Sixth IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis

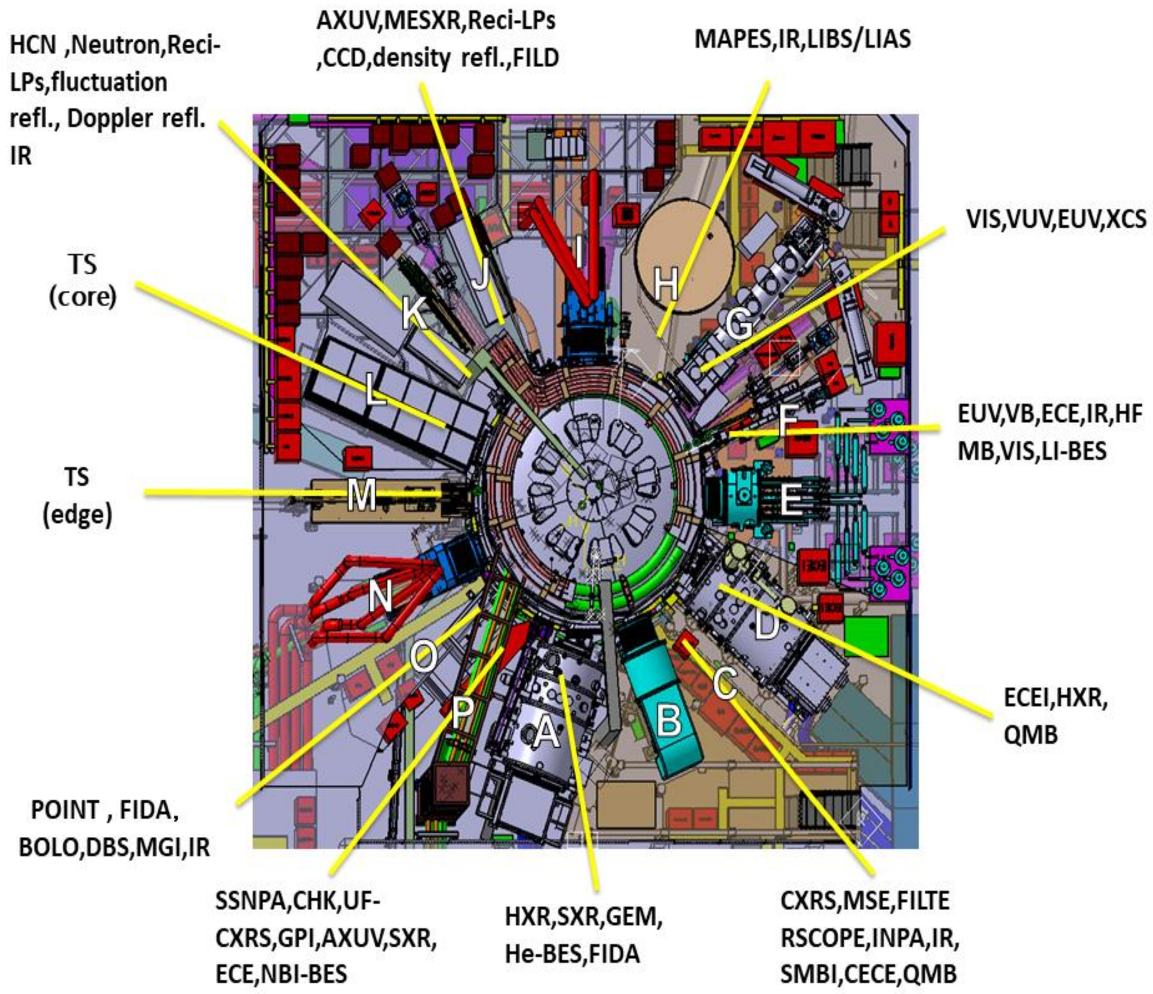
Data fusion and uncertainty quantification of density data on EAST tokamak

T. Lan[1], H. Q. Liu[1*], G. S. Li[1], S. X. Wang[1], L. F. Yang[1], J. S. Zhang[2], X. P. Xie[2], S. Z. Yuan[1], Y. Zhang[1], Q. L. Ren[1], W. X. Ding[2], Y. T. Song[1], and EAST Team.

1. Institute of Plasma Physics, Chinese Academy of Sciences, Hefei, Anhui 230031, China
2. School of Nuclear Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China

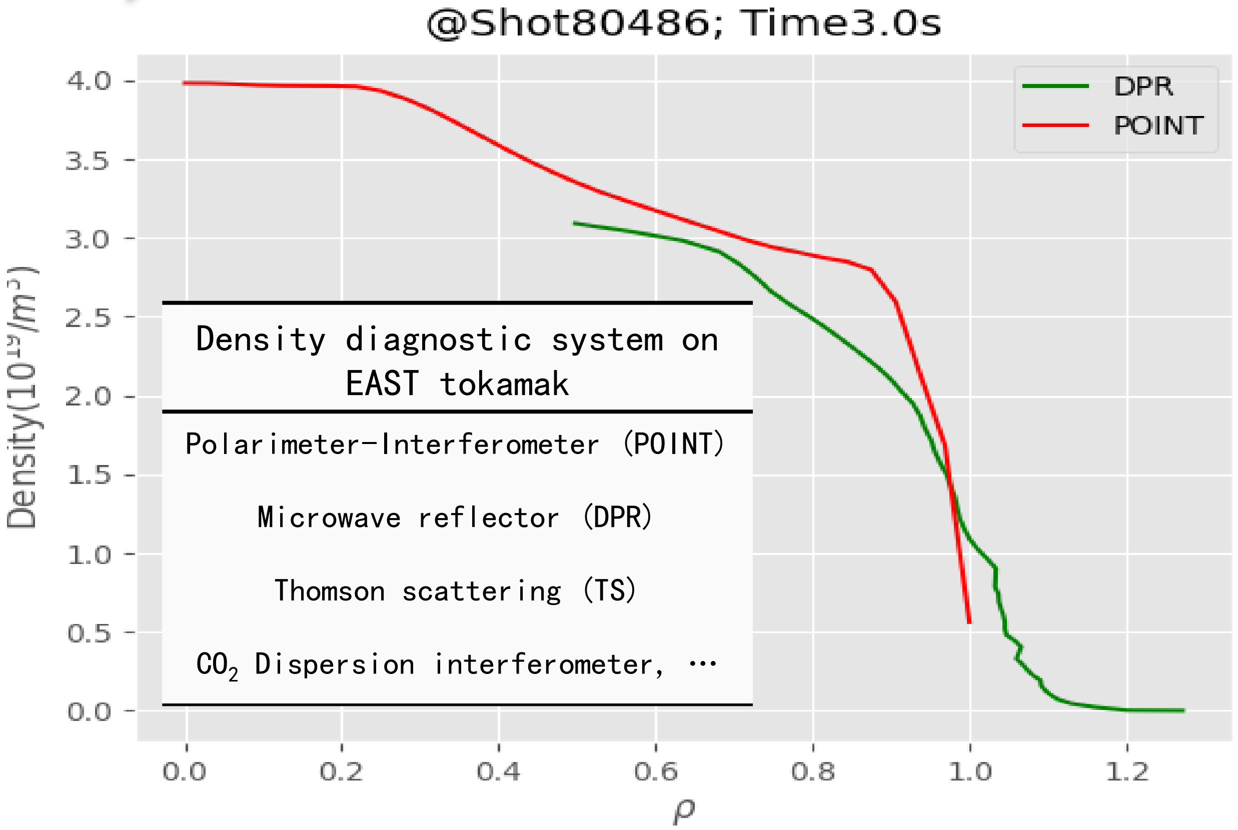
Motivation: Inconsistence between different diagnostic systems

Layout of diagnostic systems on EAST tokamak



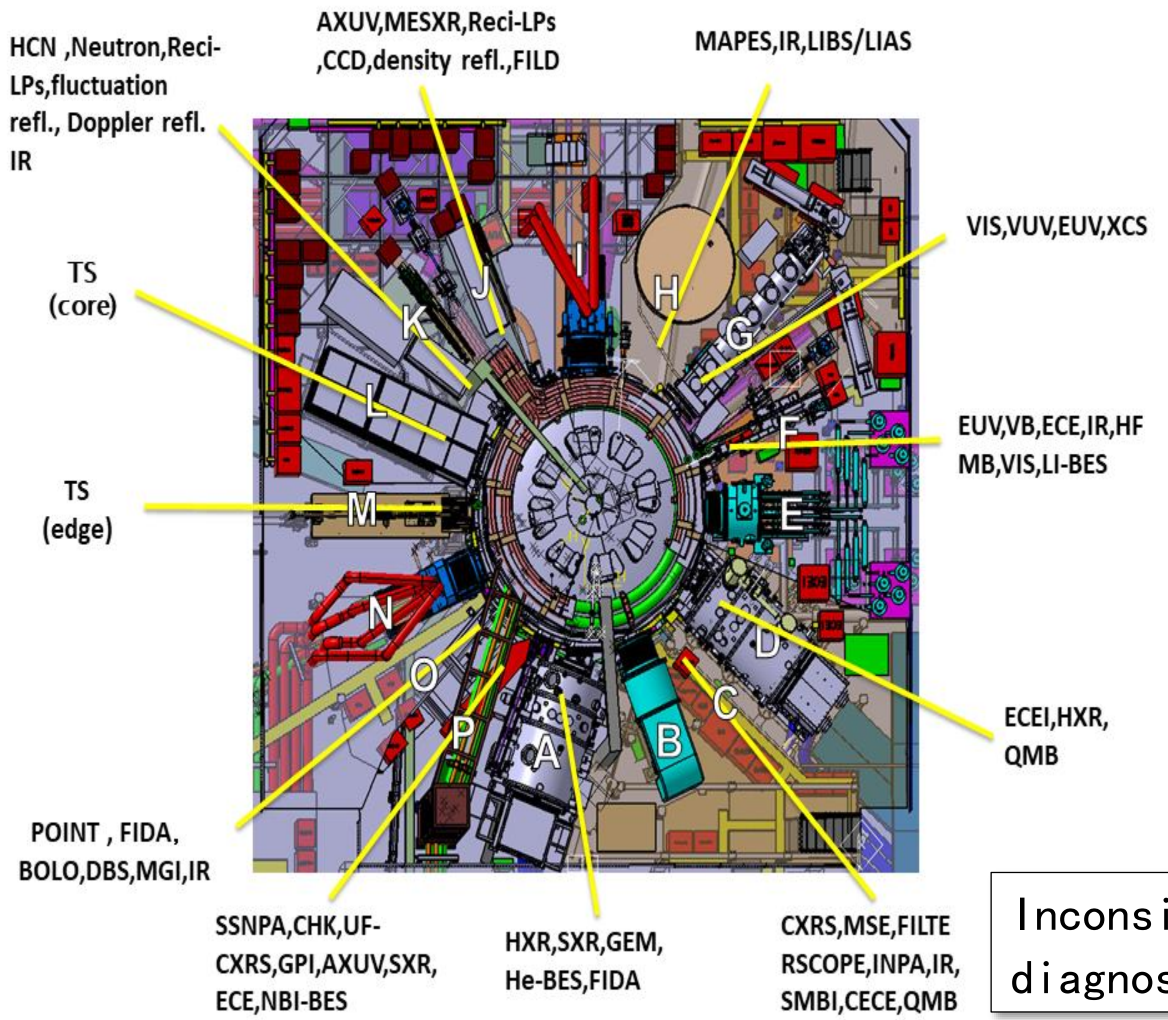
The same physical quantity measured by different systems are inconsistent

Make it difficult to use and understand data



Density diagnostic system on EAST tokamak
Polarimeter-Interferometer (POINT)
Microwave reflector (DPR)
Thomson scattering (TS)
CO_2 Dispersion interferometer, …

Measurement uncertainty leads to this inconsistence



- Environmental interference and instrument noise
- Statistical fluctuations of plasma
- Spatial sparsity of measurement channels
- Information loss during data acquisition and transmission
- Cognitive or methodological bias in data processing models

Measurement uncertainty → Inconsistence between diagnostic systems

Data fusion and uncertainty quantification

Accidental Uncertainty

- Environmental interference and instrument noise
- Statistical fluctuations of data

Methodological Uncertainty

- Spatial sparsity of measurement channels
- Information loss during measurement and transmission
- Cognitive or methodological bias in data processing models

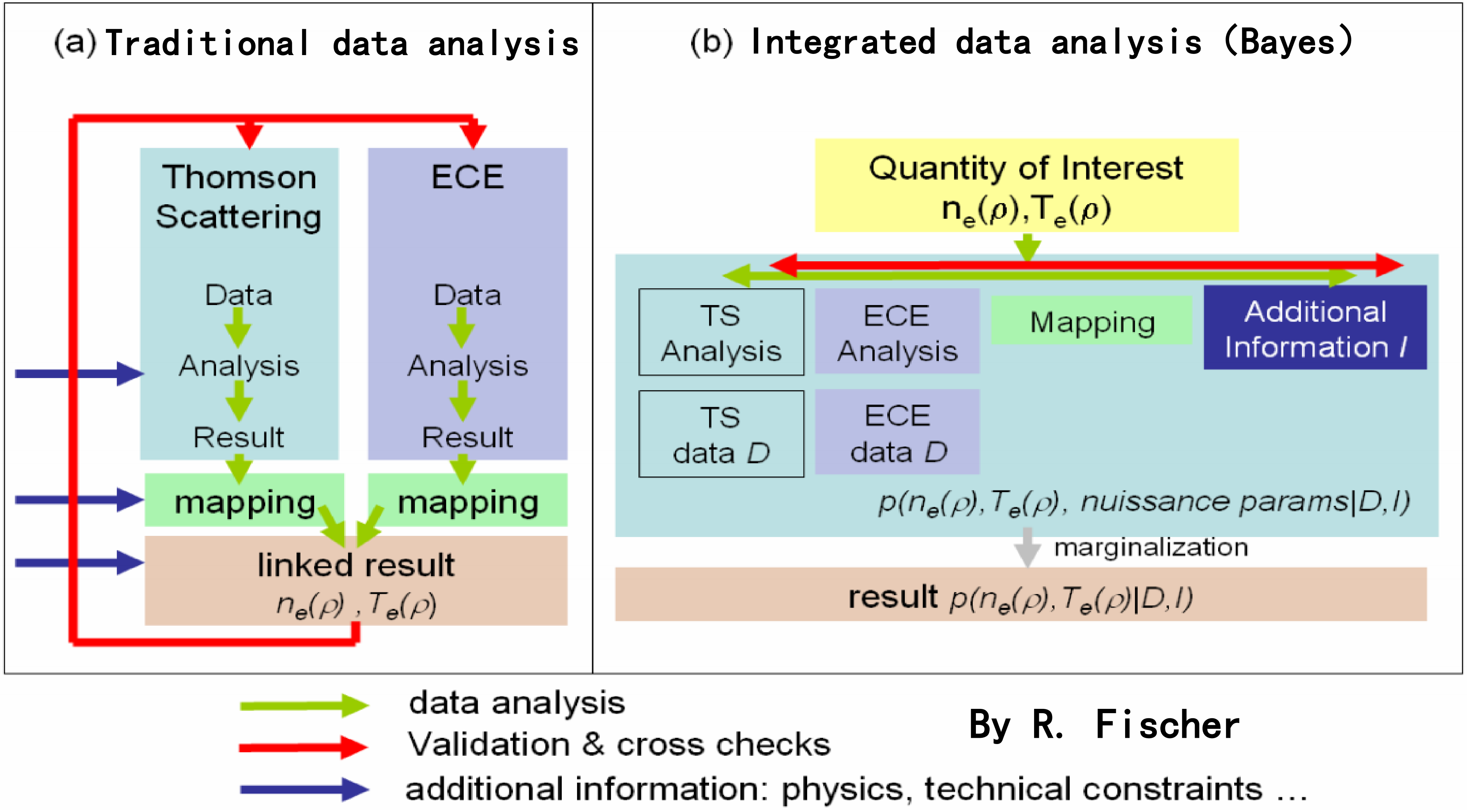
> **Data fusion**

Reduce Methodological Uncertainty

> **Uncertainty quantification (UQ)**

Reduce decision-making risks

Data fusion and UQ based on Bayesian inference



By R. Fischer

Bayes theorem

Thomas Bayes

British mathematician

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{D})}$$

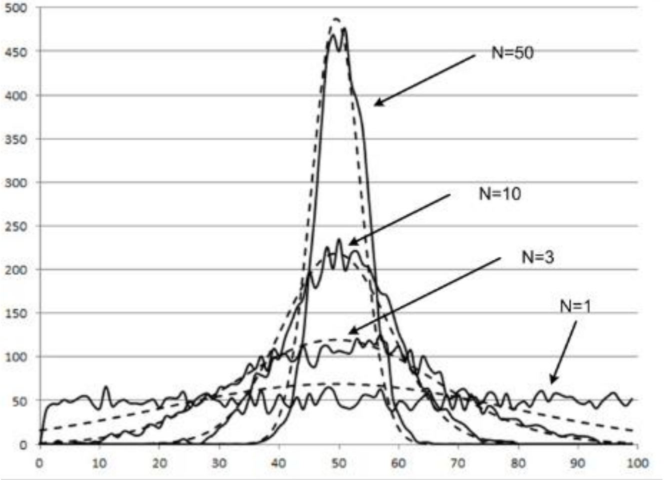
Posterior Probability

Likelihood Probability

Prior Probability

$$P(D|\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(D-f(\theta))^2}{2\sigma^2})$$

Central-limit theorem



ASIPP

Data fusion and UQ based on Bayesian inference

Plasma parameter
(Target quantity n_e)

Diagnostic Principles of Systems A, B, and C

Forward Diagnostic Model

Likelihood F_A

Likelihood F_B

Likelihood F_C

MCMC sampling

Joint probability distribution of target quantity n_e

Bayes formula

$$\mathrm{P}(n_e|D) = \frac{P(D|n_e)P(n_e)}{P(D)}$$

Data

$$\mathrm{P}(D|n_e) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(D-f(n_e))^2}{2\sigma^2})$$

Forward diagnostic model

$$P(n_e|D_{POINT}, D_{HCN}, D_{DPR}, D_{\mathrm{TS}}, \ldots.) \propto$$

$$P\left(D_{POINT-n_e}\middle|n_e\right)$$

$$\times P(D_{HCN}|n_e)$$

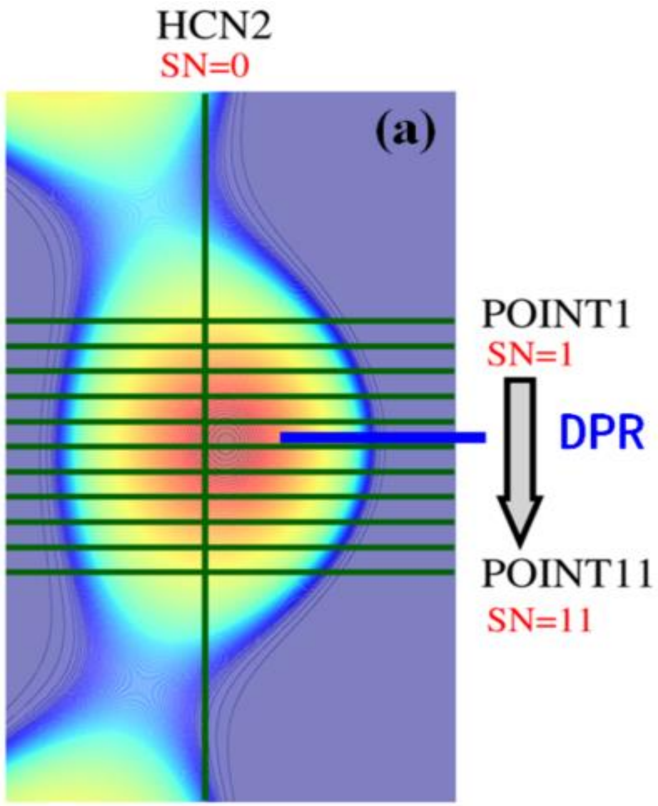
$$\times P(D_{DPR}|n_e)$$

$$\times P(D_{TS}|n_e, T_e)$$

$$\times \ldots\ldots$$

Construction of forward diagnostic model

Bayesian-based fusion of density data measured by POINT, HCN, and DPR



Basic Parameters of Diagnostic System	
POINT	Z=[-0.425, -0.34, -0.255, -0.17, -0.085, 0, 0.085, 0.17, 0.255, 0.34, 0.425] m; λ=432 μm.
HCN	R=[1.64, 1.82, 1.91]; λ=337 μm.
DPR	Z=0.03 m; f=32~110 GHz; Right-hand X-wave.

POINT

$$\begin{cases} \phi = \dfrac{\phi_L + \phi_R}{2} = C_1 \lambda \int n_e dL, \\ \alpha_F = \dfrac{\phi_L - \phi_R}{2} = C_2 \lambda^2 \int n_e B_\| dL. \end{cases}$$

DPR

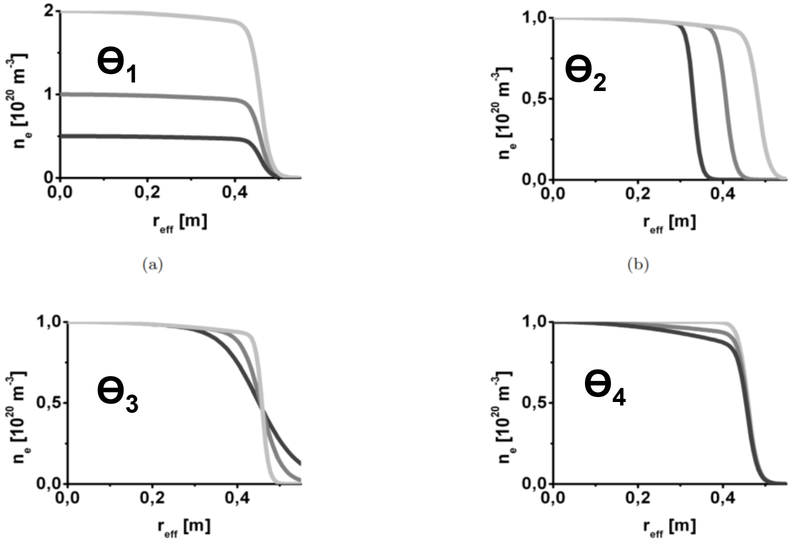
$$\omega = (\omega_{ce}^2/4 + \omega_{pe}^2)^{1/2} \pm \omega_{ce}/2.$$

$$\tau_p = \tau_{plamsma} - \tau_{wall} + \frac{2\Delta r_{vacuum}}{c}$$

Construction of forward diagnostic model

EFIT Magnetic Surface + Simulated Plasma Parameter Distribution + Diagnostic Principles

$$n_e(\mathrm{r}_{eff}) = \theta_1 \cdot 10^{20} m^{-3} \cdot \left[\frac{1-\theta_4 \cdot (\mathrm{r}_{eff}^2 / \mathrm{a}^2)}{1+(\mathrm{r}_{eff}^2/(\theta_2 \cdot a)^2)^{\theta_3}} \right]$$



(a) (b) (c) (d)

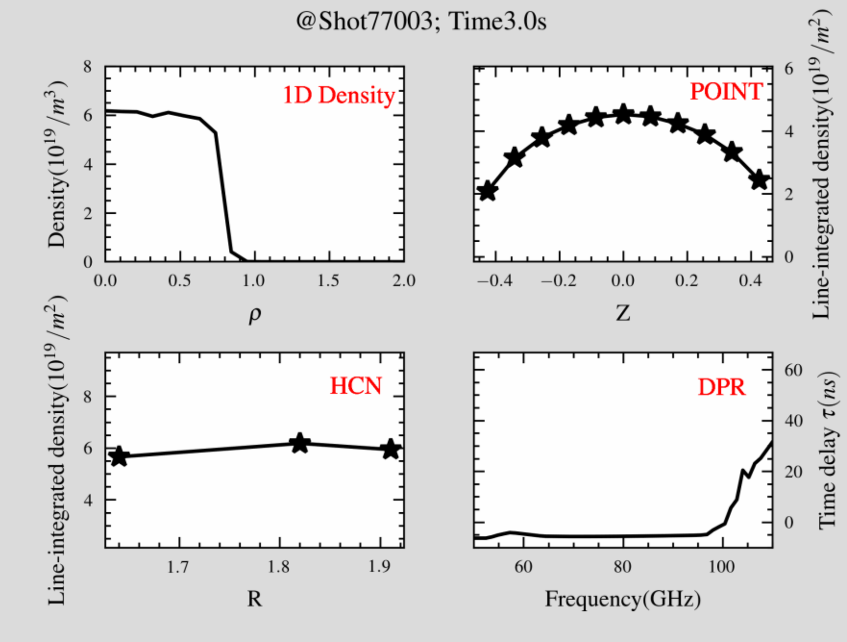
Θ_1: maximum of platform

Θ_2: position of the edge

Θ_3: decay of the edge

Θ_4: platform inclination

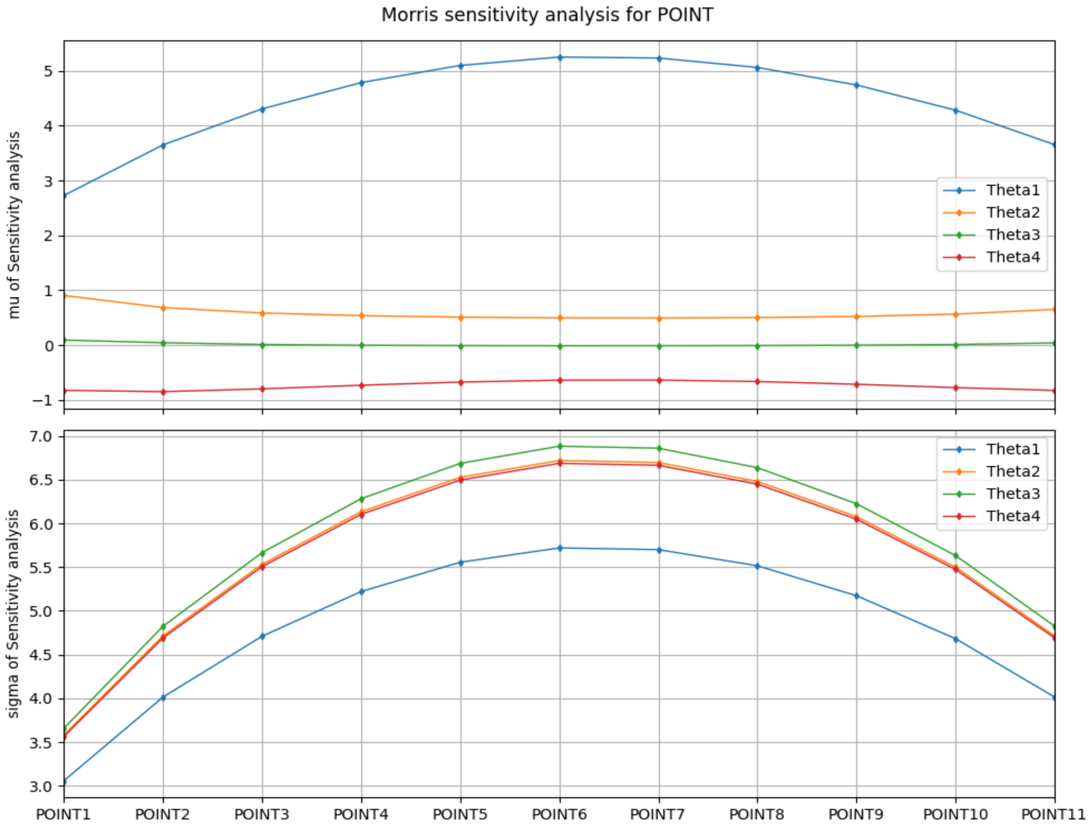
Simulated signal of POINT、HCN、DPR

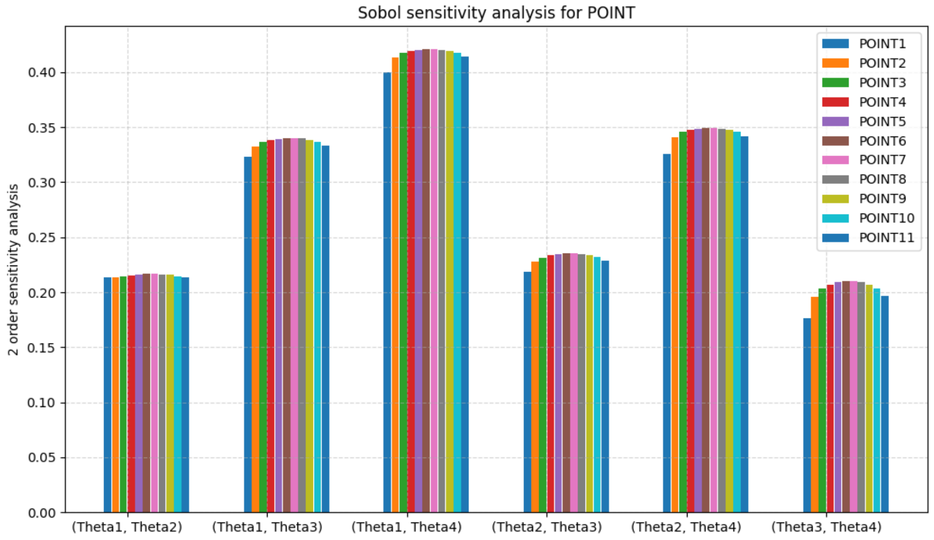


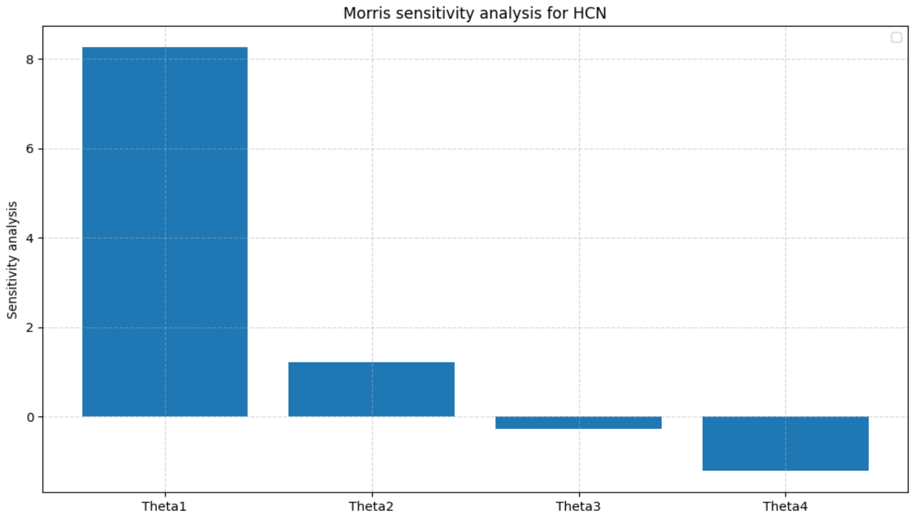
Dreier H

Sensitivity analysis

Θ_1: maximum of platform Θ_2: position of the edge
Θ_3: decay of the edge Θ_4: platform inclination







Bayesian-based fusion of simulated data from POINT, HCN, and DPR

$$n_e(r_{eff}) = \theta_1 \cdot 10^{20} m^{-3} \cdot \left[\frac{1-\theta_4 \cdot (r_{eff}^2 / a^2)}{1+(r_{eff}^2/(\theta_2 \cdot a)^2)^{\theta_3}} \right]$$

θ_1: maximum of platform

θ_2: position of the gradient

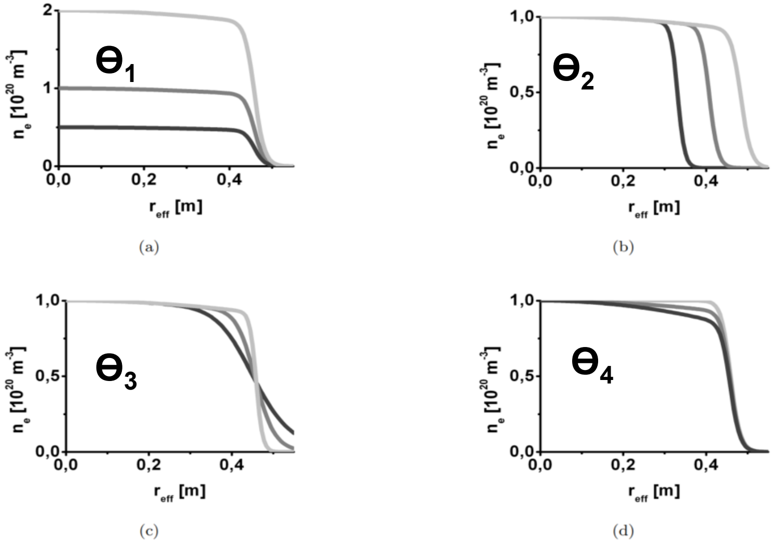
θ_3: decay at the edge

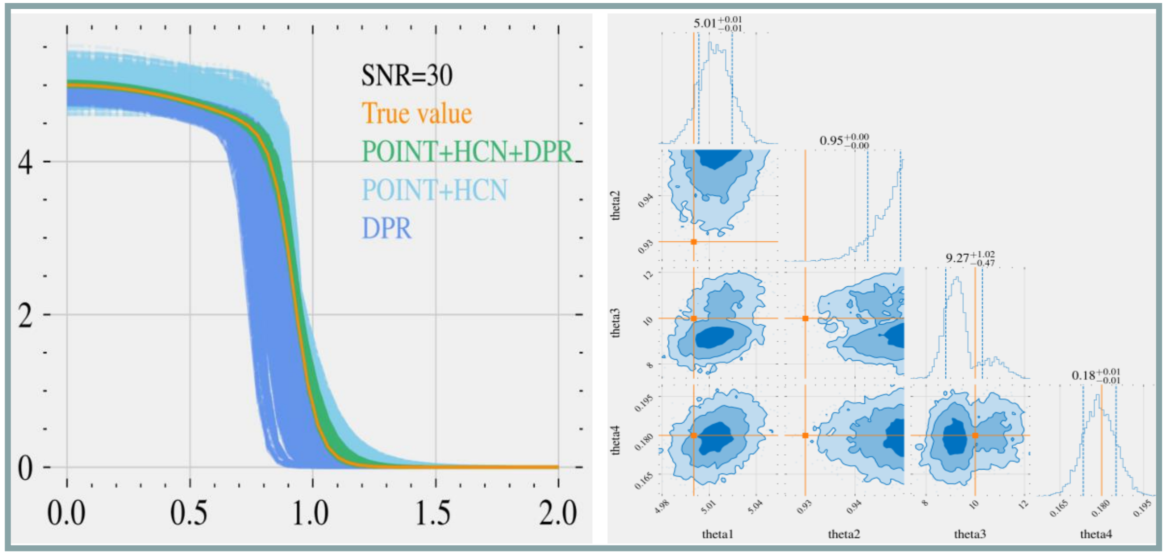
θ_4: platform inclination

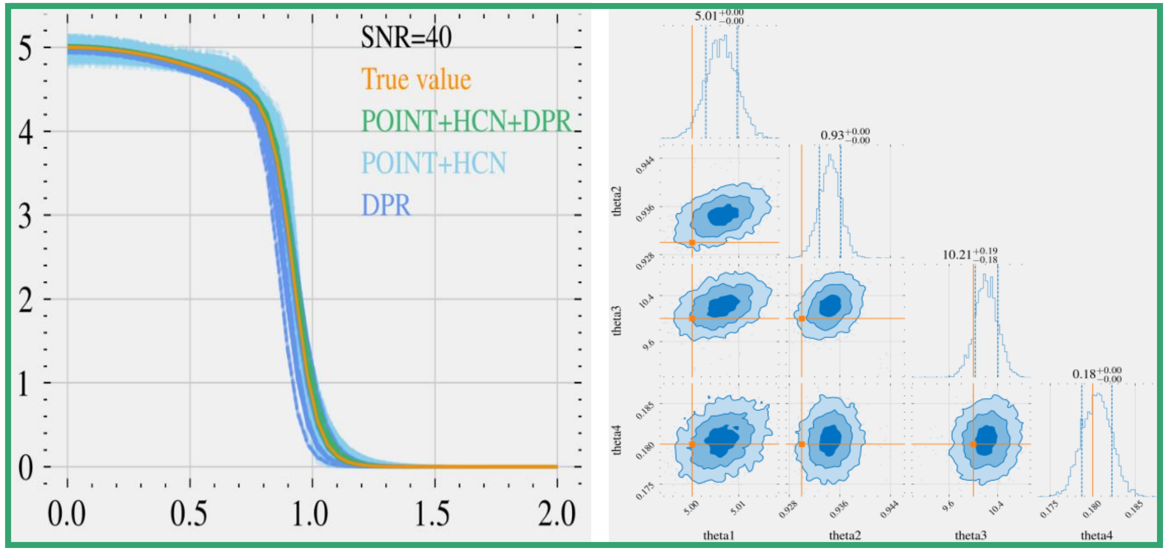
True value:

θ_1=5 θ_2=0.93

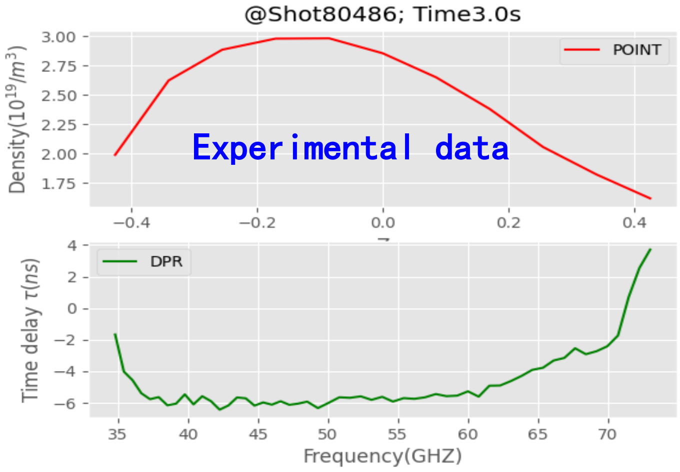
θ_3=10 θ_4=0.18

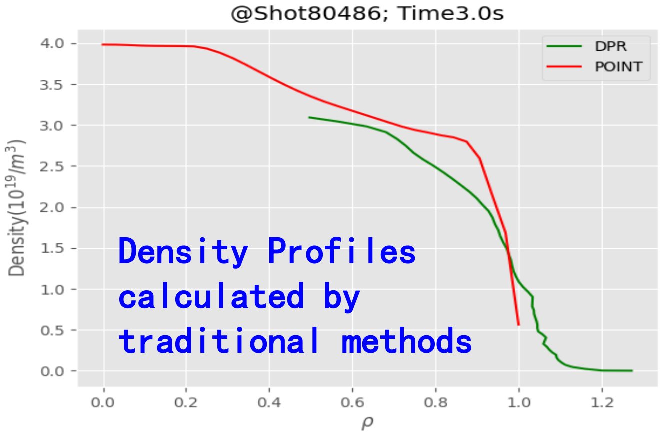




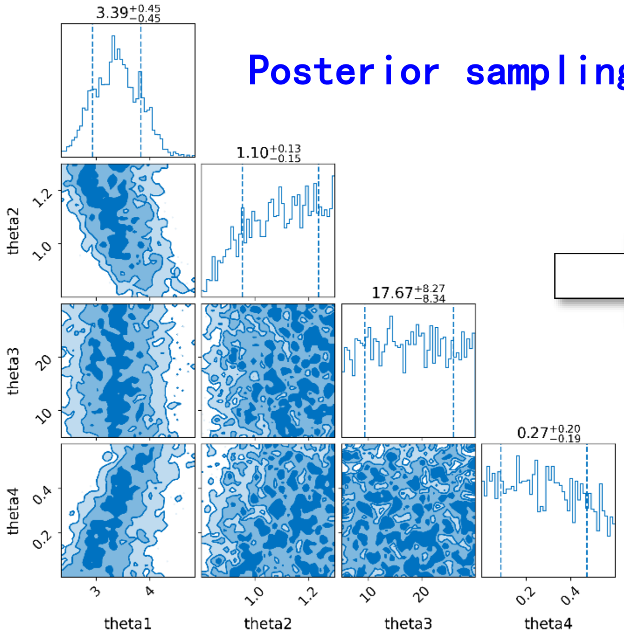


Bayesian-based inversion for experimental data of POINT system

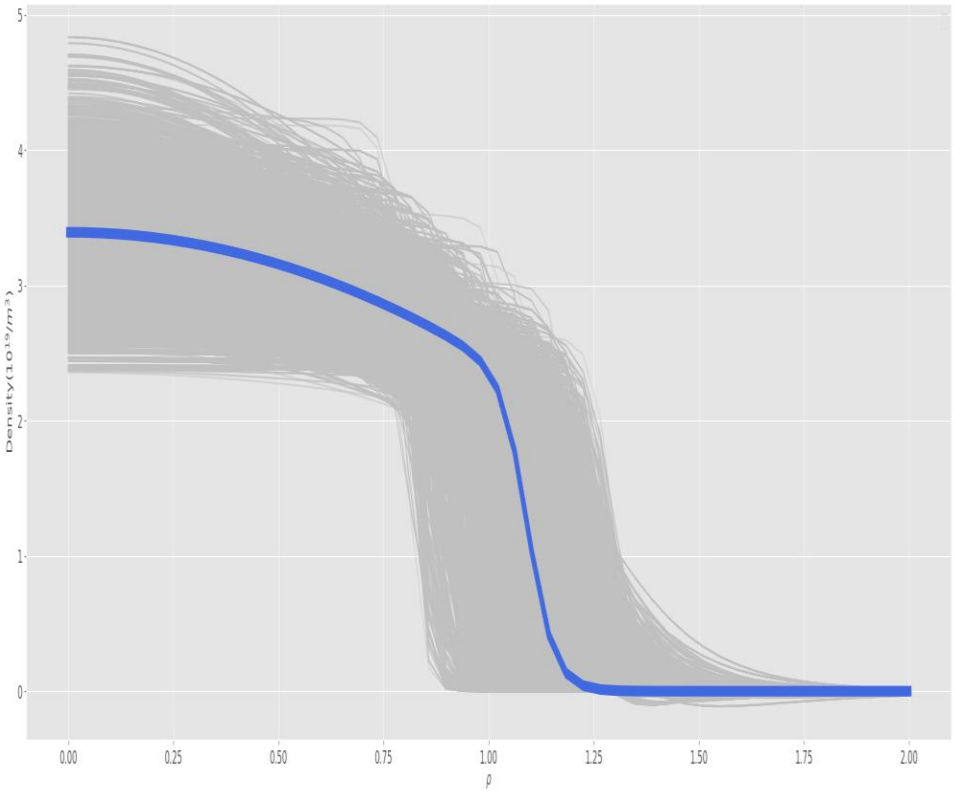




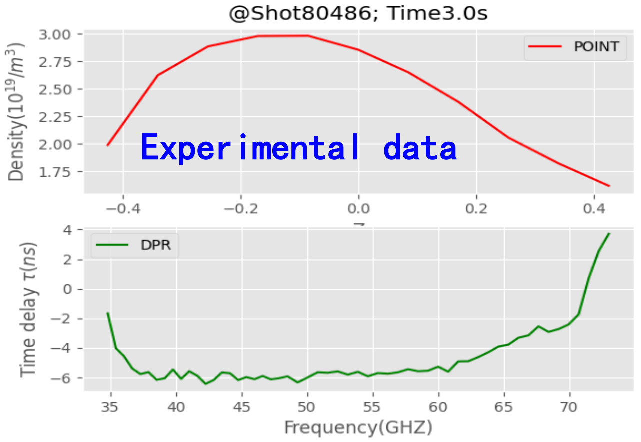
Posterior sampling

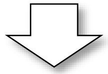


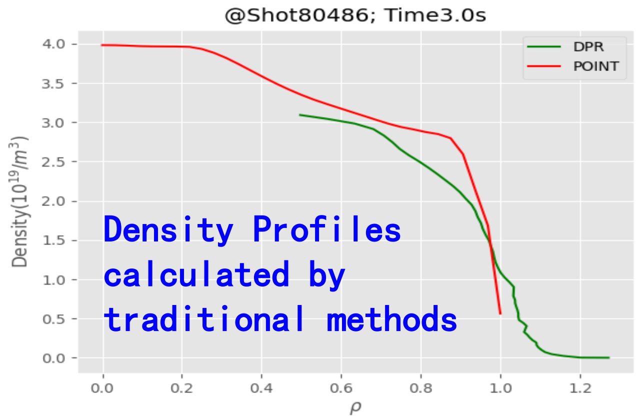
Bayesian-based inversion for POINT system

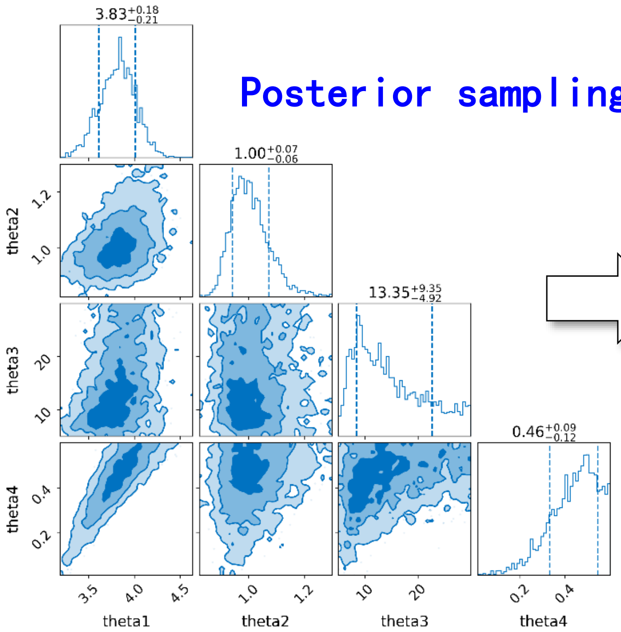


Bayesian-based fusion of experimental data from POINT and DPR

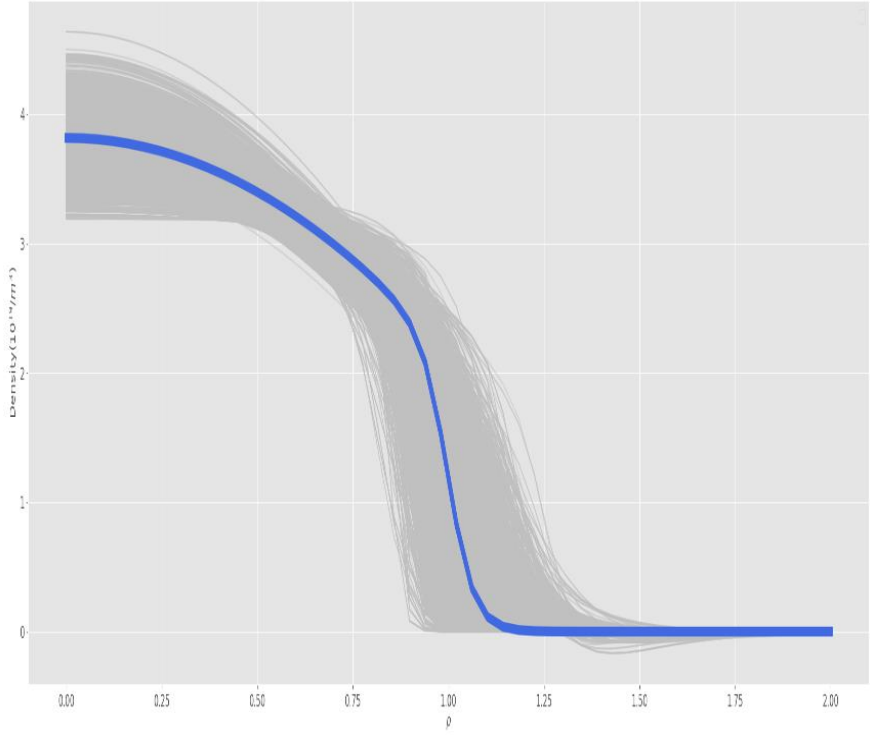








Bayesian-based fusion of density data from POINT and DPR



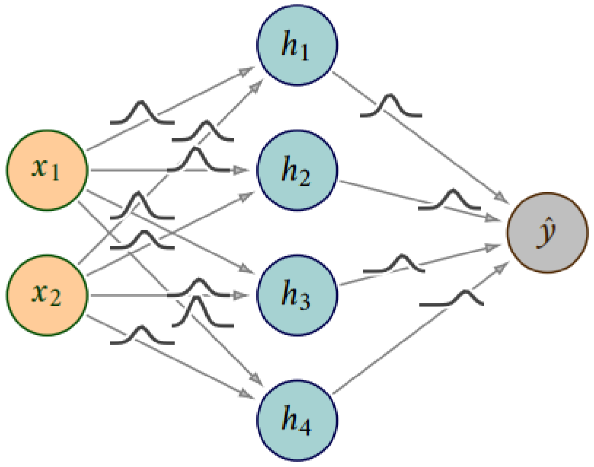
Data fusion and UQ based on deep learning

Bayesian	MC	(1) No need to change the model training process, (2) Low training complexity, (3) Easy to implement.	(1) Not very reliable for OoD data, (2) Needs multiple samplings during inference.
	MCMC	(1) Computationally more intensive compared to VI, (2) Asymptotically guarantees of producing exact samples.	(1) Very slow, (2) Fail to find poor convergence, (3) High MC error.
	VI	(1) Very fast (faster than MCMC), (2) Benefiting from stochastic optimization methods, (3) Suited to big datasets.	(1) Heavily depend on the starting point, (2) Very complicated calculations.
	BAL	(1) Able to learn from small amounts of data, (2) Able to add samples with high classification uncertainty to training.	(1) Lack of scalability to high-dimensional data, (2) Difficult to quantify loss functions.
	BBB	(1) Returning the posterior over the weights, (2) Allowing more complicated prior distributions.	(1) Requiring extra sweep over KL trade-off coefficients, (2) More parameters to train (approximately two times).
	VAE	(1) Easy to optimize its loss, (2) Mapping an input sample in the original data to latent factors.	(1) Collapse in latent space, (2) Difficult to interpret the code, (3) Low quality of the generated sample images.
Ensemble	DE	(1) Robust prediction, (2) Can be considered as base learners, (3) Limiting the dispensable sensitivity of particular training data, (4) Robust uncertainty estimates.	(1) More resource consuming, (2) Time consuming, (3) Weak performance on smaller problems.
	DEB or BDE	(1) Can perform batter than DEs in OoD settings, (2) Emulating the analytic posterior predictive.	(1) Weaker than standard DEs in not detrimental confident predictions, (2) Lazy learning procedure.

Uncertainty representation:

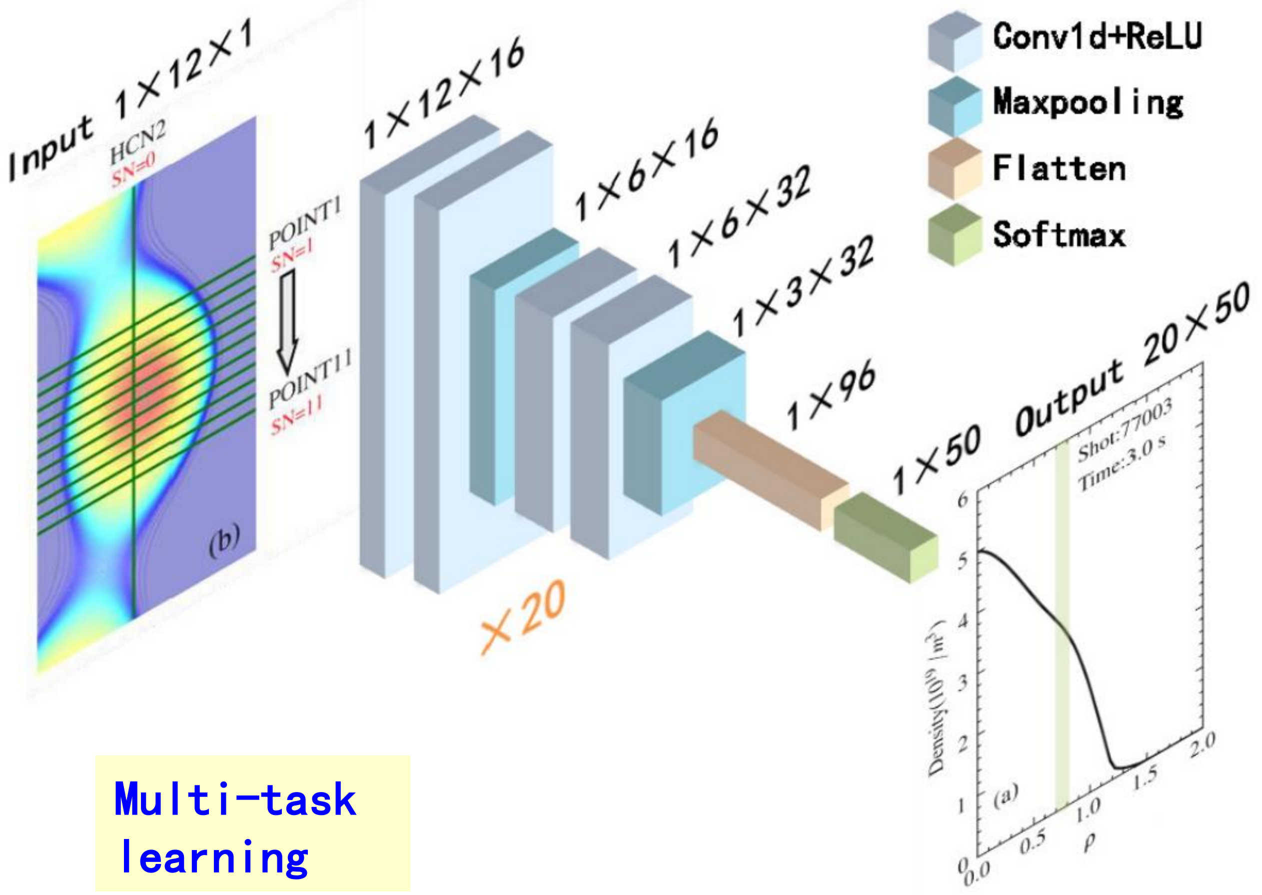
- Bayesian neural networks
- Deep ensemble network

Bayesian Neural Network



Information Fusion 76 243-297.

CNN-based fusion of density data from POINT and HCN



Multi-task learning

Advantage 1:

➢Not dependent on EFIT input

Trainable parameters: 243240

Optimizer: Adam

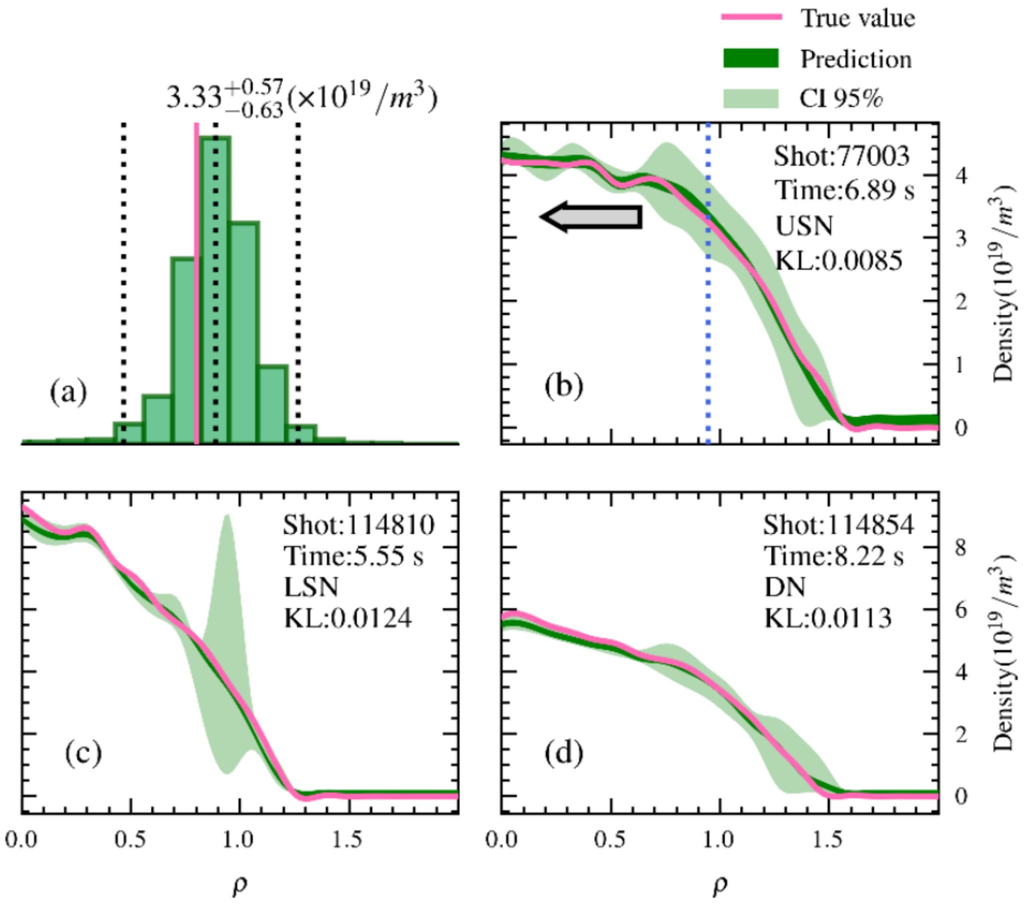
Loss function: MSE

Training set ratio: 0.8

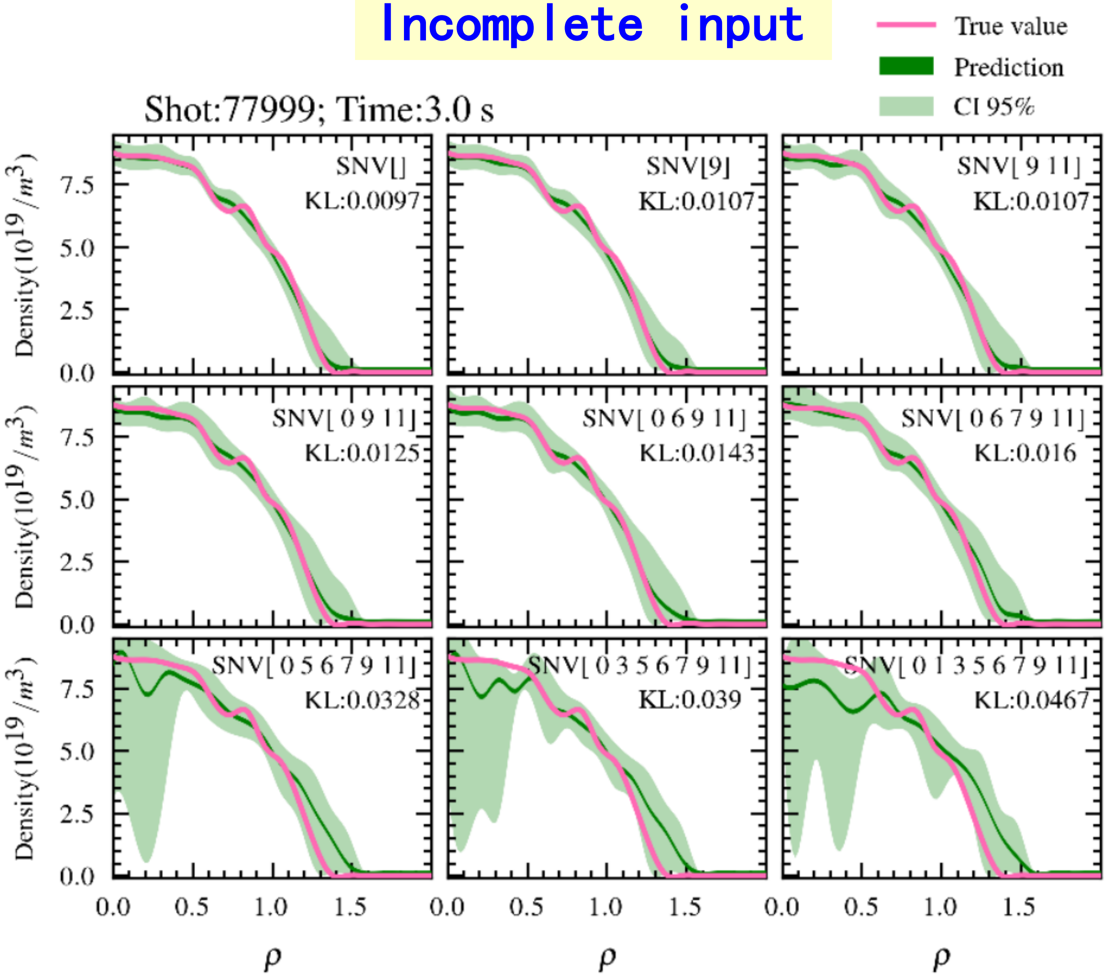
Server: Tesla V100 32GB GPUs

CNN-based fusion of density data from POINT and HCN

Complete input

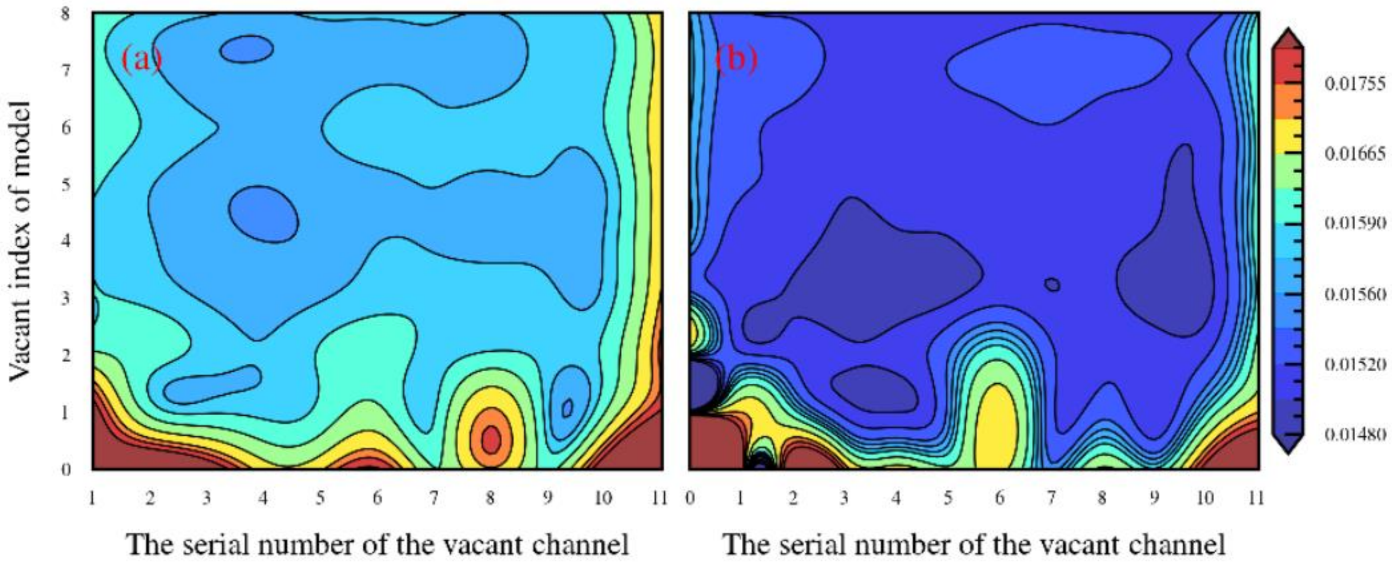


Incomplete input

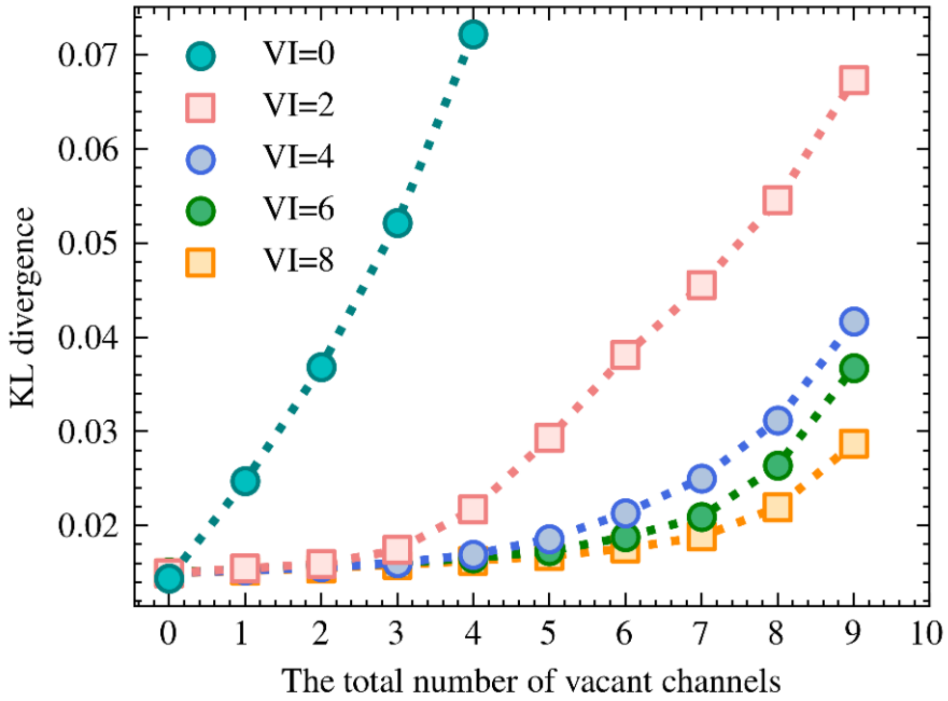


CNN-based fusion of density data from POINT and HCN

Missing 1 channel



Missing multiple channels



Advantage 2:

➢Robust to Incomplete input

Comparison of fusion methods

Bayesian inference

Advantage:

- Describes quantities and uncertainties in a probabilistic manner
- Provide an intuitive data fusion framework

Disadvantage:

- The priors significantly affect the posterior results
- Sampling process may take a relatively long time

VS

Deep learning

Advantage :

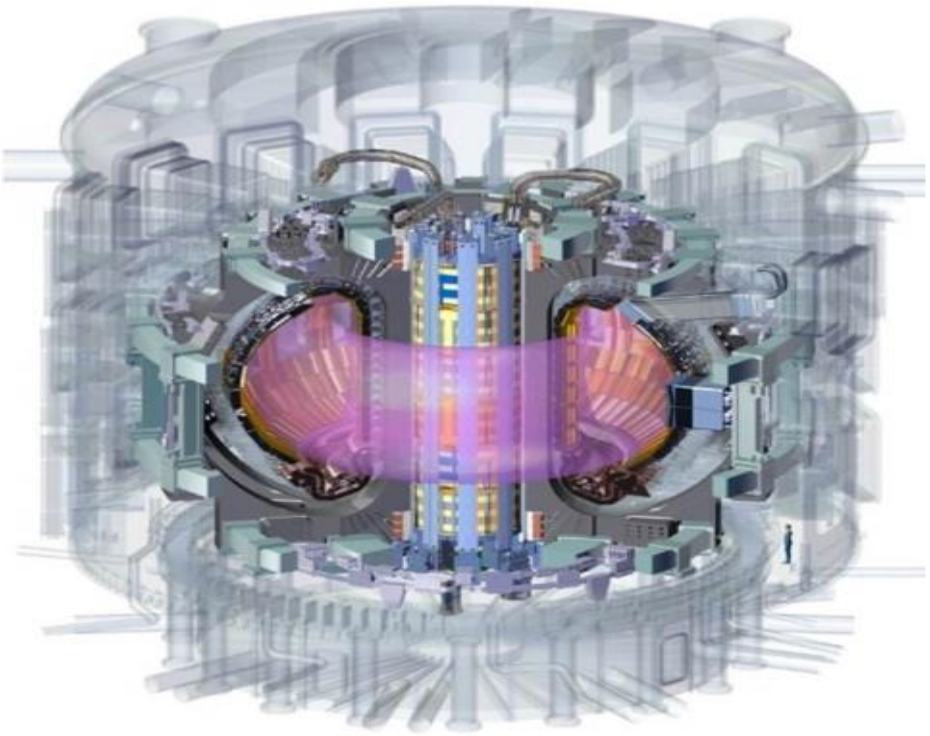
- Reduce method uncertainty brought by fusion model
- Fast, suitable for online data processing

Disadvantage:

- Model training is time consuming
- Require fusion strategy design for data with varying principles, attributes, and features

Data cleaning

- Data Cleaning: distinguish low-quality (low SNR) or useless data from high-quality (high SNR) or usable data to improve data utility and reliability.



Automatic data cleaning and missing data imputation

Complex measurement environment:

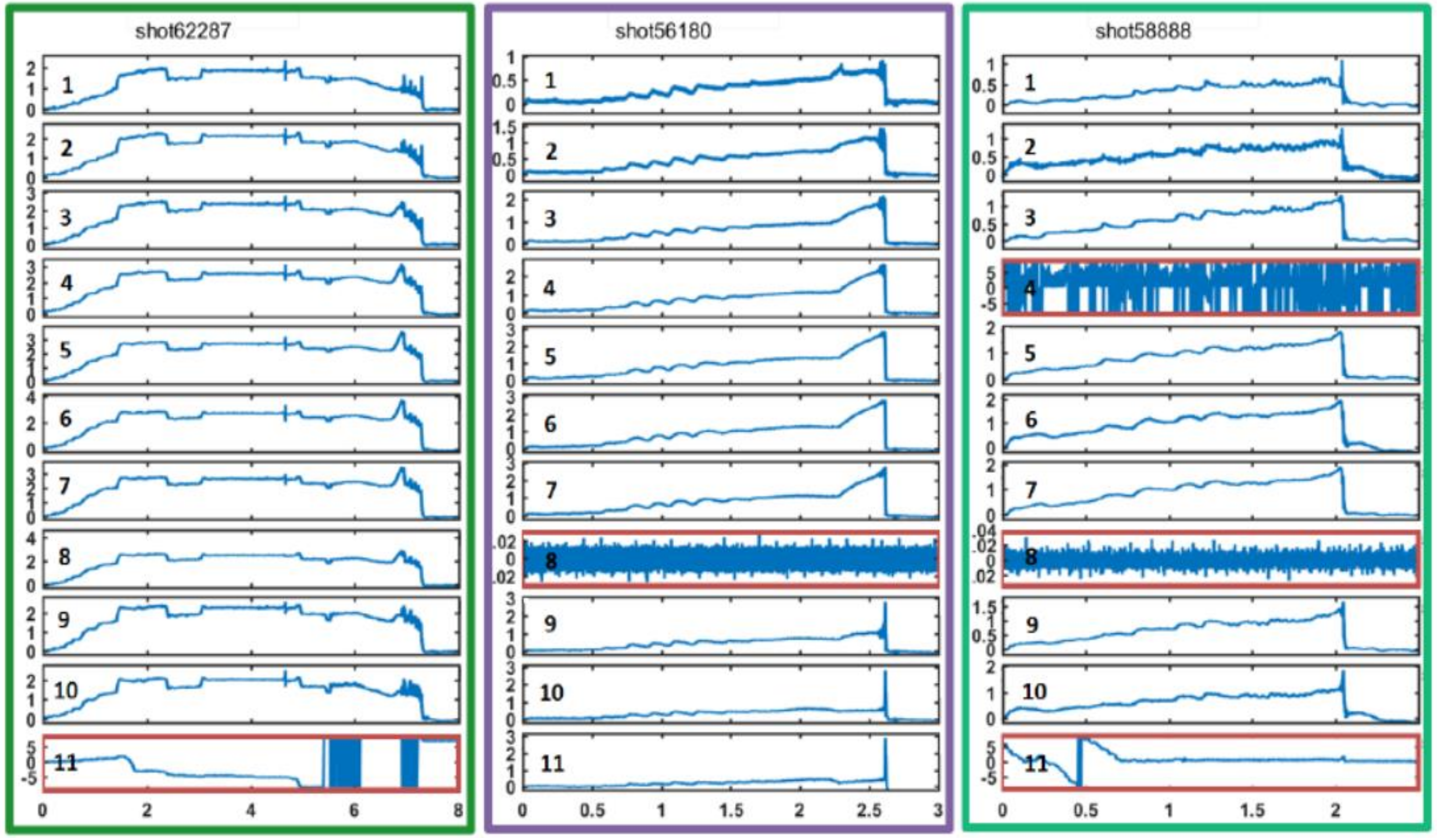
- Electromagnetic interference
- Mechanical vibration
- Neutron irradiation

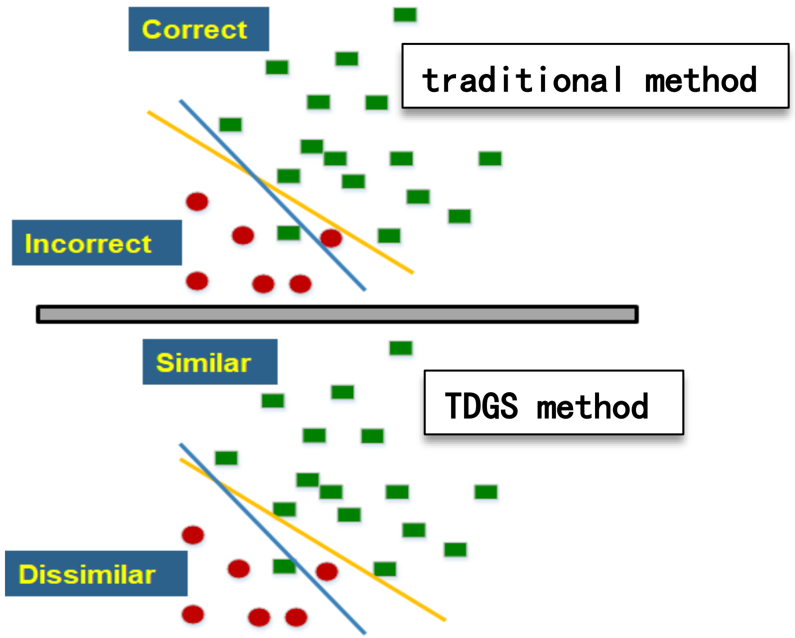
Common dirty data:

- Data missing
- Outlier
- Low signal-to-noise ratio
- Incorrect labeling
- Time displacement

Data cleaning

Line-integral density data measured by POINT system



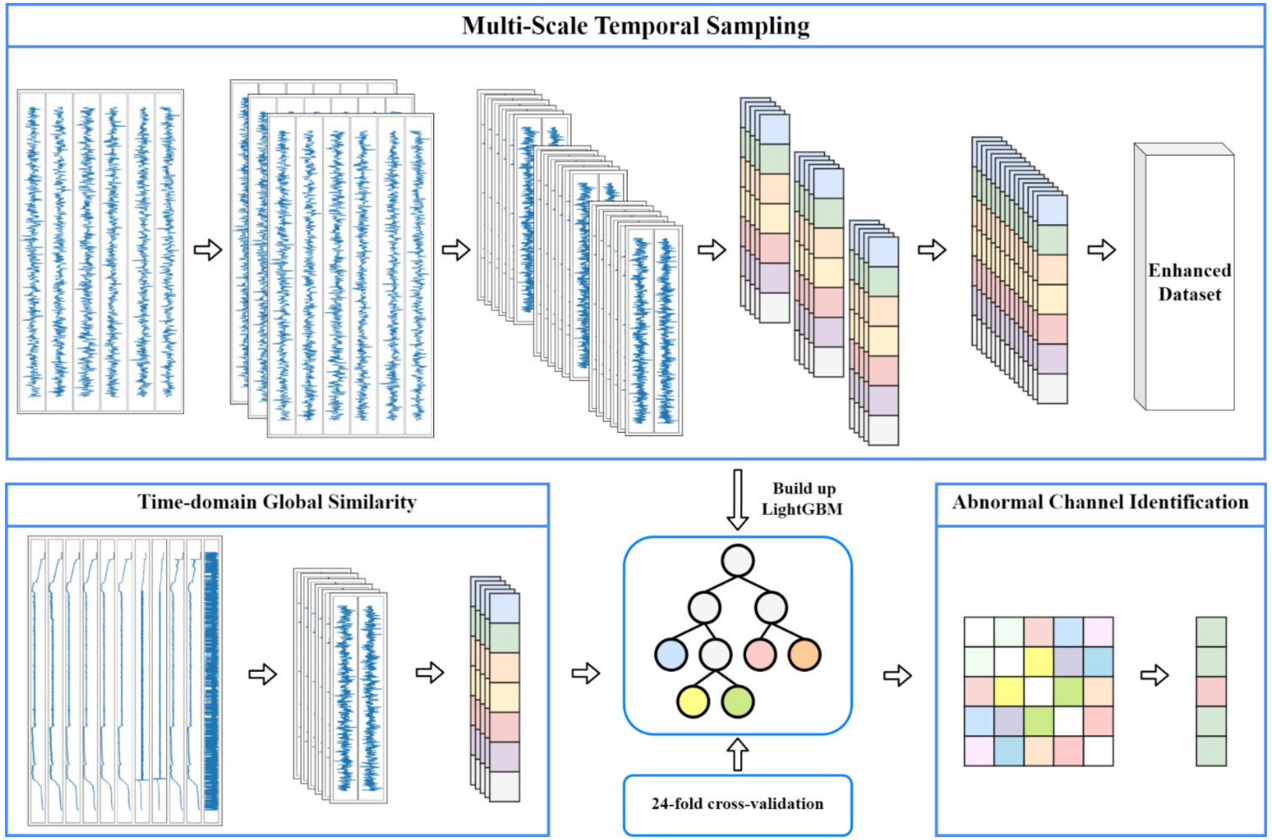


Kernel functions	Performance	Penalty factor	Window size of median filter
Linear	0.9871±0.0015	8	1000
Polynomial(3)	0.9712±0.0055	1	800
Polynomial(4)	0.9288 ±0.0139	5	2000
Rbf	0.9652±0.0036	1	1900

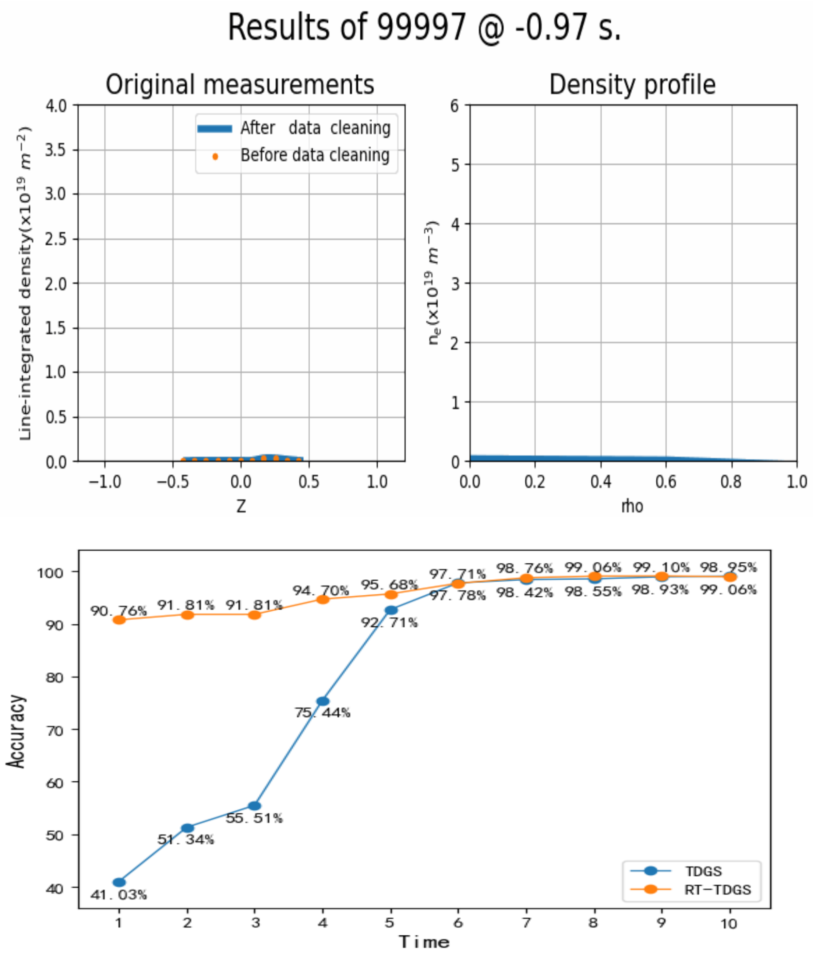
Input: pairs of data sequences; Output: similarity → correctness

Real-time data cleaning

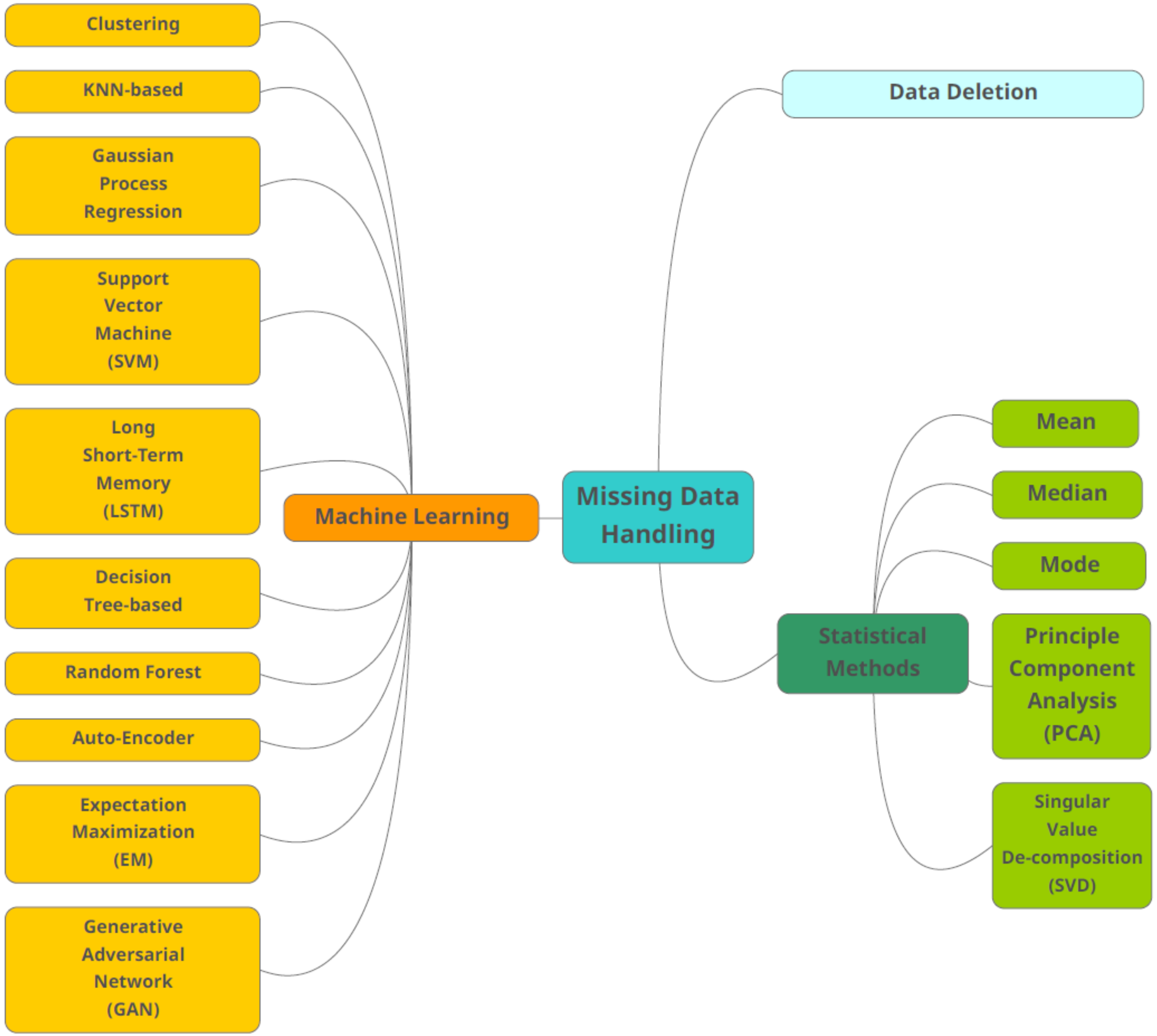
RT-TDGS (Real-Time Time-domain Global Similarity)

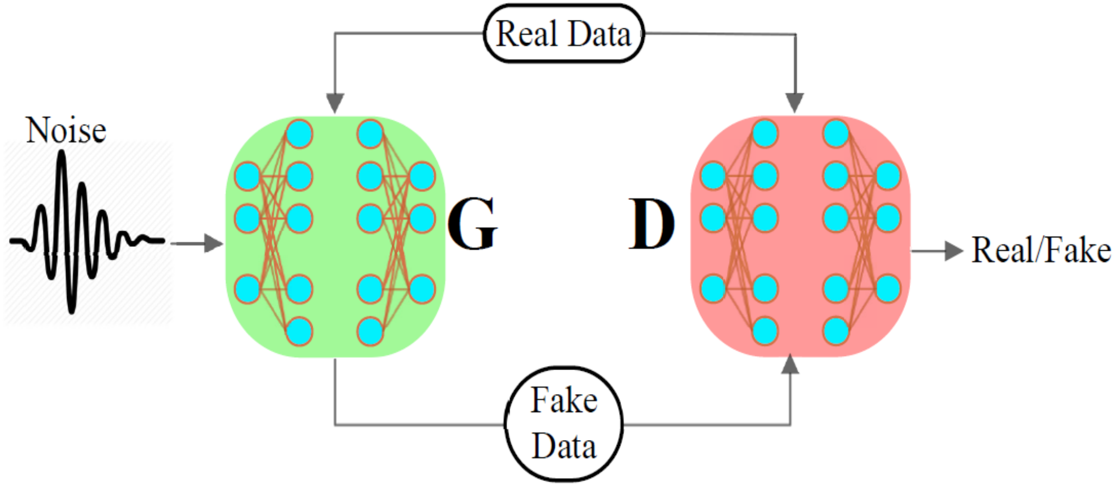


Yang L F et al poster, IAEA, 2025.



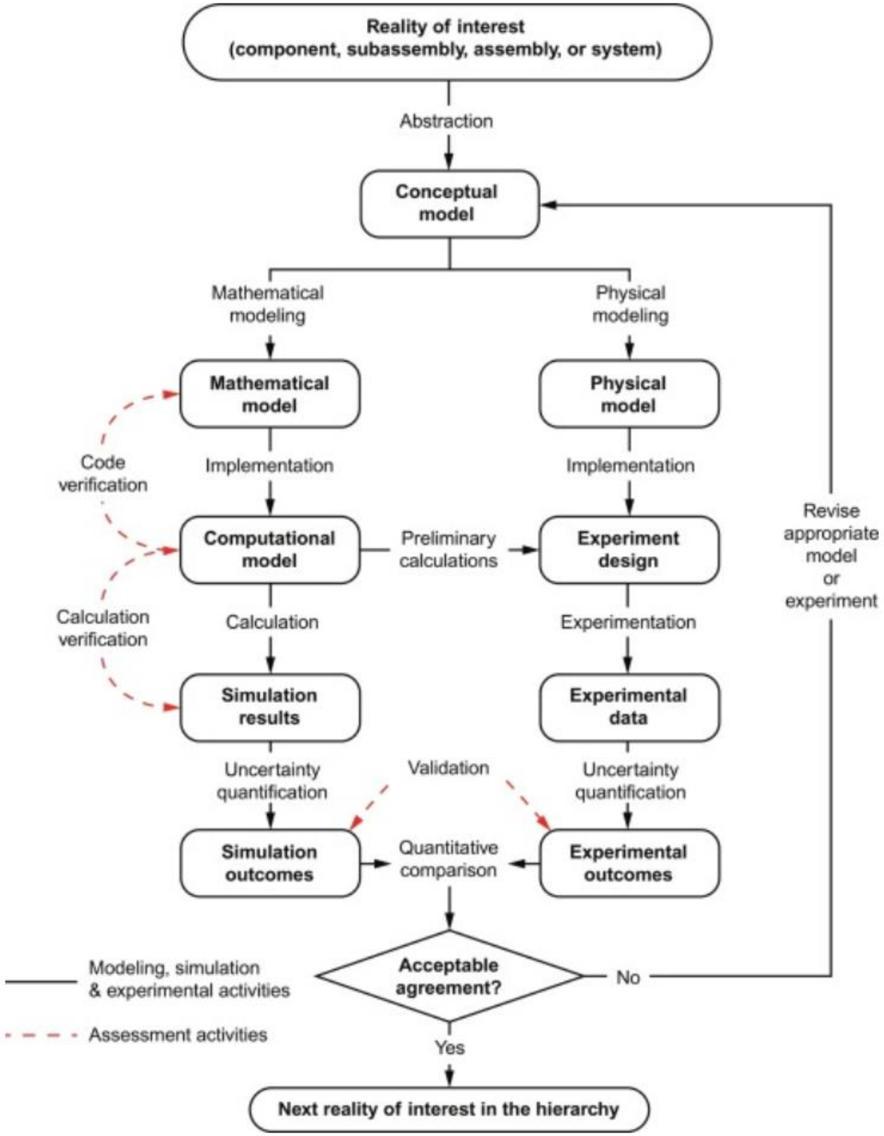
Missing data imputation





Information 13.12 (2022): 575.

Verification, Validation and Uncertainty Quantification



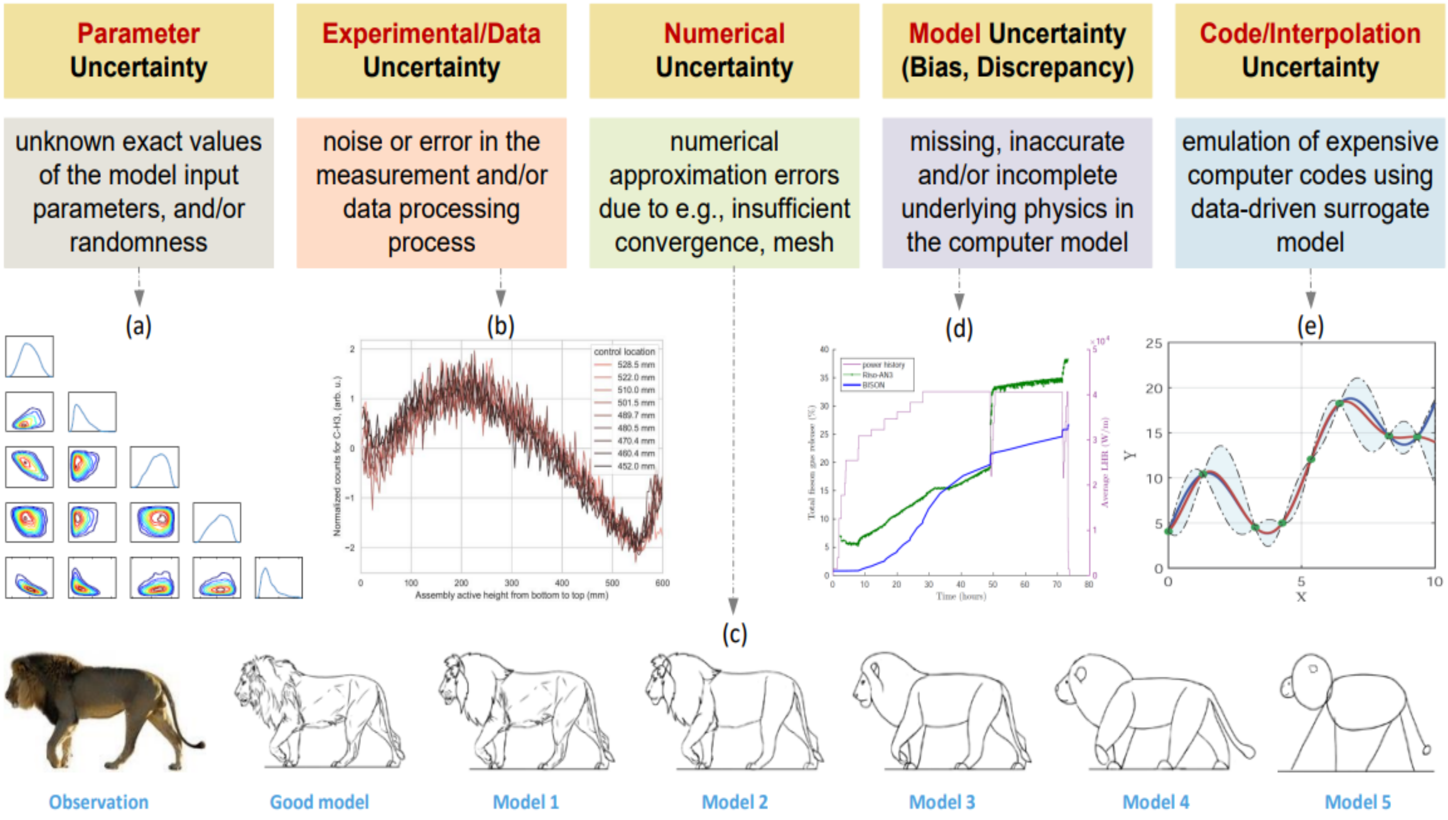


Figure 1: Illustration of uncertainty sources in physics-based M&S.

ArXiv:2503.17385 (2025).

Summary

- Measurement uncertainty brings risks to diagnostic reliability and decision reliability. These risks can be mitigated through data fusion and UQ.
- Bayesian inference and deep learning have advantages and limitations in accuracy, processing speed, and training costs. An appropriate data fusion method should be selected based on specific application scenarios.
- All fusion methods are sensitive to errors or invalid data. Data cleaning is necessary before data fusion.
- Through VVUQ, the uncertainty of experiments and simulations can be quantified, thereby enhancing the self-consistency and interpretability of experiments and simulations.

References

- Fischer R et al 2010 Fusion science and technology 58(2): 675-684.
- Dreier H. Bayesian Experimental Design-Applications in Nuclear Fusion[J]. 2007.
- Abdar M et al 2021 Information Fusion 76 243-297.
- Shahbazian R et al Information, 2022, 13(12): 575.
- Lan T et al Plasma Physics and Controlled Fusion, 2022, 64(12): 124003.
- Xie XP et al Plasma Physics and Controlled Fusion, 2025, 67(4): 045001.
- Lan T et al Computer Physics Communications, 2019, 234:159-166.
- Wu X et al arXiv:2503.17385, 2025.

ASIPP

Thank you

for your attention!