Matt McManus

New York, NY — mattmcmanus41@gmail.com — mmcmanus1.github.io/research

Research Summary

My work spans scientific machine learning (SciML), mechanistic interpretability of transformers, and sequential decision-making with RL and neuro-symbolic methods. I like problems where prior knowledge and data meet—physics in navigation, graphs in security, and leakage-aware evaluation for LLMs. At MIT I worked with **Prof. Alan Edelman** (CSAIL/Julia Lab) and **Prof. Una-May O'Reilly** (CSAIL/ALFA). Leading *MIT Pokerbots* sparked a lasting interest in RL and multi-agent strategy.

Scientific ML (with Prof. Edelman). My M.Eng. thesis develops a physics-informed neural ODE for strapdown inertial navigation in GPS-denied settings. The model embeds rigid-body kinematics, SO(3) attitude constraints, and IMU bias dynamics; trained on simulated and real IMU trajectories, it reduces 3D position RMSE by ∼60% versus a tuned EKF in walk-forward tests. I built a Julia-based simulation and evaluation harness (sensor grades, trajectories, bias drift) with CI to stress-test robustness. Encoding known physics improved out-of-sample accuracy and made the results easier to interpret. [Thesis PDF]

Interpretable Learning (mechanism and complexity). I aim to move from "what accuracy?" to "what computation happens where, and can we intervene?"

- *Mechanistic probe of transformers.* In *How Do Transformers "Do" Math?*, we tested whether a transformer trained for linear regression represents the intermediate slope w. Linear probes found stable w encodings in hidden states. We then ran causal tests—reverse probes and representational interventions that overwrite the internal estimate $w \to w'$; outputs shift predictably. This is direct evidence the model uses a specific internal variable to "do the math," not just surface heuristics.
- *Low-complexity interpolation.* In *Low-Complexity Solutions for Interpolating DNNs*, we explored training schemes that bias deep nets toward small-norm, interpolating solutions. Across architectures we observed lower test error and smaller weight norms (consistent with double descent), along with reduced intermediate-feature complexity. Shaping complexity at train time made internal computations simpler to probe and more stable to regularize.

AI for Cyber Defense (with Prof. O'Reilly). At CSAIL/ALFA I built a graph-based cyber-defense simulator and cast defense as a sequential decision problem. We benchmarked RL baselines across topologies and information regimes with ablations and failure analysis. To increase transparency, I prototyped neuro-symbolic decision layers that distill learned behaviors into human-readable rules; after head-to-head tests, we pivoted when simpler RL achieved similar performance. In parallel, I examined LLM-assisted workflows for anomaly triage and attack-path reasoning. The common thread is explainable policies that yield actionable interventions. [GPT-3 Cyber Defense Paper PDF]

Applied ML in Quantitative Settings. At Two Sigma (part-time) I developed factor-neutral cross-sectional signals, implemented feature- and learner-level decorrelation (orthogonalization, correlation-penalized loss), and validated them with leakage-safe walk-forward pipelines (rolling normalization; OOS rank IC/IR). At Bridgewater (AIA Labs) I focus on reliable LLM evaluation and RL for prompts: formal temporal/leakage constraints, calibrated scoring with dispersion/consistency diagnostics, lightweight online A/Bs, and FastAPI/Kubernetes services for reproducible runs.