Scientific Machine Learning in Astrophysics

Machine Learning for Physics; Physics for Machine Learning

Jason D. McEwen

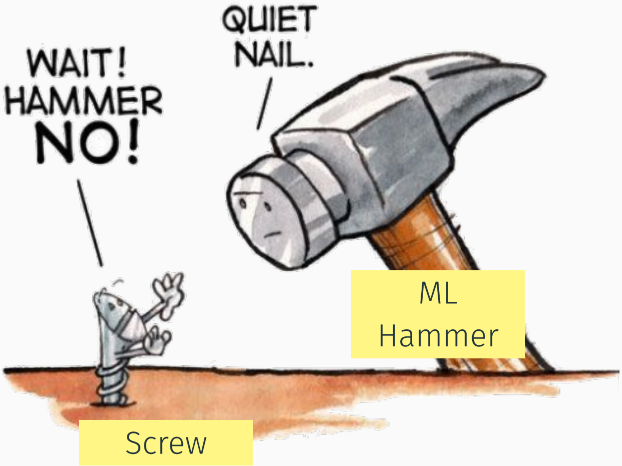
www.jasonmcewen.org

Mullard Space Science Laboratory (MSSL), University College London (UCL)

Rutherford Appleton Laboratory (RAL) Scientific Machine Learning Seminar

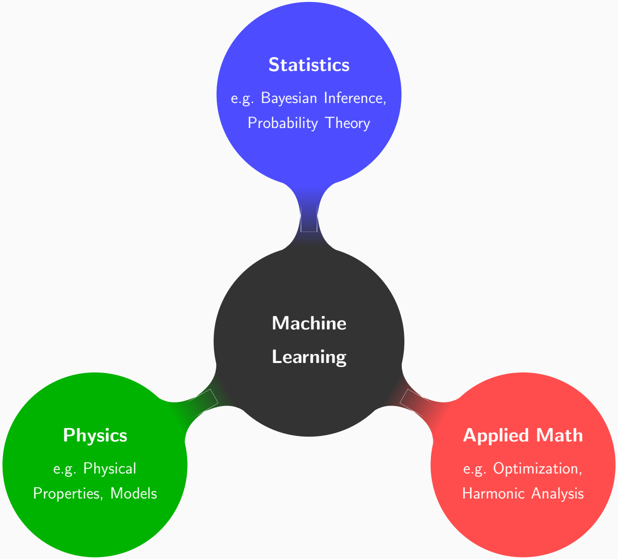
September 2023

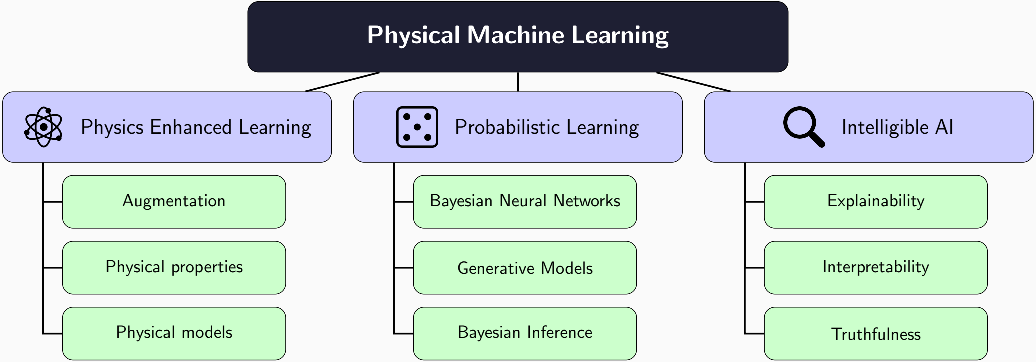
The machine learning hammer




ML Cog

Merging paradigms



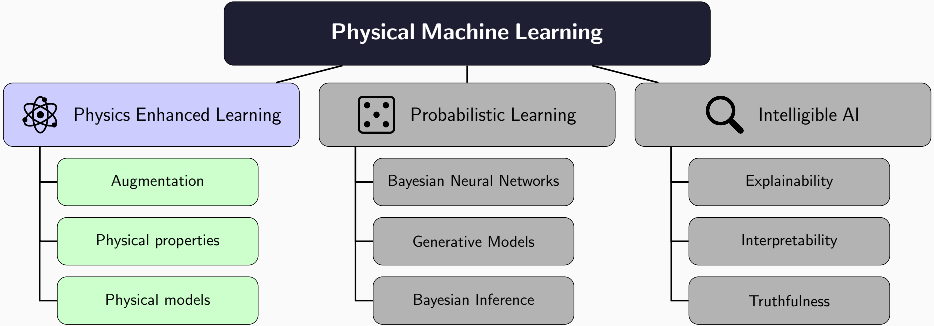

Physical Machine Learning
Physics Enhanced Learning
Augmentation
Physical properties
Physical models
Probabilistic Learning
Bayesian Neural Networks
Generative Models
Bayesian Inference
Intelligible AI
Explainability
Interpretability
Truthfulness

Physics Enhanced Learning

Physics Enhanced Learning

Embed physical understanding of the world into machine learning models.

(See review by Karniadakis *et al.* 2021.)



Augmentation

Apply **physical transformations** that data known to satisfy to augment training data $\rightsquigarrow$ ML model **learns physics through training**.

Augmentation

Apply **physical transformations** that data known to satisfy to augment training data $\rightsquigarrow$ ML model **learns physics through training**.

- Common to augment image data-set with rotations, flips, shifts, scales, contrast, ...

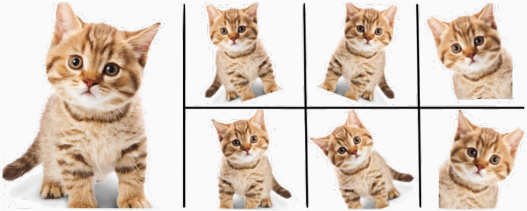
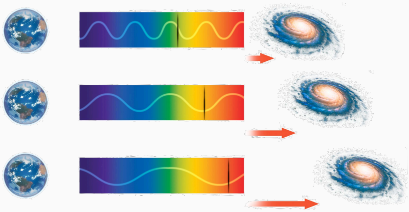


Image augmentation

Augmentation

Apply **physical transformations** that data known to satisfy to augment training data $\rightsquigarrow$ ML model **learns physics through training**.

- Redshift augmentation of supernovae observations (Boone 2019, Alves *et al.* 2022, 2023)



Redshift augmentation

Augmentation

Apply **physical transformations** that data known to satisfy to augment training data $\rightsquigarrow$ ML model **learns physics through training**.

- ▷ Data efficiency suffers: data "used" to learn physics, rather than problem.

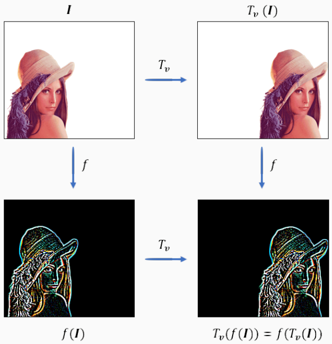
Physical properties: geometries, symmetries, conservation laws

ⓘ **Encode physical properties** of the world into ML models (e.g. geometry, symmetries, conservation laws) $\rightsquigarrow$ **Physics embedded in architecture** of ML model.

Physical properties: geometries, symmetries, conservation laws

Encode physical properties of the world into ML models (e.g. geometry, symmetries, conservation laws) ⇝ **Physics embedded in architecture** of ML model.

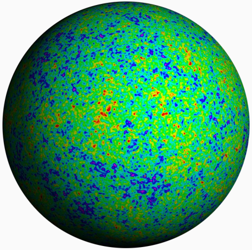
- Key factor CNNs so successful is due to encoding translational equivariance.



Physical properties: geometries, symmetries, conservation laws

Encode physical properties of the world into ML models (e.g. geometry, symmetries, conservation laws) ⇝ **Physics embedded in architecture** of ML model.

- Geometric deep learning on the sphere
  (Cobb et al. 2021; McEwen et al. 2022;
  Ocampo, Price & McEwen 2023)



CMB observed on the celestial sphere

Physical properties: geometries, symmetries, conservation laws

Encode physical properties of the world into ML models (e.g. geometry, symmetries, conservation laws) $\rightsquigarrow$ **Physics embedded in architecture** of ML model.

- Equivariant machine learning, structured like classical physics (Villar *et al.* 2021)

Orthogonal	$\mathrm{O}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = Q\,Q^\top = I_d\}$,
Rotation	$\mathrm{SO}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = Q\,Q^\top = I_d,\ \det(Q) = 1\}$
Translation	$\mathrm{T}(d) = \{w \in \mathbb{R}^d\}$
Euclidean	$\mathrm{E}(d) = \mathrm{T}(d) \rtimes \mathrm{O}(d)$
Lorentz	$\mathrm{O}(1,d) = \{Q \in \mathbb{R}^{(d+1)\times(d+1)} : Q^\top \Lambda\, Q = \Lambda,\ \Lambda = \mathrm{diag}([1,-1,\ldots,-1])\}$
Poincaré	$\mathrm{IO}(1,d) = \mathrm{T}(d+1) \rtimes \mathrm{O}(1,d)$
Permutation	$\mathrm{S}_n = \{\sigma : [n] \to [n] \text{ bijective function}\}$

Groups considered

Physical properties: geometries, symmetries, conservation laws

Encode physical properties of the world into ML models (e.g. geometry, symmetries, conservation laws) $\rightsquigarrow$ **Physics embedded in architecture** of ML model.

- Highly computationally demanding.
- Always required?

Physical properties: geometries, symmetries, conservation laws

Encode physical properties of the world into ML models (e.g. geometry, symmetries, conservation laws) $\rightsquigarrow$ **Physics embedded in architecture** of ML model.

- Highly computationally demanding.
- Always required?

- Develop efficient algorithms (e.g. Ocampo, Price & McEwen 2023).
- Inductive biases not enforced.

Physical models: PINNS and differentiable physics

Encode physical models of world into ML models:

1. Encode dynamics (differential equations) via loss functions (PINNs).
2. Embed full (differentiable) physical models inside ML model.

⇝ **Physics learned in training and embedded in model.**

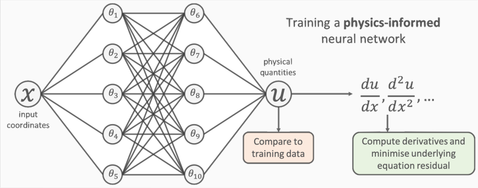
Physical models: PINNS and differentiable physics

Encode physical models of world into ML models:

1. Encode dynamics (differential equations) via loss functions (PINNs).
2. Embed full (differentiable) physical models inside ML model.

$\rightsquigarrow$ **Physics learned in training and embedded in model.**

- Physics informed neural networks (PINNs) encode differentiable equations (e.g. boundary conditions) in loss.



PINNs

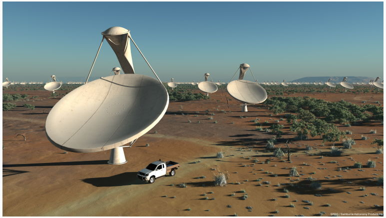
Physical models: PINNS and differentiable physics

Encode physical models of world into ML models:

1. Encode dynamics (differential equations) via loss functions (PINNs).
2. Embed full (differentiable) physical models inside ML model.

$\rightsquigarrow$ **Physics learned in training and embedded in model.**

- Differentiable physical models
  - Radio interferometric telescope (Mars *et al.* 2023, in prep.)
  - Optical PSF (Liaudat *et al.* 2023)
  - JAX-Cosmo (Campagne *et al.* 2023)



SKA (artist impression)

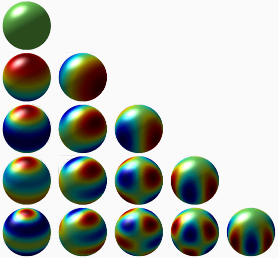
Physical models: PINNS and differentiable physics

Encode physical models of world into ML models:

1. Encode dynamics (differential equations) via loss functions (PINNs).
2. Embed full (differentiable) physical models inside ML model.

⇝ **Physics learned in training and embedded in model.**

- Differentiable mathematical methods
  - Fourier transforms
  - Spherical harmonic transforms (**`s2fft`**; Price & McEwen, in prep.)
  - Spherical wavelet transforms (**`s2wav`**; Price *et al.* in prep.)
  - Spherical scattering transforms (Mousset, Price, Allys, McEwen, in prep.)



Spherical harmonics

Physical models: PINNS and differentiable physics

Encode physical models of world into ML models:

1. Encode dynamics (differential equations) via loss functions (PINNs).
2. Embed full (differentiable) physical models inside ML model.

$\rightsquigarrow$ **Physics learned in training and embedded in model.**

- PINNs only capture limited dynamics via loss.
- Full physical models requires differentiable programming frameworks.

Physical models: PINNS and differentiable physics

Encode physical models of world into ML models:

1. Encode dynamics (differential equations) via loss functions (PINNs).
2. Embed full (differentiable) physical models inside ML model.

⇝ **Physics learned in training and embedded in model.**

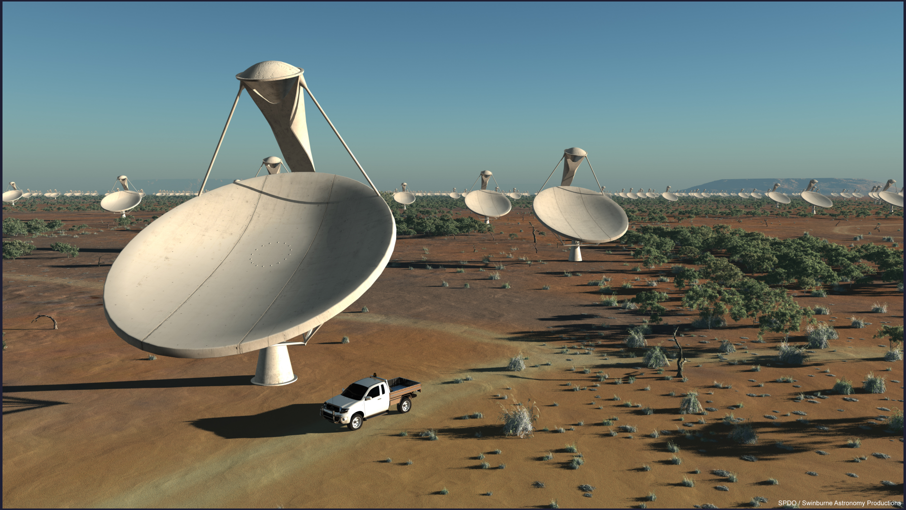
- PINNs only capture limited dynamics via loss.
- Full physical models requires differentiable programming frameworks.

- Capture full physics with differentiable models!
- Emulators also provide differentiability (e.g. `CosmoPower`; Spurio Mancini et al. 2021).
- Write new differentiable codes (e.g. `s2fft`; Price & McEwen, in prep.).

Case Study

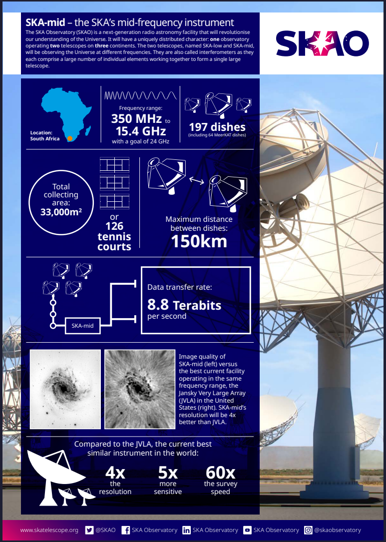
Learned interferometric imaging

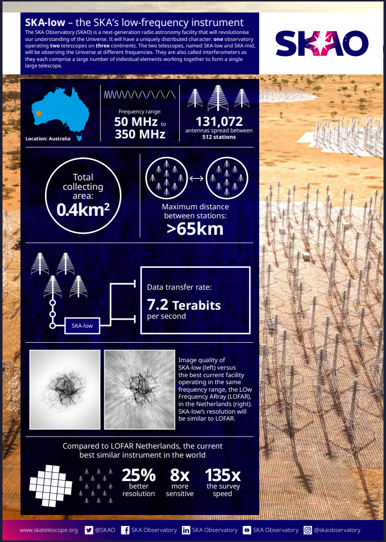
Square Kilometre Array (SKA)



SKA sites

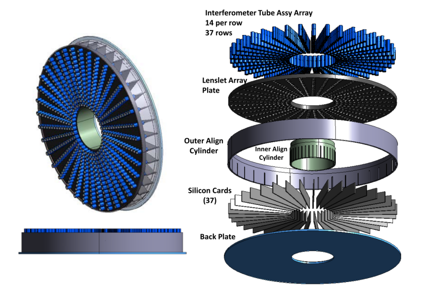
Case Study





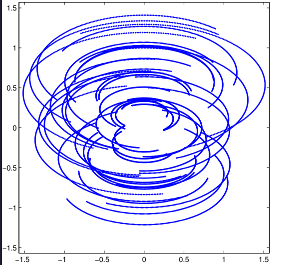
SPIDER instrument

- SPIDER is new interferometric optical imaging device developed by UC Davis and Lockheed Martin.
- Lenslet array to measure multiple interferometric baselines and photonic integrated circuits (PICs) for miniaturization.
- Reduces weight, cost and power consumption of optical telescopes.



Interferometric imaging

"Fourier"
Measurements
⇒



Recover an image from noisy and incomplete "Fourier" measurements.

Learned interferometric imaging

- Learned interferometric imaging for the SPIDER instrument (Mars *et al.* 2023)
- Learned radio interferometric imaging with varying visibility coverage (Mars *et al.* in prep.)

Code: coming soon!

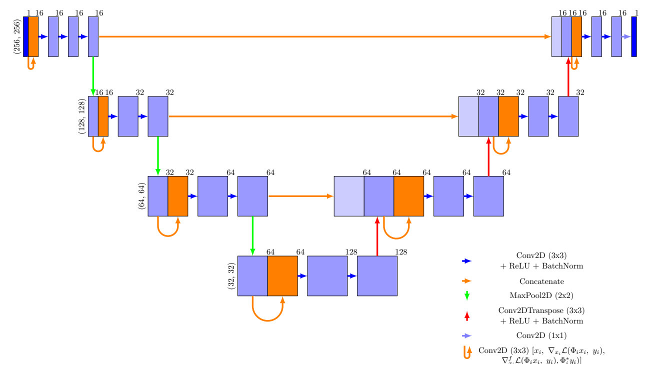
Matthijs Mars

Marta Betcke

GU-Net architecture for learned interferometric imaging

Integrate (differentiable) physical model of instrument into architecture; plus multi-resolution instrument model. (Mars *et al.* 2023, Mars *et al.* in prep.)

Transfer learning to handle measurement operator variability (telescope configuration).



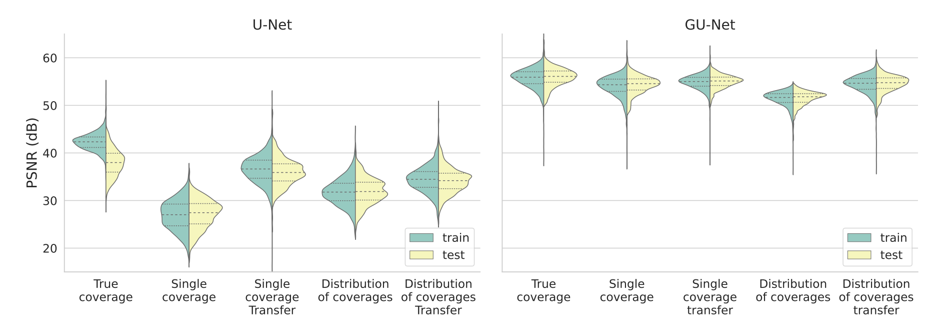
For instrument model Φ_i at resolution i, consider learned post-processing operator

$$\Lambda_{i,\theta}(x_i,\ \nabla_{x_i}\mathcal{L}(\Phi_i x_i,\ y_i), \nabla^f_{x_i}\mathcal{L}(\Phi_i x_i,\ y_i), \Phi_i^* y_i),$$

where

$$\nabla^f_{x_i}\mathcal{L}(\Phi_i x_i, y_i) \propto \Phi_i^*(W_i(\Phi_i x_i - y_i)).$$

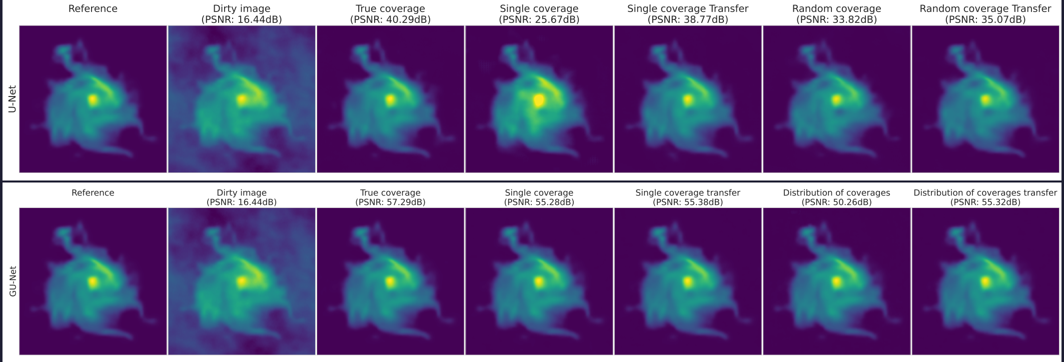
Distribution of radio interferometric reconstruction quality



Reconstruction quality (PSNR ↑) for different training strategies.

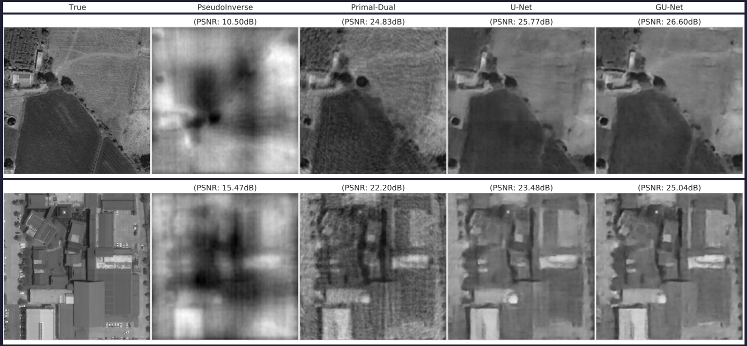
- Superior reconstruction quality by integrating physical model of instrument and more robust to measurement operator variability.
- Imaging time speed-up of 50-600× relative to classical approaches.

Reconstructed radio interferometric images



- Full end-to-end learning for radio interferometric imaging with support for varying measurement operators for the first time.

Reconstructed SPIDER images



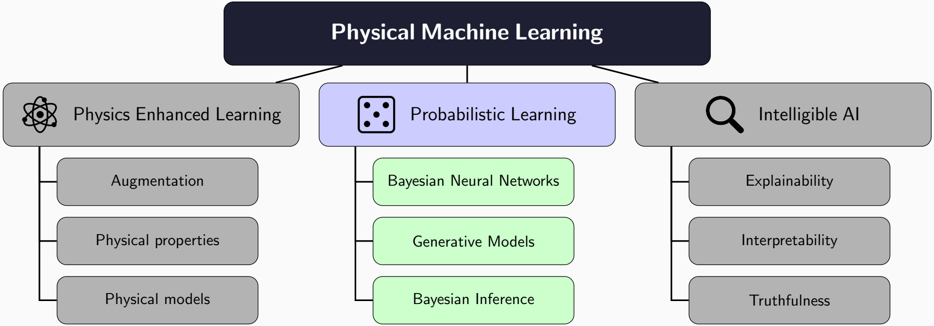
- Dramatic reduction in computational time opens up real time imaging with SPIDER for the first time.

Probabilistic Learning

Probabilistic Learning

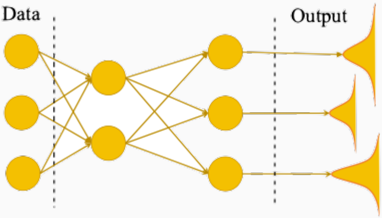
Embed a probabilistic representation of data, models and/or outputs.

(See Murray 2022.)



Bayesian neural networks for uncertainty quantification

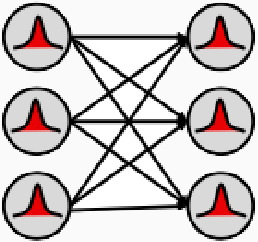
Bayesian neural networks incorporate **probabilistic representation** to quantify **uncertainty of outputs** (idea pioneered by MacKay 1992).



Bayesian neural networks for uncertainty quantification

Bayesian neural networks incorporate **probabilistic representation** to quantify **uncertainty of outputs** (idea pioneered by MacKay 1992).

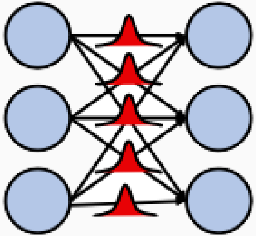
- MC Dropout (Gal & Ghahramani 2016): drop nodes probabilistically to sample an ensemble of networks.



Bayesian neural networks for uncertainty quantification

Bayesian neural networks incorporate **probabilistic representation** to quantify **uncertainty of outputs** (idea pioneered by MacKay 1992).

- Bayes by Backprop (Blundel *et al.* 2015): model distribution of weights (by variational inference).



Bayesian neural networks for uncertainty quantification

Bayesian neural networks incorporate **probabilistic representation** to quantify **uncertainty of outputs** (idea pioneered by MacKay 1992).

- Probabilistic ML frameworks (*e.g.* TensorFlow Probability).

Bayesian neural networks for uncertainty quantification

Bayesian neural networks incorporate **probabilistic representation** to quantify **uncertainty of outputs** (idea pioneered by MacKay 1992).

- Encode epistemic uncertainty of model.
- But what does the output distribution represent?
- Requires careful consideration of training data.

Bayesian neural networks for uncertainty quantification

Bayesian neural networks incorporate **probabilistic representation** to quantify **uncertainty of outputs** (idea pioneered by MacKay 1992).

- ▷ Encode epistemic uncertainty of model.
- ▷ But what does the output distribution represent?
- ▷ Requires careful consideration of training data.

- ▷ Statistical validation (hold that thought... see upcoming Truthfulness section).

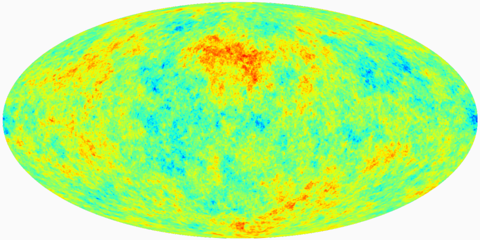
Generative models

Generative models **learn a prior distribution** from data for sampling and/or evaluating probabilities.

Generative models

Generative models **learn a prior distribution** from data for sampling and/or evaluating probabilities.

- Emulation: sample from learned prior
  (Perraudin *et al.* 2020, Allys *et al.* 2020, Price *et al.* 2023, Price *et al.* in prep.)

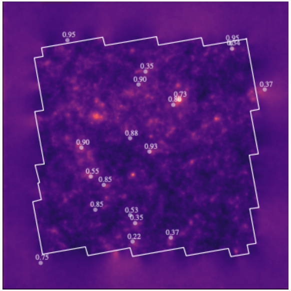


Emulated cosmic string maps
(**`stringgen`**, Price *et al.* 2023, Price *et al.* in prep.)

Generative models

Generative models **learn a prior distribution** from data for sampling and/or evaluating probabilities.

- Integrate learned priors into analysis
  (Remy *et al.* 2022, McEwen *et al.* 2023)



Learn convergence field prior
(Remy *et al.* 2022)

Generative models

Generative models **learn a prior distribution** from data for sampling and/or evaluating probabilities.

- Availability and representativeness of training data.
- Truthfulness, *e.g.* diversity of ML model often lacking.

Generative models

Generative models **learn a prior distribution** from data for sampling and/or evaluating probabilities.

- ▷ Availability and representativeness of training data.
- ▷ Truthfulness, *e.g.* diversity of ML model often lacking.

- ▷ Public datasets/benchmarks (*e.g.* BASE, IllustrisTNG, CAMELS, Quijote, CosmoGrid).
- ▷ Meta sampling to recover distribution over manifold (*e.g.* Price *et al.* 2023).
- ▷ Truthfulness (hold that thought... see upcoming Truthfulness section).

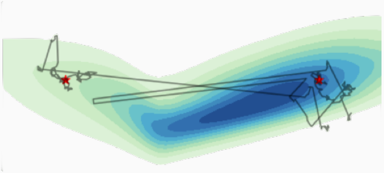
Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

- Enhanced MCMC for parameter estimation (Grabrie *et al.* 2022, Karamanis *et al.* 2022).

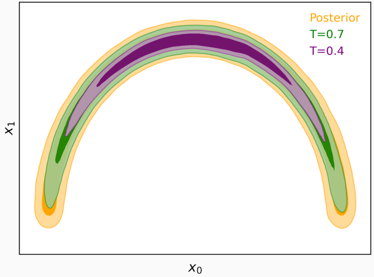


Learned proposal distributions

Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

- Enhanced Bayesian model selection (**`harmonic`**; McEwen *et al.* 2021, Polanska *et al.* 2023).

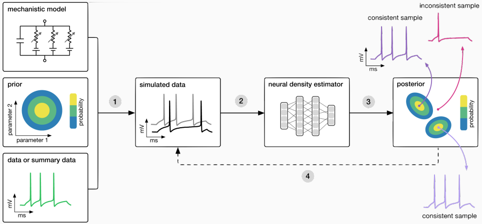


Learned harmonic mean estimator (**`harmonic`**)

Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

- ▷ Simulation-based inference (Alsing *et al.* 2018, Cranmer *et al.* 2021).
- ▷ Model selection for simulation-based inference (`harmonic`; Spurio Mancini *et al.* 2022)

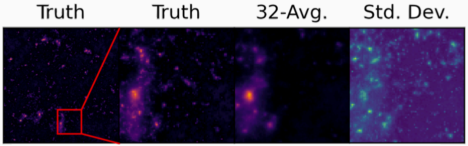


`sbi`

Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

- Variational inference
  (Whitney *et al.* in prep.)



Mass mapping with uncertainties
by variational inference

Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

- Availability and representativeness of training data.
- Cost of training.
- Truthfulness?

Bayesian inference

ML techniques can be integrated into Bayesian frameworks to **enhance accuracy and computational efficiency**, making some approaches accessible that were previously intractable.

- ▷ Availability and representativeness of training data.
- ▷ Cost of training.
- ▷ Truthfulness?

- ▷ Public datasets/benchmarks (*e.g.* BASE, IllustrisTNG, CAMELS, Quijote, CosmoGrid).
- ▷ Amortized inference (training **not** repeated for new observations).
- ▷ Integrate in Bayesian framework to provide statistical guarantees.
- ▷ Statistical validation (hold that thought... see upcoming Truthfulness section).

Case Study

Learned harmonic mean estimator for Bayesian model selection

What is the nature of dark energy?

Is the equation of state of dark energy:

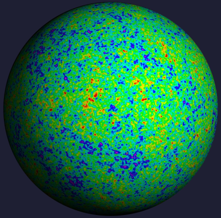
(i) constant (*i.e.* Einstein's cosmological constant) or

(ii) evolving with cosmic time?

Cosmic microwave background (CMB) radiation

Constrain nature of dark energy with observations of the cosmic microwave background (CMB) (relic radiation from the Big Bang).

Atacama Cosmology Telescope (ACT)



CMB

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{p(\theta \mid \mathbf{y}, M)}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y} \mid \theta, M)}^{\text{likelihood}} \; \overbrace{p(\theta \mid M)}^{\text{prior}}}{\underbrace{p(\mathbf{y} \mid M)}_{\text{evidence}}} = \frac{\overbrace{\mathcal{L}(\theta)}^{\text{likelihood}} \; \overbrace{\pi(\theta)}^{\text{prior}}}{\underbrace{Z}_{\text{evidence}}},$$

for parameters θ, model M and observed data $\mathbf{y}$.

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{p(\theta \mid \mathbf{y}, M)}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y} \mid \theta, M)}^{\text{likelihood}} \; \overbrace{p(\theta \mid M)}^{\text{prior}}}{\underbrace{p(\mathbf{y} \mid M)}_{\text{evidence}}} = \frac{\overbrace{\mathcal{L}(\theta)}^{\text{likelihood}} \; \overbrace{\pi(\theta)}^{\text{prior}}}{\underbrace{Z}_{\text{evidence}}},$$

for parameters θ, model M and observed data $\mathbf{y}$.

For **parameter estimation**, typically draw samples from the posterior by *Markov chain Monte Carlo (MCMC)* sampling.

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \,|\, M_j) p(M_j)}{\sum_j p(\boldsymbol{y} \,|\, M_j) p(M_j)} .$$

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_j) p(M_j)}{\sum_j p(\mathbf{y} \mid M_j) p(M_j)} .$$

For **model selection**, consider posterior model odds:

$$\frac{p(M_1 \mid \mathbf{y})}{p(M_2 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \times \frac{p(M_1)}{p(M_2)} .$$

posterior odds — Bayes factor — prior odds

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_j) p(M_j)}{\sum_j p(\mathbf{y} \mid M_j) p(M_j)} .$$

For **model selection**, consider posterior model odds:

$$\frac{p(M_1 \mid \mathbf{y})}{p(M_2 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \times \frac{p(M_1)}{p(M_2)} .$$

posterior odds Bayes factor prior odds

Must compute the **Bayesian model evidence** or **marginal likelihood** given by the normalising constant

$$z = p(\mathbf{y} \mid M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta) .$$

Bayesian inference: model selection

By Bayes' theorem for model M_j:

$$p(M_j \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, M_j) p(M_j)}{\sum_j p(\mathbf{y} \,|\, M_j) p(M_j)} \,.$$

For **model selection**, consider posterior model odds:

$$\underbrace{\frac{p(M_1 \,|\, \mathbf{y})}{p(M_2 \,|\, \mathbf{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathbf{y} \,|\, M_1)}{p(\mathbf{y} \,|\, M_2)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior odds}} \,.$$

Must compute the **Bayesian model evidence** or **marginal likelihood** given by the normalising constant

$$z = p(\mathbf{y} \,|\, M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta) \,.$$

$\rightsquigarrow$ Challenging computational problem.

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] = \frac{1}{z}$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, \boldsymbol{y})} \left[\frac{1}{\mathcal{L}(\theta)} \right] = \frac{1}{z}$$

Original harmonic mean estimator (Newton & Raftery 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)} , \quad \theta_i \sim p(\theta \,|\, \boldsymbol{y})$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, \boldsymbol{y})} \left[\frac{1}{\mathcal{L}(\theta)} \right] = \frac{1}{z}$$

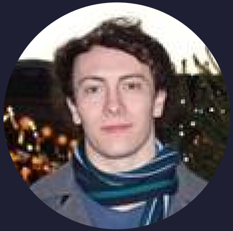
Original harmonic mean estimator (Newton & Raftery 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)}, \quad \theta_i \sim p(\theta \,|\, \boldsymbol{y})$$

Very simple approach but can fail catastrophically (Neal 1994).

- Learned harmonic mean estimator
  (McEwen *et al.* 2021)
- Bayesian model comparison for simulation-based inference
  (Spurio Mancini *et al.* 2022)
- Learned harmonic mean estimation with normalizing flows
  (Polanska *et al.* 2023)

Code: `https://github.com/astro-informatics/harmonic`



Matt Price

Alessio Spurio Mancini

Alicja Polanska

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted harmonic mean relationship (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \mid y)} \left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right] = \frac{1}{z}$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, \mathbf{y})} \left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right] = \frac{1}{z}$$

Re-targeted **harmonic mean estimator** (Gelfand & Dey 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} , \quad \theta_i \sim p(\theta \,|\, \mathbf{y})$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

***Re-targeted* harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{p(\theta \,|\, \mathbf{y})} \left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right] = \frac{1}{z}$$

***Re-targeted* harmonic mean estimator** (Gelfand & Dey 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} , \quad \theta_i \sim p(\theta \,|\, \mathbf{y})$$

$\rightsquigarrow$ How set importance sampling target distribution $\varphi(\theta)$?

How set importance sampling target distribution $\varphi(\theta)$?

Optimal target:

$$\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

(resulting estimator has zero variance).

How set importance sampling target distribution $\varphi(\theta)$?

Optimal target:

$$\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

(resulting estimator has zero variance).

But clearly **not feasible** since requires knowledge of the evidence z (recall the target must be normalised) $\rightsquigarrow$ requires problem to have been solved already!

Learned harmonic mean estimator

Learn an approximation of the optimal target distribution:

$$\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z} .$$

Learned harmonic mean estimator

Learn an approximation of the optimal target distribution:

$$\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z} .$$

- Approximation not required to be highly accurate.
- Must not have fatter tails than posterior (*e.g.* by concentrating probability mass of normalising flow).

Learned harmonic mean estimator

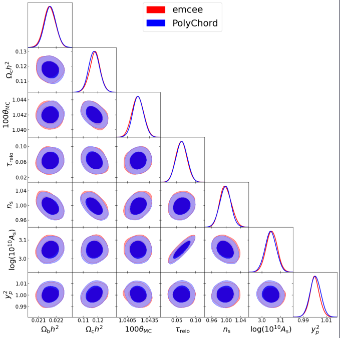
Learn an approximation of the optimal target distribution:

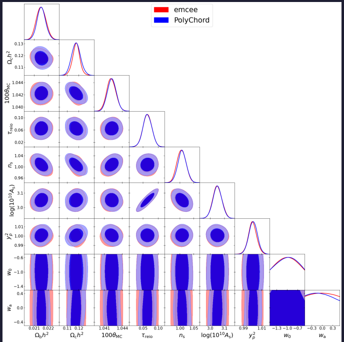
$$\varphi(\theta) \stackrel{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z} .$$

- Approximation not required to be highly accurate.
- Must not have fatter tails than posterior (*e.g.* by concentrating probability mass of normalising flow).

$\rightsquigarrow$ Solve long-standing problem by integrating ML into Bayesian framework.

What is the nature of dark energy?





Cosmological constant (LCDM):
$\log z = -168.87 \pm 0.29$

Evolving dark energy ($w_0 w_a$CDM):
$\log z = -169.32 \pm 0.25$

Bayes factor of $\Delta \log z = 0.45 \pm 0.54$: weak preference for cosmological constant (LCDM).

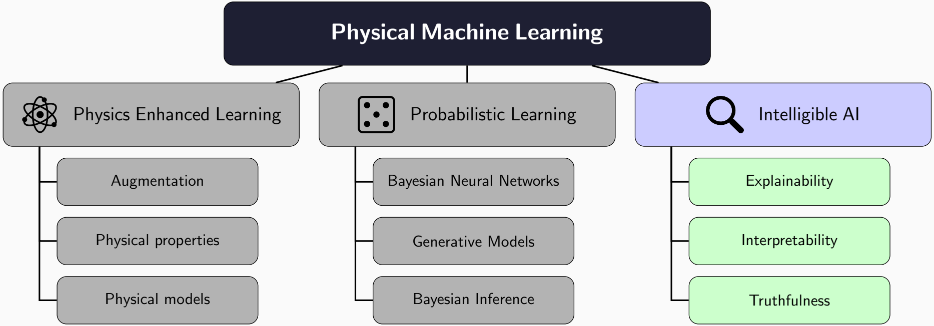
$3\times$ faster than alternative with potential to scale to considerably higher dimensions (WIP).

Intelligible AI

Intelligible AI

Machine learning methods that are able to be understood by humans.

(See Weld & Bansal 2018, Ras *et al.* 2020.)

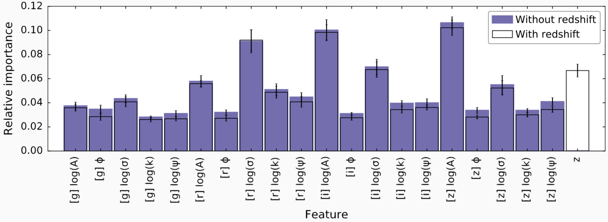


Explainability

Explainable ML techniques may or may not be interpretable themselves but their **outputs can be explained to humans.**

Explainable ML techniques may or may not be interpretable themselves but their **outputs can be explained to humans.**

- Feature importances (Lochner *et al.* 2016)

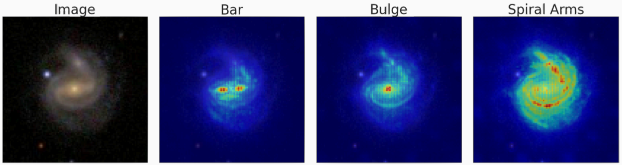


Supernova feature importances

Explainability

Explainable ML techniques may or may not be interpretable themselves but their **outputs can be explained to humans.**

- Saliency maps (Bhambra *et al.* 2022)



Galaxy saliency mapping

Explainability

Explainable ML techniques may or may not be interpretable themselves but their **outputs can be explained to humans.**

Poking the black box: may provide some explanation of outputs but humans still not able to comprehend underlying process.

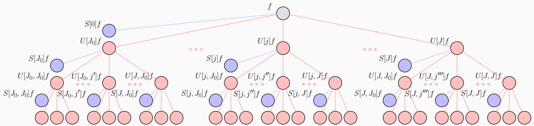
Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

- Designed models such as scattering and wavelet phase harmonic networks (Allys *et al.* 2020, Cheng *et al.* 2020, McEwen *et al.* 2022)

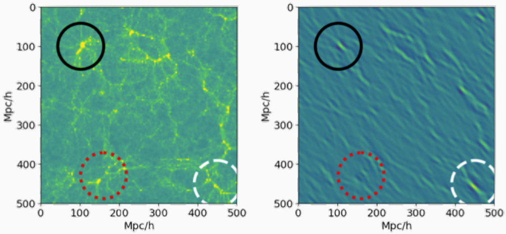


Scattering network (McEwen *et al.* 2022)

Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

- Designed models such as scattering and wavelet phase harmonic networks (Allys *et al.* 2020, Cheng *et al.* 2020, McEwen *et al.* 2022)

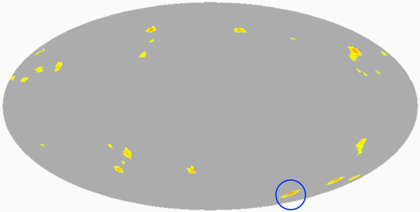


LSS features captured by wavelets (Allys *et al.* 2020)

Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

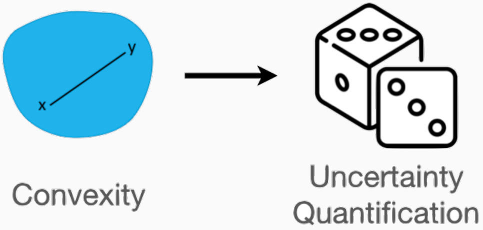
- Designed models such as scattering and wavelet phase harmonic networks (Allys *et al.* 2020, Cheng *et al.* 2020, McEwen *et al.* 2022)



First evidence that CMB cold spot due to supervoid (McEwen *et al.* 2007)

Interpretable ML models are **white boxes that can be understood by humans**.

- Interpretable constraints on ML models, *e.g.* convexity (Liaudat, McEwen *et al.* in prep.)

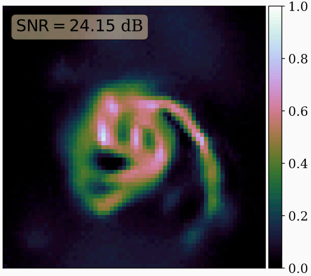


Impose convexity on learned model

Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

- Deep priors learned from training data (hybrid model-based and data-driven)
  (Remy *et al.* 2022, McEwen *et al.* 2023)



Compute Bayesian evidence for model selection
(**`proxnest`**, McEwen *et al.* 2023)

Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

- Designed models limit flexibility.
- Availability and representativeness of training data.

Interpretability

Interpretable ML models are **white boxes that can be understood by humans**.

- Designed models limit flexibility.
- Availability and representativeness of training data.

- Benefits of designed models often outweigh (minimal) reduced flexibility.
- Public datasets/benchmarks (*e.g.* IllustrisTNG, CAMELS, Quijote, CosmoGrid).
- Transfer learning, self-supervised learning.

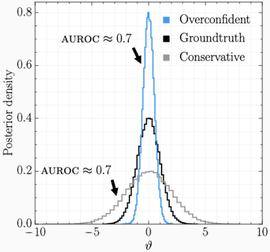
Truthfulness

Truthfulness **critical for science** in order for humans to have confidence in results of ML models. Closely coupled with a **meaningful statistical distribution** of outputs.

Truthfulness

Truthfulness **critical for science** in order for humans to have confidence in results of ML models. Closely coupled with a **meaningful statistical distribution** of outputs.

- Validity of statistical distributions (Hermans *et al.* 2022, Lemos *et al.* 2023)

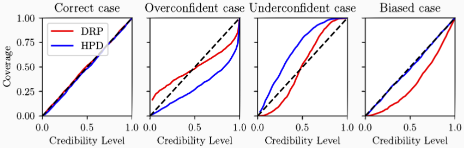


Validity of distribution (Hermans *et al.* 2022)

Truthfulness

Truthfulness **critical for science** in order for humans to have confidence in results of ML models. Closely coupled with a **meaningful statistical distribution** of outputs.

- Validity of statistical distributions (Hermans *et al.* 2022, Lemos *et al.* 2023)

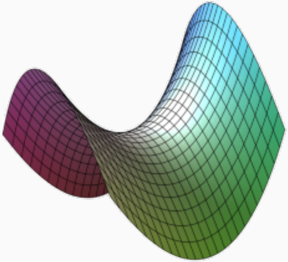


Coverage analysis (Lemos *et al.* 2023)

Truthfulness

Truthfulness **critical for science** in order for humans to have confidence in results of ML models. Closely coupled with a **meaningful statistical distribution** of outputs.

- Diversity (avoiding mode-collapse)
  (Price *et al.* 2023, Whitney *et al.* in prep.)



Recover probability distribution over full underlying manifold

Truthfulness

Truthfulness **critical for science** in order for humans to have confidence in results of ML models. Closely coupled with a **meaningful statistical distribution** of outputs.

- Uncertainties not aways meaningful.
- Diversity of ML model often lacking.

Truthfulness

Truthfulness **critical for science** in order for humans to have confidence in results of ML models. Closely coupled with a **meaningful statistical distribution** of outputs.

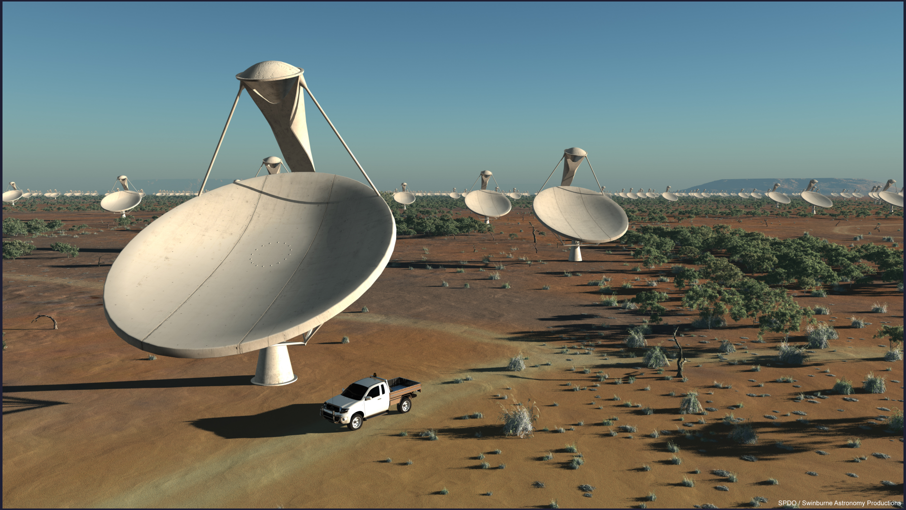
- Uncertainties not aways meaningful.
- Diversity of ML model often lacking.

- Integrate in statistical framework to inherit theoretical guarantees.
- Extensive validation tests (*e.g.* Hermans *et al.* 2022, Lemos *et al.* 2023).
- Meta sampling to recover distribution over manifold (*e.g.* Price *et al.* 2023).
- Well-posed frameworks (*e.g.* physics enhanced, probabilistic).

Case Study

Uncertainty quantification for exascale imaging

Towards exascale computing with the SKA



MAP estimation vs MCMC sampling

MAP estimation

+ Based on optimization so computationally efficient.
– Does not traditionally provide uncertainties.

MCMC sampling

– Based on sampling so computationally demanding.
+ Recover full posterior distribution.

However, based on hand-crafted priors, which are not highly expressive.

Scalable Bayesian UQ with learned data-driven priors

1. Statistical framework: Bayesian inference and MAP estimation.
2. Mathematical theory: probability concentration theorem for log-convex distributions.
3. Designed/constrained ML model: convex ML model with explicit potential.

$\rightsquigarrow$ **Scalable Bayesian UQ** with learned data-driven priors, which are highly expressive.

Scalable Bayesian UQ with learned data-driven priors

1. Statistical framework: Bayesian inference and MAP estimation.
2. Mathematical theory: probability concentration theorem for log-convex distributions.
3. Designed/constrained ML model: convex ML model with explicit potential.

$\rightsquigarrow$ **Scalable Bayesian UQ** with learned data-driven priors, which are highly expressive.

- ▷ **Interpretable method.**
- ▷ **Interpretable results.**
- ▷ **Validate** by MCMC sampling (for low-dimensional setting).

- Scalable Bayesian UQ with learned data-driven priors (Liaudat *et al.* in prep.)

Code: coming soon!

Tobias Liaudat

Marcelo Pereyra

Marta Betcke

Statistical framework: MAP estimation

Bayes Theorem (ignore normalising evidence):

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) \,, \quad \textit{i.e.} \text{ posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$p(\mathbf{y} \mid \mathbf{x}) \propto \exp\Big(-\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2/(2\sigma^2)\Big)$$

likelihood

$$p(\mathbf{x}) \propto \exp\Big(-R(\mathbf{x})\Big)$$

prior

Statistical framework: MAP estimation

Bayes Theorem (ignore normalising evidence):

$$p(\mathbf{x} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \mathbf{x}) p(\mathbf{x}) \,, \quad \textit{i.e.} \text{ posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$p(\mathbf{y} \,|\, \mathbf{x}) \propto \exp\left(-\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 / (2\sigma^2)\right) \qquad p(\mathbf{x}) \propto \exp\left(-R(\mathbf{x})\right)$$

likelihood prior

Consider log-posterior:

$$\log p(\mathbf{x} \,|\, \mathbf{y}) = -\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 / (2\sigma^2) - R(\mathbf{x}) + \text{const.}$$

Statistical framework: MAP estimation

Bayes Theorem (ignore normalising evidence):

$$p(x \mid y) \propto p(y \mid x) p(x) \;, \quad \textit{i.e.}\ \text{posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$p(y \mid x) \propto \exp\left(-\|y - \mathbf{\Phi} x\|_2^2 / (2\sigma^2)\right) \qquad p(x) \propto \exp\left(-R(x)\right)$$

likelihood prior

Consider log-posterior:

$$\log p(x \mid y) = -\|y - \mathbf{\Phi} x\|_2^2 / (2\sigma^2) - R(x) + \text{const.}$$

MAP estimator:

$$x_{\text{map}} = \arg\max_x \left[\log p(y \mid x)\right] = \arg\min_x \left[\|y - \mathbf{\Phi} x\|_2^2 + \lambda R(x)\right]$$

data fidelity regulariser

Mathematical theory: convex probability concentration

Posterior credible region:

$$p(x \in C_\alpha | y) = \int_{x \in \mathbb{R}^N} p(x|y) \mathbb{1}_{C_\alpha} \mathrm{d}x = 1 - \alpha.$$

Consider the highest posterior density (HPD) region

$$C^*_\alpha = \{x : -\log p(x) \leq \gamma_\alpha\}, \quad \text{with } \gamma_\alpha \in \mathbb{R}, \quad \text{and } p(x \in C^*_\alpha | y) = 1 - \alpha \text{ holds.}$$

Mathematical theory: convex probability concentration

Posterior credible region:

$$p(x \in C_\alpha | y) = \int_{x \in \mathbb{R}^N} p(x|y) \mathbb{1}_{C_\alpha} \mathrm{d}x = 1 - \alpha.$$

Consider the highest posterior density (HPD) region

$$C^*_\alpha = \{x : -\log p(x) \leq \gamma_\alpha\}, \quad \text{with } \gamma_\alpha \in \mathbb{R}, \quad \text{and } p(x \in C^*_\alpha | y) = 1 - \alpha \text{ holds.}$$

Theorem 3.1 (Pereyra 2017)

Suppose the posterior $p(x|y) = \exp[-f(x) - g(x)]/Z$ is log-concave on $\mathbb{R}^N$. Then, for any $\alpha \in (4\mathrm{e}^[(-N/3)], 1)$, the HPD region C^*_α is contained by

$$\hat{C}_\alpha = \left\{x : f(x) + g(x) \leq \hat{\gamma}_\alpha = f(\hat{x}_{\text{MAP}}) + g(\hat{x}_{\text{MAP}}) + \sqrt{N}\tau_\alpha + N\right\},$$

with a positive constant $\tau_\alpha = \sqrt{16 \log(3/\alpha)}$ independent of $p(x|y)$.

Mathematical theory: convex probability concentration

Posterior credible region:

$$p(x \in C_\alpha | y) = \int_{x \in \mathbb{R}^N} p(x|y) \mathbb{1}_{C_\alpha} \mathrm{d}x = 1 - \alpha.$$

Consider the highest posterior density (HPD) region

$$C^*_\alpha = \left\{x : -\log p(x) \leq \gamma_\alpha\right\}, \quad \text{with } \gamma_\alpha \in \mathbb{R}, \quad \text{and } p(x \in C^*_\alpha | y) = 1 - \alpha \text{ holds.}$$

Theorem 3.1 (Pereyra 2017)

Suppose the posterior $p(x|y) = \exp[-f(x) - g(x)]/Z$ is log-concave on $\mathbb{R}^N$. Then, for any $\alpha \in (4\mathrm{e}^[(-N/3)], 1)$, the HPD region C^*_α is contained by

$$\hat{C}_\alpha = \left\{x : f(x) + g(x) \leq \hat{\gamma}_\alpha = f(\hat{x}_{\text{MAP}}) + g(\hat{x}_{\text{MAP}}) + \sqrt{N}\tau_\alpha + N\right\},$$

with a positive constant $\tau_\alpha = \sqrt{16 \log(3/\alpha)}$ independent of $p(x|y)$.

We need only evaluate $f + g$ for the MAP estimation x_{MAP}!

Constrained ML model: convex regulariser

Adopt **neural-network-based convex regulariser** R (Goujon *et al.* 2022):

$$R(\boldsymbol{x}) = \sum_{n=1}^{N_C} \sum_{k} \psi_n \left((\boldsymbol{h}_n * \boldsymbol{x}) \, [k] \right) ,$$

- ψ_n are learned convex profile functions with Lipschitz continuous derivative;
- N_C learned convolutional filters $\boldsymbol{h}_n$.

Constrained ML model: convex regulariser

Adopt **neural-network-based convex regulariser** R (Goujon *et al.* 2022):

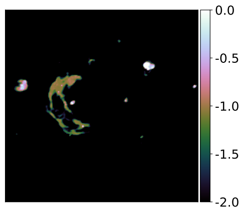
$$R(\boldsymbol{x}) = \sum_{n=1}^{N_C} \sum_{k} \psi_n \left((\boldsymbol{h}_n * \boldsymbol{x}) [k] \right),$$

- ψ_n are learned convex profile functions with Lipschitz continuous derivative;
- N_C learned convolutional filters $\boldsymbol{h}_n$.

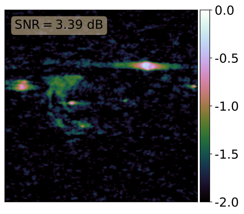
Properties:

1. **Convex** + **explicit** $\Rightarrow$ leverage convex UQ theory.
2. **Smooth regulariser with known Lipschitz constant** $\Rightarrow$ theoretical convergence guarantees.

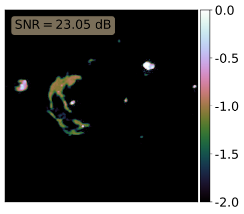
Reconstructed images



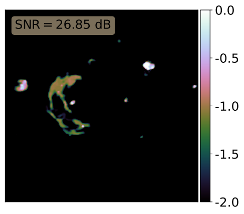
Ground truth



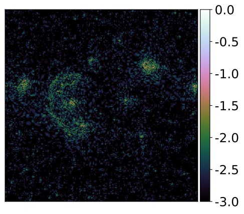
Dirty image
SNR=3.39 dB



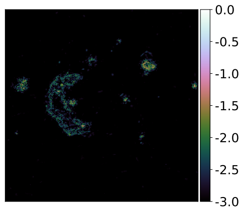
Reconstruction (classical)
SNR=23.05 dB



Reconstruction (learned)
SNR= 26.85 dB

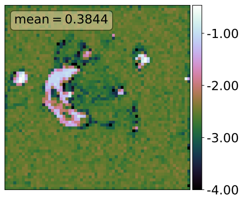


Error (classical)

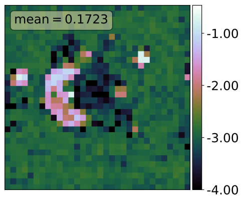


Error (learned)

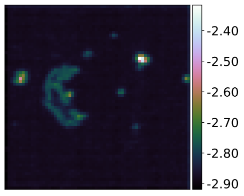
Approximate pixel-level uncertainty quantification



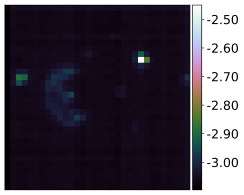
LCI
(super-pixel size 4 × 4)



LCI
(super-pixel size 8 × 8)

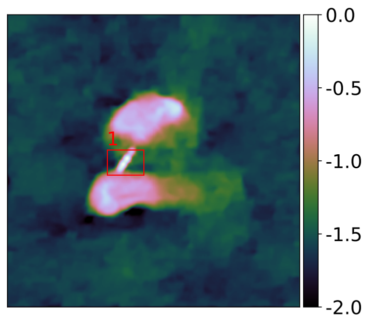


MCMC standard deviation
(super-pixel size 8 × 8)



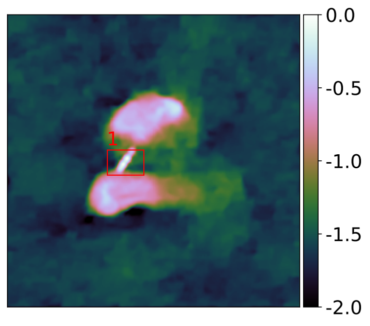
MCMC standard deviation
(super-pixel size 4 × 4)

Hypothesis testing of structure

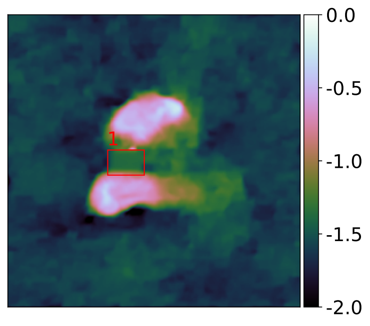


Reconstructed image

Hypothesis testing of structure

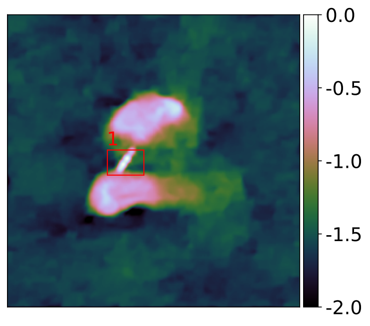


Reconstructed image

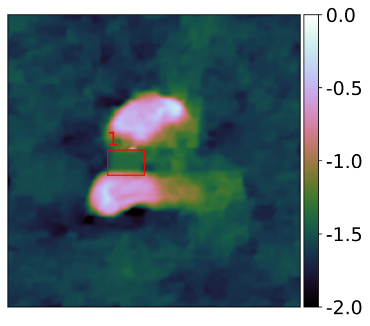


Surrogate test image (region removed)

Hypothesis testing of structure



Reconstructed image

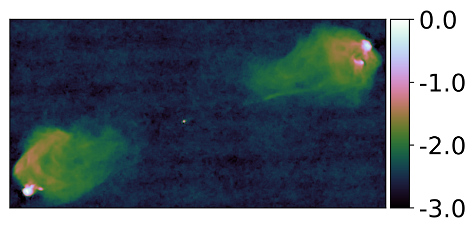


Surrogate test image (region removed)

Reject null hypothesis

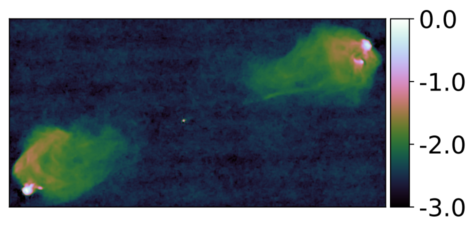
$\Rightarrow$ **structure physical**

Hypothesis testing of substructure

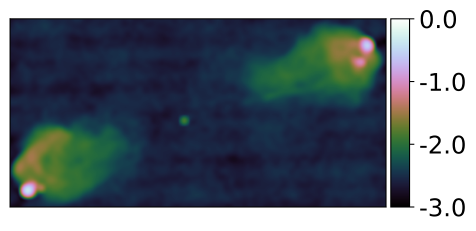


Reconstructed image

Hypothesis testing of substructure

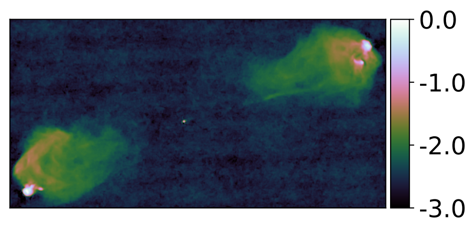


Reconstructed image

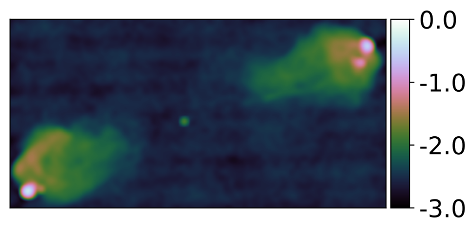


Surrogate test image (blurred)

Hypothesis testing of substructure



Reconstructed image



Surrogate test image (blurred)

Reject null hypothesis $\Rightarrow$ **substructure physical**

Scalable Bayesian UQ with learned data-driven priors

- Superior reconstruction quality by using learned data-driven prior.
- Uncertainty quantification for exascale imaging with learned priors for the first time.
- Validated by MCMC sampling (for low-dimensional setting)

Summary

Physical Machine Learning

- **Physics Enhanced Learning**
  - Augmentation
  - Physical properties
  - Physical models
- **Probabilistic Learning**
  - Bayesian Neural Networks
  - Generative Models
  - Bayesian Inference
- **Intelligible AI**
  - Explainability
  - Interpretability
  - Truthfulness

With great power comes great responsibility!