Artificial Intelligence and Cosmology

A powerful way of processing massive data

and/or

A new way of exploring theoretical models?

A truly new way of conducting research in cosmology?

J.M. Alimi

9 Septembre 2024 - Royaumont

A Brief Introduction to **Cosmology** and IA

Cosmology.

The origin,

The structure and

The evolution of the universe.

The problems are numerous:

- Nature of gravity
- Physics of the primordial universe
- Nature and dynamics of the invisible components dark matter and dark energy
- Formation of cosmic structures, galaxy formation...

In all cases, the physics (known or unknown) involved are complex and non-linear: Design, Resolution and Interprétation

A Brief Introduction to Cosmology and **IA**

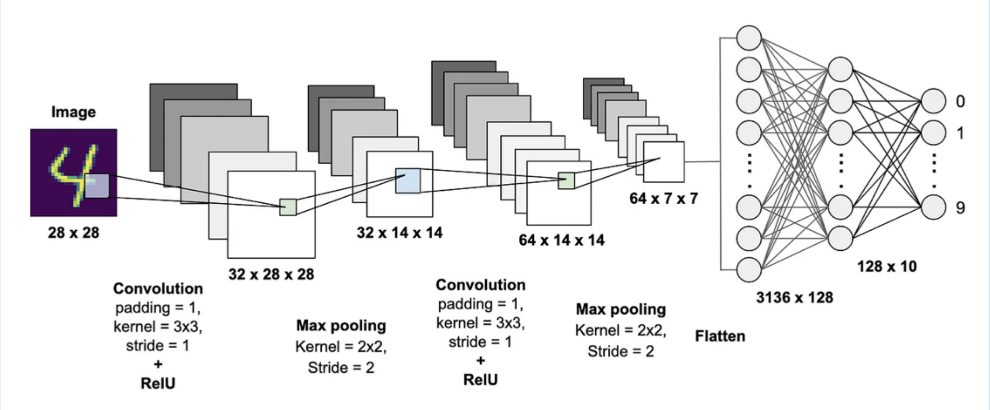
The application of **Artificial Intelligence (AI)** to cosmology (astrophysics) is (first and foremost, most often) identified as a means of solving the major challenges linked to the processing of massive data (observational data (Euclid, LLST, SKA...) or digital data).

AI helps to analyze and model these data, often providing innovative solutions to **automate** numerous **cleaning, classification and recognition tasks** in the vast volumes of data produced by modern telescopes: galaxies/stars, galaxy classification, etc.

A Brief Introduction to Cosmology and **IA**

Pixelized image analysis / Image analysis by CNN
(Convolutional Neural Network)
(reduced computation, "deep" image structure)

Convolutional neural networks (Y. LeCun 1989)
(Convolution, Pooling)
Image recognition, image segmentation
Cleaning, Classification (2010)
"Turnkey solutions", Data volume management!



What is IA, or rather, how does it work?

The Standard definition of machine learning:

- A field of study that involves developing a program's ability to learn from "data".
- The aim is to improve a program's ability to perform tasks without first being "programmed" for them.
- A distinction is made between a training or learning phase and a test or production phase.

- **The following types of learning can be distinguished: supervised, unsupervised, reinforcement learning, etc.**
- Numerous applications.

AI is more than just a data processing tool, it can be a true heuristic partner. It gives us the ability to explore complex theoretical models and generate and test hypotheses.

Machine learning as a way of thinking

Physics Inspired Machine Learning

Machine Learning and Cosmology

ML Procedure as a way of thinking: « *Physics Inspired Machine Learning* »

INPUT: DATA

- Observations, Experiments, Computations, Theories: "Data"
  - Numbers, Curves, Images, Videos…(measurements), Equations, Basic Principles, Symmetries, Patterns,…

NON LINEAR OPERATOR

OUTPUT: TARGET

Classification,
Regression,
Evolution,
Reproduction,
Inférence, ….

Sensitivity of the result according to the INPUT and the LEARNING PROCESS:
Quantity, Quality, Precision,
Physical Information,…
Non Physical Granularity of the Data …

Algorithm (DT, RF, NN, …)
Optimisation, Adaptability, Precision…

How to evaluate the quality and the precision of the output:
Metrics

Is such a conceptualization of any physical problem appropriate?
(even beyond physics?)

The construction of **the NL operator (fixed architecture)** is done "**today**" by successive **approximations (metrics)** thanks to a set of of **available initial and final data**.

Machine Learning and Cosmology

ML Procedure as a way of thinking: *« Physics Inspired Machine Learning »*

The various questions raised by the ML process are:

- adaptability to different problems
- quality of the results obtained
- interpretation of the results obtained
- ...

all come down to a cross-analysis of the four fundamental elements of the ML approach:

Data, NL operator, Prediction, Metrics.

Machine Learning and Cosmology

ML Procedure as a way of thinking: « *Physics Inspired Machine Learning* »

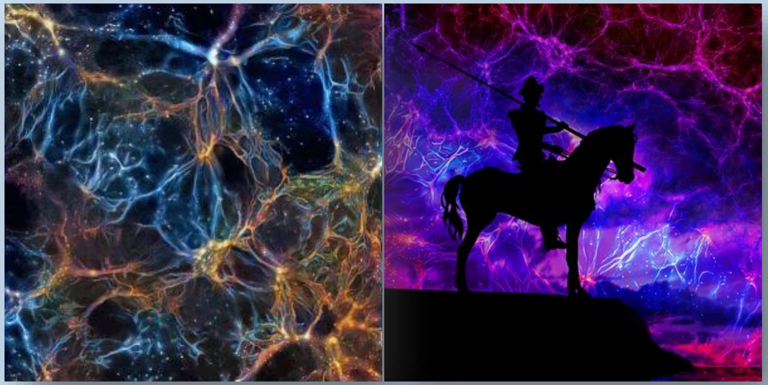
DATA:

Databases for numerical simulations (Cosmologies, Gravity, etc.) are multiplying (a prerequisite for collaboration with computing centers): Quijote, Camels, DustGrain, DEUS, ...
Quijote Simulations (Francisco Villaescusa-Navarro (Flatiron/Princeton))
https://quijote-simulations.readthedocs.io/en/latest/index.html

The Quijote simulations is a suite of more than 82,000 full N-body simulations that have been designed to accomplish two main goals:

- Quantify the information content on cosmological observables
- Provide enough statistics to train machine learning algorithms

2000 simulations correspondant à 2000 modèles cosmologiques (Ω_M, h, Ω_b, n_S, σ_8)



Inferring cosmology from
the spatial distribution of cosmic structures
(DM halos / galaxy clusters, galaxies...)

Machine Learning and Cosmology

DATA: Is it about processing as much (raw) data as possible?

Inferring cosmology from the spatial distribution of cosmic structures
(DM halos / galaxy clusters, galaxies...)
Inference and Linking Non-Gaussian statistics and non-linear dynamics.
(Optimized neural networks)

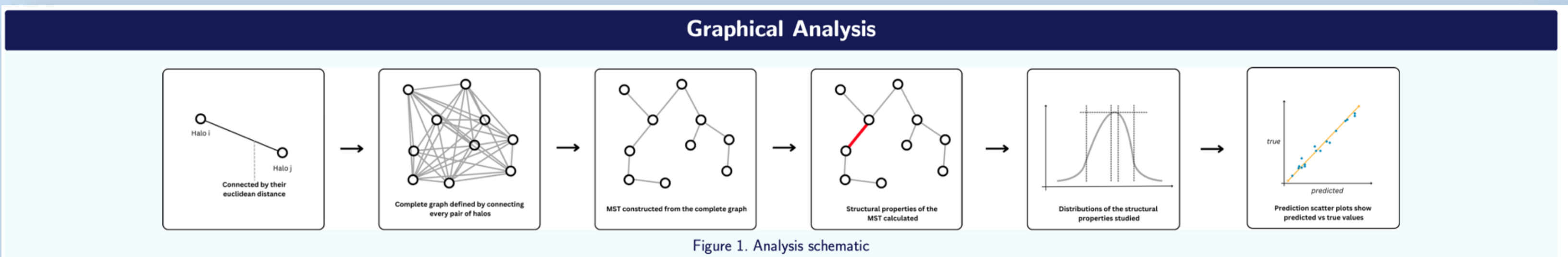
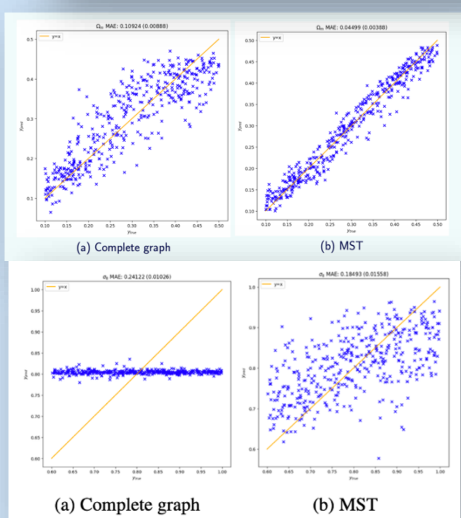


Figure 1. Analysis schematic

Ω_M

σ_8



…

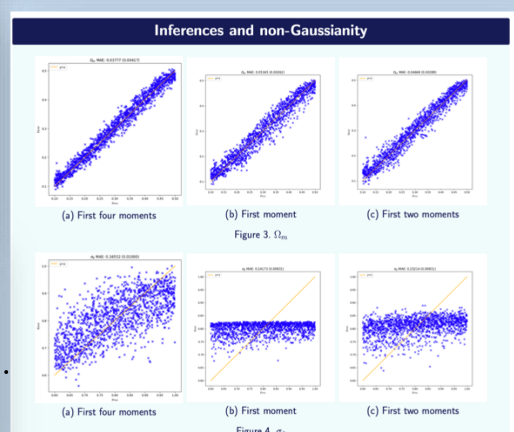


Figure 3. Ω_m

Figure 4. σ_8

…

The non-Gaussian character of the halo distributions that contains cosmological information.

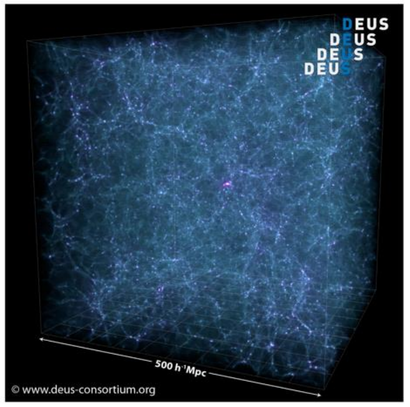
(A. Shankar & JMA 2024)

Machine Learning and Cosmology

ML Procedure as a way of thinking: *« Physics Inspired Machine Learning »*

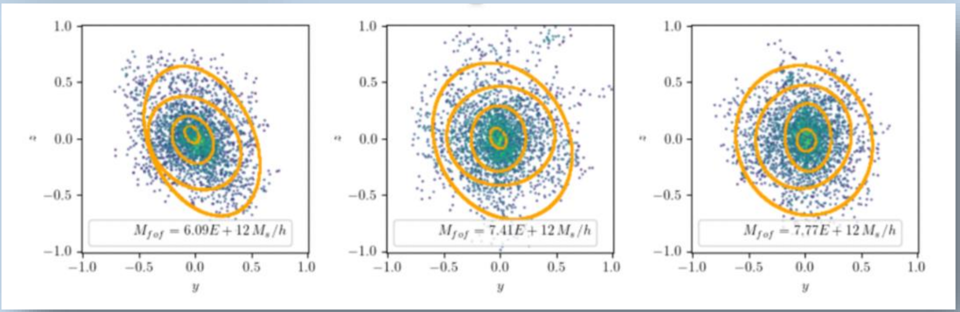
NL Operator: A multitude of algorithmic methods, elements of theory to determine which device is best suited to which problem ? **They're not just black boxes,** Not only neural networks, but also decision trees, random forests...

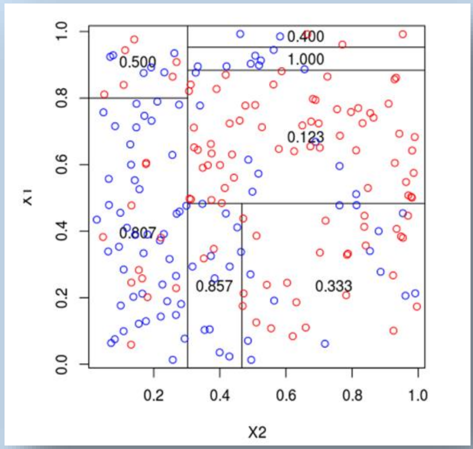
Recognize on a single DM halo the cosmological model in which it was formed



+300000 / Cosmological models. DM Halo is a complexe structure.! It forms inside the cosmic web; environmental effects, geometry of grav. collapse, dynamical status …

Decision Trees / RF Partitioning the attributes hyperspace in homogeneous regions wrt the target.





Machine Learning and Cosmology

NL Operator: Decision Tree/ Random Forest

Recognize on 1 single DM halo the cosmological model in which it was formed
Mass Profile, Shape Profile / Cosmology

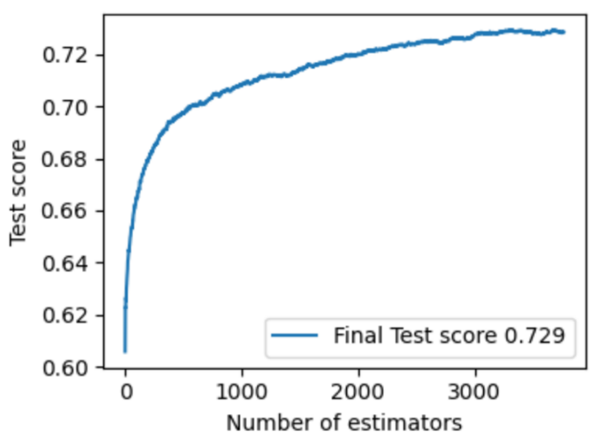
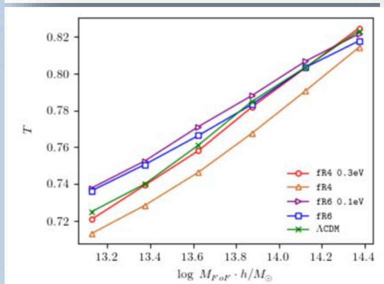
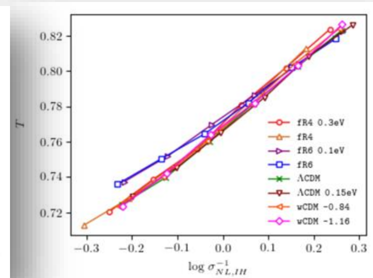


Fig. 4. Variation of the classification score (on the test set) in number of estimators in the ensemble of trees. Attributes of halos in the test set are mass, shape and velocity profiles $(M_d, a_d, b_d, c_d, a_d^V, b_d^V, c_d^V)_{d\in[0,24]}$. Early stopping criterion (see main text) yields the end of the learning as the score on a given learning sample of the training set reaches a plateau. If this out of bag sample and the test set are large enough, such a convergence should be concommitent with the convergence of the score on the test set. This appears as a plateau too giving a test score of 73 percent (around 3600 estimators).

	Mass	Position (a, b, c)	Velocity (a_V, b_V, c_V)
Mass	69% 3547	71% 4775	71% 3611
Position (a, b, c)		62% 1054	66% 1375
Velocity (a_V, b_V, c_V)			59% 1074

Table 2. Results of classification between ΛCDM and RPCDM halos, using different combination of features among mass (the sequence of $(M_\delta)_\delta$), semi-axis length (eigenvalues of shape tensors) and the eigenvalues of the velocity tensor. The presence of mass attributes is crucial to obtain a test score close to the score using all the attributes (73 percent).





...

On a large population of DM Halos (DEUS+ Dustgrain):
A new fundamental cosmological invariance in GR models and a new probe of MG theories

(R. Koskas & JMA 2024)

Machine Learning and Cosmology

NL Operator: Neural Networks:

Perceptron, MultiLayerPerceptron

1. Initialize pi weights:
2. Weighted sum calculation:

$$S = \sum (w_i \times x_i) + b$$

3. Apply activation function NL f(S) (sigmoIde, tanh, Relu...). The output becomes

$$\hat{y} = f(S)$$

4. Update weights:

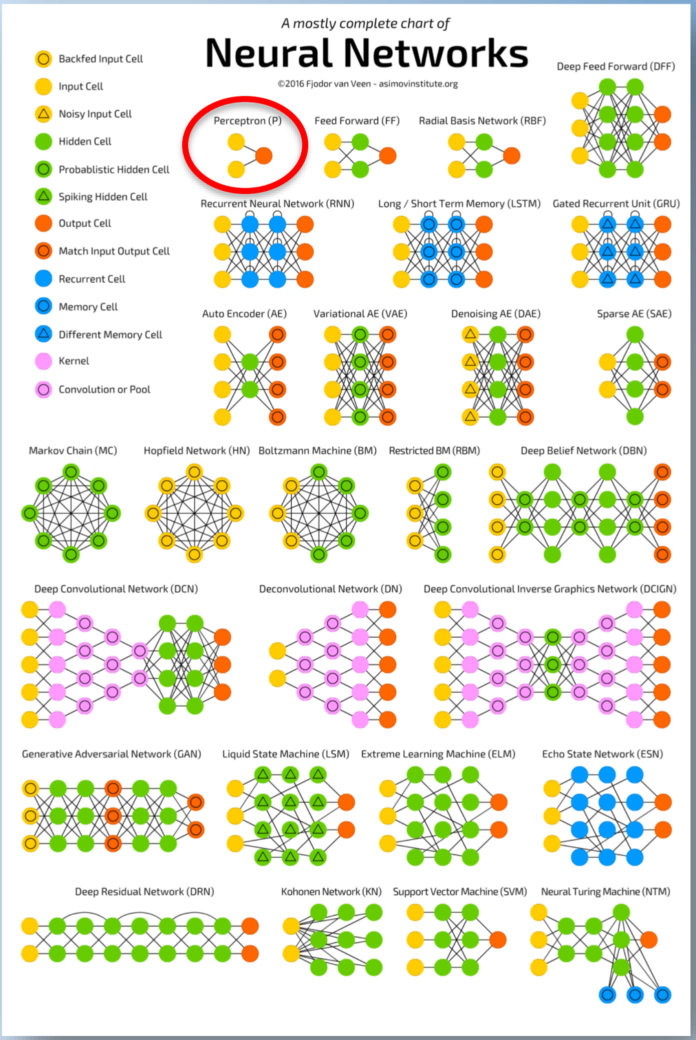
If the output predicted by the perceptron differs from the expected output, the weights are updated according to the error.

$$w_i \leftarrow w_i + \eta \times (y - \hat{y}) \times x_i$$

η is the learning rate, y is the expected output, y^ is the predicted output.

5. Iterations and convergence:

Weight updating is repeated over several iterations (or epochs) until the weights converge and the error is sufficiently small.



Architecture and Hyperparameters

Machine Learning and Cosmology

ML Procedure as a way of thinking: *« Physics Inspired Machine Learning »*

Inferring cosmology from the structure of the cosmic web

Cosmic Field NL Dynamics - Cosmic Web (Geometric Collapse Regions)

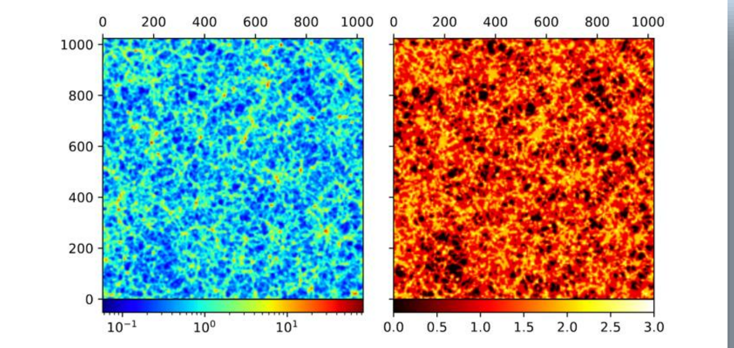
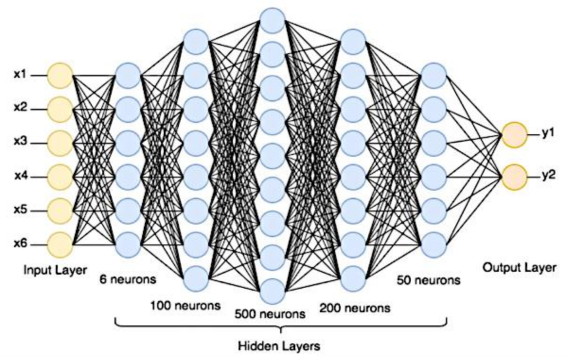
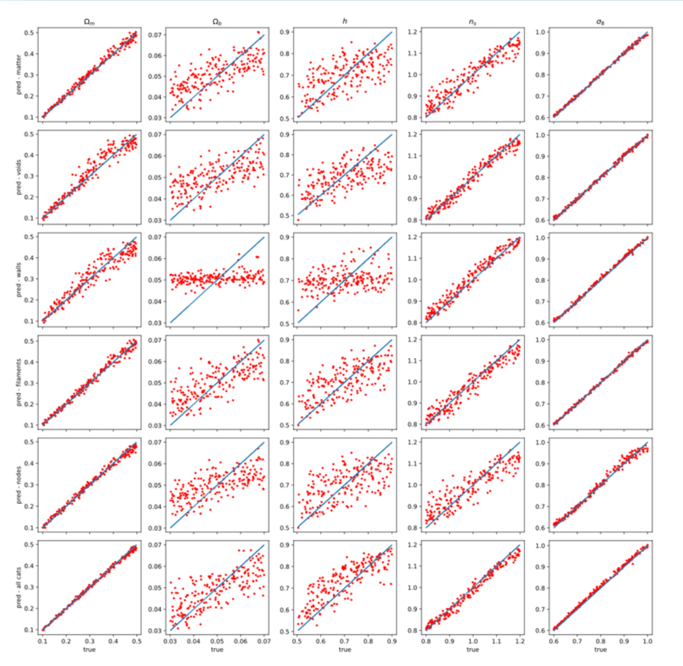


Fig. 1. *Left*: Thin slice $1000 \times 1000 \times 0.9765 (Mpch^{-1})^3$ of density field for fiducial cosmology at redshift z=0 for a smoothing length of $2Mpch^{-1}$, the colormap represents the values of the density field in each pixel. *Right:* a Corresponding number of positive tidal field's eigenvalues (*cf.* Section 2), voids, walls, filaments, and nodes correspond respectively to 0, 1, 2, and 3 positive eigenvalues.

Cosmological inference (Fields, Voids, Walls, Filaments, Nodes)
Physical observables: $_{PdFi}(\delta)$,$_{Pi}(k)$,...





Deep neural networks
Bayesian Optimization of Hyperparameters

Machine Learning and Cosmology

ML Procedure as a way of thinking: *« Physics Inspired Machine Learning »*

Inferring cosmology from the structure of the cosmic web

Cosmic Field NL Dynamics - Cosmic Web (Geometric Collapse Regions)

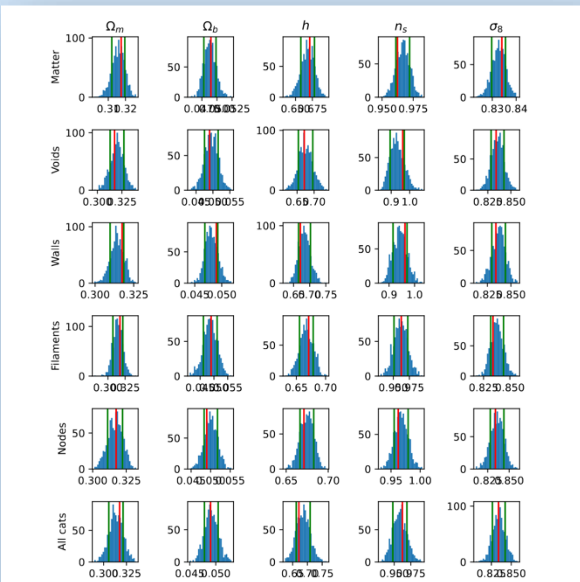
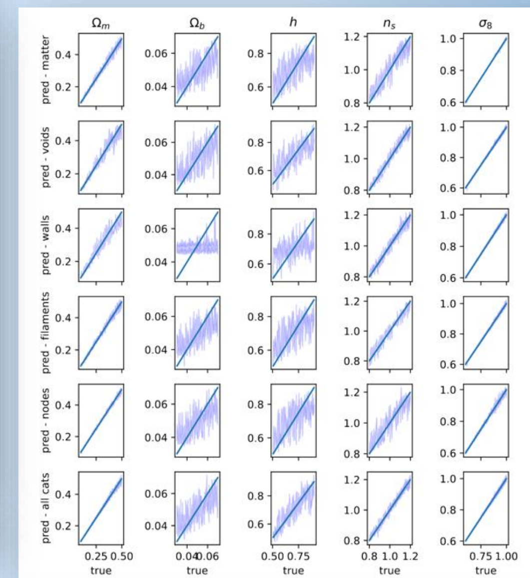


Fig. 7. Prediction intervals for the fiducial model using different observables, red vertical lines correspond to the true values while green borders indicate the $1-\sigma$ intervals.

Type		σ_{Ω_m}	σ_{Ω_b}	σ_h	σ_{n_s}	σ_{σ_8}
$P(k)$	Matière	0.0098	0.0038	0.0354	0.0238	0.0082
	Vides	0.0173	0.0055	0.0354	0.0392	0.0076
	Murs	0.0078	0.0012	0.0198	0.0133	0.0074
	Filaments	0.0092	0.0039	0.0308	0.0154	0.0061
	Noeuds	0.0061	0.0035	0.0197	0.0288	0.0088
$P(\delta)$	Matière	0.0168	0.0058	0.1244	0.0313	0.0181
	Vides	0.0196	0.0078	0.1129	0.1161	0.0201
	Murs	0.0126	0.0064	0.0989	0.0681	0.0196
	Filaments	0.0182	0.005	0.1064	0.0499	0.0157
	Noeuds	0.0132	0.0059	0.0694	0.0455	0.0177
Stat. Comb.	Moments Bruts	0.0162	0.0074	0.0945	0.0454	0.0099
	Moments	0.0244	0.0059	0.0633	0.044	0.017
	Pearson	0.0121	0.0057	0.0879	0.0331	0.0096
	$P(\delta)$	0.0146	0.0069	0.0618	0.0361	0.0102
	$P(k)$	**0.0049**	**0.001**	**0.0141**	**0.0098**	**0.0029**
Fisher - $P(k)$ individuel	Matiere	0.0969	0.0413	0.5019	**0.0132**	0.0046
	Vides P(k)	0.381	0.0234	0.295	0.2962	0.0466
	Murs	0.0752	0.0419	0.4902	0.3852	0.0903
	Filaments	0.0320	0.0189	0.2444	0.2546	0.0230
	Noeuds	0.0971	0.0481	0.6142	0.6004	0.2006
	$P_{comb}(k)$	0.0126	0.0093	0.0793	0.0319	0.0046
	WST	0.014	0.012	0.104	0.031	**0.001**
	HMF	0.007	0.037	0.230	0.100	0.007

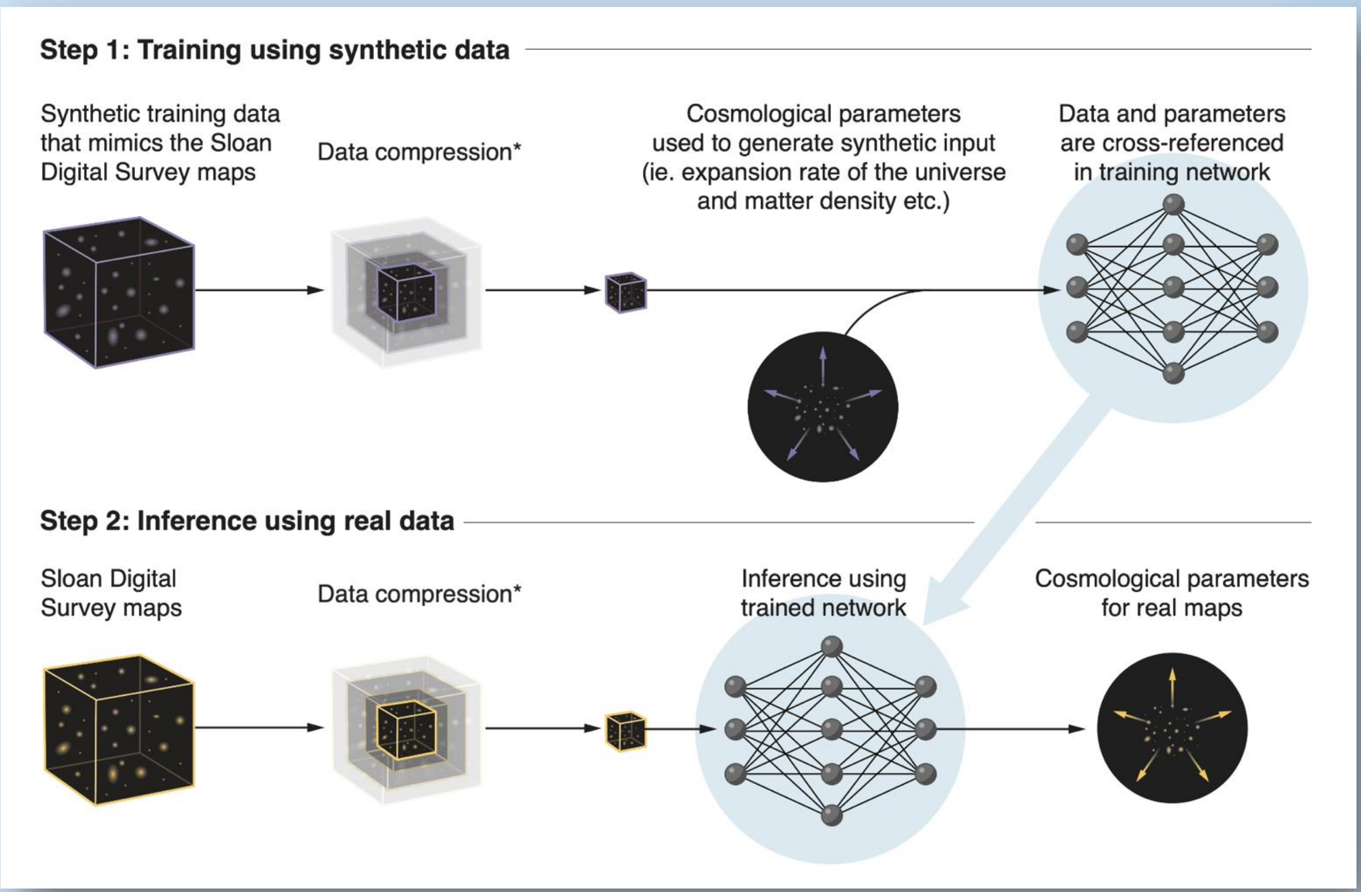


Extraction of cosmological information is powerful!
Higher accuracy than any other method on cosmology prediction (in real space)!
Cosmology and Collapse Dimensionality (dynamical approximations)

(M. Shalak & JMA 2024)

Exploring cosmological models with AI

Applications: How to compare the predictions of theoretical cosmological models with observational data



Lemos et al (2023) arxiv:2310.15256

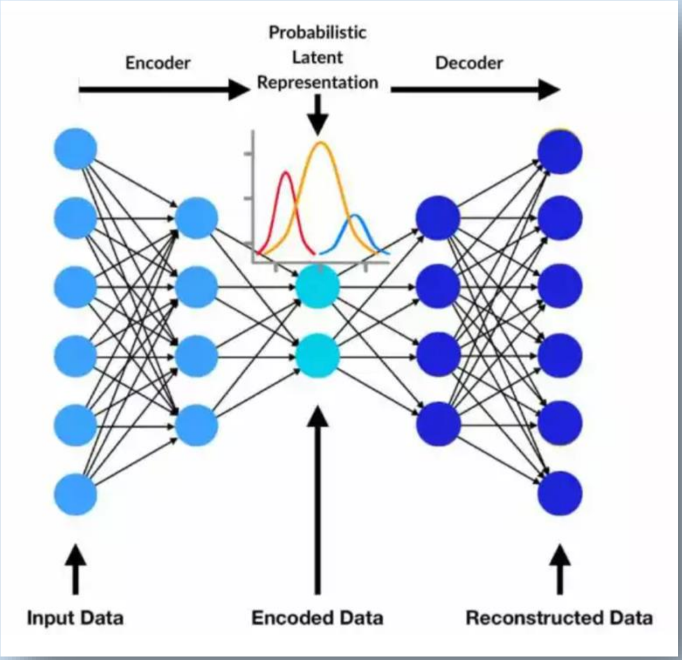
Machine Learning and Cosmology

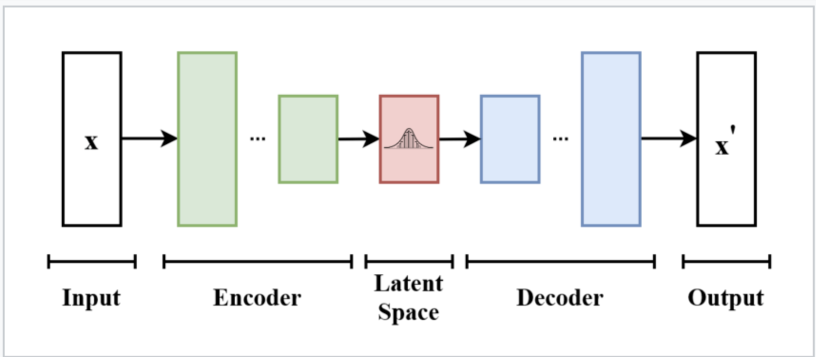
ML Procedure as a way of thinking: ***« Physics Inspired Machine Learning »***

NL operator

Generative AI: VAE (**Variational Auto-Encoder)**

Encoder-Decoder, **Latent Space**.





The basic scheme of a variational autoencoder. The model receives x as input. The encoder compresses it into the latent space. The decoder receives as input the information sampled from the latent space and produces x' as similar as possible to x.

Properties of latent space
(statistical, dynamical, geometrical...)

Cosmology and Machine Learning

Generative Models: DNN

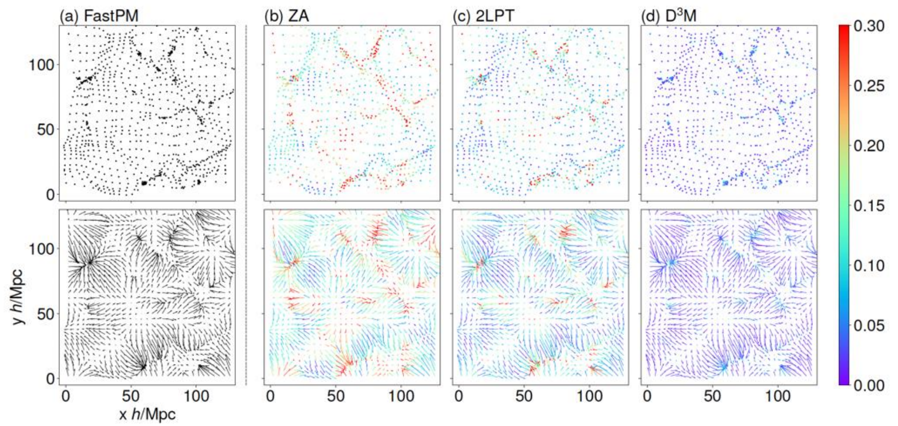
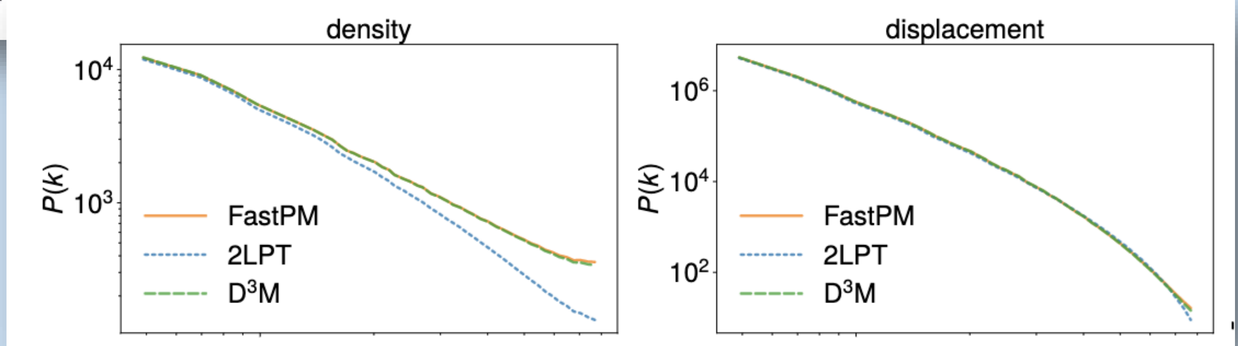


Fig. 2. The columns show 2-D slices of full particle distribution (top) and displacement vector (bottom) by various models, from left to right:
(a) FastPM: the target ground truth, a recent approximate N-body simulation scheme that is based on a particle-mesh (PM) solver ;
(b) Zel'dovich approximation (ZA): a simple linear model that evolves particle along the initial velocity vector;
(c) Second order Lagrangian perturbation theory (2LPT): a commonly used analytical approximatation;
(d) Deep learning model (D^{3}M) as presented in this work.
While FastPM (A) served as our ground truth, B–D include color for the points or vectors. The color indicates the relative difference $(q_{model} - q_{target})/q_{target}$ between the target (a) location or displacement vector and predicted distributions by various methods (b-d). The error-bar shows denser regions have a higher error for all methods. which suggests that it is harder to predict highly non-linear region correctly for all models: D^{3}M, 2LPT and ZA. Our model D^{3}M has smallest differences between predictions and ground truth among the above models (b)-(d).



… …

« Learning to Predict the Cosmological Structure Formation »
« Deep Density Displacement Model (D3M) / Neural Network »

Siyu He et al. 2019

Cosmology and Machine Learning

Generative Models: GAN and Wasserstein distances

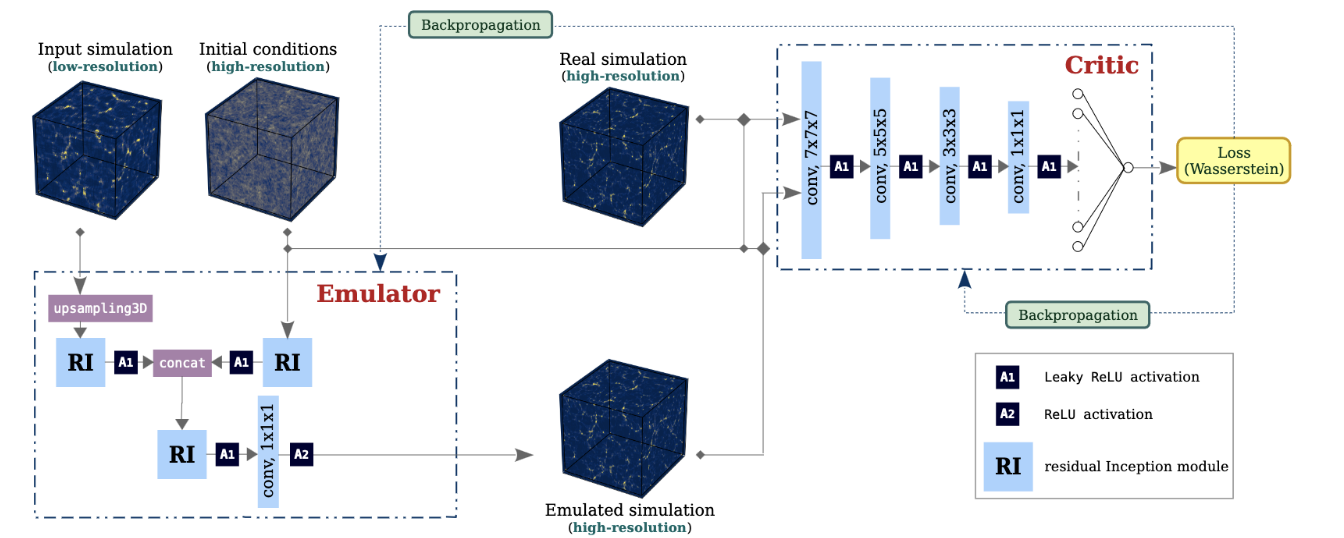
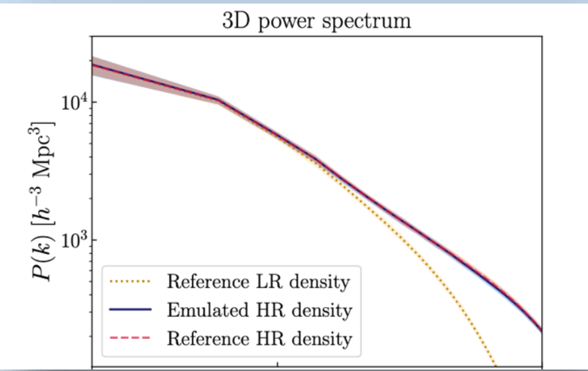


Figure 1. Schematic representation of the super-resolution emulator implemented in this work. The emulator approximates the underlying mapping of the distribution of low-resolution density field to high-resolution structures, with the input initial conditions providing an informative prior distribution from which the emulator constructs the fine structures, to yield a super-resolution field. The difference between the output of the critic for the real and emulated density fields, conditional on the initial conditions, is the approximate Wasserstein distance, which is minimized to fit the super-resolution N-body emulator.



« Super-resolution emulator of cosmological simulations using deep physical models »

Doogesh Kodi Ramanah et al 2020

Cosmology and Machine Learning

Generative Models: GAN

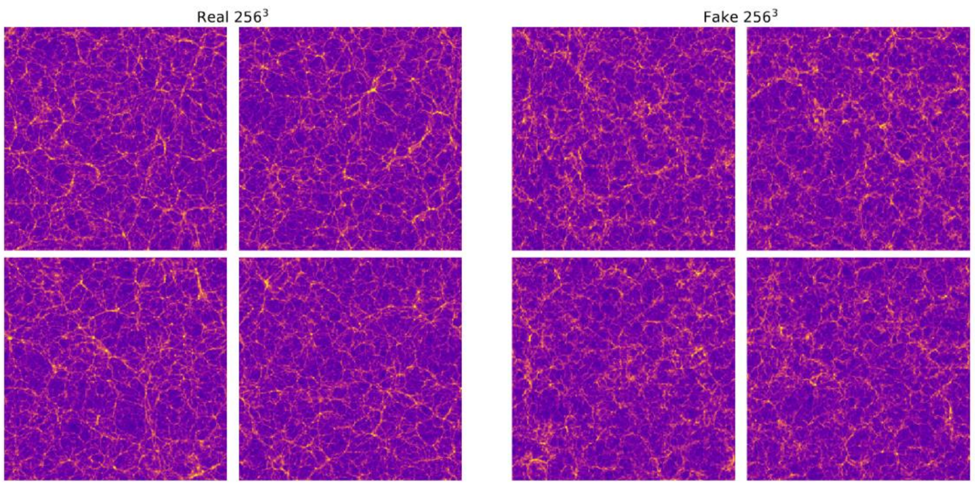


Figure 11 Middle slice from real and generated 256^3 samples. The GAN-generated samples are produced using the full multi-scale pipeline. Videos:
- 32-scale: https://youtu.be/uLwrF73wX2w
- 64-scale: https://youtu.be/xI2cUuk3DRc
- 256-scale: https://youtu.be/nWXP6DVEalA

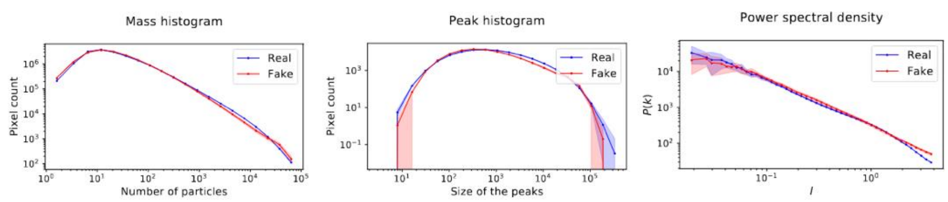


Figure 12 Summary statistics of real and GAN-generated 256^3 images using the full multi-scale pipeline. The power spectrum density is shown in units of h Mpc^{-1}, where $h = H_0/100$ corresponds to the Hubble parameter.

« Cosmological N-body simulations: a challenge for scalable generative models »

Perraudin et al. 2019

Cosmology and Machine Learning

Generative Models: CNN

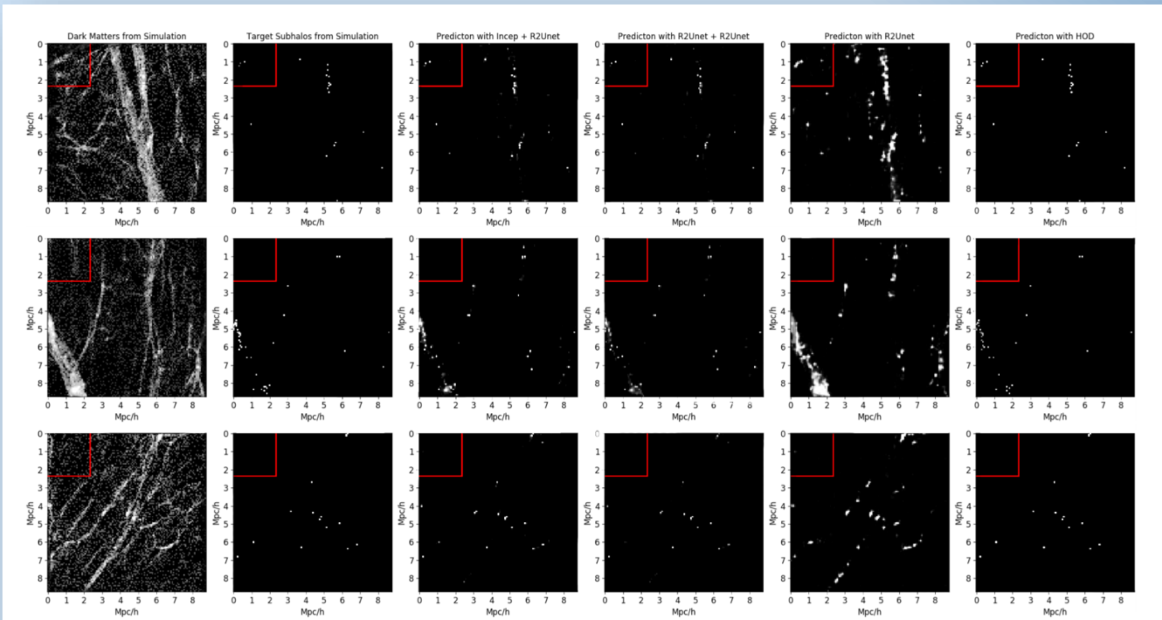
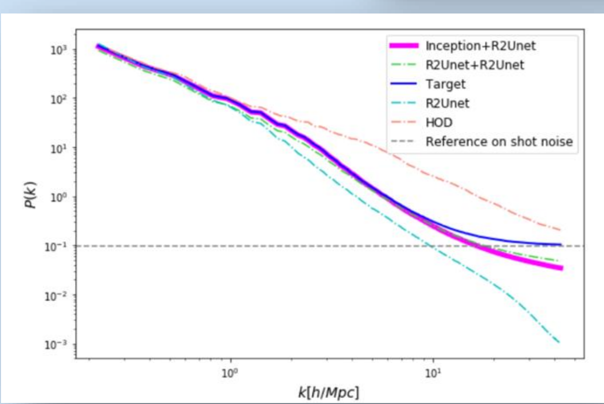


Figure 9: Visualization of slices of the simulations: first column are dark-matter halos, second column are the corresponding target galaxies. 3d and 4th columns are predictions from our two-phase models, 5th from a single-phase classifier, and last column are HOD predictions. Red square represents the size of the boxes taken as input by our models.



« From Dark Matter to Galaxies with Convolutional Networks »

Xinyue Zhang et al 2019

Exploring cosmological models with AI

Intelligent management of training data

Intelligent prediction management

Physic Inspired models in Machine Learning, introducing physical constraints explicitly in the choice of initial data, NL operator architecture, metric form, objectives.

(Semi-)analytical physical problem solving

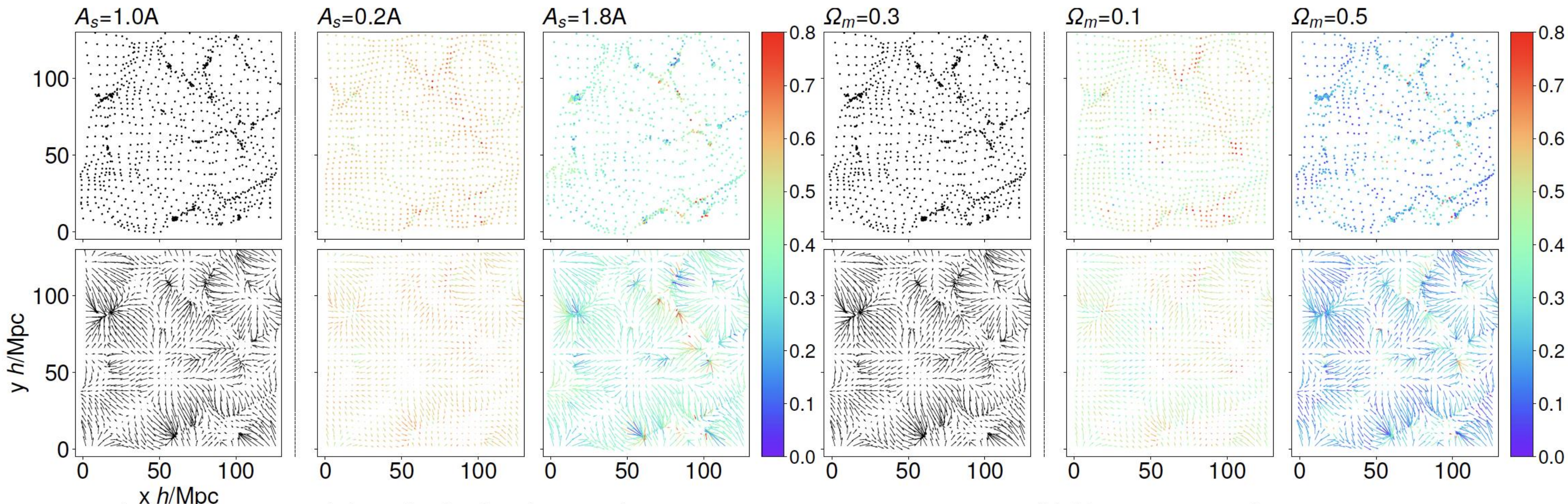
Conclusion: AI as a partner in cosmology

- **AI helps to process massive data and model the universe and its components.**

- **But it is also a heuristic partner, helping to explore and improve theoretical models and our understanding of the physics underlying cosmological properties.**

- **Future prospects: AI and cosmology synergy for new discoveries.**

Cosmology and Machine Learning



(a) Changing primordial amplitude of scalar perturbations

(b) Changing matter density