quantities

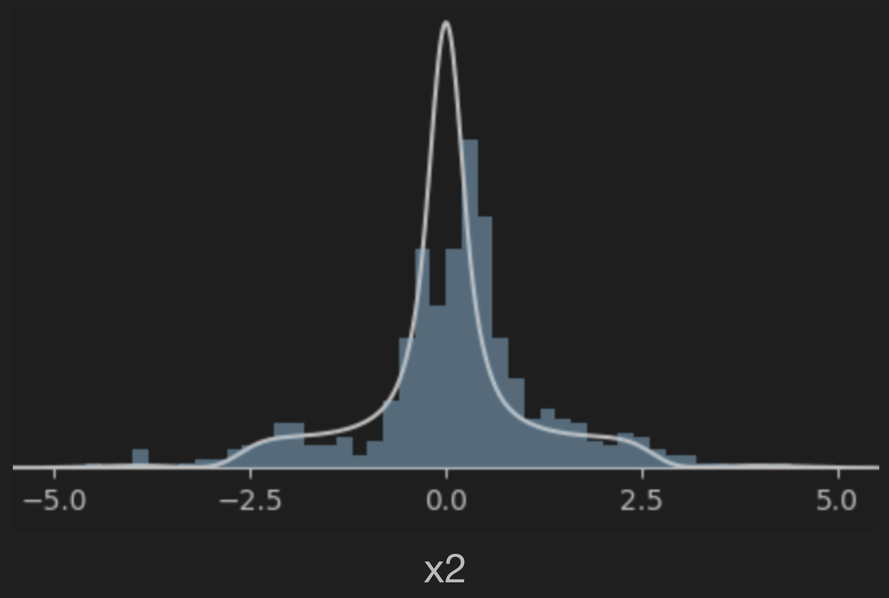
Gloeckler et al. (2024)

Flow matching: time-dependent vector field



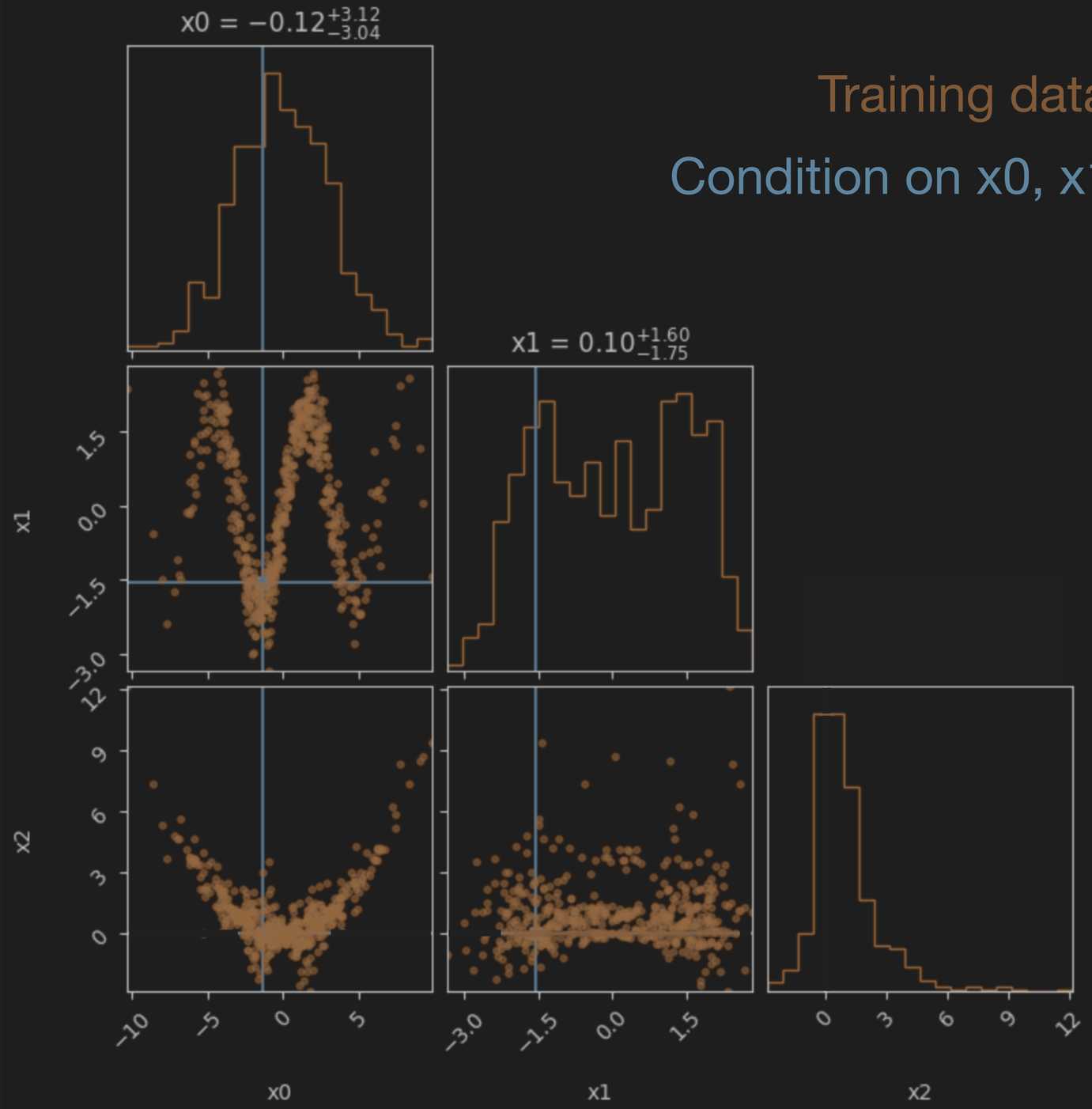
3D toy example

3D toy example



Training data

Condition on x0, x1



3D toy example

Flow not well constrained in low data regime

3D toy example

Flow not well constrained in low data regime

Guidance term towards physical manifold in low data regime

3D toy example

Flow not well constrained in low data regime

Guidance term towards physical manifold in low data regime

Baldan+2025; Bastek+2025

$$\mathscr{L} = w_{\mathrm{FM}}\mathscr{L}_{\mathrm{FM}} + w_{\mathrm{R}}\mathscr{L}_{\mathrm{R}}$$

3D toy example

Flow not well constrained in low data regime

Guidance term towards physical manifold in low data regime

Baldan+2025; Bastek+2025

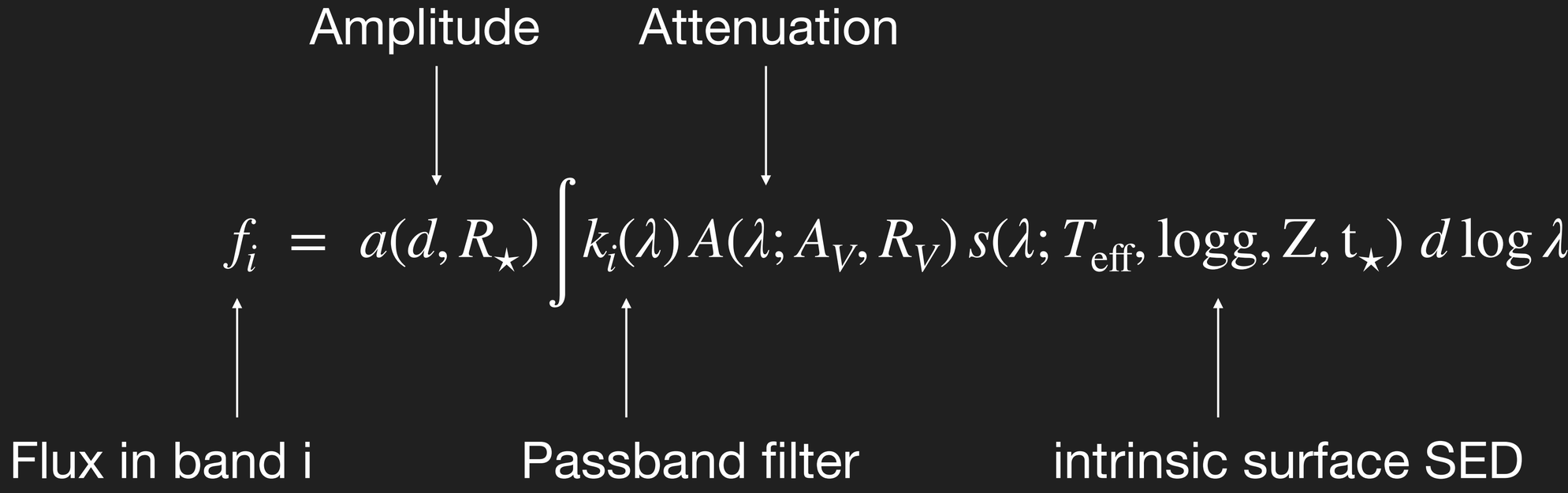
$$\mathscr{L} = w_{\mathrm{FM}}\mathscr{L}_{\mathrm{FM}} + w_{\mathrm{R}}\mathscr{L}_{\mathrm{R}}$$

$$\mathscr{L}_{\mathrm{R}} = f(x_1, \ldots, x_n)$$

E.g.: $f(x_1, \ldots, x_n) = x_1^2 + x_2^2 + x_3^2$

Guidance via measurement operator constraints

$$\mathscr{L}_{\mathrm{R}} = f(x_1, \ldots, x_n)$$



Guidance via measurement operator constraints

$\mathscr{L}_{\mathrm{R}} = f(x_1, \ldots, x_n)$

$$f_i = a(d, R_\star) \int k_i(\lambda)\, A(\lambda; A_V, R_V)\, s(\lambda; T_{\mathrm{eff}}, \mathrm{logg}, \mathrm{Z}, \mathrm{t}_\star)\, d\log\lambda$$

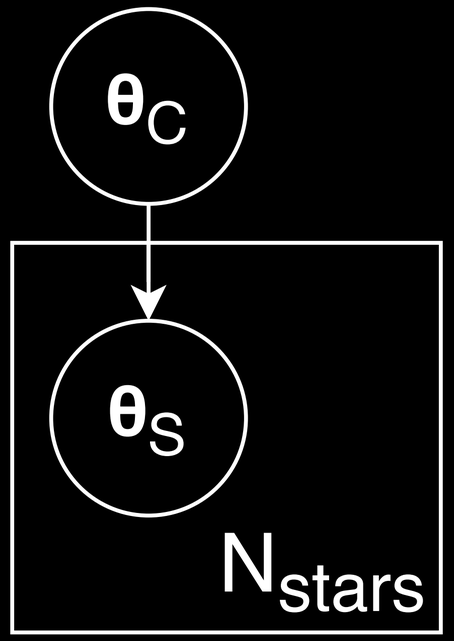
$$\updownarrow$$

$$[\mu(\lambda) + \sum_m c_m \phi_m(\lambda)]$$

$$p(\vec{x}, \vec{\theta}, \vec{c})$$

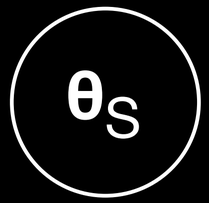
Pilot study

Clusters/young stars

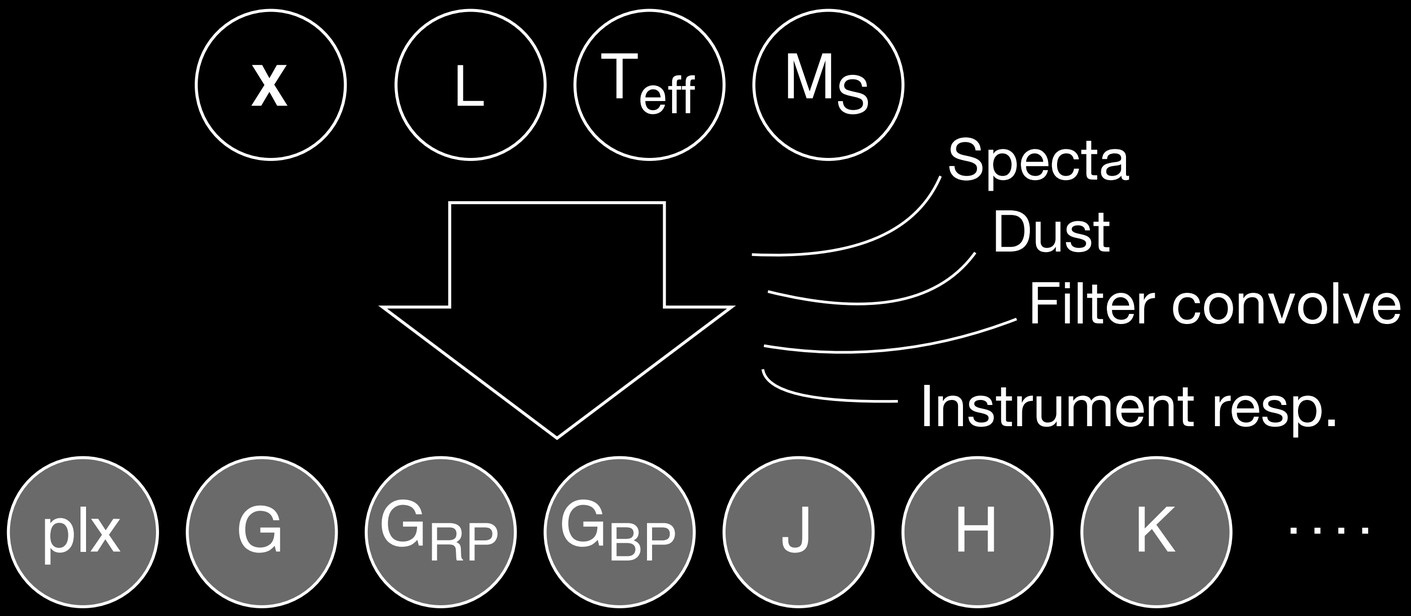


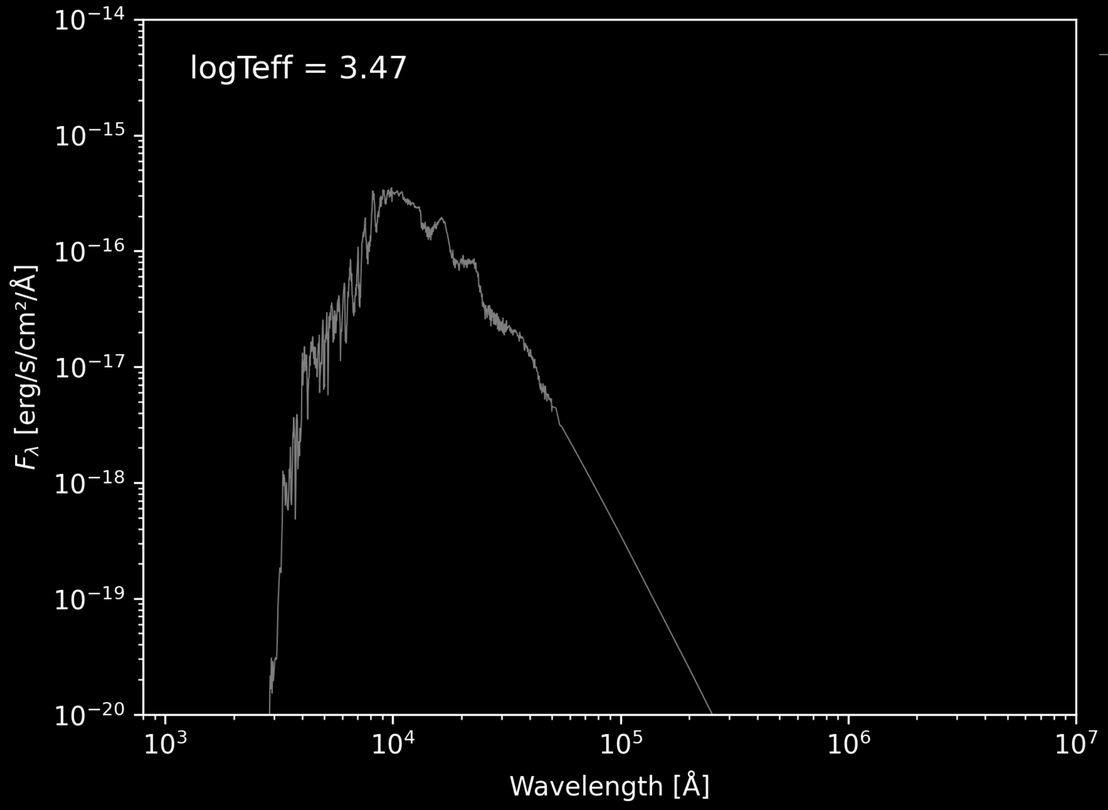
Milky Way model

- *Galaxia* code
  - Galactic structure + kinematics, SFH + chemistry evolution
- thin+thick+halo+bulge

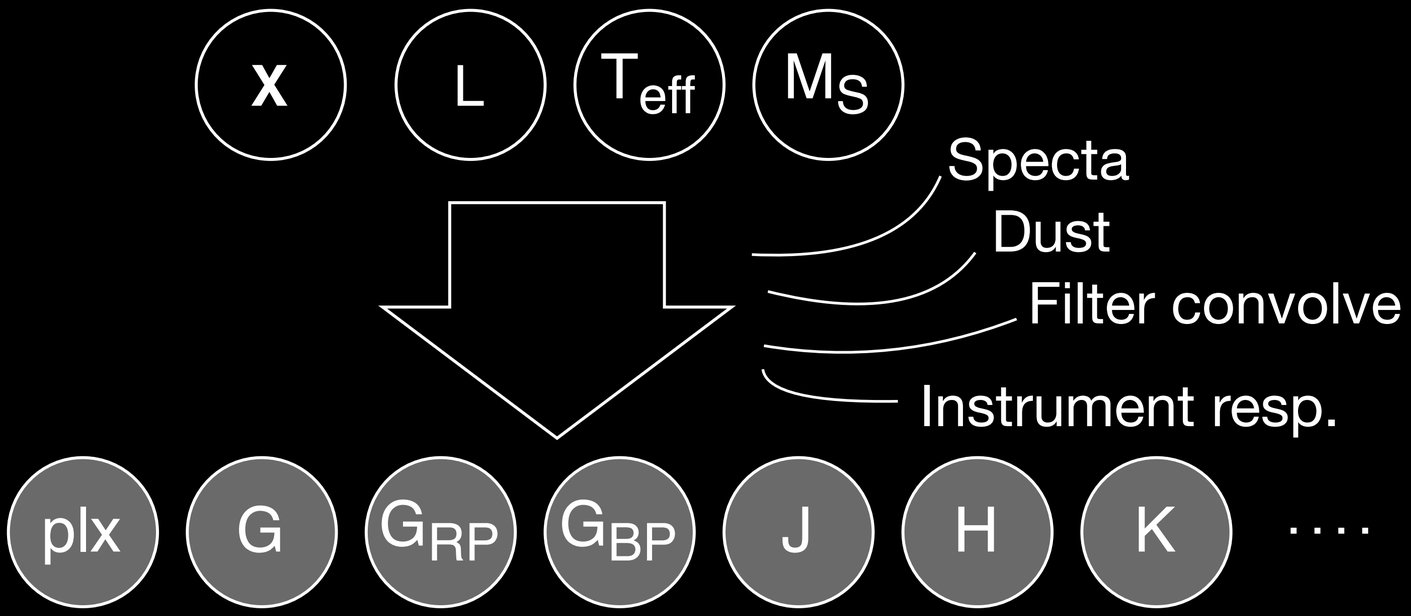


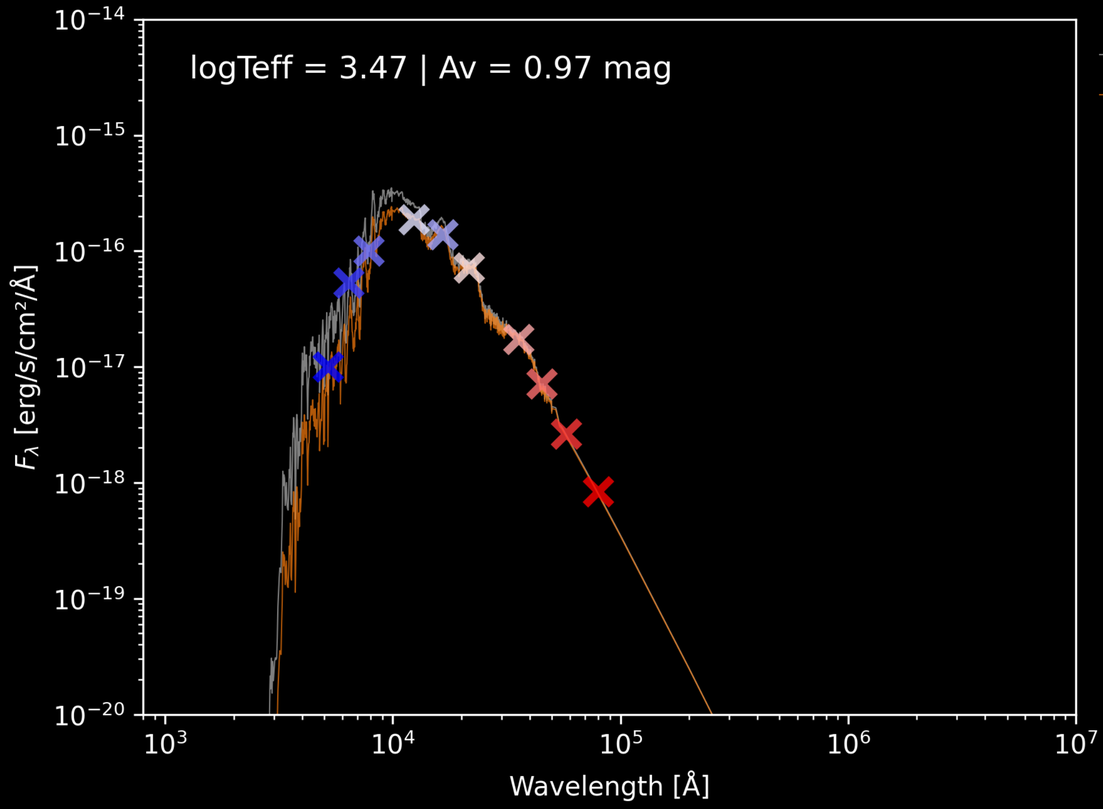
Fwd model


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]

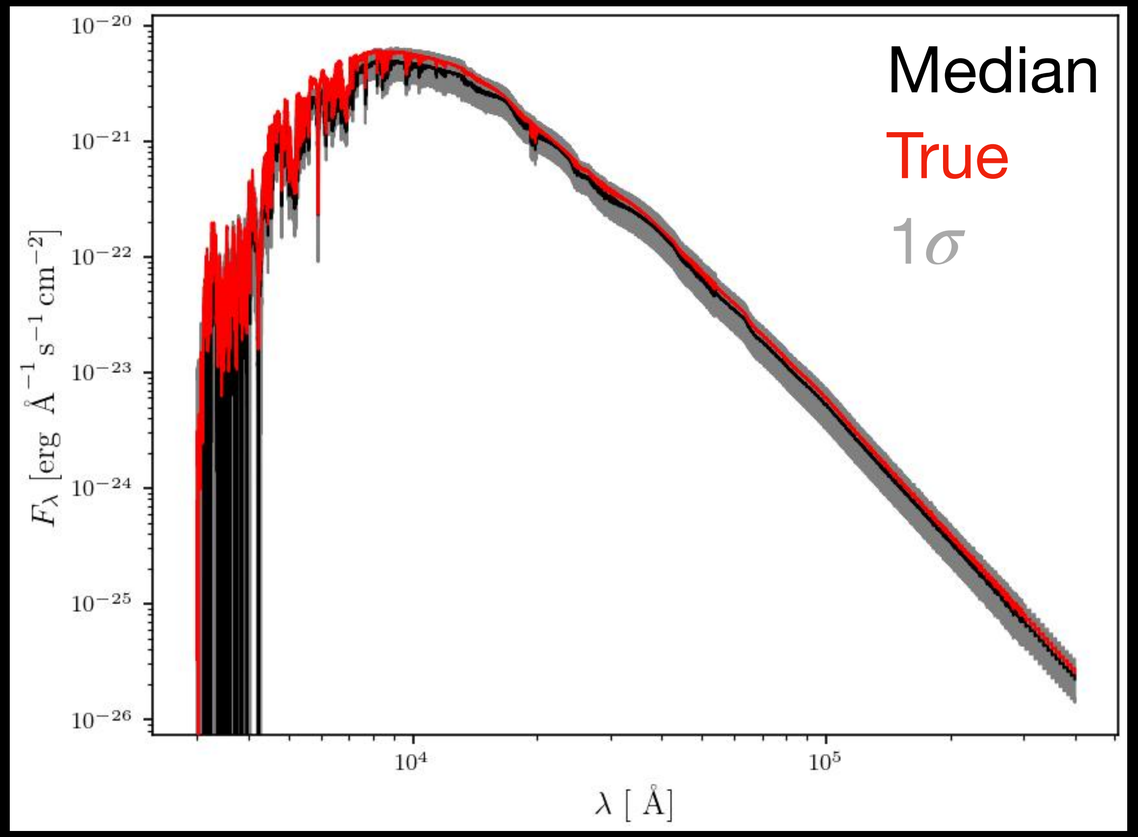
Fwd model

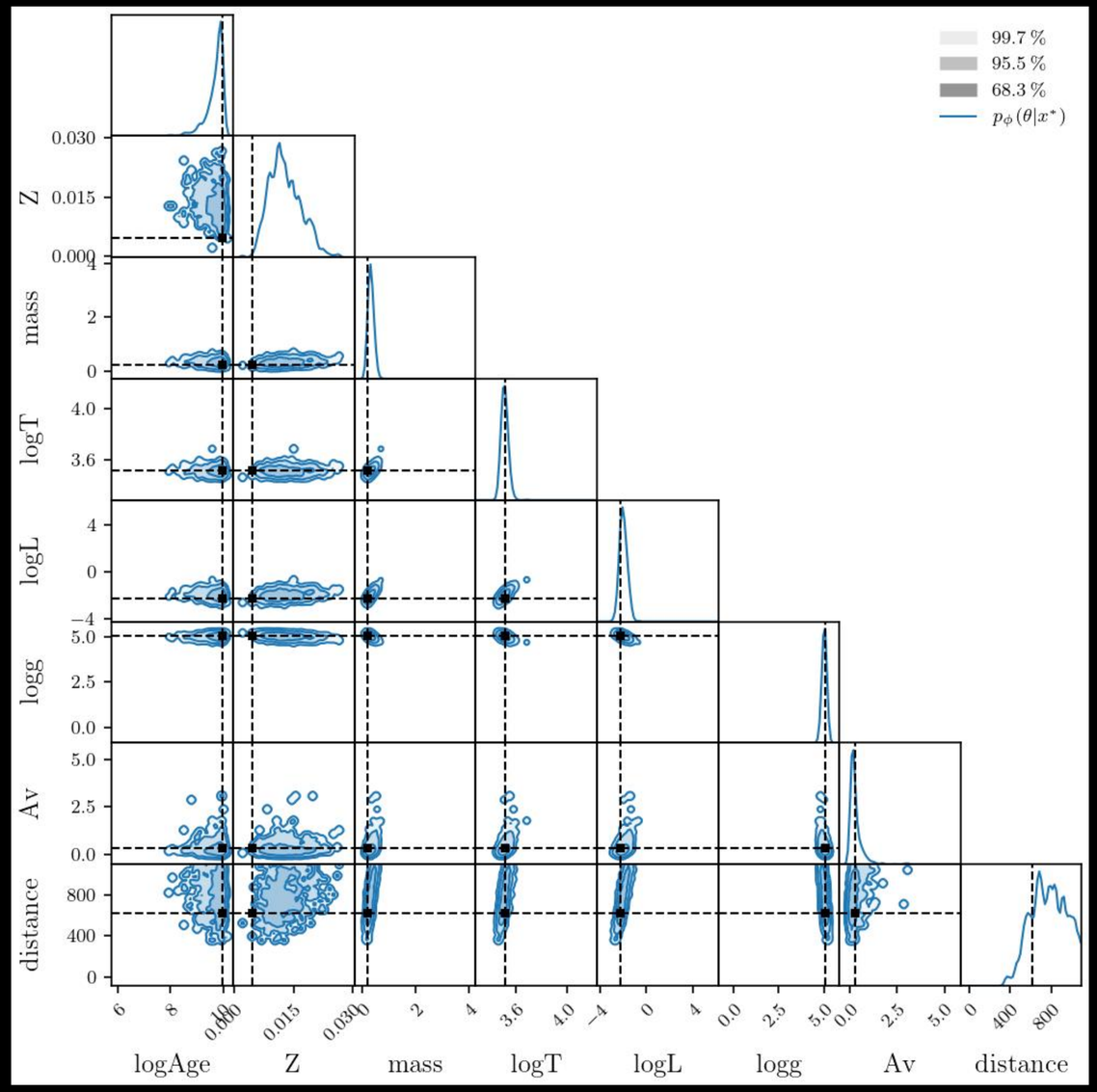

x
L
T_{eff}
M_S
Specta
Dust
Filter convolve
Instrument resp.
plx
G
G_{RP}
G_{BP}
J
H
K
....


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

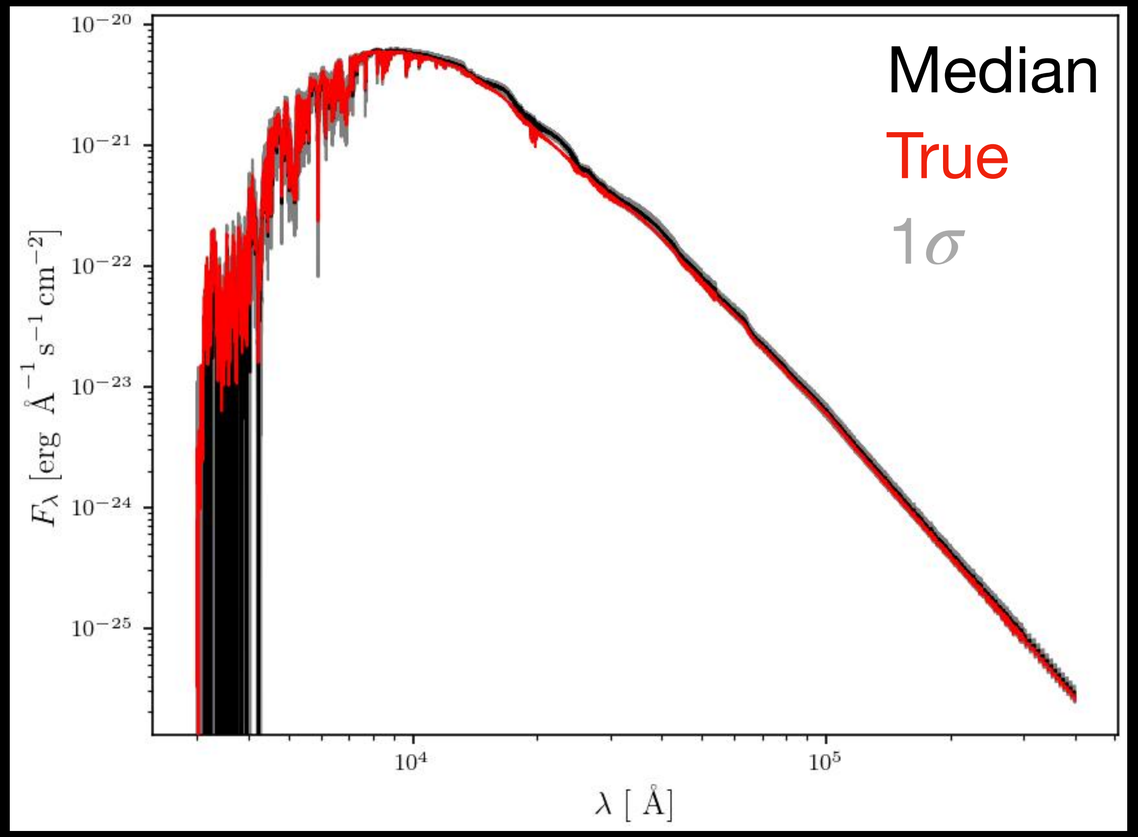
Example: low mass star

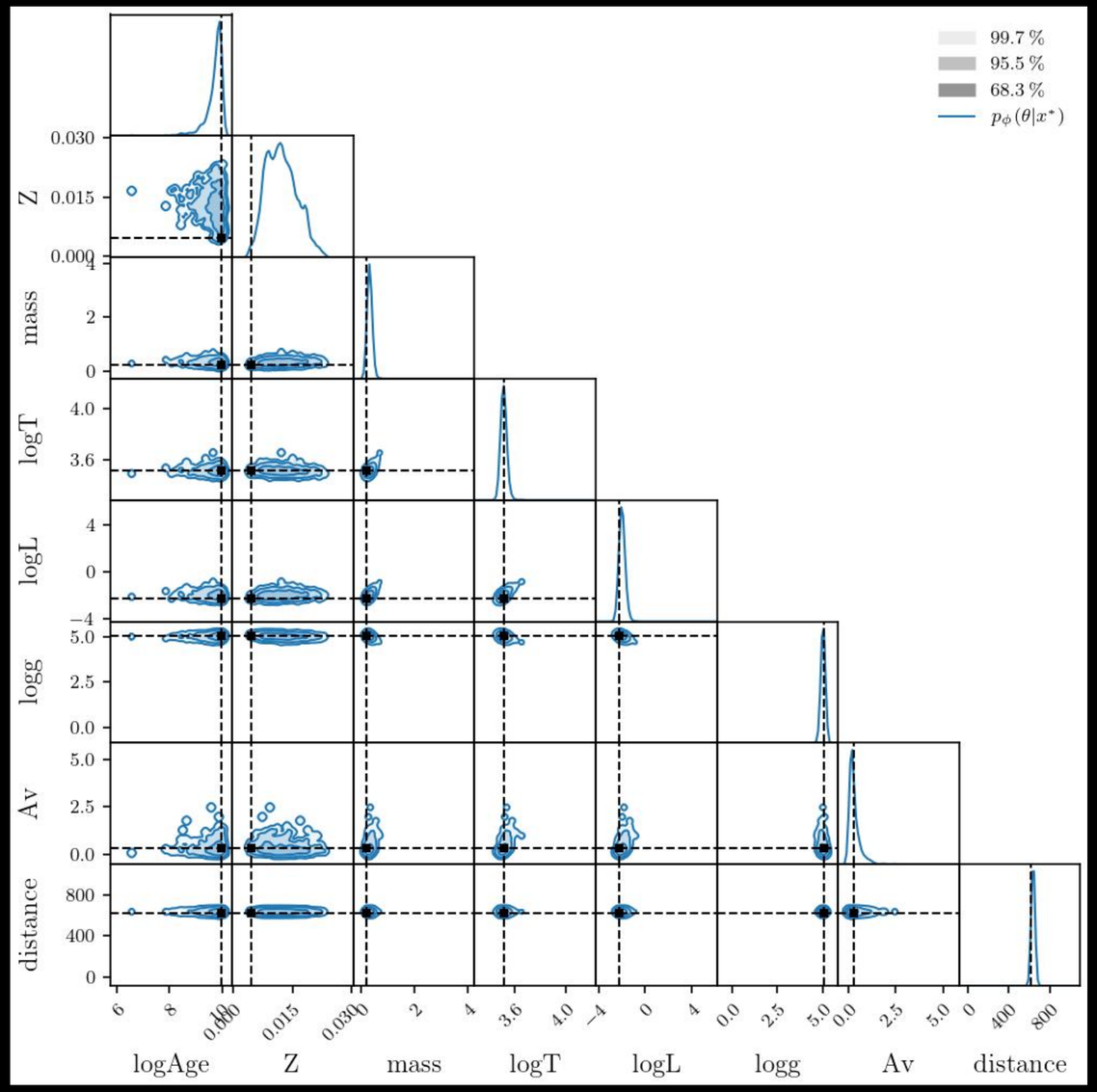
Gaia + 2MASS + low SNR parallax



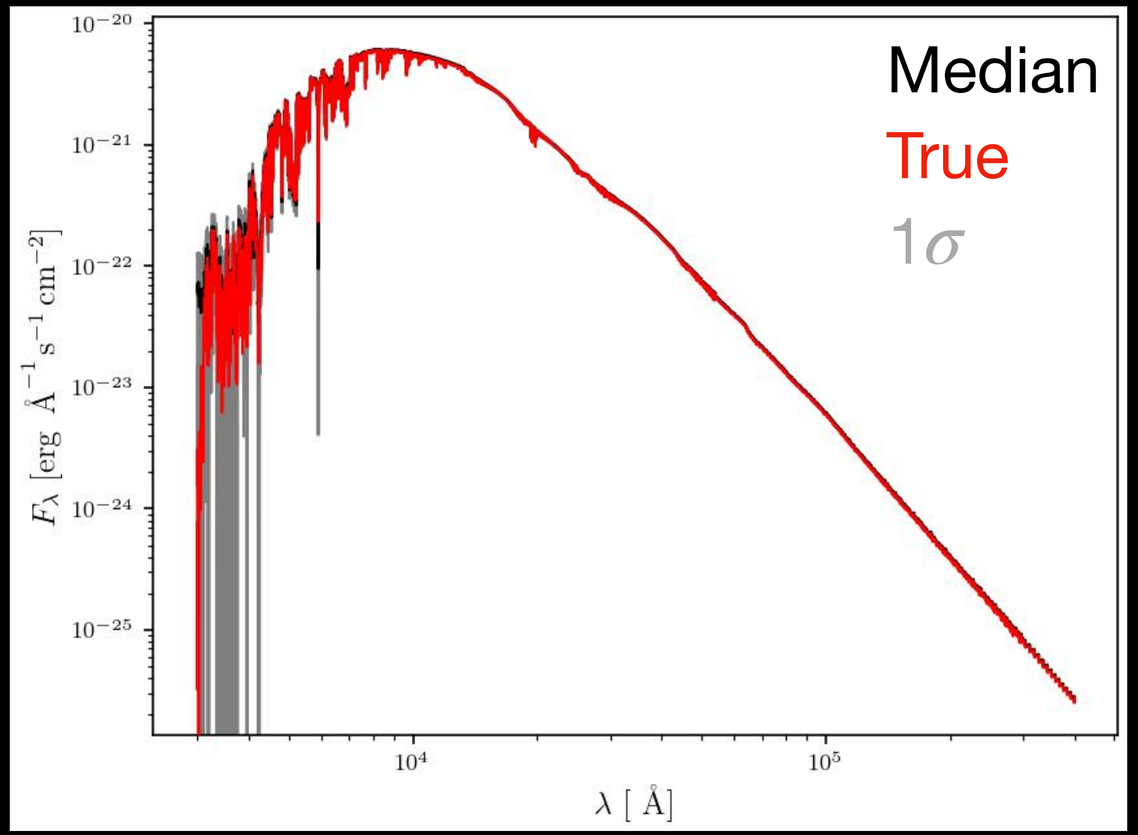


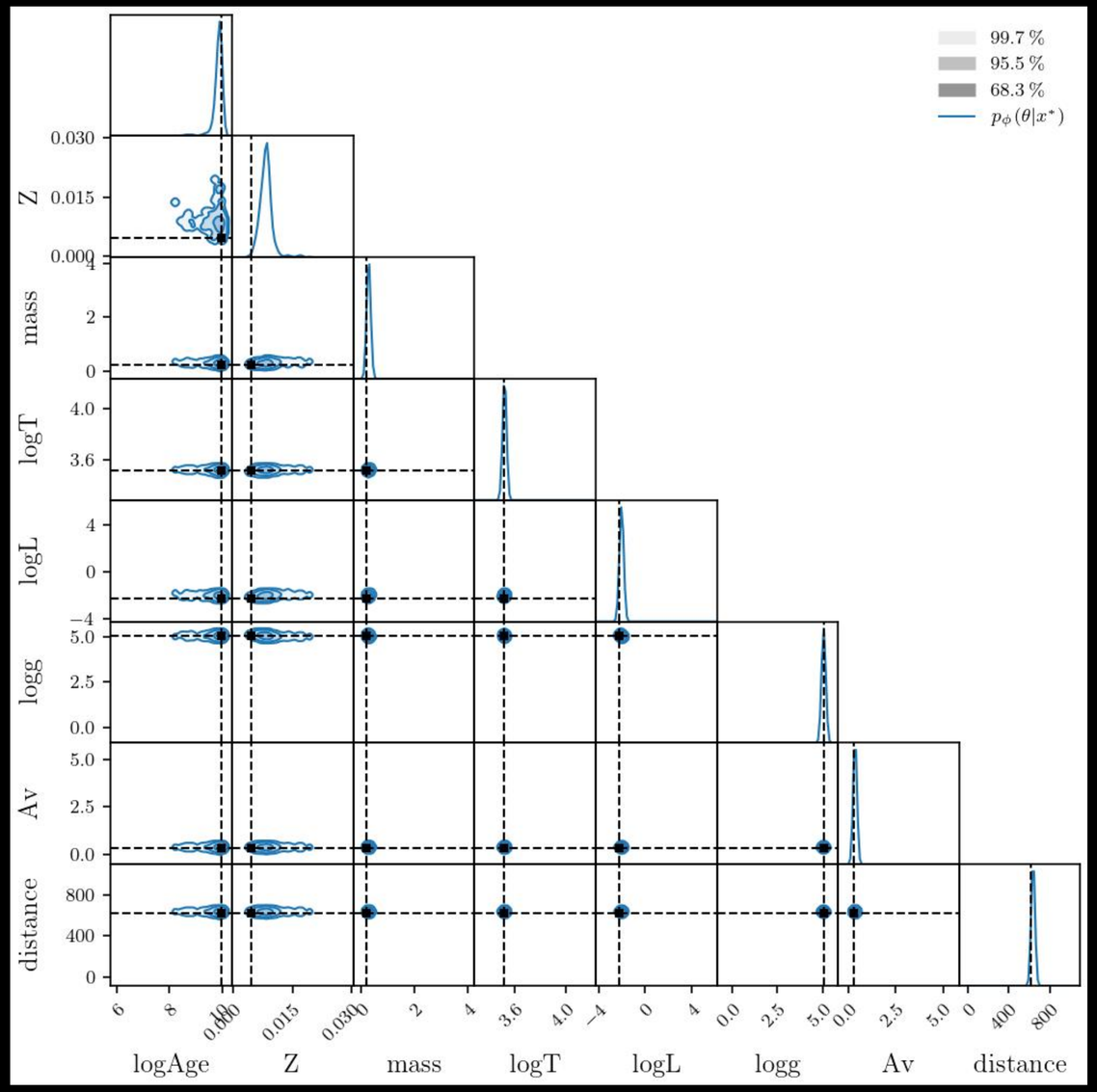
Gaia + 2MASS + high SNR parallax





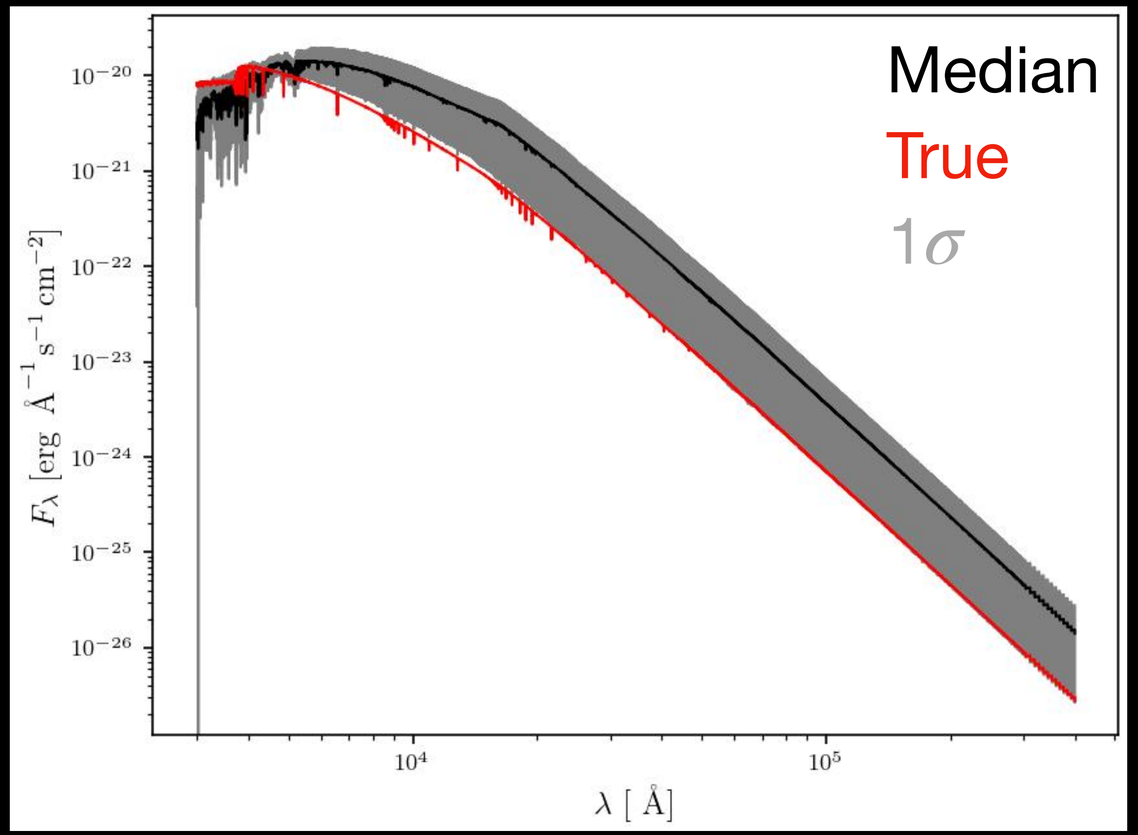
Gaia + 2MASS + high SNR parallax + LAMOST

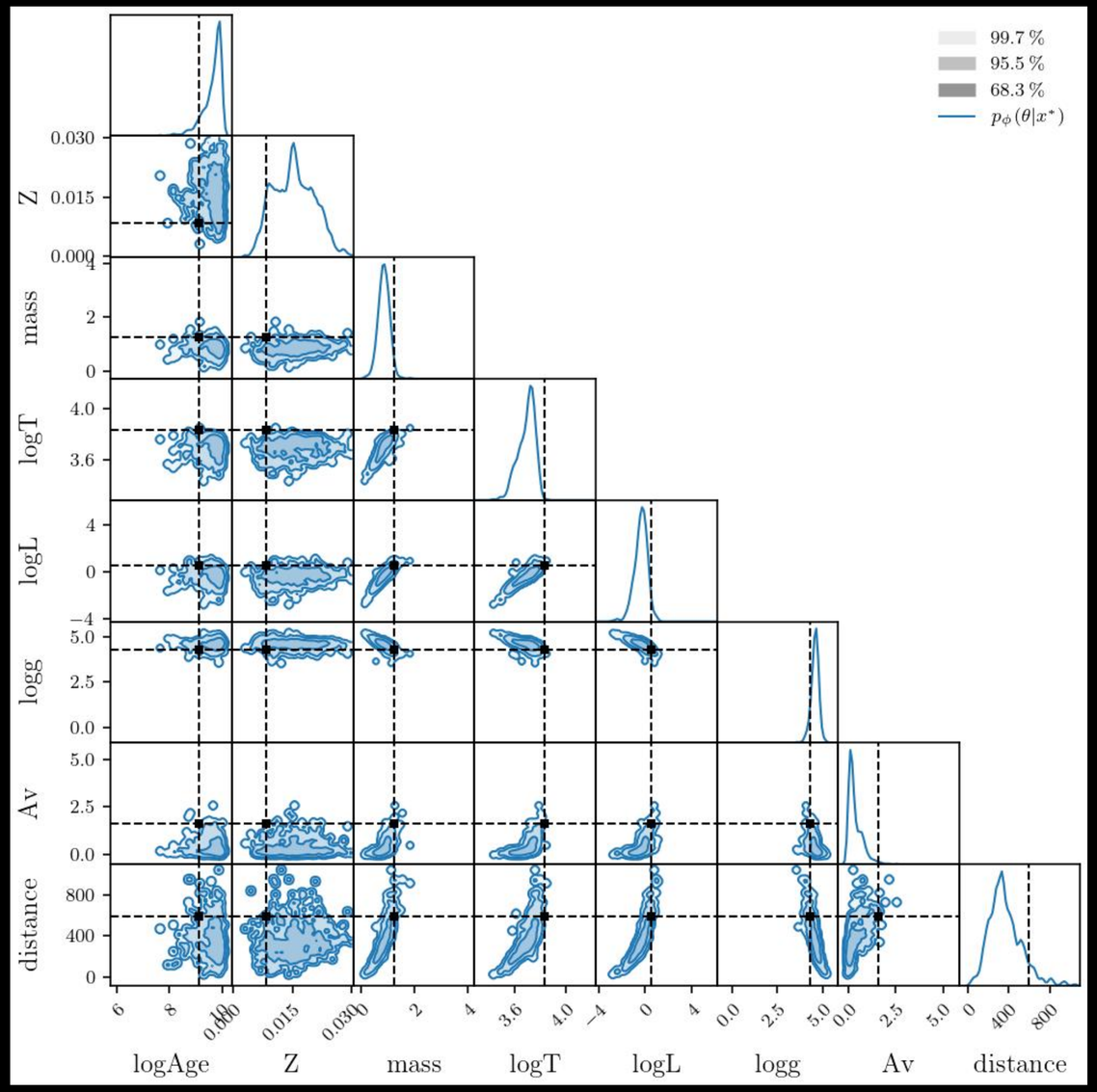




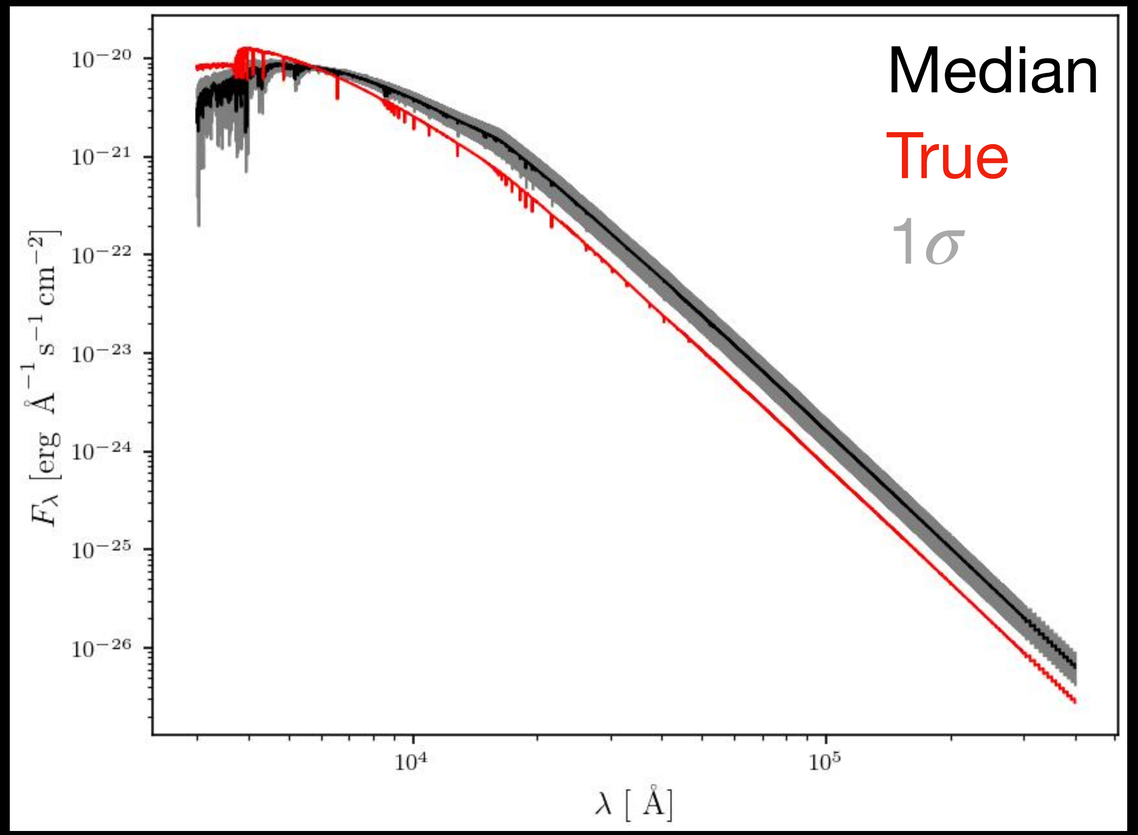
Example: high mass star

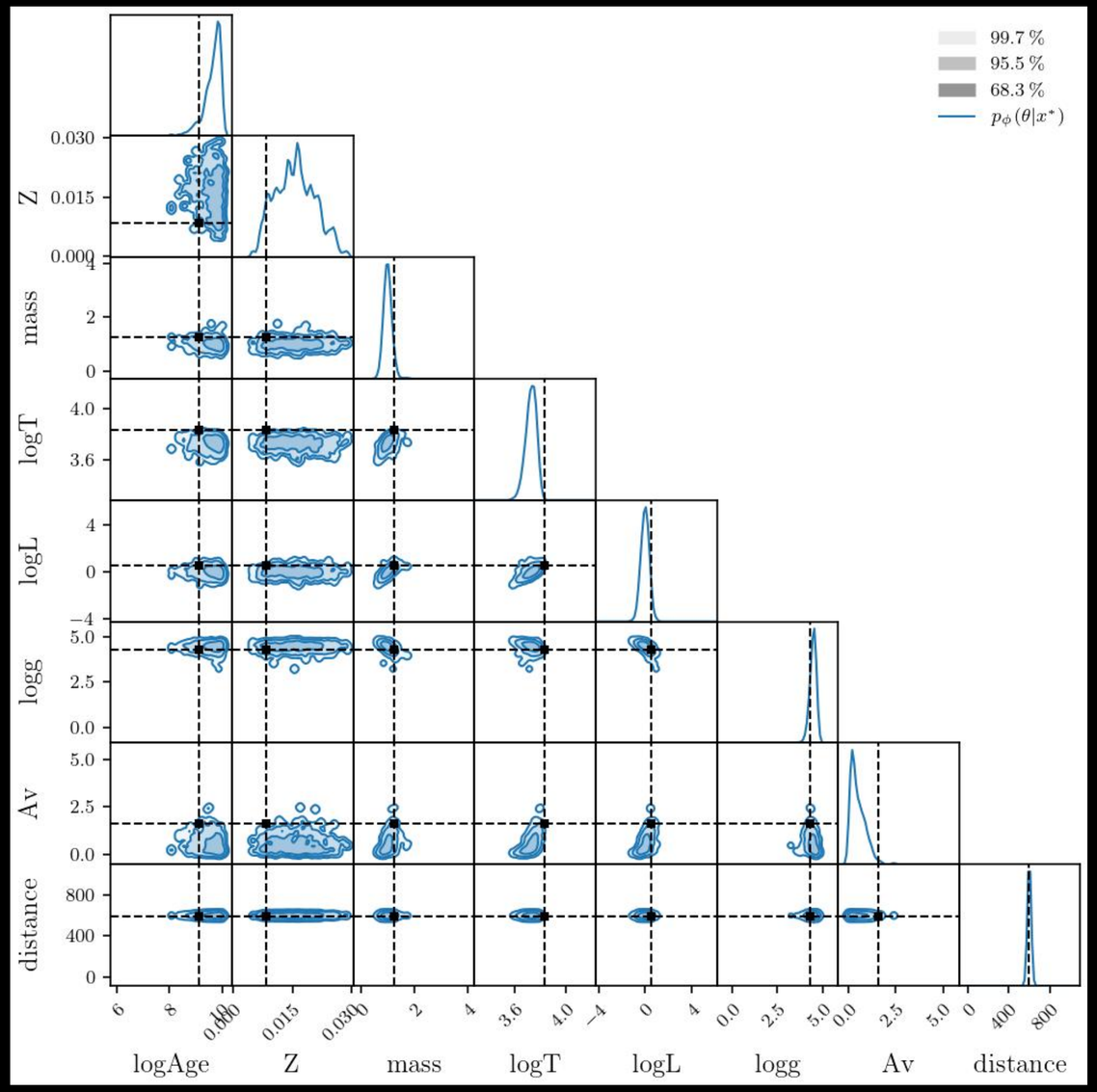
Gaia + 2MASS + low SNR parallax



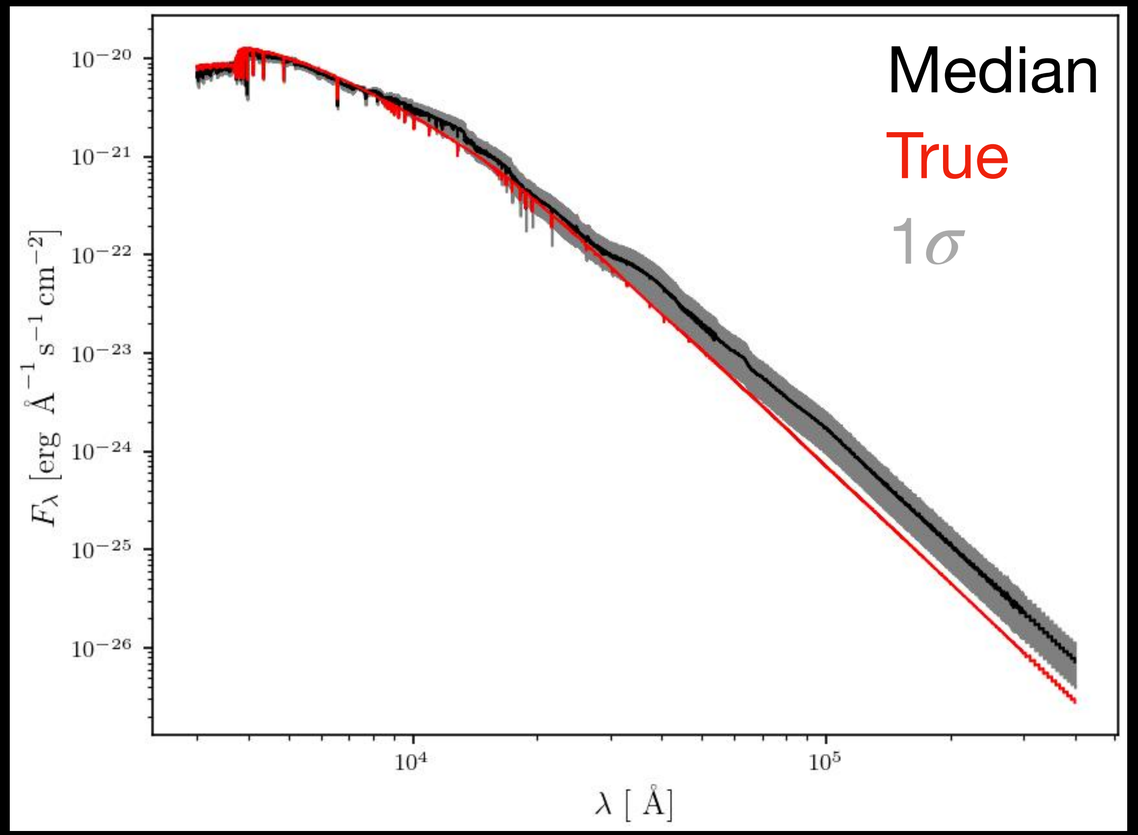


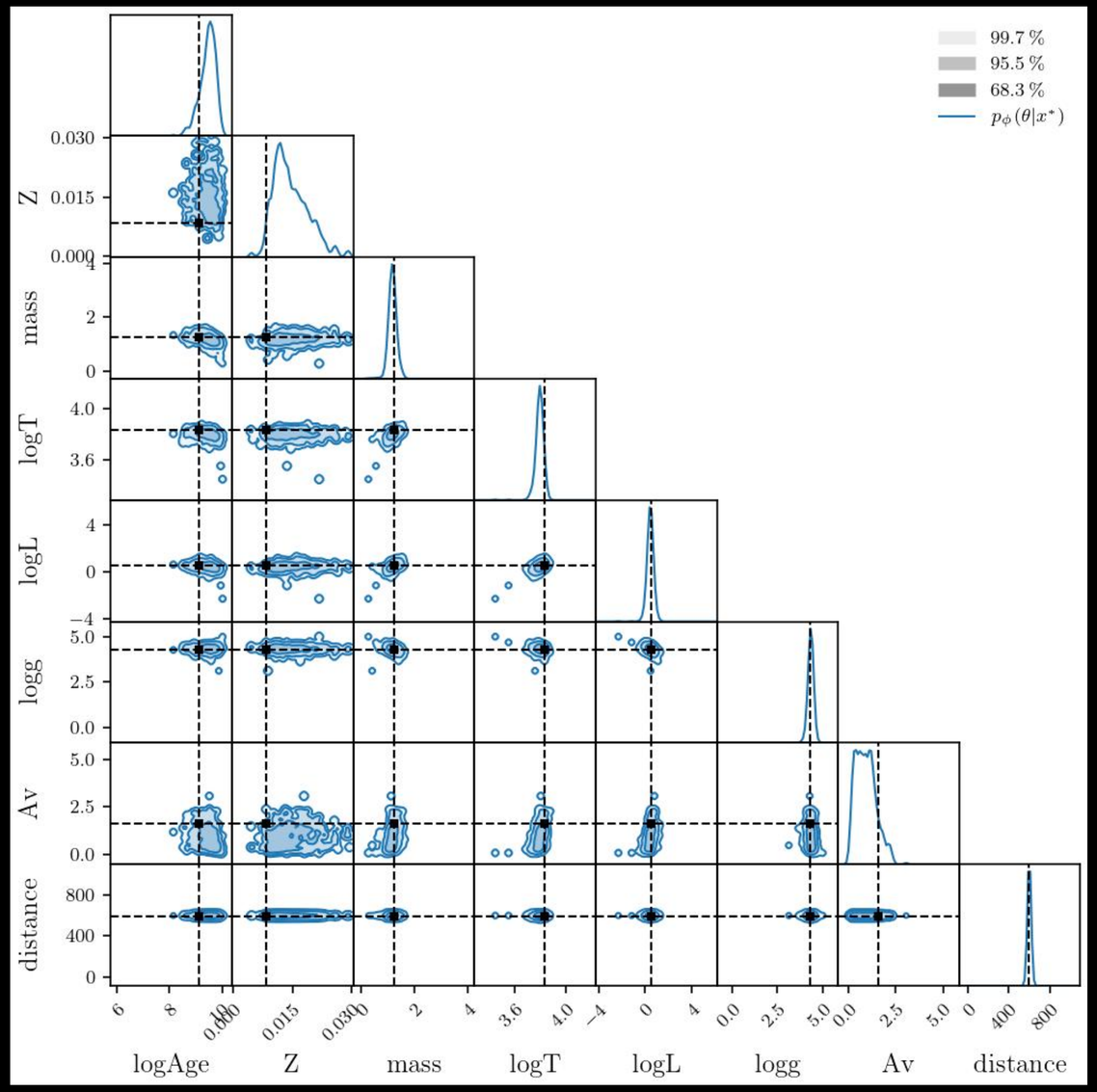
Gaia + 2MASS + high SNR parallax





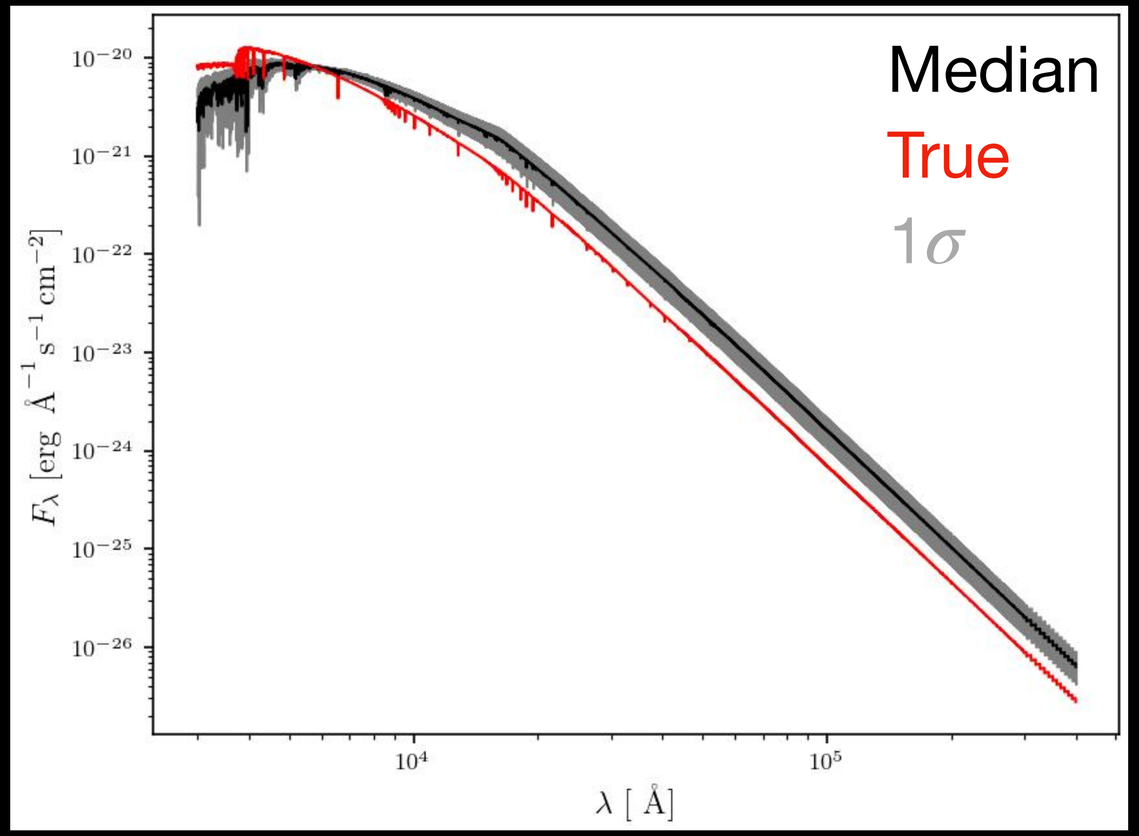
Gaia + 2MASS + high SNR parallax + LAMOST

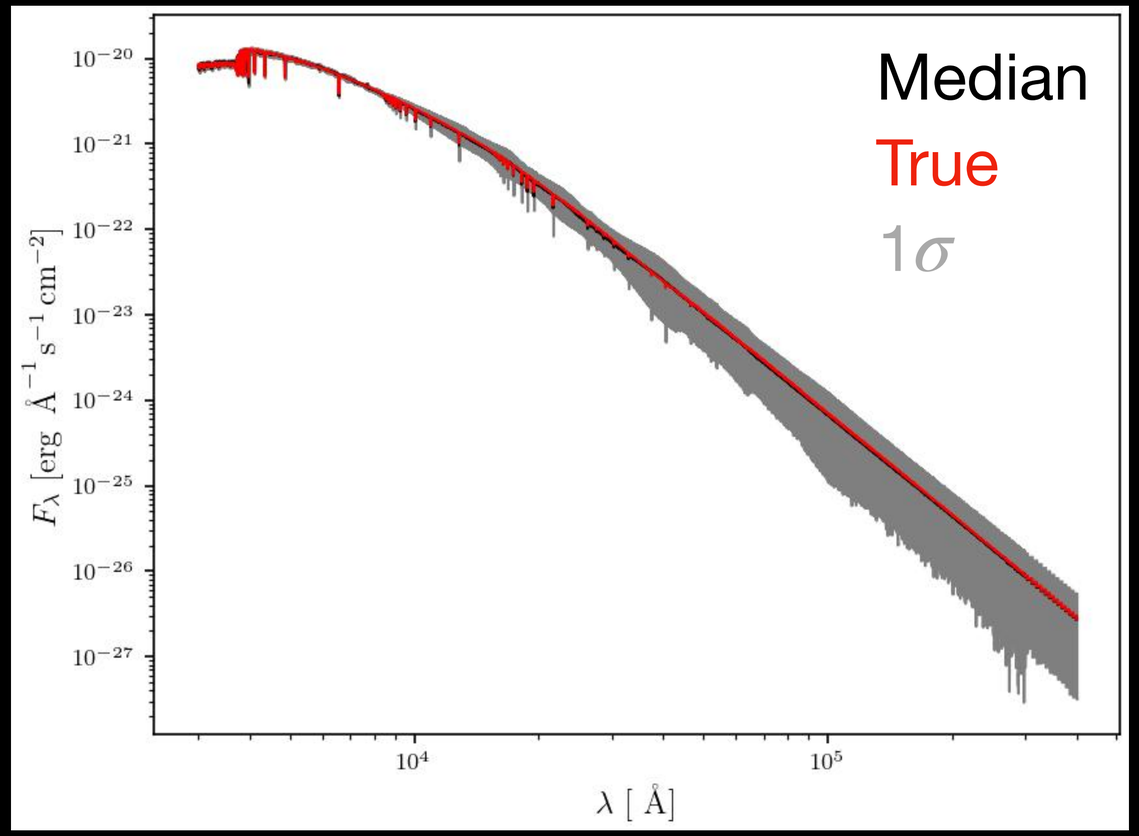




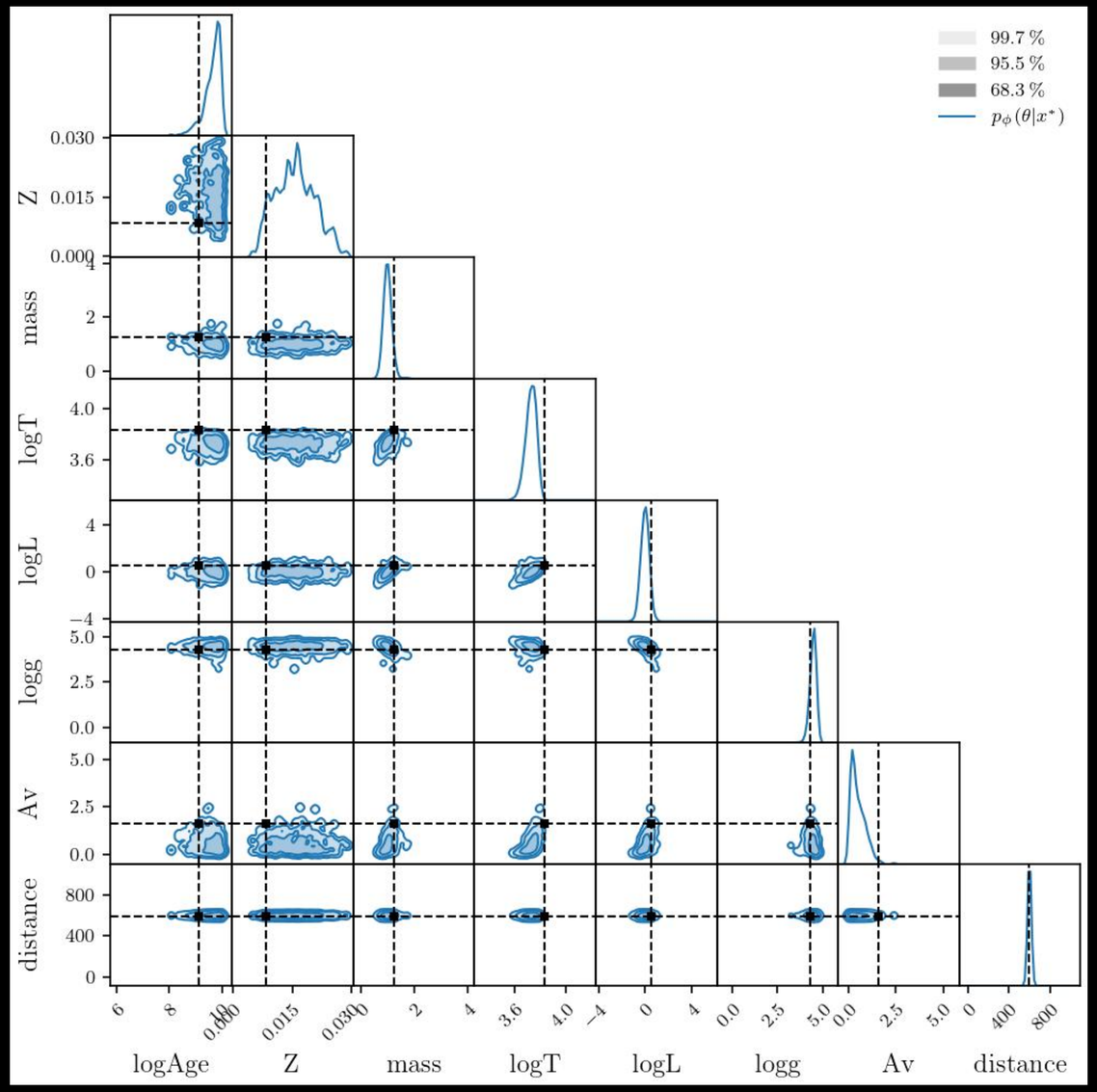
Adding the guidance term

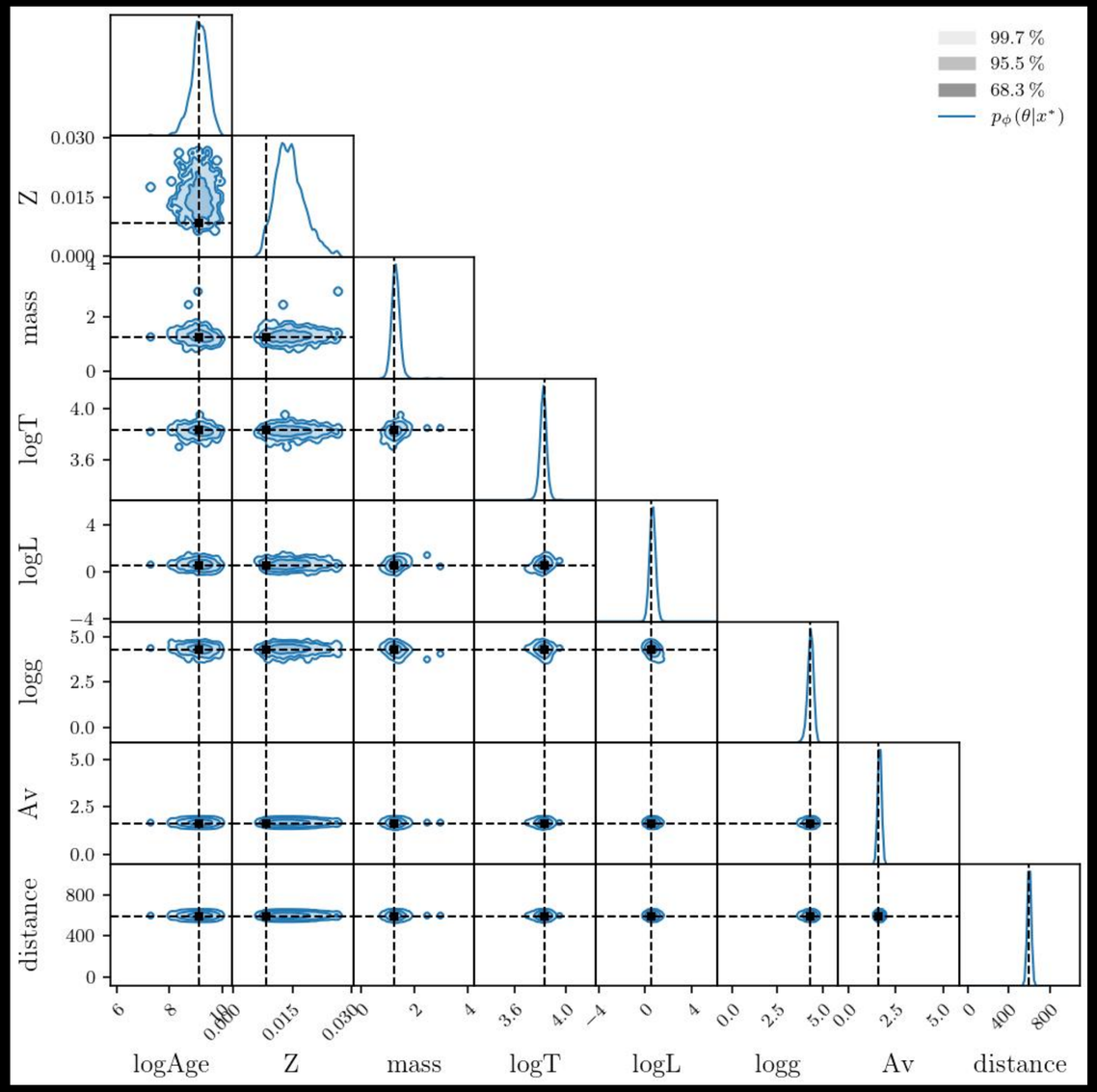
With guidance



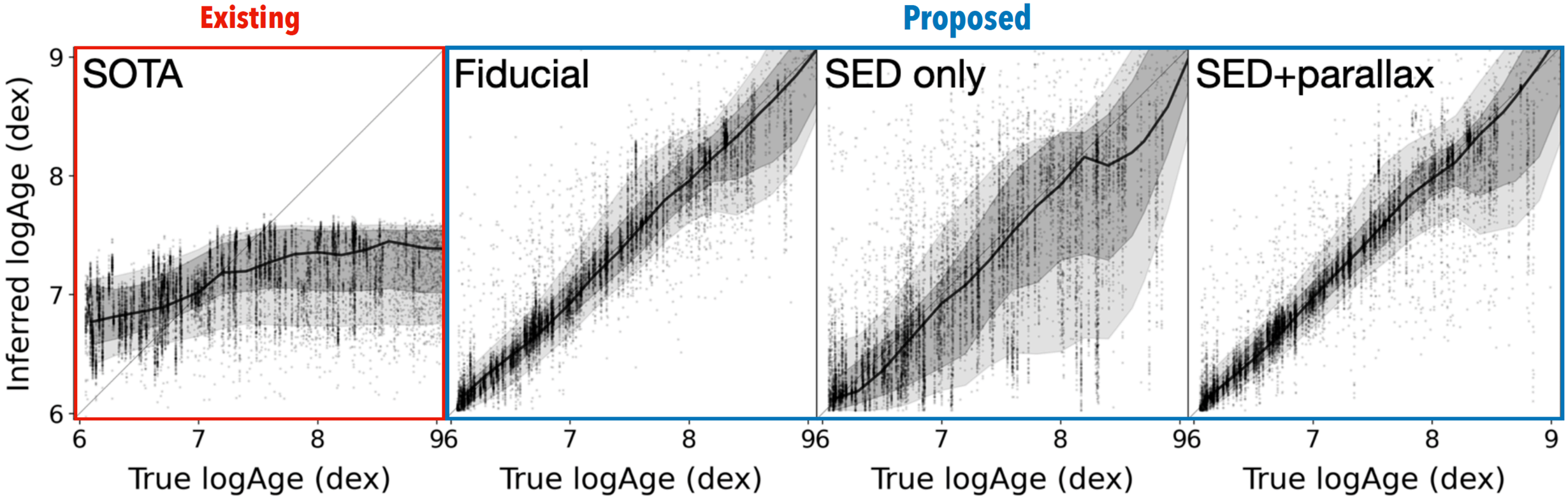


With guidance





Pilot study (Gaia + 2MASS + Av)



Summary

- SBI model using **flow matching** + **transformer model** to learn arbitrary **conditionals**
- Guidance via measurement operator constraints
- Obtain promising results on simulations

Thank you