

Pablo Lemos, Ph.D.

Machine Learning Scientist

+34 656 482 232
plemos91@gmail.com
<https://pablo-lemos.github.io>
[in plemosp](#)
[pablo-lemos](#)
ORCID: 0000-0002-4728-8473 Twitter: @PabloLemosP

Profile

Highly driven Research Scientist with 7+ years of postdoc and industry experience working in AI for science. With over 100 academic papers, including top Machine Learning conferences (ICML, NeurIPS, ICLR) and journals. Expertise includes diffusion and flow-based generative models, Bayesian inference, uncertainty quantification, and constructing large-scale experimental pipelines to probe failure modes and robustness.

Experience

2024–Present Machine Learning Scientist, *SandboxAQ*

- Developed state-of-the-art protein folding and protein–ligand affinity models.
- Led the creation of **SAIR**, the largest publicly available synthetic protein–ligand dataset; designed data-quality audits and evaluation metrics.
- Co-led a project on causal machine learning and foundation models for virtual cell modelling.
- Represented SandboxAQ at the OpenBind meeting, in No. 10 Downing Street.
- Contributed to the development and training of **OpenFold3**

2024 Research Scientist Specialist, *DreamFold*

- Designed and implemented **FoldFlow-2**: an equivariant flow-matching system for conditional protein backbone generation.
- Co-created **Iterated Denoising Energy Matching** for improved sampling from Boltzmann-like energy distributions.
- Led LLM fine-tuning and active-learning for protein sequence/data acquisition; built evaluation tooling and safety checks for generated candidates.

2022–2024 Research Scientist, *Mila / Université de Montréal*

- Researched Generative Flow Networks and diffusion-based models under Prof. Yoshua Bengio's group.
- Designed robustness and calibration experiments for large generative models applied to astrophysical inference.
- Coordinated multi-group collaborations and ran reproducible large-scale experiments.

2021–2022 Postdoctoral Research Scientist, *University of Sussex*

- Applied Bayesian neural networks and simulation-based inference to cosmological parameter estimation.
- Built evaluation pipelines for uncertainty quantification and model comparison.

2018–2021 Postdoctoral Research Scientist, *University College London*

- Developed Bayesian consistency frameworks used across the cosmology community; applied GNNs for scientific discovery.
- Led seminars, supervised students, and managed collaborative projects in Bayesian methods.

Education

2025 **Coursework**, *Harvard Medical School*, Biochemistry
2025 **Coursework**, *Harvard Medical School*, Drug Discovery and Development
2015–2018 **Ph.D., Astrophysics**, *University of Cambridge*
2013–2015 **M.Sc., Physics (GPA 3.98/4.00)**, *University of Chicago*
2009–2013 **B.Sc., Physics (Top 1%)**, *Universidad Complutense de Madrid*

Technical Skills

- **Research & Modelling:** Diffusion models, flow matching, Generative Flow Networks, energy-based sampling, Bayesian inference, uncertainty quantification, interpretability experiments.
- **Software & Systems:** Python, C++, Fortran; PyTorch, TensorFlow, JAX; Docker; Git/GitHub.
- **Domain Tools:** AlphaFold, OpenFold, RFdiffusion, ProteinMPNN, pyRosetta.
- **Other:** Dataset design (SAIR), experimental design, writing and presenting, mentorship.

Selected Publications

- Lemos, P. et al. **SAIR: Enabling Deep Learning for Protein-Ligand Interactions with a Synthetic Structural Dataset**. bioRxiv (2025).
- Sendera, M., Kim, M., Mittal, S., Lemos, P., et al. **Improved off-policy training of diffusion samplers**. NeurIPS (2024).
- Huguet, G., Vuckovic, J., Fatras, K., Thibodeau-Laufer, E., Lemos, P., et al. **Sequence-Augmented SE(3)-Flow Matching For Conditional Protein Backbone Generation**. NeurIPS (2024).
- Akhound-Sadegh, T., Rector-Brooks, J., Bose, A. J., Mittal, S., Lemos, P., et al. **Iterated denoising energy matching for sampling from Boltzmann densities**. ICML (2024).
- Lemos, P., Sharief, S., Malkin, N., Perreault-Levasseur, L. & Hezaveh, Y. **PQMass: Probabilistic Assessment of the Quality of Generative Models**. ICML (2024).
- Lahlou, S., Deleu, T., Lemos, P., Zhang, D., Volokhova, A., Hernández-García, A., Nehale Ezzine, L., Bengio, Y. & Malkin, N. **A theory of continuous generative flow networks**. ICML (2024).
- Lemos, P., Malkin, N., Handley, W., Bengio, Y., Hezaveh, Y. & Perreault-Levasseur, L. **Improving Gradient-guided Nested Sampling for Posterior Inference**. ICML (2023).
- Lemos, P., Jeffrey, N., Cranmer, M., Ho, S. & Battaglia, P. **Rediscovering orbital mechanics with machine learning**. MLST (2022).

Languages

English (fluent), Spanish (native), Galician (native), Italian (beginner), French (beginner), Portuguese (beginner).