Scientific AI for Physics Workshop

Tuesday 4 November 2025 - Friday 7 November 2025

Instituto Principia

Book of Abstracts

Contents

Luidhy Santana-Silva (CBPF): From Pixels to Classes: Deep Learning Classification of 30 Million Galaxies in the DECam Local Volume Exploration Survey 1

TBA 1

Gustavo Dalpian Universidade de São Paulo): Learning from machine learning: discovering new science for materials 1

Pedro da Costa Huot: Study of hypernuclei identification in ultra-relativistic heavy-ion collisions using a machine learning approach 1

TBA 1

Lectures 1

Manuel Szewc (Universidad de San Martin): Machine learning hadronization 1

Luis Itza Vazquez-Salazar: Molecular Simulations with/without ML 2

TBA 2

TBA 2

TBA 2

TBA 2

Phelipe Antonie Darc de Matos (CBPF): Symbolic Regression Is All You Need: From Simulations to Scaling Laws in Binary Neutron Star Mergers 2

TBA 2

André Sznajder (UERJ): Ultrafast Jet Classification for the HL-LHC 2

Joaquin Armijo (Universidade de São Paulo): DeepLensingFlow: scored and flow-based networks for Weak-lensing map statistics 3

TBA 3

Luis Itza Vazquez-Salazar: Graph Diffusion Models 3

Daniel López-Cano (universidade de São Paulo): Semi-Supervised Domain Adaptation for Sim-to-Obs Astrophysics: DESI⟶J-PAS 3

TBA 3

Raphael Cóbe (UNESP): From Research to Impact: AI2's Model for Digital Social Innovation . . . 3

Rafael Bicudo Ribeiro (Universidade de São Paulo): Physically-oriented clustering of conformations in multi-molecule systems . . . 3

TBA . . . 4

From Pixels to Classes: Deep Learning Classification of 30 Million Galaxies in the DECam Local Volume Exploration Survey . . . 4

Symbolic Regression Is All You Need: From Simulations to Scaling Laws in Binary Neutron Star Mergers . . . 4

Study of hypernuclei identification in ultra-relativistic heavy-ion collisions using a machine learning approach . . . 5

DeepLensingFlow: scored and flow-based networks for Weak-lensing map statistics . . . 5

emi-Supervised Domain Adaptation for Sim-to-Obs Astrophysics: DESI⟶J-PAS . . . 6

Hadronization and Machine Learning . . . 6

Physically-oriented clustering of conformations in multi-molecule systems . . . 6

1

Luidhy Santana-Silva (CBPF): From Pixels to Classes: Deep Learning Classification of 30 Million Galaxies in the DECam Local Volume Exploration Survey

2

TBA

3

Gustavo Dalpian Universidade de São Paulo): Learning from machine learning: discovering new science for materials

4

Pedro da Costa Huot: Study of hypernuclei identification in ultra-relativistic heavy-ion collisions using a machine learning approach

5

TBA

6

Lectures

7

Manuel Szewc (Universidad de San Martin): Machine learning hadronization

8

Luis Itza Vazquez-Salazar: Molecular Simulations with/without ML

9

TBA

10

TBA

11

TBA

12

TBA

13

Phelipe Antonie Darc de Matos (CBPF): Symbolic Regression Is All You Need: From Simulations to Scaling Laws in Binary Neutron Star Mergers

14

TBA

15

André Sznajder (UERJ): Ultrafast Jet Classification for the HL-LHC

16

Joaquin Armijo (Universidade de São Paulo): DeepLensingFlow: scored and flow-based networks for Weak-lensing map statistics

17

TBA

18

Luis Itza Vazquez-Salazar: Graph Diffusion Models

19

Daniel López-Cano (universidade de São Paulo): Semi-Supervised Domain Adaptation for Sim-to-Obs Astrophysics: DESI→J-PAS

20

TBA

21

Raphael Cóbe (UNESP): From Research to Impact: AI2's Model for Digital Social Innovation

22

Rafael Bicudo Ribeiro (Universidade de São Paulo): Physically-oriented clustering of conformations in multi-molecule systems

23

TBA

24

From Pixels to Classes: Deep Learning Classification of 30 Million Galaxies in the DECam Local Volume Exploration Survey

Author: Luidhy Santana-Silva[1]

[1] *Centro Brasileiro de Pesquisas Fisicas*

Understanding the formation and evolution of galaxies over cosmic time requires a comprehensive analysis of their morphologies, especially because morphological features are strongly connected to other galaxy properties such as stellar populations, environments, and kinematics. However, the growing size of modern sky surveys has resulted in massive volumes of unclassified galaxies, making traditional morphological analysis increasingly difficult and time-consuming. In this work, we employ Convolutional Neural Networks (CNNs) to construct a morphological catalog from galaxy images obtained by the DECam Local Volume Exploration Survey (DELVE). To ensure a reliable training sample, we use a subset of 314,000 galaxies from the Galaxy Zoo DECaLS project (Walmsley et al. 2021), allowing us to define a robust training set of approximately 98,000 galaxies classified into four morphological classes: elliptical, lenticular, spiral, and mergers. We compare our CNN model to widely used architectures from the literature and show that our model outperforms them in both computational efficiency and classification accuracy across all morphological types. Our model achieves precision scores of 97%, 98%, 99%, and 92% for elliptical, lenticular, spiral, and merger galaxies, respectively. Applying this trained model to previously unclassified data from DELVE, we generate a morphological catalog covering approximately 13 square degrees down to r-band magnitude 21.5, comprising around 30 million galaxies. The completion and public release of this catalog will not only enhance our understanding of galaxy evolution but also provide a valuable resource for the broader astronomical community.

25

Symbolic Regression Is All You Need: From Simulations to Scaling Laws in Binary Neutron Star Mergers

Author: Phelipe Antonie Darc De Matos[1]

Co-authors: Bernardo Fraga [2]; Charles Kilpatrick [3]; Clecio R De Bom [2]; Gabriel S. Teixeira [2]

[1] *Centro Brasileiro De Pesquisas Físicas*

[2] *CBPF*

[3] *Northwestern university*

Gravitational wave sources with electromagnetic counterparts have highlighted the need for predictive, interpretable models linking the parameters of compact binary systems to post-merger remnants and mass outflows. In this work, we explore AI-driven symbolic regression (SR) frameworks to derive updated analytical relations for disk ejecta mass in binary neutron star mergers, trained on state-of-the-art numerical relativity simulations. Our method reveals a set of compact equations that outperform existing fitting formulae across multiple statistical metrics while remaining physically interpretable. Notably, SR also enables alternative predictor sets (e.g., $\{M_1, M_2, \tilde{\Lambda}\}$) that match or exceed the accuracy of models relying solely on compactness of the lightest neutron star (C_1), enabling new parameter constraints from electromagnetic observations. Unlike traditional black-box machine learning models, these closed-form expressions generalize robustly to regions of the parameter space not represented in the training data, offering a physics-informed tool for multimessenger observations and constraints on the neutron star equation of state.

26

Study of hypernuclei identification in ultra-relativistic heavy-ion collisions using a machine learning approach

Authors: Alexandre Alarcon Do Passo Suaide[1]; Pedro Da Costa Huot[2]

[1] *Universidade de Sao Paulo (BR)*

[2] *Universidade de Sao Paulo (USP) (BR)*

We present a study on the application of different machine learning algorithms for the identification of hypernuclei produced in heavy-ion collisions, particularly those with mass numbers A = 3 to A = 5. The study focuses on three supervised learning algorithms - Boosted Decision Trees, Support Vector Machines, and Artificial Neural Networks - which were trained to distinguish true hypernuclei candidates from combinatorial background using topological and kinematic variables of their decay products. The results demonstrate that these techniques significantly improve the background rejection of the selected hypernuclei candidates compared to traditional identification methods, thereby enhancing both the significance and precision of the measurements.

27

DeepLensingFlow: scored and flow-based networks for Weak-lensing map statistics

Author: Joaquin Armijo[1]

[1] *IFUSP*

Wide-field astronomical surveys provide unprecedented data that allow us to reconstruct the gravitational lensing maps tracing the large-scale distribution of matter in the Universe. In the weak lensing regime, these maps serve as a powerful probe of the Lambda-CDM cosmological model. However, their high dimensionality (millions of correlated pixels) poses significant challenges for traditional statistical analyses. In this talk, I will present DeepLensingFlow, a framework that uses advanced generative machine learning models, including normalizing flows and diffusion models,

to learn and reproduce the statistical distribution of weak-lensing convergence maps. I will demonstrate how these models can generate novel, physically consistent maps that recover key summary statistics, and discuss their applicability to upcoming surveys such as Rubin-LSST.

28

emi-Supervised Domain Adaptation for Sim-to-Obs Astrophysics: DESI→J-PAS

Author: Daniel López-Cano[1]

[1] *Instituto de Física da Universidade de São Paulo (IFUSP)*

Modern ML models trained on simulations often degrade on real data because of domain shift. I will present a semi-supervised domain adaptation (SSDA) pipeline that transfers a four-class pseudo-spectral classifier (high-z QSOs, low-z QSOs, galaxies, stars) from abundant DESI→J-PAS mocks (˜1.5M) to real J-PAS observations using only a small labeled J-PAS subset. The method pretrains on mocks, then freezes the classification head and adapts the encoder with balanced cross-entropy, using J-PAS labels to guide class-conditional alignment. On a held-out J-PAS test set, SSDA improves macro-F1 to 0.82 compared to 0.79 (target-only baseline with the same label budget) and 0.73 (zero-shot mocks). Gains concentrate in quasars, e.g., high-z QSO F1 rises to 0.66 (vs. 0.55/0.37), reflecting reduced confusion near z≈2.1 and better separation from compact galaxies. I will discuss why these gains occur, remaining degeneracies in narrow-band pseudo-spectra, and how modest target supervision enables reliable, label-efficient sim-to-obs transfer for target selection and AGN searches. Code and configs will be shared for straightforward reuse.

Lectures 1/3 / 29

Hadronization and Machine Learning

Author: Manuel Szewc[None]

Hadronization, the transition from unobservable partons to measurable hadrons, is a key component of how the Standard Model of particle physics explains current data. However, due to its intrinsically non-perturbative nature, it remains challenging to model from first principles. In particle physics, where simulators are needed to relate theory and collider experiments, Monte Carlo event generators have incorporated hadronization with great success via a series of sophisticated fine-tuned empirical models. Nonetheless, current and future collider experiments are pushing simulators to their limits. Motivated by these difficulties, in these lectures I'll present proposed alternatives where the empirical model is replaced by a surrogate data-trainable Machine Learning-based model. This model should be physics-based, fit available data and be made part of existing Monte Carlo simulators so as to be usable by the community.

30

Physically-oriented clustering of conformations in multi-molecule systems

Author: Rafael Ribeiro[1]

[1] *Universidade de São Paulo*

Unsupervised machine learning techniques are widely used to analyze the extensive data generated by molecular modeling. In particular, some tools have been developed to cluster configurations from classical simulations with a standard focus on individual units, ranging from small molecules to complex proteins. Since the standard approach computes the root-mean-square deviation (RMSD) of atomic positions, accounting for atom permutations is crucial to optimizing clustering of multi-molecule systems featuring identical molecules. To address this issue, we developed the clusttraj program, a solvent-informed clustering package that corrects inflated RMSD values by finding optimal pairings between configurations. The program combines reordering schemes with the Kabsch algorithm to minimize the RMSD of molecular configurations before running a hierarchical clustering protocol. By considering evaluation metrics, one can automatically determine the optimal threshold and compare available linkage schemes. The program's capabilities will be illustrated by considering various systems, ranging from pure water clusters to solvated proteins and oligomer chains in different solvents. Ultimately, we reduce the data complexity and obtain a subset of conformations whose dependence on cluster-related parameters and representativeness with respect to desired properties will be discussed. clusttraj is implemented as a Python library and can be used to cluster generic ensembles of molecular configurations beyond solute–solvent systems.