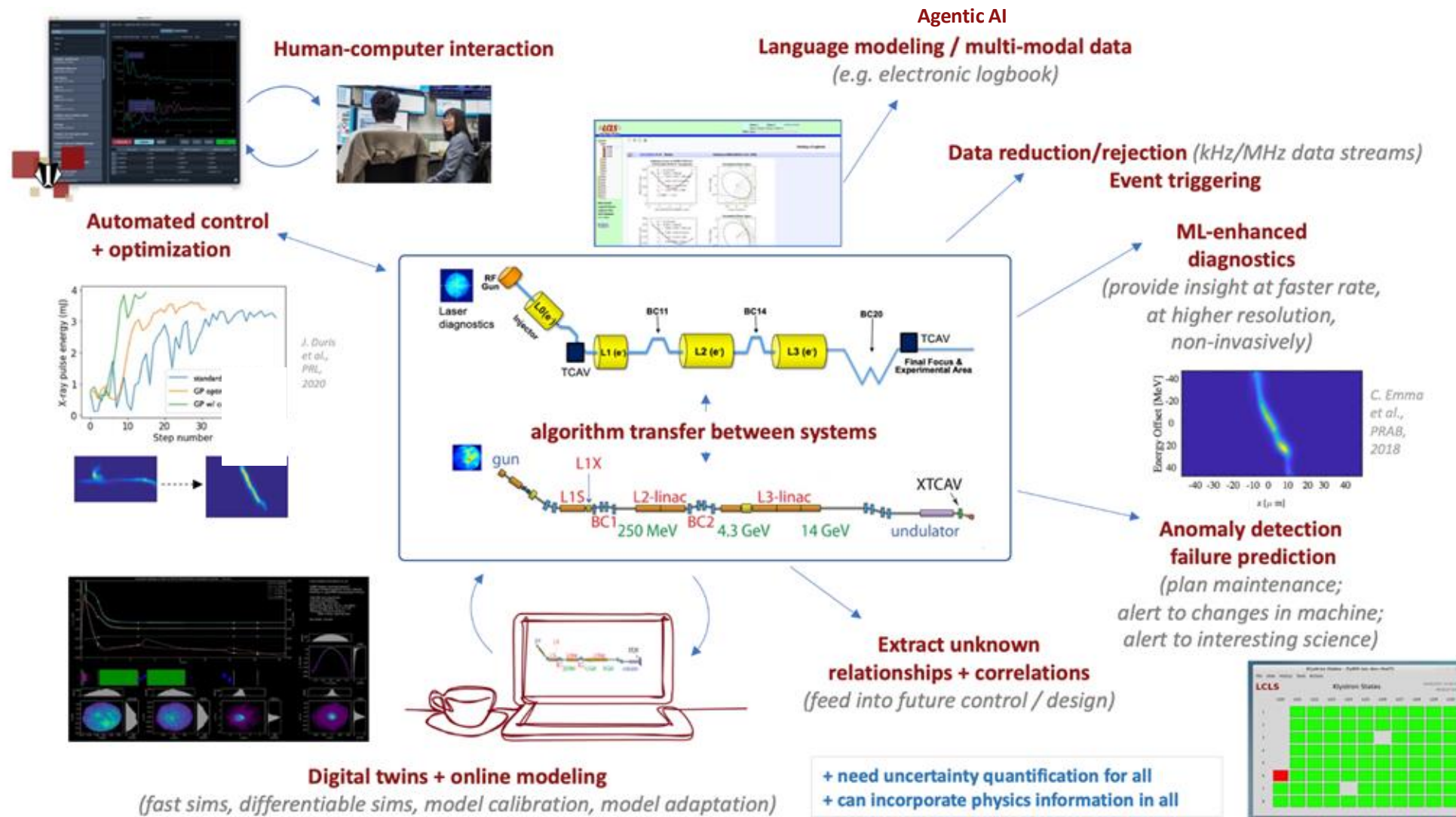


AI/ML in Operations and Design: Considerations for Future Accelerators

Auralee Edelen

US FCC week, September 10, 2026

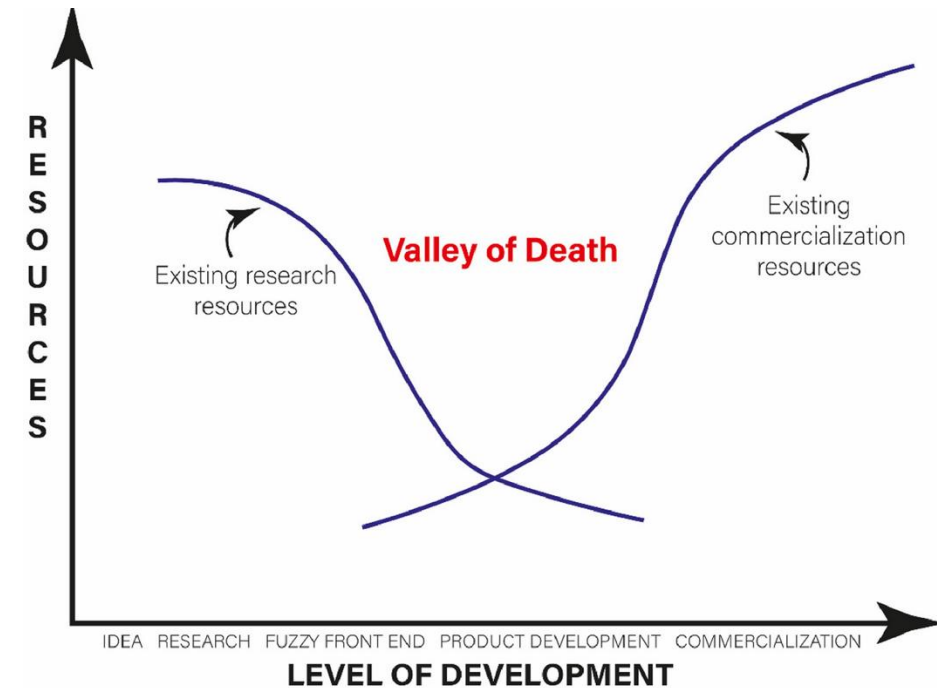
AI/ML and Operations Today



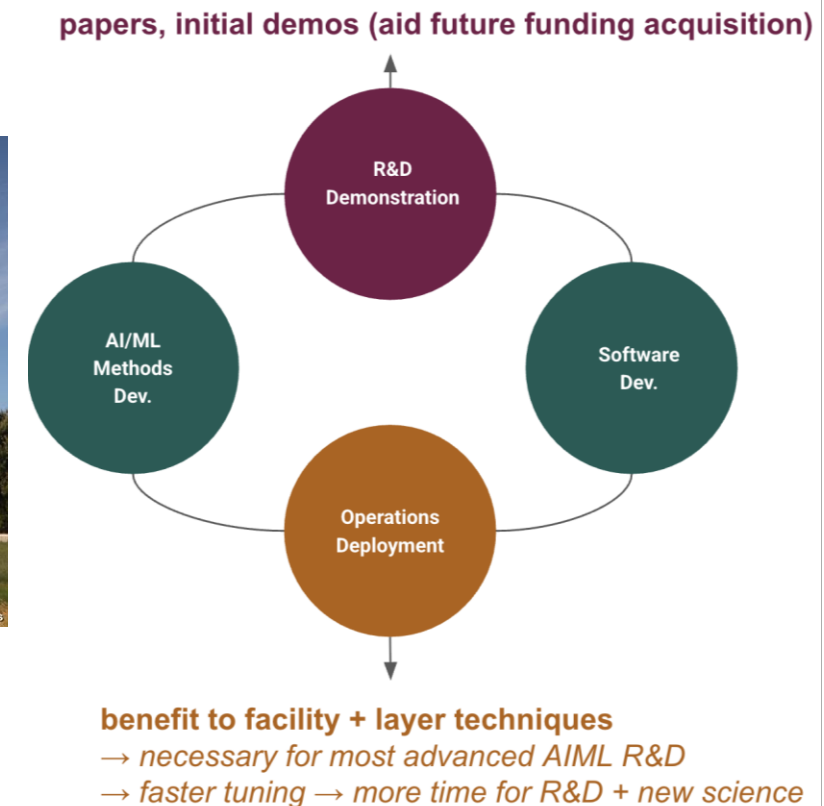
Substantial progress on individual tasks / applications → beginning to get incorporated into operations
Barriers: control system infrastructure, compute, software engineering, scattered documentation / knowledge

AI/ML and Operations Today

Biggest blockers are not AI/ML maturity but rather infrastructure and continuity of effort, coordination, and stable funding to address infrastructure limitations ...



Source: <https://www.sciencedirect.com/science/article/pii/S2666188822000119>

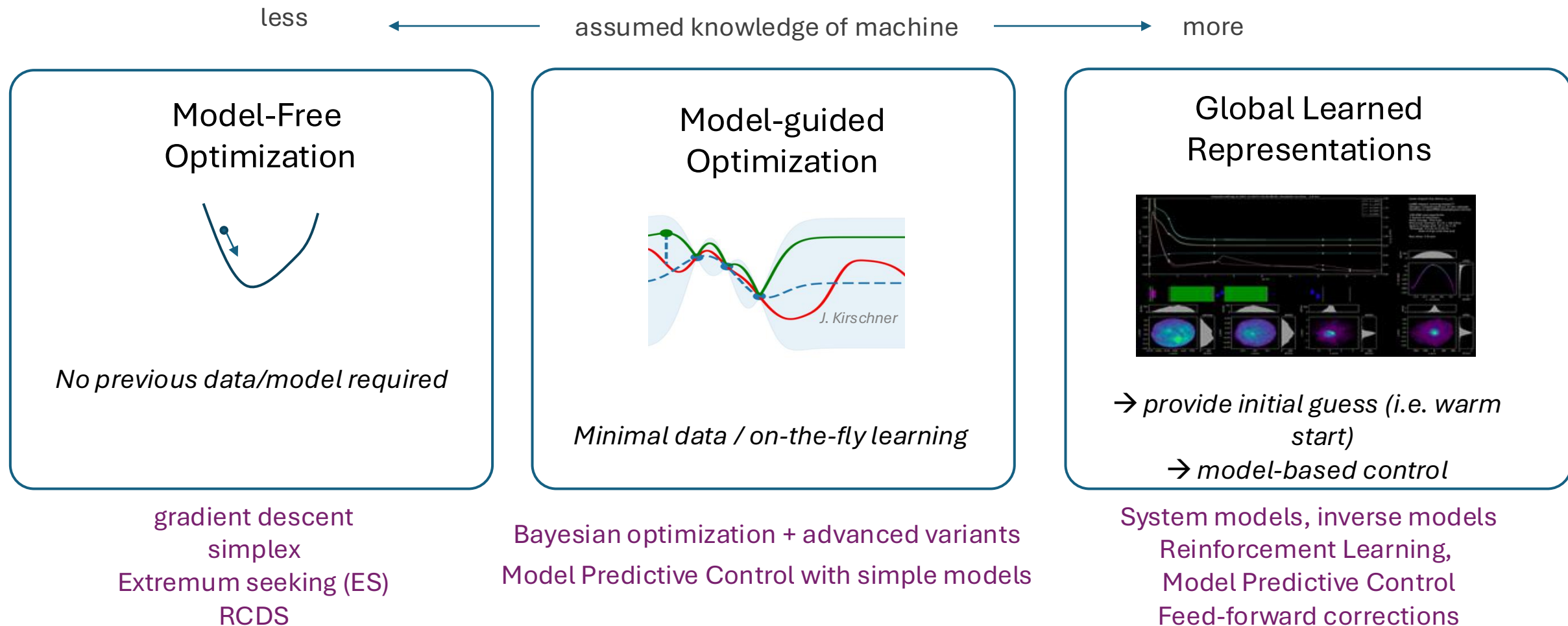


Any future accelerator has the opportunity to head-off these infrastructure challenges during the design process and make integration of advanced computational and AI/ML tools easier

examples of what's possible today ...

Control approaches leverage different amounts of data / previous knowledge

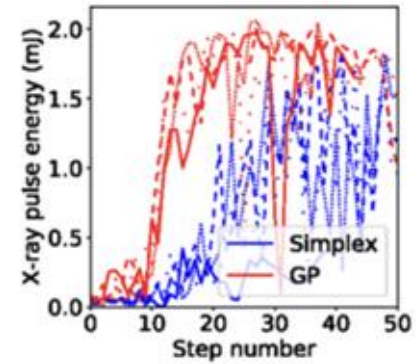
→ suitable under different circumstances and need a combined strategy



General strategy: start with sample-efficient methods that do well on new systems, then build up to more data-intensive and heavily model-informed approaches.

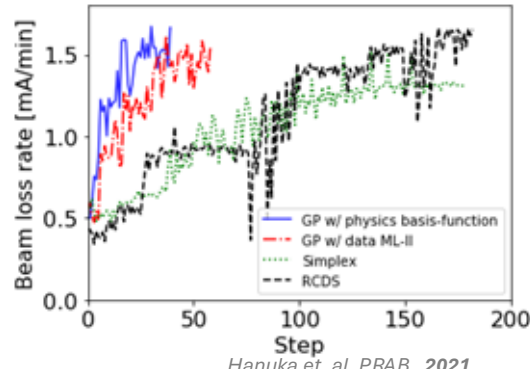
Some ML-based tuning techniques are operations-ready and highly transferrable

FEL pulse energy at LCLS



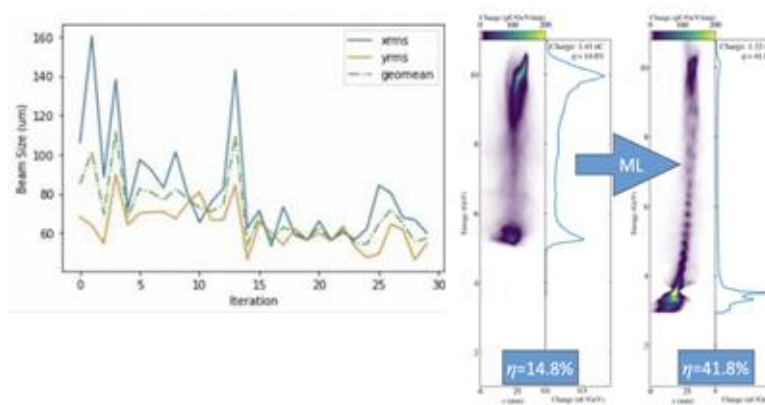
Duris et. al. PRL, 2020

Loss rate at SPEAR3

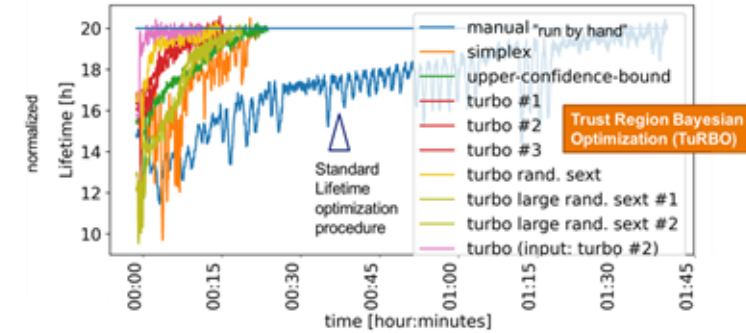


Hanuka et. al. PRAB, 2021

Sextupole tuning at FACET-II >2x efficiency of acceleration in plasma

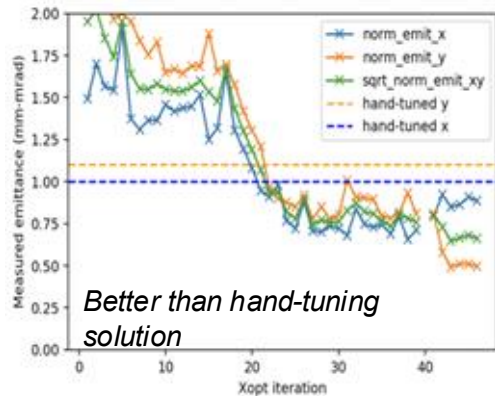


50x faster tuning and best observed lifetime at ESRF



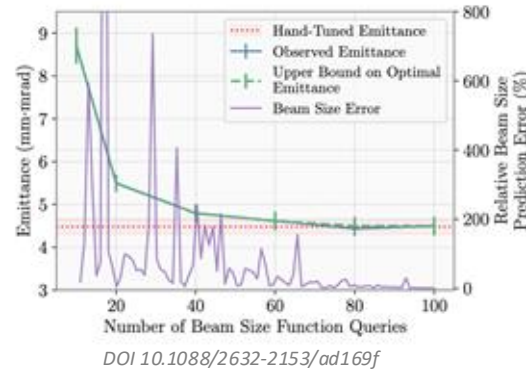
Liuzzo, ICALEPCS 2023 doi:10.18429/JACoW-ICALEPCS2023-MO3AO01

Emittance @ LCLS-II injector



Better than hand-tuning solution

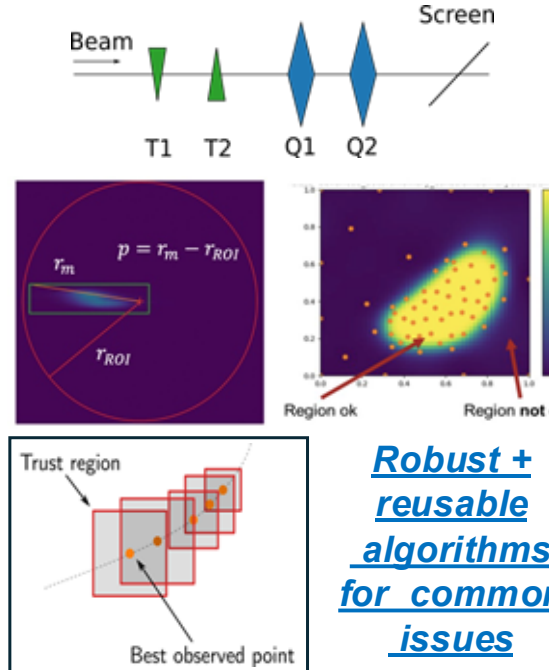
20x faster emittance tuning (BAX)



DOI 10.1088/2632-2153/ad169f

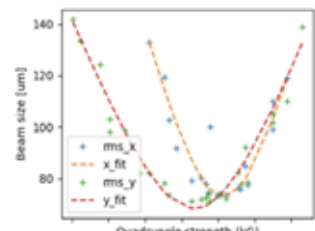
Automated beam alignment at AWA

- 20-30 minutes by hand
- 5 minutes with BAX

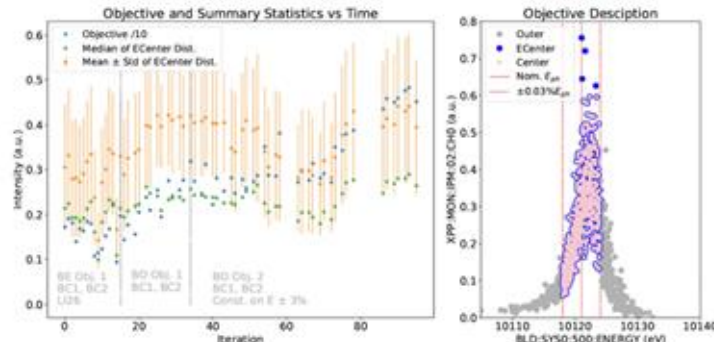


Robust + reusable algorithms for common issues

Automatic emittance measurement with BE



Roussel, Instruments 2023, 7(3), 29



Roussel, Nat Comm 12, 5612 (2021)

PHYSICAL REVIEW ACCELERATORS AND BEAMS 27, 084801 (2024)

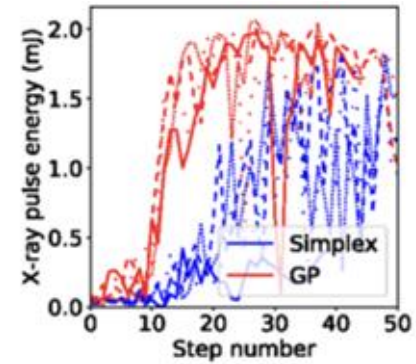
Review Article

Bayesian optimization algorithms for accelerator physics

- Ryan Roussel^{1,*}, Auralee L. Edelen¹, Tobias Boltz², Dylan Kennedy¹, Zhe Zhang¹, Fuhao Ji¹, Xiaobiao Huang³, Daniel Ratner⁴, Andrea Santamaria Garcia⁵, Chenran Xu⁶, Jan Kaiser⁷, Angel Ferran Pousa³, Annika Eichler^{3,4}, Jannis O. Lübsen⁴, Natalie M. Isenberg⁵, Yuan Gao⁵, Nikita Kuklev⁶, Jose Martinez⁶, Brahim Mustapha⁶, Verena Kain⁷, Christopher Mayes⁸, Weijian Lin⁹, Simone Maria Liuzzo¹⁰, Jason St. John¹¹, Matthew J. V. Streeter¹², Remi Lehe¹³, and Willie Neiswanger¹⁴
- ¹SLAC National Laboratory, Menlo Park, California 94025, USA
²Karlsruhe Institute of Technology, Karlsruhe, Germany
³Deutsches Elektronen-Synchrotron DESY, Germany
⁴Hamburg University of Technology, 21073 Hamburg, Germany
⁵Brookhaven National Laboratory, Upton, New York 11973, USA
⁶Argonne National Laboratory, Lemont, Illinois 60439, USA
⁷European Organization for Nuclear Research, Geneva, Switzerland
⁸xLight, Palo Alto, California 94306, USA
⁹Cornell University, Ithaca, New York 14853, USA
¹⁰European Synchrotron Radiation Facility, Grenoble, France
¹¹Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA
¹²Queen's University Belfast, Belfast, Northern Ireland, United Kingdom
¹³Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
¹⁴Stanford University, Stanford, California 94305, USA

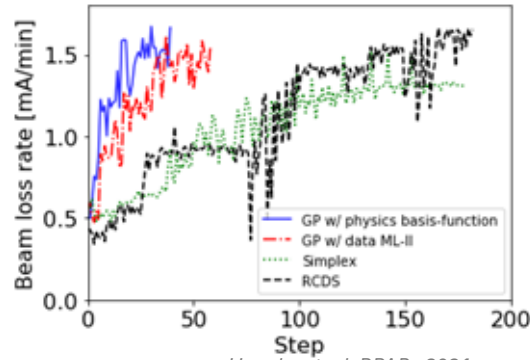
Some ML-based tuning techniques are operations-ready and highly transferrable

FEL pulse energy at LCLS



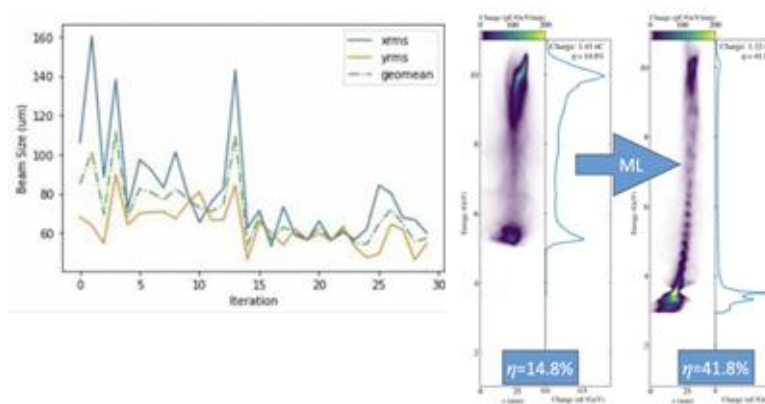
Duris et. al. PRL, 2020

Loss rate at SPEAR3

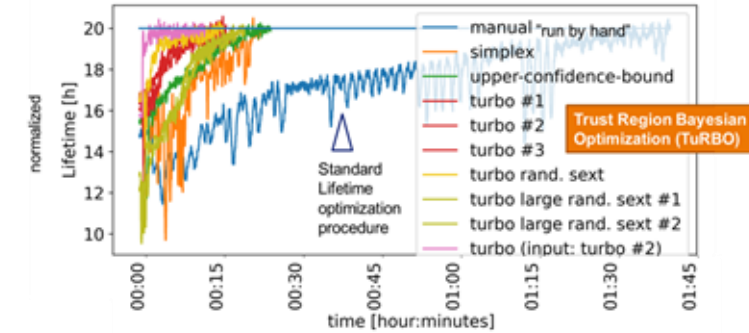


Hanuka et. al. PRAB, 2021

Sextupole tuning at FACET-II >2x efficiency of acceleration in plasma

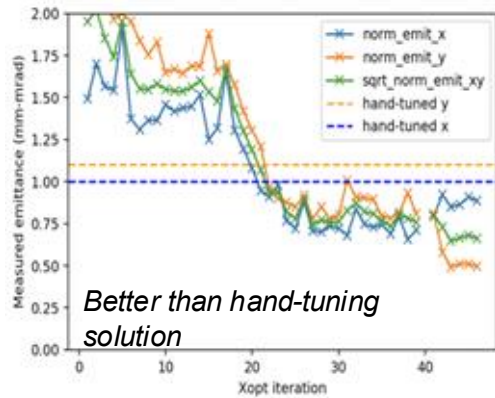


50x faster tuning and best observed lifetime at ESRF



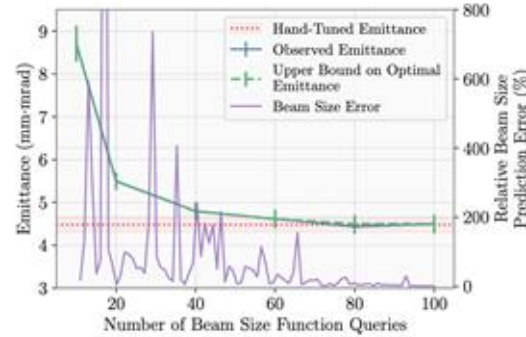
Liuzzo, ICALEPCS 2023 doi:10.18429/JACoW-ICALEPCS2023-MO3AO01

Emittance @ LCLS-II injector



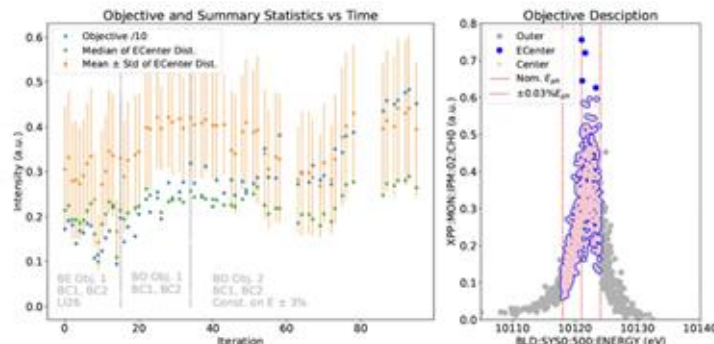
Better than hand-tuning solution

20x faster emittance tuning (BAX)



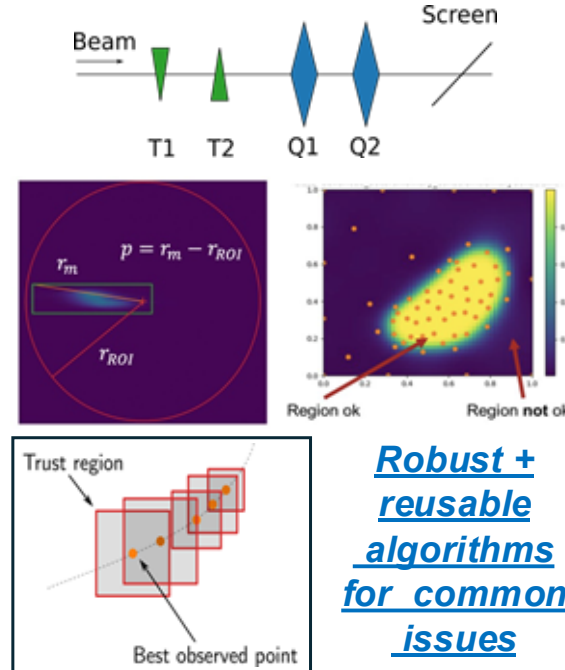
DOI 10.1088/2632-2153/ad169f

Tuning on monochromator signal

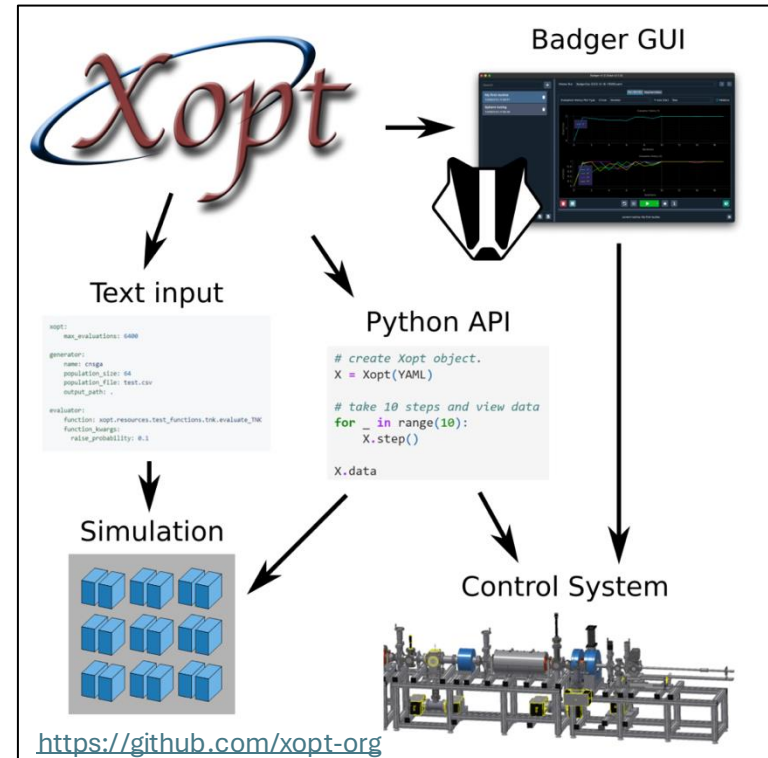


Automated beam alignment at AWA

- 20-30 minutes by hand
- 5 minutes with BAX



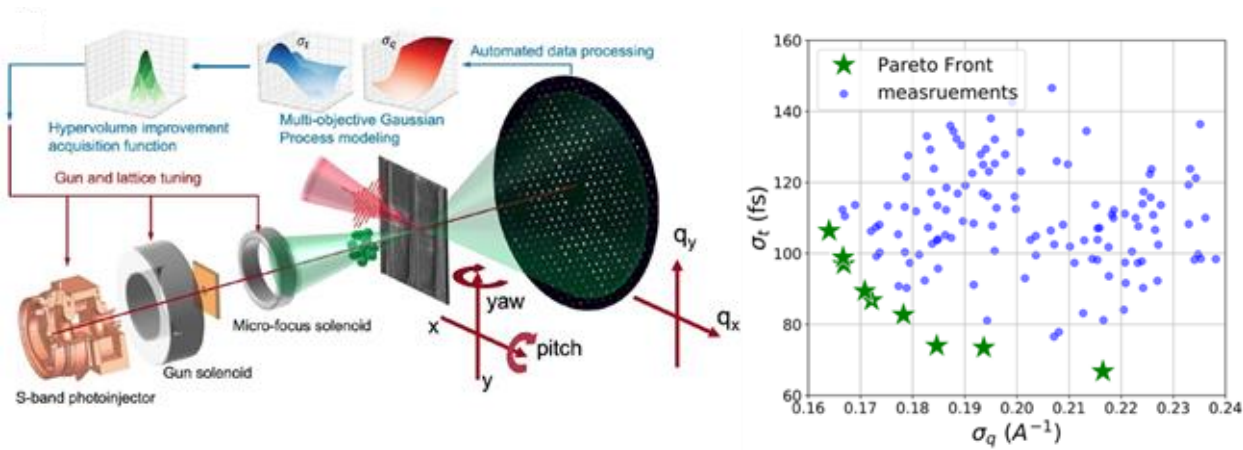
Robust + reusable algorithms for common issues



<https://github.com/xopt-org>

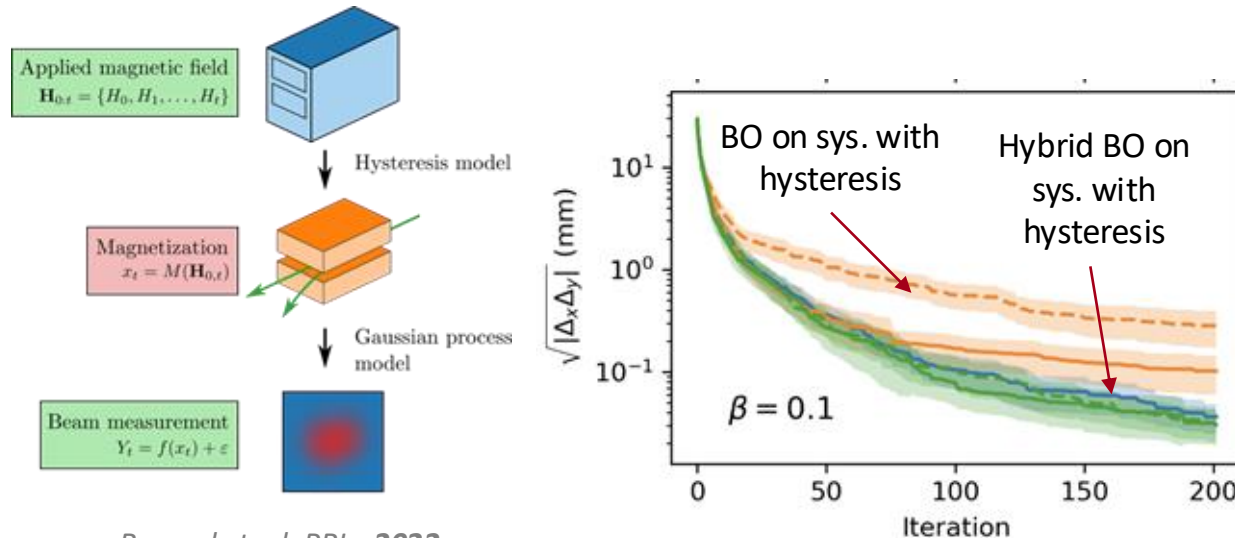
Some ML-based tuning techniques are operations-ready and highly transferrable

Multi-objective BO $\rightarrow$ experimental Pareto front



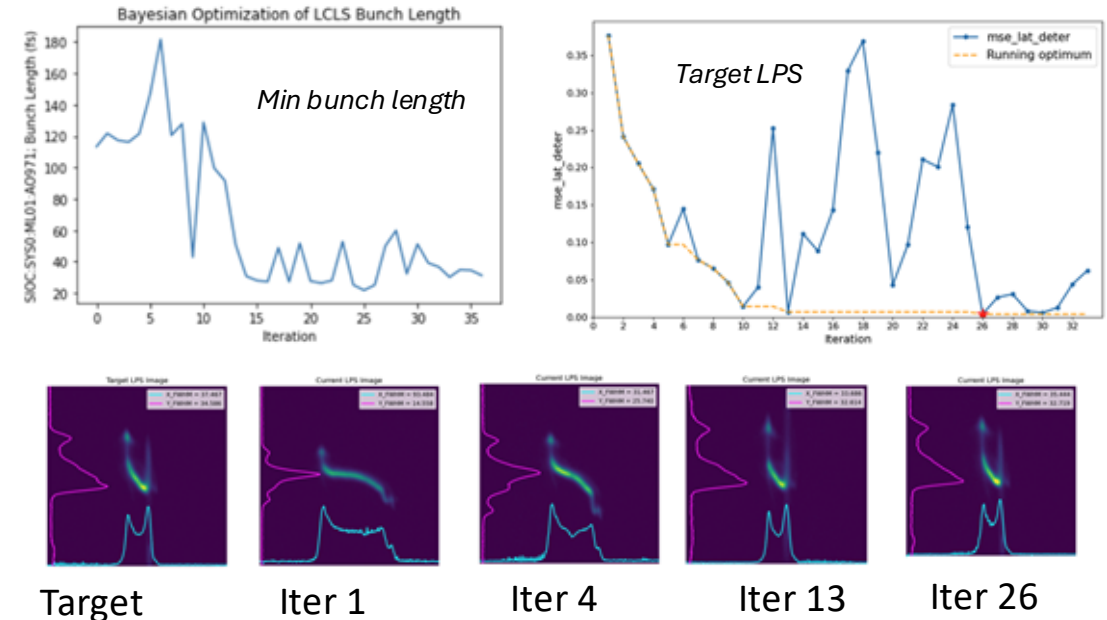
F. Ji, et al., Nat. Commun. 15, 4726 (2024)

Hysteresis-aware magnet tuning

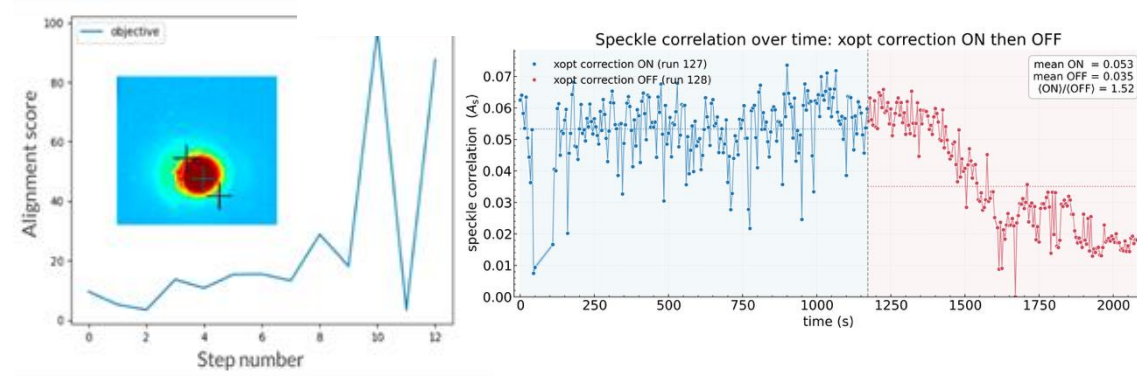


Roussel et. al. PRL, 2022

Longitudinal phase space tuning



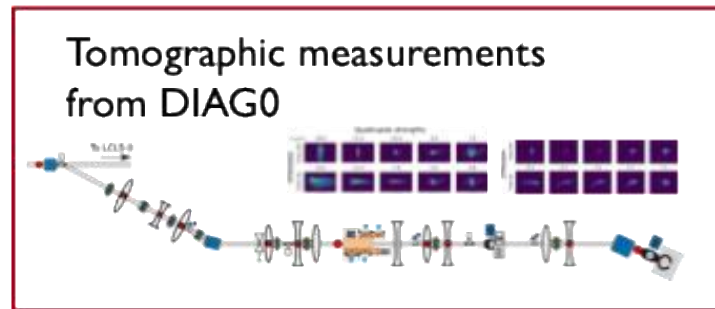
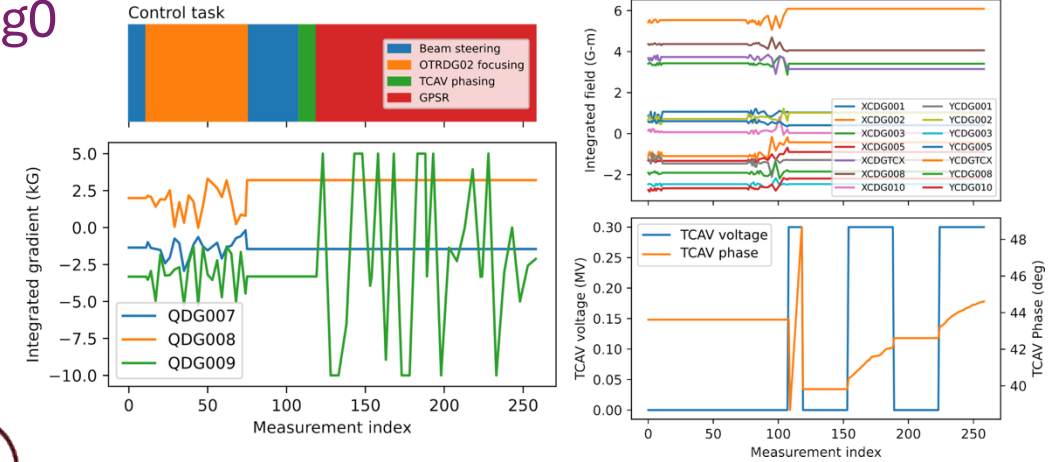
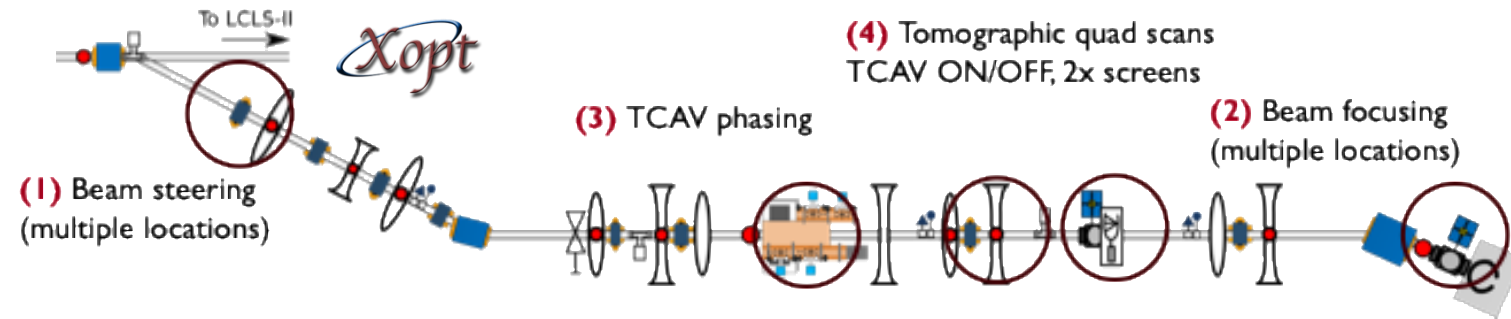
Photon beamline alignment and drift correction



ML + Digital Twin + HPC → unprecedented continuous insight into 6D phase space

Automatic tuning and 6D phase space recon. @ LCLS-II Diag0

→ ~5 minutes online, continuous estimate of phase space

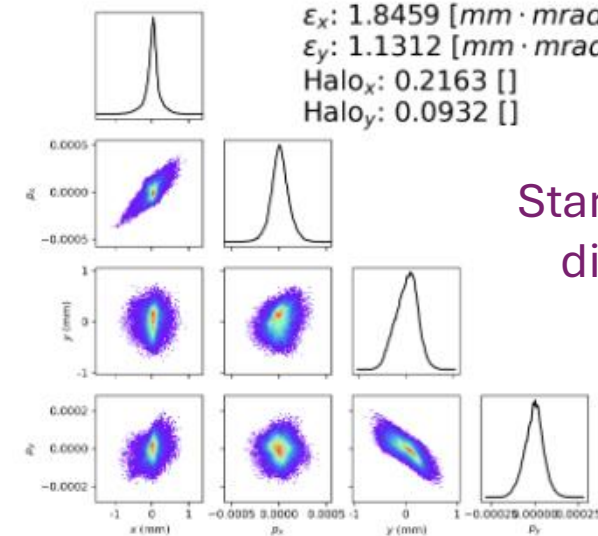
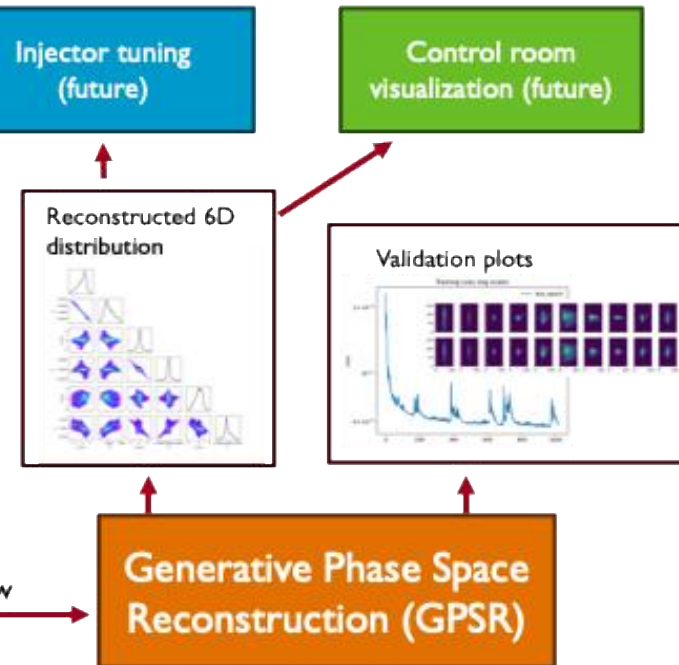


Streaming data transfer

SLAC S3DF SLAC SHARED SCIENCE DATA FACILITY



Triggers analysis workflow



Started testing tuning directly on recon. (core + halo)

Autonomous Injector Startup

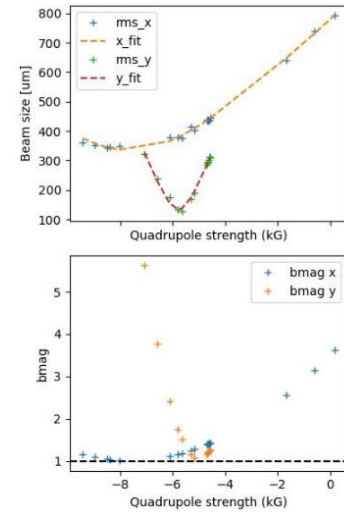


Autonomous startup of FACET-II injector: ~1hr + to 15-20 min without human intervention

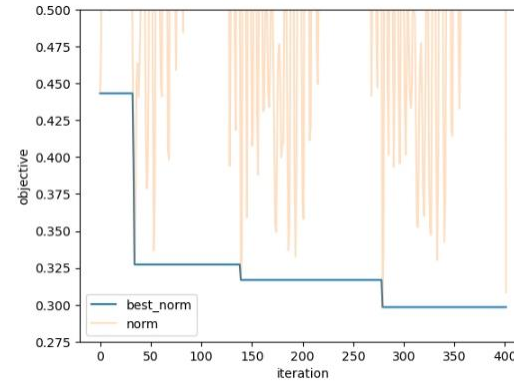
Builds on automation of individual steps in Xopt → combined with Osprey for agent-driven orchestration

Emittance

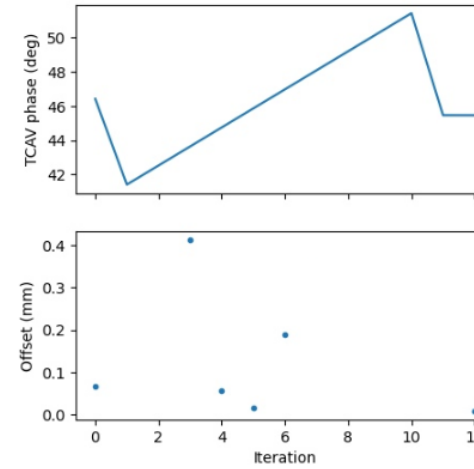
optimization +
matching (7.5 min)



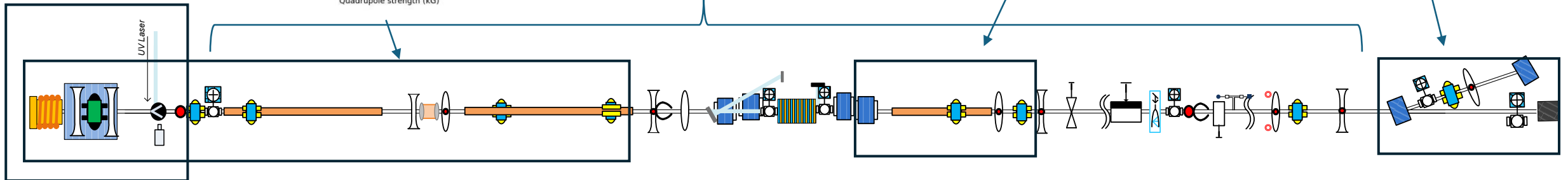
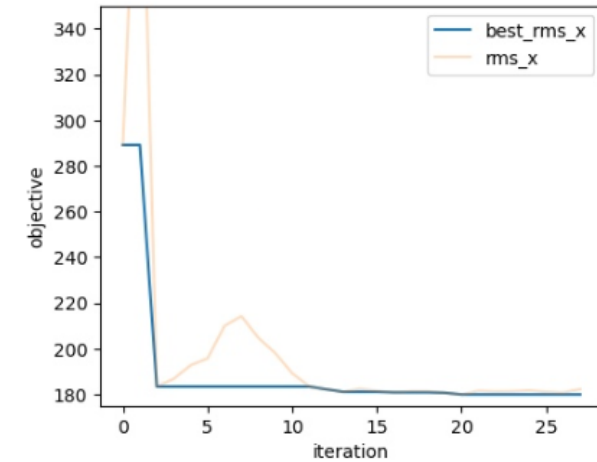
Beam steering (3 min)



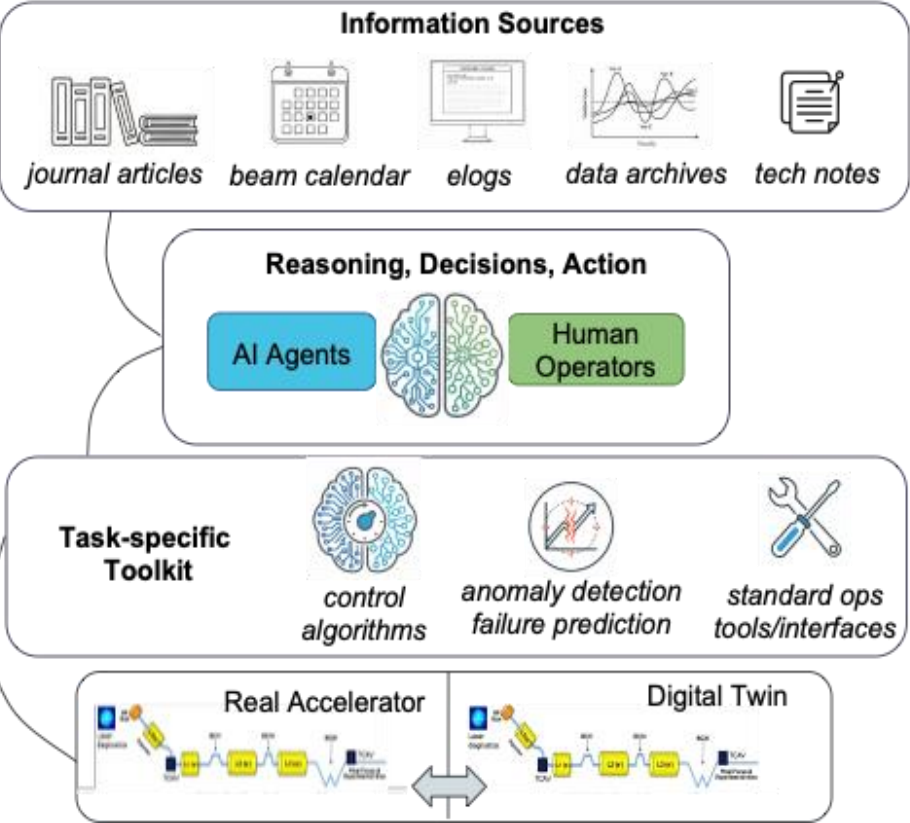
TCAV Phasing (7 sec)



Energy spread
minimization (2 min)



Agent-driven workflows



Osprey demo @LCLS

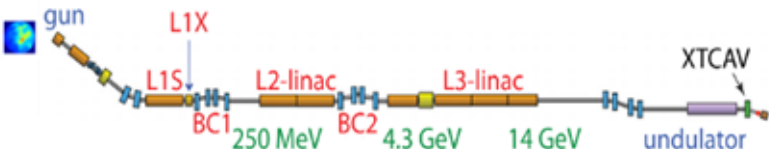


New optimization routine

natural language description of tuning task



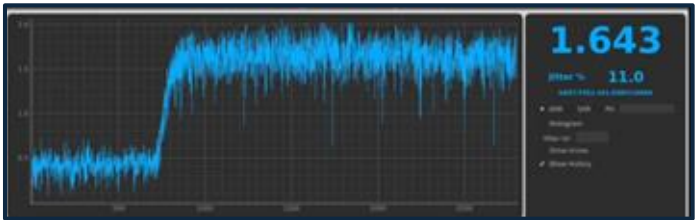
Badger run history



Genesis Mission

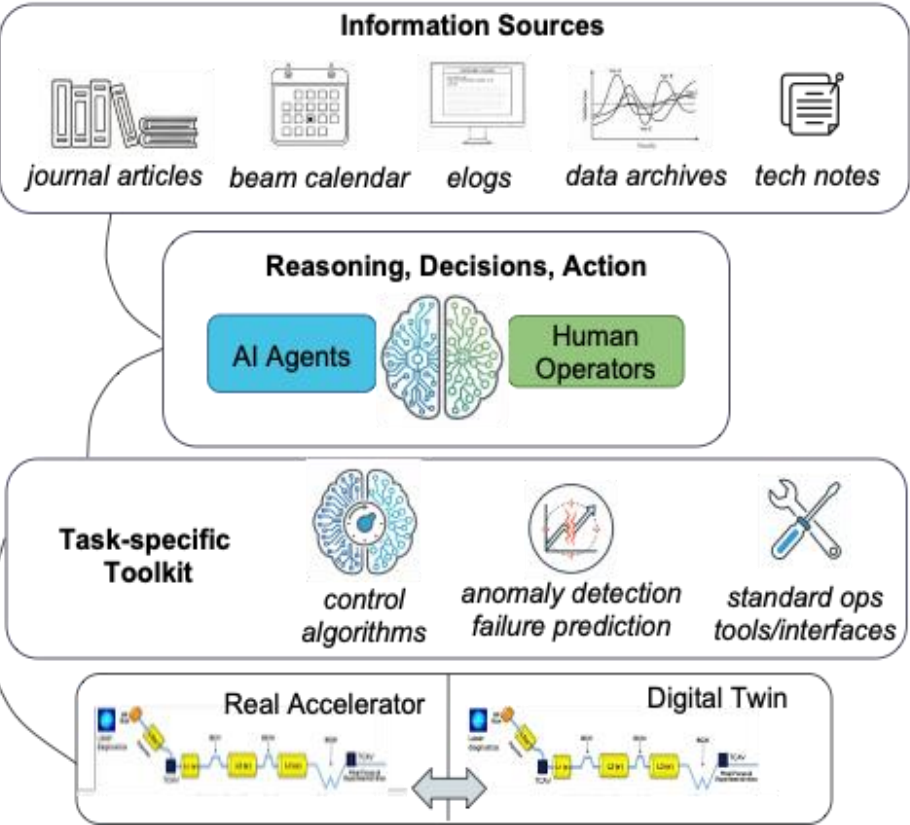


Automatic execution on machine via Badger tuning tool



Rapid improvement in FEL performance
Prompt to completion: ~ 10 minutes

Agent-driven workflows



Similar workflows have been demonstrated for design:

- instructions on task goals
- expert guidance on design workflow
- ability to run physics simulations on HPC, and code manuals

→ Agent autonomously completes task with equivalent quality solution

Osprey demo @LCLS

Per-Knob Bounds (updated)

Knob	Current (live)	Env Soft	HW Limits	Proposed Range (x2.4)
QUAD-LTUN-620-BC1RL	-48.12	(none)	[-112, 112]	[-42.52, -37.72]
QUAD-LTUN-640-BC1RL	47.61	(none)	[-112, 112]	[45.21, 50.01]
QUAD-LTUN-660-BC1RL	-65.54	(none)	[-112, 112]	[-67.94, -63.14]
QUAD-LTUN-680-BC1RL	47.67	(none)	[-112, 112]	[45.27, 50.07]

All within hardware limits ✓ | All current values centered in proposed ranges ✓

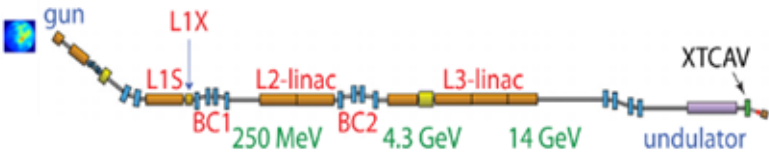
Step 5 - Proposed Routine

Field	Value
Name	KRR FEL tune-up LTUN matching quads v2
Env	1x1x (cu_har)
Knobs	4 LTUN matching quads (x2.4 around current)
Objective	pulse_intensity_pho + MAXIMIZE
Algorithm	Expected Improvement (128 MC samples, LBFGS, 20 restarts)
Initial points	1 current + 3 random (30% fill)
Based on	Apr 15 best run + updated live machine state

User answered Claude's questions:
- Ready to save this updated routine? The ranges are re-centered on the current machine state (which has shifted from the previous optimization). → Save as proposed
- Calling badger-archive... (ctrl+e to expand)

New optimization routine

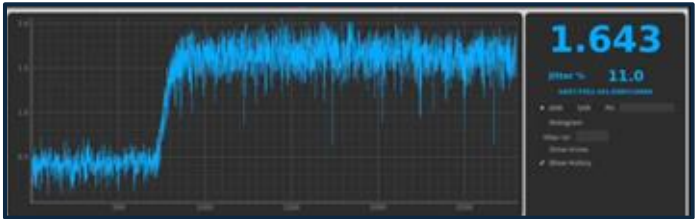
natural language description of tuning task



Badger run history



Automatic execution on machine via Badger tuning tool



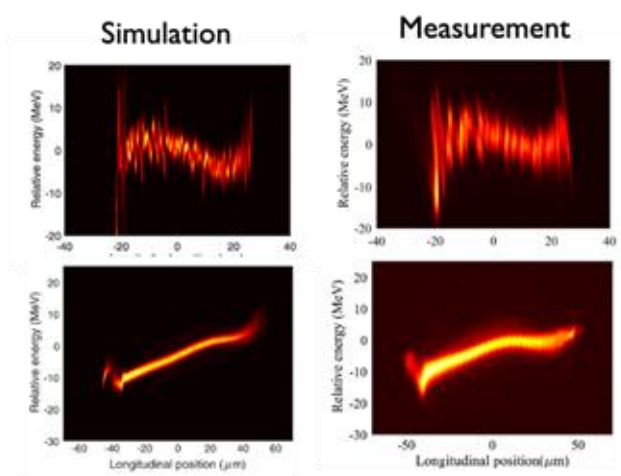
Rapid improvement in FEL performance
Prompt to completion: ~ 10 minutes



“Please go to two bunch mode and do a complete scan of the beam timing and energy separation across the valid operating range”

Unprecedented speed, accuracy, and scope in system modeling is becoming routine

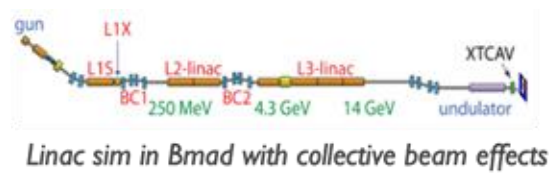
Accelerator simulations that include nonlinear and collective effects are powerful tools, but they can be computationally expensive



10 hours on thousands of cores at NERSC!

J. Qiang, et al., PRSTAB30, 054402, 2017

ML models provide fast approximations to simulations (“surrogate models”)



Linac sim in Bmad with collective beam effects

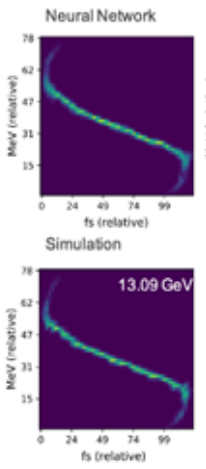
Scan of 6 settings in simulation

Variable	Min	Max	Nominal	Unit
L1 Phase	-40	-20	-25.1	deg
L2 Phase	-50	0	-41.4	deg
L3 Phase	-10	10	0	deg
L1 Voltage	50	110	100	percent
L2 Voltage	50	110	100	percent
L3 Voltage	50	110	100	percent

< ms execution speed

10⁶ times speedup

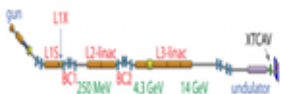
[Edelen et al., NeurIPS 2019](#)



Bringing simulation tools from HPC systems to online/local compute



Online prediction
Model-based control



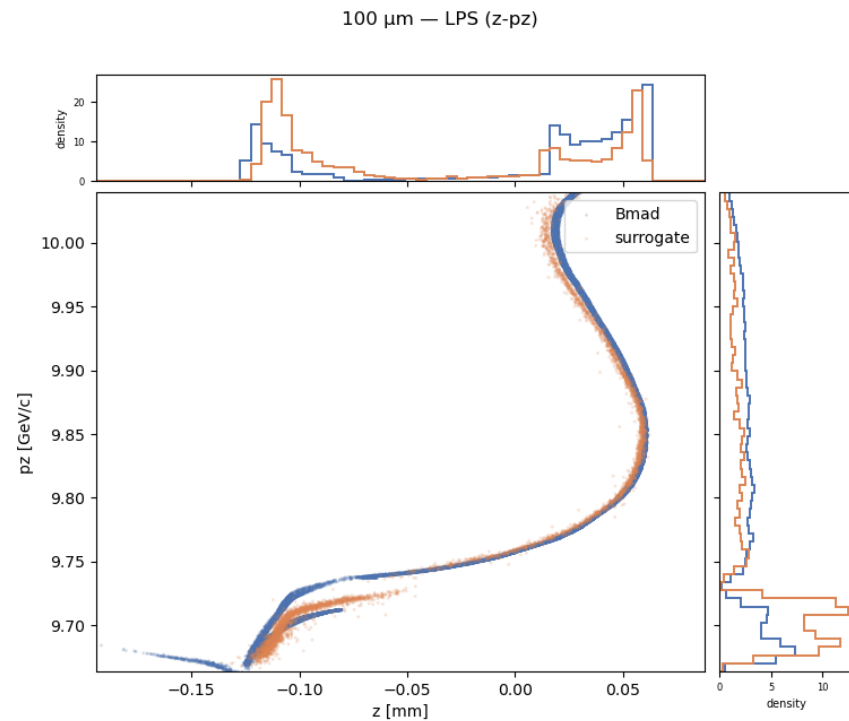
Control prototyping
Experiment planning

Accelerator community is making steady rapid progress on hybrid ML / physics simulations and differentiable simulations
→ ready to be used in design workflows today + already used in controls development

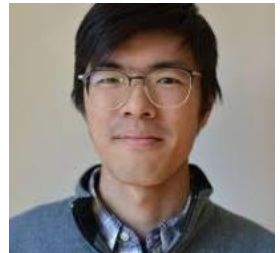
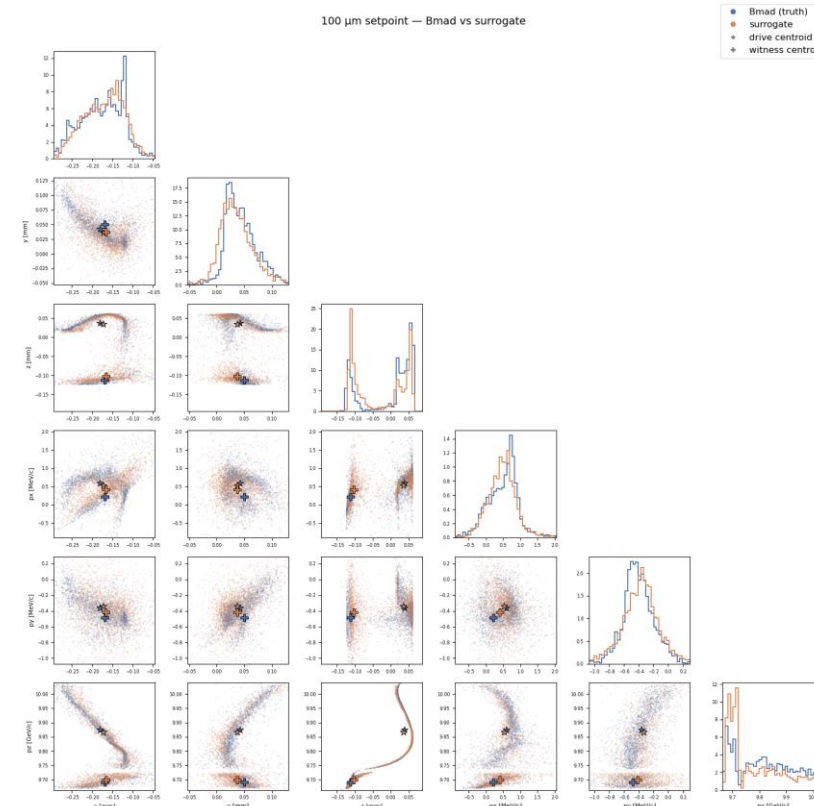
Unprecedented speed, accuracy, and scope in system modeling is becoming routine

Example:

FACET-II start-to-end simulations → surrogate to quickly predict 6D phase space at IP → train RL controller for phase space control → address sim2real and robustness prior to experimental tests



RL controller slewing two bunch separation



Ryan Wu

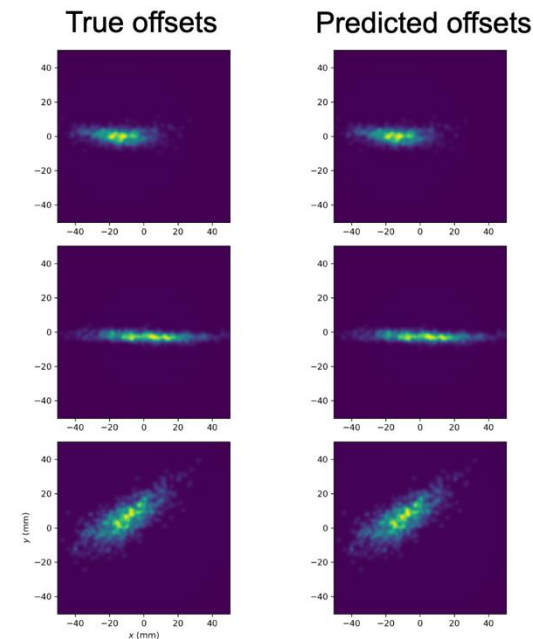
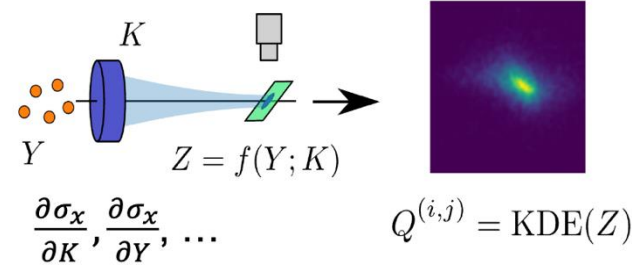
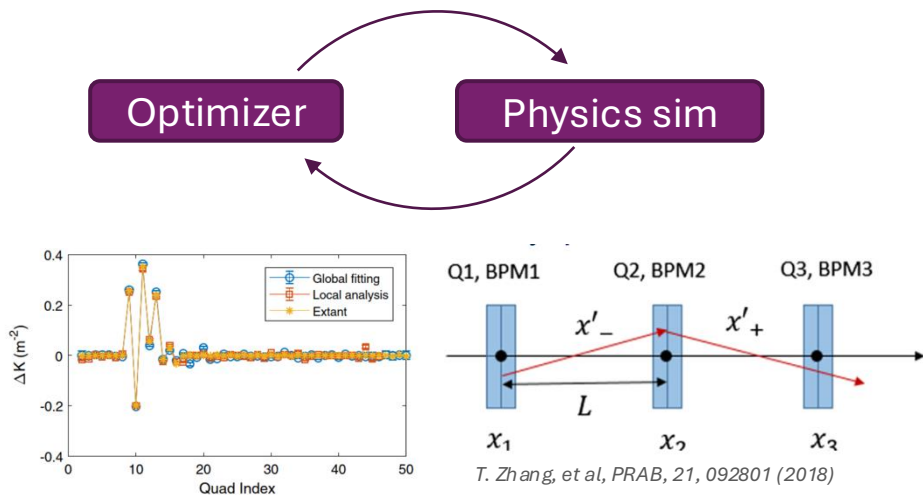
With S2E simulations in hand, entire process (surrogate training to RL controller development) took a few weeks of work from a graduate student interfacing with SLAC team

Accelerator community is making steady rapid progress on hybrid ML / physics simulations and differentiable simulations
→ ready to be used in design workflows today + already used in controls development

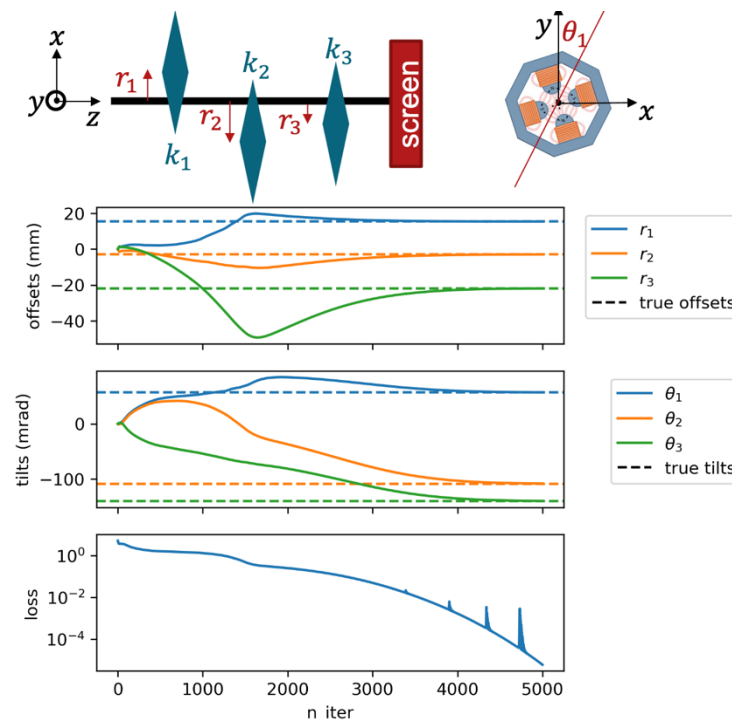
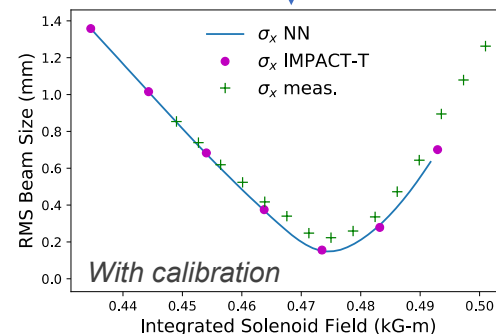
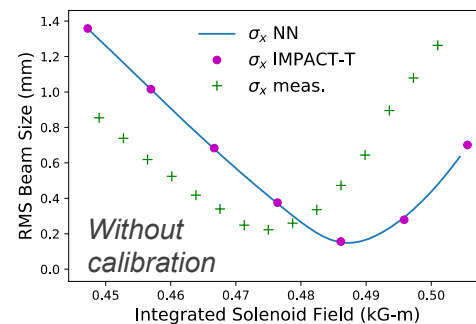
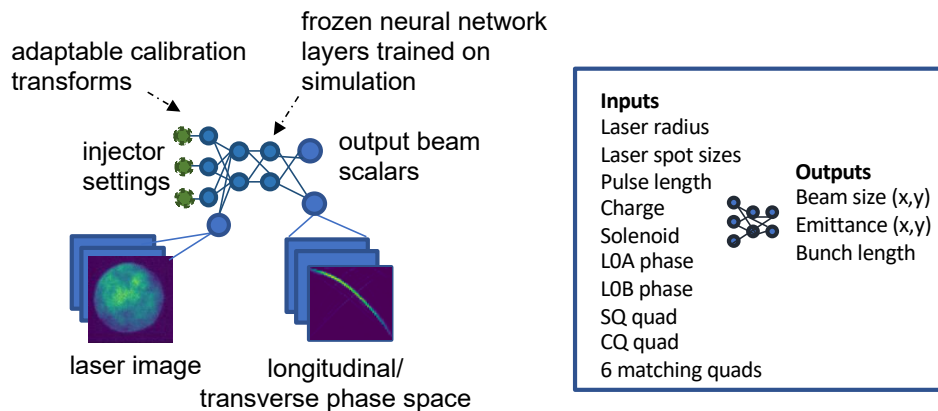
Sim2Real / model calibration methods are maturing

Classical Analytic and Brute Force Methods

Differentiable Physics



ML model trained on simulation
→ transfer learning to measured data



Looking forward: future machines and design

(AI-assisted design and designing with AI/ML in operations in mind)

Common Issues Going from Design to Operation

- **Complex and elegant design works beautifully in simulation**

→ very difficult to operate in reality depending on tools available

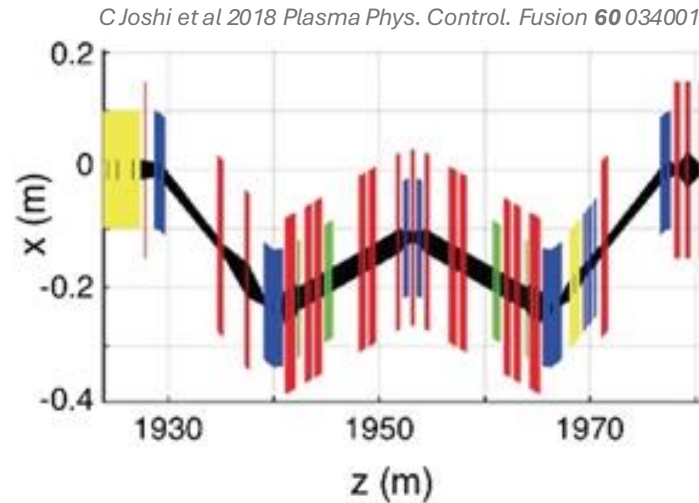
- **Sub-systems don't interact as expected / idealized once linked together** (often without prior detailed system-level sims)

→ impacts performance in unexpected ways that need to be fixed afterward

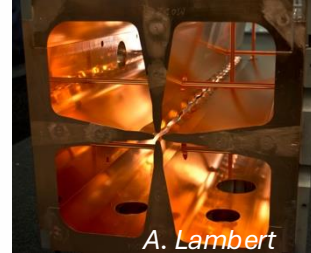
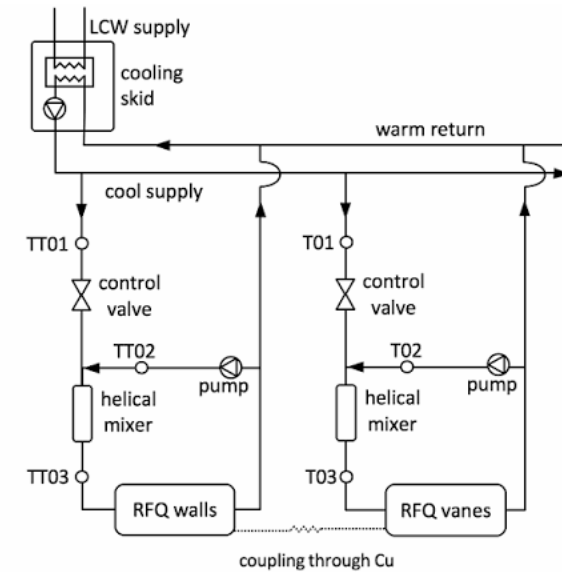
- **Controls / diagnostics treated as an afterthought**

→ kick the can down the road to commissioning or consider cutting

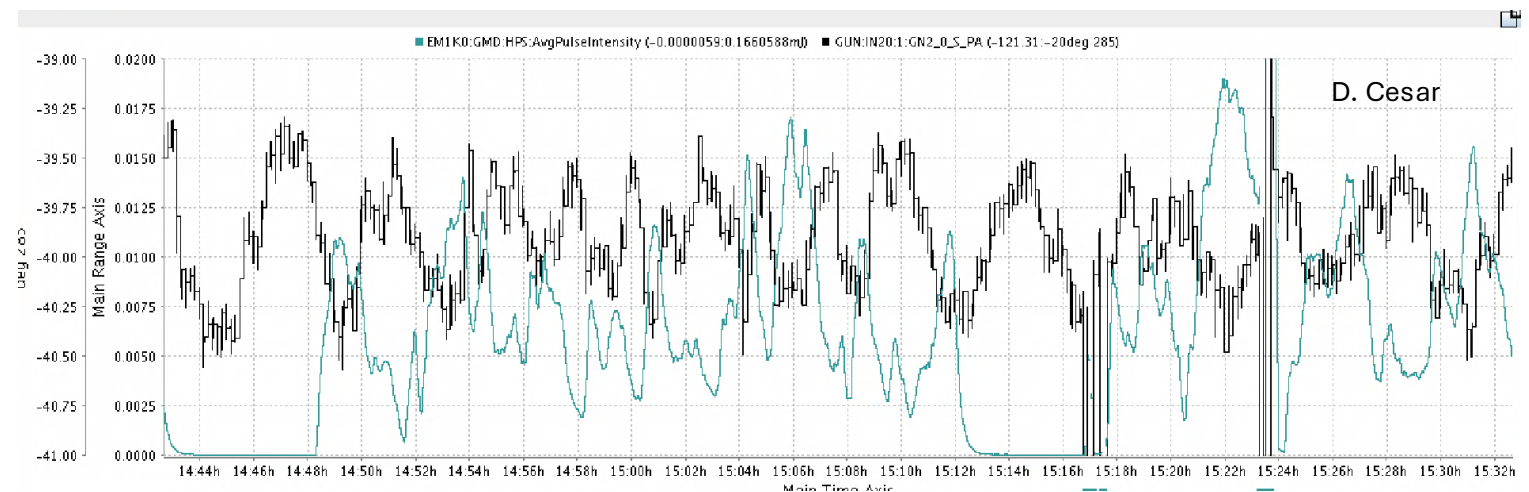
(do we really need <xyz>?)



*FACET-II W-chicane
(eventual tuning solution with ML!)*



PIP-II injector test RFQ



LLRF system oscillations and FEL pulse intensity (correlation, unknown cause)

Road Ahead: piecemeal design → holistic design

Historical approach: each subsystem designed/optimized separately

Compute, fast ML surrogate models, and ease of sharing information/tools across groups makes more holistic design possible:

- Start-to-end
(particle beam source to experiment)
 - *totally feasible today*
 - *limiter is people's time for coordination + organization structures / incentives*
- Top-to-bottom
(high-level beam dynamics to low level facility systems – cooling, cavity design, LLRF, etc)
 - *will require development / new ways to link simulations*

Enables performance improvements (system-wide optimization) and reduces risk

Good community convergence on shared modular/interoperable simulation tools and standard interfaces + formats

See J.L. Vay's talk, SciDAC5 virtual accelerators, BLAST codes, PALS / OpenPMD, LUME, MOAT, etc)

Reviews in Physics 10 (2023) 100085

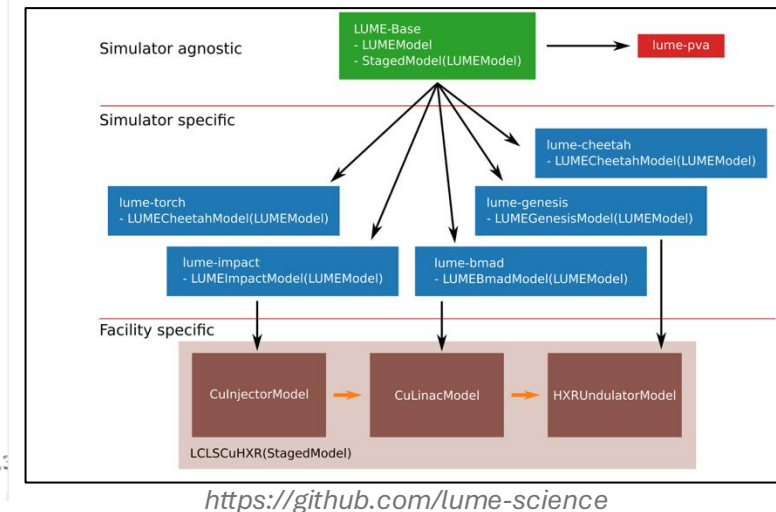
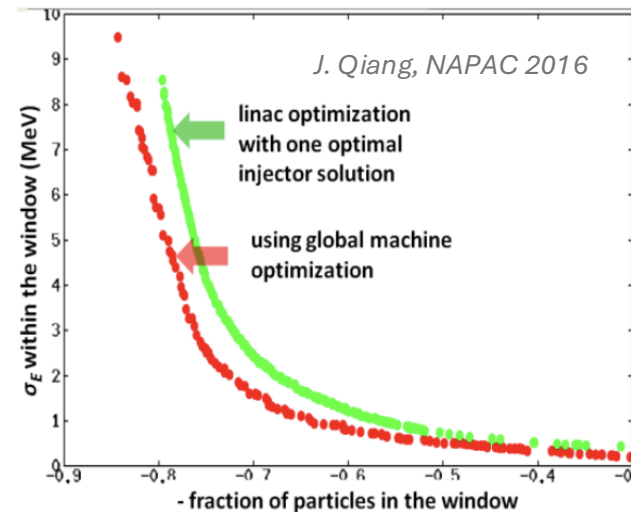
Contents lists available at ScienceDirect

Reviews in Physics

journal homepage: www.elsevier.com/locate/revip

Toward the end-to-end optimization of particle physics instruments with differentiable programming

Check for updates



Road Ahead: Co-design of lattice, controls and diagnostics

Historical approach:

Controls and diagnostics are 2nd class citizens, tacked on top of beam dynamics-driven design

Road ahead:

Control algorithms, diagnostics, and lattice / beam dynamics are optimized together as coupled systems (co-design)

→ Assess performance based on how the machine will actually operate in practice!

→ Requires comprehensive S2E modeling, HPC integration, efficient modeling and search methods (AI/ML)

Many opportunities to test effectiveness of new design / co-design processes with existing facilities

→ e.g. co-design of novel controls + diagnostics + sim2real schemes: test at FACET-II experimentally

“Operations” simulation trials and joint optimization with beam physics + component design

Control algorithms

Diagnostics placement / type

Machine model calibration methods

Lattice components placement/type (quads, sextupoles)

software

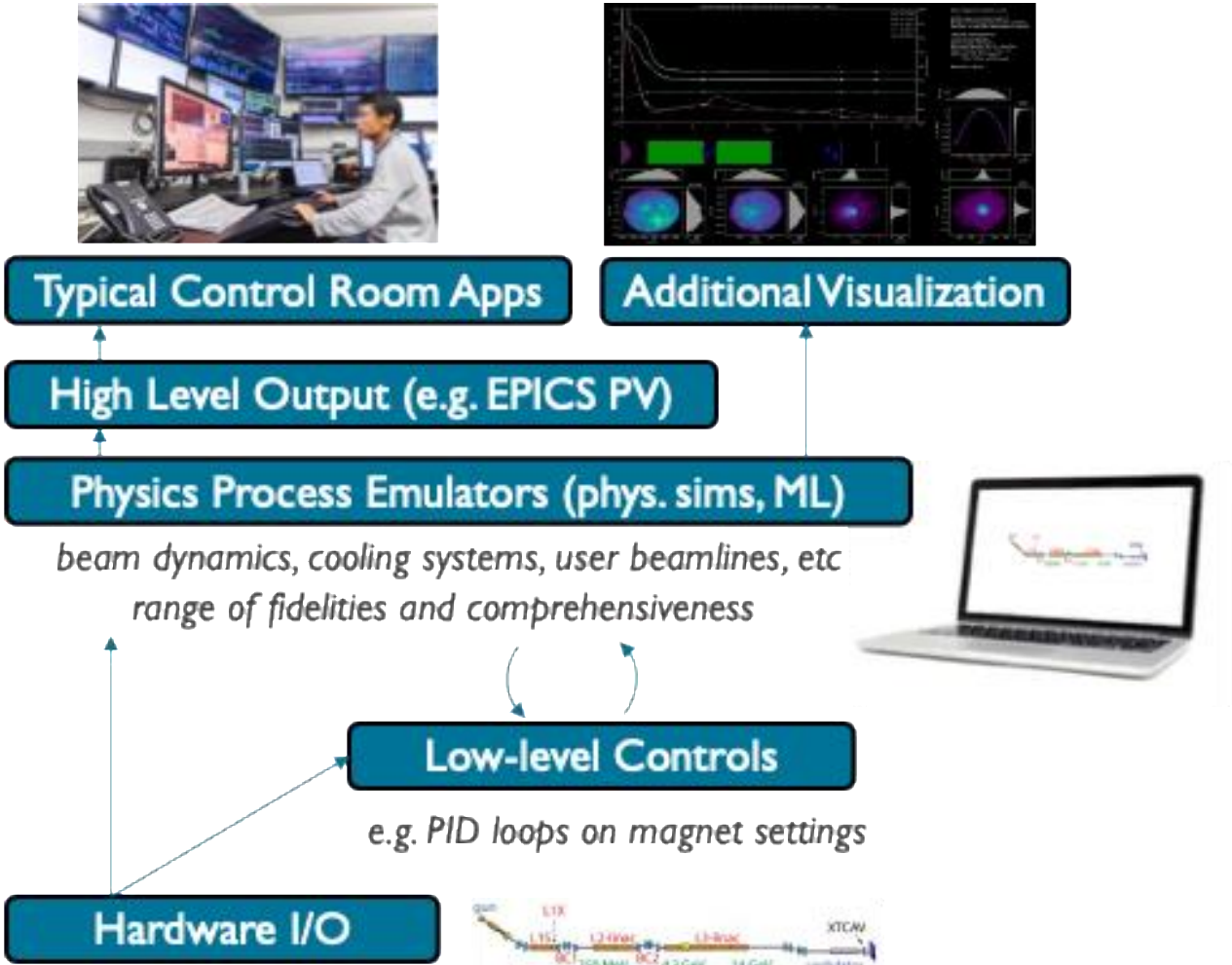
hardware/firmware

*Simplified view: ideally also includes controls + firmware
e.g. LLRF control on FPGA, detector analysis in edge devices*



Genesis is building out infrastructure and tools to make this type of application feasible at large scale (ModCon/AmsC)

Road Ahead: Virtual Accelerators → Digital Twins



Virtual accelerator – comprehensive offline model
Digital twin – adaptive online model used in decision support/control

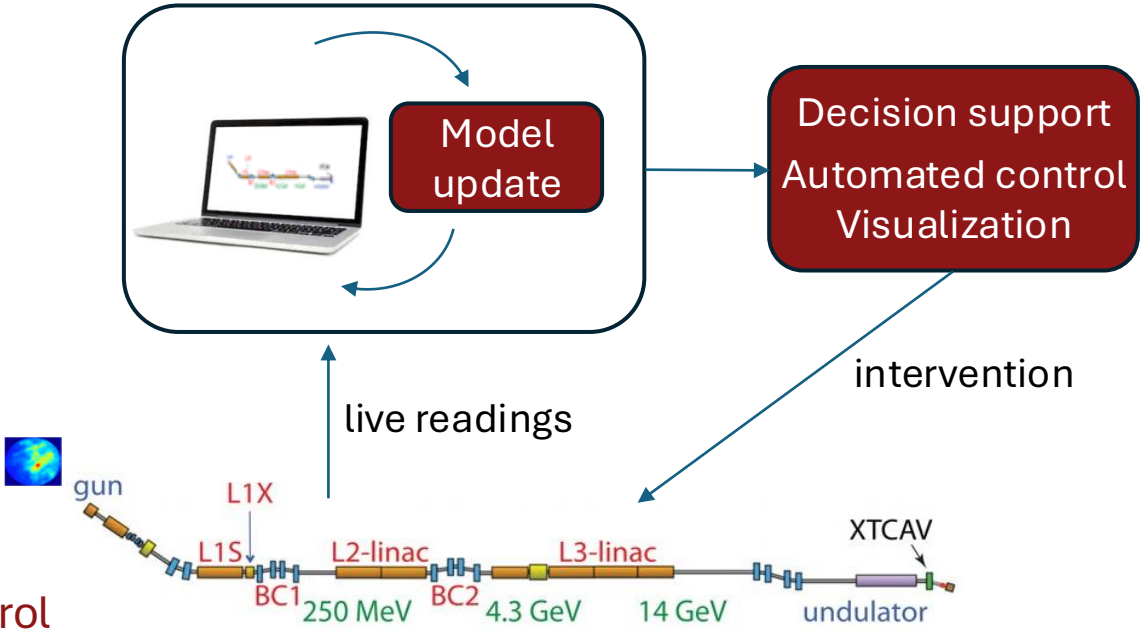
NATIONAL ACADEMIES

Sciences
Engineering
Medicine

NATIONAL ACADEMIES PRESS
Washington, DC

Foundational Research
Gaps and Future Directions
for Digital Twins

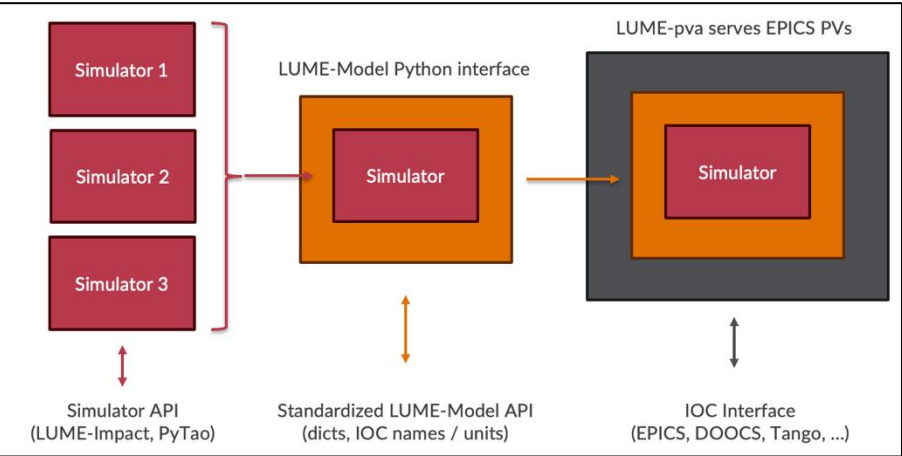
<https://nap.nationalacademies.org/catalog/26894/foundational-research-gaps-and-future-directions-for-digital-twins>



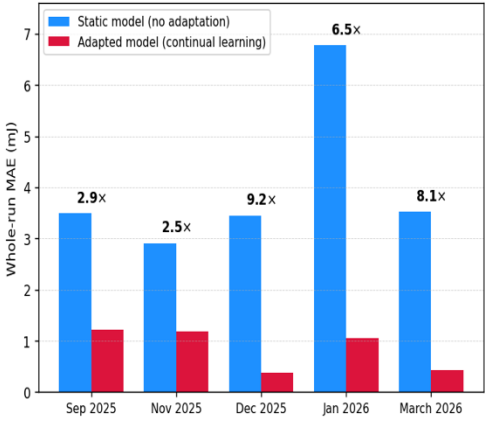
Uses in operation: model-based tuning (faster, more precise), virtual diagnostics, algorithm/software testing
Uses in design: full start-to-end co-design, high level software stack development, realistic operations testing

DT/VA Examples from LCLS/LCLS-II

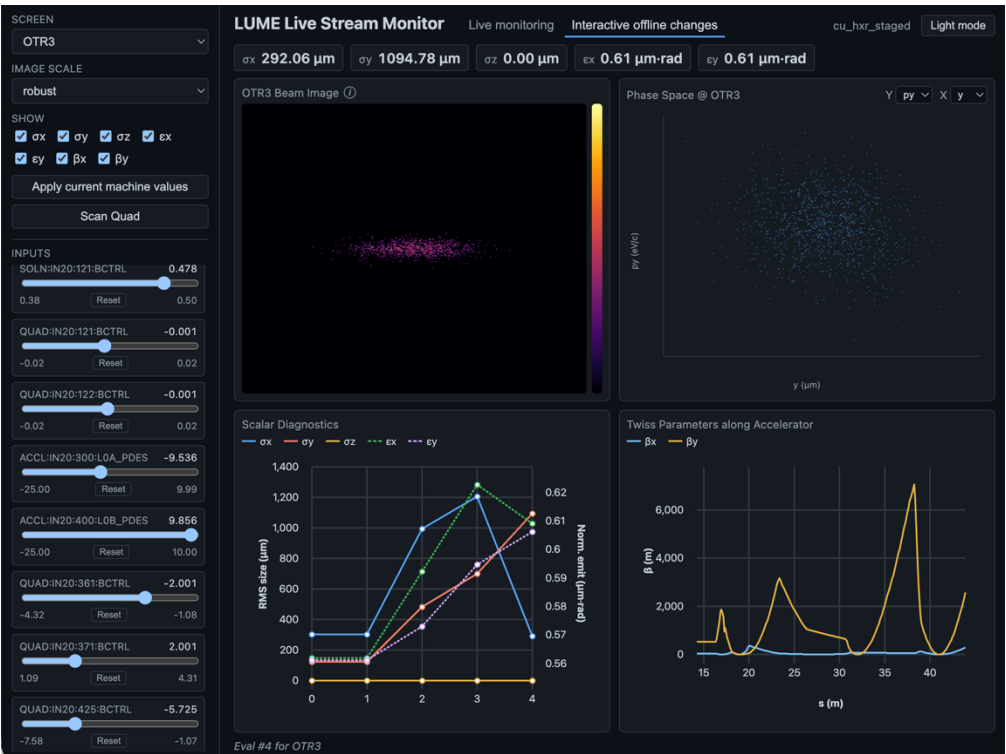
Modular software suite for DT/VA hosted on SLAC cluster (S3DF)



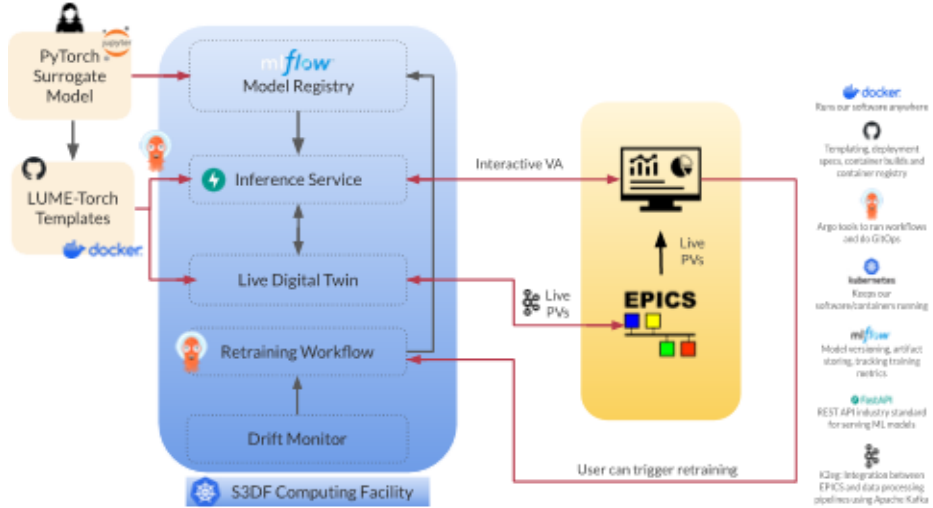
Modular/interoperable approach:
easily swap simulators + ML models



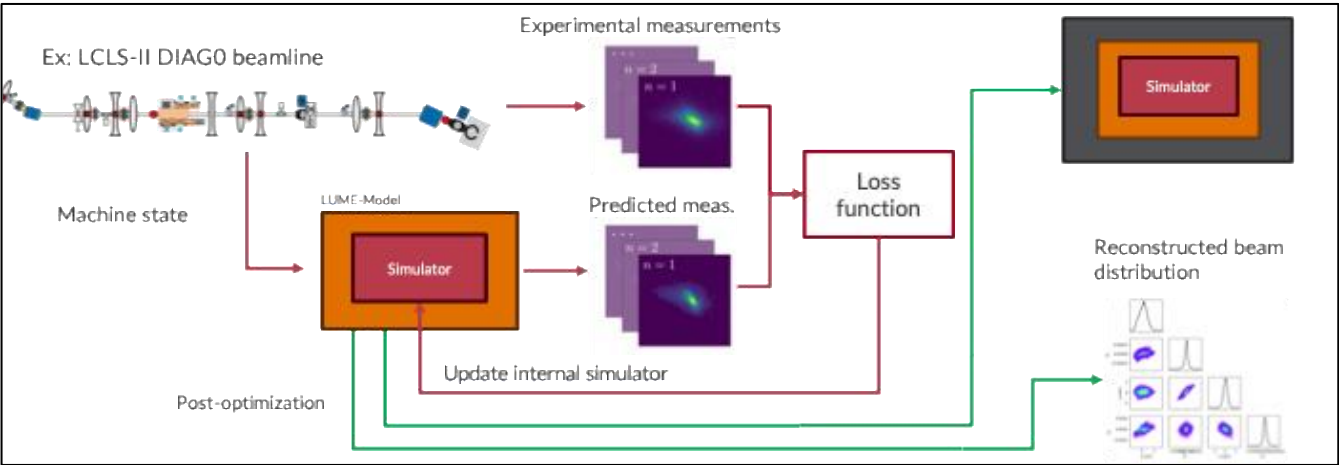
Continual learning yields
improved predictions over time



Continuously updating LUME-Model from live snapshots

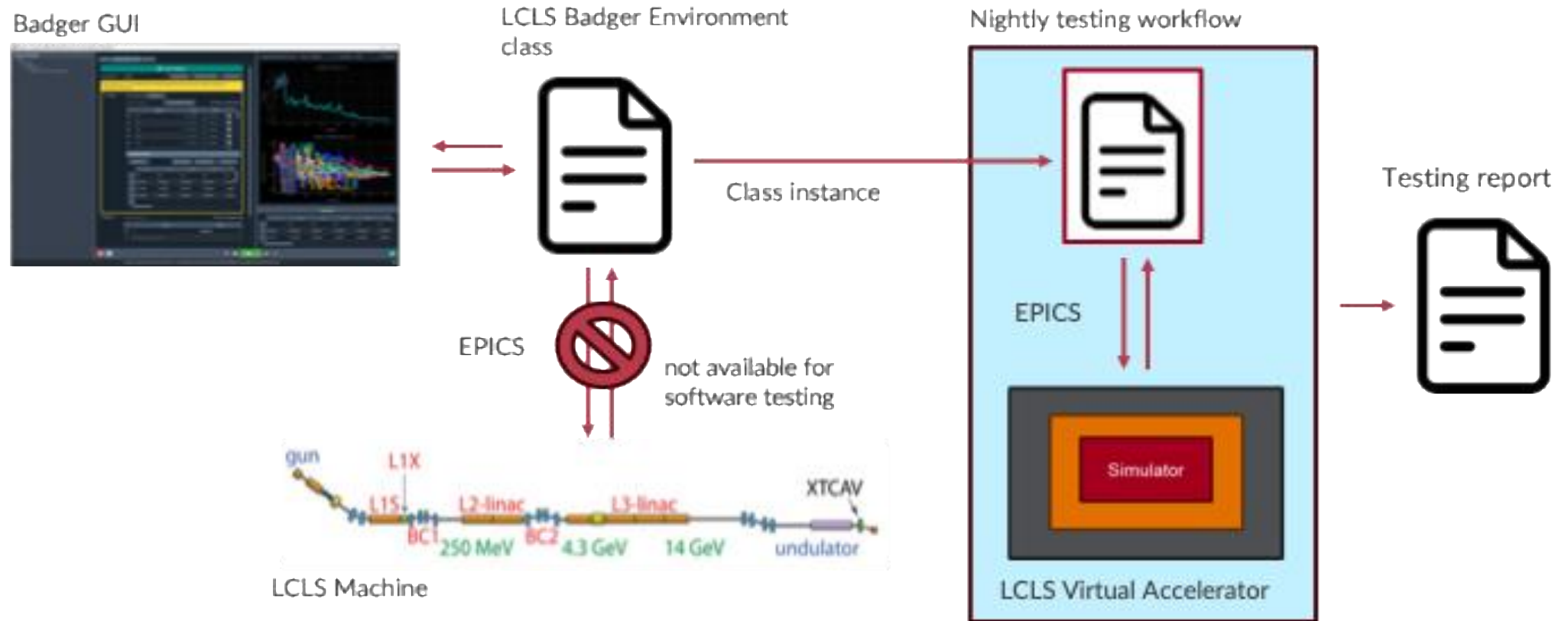


Templating and modular workflow for
ML model deployment / retraining



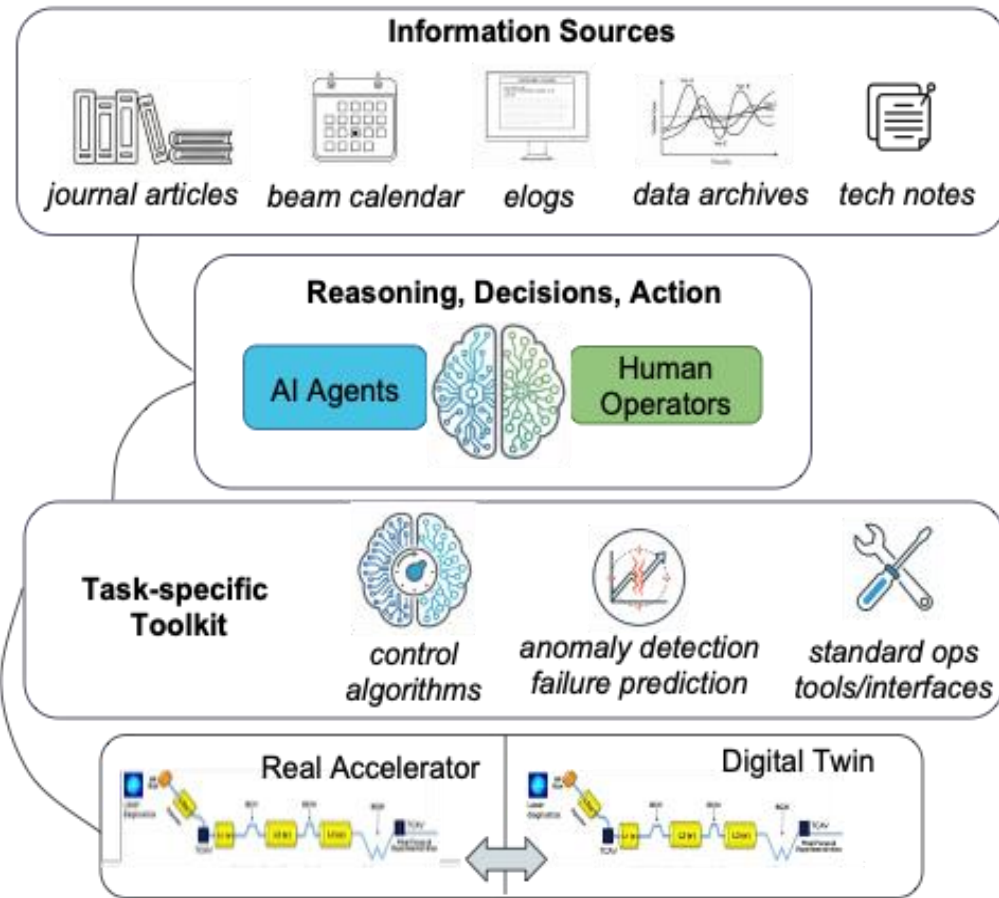
Differentiable DT/VA for model calibration and phase space reconstruction

DT/VA Examples from LCLS/LCLS-II



Virtual accelerator critical for algorithm/software development and testing

Road Ahead: Agentic AI in Operations and Design



Operation:

- Expect that AI agents will become essential tools in operation
→ *design the infrastructure / control system for both human and agent useability*
- Biggest barriers to use are mundane: *information systems, documentation, performant access (e.g. slow data retrieval), HPC integration with controls*

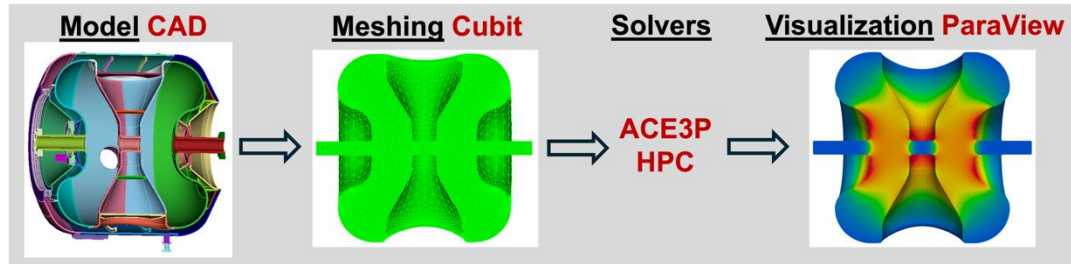
Design:

- Agents will be able to run complex workflows and leverage large-scale compute to iterate more quickly and more broadly on design than we can today
 - *Faster design process + higher performance designs*
 - *All the challenging co-design and holistic system design workflows in previous slides are ripe for agents*

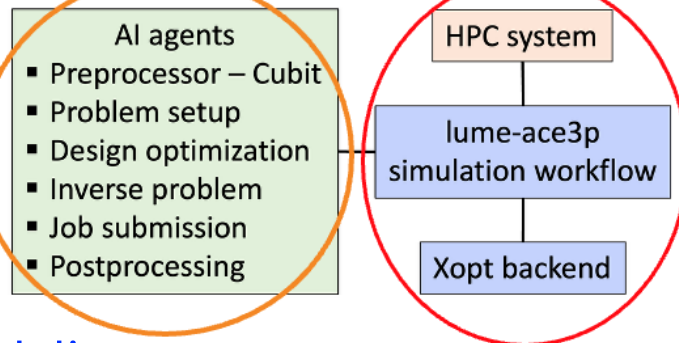
Example: Agentic Workflow for Accelerator Structure Design and Optimization



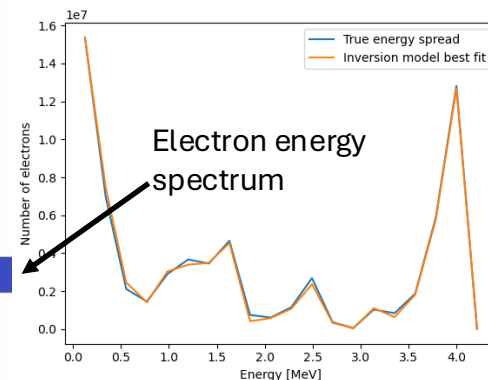
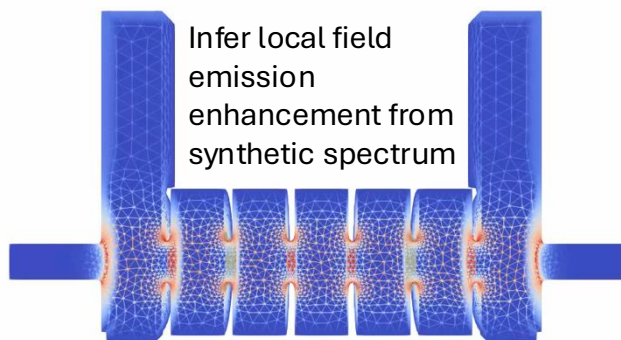
Multi-agent design assistant for ACE3P simulation workflow



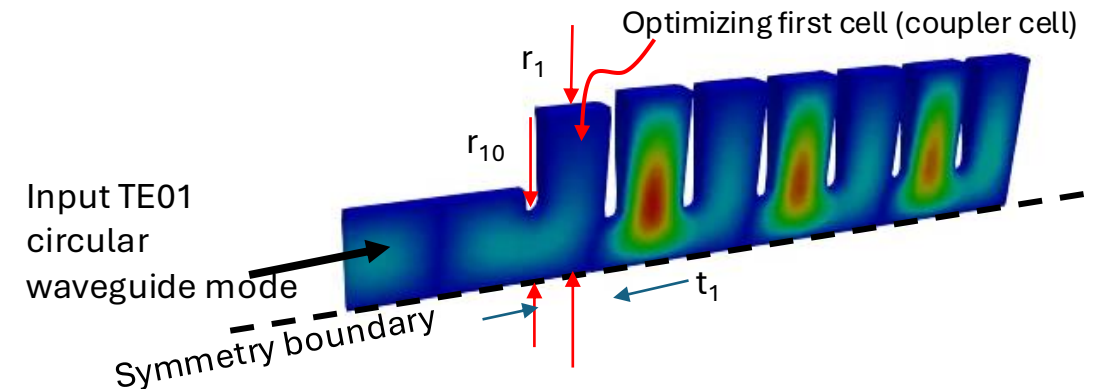
Leverage CADSAT (CAD to Simulation Agentic Toolkit) project under Genesis/AI4NS lighthouse initiative at LLNL (MADA) & SNL (Cubit).



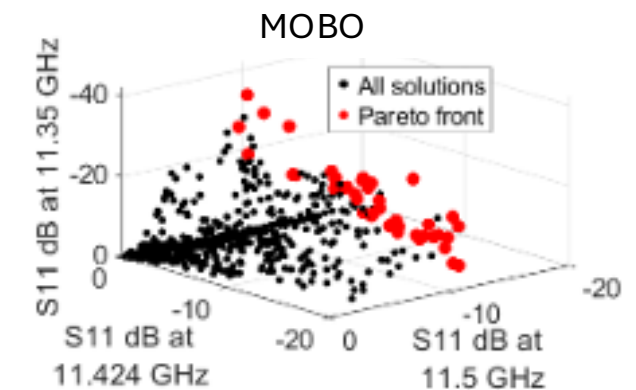
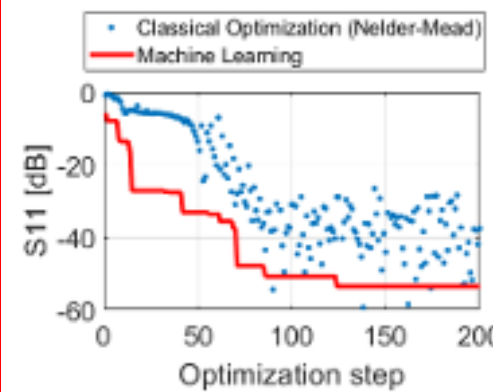
Inverse surrogate modeling



ACE3P-LUME-Xopt framework

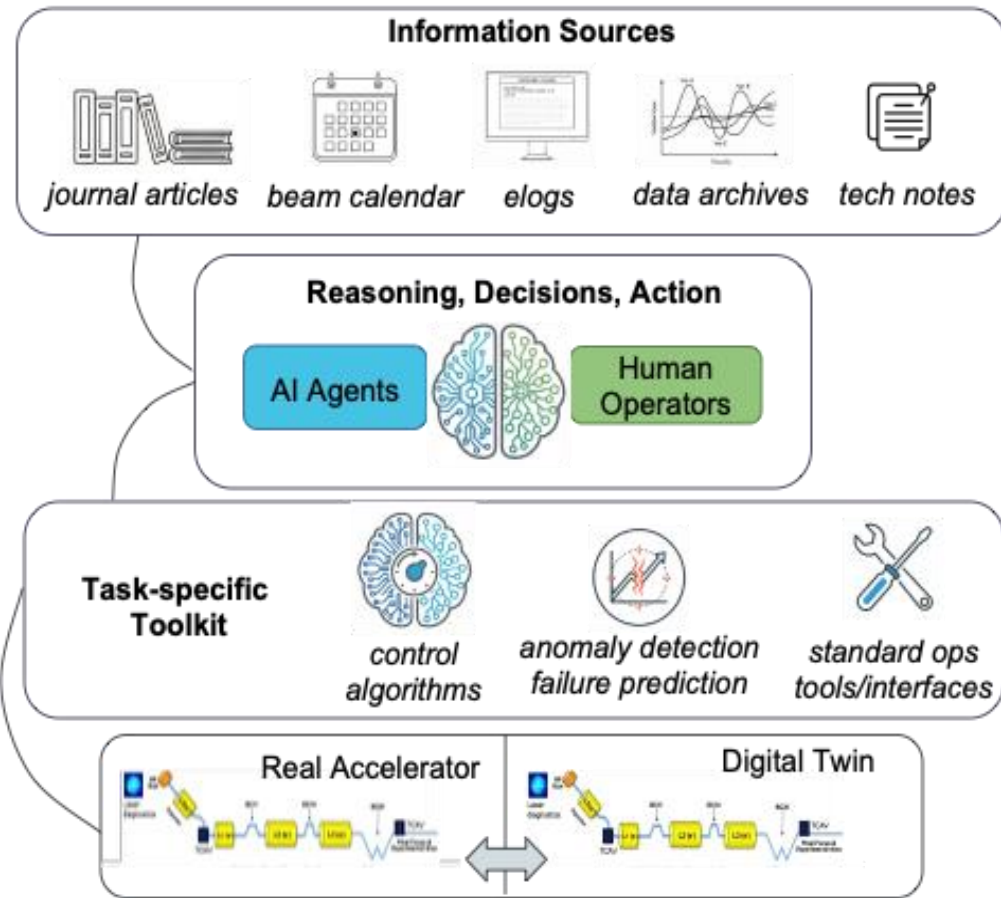


Single and multi-objective optimization with Bayesian Gaussian Process models



Faster optimal fronts via Machine Learning Exploration

Road Ahead: Agentic AI in Operations and Design



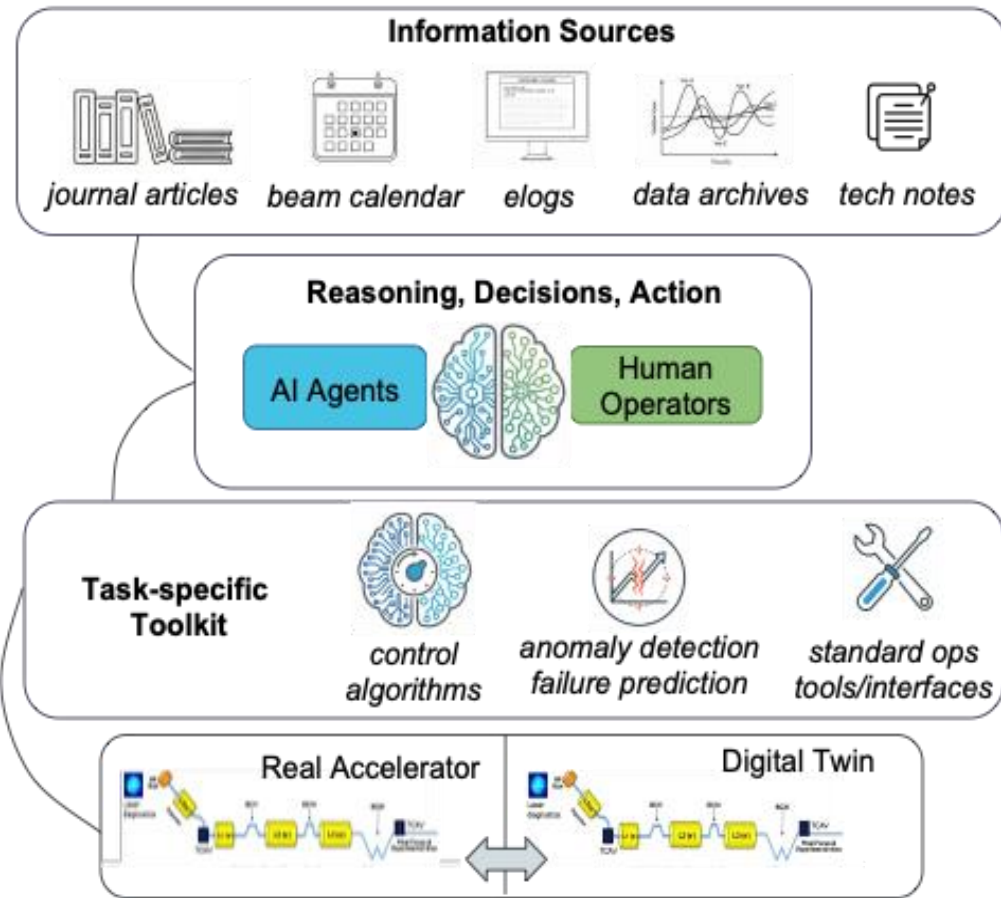
More speculative / will require development:

- Aid coordination of large teams + large complex information sets (drawings, design docs, etc)?

E.g. identify design conflicts and alert teams to updates in quasi-real time as design evolves

- Gather information from broad literature to identify unexplored opportunities and possible pitfalls based on previously built machines?
- Totally radical designs previously unachievable?

Road Ahead: Agentic AI in Operations and Design



More speculative / will require development:

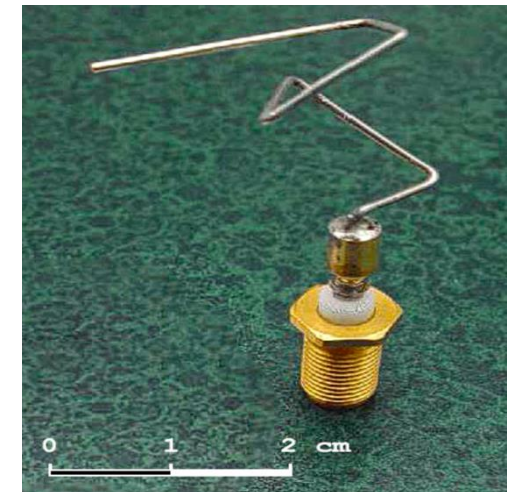
- Aid coordination of large teams + large complex information sets (drawings, design docs, etc)?

E.g. identify design conflicts and alert teams to updates in quasi-real time as design evolves

- Gather information from broad literature to identify unexplored opportunities and possible pitfalls based on previously built machines?
- Totally radical designs previously unachievable?

If this seems controversial or uncomfortable: consider the community's adoption of MOGA (multi-objective genetic algorithms)

→ this is simply a next advance in the computational tools at our disposal



Practicalities / infrastructure to consider in design of an AI-ready machine

- Data storage, tagging, movement, access / performance of access
- Middle layers / abstractions + ease of interfacing high level applications with devices
- Information systems (e.g. elog, device databases, operations wikis, documentation) and AI-ready interfaces
- Flexible software environments and processes to support rapid adaptation to new tools
- HPC integration, networking, security with local and offsite HPC
- Comprehensive virtual accelerator available from beginning and integrated into as-built digital twin for machine → available for full software stack testing

The screenshot shows a web interface for a 'Dedicated AI/ML Accelerator Test Facility'. The header is blue with the title. Below it, the dates '15-17 Jul 2026' and location 'SLAC' are shown, along with a search bar. A left sidebar contains navigation links: Overview, Scientific Program, Call for Abstracts, Timetable (selected), Contribution List, Registration, Participant List, SLAC Site Access, Venue + Meeting Room, and Contact. The main area is titled 'Timetable' and shows a date selector for 'Wed 15/07', 'Thu 16/07', 'Fri 17/07', and 'All days'. There are buttons for 'Print', 'PDF', 'Full screen', 'Detailed view', and 'Filter'. A 'Session legend' indicates 'Discussion' (grey) and 'Operation of scientific systems' (green). The timetable itself shows sessions starting at 08:00. The first session is 'Intro' (grey) from 08:30 to 08:40. From 09:00 onwards, sessions are green, indicating 'Operation of scientific systems'. These include: 'AI/ML Needs for a Commercial EUV Light Source' (09:00-09:20) by Christopher Pierce; 'Online Deployment and Operational Experience of ML-Based Beam Tuning at the ATLAS Heavy Ion Linac' (09:40-10:00) by Adwait Ravichandran; 'Outlook for AI/ML tools for SNS' (09:40-10:00) by Dr Sasha Zhukov; 'Practical Requirements for AI in Accelerator Operations' (10:00-10:20) by Benjamin Ripman; and 'Genesis Activities and MOAT' (10:20-10:40) by Jean-Luc Vay. All sessions are located at '48/1-112A/B/C/D - Redwood A/B/C/D, SLAC'.

See “AI test facility” workshop presentations and forthcoming white paper:
<https://indico.slac.stanford.edu/event/10526/timetable/#20260715.detailed>



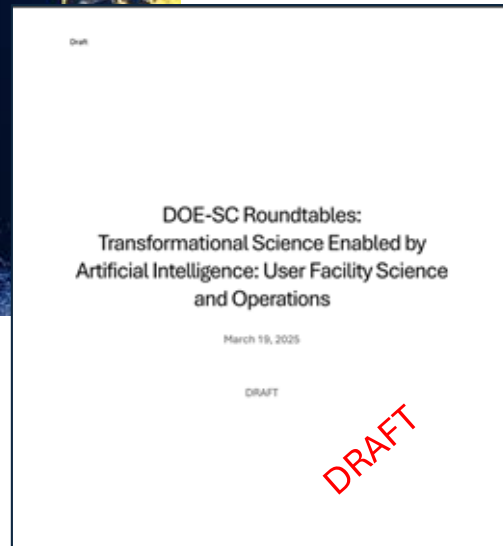
2019



2023



2023



2025

DOE-wide, highly-coordinated initiatives

- Ambitious vision to accelerate science and engineering
- Aim to leverage the full capabilities and talent of DOE

Follows years of community-driven info. gathering and strategizing

Summary / Conclusions

- **Opportunities to design with realistic operations in mind + AI/ML based on experience today**

- Numerous AI/ML tools are being integrated into operations today
- Common issues/barriers stemming from infrastructure deficits and lack of attention to software development are known
- See forthcoming “AI/ML integrated facilities” white paper with broad community input on what will be needed to fully integrate AI/ML into current and future accelerator facilities

→ *Worth taking the above into account when considering future designs*

- **Move away from piecemeal design**

- Co-design of controls / beam physics components / diagnostics
- Holistic start-to-end and top-to-bottom holistic design

→ *Likely yield performance improvements*

→ *Reduces risk and helps in critical decision making (e.g. which diagnostics can actually be discarded)*

→ *Will require broader community coordination; can start with simpler integrations and expand (integrate simulation and design of LLRF, cryo/cooling, experiment / detectors, controls, components such as rf cavity shapes, beam dynamics / lattice)*

- **Agentic AI coupled to improvements in simulation infrastructure (hybrid physics and ML modeling and differentiable sims) is going to be radically transformative for design of future accelerators**

→ *Faster design iteration, wider set of system components / performance metrics examined simultaneously, radically novel designs informed by literature gaps / broader set of aggregated knowledge*

→ *Can start today with low-hanging fruit (e.g. simple co-design) and move toward more complex integrations later (e.g. coordinating large-scale teams + across subsystems)*

- **Many opportunities to test design → operations processes and new tools with existing facilities**

→ *e.g. co-design of novel controls + diagnostics + sim2real schemes → test at FACET-II*

- **Coordinated development efforts on tools, techniques, and software (e.g. ModCon MOAT, SciDAC5 / CAMPA, AmSC, PALS) will help unlock more ambitious capabilities + accelerate progress across all accelerator-based science missions**

Thanks to many team members and collaborators across institutions on work/examples shown!

Non-exhaustive list:



Ryan Roussel
Sara Miskovich
Zhe Zhang
Dylan Kennedy
Zihan Zhu
Kiley Gruenberg
Zack Bushmann
Fuhao Ji
Ernest Williams
Claudio Emma
Brendan O'Shea
Jenny Morgan
Gopika Bhardwaj
Claudio Bisegni
Jesse Bellister
Sean Gasiorowski

Agostino Marinelli
Spencer Gessner
David Cesar
Aaron Reed
Keegan Downham
Jinseo Park
Fred Poitevan
Jana Thayer
Matt Seaberg
Yanwen Sun
Anne Sakdinawat
Chris Garnier
An Le
Nolan Stelter
J.P. Gonzalez-Aguillera
Alberto Lutman
Heinz-Deiter Nuhn
Samuel Klein



Seongyeol Kim
Gyujin Kim
Myung-Hoon Cho
Chi Hyun Shim
Hoon Heo
Haeryong Yang



Ryan Wu
Mykel Kochenderfer
Eric Darve



Daniel Ratner
Kishan Rajput
Malachi Schram
Todd Satogata



Philippe Piot
Brahim Mustapha
Jose Martinez
John Power
Eric Wisniewski
Alex Ody
Chenran Xu



Jan Kaiser
Annika Eichler
Christian Hesse



Kathryn Baker
Mateusz Leputa
Matt King
Alexander Brynes



Paul Moeller
Rob Nagler
Jonathan Edelen



Jean-Luc Vay
Remi Lehe
Thorsten Hellert
Gianluca Martino
Axel Heubl
Reva Jambunathan
Johannes Blashke



Lucy Lin
Kevin Brown
Georg Hoffstaetter
Max Rakitin



backups

Genesis MOAT (Multi-Office Accelerator Team)



- **Brings together DOE BES, HEP and NP**
 - 7+2 (PNNL, LANL) national labs (+ academia + industry partnerships + more labs)
- **Treats nation's accelerator ecosystem as a single, intelligent system**
 - Using AI to connect expertise across labs and offices
⇒ move faster, smarter, and with far greater impact
 - Views large accelerators as complex, dynamic socio-technical systems
 - distributed teams of teams
 - thousands of physical components
 - vast stores of scientific knowledge
 - rigorous safety and management frameworks
- **Several focus areas: Design, Operation, Core Capabilities / Platform Dev.**
- **Genesis Platform provides the interconnecting tissue of computing + AI capabilities**



BERKELEY LAB



A Unified and Standardized API with LUME

