Likelihood-Free Frequentist Inference

Bridging Classical Statistics and Machine Learning in Simulator-Based Inference

Ann B. Lee

Department of Statistics & Data Science / Machine Learning Department

Carnegie Mellon University

Collaborators: Luca Masserano (CMU); Nic Dalmasso (JP Morgan); Rafael Izbicki (UFSCar); Mikael Kuusela (CMU); Tommaso Dorigo (Padova)

The Interplay Between Theory/Models and Data

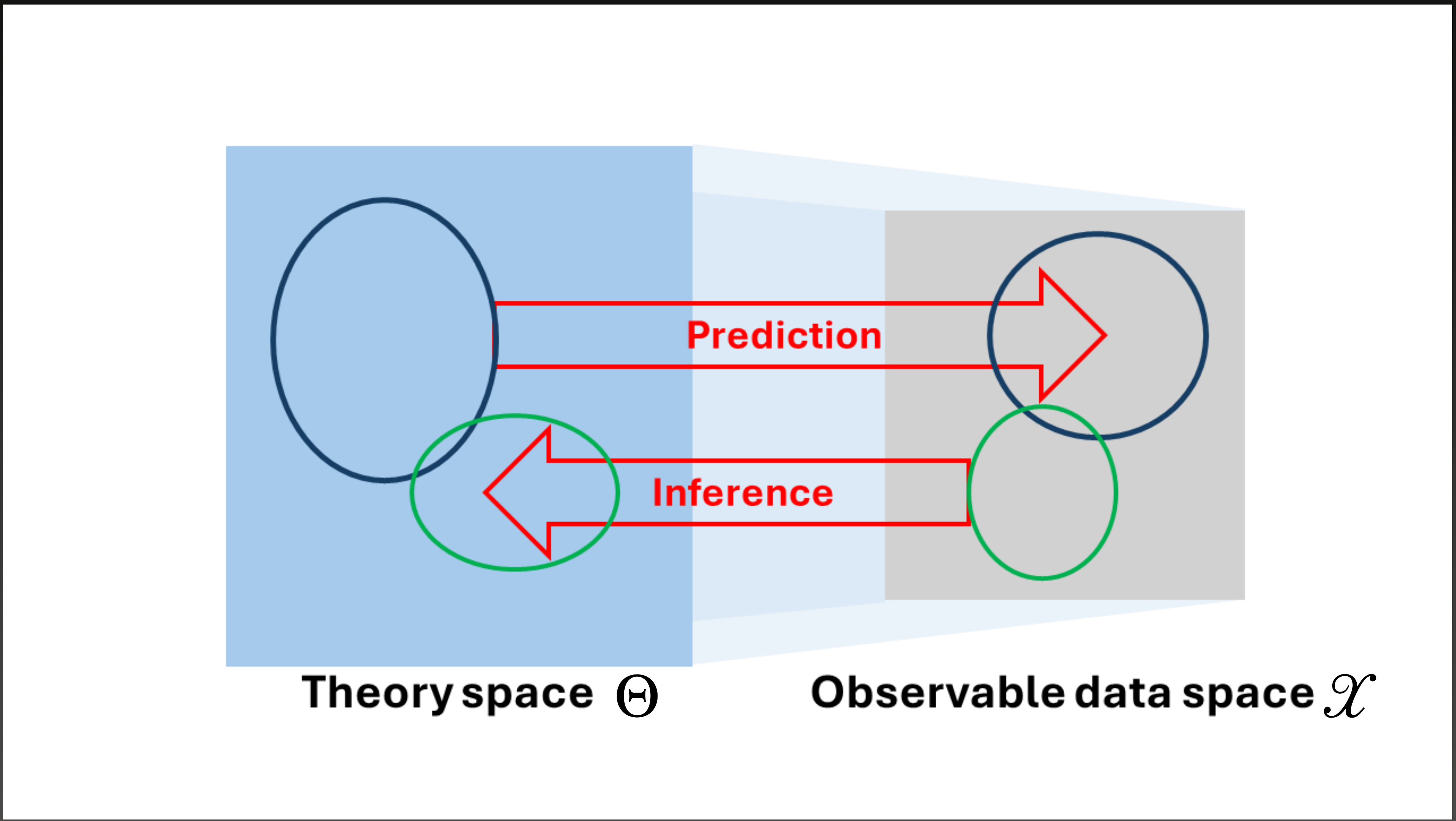
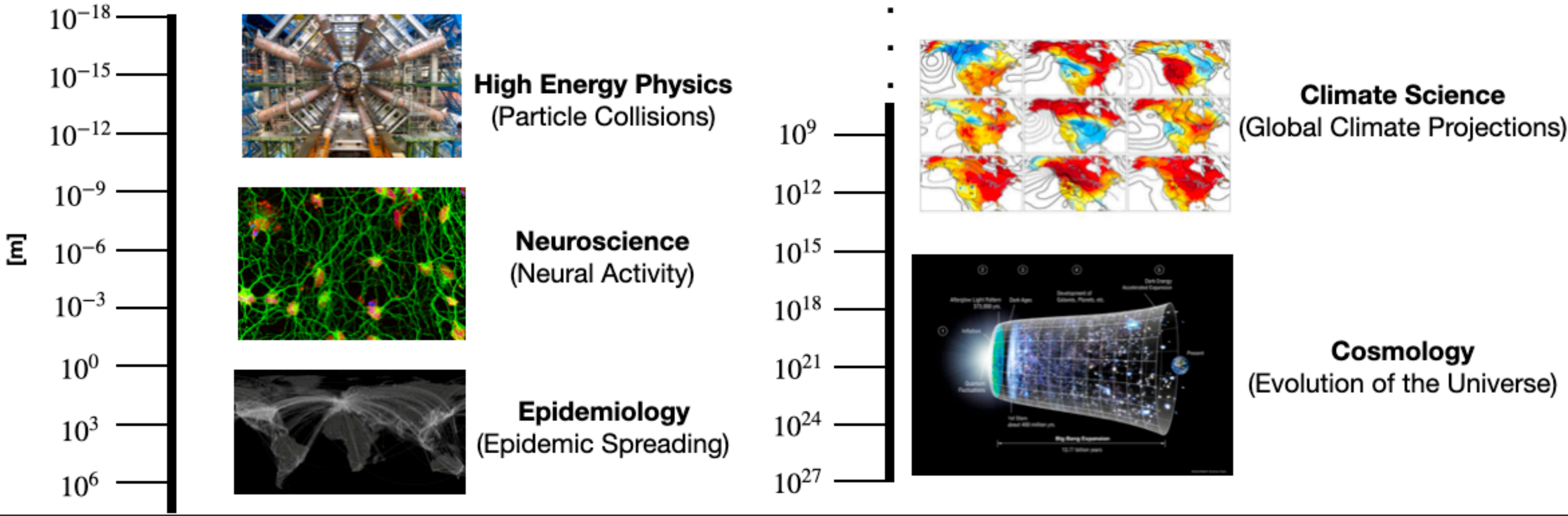


Figure credit: Tommaso Dorigo

"Theory" in the Form of Simulators



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 $\theta \in \Theta$ are internal parameters that determine measurable data $\mathscr{D} \in \mathscr{X}$

Taxonomy of Different Types of Simulators

Image credit: Kyle Cranmer

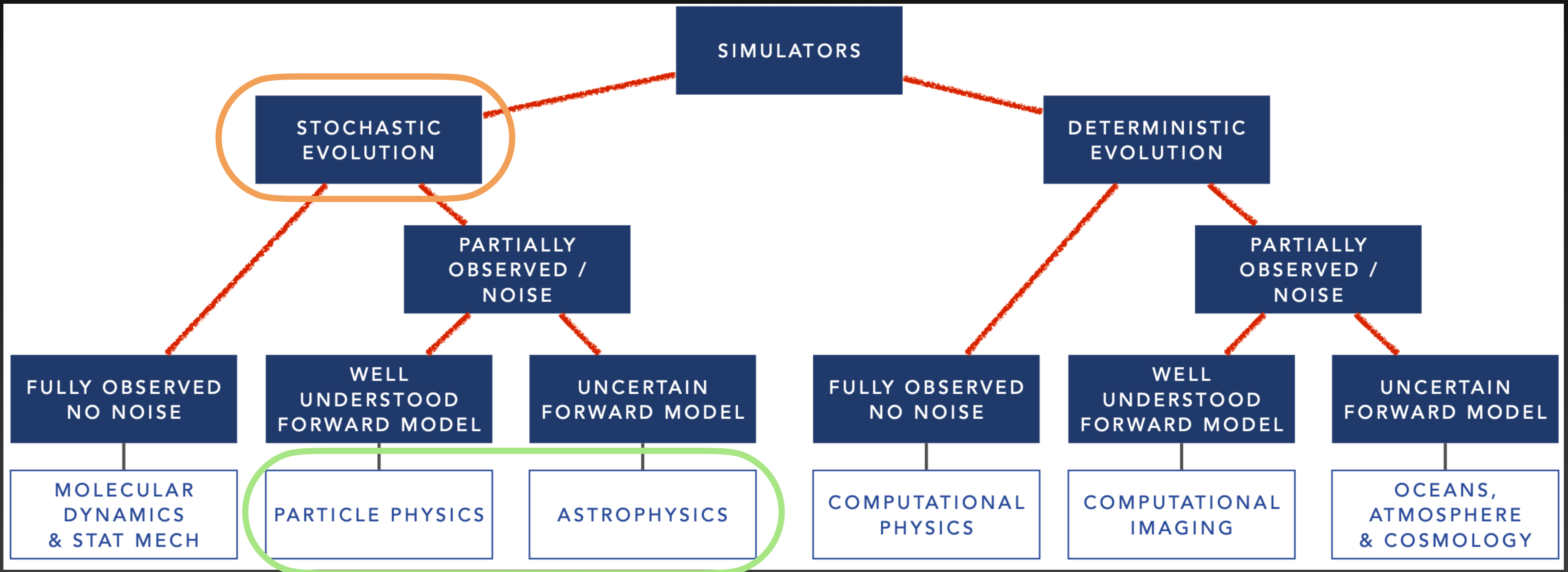


Figure credit: Kyle Cranmer

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

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

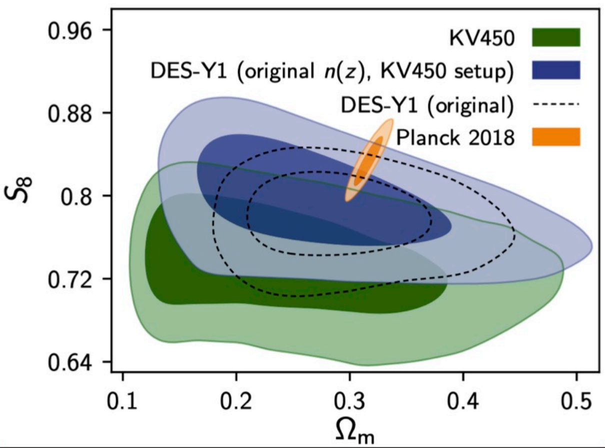
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D};\theta)$

or

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

↓

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



Simulate $(\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B)$,
where $\theta_i \sim \pi(\theta)$, $\mathcal{D}_i = \{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}\} \sim F_{\theta_i}$

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

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

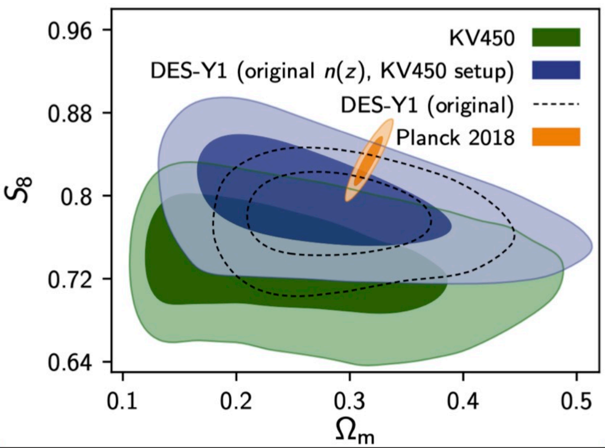
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D};\theta)$

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?

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

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

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

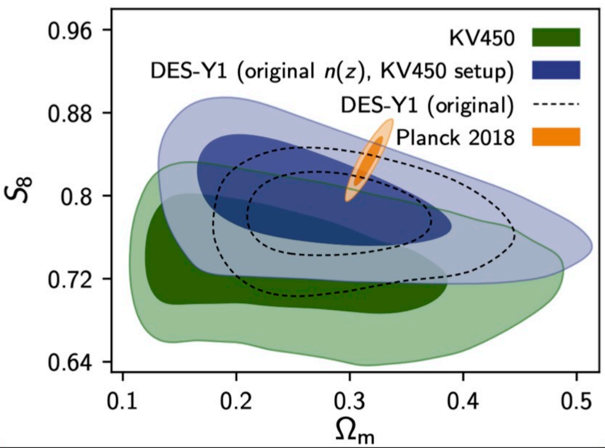
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D};\theta)$

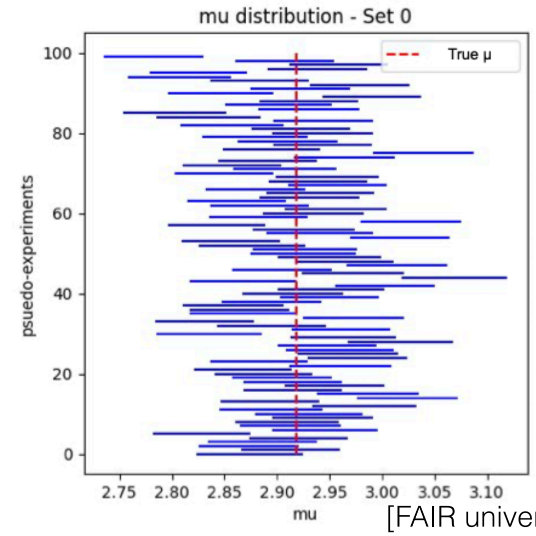
or

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

↓

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





[FAIR universe white paper]

(a) *Coverage plot*: all the predicted intervals (blue lines) for each pseudo experiment generated for a given μ_{true} (vertical dotted line).

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

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

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

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

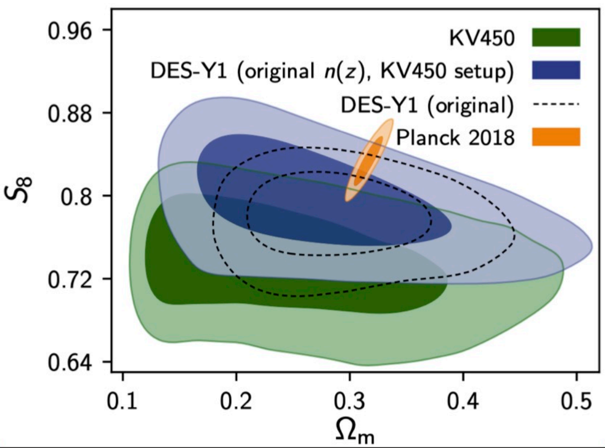
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D};\theta)$

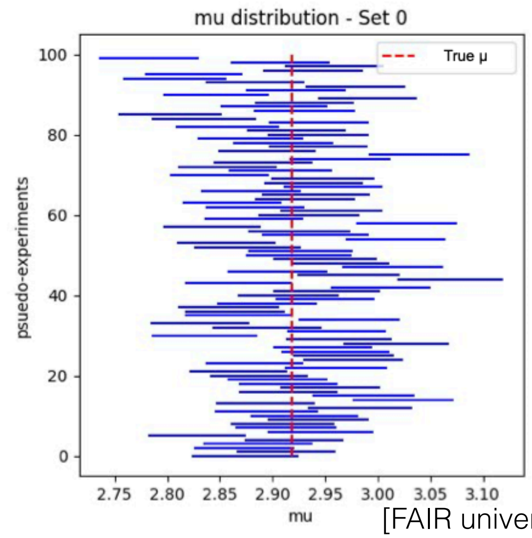
or

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

↓

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





[FAIR universe white paper]

(a) *Coverage plot*: all the predicted intervals (blue lines) for each pseudo experiment generated for a given μ_{true} (vertical dotted line).

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

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

One more note: the number of events can be small …

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

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

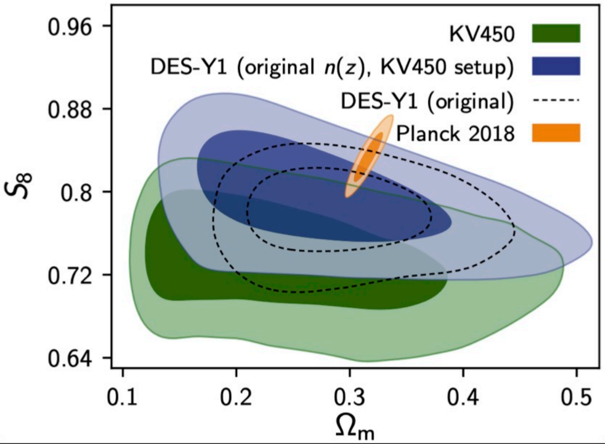
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D};\theta)$

or

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

↓

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



Simulate $(\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B)$,
where $\theta_i \sim \pi(\theta)$, $\mathcal{D}_i = \{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}\} \sim F_{\theta_i}$

- In standard frequentist statistics, n is large. Typical for HEP collider experiments
- There are also many applications (in e.g. astronomy) where n is small.
  - E.g. n=1 → single observation x from θ^*

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

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

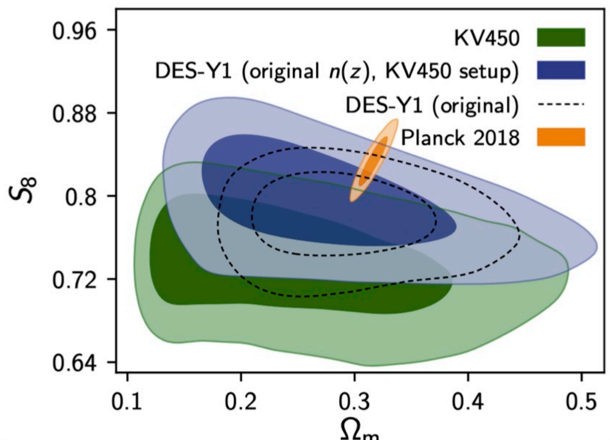
i) **Simulator** implicitly encoding $\mathscr{L}(\mathscr{D};\theta)$

or

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

↓

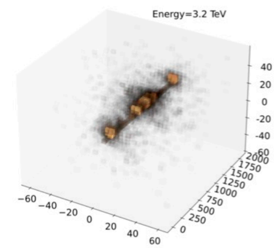
Infer internal parameters/labels of interest with measures of uncertainty.



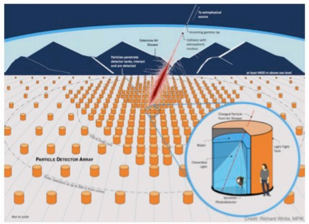
Some examples (all local parameters):

θ $\quad$ $X = \mathscr{D}$

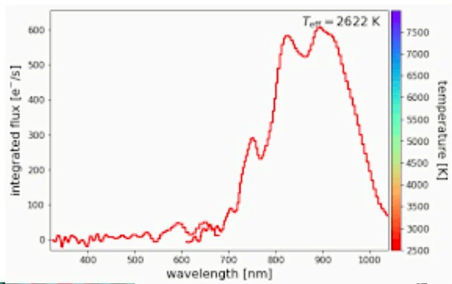
Energy of subatomic particle



Identity, orientation and energy of cosmic-ray showers

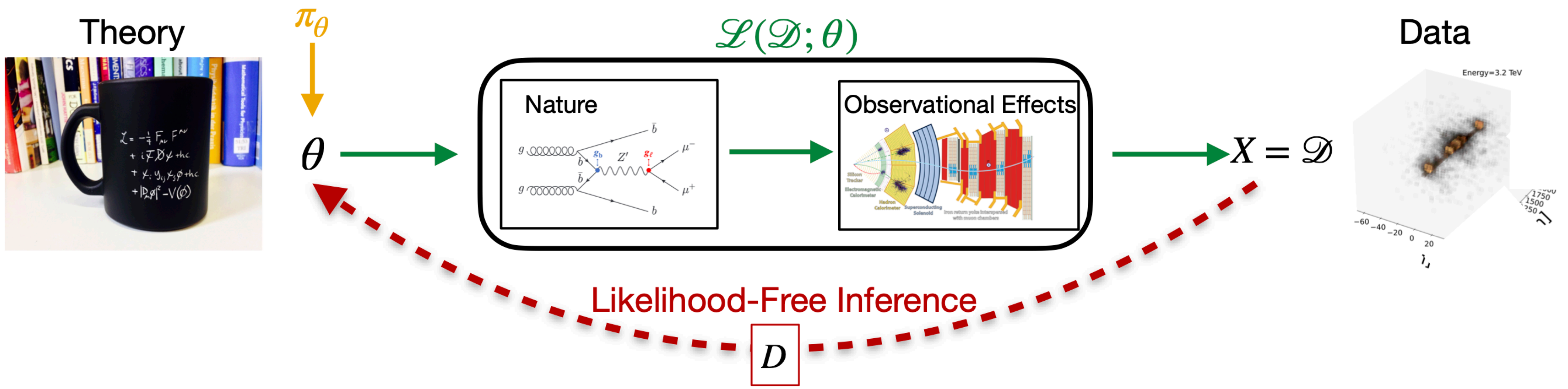


Stellar labels (e.g., mass, age, composition)



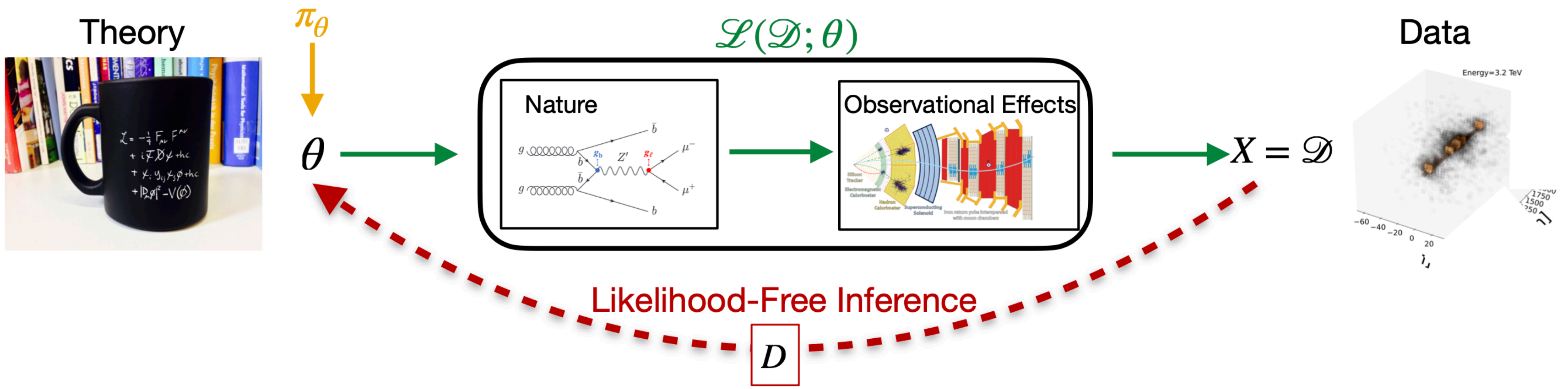
Slide credit: Luca Masserano

Complex Scientific Inference is Often “Likelihood-Free”



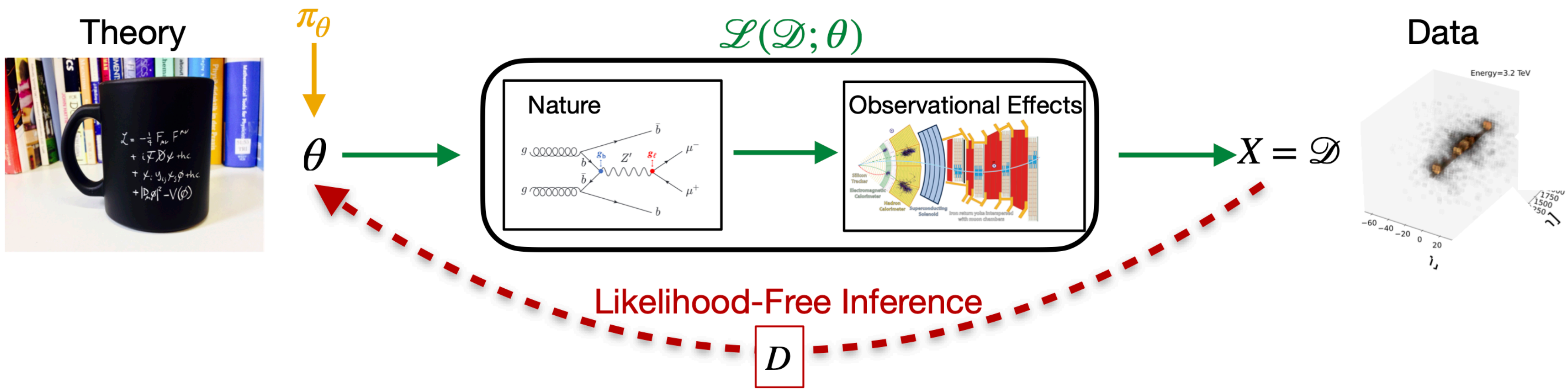
- Suppose we have knowledge of data-generating process $\theta \mapsto \mathscr{D}$ e.g. via a “high-fidelity simulation”
- But **likelihood is intractable:** e.g, $p(x \mid \theta) = \int p(x \mid z)p(z \mid \theta)\mathrm{d}z$, where z are latent variables
- Inference (inverse problem) is hard: given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

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

$L(\mathcal{D};\theta)$ same, but $\pi(\theta) \neq p_{\text{target}}(\theta)$

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.

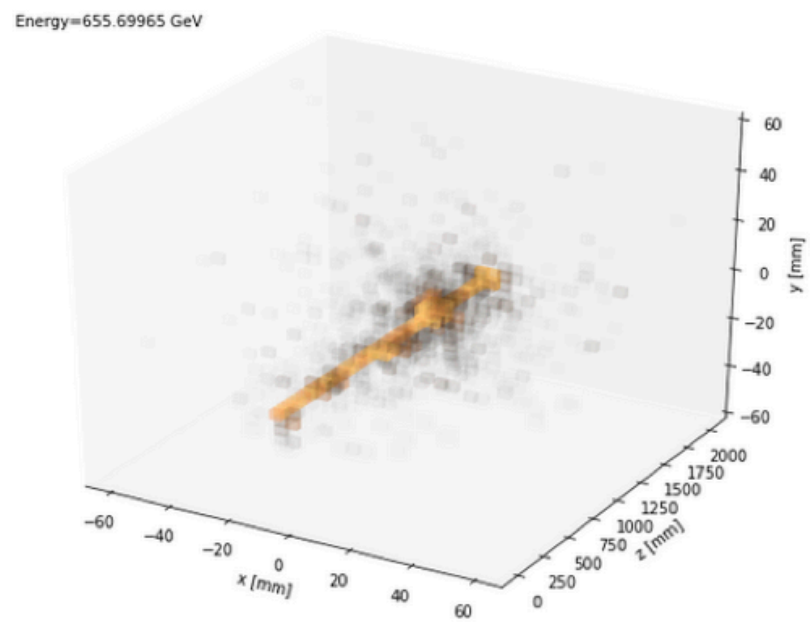


Figure 4: Muon entering the calorimeter in z direction.

[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

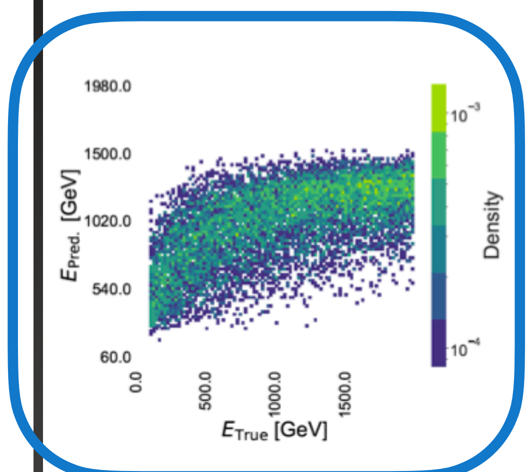


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

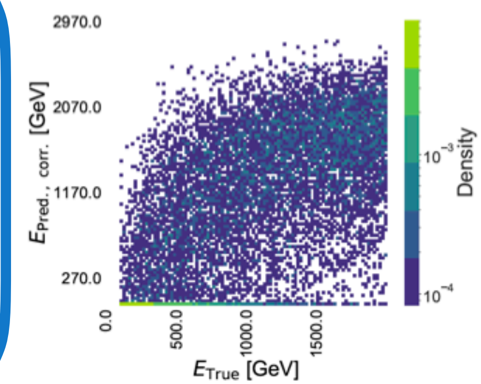


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

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

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

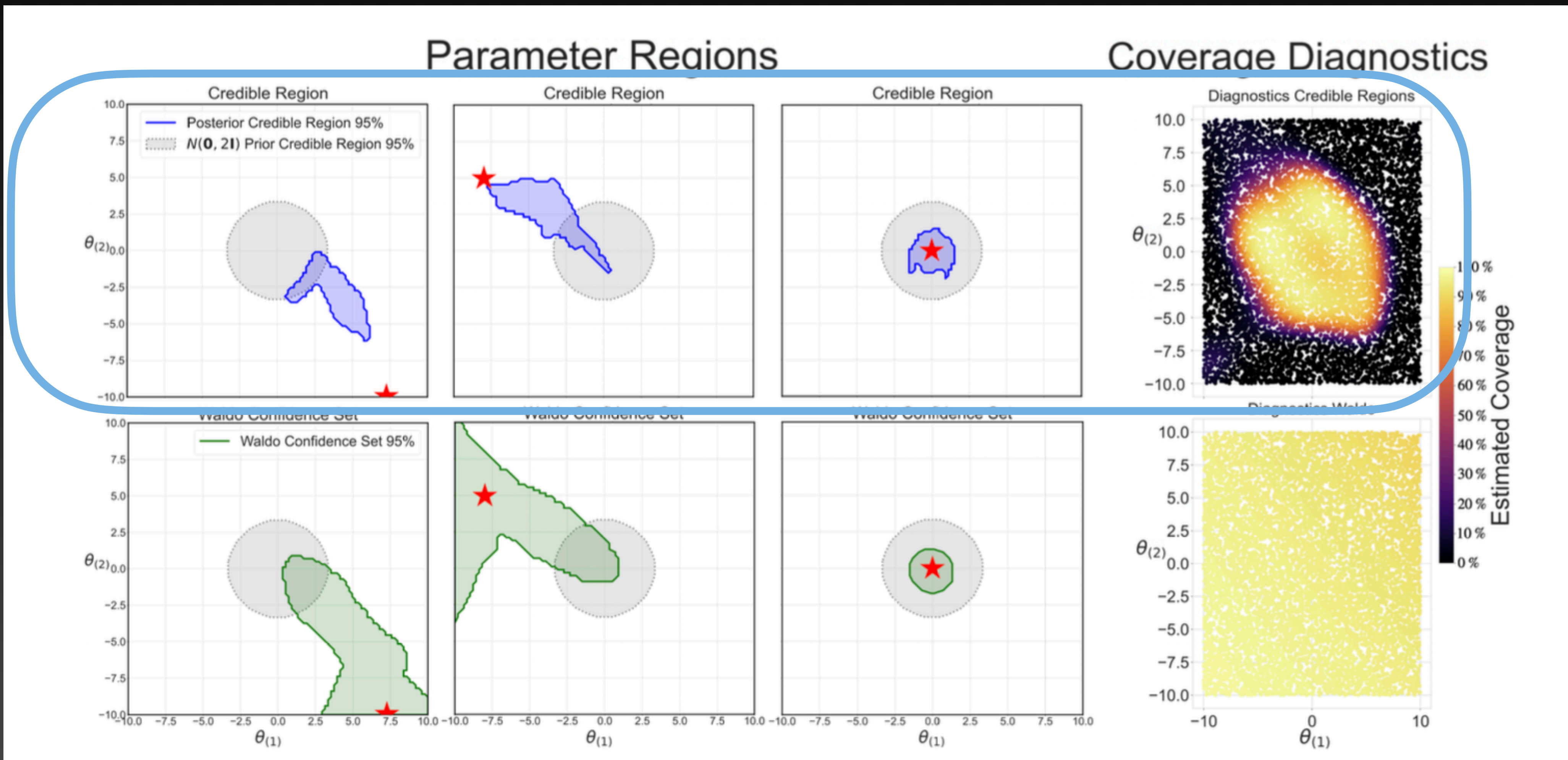
We present extensive empirical evidence showing that current Bayesian simulation-based inference algorithms are inadequate for the falsificationist methodology of scientific inquiry. Our results collected through months of experimental computations show that all benchmarked algorithms – (S)NPE, (S)NRE, 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.,

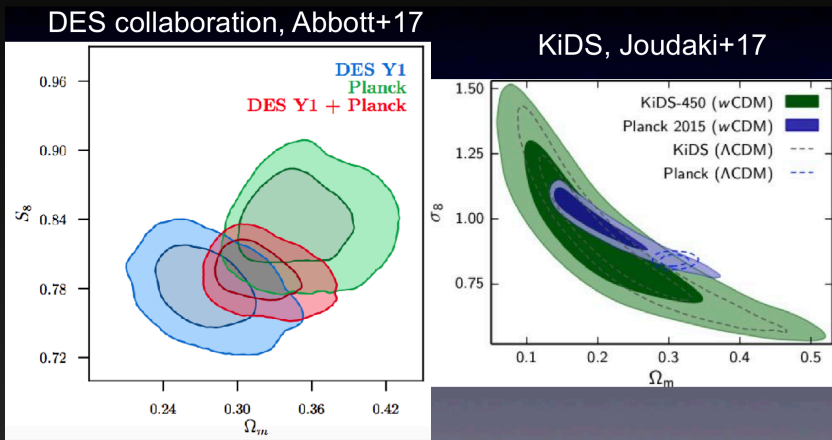
Ex: Credible Regions from Neural (NF) Posteriors

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



Blue contours: 95% credible regions from Normalizing Flows
(overly confident when prior is mismatched with true parameter)

How about Frequentist LFI Approaches?

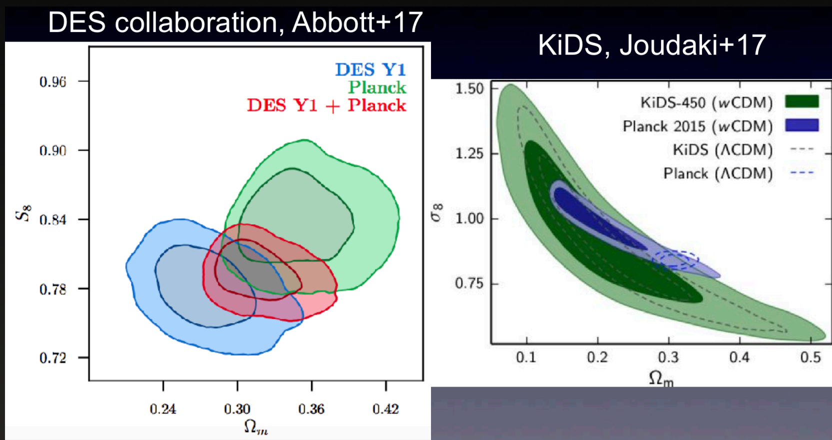


Robust coverage guarantees under shifting priors (for all θ, and for finite n)?

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

- Frequentist approaches (that estimate likelihoods or likelihood ratios) are *by construction* robust to prior prob shift
- However, most such approaches
  - rely on asymptotic assumptions (e.g. Wilks 1938) and regularity conditions
    - can't handle e.g. n=1 → single observation from θ^*
    - do not check instance-wise coverage across entire parameter space

How about Frequentist LFI Approaches?



Robust coverage guarantees under shifting priors (for all θ, and for finite n)?

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

- Frequentist approaches (that estimate likelihoods or likelihood ratios) are *by construction* robust to prior prob shift
- However, most such approaches
  - rely on asymptotic assumptions (e.g. Wilks 1938) and regularity conditions
    - can't handle e.g. n=1 → single observation from θ^*
  - lack practical tools for checking coverage across entire parameter space

Can we have it all?

Robust coverage guarantees even for small sample sizes and shifting priors ("systematics") for all $\theta \in \Theta$

Diagnostics across the entire parameter space.

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

* *All done by leveraging* the arsenal of ML/AI tools "as is" (same network architecture and same loss functions, etc)

** Modular procedures: you plug in your favorite SBI results for estimating likelihoods, posteriors or density ratios (NLE, NPE,NRE) $\Longrightarrow$ theoretical guarantees

Can we have it all?

Robust coverage guarantees even for small sample sizes and shifting priors ("systematics") for all $\theta \in \Theta$

Diagnostics across the entire parameter space.

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

* *All done by leveraging* the arsenal of ML/AI tools "as is" (same network architecture and same loss functions, etc)

** Modular procedures: you plug in your favorite SBI results for estimating likelihoods, posteriors or density ratios (NLE, NPE,NRE) $\Longrightarrow$ theoretical guarantees

General Inference Machinery for LFI

arXiv:2107.03920 (EJS 2024) 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

Confidence Sets by Inverting Tests

Theorem (Equivalence of tests and confidence sets (Neyman 1937))

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

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

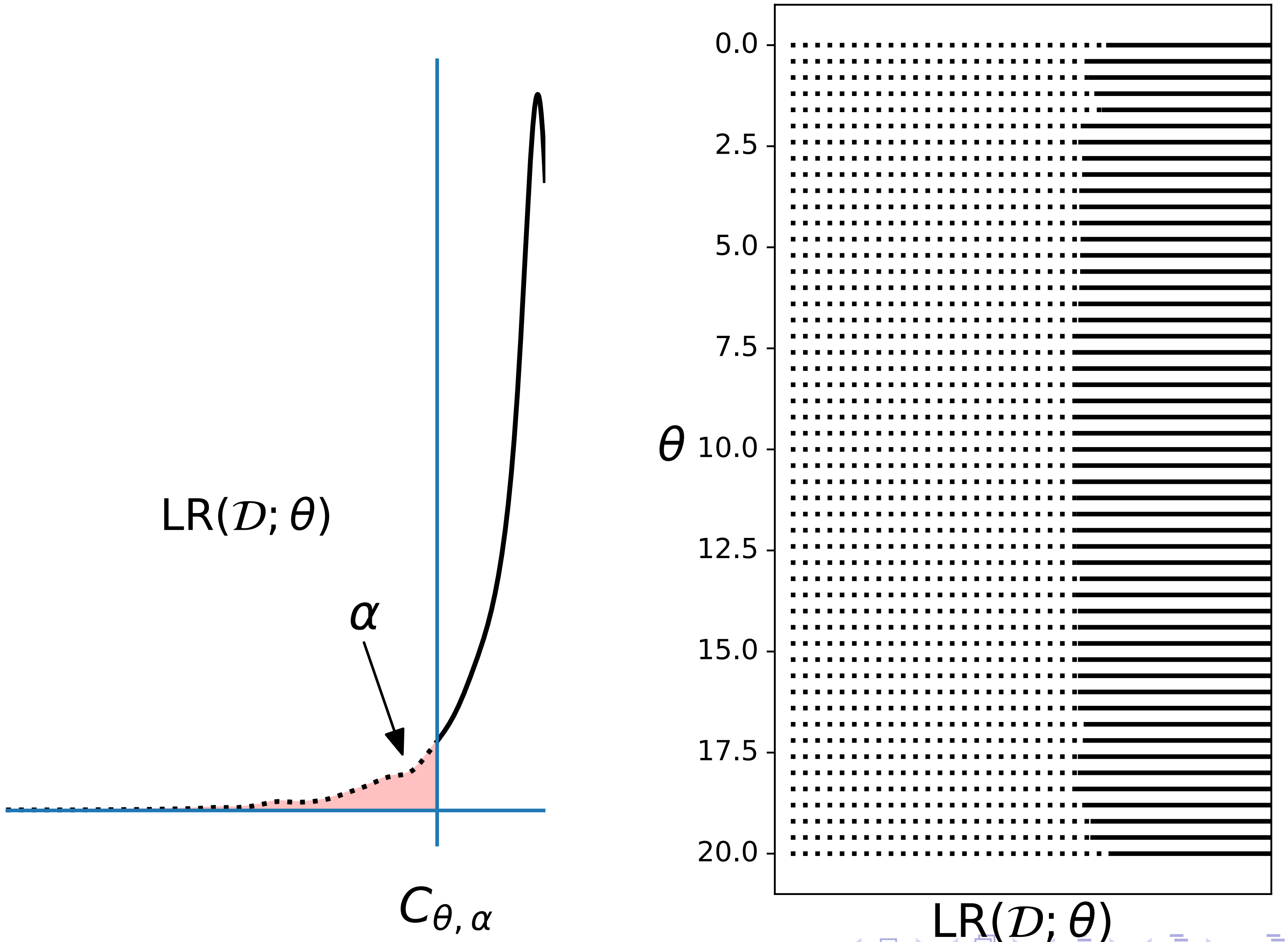
for every θ_0 in the parameter space.

Key ingredients:

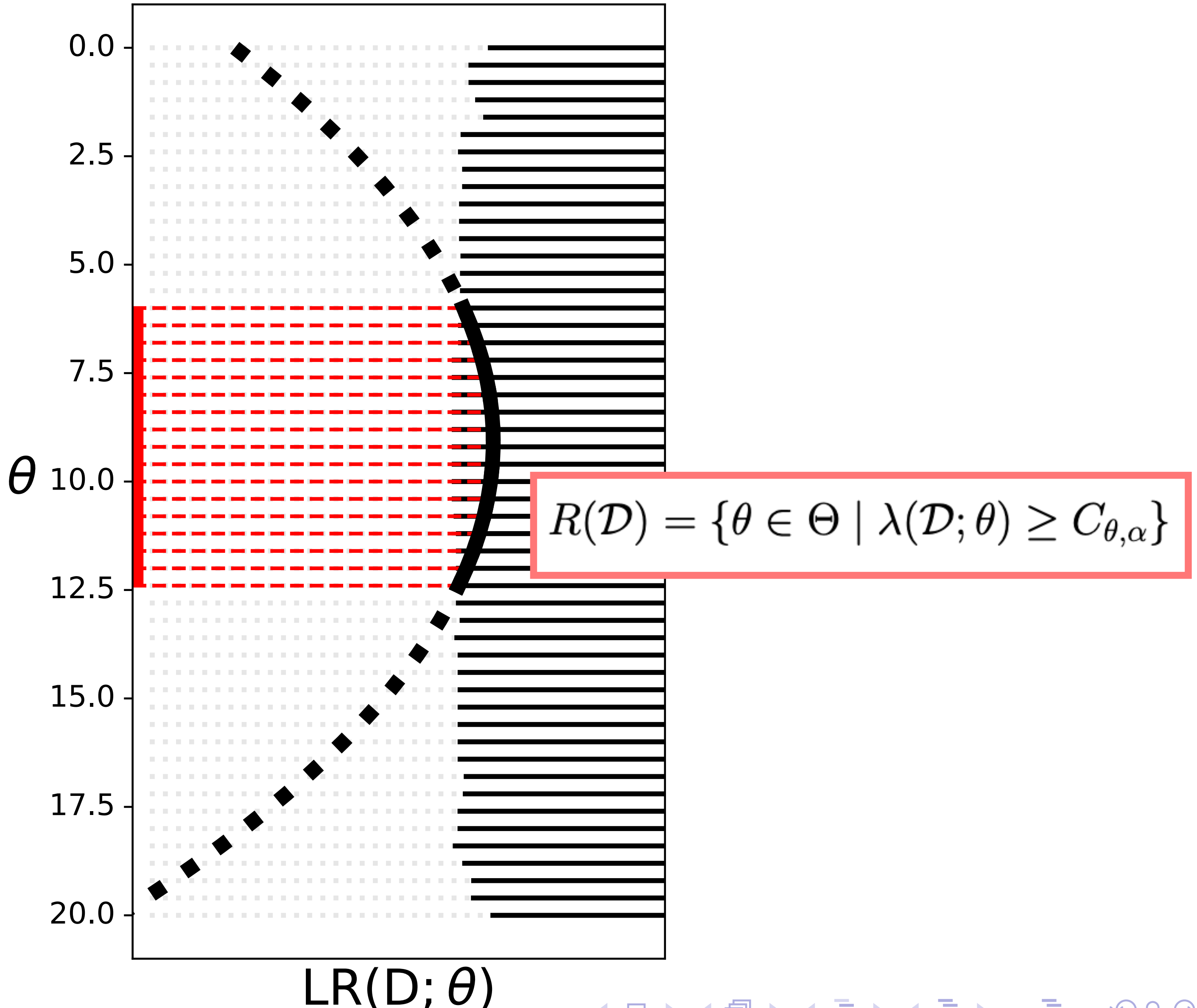
- data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$
- a test statistic, such as the likelihood ratio statistic $\lambda(\mathcal{D}; \theta_0)$
- an α-level critical value $C_{\theta_0,\alpha}$

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

1. For every θ in your parameter space: find the rejection region for test statistic $\lambda(\mathcal{D}, \theta)$



2. Observe data $\mathcal{D} = \mathrm{D}$: construct confidence set of θ by comparing $\lambda(\mathrm{D};\theta)$ and $C_{\theta,\alpha}$



How Do we Turn the Neyman Construction and Validation into Practical Procedures?

The Neyman construction requires one to test

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

for **every** $\theta_0 \in \Theta$.

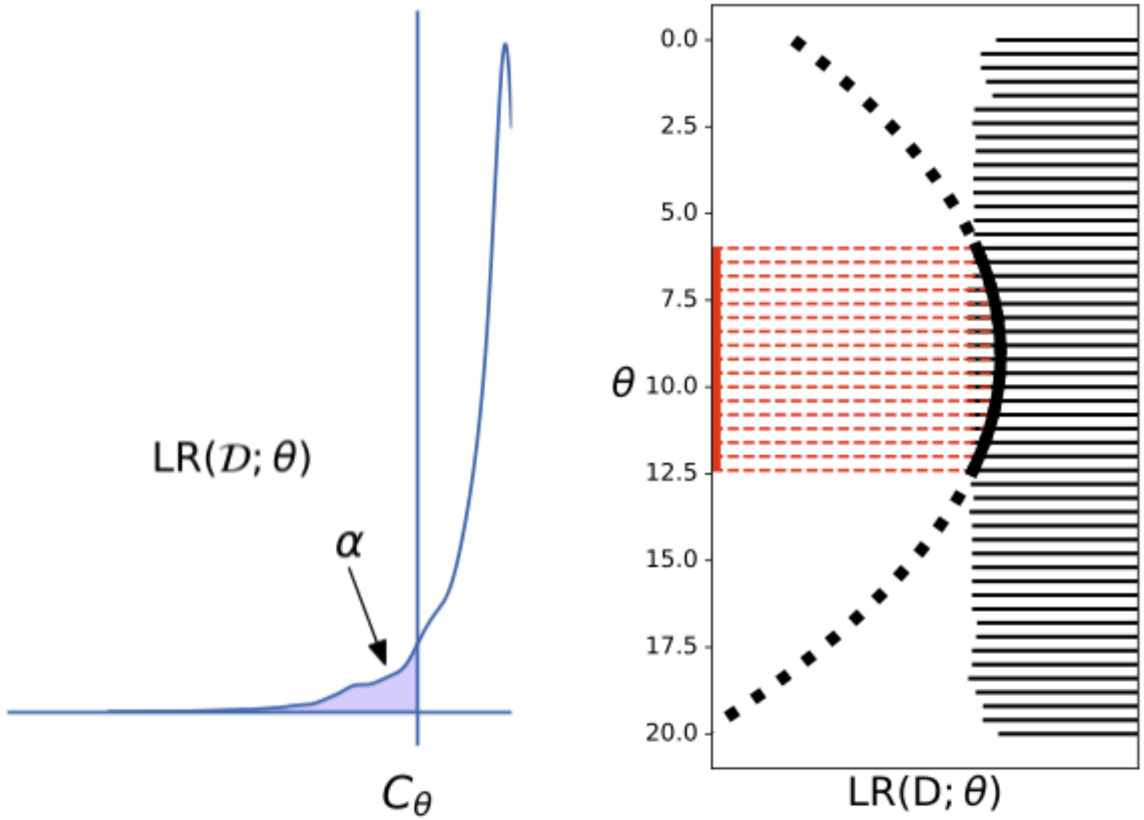
Key insight:

1. Test statistic $\lambda(\mathcal{D}; \theta)$
2. Critical values $C_{\theta_0,\alpha}$ or p-values $p(D; \theta_0)$ of the test
3. Coverage $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\right)$ of the constructed confidence set

are **conditional distribution functions** of the (unknown) parameters, and often vary smoothly across the parameter space Θ.

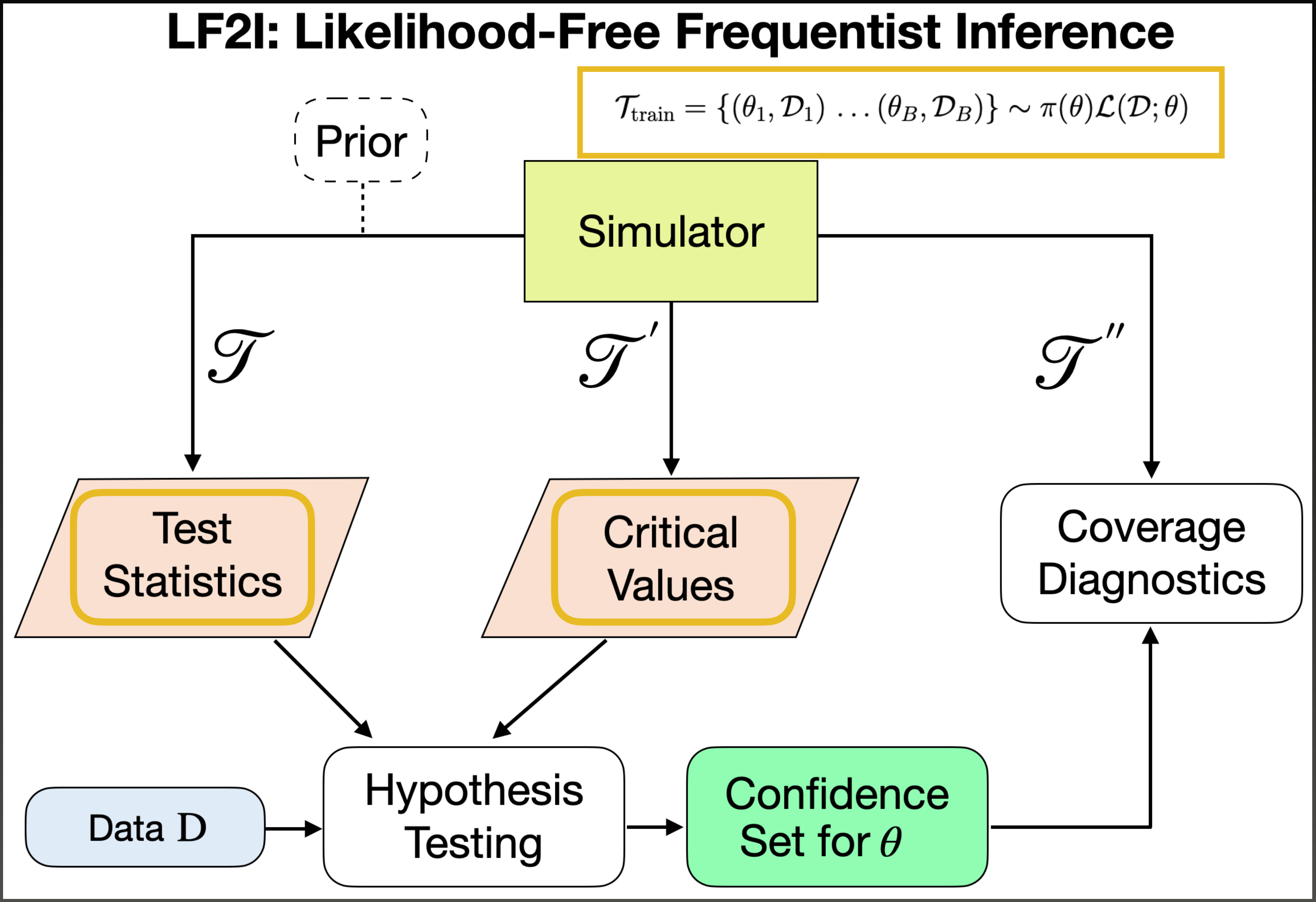
Efficient Construction of Finite-Sample Confidence Sets

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



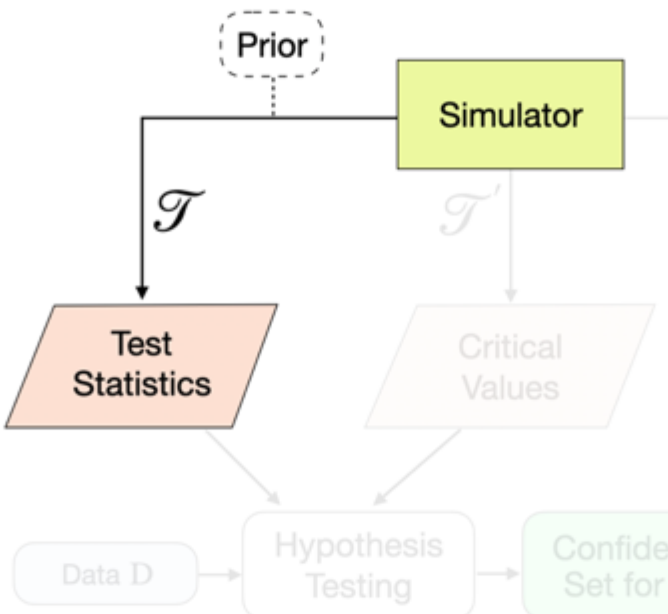
What Test Statistic?

- Derive test statistics from likelihood or LR estimates:
  - → 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]
- Derive test statistics from posteriors or predictions:
  - → "WALDO" (modified Wald test statistic) [Masserano et al 2022, arXiv:2205.15680]
  - → "Frequentist-Bayes sets" [Masserano, Carzon et al 2024-]

What Test Statistic?

- Derive test statistics from likelihood or LR estimates:
  - → 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]
- Derive test statistics from posteriors or predictions:
  - → "WALDO" (modified Wald test statistic) [Masserano et al 2022, arXiv:2205.15680]
  - → "Bayes-Frequentist sets" [Masserano, Carzon et al 2024-]

Likelihood-Based Test Statistics*



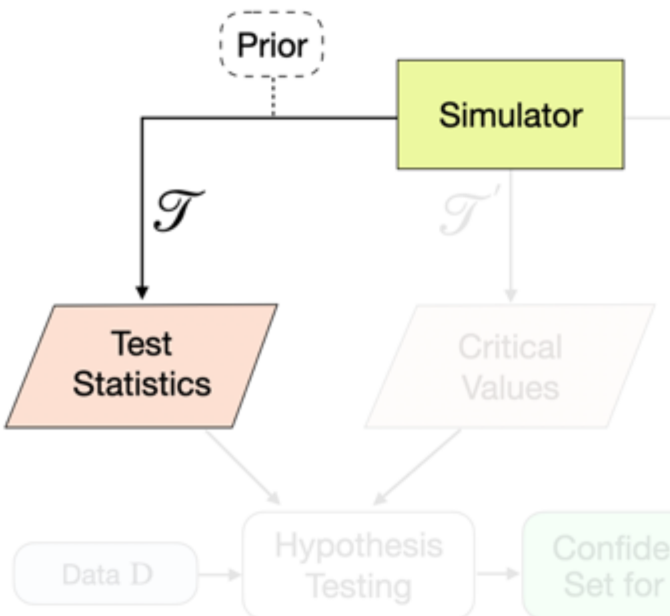
- Probabilistic classifier $h : (\theta, X) \mapsto \mathbb{P}(Y = 1 \,|\, X, \theta)$
- Simulate two sets:
  - $\{\theta_i, X_i, Y_i = 1\}_{i=1}^{B/2}$, where $\theta \sim \pi_\theta, X \,|\, \theta \sim F_\theta$
  - $\{\theta_j, X_j, Y_j = 0\}_{j=1}^{B/2}$, where $\theta \sim \pi_\theta, X \,|\, \theta \sim G$

E.g., empirica
marginal!

- Let odds $\mathbb{O}(X; \theta) := \dfrac{\mathbb{P}(Y = 1 \,|\, X, \theta)}{\mathbb{P}(Y = 0 \,|\, X, \theta)} = \dfrac{p(X \mid \theta)}{g(X)} \propto \mathscr{L}(\theta; X)$

* E.g.: Izbicki et al 2013; Cranmer et al 2015; Dalmasso et al 2020 [arXiv:2002.10399](https://arxiv.org/abs/2002.10399))

Likelihood-Based Test Statistics*



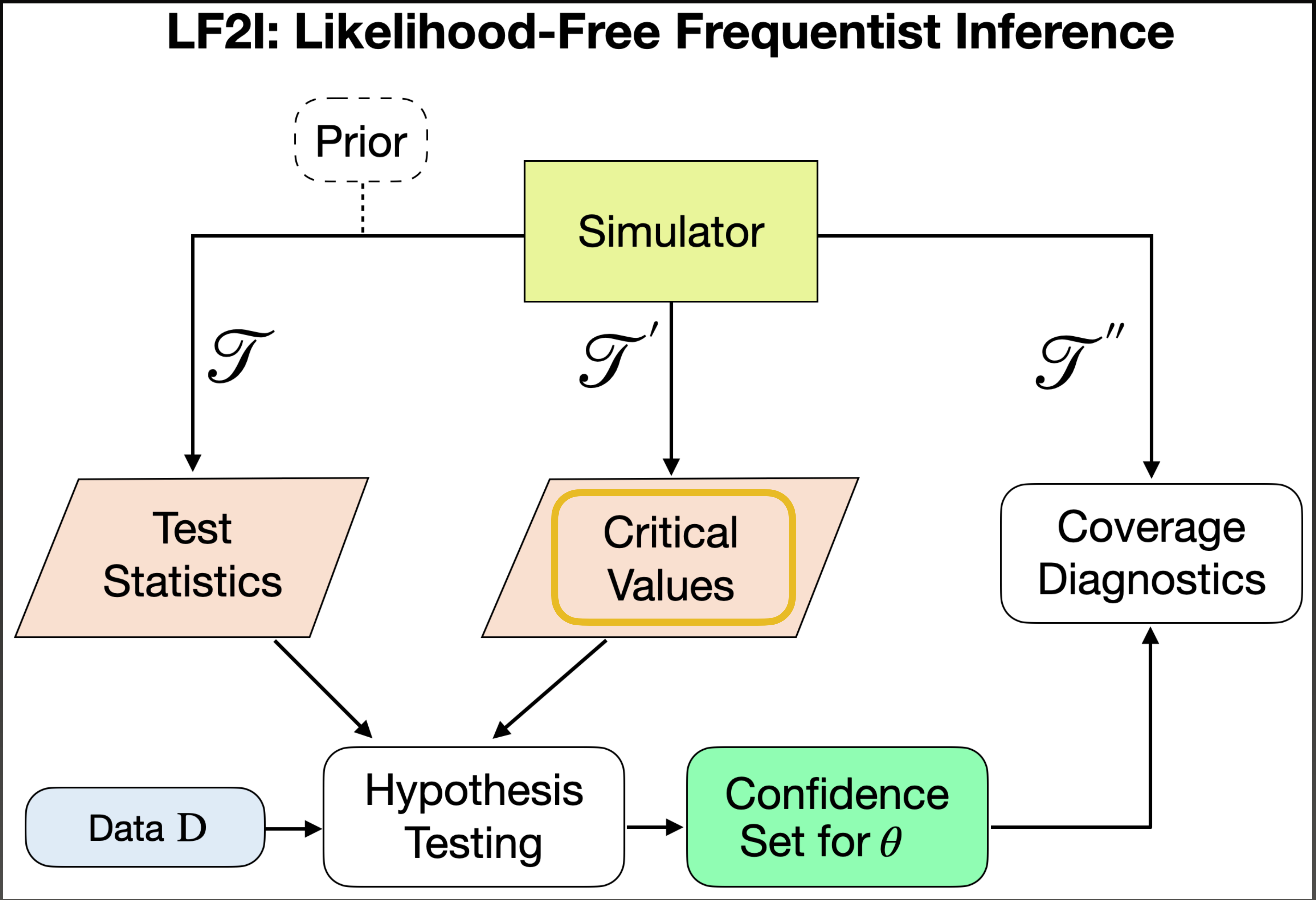
- ☐ Probabilistic classifier $h : (\theta, X) \mapsto \mathbb{P}(Y = 1 \,|\, X, \theta)$
- ☐ Simulate two sets:
  - $\{\theta_i, X_i, Y_i = 1\}_{i=1}^{B/2}$, where $\theta \sim \pi_\theta, X \,|\, \theta \sim F_\theta$
  - $\{\theta_j, X_j, Y_j = 0\}_{j=1}^{B/2}$, where $\theta \sim \pi_\theta, X \,|\, \theta \sim G$

E.g., empirica marginal!

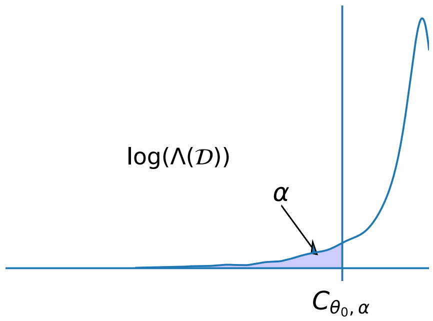
- ☐ Let odds $\mathbb{O}(X;\theta) := \dfrac{\mathbb{P}(Y = 1 \,|\, X, \theta)}{\mathbb{P}(Y = 0 \,|\, X, \theta)} = \dfrac{p(X \mid \theta)}{g(X)} \propto \mathcal{L}(\theta; X)$
- ☐ For $\mathcal{D} = (X_1, \ldots, X_n)$. Test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
  - **ACORE.** $\tau(\mathcal{D};\theta_0) := \dfrac{\prod_{i=1}^{n} \mathbb{O}(X_i;\theta_0)}{\sup_\theta \prod_{i=1}^{n} \mathbb{O}(X_i;\theta)}$
  - **BFF.** $\tau(\mathcal{D};\theta_0) := \dfrac{\prod_{i=1}^{n} \mathbb{O}(X_i;\theta_0)}{\int \prod_{i=1}^{n} \mathbb{O}(X_i;\theta) d\pi(\theta)}$

* E.g.: Izbicki et al 2013; Cranmer et al 2015; Dalmasso et al 2020 arXiv:2002.10399)

Center Branch: Estimate Critical Values



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



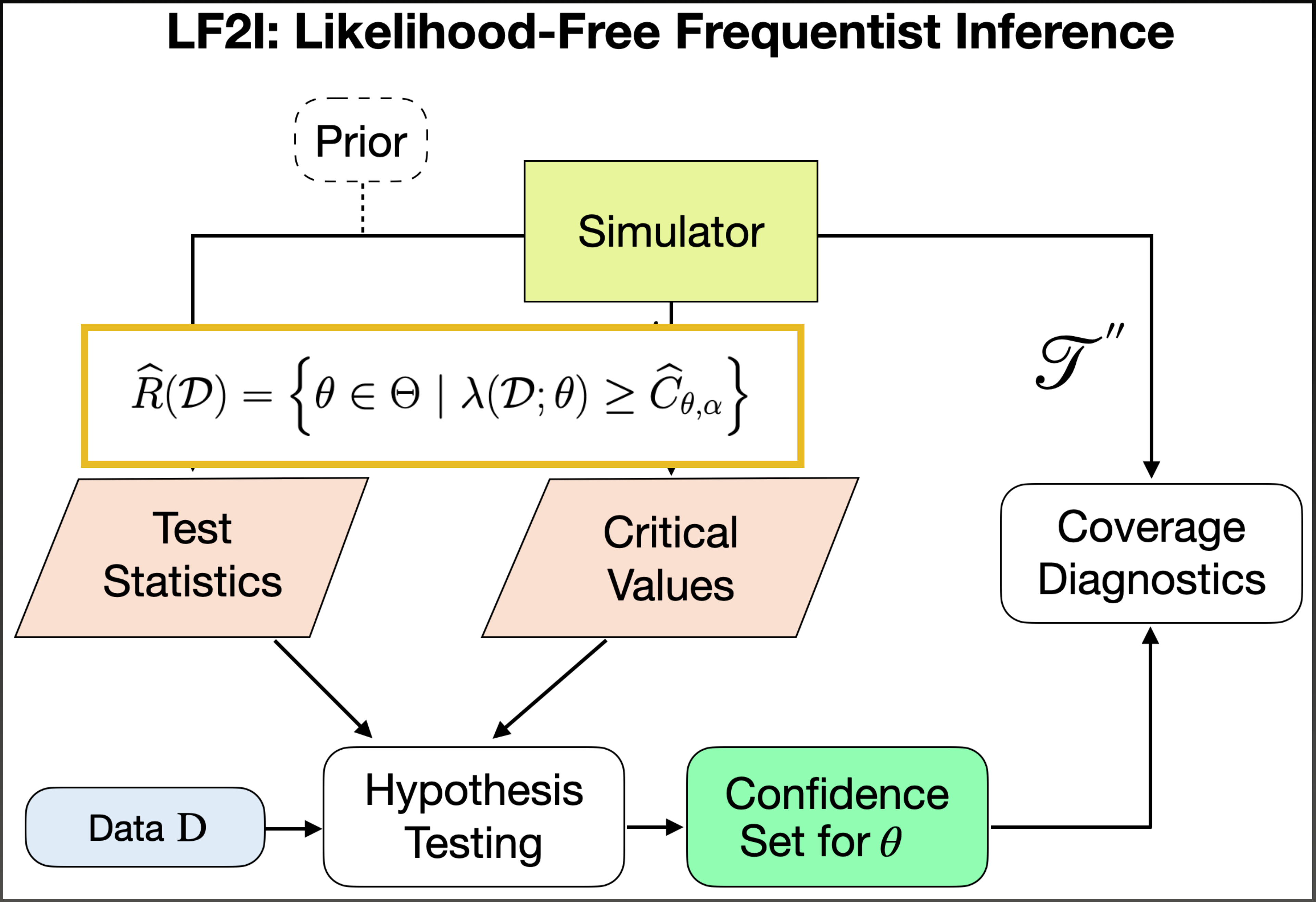
To control Type I error at level α:
Reject $H_0 : \theta = \theta_0$ when $\lambda(\mathcal{D};\theta_0) < C_{\theta_0,\alpha}$, where

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

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,

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

where C^* is such that

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

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.

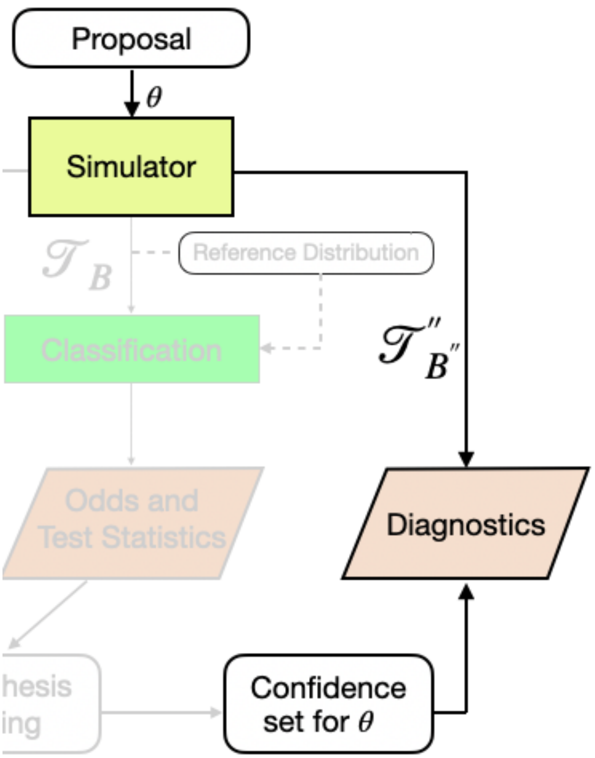
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

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.

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

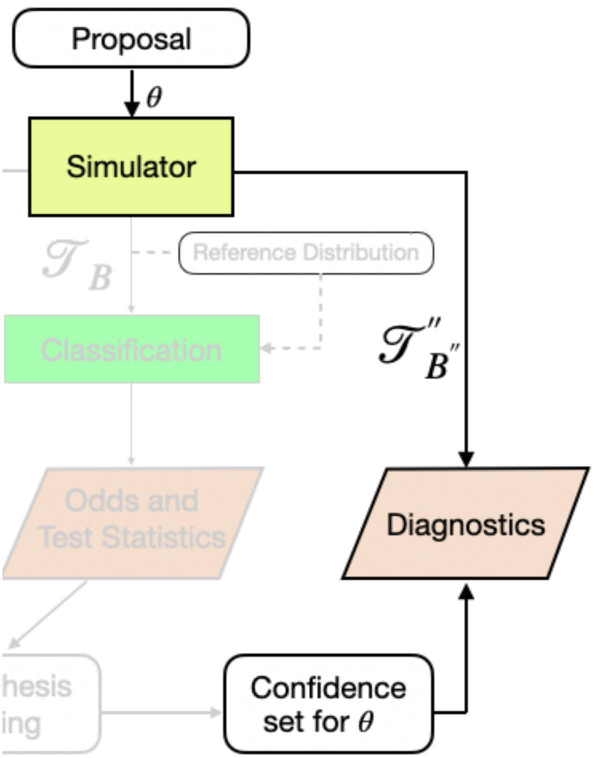
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

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

Note:

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

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



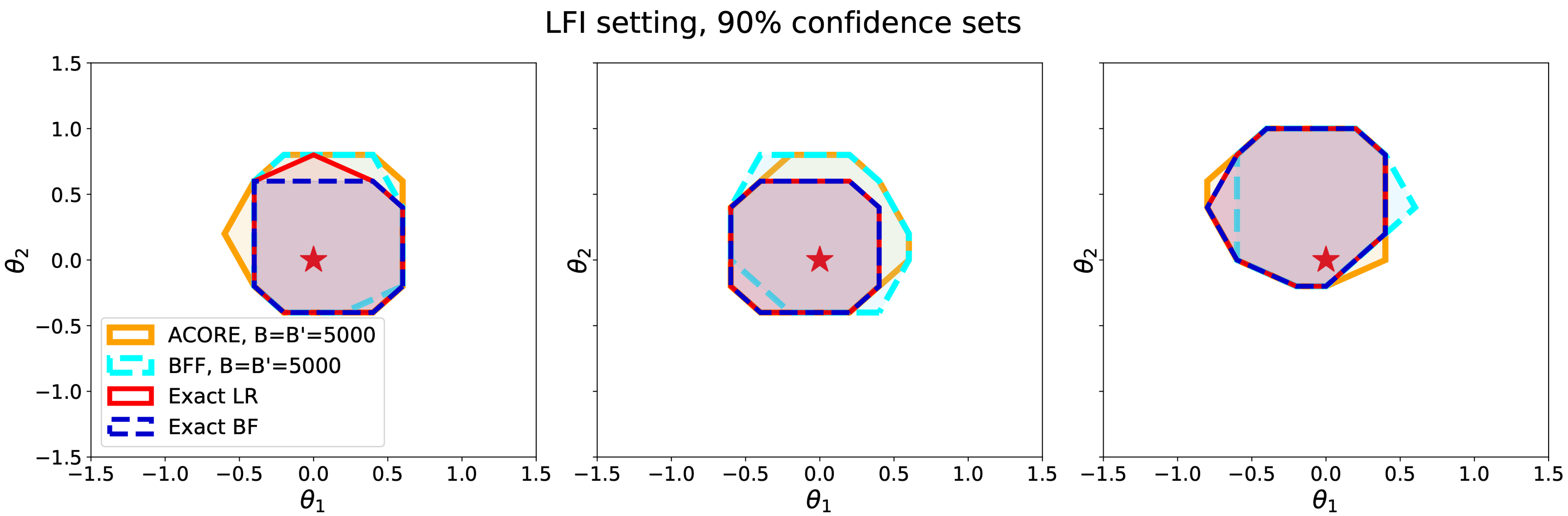
1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $\widehat{R}(\mathcal{D}_i)$
3. For $\{\theta_i, \widehat{R}(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i := \mathbb{I}(\theta_i \in \widehat{R}(\mathcal{D}_i))$ on θ_i.

Independent check of coverage across parameter space

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

Ex: Construct Confidence Sets (MVG data)

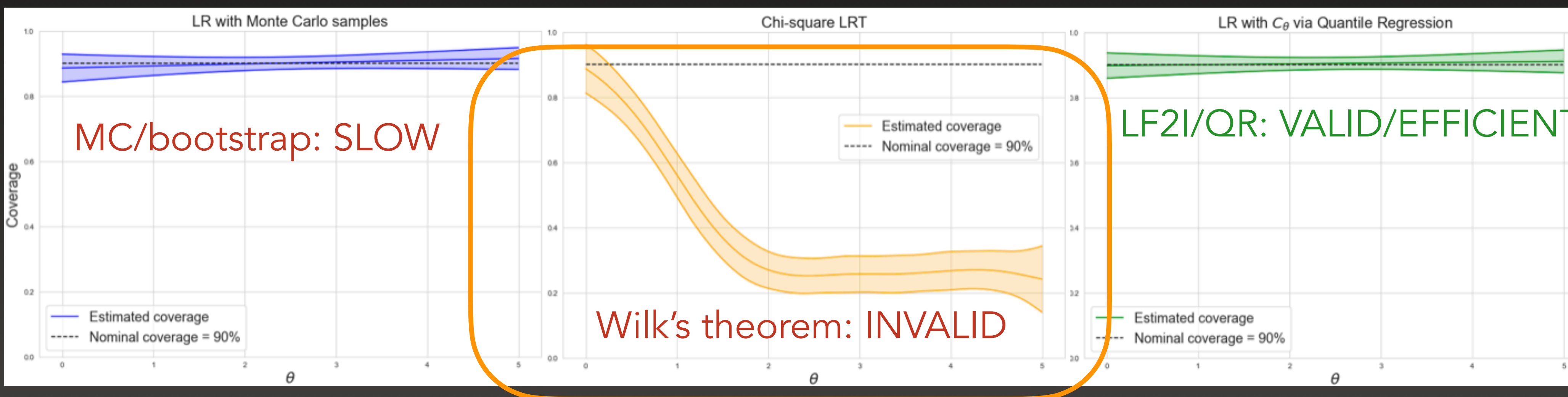
$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N(\boldsymbol{\theta}, \mathbf{I}_d), \text{ where } n = 10, \ \boldsymbol{\theta} = \mathbf{0}$$



For d<10, ACORE (estimate LRT) and BF (estimate BF) confidence sets (for B=B'=5000) are similar in size to the Exact LR confidence

Ex: 1D Gaussian mixture model with n=1000 (Diagnostics Across the Parameter Space)

$$X_1, \ldots, X_n \sim 0.5N(\theta, 1) + 0.5N(-\theta, 1)$$



(Left) LR with 1000 MC simulations at each θ on a fine grid in 1D

(Center) Assume chi-squared distribution of LR statistic

(Right) LR with quantile regression with B'=1000 simulations total

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.

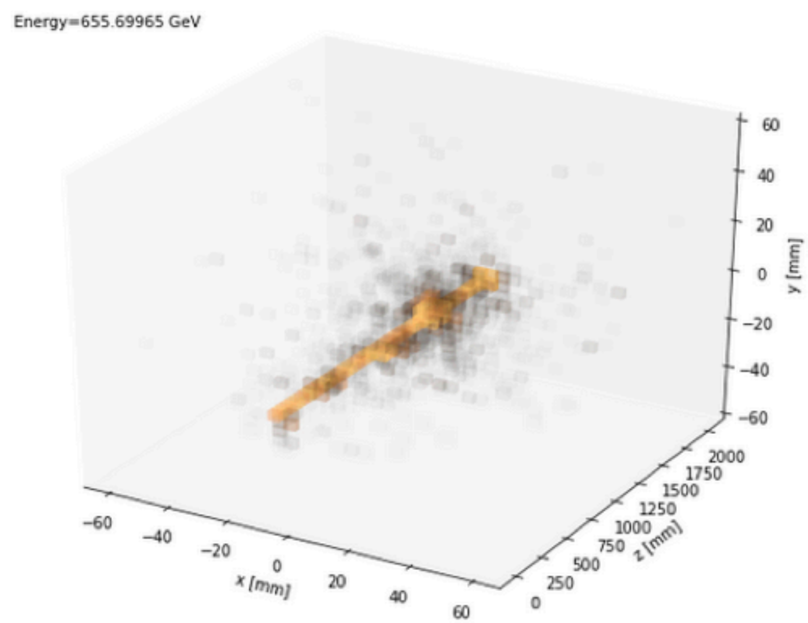


Figure 4: Muon entering the calorimeter in z direction.

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

1. Bias

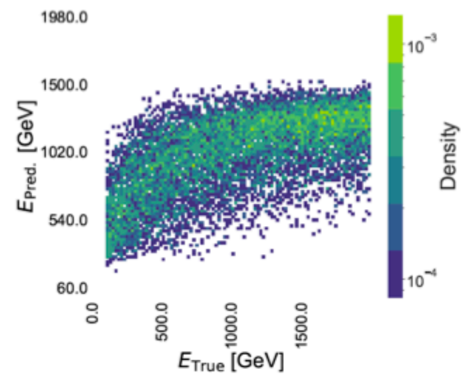


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

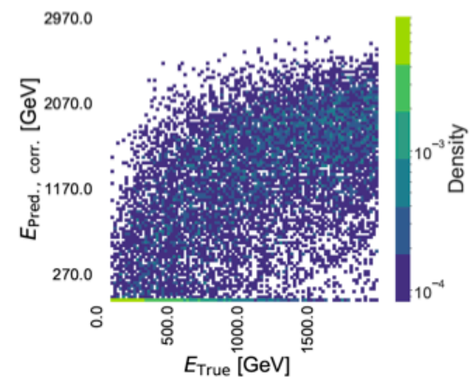


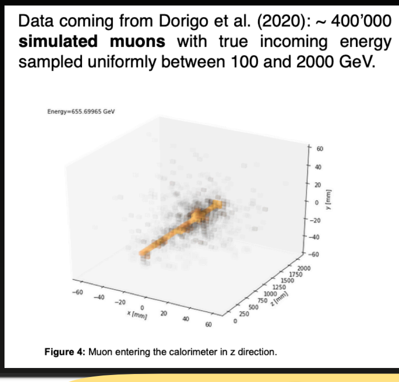
Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

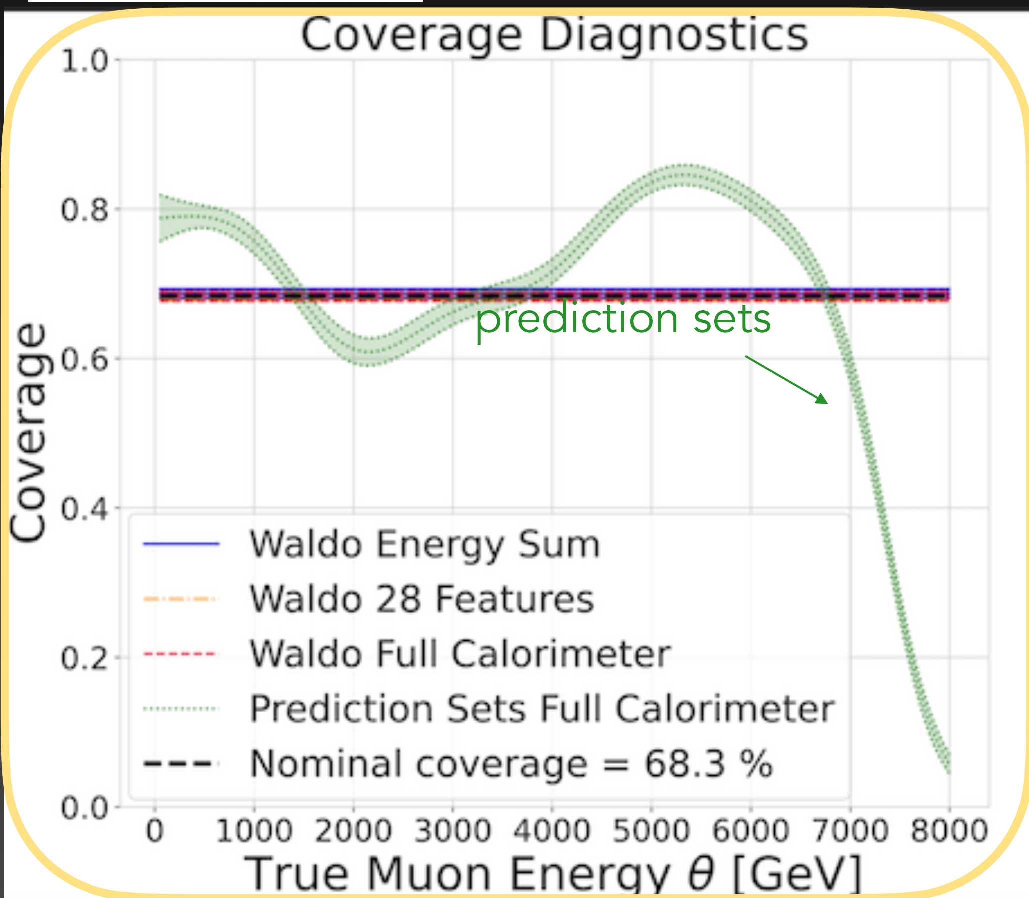
$$\mathbb{E}[\theta | X] \neq \theta^\star$$

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

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

Back to muon energy calorimeter problem:
LF2I/Waldo Confidence Sets Derived from CNN Predictions: Robust Coverage Across the Parameter Space





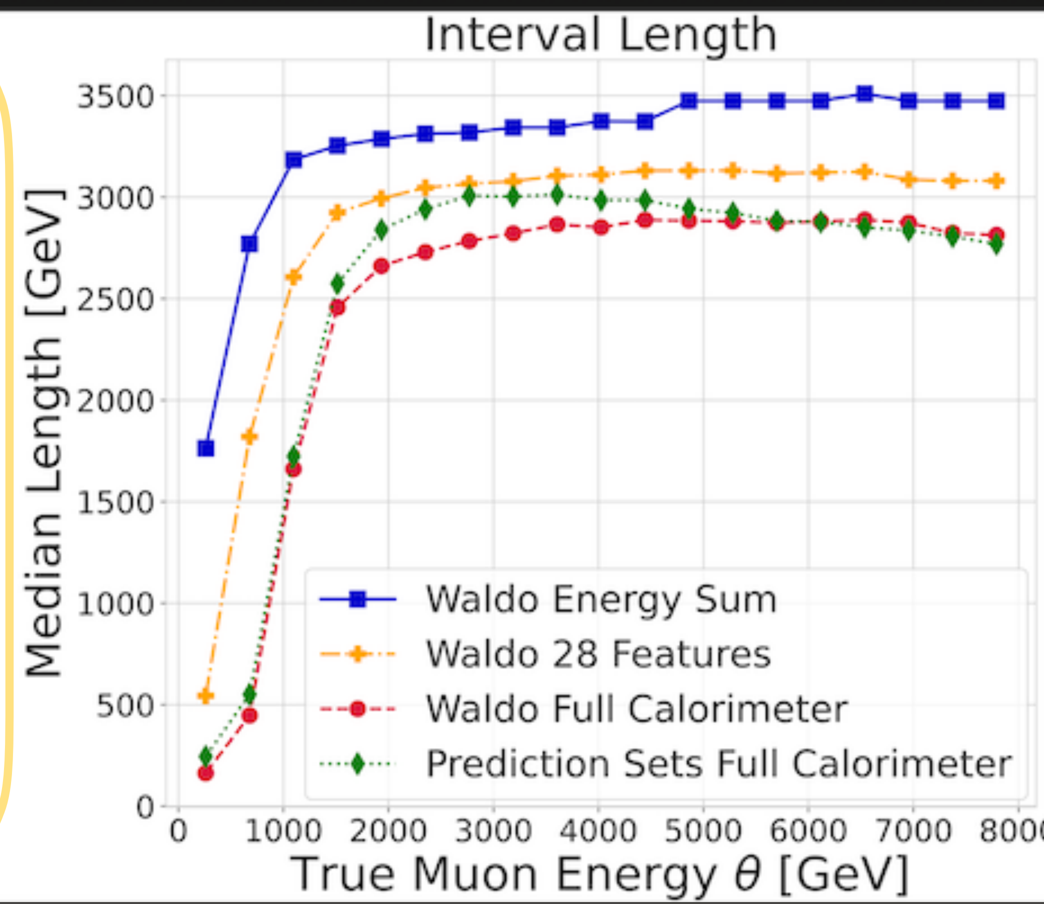


Figure credit: Luca Masserano

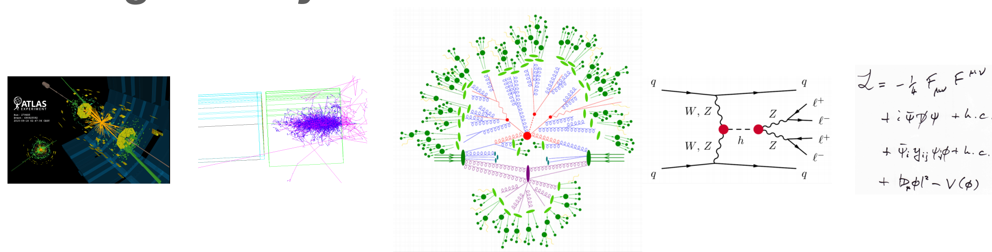
[arXiv:2205.15680](https://arxiv.org/abs/2205.15680) (AISTATS 2023)

But what if we have >1,000 nuisance parameters?

The parameters θ

One more issue: the "theory" space is not the only thing effecting the data

- **every step of the forward process** **comes with its own parameters**
  (we understand the process generally but need additional knobs to model the data)



$p(x|z_d)$ $\quad p(z_d|z_h)$ $\quad p(z_h|z_p)$ $\quad p(z_p|\theta)$

$$p(x|\theta) = \int \mathrm{d}z_d \mathrm{d}z_h \mathrm{d}z_p \quad p(x|z_d,\boxed{\theta_x}) \quad p(z_d|z_h,\boxed{\theta_d}) \quad p(z_h|z_p,\boxed{\theta_h}) \quad p(z_p|\boxed{\theta_p},\boxed{\theta_{\mathrm{th}}})$$

nuisance parameters

core "theory" parameters of inferest (e.g. "Higgs Mass")

Slide credit: Lukas Heinrich

Critical Value Estimation is Difficult with Many NPs

To guarantee frequentist coverage by Neyman's inversion technique, we need to test null hypotheses

$$H_{0,\mu_0} : \mu = \mu_0 \quad \text{versus} \quad H_{1,\mu_0} : \mu \neq \mu_0 \quad \text{for } \mu_0 \in \mathcal{M}$$

by comparing test statistics to the cutoffs $\widehat{C}_{\mu_0} := \inf_{\nu \in \mathcal{N}} \widehat{C}_{(\mu_0,\nu)}$.

That is, one needs to control the type I error at each μ_0 for *all* possible values of the nuisance parameters.

Can lead to numerically unwieldy and costly computations if the number of nuisance parameters is large (>10 NPs).

Two Popular Approaches to Systematics

Hybrid Approaches to Critical Value Estimation

- h-ACORE: Hybrid Resampling or Profiling[1] of Nuisance Parameters
  - Compare ACORE test statistic with the *hybrid cut-off*

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\Lambda(\mathcal{D};\mu_0)\big|(\mu_0,\widehat{\nu}_{\mu_0})}\left(\alpha \;\big|\mu_0, \widehat{\nu}_{\mu_0}\right)$$

  where the quantile regression is based on a train sample $\mathcal{T}'$ generated at *fixed* $\widehat{\nu}_{\mu_0}$.

- h-BFF: Integration of Nuisance Parameters
  - Compare BFF test statistic with the *approximate cut-off*

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\tau(\mathcal{D};\mu_0)\big|\mu_0}\left(\alpha \mid \mu_0\right)$$

  where we draw the train sample $\mathcal{T}'$ from the entire parameter space $\Theta = \mathcal{M} \times \mathcal{N}$, but apply quantile regression *using μ only*

[1] Van der Vaart, 2000; Chuang & Lai, 2000; Feldman, 2000; Sen et al. 2009

Assessing Confidence Sets

- *"For small sample sizes, there is no theorem as to whether profiling or marginalization will give better frequentist coverage for the parameter of interest"* (Cousins 2018)
- Our LF2I diagnostic tool can
  - provide guidance as to which method to choose for the problem at hand, and
  - pinpoint regions of parameter space where inference may be unreliable, e.g., under/over-confident.

Assessing Confidence Sets

- *"For small sample sizes, there is no theorem as to whether profiling or marginalization will give better frequentist coverage for the parameter of interest"* (Cousins 2018)
- In general settings, our LF2I diagnostic tool can
  - provide guidance as to which method to choose for the problem at hand, and
  - pinpoint regions of parameter space where inference may be unreliable, e.g., under/over-confident.

Ex: Diagnostics for Classical "On-Off" Problem

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

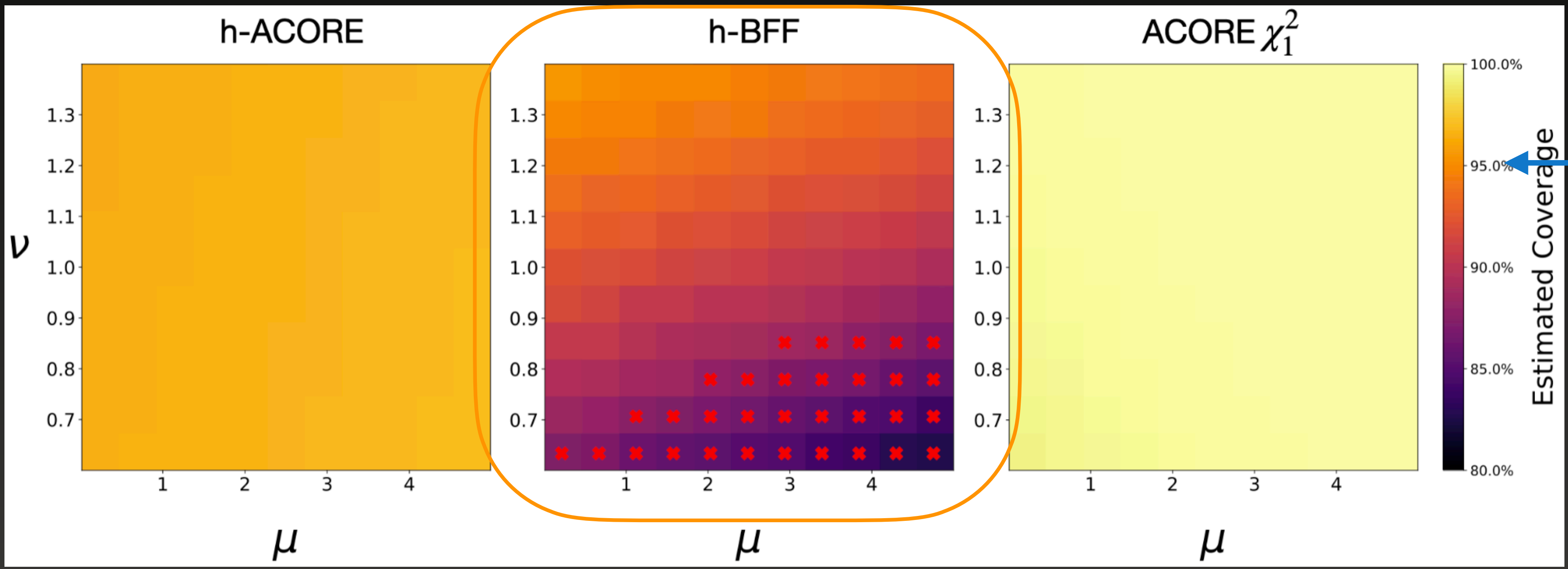
- Simultaneous measurements of two Poisson processes

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

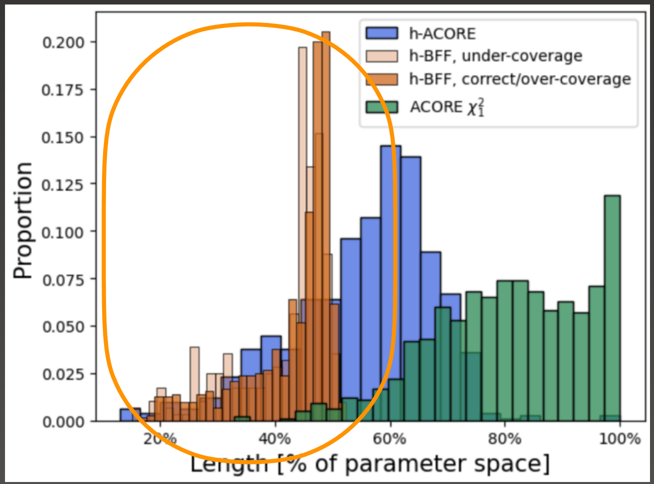
- N_B is the # of events in the background region (expected background count b)
- N_S is the # of events in the contaminated signal region (expected signal count s)
- Unknown parameters:
  - signal strength-POI (μ); scaling factor-NP (ν)
  - [L. Heinrich 2022] Set hyper-parameters at s=15, b=70, τ=1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

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

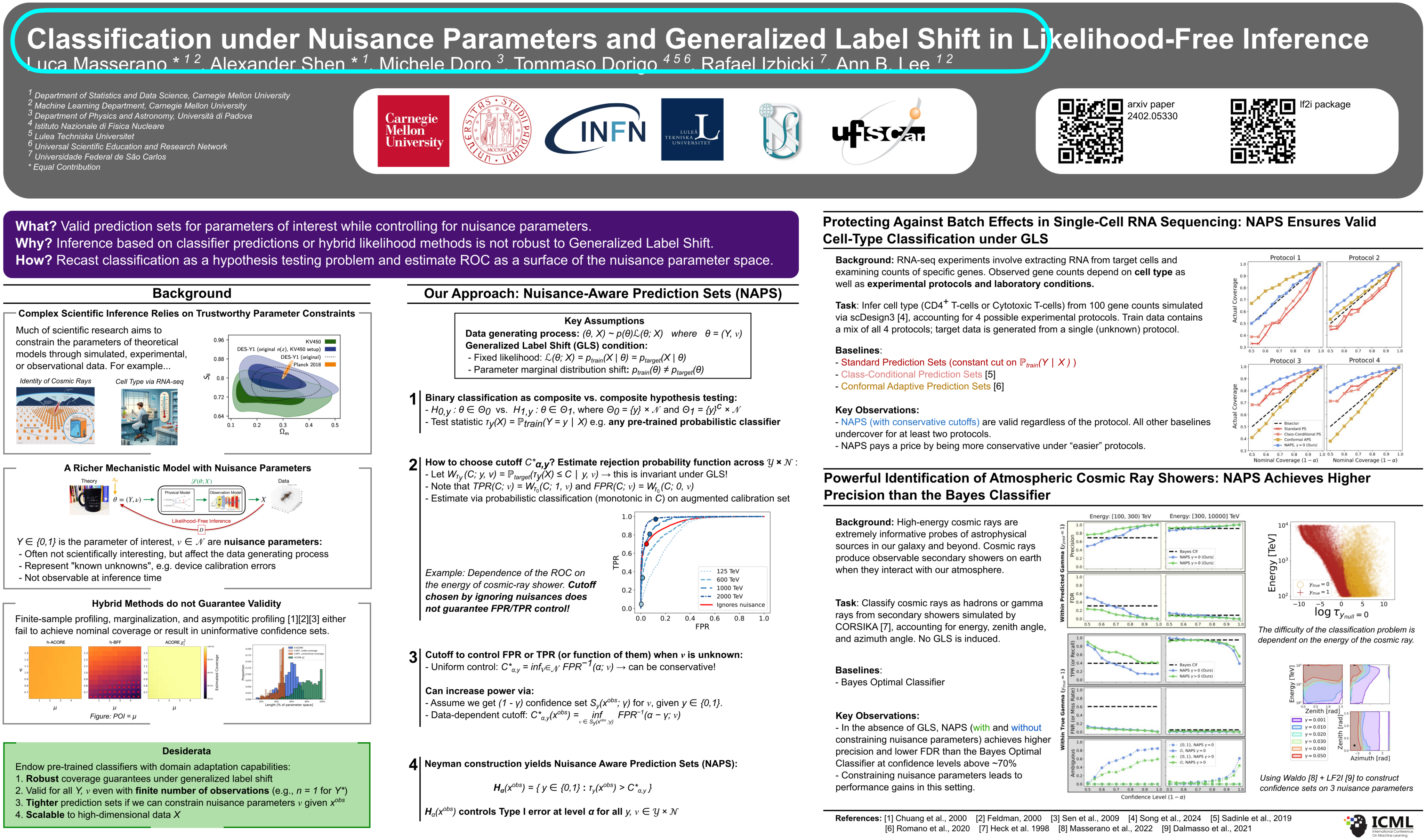
Left: LRT with profiling; Center: marginalization; Right: chi-square)



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

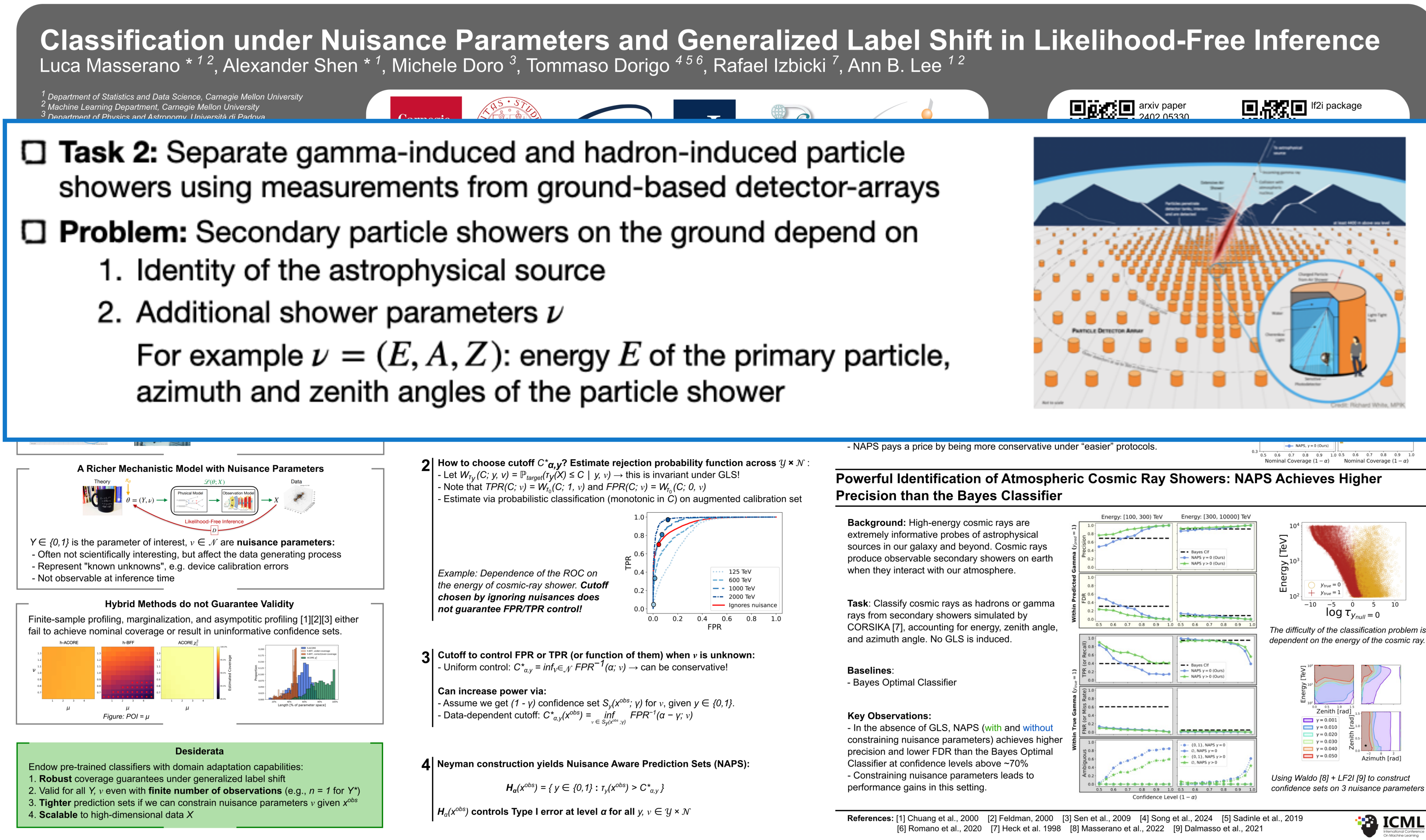


Finally, there are also nuisance-aware alternatives with coverage guarantees under shifting priors



arXiv:2402.05330 (ICML 2024)

Finally, there are also nuisance-aware alternatives with coverage guarantees under shifting priors



- **Task 2:** Separate gamma-induced and hadron-induced particle showers using measurements from ground-based detector-arrays
- **Problem:** Secondary particle showers on the ground depend on
  1. Identity of the astrophysical source
  2. Additional shower parameters ν

  For example $\nu = (E, A, Z)$: energy E of the primary particle, azimuth and zenith angles of the particle shower

arXiv:2402.05330 (ICML 2024)

Finally, there are also nuisance-aware alternatives with coverage guarantees under shifting priors

Classification under Nuisance Parameters and Generalized Label Shift in Likelihood-Free Inference

Luca Masserano * [1,2], Alexander Shen * [1], Michele Doro [3], Tommaso Dorigo [4,5,6], Rafael Izbicki [7], Ann B. Lee [1,2]

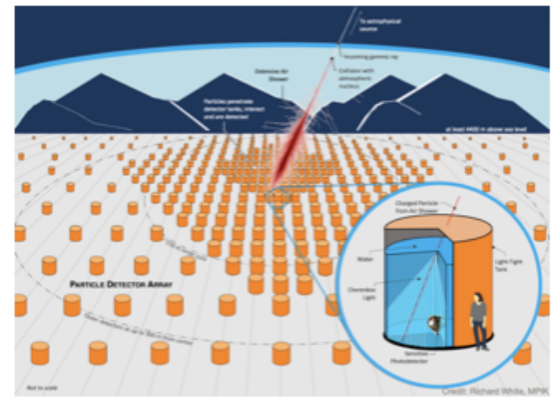
[1] Department of Statistics and Data Science, Carnegie Mellon University
[2] Machine Learning Department, Carnegie Mellon University
[3] Department of Physics and Astronomy, Università di Padova

arxiv paper 2402.05330

lf2i package

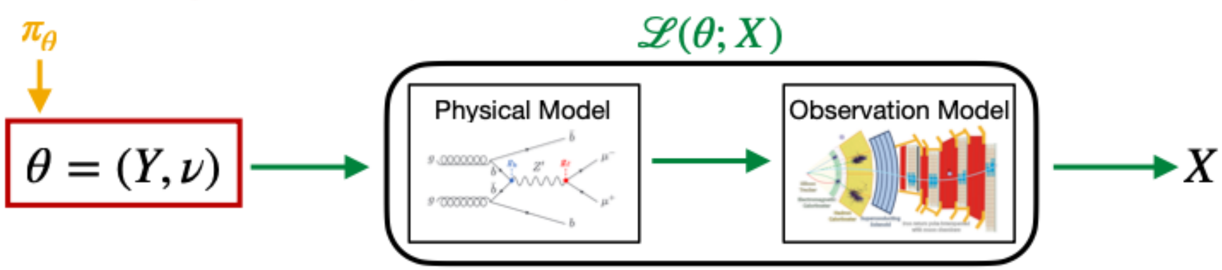
- ❑ **Task 2:** Separate gamma-induced and hadron-induced particle showers using measurements from ground-based detector-arrays
- ❑ **Problem:** Secondary particle showers on the ground depend on
    1. Identity of the astrophysical source
    2. Additional shower parameters ν

    For example $\nu = (E, A, Z)$: energy E of the primary particle, azimuth and zenith angles of the particle shower



- NAPS pays a price by being more conservative under "easier" protocols.

Masserano, Shen, Doro, Dorigo, Izbicki, Lee (ICML 2024)



- ❑ $Y \in \{0,1\}$ is the parameter of interest, $\nu \in \mathscr{N}$ are nuisance parameters not of direct interest
- ❑ **Nuisances are not observed** at the inference stage, but we can model their effect via simulations
- ❑ We assume $\mathscr{L}(\theta; X) = p(X \mid y, \nu)$ does not change but possibly $\pi_{train}(Y, \nu) \neq \pi_{target}(Y, \nu) \rightarrow$ GLS

arXiv:2402.05330 (ICML 2024)

Finally, there are also nuisance-aware alternatives with coverage guarantees under shifting priors

Classification under Nuisance Parameters and Generalized Label Shift in Likelihood-Free Inference

Luca Masserano * [1,2], Alexander Shen * [1], Michele Doro [3], Tommaso Dorigo [4,5,6], Rafael Izbicki [7], Ann B. Lee [1,2]

[1] Department of Statistics and Data Science, Carnegie Mellon University
[2] Machine Learning Department, Carnegie Mellon University
[3] Department of Physics and Astronomy, Università di Padova
[4] Istituto Nazionale di Fisica Nucleare
[5] Lulea Techniska Universitet
[6] Universal Scientific Education and Research Network
[7] Universidade Federal de São Carlos
* Equal Contribution

arxiv paper 2402.05330

lf2i package

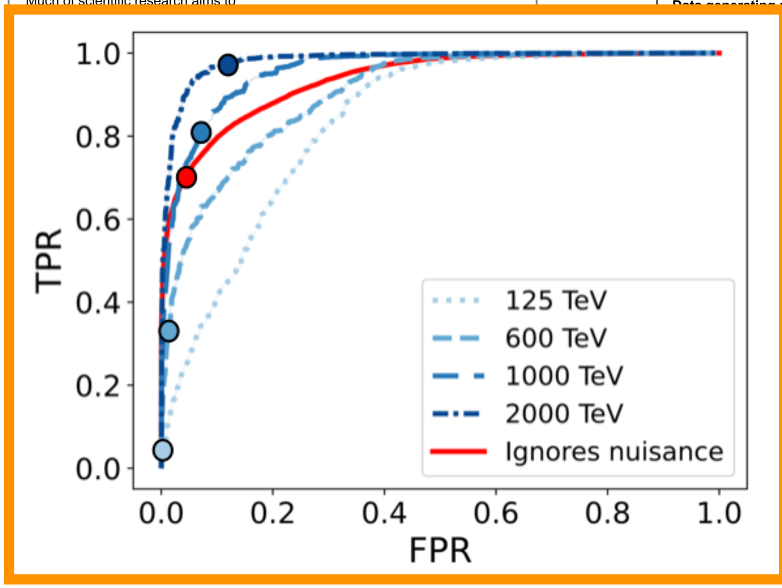
What? Valid prediction sets for parameters of interest while controlling for nuisance parameters.
Why? Inference based on classifier predictions or hybrid likelihood methods is not robust to Generalized Label Shift.
How? Recast classification as a hypothesis testing problem and estimate ROC as a surface of the nuisance parameter space.

Background

Complex Scientific Inference Relies on Trustworthy Parameter Constraints

Our Approach: Nuisance-Aware Prediction Sets (NAPS)

Key Assumptions



Estimate the ROC as a function of both the POI and NPs

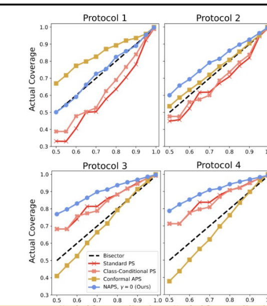
Protecting Against Batch Effects in Single-Cell RNA Sequencing: NAPS Ensures Valid Cell-Type Classification under GLS

Background: RNA-seq experiments involve extracting RNA from target cells and examining counts of specific genes. Observed gene counts depend on **cell type** as well as **experimental protocols and laboratory conditions.**

Task: Infer cell type (CD4+ T-cells or Cytotoxic T-cells) from 100 gene counts simulated via scDesign3 [4], accounting for 4 possible experimental protocols. Train data contains a mix of all 4 protocols; target data is generated from a single (unknown) protocol.

Baselines:
- Standard Prediction Sets (constant cut on $\mathbb{P}_{train}(Y \mid X)$)
- Class-Conditional Prediction Sets [5]
- Conformal Adaptive Prediction Sets [6]

Key Observations:
- NAPS (with conservative cutoffs) are valid regardless of the protocol. All other baselines undercover for at least two protocols.
- NAPS pays a price by being more conservative under "easier" protocols.



How do we increase power? Restrict search to subset of $\mathcal{N}$:

- Assume we get $(1-\gamma)$ confidence set $S_y(x^{obs};\gamma)$ for ν, given $y \in \{0,1\}$.
- Data-dependent cutoff: $C^\star_{\alpha,y}(x^{obs}) = \inf_{\nu \in S_y(x^{obs};\gamma)} FPR^{-1}(\alpha-\gamma;\nu)$

Theorem 1 (Nuisance-aware cutoffs for FPR control). Let $\alpha \in (0,1)$ and $\gamma \in [0,\alpha)$, and let $S_y(X;\gamma)$ be a valid confidence set for ν as in Definition 1. Define the nuisance-aware cutoff to be

$$C^\star_{\alpha,y}(X) = \inf_{\nu \in S(X;\gamma)} FPR^{-1}(\alpha-\gamma; y, \nu).$$

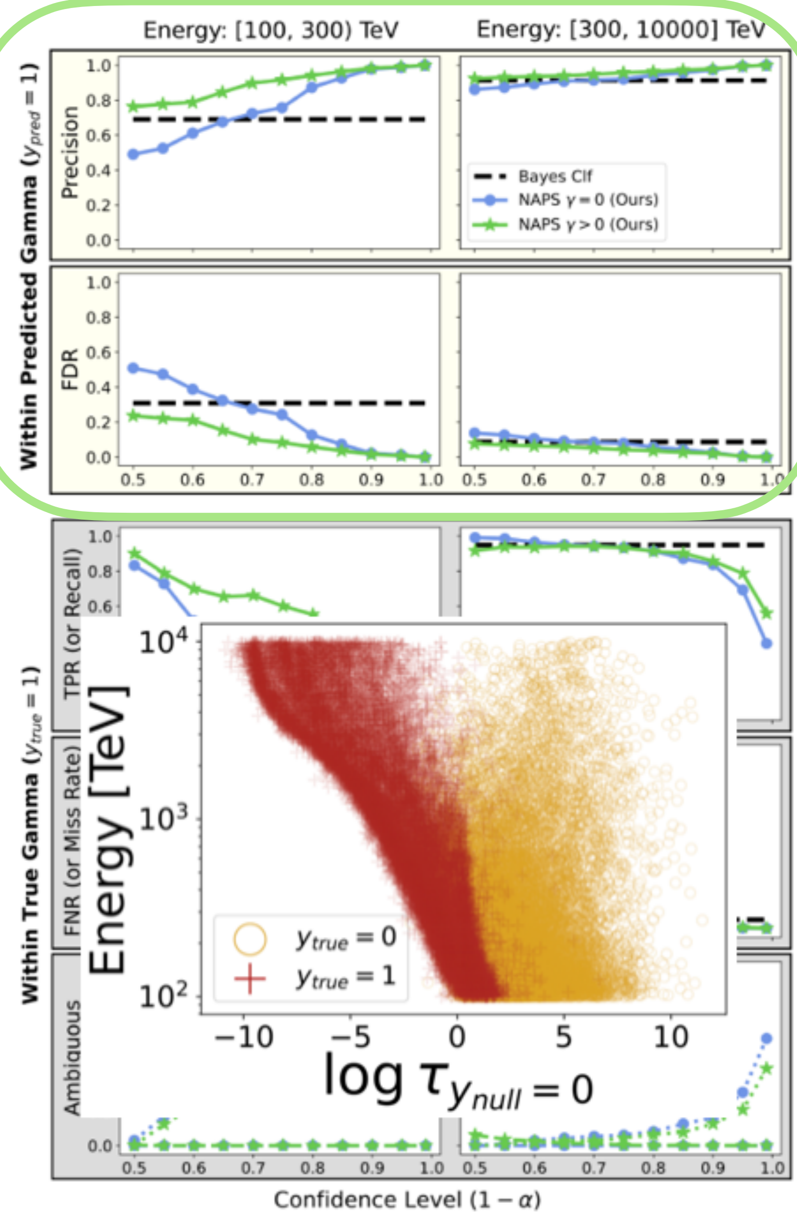
Then, for all $\nu \in \mathcal{N}$, we have FPR control at the desired level:

$$\mathbb{P}_{target}\left(\tau_y(X) \le C^\star_{\alpha,y}(X) \mid y, \nu\right) \le \alpha$$

Atmospheric Cosmic Ray Showers

- **Task:** Separate gamma-induced particle showers from hadron-induced showers using measurements from ground-based detector arrays
- Need to account for additional shower nuisance parameters: energy, azimuth angle, zenith angle
- Get confidence sets for ν using LF2I + Waldo (Masserano et al. 2022), then NAPS

- **Results:**
  - NAPS (with conservative cutoffs; blue) achieves good precision and low FDR at high confidence levels, but tends to be conservative at lower ones
  - NAPS (with data-dependent cutoffs; green) increases performance with uniformly better results
  - The set-valued classifier returns ambiguous prediction sets when it is uncertain on the output

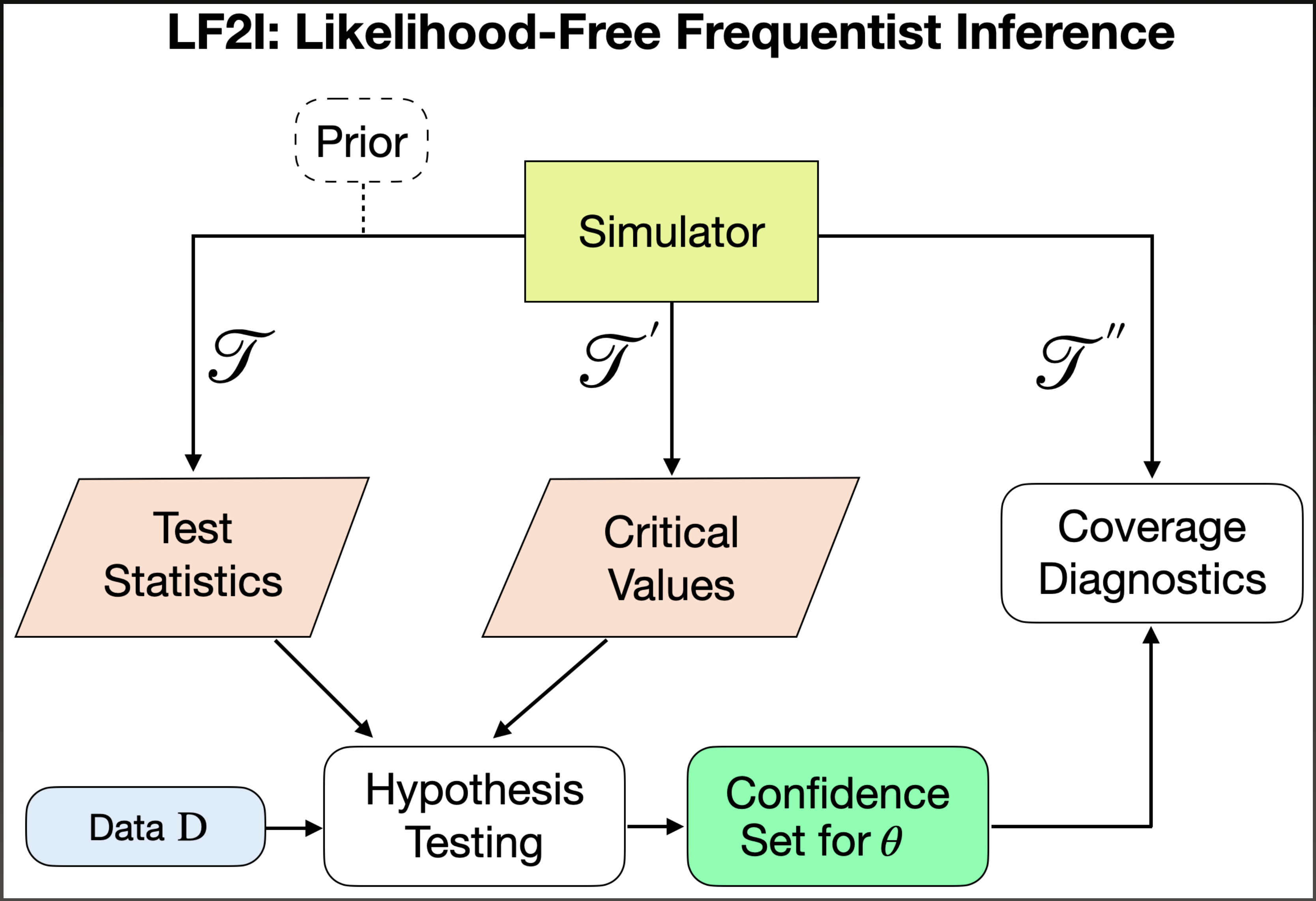


Take-Away: LF2I (inverse problem)

- Credible regions and prediction sets do not necessarily reflect where the true parameter is for inverse problems, esp for incompatible or shifting priors ("systematics")
- With LF2I we can construct confidence sets with robust coverage guarantees even for finite samples and shifting priors
- LF2I is fully modular: Plug in your favorite SBI results (for estimating likelihoods, LRs, posteriors, predictions, etc), calibrate and run diagnostics across the entire parameter space.

- LF2I is a fully modular and amortized framework

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



Take-Away: LF2I (inverse problem)

- Validity and Diagnostics: Any existing or new test statistic can be used to create valid confidence sets and run diagnostics.

- Prior Independence: LF2I guarantees (approximate) conditional coverage regardless of prior

- Power: Hardest to achieve in practice. Area where most statistical and computational advances will take place.

- ACORE (Approximate Computation via Odds Ratio Estimation):

$$\widehat{\Lambda}(\mathcal{D};\theta_0) := \log \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\sup_{\theta\in\Theta} \prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)}$$

- BFF (Bayesian Frequentist Factor):

$$\widehat{\tau}(\mathcal{D};\theta_0) := \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\int_{\Theta} \prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)\, d\pi_{\tau}(\theta)}.$$

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

Acknowledgments

- Nic Dalmasso (JP Morgan AI)
- Rafael Izbicki (UFSCar)

original LF2I framework

- Luca Masserano (CMU)

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

- Mikael Kuusela (CMU)
- Tommaso Dorigo (INFN/Padova)
- Alex Shen (CMU)

This work is funded in part by NSF DMS-2053804