Book of Abstracts

GAMM Workshop

Computational and Mathematical Methods in Data Science

Center for Data and Simulation Science

May 4–5, 2023
Center for Data and Simulation Science
University of Cologne
Cologne, Germany

University of Cologne

2023

Thursday – May 4		Chair
12:00 – 12:20	**Registration & Opening**	
12:20 – 13:15	**Alexander Heinlein:** Neural networks with physical constraints — Domain decomposition-based network architectures, and model order reduction	A. Klawonn
13:15 – 13:35	*20 Minute Break*	
13:35 – 14:00	**Claudia Drygala:** Learning from Chaos	A. Heinlein
14:00 – 14:25	**Renzhi Tian:** Data-driven turbulence modelling using Gene Expression Programming	A. Heinlein
14:25 – 14:55	*30 Minute Break*	
14:55 – 15:50	**Karen Veroy-Grepl:** Model Order Reduction in the Multi-Scale Materials Setting	S. Peitz
15:50 – 15:55	*5 Minute Break*	
15:55 – 16:20	**Janine Weber:** A Domain Decomposition-Based CNN-DNN Architecture for Model Parallel Training Applied to Image Recognition Problems	M. Stoll
16:20 – 16:45	**Kira Maag:** Out-of-Distribution Segmentation via Pixel-wise Gradient Uncertainty	M. Stoll
16:45 – 17:00	*15 Minute Break*	
17:00 – 17:25	**Reyhaneh Abbasi:** An improved detection and classification method for mouse ultrasonic vocalizations	A. Klawonn
17:25 – 17:50	**Martin Stoll:** Efficient linear algebra for training Gaussian processes	A. Klawonn
17:50 – 18:15	~~**Darlington S. David:** Breast Cancer Prediction using Machine Learning Algorithms – A Deep Learning Approach~~ *(cancelled)*	A. Klawonn
18:15 – 18:30	**Meeting of the GAMM Activity Group "Computational and Mathematical Methods in Data Science"**	
19:30	*Dinner (Brauerei Päffgen)*	

Friday – May 5		Chair
09:00 – 09:55	**Aleksandar Bojchevski:** Machine Learning with Guarantees	H. Gottschalk
09:55 – 10:00	*5 Minute Break*	
10:00 – 10:25	**Pierre-François Massiani:** Safe Value Functions	A. Bojchevski
10:25 – 10:50	**Hanno Gottschalk:** LU-Net: Invertible Neural Networks Based on Matrix Factorization	A. Bojchevski
10:50 – 11:10	*20 Minute Break*	
11:10 – 11:35	**Christian Staerk:** Adaptive sampling and variable selection strategies for high-dimensional genetic data	M. Lanser
11:35 – 12:00	**Paolo Climaco:** Investigating the effects of minimising the training set fill distance in machine learning regression	M. Lanser
12:00 – 12:05	*5 Minute Break*	
12:05 – 13:00	**Sebastian Peitz:** Sample efficiency in data-driven Model Predictive Control and Reinforcement Learning	M. Stoll
13:00	*Farewell & Light Lunch*	

An improved detection and classification method for mouse ultrasonic vocalizations

Reyhaneh Abbasi[a,b,c], Peter Balazs[a], Maria Adelaide Marconi[b], Doris Nicolakis[b], Sarah M. Zala[b], and Dustin J. Penn[b]

[a] *Acoustic Research Institute, Austrian Academy of Sciences, Vienna, Austria*
[b] *Konrad Lorenz Institute of Ethology, Department of Interdisciplinary Life Sciences, University of Veterinary Medicine, Vienna, Austria*
[c] *Vienna Doctoral School of Cognition, Behaviour and Neuroscience, University of Vienna, Vienna, Austria*

House mice and other rodents emit complex ultrasonic vocalizations (USVs) to communicate in various contexts including social and sexual interactions. These vocalizations are increasingly investigated in research on animal communication and as a phenotype for studying the genetic basis of autism and speech disorders. Rodents emit USVs in discrete units called syllables or calls. USV syllables are separated by gaps of silence and they have been classified into several different categories by researchers visually inspecting spectrograms. Because manual methods for analyzing USVs are extremely time-consuming, several methods have been recently developed for automatically detecting and classifying USVs. Here we evaluate their advantages and disadvantages in a systematic comparison, while also presenting a new approach. This study aims to 1) determine the most efficient USV detection tool among the existing methods, and 2) develop a classification model that is more generalizable than existing methods. We compared the performance of four detection methods in an out-of-the-box approach, pretrained DeepSqueak detector, MUPET, USVSEG, and the Automatic Mouse Ultrasound Detector (A-MUD). A-MUD outperformed the other methods in terms of true positive rates and false detection rates. For automating the classification of USVs, we developed BootSnap for supervised classification, which combines bootstrapping on Gammatone Spectrograms and Convolutional Neural Networks algorithms with Snapshot ensemble learning. It successfully classified calls into 12 types, including a new class of false positives that is useful for detection refinement. BootSnap outperformed the pretrained and retrained state-of-the-art tool, and thus it is more generalizable [1].

References

[1] Abbasi R, Balazs P, Marconi MA, Nicolakis D, Zala SM, Penn DJ. Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap). PLoS Computational Biology. 2022 May 12;18(5):e1010049.

Machine Learning with Guarantees

Aleksandar Bojchevski[a]

[a] *University of Cologne*

From healthcare to natural disaster prediction, high-stakes applications increasingly rely on machine learning models. Yet, most models are unreliable. They can be vulnerable to manipulation and unpredictable on inputs that slightly deviate from their training data. To make them trustworthy, we need provable guarantees. In this talk, we will explore two kinds of guarantees: robustness certificates and conformal prediction. First, we will derive certificates that guarantee stability under worst-case adversarial perturbations, focusing on the model-agnostic randomized smoothing technique. Next, we will discuss conformal prediction to equip models with prediction sets that cover the true label with high probability. The prediction set size reflects the model's uncertainty. To conclude, we will provide an overview of guarantees for other trustworthiness aspects such as privacy and fairness.

Investigating the effects of minimising the training set fill distance in machine learning regression

Paolo Climaco[a], Jochen Garcke[a,b]

[a] *Institut für Numerische Simulation, Universität Bonn, Bonn, Germany*
[b] *Fraunhofer SCAI, Sankt Augustin, Germany*

Machine learning (ML) regression methods are powerful tools leveraging large datasets for nonlinear function approximation in high-dimensional domains. Unfortunately, learning from large datasets may be unfeasible due to computational limitations and the cost of producing labels for the data points, as in the case of quantum-chemistry applications. Therefore, an important task in scientific ML is sampling small training sets from large pools of unlabelled data points to maximise models' performance while keeping the computational effort of the training and labelling processes low. In this talk, we analyse the advantages of a common approach for training set sampling aiming to minimise the fill distance of the selected set, which can be considered a measure of the data distribution. We provide a bound for the maximum expected error of the loss function, conditional to the knowledge of the data features, depending linearly on the fill distance. Next, we show that minimising such a bound by minimising the training set fill distance reduces the worst-case approximation error of several ML models, which can be interpreted as an indicator of the robustness of a model's approximation. We perform experiments considering various ML techniques, such as Kernel methods and Neural Networks. The application context is quantum-chemistry, where different datasets have been generated to develop effective ML techniques and approximate the functions mapping high-dimensional representations of molecules to their physical and chemical properties.

Breast Cancer Prediction using Machine Learning Algorithms – A Deep Learning Approach

Darlington S. David[a]

[a] *School of Mathematics, Georgia Institute of Technology, Atlanta, USA*

Breast Cancer is the deadliest and commonly diagnosed cancer in women globally. Early diagnosis and treatment of breast cancer increases the chance of a five-year survival rate by 99 %. Recent technological and computational advancement have led to the discovery of machine learning algorithms for the analysis of complex data. Machine learning algorithms have been widely applied for the analysis of breast cancer data. In this paper, we propose to implement machine learning algorithms, using a deep learning approach, for automatic detection and prediction of breast cancer using mammogram images. To achieve this, we implement transfer learning on a deep learning algorithm called Convolutional Neural Network (CNN). Two datasets of breast cancer images are analyzed using three CNN models in existing deep learning frameworks. The models perform a binary and multiclass classification task on the images. Experimental results showed that CNN models can accurately identify and predict breast cancer when provided with a large and balanced dataset.

Learning from Chaos

Claudia Drygala[a], Francesca di Mare[b], and Hanno Gottschalk[c]

[a] *University of Wuppertal, School of Mathematics and Natural Sciences, IMACM & IZMD, e-mail: drygala@uni-wuppertal.de*

[b] *Ruhr University Bochum, Department of Mechanical Engineering, Chair of Thermal Turbomachines and Aero Engines, e-mail: francesca.dimare@ruhr-uni-bochum.de*

[c] *Technical University Berlin, Institute of Mathematics, e-mail: gottschalk@math.tu-berlin.de*

Chaotic deterministic systems are characterized by a high degree of nonlinearity and unsteady, aperiodic behavior that is sensitive to initial conditions [1]. Numerical simulations of these systems often require the solution of partial differential equations on very fine computational meshes with small time steps. For engineering problems (e.g., high Reynolds numbers in turbulent flow simulations), the computational effort is unfeasible in a reasonable time despite rapidly increasing computer power [2]. To overcome this problem, we apply generative adversarial networks (GAN) as a mathematically well-founded approach [3] for the synthetic modeling of sample state snapshots from the invariant measure of chaotic deterministic systems at different hierarchy levels. With the increase of the complexity of the chaotic systems, the requirements for the GAN architecture also increase. Starting with a Vanilla GAN, we synthesize data points of the three-dimensional trajectory of the Lorenz attractor and proceed to the four-dimensional problem of the double pendulum. This is followed by modeling the turbulent flow around a cylinder using a deep convolutional GAN (DCGAN). We end with the synthesis of the flow around a low-pressure turbine stator using the conditional DCGAN pix2pixHD conditioned on the position of a rotating wake in front of the stator as the most complex example of a chaotic deterministic system. Furthermore, the ability of the conditional GAN to generalize over changes in geometry is demonstrated by generating turbulent flow fields for wake positions not included in the training data [4]. We compare the statistical properties of the synthesized state snapshots with those obtained by classical numerical methods. For the generative modeling of turbulence, we use fields of velocity fluctuations obtained from large-eddy simulations (LES) as training data. We show that GAN are efficient for simulating turbulence with a moderate amount of training data. The GAN training and inference times are significantly reduced compared to LES, while still providing turbulent flows with high resolution.

References

[1] S.H. Strogatz (2018). "Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering". CRC press.

[2] B.Winhart, M. Sinkwitz, A. Schramm, P. Post and F. di Mare, "Large Eddy Simulation of Periodic Wake Impact on Boundary Layer Transition Mechanisms on a Highly Loaded Low-Pressure Turbine Blade." Proceedings of the ASME Turbo Expo 2020: Turbomachinery Technical Conference and Exposition. Volume 2E: Turbomachinery. Virtual, Online. September 21–25, 2020. V02ET41A013. ASME. https://doi.org/10.1115/GT2020-14555.

[3] C. Drygala, B. Winhart, F. di Mare, and H. Gottschalk, "Generative modeling of turbulence", Physics of Fluids 34, 035114 (2022) https://doi.org/10.1063/5.0082562.

[4] C. Drygala, F. di Mare, and H. Gottschalk (2023). "Generalization capabilities of conditional GAN for turbulent flow under changes of geometry". arXiv preprint arXiv:2302.09945 - accepted at EUROGEN (International Conference on Evolutionary and Deterministic Methods for Design, Optimization and Control) 2023.

LU-Net: Invertible Neural Networks Based on Matrix Factorization

Robin Chan[a,b], **Hanno Gottschalk**[a], and Sarina Penquitt[c]

[a] *Institute of Mathematics, TU Berlin, Germany*
[b] *Faculty of Technology, University of Bielefeld, Germany*
[c] *School of Mathematics and Science, University of Wuppertal, Germany*

LU-Net is a simple and fast architecture for invertible neural networks (INN) that is based on the factorization of quadratic weight matrices $\mathsf{A} = \mathsf{LU}$, where L is a lower triangular matrix with ones on the diagonal and U an upper triangular matrix. Instead of learning a fully occupied matrix A, we learn L and U separately. If combined with an invertible activation function, such layers can easily be inverted whenever the diagonal entries of U are different from zero. Also, the computation of the determinant of the Jacobian matrix of such layers is cheap. Consequently, the LU architecture allows for cheap computation of the likelihood via the change of variables formula and can be trained according to the maximum likelihood principle. In our numerical experiments, we test the LU-net architecture as generative model on several academic datasets. We also provide a detailed comparison with conventional invertible neural networks in terms of performance, training as well as run time.

References

[1] R. Chan, S. Penquitt and H. Gottschalk. LU-Net: Invertible Neural Networks Based on Matrix Factorization, arXiv preprint 2023, arXiv:2302.10524

Neural networks with physical constraints – Domain decomposition-based network architectures, and model order reduction

Alexander Heinlein[a]

[a] *Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands.*
a.heinlein@tudelft.nl

Scientific machine learning (SciML) is a rapidly evolving field of research that combines techniques from scientific computing and machine learning. A major branch of SciML is the approximation of the solutions of partial differential equations (PDEs) using neural networks. The network models be can trained in a data-driven and/or physics-informed way, that is, using reference data (from simulations or measurements) or a loss function based on the PDE, respectively.

In physics-informed neural networks (PINNs) [4], simple feedforward neural networks are employed to discretize the PDEs, and a single network is trained to approximate the solution of one specific boundary value problem. The loss function may include a combination of data and the residual of the PDE. Challenging applications, such as multiscale problems, require neural networks with high capacity, and the training is often not robust and may take large iteration counts. Therefore, in the first part of the talk, domain decomposition-based network architectures improving the training performance using the finite basis physics-informed neural network (FBPINN) approach [3, 1] will be discussed. It is based on joint work with Victorita Dolean (University of Strathclyde, Côte d'Azur University), Siddhartha Mishra, and Ben Moseley (ETH Zürich).

In the second part of the talk, surrogate models for computational fluid dynamics (CFD) simulations based on convolutional neural networks (CNNs) [2] will be discussed. In particular, the network is trained to approximate a solution operator, taking a representation of the geometry as input and the solution field(s) as output. In contrast to the classical PINN approach, a single network is trained to approximate a variety of boundary value problems. This makes the approach potentially very efficient. As in the PINN approach, data as well as the residual of the PDE may be used in the loss function for training the network. The second part of the talk is based on joint work with Matthias Eichinger, Viktor Grimm, and Axel Klawonn (University of Cologne).

References

[1] V. Dolean, A. Heinlein, S. Mishra, and B. Moseley. Finite basis physics-informed neural networks as a Schwarz domain decomposition method, November 2022. arXiv:2211.05560.

[2] M. Eichinger, A. Heinlein, and A. Klawonn. Surrogate convolutional neural network models for steady computational fluid dynamics simulations. *Electronic Transactions on Numerical Analysis*, 56:235–255, 2022.

[3] B. Moseley, A. Markham, and T. Nissen-Meyer. Finite Basis Physics-Informed Neural Networks (FBPINNs): a scalable domain decomposition approach for solving differential equations, July 2021. arXiv:2107.07871.

[4] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Out-of-Distribution Segmentation via Pixel-wise Gradient Uncertainty

Kira Maag*[a], and Tobias Riedlinger*[b]

[a] *Ruhr University Bochum, Germany*
[b] *University of Wuppertal, Germany*

In recent years, deep neural networks have defined the state-of-the-art in semantic segmentation where their predictions are constrained to a predefined set of semantic classes. They are to be deployed in applications such as automated driving, although their categorically confined expressive power runs contrary to such open world scenarios. Thus, the detection and segmentation of objects from outside their predefined semantic space, i.e., out-of-distribution (OoD) objects, is of highest interest. Since uncertainty estimation methods like softmax entropy or Bayesian models are sensitive to erroneous predictions, these methods are a natural baseline for OoD detection. Here, we present a method for obtaining uncertainty scores from pixel-wise loss gradients which can be computed efficiently during inference [1]. Our approach is simple to implement for a large class of models, does not require any additional training or auxiliary data and can be readily used on pre-trained segmentation models. Our experiments show the ability of our method to identify wrong pixel classifications and to estimate prediction quality. In particular, we observe superior performance in terms of OoD segmentation to comparable baselines on the SegmentMeIfYouCan benchmark [2], clearly outperforming methods which are similarly flexible to implement.

References

[1] K. Maag and T. Riedlinger. Pixel-wise Gradient Uncertainty for Convolutional Neural Networks applied to Out-of-Distribution Segmentation. 2023.

[2] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann and M. Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.

* equal contribution

Safe Value Functions

Pierre-François Massiani[a], Steve Heim[b], Friedrich Solowjow[a], Sebastian Trimpe[a]

[a] *Institute for Data Science in Mechanical Engineering, RWTH Aachen University, 52068 Aachen, Germany*

[b] *Biomimetic Robotics Lab, Department of Mechanical Engineering, Massachussets Institute of Technology, 02139 Cambridge, USA*

Safety constraints and optimality are important but sometimes conflicting criteria for controllers. Although these criteria are often solved separately with different tools to maintain formal guarantees, it is also common practice in reinforcement learning to simply modify reward functions by penalizing failures, with the penalty treated as a mere heuristic.

In this talk, I will introduce *Safe Value Functions* (SVFs): value functions that are both optimal for a given task, and enforce safety constraints. I will present their relationship with penalties and show that failure penalization naturally gives rise to SVFs. There is an upper-unbounded interval of penalties that achieve an SVF; high penalties do not harm optimality. The analysis relies on understanding when optimal control stabilizes the *viability kernel*, i.e., the largest safe set.

Although it is often intractable to compute the minimum required penalty, SVFs reveal clear structure of how the penalty, rewards, discount factor, and dynamics interact. This insight suggests practical, theory-guided heuristics to design reward functions for optimal control problems where safety is important.

References

[1] P.-F. Massiani, S. Heim. F. Solowjow, S. Trimpe, *Safe Value Functions*, IEEE Transactions on Automatic Control, 2023.

Sample efficiency in data-driven Model Predictive Control and Reinforcement Learning

Sebastian Peitz[a], Katharina Bieker[b], Jan Stenner[a], Vikas Chidananda[a], Oliver Wallscheid[c], Steven L. Brunton[d], Kunihiko Taira[e]

[a] *Department of Computer Science, Paderborn University, Paderborn, Germany*
[b] *Department of Computer Science, LMU Munich, Munich, Germany*
[c] *Department of Electrical Engineering, Paderborn University, Paderborn, Germany*
[d] *University of Washington, Seattle, WA, USA*
[e] *UCLA, Los Angeles, CA, USA*

As in almost every other branch of science, the advances in data science and machine learning have also resulted in improved modeling, simulation and control of nonlinear dynamical systems, prominent examples being autonomous driving or the control of complex chemical processes. However, many of these approaches face the issues that they (1) do not have strong performance guarantees and (2) often tend to be very data hungry. In this presentation, we discuss different approaches to improve the sample efficiency in data-driven feedback control. We address both model predictive control (MPC) as well as reinforcement learning (RL). In MPC, learning an accurate surrogate model is paramount for the performance. Exploiting techniques from mixed-integer control, we show that one can leverage performance guarantees of autonomous systems - which have been studied much more extensively - to obtain related error bounds for control problems. In RL, we address both the usage of surrogate models as well as the exploitation of system symmetries to improve sample efficiency. We demonstrate our findings using several example systems governed by partial differential equations.

Adaptive sampling and variable selection strategies for high-dimensional genetic data

Christian Staerk[a]

[a] *Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn*

With the growing availability of high-dimensional genetic data, data-driven variable selection methods play an increasingly important role in genetic epidemiology. Scalable statistical learning methods are needed to effectively explore the high-dimensional space of possible models with thousands of potential genetic variables.

I will present an overview of recent adaptive sampling strategies which can be incorporated in various variable selection approaches, such as ℓ_0-type regularization, statistical boosting and Bayesian variable selection. The main underlying idea is to exploit the sparsity via adaptive stochastic searches, which focus on those variables that have proven to be "important" in previous iterations of the algorithms. In a first approach, the Adaptive Subspace (AdaSub) algorithm [1] tackles the high-dimensional discrete optimization problem induced by ℓ_0-type selection criteria by solving several low-dimensional sub-problems in an adaptive way, where the probability of each variable to be included in a new sub-problem is sequentially adjusted based on its selection frequency in previous sub-problems. In a statistical boosting approach, the AdaSubBoost algorithm [2] incorporates an adaptive random preselection of multivariable base-learners in each iteration, focusing on base-learners which were also predictive in previous iterations. Finally, the Metropolized AdaSub (MAdaSub) algorithm [3] is an adaptive Markov Chain Monte Carlo (MCMC) approach for Bayesian variable selection, where the individual proposal probabilities of the covariates are sequentially updated so that they converge against the posterior inclusion probabilities. Despite the continuing adaptation of the proposal probabilities, MAdaSub is ergodic, i.e. in the limit it samples from the full posterior model distribution. In gene expression data applications with more than 20,000 genes, MAdaSub can effectively sample from high-dimensional and multimodal posterior distributions.

While the adaptive variable selection methods described above, based on ℓ_0-type regularization, boosting and Bayesian variable selection, are primarily designed for sparse high-dimensional settings, I will also highlight recent approaches [4, 5] and open challenges for modelling polygenic traits based on large-scale genotype data with many influential genetic variants and large sample sizes.

References

[1] Staerk, C., Kateri, M., & Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. Electronic Journal of Statistics, 15(1), 830–879.

[2] Staerk, C., & Mayr, A. (2021). Randomized boosting with multivariable base-learners for high-dimensional variable selection and prediction. BMC Bioinformatics, 22(441).

[3] Staerk, C., Kateri, M., & Ntzoufras, I. (2022). A Metropolized adaptive subspace algorithm for high-dimensional Bayesian variable selection. Bayesian Analysis, Advance Publication, 1-31. `https://doi.org/10.1214/22-BA1351`

[4] Maj, C., Staerk, C., Borisov, O., Klinkhammer, H., Yeung, M. W., Krawitz, P., & Mayr, A. (2022). Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. Genetic Epidemiology, 46, 589–603.

[5] Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P. M., & Mayr, A. (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. Frontiers in Genetics, 13(1076440).

Efficient linear algebra for training Gaussian processes

Martin Stoll[a]

[a] *Department of Mathematics, TU Chemnitz*

In this talk I will review some of the numerical challenges that arise when training Gaussian processes. We will start from some basic definitions of these popular models within machine learning and then arise at two problems often encountered in scientific computing. Namely, the solution of a linear system and the evaluation of a matrix function. We focus on the particular case that the kernel matrix can be interpreted as a tensor for non-trivial rank and as such require numerical methods tailored for tensor operations.

Data-driven turbulence modelling using Gene Expression Programming

Renzhi Tian[a], Richard Dwight[a], and Stefan Hickel[a]

[a] *Aerodynamics Group, Faculty of Aerospace, Delft University of Technology*

Gene Expression Programming (GEP) is an established Evolutionary Algorithm (EA) used to search for expressions regressing a data-set [1]. The last decades have seen rapid progress and development in GEP. It has been successfully applied to classification-, regression-, and time series prediction problems. It has the benefit of producing explicit algebraic expressions which are concise and interpretable, in contrast to neural-networks and other highly parameterized models.

Turbulent flows are ubiquitous in aerospace engineering applications, wind energy, and many other fields. The numerical prediction of turbulent motion is highly challenging because of the wide range of spatial and temporal scales involved. Efficiently predicting turbulence with Computational Fluid Dynamics (CFD) therefore requires models for the small scales. Traditional turbulence modelling consists of physically motivated models, with a few parameters, calibrated on a small number of flow cases. Increasing Machine Learning (ML) methods are used to increase the model parameterization and fit large amounts of data. Since these models are running inside a CFD code, with physical models for mass, momentum and energy transport, they are required to be efficient in evaluation, well-behaved in extrapolation, and interpretable, and they must not destabilize the simulation.

GEP's promising potential for turbulence modelling has been put in the spotlight by Weatheritt et al. [2]. In the present study, we develop an efficient GEP platform for turbulence modelling, attempting to reproduce the capability of Weatheritt et al., before attempting to improve the algorithms, especially with respect to reducing the number of code evaluations required to fit a model.

Currently, our algorithm is dimension-aware and capable of optimizing model constants with a gradient-based method running inside GEP. We demonstrate this algorithm for a simple flow consisting of periodic hills, in two settings: (i) an *a priori* setting, where the required model-corrective fields are known in advance [3], and must only be regressed in terms of the flow quantities; and (ii) an *a posteriori* setting, where the best model running *in situ* in the CFD code is found. We later aim to use a multi-fidelity training approach, using known corrective fields as a proxy for *in situ* model fitness, without running the full solver at every step.

References

[1] C. Ferreira. Gene Expression Programming: Mathematical modelling by an artificial intelligence. Springer, 2006.

[2] J. Weatheritt and R. Sandberg. A novel evolutionary algorithm applied to algebraic modifications of the RANS stress–strain relationship. Journal of Computational Physics, 325 (2016): pp.22-37.

[3] M. Schmelzer, R. P. Dwight, and P. Cinnella. Discovery of algebraic Reynolds-stress models using sparse symbolic regression. Flow, Turbulence and Combustion, 104 (2022): pp.579-603.

Model Order Reduction in the Multi-Scale Materials Setting

(alphabetically) H. Bansal, T. Guo, Y. Hong, and **K. Veroy**[a]

[a] *Centre for Analysis, Scientific Computing and Applications, Eindhoven University of Technology (TU/e)*

Two-scale simulations are often employed to analyze the effect of the microstructure on a component's macroscopic properties. Understanding these structure–property relations is essential in the optimal design of materials, or to enable (for example) estimation of microstructure parameters through macroscale measurements. However, these two-scale simulations are typically computationally expensive and infeasible in multi-query contexts such as optimization and inverse problems. To make such analyses amenable, the microscopic simulations can be replaced by inexpensive, parametric surrogate models. In this talk, we (1) present some recent work on a non-intrusive reduced basis method to construct inexpensive surrogates for parametrized microscale problems, and (2) highlight difficulties for model order reduction presented by highly nonlinear constitutive relations in multi-scale problems in mechanics.

Acknowledgments: *This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 818473).*

A Domain Decomposition-Based CNN-DNN Architecture for Model Parallel Training Applied to Image Recognition Problems

Axel Klawonn[a], Martin Lanser[a], and **Janine Weber**[a]

[a] *Department of Mathematics and Computer Science, University of Cologne*

Deep neural networks (DNNs) and, in particular, convolutional neural networks (CNNs) have brought significant advances in a wide range of modern computer application problems. However, the increasing availability of large amounts of datasets as well as the increasing available computational power of modern computers lead to a steady growth in the complexity and size of DNN and CNN models, respectively, and thus, to longer training times. Hence, various methods and attempts have been developed to accelerate and parallelize the training of complex network architectures. In this talk, a novel domain decomposition-based CNN-DNN architecture is presented which naturally supports a model parallel training strategy. Experimental results for different 2D image classification problems are shown as well as for a face recognition problem, and for a classification problem for 3D computer tomography (CT) scans. The results show that the proposed approach can significantly accelerate the required training time compared to the global model and, additionally, can also help to improve the accuracy of the underlying classification problem.