PNAS RESEARCH ARTICLE | PHYSICS OPEN ACCESS

Data-driven Mori–Zwanzig modeling of Lagrangian particle dynamics in turbulent flows

Xander M. de Wit[a,b,c,1], Alessandro Gabbana[b,c,d], Michael Woodward[b,c], Yen Ting Lin[e], Federico Toschi[a,f], and Daniel Livescu[b,1]

Affiliations are included on p. 8.

Edited by Detlef Lohse, University of Twente, Enschede, Netherlands; received September 10, 2025; accepted February 6, 2026, by Editorial Board Member Daan Frenkel

The dynamics of Lagrangian particles in turbulence play a crucial role in mixing, transport, and dispersion in complex flows. Their trajectories exhibit highly nontrivial statistical behavior, motivating the development of surrogate models that can reproduce these trajectories without incurring the high computational cost of direct numerical simulations of the full Eulerian field. This task is particularly challenging because reduced-order models typically lack access to the full set of interactions with the underlying turbulent field. Novel data-driven machine learning techniques can be powerful in capturing and reproducing complex statistics of the reduced-order/surrogate dynamics. In this work, we show how one can learn a surrogate dynamical system that is able to evolve a turbulent Lagrangian trajectory in a way that is point-wise accurate for short-time predictions (with respect to Kolmogorov time) and stable and statistically accurate at long times. This approach is based on the Mori–Zwanzig formalism, which prescribes a mathematical decomposition of the full dynamical system into resolved dynamics that depend on the current state and the past history of a reduced set of observables, and the unresolved orthogonal dynamics due to unresolved degrees of freedom of the initial state. We show how by training this reduced order model on a point-wise error metric on short time-prediction, we are able to correctly learn the dynamics of Lagrangian turbulence, such that also the long-time statistical behavior is stably recovered at test time. This opens up a range of applications, for example, for the control of active Lagrangian agents in turbulence.

turbulence | machine learning | Mori–Zwanzig

Significance

Modeling the motion of small particles in turbulent flows is central to a wide range of scientific and engineering problems ranging from cloud formation and pollutant transport to manufacturing and nuclear fusion. Yet, despite a century of research, reproducing the statistical features of these Lagrangian trajectories has remained a formidable challenge for traditional modeling approaches. We introduce a data-driven autoregressive framework that captures the dynamics of Lagrangian turbulence within a surrogate dynamical system. This advance demonstrates that machine learning can overcome longstanding barriers in modeling chaotic particle motion, opening opportunities for prediction, design, and control of turbulent processes across science and engineering.

Understanding and predicting the trajectories of particle tracers advected by fully developed turbulent flows is a central challenge in fluid dynamics, with applications spanning a wide range of fields, including pollutant dispersion, cloud microphysics, aerosol dynamics in atmospheric chemistry, industrial processes, combustion processes, and biophysical systems (1–3). Progress in this direction has proven difficult due to the chaotic and multiscale nature of turbulence, as the dynamics of particles is influenced by complex interactions across a wide range of spatial and temporal scales (4–6).

A natural framework for studying this problem is the Lagrangian description of fluid motion, where the focus is on following individual particles as they are transported by the fluid flow. In this context, Lagrangian tracers represent idealized point-like and massless particles, whose instantaneous velocities match the local Eulerian velocity field $\boldsymbol{u}(\boldsymbol{r}, t)$, such that their dynamics obeys $\boldsymbol{v}(t) \equiv \dot{\boldsymbol{x}}(t) = \boldsymbol{u}(\boldsymbol{x}(t), t)$. As Lagrangian tracers travel the entire range of multiscale turbulent fluctuations, they recover many of the hallmarks of the statistics of turbulence (6), as the Lagrangian and Eulerian descriptions are closely connected. In particular, the small-scale dynamics of velocity structures is associated with particle accelerations, which are highly intermittent and exhibit extreme fluctuations. This manifests itself most strikingly through the probability density functions (PDFs) of Lagrangian accelerations encompassing heavy, non-Gaussian tails (7–9).

From the numerical point of view, turbulent Lagrangian dynamics is typically studied using direct numerical simulations (DNS). There, Lagrangian trajectories of particles in turbulent flows are computed from the fully resolved velocity field, which involves solving the full Navier–Stokes equations numerically without modeling approximations (2, 6, 10).

While DNS provides the most accurate trajectories, the associated high computational costs have motivated the development of a wide range of reduced order models over the past decades, which in general do not require the explicit solution of the underlying flow field, and yet are successful in preserving (some of) the statistical features of Lagrangian

Author contributions: M.W., Y.T.L., F.T., and D.L. designed research; X.M.d.W. and A.G. performed research; X.M.d.W., A.G., and Y.T.L. contributed new reagents/analytic tools; X.M.d.W. analyzed data; and X.M.d.W., A.G., and M.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.L. is a guest editor invited by the Editorial Board.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: x.m.d.wit@tue.nl or livescu@lanl.gov.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2525390123/-/DCSupplemental.

Published March 25, 2026.

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

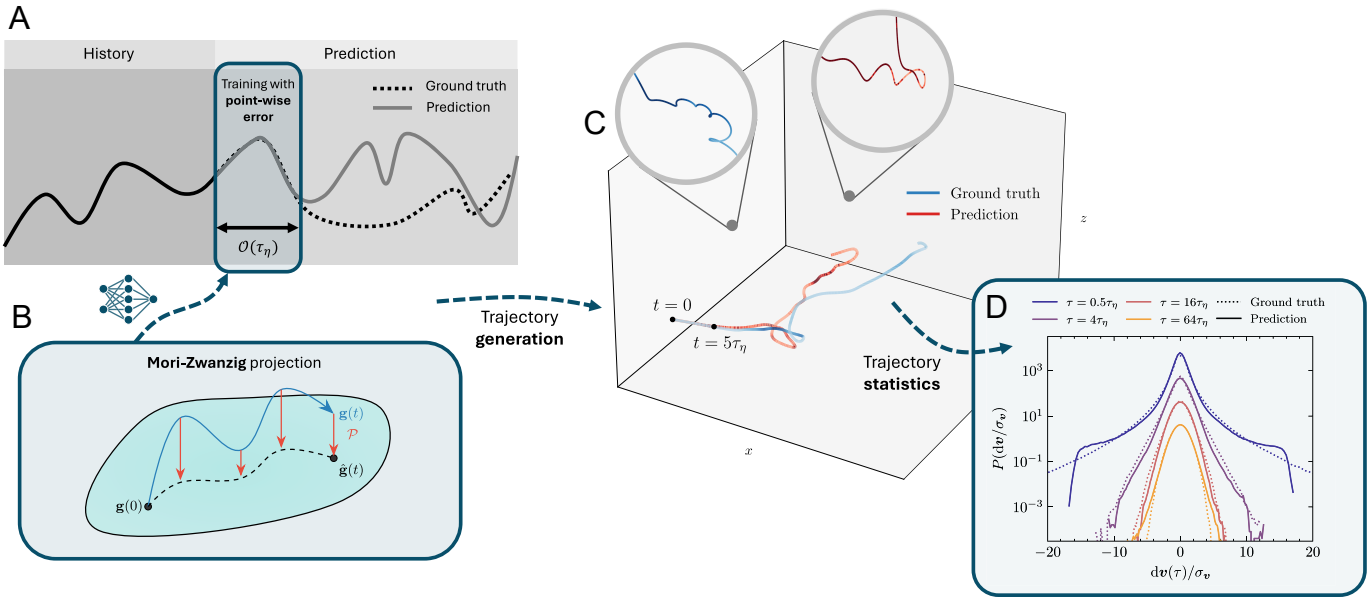

Fig. 1. (*A*) Schematic depiction of the learning task for the Mori–Zwanzig (MZ) surrogate model: Feeding a limited ground-truth history, the MZ model aims to predict the continuation of the trajectory in a way that is point-wise accurate at short times (w.r.t. Kolmogorov time τ_η) and statistically accurate at longer times. (*B*) Sketch of the data-driven MZ approach: We begin with the true evolution of the observables $\boldsymbol{g}(t)$ described by the generalized Langevin equation (GLE, Eq. **4**). Applying the projection operator $\mathcal{P}$ eliminates the orthogonal-dynamics contribution, so the resulting GLE is formulated solely in the projected subspace $\hat{\boldsymbol{g}} := \mathcal{P}\boldsymbol{g}$. (*C*) Exemplary 3D Lagrangian trajectories, comparing ground-truth data (blue) with its corresponding MZ prediction (red). Trajectories are colored by the local vorticity magnitude (darker color means larger vorticity). One can appreciate the development of intermittent vortex-filament structures in certain trajectories (*Insets*). (*D*) PDFs of Lagrangian velocity differences, normalized by their SD, for different time lags τ. PDFs are vertically shifted for readability.

turbulence. Many of such approaches rely on stochastic models in which, following a Langevin-like approach, particle velocity and/or acceleration are modeled as a stochastic process, constrained a priori to comply with certain turbulence statistics (11). Extensions based on multifractal models further refine these approaches by embedding intermittency effects (12–16) or multiparticle correlations (17). Other notable alternatives are given by kinematic simulations (KS) (18, 19), which generate synthetic velocity fields as a superposition of random Fourier modes with prescribed energy spectra, and shell models, which provide a reduced-order representation of turbulent energy transfer in Fourier space, which can in turn be translated to the Lagrangian viewpoint (20).

Although these modeling approaches can capture certain statistical features of turbulence, they often fail to reproduce the full complexity of Lagrangian turbulent statistics as observed in DNS. This has motivated the development of reduced-order models based on data-driven approaches (21, 22) in recent years, enabled by the increasing availability of high-resolution data from DNS (23–25) and experimental measurements, such as those obtained through interpolated Particle Image Velocimetry (PIV) (26–28). Notable examples include diffusion models, which have shown remarkable success in generating synthetic trajectories while preserving key statistical properties, including accurate multiscale properties beyond the restricted range where pure power laws are observed (29, 30). Such data-driven approaches thereby outperform traditional approaches based on stochastic models in terms of fidelity of the statistics. However, existing diffusion-based approaches still face important limitations, particularly because of the absence of an explicit underlying dynamical evolution. Specifically, diffusion models—being probabilistic generative models—synthesize entire trajectories through a stochastic sampling procedure, rather than learning the time-evolution equations that govern the underlying stochastic process. As a result, while these models are effective for trajectory synthesis, their applicability to problems requiring a true (surrogate) dynamical system that can be used in an autoregressive way remains an open challenge.

In this work, we utilize a mathematical formalism based on nonequilibrium statistical mechanics and dynamical systems theory for the data-driven discovery of an autoregressive reduced Lagrangian model for turbulence. The method introduced here is capable of learning a surrogate dynamical system which, provided with a short time history of the system, allows for accurate point-wise predictions within time scales comparable to the Kolmogorov scale τ_η, and remains stable and statistically accurate on longer time scales (cf. Fig. 1*A*). The backbone of the method relies on a unified data-driven approach combining the Mori–Zwanzig formalism (31, 32) and Takens' time-delay embedding (33), two established memory embeddings for the systematic development of closure modeling.

Crucially, we show that training the model over relatively short time intervals, on the order of the Kolmogorov time scale, and using a simple point-wise loss function such as the mean squared error, suffices to capture the underlying physical mechanisms driving the dynamics. Then, because the dynamics is correctly learned, at test time, the model is able to accurately recover the long-time statistics of turbulent Lagrangian accelerations and velocity gradients, including their probability distributions and (auto-)correlation structures, without incorporating any explicit statistical constraints in the training process.

We further highlight the robustness of the approach in addressing a common limitation of autoregressive, data-driven models: long-term stability. Despite being trained only on short trajectories, the model produces statistically accurate and stable predictions over significantly longer time horizons, even when initialized with highly out-of-distribution unphysical states (e.g., zero velocity perturbed by random noise).

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

Training Dataset. We train the model on a comprehensive dataset of Lagrangian trajectories recently introduced in ref. 34. The data were obtained from direct numerical simulations of the three-dimensional incompressible Navier–Stokes equations under conditions of statistically stationary, homogeneous, and isotropic turbulence. The simulations were performed in a triply periodic cubic domain of size $L = 2\pi$, discretized on a $1{,}024^3$ grid employing a standard pseudospectral method with full dealiasing (35). The database provides high-resolution time-resolved measurements of particle positions, velocities, and accelerations, along with the velocity gradient tensor evaluated at the particle locations. Data are sampled over a total time span of $T = 195\,\tau_\eta$, with a temporal resolution of $\Delta t = 0.1\,\tau_\eta$, where τ_η denotes the Kolmogorov timescale, which is the fastest dynamical timescale in turbulence. Our proposed model used the same timestep in order to resolve all dynamically active timescales. The corresponding Taylor-scale Reynolds number is $\mathrm{Re}_\lambda \approx 310$, with estimated Lagrangian integral timescale $T_L \approx 310\,\tau_\eta$.

The full dataset is split into disjoint sets: a training set (65%) used for optimization of the model weights, a validation set (28%) used for early stopping to avoid overfitting at training time and a test set (7%) that is exclusively used for evaluation of the model after training, as laid out in *Model Evaluation*. Trajectories in the training set and validation set are furthermore split in time disjointly, since only short roll-outs need to be performed at training time. This removes any potential explicit temporal bias in the training set.

Model Architecture and Training. The model is based on a dynamical equation with a finite memory that is used to account for the unresolved dynamics. It evolves the particle acceleration $\boldsymbol{a}(t)$ and local velocity gradients $\boldsymbol{\nabla u}(t)$ at the position of the particle, while the particle velocity $\boldsymbol{v}(t)$ and position $\boldsymbol{x}(t)$ are recovered through integration. The autoregressive kernels are parameterized with fully connected neural networks that are regressed on the training data.

Our method follows the principled approaches to extract Markov and memory kernels in the Mori–Zwanzig formalism, established in refs. 36 and 37. In short, training the Mori–Zwanzig model with time-delay embeddings is done in a cascaded fashion. The first MZ kernel (i.e. the Markov term) is fitted by minimizing the error after taking in the current state and its time-delay embeddings and predicting the next time step. Then, keeping the first MZ kernel fixed, the residual error is minimized by fitting the second MZ kernel, which takes in the state that is time-shifted into the history by one step. This process repeats for the full MZ memory length. Essential to the MZ formalism is that this recursive procedure is based on the satisfaction of the Dissipation-Fluctuation Theorem which relates the two-time correlation of the residual to the memory kernels.

The complete details on the derivation of the model and its architecture, the choice of hyperparameters and the training procedure are provided in *Materials and Methods*.

Results

Model Evaluation. To evaluate the performance of the surrogate model, we begin by assessing the convergence of the model predictive accuracy as a function of the size of the training set. Fig. 2 shows the PDF of particle acceleration, comparing the results obtained when training over 50K, 500K, and 5M trajectories. As evident from the figure, increasing the amount of training data leads to systematic improvements in the quality of the predictions, particularly in the accurate reconstruction of the heavy tails of the PDF, which are associated with extreme acceleration events. Based on this analysis, all subsequent results are reported for the model trained on the largest dataset of 5M trajectories.

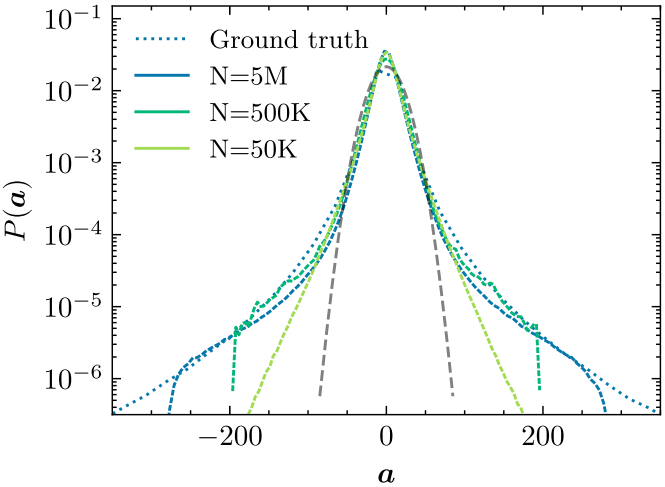

Fig. 2. Convergence of the accuracy of the predicted PDF of particle acceleration with increasing number of trajectories in the training set from 50K to 5M. The dashed line indicates a Gaussian distribution with the same variance.

We then evaluate the statistical performance of the model on key observables in the context of the dynamics of Lagrangian particles in turbulent flows. As highlighted in the introduction, the PDF of particle accelerations exhibits some of the most extreme fluctuations found in turbulence. In Fig. 3*A*, we compare the reference PDF computed over the entire dataset of ref. 34 with the long-term predictions on $\mathcal{O}(10^2\tau_\eta)$ generated by our model. The model faithfully reproduces the heavy tails of the distribution, indicating its ability to capture the intermittent and strongly nonlinear dynamics underlying particle acceleration. We find that the second moment agrees within 2% and the fourth moment agrees within 14%.

In Fig. 3*D*, we take into consideration the particle acceleration autocorrelation function, formally defined as

$$\mathrm{ACF}_{\boldsymbol{a}} = \frac{\langle \boldsymbol{a}(t_0 + t) \cdot \boldsymbol{a}(t_0) \rangle}{\langle |\boldsymbol{a}(t_0)|^2 \rangle} \qquad [1]$$

with $\langle \ldots \rangle$ denoting an ensemble average over all particles and over time origin t_0. The results are in satisfactory agreement with the ground-truth, indicating that the statistics of the temporal structure of the Lagrangian acceleration are properly captured by the surrogate model. A similar analysis can be performed for the predicted velocity gradient, and is reported in *SI Appendix*, section C.

To investigate the importance of memory, we compare the prediction of our model with Mori–Zwanzig memory and time-delay embedding to a model that has no memory in Fig 3 *A* and *D*. To ensure a fair comparison, the memoryless model has the same number of weights, the same architecture, and the same number of kernels, except that all kernels are purely Markovian and the state is not enhanced with time-delay embeddings. The results show that such a memoryless model falls short in terms of fidelity of the resulting distribution and particularly cannot capture the correct temporal correlation structure of the dynamics. In *SI Appendix*, section D, we show that even when

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

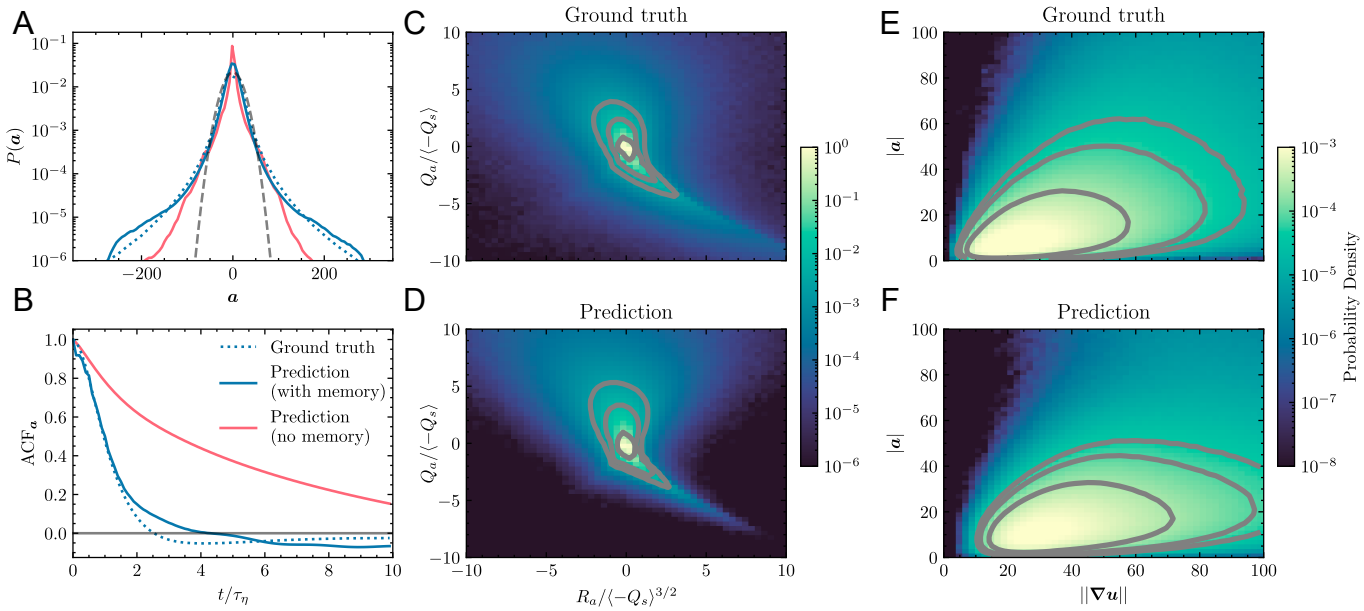

Fig. 3. Summary statistics of the long time prediction of the MZ model, compared with the ground truth. Provided are the acceleration PDF (*A*), its autocorrelation (*B*), the Q-R diagram of the velocity gradient at the position of the particle (*C* and *D*) and the joint PDF of the magnitude of acceleration and velocity gradient (*E* and *F*). The dashed line in (*A*) indicates a Gaussian distribution with the same variance. The lines in (*C*–*F*) correspond to level contours of 70%, 90%, and 95% probability percentiles.

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

further increasing the network size of the memoryless model, it cannot get appreciably closer to the performance of the memory-based approach. This underpins the principal importance of including memory in order to capture the effect of the unresolved dynamics in the reduced order model.

In order to evaluate the quality of the predictions for the fluid velocity gradient, in Fig. 3 *B* and *E* we show the so-called Q-R plot, a diagnostic tool commonly used to study the local topology of turbulence, in particular the relationship between the rate of strain and vorticity. For this analysis, we start by introducing a decomposition of the velocity gradient tensor $\nabla \boldsymbol{u} = \boldsymbol{S} + \boldsymbol{A}$ into its symmetric part $\boldsymbol{S}$, the strain-rate tensor, and its antisymmetric part $\boldsymbol{A}$, the rotation-rate tensor, with

$$\boldsymbol{S} = \tfrac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T), \quad \boldsymbol{A} = \tfrac{1}{2}(\nabla \boldsymbol{u} - (\nabla \boldsymbol{u})^T). \qquad [2]$$

Based on this decomposition, we introduce the quantities

$$Q_s = -\tfrac{1}{2}\mathrm{Tr}[\boldsymbol{S}^2], \quad Q_a = -\tfrac{1}{2}\mathrm{Tr}[\boldsymbol{A}^2], \quad R_a = -\tfrac{1}{3}\mathrm{Tr}[\boldsymbol{A}^3]. \qquad [3]$$

Here, Q_s represents the local rate of strain intensity, Q_a measures the local rotation intensity, and R_a characterizes the third-order moment of the rotational component. In the figure, the quantities Q_a and R_a are normalized by suitable powers of the average strain contribution $\langle -Q_s \rangle$ (since Q_s is negative definite in incompressible HIT), averaging over different time stamps and over particles, such to define nondimensional quantities, and plotted as joint PDF. We observe a good qualitative agreement with ground-truth data, with both figures highlighting the characteristic teardrop shape, reflecting the competition between vortical and straining motions. This offers a convincing validation for the physical fidelity of the predicted velocity gradients. Additionally, we verify whether the correlation between acceleration and velocity gradient is properly captured. To that extent, in Fig. 3 *C* and *F* we analyze joint PDF of the magnitude of acceleration $|\boldsymbol{a}|$ and of the velocity gradient $||\nabla \boldsymbol{u}||$. The figure shows good qualitative agreement between the joint PDF in the prediction and the ground truth. Indeed, events of high velocity gradient are positively correlated with high acceleration, indicating that when the particle encounters e.g. a region of high vorticity, it will experience a large acceleration and vice-versa, which is well resolved by the surrogate model.

The accurate reproduction of the heavy-tailed acceleration distribution, the proper time-correlation structure, and the correct velocity gradient statistics provide convincing evidence that the surrogate model is capable of faithfully capturing the complex statistical properties of Lagrangian turbulence, and most notably, its intermittency. As a consequence, at the more qualitative level, in examples of the generated trajectories, one can observe instances that resemble encounters with vortex filaments—a feature often associated with turbulence intermittency (see example in Fig. 1*C*).

In Fig. 4, we investigate the temporal robustness of the model predictions. The *Upper* panel shows the MSE computed at fixed time steps, with time on the horizontal axis scaled by τ_η. Predictions are compared against the ground truth, where the model is supplied with memory terms extracted from the corresponding ground truth trajectories. Initially, the MSE grows gradually up to approximately one Kolmogorov time, consistent with the design of the loss function used for training (Eq. **17**). Beyond this range, the error increases before eventually saturating and fluctuating around a steady value at large times. Indeed, owing to the chaotic nature of turbulence, it is not feasible to achieve point-wise agreement between predicted and true trajectories much beyond the Kolmogorov time scale.

In the *Lower* panel, we present the PDF of the acceleration evaluated over three selected time windows. As anticipated from the MSE behavior, the PDF computed over the interval $(0, 1)\tau_\eta$ shows the best agreement with the ground truth, as all trajectories start from the ground truth statistics that result from the true turbulent attractor. Crucially, the figure shows that statistically, although the model was trained only on short trajectories within $(0, 0.5)\tau_\eta$, it continues to yield accurate statistics even at moderate, $(12, 13)\tau_\eta$, and long, $(130, 131)\tau_\eta$, time scales.

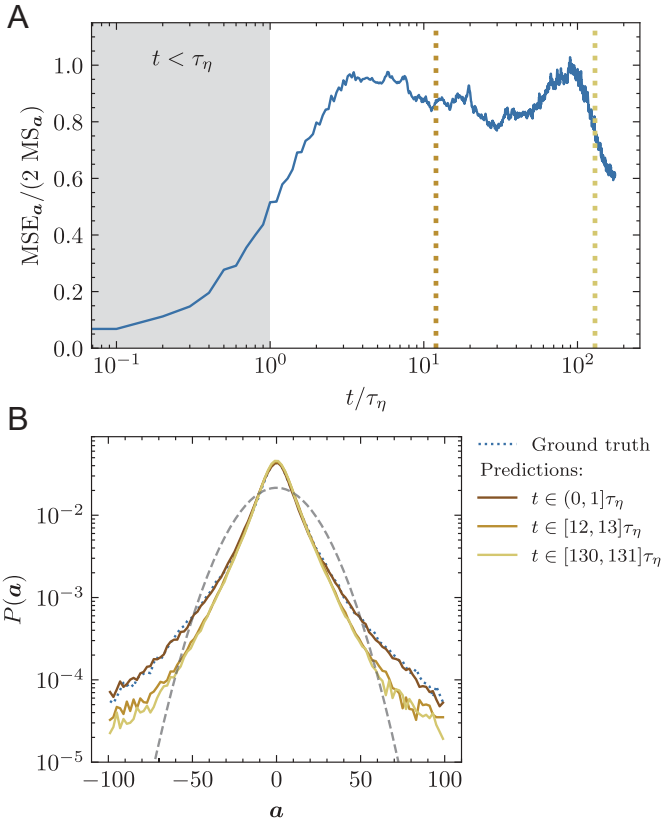

Fig. 4. Temporal robustness of the MZ model. (*A*) the mean-squared error (MSE) of acceleration, normalized by twice the mean square (MS) acceleration, as a function of time. (*B*) PDFs of acceleration computed over different time windows. The dashed line indicates a Gaussian distribution with the same variance.

This result indicates that the model successfully captures the essential physical mechanisms governing the system dynamics, despite being trained on a short time horizon. It shows that as the dynamics of the system evolves, it is able to keep the trajectories close to the true turbulent attractor, thereby producing the correct Lagrangian statistics. This confirms that the model achieves high accuracy on the point-wise prediction over short time scales, of the order of the Kolmogorov time τ_η, while sustaining statistically reliable behavior at longer times.

Generating Trajectories from Noise. So far, all the analyses presented have assumed that the initial conditions, as well as the memory terms used to initialize the dynamics, are drawn from physically consistent data extracted from the reference dataset. We now consider a more challenging test case in which the system is initialized from a fluid at rest, with only slight perturbations introduced in the initial values of acceleration and velocity gradients to statistically separate individual particles.

In Fig. 5*A*, we plot the Kullback–Leibler (KL) divergence between the model and the ground truth, evaluated at different time steps and expressed in terms of t/t_η. We observe that after a relatively short transient phase, lasting approximately $\mathcal{O}(10\tau_\eta)$, the KL divergence drops to values very close to zero. This is also reflected in the results shown in Fig. 5*B*, where we plot the PDF of acceleration over three different time windows, corresponding to early-stage, intermediate, and long-time dynamics.

Remarkably, after the initial transient phase, and despite the out-of-distribution starting conditions, the model successfully drives the dynamics onto the correct attractor that reproduces the turbulent statistics. This behavior provides strong further evidence that the model has learned the correct underlying physical mechanisms governing the system dynamics. From a practical point of view, this opens up the possibility to generate statistically adequate Lagrangian trajectories even without supplying ground truth initial conditions.

Discussion

A central challenge in modeling dynamical systems from a Lagrangian perspective lies in accurately describing the evolution of tracer particles when only a limited number of degrees of freedom are accessible, as opposed to fully resolved dynamics provided by DNS. Such reduced representations are common in real-world scenarios; for example, when tracking pollutants transported by environmental flows, it is often only feasible to track a sparse set of particles, hence with limited information about the entire flow field. This incomplete observation complicates model fidelity, requiring careful application of embedding and closure techniques to reconstruct or infer the unobserved dynamics and enable reliable predictions.

In this work, we have demonstrated that coupling the data-driven Mori–Zwanzig formalism with time-delay embedding yields a powerful framework for learning the complex time dynamics of Lagrangian particle trajectories in turbulent flows. This technique results in a general approach with enhanced expressive capabilities, allowing the model to reconstruct the influence of unresolved degrees of freedom from sequences of past observations. In doing so, the model is able to infer the underlying phase-space structure of the dynamics without requiring explicit access to all variables of the full-order system.

The resulting reduced-order model leverages intrinsic memory of the system to accurately capture multiscale phenomena. We show that by incorporating these memory effects and training the model over many trajectories with a short-time horizon—on the order of the Kolmogorov time scale—it is possible to capture the correct physical mechanisms driving the dynamics.

As a result, the model achieves statistically accurate predictions over significantly longer timescales. This is demonstrated through comparisons of the PDF of acceleration and the autocorrelation of acceleration against ground truth data from DNS. We also validate the model performance by examining the correlations with local velocity gradients. Furthermore, the model exhibits robust behavior even when initialized from out-of-distribution conditions, such as zero velocity perturbed by noise, highlighting its ability to self-correct and reconstruct realistic trajectories.

We point out that, owing to the universality of small-scale turbulence, this model also holds potential to be used for turbulent flows away from the idealized setting of homogeneous isotropic turbulence. The advection due to resolved mean flows could then be superposed onto the turbulent fluctuations provided by the model in this work to yield an effective Lagrangian large-eddy simulation. In future work, we plan to evaluate the capability of the model for training at higher Reynolds number, extending to larger datasets, and further expanding the generalizability of our model.

This study opens up several promising avenues for future applications. One particularly relevant scenario involves the real-time control of active particles in turbulent environments (38, 39). In such contexts, implementing model predictive control requires rapid and repeated evaluation of the system response. Running

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

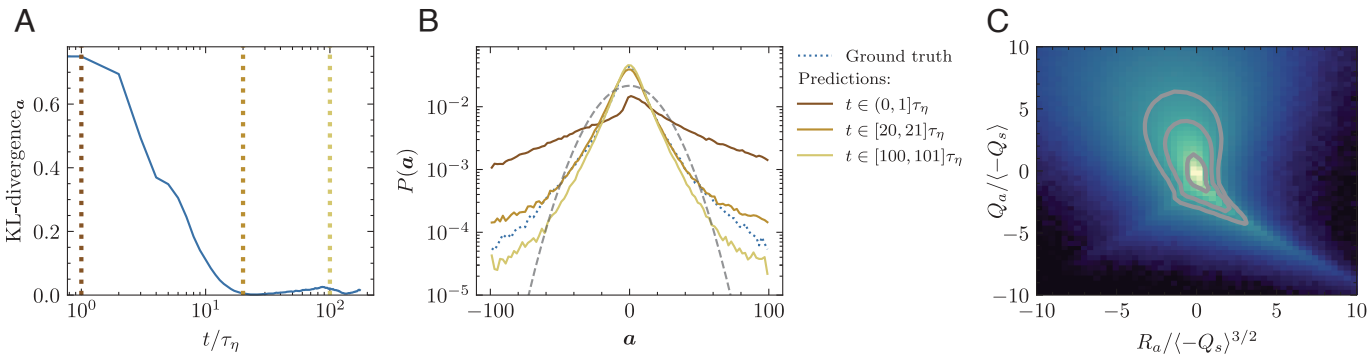

Fig. 5. Model predictions when starting from statistically perturbed zero initial conditions. (*A*) KL-divergence $D_{\mathrm{KL}}(\text{ground-truth}\,||\,\text{model})$ based on the distribution of particle acceleration as a function of time throughout the predicted trajectories. (*B*) corresponding PDFs of acceleration for selected time windows. The dashed line indicates a Gaussian distribution with the same variance. (*C*) Q-R diagram of the predicted velocity gradient at the particle position, averaged from $t > 20\tau_\eta$.

fully resolved flow simulations in real time is computationally prohibitive, whereas evaluating a reliable surrogate offers a practical and efficient alternative. By capturing the essential dynamics with significantly reduced computational cost, data-driven reduced-order models could enable real-time feedback control and decision-making in complex, unsteady flow environments.

Materials and Methods

The Mori–Zwanzig Formalism. For completeness, and to fix notation, we provide a summary of the Mori–Zwanzig formalism and the data-driven method used in this work. Originally introduced in statistical mechanics nearly sixty years ago (31, 32), the Mori–Zwanzig (MZ) formalism provides a rigorous mathematical framework for constructing non-Markovian reduced-order models, capturing the influence of unresolved variables through a memory kernel and orthogonal dynamics (31, 32). The resulting reduced-order model, known as the Generalized Langevin Equation (GLE), is an integro-differential equation that governs the dynamics of selected observables (resolved variables) and includes a Markovian term, a memory-dependent term, and an orthogonal dynamics (or noise) term (40). This projection operator formalism, and the resulting GLE, can be found in a variety of applications, including fluid mechanics (41, 42) and molecular dynamics (43, 44). The memory term quantifies the interactions between resolved and unresolved dynamics. This memory effect is highly sensitive to both the choice of observables and the projection operator, and thus classical approaches (45, 46) to analytical derivation of the Mori–Zwanzig operators have been especially challenging.

Recent advances have led to the development of data-driven Mori–Zwanzig methodologies (47–49), extending approximate Koopmanian learning and demonstrating improved performance compared to traditional methods such as Dynamic Mode Decomposition (DMD), Extended DMD (EDMD), and Higher-Order DMD. Furthermore, by considering regression as the projection operator (36) $\mathcal{P}$, nonlinear (neural network) based models for the MZ operators have been formalized. These approaches have shown promising results in applications involving stationary homogeneous isotropic turbulence (50) and boundary layer transition phenomena (37, 48, 51) as well as transient flows (52).

We now provide a general outline of the MZ framework and the data-driven method following the work of ref. 36. The goal of the Mori–Zwanzig procedure is to construct evolutionary equations for a reduced set of observables $\boldsymbol{g} \in \mathbb{R}^M$, $M \ll N$, which can be any L^2-integrable function of the state, also referred to as the resolved variables, where the full state $\boldsymbol{p} \in \mathbb{R}^N$ is intractable to resolve. The Mori–Zwanzig formalism then proceeds with a specified projection operator $\mathcal{P}$, which maps a function of the full-space configuration, $\boldsymbol{g} : \mathbb{R}^N \to \mathbb{R}^M$, to a function of only the resolved observables $\mathcal{P}\boldsymbol{g} : \mathbb{R}^M \to \mathbb{R}^M$. An operator algebraic derivation results in the generalized Langevin equation (GLE) (32):

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(t) = \boldsymbol{M}(\boldsymbol{g}(t)) - \int_0^t \boldsymbol{K}(t-s, \boldsymbol{g}(s))\mathrm{d}s + \boldsymbol{F}(t), \qquad [4]$$

which governs the time evolution of the resolved observables $\boldsymbol{g}(t)$ through the Markovian term $\boldsymbol{M}$ and memory term $\boldsymbol{K}$. The GLE is a closed and exact integro-differential equation for the resolved variables, although it is now nonlocal in time and contains an orthogonal dynamics term $\boldsymbol{F}$ (often referred to as noise term) that depends on time and the initial full state.

The GLE can then be written in discrete-time (53–56), which prescribes the exact and closed set of non-Markovian evolutionary equations for the observables given any initial condition of states as

$$\boldsymbol{g}_{n+1} = \boldsymbol{\Omega}_0(\boldsymbol{g}_n) + \sum_{k=1}^{n} \boldsymbol{\Omega}_k(\boldsymbol{g}_{n-k}) + \boldsymbol{W}_n. \qquad [5]$$

The discrete time GLE Eq. **5** states that the vector of observables $\boldsymbol{g}_{n+1}$ at time $n+1$ evolves according to three parts: 1) a *Markovian* operator: $\boldsymbol{\Omega}_0 : \mathbb{R}^M \to \mathbb{R}^M$ which only depends on the observables at the previous time step n, 2) the *memory kernel*: the series of operators $\boldsymbol{\Omega}_k : \mathbb{R}^M \to \mathbb{R}^M$ depending on observables with a time lag k, and 3) the *orthogonal dynamics*: $\boldsymbol{W}_n : \mathbb{R}^N \to \mathbb{R}^M$ depending on the full (initial) state. The above GLE is general for any projection operator (see ref. 36 for a detailed discussion).

The operators $\boldsymbol{\Omega}_k$ and $\boldsymbol{W}_n$ depend on the choice of the projection operator $\mathcal{P}$, the choice of observables $\boldsymbol{g}$, and the finite-time (Δ) Koopman operator $\mathcal{K}$ (53, 57):

$$\boldsymbol{\Omega}_n := \mathcal{P}\mathcal{K}[(1-\mathcal{P})\mathcal{K}]^n, \qquad [6a]$$

$$\boldsymbol{W}_n := [(1-\mathcal{P})\mathcal{K}]^{n+1}\boldsymbol{g}. \qquad [6b]$$

An essential aspect of the Mori–Zwanzig formalism, often overlooked in modeling papers, is that the memory kernel and orthogonal dynamics are not independent. This general relationship between the memory kernel $\boldsymbol{\Omega}$ and the two-time correlation of orthogonal dynamics $\boldsymbol{W}$, commonly known as the generalized fluctuation–dissipation (GFD) relationship (47, 53, 58) can be expressed in the discrete form as

$$\boldsymbol{\Omega}_n\boldsymbol{g} = \mathcal{P}\mathcal{K}\boldsymbol{W}_{n-1}, \quad \forall n \in \mathbb{N}, \qquad [7]$$

where $\mathcal{K}$ is the discrete time Koopman operator. This can be derived directly from Eq. **6**. The approximation, ultimately, then lies in neglecting the orthogonal dynamics $\boldsymbol{W}_n$ in Eq. **5**, yielding the so-called truncated GLE. This ensures that the dynamics becomes a closed system that depends solely on the resolved variables $\boldsymbol{g}$ themselves, as enabled through the projection operator.

In summary, there are three key elements that one should aim to achieve with a data-driven MZ procedure: 1) learn the operators that satisfy the GLE, 2) use a well-defined projection operator, and 3) satisfy the GFD.

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

By extending the Dyson identity we utilize a discrete-time GLE that accommodates configurations sampled at irregular intervals of time. The Dyson identity is used in the derivation of the GLE. It is an operator identity for splitting the exponential of a sum of operators. For two operators A and B, it states

$$e^{t(A+B)} = e^{tA} + \int_0^t d\tau\, e^{(t-\tau)(A+B)}\, B\, e^{\tau A}. \quad [8]$$

In the context of the Mori–Zwanzig, one decomposes the Liouville operator $\mathcal{L}$ as

$$\mathcal{L} = \mathcal{Q}\mathcal{L} + \mathcal{P}\mathcal{L}, \quad [9]$$

where $\mathcal{Q} = 1 - \mathcal{P}$. The Dyson identity then takes the form

$$e^{t\mathcal{L}} = e^{t\mathcal{Q}\mathcal{L}} + \int_0^t d\tau\, e^{(t-\tau)\mathcal{L}}\, \mathcal{P}\mathcal{L}\, e^{\tau\mathcal{Q}\mathcal{L}}. \quad [10]$$

In discrete form, the Dyson identity becomes (53, 54):

$$\mathcal{K}^{n+1} = \sum_{k=0}^{n} \mathcal{K}^{n-k}\mathcal{P}\mathcal{K}(\mathcal{Q}\mathcal{K})^k + (\mathcal{Q}\mathcal{K})^{n+1}, \quad [11]$$

where $\mathcal{K}$ is the discrete time Koopman operator $\mathcal{K} \equiv e^{\Delta t\mathcal{L}}$. Specifically, the projection operators associated with each sampling instance can differ, unlike conventional MZ treatments that assume a single fixed projector. This generalization enables a data-driven construction with variable time steps.

Several data-driven methods have been proposed for estimating the MZ operators (36, 53, 57, 59–62). In the data-driven MZ method (36) used in this work, we assume a finite length memory h. The MZ operators $\mathbf{\Omega}_k$ (for $k \in \{0, 1, \ldots, h\}$) in the truncated GLE are learned with Neural Networks, enforcing the GFD relation by following ref. 36, and using regression as the projection operator. Additionally, we take advantage of the variable projection method in order to explore a GLE with variable time steps. For more details on the derivation and algorithms, the reader is referred to ref. 36.

Time-Delay Embedding. To further increase the expressiveness of the model, the MZ operators can be extended by incorporating additional time-delay embeddings. In this formulation, every memory kernel acts not only on the resolved state at a given time but also on an augmented state vector that includes a finite sequence of past states, referred to as time-delay embeddings. The theoretical foundation for this approach was rigorously formalized by Takens' embedding theorem (33).

Since MZ memory provides a general term to account for how the unresolved variables interact with the resolved observables, MZ memory and delay embeddings are not mutually exclusive. Integrating time-delay embedding with the Mori-Zwanzig formalism offers a natural and systematic strategy to address unresolved scales or variables, explicitly introducing memory effects while simultaneously enlarging the functional space for regression. For example, Lin et al. (36) showed that their combination can produce optimal predictive performance. Including MZ memory can thus alleviate the need to determine the embedding dimension of the strange attractor assumed by Takens' theorem. Consequently, this combined framework can produce more accurate, robust, and interpretable reduced-order models.

Data-Driven Mori–Zwanzig for Lagrangian Turbulence. We now introduce the data-driven Mori–Zwanzig formalism augmented with time-delay embedding in the context of Lagrangian turbulence. The resulting reduced-order model explicitly governs the temporal evolution of particle acceleration $\boldsymbol{a}$; where particle position and velocity are directly obtained via temporal integration. Furthermore, we include local fluid velocity gradients $\boldsymbol{\nabla u}$ at the position of the particle in the model to encode the knowledge of local small-scale structures of turbulent flow (39), such as strain and vorticity, which strongly influence particle dynamics. Formally, the model is then defined as

$$\begin{pmatrix} \boldsymbol{a}_{n+1} \\ \boldsymbol{\nabla u}_{n+1} \end{pmatrix} = \mathrm{MZ}_h(\boldsymbol{a}_{n,\ldots,n-h}, \boldsymbol{\nabla u}_{n,\ldots,n-h}), \quad [12a]$$

$$\boldsymbol{v}_{n+1} = \boldsymbol{v}_n + \boldsymbol{a}_n \Delta t, \quad [12b]$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{v}_n \Delta t, \quad [12c]$$

where the right-hand side of Eq. **12a** consists of the superposition between a Markovian kernel and a finite number h of memory kernels as

$$\mathrm{MZ}_h(\boldsymbol{a}_{n,\ldots,n-h}, \boldsymbol{\nabla u}_{n,\ldots,n-h}) = \overbrace{\mathbf{\Omega}_0(\boldsymbol{a}_n, \boldsymbol{\nabla u}_n)}^{\text{Markovian kernel}} + \sum_{k=1}^{h} \underbrace{\mathbf{\Omega}_k(\boldsymbol{a}_{n-k}, \boldsymbol{\nabla u}_{n-k})}_{\text{memory kernels}}. \quad [13]$$

The next key ingredient is the data-driven learning of the MZ operators $\mathbf{\Omega}_k$. Following the approach described in ref. 36, we define the projection operator in the context of regression. Specifically, the method consists of incrementally minimizing the orthogonal dynamics (i.e., residual noise) by explicitly enforcing the generalized fluctuation-dissipation relation, Eq. **7**, a necessary condition for consistency with the MZ formalism. By interpreting regression as the process of forming the projection operator, and leveraging the property that the projection $\mathcal{P}$ annihilates the noise term ($\mathcal{P}\boldsymbol{W} = 0$), we can express the learned model directly in terms of projected quantities, without explicitly modeling the noise. However, in this case, it is not exactly closed, but only approximately so by the memory kernels.

In practice, we perform the regression through deep learning, where we learn the kernels $\mathbf{\Omega}_k$, parameterized as fully connected multilayer perceptron (MLP) neural network (NN). Before going into details on the training procedure, we consider a generalization of Eq. **12a** where the MZ paradigm is coupled with d Takens' time-delay embeddings. The resulting formulation is then:

$$\mathrm{MZ}_h(\boldsymbol{a}_{n,\ldots,n-\kappa(h)-d}, \boldsymbol{\nabla u}_{n,\ldots,n-\kappa(h)-d}) = \Omega_0(\boldsymbol{a}_{n,\ldots,n-d}, \boldsymbol{\nabla u}_{n,\ldots,n-d}) + \sum_{k=1}^{h} \Omega_k(\boldsymbol{a}_{n-\kappa(k),\ldots,n-\kappa(k)-d}, \boldsymbol{\nabla u}_{n-\kappa(k),\ldots,n-\kappa(k)-d}). \quad [14]$$

In the above, we also allow for nonlinearly time-spaced memory kernels, via the mapping $\kappa(k)$.

In the DNS, the mean flow is typically explicitly enforced to be zero, which ensures that both Eulerian and Lagrangian velocities always remain statistically bound with some spread around a zero mean. However, this zero mean condition is not inherently enforced in the surrogate model, leading to the possibility of unbounded, diffusively growing velocities. To enforce a zero mean condition, we thus apply a small linear damping based on the Lagrangian velocity to the predicted acceleration, replacing Eq. **12a** by

$$\begin{pmatrix} \boldsymbol{a}_{n+1} \\ \boldsymbol{\nabla u}_{n+1} \end{pmatrix} = \mathrm{MZ}_h(\boldsymbol{a}_{n,\ldots,n-h}, \boldsymbol{\nabla u}_{n,\ldots,n-h}) - \begin{pmatrix} \alpha \boldsymbol{v}_n \\ \boldsymbol{0} \end{pmatrix}. \quad [15]$$

The damping factor α should be tuned in such a way that its action remains restricted to the longest time scales. In practice this means that the damping time scale α^{-1} should be on the order of the integral timescale of turbulence to ensure that the short time and inertial range dynamics remain unaffected. The numerical analysis of the effect of the damping term is provided in *SI Appendix*, section B.

Table 1. Hyperparameters used for the MZ model

Number of hidden layers	9
Neurons per layer	768
Model size	5M
Total number of weights	5×10^6
Number of epochs	500
Unroll horizon L_T	5
MZ history h	5
Delay embeddings d	40
Time-step Δt	$0.1\tau_\eta$
Damping α	0.5

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

Memory in Lagrangian Turbulence. Temporal correlations in Lagrangian trajectories in turbulence can in principle extend from the (sub-)Kolmogorov scales all the way to the integral time scales. This offers a practical challenge for modeling high-Reynolds flows, as one needs to take correlations into account over a large dynamic range of scales. We therefore opt for a logarithmic spacing of memory terms $\kappa(k) = \lambda^{k-1}$, where we choose $\lambda = 2$, allowing us to account for correlation over a large range of time-scales while limiting the number of memory terms. Since the accelerations that we are modeling are related to the velocity gradient field, they are dominated by contributions from the small scales. Hence, to maximize the information from the short-time correlations, we keep the Takens' time-delay embedding linearly spaced, maintaining a high resolution of the short-time history, on the order of the Kolmogorov scale.

Hyperparameters and Loss Function. In order to define the training procedure for establishing the kernels $\mathbf{\Omega}_k$ discussed in the previous section, we introduce the loss function, defined as the sum of the mean-squared error computed on the prediction of particle acceleration and fluid velocity gradient:

$$\mathcal{L} = \mathrm{MSE}_{\boldsymbol{a}} + \mathrm{MSE}_{\boldsymbol{\nabla u}}. \qquad [16]$$

We employ here a technique known in the literature as *unrolled training* or *trajectory learning* (63), which consists of applying the model to make predictions up to L_T consecutive time steps at training time, backpropagating the gradients through all (or just a few of) such time steps. We weight the contributions of each of these time steps in the loss, giving more importance to shorter-term predictions, following the same approach as put forward in ref. 64. We thus train the model on the following error metric, with MSE defined for a generic observable $\boldsymbol{g}$ as

$$\mathrm{MSE}_g = \sum_{i=1}^{L_T} w_i \left(\tilde{\boldsymbol{g}}_i - \boldsymbol{g}_i\right)^2, \qquad [17]$$

where $\tilde{\boldsymbol{g}}_i$ is the prediction of the model, and weights w_i decay exponentially with rate factor γ as $w_i = \exp(-\gamma i/L_t)$. The unroll horizon L_t should be chosen on the order of the Kolmogorov time, such that one can expect to be able to make predictions that are point-wise accurate on this time scale, minimizing the mean-squared error. Optimization is done with stochastic gradient descent using the Adam optimizer and with early stopping based on validation loss to avoid overfitting.

The full choice of all hyperparameters employed in this work is provided in Table 1. Moreover, in *SI Appendix*, section A, we report an analysis of the impact of memory lengths and unroll horizon parameters.

Data, Materials, and Software Availability. Previously published data were used for this work (arXiv:2303.08662) (34).

ACKNOWLEDGMENTS. This work was supported in part by the Netherlands Organization for Scientific Research (NWO) through the use of supercomputer facilities (Snellius) under Grant Nos. 2023.026 and 2025.008. This publication is part of the project "Shaping turbulence with smart particles" with Project No. OCENW.GROOT.2019.031 of the research program Open Competitie ENW XL which is (partly) financed by the Dutch Research Council (NWO). This work is coauthored by employees of Triad National Security, LLC which operates Los Alamos National Laboratory (LANL) under Contract No.89233218CNA000001 with the U.S. Department of Energy/National Nuclear Security Administration. A.G. gratefully acknowledges the support of the LANL/LDRD Program under project number 20240740PRD and the Center for Non-Linear Studies for this work. Y.T.L. and D.L. were supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research's Applied Mathematics Competitive Portfolios program. X.M.d.W., M.W., and D.L. also acknowledge the support from the LANL/LDRD Program under project number 20240318ER.

Author affiliations: [a]Department of Applied Physics and Science Education, Fluids and Flows group and J.M. Burgers Center for Fluid Mechanics, Eindhoven University of Technology, Eindhoven 5600 MB, Netherlands; [b]Computational Physics and Methods Group (CCS-2), Los Alamos National Laboratory, Los Alamos, NM 87545; [c]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545; [d]Department of Physics and Earth Science, University of Ferrara, INFN Ferrara, Ferrara 44121, Italy; [e]Information Science Group (CCS-3), Los Alamos National Laboratory, Los Alamos, NM 87545; and [f]Instituto per le Applicazioni del Calcolo, Consiglio Nazionale delle Ricerche, Rome I-00185, Italy

1. S. B. Pope, *Turbulent Flows* (Cambridge University Press, 2000).
2. P. K. Yeung, Lagrangian investigations of turbulence. *Annu. Rev. Fluid Mech.* **34**, 115–142 (2002).
3. G. Falkovich, A. Fouxon, M. Stepanov, Acceleration of rain initiation by cloud turbulence. *Nature* **419**, 151–154 (2002).
4. B. Sawford, Turbulent relative dispersion. *Annu. Rev. Fluid Mech.* **33**, 289–317 (2001).
5. J. Bec *et al.*, Lyapunov exponents of heavy particles in turbulence. *Phys. Fluids* **18**, 091702 (2006).
6. F. Toschi, E. Bodenschatz, Lagrangian properties of particles in turbulence. *Annu. Rev. Fluid Mech.* **41**, 375–404 (2009).
7. A. La Porta, G. A. Voth, A. M. Crawford, J. Alexander, E. Bodenschatz, Fluid particle accelerations in fully developed turbulence. *Nature* **409**, 1017–1019 (2001).
8. G. A. Voth, A. La Porta, A. M. Crawford, J. Alexander, E. Bodenschatz, Measurement of particle accelerations in fully developed turbulence. *J. Fluid Mech.* **469**, 121–160 (2002).
9. N. Mordant, A. M. Crawford, E. Bodenschatz, Experimental Lagrangian acceleration probability density function measurement. *Phys. D Nonlinear Phenom.* **193**, 245–251 (2004).
10. P. K. Yeung, S. B. Pope, An algorithm for tracking fluid particles in numerical simulations of homogeneous turbulence. *J. Comput. Phys.* **79**, 373–416 (1988).
11. B. L. Sawford, Reynolds number effects in Lagrangian stochastic models of turbulent dispersion. *Phys. Fluids A Fluid Dyn.* **3**, 1577–1586 (1991).
12. J. D. Wilson, B. L. Sawford, Review of Lagrangian stochastic models for trajectories in the turbulent atmosphere. *Bound. Layer Meteorol.* **78**, 191–210 (1996).
13. L. Biferale, G. Boffetta, A. Celani, A. Lanotte, F. Toschi, Particle trapping in three-dimensional fully developed turbulence. *Phys. Fluids* **17**, 021701 (2005).
14. A. G. Lamorgese, S. B. Pope, P. K. Yeung, B. L. Sawford, A conditionally cubic-Gaussian stochastic Lagrangian model for acceleration in isotropic turbulence. *J. Fluid Mech.* **582**, 423–448 (2007).
15. J. P. Minier, S. Chibbaro, S. B. Pope, Guidelines for the formulation of Lagrangian stochastic models for particle simulations of single-phase and dispersed two-phase turbulent flows. *Phys. Fluids* **26**, 113303 (2014).
16. B. Viggiano *et al.*, Modelling Lagrangian velocity and acceleration in turbulent flows as infinitely differentiable stochastic processes. *J. Fluid Mech.* **900**, A27 (2020).
17. I. M. Mazzitelli, F. Toschi, A. S. Lanotte, An accurate and efficient Lagrangian sub-grid model. *Phys. Fluids* **26**, 095101 (2014).
18. J. C. H. Fung, J. C. R. Hunt, N. A. Malik, R. Perkins, Kinematic simulation of homogeneous turbulence by unsteady random fourier modes. *J. Fluid Mech.* **236**, 281–318 (1992).
19. S. Murray, M. Lightstone, S. Tullis, Single-particle Lagrangian and structure statistics in kinematically simulated particle-laden turbulent flows. *Phys. Fluids* **28**, 033302 (2016).
20. L. Biferale, Shell models of energy cascade in turbulence. *Annu. Rev. Fluid Mech.* **35**, 441–468 (2003).
21. J. Friedrich, J. Peinke, A. Pumir, R. Grauer, Explicit construction of joint multipoint statistics in complex systems. *J. Phys. Complex.* **2**, 045006 (2021).
22. M. Sinhuber, J. Friedrich, R. Grauer, M. Wilczek, Multi-level stochastic refinement for complex time series and fields: A data-driven approach. *New J. Phys.* **23**, 063063 (2021).
23. Y. Kaneda, T. Ishihara, M. Yokokawa, K. Itakura, A. Uno, Energy dissipation rate and energy spectrum in high resolution direct numerical simulations of turbulence in a periodic box. *Phys. Fluids* **15**, L21–L24 (2003).
24. P. K. Yeung, X. M. Zhai, K. R. Sreenivasan, Extreme events in computational turbulence. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12633–12638 (2015).
25. P. K. Yeung, K. Ravikumar, Advancing understanding of turbulence through extreme-scale computation: Intermittency and simulations at large problem sizes. *Phys. Rev. Fluids* **5**, 110517 (2020).
26. D. Schanz, S. Gesemann, A. Schröder, Shake-the-box: Lagrangian particle tracking at high particle image densities. *Exp. Fluids* **57**, 1–27 (2016).
27. N. Machicoane, *Recent Developments in Particle Tracking Diagnostics for Turbulence Research* F. Toschi, M. Sega *et al.*, Eds. (Springer International Publishing, Cham, 2019), pp. 177–209.
28. A. Schröder, D. Schanz, 3D Lagrangian particle tracking in fluid mechanics. *Annu. Rev. Fluid Mech.* **55**, 511–540 (2023).
29. T. Li, L. Biferale, F. Bonaccorso, M. A. Scarpolini, M. Buzzicotti, Synthetic Lagrangian turbulence by generative diffusion models. *Nat. Mach. Intell.* **6**, 393–403 (2024).
30. T. Li, S. Tommasi, M. Buzzicotti, F. Bonaccorso, L. Biferale, Generative diffusion models for synthetic trajectories of heavy and light particles in turbulence. *Int. J. Multiph. Flow* **181**, 104980 (2024).
31. H. Mori, Transport, collective motion, and Brownian motion. *Prog. Theor. Phys.* **33**, 423–455 (1965).
32. R. Zwanzig, Nonlinear generalized Langevin equations. *J. Stat. Phys.* **9**, 215–220 (1973).
33. F. Takens, "Detecting strange attractors in turbulence" in *Dynamical Systems and Turbulence, Warwick 1980*, D. Rand, L. S. Young, Eds. (Springer, Berlin Heidelberg, Berlin, Heidelberg, 1981), pp. 366–381.
34. L. Biferale, F. Bonaccorso, M. Buzzicotti, C. Calascibetta, TURB-Lagr. A database of 3D Lagrangian trajectories in homogeneous and isotropic turbulence. arXiv. https://arxiv.org/abs/2303.08662.
35. X. M. de Wit, R. P. J. Kunnen, H. J. H. Clercx, F. Toschi, Efficient point-based simulation of four-way coupled particles in turbulence at high number density. *Phys. Rev. E* **110**, 015301 (2024).
36. Y. T. Lin, Y. Tian, D. Perez, D. Livescu, Regression-based projection for learning Mori-Zwanzig operators. *SIAM J. Appl. Dyn. Syst.* **22**, 2890–2926 (2023).

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.

37. M. Woodward *et al*., "Data-driven Mori-Zwanzig: Reduced order modeling of sparse sensors measurements for boundary layer transition" in *AIAA AVIATION 2023 Forum* (2023), p. 4256.
38. C. Calascibetta, L. Biferale, F. Borra, A. Celani, M. Cencini, Optimal tracking strategies in a turbulent flow. *Commun. Phys.* **6**, 256 (2023).
39. Y. Jiao, H. Hang, J. Merel, E. Kanso, Sensing flow gradients is necessary for learning autonomous underwater navigation. *Nat. Commun.* **16**, 3044 (2025).
40. H. Grabert, *Projection Operator Techniques in Nonequilibrium Statistical Mechanics, Springer Tracts in Modern Physics* (Springer-Verlag, Berlin, Heidelberg, 1982), vol. 95.
41. E. J. Parish, K. Duraisamy, Non-Markovian closure models for large eddy simulations using the Mori-Zwanzig formalism. *Phys. Rev. Fluids* **2**, 014604 (2017).
42. S. K. Falkena, C. Quinn, J. Sieber, J. Frank, H. A. Dijkstra, Derivation of delay equation climate models using the Mori-Zwanzig formalism. *Proc. R. Soc. A* **475**, 20190075 (2019).
43. Z. Li, X. Bian, X. Li, G. E. Karniadakis, Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. *J. Chem. Phys.* **143**, 243128 (2015).
44. P. Español, P. Warren, Statistical mechanics of dissipative particle dynamics. *Europhys. Lett.* **30**, 191 (1995).
45. A. Chorin, P. Stinis, Problem reduction, renormalization, and memory. *Commun. Appl. Math. Comput. Sci.* **1**, 1–27 (2007).
46. Y. Zhu, "Mori-Zwanzig equation: Theory and applications," Ph.D. thesis, UC Santa Cruz (2019). https://escholarship.org/uc/item/0mg7h7qg.
47. Y. Lin, Y. Tian, D. Livescu, M. Anghel, Data-driven learning for the Mori-Zwanzig formalism: A generalization of the Koopman learning framework. *SIAM J. Appl. Dyn. Syst.* **20**, 2558–2601 (2021).
48. M. Woodward *et al*., Mori-Zwanzig modal decomposition. arXiv [Preprint] (2023). https://arxiv.org/abs/2311.09524.
49. Y. Zhu, H. Lei, Effective Mori-Zwanzig equation for the reduced-order modeling of stochastic systems. *Discret. Contin. Dyn. Syst.* **15**, 959–982 (2022).
50. Y. Tian, Y. Lin, M. Anghel, D. Livescu, Data-driven learning of Mori-Zwanzig operators for isotropic turbulence. *Phys. Fluids* **33**, 125118 (2021).
51. M. Woodward *et al*., "Data-driven Mori-Zwanzig: Approaching a reduced order model for hypersonic boundary layer transition" in *AIAA SCITECH 2023 Forum* (American Institute of Aeronautics and Astronautics Inc., 2023).
52. M. Woodward *et al*., "Mori-Zwanzig modal decomposition: Applications and analysis of laminar-turbulent boundary layer transition" in *AIAA AVIATION FORUM AND ASCEND 2024* (American Institute of Aeronautics and Astronautics, 2024).
53. K. K. Lin, F. Lu, Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism. *J. Comput. Phys.* **424**, 109864 (2021).
54. E. Darve, J. Solomon, A. Kia, Computing generalized Langevin equations and generalized Fokker-Planck equations. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10884–10889 (2009).
55. F. Gilani, D. Giannakis, J. Harlim, Kernel-based prediction of non-markovian time series. *Phys. D Nonlinear Phenom.* **418**, 132829 (2021).
56. Z. She, P. Ge, H. Lei, Data-driven construction of stochastic reduced dynamics encoded with non-Markovian features. *J. Chem. Phys.* **158**, 034102 (2023).
57. Z. Li, H. Lee, E. Darve, G. Karniadakis, Computing the non-Markovian coarse-grained interactions derived from the Mori-Zwanzig formalism in molecular systems: Application to polymer melts. *J. Chem. Phys.* **146**, 14104 (2017).
58. R. Kubo, The fluctuation-dissipation theorem. *Rep. Prog. Phys.* **29**, 255 (1966).
59. A. J. Chorin, O. H. Hald, R. Kupferman, Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 2968–2973 (2000).
60. A. J. Chorin, O. H. Hald, R. Kupferman, Optimal prediction with memory. *Phys. D Nonlinear Phenom.* **166**, 239–257 (2002).
61. D. González, F. Chinesta, E. Cueto, Learning non-Markovian physics from data. *J. Comput. Phys.* **428**, 1–14 (2020).
62. Y. Zhu, J. M. Dominy, D. Venturi, On the estimation of the Mori-Zwanzig memory integral. *J. Math. Phys.* **59**, 103501 (2018).
63. J. Brandstetter, D. E. Worrall, M. Welling, "Message passing neural PDE solvers" in *International Conference on Learning Representations* (2022).
64. X. Wang *et al*., "CARD: Channel aligned robust blend transformer for time series forecasting" in *International Conference on Learning Representations* (2024).

Downloaded from https://www.pnas.org by UNIVERSITA STUDI FERRARA on March 26, 2026 from IP address 192.84.144.10.