Decoding Collective Dynamics: Machine Learning Insights into Active Matter Simulations

Project Proposal 01 — Predicting Emergent Dynamics

Problem Statement

Active matter systems—composed of self-propelled, interacting particles—exhibit complex emergent behaviors such as collective motion, vortical turbulence, and spontaneous symmetry breaking. Understanding and predicting the evolution of such systems typically require expensive numerical simulations. Yet, many macroscopic features of interest (e.g., global vorticity, alignment order, kinetic energy) can be summarized as time-series observables extracted from large-scale simulations. Efficiently forecasting their short-term evolution could provide both physical insight and data-driven benchmarks for reduced active-fluid models.

Proposed Approach

Using **The Well Dataset**, which provides 24 repetitions of 81-frame simulations of active rod-like particles on a 256×256 grid across alignment and dipole-strength sweeps, this project focuses on the **forecasting of global observables** derived from the full fields.
The workflow is structured into three main components:

1. **Feature extraction & baseline exploratory analysis (EDA)**
    - Compute time-series of global observables: mean vorticity, polarization (order parameter), and kinetic energy.
    - Visualize temporal trends, compute autocorrelation, and assess characteristic timescales.
2. **Problem framing & baseline forecasting models**
    - Define fixed-window forecasting (e.g., input = 70 steps → output = next 11 steps).
    - Implement baselines: **persistence model** (copy last value) and **linear regression** on lagged inputs.
    - Evaluate using **MSE** and **MAE**, establishing reference performance.
3. **Advanced model plan**

- Review literature on **RNNs, LSTMs, and attention-based models** for physics-informed time-series forecasting.

- Draft model specifications: input/output tensor shapes, loss function (e.g., multi-step MSE), training protocol, and validation strategy (e.g., train/test split over simulations).

Hypotheses

- (H1) Short-term dynamics of global observables are low-dimensional and predictable with simple autoregressive structure.

- (H2) Neural architectures incorporating temporal memory (LSTM/attention) will outperform linear baselines, especially near dynamical transitions.

- (H3) Forecast skill will depend systematically on control parameters (alignment and dipole strength), revealing interpretable dynamical regimes.

Integration into Joint Manuscript

This component contributes to the **temporal forecasting and analysis of macroscopic observables** section of the joint paper *"Decoding Collective Dynamics."*

- Provides **quantitative baselines** and **predictive metrics** for reduced modeling efforts.

- Supplies **summary plots and error analyses** for inclusion alongside field-level machine-learning results.

- Serves as a reproducible module within *The Well* processing pipeline, bridging microscopic simulation data and system-level temporal behavior.

Outcome: A compact, data-driven study of predictability in active matter observables, forming the time-series analysis backbone for the collaborative manuscript.

The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning

Abstract

Machine learning based surrogate models offer researchers powerful tools for accelerating simulation-based workflows. However, as standard datasets in this space often cover small classes of physical behavior, it can be difficult to evaluate the efficacy of new approaches. To address this gap, we introduce *the Well:* a large-scale collection of datasets containing numerical simulations of a wide variety of spatiotemporal physical systems. The Well draws from domain experts and numerical software developers to provide 15TB of data across 16 datasets covering diverse domains such as biological systems, fluid dynamics, acoustic scattering, as well as magneto-hydrodynamic simulations of extra-galactic fluids or supernova explosions. These datasets can be used individually or as part of a broader benchmark suite. To facilitate usage of the Well, we provide a unified PyTorch interface for training and evaluating models. We demonstrate the function of this library by introducing example baselines that highlight the new challenges posed by the complex dynamics of the Well. The code and data is available at `https://github.com/PolymathicAI/the_well`.

C.2 `active_matter`

Description of the physical phenomenon. We are interested in studying the dynamics of N active particles of length ℓ and thickness b (aspect ratio $\ell/b \gg 1$) immersed in a Stokes fluid with cubic volume V. In large particle limit, continuum kinetic theories describing the evolution of the distribution function $\Psi(\mathbf{x},\mathbf{p},t)$ have proven to be useful tools for analyzing and simulating particle suspensions [116, 117]. The Smoluchowski equation governs Ψ's evolution, ensuring particle number conservation,

$$\frac{\partial \Psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\dot{\mathbf{x}}\Psi) + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\Psi) = 0, \tag{4}$$

where the conformational fluxes $\dot{\mathbf{x}}$ and $\dot{\mathbf{p}}$ are obtained from the dynamics of a single particle in a background flow $\mathbf{u}(\mathbf{x},t)$. The moments of Ψ yield the concentration field $c = \langle 1 \rangle$, polarity field $\mathbf{n} = \langle \mathbf{p} \rangle / c$, and nematic order parameter $\mathbf{Q} = \langle \mathbf{pp} \rangle / c$, with $\langle f \rangle = \int_{|\mathbf{p}|=1} f\Psi \, d\mathbf{p}$. For dense suspensions, the conformational fluxes are

$$\dot{\mathbf{x}} = \mathbf{u} - d_T \nabla_{\mathbf{x}} \log \Psi; \quad \dot{\mathbf{p}} = (\mathbf{I} - \mathbf{pp}) \cdot (\nabla \mathbf{u} + 2\zeta \mathbf{D}) \cdot \mathbf{p} - d_R \nabla_{\mathbf{p}} \log \Psi. \tag{5}$$

Here d_T and d_R are dimensionless translational and rotational diffusion constants, ζ is the strength of particle alignment through steric interactions, and $\mathbf{D} = \langle \mathbf{pp} \rangle$ is the second-moment tensor. The Smoluchowski equation is coupled to the Stokes flow as

$$-\Delta \mathbf{u} + \nabla P = \nabla \cdot \mathbf{\Sigma}, \nabla \cdot \mathbf{u} = 0, \tag{6}$$

$$\mathbf{\Sigma} = \alpha \mathbf{D} + \beta \mathbf{S} : \mathbf{E} - 2\zeta\beta(\mathbf{D} \cdot \mathbf{D} - \mathbf{S} : \mathbf{D}). \tag{7}$$

Here $P(\mathbf{x},t)$ is the fluid pressure, α is the dimensionless active dipole strength, β characterizes the particle density, $\mathbf{E} = [\nabla \mathbf{u} + \nabla \mathbf{u}^\top]/2$ is the symmetric rate-of-strain tensor, and $\mathbf{S} = \langle \mathbf{pppp} \rangle$ is the fourth-moment tensor. The stress tensor $\mathbf{\Sigma}$ in Eq. (7) includes contributions from active dipole strength, particle rigidity, and local steric torques. Despite the fact that kinetic theories are consistent with microscopic details and are amenable to analytical treatment, they are not immune from computational challenges. For instance, in dense suspensions with strong alignment interactions (high ζ), the cost to resolve the orientation field $\mathbf{p}$ is prohibitively high even in 2D. Though approximate coarse-grained models that track only low-order moments exist, they rely on phenomenological [118][119] or learned corrections [120] to close the system. This underscores the need for fast, high-fidelity, data-efficient physical surrogate models to track and predict the evolution of few low-order moments. An autoregressive surrogate model can efficiently screen the high-dimensional parameter space of complex active matter systems and help design self-organizing materials that switch between nontrivial dynamical states in response to external actuation or varying parameters.

Simulation details. We numerically close the system of equations (4)-(6) using pseudo-spectral discretization where Fourier differentiation is used to evaluate the derivatives with respect to space and particle orientation. We use the second order implicit-explicit backward differentiation time-stepping scheme (SBDF2), where the linear terms are handled implicitly and the nonlinear terms explicitly with time-step $\Delta t = 0.0004$. The numerical simulations are performed in a periodic square domain of length $L = 10$ with 256^2 spatial modes and 256 orientational modes. The simulation code is available at `https://github.com/SuryanarayanaMK/Learning_closures/tree/master`. The approximate time to generate the data is 20 minutes per simulation on an A100 80GB GPU in `fp64` precision. In total, this is about 75 hours of simulation.

Varied Physical Parameters. $\alpha \in \{-1,-2,-3,-4,-5\}$ $\beta = 0.8$; $\zeta \in \{1,3,5,7,9,11,13,15,17\}$.

Fields present in the data. concentration (scalar field), velocity (vector field), orientation tensor (tensor field), strain-rate tensor (tensor field).

References to cite when using these simulations: [120].

```
Number of simulation repetitions: 24
Size of each repetition: torch.Size([81, 256, 256, 11])
Field names: ['concentration', 'velocity_x', 'velocity_y', 'D_xx', 'D_xy', 'D_yx', 'D_yy', 'E_xx', 'E_xy', 'E_yx', 'E_yy']
```