

ARTIFICIAL INTELLIGENCE FOR SCIENCE

Sašo Džeroski

Head of Department of knowledge technologies
Jožef Stefan Institute, Ljubljana, Slovenia

ARTIFICIAL INTELLIGENCE FOR SCIENCE: The project

*The project will focus on the **development of AI approaches** along four major directions, following requirements that AI methods have to meet to be used in science.*

*It will also address **a variety of applications of AI in science** in physical sciences and engineering (PE) and life sciences (LS), demonstrating the utility of the developed methods.*

In PE area, we will focus on

- *AI in materials science/ engineering (PE11) and*
- *Mathematics (PE1)*

In LS, we will use AI in

- *Immunology (LS6), and broader medicine (LS7), but also in*
- *Plant biology, environmental biology and ecology (LS8).*



ARTIFICIAL INTELLIGENCE FOR SCIENCE: The people

The consortium consists of five partners (3 research institutes and 3 universities), the best research organizations in Slovenia

- *Jožef Stefan Institute (IJS)*
- *University of Ljubljana (UL)*
- *National Institute of Chemistry (KI)*
- *University of Maribor (UM)*
- *National Institute of Biology (NIB)*

The first two carry out development of AI methods and the last three focus on applications of AI in different sciences

Many different groups from these institutions are involved, listed below, together with the group leaders



ARTIFICIAL INTELLIGENCE FOR SCIENCE: The people

Jožef Stefan Institute (IJS)

- *IJS-E8, Dept. of Knowledge Technologies: **Saso Dzeroski, Dragi Kocev***
- *IJS-E3, Dept. for Artificial Intelligence: **Dunja Mladenic***
- *IJS-E7, Computer Systems Dept.: **Tome Eftimov***
- *IJS-E9, Dept. of Intelligent Systems: **Tea Tusar, Bogdan Filipic***
- *+ A number of other departments at no cost to the project (E2, **P. Boshkoski**; K7: **S. Sturm**; F9: **B. Kersevan**; CMI: **J. Javorsek**)*

National Institute of Chemistry (KI)

- *KI-D12, Synthetic Biology and Immunology: **Roman Jerala***
- *KI-D10, Materials Chemistry: **Nejc Hodnik, Dusan Strmcnik***

National Institute of Biology (NIB)

- *Dept. of Biotech. & Systems Biology: **Kristina Gruden***



ARTIFICIAL INTELLIGENCE FOR SCIENCE: The people

University of Ljubljana (UL)

- *UL-FRI, Faculty of Computer and Information Science*
 - *FRI-LUI, Lab. of Artificial Intelligence: **Aleksander Sadikov, Vida Groznik***
 - *FRI-LBI, Bioinformatics: **Tomaz Curk***
 - *FRI-LUVSS, Visual Cognitive Systems: **Matej Kristan, Daniel Skocaj***
- *UL-FMF, Mathematics and Physics: **Ljupco Todorovski***
- *UL-FGG, Civil and Geodetic Engineering: **Natasa Atanasova***

University of Maribor (UM)

- *UM-FZV, Faculty of Health Sciences: **Gregor Stiglic***
- *UM-FKKT, Chemistry and Chemical Technology: **Urban Bren***



ARTIFICIAL INTELLIGENCE FOR SCIENCE: Content

The project has four work packages (WPs) centered on the development of different AI methods & the use of these classes of methods to problems from different areas of science.

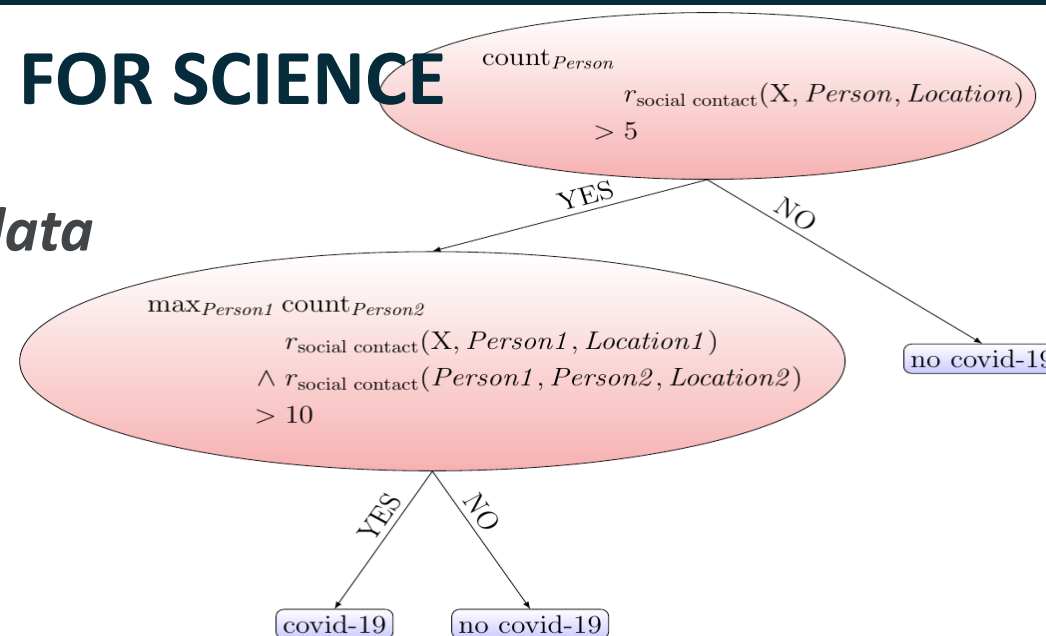
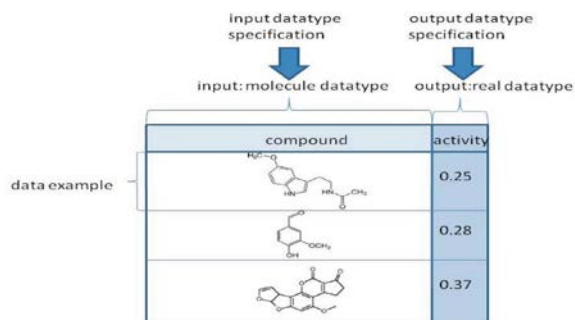
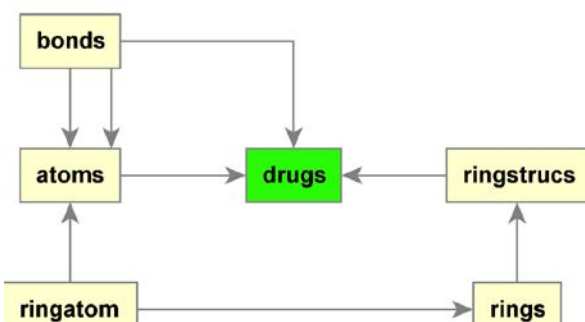
- *A) Explainable ML for science,*
- *B) Foundation models for science,*
- *C) Automated scientific modelling, and*
- *D) Semantic technologies for open science.*

Each WP has objectives and tasks regarding AI method development. Each WP also has objectives and tasks regarding the applications of the AI methods in that class to different problems in science.



A. EXPLAINABLE MACHINE LEARNING FOR SCIENCE

Learning interpretable models from complex data

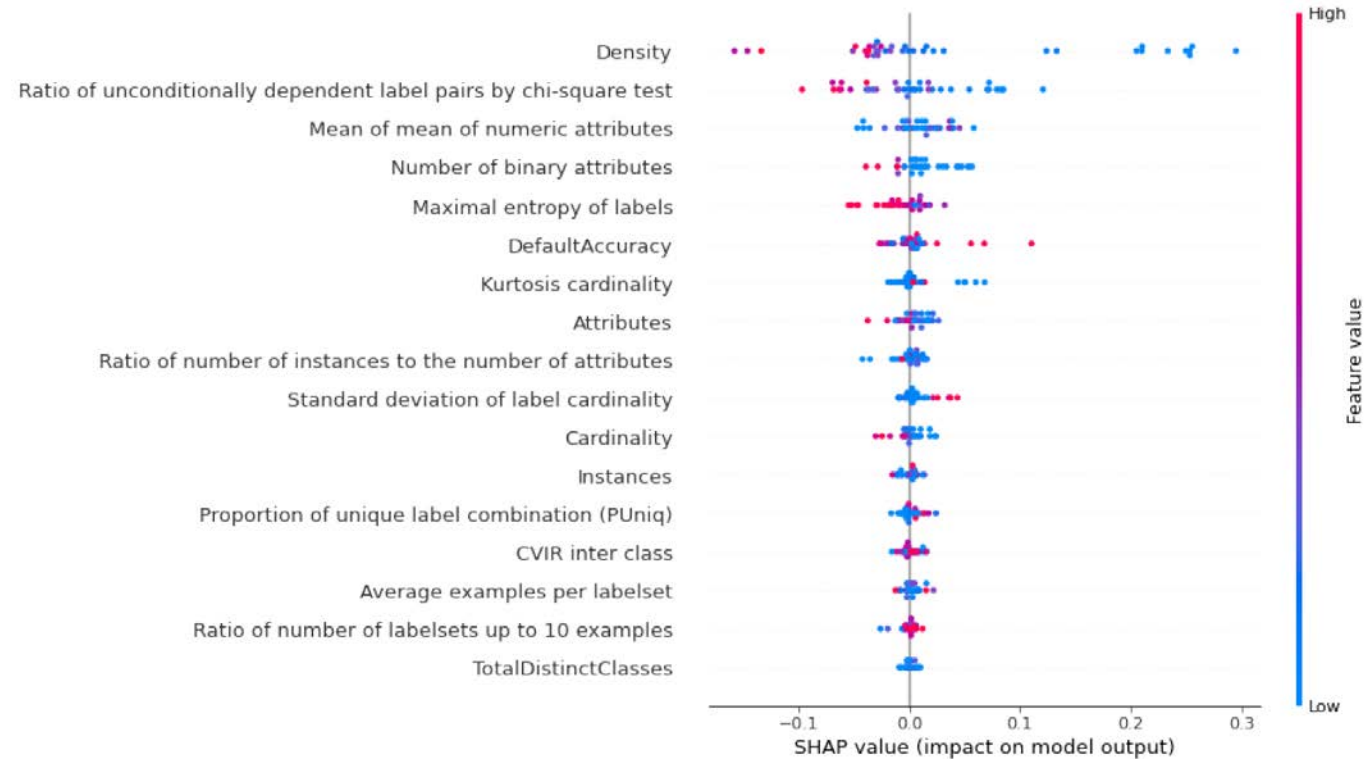
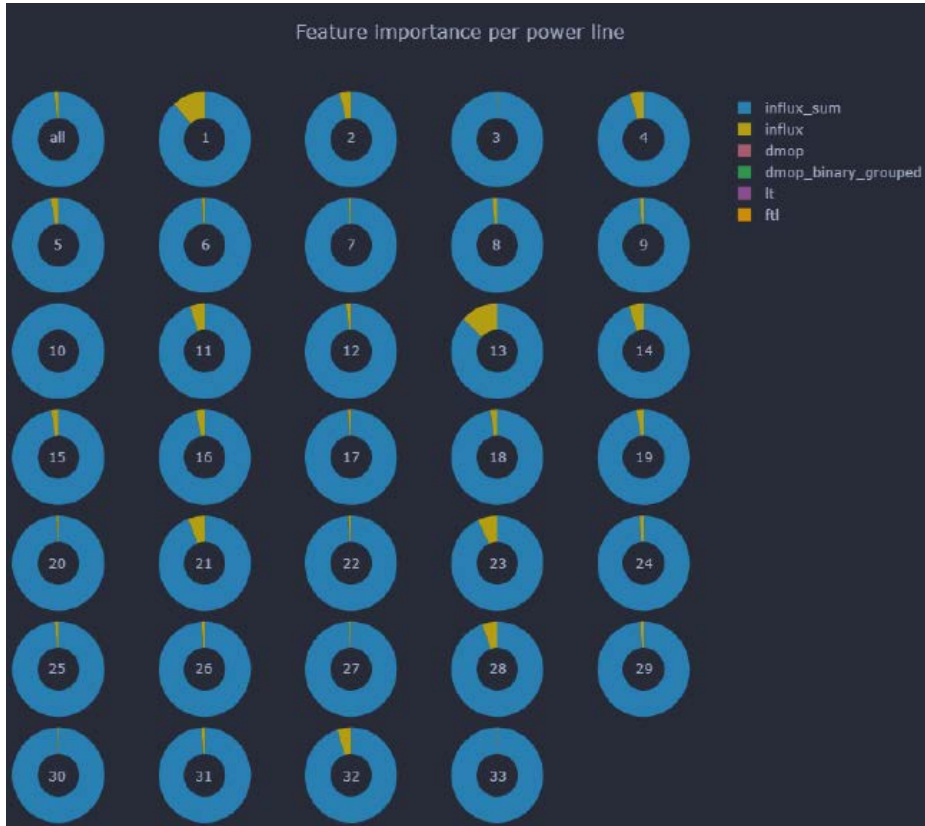


	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	?
Example 3	1	FALSE	0.08	0.07	?
Example 4	2	TRUE	0.49	0.69	Yes
Example 5	3	TRUE	0.49	0.69	No
Example 6	4	FALSE	0.08	0.07	?
...

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	0.68	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	0.10	1.69	7.57
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	3.51	2.50
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
...

A. EXPLAINABLE MACHINE LEARNING FOR SCIENCE

Explaining (uninterpretable models) and their predictions



A) EXPLAINABLE ML FOR SCIENCE

Methods. *Develop explainable ML methods for learning interpretable models from complex data, incl. methods that integrate neural and symbolic approaches, methods for explaining model predictions, and methods for monitoring the development of trends in bibliographic databases.*

Task A1: *Methods for learning interpretable models from complex data*

Task A2: *Methods for explaining models and*

Task A3: *Analyzing text and graph data to monitor the development of scientific fields*

A) EXPLAINABLE ML FOR SCIENCE




Applications. Apply explainable ML methods for learning in complex settings to different problems in the sciences, such as monitoring the development of scientific fields, design of gene therapy, design of novel drugs and formalization of mathematics/ discovery of new mathematical knowledge.

Task A4: Explainable ML for the design of therapeutics in gene therapy

Task A5: Explainable ML for drug design

Task A6: Explainable ML for mathematics

Discovery of Exact Equations for Integer Sequences

Boštjan Gec^{1,2,*} , Sašo Džeroski¹  and Ljupčo Todorovski^{1,3} 

¹ Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia; saso.dzeroski@ijs.si (S.D.); ljupco.todorovski@fmf.uni-lj.si (L.T.)

² Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia

³ Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia

* Correspondence: bostjan.gec@ijs.si

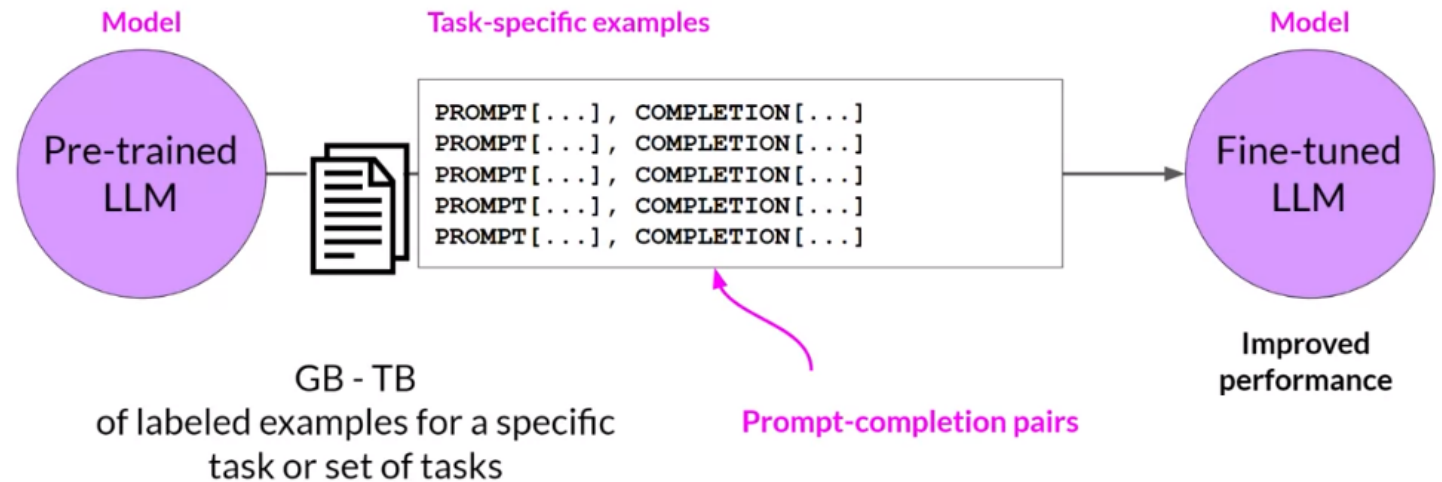
Abstract: Equation discovery, also known as symbolic regression, is the field of machine learning that studies algorithms for discovering quantitative laws, expressed as closed-form equations or formulas, in collections of observed data. The latter is expected to come from measurements of physical systems and, therefore, noisy, moving the focus of equation discovery algorithms towards discovering approximate equations. These loosely match the noisy observed data, rendering them inappropriate for applications in mathematics. In this article, we introduce *Diofantos*, an algorithm for discovering equations in the ring of integers that exactly match the training data. *Diofantos* is based on a reformulation of the equation discovery task into the task of solving linear Diophantine equations. We empirically evaluate the performance of *Diofantos* on reconstructing known equations for more than 27,000 sequences from the online encyclopedia of integer sequences, OEIS. *Diofantos* successfully reconstructs more than 90% of these equations and clearly outperforms SINDy, a state-of-the-art method for discovering approximate equations, that achieves a reconstruction rate of less than 70%.

B. FOUNDATION MODELS FOR SCIENCE

Foundation models (FMs) are large models that are generated by applying ML (typically deep learning) to a broad collection of data at scale and can be adapted for use in a wide range of downstream tasks; LLMs are a prime example

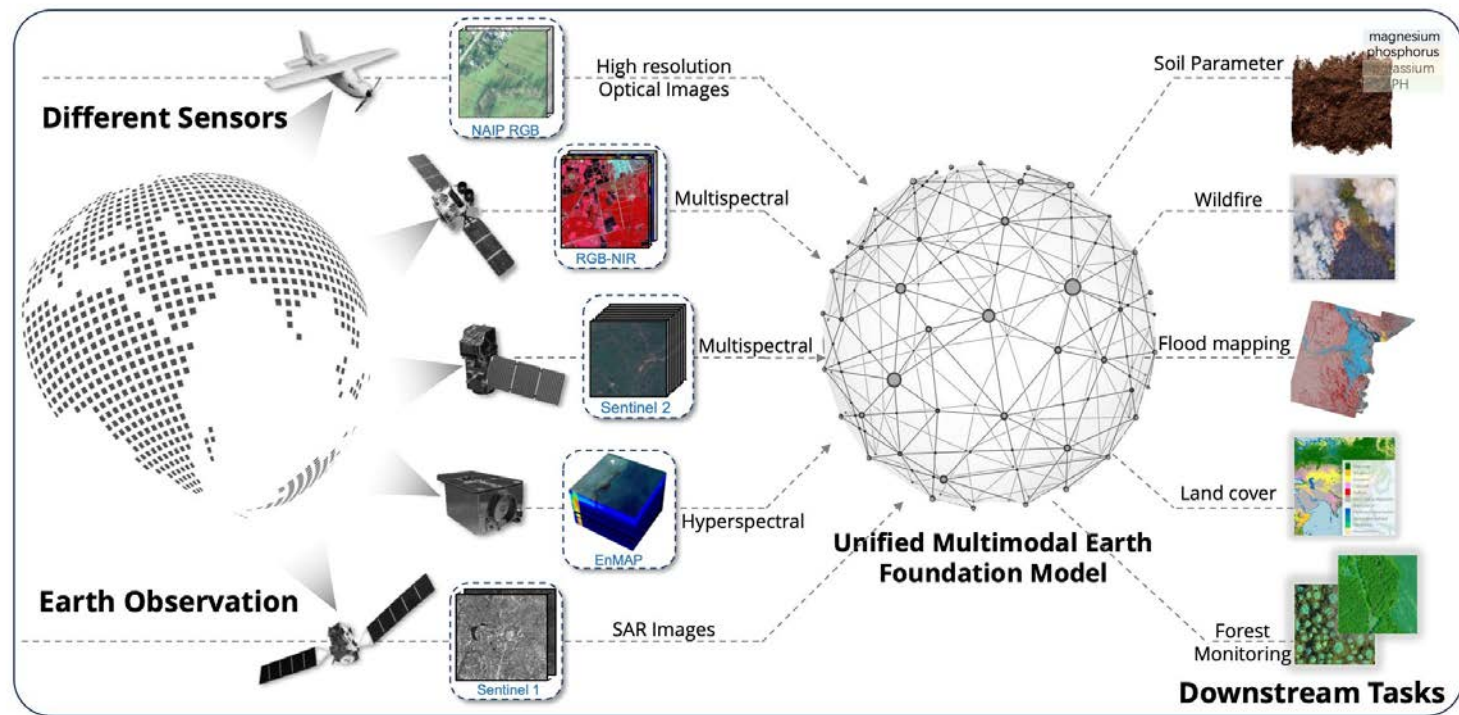
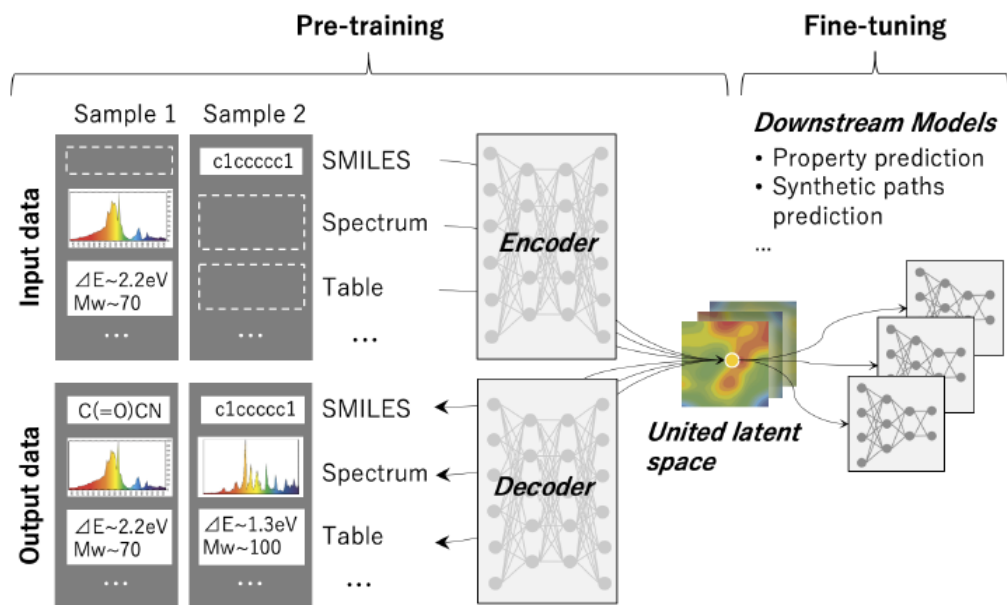
LLM fine-tuning at a high level

LLM fine-tuning



B. FOUNDATION MODELS FOR SCIENCE

Multi-modal foundation models (FMs)



B) FOUNDATION MODELS FOR SCIENCE

Methods. *Develop (methodology for pre-training and fine-tuning) multimodal foundation models, to be used for different combinations of modalities on different downstream tasks in different domains of science.*

Task B1: *Vision-language models for anomaly detection*

Task B2: *Multimodal FMs that support additional modalities*

Task B3: *Aligning different modalities into a single space*

B) FOUNDATION MODELS FOR SCIENCE

Applications. Learn and apply multimodal foundation models in different scientific domains, including medicine and healthcare (and broader life sciences, LS), as well as materials science (PE).

Task B4: Multimodal foundation models in medicine and healthcare

Task B5: Multimodal foundation models in the life sciences

Task B6: Multimodal foundation models in materials science

B. FOUNDATION MODELS FOR SCIENCE (Example from nutrition)

LLMs can be adapted (with own data) to specific domains: As an example, we have
Adaptation of the Llama 3 model to the nutrition domain

Model is fine-tuned on several nutrition datasets to be able to:

- Extract food related named entities (NE): NER (recognition) / NEL (linking)
- Classify food entities according to several food taxonomies (e.g., FoodON)
- Retrieve nutritional values for ingredients and recipes

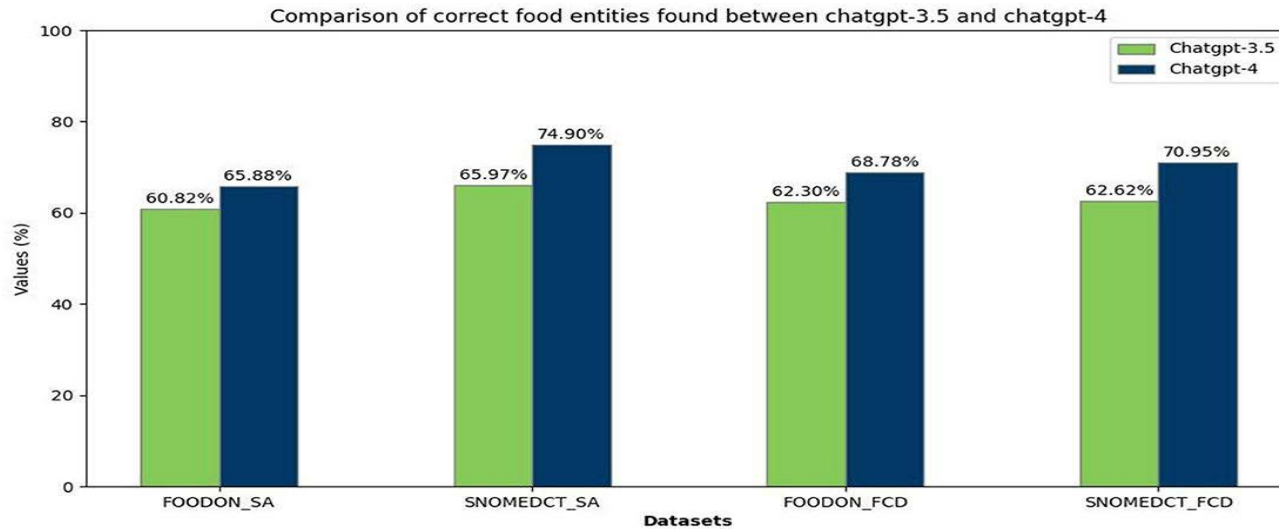
EXAMPLE:

Input: Compute the nutrient values per 100 grams in a recipe with the following ingredients: 250 g cream, whipped, cream topping, pressurized, 250 g yogurt, greek, plain, nonfat, 50 g sugars, powdered.

Output: Nutrient values per 100 g listed: energy - 179.00, fat - 10.28, protein - 6.09, salt - 0.05, saturates - 6.34, sugars - 14.00

B. FOUNDATION MODELS FOR SCIENCE (Example from nutrition)

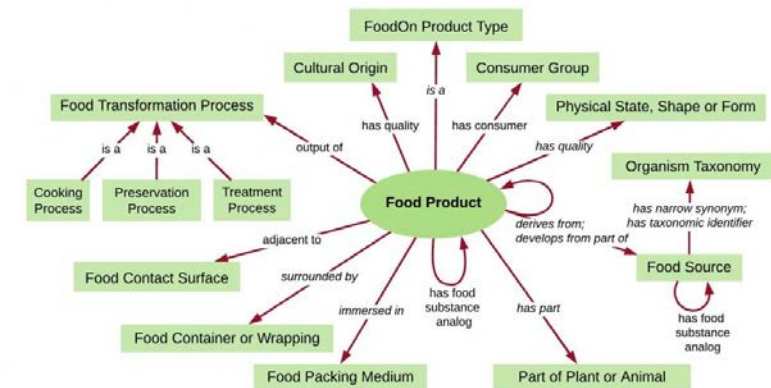
LLMs fine tuned with own data for NER (results):



LLMs fine tuned with own data for NEL (results):

ChatGPT results: 0.0

NEL: Excessive **salt** [00002 (FOODB)] intake has been associated with a higher incidence of **heart disease** [0001(UMLS)].



B. FOUNDATION MODELS FOR SCIENCE (Example from nutrition)

LLMs fine tuned with own data for NEL:

- Multi-task fine-tuning of open LLMs
 - NER
 - NEL
 - Recipe nutrient value prediction
 - Predict food traffic light system

Own datasets crucial for fine-tuning

NEL - FoodON

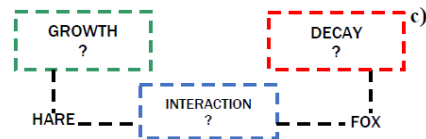
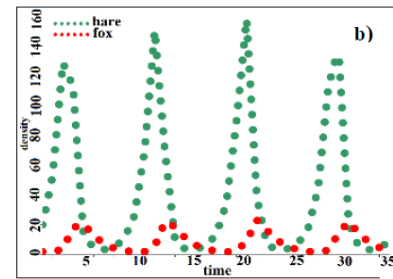
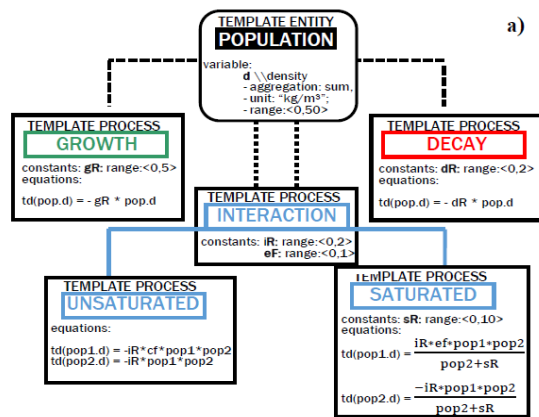
Test Fold	F1
0	0,941
1	0,942
2	0,942
3	0,944
4	0,943

Nutrient value prediction

Test Fold	F1
0	0,965
1	0,966
2	0,963
3	0,963
4	0,963

C. AUTOMATED SCIENTIFIC MODELLING

Learning scientific models in the form of equations from data and domain knowledge



d)

```

entity hare : Population; {
  vars: d {role endogenous; initial:20}; }
entity fox : Population {
  vars: d {role endogenous; initial:2}; }
process growth (hare): Growth
{ consts: gR=2.5;}
process decay (fox): Decay
{ consts: dR=1.2;}
process predator_prey (fox,hare): UnsaturatedPP
{ consts: eF = 0.1, iR = 0.3;}

```

e)

$$\frac{d}{dt} hare_d = 2.5 * hare_d - 0.3 - hare_d * fox_d$$

$$\frac{d}{dt} fox_d = 0.1 * 0.3 * hare_d * fox_d - 1.2 * fox_d$$

C) AUTOMATED SCIENTIFIC MODELLING

Methods. *Develop AI methods for learning scientific models represented as different kinds of equations, from both data and domain knowledge, using symbolic and neural approaches for good fit and interpretability.*

Task C1: *Equation discovery with attribute grammars*

Task C2: *Discovering different kinds of differential equations*

Task C3: *Neuro-symbolic equation discovery*



C) AUTOMATED SCIENTIFIC MODELLING

Applications. Apply AI methods for learning scientific models in the form of equations, from scientific data in different domains: plant biology and ecology (LS), as well as electrochemistry and materials science (PE).

Task C4: Estimating reaction rates in a stress signalling network

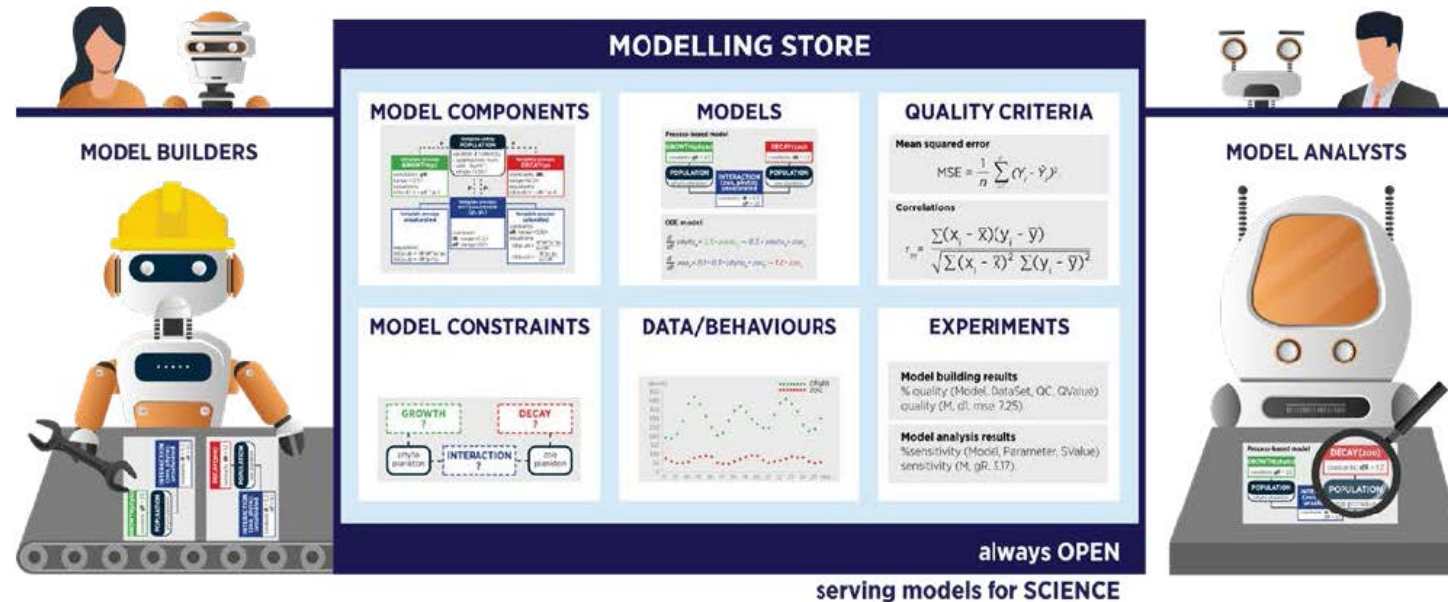
Task C5: Integrated modelling of aquatic ecosystems

**Task C6: Equation discovery in electrocatalysis/
electrochemistry**



D. SEMANTIC TECHNOLOGIES FOR OPEN SCIENCE

Not only data should be FAIR, but all of the artefacts of scientific modelling (e.g. models): We need to represent, annotate and store them, so that they can be found and re-used



D) SEMANTIC TECHNOLOGIES FOR OPEN SCIENCE

Methods. *Develop semantic resources describing the experimental computer science branches of ML and optimization (incl. descriptions of methods, tasks and performance), as well as semantic resources to support AI applications in different scientific domains.*

Task D1: *Semantic resources for complex data/ tasks and advanced ML methods/ models*

Task D2: *Semantic resources for single- and (constrained) multi-objective optimization*

Task D3: *Semantic resources to support AI applications in specific scientific domains*

D) SEMANTIC TECHNOLOGIES FOR OPEN SCIENCE

Applications. Apply explainable ML methods to relate properties of tasks/problems and properties/configurations of algorithms to algorithm performance in the areas of ML (considering more complex ML tasks) and the area of optimization (considering SOO, MOO and CMO problems).

Task D4: AutoML for more complex ML tasks, e.g., multi-target prediction and semi-supervised learning

Task D5: AutoOPT on single-objective optimization experiment databases

Task D5: AutoOPT on multi-objective optimization experiment databases

Synergy with other projects, e.g., at JSI-E8

Projects on the topic of materials science

- *4D STEM of energy related materials down to quantum level (ARIS large project, Led by JSI-K5, Andreja Bencan Golob)*
- *Fundamental understanding of Hydrogen Evolution Reaction for a new generation of nickel-based electro-catalysts in alkaline water and chlor-alkali electrolysis (ARIS large project, Led by KI, Dusan Strmcnik)*
- *DAEMON: Data-driven Applications towards the Engineering of functional Materials: an Open Network (EU, COST)*



Synergy with other projects, e.g., at JSI

Other EU Projects

- *ELIAS: European Lighthouse of AI for sustainability (E8, E3)*
 - *Design (of materials)*
 - *Automated modelling of dynamical systems from data and domain knowledge*
 - *Surrogate modelling (for speeding up simulations)*
 - *Synergy with WP C of AI4Sci*
- *Leveraging Benchmarking Data for Automated Machine Learning and Optimization (E7, E8)*
 - *ERA Chair*
 - *Synergy with WP D of AI4Sci*



New project that includes JSI and UL

Large Language Models for the EU

- *WP3: Catalogue for Large Language Tools and AI Models*

The activity consists of making large language models (LLMs), and more generally language technology tools, available for the largest possible community. This will be achieved by providing access to an online catalogue that aims to facilitate the discovery of models, as well as the understanding of their specific requirements and constraints. The objective of this activity is to incentive the ecosystem to produce models optimized for use in a specific language, sector or use case, and to reduce the barriers to entry into the use of these models for SMEs.



International Meetings in the Area of AI for Science

- *The Discovery Science Conference*
- *AAAI Symposia Series*
 - *2023 Spring Symposium: Computational Approaches to Scientific Discovery*
 - *2024 Fall Symposium: Integrated Approaches to Computational Scientific Discovery*
- *The Nobel-Turing Workshop Series*
- *ICML Workshops on AI for Science*
- *NeurIPS Workshops on AI for Science*



DS 2025

Ljubljana, Slovenia
September 22-26, 2025

Artificial Intelligence for Science

An extended edition of the **Discovery Science** conference



DS-2025: The program

An extended edition of DS, devoted to AI in Science

Special sessions, e.g., on:

- Explainable AI 4 Science
- Symbolic regression 4 Science
- Semantic technologies 4 Science
- Foundation models 4 Science
- Physics informed NNs 4 Science



One-day parallel events on:

- AI & Environmental sciences
- AI & Life sciences
- AI & HPC
- AI & Materials science
- AI & Physics

Our AI4Sci project is timely!



European
Commission



National Policies for AI in Science

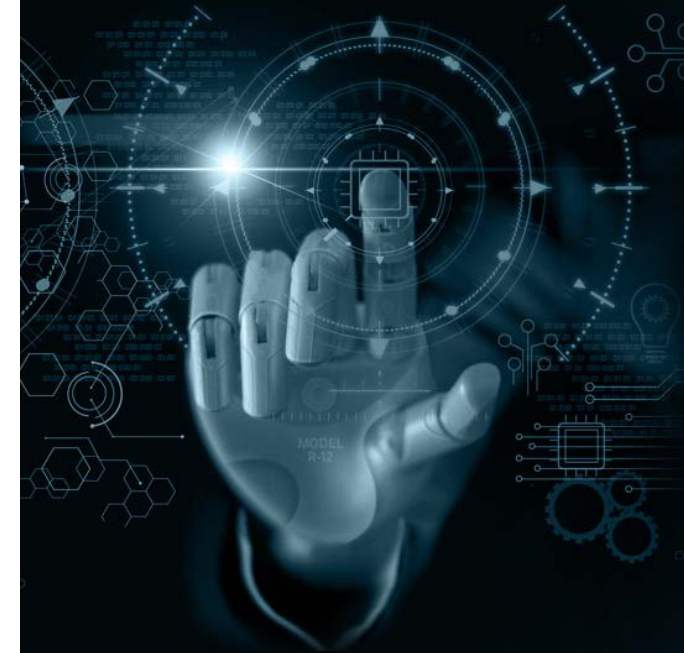
#HorizonEU

POLICY SUPPORT FACILITY (PSF) CHALLENGE - MUTUAL LEARNING EXERCISE (MLE)

European
Commission

Scientific Advice Mechanism

Successful and timely uptake of
Artificial Intelligence
in science in the EU





MANY THANKS
FOR YOUR ATTENTION!