Jakub Vrábel

🌐 jvrabel.github.io | in Linkedin | ✉ jakub.vrabel@ceitec.vutbr.cz | 🎓 GoogleScholar

Bio

I am a postdoctoral researcher working on foundational machine learning and AI safety, with a focus on loss-landscape geometry/connectivity/symmetries, interpretability, and security. I have a Ph.D. in appl. physics, where I specialized in interpretable ML for spectroscopic data and physics-inspired learning.

Education

2019 - 2025	Ph.D. degree in applied physics	Central European Institute of Technology (CZ)
2024 (3.5 mo.)	visiting grad. student (w/ D. Krueger)	CBL, University of Cambridge (UK)
2021 (3 mo.)	visiting grad. student (w/ H. Lohninger)	Technische Universität Wien (AT)
2017 - 2019	Master's deg. - appl. physics (*with distinction*)	FME, Brno University of Technology
2016 - 2017	additional Bc. studies - appl. physics	FME, Brno University of Technology
2013 - 2016	Bachelor's degree - mechatronics	FME, Brno University of Technology
extra courses:	statistical physics of fields, GTR, advanced quantum mech., Lie groups, data mining	

Work Experience

2025 - present	Visiting research fellow – AI Governance Initiative, University of Oxford (UK), 3 mo. *Topic: automated interpretability + hallucination mitigation in LLMs*
2016 - present	Research associate – CEITEC (CZ), Group of Laser Spectroscopy *Topic: machine learning for spectroscopic data, laser-induced plasma physics*
2018	Research intern – Complutense University of Madrid (ES), 6 months *Topic: characterization of laser-induced plasmas (with prof. Caceres)*
2015 - 2017	Multimedia specialist – Moravia IT (now RWS) (CZ), part-time *Localization of Microsoft Office (video editing, vector/raster graphics, scripting)*

Summer schools

2025, 2024, 2023	ELLIS Summer School - Probabilistic ML (Cambridge, UK)
2024, 2023, 2022	IAIFI Summer School - Physics and AI (Cambridge/Boston, MA, USA)
2022	Arnold Sommerfeld Centre Summer School - Physics meets AI (Munich, DE)
2022	Erwin Schrödinger Institute Summer School – ML for materials (Vienna, AT)
2021	ML in Quantum Physics and Chemistry Summer School (Warsaw, PL)

Selected Publications

2024   *Input Space Mode Connectivity in Deep Neural Networks.* **Vrábel, J.**, Shem-Ur, O., Oz, Y., Krueger, D. ICLR 2025, SciForDL NeurIPS 2024 Workshop (oral), tweet.

2024   *Interpreting convolutional neural network classifiers applied to laser-induced breakdown optical emission spectra.* Képeš, E., **Vrábel, J.**, et al. Talanta, DOI: 10.1016/j.talanta.2023.124946.

2023   *Spectral library transfer between distinct Laser-Induced Breakdown Spectroscopy systems trained on simultaneous measurements.* **Vrábel, J.**, et al. JAAS, DOI: 10.1039/D2JA00406B.

2020   *Classification of challenging Laser-Induced Breakdown Spectroscopy soil sample data-EMSLIBS contest.* **Vrábel, J.**, et al. Spectrochim. Acta B, DOI: 10.1016/j.sab.2020.105872.

2020   *Restricted Boltzmann Machine method for dimensionality reduction of large spectroscopic data.* **Vrábel, J.**, et al. Spectrochim. Acta B, DOI: 10.1016/j.sab.2020.105849.

*   a complete list of publications can be found on Google Scholar

Stats

- H-index: 12 (450+ citations)
- 15 journal articles
- 8 conference talks, 5 posters
- P.I. of 5 student projects (funding 50,000 €)
- R.A. in 4 projects (tot. funding 500,000+ €)
- superv. of 7 students (undergrad., internship)

Conferences

2025	ICLR (Singapore)	poster
2024	IAIFI workshop (Cambridge, MA, USA)	oral presentation
2024	Youth in High Dimensions (Trieste, IT)	poster
2023	IAIFI workshop (Boston, USA)	poster
2022	SCIX (Covington - KY, USA)	oral presentation
2022	LEA (Tokyo, JP)	(online) oral presentation
2022	IAIFI workshop (Boston, USA)	poster
2022	LIBS2022 (Bari, IT)	oral presentation
2021	EMSLIBS (Gijon, ES)	(online) oral presentation
2020	LIBS2020 (Kyoto, JP)	(online) oral presentation
2020	IOM-LIBS (fully-online)	(online) oral presentation
2019	EMSLIBS (Brno, CZ)	poster and oral presentation
2017	EMSLIBS (Pisa, IT)	poster

Skills

machine learning (deep learning, kernel methods), statistical physics, LLMs, computer vision (YOLO), Python (JAX, Pytorch, scikit-learn), R, MATLAB, GitHub, HPC, cloud & GPU computing (Azure)

Other Achievements & Activities

rector's prize (2024), Josef Hlávka award (2023), 2nd place in the Competition of Young Spectroscopists (IMMSS, 2023), rector's honorable mention (2019), involved in 3 national-wide and 2 international projects, teaching seminars for Physics 1 & 2.

Ongoing Work

1. *Mitigating contextual hallucinations in LLMs* (with Fazl Barez). LLMs may hallucinate even with correct information available, often by missing key context or overruling retrieved facts with parametric knowledge. We address this via prefix optimization to improve context use in RAG. My role: idea, concept, experiments, theory.
2. *LLM knowledge distillation* (with David Krueger). In current large language models (LLMs), selectively *blacklisting* or *unlearning* (potentially harmful) capabilities from general-purpose models is arguably infeasible. We propose and test an alternative approach to build a "narrow AI" based on knowledge distillation. My role: numerical experiments, concept design.
3. *Neural tangent kernel (NTK) and parameter space symmetries* (with O. Shem-Ur and Y. Oz). We use the NTK framework to study parameter space symmetries and explain the mechanisms behind their role in improving convergence and generalization. My role: numerical experiments.

Further career plan

With a deep passion for science, particularly ML and physics, I am driven to learn and explore these areas continuously. My aim is to make valuable contributions to the scientific community in foundational ML/AI. My next objective is to find a postdoc position that aligns with my interests and aspirations.