# **Deep Learning and Numerical Simulations**

(in cosmology and galaxy formation)

Rupert Croft (Carnegie Mellon)

Tiziana Di Matteo Yin Li Xiaowen Zhang Patrick Lachance Yueying Ni Simeon Bird Yu Feng Meris Sipp Ankita Dasgupta Runxuan Wang

#### Definitions:

Numerical Simulation: computer model starting from cosmological initial conditions and including at least the physics of gravity and dark matter.

Deep Learning: branch of Machine Learning that uses Neural Networks models that learn from large amounts of training data and then make predictions Imagine the perfect simulation + deep learning machine... what do we want from it?





simulates huge volumes at high resolution



trustable with uncertainty quantitication



simulates all physics + physics-based



fast with minimal storage needs



simulates all models





perfect subgrid



allows inference



low energy use



explainable

Neural networks make everything better...



(except spelling, apparently) ... where shall we add Deep Learning to our simulation workflow?

NN have been tried everywhere in simulation workflow:



NN have been tried everywhere in simulation workflow:



NN have been tried everywhere in simulation workflow:



OK, NN, write me my NSF grant report, but it's got to rhyme:

# grant proposal the mc



ChatGPT

With NSF's embrace, in astrophysics we delved, Deep learning and cosmos, our realms beautifully melded. Simulations grand, galaxies in formation danced, Gravity's embrace, hydrodynamics enhanced. subgrid Stars birthed in the cosmic ballet, black holes in cosmic trance, A symphony of data, inferences did advance. Simulation-based, our insights did gleam, Comparing to observations, like a cosmic dream. Inferred parameters, astrophysical delight, NSF's support, our journey took flight. Cosmological mysteries, unraveled in the night, With gratitude profound, our grant report takes flight.

No no – I need it written in the style of a pirate...



#### 🚳 ChatGPT

ې prc Arr matey, with the wind in our sails and the hearty support of the NSF, we embarked on a cosmic adventure in the realm of astrophysics and deep learning! With the precision of a sextant, we crafted grand computer simulations of galaxy formation, navigatin' through the treacherous waters of gravity, hydrodynamics, star birth, and black holes. Yarrr, our ship of knowledge set sail on the vast cosmic sea, employin' simulation-based inference like a trusty compass. We compared our digital treasures to the celestial charts, and lo and behold, we deduced the values of cosmological and astrophysical parameters as if decipherin' a long-lost map. The NSF be our steadfast anchor, and with their support, we have unveiled cosmic secrets as if plunderin' a hidden chest of astrophysical gold. Aye aye, the stars themselves be singin' our shanty of discovery!

Let's next compare recent DL developments to those in cosmological simulations:

#### AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance

Handwriting recognition 🔵 Speech recognition 😑 Image recognition 🛑 Reading comprehension

Language understanding 🔵 Common sense completion 🥚 Grade school math 🛑 Code generation



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point", which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: ContextualAI



https://github.com/florent-leclercq/Moore\_law\_cosmosims

https://www.illustris-project.org/

# The number of data points used to train AI models has increased dramatically over the last seventy years

Number of data points used to train notable AI models



Training data size refers to the amount or quantity of data that is used to train an AI model, indicating the number of examples or instances available for the model to learn from. Each domain has a specific data point input unit, such as images to train vision models, words for language models and timesteps for games models. This means systems can only be compared directly within the same domain.

Chart: Will Henshall for TIME · Source: Epoch via Our World in Data

#### TIME

DL model training dataset sizes have caught up with the largest simulations:



#### 2D vs 3D

There is a vast literature concerning DL and 2D images.

The ill-fated Metaverse stimulated 3D generative DL for a while.

2



Generative Adversarial Network (Goodfellow 2014)

#### First Generative AI cosmological sims (of 2D slices)



Rodriguez+ 2018 (arXiv:1801.09070)



#### First Generative AI cosmological sims (of 2D slices)



Rodriguez+ 2018 (arXiv:1801.09070)

# N-body simulation samples Generative Adversarial Network (GAN) samples

Rodriguez+ 2018 (arXiv:1801.09070)



#### First Generative AI cosmological sims (of 2D slices)



Other generative DL architectures can be used e.g., Denoising Diffusion (Sohl-Dickstein+ 2015), or Poisson-Flow (Xu+ 2022, arXiv:2209.11178)

#### First Generative AI cosmological sims (of 2D slices)



Rodriguez+ 2018 (arXiv:1801.09070)

#### Generator vs emulator (field level):

## Correlation coefficient of individual Fourier modes



#### Generator vs emulator (field level):

### Correlation coefficient of individual Fourier modes



remember that both generator and emulator need to be trained with data

k (h/Mpc)

Apart from field level, can have a NN emulator for summary statistics:



CAMELS project

Cosmology and Astrophysics with MachinE Learning Simulations

www.camel-simulations.org/

(see e.g., talk by Daniel Angles-Alcazar this morning) First 3D **<u>particle-based</u>** emulation of simulations (He+ 2019 <u>arXiv:1811.06533</u>)

"Learning structure formation"





#### Reminder: what do we want from our DL simulator?





simulates huge volumes at high resolution



fast with minimal storage needs

How to achieve this:

- Run a computationally cheap low resolution simulation
- Use DL to add in the finer details

roughly balance time taken



#### 3D voxel based super-resolution emulator: Kodi Ramanah+ 2020 (arXiv:2001.05519)



(factor of 2 increase in spatial resolution)

(This is GAN, but DDM also possible: Schanz+ 2023 arXiv:2310.06929 Rouhiainen+ 2023 arXIv: 2311.05217)



#### Li+ 2021 (arXiv:2010.06608)

- Particle-based super-resolution
- Output format same as simulation
- Factor of 8 spatial enhancement, 512 in mass
- 5120<sup>3</sup> SR simulation takes 16 hours on 1 GPU

#### SR simulation can generate subhalos:

low resolution N-body simulation

high resolution N-body simulation for comparison LR  $M_h = 1 \times 10^{13} M_{\odot}$  $M_h = 6 \times 10^{13} M_{\odot}$  $M_h = 4 \times 10^{13} M_{\odot}$  $M_h = 3 \times 10^{14} M_{\odot}$ z=2z=0z=2z=0HR  $M_h = 6 \times 10^{13} M_{\odot}$  $M_{\rm D}=4\times 10^{13}M_{\odot}$  $M_h = 1 \times 10^{13} M_{\odot}$  $M_h = 3 \times 10^{14} M_{\odot}$ z=2 z=0z=0SR  $M_h = 1 \times 10^{13} M_{\odot}$  $M_h=6\times 10^{13}M_\odot$  $M_h = 4 \times 10^{13} M_{\odot}$  $M_h = 3 \times 10^{14} M_{\odot}$ 

z=0

small halos are in different places in HR and SR

Ni+ 2021 (arXiv:2105.01016)

super-resolution

A use case for AI SR: simulations of Fuzzy Dark Matter that incorporate quantum pressure



low res





hi res





super res



these SR simulations don't yet use quantum pressure in the training set but if they did it could massively speed up the extremely time consuming FDM runs

Sipp, Lachance+ 2023 (arXiv:2210.12907)

Can train the NN using data from multiple redshifts + the redshift itself ("style")



style

**Z**2

if we keep the same noise, the time evolution is consistent

#### Zhang+ 2023 (arXiv:2305.12222)

 $z = 1.15 \sim 0.45$ 

Lbox = 5 Mpc/h

~ 0.91

#### power spectrum of field vs true HR simulation at different redshifts

10<sup>0</sup>

*k* [*h* Mpc<sup>-1</sup>]

10<sup>1</sup>



0.8

 $10^{-1}$ 

a forming cluster in the SR simulation (with subhalo tracks)

#### Example halo merger trees from the HR and SR simulations



Zhang+ 2023 (arXiv:2305.12222)

#### Next step: SR emulator (see Xiaowen Zhang and Patrick Lachance's poster)



correlation coefficient of individual Fourier modes between SR and HR



As the output of the SR has the same format (MP-Gadget) as a simulation, we can evolve it forward with and N-body code to see what happens:





#### Dagupta+ 2023



The work we have talked about was all dark matter + gravity only, but super-resolution can work for hydrodynamics too:

simulates all physics



Intergalactic medium eulerian hydrodynamics



The Lyman-alpha forest is a simple physical system which can be simulated at the field level using SR techniques. Here are summary statistics from Lyman-alpha spectra.



Jacobus+ 2023 arXiv:2308.02637



What about smaller scale hydrodynamics?

Turbulence?

perfect subgrid



What about smaller scale hydrodynamics?

Turbulence?

Clard-Mobil OCORACLES

Honda 2023 Formula One winning engine modelled with Physics Informed Enhanced Super-Resolution GAN (PIERSGAN: Bode+21 d-nb.info/123852396X/34)

perfect subgrid



air-fuel fraction

dissipation rate

- 1. Use the PIESRGAN to reconstruct  $\Phi^n_R$  from  $\Phi^n_{LES}$ .
- 2. (Only for nonuniversal quantities) Use  $\Phi_{\rm R}^n$  to update the scalar fields of  $\Phi$  to  $\Phi_{\rm R}^{n;{\rm update}}$  by solving the unfiltered scalar equations on the mesh of  $\Phi_{\rm R}^n$ .
- 3. Use  $\Phi_{\rm R}^{n;{\rm update}}$  to estimate the unclosed terms  $\Psi_{\rm LES}^{n}$  in the LES equations of  $\Phi$  for all fields by evaluating the local terms with  $\Phi_{\rm R}^{n;{\rm update}}$  and applying a filter operator.
- 4. Use  $\Psi_{\text{LES}}^n$  and  $\Phi_{\text{LES}}^n$  to advance the LES equations of  $\Phi$  to  $\Phi_{\text{LES}}^{n+1}$ .

For each timestep, reconstruct fine scales from Large Eddy Simulation (LES) and then advance time evolution of PIERSGAN turbulence and combustion tracks direct numerical simulation (DNS)



universal character on smallest scales



Hirashima + (2023) arXiv: 2311.08460

Gibbs sampling the super-resolution SN grid simulation with particles gives correct thermal energy vs hires:

#### low res vs high res

reconstructed vs high res



Hirashima + (2023)



allows inference



trustable with uncertainty quantitication

#### Simulation-Based Inference and Deep Learning



use neural networks to approximate the function that relates an observable (with unknown likelihood) to the values of the parameters

simulated data

mode

amortized

posterior

posterior

Sc

density

(e.g. normalizing flows)

(Plot: Avecilla+ 21 arXiv:2021.09.30.462581)

#### An example of SBI: SIMBIG



0.45

Test of robustness using different forward models in SIMBIG:



posterior mean - true value of parameter

different forward models -

Abacus Summit, Quijote, vary HOD

remove bias by cutting k>0.3 h/Mpc

Régaldo-Saint Blancard + 2023

#### Variation of forward models with cosmological hydrodynamics: CAMELS



 $\log(\Sigma/\Sigma_0)$ 

4

Cosmology and Astrophysics with MachinE Learning Simulations

#### CAMELS project

#### www.camel-simulations.org/

(see e.g., talk by Daniel Angles-Alcazar this morning)



6

5

7

CAMELS test of robustness: recovery of cosmological parameters after training on a different forward model:



Villaescusa-Navarro+ 2021 (arXiv:2109.09747)



 $\Omega_{\mathsf{M}}$  from a single galaxy

bias caused by baryonic effects on dark matter being different in different forward models Inferring the WDM mass from 2d slices with a CNN:



Rose+ 2023 (arXiv:2304.14432)

robustness- varying cosmology, simulation resolution, image resolution



#### Getting DM from stellar field with diffusion model: Park+ 2023 (arXiv:2311.08558)



input

true dark matter field

DDM dark matter field

marginalize over astrophysics –but as usual need lots of training data

#### SBI from 1D data – getting temperature from thermal smoothing of Lya forest:





Wang+ 2022 (arXiv:2112.05721)

Nayak+ 2023 (arXiv:2311.02167) grid of training sims

Field level does factor of factor 1.71 better than P(k)+PDF at inferring therrnal parameters



Again: do we have enough training data?

doesn't include systematics, both astrophysical and instrumental... For inference we need to compute gradients with respect to inputs. Do this for all operators of a simulation and we have a differentiable simulation.

DL backpropagation then lets gradient information flow from a simulator into an NN and vice versa

Example: Flow PM, Modi 21+ (arXiv:2010.11847)

A TensorFlow enabled Particle-Mesh code running on GPUs



reconstruction of the initial field

- Both analytic differentiation and automatic differentiation backpropagate the gradients through the whole history, thus requires saving the states of a simulation at all time steps in memory.
- A solution is to use the adjoint variables, are dual to the regular variables but evolve backwards in time.
- To compute the gradients we only need to evolve a simulation forward, and then backward jointly with its dual adjoint equations.



We haven't really talked about one side of DL problems:

We also wanted our DL+simulation machine to be/have:



trustable with uncertainty quantitication



explainable



low energy use



#### **Responsible AI in the Natural Sciences** A hybrid mini workshop

#### May 8th, 2023, 1 - 5:30pm | Gates Hillman Center Room 6115 | Zoom

#### We invite researchers in:

#### To discuss:

Chemistry +Everyone welcome

**Ethics of AI** 

#### **Register Here**

Enter the Human vs **Generative AI contest** (with prizes!)













Steinn Sigurdsson Savannah Thais Columbia. DSI Penn State, arXiv



NSF AI Planning Institute For Data-Driven Discovery in Physics



#### (sites.google.com/view/rainscmu)



Hoda Heidari CMU, Responsible Al Ahmad Tafti Pitt, HexAl

Block Center for

**Technology and Society** 

Hima Lakkaraju Harvard, TrustML



#### AI generated

or

real?



#### AI generated

or

real?



#### AI generated

or

real?

#### Answers:





AI generated

AI generated



low energy use Astrophysics simulations already use enormous amounts of energy: DL could make this worse.

Greenhouse gases associated with astronomy:



Stevens+ 2020 (for Australian astronomers)

: 400 million core hours for 700 researchers

Astrophysics simulations already use enormous amounts of energy: DL could make this worse.

Greenhouse gases associated with astronomy:



Supercomputing use per average astronomer equivalent to driving three Ford F150s

French translation: 39 Citroen 2CV (based on gas mileage and US vs France average miles)



There is room to improve: AlphaGo uses 50,000x more energy than Lee Sedol, World Go Champion

power consumption:

10<sup>6</sup> watts





power consumption:

20 watts

How to reduce carbon footprint of DL:

- Use pre-trained models as much as possible
- Track CO<sub>2</sub> use with a tool like CodeCarbon (<u>mlco2.github.io/codecarbon</u>)
- Train models during the day when energy is cleaner (more solar).
- Don't grid search hyperparameters- use Bayesian optimization (e.g., using RayTune, <u>ray.io</u>)

#### The Future:

• What do we use Generative simulations for ?

Best use cases? mocks, subgrid

- SBI measure parameters, but also extract knowledge astrophysics parameters are a start
- negative outlook: what about data, training?

• positive outlook: we as scientists can help – analytic models – for summary statistics but also for some fields e.g., EFT.