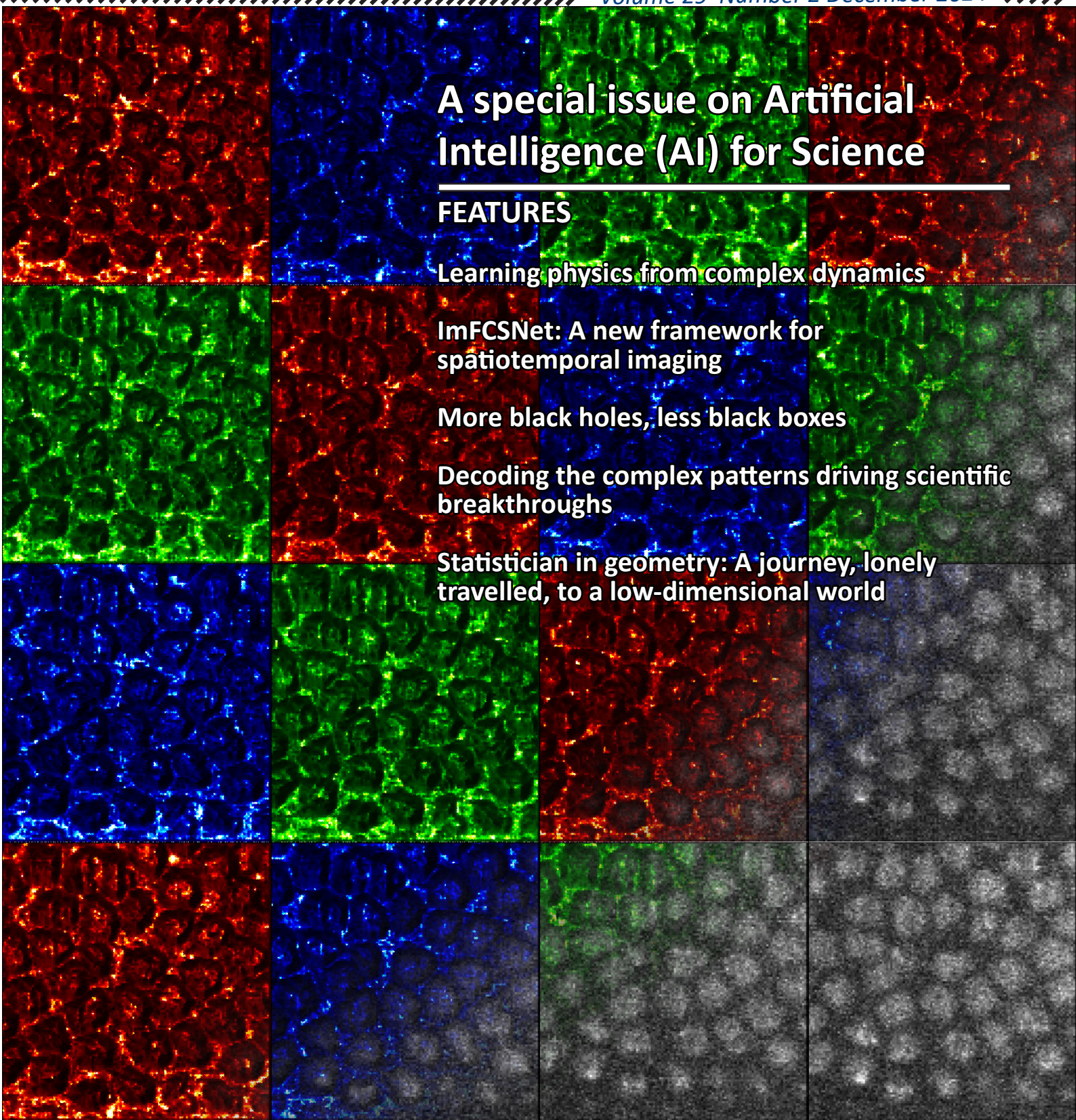
NUS National University of Singapore | Faculty of Science

# Advances in Science

Volume 29 Number 2 December 2024



## A special issue on Artificial Intelligence (AI) for Science

FEATURES

# Table of Contents

## RESEARCH FEATURES




# Advances in Science

The Faculty of Science conducts basic and applied experimental, theoretical and simulation research over a broad spectrum of science, mathematics and technology domains. We cover most of the key fields in biological sciences, chemistry, physics, pharmacy, pharmaceutical sciences, food sciences, natural history, mathematics, statistics and data science.

Advances in Science is published online twice a year. It is written for a broad scientific audience interested to keep up with some of the key areas of science pioneered by researchers at the Faculty of Science.





For further information on the research in this newsletter, please contact:

Editor: SOH Kok Hoe (kok.hoe@nus.edu.sg)
Consultants: Professor CHEN Wei (chmcw@nus.edu.sg)
Associate Professor LIOU Yih-Cherng (dbslyc@nus.edu.sg)

Dean's Office, Faculty of Science
National University of Singapore
Blk S16, Level 9, Science Drive 2
Singapore 117546


For the latest research news, please refer to:
URL: www.science.nus.edu.sg/research/research-news

**On the cover:** Deep learning enabled functional fluorescence imaging. The grayscale images display the raw channel intensities taken in time. The red, green, and blue channels represent parameter maps, extracted by specially developed convolutional neural networks.

# Learning physics from complex dynamics

Machine learning offers a new way to discover macroscopic physical laws of complex dynamical systems

## Introduction

Laws of physics are succinct descriptions of nature that are amenable to human understanding and offer deep insight into the physical phenomena. However, for a highly complex system where there is little prior knowledge, it may be very difficult to discover such descriptions. With advent of data-driven technology, machine learning offers an alternative approach to achieve this.

There are two generic approaches to describe the dynamics of complex systems: the microscopic approach which models the fine-grained dynamics of each component of the system and their interactions (e.g. molecular dynamics), and the macroscopic approach which focuses on the evolution of observable quantities that represent the coarse-grained, large-scale behaviour (e.g. thermodynamics). The microscopic approach has the advantage of accuracy, but it is computationally prohibitive for most realistic physical systems of interest. On the other hand, the macroscopic approach, which provides a relation between a small number of observable and interpretable parameters, thus being very efficient even for highly complex systems. However, constructing such descriptions require deep theoretical knowledge or extensive experimentation, and they have only been constructed for a handful of idealistic systems. With the growing availability of computational and experimental data, it would be extremely beneficial if this process can be automated.

In their recent work [1], Dr Li Qianxiao and his colleagues developed a machine-learning based method to automatically extract macroscopic physical principles from observations of trajectory data, giving rise to an alternative approach to construct thermodynamic descriptions for dissipative systems.

## Methodology

The approach is called Stochastic OnsagerNet (Fig. 1), which integrates machine learning with statistical physics to develop macroscopic models for complex, stochastic, and dissipative dynamical systems. Starting with certain macroscopic quantities of interest, the method introduces a small set of auxiliary closure variables, which facilitate the creation of a data-driven yet physically grounded dynamical equation governing their collective time evolution. This approach is inherently interpretable, drawing from Lars Onsager's work on non-equilibrium statistical physics [2]. For example, it generates an effective potential analogous to free energy, yielding a data-driven equation of state. Unlike traditional phenomenological models that assume simplified forms for free energy, here the free energy is approximated using neural networks specifically designed to learn from data.

## Application to polymer stretching dynamics

The effectiveness of the method is demonstrated by studying a well-known problem in polymer rheology: the stretching of a long polymer chain under externally applied forces. While the microscopic physics governing each particle in the polymer chain is straightforward and well understood, predicting the macroscopic behaviour, such as the chain's length over time, poses significant challenges. This difficulty arises from the complex interactions between thermal noise, polymer configurations, and the applied forces. Stochastic OnsagerNet is shown to automatically generate accurate and interpretable thermodynamic models, enabling the understanding, prediction, and even manipulation of polymer behaviour. The model's accuracy was further validated through experiments involving electrokinetic stretching of DNA molecules in a microfluidic channel. The results showed that the learned model not only captures the heterogeneity in the stretching dynamics but also accounts for the fluctuations around the stretched state.

## Discussion

The method's generality allows it to be applied to a wide range of stochastic and dissipative dynamical processes, such as the self-healing of materials, the relaxation of magnetic systems, and the spread of diseases. Its core strength lies in providing a systematic framework for analysing the macroscopic behaviour of complex dynamics. However, the method does have limitations. It is specifically designed for noisy, dissipative systems that can be modelled using Onsager-type dynamical equations, which restricts its effectiveness in capturing other types of dynamics, such as chaotic or highly excited systems. Moreover, the approach depends heavily on large, high-quality observational datasets for training, which necessitates high-throughput experiments, like those developed for polymer stretching dynamics.

An exciting future prospect for the method is to "close the loop" between experimentation, learning, and control. This would involve developing data-driven strategies for manipulating the learned macroscopic dynamics, such as adjusting external conditions to alter polymer behaviour. By doing so, the method could enable more efficient and targeted data collection while also

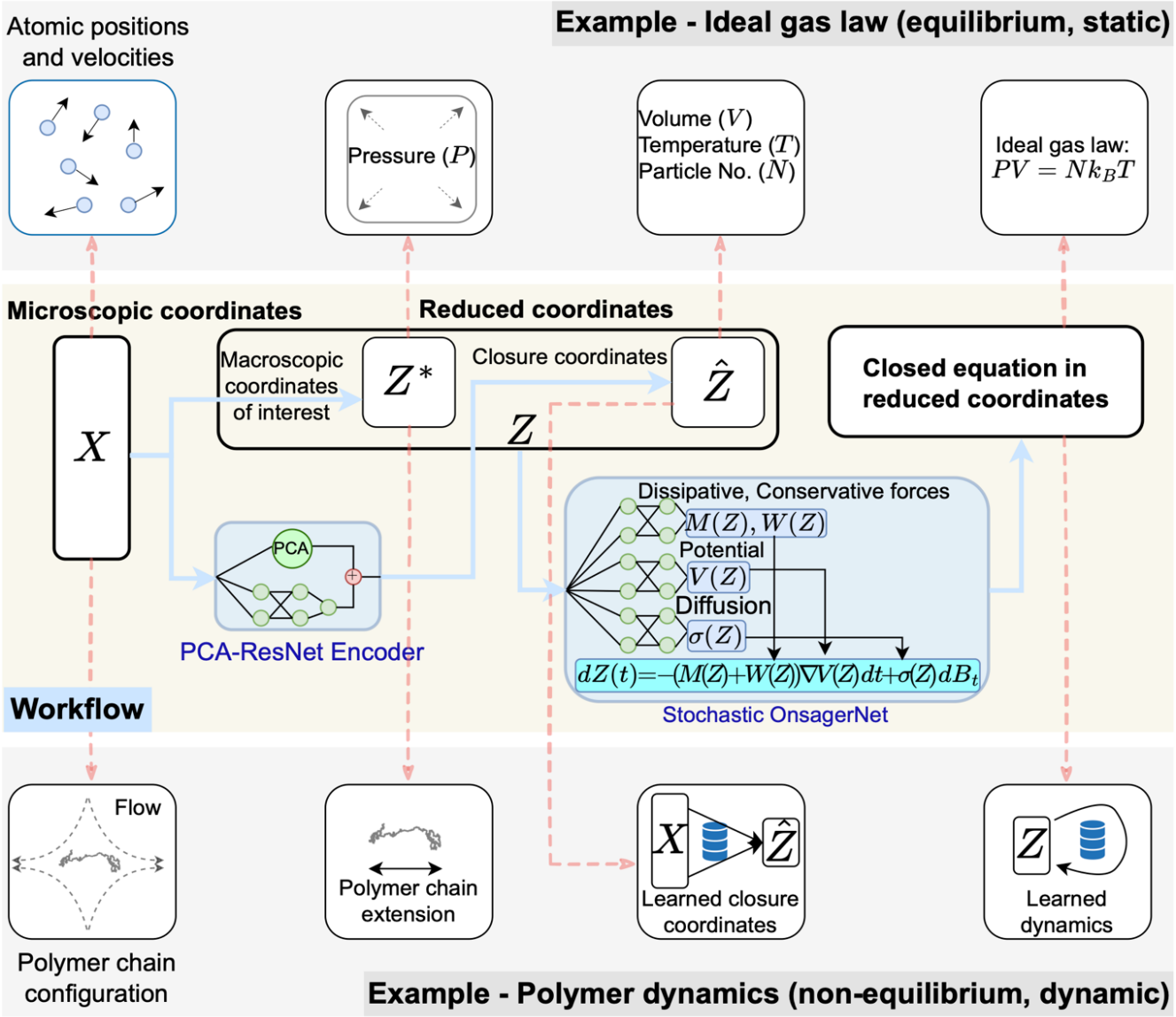
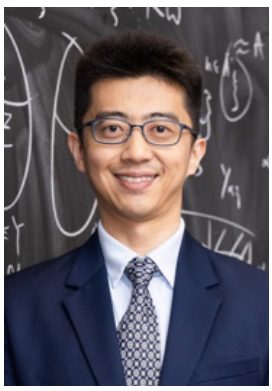



Figure 1: Overall workflow of Stochastic OnsagerNet. Given a complex dynamical process described by microscopic coordinates X, we are interested in the dynamics of macroscopic coordinates Z*. This requires the construction of closure coordinates Ẑ and a closed equation for the combined reduced coordinates Z = (Z*, Ẑ). The classical ideal-gas law is one illustration of this process (top panel); for general non-equilibrium, dynamic systems such as polymers (bottom panel), carrying out this workflow from theory is challenging. Our method (middle panel) simultaneously constructs the closure coordinates and models their temporal evolution using a combination of the generalized Onsager principle and deep learning. [Credit: Nature Computational Science]

supporting real-world applications that require fast and accurate control of macroscopic behaviours.

**For more details, please visit:**
**https://blog.nus.edu.sg/qianxiaoli/**

**Qianxiao LI is an Assistant Professor in the Department of Mathematics, and a principal investigator in the Institute for Functional Intelligent Materials, at the National University of Singapore. He graduated with a B.A. in mathematics from the University of Cambridge and a Ph.D. in applied mathematics from Princeton University. His research interests include the interplay of machine learning and dynamical systems, control theory, stochastic optimisation algorithms and data-driven methods for science and engineering.**



References

[1] Chen XL, et al., "Constructing custom thermodynamics using deep learning" Nature Computational Science Volume: 4 Issue: 1 page 66 - 85 DOI: 10.1038/s43588-023-00581-5 Published: 2024.

[2] Onsager L, "Reciprocal relations in irreversible processes. I.", Physical Review Volume: 37 Issue: 4 Page: 405-426 DOI:10.1103/PhysRev.37.405 Published: 1931.

# ImFCSNet: A new framework for spatiotemporal imaging

Analysing biomolecular processes through deep learning is faster while requiring less data

## Introduction

Imaging is an indispensable building block of modern science. Since the introduction of the compound microscope, imaging has become a standard tool in many scientists' arsenals. The ability to see beyond the magnification limit of the naked eye has been key to unlocking many insights that would have otherwise been inaccessible.

Most early microscopy studies focused on the analysis of structure. However, there is only so much information you can glean from a single static image. One notable finding was in 1827, when botanist Robert Brown used microscopy to observe how pollen moves randomly while in water, something that was only possible with observation through time. This then formed the basis of Albert Einstein's seminal work on the Brownian diffusion.

In the modern day, technological advances in camera and detector technology as well as lasers as illumination sources, provide high spatiotemporal resolution at single-molecule sensitivity in microscopy. By observing systems at high spatiotemporal resolution over long times we gain information on structure and dynamics over many scales in a single measurement.

Similarly, the field of deep learning has gone through significant transformations over the past 80-odd years. Starting from the basic McCulloch-Pitts model of a single neuron, computer scientists have designed digital neural networks, which are capable of "learning" to master a variety of different tasks through an iterative process called "training". While early neural network models were constrained by the computational limitations of the early 1980s, the same theoretical underpinnings have stood the test of time, leading to the artificial intelligence boom of the 2020s, with powerful models such as OpenAI's GPT and Google's Gemini series becoming synonymous with "artificial intelligence" in the public consciousness.

One particularly interesting subset of deep learning is computer vision, which can trace its roots directly to the field of imaging. We have come a long way from the basic character recognition LeNet of 1989, with modern models showing strong capabilities in visual understanding.

In this article, we describe our attempts to incorporate deep learning approaches into our imaging pipelines, granting us more insights from image time series than would be possible with static images.

## Fluorescence correlation spectroscopy

Our work primarily focuses on fluorescence correlation spectroscopy (FCS), a mature, widely used statistical analysis tool designed to explain underlying molecular processes or interactions within a sample with single molecule sensitivity.

Originally, FCS was conducted as single-point experiments, but modern cameras with fast frame capture capabilities have made it possible to do Imaging FCS, which extends classical point-based FCS to cover larger observation areas, allowing us to probe how molecular dynamics change over biologically relevant spatial scales.

FCS has been proven to be capable of extracting information from signal fluctuations that otherwise appear like noise. As shown by Manfred Eigen in his Nobel Prize lecture, the fluctuations from equilibrium present in any measurement contain information about the underlying system. In the context of FCS, this information is represented as the autocorrelation function (ACF), which describes the similarity (the titular "correlation") across different lag times. The ACF can then be fitted to a mathematically derived fit function to determine characteristics of the system, such as the number of fluorescently labelled particles passing through the volume, the diffusion coefficient, or speeds at which these particles are moving.

However, there are limitations to the ACF-based fitting approach. As a statistical fitting process, the FCS measurements need to be sufficiently long to obtain robust estimates of the underlying dynamics. In addition, FCS requires as an input, a model of the molecular process to be observed. However, in many cases for these models, no analytic fitting function can be derived.

## ImFCSNet

In our recent work, we show that deep learning can serve as a viable alternative for Imaging FCS data fitting. Imaging FCS works with video-like data, making it inherently compatible with computer vision techniques. We designed the ImFCSNet architecture, a custom convolutional neural network (CNN) architecture designed specifically for investigating the diffusion characteristics in Imaging FCS data [1]. Through experimental verification, we find that ImFCSNet requires less data, is robust to defocusing, and is magnitudes faster than conventional FCS fitting.

Here, we describe three notable aspects of ImFCSNet's design, and how we believe they can help advance the field of FCS imaging.

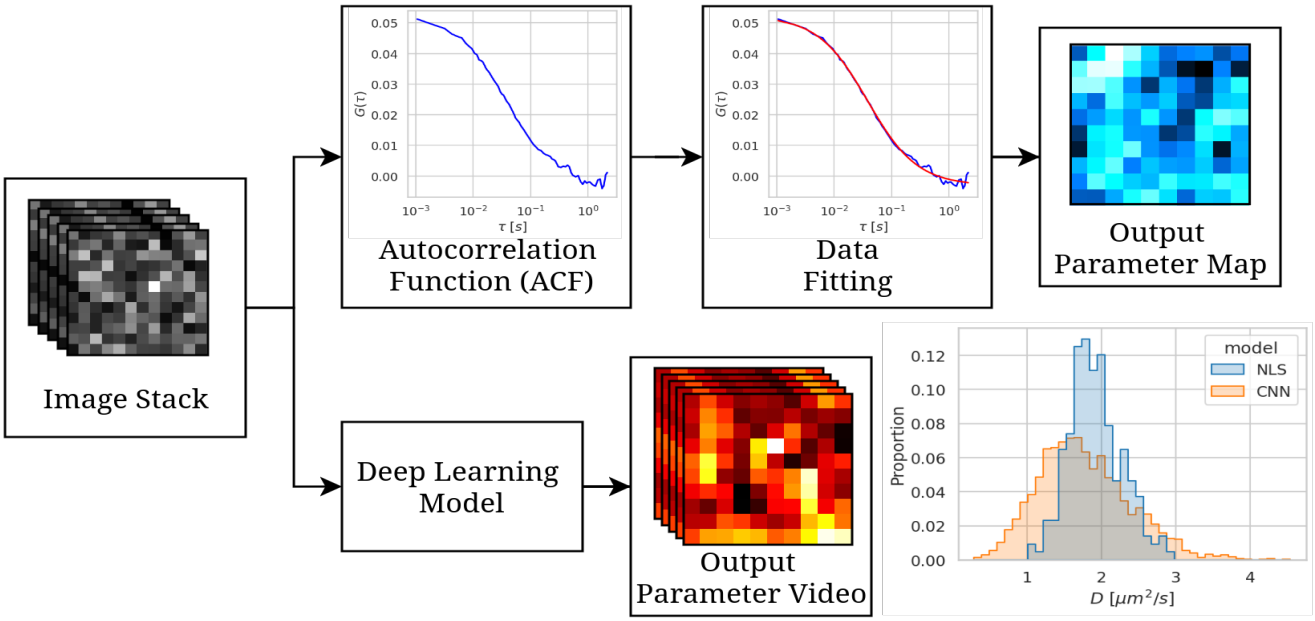



Figure 1: Workflow for the ImFCSNet, which can efficiently extract molecular dynamics parameters from raw imaging data.

First, ImFCSNet is trained end-to-end with no intermediate data processing steps. CNNs are capable of learning complex, hierarchical mappings between the input data and the provided training labels. With ImFCSNet, we use simulations as a source of training data. This circumvents the data-hungry nature of deep learning, while also giving us labels that are verifiably true. Unlike FCS, we skip the ACF representation, and train to directly extract the underlying target parameters directly from the raw input data. This allows ImFCSNet to sidestep the limitations of ACF fitting with regards to minimum measurement time, and the fit model dependence, and unlocks the possibilities of understanding how molecular dynamics evolve at shorter time scales.

Next, ImFCSNet is extensible. While our initial targets with ImFCSNet revolved around the diffusion coefficient, we can leverage the flexibility of deep learning to extend this to further applications. Our recent work shows that the same ImFCSNet architecture initially proposed for determining the diffusion coefficient can be adapted to predict the number of particles with minimal changes to the training protocol, which is compelling evidence that the same "recipe" is versatile enough to be applied to different tasks [2].

Finally, ImFCSNet is made to be easy to use and adapt. It does not matter if a tool is proven to be superior if it is inaccessible. While CNN training is generally a complex web of abstractions and design decisions, ease of use should always be considered. This is particularly true with ImFCSNet, as simulation parameters vary across different acquisition setups. We take inspiration from widely used deep learning frameworks and focus on accessibility with our codebase. In theory, any user should be able to train and use their own custom ImFCSNet tuned for their system, even without deep learning expertise.

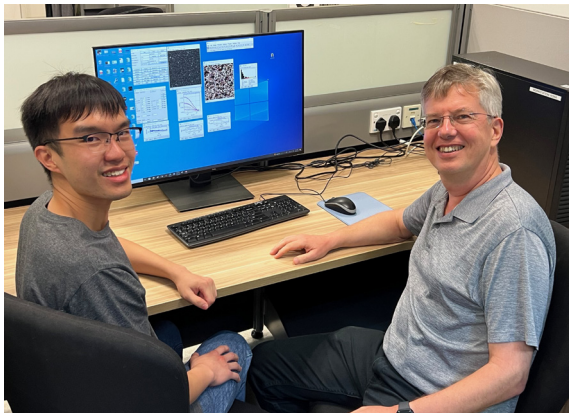
## Future work

While ImFCSNet has proven to be a viable tool, there is still work to be done. FCS is capable of extracting much more information beyond the diffusion coefficient and concentration, and ImFCSNet still has some ways to go before it can fully replicate FCS' capabilities, let alone surpass it. However, our work shows that deep learning can supplement Imaging FCS, opening the door to new possibilities in data analysis.

**For more details, please visit: https://www.dbs.nus.edu.sg/staffs/thorsten-wohland/**

**Thorsten WOHLAND is a Professor with the Departments of Biological Sciences and Chemistry at the National University of Singapore (NUS). He is also the director for the Centre for Bioimaging Sciences, NUS. He works on the development and application of fluorescence spectroscopy.**

**SIM Shao Ren is a Ph.D. student in Professor Wohland's laboratory. He is a computational biologist with a focus on deep learning, and was previously a deep learning engineer working on biomedical computer vision.**



References

[1] WH Tang, et al., "Deep learning reduces data requirements and allows real-time measurements in imaging FCS" Biophysical Journal Volume: 123 Issue: 6 Page:655-666 DOI: 10.1016/j.bpj.2023.11.3403 Published: 2024.

[2] T Wohland, et. al., "FCS videos: Fluorescence correlation spectroscopy in space and time" Biochimica ET Biophysica Acta-General Subjects Volume: 1868 Issue: 11 DOI: 10.1016/j.bbagen.2024.130716 Published: 2024.

# More black holes, less black boxes

Statistically principled learning for inverse problems in gravitational-wave astronomy

## Background

Astronomy is the modern embodiment of humankind's enduring fascination with the Universe. Over the past century, technological advances have expanded our view of the cosmos (from visible light to the entire electromagnetic spectrum), and with it the range and detail of our astronomical observations. More recently, however, the historic 2015 detection of gravitational waves (GWs) from two merging black holes has opened up an entirely different channel through which to study the Universe.

The nascent field of GW astronomy allows us to observe phenomena we would otherwise be blind to, and provides exciting new insights into others that are already visible. GW signals from binary systems of black holes or neutron stars are now routinely detected by ground-based interferometers such as LIGO in the kilohertz frequency band. Pulsar timing arrays are also poised to discover nanohertz signals from supermassive black holes, while space-based observatories such as the ESA-NASA mission LISA will cover the source-rich millihertz range in the next decade.

An important aspect of GW astronomy is solving inverse problems, i.e., determining the properties of astrophysical sources from their GW signals. This involves constructing complex forward models for possible signals by solving the equations of general relativity, as well as using these forward models in data-analysis algorithms to extract and characterise actual signals in detector data. Many challenges hinder both such tasks, which calls for researchers to devise novel computational and statistical techniques in their solutions.

Machine learning (ML) is increasingly used to confront said challenges – although it faces unique hurdles in this field, such as noise-dominated data and the need for high precision in modelling. To establish the scientific viability of ML methods in solving GW inverse problems, it is also crucial to clarify how they relate to the existing theoretical and computational framework for GW data analysis, which is already rigorously founded on well-understood principles from the broader fields of signal processing and Bayesian inference.

## The status of ML in GW astronomy

ML (in particular deep learning) has had a broad impact on the field of GW astronomy. Since a triad of initial papers from 2018 exploring the usage of deep learning in GW data analysis, there has been an exponential explosion of literature on the topic. Early work dealt only with the extraction and classification of GW signals in data. However, despite the considerable attention paid to such techniques, most are still not accepted as standard in GW astronomy – being either statistically unprincipled, or uncompetitive, or both.

In 2019, my collaborators and I were among the first to showcase the viability of deep learning for both aspects of the GW inverse problem: forward modelling [1] and Bayesian inference [2]. Our techniques explicitly augment or emulate the existing GW analysis framework, and are thus more defensible on rigour and principle. This stands in contrast to much of the literature on deep learning in science, which often adopts a "black-box" approach to learning that prioritises fast and accurate estimation or prediction – but with less regard for the quantification of uncertainty, or the statistical significance of conclusions.

## Forward modelling: Learning to describe GW signals

The most common sources for GW astronomy involve the gradual inward spiralling and eventual merger of a binary system whose components are extremely massive and dense objects, e.g., black holes. Such systems are strongly gravitating with highly non-linear dynamics, and can only be modelled accurately through the difficult task of solving Einstein's equations of general relativity. Forward models for the predicted GW signals ("waveforms") from these sources then typically contain numerical calculations that are highly precise but computationally expensive. This is at odds with the nature of the data-analysis algorithms used in GW astronomy, which rely heavily on Monte Carlo simulations of waveforms – i.e., generating them in bulk and rapidly.

Fast but sufficiently accurate fits to waveform calculations are thus often required, in order to construct approximate forward models that can feasibly be used in data analysis. In [1], we were the first to harness the powerful fitting capability of deep neural networks for this purpose, by combining them with a classical dimensional-reduction technique that has been used in GW modelling to good effect. We trained a neural network to reproduce the reduced representation of waveforms from a simple class of binary source, and verified that it could attain comparable speed and accuracy to other (non-learning) fitting algorithms. This work showed that neural networks are a viable option for fitting tasks in GW modelling, and posited that they would scale well to problems of higher dimensionality and complexity.

## Bayesian inference: Learning about GW signals in data

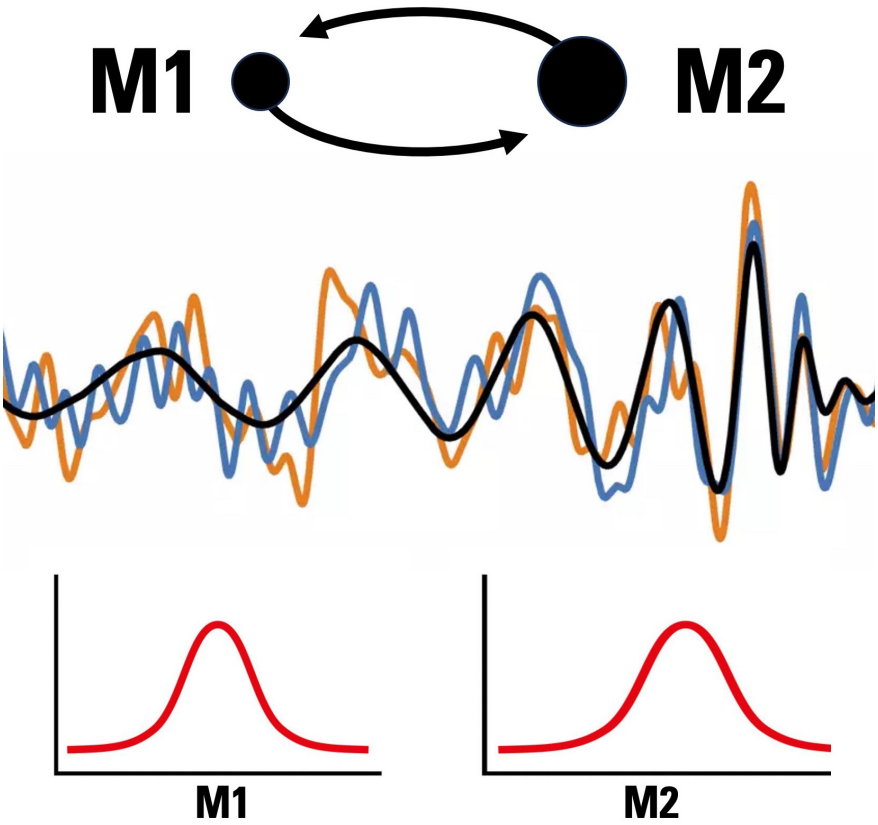



Figure 1: In the archetypal inverse problem for gravitational-wave astronomy, Einstein's equations are used to model an astrophysical binary (top) and to predict its gravitational-wave signal (middle; black curve), which is then compared against detector data (middle; coloured curves) to infer source parameters such as the binary masses (bottom).

If a GW signal is present in detector data, the astrophysical parameters of its source can be inferred through posterior estimation – the mapping out of the posterior probability distribution for the parameters. In GW astronomy, this task must typically be performed with random sampling algorithms such as Markov chain Monte Carlo methods. Unfortunately, the posterior in GW data analysis is notoriously difficult to sample from, as it has probability tails that are both "heavy" (not exponentially bounded) and highly multi-modal. Furthermore, the raw cost of evaluating the forward model at each iteration of the sampling algorithm can also be a computational bottleneck. Both factors mean that Bayesian inference for even just a single GW source is a time-consuming endeavour, often taking hours to days.

One promising way that ML can help is simulation-based inference. This strategy "amortises" the cost of inference by training a model on a large set of simulated GW data streams (signal plus noise), in a way that it can later predict the posterior for an actual data stream almost instantly. In [2], we introduced such methods to GW data analysis by training a neural network to directly output the posterior, given some data containing a signal. The result also highlighted for the first time that neural networks could be used in a statistically principled way within the canonical framework of GW data analysis. Alongside two other independent research groups, our early work laid the foundations for the ongoing interest in applying such techniques to completely front-load the computational cost of GW inference.

## Future directions

While our research has demonstrated the viability of using ML and deep learning for scientific inverse problems in GW astronomy, much remains to be done before such methods become standard in the field. An important next step is to improve the efficacy of ML in fitting forward models, specifically by investigating the impact of feature engineering and architecture design in deep neural networks. Another direction is to use ML to perform posterior estimation not just efficiently, but also in ways that are robust against potential errors in forward models. The development of methods along these lines will encourage a wider acceptance and uptake of ML in GW astronomy, and can even help to address some of the greater data-analysis challenges posed by next-generation GW detectors.

**For more details, please visit: https://www.physics.nus.edu.sg/faculty/chua-alvin-jk/**

**Alvin Chua is an Assistant Professor in the Departments of Physics, Mathematics, and Statistics & Data Science at the National University of Singapore. He obtained his Ph.D. from the University of Cambridge, and has held postdoctoral appointments at the NASA Jet Propulsion Laboratory and the California Institute of Technology. His current research interests are in gravitational-wave astrophysics and data analysis; data science and machine learning; as well as applied and computational statistics.**

References

[1] Chua AJK*; Galley CR*; Vallisneri M*, "Reduced-Order Modeling with Artificial Neurons for Gravitational-Wave Inference", Physical Review Letters Volume: 122 Issue: 21 Article Number: 211101 DOI: 10.1103/PhysRevLett.122.211101 Published: 2019.

[2] Chua AJK*; Vallisneri M*, "Learning Bayesian Posteriors with Neural Networks for Gravitational-Wave Inference" Physical Review Letters Volume:124 Issue: 4 DOI:10.1103/PhysRevLett.124.041102 Published: 2020.

# Decoding the complex patterns driving scientific breakthroughs

Harnessing machine learning to decode disorder, complexity and adaptability in science

## Why should scientists care about machine learning?

Nature's creativity is astonishing, and yet it all stems from a relatively small set of fundamental laws and mechanisms, some already discovered and others still waiting to be uncovered. The pace of these discoveries has quickened thanks to advances in machine learning (ML), improved scientific instruments, and the explosion of high-quality data. Together, these elements are crystallising into a specialised kind of discovery-AI (artificial intelligence) designed to accelerate scientific breakthroughs. Unlike other forms of AI, which may be generalised or focused on tasks like image recognition, this discovery-AI is deeply grounded in precise, logical knowledge built up over centuries of scientific investigation. This makes it a unique opportunity to blend data, models, and algorithms in a way that could reshape how we think about AI.

This article explains how discovery-AI can help us wrestle (even tame) a "three-headed beast" in scientific discovery.

## The temperature-driven "three-headed beast" that rules our universe: Disorder, complexity, and adaptability

Disorder is a fundamental aspect of diversity and creativity. While some of us view disorder as a nuisance in our daily lives, disorder is crucial for the richness and complexity in our natural world. For instance, life evolved from the diverse modifications of and interactions among molecular building blocks. Without this inherent disorder, life would not adapt to changing environments, a key ingredient in evolution.

In materials science, controlled disorder creates complexity that enables new properties. Pure, perfect crystals lack the functionalities needed in modern electronics. Introducing disorder, such as adding different atoms or creating irregularities manipulates electricity and energy flow, enabling the development of devices like transistors and solar cells. This controlled disorder allows for adaptable and efficient materials.

Disorder is also an integral aspect of complexity and adaptability in natural ecosystems. Species interact with some unpredictability and diversity. This disorder ensures no two ecosystems are alike, fostering biodiversity and allowing ecosystems to thrive despite changing conditions.

Scientists typically measure disorder using the statistical concept of entropy. You can think of entropy as "accessible disorder". Disorder is a source of complexity and diversity in natural and human-made systems. It underpins adaptability, creativity, and resilience. Entropy, though often seen as a hindrance, is essential to system evolution, offering endless opportunities for discovery and innovation.

## More is different… But how?

Novel phenomena often emerge from the complex interactions of many simple elements. In physics, this idea is sometimes summed up by the phrase, "more is different". It means that when you look at a whole system, like a flock of birds or a chemical reaction, new behaviours emerge that you would not predict just by studying individual parts.

Complexity in natural systems arises when many interactions happen simultaneously. These interactions are not simple or straightforward, which makes it challenging to capture the system's behaviour with traditional mathematical models.

For instance, a group of interacting particles or cells becomes overwhelmingly complex with dozens of them, each influencing the others non-linearly and unpredictably. It is like trying to predict how a group of people will behave based on just a few conversations—they might share information, react to one another, or influence each other indirectly.

This intermediate scale of interactions is where complexity thrives. The system is not just a collection of simple parts, nor is it a smooth, easily averaged whole. The result is a level of unpredictability that is difficult to describe succinctly with equations alone. Traditional methods may struggle because they rely on approximations that smooth out the details where the true complexity lies.

Powerfully, ML now routinely extracts curious patterns in complex, adaptive systems that humans used to dismiss as disordered. ML excels at handling vast amounts of data and detecting patterns in non-linear and non-reciprocal interactions. By learning directly from the data, ML models identify subtle, complex relationships that drive system behaviour, offering insights where conventional approaches fail. It enables us to navigate system complexity, finding structure and predictability in otherwise chaotic and disordered systems.

## Discovering emergent phenomena from motif hierarchies in complex materials

To grasp how machine learning aids in understanding complex materials, let us consider language. Natural language exhibits patterns, such as word combinations like "machine learning" or "natural language". Linguists call

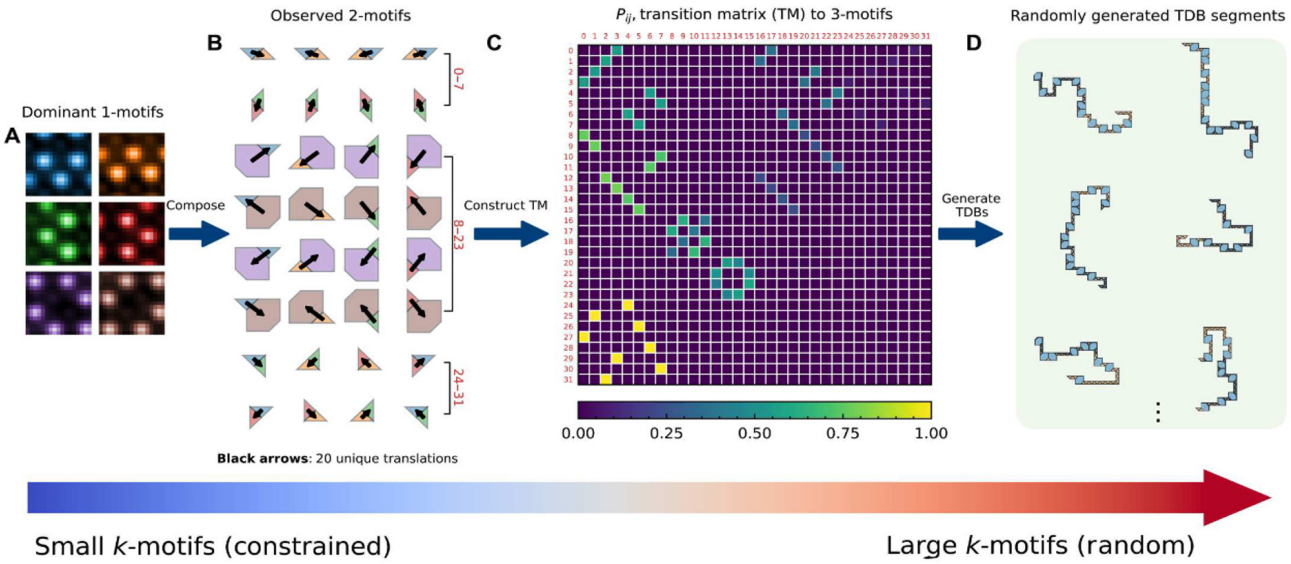



Figure 1: Hierarchy of structural motifs learned from a piezoelectric material – like an *n*-gram language model. (A) The structural motifs along a domain boundary can sometimes be reduced into simple shapes such as hexagons and triangles. (B) These motifs can be paired with various rules as 2-motifs. (C) Three-motifs can be extended as 3-motifs, abstracted here as a transition matrix where different 2-motifs (rows) continue into different 3-motifs (columns). (D) This transition matrix can be applied repeatedly, like the odds of a 6-sided dice, to generate random domain boundaries which match those observed experimentally. Figure taken from [2].

these patterns *n*-grams (where *n* counts the number of words in a sequence), which help grasp basic grammar rules. Despite their structured nature, language remains flexible, allowing for word rearrangements, synonyms, and meaningful sentence creation. This diversity adheres to language rules while fostering creativity and variation.

In materials science, atomic arrangements resemble "motifs", this is similar to words in language. Transition metal dichalcogenides, catalysts like polyoxometalates [1], and piezoelectric materials [2](see Figure 1) possess repeating atomic structures that form motifs. These motifs can combine and interact, creating a hierarchy of arrangements. Machine learning models can identify these atomic motifs, revealing their interactions and variations.

For instance, a motif might involve a specific metal atom arrangement surrounded by oxygen, repeating in polyoxometalates. Depending on these motif arrangements and interactions, the material exhibits diverse properties, akin to how rearranging words changes the meaning of a sentence. By decoding these patterns and hierarchies, we gain insights into the local atomic structure and the material's "story," enabling the design of tailored materials, such as improved catalysts or efficient electronic devices.

## Emergent phenomena in a hierarchy of spatiotemporal motifs in self-organised cellular systems

Imagine a busy cityscape with cars, people, and activities happening all at once. Without paying much attention to the details, this scene may look chaotic. But if you look closely, you will notice patterns such as how people move through streets, gather at specific times, or form queues. There are hidden rules and behaviours that bring order to what seems like chaos.

Cells in our body work in a similar way, but instead of observing cars and people, we are looking at the shapes and movements of biological cells. These cells often interact, cluster, and rearrange themselves, creating dynamic patterns. The challenge in studying this is that we cannot always measure the chemical signals or understand the exact "intentions" of each cell. Instead, we mostly observe their morphology—their shapes—and how they move and change over time. From these movements and changes, we must infer their functions and underlying mechanisms, often without knowing the specific interactions happening inside or between the cells.

Machine learning helps us decode these spatiotemporal patterns by identifying recurring shapes, movements, and arrangements of cells. Think of it like learning a language, where the shape and motion of each cell are like words, and how they change or interact over time forms a kind of grammar. By analysing this "language," ML can help us infer what might be driving these behaviours, even if we do not have direct access to the chemical signals or interactions (see Figure 2).

Uncovering these spatiotemporal motifs can lead to breakthroughs in our understanding of cell behaviour, which is crucial for fields like developmental biology, cancer research, and tissue engineering. Recognising patterns in a city helps improve traffic flow, and similarly, understanding how cells change shape and interact can give

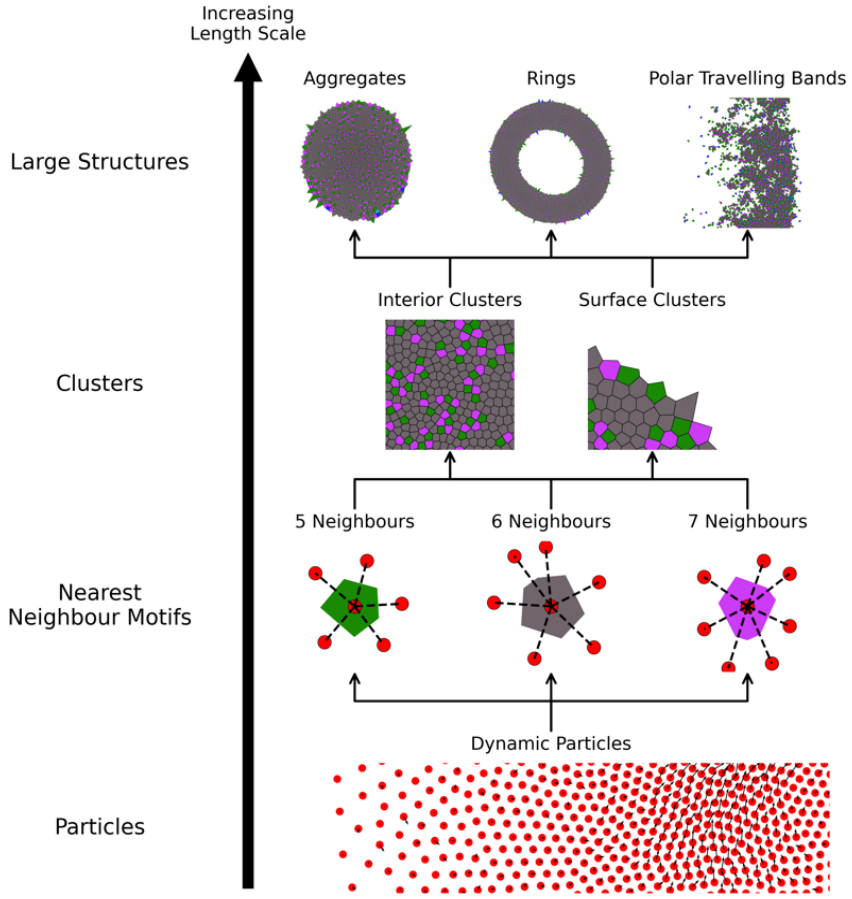



Figure 2: Learning the language of complex interactions between biological cells (image to be read from bottom to top). Cells with complex interaction rules, abstracted here as dynamically moving particles, often form dense nearest-neighbour (NN) motifs (5-NN, 6-NN, 7-NN, etc.). These NN motifs, in turn, occur in various motif clusters, whose behaviours determine whether they eventually grow into cellular aggregates, rings, or larger bands that spontaneously migrate in persistent directions. Figure by LIM Ying Chen.

us new insights into how to influence or control these processes, leading to better therapies and treatments.

## Interpretable machine learning fuels our collective curiosities

Humans have always been curious about the world, driven by a desire to explore, understand, and explain how things work. Scientific research, in broad strokes, professionalises this curiosity. This curiosity has led to powerful discoveries about regularities and irregularities in our world: from deep connections and abstract patterns in mathematics, to the laws of physics that describe the origin of forces and particles, to the grip of thermodynamics on our chemistry, to the inner workings of cells.

Science has always been driven by data from experimentation. ML is now used alongside traditional methods to help make sense of the vast amounts of data produced by modern instrumentation. ML allows computers to learn from data and find patterns that humans simply cannot fathom. This is especially useful in scientific research, where complex systems and interactions abound.

For discovery-AI to transform science, its complex ML models need to be interpretable, so that they can explain what they discover in a way that humans understand. This is not just about building smarter tools—it is about figuring out how AI can help us discover new scientific principles that are beyond our imaginations.

**For more details, please visit: https://www.dbs.nus.edu.sg/staffs/duane-loh/**

**Duane LOH is an Associate Professor in the Department of Physics and the Department of Biological Sciences at the National University of Singapore, as well as a Principal Investigator at the NUS Centre for Bio-imaging Sciences. Duane's research combines statistical physics, optics, and machine learning to create novel "computational lenses" that help in seeing and navigating the fleeting and chaotic nanometre-size world. His research also uses machine learning to help co-create a "language model of disorder" in complex materials, interacting cells, and the spread of vector-borne diseases in human populations.**

References

[1] Dan JD; Zhao XX; Ning SC; Lu J; Loh KP; He Q; Loh ND*; Pennycook SJ*, "Learning motifs and their hierarchies in atomic resolution microscopy" Science Advances Volume: 8 Issue: 15 DOI: 10.1126/sciadv.abk1005 Article Number: eabk1005 Published: 2022.

[2] Dan JD*; Waqar M; Erofeev I; Yao K; Wang J; Pennycook SJ; Loh ND*, "A multiscale generative model to understand disorder in domain boundaries" Science Advances Volume:9 Issue: 42 DOI: 10.1126/sciadv.adj0904 Article Number: eadj0904 Published: 2023.

# Statistician in geometry: A journey, lonely travelled, to a low-dimensional world

Unveiling the intricate dance between statistics and geometry to transform data analysis

## Introduction

In the immense landscape of data that surrounds us, understanding the role of dimensions—attributes that describe each data point—is essential. As technology evolves, we increasingly encounter datasets with hundreds or even tens of thousands of dimensions, presenting both opportunities for deeper insights and significant analytical challenges. This complexity necessitates dimensionality reduction, which simplifies data by transforming it from a high-dimensional space to a lower-dimensional one, preserving key properties. This process not only aids in visualising and interpreting data but also reduces storage and computational demands, and importantly, it helps eliminate noise and irrelevant features, clarifying the insights derived from the data.

For years, statistics has been deeply rooted in linearity, which often falls short in capturing the complex nature of real-world data. Mathematically, central limit theorems on non-linear spaces can exhibit unusual asymptotics depending on the underlying geometry, such as when mass near the cut locus of the population mean pulls on sample means, causing them to converge more slowly than expected. The interplay between the data distribution and its underlying geometry gives rise to central limit theorems that reveal two intriguing phenomena: “smeariness” (slow convergence) and “stickiness” (fast convergence). This behaviour reflects both aspects of the curvature and the measure, therefore thus defines a much higher ceiling for researchers to pursue, in a fundamental way. Our team, working with the world-renowned geometer, S.-T YAU and his institute, aims to push the geometry-aware statistics research boundary to a new level beyond linearity. We specialise in nonlinear dimensionality reduction, designed to address complex relationships between dimensions that linear methods like principal component analysis fail to resolve. These nonlinear techniques excel at deciphering complex patterns, allowing us to explore the inherent geometry and topology of data hidden within high-dimensional spaces and reveal insights that traditional approaches might miss.

As the field of Artificial Intelligence evolves, our ability to process vast amounts of data has improved, yet challenges from extreme dimensionality persist. Machine learning models, heavily reliant on extensive large models and substantial computational resources, are increasingly becoming resource constrained. Our approach, which emphasises nonlinear structures in high-dimensional data, presents a promising solution. By integrating advanced mathematical strategies, we aim to reduce dependency on extensive resources and enhance the efficiency of data processing, offering a more sustainable path forward in high dimensional data analysis.

## Learning the data manifold

From the perspective of the interplay between geometry and statistics, high-dimensional data often cluster around simpler, hidden structures termed “manifolds”, and understanding this interplay between geometry and statistics is crucial for deciphering complex data. Manifold learning, which can be categorised into manifold embedding, manifold denoising, and manifold fitting, serves this purpose. Manifold embedding techniques reduce high-dimensional datasets to lower-dimensional forms, preserving essential relationships like distances between data points. Manifold denoising cleans the data by removing outliers that deviate from the expected manifold configuration. However, these methods often fail to capture the full geometric details. Our focus, manifold fitting, reconstructs a smooth manifold that accurately reflects the underlying low-dimensional structure, capturing both the geometric and topological properties of the data for a more comprehensive understanding.

Yao’s original work starts from the introduction of the concept called “principal flow”, a non-linear generalisation of principal component analysis on manifolds. This innovation uses a one-dimensional smooth curve to trace the primary variations within the data, thereby increasing the accuracy and efficiency of nonlinear data analysis. He further invented new concepts such as “principal sub-manifold” and “fixed boundary flow” by combining differential geometry with modern statistics to deal with data that has higher latent dimension or that have given boundaries.

To address the ubiquitous noise in real data more effectively, the group has developed a series of state-of-the-art fitting methods (Yao et al. (2024, a,b)) to study the latent manifold. For each sample point, we first identify the main direction of the noise and then “push” the sample towards the latent manifold along this direction. This technique leverages the local geometric structure of each sample, which has been theoretically proven to effectively reduce noise and simplify data representation through nonlinear dimensionality reduction. Crucially, as all sample points remain in their original space, this method can be seamlessly integrated into complex workflows, such as neural networks and bioinformatics pipelines, enhancing their efficiency and accuracy.

## Explore the world of genes with manifold

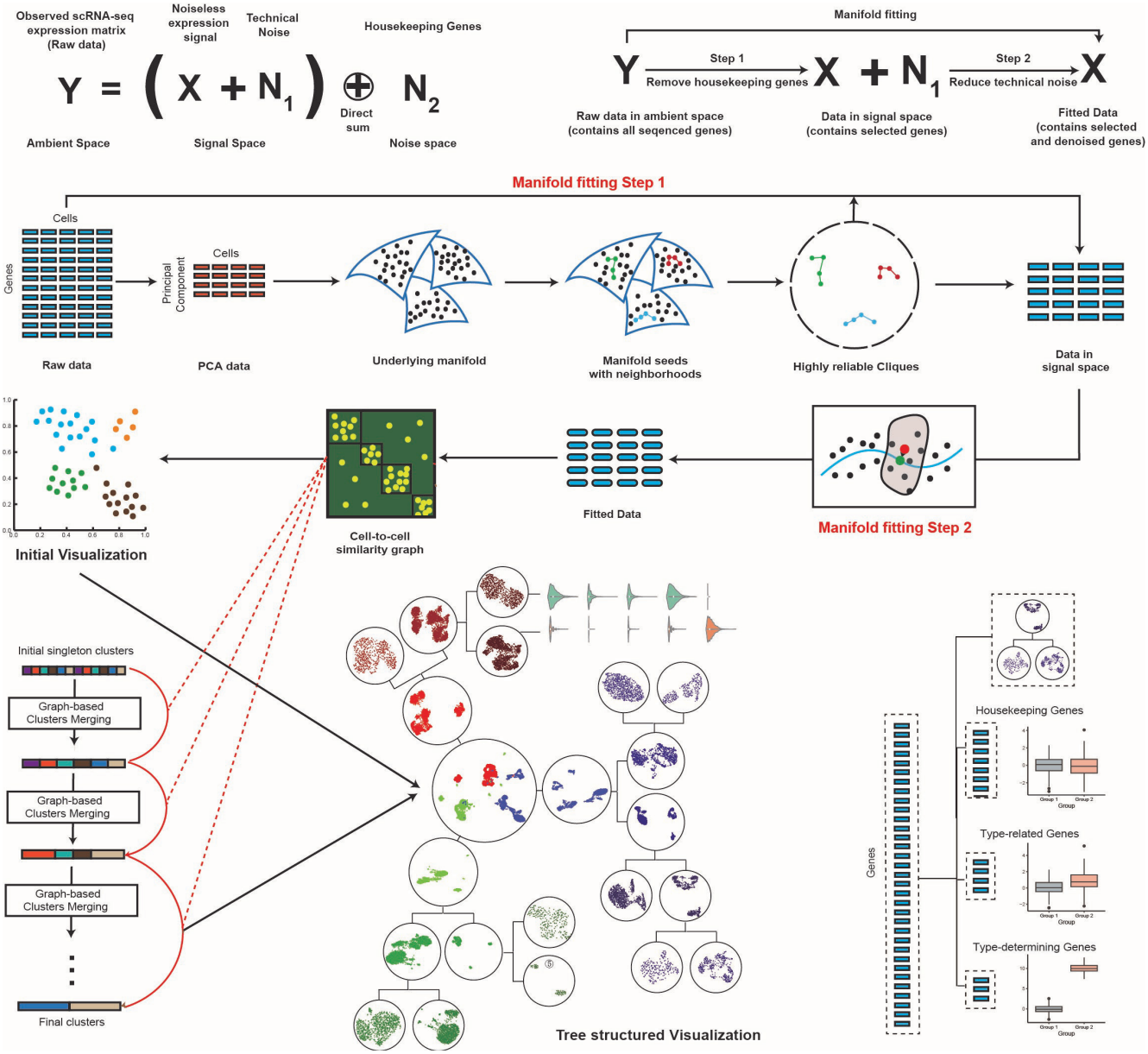



Figure 1: Building a Cell Atlas Using the Manifold Fitting Framework: Raw data incorporating manifold-based information undergoes clustering through a graph-based method. This process organizes cells into a tree-structured classification, facilitating the identification of novel cell types and key genes responsible for cell differentiation.

Single-cell RNA sequencing (scRNA-seq) has revolutionised genomic research by providing a detailed view of the genetic makeup of individual cells, enhancing our understanding of cellular interactions, diversity, and development. This technology traces variations in gene expression, shedding light on diseases such as diabetes, Alzheimer's disease, and cancer, and supports advances in multi-omics analysis and spatial transcriptomics. However, scRNA-seq faces significant challenges, including biological noise from natural cellular variability and measurement errors associated with sequencing techniques. These issues can hinder the accurate interpretation of biological data.

To tackle these challenges, we developed a framework that utilises a manifold fitting method to analyze scRNA-seq data. This process begins with a data transformation to enhance the signal-to-noise ratio, reducing gene expression variability and correcting for batch effects. An unsupervised approach is used to adaptively select the most suitable method for each dataset. A manifold fitting algorithm employing a shared nearest neighbour metric then efficiently defines data neighbourhoods, streamlining the process and expediting data handling. This reduces overlaps within groups and enhances separation between them, facilitating more effective clustering.

Our framework applies a variety of fast clustering algorithms tailored to the complex nature of scRNA-seq datasets, selecting the optimal clustering by comparing the similarity of cell types within and between clusters for precise cellular function and type analysis. Additionally, this method adeptly reveals the hidden low-dimensional structure of the data, effectively countering both technical variability

and biological noise. Numerous experiments have demonstrated that our approach not only more effectively recovers distorted RNA expression data but also enhances clustering capabilities beyond existing techniques, significantly improving the depth and accuracy of single-cell genomic analysis. Furthermore, our method can identify potential subclasses of cells, offering new insights for ongoing scientific exploration.

We are also developing another workflow based on manifold fitting that will revolutionise cell type identification and cell atlas construction. This new workflow promises to provide new insights into cellular function and organisation, potentially setting new standards in the field. I have been scheduled to deliver a 60-minute plenary lecture on this progress at the incoming International Congress of Chinese Mathematicians (ICCM) in Shanghai in 2025.

## The next steps

Having established a robust theoretical foundation, we have demonstrated the effectiveness of manifold fitting in harnessing the low-dimensional structures hidden within data to enhance its analytical capabilities. This technique has been successfully integrated with neural networks and scRNA-seq, showcasing its efficacy in complex genomic studies.

Looking forward, we are exploring the application of manifold fitting to human metabolomics data, aiming to advance precision medicine by deepening our understanding of metabolic processes and their individual variations. Additionally, since manifold fitting can be seamlessly integrated as a module into various existing workflows, we plan to extend its application across diverse fields that handle large, high-dimensional data sets. This expansion is expected to significantly boost the analytical capabilities of current models and reduce the computational resources required, thereby driving more efficient and sustainable progress across multiple scientific domains.

**For more details, please visit:**
**https://zhigang-yao.github.io/**

**Zhigang YAO is an Associate Professor in the Department of Statistics and Data Science at the National University of Singapore, where he also holds courtesy appointment with the Department of Mathematics and is a Faculty Affiliate of the Institute of Data Science. He proactively promotes the emerging research direction in the interface of statistics and geometry at the international level.**

References

[1] Yao ZG*; Su JJ*; Yau ST*, "Manifold fitting with CycleGAN" Proceedings of The National Academy of Sciences of The United States of America Volume: 121 Issue: 5 DOI: 10.1073/pnas.2311436121 Article number: e2311436121 Published: 2024.

[2] Yao ZG*; Li BJ*; Lu YK*; Yau ST*, "Single-cell analysis via manifold fitting: A framework for RNA clustering and beyond" Proceedings of The National Academy of Sciences of The United States of America Volume: 121 Issue: 37 DOI: 10.1073/pnas.2400002121 Article Number: e2400002121 Published: 2024.


NUS
National University
of Singapore
Faculty of
Science