# Physics Steering: Causal Control of Cross-Domain Concepts in a Physics Foundation Model

**Rio Alexa Fear**
University of Cambridge
raf74@cam.ac.uk

**Payel Mukhopadhyay**
University of Cambridge
pm858@cam.ac.uk

**Michael McCabe**
NYU & Simons Foundation
mmccabe@simonsfoundation.org

**Alberto Bietti**
Flatiron Institute
abietti@flatironinstitute.org

**Miles Cranmer**
University of Cambridge
mc2473@cam.ac.uk

**The PolymathicAI Collaboration**

## Abstract

Recent advances in mechanistic interpretability have revealed that large language models (LLMs) develop internal representations corresponding not only to concrete entities but also distinct, human-understandable abstract concepts and behaviour. Moreover, these hidden features can be directly manipulated to steer model behaviour. However, it remains an open question whether this phenomenon is unique to models trained on inherently structured data (ie. language, images) or if it is a general property of foundation models. In this work, we investigate the internal representations of a large physics-focused foundation model. Inspired by recent work identifying single directions in activation space for complex behaviours in LLMs, we extract activation vectors from the model during forward passes over simulation datasets for different physical regimes. We then compute "delta" representations between the two regimes. These delta tensors act as concept directions in activation space, encoding specific physical features. By injecting these concept directions back into the model during inference, we can steer its predictions, demonstrating causal control over physical behaviours, such as inducing or removing some particular physical feature from a simulation. These results suggest that scientific foundation models learn generalised representations of physical principles. They do not merely rely on superficial correlations and patterns in the simulations. Our findings open new avenues for understanding and controlling scientific foundation models and has implications for AI-enabled scientific discovery.

## 1 Introduction

Recent advances in the field of interpretability have enhanced our comprehension of how foundation models function. Methods such as probing [38, 39] and Sparse Autoencoders (SAEs) [4, 8], primarily designed for large language models (LLMs), have uncovered that these models often form internal representations, or hidden features, that closely resemble human concepts [9, 5, 14]. A wide range of features from descriptive nouns [11], to abstract meta-concepts, and more [35, 21] are well

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

documented in the literature. Recent studies have even indicated that intricate behaviours such as refusal can be mediated by a single direction in activation space [1]. Importantly, these features are not merely correlational; interventions such as activation steering demonstrate that they have a causal influence on model behaviour [36, 40, 20].

The scientific community is increasingly leveraging large-scale models developed on extensive datasets across diverse domains, including chemistry [7, 3], astronomy [25, 19, 30], climate science [26], and healthcare [18]. However, while there has been rapid progress in interpretability research for LLMs, the internal representations of foundation models trained on scientific data remain largely unexplored. A key open question is whether — in a manner similar to LLMs — simulation models form interpretable representations which align with fundamental physical laws and principles, or if they depend on superficial correlations and patterns in the data.

This paper begins to tackle these questions by applying interpretability techniques, adapted from LLM studies, to a state-of-the-art transformer model pretrained on The Well [27], a large and varied collection of PDE simulations. Our research aims to determine if this physics foundation model yields interpretable internal representations of physical phenomena and whether these representations can be causally manipulated. Inspired by the method of identifying single direction "concept vectors" described by [1, 40], we employ a modified version of the technique to determine directions in the model's activation space which correspond to specific physical concepts. We inject these concept vectors during the model's forward pass to achieve activation steering [36] and thereby assess their causal impact on the resulting simulations.

Our contributions include:

- A methodology for extracting interpretable physical concept features from transformer-based physics models.
- We compute single-direction "delta" tensors between activations from contrasting physical regimes.
- We show that intervention along these directions causally steers model predictions in interpretable ways.
- We provide evidence that concept features a transferable between unrelated physical systems, suggesting that neural networks learn transferable abstract concepts across different physics domains.

### 1.1 Background

**The Physics Foundation Model** The model we investigate is a large vision transformer [37, 17] based foundation model designed for spatiotemporal surrogate modelling of physical systems described by PDEs. This model has been pretrained on a large range of complex and diverse datasets present in the Well collection [27]. It builds upon similar physics foundation model approaches introduced by [13, 16, 6, 22]. In short, the model is trained autoregressively to predict the next state of a physical system given a sequence of previous states. A key aspect of this pretraining is aiming to learn broadly useful representations of physical dynamics and facilitate transfer learning.

**The Well** [27] is a large-scale (15TB) benchmark dataset comprising 16 distinct numerical simulations curated in collaboration with domain experts. It spans diverse fields including fluid dynamics (e.g., Rayleigh-Bénard convection, Shear Flow, Magneto-hydrodynamics), astrophysics (e.g., Supernovae, Post-neutron star mergers), acoustic scattering, and even biological systems; amongst other things.

The data is provided as sequences of snapshots on uniform grids, and for each simulation includes multiple trajectories with varying initial conditions or physical parameters. The Well provides the diverse, high-quality data necessary for the physics foundation model to learn representations that generalize across physical domains and provides a challenging benchmark for evaluating generalization and transfer in scientific ML [34]. The Well also serves as the testbed for our interpretability investigations.

### 1.2 Interpretability

Mechanistic interpretability aims to reverse engineer neural networks into human-understandable algorithms [28, 24]. Several key interpretability hypotheses underpin this work, these are briefly covered below.

**Linear representation hypothesis** posits that features (i.e., concepts) are represented linearly as directions in a models activation space [10, 2, 29].

**Polysemanticity** refers to the theory that deep learning models can represent more features than the dimensionality of their activation space would suggest. Models achieve this by assigning multiple, potentially unrelated, features to a single neuron (*polysemanticity*) and representing features in non-orthogonal directions (*superposition*) [32, 10, 15, 2].

An unfortunate side effect of polysemanticity is that it complicates interpretation, as individual polysemantic neurons are generally not easily interpretable. Various techniques (e.g. SAEs) aim to address this by creating new representations of internal activations where neurons and features have a 1 to 1 relationship, resulting in *monosemantic* features [4, 35].

It should be noted that a monosemantic feature is not necessarily a feature which makes sense to a human being. For instance, one can imagine an LLM learning a monosemantic feature based purely on complex correlations between tokens which bears no relationship whatsoever to any human concept. After all, a feature is just a reusable, statistically independent component of a dataset that a model happens to find useful. Yet recent interpretability research has demonstrated that LLMs (and vision transformers) *do* indeed learn a multitude of human-interpretable features. [23, 39].

Many researchers argue that concept-based features emerge as a side effect of structure which is intrinsic to the training data. Language possesses inherent syntactic and semantic hierarchies that reflect human concepts, therefore text data provides a rich, human-understandable symbolic structure that models can pick up directly from the text. It is thus perhaps unsurprising that language data should lend itself to the formation of meaningful high-level abstractions and concept features in LLMs.

Numerical physics data, on the other hand, lacks an explicit concept structure. Instead, this structure can only arise indirectly through the abstraction of underlying governing rules, which a model must first infer. Therefore, for a foundation model trained on physics data, there is less *a priori* reason to assume its internal representations will correspond to human concepts

**Activation steering.** Beyond passive observation, interpretability aims to achieve causal understanding. Activation steering is a causal intervention technique where a precomputed vector, representing a concept, is added to the models activations at a specific layer during a forward pass. If the concept vector is meaningful and the intervention is successful, the models output will change in a manner which is consistent with the concept. This serves to test the causal link between activation directions and the models behaviour [36, 40, 20, 35]. Activation steering has been used to control stylistic attributes, factual recall, and more.

**Single Direction Steering.** Our work draws on the approach outlined by [1], which showed that complex behaviours in LLMs, such as refusal behaviour, can be identified with a single direction in activation space. This direction can be found by the computation of concept *deltas*, that is, by finding differences between model activations for different inputs (e.g., toxic vs. non-toxic text), one can identify directions in activation space that correspond to specific concepts. These directions can then be added or subtracted from activations during inference to steer the model's behaviour.

## 2 Methodology

Our methodology consists of four main steps: (1) selection of contrasting simulation files representing two distinct physical regimes; (2) extraction of activations from forward passes of the model across several examples from each regime; (3) calculation of "delta" concept directions; and (4) injection of concept directions to steer model outputs.

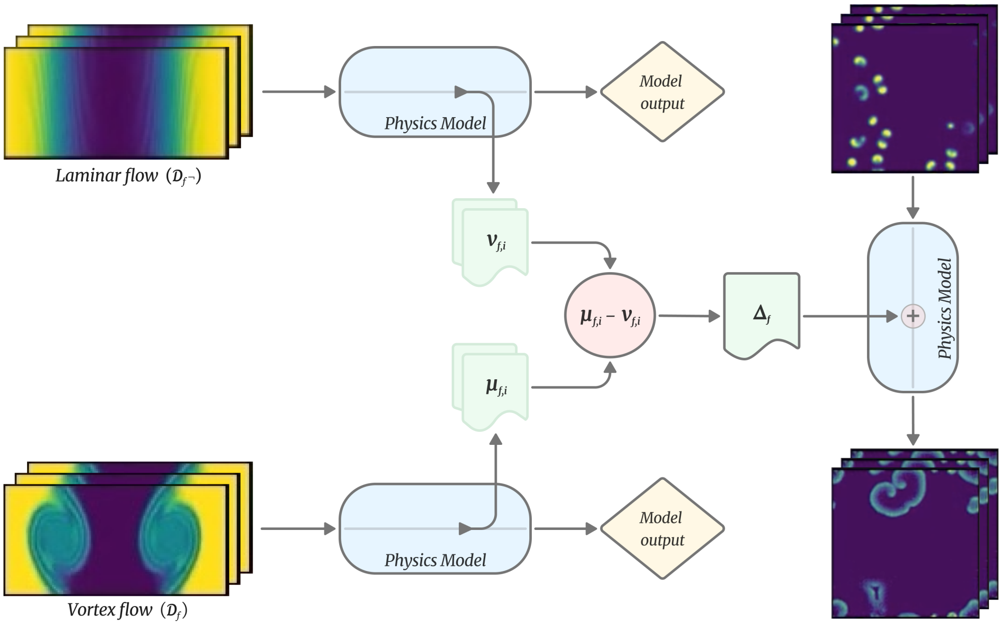

**Figure 1:** *Schematic illustration of methodology. Activations are first extracted from the physics model during forward passes over input segments that exhibit physical feature $f$, yielding activations $\boldsymbol{\mu}_{f,i}$, and from segments lacking the feature, $f^{\neg}$, yielding $\boldsymbol{\nu}_{f,i}$. The difference between these activations, $\boldsymbol{\Delta}_f$, is then injected back into the model during inference to steer future results.*

We investigate whether single-direction activation interventions, termed "delta steering", can be used to understand and control the internal representations of physical phenomena within the physics foundation model, we therefore investigate the hypothesis that physical concepts are linearly represented in the latent space of physics foundation models. Let $\mathbf{a}$ denote the activation tensor for a particular transformer block of a physics foundation model. We seek to identify direction $\boldsymbol{\Delta}_f$ in activation space such that the intervention $\mathbf{a} \rightarrow \mathbf{a} + \alpha\boldsymbol{\Delta}_f$ for scalar $\alpha > 0$ causally steers the model's predictions toward a desired physical feature $f$. The methodology employed consists of four primary steps, adapted from techniques used in LLM interpretability.

1. **Selection of Contrasting Simulation Files:** We create two groups of simulations taken from The Well such that the groups represent two distinct regimes of one physical system, with the difference between them being some physical feature which has visually distinguishable macro-scale effects, complex dynamics emerging from micro-dynamics, and the existence analogous structures across different phenomena to enable transferability studies. To meet these criteria, we focused our initial investigations on vorticity within the Shear Flow dataset, chosen for its well-understood physics and distinct visual features.
2. **Activation Extraction:** Activations were extracted from the model during forward passes over selected input segments from the simulation trajectories. PyTorch [31] hooks were used to capture the activations from a specific model layer. For this work, we chose the final transformer block, hypothesised to be the most likely to contain abstract representations of physical dynamics [12]. However, noting the observations in [33], we suspect that features in intermediate-to-late layers might achieve similar effects, and exploring this layer-dependence would be an interesting direction for future work. The model was run in rollout mode, processing windowed segments of consecutive timesteps. For each input, we extract the activation tensor $\mathbf{a} \in \mathcal{A} \subseteq \mathbb{R}^{T \times C \times W \times H}$, where $\mathcal{A}$ is the activation space, $T$ is the sequence length, $C$ the channel/feature dimension, and $W$, $H$ are spatial dimensions (width and height).
3. **Calculation of Concept Directions:** The saved tensors were averaged across each group resulting in an average laminar flow tensor and an average vortex flow tensor. The "delta tensor", or concept direction, were then computed by taking the difference between the two averaged activation tensors.

Let $\mathcal{D}_f$ denote the dataset of activation tensors extracted from input segments that exhibit physical feature $f$, and let $\mathcal{D}_{f^\neg}$ denote the dataset from input segments that lack feature $f$ or exhibit the opposite of feature $f$. To identify the direction corresponding to the physical feature $f$, we first normalize and then average the activations.

For each activation position $i = (t, w, h) \in \mathcal{I}$, where $\mathcal{I}$ is the set of all activation positions in the model's representation and these positions correspond to spatiotemporal locations in the physical simulation, we normalize the activations:

$$\hat{\mathbf{a}}_i = \frac{\mathbf{a}_i - \bar{\mathbf{a}}_i}{\sigma_i} \tag{1}$$

where $\bar{\mathbf{a}}_i$ and $\sigma_i$ are the mean and standard deviation across the training data at position $i$. We then compute the mean normalized activations for each dataset:

$$\boldsymbol{\mu}_{f,i} := \frac{1}{|\mathcal{D}_f^{\text{train}}|} \sum_{\mathbf{a} \in \mathcal{D}_f^{\text{train}}} \hat{\mathbf{a}}_i, \qquad \boldsymbol{\nu}_{f,i} := \frac{1}{|\mathcal{D}_{f^\neg}^{\text{train}}|} \sum_{\mathbf{a} \in \mathcal{D}_{f^\neg}^{\text{train}}} \hat{\mathbf{a}}_i \tag{2}$$

and we compute the concept direction as the difference between averaged activations:

$$\Delta_{f,i} := \boldsymbol{\mu}_{f,i} - \boldsymbol{\nu}_{f,i} \tag{3}$$

yielding the full concept direction tensor $\boldsymbol{\Delta}_f \in \mathcal{A}$. This direction is interpreted as encoding the concept of physical feature $f$ in activation space.

For cross-domain transfer experiments where spatial structures may not align between different physical systems, we also compute a spatially-averaged concept direction:

$$\overline{\boldsymbol{\Delta}}_f := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \Delta_{f,i} \tag{4}$$

This spatially-averaged direction $\overline{\boldsymbol{\Delta}}_f \in \mathbb{R}^C$ preserves only the channel-wise concept information.

4. **Activation Steering (Injection of Concept Directions):** To test the causal influence of these concept directions, they were injected back into the model during inference. Using a forward hook at the same target layer, the original activations $\mathbf{a}$ were modified by addition of the concept direction. The modified activations $\mathbf{a}'$ were calculated with steering function $\mathbf{s} : \mathcal{A} \times \mathcal{A} \times \mathbb{R} \to \mathcal{A}$ where $\mathbf{a}, \boldsymbol{\Delta}_f \in \mathcal{A}$ and $\alpha \in \mathbb{R}$:

$$\mathbf{s}(\mathbf{a}, \boldsymbol{\Delta}_f, \alpha) := \mathbf{a} + \alpha \|\mathbf{a}\|^2 \frac{\boldsymbol{\Delta}_f}{\|\boldsymbol{\Delta}_f\|^2} \tag{5}$$

where $\alpha$ is a scaling factor. The output was then renormalised to preserve the original norm of $\mathbf{a}$. This intervention was applied across all tokens and time steps.

# 3 Experiments

## 3.1 Progressive Suppression of a Physical Feature

The most straightforward method of testing the causal influence of the concept direction is to suppress the physical feature $a$ in the output simulation by means of subtraction in eq. (5). However, for ease of interpretation, the choice of the physical feature $a$ is crucial.

**Result.** The effect of negative activation steering was visually striking. Whereas the unmodified simulation displayed two prominent vortical structures, the steered simulations showed a progressive suppression of these features with increasing $\alpha$. The flow was instead transformed into a smooth, parallel state characteristic of a laminar regime. The successful laminarisation of the flow is an encouraging first sign that our method can precisely target and remove specific complex phenomena from a simulation.

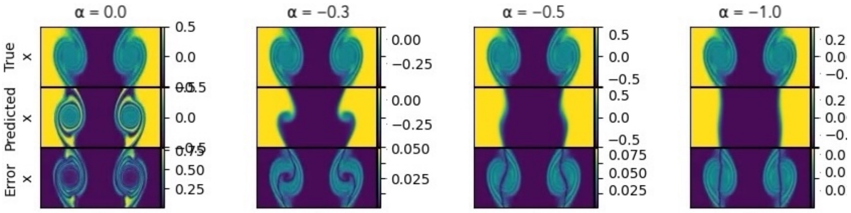

**Figure 2:** *Negative* $\mathbf{\Delta}_{vortex}$ *injection into shear flow vortex regime, for* $\alpha$ *values of 0, 0.3, 0.5 and 1.0. Frame: 64.*

### 3.2 Continuous induction of a Physical Feature

A natural subsequent question to ask is whether the opposite intervention will be similarly effective — can the addition of the same concept vector to the activations at layer $l$ give rise to the associated physical feature in the output simulation?

Successful inducement of a feature is a notably higher bar to pass than simple suppression because the model predicts the token deltas at each time step, rather than the entire state of tokens, so it is conceivable that the feature suppression intervention may not truly be targeting a feature representing a physical characteristic. It may instead be setting the prediction deltas to zero for a range of tokens, thereby resulting in the initial simulation state (i.e., laminar flow), persisting throughout the model rollout window.

To address this concern we repeated the suppression procedure but with the sign reversed in eq. (5).

**Result.** Positive injection of the learned vortex direction during inference on shear-flow simulations in the laminar regime reliably induced vortical structures, with the effect scaling with the steering strength $\alpha$: small injections ($\alpha \approx 0.1$ to $0.4$) produced subtle perturbations and incipient rotation, while moderate injections ($\alpha \approx 0.4$ to $0.5$) yielded well-formed vortices.

This result further validates the physical interpretation of the vortex direction and suggests that the extracted direction may encode a meaningful, controllable physical feature capable of introducing vortex formation into an otherwise laminar regime.

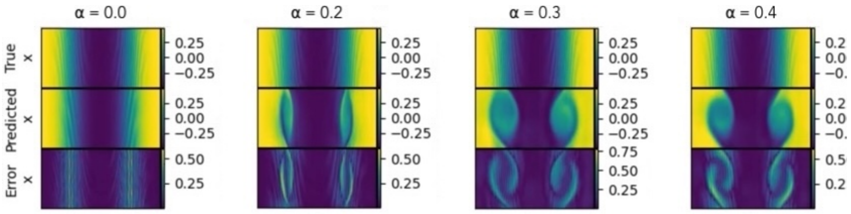

**Figure 3:** *Positive* $\mathbf{\Delta}_{vortex}$ *injection into shear flow laminar regime, for* $\alpha$ *values of 0, 0.2, 0.3 and 0.4. Frame: 64.*

The success of the concept induction experiment raises a question: given that vortices are being introduced to a simulation in which they would not normally arise, by what mechanism does the model transform a nominally laminar flow to produce a vortex? Is it applying a physically valid initial perturbation and then correctly simulating the ongoing natural evolution? Is it simulating a modified but self-consistent version of the physics? Or is it cosmetically shifting the output to look more like the target concept?

### 3.3 Additional Physical Features

Given the success of both suppression and induction of the vorticity concept direction we next asked whether an alternative, very different concept can be found? Where a vortex is a localised phenomenon which is defined by its structure, we now aim to isolate a concept direction that represents process-based phenomenon, which is not defined by a specific structure or confined to a particular location:

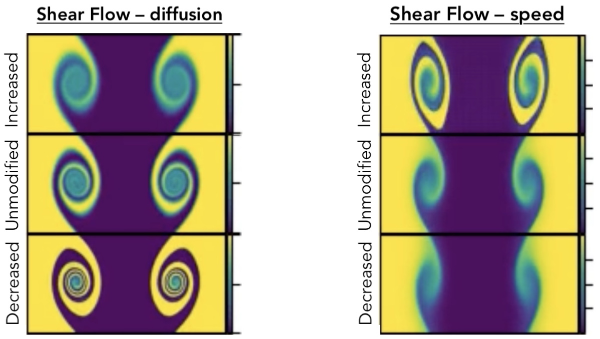

**Figure 4:** *On the left tracer fields for* $\Delta_{diffusion}$ *injection into Shear Flow vortex regime with (top)* $\alpha = 0.1$ *and (bottom)* $\alpha = -0.1$. *On the right tracer fields for* $\Delta_{speed}$ *injection into Shear Flow vortex regime with (top)* $\alpha = 0.1$ *and (bottom)* $\alpha = -0.1$. *Frame(left): 30, Frame(right): 24.*

### 3.3.1 Diffusion

Using the same Shear Flow simulation and the same extraction and injection methodology we computed the diffusion delta direction as the difference between the averaged activations for several high molecular diffusion and low molecular diffusion Shear Flow data files – that is, two groups of Shear Flow simulations with identical Reynolds numbers but different sets of Schmidt numbers.

**Result.** We discover analogous results for diffusion phenomena, with the diffusion direction encoding meaningful information about diffusion processes which can be manipulated causally. In fig. 4 addition of the diffusion direction presented itself as a more diffuse looking fluid interface, while subtraction led to a more sharply defined interface. In appendix fig. 9 addition also leads to larger, more spread out core pressure minima and y-velocity high/low zones, plus smoother x-velocity gradients; subtraction on the other hand leads to a reduction in the size of the same regions, along with sharper x-velocity gradients.

### 3.3.2 Temporal

After isolating a structural feature (vorticity) and a process-based one (diffusion), we investigated whether a more fundamental simulation property – its temporal progression – could be similarly controlled. To create a "speed" feature, we used the same extraction and injection methodology on two Shear Flow simulations that were physically identical but sampled at different frame rates. The delta direction was computed as the difference between the mean activations of a high-frame-rate (fast) simulation and a low-frame-rate (slow) one.

**Result.** Injecting the "speed" direction with a positive steering coefficient caused the vortex to form much earlier in the rollout window. Conversely, subtracting the direction delayed the formation of the vortex. This can seen by the fact that the vortex in the top right of fig. 4 is larger and more well developed compared to the lower right image where the vortex has barely formed by the same video frame.

### 3.4 Feature Transfer Between Physical Systems

A final important question arises regarding the nature and usefulness of the discovered concept directions: Are these concept directions specific to the Shear Flow dataset which was used to derive them, or do they represent a more general physical understanding learned by the physics foundation model?

We thus tested the transferability of the vorticity and speed features by applying the delta concept injection derived from the Shear Flow datasets to three alternative Well datasets, ordered by increasing dissimilarity from the Shear Flow dataset.

1. **Rayleigh-Bénard Convection**: An alternative fluid dynamics dataset which models fluid heated from below and cooled from above, creating convection patterns.
2. **Euler Quadrant**: A second alternative fluid dynamics dataset which seeks to simulate two compressible, inviscid gas species governed by the Euler equations.
3. **Gray-Scott Reaction-Diffusion**: An entirely unrelated system outside the field of fluid dynamics. This dataset contains simulations of a chemical reaction-diffusion system which produces various pattern formations, including gliders, spots, spirals, and mazes depending on parameter settings.

For within-domain steering (e.g., shear flow to shear flow), we use the full concept direction tensor $\mathbf{\Delta}_f$ that preserves spatial structure. However, when transferring concept directions between different physical systems, the activation tensors may have different spatial dimensions. To address this, we employ two strategies:

- **Spatial averaging**: Using the spatially-averaged concept direction $\overline{\mathbf{\Delta}}_f$ defined in Equation (5), which preserves only channel-wise information. This approach assumes that the physical concept is encoded primarily in the channel dimensions rather than specific spatial patterns.
- **Spatial alignment**: When spatial dimensions are similar (differing by at most one element), we pad or interpolate to match dimensions, preserving spatial structure. Interpolation and padding produced nearly identical results, so we describe the results below in terms of the inclusion or non-inclusion of spatial dimensions.

Our experiments show that spatial averaging generally produces more interpretable and physically consistent results for cross-domain transfer, as it extracts the abstract concept independent of system-specific spatial configurations.

**Rayleigh-Bénard Vorticity Transfer.** In the first concept transfer experiment we see a clear illustration of the impact which the presence of spatial dimensions in the steering tensor can have. Across each of the Rayleigh-Bénard results we see that the intervention appears to manifest as moderate changes to convection in the buoyancy field, in addition to comparatively extreme shifts in the pressure field. Two primary observations jump out to the viewer: Firstly, when the spatial dimensions are not included there is an increase in convection (that is, convection patterns appear earlier and are larger) with positive steering, and a corresponding decrease with negative steering. Secondly, when the spatial dimensions are included the simple positive direction = increase and negative direction = decrease relationship disappears. Instead *both* directions produce an increase in convection in the buoyancy field, along with large high and low pressure zones which appear to be inverted between the two results.

**Euler Vorticity Transfer.** Here we see a more straightforward result: an increase in the size and number of rotational flow features in the positive steering direction, especially at shock interfaces. Conversely, in the negative direction we see a decrease in size and number of rotational flow features. It is interesting to note that the shock interfaces are precisely where one would expect vortices to show up in a physically real scenario.

**Euler Speed Transfer.** In the second Euler transfer result it is immediately apparent that the shock fronts move faster with positive steering and slower with negative steering. In fig. 6 the effect can most readily be observed by comparing the position of the vertical shock line along the bottom of each of the three images. In the positive (top) image it is further along than the unmodified (middle) image, which in turn is further along than the negative (bottom) image. Another notable observation is that the addition of the speed direction has led to the creation of rotational features along both sides of the thick yellow shock front in the top right of the image.

**Gray-Scott Vorticity Transfer.** Of all our results, the most surprising was vorticity steering in the Gray-Scott "gliders" simulation, a physical system which is defined by interactions between two chemical species ("A" and "B") and where the concept of a fluid vortex does not apply. Despite this, we find that positive vortex steering induced the transformation of gliders in the chemical concentration fields into spiral patterns very reminiscent of those normally found in a "spirals" type Gray-Scott Reaction Diffusion system.

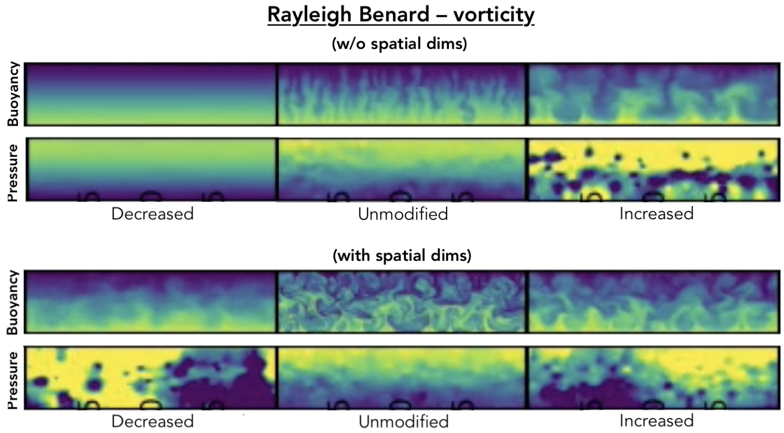

**Figure 5:** *Transfer of $\boldsymbol{\Delta}_{vortex}$ concept injection to Rayleigh-Bénard simulations. Pressure and buoyancy fields for (top) averaging over spatial dimensions: (left) $\alpha = -0.1$, (centre) $\alpha = 0.0$, (right) $\alpha = 0.1$; (bottom) including spatial dimensions (no averaging): (left) $\alpha = -0.1$, (centre) $\alpha = 0.0$, (right) $\alpha = 0.1$. Frame(top): 40, Frame(bottom): 50.*

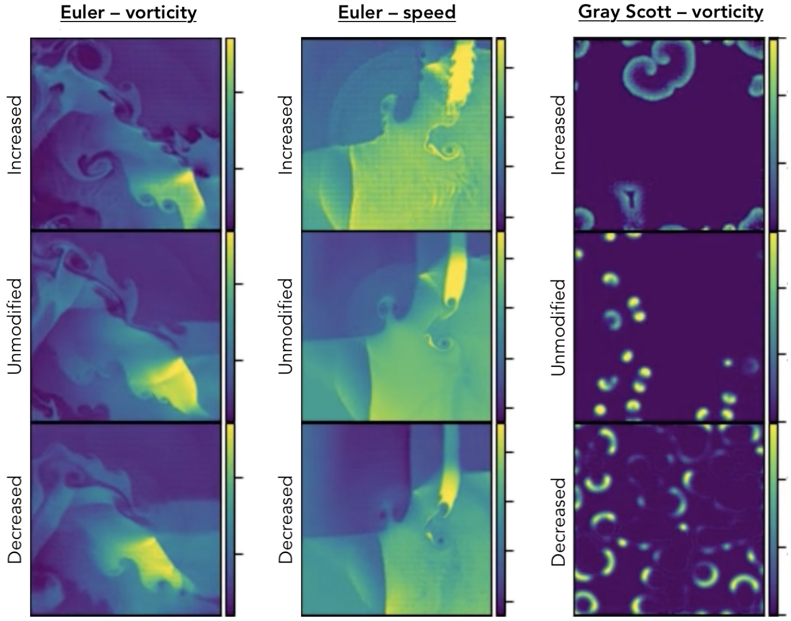

**Figure 6:** *(Left) Density field for $\overline{\boldsymbol{\Delta}}_{vortex}$ injection into Euler quadrants. (Middle) Density field for $\overline{\boldsymbol{\Delta}}_{speed}$ injection into Euler quadrants. (Right) Chemical species B for $\overline{\boldsymbol{\Delta}}_{vortex}$ injection into Gray-Scott reaction diffusion. All computed by averaging over spatial dimensions, with $\alpha = 0.1$ (top), $\alpha = 0.0$ (middle), and $\alpha = -0.1$ (bottom). Frames: 50 (left), 28 (middle), 48 (right).*

## 4 Discussion

These experiments show that these concept directions are not merely correlational but have a causal effect on the simulation. Within the Shear Flow dataset, adding the vortex direction induced vortical structures in a laminar flow, while subtracting it suppressed existing vortices.

Interestingly, these concept directions appear to generalise across different physical systems. The vortex direction, derived entirely from Shear Flow simulations, introduced broadly analogous rotational structures when transferred to other fluid dynamics datasets like Rayleigh-Bénard convection and Euler quadrant flows. Most remarkably, when applied to the Gray-Scott reaction-diffusion system—a chemical system where fluid vortices are not physically defined—the same intervention produced spiral patterns. Results which suggest that the model may have learned an abstract representation of "rotation" or "spiralling" that transcends any specific physical domain.

A key open problem — and a limitation of this work — is the question of the physicality of the steered results. As a general rule we find that inclusion of spatial dimensions in a transfer steering tensor results in more physically unrealistic results in the secondary visualisation fields, but when those dimensions are averaged over and dropped the results often appear physically plausible. Having said that, it is hard to define what a "reasonable" and "physically plausible" result should actually look like in this context since by its very nature we are aiming to introduce a physical feature into a system where that feature should not naturally be present. Further work is needed to definitively answer the question; for now, we have noted some encouraging observations:

1. Steering modifies all fields in the simulation in a physically self-consistent manner.
2. The effects of steering tend to show up in physically expected places (e.g. vortices appearing along shock fronts in Euler, gliders transforming into spirals in Gray-Scott; plus more diffuse flow and smoother gradients in the diffusion steering experiment.).

We have also observed that the distance of the simulation's initial conditions from the desired physical regime is important. For example, when performing positive vorticity steering on laminar shear flow with varying Reynolds and Schmidt numbers, we find that the further the initial conditions are from the vortex regime, the harder it will be (i.e. the higher alpha will need to be) to finally force a vortex into the simulation. By the time you succeed (if you do at all), the high alpha will have caused the y velocity and pressure fields to become completely distorted and unphysical. Conversely, if the initial conditions are such that the simulation is on the brink of the vortex regime, only a small nudge will be required, leaving all of the fields looking far more reasonable.

## 5 Conclusion

Our work demonstrates that interpretability techniques from LLMs can be successfully adapted to scientific foundation models. By calculating the difference in mean activations between contrasting physical regimes—a method we term "delta steering"—we isolate single directions in the latent space of the physics foundation model that correspond to specific physical concepts like vorticity, diffusion, and simulation speed. Injecting these directions during inference provides direct, causal control over the model's predictions, allowing us to manipulate physical behaviours in silico.

These findings provide early evidence that scientific foundation models, much like LLMs, develop abstract, domain-general representations of fundamental concepts. The success of our simple difference-of-means approach suggests that these core physical concepts are represented strongly and linearly in the model's activation space, aligning with the linear representation hypothesis.

The emergence of steerable, interpretable features in scientific models has significant implications. It increases our confidence that these models are learning genuine physical principles rather than superficial correlations. It opens new avenues for interacting with simulations: we can perform counterfactual exploration ("what if this flow were more diffuse?"), correct simulation errors in real-time, and audit a model's understanding of physics by testing its response to targeted interventions.

## Acknowledgments

We thank Schmidt Sciences and the Simons Foundation for their support. We are grateful to Shirley Ho, Francois Lanusse, and the rest of the PolymathicAI team for valuable discussions and feedback. We also thank Neel Nanda, Rich Turner, and Max Welling for valuable discussions.

## References

[1] Andy Arditi, Oscar Obeso, Leo Rimsky, Jérémy Scheurer, Paul Christiano, and Ryan Leahy. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[2] Sanjeev Arora et al. Linear algebraic structure of word embeddings. *arXiv preprint arXiv:1805.05918*, 2018.

[3] Andres M Bran et al. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[4] Trenton Bricken et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Sascha Girish, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Yadi Cao, Yuxuan Liu, Liu Yang, Rose Yu, Hayden Schaeffer, and Stanley Osher. Vicon: Vision in-context operator networks for multi-physics fluid dynamics prediction, 2025.

[7] Seyone Chithrananda et al. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.06863*, 2020.

[8] Hoagy Cunningham et al. Sparse autoencoders find highly interpretable directions. *Transformer Circuits Thread*, 2023.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[11] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[12] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2021.

[13] Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. Dpot: Auto-regressive denoising operator transformer for large-scale pde pre-training, 2024.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[15] Tom Henighan et al. Superposition, memorization, and double descent. *Transformer Circuits Thread*, 2022.

[16] Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes, 2024.

[17] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[18] Lavender Yao Jiang et al. Health system-scale language models are all-purpose prediction engines. *Nature*, pages 1–6, 2023.

[19] Henry W Leung and Jo Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *arXiv preprint arXiv:2306.07329*, 2023.

[20] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

[21] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[22] Michael McCabe, Bruno Blancard, Liam Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

[23] Kevin Meng et al. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35, 2022.

[24] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[25] Tuan Nguyen et al. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2311.16446*, 2023.

[26] Tung Nguyen et al. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[27] Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Keaton Burns, Stuart Dalziel, Drummond Fielding, Daniel Fortunato, Jared Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan Nixon, Jeff Shen, Romain Watteaux, Bruno Blancard, François Rozet, Liam Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. *arXiv preprint arXiv:2412.00568*, 2024.

[28] Chris Olah et al. Zoom in: An introduction to circuits. *Distill*, 5(3):e24, 2020.

[29] Kiho Park, Yo Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[30] Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, and Shirley Ho. Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, June 2024.

[31] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library, 2019.

[32] Adam Scherlis et al. Polysemanticity and capacity in neural networks. *arXiv preprint*, 2023.

[33] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

[34] Makoto Takamoto et al. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

[35] Adly Templeton et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

[36] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

[37] Ashish Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Jesse Vig et al. Causal mediation analysis for interpreting neural nlp models. *arXiv preprint arXiv:2004.12265*, 2020.

[39] Kevin Wang et al. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[40] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

# A Additional Plots

## A.1 Spatial Dimensions for Within-Domain Steering

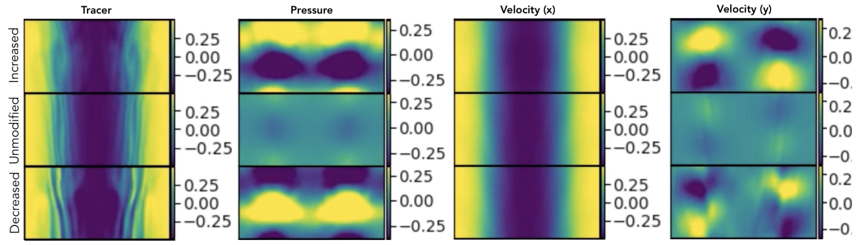

**Figure 7:** *Trace, pressure, x velocity and y velocity for $\Delta_{vortex}$ injection into shear flow laminar regime with (top) $\alpha = 0.2$, (middle) $\alpha = 0.0$ and (bottom) $\alpha = -0.2$. Frame: 64.*

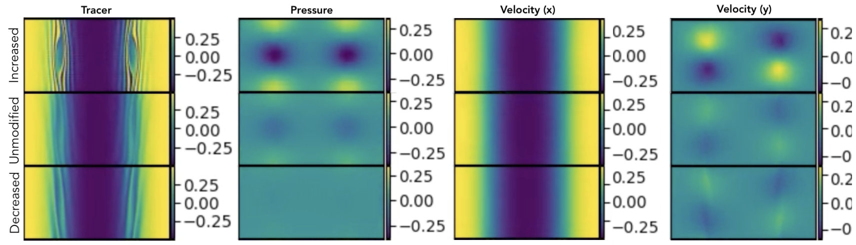

**Figure 8:** *Trace, pressure, x velocity and y velocity for $\overline{\Delta}_{vortex}$ injection into shear flow laminar regime with (top) $\alpha = 0.7$, (middle) $\alpha = 0.0$ and (bottom) $\alpha = -0.5$. Frame: 64.*

In fig. 7 we visualise the effect of shear flow vorticity steering with spatial dimensions on related physical fields beyond the primary tracer field. It is clear that the activation steering does not merely alter the tracer output, but introduces coordinated changes across the pressure and velocity fields that on the surface seem consistent with real vortex dynamics. Coordinated changes which evolve and persist throughout the rollout window (frame 64 being the final frame of the rollout window). These modifications are bidirectional, with positive and negative steering causing opposite changes to the pressure and y velocity fields. It should be noted that this result is not consistent with the interpretation of a linear steering response, where positive steering amplifies the concept feature and negative steering suppresses it. This is because while the positive steering result looks as one might expect, the negative steering result does not, since if the negative steering was causing a decrease in vorticity we would expect a flattening of the pressure field, not an inversion.

In fig. 8 we visualise the effect of shear flow vorticity steering *without* spatial dimensions. Two key differences can be see with the previous fig. 7. Firstly, the primary tracer displays a smaller, less well-formed vortex in the positive direction (despite a higher value of $\alpha$ than fig. 7) and the negative direction displays a more natural-looking laminar flow. Secondly, the changes to the secondary fields are more subtle in the positive direction and no longer inverted in the negative direction. A result which *is* consistent with the interpretation of a linear steering response.

These results are comparable to the Rayleigh-Bénard results in fig. 5, and other experiments we have performed:

- When spatial dimensions are averaged ($\overline{\mathbf{\Delta}}_f$) over results are consistent with the interpretation of a linear steering response.
- When this does not occur and spatial dimensions are left in place ($\mathbf{\Delta}_f$) steering has a tendency to induce mirrored field changes in positive and negative steering.
- $\overline{\mathbf{\Delta}}_f$ steering usually produces more natural-looking results.
- $\mathbf{\Delta}_f$ steering, on the other hand, has the capability to produce more extreme changes, such as the creation of large, intricate vorticities, but at the expense of a less natural-looking final result across all fields.

### A.2 Shear Flow Diffusion Steering – Additional Fields

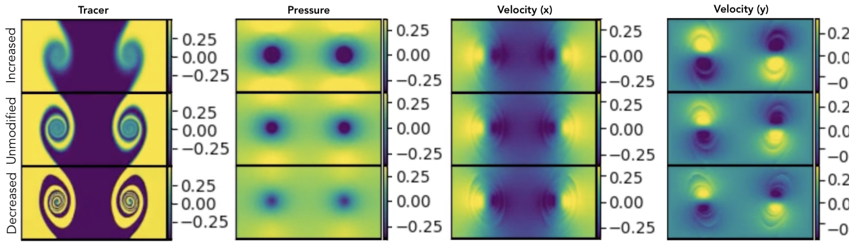

**Figure 9:** *Trace, pressure, x velocity and y velocity for $\mathbf{\Delta}_{diffusion}$ injection into shear flow vortex regime with (top) $\alpha = 0.2$, (middle) $\alpha = 0.0$ and (bottom) $\alpha = -0.2$. Frame: 28.*

## B Data Files

### B.1 Simulations

**Table 1:** *Simulation Data Files*

| Simulation | Datafile |
|---|---|
| fig. 2 (all) | shear_flow_Reynolds_5e4_Schmidt_2e-1 |
| fig. 3 (all) | shear_flow_Reynolds_5e5_Schmidt_1e0 |
| fig. 4 (left) | shear_flow_Reynolds_5e4_Schmidt_2e-1 |
| fig. 4 (right) | shear_flow_Reynolds_5e4_Schmidt_5e-1 |
| fig. 5 (all) | rayleigh_bernard_rayleigh_1e9_prandtl_10 |
| fig. 6 (left) | euler_multi_quadrants_openBC_gamma_1.33_H2O_20 |
| fig. 6 (middle) | euler_multi_quadrants_openBC_gamma_1.404_H2_100_Dry_air_-15 |
| fig. 6 (right) | gray_scott_reaction_diffusion_gliders_F_0.014_k_0.054 |
| fig. 7 (all) | shear_flow_Reynolds_1e5_Schmidt_2e-1 |
| fig. 8 (all) | shear_flow_Reynolds_1e5_Schmidt_2e-1 |
| fig. 9 (all) | shear_flow_Reynolds_5e4_Schmidt_2e-1 |

### B.2 Steering Tensors

**Table 2:** *Vortex Regime Simulations*

| Name | Reynolds | Schmidt |
|---|---|---|
| Shear Flow | 1e4 | 1e-1 |
| Shear Flow | 1e4 | 2e-1 |
| Shear Flow | 1e4 | 2e0 |
| Shear Flow | 1e4 | 5e-1 |
| Shear Flow | 1e4 | 5e0 |
| Shear Flow | 1e5 | 1e-1 |
| Shear Flow | 1e5 | 1e0 |
| Shear Flow | 1e5 | 2e0 |
| Shear Flow | 1e5 | 5e-1 |
| Shear Flow | 5e4 | 1e-1 |
| Shear Flow | 5e4 | 1e0 |
| Shear Flow | 5e4 | 1e1 |
| Shear Flow | 5e4 | 2e0 |
| Shear Flow | 5e4 | 5e-1 |
| Shear Flow | 5e4 | 5e0 |
| Shear Flow | 5e5 | 1e0 |
| Shear Flow | 5e5 | 2e-1 |
| Shear Flow | 5e5 | 5e0 |

**Table 3:** *Laminar Regime Simulations*

| Name | Reynolds | Schmidt |
|---|---|---|
| Shear Flow | 1e4 | 1e0 |
| Shear Flow | 1e4 | 1e1 |
| Shear Flow | 1e5 | 1e1 |
| Shear Flow | 1e5 | 2e-1 |
| Shear Flow | 1e5 | 5e0 |
| Shear Flow | 5e4 | 2e-1 |
| Shear Flow | 5e5 | 1e-1 |
| Shear Flow | 5e5 | 1e1 |
| Shear Flow | 5e5 | 2e0 |
| Shear Flow | 5e5 | 5e-1 |

**Table 4:** *High Diffusion Simulation*

| Name | Regime | Reynolds | Schmidt | Viscosity | Diffusion |
|---|---|---|---|---|---|
| Shear Flow | single vortex | 5e4 | 2e-1 | 2.00e-05 | 1.00e-04 |

**Table 5:** *Low Diffusion Simulation*

| Name | Regime | Reynolds | Schmidt | Viscosity | Diffusion |
|---|---|---|---|---|---|
| Shear Flow | single vortex | 5e4 | 1e1 | 2.00e-05 | 2.00e-06 |

**Table 6:** *High Speed Simulations*

| Name | Reynolds | Schmidt | dt_stride |
|---|---|---|---|
| Shear Flow | 1e4 | 1e0 | 2 |
| Shear Flow | 1e4 | 1e1 | 2 |
| Shear Flow | 1e5 | 1e1 | 2 |
| Shear Flow | 1e5 | 2e-1 | 2 |
| Shear Flow | 1e5 | 5e0 | 2 |
| Shear Flow | 5e4 | 2e-1 | 2 |
| Shear Flow | 5e5 | 1e-1 | 2 |
| Shear Flow | 5e5 | 1e1 | 2 |
| Shear Flow | 5e5 | 2e0 | 2 |
| Shear Flow | 5e5 | 5e-1 | 2 |

**Table 7:** *Low Speed Simulations*

| Name | Reynolds | Schmidt | dt_stride |
|---|---|---|---|
| Shear Flow | 1e4 | 1e0 | 1 |
| Shear Flow | 1e4 | 1e1 | 1 |
| Shear Flow | 1e5 | 1e1 | 1 |
| Shear Flow | 1e5 | 2e-1 | 1 |
| Shear Flow | 1e5 | 5e0 | 1 |
| Shear Flow | 5e4 | 2e-1 | 1 |
| Shear Flow | 5e5 | 1e-1 | 1 |
| Shear Flow | 5e5 | 1e1 | 1 |
| Shear Flow | 5e5 | 2e0 | 1 |
| Shear Flow | 5e5 | 5e-1 | 1 |