Challenges and Innovations in Learning from Heterogeneous Data in Fundamental Physics

C.Biscarat[1], S.Caillou[1], J.Stark[1]

[1]L2I Toulouse, CNRS/IN2P3, Université de Toulouse

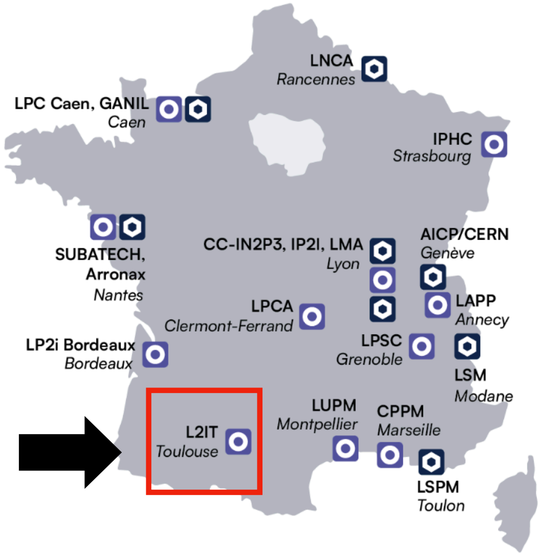
EuCAIFCon2025

AI for science, science for AI

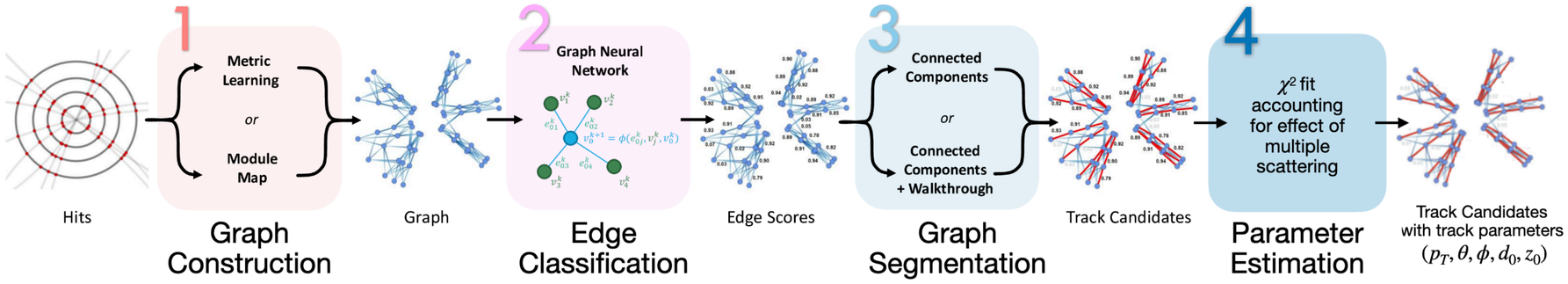
Université de Toulouse

A word on the speaker

Sylvain Caillou
Computer Science Staff Research Engineer
at Laboratoire des 2 Infinis (L2IT), Toulouse, FR
A new lab with a strong focus on AI for the 2 Infinite Physics



Work mainly on Track pattern recognition using Graph Neural Networks (GNNs) for ATLAS ITk



ATLAS EXPERIMENT

EuCAIFCon2025

Harnessing AI and ML Innovations for High-Luminosity LHC: Transitioning from R&D to Production

Jun 16, 2025, 3:55 PM
25m
T1a+T1b

Highlight talk | Inference & Uncertai... | Highlight talks

WorkShop on “Heterogeneous Data and Large Language Model”

~80 AI experts gathered over 4 days in Toulouse, 28 talks including 6 keynote adresses

WorkShop on “Heterogeneous Data and Large Language Model”

~80 AI experts gathered over 4 days in Toulouse, 28 talks including 6 keynote adresses

Bring people together: Participants from major international physics collaborations (CERN, LISA) and the Toulouse AI ecosystem (3IA ANITI, local industry, HPC stakeholders)

Deliberately interdisciplinary: fundamental physics, Earth observation, biology, fluid mechanics, weather forecasting, cognitive science, and more—encouraging cross-field idea sharing and collaboration

AI for science, science for AI

Heterogeneous Data and Large Representation Models in Science

30 September 2024 to 3 October 2024 Toulouse, France

Europe/Paris timezone

Enter your search term

Overview
- Keynote speakers
- Contribution List
- Timetable week view
- Timetable detailed view
- Instructions to presenters
- Participant List
- Committees & Support
- Workshop Venue
  - Getting to Toulouse
  - Accomodations
- Social events
- Workshop poster
- Acknowledgments
- Call for Abstracts
- Registration
- Privacy Information

Contact

aissai_heterogeneous-d...

The third workshop of the AISSAI semester on "Artificial Intelligence for the two infinites" will be held in Toulouse, France from September 30th to October 3rd, 2024. This workshop will be devoted to Heterogeneous Data and Large Representation Models in Science.

The workshop came to its end. Thanks for being with us and making this event lively, collaborative and friendly!

Thank you all for you participation!

More is to come: pictures of the event, recordings of the talks, and your input to the paper we intend to write. We will keep in touch with you!

Institut Supérieur de l'Aéronautique et de l'Espace

SUPAERO

CERFACS

CENTRE EUROPÉEN DE RECHERCHE ET DE FORMATION AVANCÉE EN CALCUL SCIENTIFIQUE

CerCo UMR5549

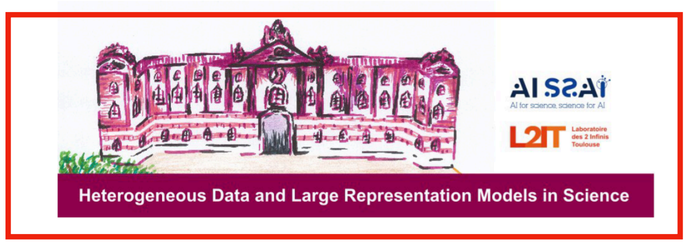
METEO FRANCE

À VOS CÔTÉS, DANS UN CLIMAT QUI CHANGE

Thank you to AISSAI for making this workshop possible!

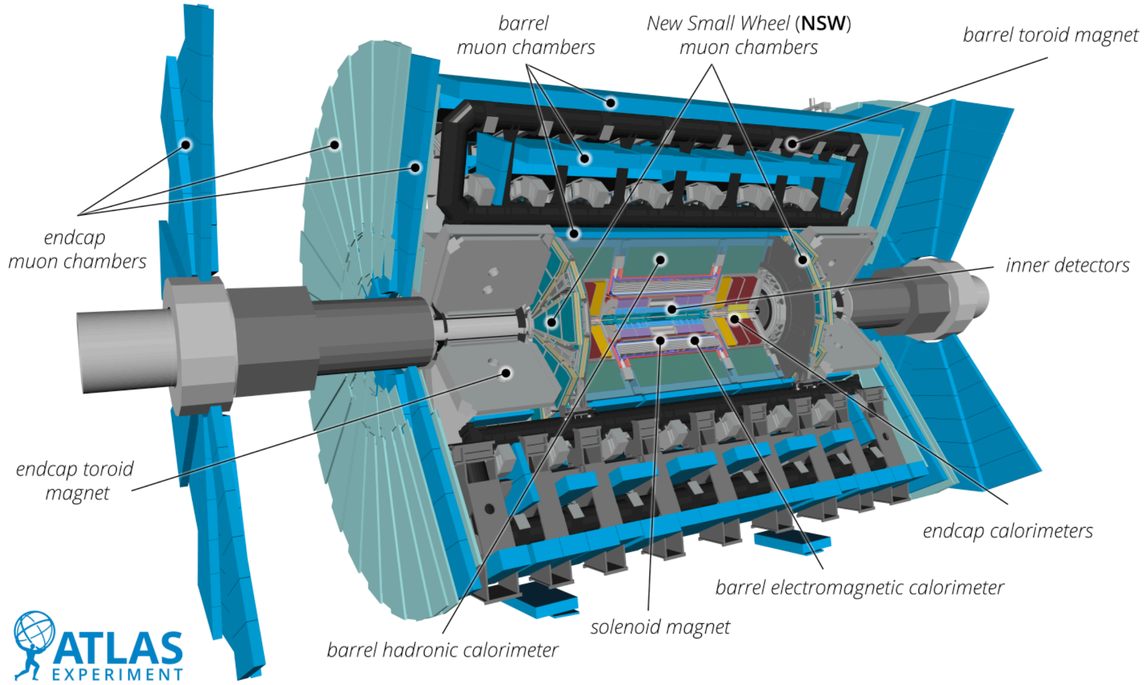
The **AISSAI center** was created in 2021 as a part of the CNRS National strategy in AI

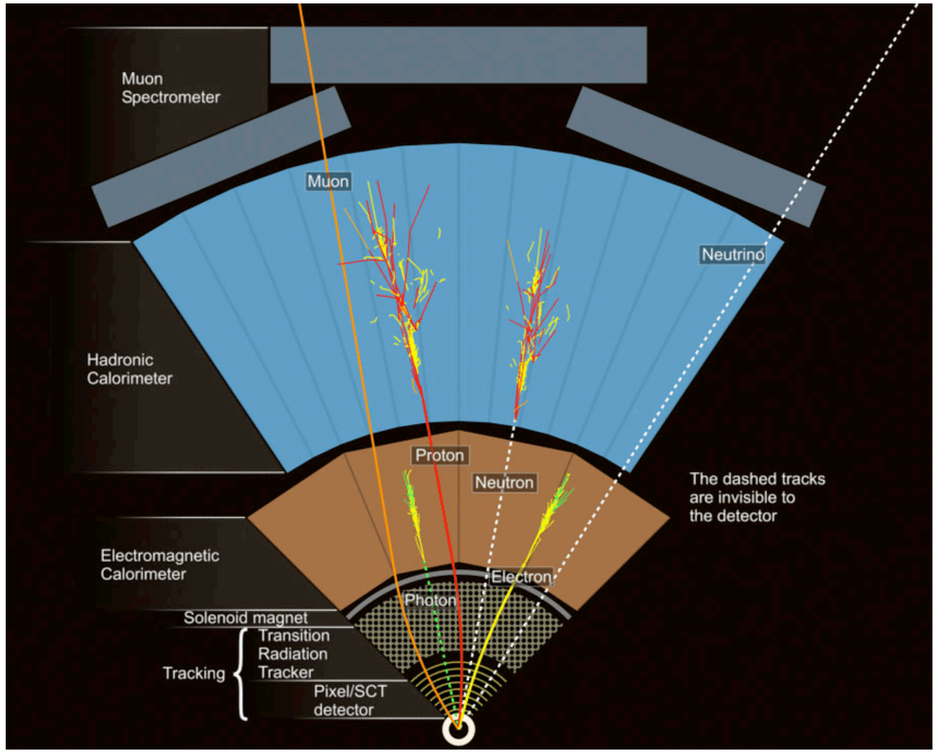
Main objective: structure and organize cross-functional actions involving all CNRS institutes at the interfaces with AI.

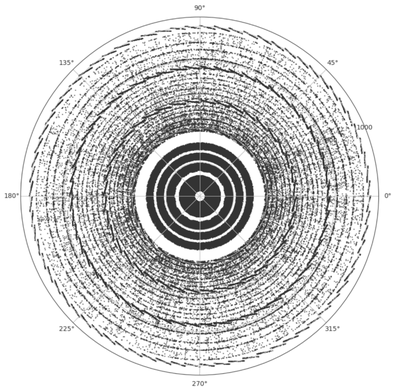


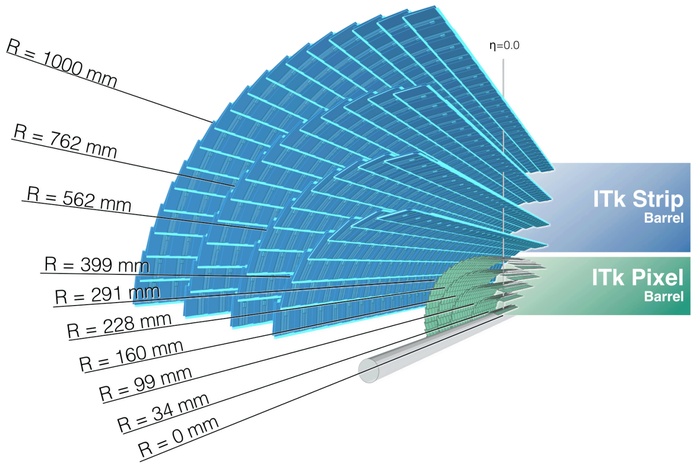
Our workshop was a part of the Artificial intelligence for the two infinites AISSAI thematic trimester

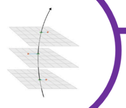
Heterogeneous data in particle physics detectors





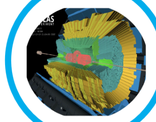






Tracks

- Left by charged particles going through inner detector
- Information on trajectories, hits in layers etc
- Neutral particles don't interact



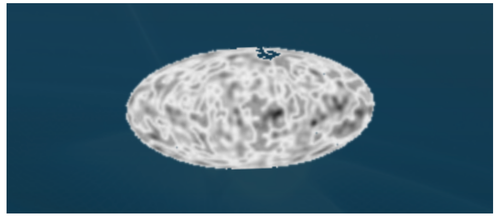
Calorimeter

- Measures energy of particles
- Complements tracker information for charged particles
- Neutral particles do interact

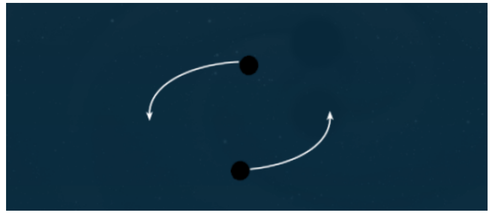
Leptons

- Electrons and muons are built from combining information from other parts of the detector
- Muons leave tracks in a dedicated muon spectrometer

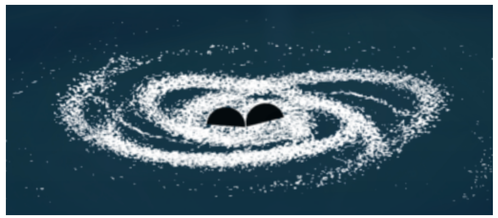
Gravitational Waves sources for LISA



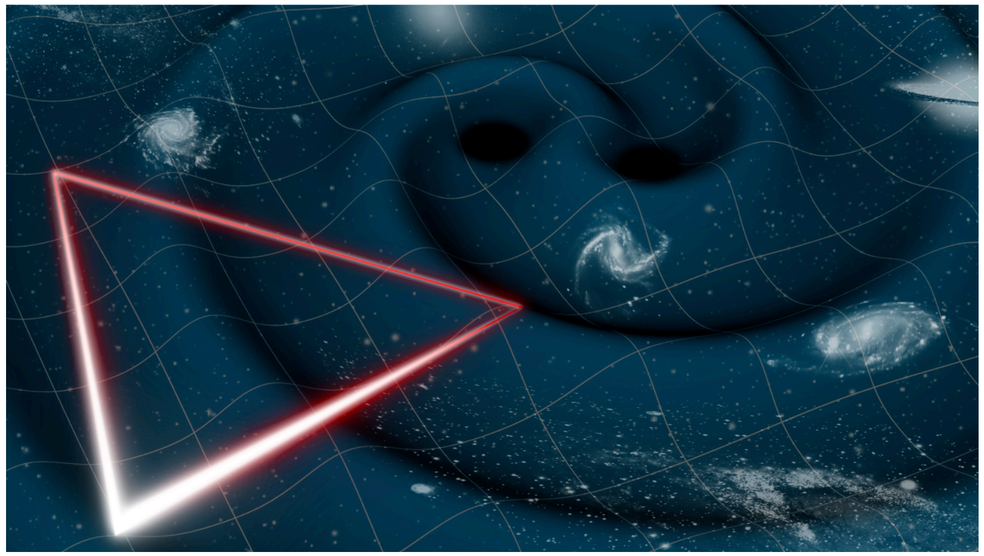
Cosmological sources stochastic backgrounds + bursts or a background generated by cosmic strings ? **Amplitude/rate very uncertain**.

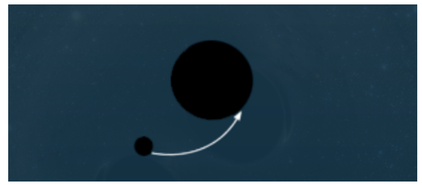


Stellar-origin-black-hole inspirals (SOBH)
O(few) yr^{-1}



Massive black hole binaries (MBHBs)
Rate uncertain, could be several tens per year





Extreme-mass-ratio inspirals (EMRIs)
Rate unknown, could be as low as ~1 yr^{-1} or as high as 1000 yr^{-1}

Ultra-compact binaries (UCBs)
~10^{7} systems
~10^{4} resolvable

Exotic sources

Signal will be thousand of Gravitational Waves coming from everywhere from hetereogenous sources (galactic binaries, mergers of massive black holes, …)

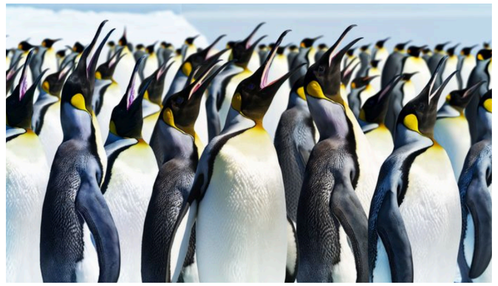
LISA Data Analysis challenge (The Global fit)

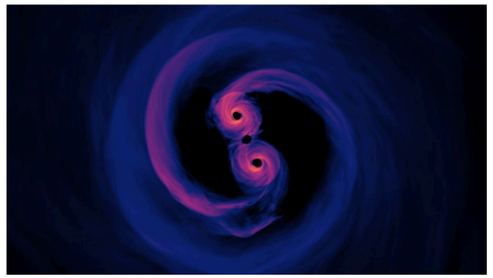
Separating overlapping Gravitational Waves signals is an extremely hard problem
(The Global fit challenge)

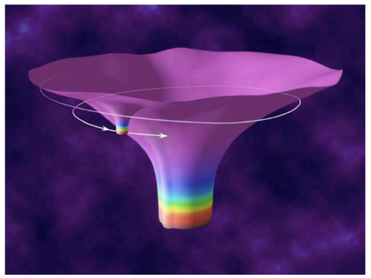
Hundreds

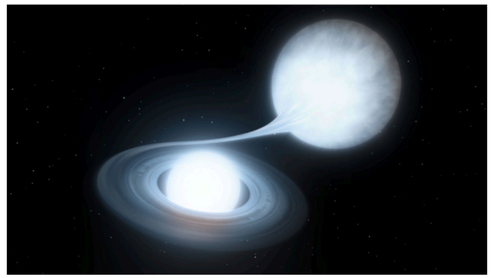
Tens to thousands

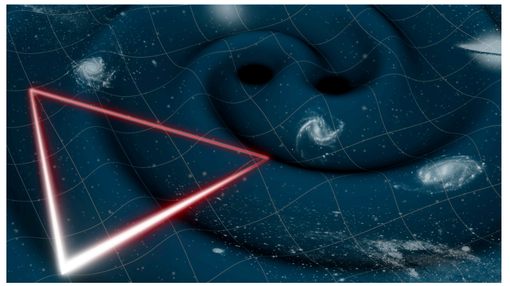
Millions







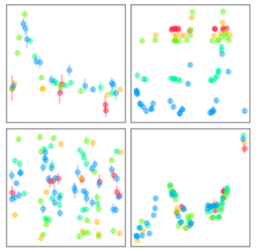
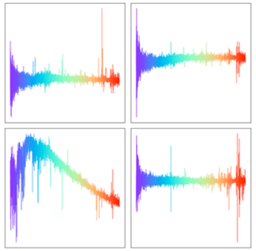


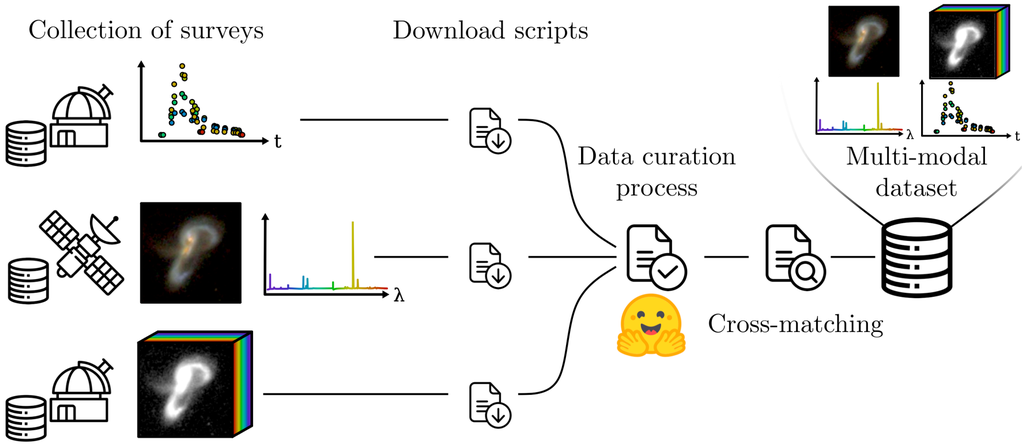


Can AI help with that ?

Heterogeneous Data in Astrophysics

Images, spectra, and time-series measurements of millions of astrophysical phenomena

	Images	Time-Series	Spectra
# examples	140M	4.5M	225M
Description	images in a variety of wavelength ranges, including optical and infrared	multivariate time-series of flux + uncertainty in different wavelength ranges	flux as a function of wavelength
Tasks	galaxy classification, physical property estimation	time-series classification, redshift estimation	physical property estimation
Examples			



The Multimodal Universe: 100TB of Astronomical Scientific Data

- **Goal**: Assemble the first large-scale multi-modal dataset for machine learning in astrophysics.
- Main pillars:
  - Engage with a broad community of AI+Astro experts.
  - Adopt **standardized conventions for storing and accessing data** and metadata through mainstream tools (e.g. Hugging Face Datasets).
  - Target large astronomical surveys, varied types of instruments, many **different astrophysics sub-fields**.

Polymathic

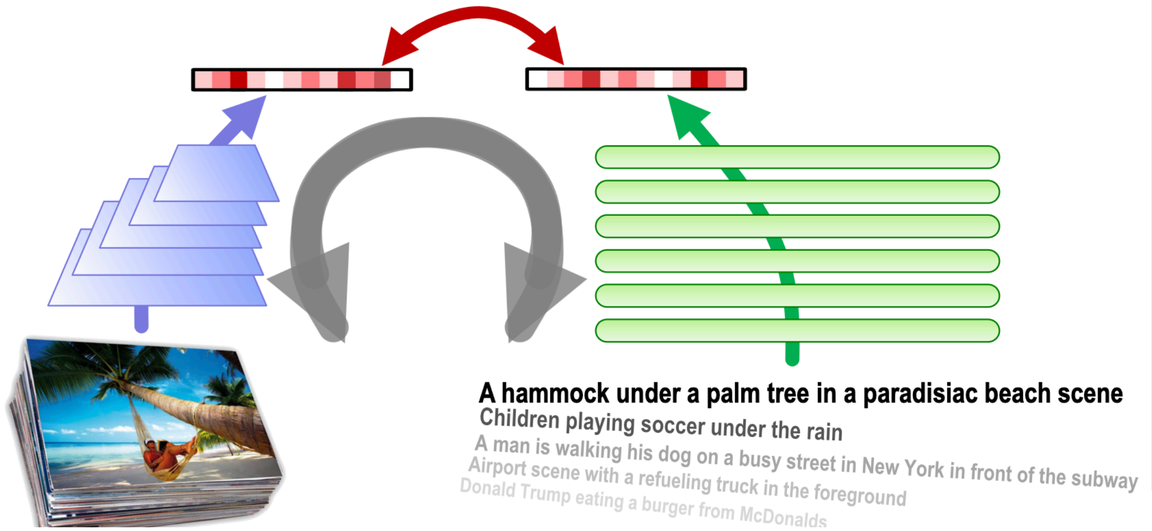
Can we build AI models which leverage information from heterogeneous datasets and learn better representations of the physic objects?

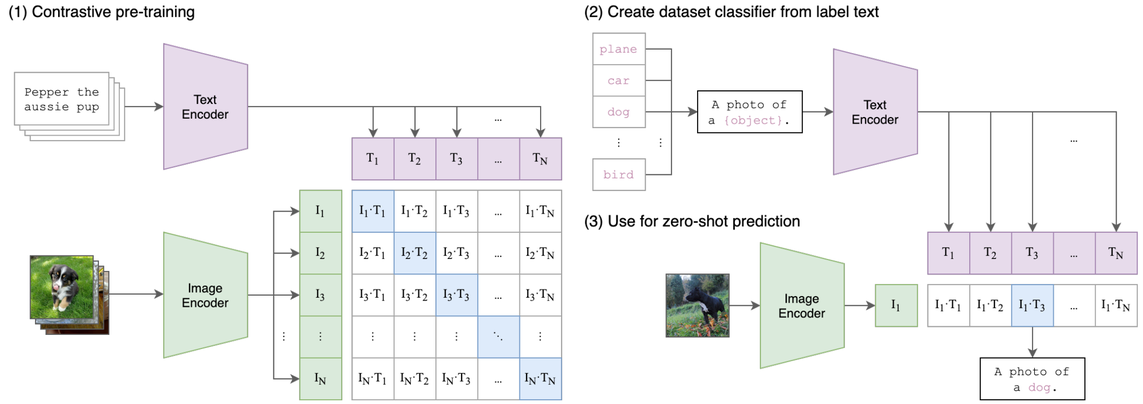
Multimodality Architectures: State of the Art

(Self)

Massively supervised learning (« brute-force »):

- **Representation alignment:** **CLIP (OpenAI)** **400M image-text pairs**
- **Image-to-Text (captioning):** **CoCa (Google)** **4B**
- **Text-to-Image:** **DALL-E3 (OpenAI)** **>5B**
  Imagen (Google)





CLIP: Contrastive Language-Image Pre-training

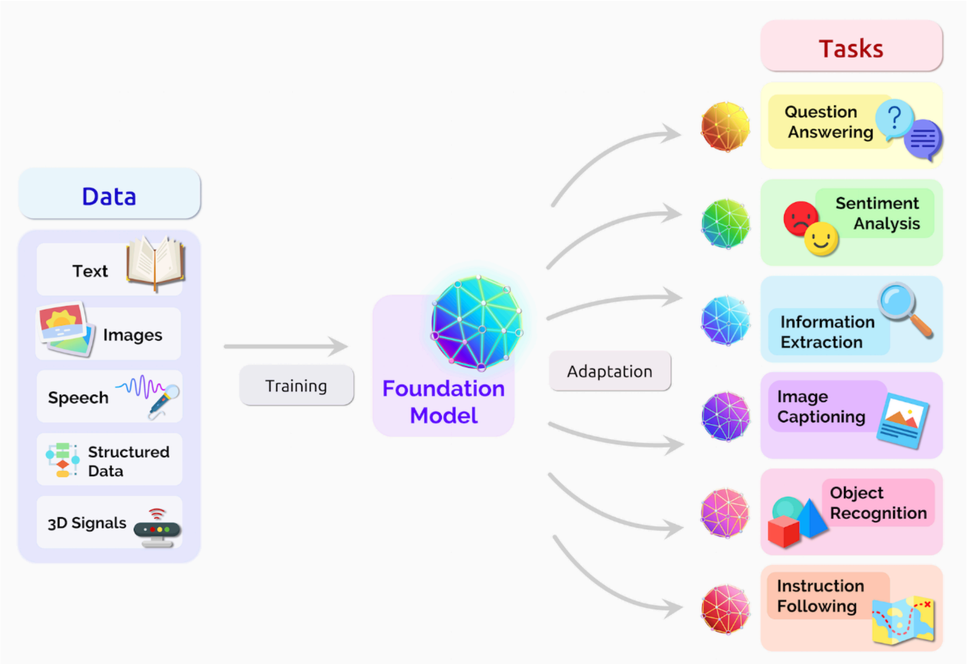
The rise of foundation models

Foundation model approach

- **Pretrain** models on pretext tasks, with self-supervision, on very large scale datasets.
- **Adapt** pretrained models to downstream tasks (Transfert Learning).
- **Combine** pretrained modules in more complex systems.

Foundation model take advantage of **multimodality**:

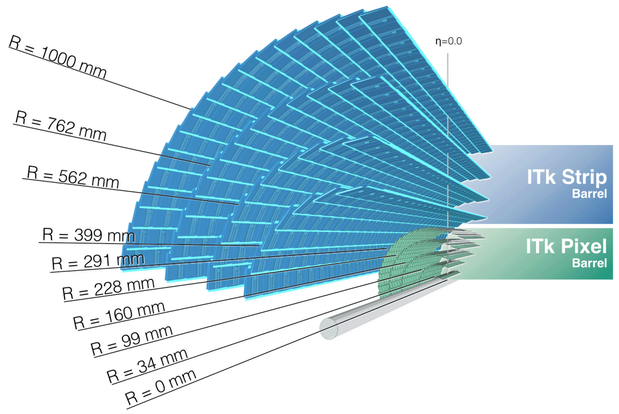
- Self-supervised training by contrastive cross modal tasks
- Enrich data representation in latent spaces by semantic grounding between modalities

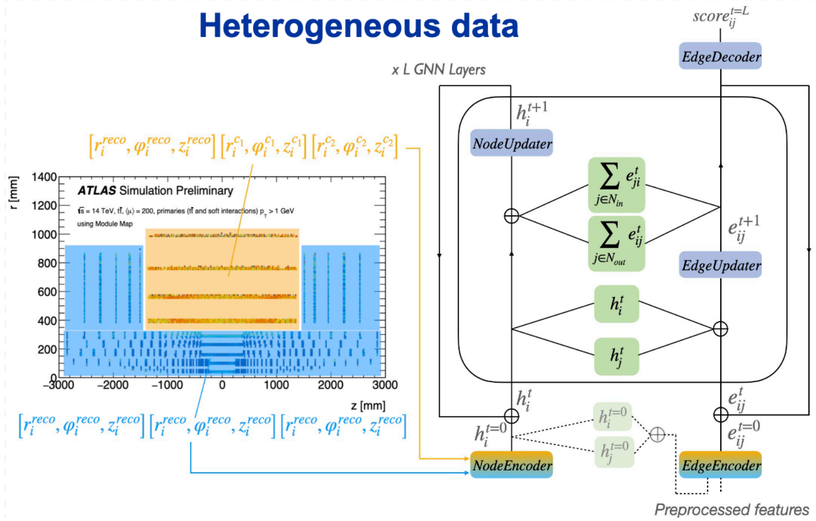


On the Opportunities and Risks of Foundation Models, Bommasani et al 2021

Learning from heterogeneous data in ATLAS

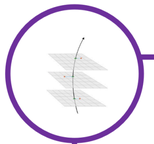
Graph Neural Networks for track reconstruction in the ATLAS ITk detector, Minh Tuan Pham o.b.o GNN4ITk team





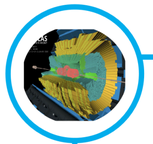
GNN deals with heterogeneous inputs from the Pixel and the Strip sub-systems in the ATLAS ITk

Heterogeneous, Multi-Task Models for Flavour Tagging in ATLAS Jackson Barr o.b.o the SALT team



Tracks

- Left by charged particles going through inner detector
- Information on trajectories, hits in layers etc
- Neutral particles don't interact

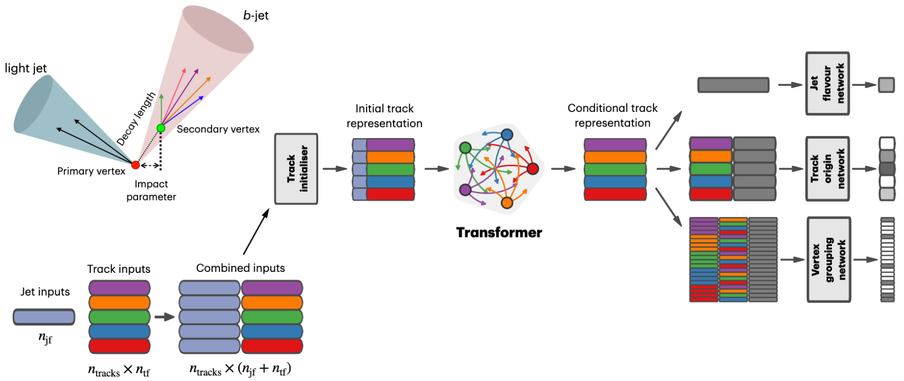


Calorimeter

- Measures energy of particles
- Complements tracker information for charged particles
- Neutral particles do interact

Leptons

- Electrons and muons are built from combining information from other parts of the detector
- Muons leave tracks in a dedicated muon spectrometer



Current generation of flavour tagging model in production in ATLAS, called GN2, is a transformer based model

Keynote Adress: Foundation models for high energy physics, Anna Hallin

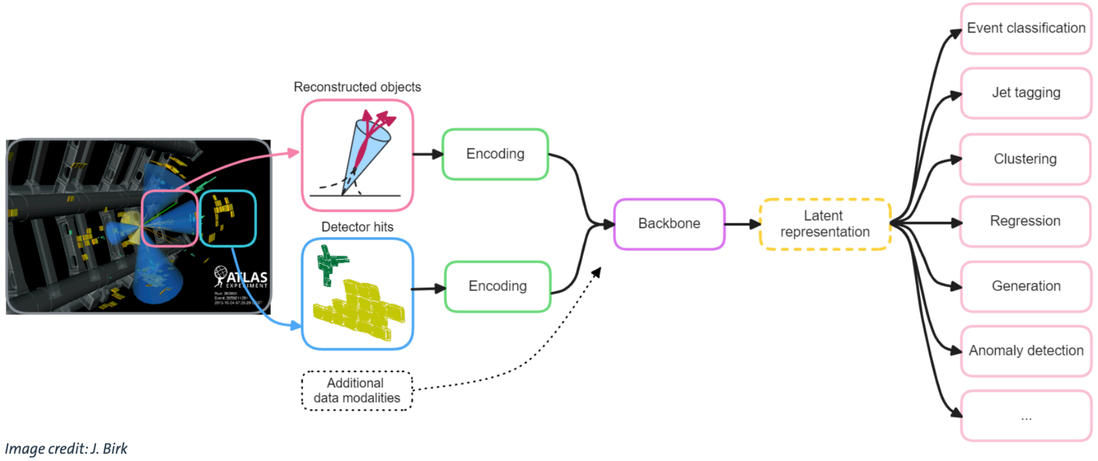


Image credit: J. Birk

Review of existing foundations models in HEP

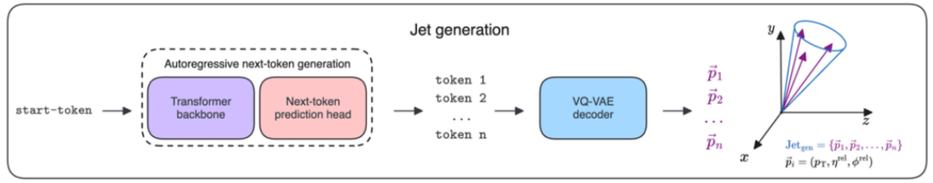
Name	Pre-training goal	Architecture	Loss	Downstream tasks
ParT	Classification	Transformer	Cross-entropy class labels	Classification on different dataset
MPM	Predict masked out tokens (surrogate task)	Transformer	Cross-entropy masked token prediction	Classification (tagging, anomaly detection)
OmniJet-α	Next token prediction (generation)	Transformer	Cross-entropy next token prediction	Classification (tagging), Generation (unconditional)
OmniLearn	Generation + classification	Transformer + diffusion	Cross-entropy class labels + diffusion velocity parameter	Classification (tagging: different dataset, different experiment, different collision type; anomaly detection), Generation (conditional), Reweighting and unfolding

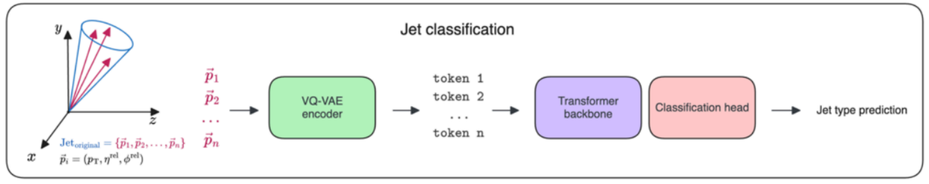
$\text{Jet} = \{p_1, p_2, \ldots, p_N\}$

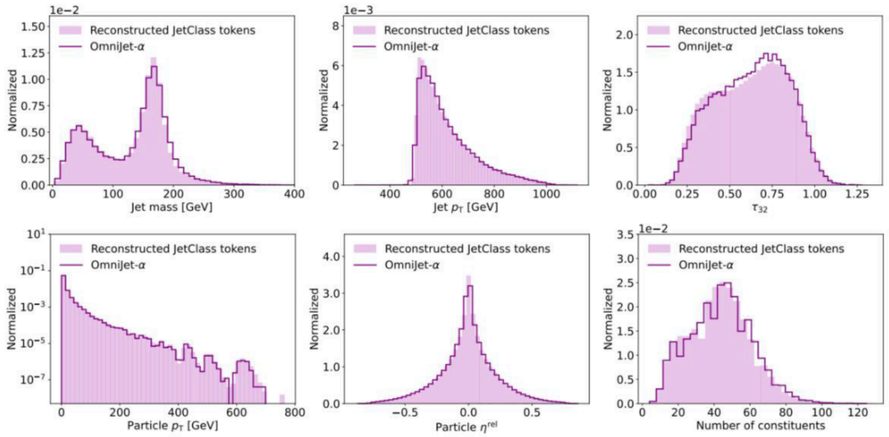
$p_i = \{p_T, \eta, \phi, \text{PID}, \text{charge}, \ldots\} \rightarrow \text{token}_i$

Jets as **sequences of integers**:

$\{< \text{start token} >, \text{token}_1, \text{token}_2, \ldots, \text{token}_N, < \text{stop token} >$







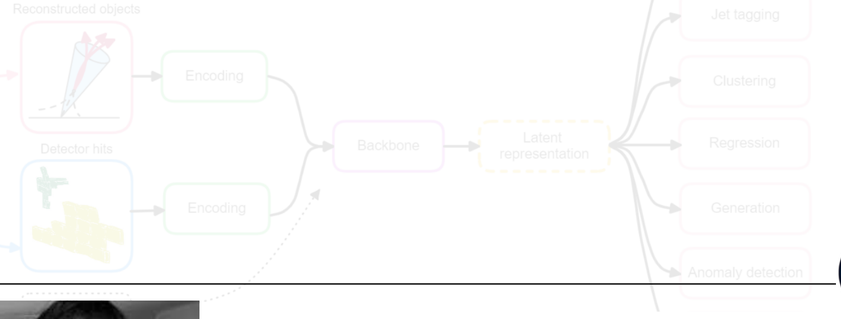
Focus on OmniJet-α

Uses a transformer for generative pretraining based on the GPT-1 architecture with next-token-prediction as training target. $p(xj|xj–1, \ldots, x1, < \text{start token} >)$

Generation : Generally good agreement to truth distribution

Able to task-switch: unsupervised full jet generation to supervised classification

Image credit: J. Birk



Name	Pre-training goal	Architecture	Loss	Downstream tasks
ParT	Classification	Transformer	Cross-entropy class labels	Classification on different dataset
MPM	Predict masked out tokens (surrogate task)	Transformer	Cross-entropy masked token prediction	Classification (tagging, anomaly detection)
OmniJet-α	Next token prediction	Transformer	Cross-entropy next token prediction	Classification (tagging), Generation (unconditional)
			Cross-entropy class labels + diffusion velocity parameter	Classification (tagging; different dataset, different experiment, different collision type; anomaly detection), Generation (conditional), Reweighting and unfolding

EuCAIFCon2025

Foundation models for HEP

Jun 19, 2025, 11:00 AM

25m

T1a+T1b

Highlight talk | Foundation Models | Highlight talks

$\text{Jet} = \{p_1, p_2, \ldots, p_N\}$

$p_i = \{p_T, \eta, \phi, \text{PID}, \text{charge}, \ldots\} \rightarrow \text{token}_i$

Jets as **sequences of integers**:

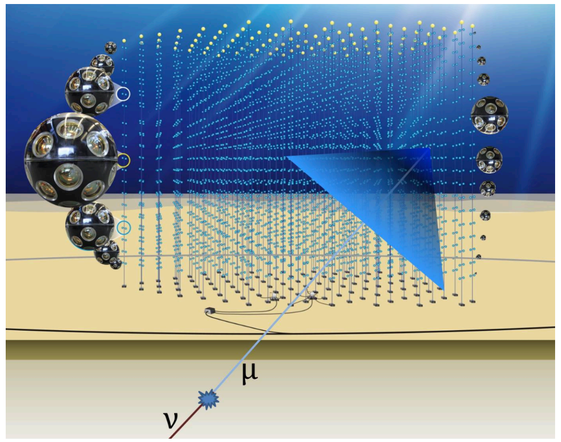
$\{< \text{start token} >, \text{token}_1, \text{token}_2, \ldots, \text{token}_N, < \text{stop token} >\}$

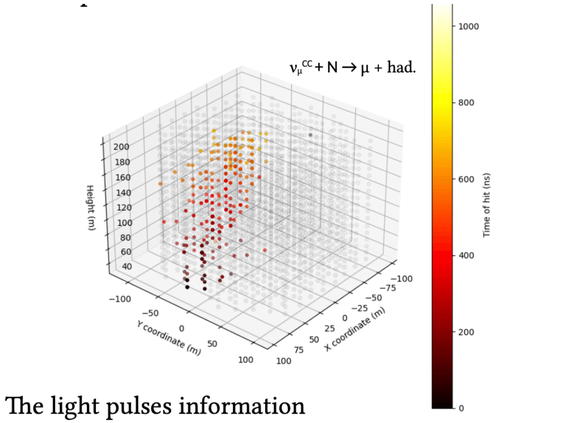
Foundation models in Neutrino Physics

DL classification
Neutrino vs Background

DL classification
Track-like vs Shower-like

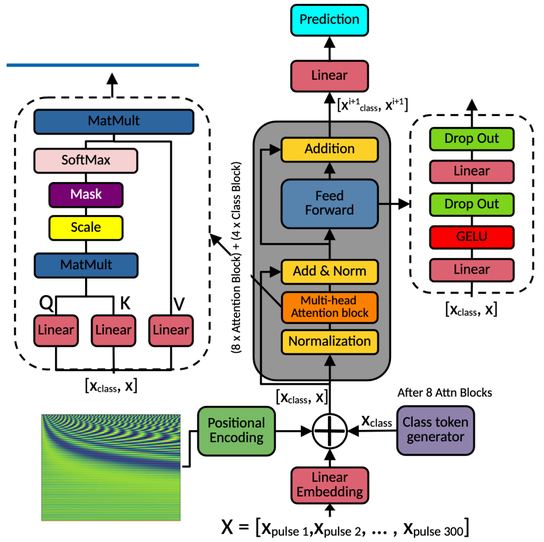
Track and shower reconstruction





The light pulses information

$$X_{pulse} = [pos_x, pos_y, pos_z, dir_x, dir_y, dir_z, t, ToT]$$

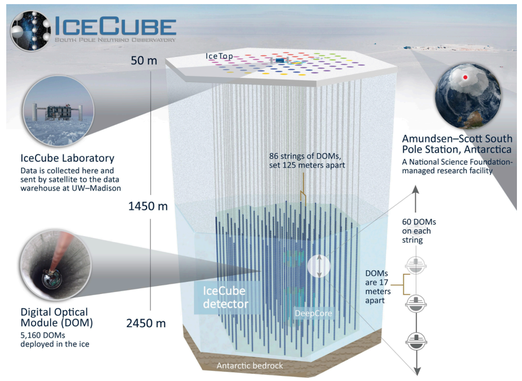


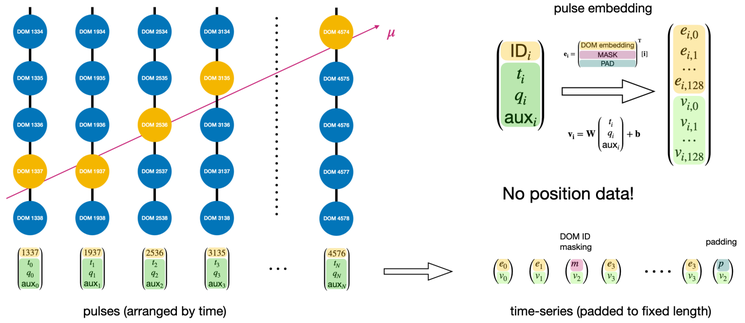
$X = [X_{pulse\ 1}, X_{pulse\ 2}, \dots, X_{pulse\ 300}]$

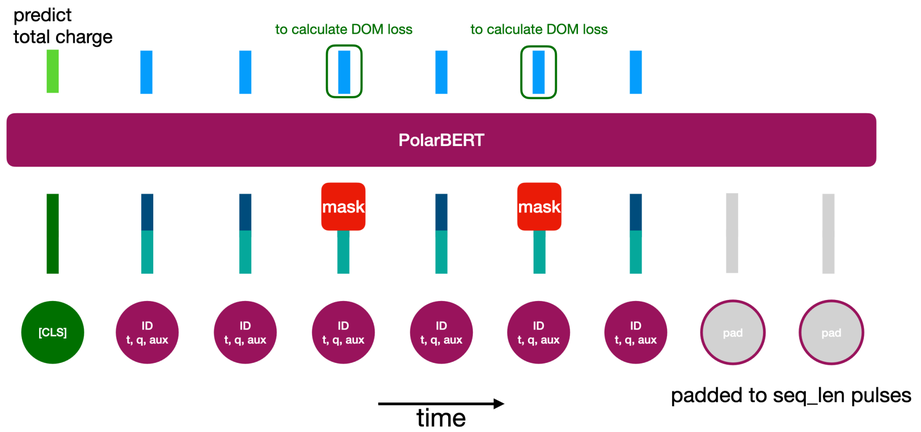
Foundation model in KM3NeT?

- learns as the detectors grow
- can handle multiple geometries
- can handle both KM3NeT/ORCA and KM3NeT/ARCA
- classification ↔ reconstruction

Small thinks big: transfer learning in KM3NeT/ORCA for neutrino event reconstruction, Iván Mozún Mateo, Antonin Vacheret o.b.o the KM3NeT collaboration







PolarBERT: A Foundation Model for IceCube, Inar Timiryasov, Jean-Loup Tastet, Oleg Ruchayskiy, Niels Bohr Institute and DIKU, University of Copenhagen

DL classification
Neutrino vs Background

DL classification
Track-like vs Shower-like

Track and shower reconstruction

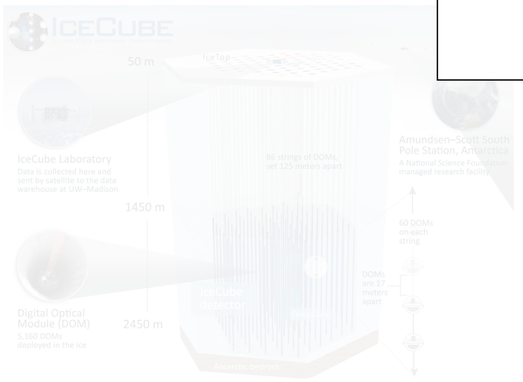
EuCAIFCon2025

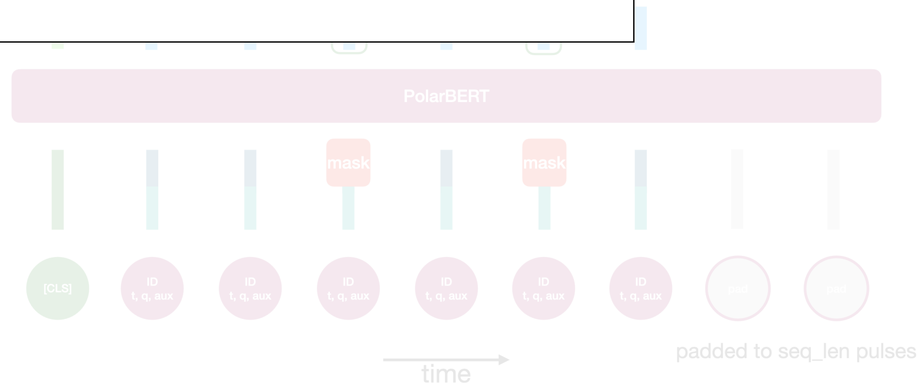
Iván Mozún Mateo,
Antonin Vacheret

⚡ Enhancing low energy reconstruction and classification in KM3NeT/ORCA with transformers

Jun 17, 2025, 2:21 PM
3m
T1c

Poster Session A | Patterns & Anomalies | Patterns & Anomalies



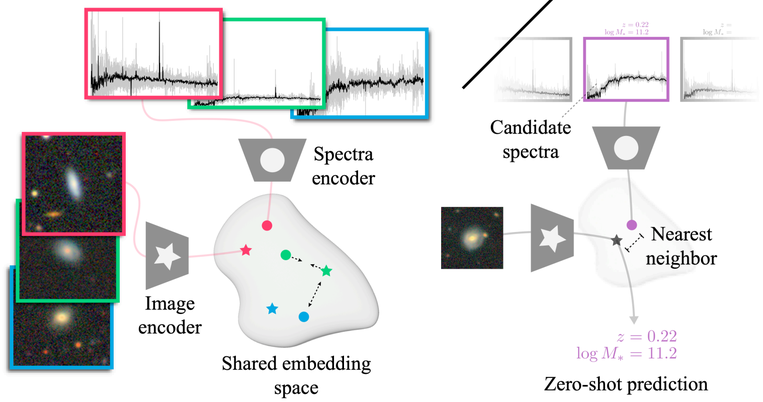


Keynote Address: Multimodal Pretraining for Astrophysical Foundation Models, Francois Lanusse

Most Specific — AstroCLIP — "Massively Multi-Modal Large Data Model for Astrophysics" — Most General

Independent models for every type of observation

Single model capable of processing all types of observations



AstroCLIP

Cross-Modal Pre-Training for Astronomical Foundation Models

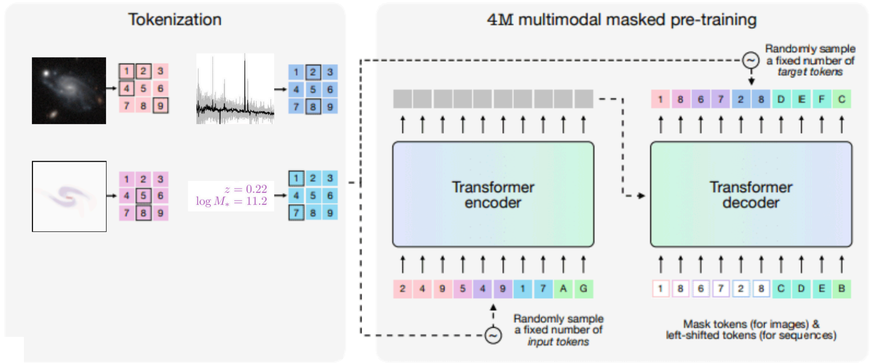
astro-ph.IM arXiv:2310.03024

Polymathic

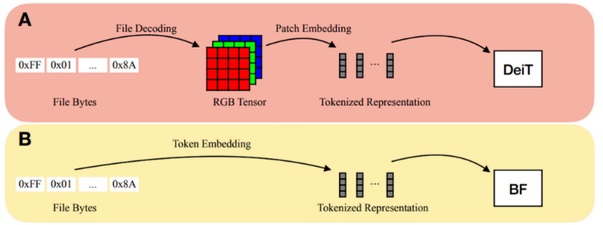
Project led by Francois Lanusse, Liam Parker, Leopoldo Sarra, Siavash Golkar, Miles Cranmer

Accepted contribution at the NeurIPS 2023 AI4Science Workshop

Published in the Monthly Notices of Royal Astronomical Society



Mizrahi et al. (2023)



Bytes Are All You Need (Horton et al. 2023)

Next Step: Any-to-Any Modeling on Scientific Data !

Most Specific — AstroCLIP

"Massively Multi-Modal Large Data Model for Astrophysics"

Most General

Independent models for every type of observation

Single model capable of processing all types of observations

Bytes Are All You Need (Horton et al. 2023)

EuCAIFCon2025

Summary talk: Cosmology

Jun 20, 2025, 11:15 AM

45m

T1a+T1b

Plenary talk

Summary talk

AstroCLIP

Cross-Modal Pre-Training for Astronomical Foundation Models

astro-ph.IM · arXiv:2310.03024

Polymathic

Project led by Francois Lanusse, Liam Parker, Leopoldo Sarra, Siavash Golkar, Miles Cranmer

Accepted contribution at the NeurIPS 2023 AI4Science Workshop

Published in the Monthly Notices of Royal Astronomical Society

Keynote Adress: Deep Learning & the Global Workspace Theory, Ruffin Van Rullen

Rufin Van Rullen
CNRS, The Brain and Cognition Research Center (CerCo), Artificial and Natural Intelligence Toulouse Institute (ANITI), Université de Toulouse, France.

Multimodal architectures: Limits

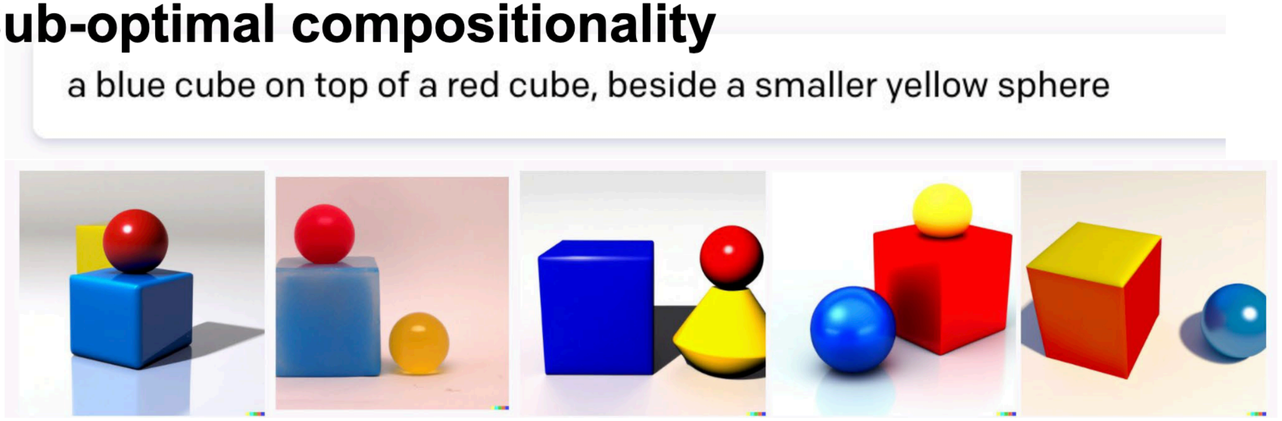
- **Hungry for data & computational resources**

  (a child probably gets <<1M of explicit supervision examples)

- **Sub-optimal multimodal grounding**

  Devillers et al (CoNLL 2021): *Does language help generalization in vision models?*

- **Sub-optimal compositionality**



➔ **Solution A: bigger models, trained with even more data?**

➔ **Solution B: change of paradigm?**

The Global Workspace Theory

Bernard J. Baars (1993)
former Senior Fellow in Theoretical Neurobiology at the Neurosciences Institute in San Diego, US.

Access through **your organization** Purchase PDF

ELSEVIER

Progress in Brain Research
Volume 150, 2005, Pages 45-53

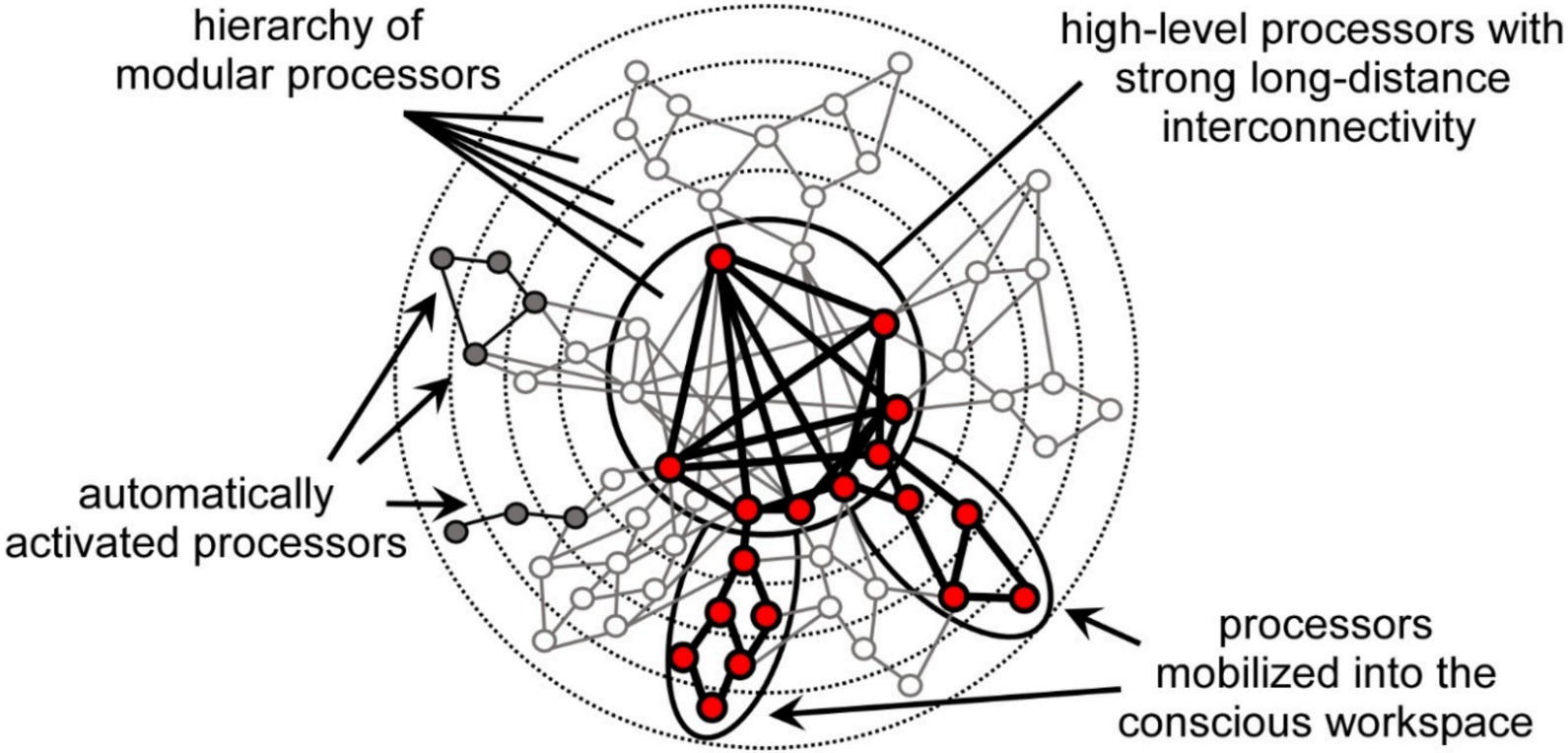
Global workspace theory of consciousness: toward a cognitive neuroscience of human experience

Bernard J. Baars

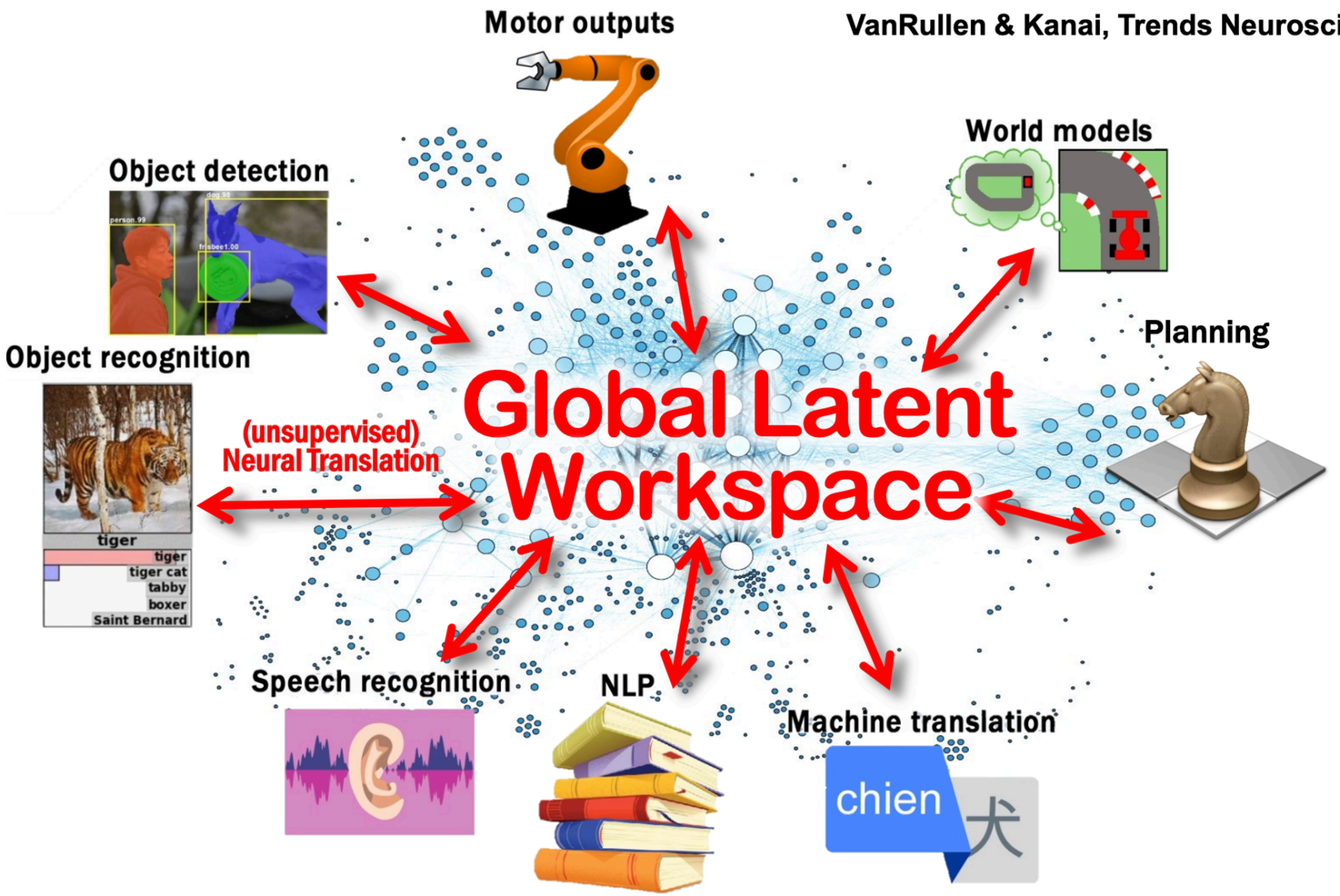
Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/S0079-6123(05)50004-9 Get rights and content



Deep learning and the Global Workspace Theory



Trends in Neurosciences

CellPress

Volume 44, Issue 9, September 2021, Pages 692-704

Opinion

Deep learning and the Global Workspace Theory

Rufin VanRullen [1,2], Ryota Kanai [3]

[1] The Brain and Cognition Research Center (CerCo), CNRS UMR5549, Toulouse, France
[2] Artificial and Natural Intelligence Toulouse Institute (ANITI), Université de Toulouse, France
[3] Araya Inc., Tokyo, Japan

Towards implementating GWT

Modules

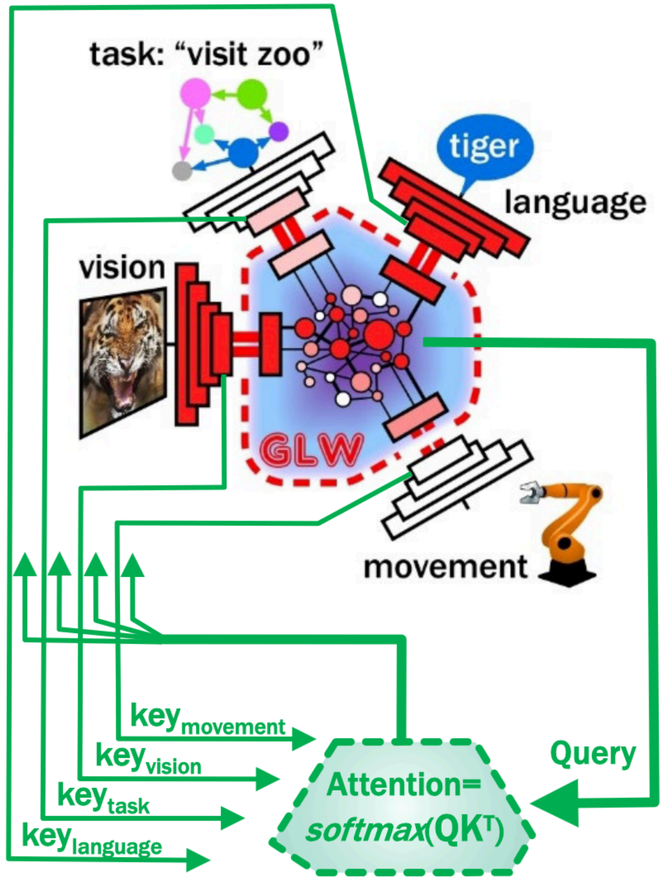
- (N≥2) Pretrained networks
  (hundreds of available choices!)
- Choice determines model functionality

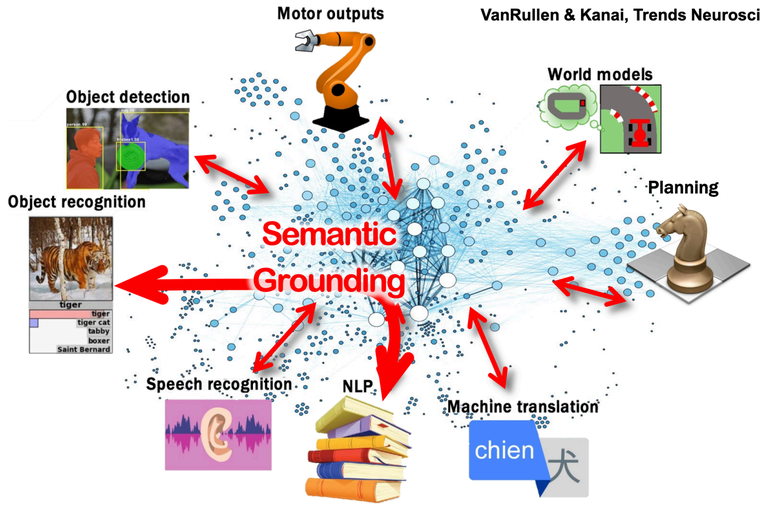
Broadcast

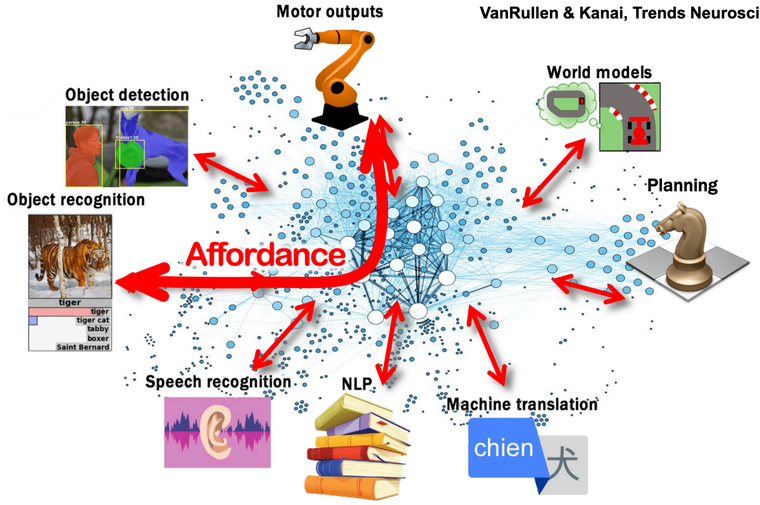
- Unsupervised neural translation
- Trained via cycle-consistency objective

Attention

- Transformer: key-query matching
- Top-down & bottom-up control







Léopold Maytié

Ph.D Student
CerCo
Université de Toulouse
Toulouse (France)

Multimodal systems with Global Workspace

Our ecosystem...

Funded by the European Union

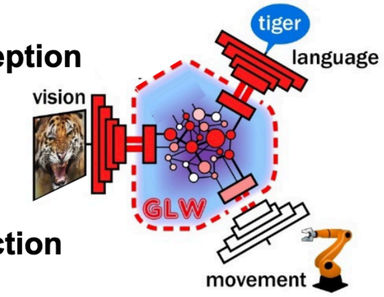
European Research Council
Established by the European Commission

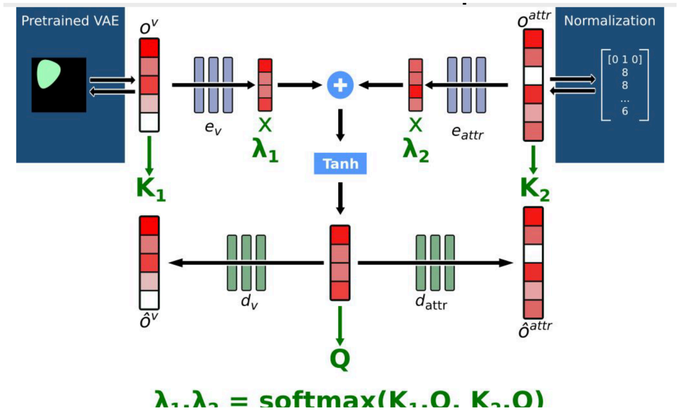
ERC Advanced Project GLOW (2023-2028)

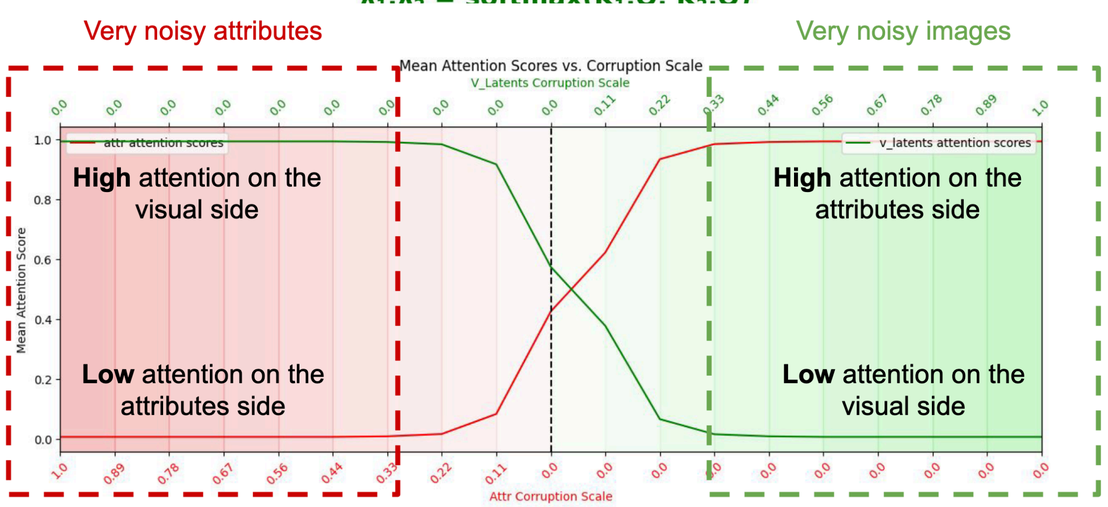
- **Develop brain-inspired multimodal deep learning systems**
- **Evaluate their use and relevance for machine learning**
- **Advance our knowledge of the brain**

ANITI Synergy Chair C3-PO (2024-2028)

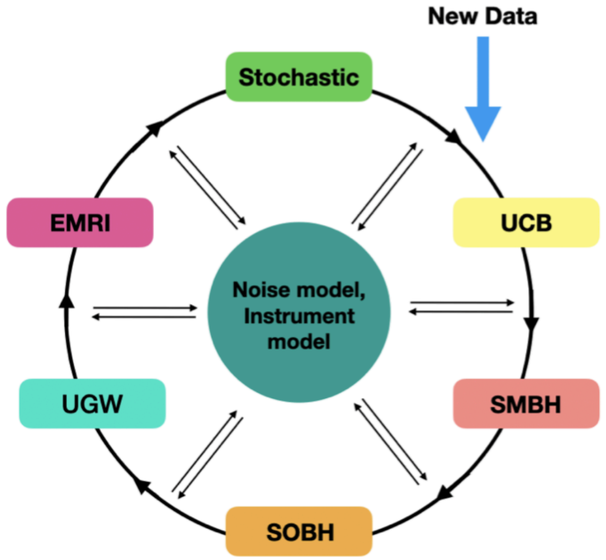
- **Cobots with Conversation, Cognition & Perception**
- **Chairs: R.VanRullen (CerCo), N. Asher (IRIT), T. Serre (Brown), O. Stasse (LAAS)**
- **Frugal multimodal robotic systems with grounded perception, language and action**







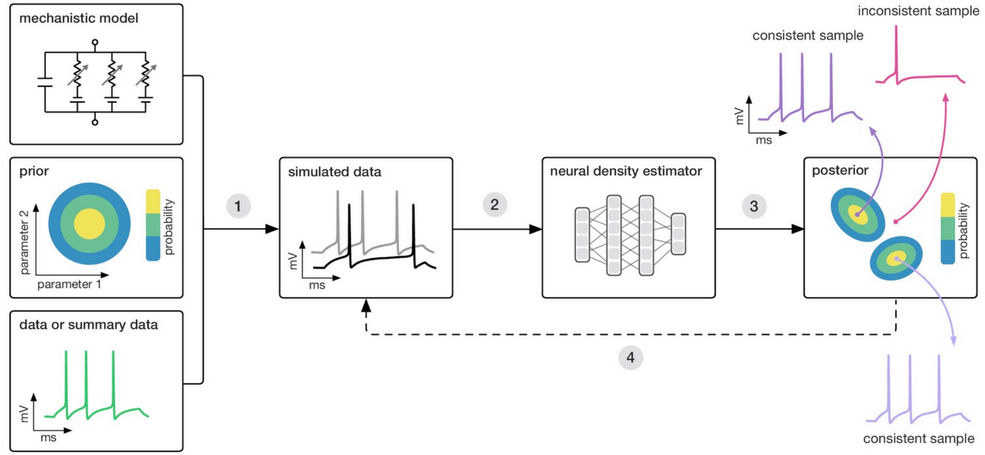
LISA Data Analysis challenge: Towards an AI-driven Global fit ?



Successful solutions to the global fit problem have used classic stochastic sampling techniques.

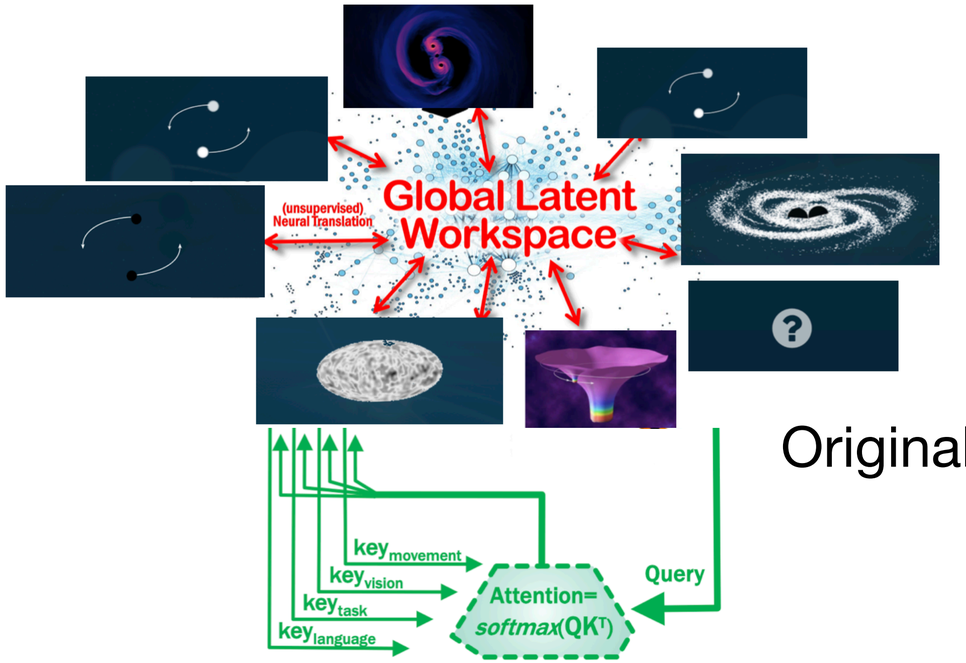
Typical strategy adopted is to iteratively update the solution for one source type and then move to the next.

Scaling and computing ressources stay a challenge



Simulation Based Inference (SBI) is explored to learn posterior for a single source of data

<u>Neural density estimation for Galactic Binaries in LISA data analysis</u>
Natalia Korsakova (2024)



Would it be possible to combine type of sources dedicated neural density estimators in a “Global Workspace” with Attention Mechanisms ??

Original idea from J. Stark

Why choose a hierarchical, modular architecture (à la Global Workspace) over a single giant, fully-tokenized Transformer?

Preserves symmetry & equivariance
Each modality keeps its own specialist encoder that respects the intrinsic geometric / group-theoretic structure of the data (e.g., CNNs for images, SE(3) GNNs for molecules), instead of flattening everything into one generic token space.

Built-in explainability
Clear boundaries between modules + a central workspace let you trace *which* specialist contributed *what* evidence and *when*, making reasoning chains inspectable rather than buried in millions of mixed-modality parameters.

Future-proof modularity
New capabilities—Retrieval-Augmented Generation, neuro-symbolic reasoning, causal-learning engines—can be plugged in as standalone processors that broadcast to / read from the workspace, without retraining the entire stack.

A structured “society of experts” scales knowledge, transparency, and extensibility better than an ever-larger monolith.

Thanks to all the keynotes !

Anna Hallin

Institute of Experimental Physics, Universität Hamburg (Germany). Particle physicist. Lead of the effort on Foundation Models in the research group for machine learning in particle physics at Hamburg. → biography

Jordi Inglada

French Space Agency (CNES, Toulouse), Center for Spatial Biosphere Studies (CESBIO, Toulouse). Main researcher in the "Artificial and Natural Intelligence Toulouse Institute" (ANITI) in the theme "Learning with little complex data, AI and phys models". → biography

François Lanusse

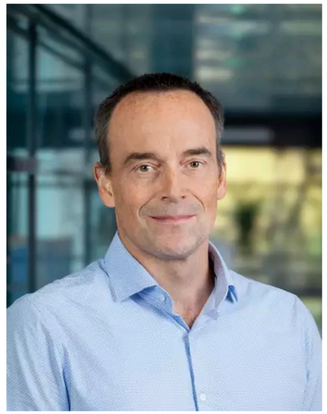
CNRS, UMR AIM / Flatiron Institut.

Michel Besserve

Empirical Inference Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany. Centers of interest: Causality and Machine Learning For Complex Systems. → biography

Andrew El-Kadi

Research Engineer at Google DeepMind working as part of the Weather effort. → biography

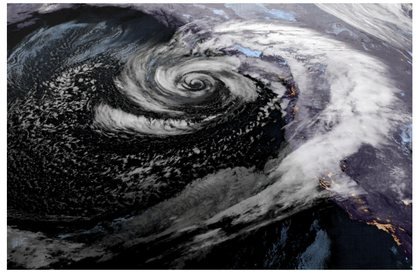


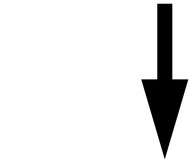
Jonathan Gair

Max Planck Institute for Gravitational Physics - Albert Einstein Institute - (AEI Potsdam, Germany). Group leader in Gravitational Waves data analysis (current LIGO and PTA detectors, future spatial mission LISA and ground-based Einstein Telescope apparatus).

Rufin VanRullen

CNRS, The Brain and Cognition Research Center (CerCo), Artificial and Natural Intelligence Toulouse Institute (ANITI), Université de Toulouse, France.





nature

Explore content · About the journal · Publish with us

nature › articles › article

Article | Open access | Published: 04 December 2024

Probabilistic weather forecasting with machine learning

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam & Matthew Willson

Nature **637**, 84–90 (2025) | Cite this article

Thanks to all the participants and the committees!

Scientific advisory committee
- Sylvain Caillou (L2IT, IN2P3, CNRS/UT3) - chair
- Alexandre Boucaud (APC, IN2P3, CNRS)
- Tobias Golling (Université de Genève)
- François Lanusse (Polymathic AI)
- Daniel Murnane (Copenhagen University)
- Thomas Oberlin (ISAE-SUPAERO, ANITI, Université de Toulouse)
- Jan Stark (L2IT, IN2P3, CNRS/UT3)
- Gordon Watts (Washington University)

Local organisation committee
- Catherine Biscarat (L2IT, IN2P3, CNRS/UT3) - chair
- Sylvain Caillou (L2IT, IN2P3, CNRS/UT3)
- Jocelyne Gauthier (L2IT, IN2P3, CNRS/UT3)
- Jan Stark (L2IT, IN2P3, CNRS/UT3)
- Jeanette Thibaut (L2IT, IN2P3, CNRS/UT3)

Scientific secretaries
- Vasco Gennari, PhD student (L2IT, CNES, CNRS/IN2P3, UT3)
- Alexandro Martone, PhD student (L2IT, IN2P3, CNRS/UT3)
- Minh-Tuan Pham, PhD student (University of Wisconsin-Madison)

jan.stark@l2it.in2p3.fr

catherine.biscarat@l2it.in2p3.fr

sylvain.caillou@l2it.in2p3.fr

EuCAIFCon2025

Do not hesitate to get in touch with us !