Bridging Dimensions: Confident Reachability for High-Dimensional Controllers

Air Force Research Laboratory | Information Directorate | Core Technical Competencies
Autonomy, Command and Control, and Decision Support (AC2)
Connectivity and Dissemination (CAD)
Cyber Science and Technology (CYB)
Processing and Exploitation (PEX)
AFRL INFORMATION INSTITUTE

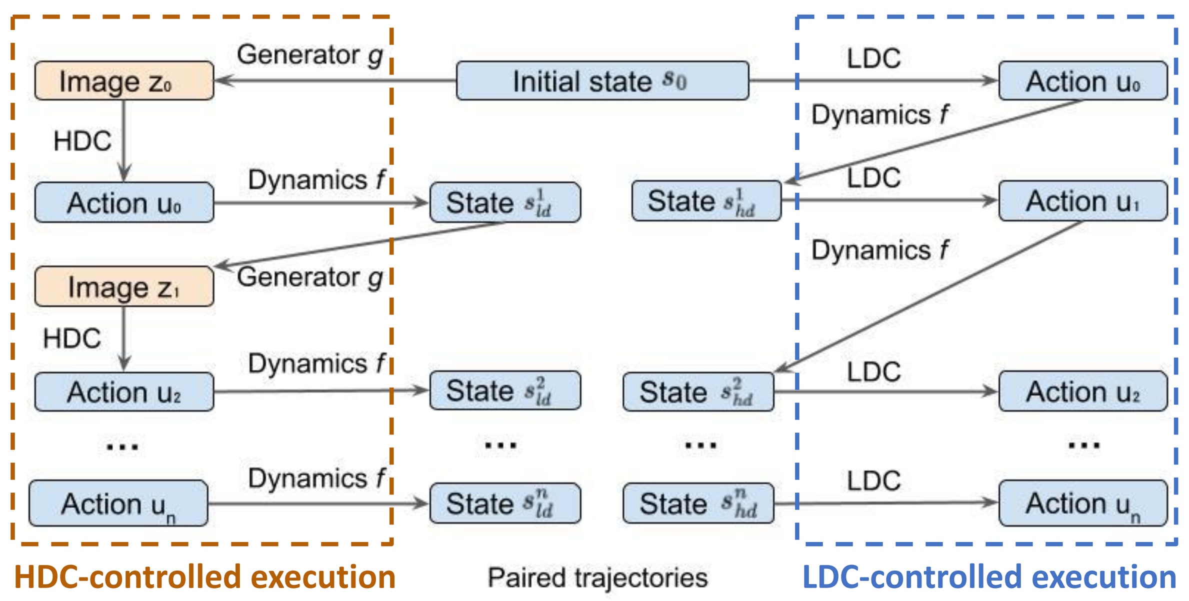
INFORMATION INSTITUTE MISSION: *Strengthen and expand information technology research, develop collaborative relationships, and increase research emphasis in areas of information technologies for the Information Directorate.*

PROBLEM

Autonomous systems, like self-driving cars and unmanned aircraft, rely on high-dimensional (e.g., vision-based) controllers (*HDC*) to perform complex and critical tasks.

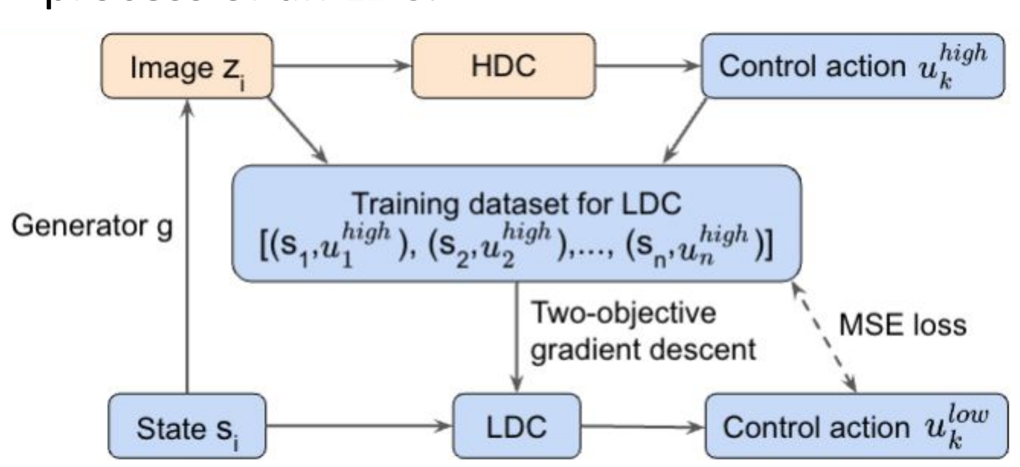
- However, the HDC-controlled systems lack *formal safety guarantees* on their behavior.

Goal: Perform reachability analysis on systems with HDCs, i.e., construct an overapproximated set of states that the system can reach from the initial set within a given time horizon. This reachable set can be intersected with goal/unsafe sets to provide a safety guarantee.



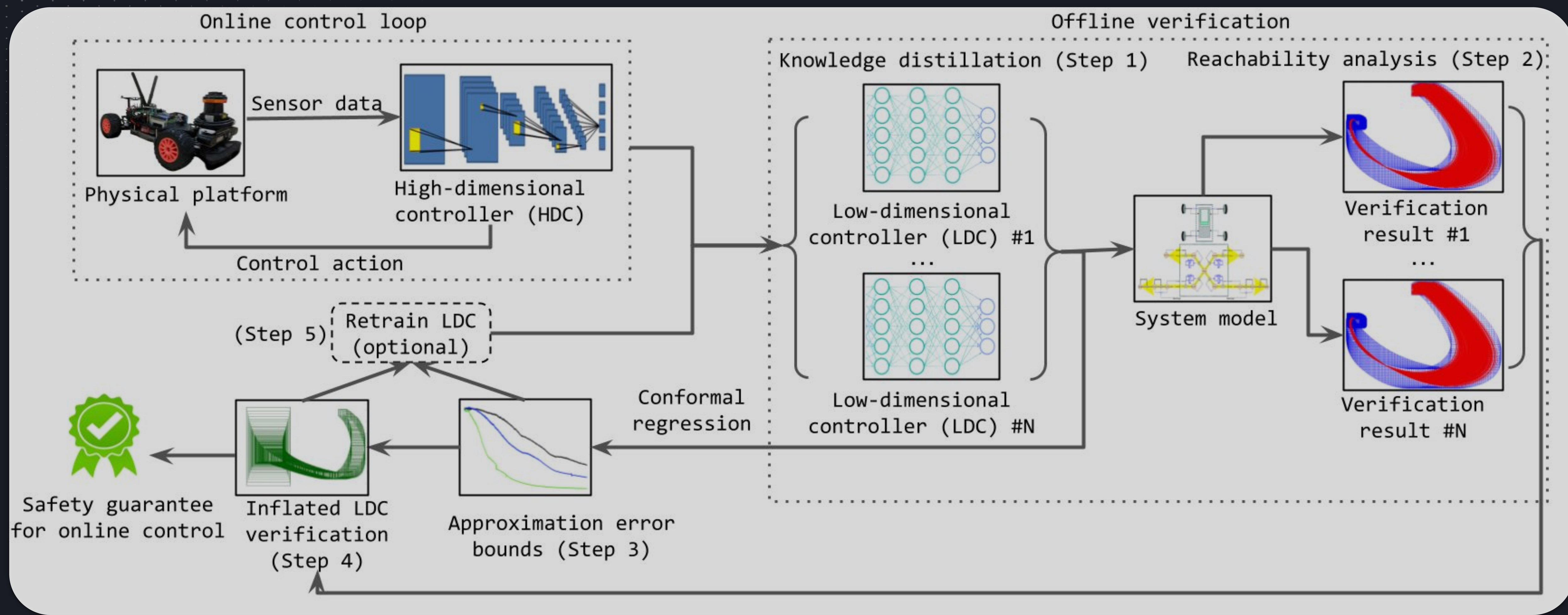
APPROACH

1. **Distill HDC knowledge:** Mimic the behavior of an HDC with multiple low-dimensional (state-based) controllers (*LDCs*). The training process of an LDC:



2. **Estimate HDC-LDC discrepancies:** Compute differences between HDC- and LDC-controlled systems. We introduce statistical upper bounds of two types: *trajectory-based* and *action-based*. Both are estimated with *conformal prediction* from labeled paired trajectories of LDC and HDC.
3. **Inflate LDC reachable sets:** We obtain an HDC reachable set by computing an LDC reachset using the *POLAR toolbox* and inflating it with either discrepancy from Step 2.

End-to-end safety verification of high-dimensional controllers:

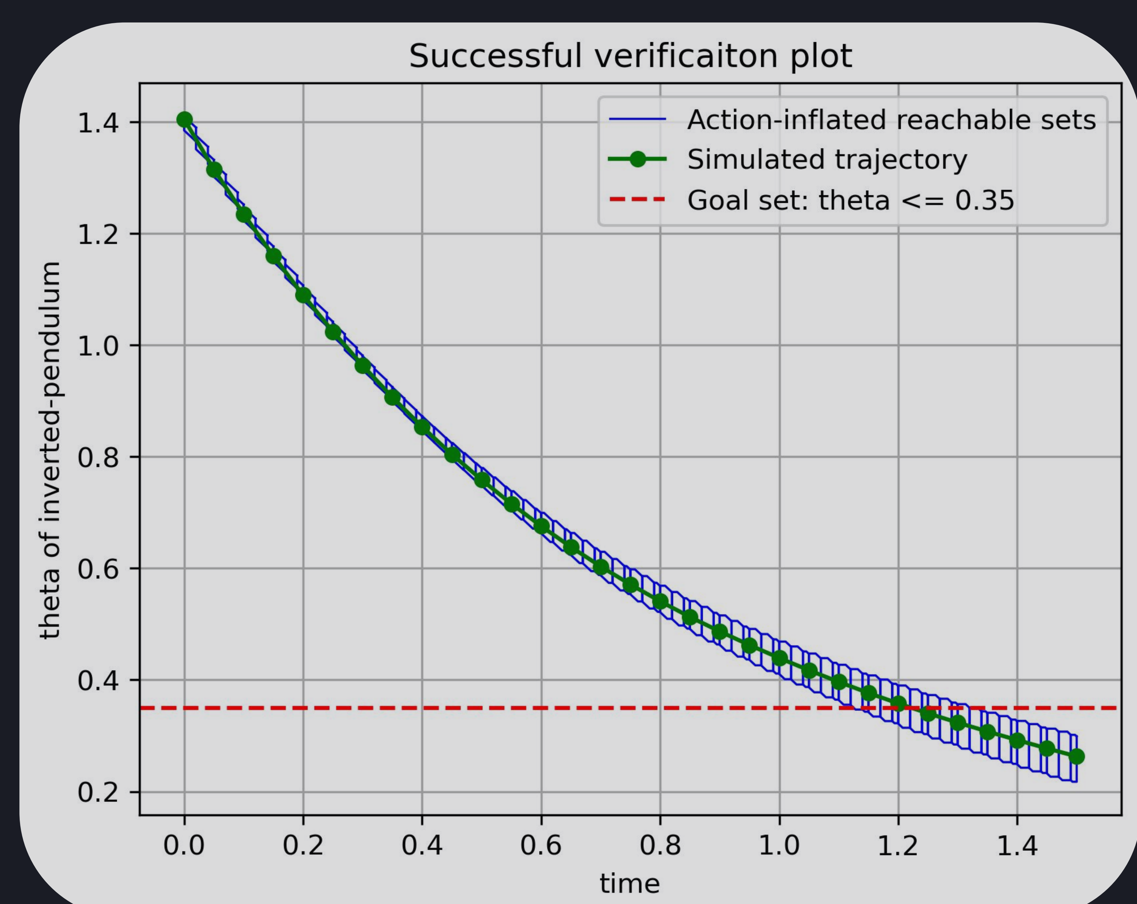


Major contributions:

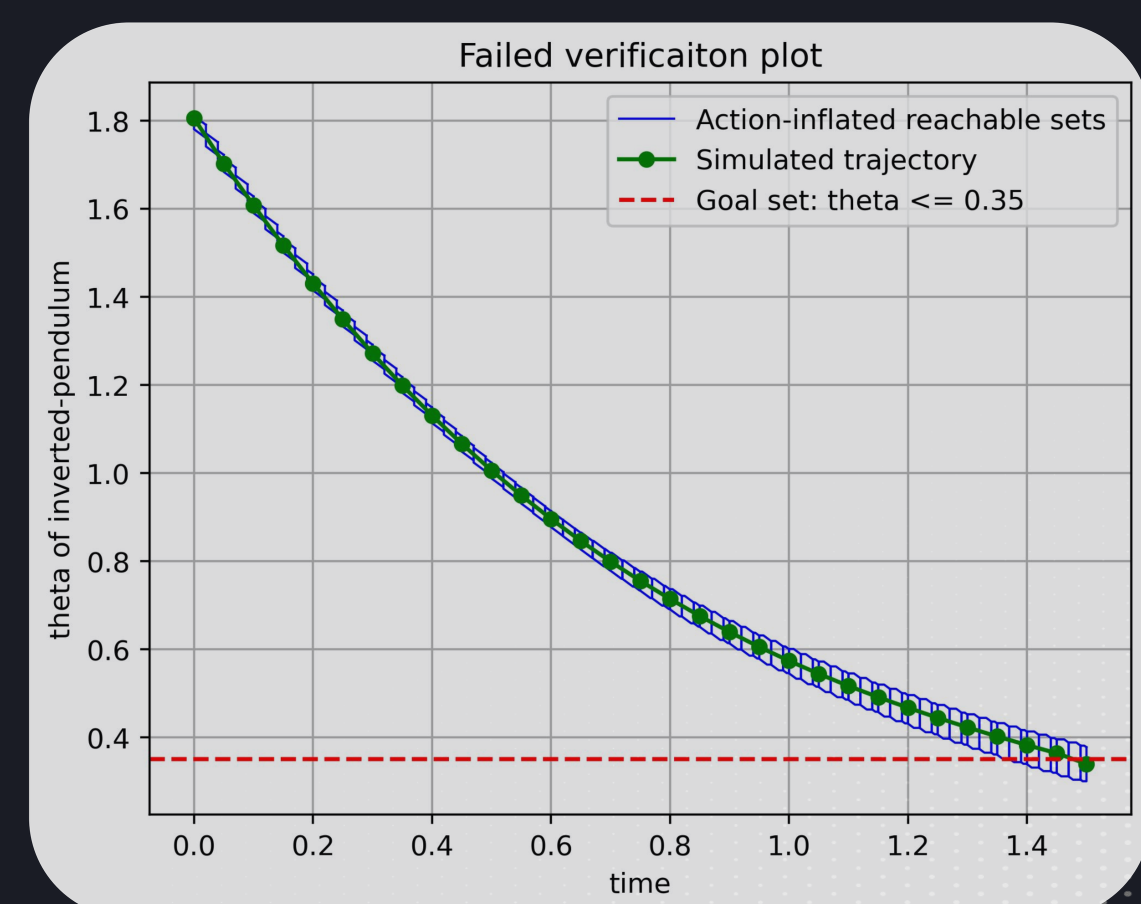
1. Reduce high-dimensional verification to the reachability analysis of multiple (4–10) *approximating low-dimensional controllers*.
2. Inflate reachable sets with statistical bounds on discrepancies (≈5%) between trajectories/actions using *conformal prediction*.
   - F1 score increased by 5–20 p.p. compared to a purely data-driven approach.

Examples of verification: *true positive* and *false negative*

ground truth and verification → safe



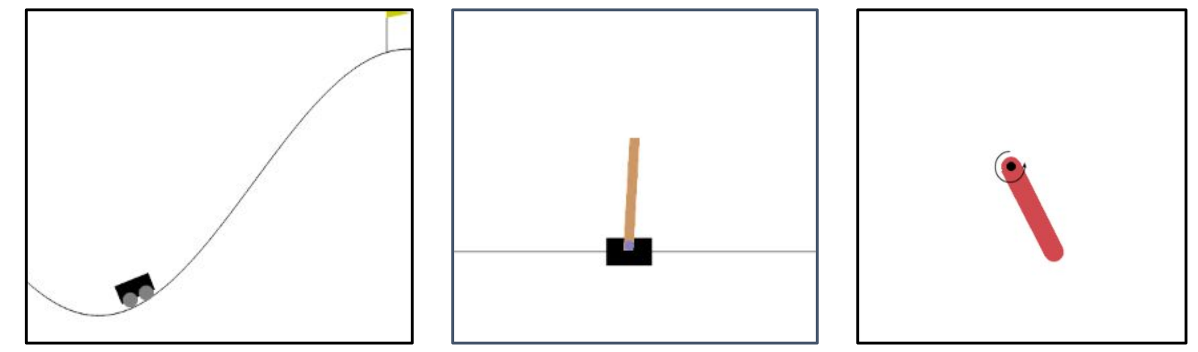
ground truth → safe, verification → unsafe



This research was supported by the Air Force Research Laboratory, Information Directorate, through the Visiting Faculty Research Program, Contract Number FA8750-20-3-1003.

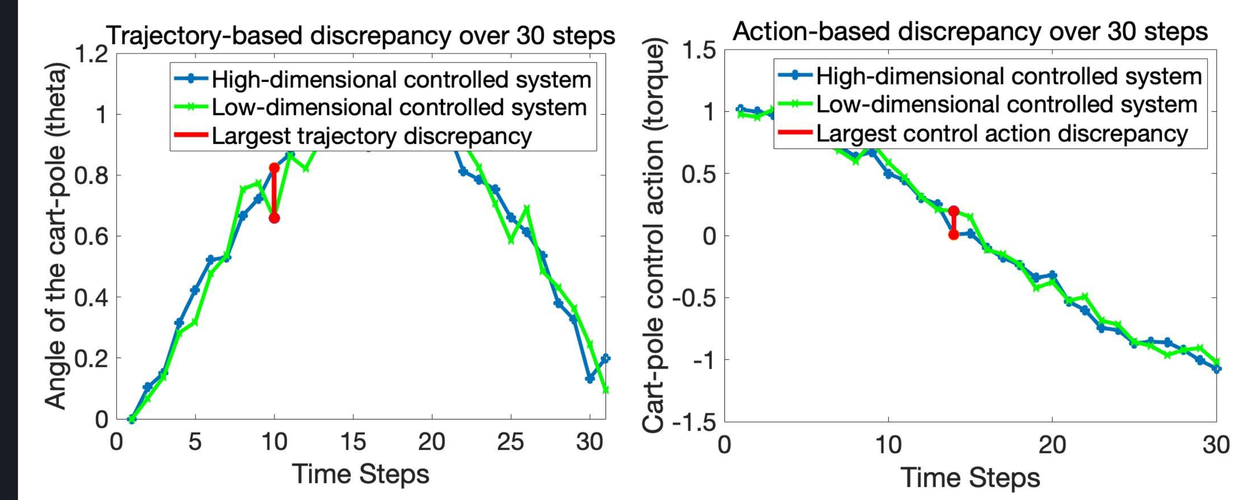
Approved for Public Release, Distribution Unlimited, AFRL-2024-5737

RESULTS: 3 CASE STUDIES



Mountain car | Cartpole | Inverted pendulum

Trajectory-based and action-based discrepancy bounds can differ significantly:



With a confidence level of 0.05, both approaches achieved a minimum precision of 0.95 and significant true positive rates. The trajectory-based multi-LDCs approach with showed best performance.

Table 1: Verification performance ($M = 4$ for IP and CP, $M = 10$ for MC).

Benchmark	Metrics	Trajectory-based approach		Action-based approach	
		1 LDC	M LDCs	1 LDC	M LDCs
Inverted Pendulum (IP)	True positive rate	0.4662	**0.7938**	0.0603	0.4050
	True negative rate	0.9976	**0.9995**	**1.0000**	**0.9999**
	Precision	0.9880	**0.9985**	**1.0000**	**0.9997**
	F1-score	0.6335	**0.8844**	0.1137	0.5765
Mountain Car (MC)	True positive rate	**0.7220**	**0.7207**	0.1050	0.2659
	True negative rate	0.9693	0.9872	**0.9964**	**1.0000**
	Precision	0.9621	0.9793	**0.9999**	**1.0000**
	F1-score	0.8249	**0.8303**	0.1900	0.4201
Cartpole (CP)	True positive rate	0.7225	**0.7450**	0.6554	0.7238
	True negative rate	0.9998	**1.0000**	**1.0000**	**1.0000**
	Precision	0.9999	**1.0000**	**1.0000**	**1.0000**
	F1-score	0.8389	**0.8539**	0.7918	0.8398

FULL PAPER

Yuang Geng, Jake Baldauf, Souradeep Dutta, Chao Huang, and Ivan Ruchkin, "*Bridging Dimensions: Confident Reachability for High-Dimensional Controllers*", in Proc. of the 26th International Symposium on Formal Methods (FM), 2024.

FUTURE WORK

- *Exhaustively* bridge HDC and LDC with satisfiability solving, without statistical bounds.
- Compute statistical bounds without sampling unlimited paired labeled trajectories.
- Develop end-to-end HDC verification toolbox.

Authors (University of Florida):
- **Ivan** Ruchkin, Asst. Professor
- Yuang Geng, PhD student

Mentors (AFRL/RITA):
- Matthew Anderson
- Steven Drager