

CONTACT
INFORMATION

King's College London
Department of Mathematics
Strand Building
Level 5 - S4.01
Tel: +44 7506052802
✉ E-mail: qinghua.zhou@kcl.ac.uk

Website: qinghua-zhou.github.io
Google Scholar: scholar.google.com/qz
Linkedin: linkedin.com/in/qinghua-zhou
Github: github.com/qinghua-zhou
Huggingface: [spaces/qinghua-zhou](https://spaces/huggingface/qinghua-zhou)

INTRO

About me: I am a researcher working in the fields of machine learning and artificial intelligence. My research interests lie in exploring pathways towards robust, stable and trustworthy AI systems. This includes the development of high-performance software and efficient scaling of large-scale simulations. My most current work is on the principled theoretical and computational analysis of modern computer vision and large language models, their [structures](#), [functions](#), [optimization](#), and methods to [edit or attack](#) them.

SELECT
PROJECTS

- **Stealth edits for large language models**

Oliver Sutton, Qinghua Zhou*, Wei Wang, Desmond Higham, Alexander Gorban, Alexander Bastounis, and Ivan Tyukin*

Advances in Neural Information Processing Systems (NeurIPS), 2024

[Paper](#) | [Code](#) | [Huggingface Demo](#) | [SIAM News Front Page Article](#)

We expose the susceptibility of modern AI models to a new class of malicious attacks and reveal a new theoretical understanding of the causes behind this. This work enables us to either introduce external model components with easily 10,000 edits/layer with almost no impact on the model or hide an attack that is virtually impossible to detect, and even if the attack is found, it is impossible to determine the triggering prompt.

- **How adversarial attacks can disrupt seemingly stable accurate classifiers**

Oliver Sutton, Qinghua Zhou, Ivan Tyukin, Alexandar Gorban, Alexander Bastounis, Desmond Higham
Neural Networks, 2024

We demonstrate a fundamental feature of classifiers working with high dimensional input data: simultaneous susceptibility of an 'accurate' model to adversarial attacks, and robustness to random perturbations of the input data. We provide theoretical framework and extensive empirical verification that using additive noise during training or testing is inefficient for eradicating or detecting adversarial examples, and insufficient for certification of robustness!

- **Multiple-instance ensemble for construction of deep heterogeneous committees for high-dimensional low-sample-size data**

Qinghua Zhou, Shuihua Wang, Hengde Zhu, Xin Zhang, Yu-Dong Zhang
Neural Networks, 2023

Learning in the high dimensional and low-sample size (HDLS) domain is recognised as one of the core challenges for modern AI. Here we introduce a novel stacking method that utilises attention pooling mechanisms for ensembles and cascades. We provide extensive empirical experiments over a range of HDLS datasets in the medical domain.

OTHER
WORKS

- **Neuromorphic tuning of feature spaces to overcome the challenge of low-sample high-dimensional data**

Qinghua Zhou, Oliver Sutton, Yu-Dong Zhang, Alexander Gorban, Valeri Makarov, Ivan Tyukin

Here we consider the scenario of mesoscopic sample-size where we cannot reliably build models with high degrees of expressivity. We present a neuromorphic algorithm capable of improving existing feature spaces via learning relevant associations in high dimensional data with high probability!

- **Quasi-orthogonality and intrinsic dimensions as principled measures of learning and generalisation**

Qinghua Zhou, Alexandar Gorban, Jonathan Bac, Andrei Zinovyev, Ivan Tyukin

We examine correlations between the accuracies of trained networks and measures of randomly initialised networks. While prior work focused on volume-based principled measures, we examine measures of intrinsic dimensionality and quasi-orthogonality. Our observations further validates prior theoretical findings, and we provides new perspectives for zero-shot NAS.

- Much more publications on [Google Scholar Page](#)

ACADEMICS	King's College London <i>Department of Mathematics, United Kingdom</i>	2022–2025
	• Postdoctoral Research Associate in Methods and Algorithms of Artificial Intelligence. • Topic: Robust, resilient, certifiable and trustworthy AI systems • Right to work in UK with Global Talent Visa • Supervisor: Prof. Ivan Tyukin	
	University of Leicester, <i>School of Computing and Mathematical Sciences, United Kingdom</i>	2019–2023
	• Ph.D. Computer Science Research. • Topic: Learning from high dimensional low-sample size data in medical applications • Graduate Teaching Assistant studentship • Supervisors: Prof. Yudong Zhang and Prof. Ivan Tyukin	
	University of Sydney, <i>School of Mathematics and Statistics, and School of Physics, Australia</i>	2016–2018
	• B.Sc., Science (Advanced Mathematics) • Majors in both Physics and Applied Mathematics • Introduction to research through asteroseismology with Prof. Tim Bedding	
OTHER EXPERIENCES	Industry Collaboration with TG-0 <i>Department of Mathematics, King's College London, and TG-0, London</i>	2024
	• Development of reservoir computing algorithms for specific hardware applications • Development of low-cost and efficient algorithmic alternatives for specialised hardware	
	Academic Teaching - Graduate Teaching Assistant <i>School of Computing and Mathematical Sciences, University of Leicester</i>	2019–2021
	• Assistance in teaching undergraduate to Masters modules. • Preparation of materials for teaching and assessments • Delivering tutorial sessions in lecture format.	
	Research Assistant Coordinator <i>Collaboration with Department of Cardiovascular Sciences, University of Leicester.</i>	2019–2020
	• Study and processing of gene expression data from the GEO database. • Development of machine learning approach for knowledge discovery. • Lead and coordinate a small group of research assistants in the same project.	

QR CODE



REFERENCES Available upon request.