2025 EMBRIO SUMMER SCHOOL

AI in Research, Data Management Best Practices, and Introduction to Biological Modeling Tools
June 9–11, 2025
📍 *Martin C. Jischke Hall of Biomedical Engineering (MJIS), Room 1097, Purdue University*

Schedule at a Glance (summary of each session providing starting on page 3)

Day 1 – Monday, June 9

Theme: Integrating AI Into Your Everyday Research Workflows
8:00 AM – 5:45 PM

- **8:00 – 8:30 AM** – Registration & Breakfast **(no food or drink allowed in Room 1097)**
- **8:30 AM** – Welcome – Jeremy Zartman, Co-PI and Co-Director, EMBRIO Institute
- **8:40 AM** – Orientation to the schedule
- **8:45 AM** – Quick Intro to Common AI Tools – Ashish, RCAC
- **9:00 – 10:00 AM** – Prompt Engineering: Foundations + Research Use Cases – Ashish, RCAC
- **10:00 – 10:15 AM** – Break
- **10:15 – 11:30 AM** – Retrieval-Augmented Generation (RAG) – Ashish, RCAC
- **11:30 AM – 1:00 PM** – Lunch (on own or in groups)
- **1:00 – 2:00 PM** – Using AI for Literature Review and Analysis (Scite.ai) – Dave Zwicky, Purdue
- **2:00 – 2:15 PM** – Break
- **2:15 – 3:30 PM** – Fine-Tuning Custom Models – Ashish, RCAC
- **3:30 – 3:45 PM** – Break
- **3:45 – 5:00 PM** – Building a Simple AI Agent – Ashish, RCAC
- **5:00 – 5:45 PM** – Debrief the Day with RCAC Staff: Q&A + Heavy Appetizers, Form teams or continue solo work; brainstorm AI use case ideas

- **Evening** – Dinner (on own or in small groups)

Day 2 – Tuesday, June 10

Themes: Developing Your AI Use Case and Data Management Best Practices

8:00 AM – 3:30 PM

- **7:30** Breakfast items available ***(no food or drink allowed in Room 1097)***
- **8:00 – 8:15 AM** – Collaborating with RCAC on integrating AI into your research

- **8:15 – 9:45 AM** – Work on Your Use Case (team or individual; support available from RCAC)
- **9:45 – 10:00 AM** – Break
- **10:00 – 11:15 AM** – EMBRIO Mini-Showcase: Uses of AI in Research – Lucas Wiese, Linlin Li, Nilay Kumar
- **11:15 AM – 1:00 PM** – Lunch
- **1:00 – 3:00 PM** – EMBRIO Data Management and Sharing Hands-On Tutorial – Scott Bolton (See Instructions for prepping your sample data and publications)
- **3:00 – 3:30 PM** – Break
- **Evening** – Dinner (on own or in small groups)

Day 3 – Wednesday, June 11

Theme: Biological Modeling Tools

9:00 AM – 1:00 PM

- **8:30** Breakfast items available ***(no food or drink allowed in Room 1097)***

- **9:00 – 10:30 AM** – CompuCell3D Cell Sorting Demo – Pedro Cenci dal Castro
- **10:30 – 10:45 AM** – Break
- **10:45 – 12:45 PM** –Tissue Forge Modeling Demo, Sharon Minsuk (See instructions for installing Miniconda pk mngr ahead of the session)
- **12:45** –– Lunch (boxed)
- Wrap-up and Evaluation (Attendees, please complete the online evaluation survey)
- Until 2:00 PM – Poster Setup for Thursday Retreat (Please use assigned poster numbers)

Details of each session

1. Prompt Engineering: Foundations + Research Use Cases: Ashish, Sr. Data Scientist, RCAC, Purdue

What we'll cover:

- Prompt types: zero-shot, few-shot, chain-of-thought
- Role vs task prompting
- Research examples: paper summarization, protocol translation, generating literature queries

Hands-on:

- Live walkthrough comparing prompt styles
- Small group prompt crafting based on research scenarios

2. Retrieval-Augmented Generation (RAG): Ashish, Sr. Data Scientist, RCAC, Purdue

What we'll cover:

- Why RAG is a better fit than fine-tuning for many research cases
- Pipeline: document chunking → embedding → retrieval → answer generation
- Tools: LangChain, FAISS/Chroma, PDFLoader

Hands-on:

- Load and chunk EMBRIO documents
- Embed and query using RAG logic
- Ask questions like: "What methods were used in dataset X?"

3. Fine-Tuning Custom Models: Ashish, Sr. Data Scientist, RCAC, Purdue

What we'll cover:

- Fine-tuning vs RAG
- Intro to LoRA, prompt tuning, and full model updates
- When fine-tuning is appropriate in research workflows

Hands-on:

- Use Hugging Face Transformers + PEFT on a toy dataset
- Fine-tune a small model (e.g., DistilBERT) and evaluate output

4. Using AI for Literature Review and Analysis: Dave Zwicky, Assoc. Prof, Science and Engineering Libraries, Purdue

What we'll cover:

- Introduction to the Scite.ai tool
- Literature review and analysis
- How to best interpret and use the results
- Avoiding pitfalls

Hands-on:

- Trying out the tool
- Sharing your results

5. Building a Simple AI Agent: Ashish, Sr. Data Scientist, RCAC, Purdue

What we'll cover:

- What an AI agent is (vs a single model call)
- Components: LLM, tools, memory, decision logic
- Use cases in research automation (e.g., summarizing papers, generating structured reports)

Hands-on:

- Build a reactive LangChain agent with a summarizer + formatter
- Let attendees modify logic for their research needs

6. Your Use Case

Work on your own or in teams to apply AI for your own use case. (RCAC and others available for help)

Prepare a 2-4 slide 5 min presentation to share your use case:

- The Problem/Idea,
- What tools, methods, inputs used,
- The Challenges encountered, did you overcome them
- What did you learn,
- What you want to learn or try next

7. **EMBRIO Mini-Showcase: Uses of AI in Research – Lucas Wiese, Linlin Li, Nilay Kumar**

This session is meant to inspire ideas regarding the value of AI use in research. It is not meant to teach you exactly how to perform these methods, but rather to provide several basic to advanced use-case examples of how EMBRIO members are using AI in their research workflows.

Presenters:
Lucas Wiese (Ph.D. Candidate, Magana, ROCkETEd Lab and Governance and Responsible AI lab (GRAIL) Center, Purdue)
Linlin Li (Senior Research Scientist, Umulis Lab: Quantitative and Systems Biology Research Group, Purdue)
Nilay Kumar (Senior Research Scientist, Pienaar Lab, Computational Systems Parmacology Lab, and Buganza-Tepole Lab, Mechanics and Mechanobiology Lab)

8. EMBRIO Data Management and Sharing Hands-On Tutorial (Scott Bolton, Research Scientist, EMBRIO and BME, Purdue)

(See Instructions for prepping your sample data and publications)

In this workshop, we will be learning about bioimaging data storage and management using EMBRIO repository resources. Topics included will be FAIR principles for bioimaging, Data Depot storage, Purdue University Research Repository (PURR) web pages, OMERO bioimaging database, and metadata handling methods. We will show how to upload and manage image data from microscopy experiments, evaluate microscopy metadata, and begin looking at higher-level experimental metadata for keyword tagging.

9. **CC3D Introduction and Demo – Pedro Cenci dal Castro (Glazier Lab, IU, Biocomplexity Institute)**

CompuCell3D – CC3D is a framework for modeling multiscale virtual tissues. It includes an editor, a player, and Python API for modularization of simulations, allowing integration of the simulations to other systems, such as parameter scan in clusters and deployment on the web via Jupyter Notebooks. The objective of this introductory course is to 1) familiarize the students with concepts of the Cellular Potts Model – CPM, such as fluctuation amplitude, constraints and energy minimization, and 2) implement a model using CC3D. For the model implementation, we will use the cell sorting demonstration, where a very simple simulation architecture is able to achieve various patterns of spatial cell organization. I will provide links to other demonstrations, simulations and learning materials.

The session will provide the following introduction:

- Overview of CompuCell3D and the Cellular Potts Model
- Energy functional and constraints
- Energy minimization
- The CPM algorithm for minimizing global energy
- Fluctuation Amplitude
- CC3D tools and applications
- CC3D plugins, capabilities and limitations

- Opening a CC3D simulation on the web browser (no installation required)
- Defining a model
- Cell Sorting model – adhesion hypothesis
- Implementing the Cell Sorting model on CC3D
- Exploring the outcomes

10. **Tissue Forge – Sharon Minsuk (Glazier Lab, IU, Biocomplexity Institute)**

(See instructions for installing Miniconda pk mngr ahead of the session)

Tissue Forge (TF) is another, very different, framework for multiscale tissue modeling. Like CC3D, it has an API for programming in Python (as well as in C or C++). It is based on a completely different concept known as "center-based" modeling. The basic units of TF modeling are particles, defined as 0-dimensional masses located in 3d space, which can exert forces on one another. By carefully defining our particles and forces, we can model across scales from the molecular to the whole tissue. We will explore these concepts, and implement a simple TF model. By implementing a cell sorting simulation, just as was done in the CC3D workshop, we'll gain an appreciation for the similarities and differences, and strengths and weaknesses, of the two modeling frameworks.

Tentative Outline:

- Overview of Tissue Forge
- Particles
  - Particle Types; basic particle properties: mass, radius, dynamics
  - Particle class hierarchy: ParticleTypeSpec, ParticleType, Particle
  - Generating particle instances
- Forces
  - Potentials
  - Types of potentials supported by Tissue Forge
  - Type-based potential binding
  - Explicit particle-pair-based bonding
  - Force vectors
- Hands-on: building a cell sorting simulation
- Topics, time permitting – participant choice (Q&A and code examples)