

Profile

Researcher at the intersection of biology and machine learning, with formal training in biochemistry, experimental medicine, and AI. Brings a proven track record of designing and leading projects, mentoring students, and developing software. Experience with training generative biomolecule models including [SqueezeProt](#), a protein language model, as well as controlling for [data leakage in pre-trained foundation models](#).

Education

Doctor of Philosophy

McGill University & Mila

Department of Electrical Engineering

Fall 2020 - Winter 2026

- Advisor: Prof. Amin Emad (amin.emad@mcgill.ca)
- Thesis: “*Addressing Pervasive Challenges in Generalizable Machine Learning Models of Protein-Protein Interaction*”
- Developed **SqueezeProt**, a fast and generalizable protein LLM achieving $5\times$ faster inference than SotA with comparable performance, using an efficient architecture and strict evaluation protocols.
- Created **INTREPPID**, a contrastive learning model for cross-species protein interaction prediction.
- Designed and implemented data pipelines for large-scale, curated protein datasets, focusing on **mitigating data leakage**.
- Published three first-author papers in **leading bioinformatics journals** (*Nature Machine Intelligence*, *Briefings in Bioinformatics*, & *Bioinformatics*), presenting at **premiere international computational biology conferences** (ISMB, CSHL NetBio).

Master of Science

McGill University & GCI

Department of Experimental Medicine

Fall 2015 - Winter 2018

- Advisor: Prof. Luke M. McCaffrey (luke.mccaffrey@mcgill.ca)
- Thesis: “*On the automated regional classification of early breast cancers*”
- Contributed to the publication of two manuscripts which have a **cumulative 184 citations**.
- **Experimental techniques** included:
 - Cell culture maintenance & Viral transfection
 - Tissue labelling (radio & fluorescent labelling) & Confocal microscopy

Bachelor of Science

McGill University

Department of Biochemistry, Honours Degree

Fall 2012 - Winter 2015

- Advisor: Prof. Alan Nepveu, Prof. Luke M. McCaffrey (alain.nepveu@mcgill.ca)
- Six semesters of undergraduate research, with a focus on DNA damage and early cancer progression.

Experience

Postdoctoral Researcher, McGill | Montréal, Québec

March 2026 - Present

- Continuing work on designing and fine-tuning protein language models specifically for protein interaction tasks.
- Designed/implemented graph partitioning and community-detection pipelines in Rust for large-scale protein interaction networks (STRING).
- Mentoring and co-ordinating a team of as many as four students.
- Reference: Amin Emad (amin.emad@mcgill.ca)

ML Developer, Coveo Solutions | Montréal, Québec

Feb - Aug 2018

- Researching deep learning applications for semantic document search. Automated feature engineering.
- Manager: Sébastien Paquet (spaquet@coveo.com).

Highlighted Publications

1. **Joseph Szyborski** and Amin Emad. 2026. "A flaw in using pretrained protein language models in protein-protein interaction inference models." *Nature Machine Intelligence*, DOI: [10.1038/s42256-025-01176-7](https://doi.org/10.1038/s42256-025-01176-7). Citations: 5
2. **Joseph Szyborski** and Amin Emad. 2024. "INTREPPPID—an Orthologue-Informed Quintuplet Network for Cross-Species Prediction of Protein-Protein Interaction." *Briefings in Bioinformatics* 25 (5): bbae405. DOI: [10.1093/bib/bbae405](https://doi.org/10.1093/bib/bbae405). Citations: 11, [4× more than average](#) papers in the field.
3. **Joseph Szyborski** and Amin Emad. 2022. "RAPPPID: Towards Generalizable Protein Interaction Prediction with AWD-LSTM Twin Networks." *Bioinformatics* 38 (16): 3958–67. DOI: [10.1093/bioinformatics/btac429](https://doi.org/10.1093/bioinformatics/btac429). Citations: 39, [7.8× more than average](#) papers in the field.

Software

[SqueezeProt](#)

A protein language model (pLM) which focuses on speed and was strictly evaluated, unlike other leading pLMs.

[PPI.bio](#)

Online interface for inferring protein-protein interactions.

[INTREPPPID](#)

Predict interactions between proteins of different organisms by their sequence using contrastive learning.

[PPI Origami](#)

CLI tool for preprocessing protein-protein interaction datasets for supervised machine learning.

[RAPPPID](#)

Predict interactions between proteins given their sequence using deep recurrent networks.

Awards and Honors

- **Postdoctoral Scholar Award** D2R Initiative (McGill University) 2025
– Value of \$140,000 (plus benefits)
- **Les Vadasz Scholar** McGill Engineering Doctoral Award 2020
– Value of \$128,000
- **Graduate Excellence Fellowship** McGill Engineering 2020
– Value of \$10,000
- **First Place**, ConUHacks Hackathon (Concordia University) 2017

Conferences

Talks: [ISMB '24](#), [ISMB '23](#), [ISMB '22](#), [GLBIO '23](#), [BIRS '22](#)

Posters: [ISMB '23](#), [GLBIO '23](#), [CSHL NetBio '23](#), [ISMB '22](#)

Skills

Languages: English (Native), French (Fluent), Italian (Conversational)

Technologies: Python, PyTorch, Hugging Face, Rust.