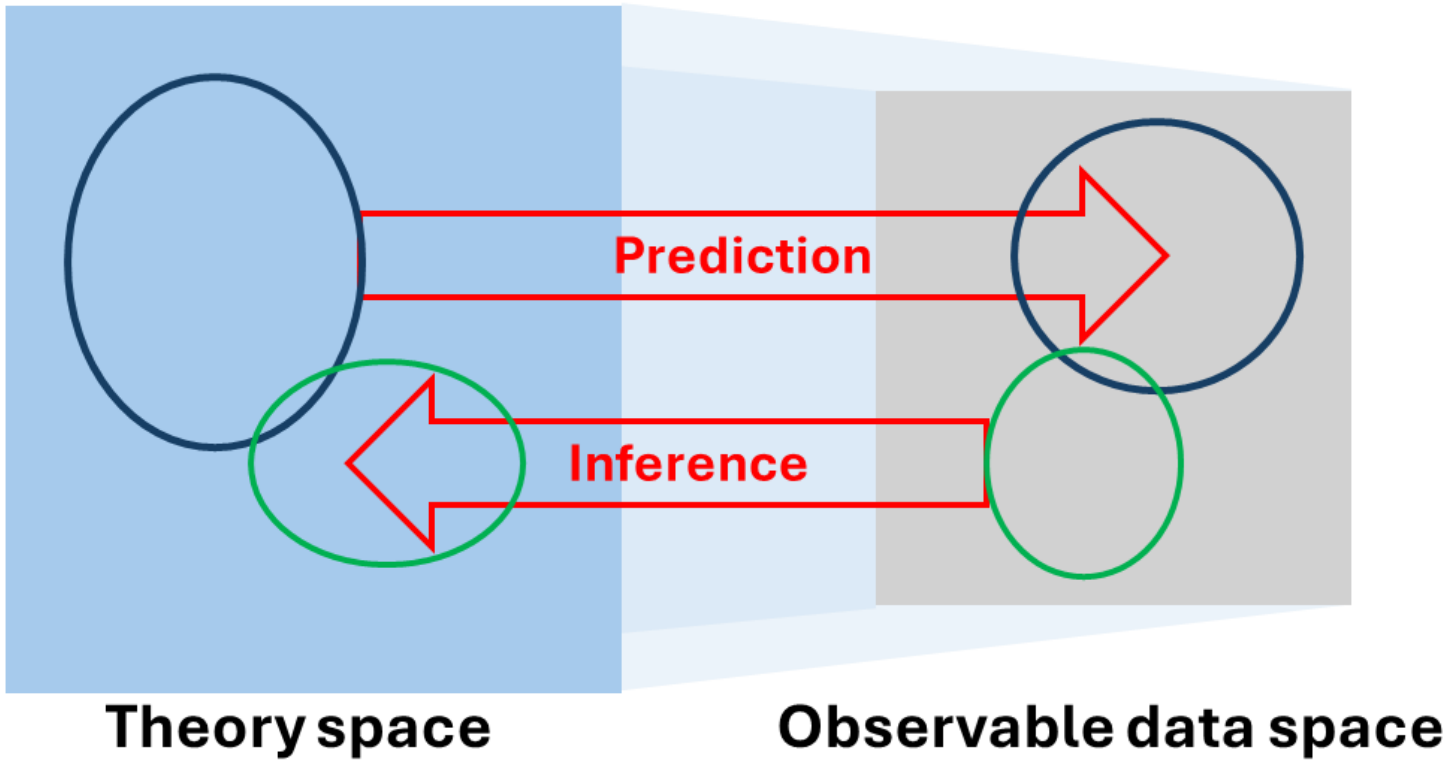
Reliable Scientific Inference with Neural Density Estimator and Predictive NNs

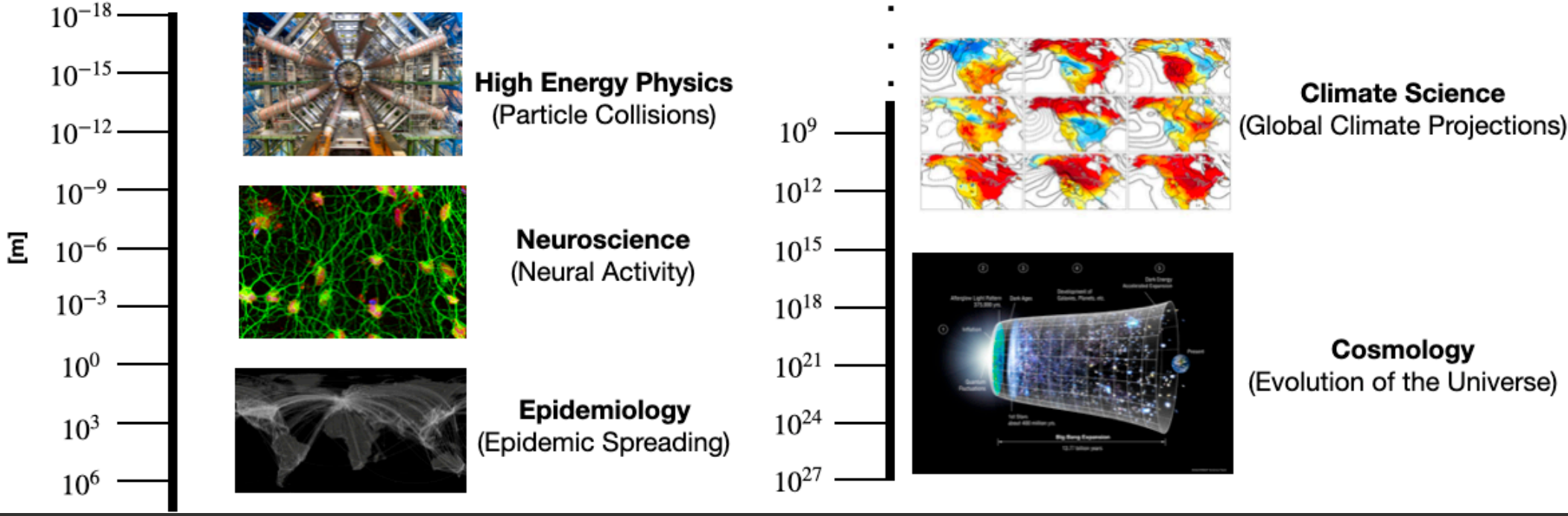
Ann B. Lee
Department of Statistics & Data Science
Carnegie Mellon University

Joint work with Luca Masserano (CMU), Nic Dalmasso (JP Morgan AI); James Carzon (CMU); Antonio Carlos Herling (CMU); Alex Shen (CMU); Rafael Izbicki (UFSCar); Mikael Kuusela (CMU); Tommaso Dorigo (Padova); Josh Speagle (U.Toronto)

The Interplay Between Theory/Models and Data



“Theory” in the Form of Simulations



Credit: Dalmasso (adapted from Cranmer et al, 2020)

- Physics-based simulator as a causal (mechanistic) model that encodes the data-generating process $\theta \mapsto \mathscr{D}$, where θ are internal parameters that determine measurable data $\mathscr{D}$

How Do We Test or Constrain Our Theory/Model Given Data?

"Labeled" data $\{\theta_i, \mathscr{D}_i\}_{i=1}^{B}$ from either

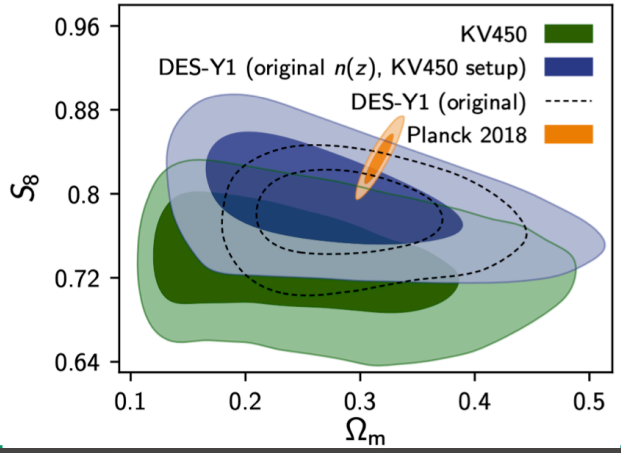
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D}; \theta)$

or

ii) Observational study with "precise" labels θ from **auxiliary measurements**

↓

Infer internal parameters/labels of interest with measures of uncertainty.



How Do We Test or Constrain Our Theory/Model Given Data?

"Labeled" data $\{\theta_i, \mathscr{D}_i\}_{i=1}^{B}$ from either

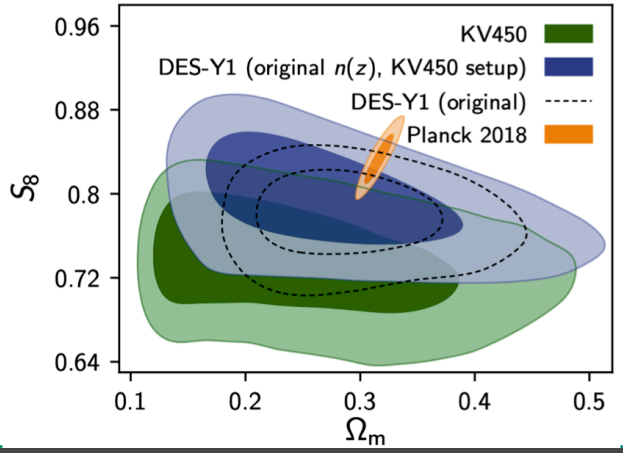
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D}; \theta)$

or

ii) Observational study with "precise" labels θ from **auxiliary measurements**

↓

Infer internal parameters/labels of interest with measures of uncertainty.



- Are we confident that these regions include the true/ unknown parameter with high probability?
- Do the sizes of the regions reflect our constraining power?

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\right) = 1 - \alpha, \quad \forall \theta \in \Theta$$

How Do We Test or Constrain Our Theory/Model Given Data?

"Labeled" data $\{\theta_i, \mathscr{D}_i\}_{i=1}^{B}$ from either

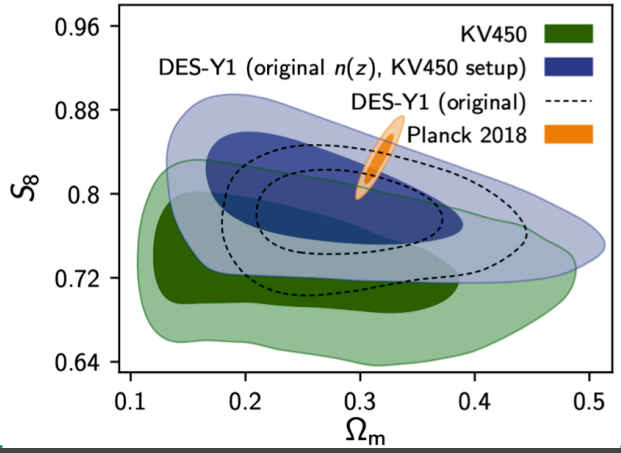
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D}; \theta)$

or

ii) Observational study with "precise" labels θ from **auxiliary measurements**

↓

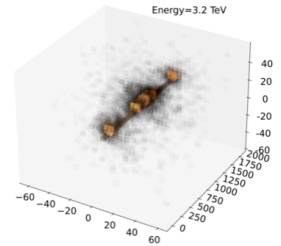
Infer internal parameters/labels of interest with measures of uncertainty.



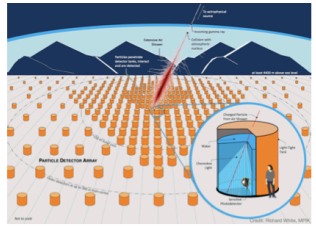
Some examples (all "local" parameters):

θ ⟷ $\mathscr{D} = X$

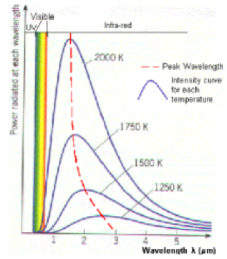
Energy of subatomic particle



Identity, orientation and energy of cosmic-ray showers



Stellar labels (e.g., mass, age, composition)



How Do We Test or Constrain Our Theory/Model Given Data?

"Labeled" data $\{\theta_i, \mathscr{D}_i\}_{i=1}^{B}$ from either

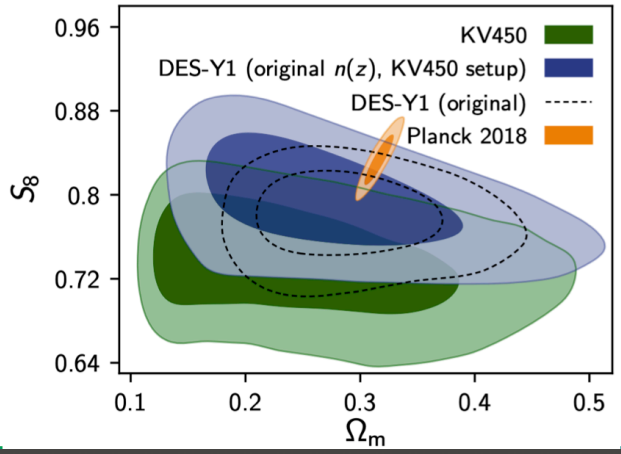
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D}; \theta)$

or

ii) Observational study with "precise" labels θ from **auxiliary measurements**

↓

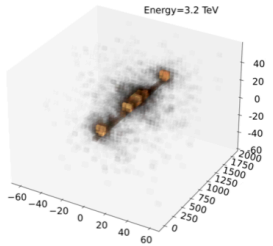
Infer internal parameters/labels of interest with measures of uncertainty.



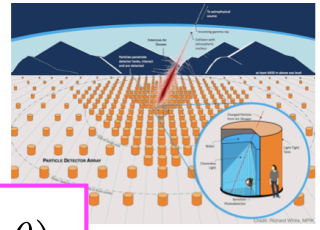
Some examples (all "local" parameters):

θ ⟷ $\mathscr{D} = X$

Energy of subatomic particle

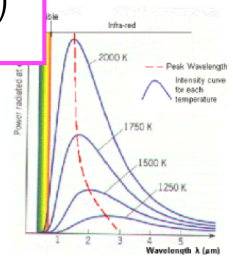


Identity, orientation and energy of cosmic-ray showers

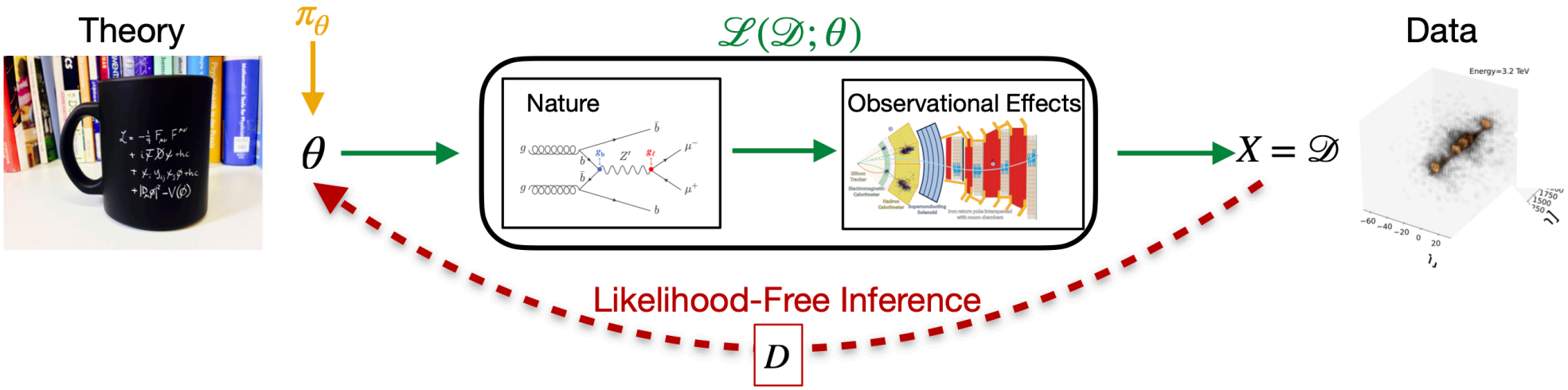


$$\mathcal{T}_{\text{train}} = \{(\theta_1, \mathbf{X}_1) \ldots (\theta_B, \mathbf{X}_B)\} \sim \pi(\theta)\mathcal{L}(\mathbf{x}; \theta)$$

Stellar labels (e.g., mass, age, composition)

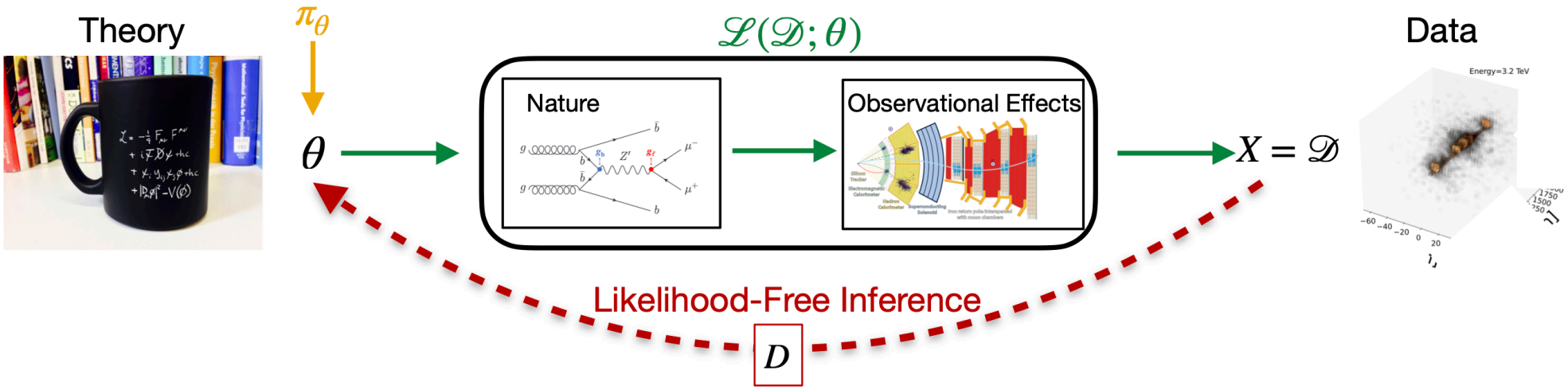


Complex Scientific Inference is Often "Likelihood-Free"



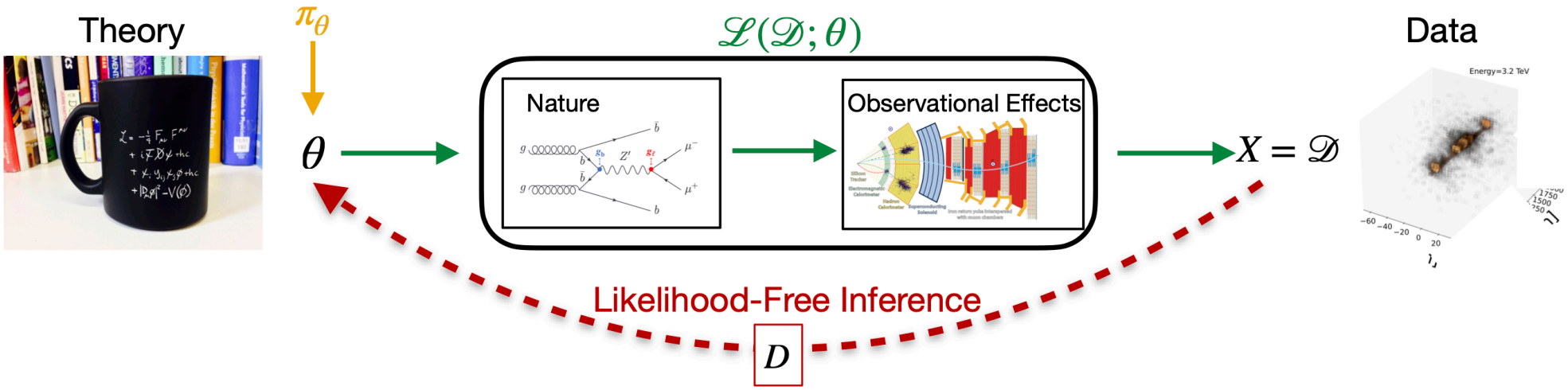
- Suppose we have knowledge of data-generating process $\theta \mapsto \mathscr{D}$ e.g. via a "high-fidelity simulation"
- But **likelihood is intractable:** e.g, $p(x \mid \theta) = \int p(x \mid z)p(z \mid \theta)\mathrm{d}z$, where z are latent variables
- Inference (inverse problem) is hard: given <u>new</u> $D = \{x_1^{obs}, \ldots, x_n^{obs}\}$, use $\{\theta_i, D_i\}_{i=1}^B$ to infer parameters $\theta^\star$

Complex Scientific Inference is Often "Likelihood-Free"



- Suppose we have knowledge of data-generating process $\theta \mapsto \mathscr{D}$ e.g. via a "high-fidelity simulation"
- But **likelihood is intractable:** e.g, $p(x \mid \theta) = \int p(x \mid z)p(z \mid \theta)\mathrm{d}z$, where z are latent variables
- Inference (inverse problem) is hard: given new $D = \{x_1^{obs}, \ldots, x_n^{obs}\}$, use $\{\theta_i, D_i\}_{i=1}^B$ to infer parameters $\theta^\star$
- Assumptions in our work regarding the data-generating process:
  1. Likelihood $\mathscr{L}(\mathscr{D};\theta)$ does not change between training and inference: no unaccounted-for model uncertainties
  2. "Prior" π_θ (i.e., how we observe train data across the parameter space) could be poorly designed

Complex Scientific Inference is Often "Likelihood-Free"



- Suppose we have knowledge of data-generating process $\theta \mapsto \mathscr{D}$ e.g. via a "high-fidelity simulation"
- But **likelihood is intractable:** e.g, $p(x \mid \theta) = \int p(x \mid z)p(z \mid \theta)\mathrm{d}z$, where z are latent variables
- Inference (inverse problem) is hard: given new $D = \{x_1^{obs}, \ldots, x_n^{obs}\}$, use $\{\theta_i, D_i\}_{i=1}^B$ to infer parameters $\theta^\star$

$$\mathcal{T}_{\text{train}} = \{(\theta_1, \mathbf{X}_1) \ldots (\theta_B, \mathbf{X}_B)\} \sim \pi(\theta)\mathcal{L}(\mathbf{x};\theta)$$
$$\mathcal{T}_{\text{target}} = \{(\theta_1^*, \mathbf{X}_1^{\text{new}}) \ldots (\theta_N^*, \mathbf{X}_N^{\text{new}})\} \sim p_{\text{target}}(\theta)\mathcal{L}(\mathbf{x};\theta)$$

Predictive Approach Can Be Very Powerful, But One Needs to Correct for Bias

[with Luca Masserano, Tommaso Dorigo, Rafael Izbicki and Mikael Kuusela]

Data coming from Dorigo et al. (2020): ~ 400'000 **simulated muons** with true incoming energy sampled uniformly between 100 and 2000 GeV.

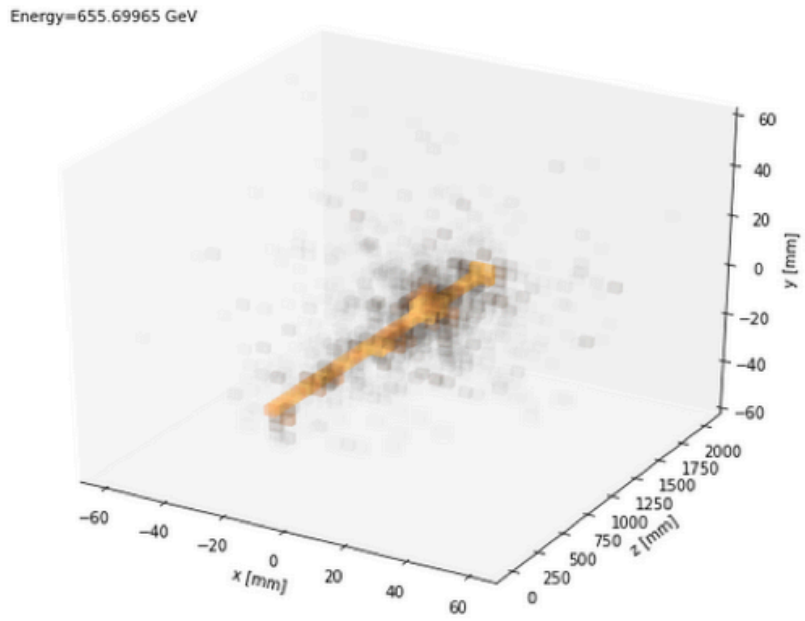


Figure 4: Muon entering the calorimeter in z direction.

[Kieseler et al., July 2021 arXiv:2107.02119]

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim r(\theta),\ \mathbf{X}|\theta \sim F_\theta$$

1. Bias

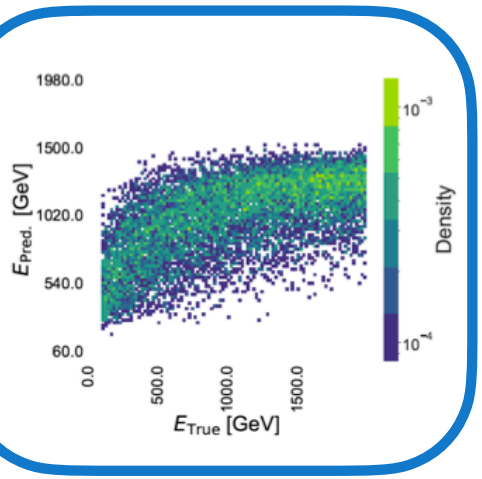


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

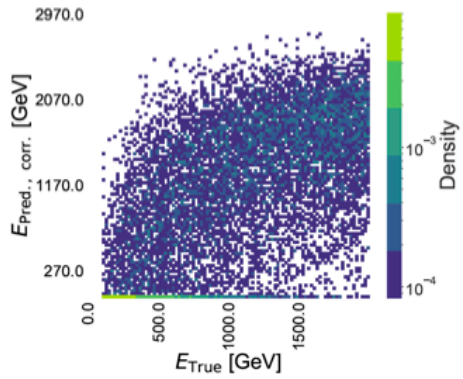


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

$$\mathbb{E}[\theta|X] \neq \theta^\star$$

Source: Dorigo et al 2020.
Slide credit: Luca Masserano

Similarly, posteriors do not guarantee coverage of internal parameters (often "over-confident")

Averting A Crisis In Simulation-Based Inference

https://arxiv.org/abs/2110.06581

Joeri Hermans*
University of Liège
joeri.hermans@doct.uliege.be

Arnaud Delaunoy*
University of Liège
a.delaunoy@uliege.be

François Rozet
University of Liège
francois.rozet@uliege.be

Antoine Wehenkel
University of Liège
antoine.wehenkel@uliege.be

Gilles Louppe
University of Liège
g.louppe@uliege.be

Abstract

We present extensive empirical evidence showing that current Bayesian simulation-based inference algorithms are inadequate for the falsificationist methodology of scientific inquiry. Our results collected through months of experimental computations show that all benchmarked algorithms – (S)NPE, (S)NRE, SNL and variants of ABC – may produce overconfident posterior approximations, which makes them demonstrably unreliable and dangerous if one's scientific goal is to constrain parameters of interest. We believe that failing to address this issue will lead to a well-founded trust crisis in simulation-based inference. For this reason, we argue that research efforts should now consider theoretical and method-

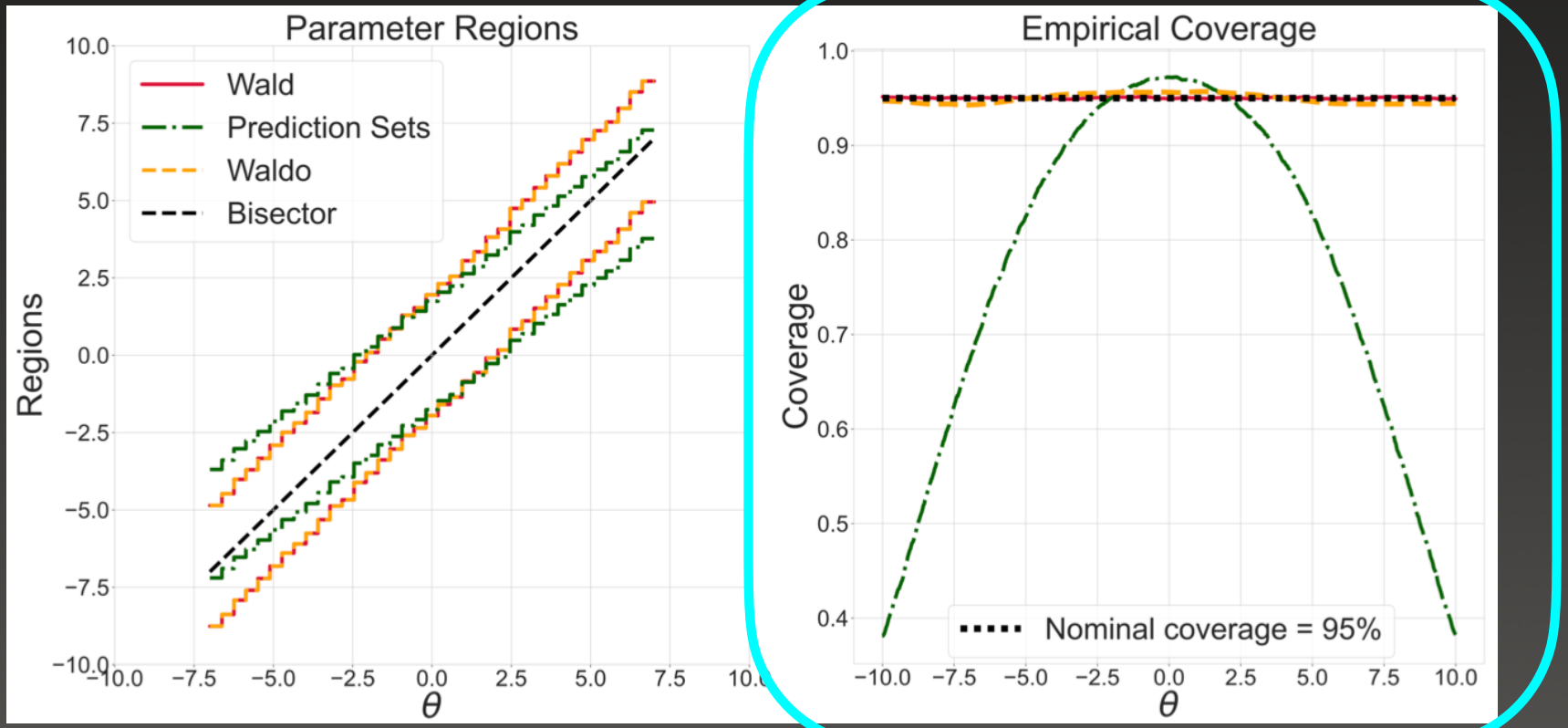
evaluation requires the often *intractable* integration of all stochastic execution paths. In this problem setting, statistical inference based on the likelihood becomes impractical. However, approximate inference remains possible by relying on likelihood-free *approximations* thanks to the increasingly accessible and effective suite of methods and software from the field of simulation-based inference (Cranmer et al., 2020).

While simulation-based inference targets domain sciences, advances in the field are mainly driven from a machine learning perspective. The field, therefore, inherits the quality assessments (Lueckmann et al., 2021) customary to the machine learning literature, such as the minimization of classical divergence criteria. Despite recent developments of post hoc diagnostics to inspect the quality of likelihood-free approximations (Cranmer et al., 2015; Brehmer et al., 2018, 2019; Hermans et al., 2021; Lueckmann et al., 2021; Talts et al.,

Ex: Coverage of Prediction and Posterior Intervals Depends on the Choice of Prior

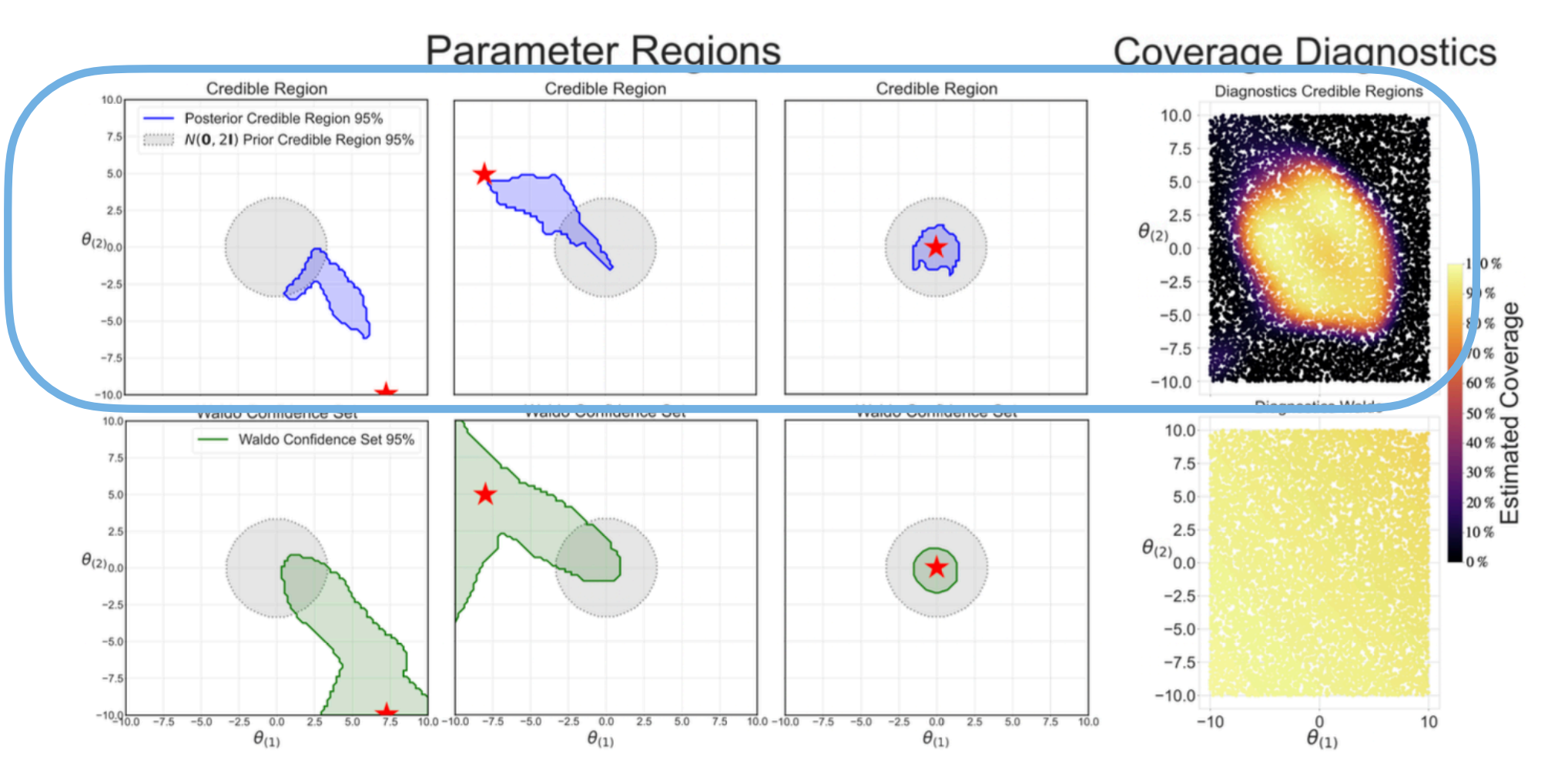
- Likelihood: $\mathscr{D} \mid \theta \sim \mathcal{N}(\theta,\ \sigma = 1)$
- Assume prior: $\theta \sim \mathcal{N}(\mu = 0, \sigma = 2)$

$\Rightarrow$ freq coverage (green curve) for 95% credible set



Ex: Credible Regions from Neural (NF) Posteriors

$$\mathcal{D}|\boldsymbol{\theta} \sim \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \odot \mathbf{I}), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^2 \text{ and } n = 1$$



Blue contours: 95% credible regions from Normalizing Flows (overly confident when prior is poorly specified)

So what does this mean for reliable scientific inference?

Is there a way we can still take advantage of neural posteriors and generative models where AI has made real breakthroughs?

Desiderata

- **Develop LFI procedures with the following properties:**
  1. Robust coverage guarantees under poorly specified priors or shifting priors
  2. Tight constraints when the prior is consistent with D_{obs}

Desiderata

- **Develop LFI procedures with the following properties:**

1. Robust coverage guarantees under poorly specified priors or shifting priors
2. Tight constraints when the prior is consistent with D_{obs}
3. Valid for all $\theta \in \Theta$, and in finite samples (e.g., $n = 1 \rightarrow$ single observation from $\theta^{\star}$)
4. Interpretable diagnostics that checks coverage across the entire parameter space; local/conditional and not just average/marginal coverage

Desiderata

- **Develop LFI procedures with the following properties:**
  1. Robust coverage guarantees under poorly specified priors or shifting priors
  2. Tight constraints when the prior is consistent with D_{obs}
  3. Valid for all $\theta \in \Theta$, and in finite samples (e.g., $n = 1 \rightarrow$ single observation from $\theta^\star$)
  4. Interpretable diagnostics that checks coverage across the entire parameter space; local/conditional and not just average/marginal coverage
- **How? Bridge advances in LFI with a principled statistical inference framework:**

  Leverage predictive NNs or NDEs and generative models to construct sets $\mathscr{R}(\mathscr{D})$ such that

  $$\mathbb{P}_{\mathscr{D}|\theta}\left(\theta \in \mathscr{R}(\mathscr{D})\right) = 1 - \alpha, \quad \forall \theta \in \Theta$$ **(Local validity)**

  and $\mathbb{E}[\,|\mathscr{R}(\mathscr{D})|\,]$ is small for target data **(High constraining power)**

Can we have it all?

Valid inference even for small sample size (e.g. n=1), and poorly specified prior.

But higher constraining power if well-specified prior.

Diagnostics across the entire parameter space.

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\right) = 1 - \alpha, \quad \forall \theta \in \Theta$$

* *All done by leveraging already existing posterior estimates, or ML predictive tools "as is".*

** Modular procedures with theoretical guarantees.

Can we have it all?

Valid inference even for small sample size (e.g. n=1), and poorly specified prior.

But higher constraining power if well-specified prior.

Diagnostics across the entire parameter space.

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\right) = 1 - \alpha, \quad \forall \theta \in \Theta$$

* *All done by leveraging already existing posterior estimates, or ML predictive tools "as is".*

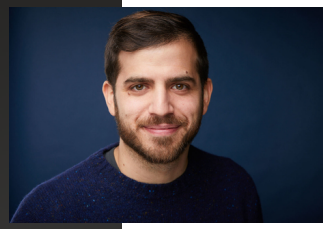
** Modular procedures with theoretical guarantees.

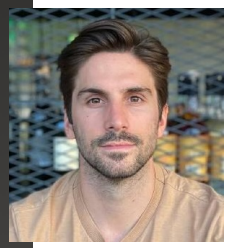
General Inference Machinery for LFI

arXiv:2107.03920 (to appear in EJS) arXIv:2002.10399 (ICML 2021)

LF2I

Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference*





Niccolò Dalmasso[1,†], Luca Masserano[2,†], David Zhao[2], Rafael Izbicki[3], Ann B. Lee[2]

[1] *Department of Statistics and Data Science, Carnegie Mellon University*
e-mail: niccolo.dalmasso@gmail.com

[2] *Department of Statistics and Data Science, Machine Learning Department, Carnegie Mellon University*
e-mail: lmassera@andrew.cmu.edu, *e-mail:* davidzhao@andrew.cmu.edu
e-mail: annlee@andrew.cmu.edu

[3] *Department of Statistics, Federal University of São Carlos*
e-mail: rafaelizbicki@gmail.com

Hypothesis Testing and Confidence Sets

Key ingredients:

- data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$
- a test statistic, such as likelihood ratio statistic $\Lambda(\mathcal{D}; \theta_0)$
- an α-level critical value $C_{\theta_0,\alpha}$

Reject the null hypothesis H_0 if $\Lambda(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$

Theorem (Neyman inversion, 1937)

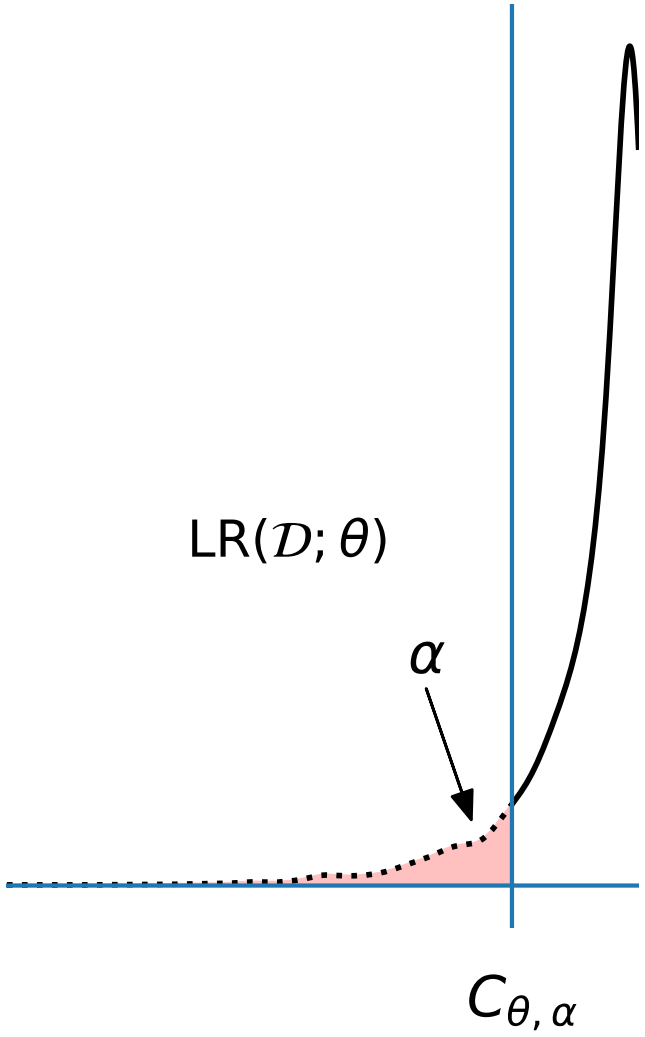
Building a $1 - \alpha$ confidence set for θ is equivalent to testing

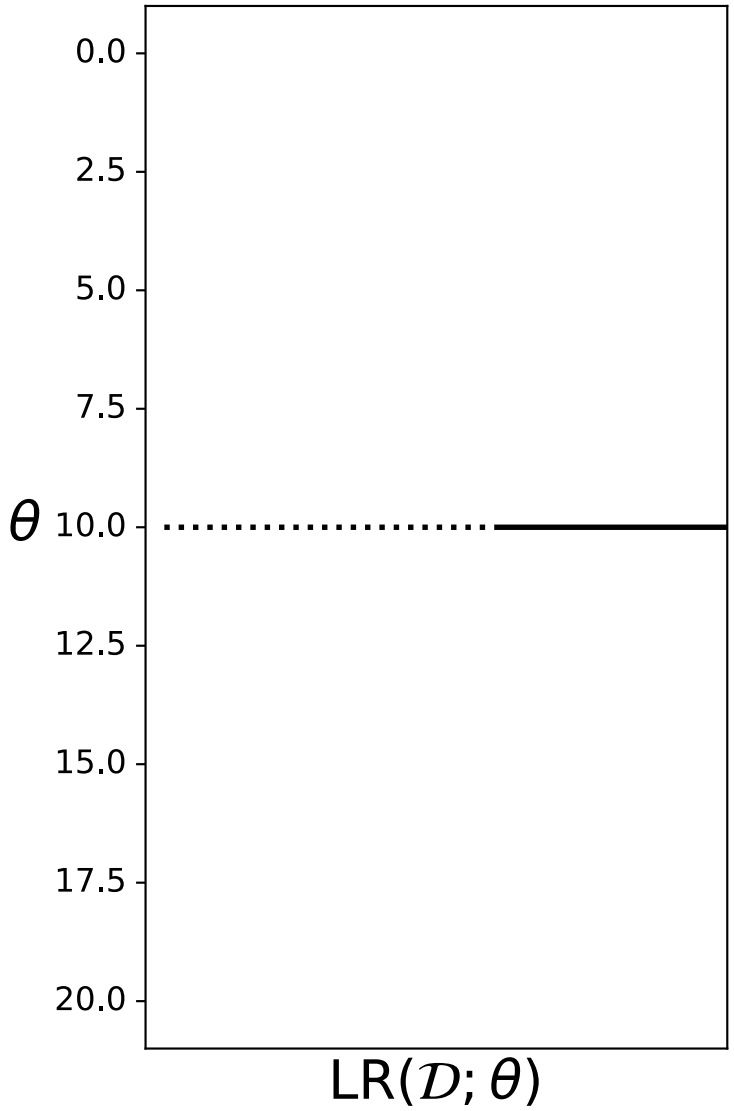
$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for θ_0 across the parameter space.

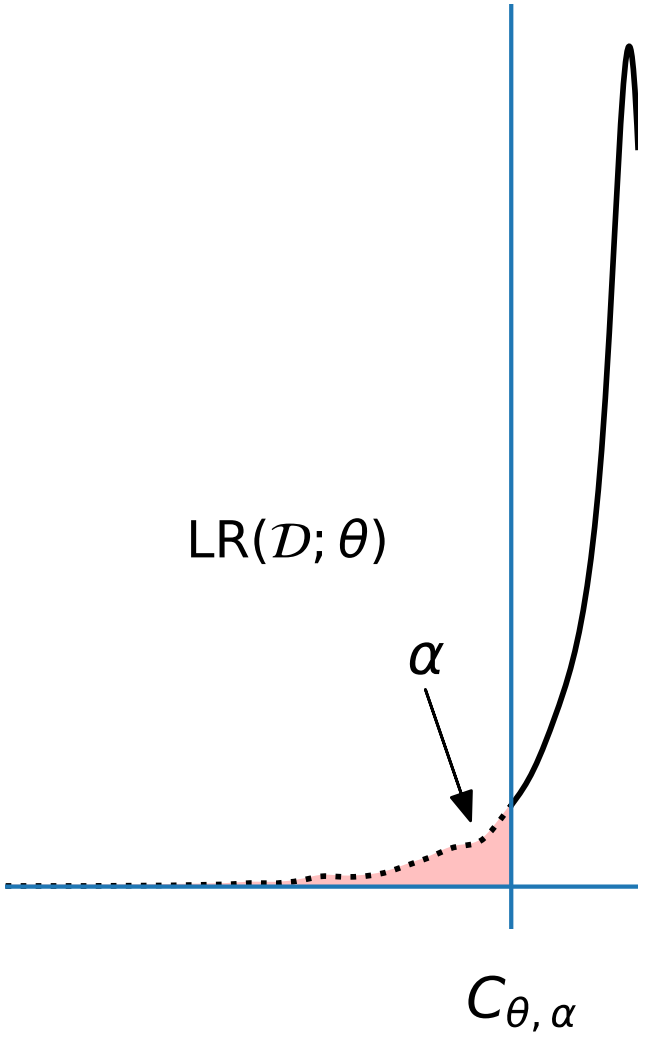
Slide credit: Nic Dalmasso

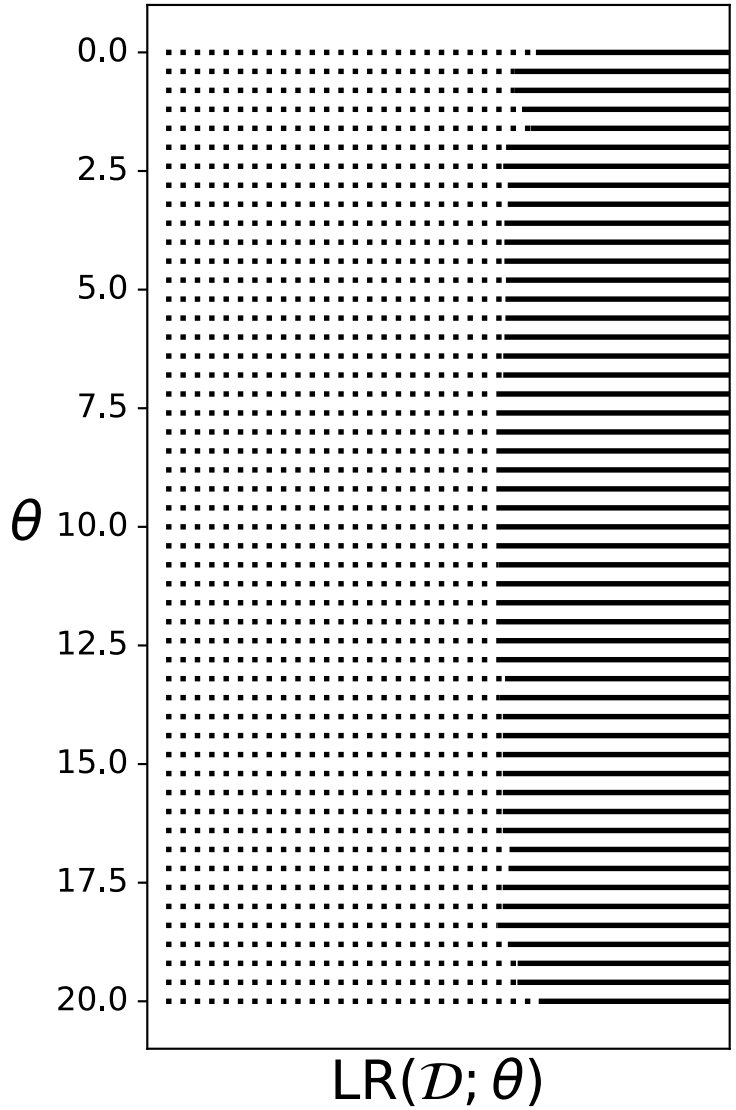
1. Fixed θ. Find the rejection region for test statistic λ.



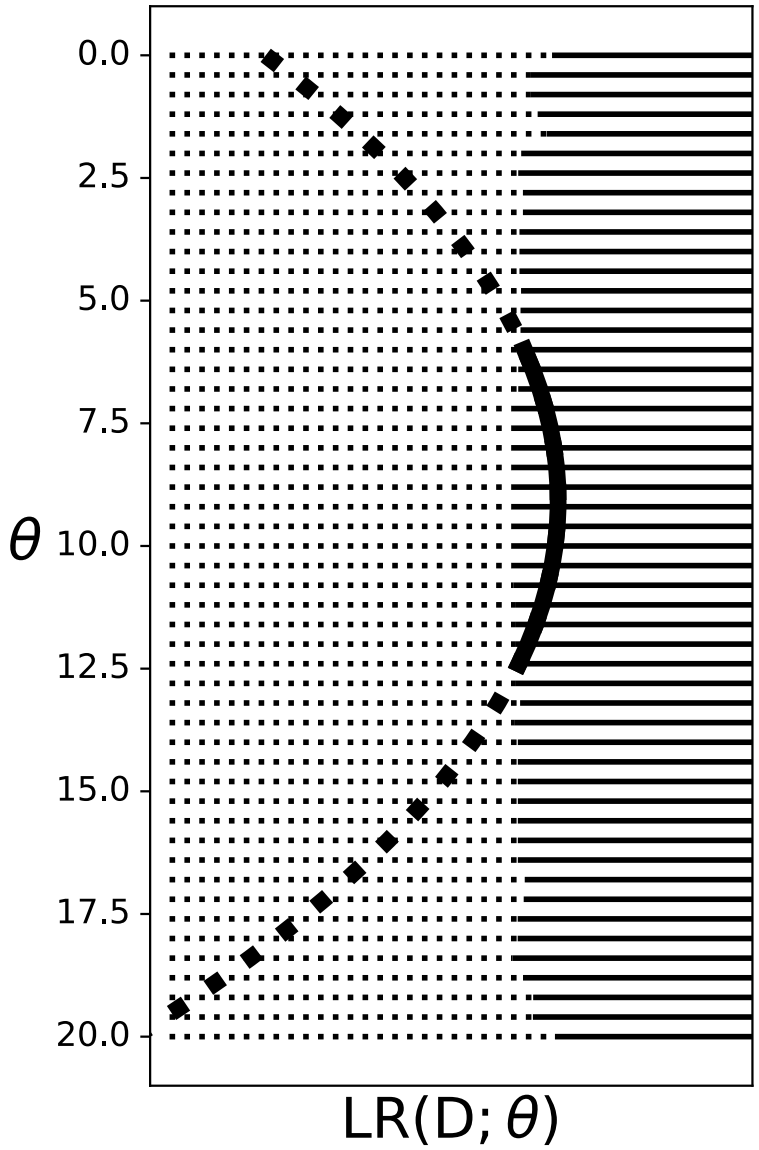


2. Repeat for every θ in parameter space.

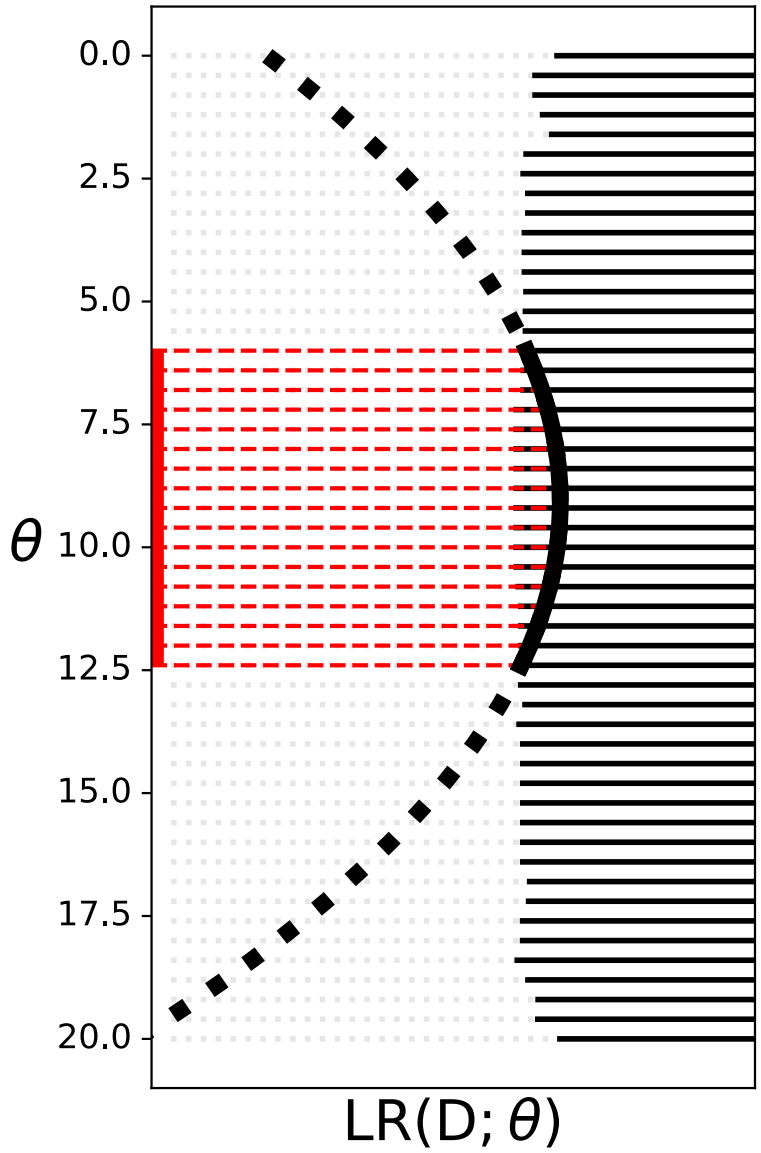




3. Observe data $\mathcal{D} = \mathrm{D}$. Evaluate $\lambda(\mathrm{D}; \theta)$.



4. Construct $(1 - \alpha)$ confidence set for θ.



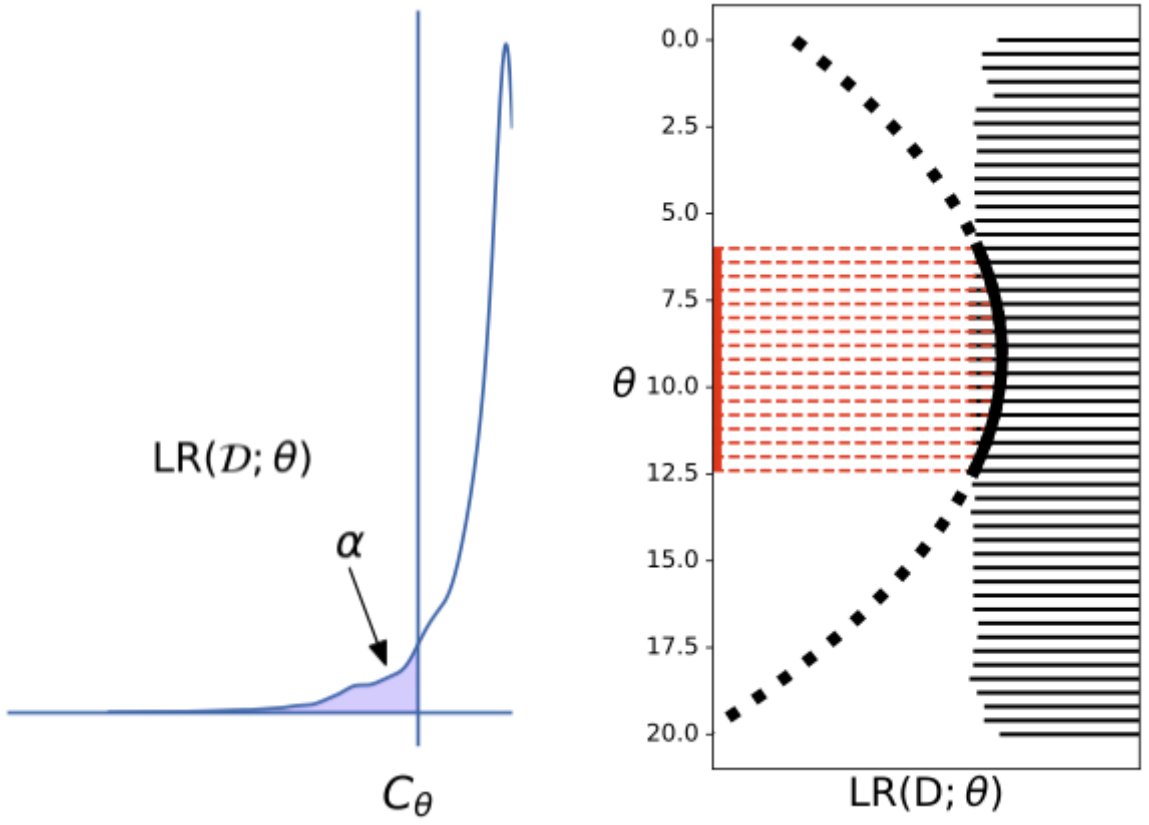
Challenges to Computing Frequentist Confidence Sets

- Inverting hypothesis tests at every parameter $\theta_0 \in \Theta$ does not appear computationally feasible…
- Hence, most (if not all) frequentist confidence sets rely on large-sample theory; e.g. assuming that the LRT statistic is approximately chi-squared distributed

However, LF2I framework requires neither large-sample theory or MC/bootstrap sampling …

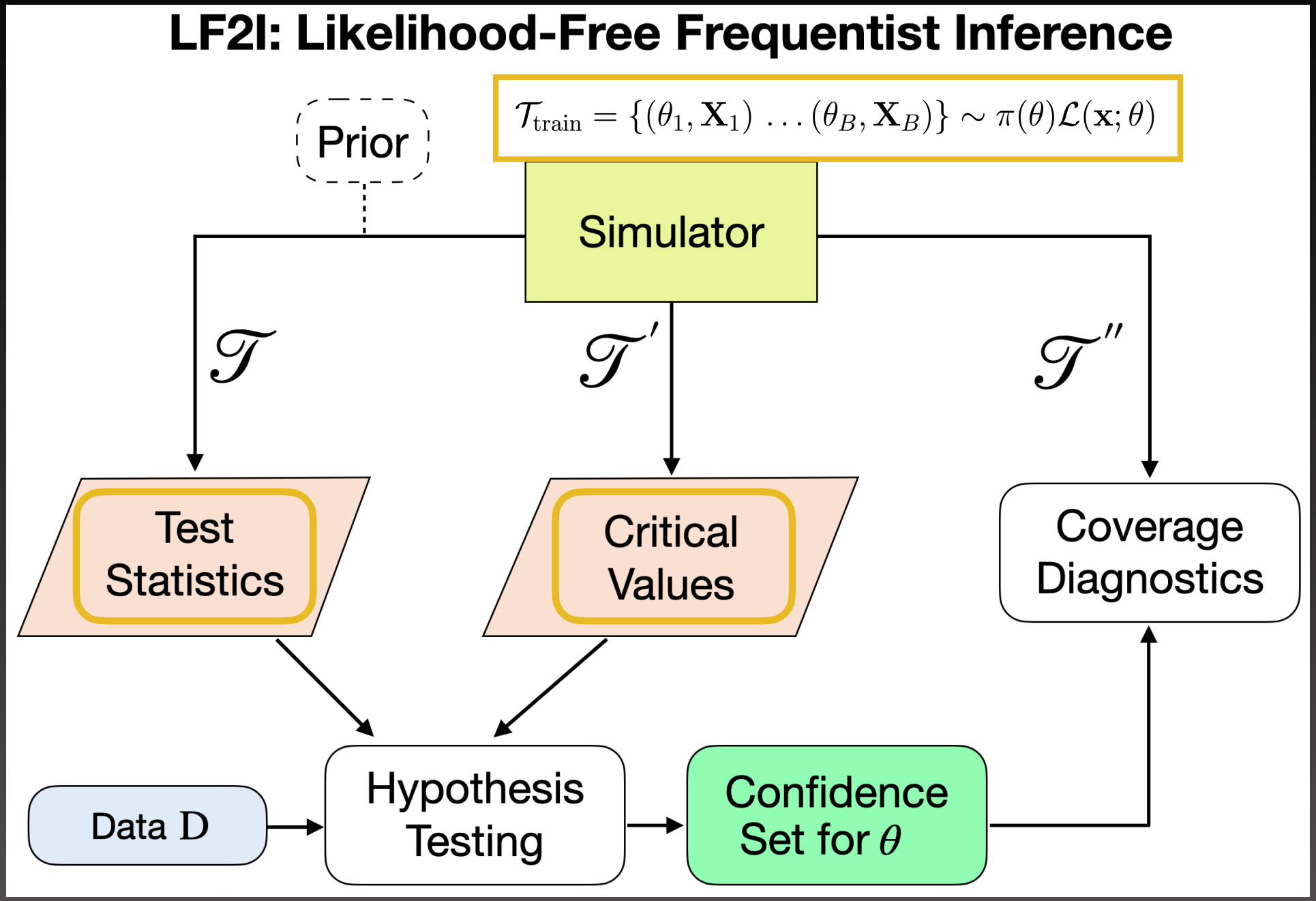
Efficient Construction of Finite-Sample Confidence Sets

LF2I

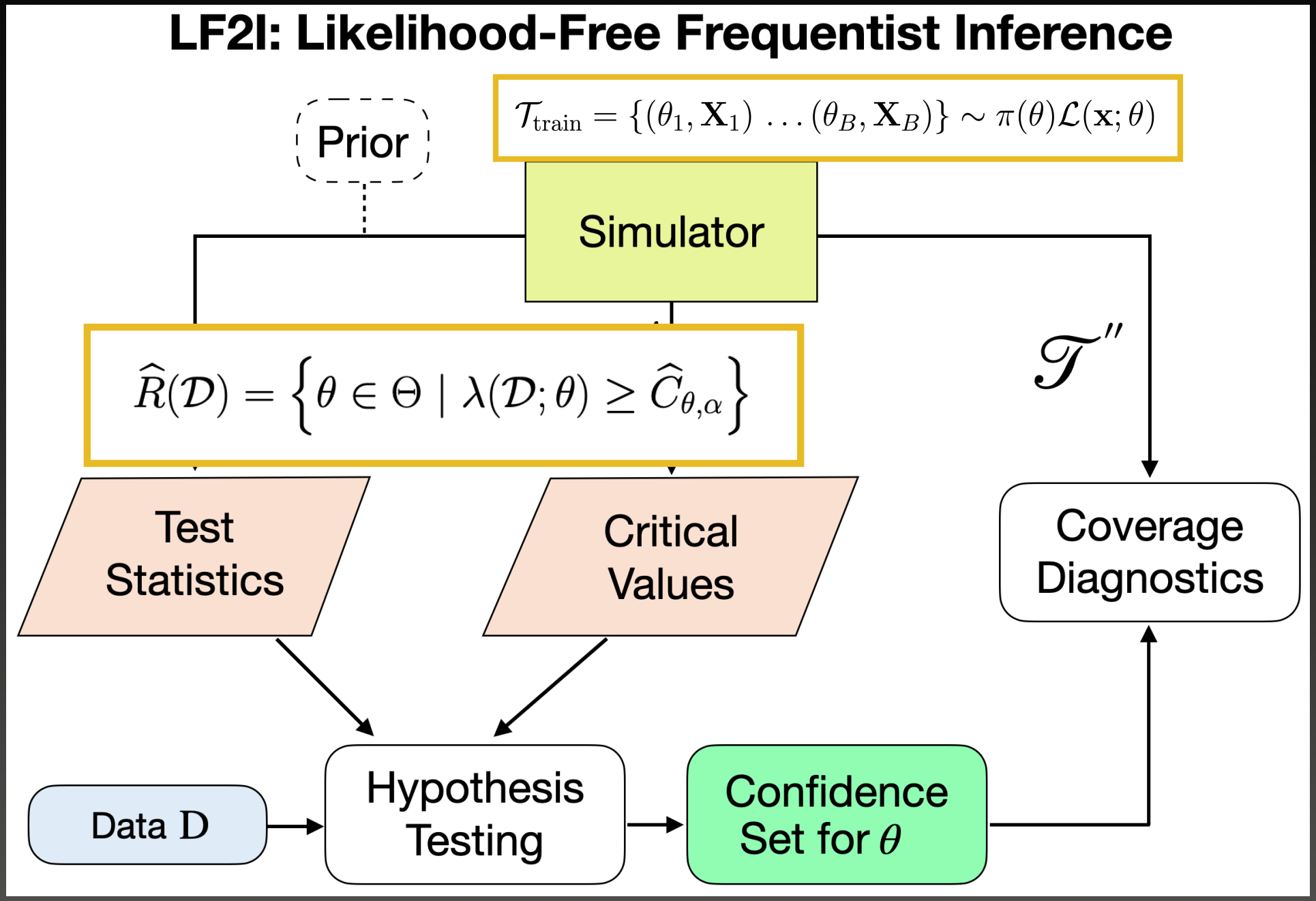


Rather than running a batch of ~~Monte Carlo~~ simulations for every null hypothesis $\theta = \theta_0$ on, e.g., a fine enough grid in Θ, we can interpolate across the parameter space using training-based ML algorithms.

Our Inference Machinery



Construct Confidence Set via Neyman Inversion



What Test Statistic?

- Originally, we were defining likelihood-based test statistics:
  - → ACORE (approximate LRT) [Izbicki et al 2013; Cranmer et al 2015; Dalmasso et al 2020, arXiv:2002.10399]
  - → BFF (approximate Bayes Factor) [Dalmasso et al 2021, arXiv:2107.03920; Heinrich 2022, arXiv: 2203.13079]
- Now, we are moving toward test statistics computed from posteriors
  - → "WALDO" (modified Wald test statistic) [Masserano et al 2022, arXiv:2205.15680]
  - → "Bayes-Frequentist sets" [Masserano, Carzon et al 2024-]

arXiv:2205.15680 (AISTATS 2023)

Simulation-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms or Posterior Estimators for Inverse Problems

Luca Masserano[1] **Tommaso Dorigo**[2] **Rafael Izbicki**[3] **Mikael Kuusela**[1] **Ann B. Lee**[1]
[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]INFN, Sezione di Padova [3]Department of Statistics, Federal University of São Carlos

Abstract

Predictive algorithms, such as deep neural net-

Theorem (Neyman 1937)

Constructing a $1-\alpha$ confidence set for θ is equivalent to testing

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for every $\theta_0 \in \Theta$.

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{\left(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0\right)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

1 INTRODUCTION

The vast majority of modern machine learning targets pre-
on problems, with algorithms such as Deep Neural
of a data-generating process with reliable measures of
uncertainty. The parameters of interest, which we denote by

certainty quantification, especially when both pa-

- **Wald** test statistic (1D case):

$$\tau^{Wald}(\mathscr{D}; \theta_0) := \frac{(\theta^{MLE} - \theta_0)^2}{\mathbb{V}[\theta^{MLE}]}$$

- **Waldo** test statistic (1D and p-D case):

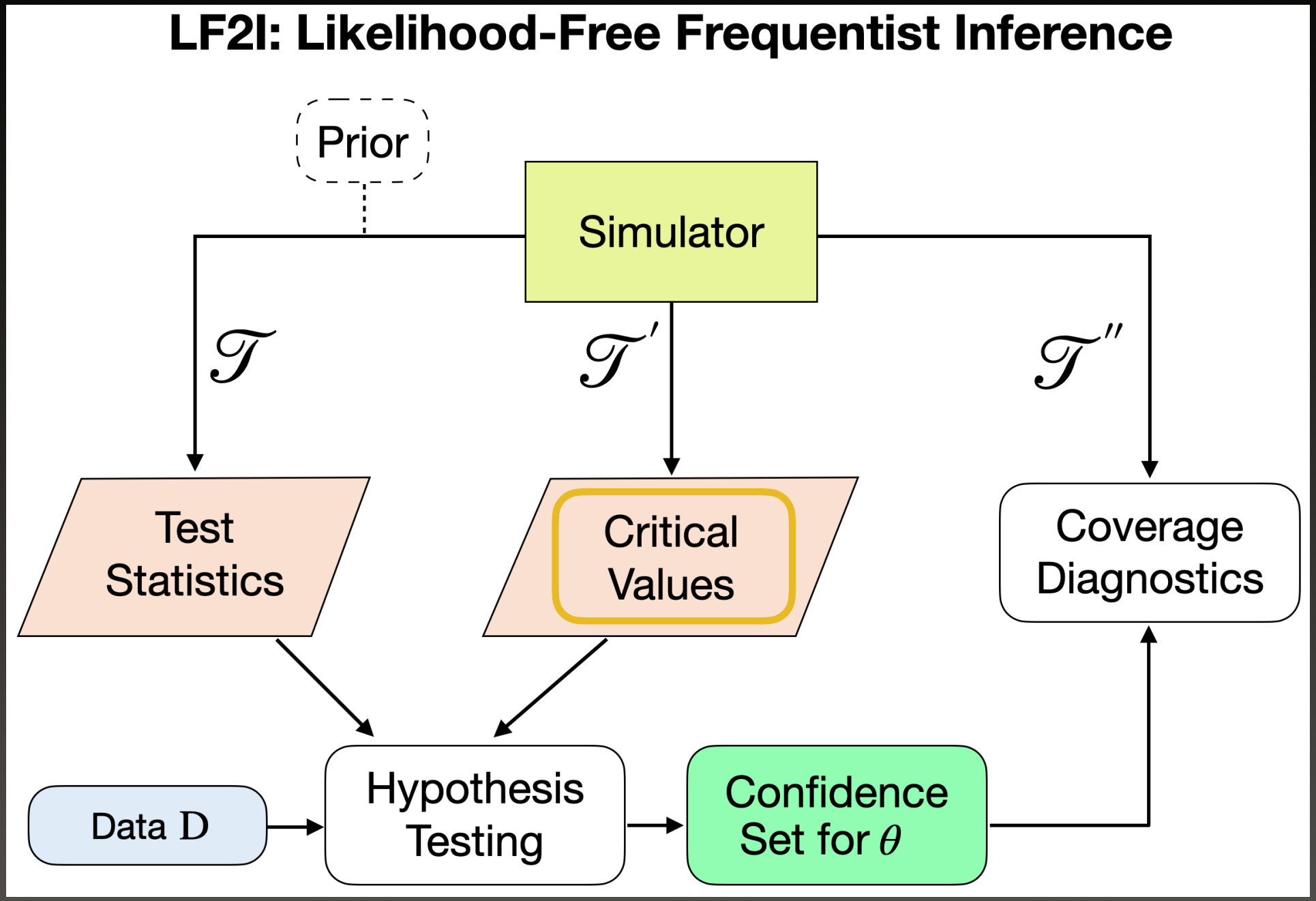
$$\tau^{Waldo}(\mathscr{D}; \theta_0) := \frac{(\mathbb{E}[\theta \,|\, \mathscr{D}] - \theta_0)^2}{\mathbb{V}[\theta \,|\, \mathscr{D}]}$$

$$\tau^{Waldo}(\mathscr{D}; \theta_0) := (\mathbb{E}[\theta \,|\, \mathscr{D}] - \theta_0)^T \mathbb{V}[\theta \,|\, \mathscr{D}]^{-1} (\mathbb{E}[\theta \,|\, \mathscr{D}] - \theta_0)$$

plicated to be evaluated explicitly. Let $\mathcal{D} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)$
denote observable data, where the "sample size" n refers

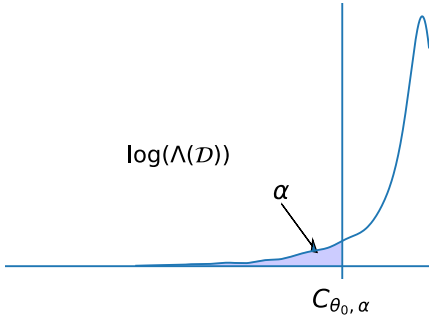
sample theory. Many simulator-based inference
(SBI) methods are indeed known to produce bi-

Center Branch: Estimate Critical Values



Estimating Critical Values $C_{\theta_0,\alpha}$

To control Type I error at level α:
Reject $H_0 : \theta = \theta_0$ when $\lambda(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$, where

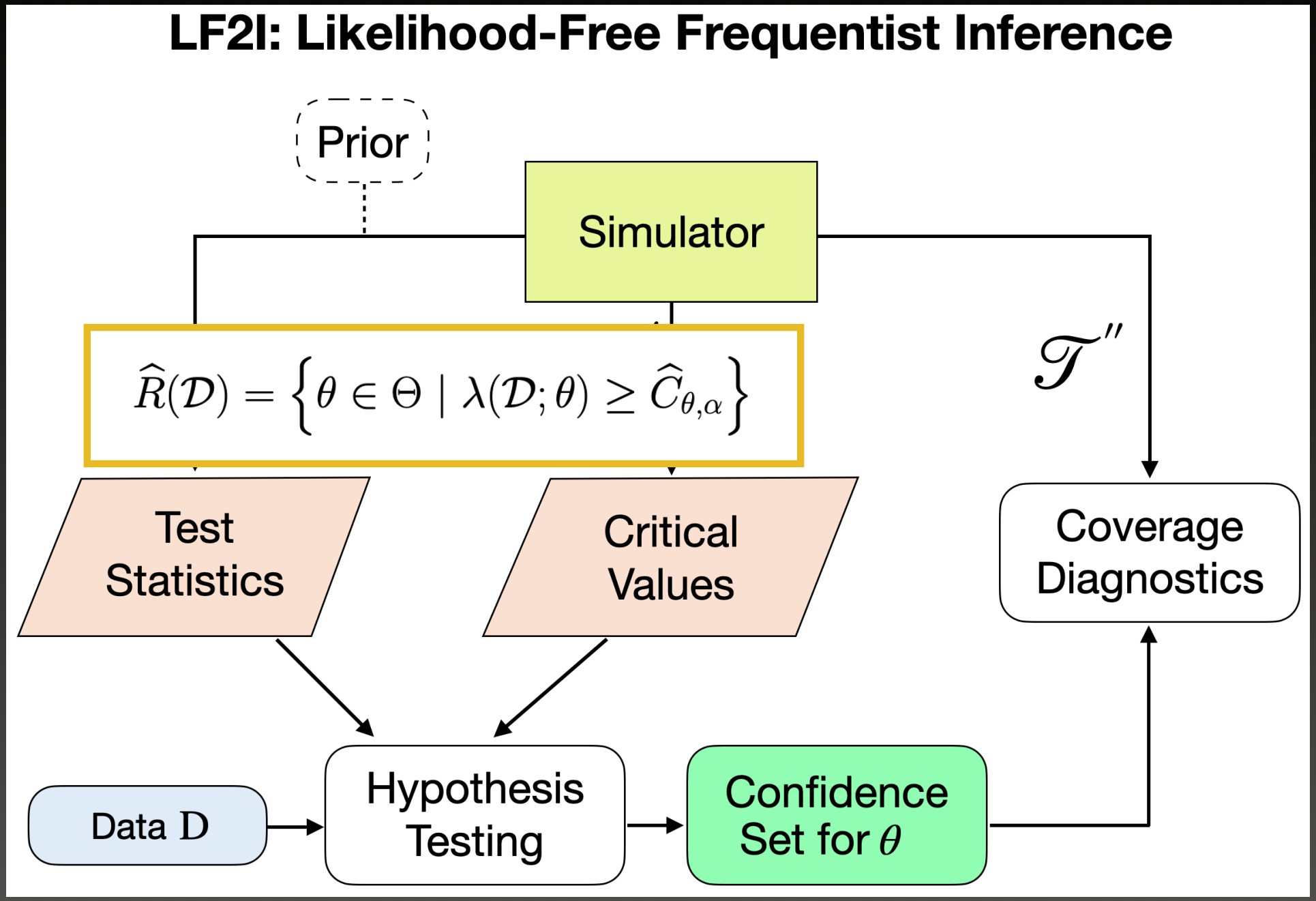


$$C_{\theta_0,\alpha} = \arg\sup_{C \in \mathbb{R}} \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0} \left(\lambda(\mathcal{D}; \theta_0) < C \right) \leq \alpha \right\}.$$

Problem: Need to compute $\mathbb{P}_{\mathcal{D}|\theta} \left(\lambda(\mathcal{D}; \theta) < C \right)$ for every $\theta \in \Theta$.

Solution: $F_{\lambda|\theta}(C \mid \theta) \equiv \mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \theta) < C \mid \theta)$ is a conditional CDF, so we can estimate its α-quantile via quantile regression $F^{-1}_{\lambda|\theta}(\alpha|\theta)$.

Construct Confidence Set via Neyman Inversion



Are the Constructed Confidence Sets Valid?

Theorem (Validity for any test statistic)

Let $C_{B'}$ be the critical value of a level-α test based on the statistic $\lambda(\mathcal{D}; \theta_0)$. Then, if the quantile regression estimator is consistent,

$$C_{B'} \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} C^*,$$

where C^* is such that

$$\mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \theta_0)) \leq C^*) = \alpha.$$

NOTE: Regardless of the number of observations n, how well we estimate the test statistic, and the choice of prior π_θ

If $B^{'}$ is large enough, we can construct a confidence set with guaranteed nominal coverage regardless of the observed sample size n.

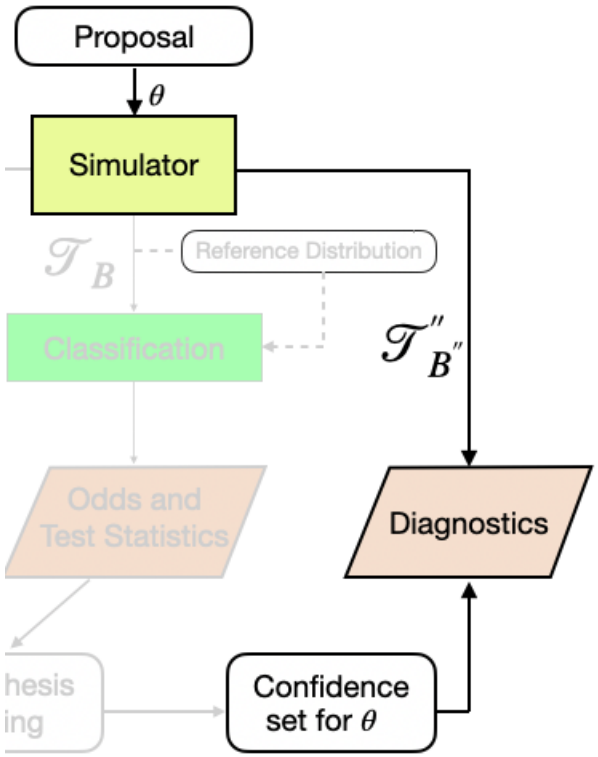
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

How do we check coverage of constructed confidence sets across Θ?

Note:

$$\widehat{R}(\mathcal{D}) = \left\{\theta \in \Theta \mid \lambda(\mathcal{D};\theta) \geq \widehat{C}_{\theta,\alpha}\right\}$$

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D}) \mid \theta\right) = \mathbb{E}_{\mathcal{D}|\theta}\left[\mathbb{I}\left(\theta \in \widehat{R}(\mathcal{D})\right) \mid \theta\right]$$



1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $\widehat{R}(\mathcal{D}_i)$
3. For $\{\theta_i, \widehat{R}(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i := \mathbb{I}(\theta_i \in \widehat{R}(\mathcal{D}_i))$ on θ_i.

Independent check of coverage across parameter space

How close is the actual coverage to the nominal confidence level $1-\alpha$?

Ex: Diagnostics for Classical "On-Off" Problem

[Lyons 2008; Cowan et al 2011; Cowan 2012; L. Heinrich 2022]

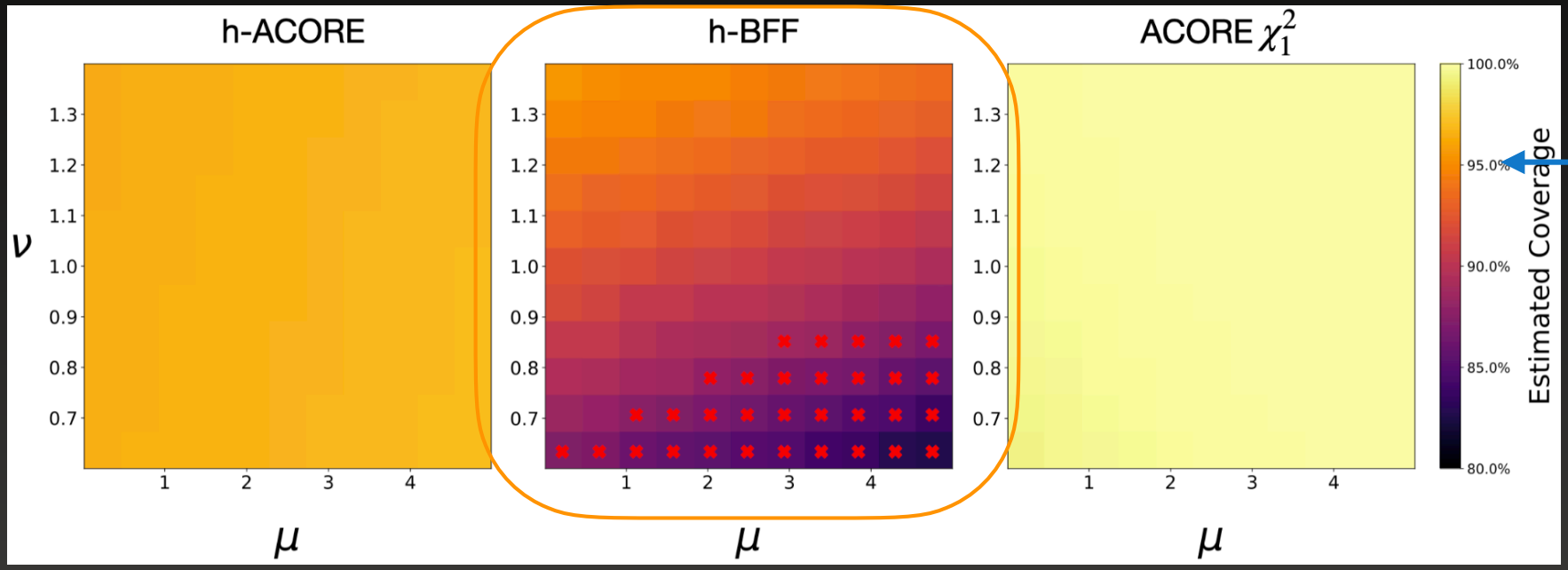
- Simultaneous measurements of two Poisson processes

$$\text{Observed data } \mathbf{X} = (N_b, N_s),$$
$$\text{where } N_b \sim \text{Pois}(\nu\tau b),\ N_s \sim \text{Pois}(\nu b + \mu s)$$

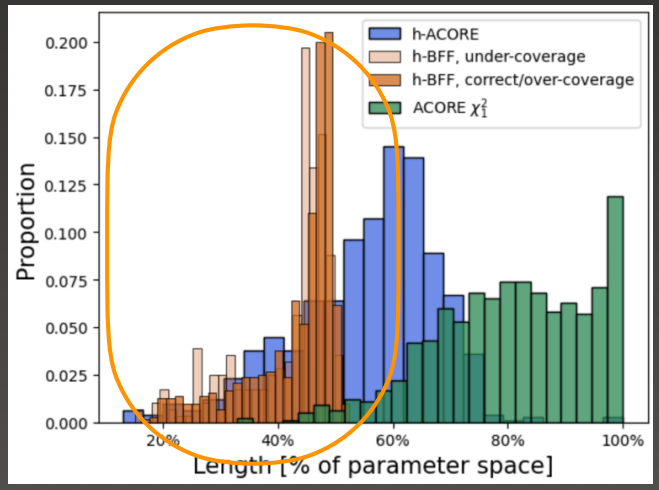
- N_B is the # of events in the background region (expected background count b)
- N_S is the # of events in the contaminated signal region (expected signal count s)
- Unknown parameters:
  - signal strength-POI (μ); scaling factor-NP (ν)
  - [L. Heinrich 2022] Set hyper-parameters at s=15, b=70, τ=1 ⇒ asymptotic regime with profiled values away from the MLE

Our diagnostic tool can identify regions in parameter space with under/over-coverage (95% nominal)

Left: LRT with profiling; Center: marginalization; Right: chi-square)



h-BFF (center top) has closest to nominal coverage with the highest constraining power (orange hist)



Back to the Problem of Calorimetric Muon Energy Measurement… [Masserano et al, AISTATS 2023]

Data coming from Dorigo et al. (2020): ~ 400'000 **simulated muons** with true incoming energy sampled uniformly between 100 and 2000 GeV.

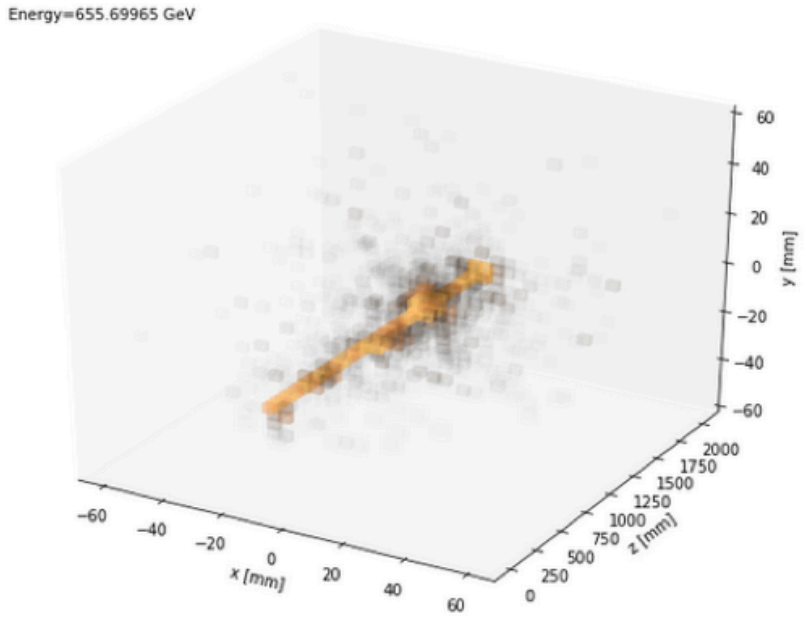


Figure 4: Muon entering the calorimeter in z direction.

[Kieseler et al., July 2021 arXiv:2107.02119]

1. Bias

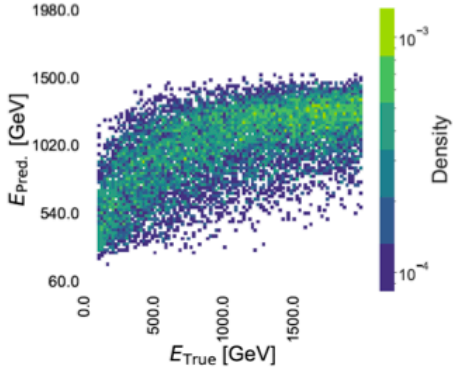


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

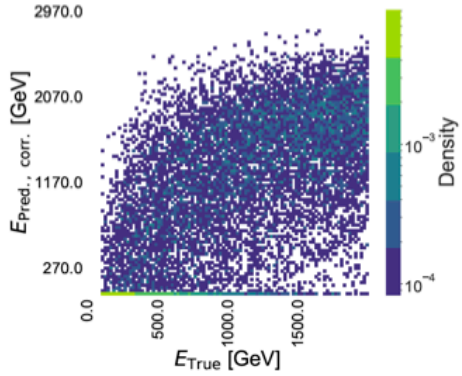


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

$$\mathbb{E}[\theta|X] \neq \theta^{\star}$$

Source: Dorigo et al 2020.
Slide credit: Luca Masserano

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim r(\theta),\ \mathbf{X}|\theta \sim F_\theta$$

Figure 6: Road transport of a structure for the ATLAS air toroids. Photo reproduced from Ref. [181].

Can we do frequentist inference for muon energy?

We are mainly interested in **two questions**:

1. Infer, from the pattern of the energy deposits in the calorimeter, how much energy the incoming muon had *and* construct a **confidence set for it with proper coverage**

   → **goal**: Reconstruct muon properties with rigorous uncertainties for downstream analyses

2. How much added value does a **high granularity of the calorimeter** cells offer over the 1D and 28D representations?

   → **goal**: devise better and more cost-effective calorimeters for future particle colliders

Slide credit: Luca Masserano

Inputs: 1D energy-sum, 28 features or full calorimeter

Prediction algorithms used

Three "nested" datasets:

1. One-dimensional energy sum: minimizer of Cross-Validation MSE loss (XGBoost)
2. 27 features + 1D energy sum: minimizer of Cross-Validation MSE loss (XGBoost)
3. Full calorimeter (51200-D) + 28 features: custom CNN (with MSE loss) from Kieseler et al. (2022)

→ We estimate $\mathbb{E}[\theta \,|\, \mathscr{D}]$ and $\mathbb{V}[\theta \,|\, \mathscr{D}]$ for each of these. Muon energy is θ

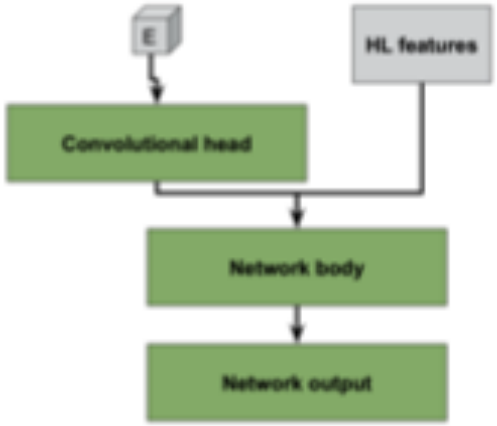


Image credit: Kieseler et al. (2022)

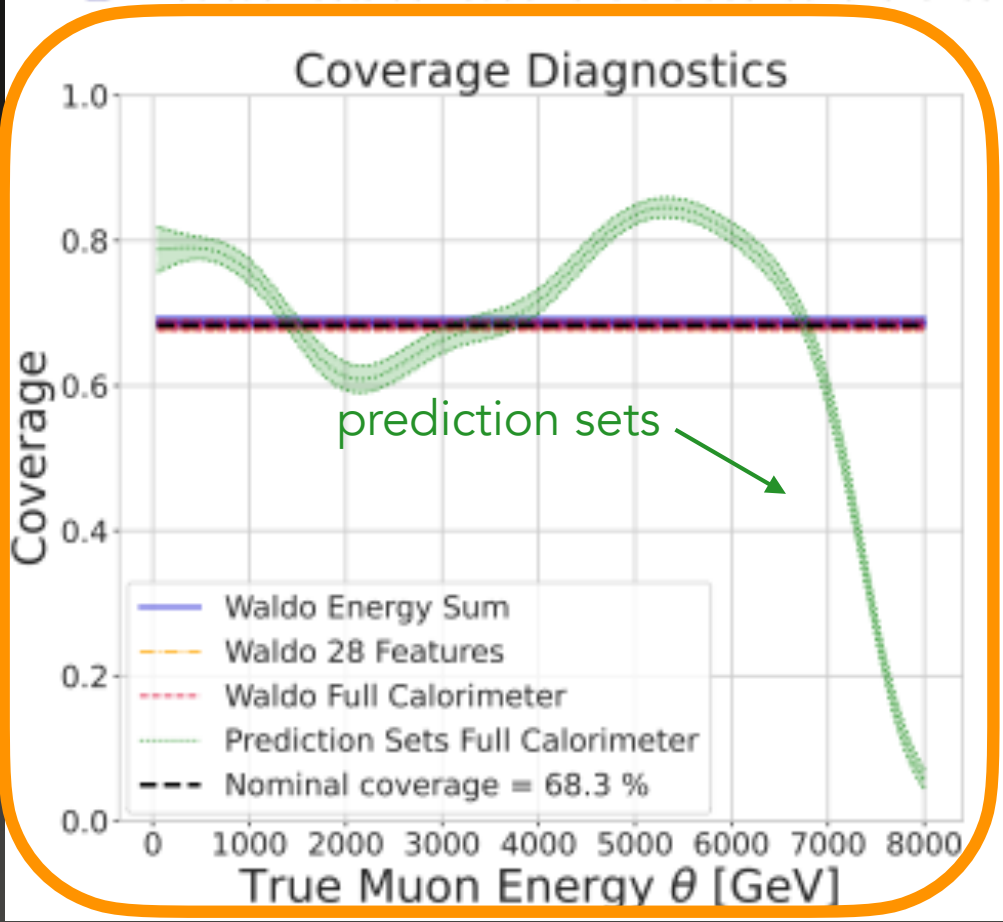
$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

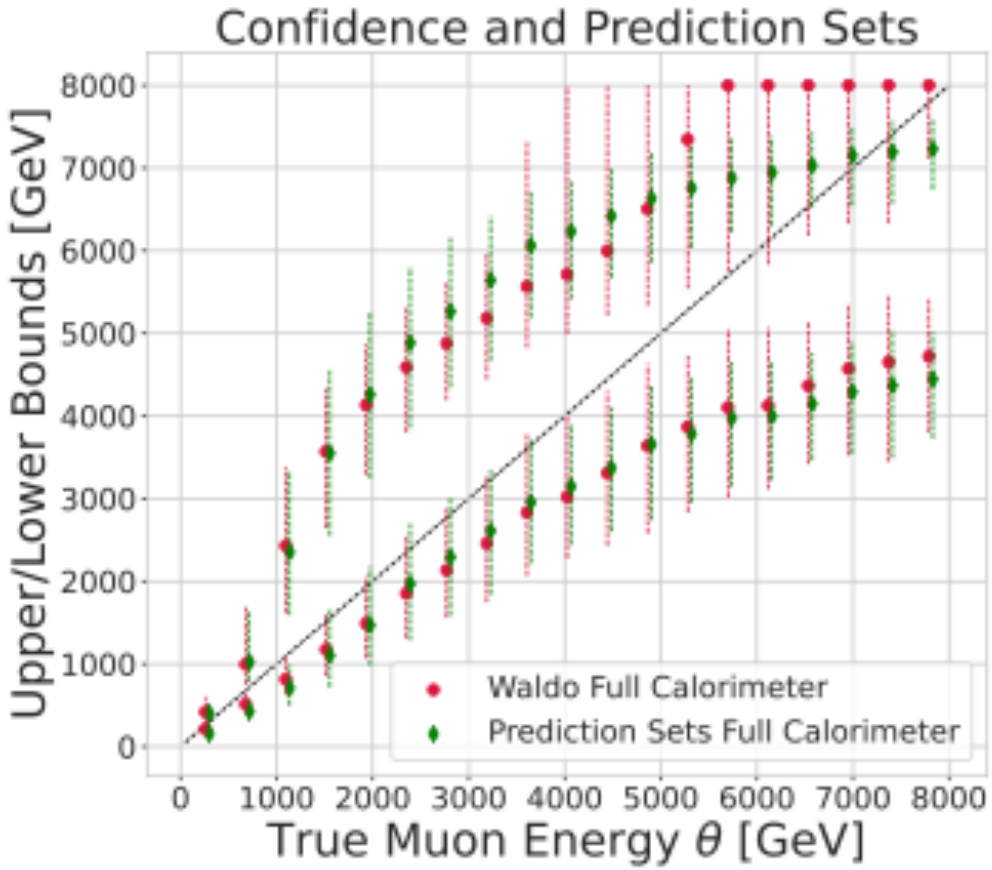
Slide credit: Luca Masserano

Valid confidence sets?

Confidence sets for muon energy have proper coverage

- Nominal coverage is achieved regardless of the dataset used
- Prediction sets do not achieve the desired level of coverage

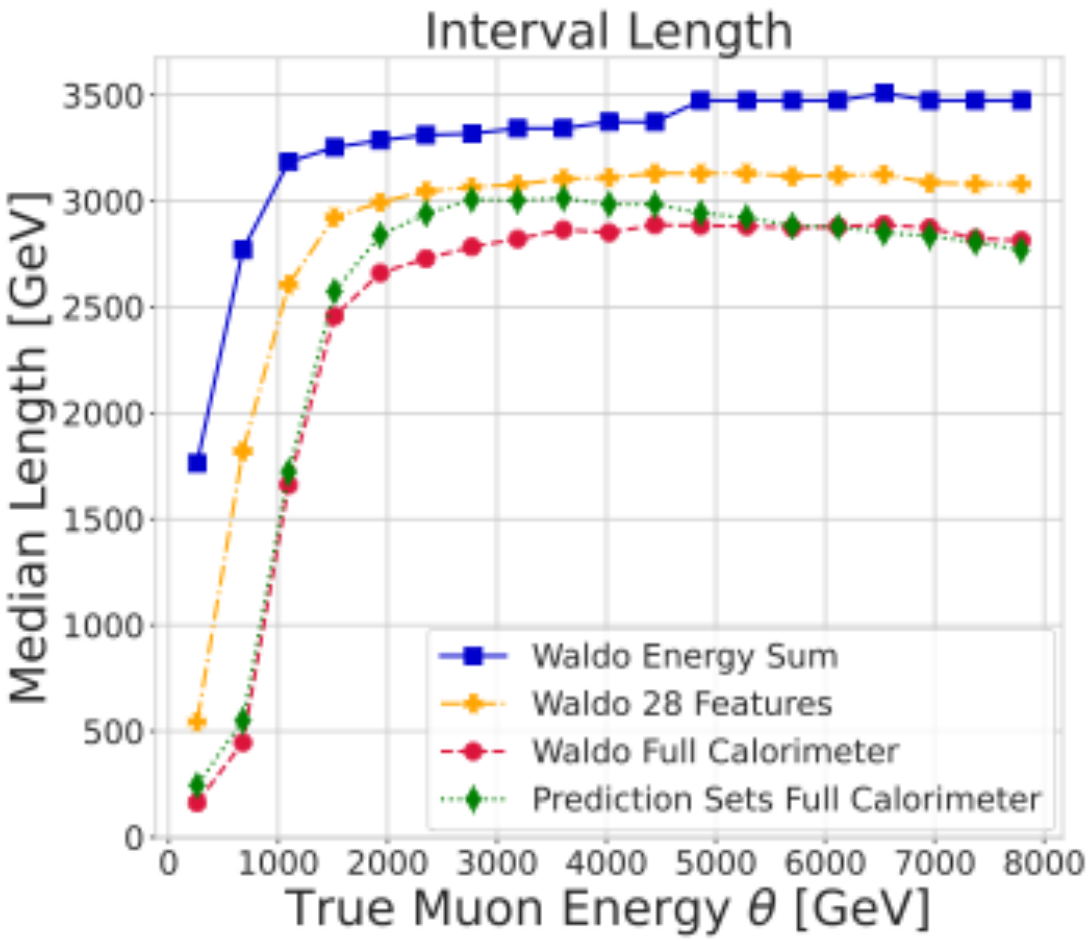




Slide credit: Luca Masserano

Constraining power?

Valuable information in high-granularity calorimeter

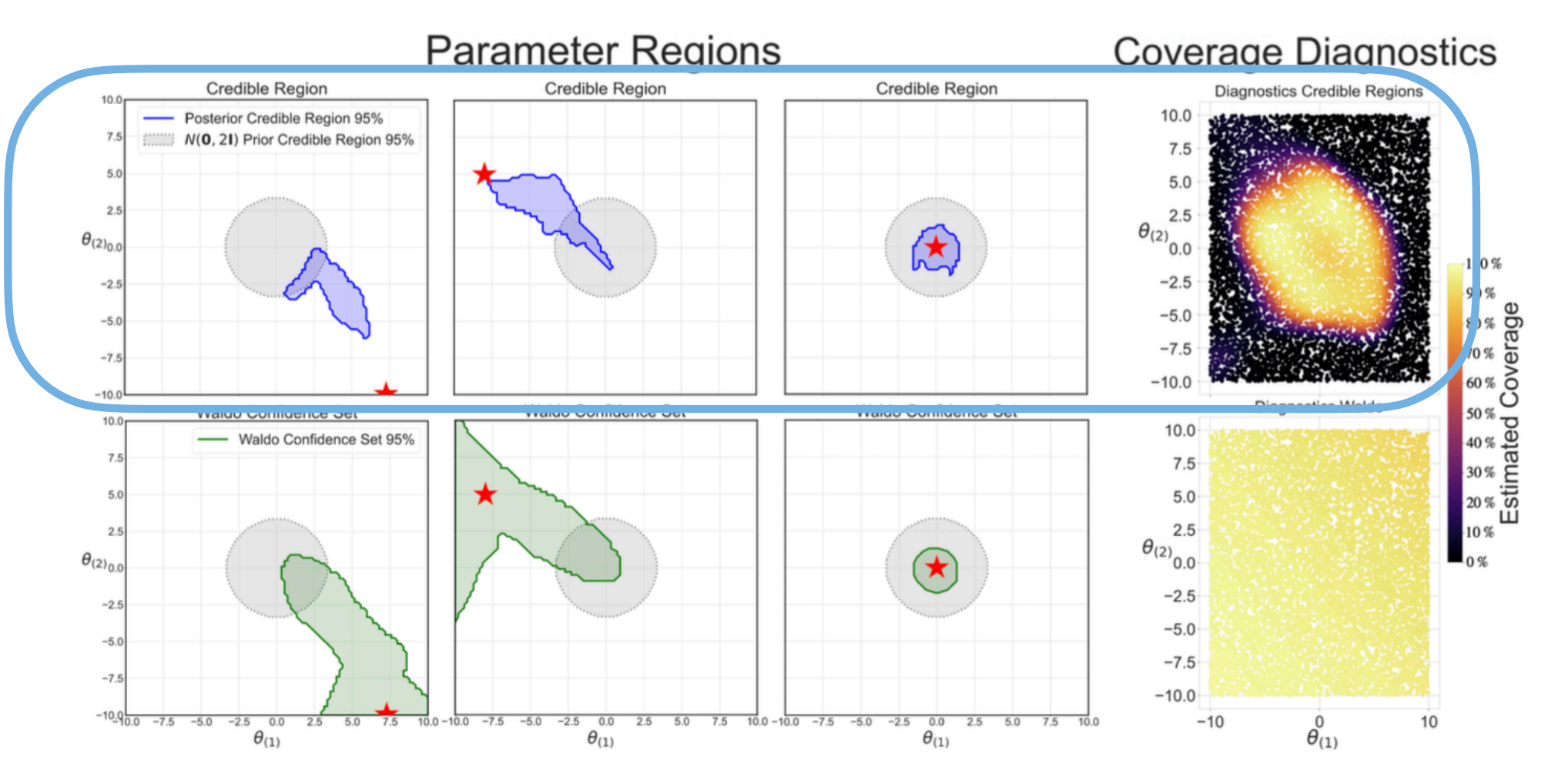


- Intervals are shorter as the data becomes higher-dimensional
- Prediction sets can even be larger than Waldo confidence sets (while also not guaranteeing coverage)

Slide credit: Luca Masserano

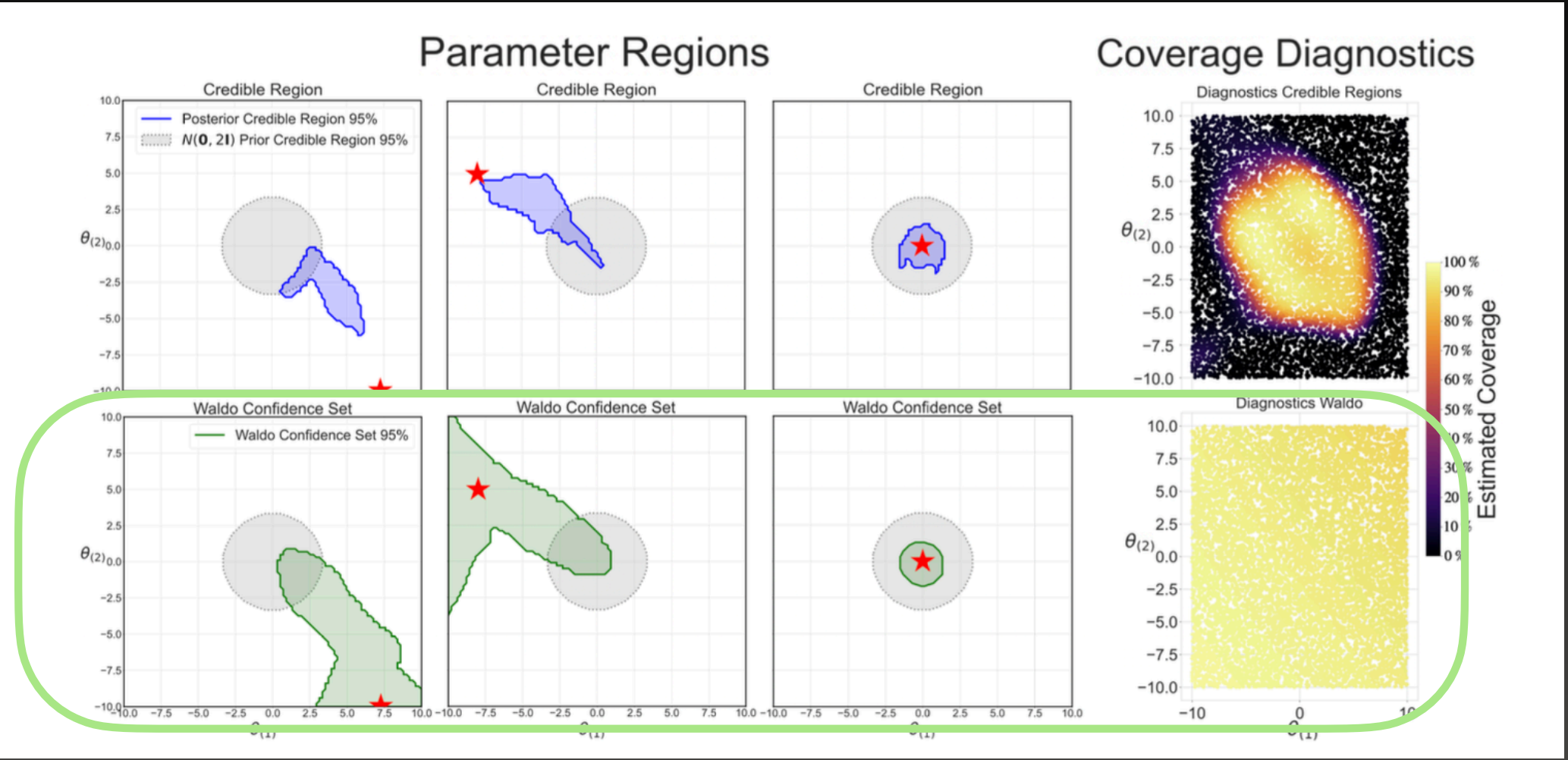
Ex: Credible Regions from Neural (NF) Posteriors

$$\mathcal{D}|\boldsymbol{\theta} \sim \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \odot \mathbf{I}), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^2 \text{ and } n = 1$$



Blue contours: 95% credible regions from Normalizing Flows (overly confident when prior is poorly specified)

Ex: LF2I/Waldo Confidence Sets Derived from the Same Neural Posteriors ⇨ Correct Coverage



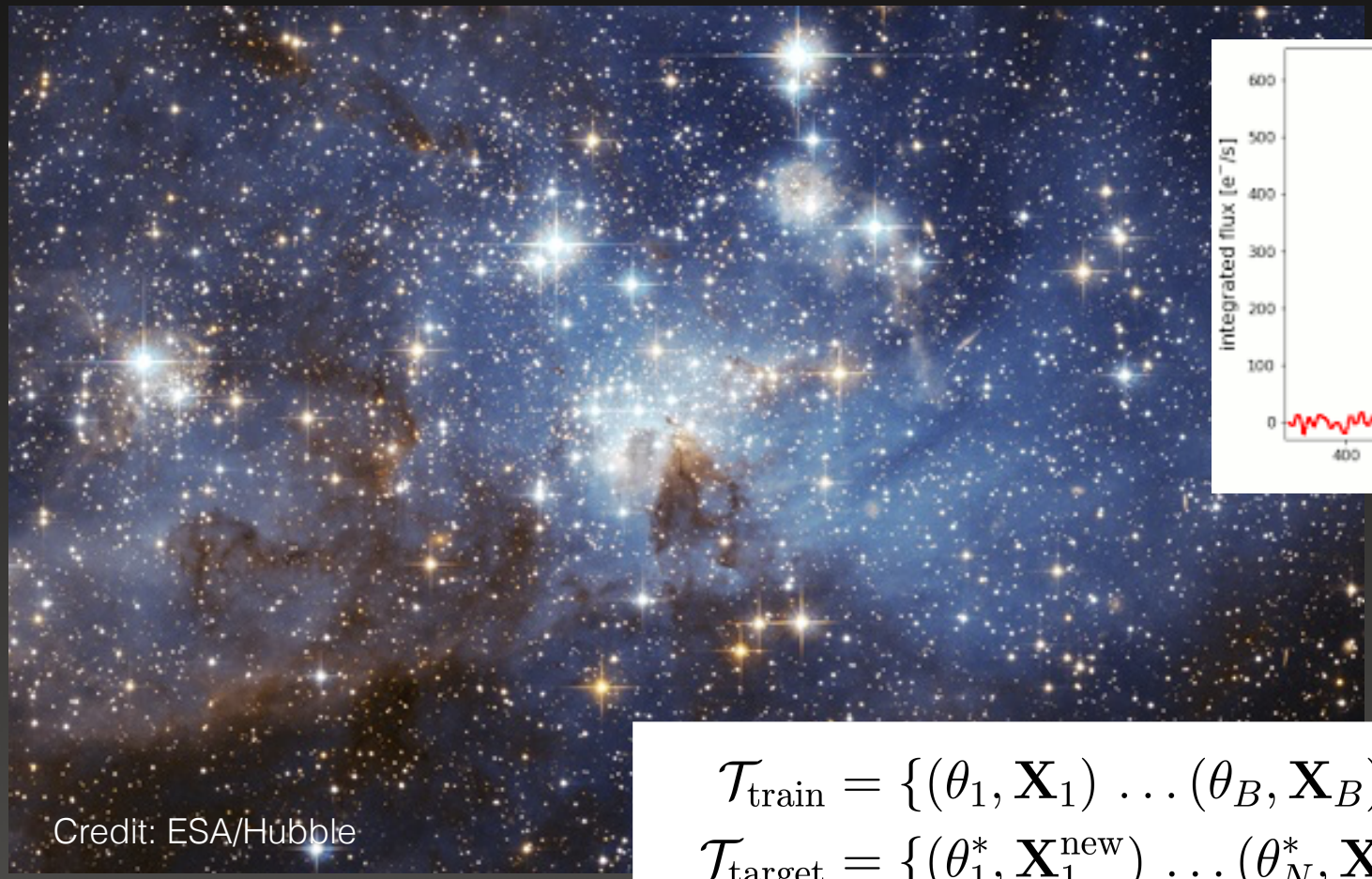
Waldo guarantees coverage everywhere, even if the prior poorly specified. Well-specified prior ⇨ power (tighter constraints)

$$\tau^{\text{WALDO}}(\mathcal{D};\boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

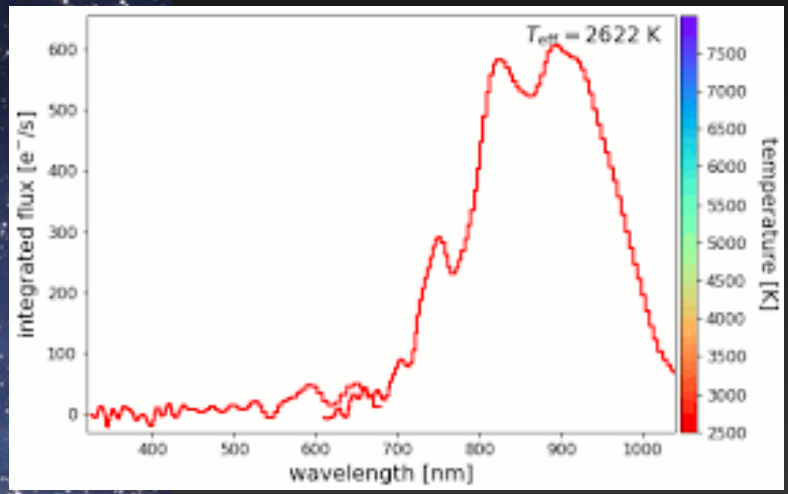
New Direction: “LF2I beyond SBI”

Project with Antonio Riberio and Professor Josh Speagle@UoT

Infer Properties of Stars in the Milky Way from Low-Resolution Spectra Using Cross-Matched Catalogs



Credit: ESA/Hubble

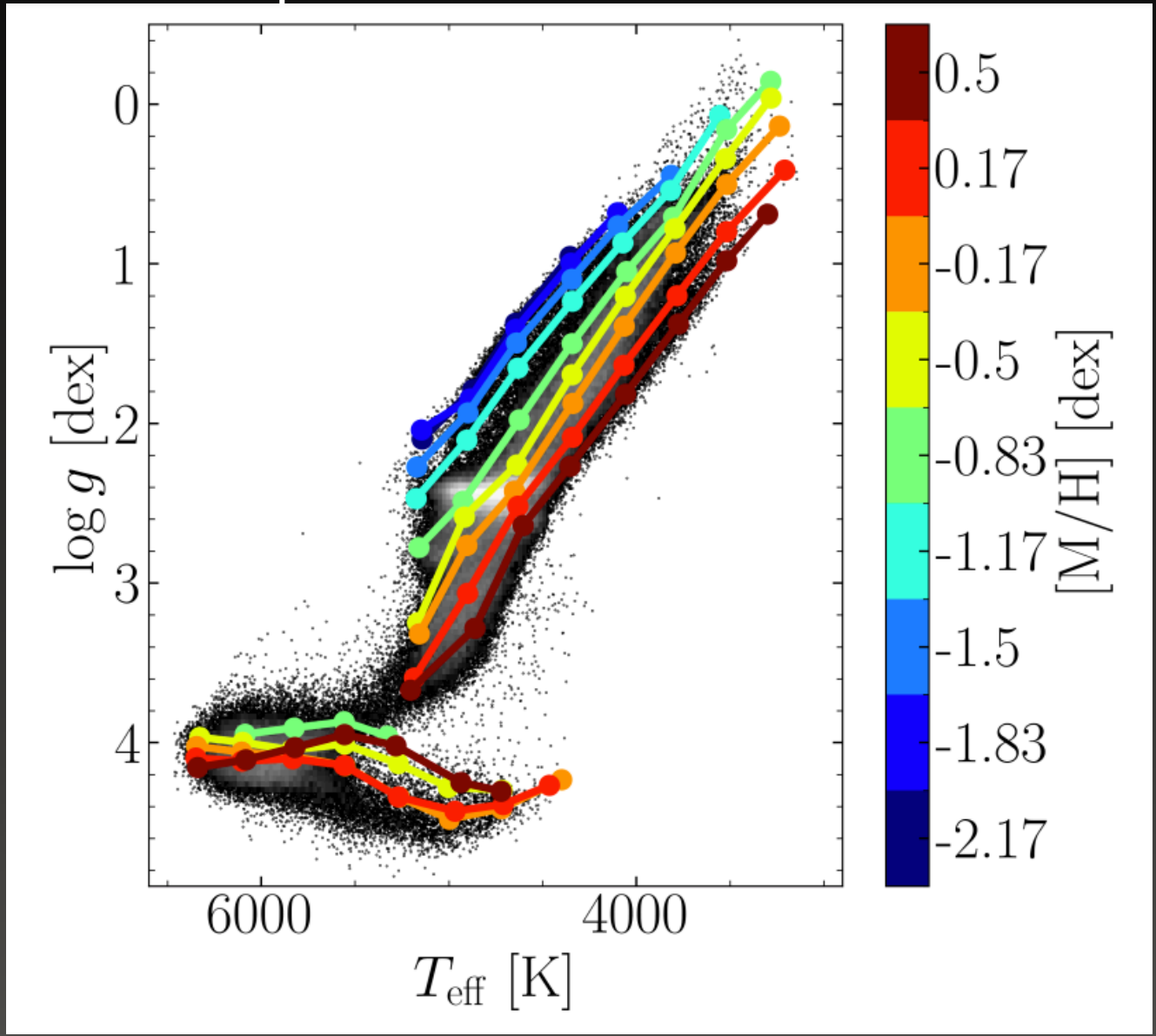


Credit: ESA/Gaia

$$\mathcal{T}_{\text{train}} = \{(\theta_1, \mathbf{X}_1) \ldots (\theta_B, \mathbf{X}_B)\} \sim \pi(\theta)\mathcal{L}(\mathbf{x};\theta)$$

$$\mathcal{T}_{\text{target}} = \{(\theta_1^*, \mathbf{X}_1^{\text{new}}) \ldots (\theta_N^*, \mathbf{X}_N^{\text{new}})\} \sim p_{\text{target}}(\theta)\mathcal{L}(\mathbf{x};\theta)$$

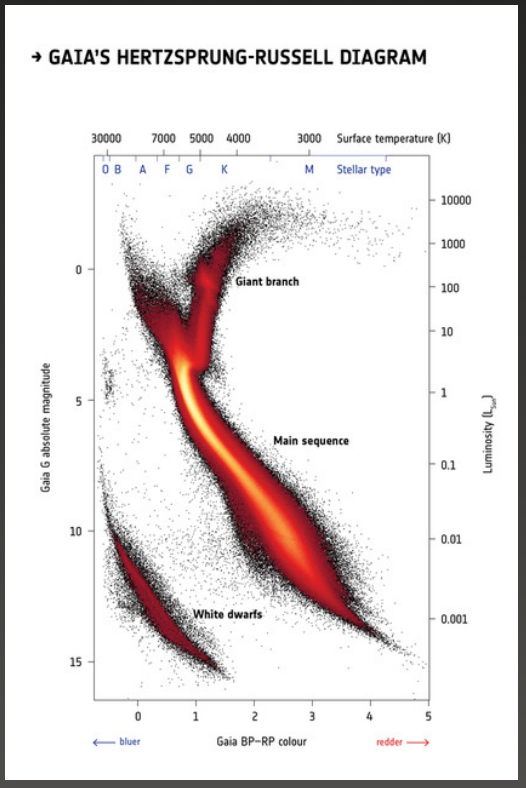
Kiehl diagram illustrating stellar evolution for ~500K stars plotted with Gaia satellite data



Laroche and Speagle (2024)

Gaia XP Spectra With Apogee Labels

- Observables X = low-resolution spectra from GAIA satellite
- Stellar labels θ (3: gravitational constant, metallicity [Fe/H], effective surface temperature) from cross-matching Gaia/Apogee catalogs. A subset (~200 K) of full XP catalog have "good labels" [Laroche & Speagle 2024]



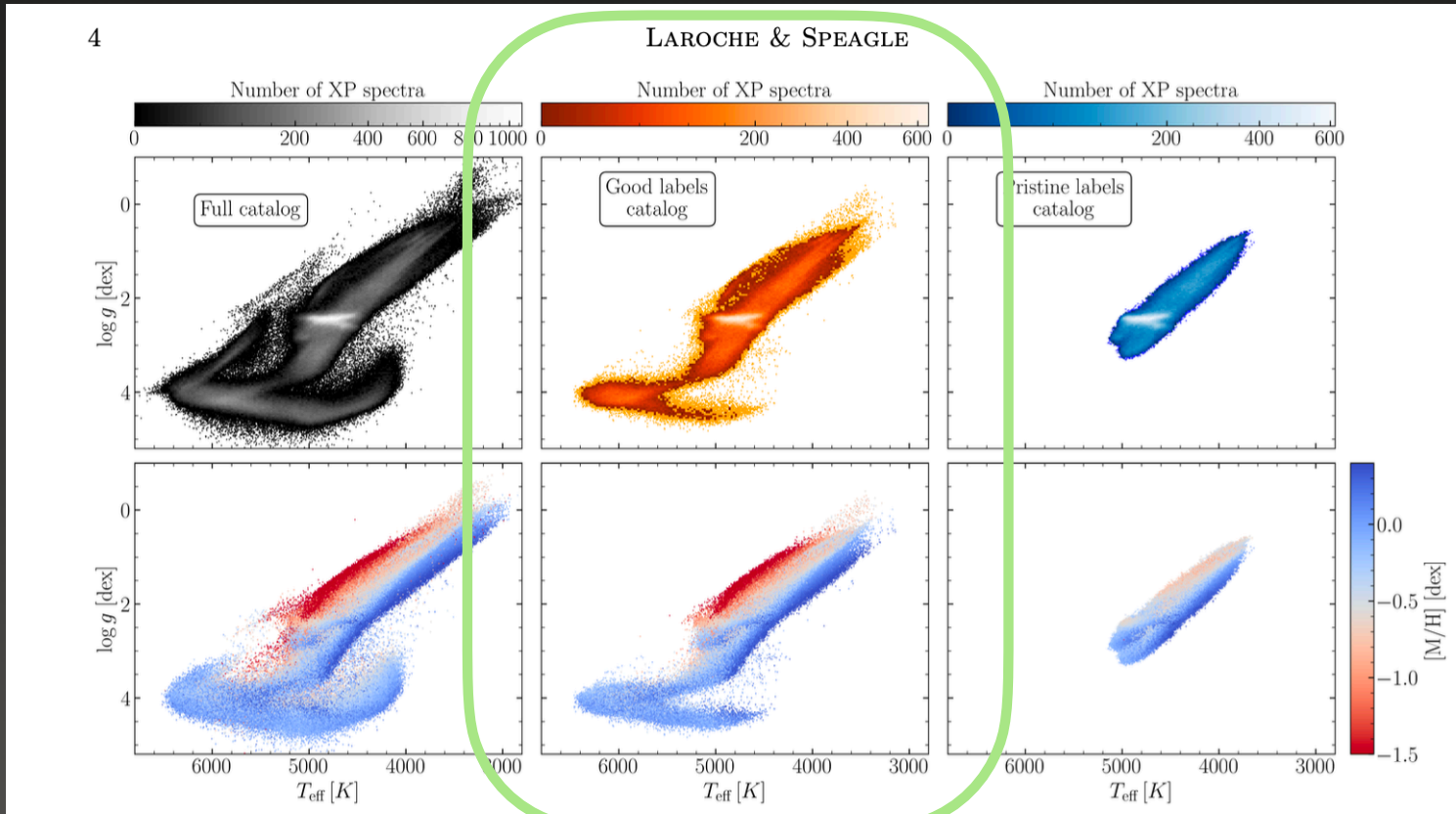
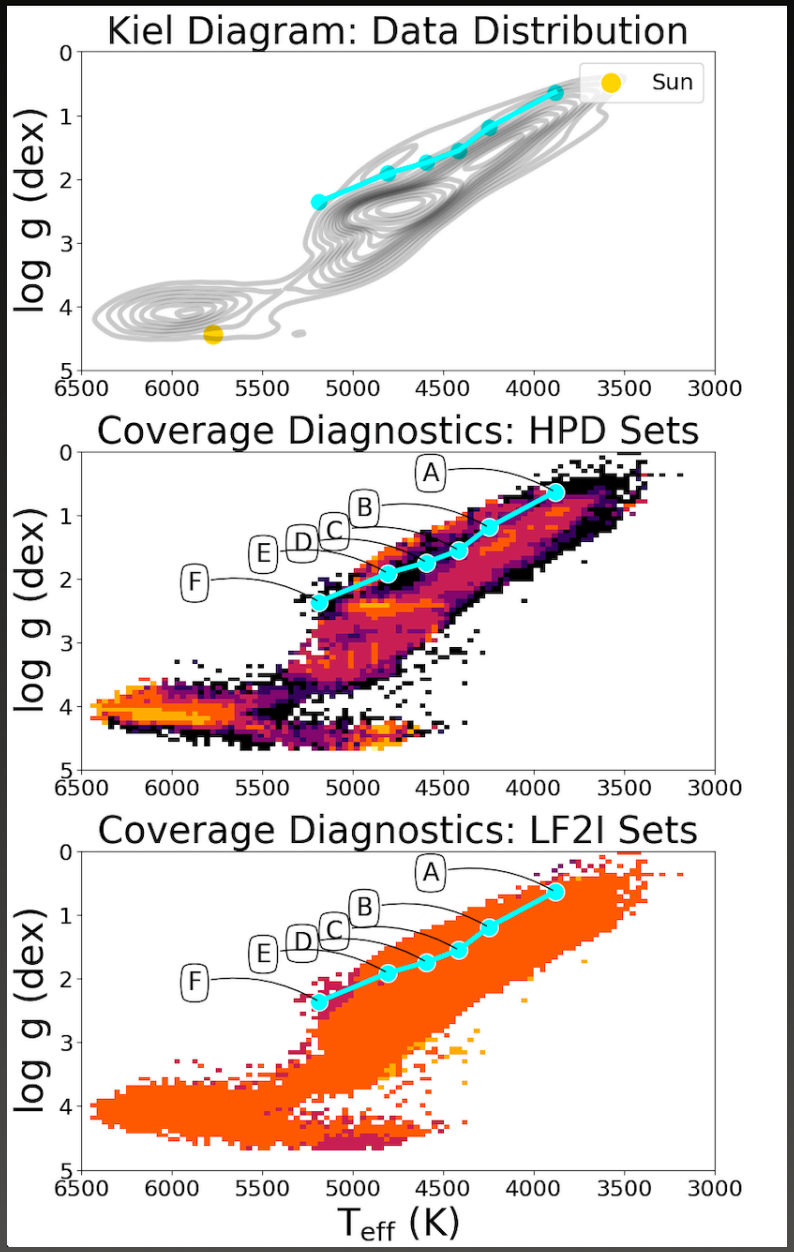
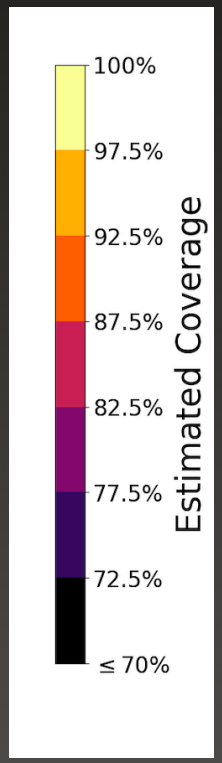


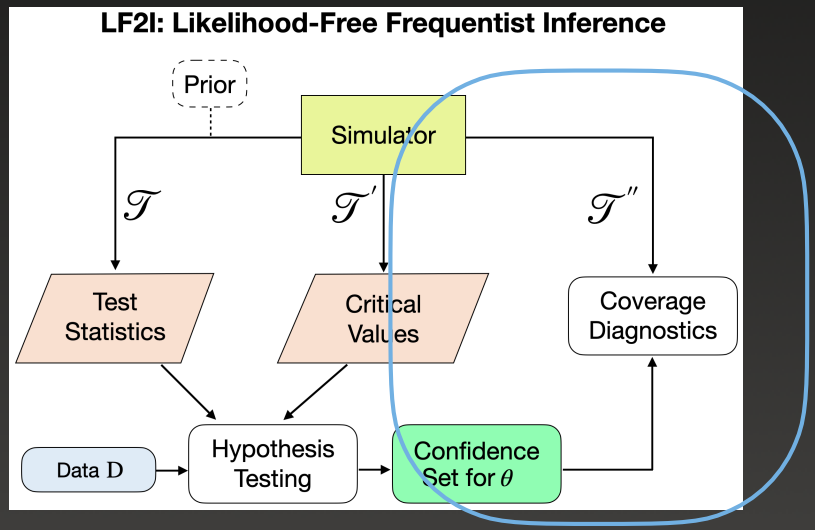
Figure 2. Kiel diagrams for the full (left), good labels (middle) and pristine labels (left) catalogs, colored by XP spectra number density (top row) and metallicity (bottom row). The increasingly restrictive quality cuts for the good labels and pristine labels

Preliminary Results (Red Giant Sequence, [Fe/H]=-1)



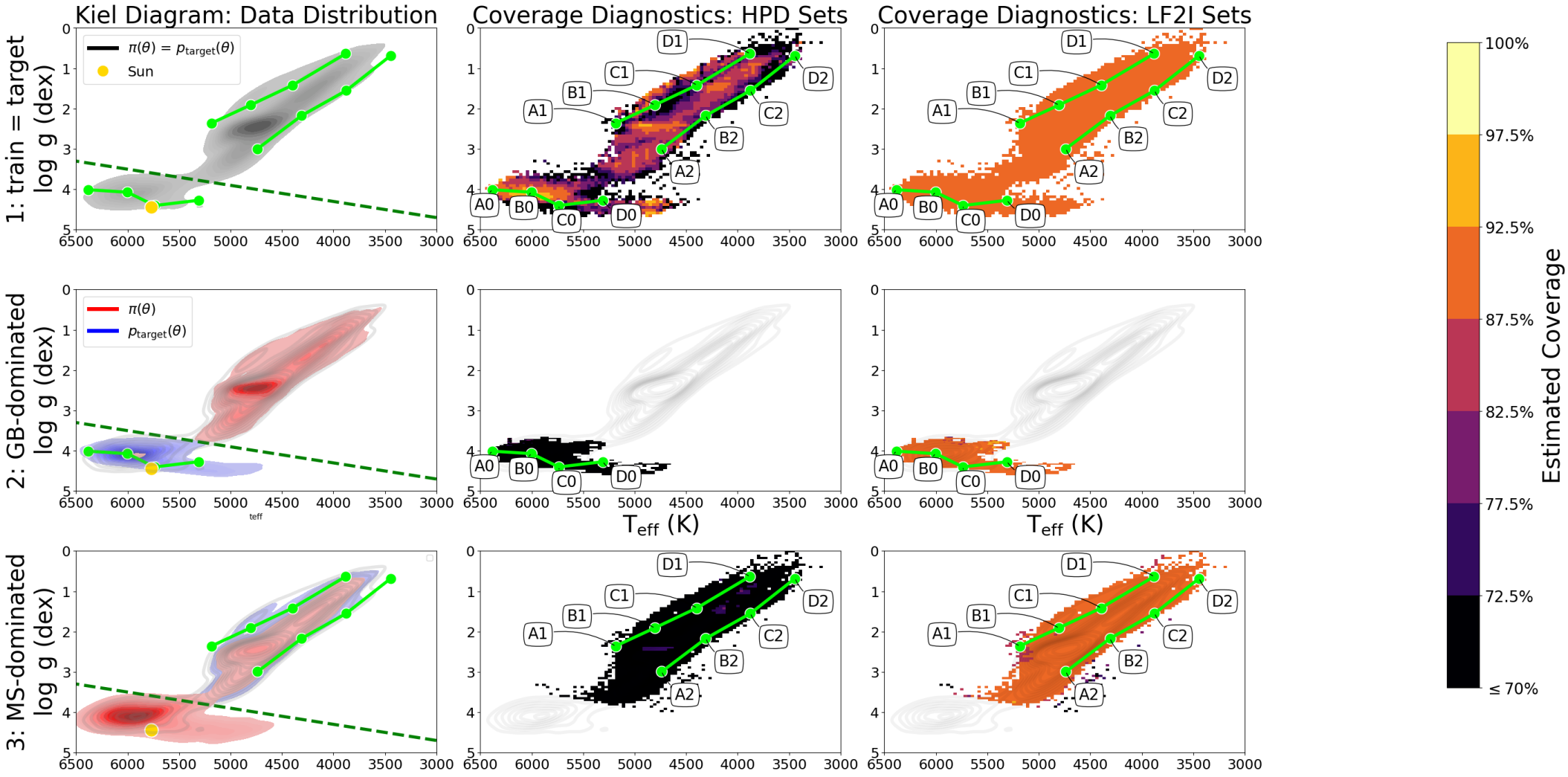


- Even w/o prior prob shift, credible regions (HPD sets) have frequentist coverage on average but not at fixed θ (typically under-cover where $\pi(\theta)$ is low)

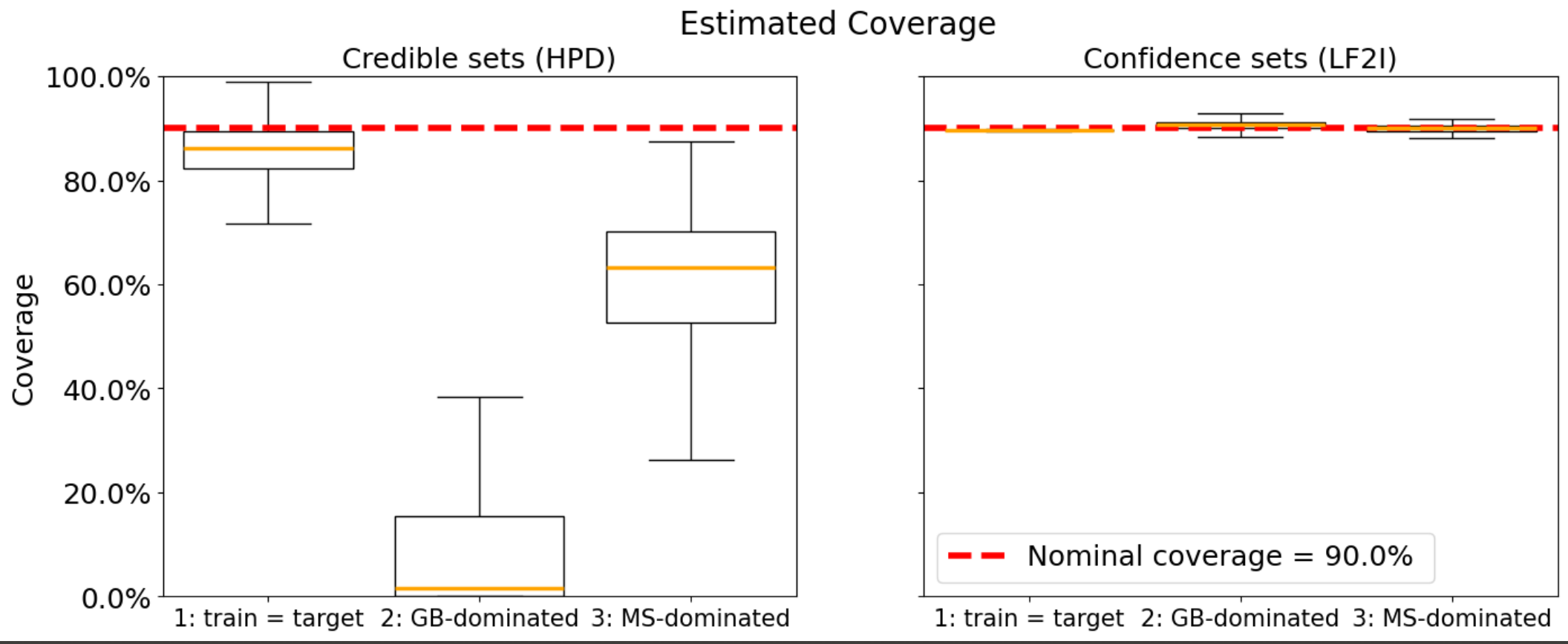


$$\mathcal{T}_{\text{train}} = \{(\theta_1, \mathbf{X}_1) \ldots (\theta_B, \mathbf{X}_B)\} \sim \pi(\theta)\mathcal{L}(\mathbf{x};\theta)$$
$$\mathcal{T}_{\text{target}} = \{(\theta_1^*, \mathbf{X}_1^{\text{new}}) \ldots (\theta_N^*, \mathbf{X}_N^{\text{new}})\} \sim p_{\text{target}}(\theta)\mathcal{L}(\mathbf{x};\theta)$$

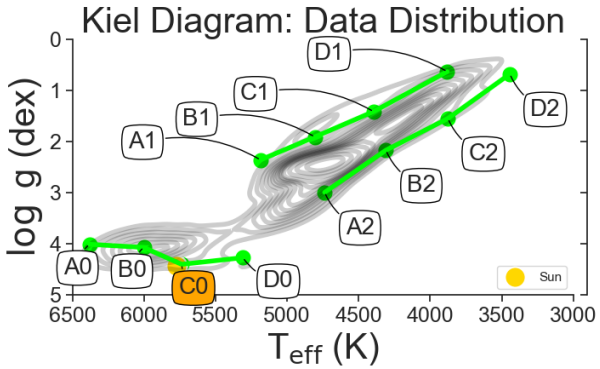
What if the Train/Target Distributions Differ?

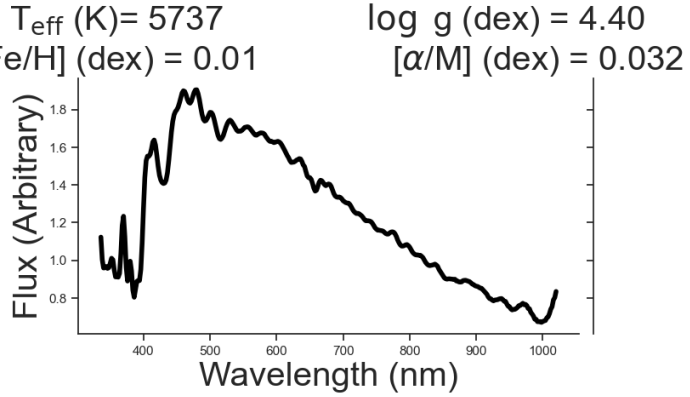


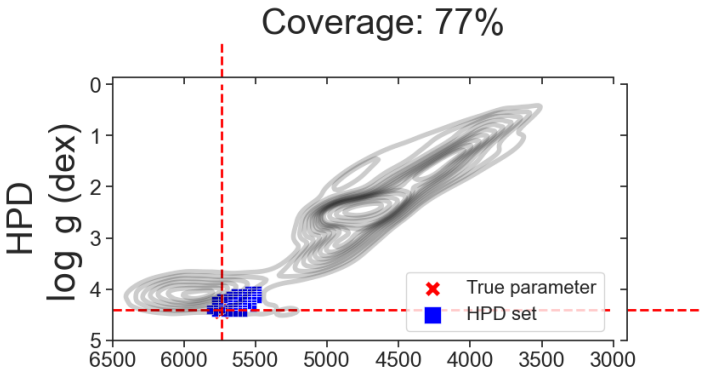
- Setting 1: train/test distributions the same
- Setting 2: GB-dominated (train on 98% GB) —- inference on MS stars
- Setting 3: MS-dominated (train on 80% MS) —- inference on GB stars

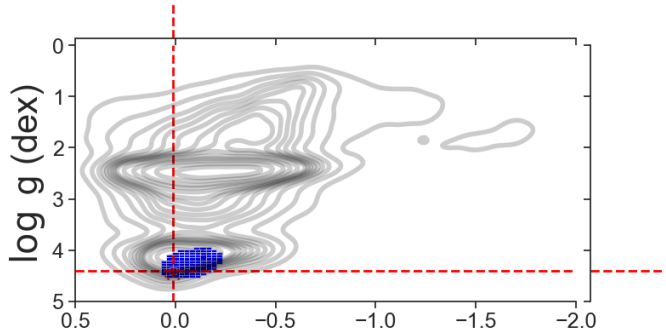

Estimated Coverage
Credible sets (HPD)
Confidence sets (LF2I)
Coverage
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
1: train = target
2: GB-dominated
3: MS-dominated
Nominal coverage = 90.0%

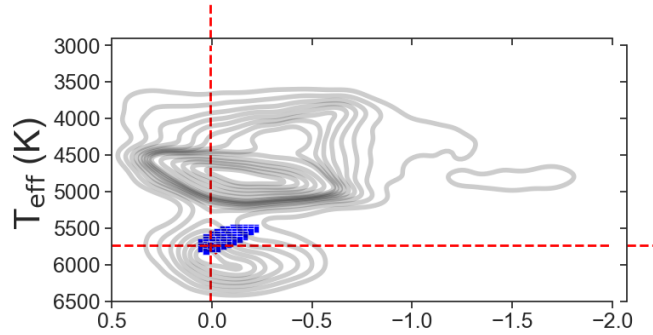
Inference for a Sun-like star (no prior prob shift)

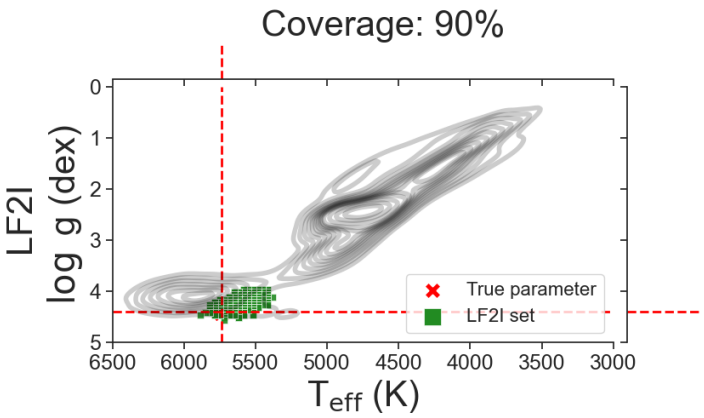


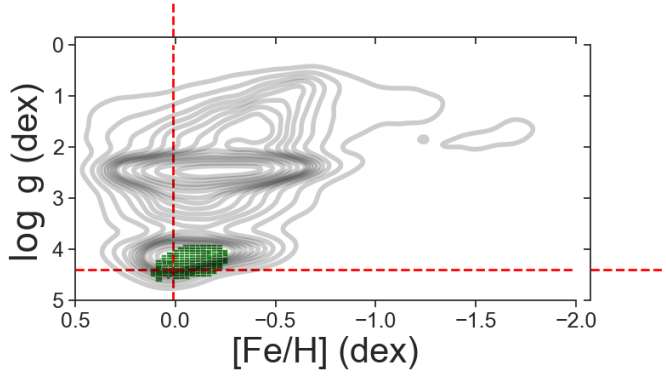


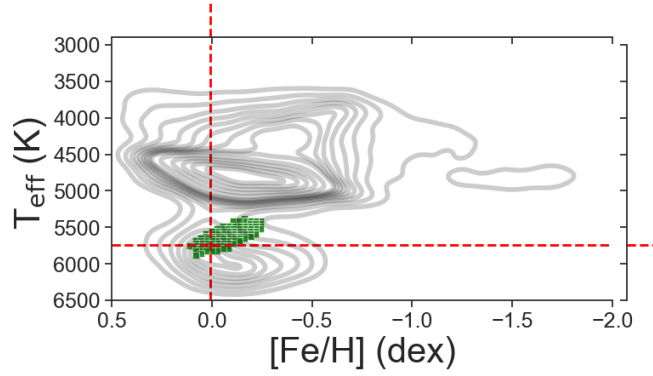




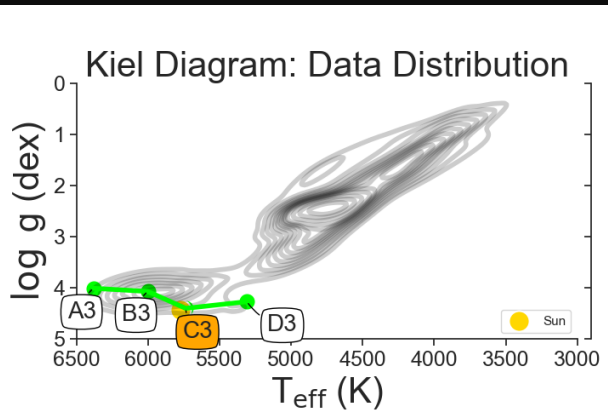


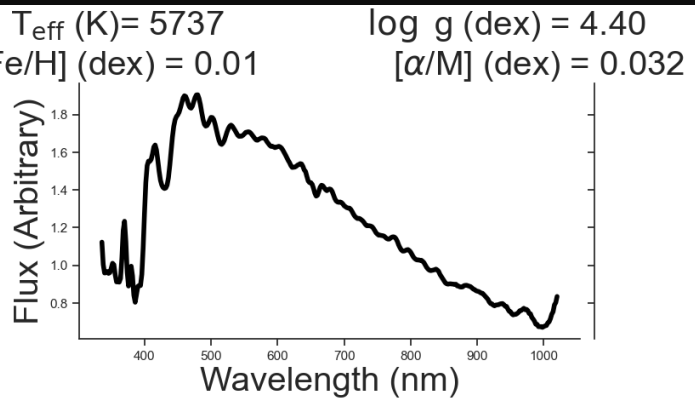


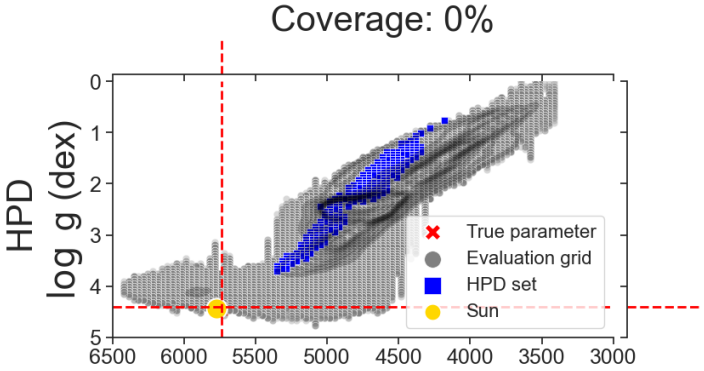


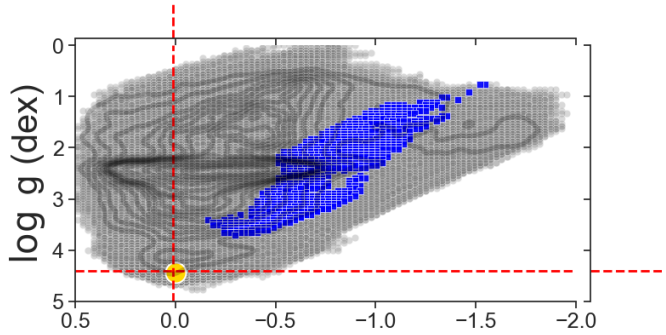


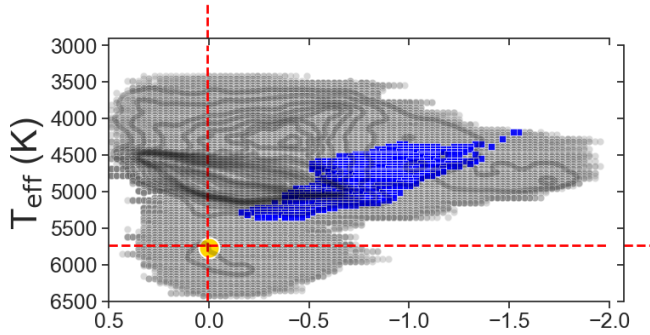
Inference for a Sun-like star (GB-dominated prior)

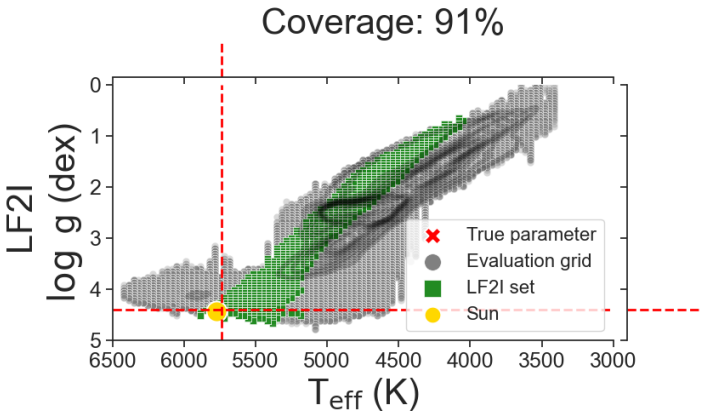


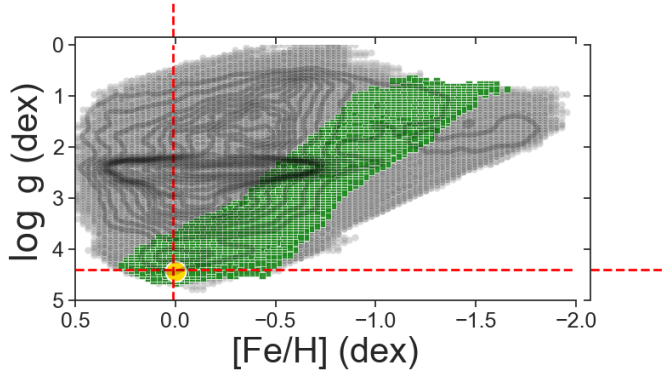


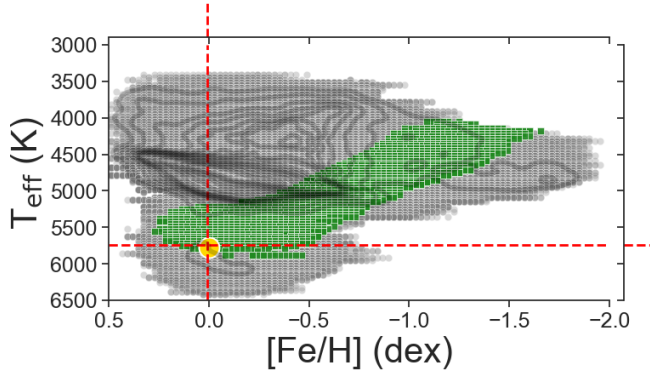










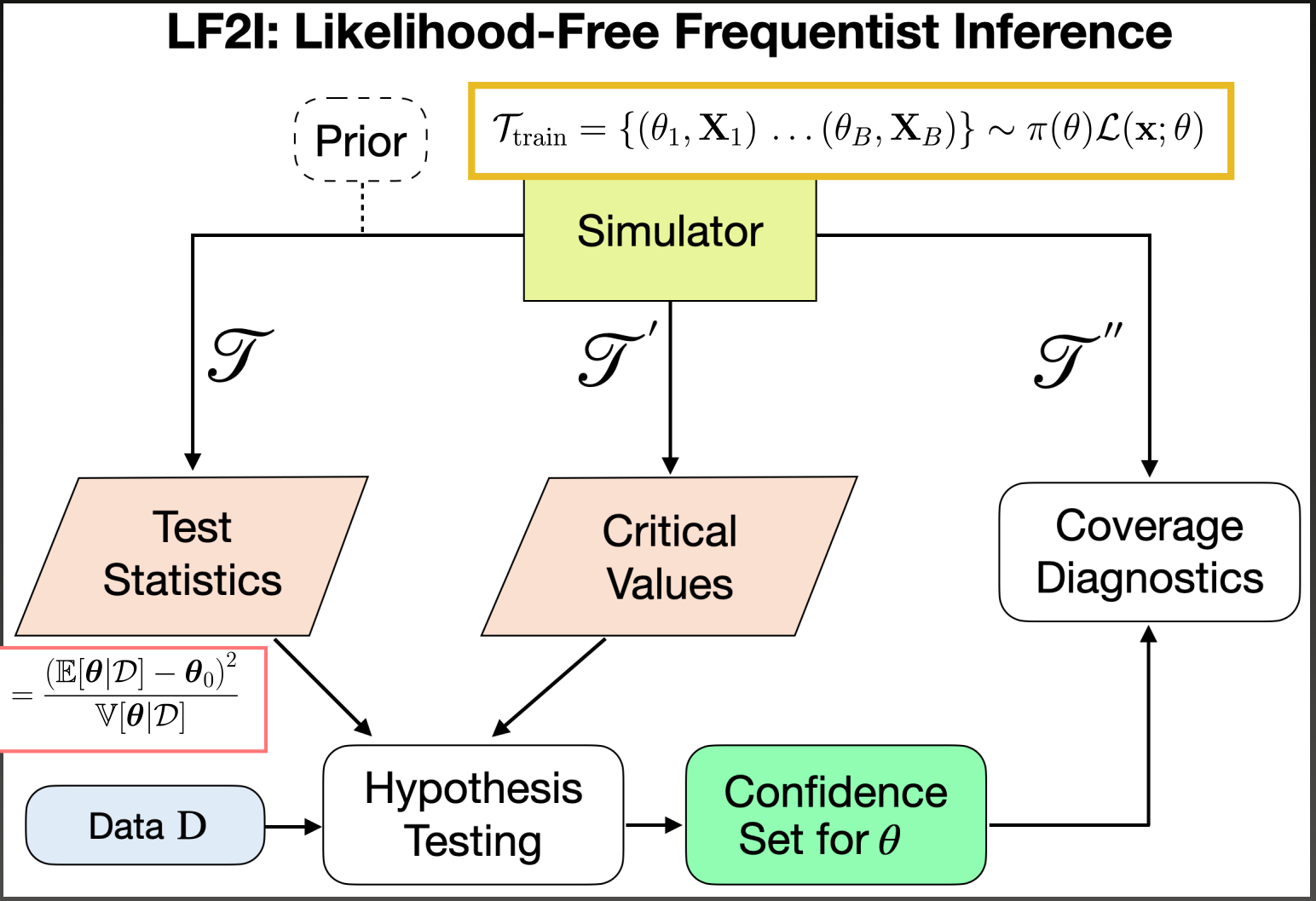


Take-Away: LF2I (inverse problem)

- Validity and diagnostics: Credible regions from NDEs do not necessarily reflect where the true parameter is
  - with LF2I we can “frequentize posteriors” to construct locally valid confidence sets for finite number of observations (n=1), and run diagnostics
- Prior Independence: nominal coverage regardless of prior (well-specified prior => power)

- LF2I is a fully modular and amortized framework. Compatible with any test statistic (likelihoods, LRs, posteriors, etc)

https://github.com/lee-group-cmu/lf2i



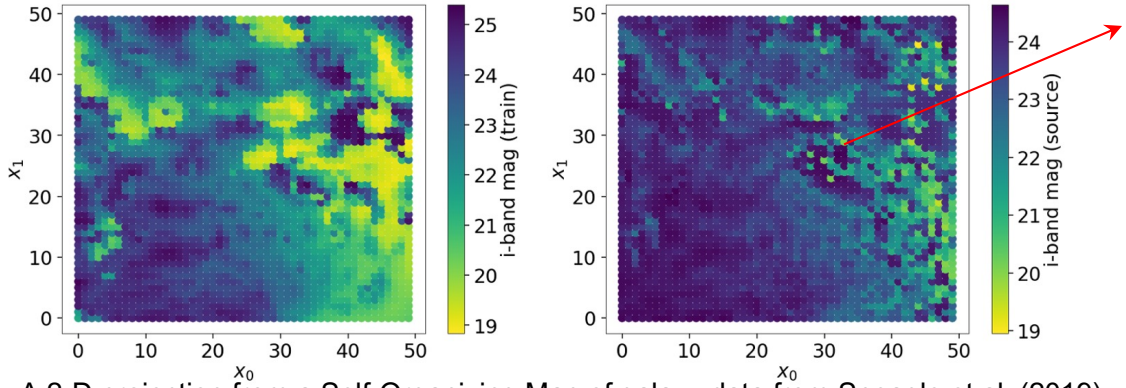
Take-Away: LF2I (inverse problem)

- Validity and diagnostics: Credible regions from NDEs do not necessarily reflect where the true parameter is
  - with LF2I we can “frequentize posteriors” to construct locally valid confidence sets for finite number of observations (n=1), and run diagnostics
- Prior Independence: nominal coverage regardless of prior (well-specified prior => power)

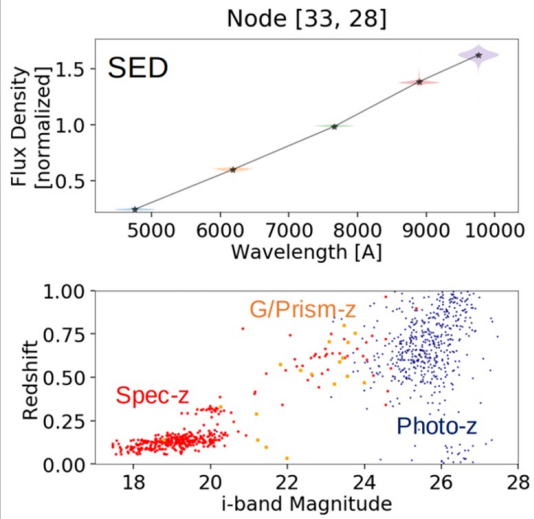
What's Next? Open Problems

Incorporating Uncertainties, Bias in ML Applications

- Many datasets have heteroscedastic uncertainties, missing/censored data, and biased subsets where ground truth labels are available (often with higher SNR/less censoring).
- Traditionally, dealing with this has involved "degrading" the training data to match the properties of the broader dataset.
- Can we account for these domain adaptations and perform robust uncertainty quantification (UQ) without doing this?



A 2-D projection from a Self-Organizing Map of galaxy data from Speagle et al. (2019) showing the mean brightness of sources in the training data (left) and target data (right). The training data is generally significantly biased to be much brighter than the target data.

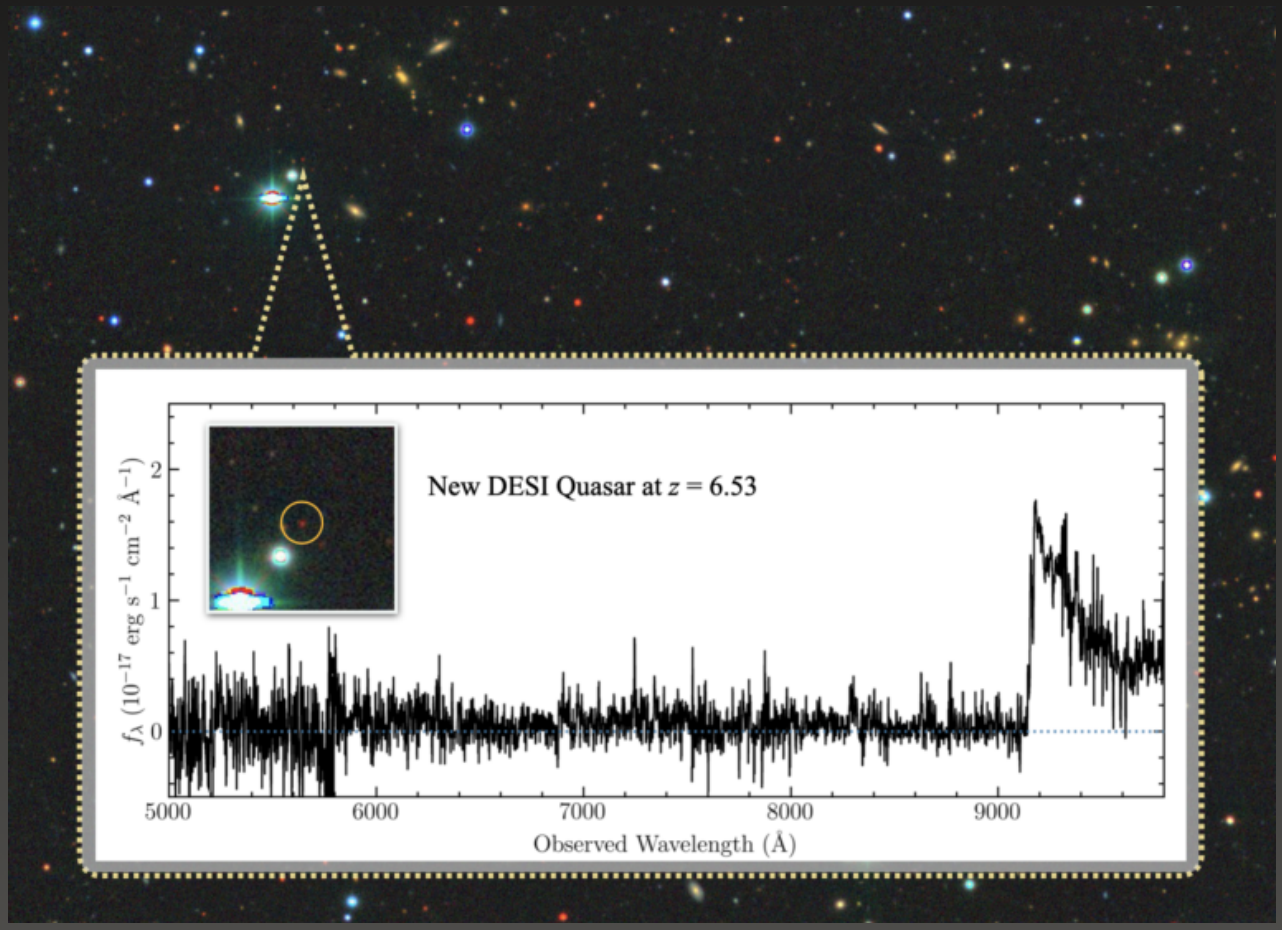


A "node" from the plot on the left showing the training data (top) along with the evolution of the ground-truth labels as a function of brightness (bottom). The dramatic evolution in labels as a function of brightness highlights a significant potential for biases in later applications.

Slide Credit: Josh Speagle and Ashley Villar

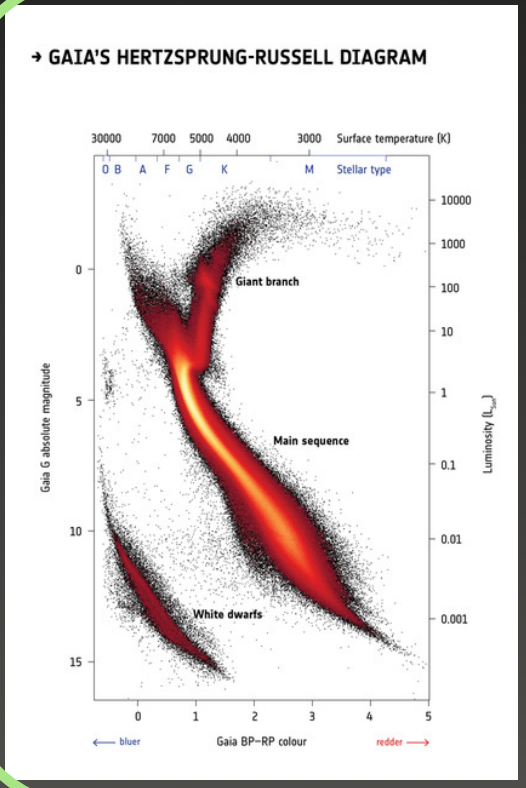
Work-In-Progress: Incorporating Measurement Errors

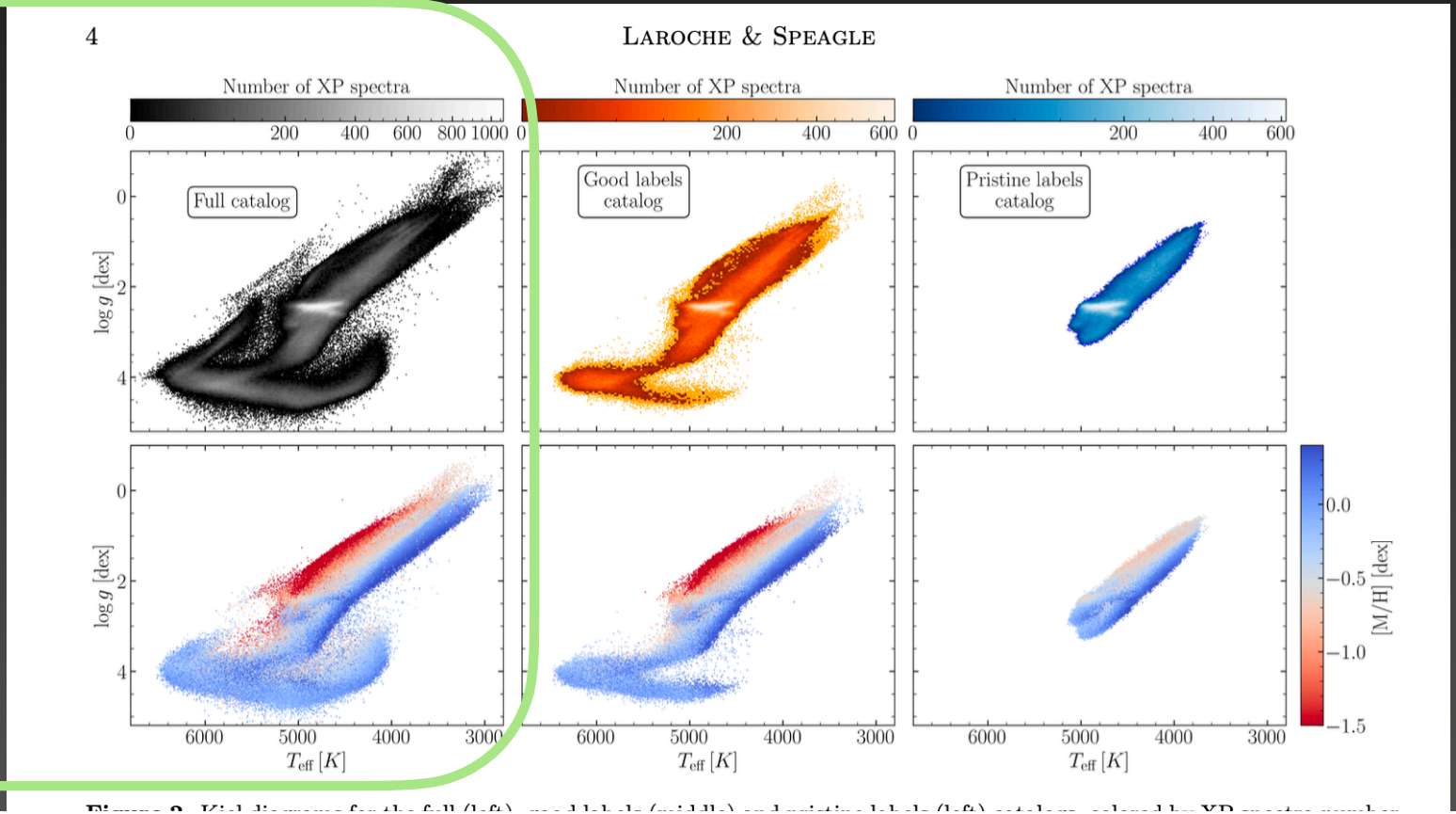
- Project 1 (SBI): Inference of galaxy star-formation histories from spectroscopic/photometric data (labels from simulations)



Work-In-Progress: Incorporating Measurement Errors

- Project 2: Inference of stellar labels from low-resolution spectra (partially labeled data from cross-matched catalogs).
  - We can handle selection bias, but need to develop methods to incorporate measurement errors from data from the GAIA satellite





Acknowledgments

- Nic Dalmasso (CMU alumni)

  original LF2I framework

- Rafael Izbicki (UFSCar)
- Luca Masserano

$$\tau^{\text{WALDO}}(\mathcal{D};\boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

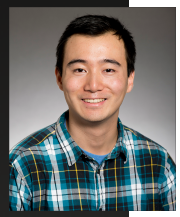
- Mikael Kuusela, Tommaso Dorigo (INFN/Padova)
- James Carzon, Antonio Carlos Herling, Alex Shen, Josh Speagle (U.Toronto)

This work is funded in part by NSF DMS-2053804

Carnegie Mellon University
Statistics & Data Science
Dietrich College of Humanities and Social Sciences

Extra Slides Start Here

Finally, if you are instead interested in calibrated PDs and posteriors (consistent with a chosen prior)…



Diagnostics for Conditional Density Models and Bayesian Inference Algorithms

UAI, PMLR (161) 2021

David Zhao[1] Niccolò Dalmasso[1] Rafael Izbicki[2] Ann B. Lee[1]

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]Department of Statistics, Federal University of São Carlos (UFSCar), S

Definition 1 (Global Consistency). *An estimate $\widehat{f}(y|\mathbf{x})$ is globally consistent with the density $f(y|\mathbf{x})$ if the following null hypothesis holds:*

$$H_0 : \widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x}) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } y \in \mathcal{Y}. \quad (1)$$

arXiv:2205.14568

CONDITIONALLY CALIBRATED PREDICTIVE DISTRIBUTIONS BY PROBABILITY-PROBABILITY MAP: APPLICATION TO GALAXY REDSHIFT ESTIMATION AND PROBABILISTIC FORECASTING

BY BIPRATEEP DEY[1,a], DAVID ZHAO[2,d], JEFFREY A. NEWMAN[1,b], BRETT H. ANDREWS[1,c], RAFAEL IZBICKI[3,e], AND ANN B. LEE[4,f]

[1]*Department of Physics and Astronomy and PITT-PACC, University of Pittsburgh,* [a]*biprateep@pitt.edu;* [b]*janewman@pitt.edu;* [c]*andrewsh@pitt.edu*

[2]*Department of Statistics and Data Sci*

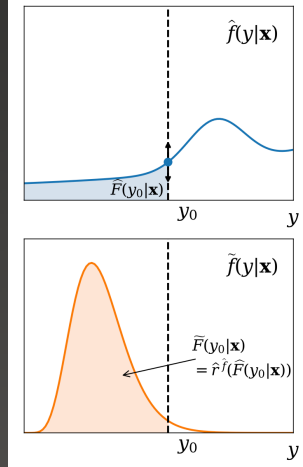
[3]*Department of Statistic*

[4]*Department of Statistics and Data S*

DEFINITION 3 (Recalibrated PD). *The recalibrated predictive distribution (PD) of Y given* $\mathbf{x}$ *is defined through a P-P map,*

$$\widetilde{F}(y|\mathbf{x}) := \widehat{r}^{\widehat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right), \quad (6)$$

where $\widehat{r}^{\widehat{f}}$ *is the regression estimator of the PIT-CDF (Equation 4).*



(b) `Cal-PIT` by Mapping Probabilities

Taxonomy of Different Types of Simulators

Image credit: Kyle Cranmer

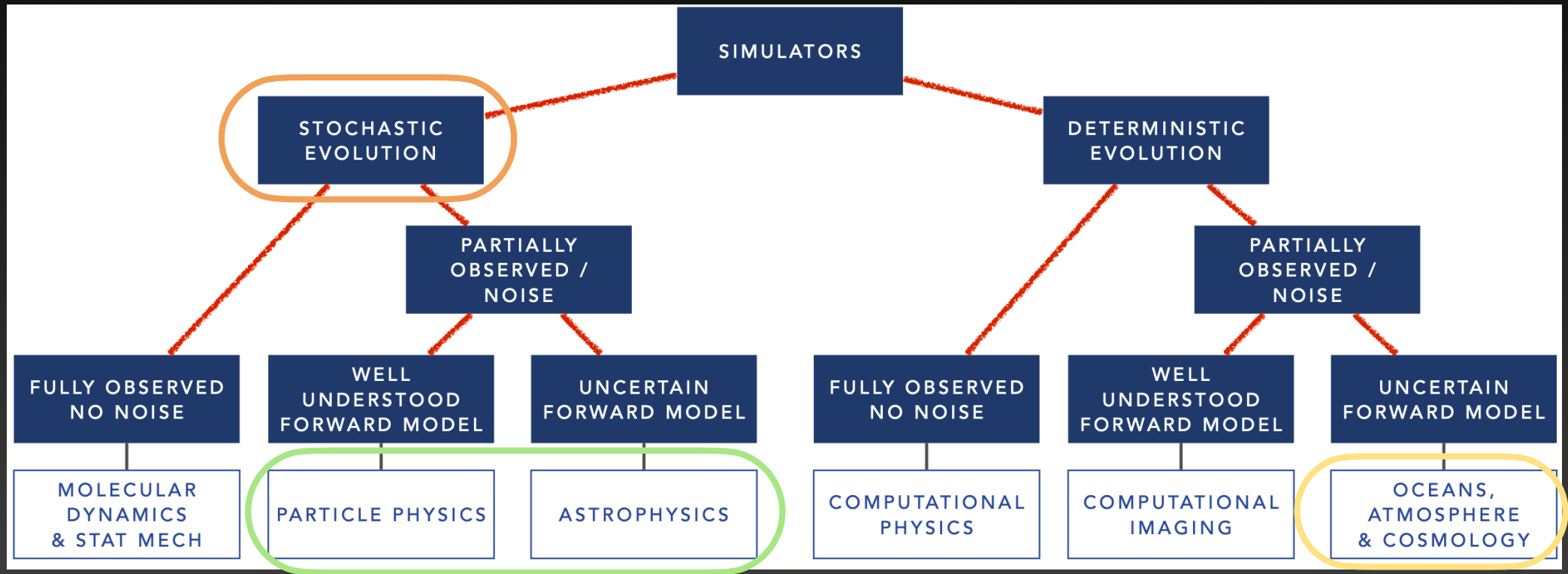
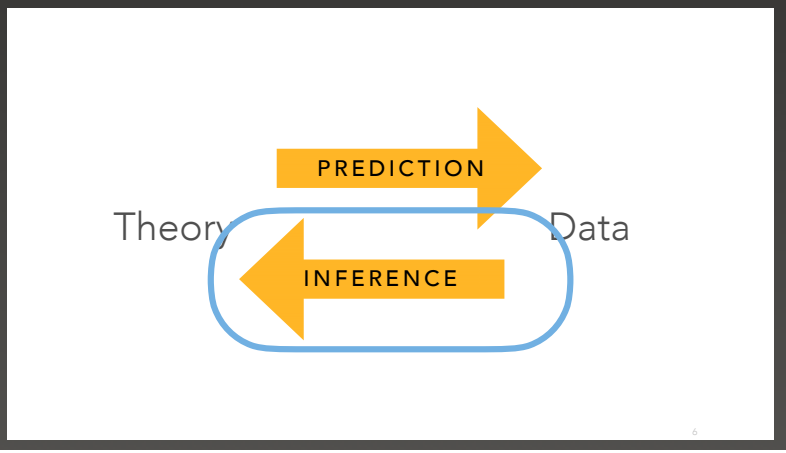


Figure credit: Kyle Cranmer

Simulation/Prediction to Scientific Inference

- Revolution in simulators and AI generative models (GANs, transformers, diffusion models etc) & high-performance predictive NNs.
- But what about scientific inference?
  - Simulators are often poorly suited for the "inverse problem" of inferring the causes behind observed phenomena.



Recent Developments in LFI are Mainly Driven by ML*

- Leverage ML algorithms to directly **estimate key inferential quantities** from simulated data

$$\{(\theta_1, \mathscr{D}_1), \ldots, (\theta_B, \mathscr{D}_B)\}, \text{ where } \theta \sim \pi_\theta(\cdot),\ \mathscr{D} \mid \theta \sim \mathscr{L}(\theta; \cdot)$$

  - **Parameters** θ (inference via point predictions) [e.g., Kieseler et al. (2022); Ho et al. (2019); Gerber and Nychka (2021)]
  - **Posteriors** $p(\theta \mid D)$ [e.g., Papamakarios et al. (2016); Lueckmann et al. (2016); Izbicki et al. (2019); Greenberg et al. (2019); Corso et al. (2023)]
  - **Likelihoods** $p(D \mid \theta)$ [e.g., Izbicki et al. (2014); Brehmer et al. (2020); Walchessen et al. (2023)]
  - **Likelihoods ratios** $p(D \mid \theta_1)/p(D \mid \theta_2)$ [e.g., Cranmer et al. (2015); Thomas et al. (2022); Hermans et al. (2020); Durkan et al. (2020); Brehmer et al. (2020)]

* Review paper on SBI by Cranmer, Brehmer, Louppe; PNAS 2019

Recent Developments in LFI are Mainly Driven by ML*

- Leverage ML algorithms to directly **estimate key inferential quantities** from simulated data

$$\{(\theta_1, \mathscr{D}_1), \ldots, (\theta_B, \mathscr{D}_B)\}, \text{ where } \theta \sim \pi_\theta(\cdot),\ \mathscr{D} \mid \theta \sim \mathscr{L}(\theta; \cdot)$$

  - **Parameters** θ (inference via point predictions) [e.g., Kieseler et al. (2022); Ho et al. (2019); Gerber and Nychka (2021)]
  - **Posteriors** $p(\theta \mid D)$ [e.g., Papamakarios et al. (2016); Lueckmann et al. (2016); Izbicki et al. (2019); Greenberg et al. (2019); Corso et al. (2023)]
  - **Likelihoods** $p(D \mid \theta)$ [e.g., Izbicki et al. (2014); Brehmer et al. (2020); Walchessen et al. (2023)]
  - **Likelihoods ratios** $p(D \mid \theta_1)/p(D \mid \theta_2)$ [e.g., Cranmer et al. (2015); Thomas et al. (2022); Hermans et al. (2020); Durkan et al. (2020); Brehmer et al. (2020)]

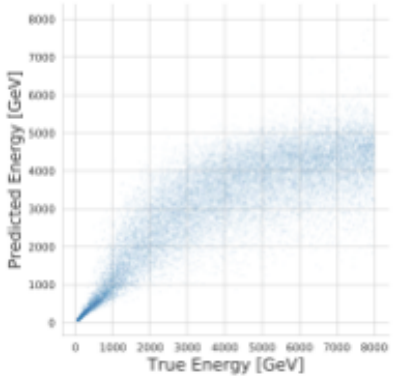
- These approaches can **handle complex, unstructured and high-dimensional data** and **approximate complicated distributions** without prior dimension reduction
- Some of them also provide **amortized inference** (train once, evaluate on many observations)

* Review paper on SBI by Cranmer, Brehmer, Louppe; PNAS 2019

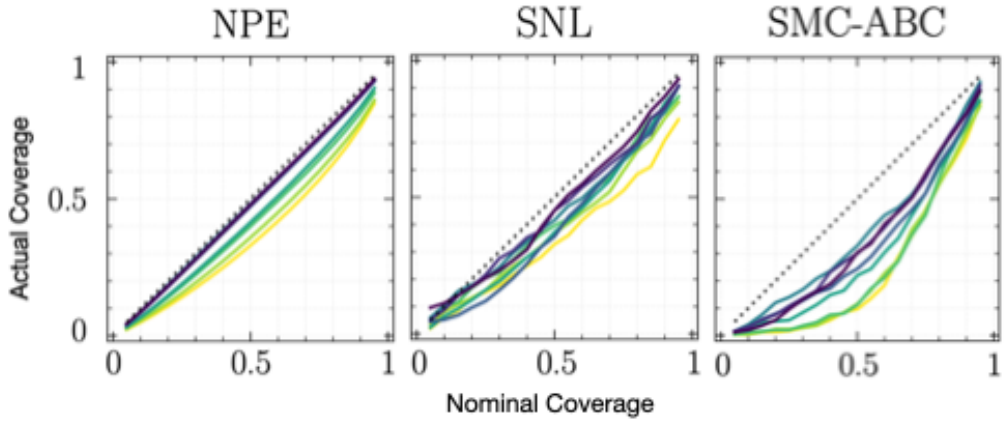
What is Missing? → Trustworthy UQ for Inverse Problems

- Do ML methods provide **reliable constraints** for internal parameters of interest?

Prediction algorithms[1]



Posterior estimators[2]

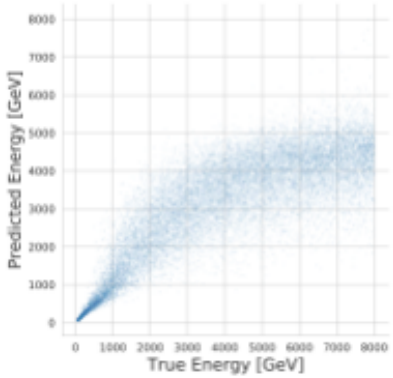


1. E.g., Kieseler et al. (2022) 2. Hermans et al. (2021)

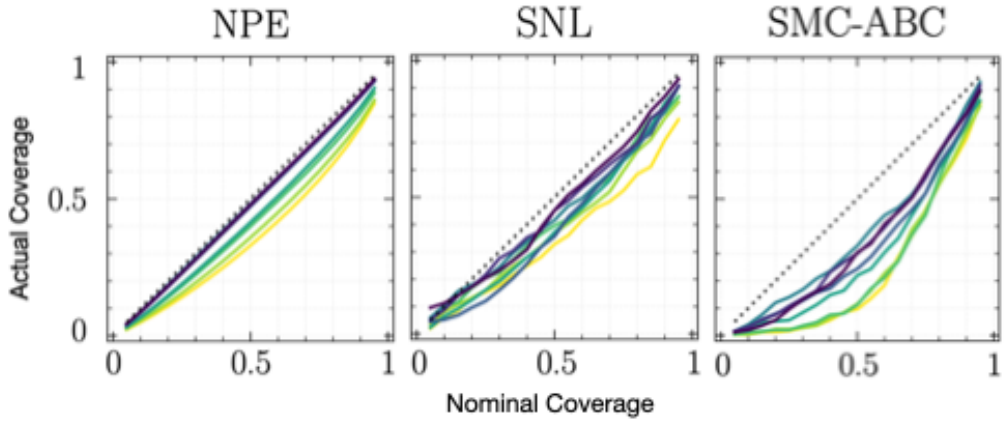
What is Missing? → Trustworthy UQ for Inverse Problems

- Do ML methods provide **reliable constraints** for internal parameters of interest?

Prediction algorithms[1]



Posterior estimators[2]

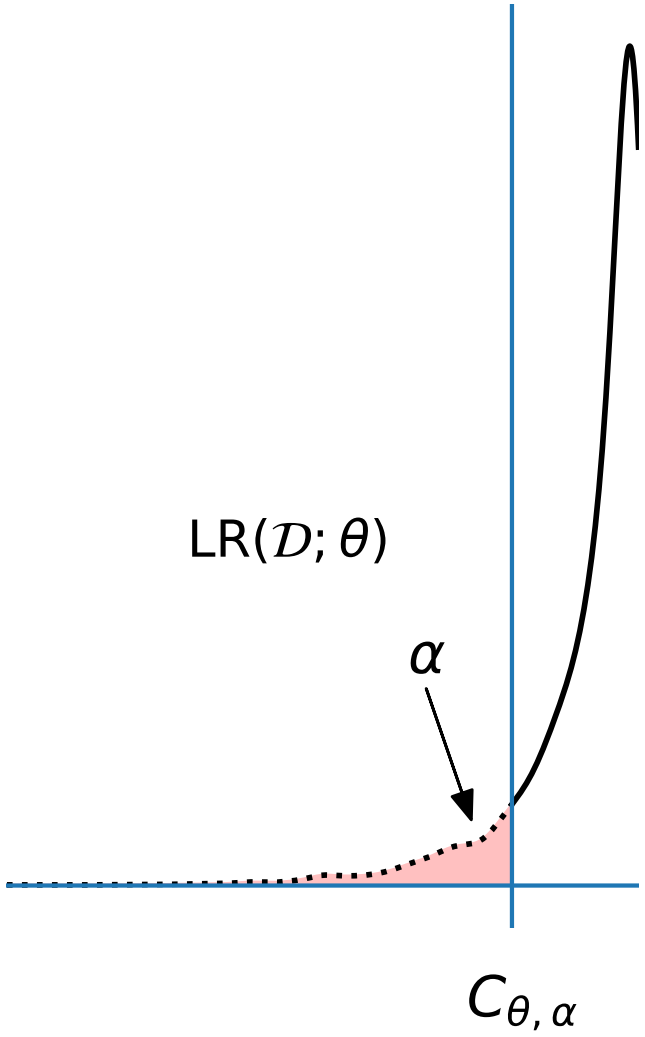


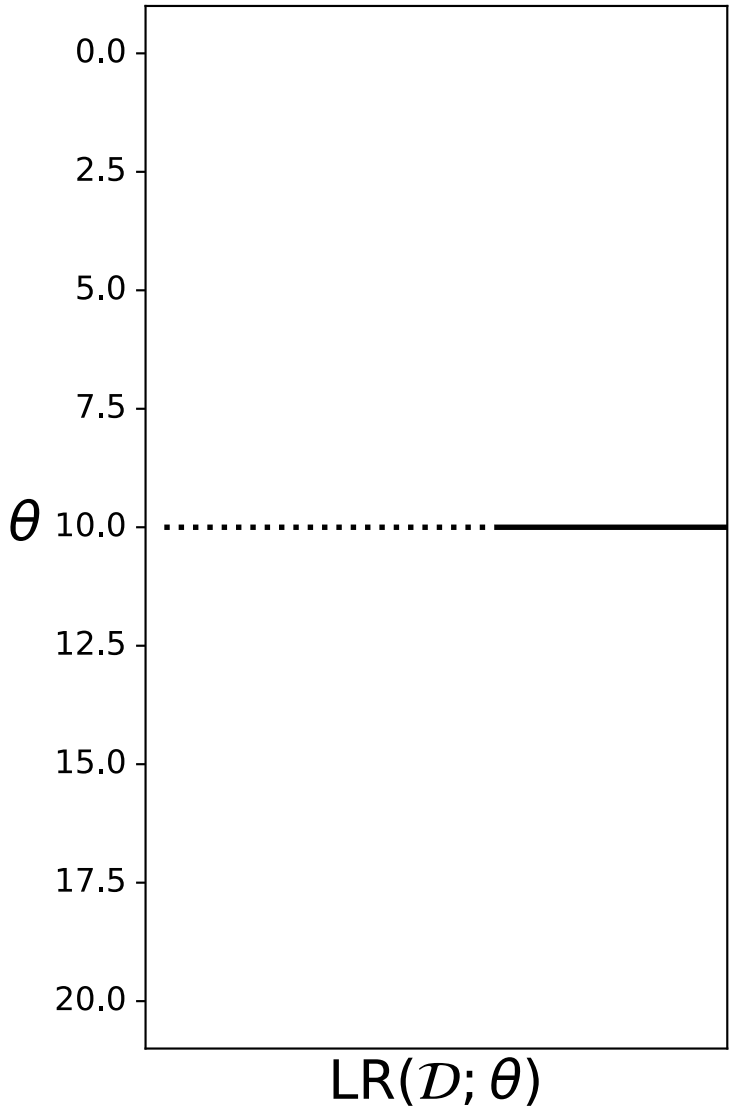
Problems:

- **ML training** vs. **Scientific Evaluation criteria:** should target trustworthy uncertainty quantification, not exactness of an approximation
- Most ML methods target **prediction**, rather than **inference** in $\theta \mapsto \mathscr{D}$ problems
- Training data sampled from prior $\theta \sim \pi_\theta$ → **possibly harmful bias** if not consistent with D_{obs}

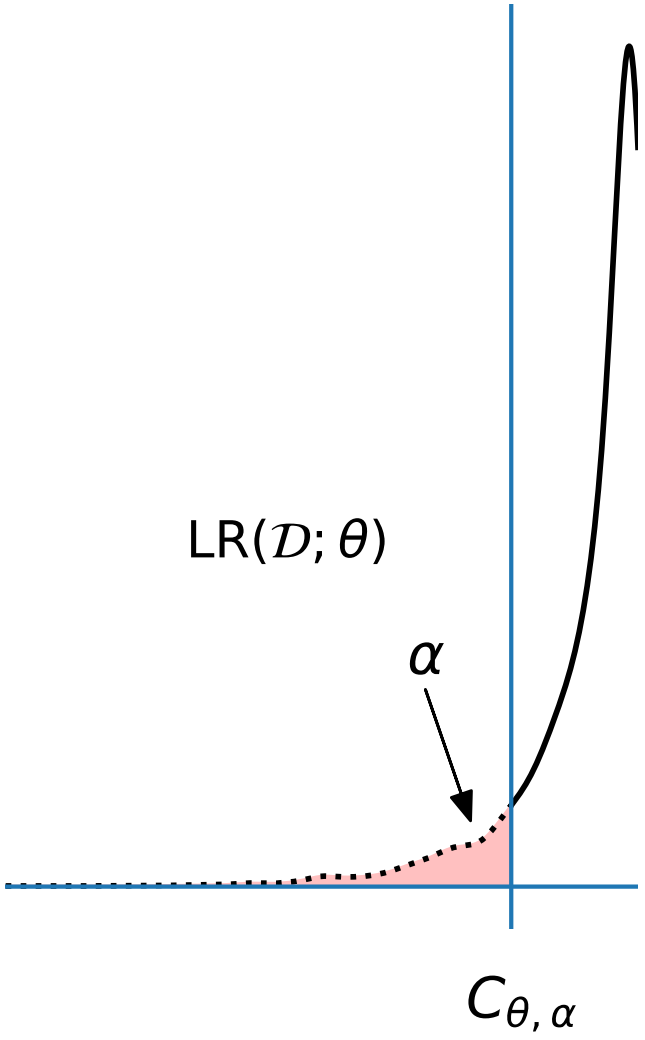
1. E.g., Kieseler et al. (2022) 2. Hermans et al. (2021)

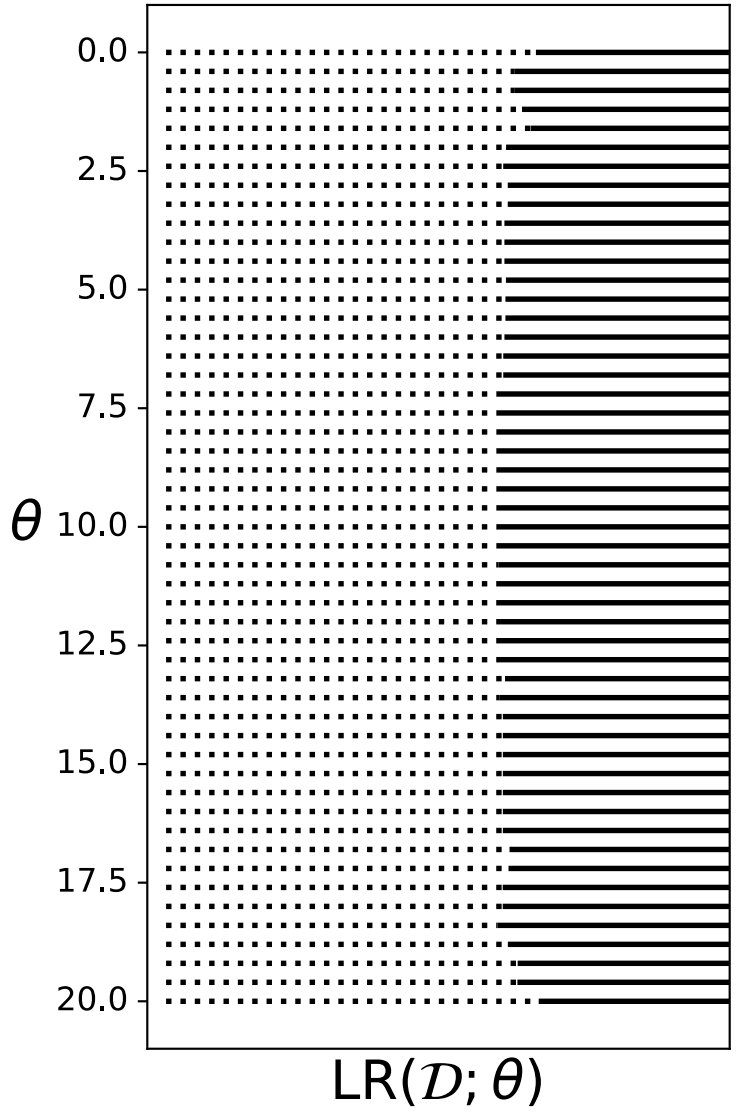
1. Fixed θ. Find the rejection region for test statistic λ.



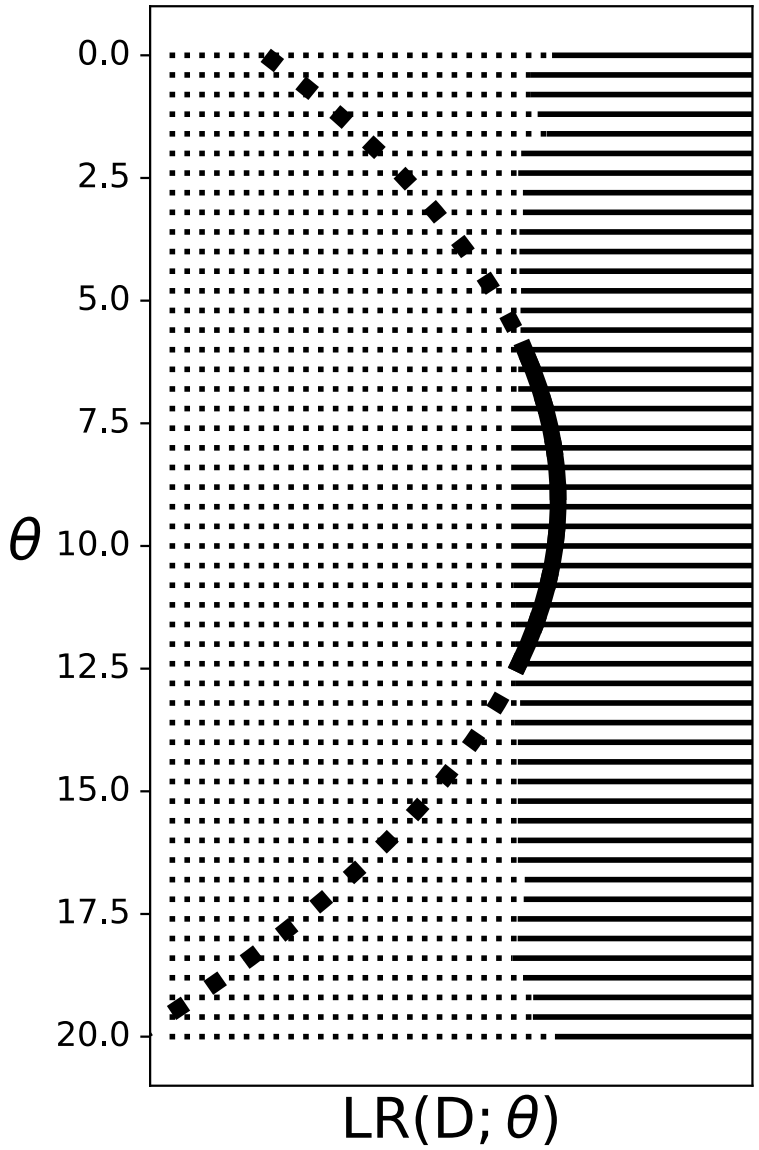


2. Repeat for every θ in parameter space.

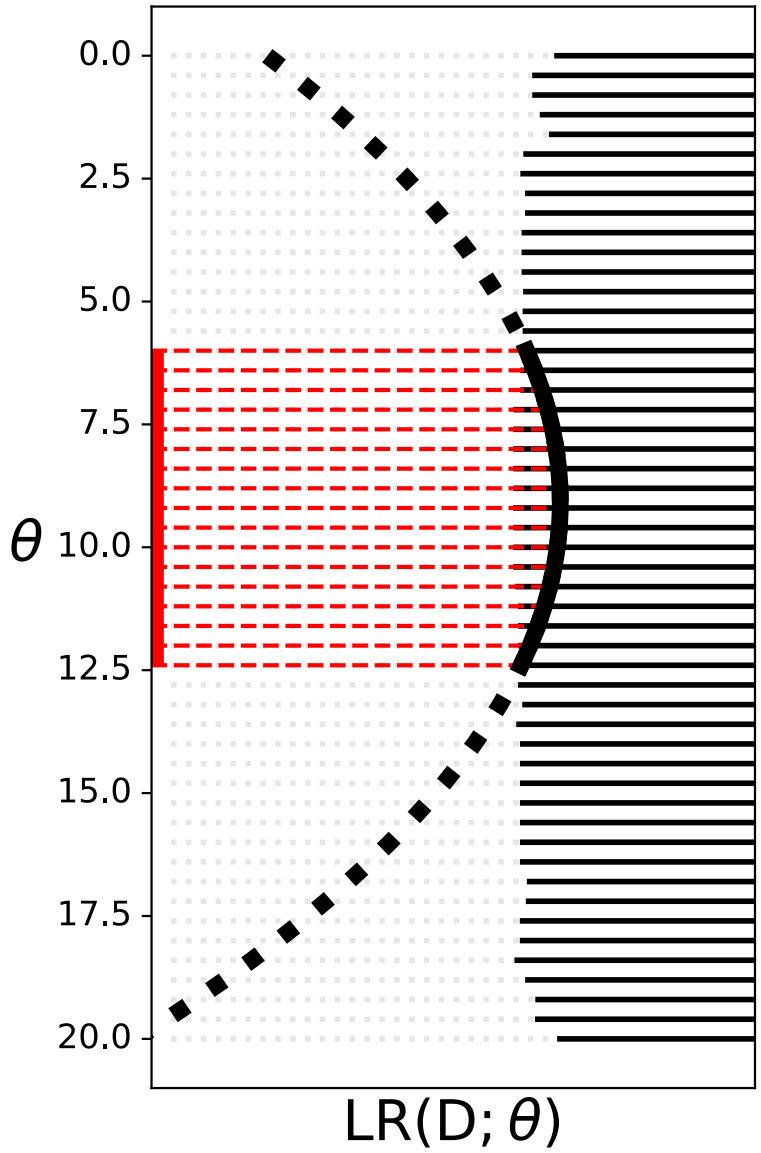




3. Observe data $\mathcal{D} = \mathrm{D}$. Evaluate $\lambda(\mathrm{D}; \theta)$.

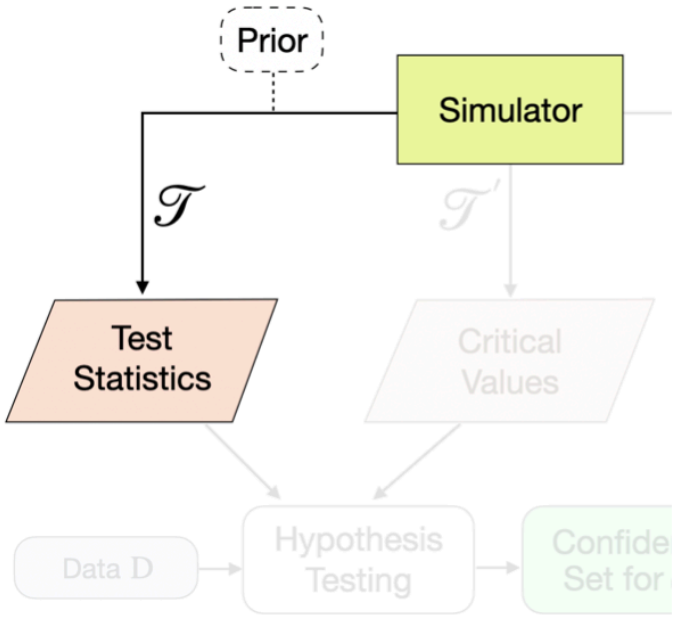


4. Construct $(1 - \alpha)$ confidence set for θ.



Leveraging Predictive NNs and Generative Models

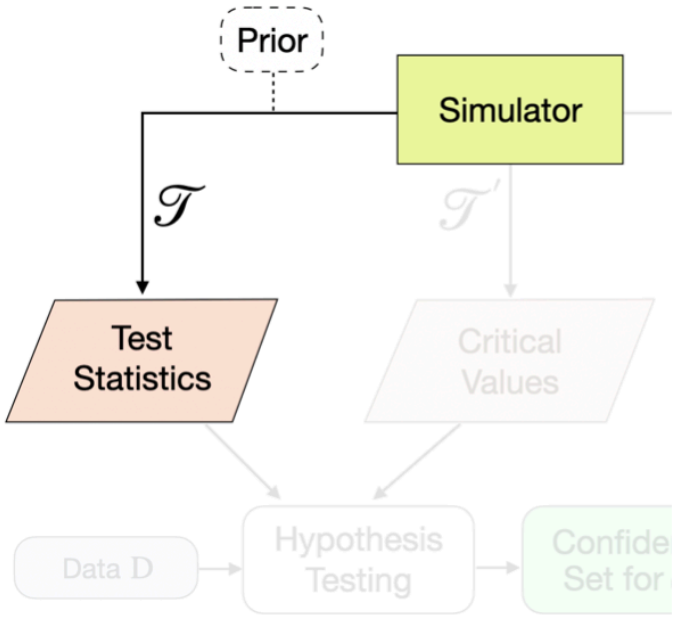
Masserano, Dorigo, Izbicki, Kuusela, Lee (AISTATS 2023)



- **Goal:** move away from likelihood-based test statistics. Why?
  - Large arsenal of AI tools for prediction and NPEs. (LR trick + maximization/integration over many parameters sometimes hard to implement in practice → loss of power)
  - LR approaches do not benefit from good priors.

Leveraging Predictive NNs and Generative Models

Masserano, Dorigo, Izbicki, Kuusela, Lee (AISTATS 2023)



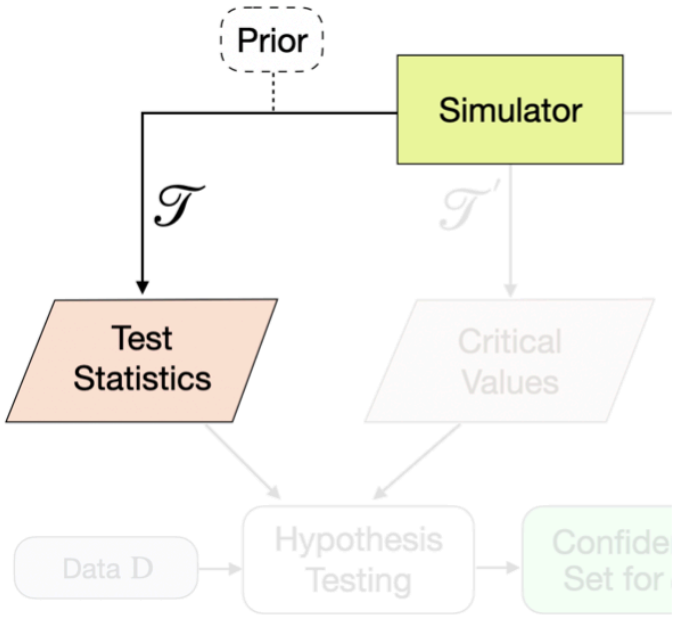
- **Goal:** move away from likelihood-based test statistics. Why?
  - Large arsenal of AI tools for prediction and NPEs. (LR trick + maximization/integration over many parameters sometimes hard to implement in practice → loss of power)
  - LR approaches do not benefit from good priors.
- From **Wald** to **Waldo** test statistic

$$\tau^{Wald}(\mathscr{D};\theta_0) := \frac{(\theta^{MLE}-\theta_0)^2}{\mathbb{V}[\theta^{MLE}]} \quad \rightarrow \quad \tau^{Waldo}(\mathscr{D};\theta_0) := \frac{(\mathbb{E}[\theta\,|\,\mathscr{D}]-\theta_0)^2}{\mathbb{V}[\theta\,|\,\mathscr{D}]}$$

- In practice:
  - **Posterior:** normalizing flows, diffusion models, etc… Approximate $\mathbb{E}[\theta\,|\,\mathscr{D}]$ and $\mathbb{V}[\theta\,|\,\mathscr{D}]$ by sampling from posterior
  - **Prediction:** under squared error loss, predictions yield $\mathbb{E}[\theta\,|\,\mathscr{D}]$. Then let, e.g., $\mathbb{V}[\theta\,|\,\mathscr{D}] = \mathbb{E}[(\theta - \mathbb{E}[\theta\,|\,\mathscr{D}])^2\,|\,\mathscr{D}]$

Leveraging Predictive NNs and Generative Models

Masserano, Dorigo, Izbicki, Kuusela, Lee (AISTATS 2023)



- **Goal:** move away from likelihood-based test statistics. Why?
  - Large arsenal of AI tools for prediction and NPEs. (LR trick + maximization/integration over many parameters sometimes hard to implement in practice → loss of power)
  - LR approaches do not benefit from good priors.
- From **Wald** to **Waldo** test statistic

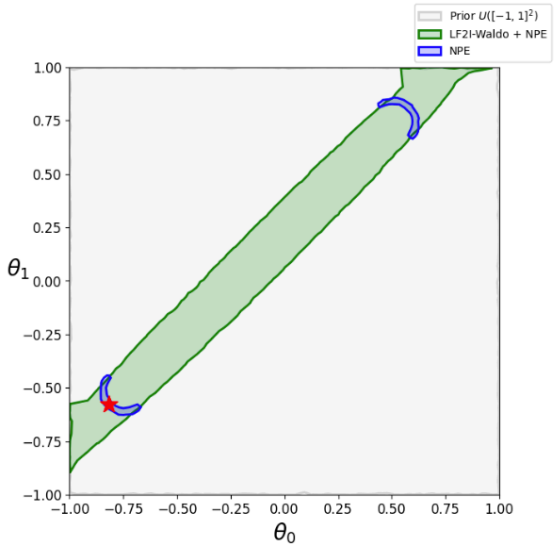
$$\tau^{Wald}(\mathscr{D};\theta_0) := \frac{(\theta^{MLE}-\theta_0)^2}{\mathbb{V}[\theta^{MLE}]} \quad \rightarrow \quad \tau^{Waldo}(\mathscr{D};\theta_0) := \frac{(\mathbb{E}[\theta\,|\,\mathscr{D}]-\theta_0)^2}{\mathbb{V}[\theta\,|\,\mathscr{D}]}$$

- In practice:
  - **Posterior:** normalizing flows, diffusion models, etc… Approximate $\mathbb{E}[\theta\,|\,\mathscr{D}]$ and $\mathbb{V}[\theta\,|\,\mathscr{D}]$ by sampling from posterior
  - **Prediction:** under squared error loss, predictions yield $\mathbb{E}[\theta\,|\,\mathscr{D}]$. Then let, e.g., $\mathbb{V}[\theta\,|\,\mathscr{D}] = \mathbb{E}[(\theta - \mathbb{E}[\theta\,|\,\mathscr{D}])^2\,|\,\mathscr{D}]$

- Quality of the prediction algorithm or posterior estimator only influences **power**
- With generative models, obtain frequentist guarantees with Bayesian posteriors

Provably optimal confidence sets from arbitrary posteriors

- Waldo suffers from a key shortcoming: it cannot handle multimodal posteriors
- Possible solution: use $p(\theta \mid X)$ directly as a test statistic. Neural density estimators allow to directly evaluate the posterior probability
- Possibly disjoint confidence sets of minimum average size:



Theorem (Informal). Let $\mathscr{R}(X) = \{\theta : h(\theta, \tau(X;\theta)) > \alpha\}$, where $h_\tau(X;\theta)$ is the p-value for the test statistic $\tau(X;\theta) = p(\theta \mid X)$. Then

$$\mathscr{R}(X) = \arg\min_{A(X)} \mathbb{E}[\,|A(X)|\,]$$

where $\mathbb{E}[\,|A(X)|\,]$ is taken with respect to the marginal $p(X)$.

- $\mathbb{E}[\,|A(X)|\,] = \int_{\mathscr{X}} \left(\int_{\mathscr{R}(x)} \mathrm{d}\theta \right) p(x)\mathrm{d}x$ and $p(x) = \int \mathscr{L}(\theta; x)\pi(\theta)\mathrm{d}\theta$