

ACRO? All Scientific Simulators are Wrong... Can AI Bridge the Gap?

Richard Gao, Daniel Gedon, Georgia Channing, Noémi Éltető, Magdalena Lederbauer

[TODO: Necessary content:

- **Workshop title**
- **Topic overview & ICML relevance**
- **Activities & schedule**
- **Invited speakers & confirmation status**
- **Workshop history**
- **Related recent ICML/NeurIPS events**
- **Organizers, bios, links, experience, conflicts**
- **Designated contact organizers**
- **Website**

]

1 Workshop Summary

[Brief description of the topics to be covered, and an explanation as to why the workshop will appeal to ICML audiences]

[TODO: Currently this is a typical workshop website blurb and would need to be rewritten for the proposal.]

Machine learning for advancing scientific discovery, also known as AI4Science, has rapidly emerged as a central research theme across disciplines and sub-communities. Scientific simulators are a central component of this development. These are mechanistic models that encode decades of domain knowledge and enable prediction, understanding, and hypothesis generation for complex real-world systems. Simulator-based workflows encompass the full scientific pipeline built around such simulators, including data handling, forward simulation, inference, uncertainty quantification, and decision-making. Nowadays, machine learning plays a crucial role across all stages of these simulator-based workflows.

Inherent to simulator-based workflows is the assumption or even the requirement that the simulator accurately represents reality. However, all simulators include simplifications and approximations of reality, leading to discrepancies between simulated and observed data. In other words, "all models are wrong, some are useful". This raises a central question: How can we make the best use of imperfect simulators? How can machine learning help mitigate the limitations of simulator-based workflows in the presence of simulator discrepancies?

This workshop focuses on advances in machine learning that address the challenges arising with scientific simulator-system discrepancies throughout simulator-based workflows. In particular, recent works that identify, diagnose, mitigate, and/or augment simulator discrepancies in machine learning for scientific applications. We invite contributions from all scientific domains that use simulators and machine learning in their workflows. Topics of interest include methods and applications addressing simulator discrepancies in

simulator-based workflows, including but not limited to:

[TODO: maybe just reduce down to 2 lines, theory + application then list in-line — For proposal, yes. For the website, we would like to keep a list. — Would need more details on the application side]

[TODO: think about the spectrum + balance from theory/methods vs. application]

- Parameter inference and simulation-based inference
- Model and component discovery
- Multi-fidelity simulation and model coupling
- Experiment design and hypothesis verification
- Generalized Bayesian inference
- LLM-assisted inference and reasoning
- Differentiable simulators and frameworks
- Simulator pipeline including data handling and summary statistics learning
- Active learning and Bayesian optimization
- Emulator and surrogate learning
- Application areas in biology, chemistry, physics, and related fields

This workshop aims to unite researchers from different communities dealing with simulator discrepancies in the simulator-based workflow. By bridging communities across machine learning and the natural sciences, we seek to foster a common understanding of how to quantify, learn from, and reduce simulator discrepancies across fields.

2 Activities and tentative schedule

[Short description and rough timetable of all planned activities (talks, posters, panels), detailing planned in-person and virtual elements]

Options to include in a possible schedule:

- 6 invited talks (30 min each)
- orals for best papers
- poster session for accepted papers
- panel with moderated Q and A, including lightning-round format (e.g., 45s answers) to maximize efficiency
- something that encourages audience participation (e.g., random debates?)
- sponsored events, e.g., mentorship lunch (pending sponsors obviously), if we can lock in a \$3-5k contribution, it would host about 50-70 people for a catered lunch...but do ICML workshops provide lunch? If so, it could be travel awards
- tutorials or demos or something? Here, something we have experience doing would go a long way for credibility

Table 1: Tentative schedule of events (so far copy past from RG’s NeurIPS workshop proposal.)

Time	Event	Time	Event
9:00 – 9:10am	Opening remarks	1:30 – 2:00pm	Invited talk 4
9:10 – 9:50am	Lightning talks	2:00 – 2:40pm	Poster session
9:50 – 10:20am	Invited talk 1	2:50 – 3:20pm	Invited talk 5
10:20 – 10:50am	Invited talk 2	3:20 – 3:50pm	Invited talk 6
11:00 – 11:50am	Discussion sessions	3:50 – 4:20pm	Un-Poster session
11:50 – 1:00pm	Networking lunch	4:30 – 5:00pm	Invited talk 7
1:00 – 1:30pm	Invited talk 3	5:10 – 5:45pm	Moderated panel

3 Invited speakers and Scientific Advisors

[List of invited speakers, specifying who has been confirmed. We strongly recommend that all speakers present in person. However, virtual talks will be allowed in special circumstances to account for hard constraints (e.g., denied visas) or to significantly increase the diversity of speakers.]

Speakers. We have secured confirmation from all 6 invited speakers, which span a range of relevant research topics involving ML + simulators, as well as backgrounds, i.e., from academic / university labs, industry labs, FROs, etc., Gender/location / seniority were also taken into consideration, with X/Y/Z m/f/d speakers, X PhD/postdoc, and X ECR (pre-tenured faculty). Confirmed speakers:

- Shirely Ho
- Jonas Köhler

Could do some break-out groups/go more domain-specific later. Have plenary talks in the morning. Also, maybe a panel. Mentorship lunch: students + big shots.

Mentorship lunch, sponsored stuff,

Scientific Advisors. Confirmed:

- Jakob Macke
- Ceclia Celmenti
- Max Welling

4 History

This is the first proposed iteration of this workshop at ICML or elsewhere.

5 Previous related workshops

[Similar past and current events at ICML and NeurIPS in the last 1-2 years, even if not organized by the present workshop organizers: New workshops are welcome to build on prior workshops if a good case is made; completely original workshops are also welcome]

Closely related workshops list. This is a selection of closely relevant workshops in ICML/ICLR/NeurIPS since 2020.

- Machine Learning and the Physical Sciences, <https://ml4physicalsciences.github.io>
- Data-driven and Differentiable Simulations, Surrogates, and Solvers (D3S3), <https://d3s3.github.io>
- Building Physically Plausible World Models, <https://openreview.net/forum?id=E2prL2W5dZV>
- Synergy of Scientific and Machine Learning Modeling, <https://syms-ml.github.io/2023/>
- Machine Learning for Astrophysics, <https://ml4astro.github.io/icml2022/>
- Deep Learning for Simulation, <https://simdl.github.io>

Proposals should be two pages long, in single-column A4 or letter format, with font size 11 or greater, excluding organizer contact details/CVs and bibliographic references

6 Organizers

[List of organizers with email addresses, web page URLs, pointers to Google Scholar or other similar citation service pages, a one-paragraph bio for each organizer, describing research expertise, and previous experience organizing scientific meetings. Unless explicitly indicated, we will assume all organizers are intending to participate in person.]

[Also brief note explicitly stating diversity considerations of organizers (in research topic, career stage, gender, location (?), academic lineage, etc.).]

1. **Richard Gao**, r.dg.gao@gmail.com, is an assistant professor at Goethe University Frankfurt. Website: rdgao.com; Google scholar: link. Research interests. Organizing experience. Contact person for communications.
2. **Daniel Gedon**, daniel.gedon@uni-tuebingen.de, is a postdoctoral fellow at Tübingen University. Website: dgedon.github.io; Google Scholar: link. Research interests. Organizing experience. Contact person for communications.
3. **Georgia Channing**, georgia.channing@huggingface.co, leads the AI for science team at Hugging Face. Website: huggingface.co/cgeorgiaw; Google Scholar: link. Research interests. Organizing experience.
4. **Noémi Éltető**, noemieteto@gmail.com, is a research scientist at Google Deepmind. Website: noemieteto.github.io; Google scholar: link. Research interests. Organizing experience.
5. **Magdalena Lederbauer**, magled@mit.edu, is a PhD student in Chemical Engineering and Computational Science and Engineering at MIT, advised by Connor Coley. Her research focuses on machine learning for chemical discovery, including molecular structure elucidation using physics-constrained neural networks and multi-modal learning approaches for extracting synthesis procedures from scientific literature at scale. She received her BSc and MSc in Chemistry from ETH Zurich. Website: mlederbauer.com; Google Scholar: link.

7 Contact information

[Names of two organizers designated as contacts for all communications]

Daniel Gedon and Richard Gao

8 Bibliography

TODO