Stewart Slocum

https://stewyslocum.com | Google Scholar | (443) 986-0896 | sslocum3@mit.edu

Education

Ph.D. Computer Science, *Massachusetts Institute of Technology* **August 2022 - Present**
Advised by Professor Dylan Hadfield-Menell

B.S. Computer Science, Applied Math and Statistics, *Johns Hopkins University* **August 2017-May 2021**
GPA: 3.76
Advised by Professor Rene Vidal

Honors & Awards

NSF GRFP Fellowship

Publications, Preprints, and Conference Presentations

[1] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, **Stewart Slocum**, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, Dylan Hadfield-Menell. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *Under review at Transactions on Machine Learning Research* (2023).

[2] Aditya Chattopadhyay, **Stewart Slocum**, Benjamin D. Haeffele, Rene Vidal, Donald Geman. "Interpretable by Design: Learning Predictors by Composing Interpretable Queries." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[3] Lauren Wheelock, **Stewart Slocum**, Jorma Gorns, Sam Sinai. "Risk-Adjusted Selection for Validation of Sequences in AAV Design Using Composite Sampling." *American Society of Gene & Cell Therapy Conference* (2021).

[4] Sam Sinai, Richard Wang, Alexander Whatley, **Stewart Slocum**, Elina Locane, Eric Kelsic. "AdaLead: A simple and robust adaptive greedy search algorithm for sequence design." *arXiv preprint arXiv:2010.02141* (2020).

[5] Vikram Mulligan, Hans Melo, Haley Merritt, **Stewart Slocum**, Brian Weitzner, Andrew Watkins, P Douglas Renfrew, Craig Pelissier, Paramjit Arora, Richard Bonneau. "Designing peptides on a quantum computer." *bioRxiv* (2019).

Work Experience

Algorithmic Alignment Group, MIT *PhD Student* **August 2022 -**
Working on robust reward learning methods for safe and aligned AI systems.

1. Building safety guardrails around LLM chatbots using Inverse Prompt Engineering, a new method that interprets prompts as observations about the true goal rather than as its definition. By inverting a principled probabilistic model of the human prompt engineer, we obtain strong out-of-distribution robustness and jailbreak defenses.

2. Building an unsupervised, multi-agent, game-theoretic mechanism for eliciting truthful responses in LLMs and advanced AI systems, leveraging the theory of Peer Prediction and Bayesian Truth Serum mechanisms.

Vision Lab, JHU *Assistant Research Scientist* **October 2020 - July 2022**
Two research projects on principled deep learning methods.

1. Worked on a new information-theoretic framework for building interpretable machine learning models through finding minimal, sufficient, interpretable explanations for model predictions. This resulted in the IEEE publication Interpretable by Design: Learning Interpretable Predictors by Composing Interpretable Queries.

2. Led a project developing physics-inspired, geometrically aware optimization algorithms for non-Lipschitz-smooth, non-convex optimization algorithms for deep learning. Contributions include proposing new algorithms, theoretical analyses, and performing large-scale deep learning experiments.

Dyno Therapeutics, *Machine Learning Intern* **June - August 2020 Full-time, - March 2021 Part-time**

Worked on sequence proposal strategies for Dyno, an ML startup designing viral vectors for gene therapy. Built algorithms using discrete optimization methods, generative models, and reinforcement learning to propose high-scoring sequences with respect to trained property predictors for the VP1 region of AAV viral capsids. Also created heuristics to validate methods and detect adversarial sequences. Working with another intern, I built the company's first end-to-end ML pipeline. My contributions were key to the design of two novel deep learning-driven libraries of AAV viral vectors with modified VP1 regions, targeting the nervous system and muscle tissue.

Work developed into an open-source benchmarking environment for biological sequence design problems called FLEXS and a methods paper AdaLead: A simple and robust adaptive greedy search algorithm for sequence design. Following this, I worked on a meta-selection algorithm inspired by Bayesian model averaging to choose the most promising sequences to synthesize out of a large pool of candidates under a variety of models and noise corruption settings. This led to an oral presentation at ASGTC 2021 (top 25% of submissions) Risk-Adjusted Selection for Validation of Sequences in AAV Design Using Composite Sampling.

NASA Goddard Space Flight Center, *Quantum Computing Research Intern* **Summer 2019**

Developed a quantum annealing solver on the D-Wave quantum computer for protein design problems. Designed and ran numerical experiments to characterize performance scaling with problem size and problem graph density. Built preprocessing heuristics that improved performance by an order of magnitude. This led to a preprint Designing Peptides on a Quantum Computer. We have since used our method to create the first quantum-designed peptide, a 16-residue molecule with an exotic mirror symmetry whose structure was confirmed experimentally. Our method is one of the first successful applications of quantum computers to a problem of non-trivial size. We are preparing this version of our work for journal submission.

NASA Goddard Space Flight Center, *Virtual Reality Software Engineering Intern* **Summer 2017 and 2018**

Built Virtual Reality scientific visualization tools to study point cloud data with applications in astrophysics, marine biology, and planetary science. See press coverage and journal publication.