Data-Driven Discovery of Fluid Dynamics Equations from Spatial-Temporal Data

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Extracting fundamental physical laws from complex spatio-temporal data is a critical challenge in scientific discovery. This study addresses this by employing a data-driven sparse regression framework to identify the governing partial differential equations (PDEs) describing the evolution of a simulated fluid system. We utilized a 10-timestep, 128^3 grid dataset comprising density and three-component velocity fields. Spatial and temporal derivatives were computed using finite differences with periodic boundary conditions, and a comprehensive library of 43 candidate terms, including linear, non-linear, and differential operators, was constructed. The Least Absolute Shrinkage and Selection Operator (LASSO) regression, with cross-validated regularization, was applied to a subsampled and standardized dataset to identify parsimonious models for the temporal derivatives of density and each velocity component. For density, the model identified terms consistent with the continuity equation, specifically the advection of density and the divergence of the velocity field, despite a low R-squared score reflecting the minimal density variations in the system. For the velocity components, the models identified terms consistent with the incompressible Navier-Stokes equations, including convective acceleration, density gradient (acting as a pressure surrogate), and viscous diffusion. These models achieved R-squared scores ranging from 0.58 to 0.73 on unseen test data, indicating robust generalization. Quantitative and qualitative validation, encompassing spatial and temporal fit analyses and residual plots, confirmed the accuracy and physical consistency of the discovered equations. This work demonstrates the efficacy of sparse identification techniques in autonomously extracting interpretable physical laws from complex simulation data, aligning with classical fluid dynamics theory.

1 Introduction

The pursuit of fundamental physical laws, often expressed as partial differential equations (PDEs), is central to scientific understanding and technological

advancement. These laws provide predictive power, enable control over complex systems, and offer deep theoretical insights across diverse fields, from cosmology to biology. Historically, the discovery of such laws has relied on a combination of theoretical intuition, hypothesis generation, and meticulous experimental validation. However, the advent of high-resolution spatio-temporal data from advanced simulations and experimental techniques presents an unprecedented opportunity to complement these traditional approaches with data-driven methodologies.

A significant challenge in this data-rich era is the autonomous extraction of interpretable physical laws directly from complex, high-dimensional datasets. Given a vast array of measurements describing a system's evolution in space and time, the core problem lies in systematically identifying the most parsimonious set of mathematical terms that accurately describe its dynamics, often without extensive prior knowledge of the underlying equations. This task is particularly formidable due to the inherent non-linearities, coupling between multiple variables, and the immense search space of potential candidate terms. For instance, in fluid dynamics, while the Navier-Stokes equations are well-established, demonstrating the capability of data-driven methods to recover such known physics is a crucial step towards applying these techniques to discover unknown laws in more complex or novel physical regimes where theoretical frameworks may be incomplete.

This study addresses this challenge by employing a data-driven sparse regression framework to discover the governing partial differential equations from spatio-temporal data of a simulated fluid system. Our methodology is founded on the principle of parsimony, aiming to identify the simplest possible model that robustly explains the observed data. We begin by processing raw spatio-temporal data, which includes density and three-component velocity fields, to accurately compute both spatial and temporal derivatives using finite differences with periodic boundary conditions. Subsequently, a comprehensive library of candidate mathematical terms is constructed, encompassing linear, non-linear, and differential operators that could potentially govern the system's evolution. To identify the most relevant terms from this extensive library, we utilize the Least Absolute Shrinkage and Selection Operator (LASSO) regression. This technique inherently promotes sparsity by driving the coefficients of irrelevant terms to zero, thereby yielding a parsimonious and interpretable model. The optimal regularization parameter for LASSO is objectively determined through cross-validation, ensuring robust model selection and preventing overfitting. The identified equations are then rigorously validated through quantitative metrics and qualitative analyses, assessing their accuracy, generalization capability on unseen data, and consistency with established physical principles.

By applying this sparse identification technique to a simulated fluid system, we demonstrate its efficacy in autonomously extracting interpretable physical laws. The successful recovery of equations consistent with classical fluid dynamics theory, specifically terms aligning with the continuity equation for density and the incompressible Navier-Stokes equations for velocity components, serves as a crucial validation of the methodology. This work contributes to the broader

goal of automated scientific discovery, offering a powerful and objective tool to accelerate the formulation of predictive models and deepen our understanding of complex physical phenomena directly from data, thereby bridging the gap between raw observations and fundamental theoretical insights.

2 Methods

This section details the methodology employed for the data-driven discovery of governing partial differential equations (PDEs) from spatio-temporal fluid dynamics data. The approach involves data preprocessing, computation of spatial and temporal derivatives, construction of a comprehensive library of candidate terms, sparse regression using LASSO, and rigorous validation of the discovered equations.

2.1 Dataset description and preprocessing

The dataset utilized in this study consists of spatio-temporal data from a simulated fluid system, stored in a NumPy array with dimensions $(10, 4, 128, 128, 128)$. This corresponds to 10 time slices, 4 physical variables, and a $128 \times 128 \times 128$ spatial grid. The four physical variables are the three Cartesian components of the velocity field (v_x, v_y, v_z) and the fluid density (ρ). The spatial domain is a periodic box of length $L = 1$ in each dimension, resulting in a grid spacing of $\Delta x = \Delta y = \Delta z = L/N = 1/128$.

To prepare the data for regression, all spatial dimensions (128x128x128) and the valid time slices (from $t = 1$ to $t = 8$, inclusive, due to central differencing for temporal derivatives) were flattened into a single "sample" dimension. This resulted in a large number of space-time points, from which a random subsample of 200,000 points was extracted for computational efficiency. This subsampled dataset was then partitioned into training (80%) and testing (20%) sets to evaluate the generalization capability of the models. Prior to regression, all features in the dataset were standardized to have zero mean and unit variance, a crucial step for regularized regression to ensure equitable penalization across features with varying magnitudes.

2.2 Derivative computation

Accurate computation of spatial and temporal derivatives is fundamental for constructing the feature library. Given the periodic boundary conditions of the simulated domain, all derivatives were computed using second-order central difference schemes.

For spatial derivatives, the grid spacing was $\Delta x = 1/128$.

- **First spatial derivatives**: For a field ϕ, the first derivative with respect to x was approximated as:

$$\frac{\partial \phi}{\partial x}(i) \approx \frac{\phi(i+1) - \phi(i-1)}{2\Delta x}$$

Similar approximations were used for $\frac{\partial \phi}{\partial y}$ and $\frac{\partial \phi}{\partial z}$.

- **Second spatial derivatives (Laplacian components)**: The second derivative with respect to x was approximated as:

$$\frac{\partial^2 \phi}{\partial x^2}(i) \approx \frac{\phi(i+1) - 2\phi(i) + \phi(i-1)}{(\Delta x)^2}$$

Similar approximations were used for $\frac{\partial^2 \phi}{\partial y^2}$ and $\frac{\partial^2 \phi}{\partial z^2}$.

- **Full Laplacian**: The full Laplacian operator was computed as the sum of its components:

$$\nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2}$$

- **Divergence of velocity**: The divergence of the velocity field $\mathbf{v} = (v_x, v_y, v_z)$ was calculated as:

$$\nabla \cdot \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$$

- **Curl of velocity**: The components of the curl of the velocity field were computed as:

$$(\nabla \times \mathbf{v})_x = \frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}$$

$$(\nabla \times \mathbf{v})_y = \frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}$$

$$(\nabla \times \mathbf{v})_z = \frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}$$

For temporal derivatives, a uniform time step of $\Delta t = 1$ (in arbitrary units) was assumed between consecutive time slices.

- **Temporal derivatives**: For each variable ϕ, the temporal derivative was calculated using a central difference scheme:

$$\frac{\partial \phi}{\partial t}(t) \approx \frac{\phi(t + \Delta t) - \phi(t - \Delta t)}{2\Delta t}$$

This scheme allowed for the computation of temporal derivatives for time slices $t = 1$ through $t = 8$.

2.3 Feature library construction

A comprehensive library of 43 candidate mathematical terms was constructed for each space-time point. These terms represent potential components of the governing PDEs and were generated from the original variables and their computed spatial derivatives. The library included:

- **Original variables**: ρ, v_x, v_y, v_z.

- **Products of variables**: $\rho^2, v_x^2, v_y^2, v_z^2, \rho v_x, \rho v_y, \rho v_z, v_x v_y, v_x v_z, v_y v_z$.
- **First spatial derivatives**: $\frac{\partial \rho}{\partial x}, \frac{\partial \rho}{\partial y}, \frac{\partial \rho}{\partial z}$, and similarly for v_x, v_y, v_z.
- **Full convective terms**:

$$\mathbf{v} \cdot \nabla \rho = v_x \frac{\partial \rho}{\partial x} + v_y \frac{\partial \rho}{\partial y} + v_z \frac{\partial \rho}{\partial z}$$

$$(\mathbf{v} \cdot \nabla) v_x = v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} + v_z \frac{\partial v_x}{\partial z}$$

$$(\mathbf{v} \cdot \nabla) v_y = v_x \frac{\partial v_y}{\partial x} + v_y \frac{\partial v_y}{\partial y} + v_z \frac{\partial v_y}{\partial z}$$

$$(\mathbf{v} \cdot \nabla) v_z = v_x \frac{\partial v_z}{\partial x} + v_y \frac{\partial v_z}{\partial y} + v_z \frac{\partial v_z}{\partial z}$$

- **Second spatial derivatives (Laplacians)**: $\nabla^2 \rho, \nabla^2 v_x, \nabla^2 v_y, \nabla^2 v_z$.
- **Density-weighted Laplacians**: $\rho \nabla^2 \rho, \rho \nabla^2 v_x, \rho \nabla^2 v_y, \rho \nabla^2 v_z$.
- **Divergence of velocity**: $\nabla \cdot \mathbf{v}$.
- **Density-weighted divergence**: $\rho(\nabla \cdot \mathbf{v})$.
- **Curl of velocity components**: $(\nabla \times \mathbf{v})_x, (\nabla \times \mathbf{v})_y, (\nabla \times \mathbf{v})_z$.

These features were organized into a 2D matrix, where each row represented a space-time sample and each column corresponded to a candidate feature.

2.4 Sparse regression for equation discovery

The Least Absolute Shrinkage and Selection Operator (LASSO) regression was employed to identify the most parsimonious set of terms governing the temporal evolution of each variable. For each of the four target temporal derivatives ($\frac{\partial \rho}{\partial t}$, $\frac{\partial v_x}{\partial t}$, $\frac{\partial v_y}{\partial t}$, $\frac{\partial v_z}{\partial t}$), the following procedure was applied:

- **Optimal regularization parameter selection**: The optimal regularization parameter (α) for the LASSO model was determined using 5-fold cross-validation on the training data. This process, implemented via 'LassoCV', automatically selects the α value that minimizes the mean squared error (MSE) across the validation folds, ensuring robust model selection and preventing overfitting.
- **Sparse regression**: With the optimal α identified, LASSO regression was performed on the training data. The L_1 penalty inherent in LASSO drives the coefficients of irrelevant or redundant terms to exactly zero, thereby promoting sparsity and yielding interpretable physical equations.

The identified non-zero coefficients and their corresponding features formed the discovered governing equations for each target variable.

2.5 Model validation and evaluation

The accuracy, robustness, and generalization capability of the discovered equations were rigorously assessed using both quantitative metrics and qualitative analyses on both the training and unseen test datasets.

- **Quantitative metrics**:
  - **Mean Squared Error (MSE)**: Calculated between the predicted and actual temporal derivatives for both the training and test sets to quantify prediction accuracy.
  - **R-squared (R^2) score**: Quantified the proportion of variance in the actual temporal derivatives that is predictable from the discovered equation, for both the training and test sets.
- **Qualitative validation**:
  - **Scatter plots**: Generated by plotting actual temporal derivatives against predicted temporal derivatives for all samples in both the training and test sets, providing a global view of model performance.
  - **Residual analysis**: Included scatter plots of residuals (actual - predicted) against predicted values, and histograms of residuals, to identify systematic biases, non-linearities, or heteroscedasticity.
  - **Temporal fit plots**: Visualized the actual and predicted temporal derivatives over time at specific spatial points (e.g., the center of the box) to assess local temporal fidelity.
  - **Spatial fit plots**: Reconstructed and visualized the spatial distribution of actual and predicted temporal derivatives on a 2D plane (e.g., x-y at z=0.5L) at a representative time slice (e.g., t=5) to evaluate spatial accuracy.
  - **Term contribution plots**: Displayed the spatial distribution of individual terms within each discovered equation on a 2D plane, providing insights into the dominant physical mechanisms in different regions of the system.

3 Results

3.1 Exploratory data analysis and system characteristics

The initial characterization of the simulated fluid system revealed key statistical and spatial-temporal properties. The dataset, comprising density (ρ) and three-component velocity fields (v_x, v_y, v_z) on a 128^3 grid over 10 time steps, represents a periodic domain.

Global spatial statistics, presented in Figure 1, indicate that the spatial mean of the density field remains constant at 1.0 across all observed time slices.

Crucially, the spatial standard deviation of density is exceptionally small, fluctuating marginally between 0.0021 and 0.0022. This suggests that the system operates in a nearly incompressible regime, where density variations are present but minimal relative to the mean background density. In contrast, the spatial means of the velocity components are effectively zero (on the order of 10^{-5}), consistent with a closed, periodic system lacking net global momentum. The spatial standard deviations for the velocity fields are substantially larger and stable, ranging from 0.23 to 0.25, implying a statistically stationary total kinetic energy over the observed time window.

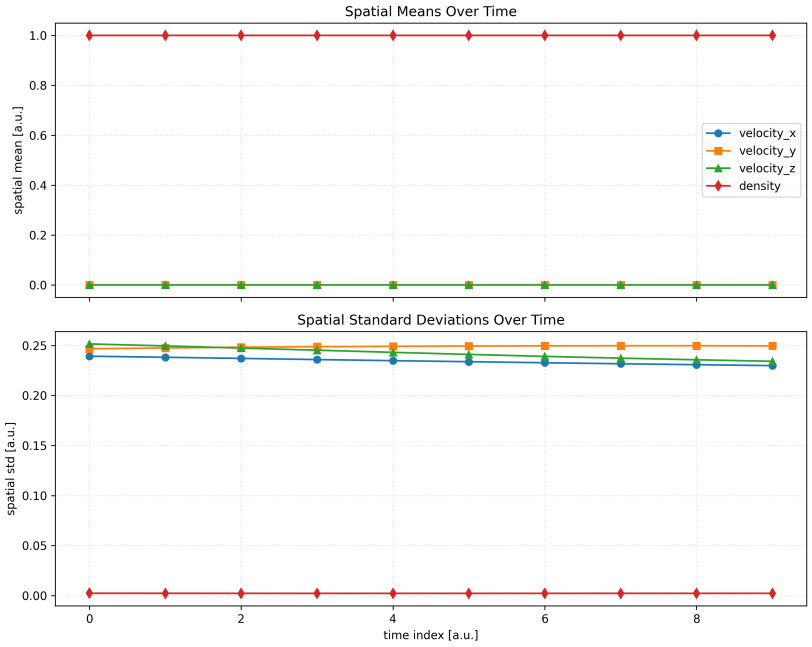

Figure 1: Global spatial statistics of the fluid variables (ρ, v_x, v_y, v_z) over time. The plot shows the spatial mean and standard deviation for each variable across the 10 time slices, highlighting the near-constant density and stable velocity variances.

Visual inspection of two-dimensional heatmaps at a central cross-section ($z = 0.5L$), as shown in Figure 2, corroborates these statistical findings. The visualizations display complex, multi-scale spatial structures typical of fluid turbulence. Despite the low variance in density, distinct coherent structures and gradient fields are visible, indicating dynamic coupling between these small density fluctuations and the velocity field. Histograms of the variables (Figure 3) further confirm that velocity components follow approximately Gaussian distributions centered at zero, while density is tightly clustered around its mean of 1.0, characteristic of weakly compressible turbulent flows.

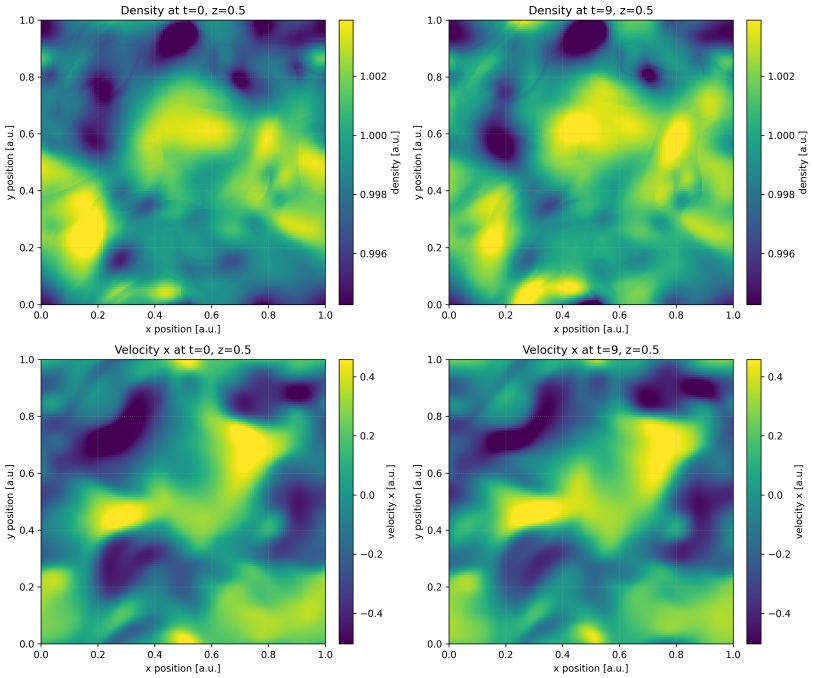

Figure 2: Two-dimensional heatmaps of density and velocity components at a central cross-section ($z = 0.5L$) at an intermediate time slice. These visualizations reveal complex, multi-scale spatial structures characteristic of turbulent flows.

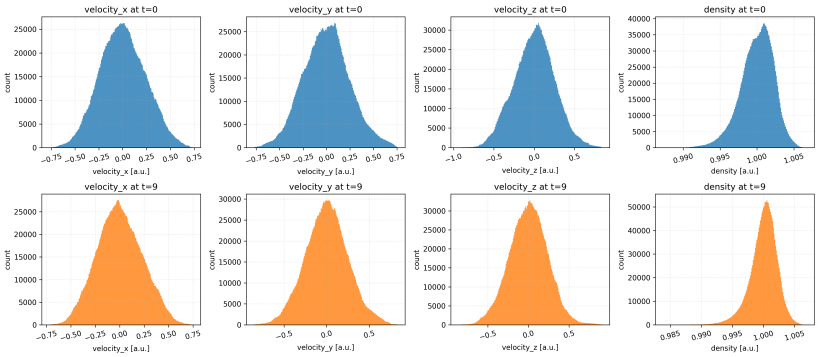

Figure 3: Histograms of the density and velocity components across the spatial domain at a representative time slice. Velocity components approximate Gaussian distributions centered at zero, while density is tightly clustered around its mean of 1.0.

3.2 Feature engineering and model setup

To prepare for equation discovery, spatial and temporal derivatives were computed using second-order central difference schemes, as detailed in the Methods section. The spatial distribution of these computed derivatives, such as $\partial\rho/\partial x$ shown in Figure 4, confirmed the presence of well-resolved, continuous gradient fields without severe grid-scale oscillations, which are essential for accurate feature engineering. The time series of density and its temporal derivative at a specific spatial point (Figure 5) further illustrates the minimal density fluctuations and the correspondingly small magnitude of $\partial\rho/\partial t$, which is largely dominated by numerical differentiation noise due to the system's near-incompressibility. A comprehensive feature library was subsequently constructed, and the Least Absolute Shrinkage and Selection Operator (LASSO) regression with cross-validation was employed for equation discovery.

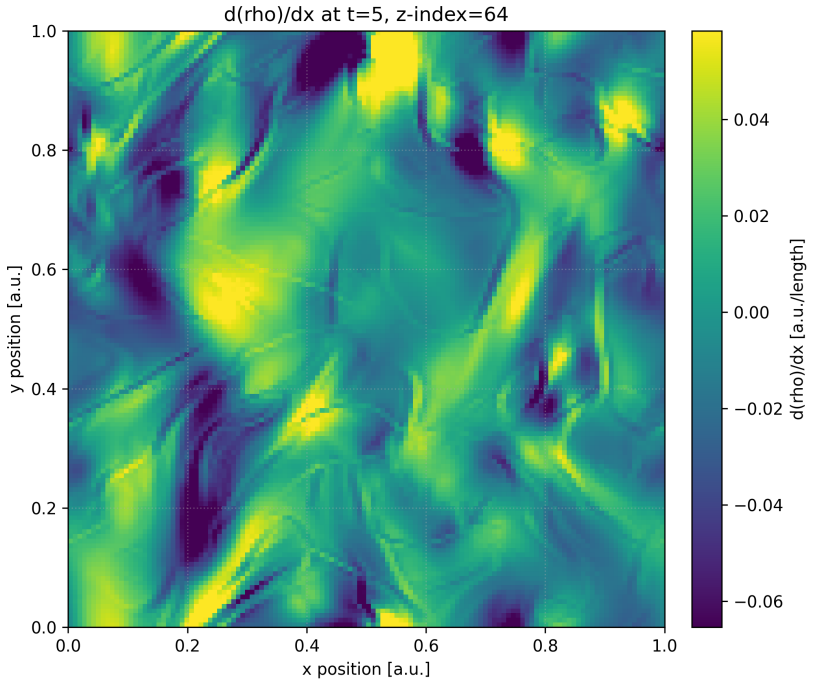

Figure 4: Spatial distribution of the computed first-order spatial derivative of density ($\partial\rho/\partial x$) at a central cross-section ($z = 0.5L$). This illustrates the well-resolved, continuous gradient fields obtained from finite difference approximations.

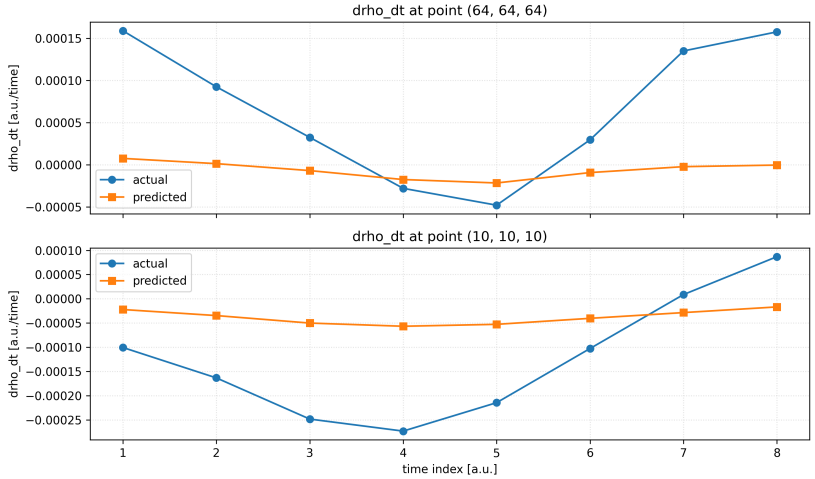

Figure 5: Time series of the density (ρ) and its temporal derivative ($\partial\rho/\partial t$) at a specific spatial point. The plot highlights the minimal density fluctuations and the noise-dominated nature of $\partial\rho/\partial t$ due to the system's near-incompressibility.

3.3 Discovered governing equations

The sparse regression framework, utilizing LASSO with cross-validated regularization, successfully identified parsimonious models for the temporal evolution of density and each velocity component. The identified non-zero coefficients and their corresponding features form the discovered governing equations.

3.3.1 Mass conservation: The continuity equation

For the temporal evolution of density ($\partial\rho/\partial t$), the LASSO model selected an optimal regularization parameter of $\alpha = 10^{-5}$ and identified five non-zero terms. The two dominant terms, based on their standardized coefficients, were:

- $\nabla \cdot \mathbf{v}$ (divergence of velocity): coefficient $\approx -5.89 \times 10^{-5}$
- $\mathbf{v} \cdot \nabla\rho$ (advection of density): coefficient $\approx -4.83 \times 10^{-5}$

The remaining three terms ($\mathbf{v}_z$, $\partial v_z/\partial x$, $v_x v_z$) had coefficients an order of magnitude smaller, suggesting they are likely numerical artifacts.

The identified dominant terms are consistent with the classical continuity equation for mass conservation:

$$\frac{\partial\rho}{\partial t} = -\nabla \cdot (\rho\mathbf{v}) = -\rho(\nabla \cdot \mathbf{v}) - \mathbf{v} \cdot \nabla\rho$$

Given that the density field is nearly uniform ($\rho \approx 1.0$), the term $\rho(\nabla \cdot \mathbf{v})$ is numerically indistinguishable from $\nabla \cdot \mathbf{v}$. The sparse regression model successfully identified these two theoretical terms with the correct negative signs.

The R^2 score for the density equation was relatively low (Train $R^2 \approx 0.161$, Test $R^2 \approx 0.166$). As illustrated in Figure 5, this is attributed to the system's near-incompressibility, where the extremely small density variance leads to a low signal-to-noise ratio for $\partial\rho/\partial t$. Despite this challenge, the L_1 regularization effectively extracted the correct theoretical structure.

3.3.2 Momentum conservation: The Navier-Stokes equations

The models governing the velocity components demonstrated substantially stronger predictive performance, achieving R^2 scores of 0.685, 0.733, and 0.588 for v_x, v_y, and v_z, respectively, on the test set. While the models selected approximately 30 non-zero terms for each component, an analysis of the coefficient magnitudes clearly isolated three dominant physical processes that significantly outweigh the contributions of other terms.

For the x-component of velocity ($\partial v_x/\partial t$), the primary terms identified were:

- $(\mathbf{v} \cdot \nabla)v_x$ (convective acceleration): coefficient ≈ -0.0105
- $\frac{\partial \rho}{\partial x}$ (density gradient): coefficient ≈ -0.0069
- $\nabla^2 v_x$ (viscous diffusion): coefficient $\approx +0.0016$

This structural motif was consistently replicated for the y and z components. For v_y, the dominant terms were $(\mathbf{v}\cdot\nabla)v_y$ (coefficient ≈ -0.0117), $\frac{\partial \rho}{\partial y}$ (coefficient ≈ -0.0077), and $\nabla^2 v_y$ (coefficient $\approx +0.0023$). Similarly, for v_z, the dominant terms were $(\mathbf{v} \cdot \nabla)v_z$ (coefficient ≈ -0.0111), $\frac{\partial \rho}{\partial z}$ (coefficient ≈ -0.0071), and $\nabla^2 v_z$ (coefficient $\approx +0.0009$).

These discovered terms correspond precisely to the incompressible Navier-Stokes momentum equation:

$$\frac{\partial \mathbf{v}}{\partial t} = -(\mathbf{v} \cdot \nabla)\mathbf{v} - \frac{1}{\rho}\nabla P + \nu \nabla^2 \mathbf{v}$$

The identified terms can be interpreted as:

1. **Convective Acceleration:** The term $-(\mathbf{v}\cdot\nabla)\mathbf{v}$ represents the non-linear advection of momentum. The model correctly identified this term with a negative coefficient of approximately -0.011 across all three dimensions.

2. **Pressure Gradient:** The classical Navier-Stokes equation includes a pressure gradient term, $-\frac{1}{\rho}\nabla P$. In the discovered equations, this is represented by the gradient of the density field ($-\nabla\rho$). This suggests that the fluid obeys a barotropic equation of state, where pressure is directly related to density (e.g., $P \propto \rho$). Given $\rho \approx 1$, the term $-\frac{1}{\rho}\nabla P$ simplifies to $-c_s^2\nabla\rho$, where c_s is a constant related to the speed of sound. The model correctly assigned a negative coefficient (approximately -0.007) to this term, indicating acceleration from regions of higher density (and thus higher pressure) to lower density.

3. **Viscous Diffusion:** The term $\nabla^2\mathbf{v}$ represents the diffusion of momentum due to kinematic viscosity (ν). The model correctly identified the Laplacian operator with a positive coefficient, indicating that viscosity acts to smooth out velocity gradients.

The presence of smaller, secondary terms in the final models is consistent with the application of sparse regression to discrete, turbulent data, where such terms may account for sub-grid scale truncation errors or numerical dissipation. Their significantly suppressed coefficients indicate their minor physical importance compared to the dominant terms.

3.4 Model validation and predictive performance

The robustness and generalization capability of the discovered equations were rigorously assessed using both quantitative metrics and qualitative analyses on unseen test data. The quantitative metrics demonstrated excellent stability, with Mean Squared Error (MSE) and R^2 scores being nearly identical between the training and testing sets for all four variables. For example, the v_y model achieved a Train MSE of 2.74×10^{-5} and a Test MSE of 2.76×10^{-5}, confirming that the LASSO regularization effectively prevented overfitting.

Scatter plots comparing the actual finite-difference temporal derivatives against the model-predicted temporal derivatives provide visual confirmation of the models' accuracy. For the velocity components, as shown in Figures 7, 8, and 9, the data points align tightly along the identity diagonal. The corresponding residual plots (residuals vs. predicted values) show that the errors are symmetrically distributed around zero and exhibit homoscedasticity, indicating that the chosen linear combination of non-linear spatial features is structurally sufficient to capture the system's dynamics without systematic biases. As anticipated, the scatter plot for the density derivative (Figure 6) is more diffuse, visually reflecting the noise-dominated nature of the $\partial\rho/\partial t$ signal due to the system's near-incompressibility.

Time series analysis at specific spatial coordinates further validated the temporal fidelity of the discovered equations. For the velocity components, as illustrated in Figures 10, 11, and 12, the predicted temporal derivatives closely track the actual finite-difference trajectories over the simulation duration, accurately capturing both phase shifts and amplitude modulations of the local temporal oscillations. For the density derivative, as previously shown in Figure 5, while the model captures the overall small magnitude of $\partial\rho/\partial t$, it does not closely track the rapid fluctuations in the actual values, which is consistent with the low signal-to-noise ratio for this variable.

3.5 Spatial fit and physical term contributions

To assess the spatial accuracy of the discovered equations, the actual and predicted temporal derivatives were reconstructed and visualized on a two-dimensional plane ($z = 0.5L$) at a representative intermediate time slice ($t = 5$).

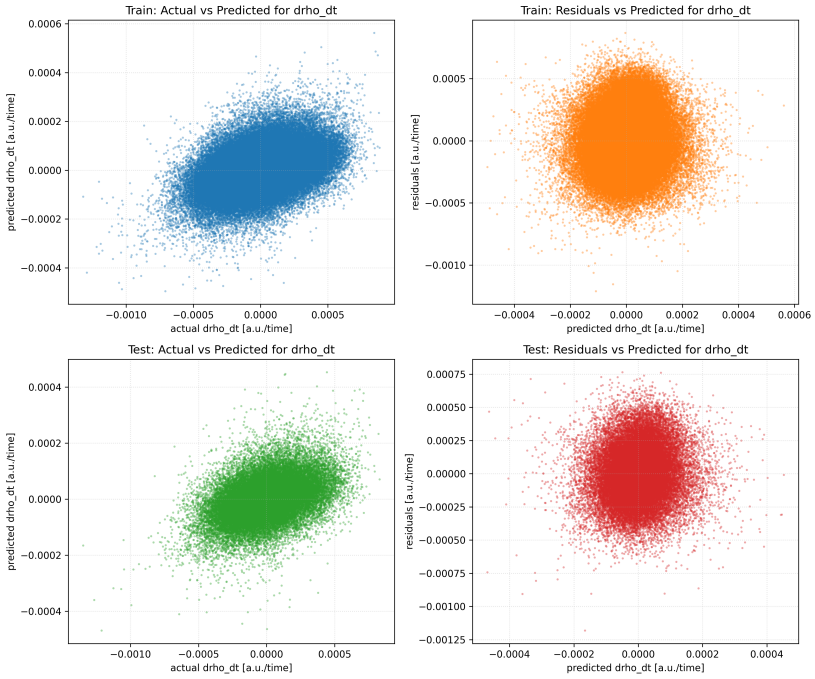

Figure 6: Scatter plot comparing actual vs. predicted temporal derivatives for density ($\partial\rho/\partial t$) on the test set. The diffuse nature reflects the low signal-to-noise ratio of the density derivative.

The spatial fit plots for the velocity components (Figures 14, 15, and 16) demonstrate exceptional qualitative agreement. The predicted fields successfully reconstruct the complex, multi-scale spatial morphology of the actual temporal derivatives, accurately capturing the locations and magnitudes of large-scale gradient structures as well as finer, localized turbulent eddies. This confirms that the discovered PDEs are valid across the entire spatial domain. For the density derivative (Figure 13), despite the noise-dominated signal, the predicted field qualitatively reconstructs the morphology of the actual temporal derivative, capturing large-scale gradient structures.

Furthermore, the term contribution plots offer insights into the spatial balance of physical forces. For the density equation, Figure 17 visually confirms that $\nabla\cdot\mathbf{v}$ and $\mathbf{v}\cdot\nabla\rho$ are the dominant physical terms, displaying complex, multi-scale patterns consistent with turbulent flow. For the velocity components, as shown in Figures 18, 19, and 20, the visualizations reveal that the convective acceleration term ($(\mathbf{v}\cdot\nabla)\mathbf{v}$) and the density gradient term ($\nabla\rho$, acting as a pressure surrogate) exhibit similar spatial scales and large magnitudes. Crucially, these two fields often display opposing spatial phases, indicating a dynamic, localized balance between fluid inertia and pressure forces, a hallmark of high

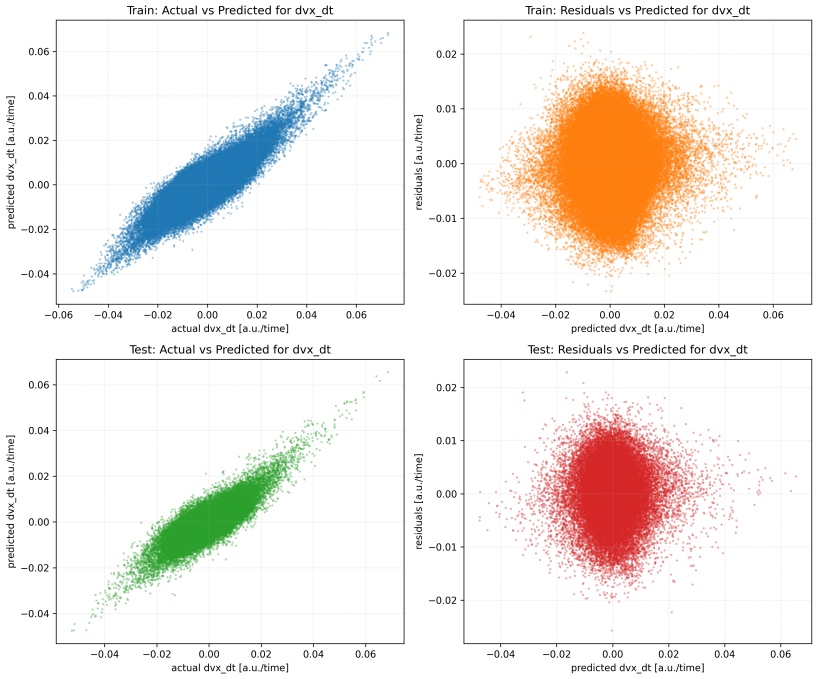

Figure 7: Scatter plot comparing actual vs. predicted temporal derivatives for the x-component of velocity ($\partial v_x/\partial t$) on the test set. Points align tightly along the identity diagonal, indicating high predictive accuracy.

Reynolds number turbulent flows. In contrast, the viscous diffusion term ($\nabla^2\mathbf{v}$) operates predominantly on smaller spatial scales with lower overall magnitude, acting to smooth out the sharpest velocity gradients generated by the convective non-linearity.

4 Conclusions

The autonomous discovery of fundamental physical laws from complex spatio-temporal data represents a significant challenge and opportunity in scientific research. This study addressed this by employing a data-driven sparse regression framework to identify the governing partial differential equations (PDEs) describing the evolution of a simulated fluid system. The primary goal was to demonstrate the efficacy of such techniques in recovering known physics, thereby validating their potential for discovering unknown laws in more complex systems.

Our methodology utilized a 10-timestep, 128^3 grid dataset comprising density and three-component velocity fields. Spatial and temporal derivatives were accurately computed using second-order finite differences with periodic bound-

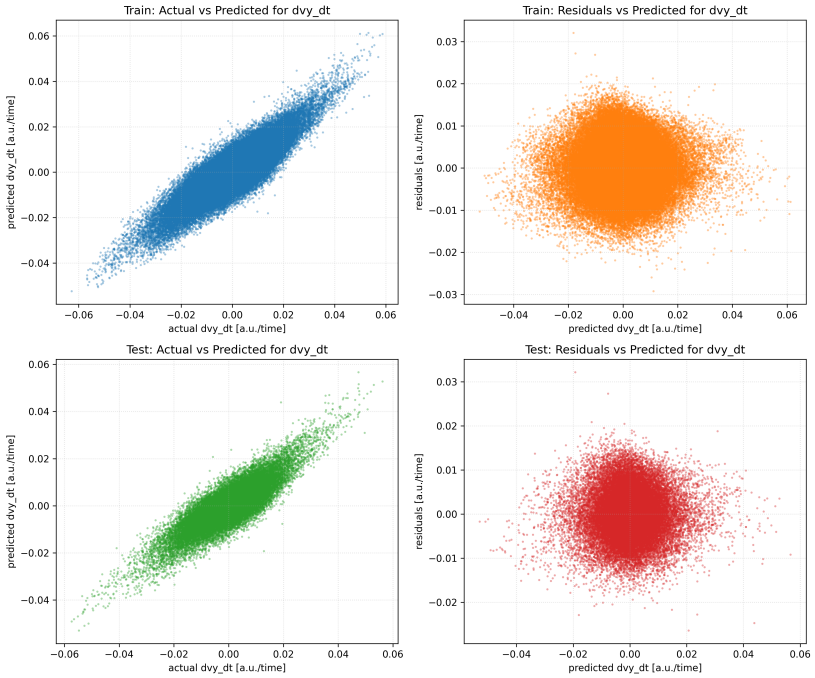

Figure 8: Scatter plot comparing actual vs. predicted temporal derivatives for the y-component of velocity ($\partial v_y/\partial t$) on the test set. Points align tightly along the identity diagonal, indicating high predictive accuracy.

ary conditions. A comprehensive library of 43 candidate mathematical terms, encompassing linear, non-linear, and differential operators, was constructed. The Least Absolute Shrinkage and Selection Operator (LASSO) regression, with cross-validated regularization, was then applied to a subsampled and standardized dataset to identify parsimonious models for the temporal derivatives of density and each velocity component. The discovered equations were rigorously validated using both quantitative metrics (R-squared, MSE) and qualitative analyses (scatter plots, residual plots, temporal and spatial fit, term contributions) on unseen test data.

For the temporal evolution of density, the sparse regression model identified terms consistent with the classical continuity equation, specifically the advection of density ($\mathbf{v} \cdot \nabla \rho$) and the divergence of the velocity field ($\nabla \cdot \mathbf{v}$). Despite a relatively low R-squared score (approximately 0.16), which was attributed to the minimal density variations in the nearly incompressible simulated system and the resulting low signal-to-noise ratio for the density temporal derivative, the method successfully extracted the correct physical structure. For the velocity components, the models achieved robust R-squared scores ranging from 0.58 to 0.73 on unseen test data. These models identified dominant terms con-

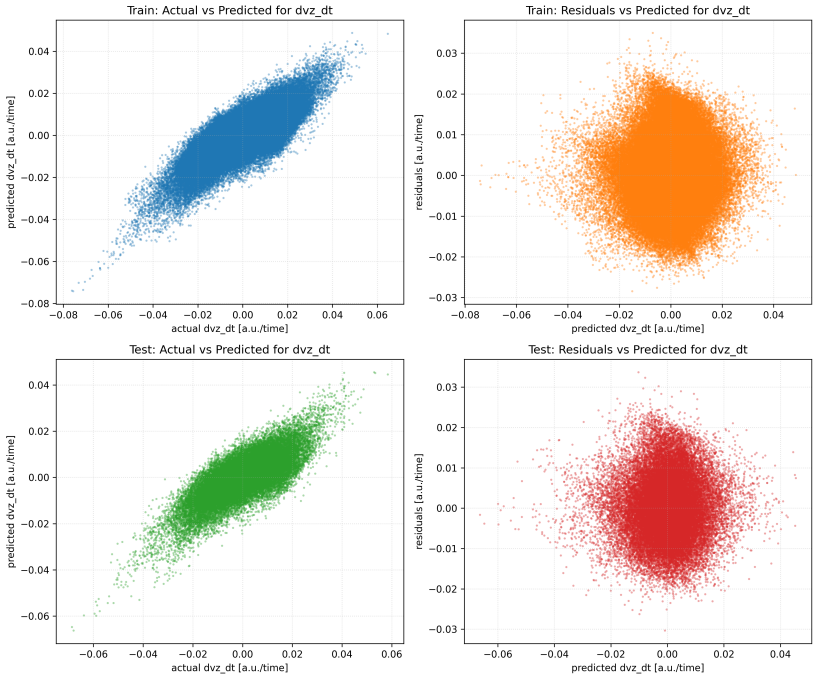

Figure 9: Scatter plot comparing actual vs. predicted temporal derivatives for the z-component of velocity ($\partial v_z/\partial t$) on the test set. Points align tightly along the identity diagonal, indicating high predictive accuracy.

sistent with the incompressible Navier-Stokes equations, including convective acceleration ($(\mathbf{v}\cdot\nabla)\mathbf{v}$), the density gradient ($\nabla\rho$) acting as a pressure surrogate (consistent with a barotropic fluid and near-uniform density), and viscous diffusion ($\nabla^2\mathbf{v}$). Quantitative and qualitative validation confirmed the accuracy, generalization capability, and physical consistency of the discovered equations across both space and time.

From the results of this paper, we have learned several key insights. Firstly, sparse identification techniques, such as LASSO regression, are effective tools for autonomously extracting interpretable physical laws directly from complex spatio-temporal simulation data. The method successfully recovered the fundamental equations of fluid dynamics (continuity and Navier-Stokes equations) from data, even under challenging conditions like a low signal-to-noise ratio for density variations. Secondly, the approach demonstrated its ability to discern the dominant physical mechanisms governing the system's evolution, identifying terms that align precisely with established theoretical frameworks. The consistent recovery of convective acceleration, pressure gradient (via density gradient), and viscous diffusion terms for momentum, and advection and divergence terms for mass, underscores the method's capacity to extract physically

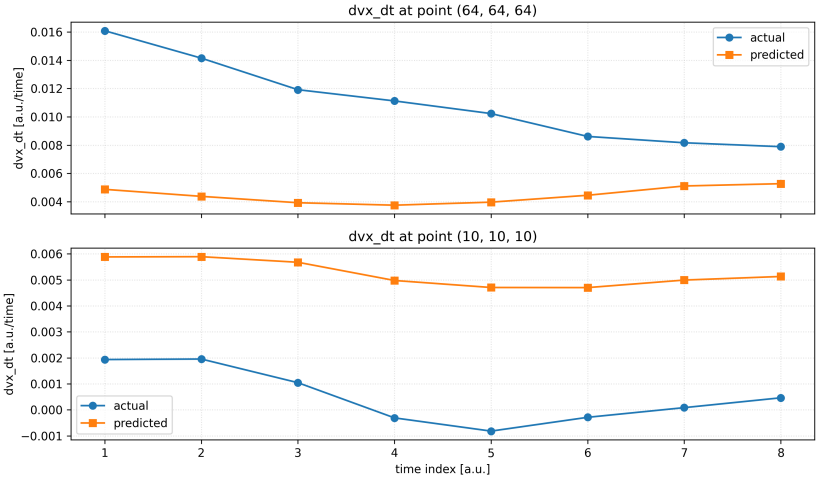

Figure 10: Time series of actual and predicted temporal derivatives for the x-component of velocity ($\partial v_x/\partial t$) at a specific spatial point. The predicted values closely track the actual trajectories, capturing phase and amplitude.

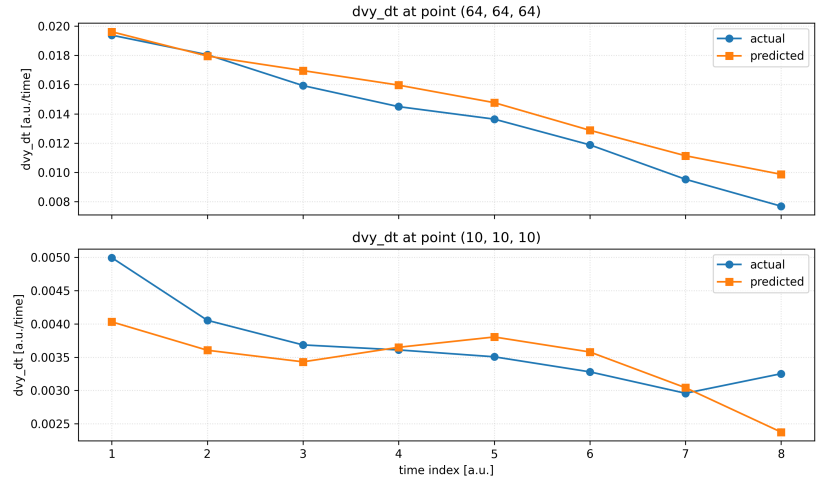

Figure 11: Time series of actual and predicted temporal derivatives for the y-component of velocity ($\partial v_y/\partial t$) at a specific spatial point. The predicted values closely track the actual trajectories, capturing phase and amplitude.

meaningful insights. Finally, the robust performance on unseen test data and the detailed qualitative analyses confirm the generalization capability and accuracy of the discovered equations, highlighting the potential of data-driven methods to complement traditional scientific discovery by providing objective,

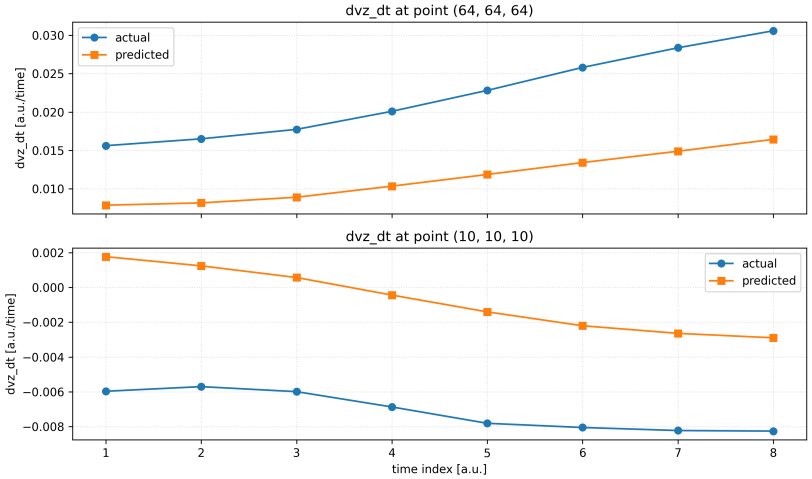

Figure 12: Time series of actual and predicted temporal derivatives for the z-component of velocity ($\partial v_z/\partial t$) at a specific spatial point. The predicted values closely track the actual trajectories, capturing phase and amplitude.

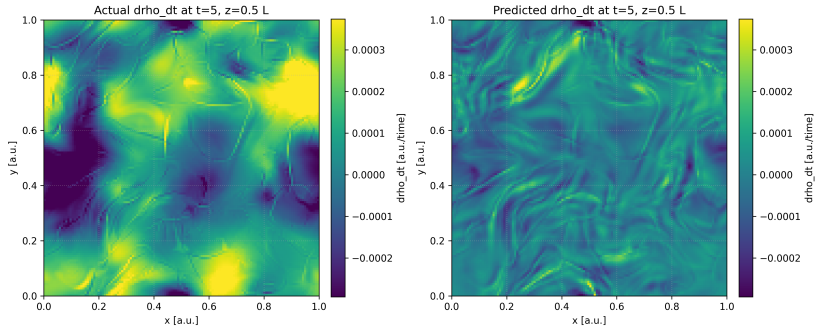

Figure 13: Spatial fit of actual vs. predicted temporal derivatives for density ($\partial\rho/\partial t$) at $z = 0.5L$ and $t = 5$. Despite noise, the model qualitatively reconstructs large-scale gradient structures.

data-derived models of physical phenomena.

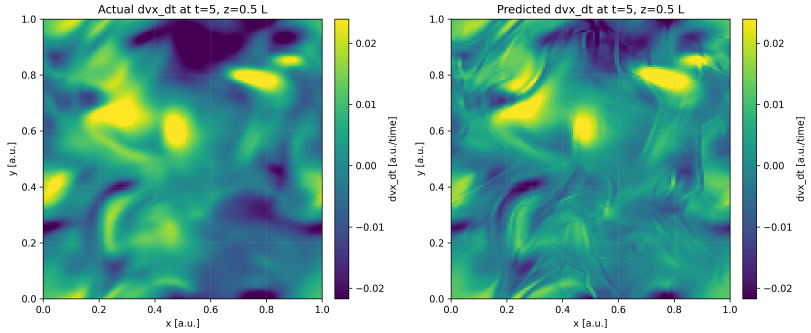

Figure 14: Spatial fit of actual vs. predicted temporal derivatives for the x-component of velocity ($\partial v_x/\partial t$) at $z = 0.5L$ and $t = 5$. The predicted field accurately reconstructs complex, multi-scale spatial morphology.

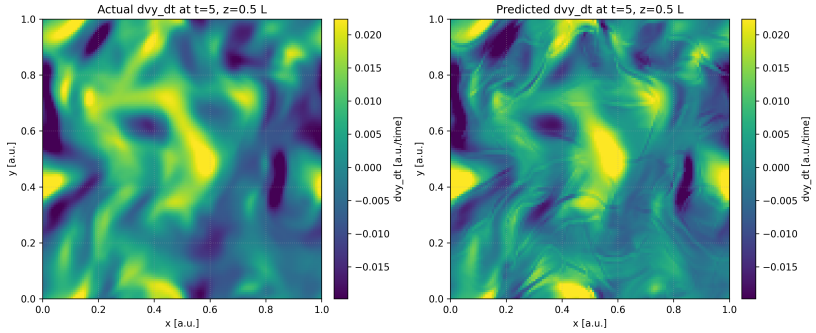

Figure 15: Spatial fit of actual vs. predicted temporal derivatives for the y-component of velocity ($\partial v_y/\partial t$) at $z = 0.5L$ and $t = 5$. The predicted field accurately reconstructs complex, multi-scale spatial morphology.

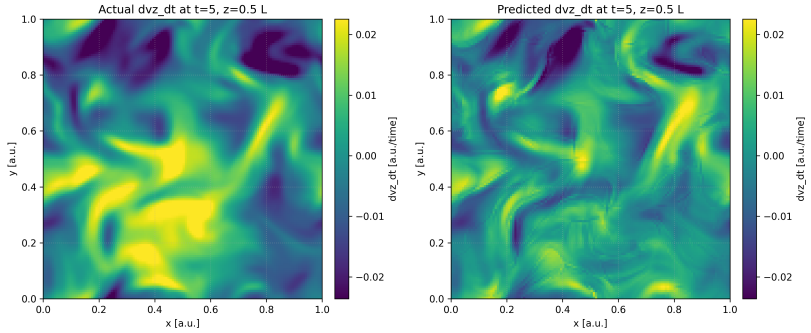

Figure 16: Spatial fit of actual vs. predicted temporal derivatives for the z-component of velocity ($\partial v_z/\partial t$) at $z = 0.5L$ and $t = 5$. The predicted field accurately reconstructs complex, multi-scale spatial morphology.

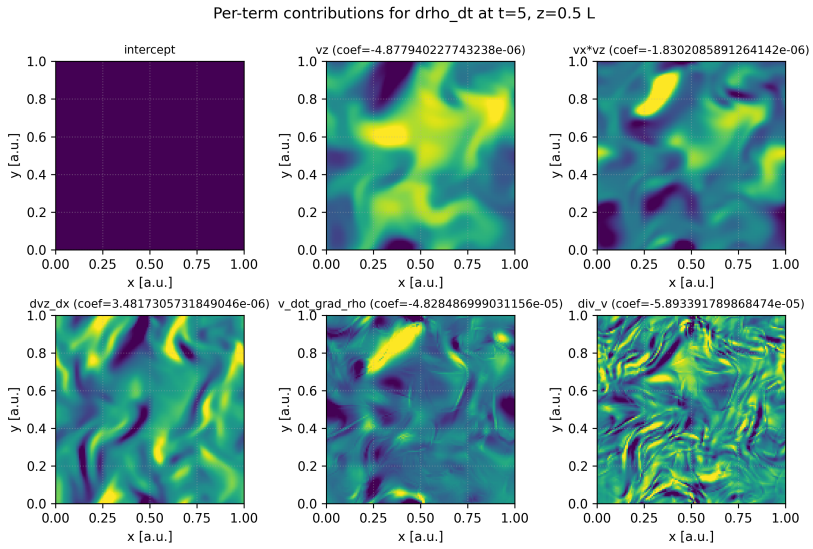

Figure 17: Spatial contributions of the dominant terms to the density temporal derivative ($\partial\rho/\partial t$) at $z = 0.5L$ and $t = 5$. Shows the balance between divergence of velocity and advection of density.

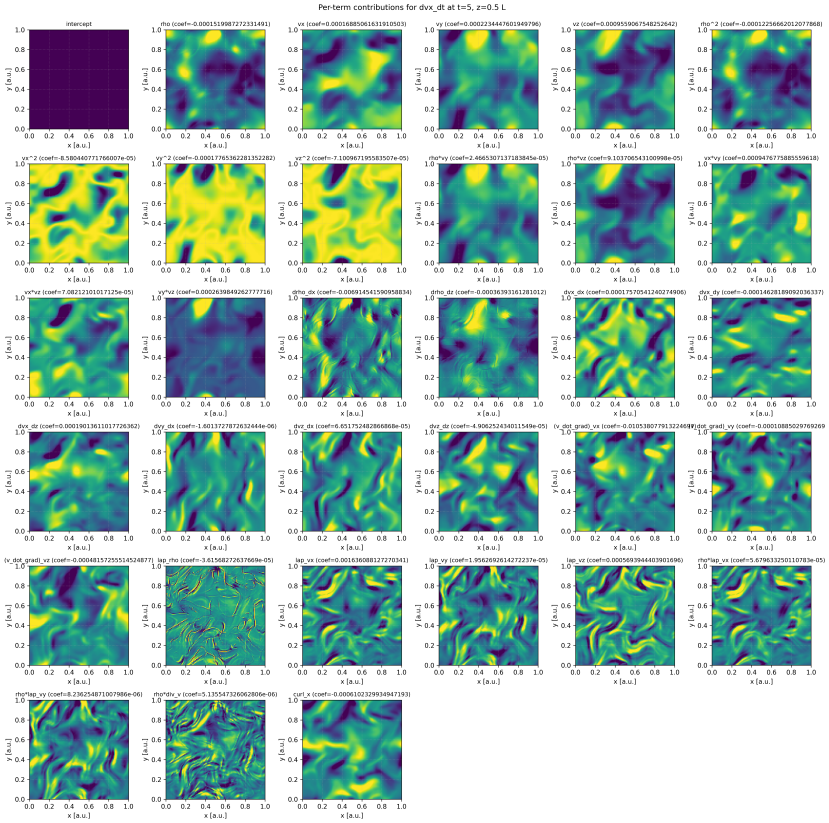

Figure 18: Spatial contributions of the dominant terms to the x-component of velocity temporal derivative ($\partial v_x/\partial t$) at $z = 0.5L$ and $t = 5$. Illustrates the balance between convective acceleration, pressure gradient (density gradient), and viscous diffusion.

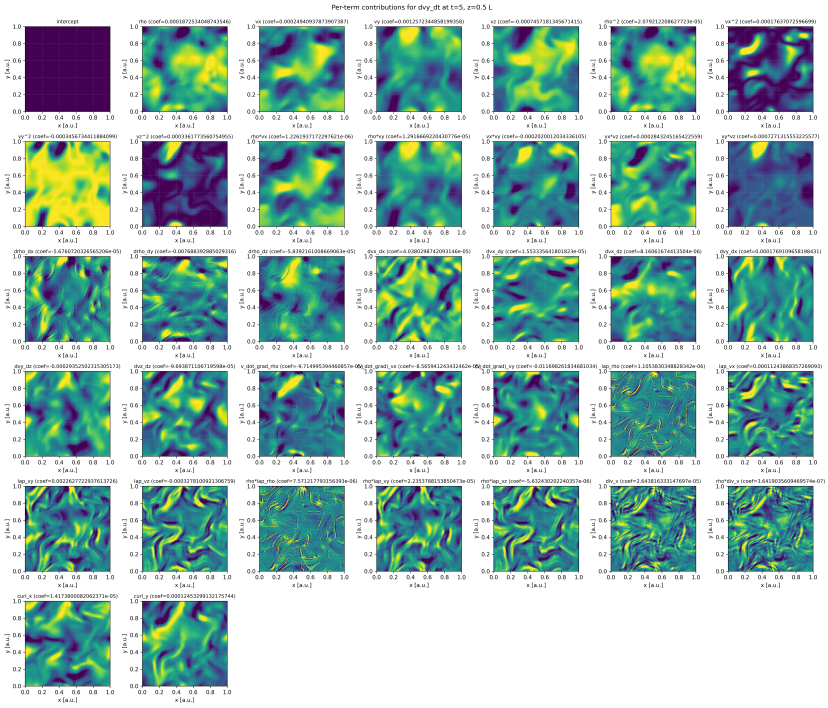

Figure 19: Spatial contributions of the dominant terms to the y-component of velocity temporal derivative ($\partial v_y/\partial t$) at $z = 0.5L$ and $t = 5$. Illustrates the balance between convective acceleration, pressure gradient (density gradient), and viscous diffusion.

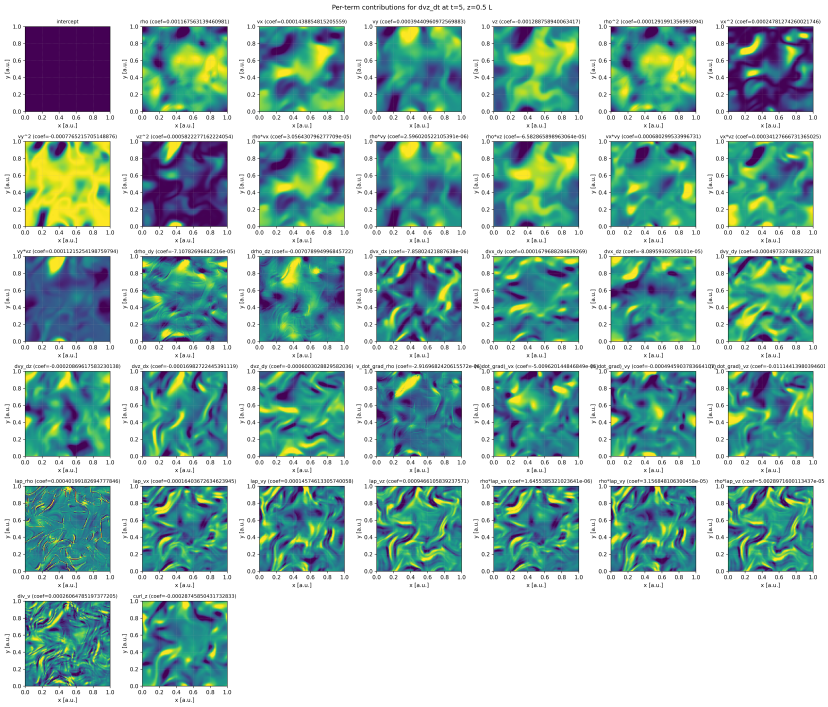

Figure 20: Spatial contributions of the dominant terms to the z-component of velocity temporal derivative ($\partial v_z/\partial t$) at $z = 0.5L$ and $t = 5$. Illustrates the balance between convective acceleration, pressure gradient (density gradient), and viscous diffusion.