COSEAL 2026

Workshop Schedule Overview: May 18-20, 2026

Monday, May 18

TIME	SESSION
08:00 - 09:00	Arrival & Coffee
09:00 - 09:30	Opening Session
09:30 - 10:30	Oral presentations 1
10:30 - 11:00	Coffee break
11:00 - 12:30	Poster session 1
12:30 - 13:30	Lunch break
13:30 - 15:00	Oral presentations 2
15:00 - 15:30	Coffee break
15:30 - 17:00	Poster session 2
17:30 - 20:30	Welcome reception @ RWTH SkyLounge

Tuesday, May 19

TIME	SESSION
08:00 - 09:00	Arrival & Coffee
09:00 - 10:30	Oral presentations 3
10:30 - 11:00	Coffee break
11:00 - 12:30	Poster session 3

12:30 - 13:30	Lunch break
13:30 - 15:00	Breakout session
15:00 - 15:30	Coffee break
15:30 - 17:00	Poster session 4
17:30 - 19:00	Dom tour
19:00 - 22:00	Workshop dinner @ Restaurant Elisenbrunnen

Wednesday, May 20

TIME	SESSION
08:00 - 09:00	Arrival & Coffee
09:00 - 11:00	Special Session: 50 Years of Algorithm Selection
11:00 - 11:30	Coffee break
11:30 - 13:00	Poster session 5
13:00 - 13:30	Closing Session

📋 Detailed Program

Full schedule with all presentations and timings

Monday, May 18

Arrival & Coffee

🕒 08:00 - 09:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Opening Session

🕒 09:00 - 09:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Oral presentations 1

🕒 09:30 - 10:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

09:30 - 09:50 (15+5 min)

Towards Robust AutoML

Jan van Rijn

09:50 - 10:10 (15+5 min)

Interactive Automated Machine Learning

Lukas Fehring, many others

From Bayesian Optimization over ensembled models to recent Agentic frameworks, AutoML systems are effective and can potentially outperform expert-configured solutions. At the same time, industry practitioners remain hesitant to adopt AutoML solutions, reporting a lack of agency and oversight. As a result, practitioners prefer manual model selection and hyperparameter optimization. To synergize AutoML with human expertise and tackle industry practitioners' concerns, recent research has focused on formalizing frameworks to integrate external input into the optimization loop. Examples of expert input include steering through prior information and allowing experts to reject suggested configurations. In this talk, we will discuss the current state of the field and opportunities to include agentic systems not only as independent AutoML systems but also as informative prior sources.

10:10 - 10:15 (3+2 min)

Reducing overtuning in hyperparameter optimization: An empirical evaluation of mitigation strategies

Sietse Schröder, Mitra Baratchi, Bernd Bischl, Matthias Feurer, Jan N. van Rijn

Hyperparameter optimization (HPO) aims to identify configurations that generalize well to unseen data, typically by minimizing validation error estimates obtained via resampling, such as holdout or cross-validation. Recent work has shown that excessive optimization of these stochastic validation estimates can lead to overtuning (Schneider et al., 2025), a degradation in true generalization performance despite continued improvements in validation performance. In this work, overtuning was observed across a wide range of HPO benchmark studies. In this ongoing work, we conduct a large-scale empirical study of how overtuning can be mitigated in practice. We systematically investigate a range of mitigation strategies, spanning resampling strategies, incumbent selection rules, and modifications to the optimizer, while revisiting the severity and determinants of overtuning on the TabArena benchmark (Erickson et al., 2025). Preliminary results suggest that certain mitigation strategies are effective at reducing overtuning in settings where overtuning is most pronounced, such as small-data regimes or when a limited resampling budget (e.g., holdout) is used. However, the same strategies may negatively impact generalization performance when overtuning is weak or absent, highlighting the importance of deeper insight into when and how overtuning mitigation should be applied. References Schneider, L., Bischl, B., & Feurer, M. (2025). Overtuning in Hyperparameter Optimization. In International Conference on Automated Machine Learning (pp. 17-1). PMLR. Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., & Hutter, F. (2025). Tabarena: A living benchmark for machine learning on tabular data. arXiv preprint arXiv:2506.16791.

10:15 - 10:20 (3+2 min)

SORE : Self-Optimizing Regularized Ensemble for Evolving Data Streams

Daniel Nowak Assis, Carola Doerr, Maroua Bahri

10:20 - 10:25 (3+2 min)

Rethinking Multilingual Embedding Model Ranking Across Learning Tasks and Languages

Ana Gjorgjevikj, Barbara Koroušić Seljak, Tome Eftimov

Text embeddings are a core component of modern natural language processing systems, enabling learning tasks such as classification, clustering, retrieval, semantic textual similarity. Many industry applications, including semantic search over text, recommender systems, and retrieval-augmented generation (RAG) pipelines, depend on vector databases built from these embeddings. However, embedding models still struggle to generalize reliably across tasks and languages, making model selection inherently task/language-dependent. Reliable insights into model generalization abilities are only possible through multilingual and multi-task platforms that incorporate variety of datasets and apply robust aggregation of model performance scores across those datasets. When datasets are highly correlated or evaluated using incomparable performance metrics, the commonly used average-based performance aggregation (employed in popular platforms such as GLUE, SuperGLUE, Big-Bench, Hugging Face's Open LLM Leaderboard, and MTEB version 1) can result in biased model rankings. Using a single score aggregation strategy is also insufficient for robust model selection. In this talk, we will present results from a comprehensive, language-aware study of the MTEB Multilingual v2 platform, offering practical guidance for language-specific model selection across approximately 230 languages. We perform both (i) language-specific, task-specific analysis and (ii) language-specific, task-agnostic analysis, ensuring robustness to both dataset composition and performance aggregation strategy selection. Our detailed analysis of five of the most widely used languages worldwide shows that large-scale LLM-based embedding models emerge as robust top performers across most, but not all tasks (e.g., retrieval), and that only a small subset of these models consistently generalizes well across different tasks, aggregation strategies, and dataset compositions.

10:25 - 10:30 (3+2 min)

Automatic Configuration of 3D Kronecker Sequences for Low Star Discrepancy

Imène Ait Abderrahim, Carola Doerr, Martin Durand

The L∞ star discrepancy of a point set P measures how well spread the points in P are. It is defined as the maximum absolute difference between the volume of a box anchored in the origin and the proportion of points from the P falling inside it. The construction of low star discrepancy point sets has traditionally relied on number-theoretic sequences such as Halton, Sobol′, and Kronecker constructions, which offer strong asymptotic guarantees but limited flexibility for fixed dimensions and sample sizes. To address this limitation, research on the optimization-driven construction of low star discrepancy point sets tailored to specific dimensions and sample sizes has received increased attention in recent years. In this work, we propose using automatic algorithm configuration to tune the parameters of Kronecker sequences, a generic tool to generate low-discrepancy point sets. The identified parameters enable us to produce 3-dimensional point sets that outperform state-of-the-art constructions for sets of at least 500 points, and are competitive with the best known point sets for smaller sizes. Interestingly, our experiments show that it is possible to discover parameter values that consistently produce good point sets, independent of the size even if it is possible to obtain size-specific parameters yielding even better sets for the chosen number of points."

Coffee break

🕒 10:30 - 11:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Poster session 1

🕒 11:00 - 12:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

In-Context Decision Making for Optimizing Complex AutoML Pipelines

Amir Rezaei Balef, Katharina Eggensperger

Combined Algorithm Selection and Hyperparameter Optimization (CASH) has been fundamental to traditional AutoML systems. However, with the advancements of pre-trained models, modern ML workflows go beyond hyperparameter optimization and often require fine-tuning, ensembling, and other adaptation techniques. While the core challenge of identifying the best-performing model for a downstream task remains, the increasing heterogeneity of ML pipelines demands novel AutoML approaches. This work extends the CASH framework to select and adapt modern ML pipelines. We propose PS-PFN to efficiently explore and exploit adapting ML pipelines by extending Posterior Sampling (PS) to the max k-armed bandit problem setup. PS-PFN leverages prior-data fitted networks (PFNs) to efficiently estimate the posterior distribution of the maximal value via in-context learning. We show how to extend this method to consider varying costs of pulling arms and to use different PFNs to model reward distributions individually per arm. Experimental results on one novel and two existing standard benchmark tasks demonstrate the superior performance of PS-PFN compared to other bandit and AutoML strategies. We make our code and data available at https://github.com/amirbalef/CASHPlus.

α-PFN: In-Context Learning Entropy Search

Tom Julian Viering, Steven Adriaensen, Herilalaina Rakotoarison, Samuel Müller, Carl Hvarfner, Frank Hutter, Eytan Bakshy

We show how Prior-data Fitted Networks (PFNs) can be adapted to efficiently predict Entropy Search (ES), an information-theoretic acquisition function. PFNs were previously shown to be able to accurately approximate Gaussian Process (GP) predictions. To approximate ES we extend them to condition on information about the optimum of the underlying function. Conditioning on this information is not straightforward and previous methods relied on complex, handcrafted, and/or computationally heavy approximations. PFNs, however, offer learned approximations that require just a single forward pass. Additionally, we train alpha-PFN, a new type of PFN model, on the information gains predicted by the first, letting us directly predict the value of the acquisition function in a single forward pass, effectively avoiding the traditional sampling-based approximations. This approach makes using Entropy Search and its variations straightforward and efficient in practice. We validate our approach empirically on synthetic GP samples of up to six dimensions, where the alpha-PFN matches or improves upon the regrets obtained by current approximations to predictive and joint Entropy Search, at a reduced computational cost. While this provides an initial proof of concept, the real potential of our method lies in its ability to efficiently perform Entropy Search for arbitrary function priors, unlike the current GP-specific approximations.

Local Entropy Search over Descent Sequences for Bayesian Optimization

David Stenger, Armin Lindicke, Alexander von Rohr, Sebastian Trimpe

Searching large and complex design spaces for a global optimum can be infeasible and unnecessary. A practical alternative is to iteratively refine the neighborhood of an initial design using local optimization methods such as gradient descent. We propose local entropy search (LES), a Bayesian optimization paradigm that explicitly targets the solutions reachable by the descent sequences of iterative optimizers. The algorithm propagates the posterior belief over the objective through the optimizer, resulting in a probability distribution over descent sequences. It then selects the next evaluation by maximizing mutual information with that distribution, using a combination of analytic entropy calculations and Monte-Carlo sampling of descent sequences. Empirical results on high-complexity synthetic objectives and benchmark problems show that LES achieves strong sample efficiency compared to existing local and global Bayesian optimization methods.

Preferential Bayesian Optimization with Crash Feedback

Johanna Menn, David Stenger, Sebastian Trimpe

Bayesian optimization is a popular black-box optimization method for parameter learning in control and robotics. It typically requires an objective function that reflects the user's optimization goal. However, in practical applications, this objective function is often inaccessible due to complex or unmeasurable performance metrics. Preferential Bayesian optimization (PBO) overcomes this limitation by leveraging human feedback through pairwise comparisons, eliminating the need for explicit performance quantification. When applying PBO to hardware systems, such as in quadcopter control, crashes can cause time-consuming experimental resets, wear and tear, or otherwise undesired outcomes. Standard PBO methods cannot incorporate feedback from such crashed experiments, resulting in the exploration of parameters that frequently lead to experimental crashes. We thus introduce CrashPBO, a user-friendly mechanism that enables users to both express preferences and report crashes during the optimization process. Benchmarking on synthetic functions shows that this mechanism reduces crashes by 63 % and increases data efficiency. Through experiments on three robotics platforms, we demonstrate the wide applicability and transferability of CrashPBO, highlighting that it provides a flexible, user-friendly framework for parameter learning with human feedback on preferences and crashes.

Hyperparameter Optimization via Interacting with Probabilistic Circuits

Jonas Seng, Fabrizio Ventola, Zhongjie Yu, Kristian Kersting

Despite the growing interest in designing truly interactive hyperparameter optimization (HPO) methods, to date, only a few allow to include human feedback. Existing interactive Bayesian optimization (BO) methods incorporate human beliefs by weighting the acquisition function with a user-defined prior distribution. However, in light of the non-trivial inner optimization of the acquisition function prevalent in BO, such weighting schemes do not always accurately reflect given user beliefs. We introduce a novel BO approach leveraging tractable probabilistic models named probabilistic circuits (PCs) as a surrogate model. PCs encode a tractable joint distribution over the hybrid hyperparameter space and evaluation scores. They enable exact conditional inference and sampling. Based on conditional sampling, we construct a novel selection policy that enables an acquisition function-free generation of candidate points (thereby eliminating the need for an additional inner-loop optimization) and ensures that user beliefs are reflected accurately in the selection policy. We provide a theoretical analysis and an extensive empirical evaluation, demonstrating that our method achieves state-of-the-art performance in standard HPO and outperforms interactive BO baselines in interactive HPO.

Reducing overtuning in hyperparameter optimization: An empirical evaluation of mitigation strategies

Sietse Schröder, Mitra Baratchi, Bernd Bischl, Matthias Feurer, Jan N. van Rijn

Hyperparameter optimization (HPO) aims to identify configurations that generalize well to unseen data, typically by minimizing validation error estimates obtained via resampling, such as holdout or cross-validation. Recent work has shown that excessive optimization of these stochastic validation estimates can lead to overtuning (Schneider et al., 2025), a degradation in true generalization performance despite continued improvements in validation performance. In this work, overtuning was observed across a wide range of HPO benchmark studies. In this ongoing work, we conduct a large-scale empirical study of how overtuning can be mitigated in practice. We systematically investigate a range of mitigation strategies, spanning resampling strategies, incumbent selection rules, and modifications to the optimizer, while revisiting the severity and determinants of overtuning on the TabArena benchmark (Erickson et al., 2025). Preliminary results suggest that certain mitigation strategies are effective at reducing overtuning in settings where overtuning is most pronounced, such as small-data regimes or when a limited resampling budget (e.g., holdout) is used. However, the same strategies may negatively impact generalization performance when overtuning is weak or absent, highlighting the importance of deeper insight into when and how overtuning mitigation should be applied. References Schneider, L., Bischl, B., & Feurer, M. (2025). Overtuning in Hyperparameter Optimization. In International Conference on Automated Machine Learning (pp. 17-1). PMLR. Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., & Hutter, F. (2025). Tabarena: A living benchmark for machine learning on tabular data. arXiv preprint arXiv:2506.16791.

Assessing Overfitting in Federated Learning Systems: an Empirical Study in Tabular Data

Omar Mohammed

Overfitting assessment in Federated Learning (FL) is ambiguous because training produces both client-local and aggregated global model states, and evaluation can target different data distributions. This means that the standard definition of overfitting is not suitable for FL. We formalize overfitting in FL by stating whether we evaluate a client-local model or the aggregated global model, and whether the evaluation uses a client-held-out test split or the shared global test split. We then propose a round-wise summary measure that aggregates per-round overfitting values into a single score using mean area-under-the-curve (mAUC). Experiments on 20 tabular datasets comparing FedAvg and FedProx show that the association between client-local overfitting and global-model overfitting is strongly regime-dependent: it is high in small federations but can become very weak in larger federations. Our results further reveal that some global overfitting measures are redundant, while measures that capture how overfitting varies across clients provide additional, non-redundant information. In summary, the results support the importance of adapting the standard view of overfitting to FL.

Risk-Averse Decision-Making in Multi-Armed Bandits

Sabrina Khurshid, Mohammed Shahid Abdulla, Gourab Ghatak.

Sharpe Ratio (SR) is a critical parameter in characterizing financial time series as it jointly considers the reward and the volatility of any stock/portfolio through its variance. Deriving online algorithms for optimizing the SR is particularly challenging since even offline policies experience constant regret with respect to the best expert Even-Dar et al (2006). Thus, instead of optimizing the usual definition of SR, we optimize regularized square SR (RSSR). We consider two settings for the RSSR, Regret Minimization (RM) and Best Arm Identification (BAI). In this regard, we propose a novel multi-armed bandit (MAB) algorithm for RM called UCB-RSSR for RSSR maximization. We derive a path-dependent concentration bound for the estimate of the RSSR. Based on that, we derive the regret guarantees of UCB-RSSR and show that it evolves as O(log n) for the two-armed bandit case played for a horizon n. We also consider a fixed budget setting for well-known BAI algorithms, i.e., sequential halving and successive rejects, and propose SHVV, SHSR, and SuRSR algorithms. We also derive the upper bound for the error probability of all proposed BAI algorithms. We demonstrate that UCB-RSSR outperforms the only other known SR optimizing bandit algorithm, U-UCB Cassel et al (2023). We also establish its efficacy with respect to other benchmarks derived from the GRA-UCB and MVTS algorithms. We further demonstrate the performance of proposed BAI algorithms for multiple different setups. Our research highlights that our proposed algorithms will find extensive applications in riskaware portfolio management problems. Consequently, our research highlights that our proposed algorithms will find extensive applications in risk-aware portfolio management problems.

TBD

Automated Algorithm Configuration for Neural Network Robustness Verification

Konstantin Kaulen, Holger Hoos

SORE : Self-Optimizing Regularized Ensemble for Evolving Data Streams

Daniel Nowak Assis, Carola Doerr, Maroua Bahri

Automatic Configuration of 3D Kronecker Sequences for Low Star Discrepancy

Imène Ait Abderrahim, Carola Doerr, Martin Durand

The L∞ star discrepancy of a point set P measures how well spread the points in P are. It is defined as the maximum absolute difference between the volume of a box anchored in the origin and the proportion of points from the P falling inside it. The construction of low star discrepancy point sets has traditionally relied on number-theoretic sequences such as Halton, Sobol', and Kronecker constructions, which offer strong asymptotic guarantees but limited flexibility for fixed dimensions and sample sizes. To address this limitation, research on the optimization-driven construction of low star discrepancy point sets tailored to specific dimensions and sample sizes has received increased attention in recent years. In this work, we propose using automatic algorithm configuration to tune the parameters of Kronecker sequences, a generic tool to generate low-discrepancy point sets. The identified parameters enable us to produce 3-dimensional point sets that outperform state-of-the-art constructions for sets of at least 500 points, and are competitive with the best known point sets for smaller sizes. Interestingly, our experiments show that it is possible to discover parameter values that consistently produce good point sets, independent of the size even if it is possible to obtain size-specific parameters yielding even better sets for the chosen number of points."

Rethinking Multilingual Embedding Model Ranking Across Learning Tasks and Languages

Ana Gjorgjevikj, Barbara Koroušić Seljak, Tome Eftimov

Text embeddings are a core component of modern natural language processing systems, enabling learning tasks such as classification, clustering, retrieval, semantic textual similarity. Many industry applications, including semantic search over text, recommender systems, and retrieval-augmented generation (RAG) pipelines, depend on vector databases built from these embeddings. However, embedding models still struggle to generalize reliably across tasks and languages, making model selection inherently task/language-dependent. Reliable insights into model generalization abilities are only possible through multilingual and multi-task platforms that incorporate variety of datasets and apply robust aggregation of model performance scores across those datasets. When datasets are highly correlated or evaluated using incomparable performance metrics, the commonly used average-based performance aggregation (employed in popular platforms such as GLUE, SuperGLUE, Big-Bench, Hugging Face's Open LLM Leaderboard, and MTEB version 1) can result in biased model rankings. Using a single score aggregation strategy is also insufficient for robust model selection. In this talk, we will present results from a comprehensive, language-aware study of the MTEB Multilingual v2 platform, offering practical guidance for language-specific model selection across approximately 230 languages. We perform both (i) language-specific, task-specific analysis and (ii) language-specific, task-agnostic analysis, ensuring robustness to both dataset composition and performance aggregation strategy selection. Our detailed analysis of five of the most widely used languages worldwide shows that large-scale LLM-based embedding models emerge as robust top performers across most, but not all tasks (e.g., retrieval), and that only a small subset of these models consistently generalizes well across different tasks, aggregation strategies, and dataset compositions.

Lunch break

🕒 12:30 - 13:30

Oral presentations 2

🕒 13:30 - 15:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

13:30 - 13:50 (15+5 min)

Towards a Foundation Model for Meta-algorithmics

Steven Adriaensen

Meta-algorithmics aims to automate algorithm selection and configuration by predicting algorithm performance from data. In practice, this still relies on narrowly scoped performance models trained on carefully curated, task-specific datasets tied to particular algorithms, instances, implementations, and execution environments. Such data is expensive to collect and rarely transferable. The resulting predictors are tightly coupled to the benchmarks they were trained on, limiting practical use and making them highly sensitive to changes in setup. Moreover, modeling performance per algorithm in isolation obscures relationships between algorithms and leads to weak or misleading uncertainty estimates over relative performance, which is the quantity that ultimately matters for robust, risk-aware decision making. This talk explores a different direction inspired by the rise of foundation models in machine learning. I will sketch a vision in which a single pretrained model adapts to new algorithmic decision problems through in-context learning rather than retraining. Instead of producing point predictions for individual algorithms, such a model would reason directly about uncertainty, relative performance, and trade-offs, conditioned on whatever performance data and contextual information is available at inference time. While this perspective is still exploratory, it suggests a path toward more data-efficient, flexible, and robust algorithmic decision making, and raises questions about how far the foundation-model paradigm can be pushed to revolutionize meta-algorithmics.

13:50 - 14:10 (15+5 min)

On the Journey of Dynamic Algorithm Configuration of Bayesian Optimization

Carolin Benjamins, many others

The prowess of Bayesian Optimization (BO) crucially depends on the exploration-exploitation trade-off (EETO) as expressed by the acquisition function (AF). To date, the choice of the AF, and parameters controlling the EETO, have been primarily static. However, manual heuristics explicitly adapting the EETO demonstrate their benefit, but either cannot adapt to the task, or are limited due to their handcrafted nature. In this talk, we venture into the development of Dynamic Algorithm Configuration (DAC) policies that control such a trade-off. Our goal is to design a data-driven policy, dynamically adjusting the EETO of BO during each single run. From switching AFs, via a meta-learned AF schedule selector, to our manual heuristic SAWEI, we explore the potential of such policies. As a next step, we explore how to learn such a policy, which requires the careful formulation of BO as a contextual Markov Decision Process (cMDP), enabling to learn a policy based on state features of BO, whilst interacting with different target objective functions during training. The maxim is, that whenever an iterative algorithm has a hyperparameter determining its behavior, adapt it over time!

14:10 - 14:30 (15+5 min)

Data Needs Tuning Too: Reassessing Machine Learning Performance Through a Data-Centric Lens

Sasa Mladenovic, Marius Lindauer, Carola Doerr

We show that current benchmarking practices in machine learning underemphasize the impact of data preparation on performance.

14:30 - 14:35 (3+2 min)

Automated Black-Box Optimization of Binary Oxide Mask Generators Using CMA-ES for Realistic Nuclear Fuel Simulations

Vojtěch Bláha, Jaroslav Knotek, Jan Blažek, Martin Holeňa

Black-box optimization plays a pivotal role in automatically configuring complex simulation pipelines where gradient information is unavailable and evaluations are computationally expensive. In this work, we formulate the automatic calibration of a parametrized synthetic binary oxide mask generator as a challenging black-box optimization problem and compare the performance of CMA-ES and its advanced variant SLSQP-LQ-CMA-ES on this task. We construct a generator with 11 parameters that synthesizes binary masks of oxide distributions on nuclear fuel cross-sections through a combination of Gaussian basis functions. To assess quality, we define a loss based on matching 20 generated masks with 20 real experimental masks using 31 descriptive features (e.g., convexity, spatial extent, edge complexity). The resulting objective function is expensive to evaluate (~2.25 h per evaluation), highlighting the need for efficient black-box optimization. Our empirical evaluation shows that both CMA-ES and SLSQP-LQ-CMA-ES are capable of identifying high-quality parameter configurations for this moderately dimensional continuous optimization problem. The proposed pipeline enables fully automated hyperparameter discovery and supports adaptation across different fuel types and oxidation stages without manual tuning. This aligns the reported research with the workshop's focus on algorithm configuration, hyperparameter optimization and AutoML. At the same time, black-box optimization connects our research also to explorative landscape analysis, and through it to meta-learning.

14:35 - 14:40 (3+2 min)

GPBench: Improving Benchmarking in Genetic Programming

Roman Kalkreuth, Marie Anastacio, Anja Jankovic, Julian Dierkes, Fabrício Olivetti de França, Holger Hoos

Genetic Programming (GP) is a search paradigm originally proposed for program synthesis but capable of tackling multiple problem domains. Over the years, GP has evolved, with many proposed variations, especially in how they represent a solution. Current GP benchmarking initiatives are fragmented, while these different representations are rarely compared to each other. As a consequence, the community is unaware of the benefits of each representation and in which situation they perform best. To alleviate this issue, we introduce GPBench, a community-driven project that aims at improving benchmarking in GP. As a first step, we present a unified framework, dubbed TinyverseGP, which facilitates benchmarking of multiple representations and problem domains, including symbolic regression, logic synthesis and policy search.

14:40 - 14:45 (3+2 min)

Hyperspherical Simplex Encoding for Continuous Optimization of Discrete Search Spaces

Anna Franke, Aaron Klein, Pascal Kerschke

Many (multi-objective) optimization algorithms - such as Bayesian optimization and evolutionary approaches (EA) - are primarily designed for continuous search spaces. When applied to problems containing categorical variables, these methods require either specialized variation operators (in the case of EAs) or an appropriate encoding scheme that embeds discrete choices into a continuous domain. Common approaches include integer encoding, which introduces an artificial ordinal structure that may mislead model-based algorithms, as they can exploit non-existent relationships implied by the imposed ordering. Another widely used approach is one-hot encoding, which substantially increases the dimensionality of the search space and thus may degrade the algorithms' performance, especially in the case of evolutionary operators, due to the larger search space. Categorical search spaces take a central role in hyperparameter optimization and, in particular, in Neural Architecture Search (NAS), where architectural decisions are discrete by nature. Therefore, efficient and structure-preserving encodings are essential for scalable (multi-objective) optimization in this domain. In this work, we introduce a geometric encoding scheme for categorical variables. A categorical variable with n levels is represented as the vertices of a regular (n−1)-dimensional simplex (e.g., triangle, tetrahedron), embedded on an (n−1)-dimensional hypersphere. By exploiting the spherical structure, each category can be expressed using (n−2)-dimensional spherical coordinates, enabling continuous optimization approaches to search directly along the hypersphere. Consequently, the proposed approach yields a lower-dimensional continuous representation compared to one-hot encoding and avoids the artificial ordering imposed by integer encoding. The proposed encoding is assessed using a simple multi-objective evolutionary algorithm and compared against integer and one-hot encodings. Performance is assessed using the hypervolume indicator.

14:45 - 14:50 (3+2 min)

Sustainable evaluation of the state of the art

Marie Anastacio, Ashlin Iser, Théo Matricon

In an active field of research such as artificial intelligence, the state of the art is in perpetual movement and the empirical evaluation of new methods requires to run them on many benchmarks instances or datasets. However, this process is time-consuming and energy-consuming, often requiring several CPU years of computation to evaluate the impact on performance of a single idea. To avoid this bottleneck, developpers can rely on smaller sets of handpicked benchmarks, leading them to overfit to the ones they choose. In recent years, several methods have been developed in the context of automated reasoning to select instances that provide enough statistical evidence to evaluate the relative performance of algorithms. However, these methods are not yest readily available in an easy-to-use form and thus seldomly used. We present a tool for sustainable benchmarking of new methods and invite the community to use it and contribute to it.

Coffee break

🕒 15:00 - 15:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Poster session 2

🕒 15:30 - 17:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Quality Diversity Optimization for Evolving Differentiating Graph Structures

Saba Sadeghi Ahouei, Aneta Neumann, Frank Neumann

We introduce a quality diversity optimization approach to evolve graph structures that are easy to solve for one algorithm but hard for another. Our new MAP-Elites algorithm generates these differentiating graphs by evolving graphs' topologies using graph generators. We conduct a comprehensive experimental investigation of chance-constrained maximum coverage problems with deterministic and stochastic weights on the vertices as an example of graph problems. The experimental results demonstrate that our method effectively evolves differentiating benchmarks and reveals the significant impact of selected structural features on problem difficulty.

Learning to Assess the Reliability of Number-of-Runs Estimation in Stochastic Optimization

Sara Gjorgjieva, Eva Tuba, Tome Eftimov

In large-scale benchmarking of stochastic optimization algorithms, the key challenge is no longer whether repeated runs are needed for reliability, but how to determine when sufficient evidence has been collected without incurring unnecessary computational cost. We study a learning-based extension of a recent empirical online heuristic that adaptively estimates the required number of runs using outlier handling and skewness-based symmetry checks. Using annotated outcomes from 132,000 Nevergrad runs on COCO (24 problems in 20 dimensions, 10 instances each, 11 optimizers), we train classifiers on 23 statistical, energy-free, and shape and stability features to predict whether a run-number estimate is reliable, prioritizing detection of incorrect estimates via minority-class recall. Across five experimental setups, ensemble models achieve high minority-class recall (typically ≥ 0.8), and cross-optimizer aggregation yields reliable models, enabling online identification of potentially unsafe early stopping decisions.

Similarity-based Portfolio Construction for Black-box Optimization

Catalin-Viorel Dinu, Diederick Vermetten, Carola Doerr

In black-box optimization, a central question is which algorithm to use to solve a given, previously unseen, problem. Selecting a single algorithm, however, entails inherent risks: inaccuracies in the selector may lead to poor choices, and even well-performing algorithms with high variance can yield unsatisfactory results in a single run. A natural remedy is to split the evaluation budget across multiple runs of potentially different algorithms. Such sequential algorithm portfolios benefit from variance reduction and complementarities between algorithms, often outperforming approaches that allocate the entire budget to a single solver. While effective portfolios can be constructed post-hoc, transferring this idea to the algorithm selection setting is non-trivial. We show that a naive portfolio constructed over the full training set already outperforms the strongest traditional baseline, the virtual best solver. We then propose a simple yet effective k-nearest-neighbor–based finetuning approach to construct portfolios tailored to unseen instances, yielding further improvements and highlighting the effectiveness of portfolio selection in fixed-budget black-box optimization.

Beyond Human Heuristics: Algorithmic Discovery for Optimization

John Abela

Many optimization problems admit enormous spaces of possible heuristics and approximation algorithms. Because Kolmogorov complexity over problem (and algorithm) encodings is unbounded, it is plausible that a substantial fraction of useful algorithms are too intricate—or simply too unlikely—to be found by unaided human search. This poster motivates algorithmic discovery as a computational alternative: define a constrained domain-specific language (DSL) for candidate heuristics, couple it with an automated evaluator, and use search (evolutionary methods, reinforcement learning, and/or LLM-guided proposal) to explore large code spaces under strict performance budgets. Recent results suggest this paradigm can exceed human-designed baselines in multiple domains, including self-play reinforcement learning in Go (AlphaGo Zero), discovery of faster matrix multiplication schemes (AlphaTensor), and improved low-level sorting routines (AlphaDev), as well as LLM-plus-evolution program search (FunSearch).

Efficient Heteroscedastic Bayesian Optimization for Risk-Aware Automated Reinforcement Learning

Mingxuan Che, Tsung-Yuan Tseng, Alexander von Rohr, Theresa Eimer, Marius Lindauer

Reinforcement learning (RL) has shown remarkable success across various complex tasks. However, the learning outcome of RL agents exhibits significant stochasticity and the distribution over the learning outcomes heavily dependents on hyperparameter selection. We propose an efficient and risk-aware heteroscedastic Bayesian optimization (ERAHBO) algorithm that models both the mean and variance of the underlying distribution as functions of the inputs. Our approach aims to efficiently identify hyperparameters that yield learning outcomes characterized by high mean and low variance. We demonstrate our method's effectiveness against popular BO algorithms across diverse RL agents and tasks.

When Are RL Hyperparameters Benign? A Landscape Study in Offline Goal-Conditioned RL

Jan Malte Töpperwien, Aditya Mohan, Marius Lindauer

Deep reinforcement learning (RL) is expensive to reproduce and deploy because performance can hinge on potentially changing optimal hyperparameter configurations. A central question is whether this brittleness is intrinsic to RL or induced by non-stationary training signals, especially temporal-difference (TD) bootstrapping. In this paper, we use offline goal-conditioned RL (GCRL) as a controlled testbed, where the data distribution is fixed and non-stationarity can be introduced explicitly via dataset mixtures and scheduled phase shifts in data quality. Across non-TD, goal-structured objectives (CRL/QRL) and a TD method (HIQL), we find a consistent separation. With fixed datasets, hyperparameter landscapes are far more benign than commonly reported in online RL, exhibiting broad near-optimal regions. Under scheduled non-stationarities, TD learning becomes markedly more fragile: phase-wise optima and hyperparameter importances drift sharply, yielding narrow, moving basins. By contrast, CRL/QRL retain stable basins once the datasets contain modest expert content (~ 20%), with sensitivity concentrated on a small set of hyperparameters. These results demonstrate that hyperparameter brittleness within a training run is not inherent to goal-conditioned RL: it is largely a bootstrapping effect, which can be avoided by using stationary non-TD objectives. When such objectives are not feasible, a concrete way of mitigating the noise in TD learning is through dynamic hyperparameter configuration during training.

Learning the Dynamic Exploration-Exploitation Trade-Off in Bayesian Optimization with Reinforcement Learning

Carolin Benjamins, Thibaut Klenke, Leona Hennig, Theresa Eimer, Carola Doerr, Marius Lindauer

The prowess of Bayesian Optimization (BO) crucially depends on the exploration-exploitation trade-off (EETO) as expressed by the acquisition function (AF). To date, the choice of the AF, and parameters controlling the EETO, have been primarily static. However, manual heuristics explicitly adapting the EETO demonstrate their benefit, but either cannot adapt to the task, or are limited due to their handcrafted nature. Our goal is to design a data-driven policy, dynamically adjusting the EETO of BO during each single run. For this, we follow the paradigm of Dynamic Algorithm Configuration (DAC) to meta-learn such a policy. The key is the careful formulation of BO as a contextual Markov Decision Process (cMDP), enabling to learn a policy based on state features of BO, whilst interacting with different target objective functions during training. We evaluate our meta-learned policies on a range of commonly used synthetic functions, BBOB, on tasks from natural sciences and engineering, as well as on the mixed-space hyperparameter optimization benchmark YAHPO-Gym. Our learned policies exhibit superior performance compared to the static EETO, and generalization capabilities across dimensions and domains.

Does Architecture Matter in Deep Reinforcement Learning? Evidence from Policy Network Design in Differentiable Inventory Simulators

Saied Farham-Nia, Ulrich W. Thonemann

Deep reinforcement learning (DRL) has become a prominent approach for solving complex sequential decision problems, yet the role of policy network architecture remains insufficiently understood, particularly in structured operations management settings. This paper investigates architectural choices in policy networks when applying direct backpropagation through differentiable simulators to inventory control problems. We study three canonical inventory settings—backlogged demand, lost sales, and demand substitution—using end-to-end differentiable simulators that enable gradient-based policy optimization without value-function approximation or sampling-based policy gradient methods. We compare a multi-layer perceptron (MLP) policy network with a substantially simpler linear policy architecture. Both policies are trained via direct stochastic gradient descent on the infinite-horizon reward by backpropagating through simulated trajectories. Our empirical results show that the linear policy consistently attains near-optimal performance across all settings, achieving less than a 1% optimality gap relative to benchmark solutions and yielding tight confidence intervals. Despite its simplicity, the linear architecture performs comparably to, and in some cases only marginally below, the deeper MLP model, while requiring significantly less computational time and substantially reduced hyperparameter tuning. These findings suggest that in systems with known dynamics and differentiable environments, such as inventory control problems, increasing architectural complexity may provide only limited additional benefit. Instead, leveraging structural properties and simulator differentiability can produce efficient, stable, and high-performing solutions using remarkably simple policy classes. Our results contribute to a clearer understanding of when deep architectures are genuinely necessary in DRL and when simpler models are sufficient.

Configuring LLM-Generated Heuristics via Instance-Specific Deep Reinforcement Learning

Xiaowei Liu, Kevin Tierney

Designing effective optimization algorithms requires both discovering high-quality heuristic functions and fine-tuned parameter configurations. Large language models (LLMs) have recently demonstrated promising performance in automated heuristic discovery, but configuring the parameters of these generated heuristics remains a challenge. Existing work relies on prompt tuning rather than integrating automated algorithm configuration with LLM-generated heuristics. This paper proposes a method that employs deep reinforcement learning to configure generated heuristic operators in an instance-specific manner. The method combines LLM-based heuristic refinement with local search to minimise the modification in the parameter space. We evaluate the proposed approach on vehicle routing problems.

BONO-Bench: A Comprehensive Test Suite for Bi-objective Numerical Optimization with Traceable Pareto Sets

Lennart Schäpermeier, Pascal Kerschke

The evaluation of heuristic optimizers on test problems, better known as benchmarking, is a cornerstone of research in multi-objective optimization. However, most test problems used in benchmarking numerical multi-objective black-box optimizers come from one of two flawed approaches: On the one hand, problems are constructed manually, which result in problems with well-understood optimal solutions, but unrealistic properties and biases. On the other hand, more realistic and complex single-objective problems are composited into multi-objective problems, but with a lack of control and understanding of problem properties.This paper proposes an extensive problem generation approach for bi-objective numerical optimization problems consisting of the combination of theoretically well-understood convex-quadratic functions into unimodal and multimodal landscapes with and without global structure. It supports configuration of test problem properties, such as the number of decision variables, local optima, Pareto front shape, plateaus in the objective space, or degree of conditioning, while maintaining theoretical tractability: The optimal front can be approximated to an arbitrary degree of precision regarding Pareto-compliant performance indicators such as the hypervolume or the exact R2 indicator. To demonstrate the generator's capabilities, a test suite of 20 problem categories, called BONO-Bench, is created and subsequently used as a basis of an illustrative benchmark study. Finally, the general approach underlying our proposed generator, together with the associated test suite, is publicly released in the Python package bonobench to facilitate reproducible benchmarking.

GPBench: Improving Benchmarking in Genetic Programming

Roman Kalkreuth, Marie Anastacio, Anja Jankovic, Julian Dierkes, Fabrício Olivetti de França and Holger Hoos

Genetic Programming (GP) is a search paradigm originally proposed for program synthesis but capable of tackling multiple problem domains. Over the years, GP has evolved, with many proposed variations, especially in how they represent a solution. Current GP benchmarking initiatives are fragmented, while these different representations are rarely compared to each other. As a consequence, the community is unaware of the benefits of each representation and in which situation they perform best. To alleviate this issue, we introduce GPBench, a community-driven project that aims at improving benchmarking in GP. As a first step, we present a unified framework, dubbed TinyverseGP, which facilitates benchmarking of multiple representations and problem domains, including symbolic regression, logic synthesis and policy search.

On the Influence of the Feature Computation Budget on Per-Instance Algorithm Selection for Black-Box Optimization

Koen van der Blom, Diederick Vermetten

Per-instance algorithm selection (PIAS) takes advantage of complementarity between a set of algorithms by deciding which algorithm to run on a given instance. This decision is based on features of the instances, which, in the context of black-box optimization (BBO), require a part of the optimization budget to be computed. This raises two questions: (a) from which fraction of the budget spent on feature computation does PIAS become worth it for BBO, and (b) which fraction of the budget optimizes the trade-off between feature accuracy and PIAS performance. To this end, we perform a broad study where PIAS with varying sampling budgets for feature computation is compared to the single best algorithm on a broad range of algorithm selection scenarios. These scenarios consist of two portfolio sizes, three problem sets, 4 dimensionalities, and 10 target budgets. We find that PIAS is viable for the majority of tested scenarios, even when as much as a quarter of the total budget is spent on feature computation. The trade-off for the fraction of the budget spent on feature computation to maximize the benefit of PIAS is highly dependent on the AS scenarios. Further, on average 20 percent of PIAS loss to the virtual best solver is explained by the budget spent on feature computation, highlighting the importance of properly accounting for the feature budget.

Discovering Interpretable Multi-Parameter Control Policies for Evolutionary Algorithms Using Deep Reinforcement Learning

Tai Nguyen, Phong Le, Carola Doerr, Nguyen Dang

While the application of Deep Reinforcement Learning (deep-RL) for parameter control in evolutionary algorithms has grown rapidly in recent years, rigorous theoretical analysis of parameter control remains largely restricted to single-parameter settings. This disparity stems from the significant challenge of deriving effective, interpretable multi-parameter control policies that are amenable to theoretical study. In this work, we demonstrate how deep-RL can be leveraged to overcome this barrier, using the (1+(\lambda,\lambda))-genetic algorithm optimizing OneMax -- one of the few problems where a super-constant speedup of dynamic control has been formally proven, as a representative case study. We first address the inherent difficulties of applying deep-RL to this multi-parameter control landscape, showing that standard approaches often struggle to converge. To enable effective learning in this complex combinatorial action space, we introduce a set of algorithm-agnostic enhancements targeting action-space decomposition, reward shifting, and credit assignment over long horizons. Equipped with these enhancements, we evaluate commonly used deep-RL methods and demonstrate that Double Deep Q-Networks (DDQN) uniquely avoid the policy collapse observed in Proximal Policy Optimization (PPO), yielding high-quality behavioral trajectories suitable for downstream analysis. Crucially, we move beyond the black-box nature of neural networks by distilling the learned behaviors into a transparent, symbolic control policy. This derived policy does not only offer interpretability for future theoretical analysis but also yields exceptional performance, consistently outperforming existing baselines across a wide range of problem sizes.

Welcome reception @ RWTH SkyLounge

🕒 17:30 - 20:30 📍 RWTH Hauptgebäude, Templergraben 55, 52062 Aachen

Tuesday, May 19

Arrival & Coffee

🕒 08:00 - 09:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Oral presentations 3

🕒 09:00 - 10:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

09:00 - 09:20 (15+5 min)

Explainable decision-making: combinatorial optimization and counterfactual explanations

Pieter Leyman

In the fields of Operations Research and Computer Science, a primary objective is to develop methods for solving complex decision-making problems, often formulated as combinatorial optimization problems (COPs). These problems are typically NP-hard, meaning the best-known exact algorithms require computational time which grows exponentially with problem size, and furthermore provide no insights into decisions as part of the optimal solution. In essence, the individual decisions made at an instance level are not explained. While the existing field on explainable AI (XAI) has flourished in the past decade, it has focused almost exclusively on explaining decisions for classification, while COPs have been left out in the cold. In this talk I will highlight the major differences between explaining classification and COP decisions, considering factuals (current situation), counterfactuals (what needs explaining) and CFEs. Based on the classical 0-1 knapsack problem, I will furthermore propose a taxonomy for CFEs in a COP context, to argue that the search for optimal CFEs is a COP in its own right. Finally, I will share some preliminary results on what is involved in solving an explaining COP (X-COP) compared to solving its COP counterpart.

09:20 - 09:40 (15+5 min)

What to Monitor and How to React? Exploring Drift Detection and Reaction for Algorithm Selection under Streaming Problem Instances

Margherita Battistotti, Manuel Lopéz-Ibañez, Julia Handl, Kate Smith-Miles, Mario Andrés Muñoz.

Streams of optimization instances arise in many practical applications, yet they remain relatively underexplored in the AAS and AAC literature. While a few recent approaches address streaming scenarios, they rely on specific assumptions about the characteristics of the stream. Despite differences in methods and assumptions, a key common challenge when facing streams is the detection of and reaction to drift. Using synthetically generated instance streams designed to reflect a range of different scenarios and employing an established drift detector, we explore alternative strategies for drift detection and reaction within an AAS framework. What should be monitored to detect drift and trigger a reaction, instance feature values or algorithm selection accuracy? Once drift is detected, how should the model be updated: by retraining on all past data, by focusing only on recent instances, or by weighting instances by 'importance', independently of their arrival? Although these design choices depend on the characteristics of the stream, we seek strategies that are robust when such characteristics are unknown.

09:40 - 10:00 (15+5 min)

Beyond Scaling Laws: The Impact of Design Choices on Downstream Performance in Large Language Models

Ana Nikolikj, Emmy Liu, Kiril Gashteovski, Tome Eftimov

10:00 - 10:05 (3+2 min)

OPL and friends: Libraries for optimisation problems, algorithms, and feature sets

Koen van der Blom, the OPL team

To benchmark, and to build algorithm selectors that represent the state of the art, we need to know what we have in terms of benchmark problems and algorithms to solve them. In black-box optimisation, this is an issue. Except for a small set of popular tools, we don't know what we have, and thus, we don't know what the state of the art in the field is. In order to get an overview of relevant optimisation problems, the optimisation problem library (OPL) was created. This library catalogues individual problems, benchmark suites, and instance generators. Each entry is annotated with high-level properties that describe the problems. This includes, for example, the number of decision variables and their types, the number of objectives, and whether there are constraints, multiple fidelities, noise, or other properties. This allows follow-up questions, such as: Which algorithm can handle these properties? A similar library that collects algorithms would allow us to get an overview of the options to solve a problem with a given set of properties. A deeper understanding of problem instances and their (dis)similarities, as well as algorithm selection systems, are facilitated by a library with feature sets, again annotated with the properties they can handle. Together, these libraries form the necessary foundations to take the next step towards thorough benchmarks and algorithm selection systems that accurately represent the state of the art and enable us to identify research gaps.

10:05 - 10:10 (3+2 min)

Per-class Algorithm Selection for Black-box Optimisation

Koen van der Blom, Carola Doerr

Algorithm selection can help both experts and non-experts to choose more effective algorithms for their problems. Commonly used per-instance algorithm selectors rely on instance-specific features, which require function evaluations to compute for black-box optimisation problems, and need to be defined for the domain of interest. If such features are not defined, or no compute budget is available for feature extraction, per-instance algorithm selection cannot be used. We propose a per-class automated algorithm selection framework that uses a priori known features, instead of features that require function evaluations to compute. With this approach, we can perform selection for classes of instances that have the same known feature values, instead of on a per-instance basis. The framework is general for all black-box optimisation problems, and can be implemented with any combination of known features, such as the problem dimensionality, the number of objectives, the types of decision variables, and the available computational budget. An implementation of our framework for single-objective continuous problems is compared to the hand-designed selector NGOpt from the Nevergrad platform. The results show that our automated framework is able to construct a highly effective selector, that outperforms NGOpt on the two considered test sets.

10:10 - 10:15 (3+2 min)

Specialising Algorithm Selectors Over Time for Recurring Problems

Koen van der Blom, Vanessa Volz, Carola Doerr

Black-box optimisation problems are typically solved using a single, a priori chosen algorithm. However, when faced with a stream of similar problem instances, learning how adjust the algorithm choice or configuration over time can lead to substantial performance improvements. Most existing work focuses on how optimisation methods can improve during an optimisation run (e.g., self-adaptation, dynamic algorithm selection). In this work, we look at how to improve from one run to the next. We propose a method to specialise an algorithm selector, trained on a broad set of problems, for a single recurring problem, by learning over time from a stream of problem instances that have to be optimised. This is achieved by performing problem classification with best-so-far performance trajectories, and retraining the selector for the identified problem(s). The use of best-so-far performance trajectories ensures applicability even for settings where no features can be computed. We describe this problem in detail, and implement a first end-to-end approach to serve as baseline for future work. Experimental results on the unconstrained single-objective noise-free continuous BBOB problems show clear improvements over a static selector for problems with 10 and 100 dimensions, while for 2 dimensions results are competitive, but not substantially better.

10:15 - 10:20 (3+2 min)

Graph-Attributed Meta-Features for Automated Algorithm Selection in Spatial Domains

Maamar Arbouz

Selecting the optimal machine learning algorithm for geospatial tasks often depends on the underlying spatial autocorrelation and structural heterogeneity of the data. Traditional meta-features (statistical descriptors) often fail to capture these higher-order interactions. This poster presents a novel approach to Algorithm Selection using Graph Neural Networks (GNNs) as meta-learners. By representing datasets as attributed superpixel graphs (based on GraphSAGE architectures), we extract structural embeddings that serve as "signatures" for specific problem instances. We demonstrate that these graph-based meta-features significantly improve the accuracy of performance prediction models compared to standard statistical baselines. We invite discussions on how this framework can be generalized to automated neural architecture search (NAS) within the COSEAL scope.

10:20 - 10:25 (3+2 min)

Comparing MechBench with BBOB via Exploratory Landscape Analysis and Performance Trajectories

Iván Olarte Rodríguez, Maria Laura Santoni, Fabian Duddeck, Carola Doerr, Thomas Bäck, Elena Raponi

Benchmarking plays a central role in the evaluation and development of optimization algorithms. The recently developed MECHBench suite introduces test problems derived from structural mechanics applications. In this work, we analyze and compare the problem instances from MechBench with those of the classical blackbox optimization benchmarking (BBOB) test suite using Exploratory Landscape Analysis (ELA) and algorithm performance profiles. Our results indicate that the two suites exhibit differences both in their landscape characteristics and in the performance of optimization algorithms. Since the MechBench problems are based on numerical simulations and thus computationally costly, we focus on Bayesian optimization methods—Vanilla-BO, TuRBO1, HEBO, and BAxUS—that are known to perform well under small budgets, complemented by CMA-ES as a population-based heuristic baseline. We find that algorithm performance and rankings vary across the two suites and dimensions (5, 10, and 20), underlining the importance of evaluating optimization methods not only on synthetic test functions but also on benchmarks that better reflect real-world scenarios.

10:25 - 10:30 (3+2 min)

Benchmarking and Transfer Learning for Hyperparameter Optimization of Graph Neural Networks

Marek Dědič, Michal Bělohlávek

Graph Neural Networks (GNNs) rely critically on effective hyperparameter optimization (HPO), but comprehensive HPO benchmarks for graph learning tasks remain limited. This paper addresses this gap by presenting an extensive comparison of HPO strategies for GNNs, ranging from simple random search to sophisticated Sequential Model-Based Optimization (SMBO) techniques like Bayesian Optimization (BO) and Tree-structured Parzen Estimators (TPE). Furthermore, we explore meta-learning using transfer learning, investigating its potential to accelerate HPO on new tasks by leveraging knowledge from past runs. Our results provide a practical comparative guide and demonstrate the viability of using meta-learned knowledge to significantly accelerate GNN HPO.

10:30 - 10:35 (3+2 min)

When More Cores Are Not Enough: Negative Results on Parallel Constraint Solving and Static Portfolios

Alessio Pellegrino

Parallelism is widely assumed to improve performance of combinatorial optimisation solvers, yet in practice, additional cores can yield limited gains or even degrade performance. In this paper, we study the effectiveness of parallel constraint solver configurations using as benchmarks the instances of the last MiniZinc Challenges, i.e., the international constraint solving competition. We evaluate the best freely available CP solvers across multiple core counts and observe that, while parallelism often reduces runtime, non-linear and instance-dependent behaviour is common because every solver exhibits cases where the sequential configuration performs better, and super-linear speedups occur only rarely. We then investigate whether simple, interpretable machine learning models can predict when allocating additional cores is beneficial and whether static portfolios of solvers can outperform the single best solver of the portfolio. Our results are largely negative: models often fail to beat a majority baseline and indicate that generalisation across unseen problem classes is difficult. Moreover, due to the existence of a dominant solver in the MiniZinc Challenge, although a static portfolio can appear advantageous in theory, shared-resource contention introduces a substantial gap that typically erases any benefit in practice. Our findings suggest that straightforward approaches to exploiting parallelism (i.e., scaling solvers blindly or running static portfolios) are insufficient, and that effective portfolio solving will likely require interference-aware dynamic selection strategies to jointly reason about solver choice and resource allocation.

Coffee break

🕒 10:30 - 11:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Poster session 3

🕒 11:00 - 12:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Rethinking Evaluation Paradigms in IBP-based Certified Training

Konstantin Kaulen, Hadar Shavit, Holger Hoos

Deep neural networks achieve strong performance on many supervised learning tasks but remain vulnerable to adversarial perturbations. Neural network verification provides mathematically rigorous robustness guarantees, yet at substantial computational cost. To mitigate this, certified training techniques optimise for verifiable robustness during training, typically inducing a trade-off between natural and certified accuracy controlled by method-specific hyperparameters. Because these metrics are inherently conflicting, the common practice of reporting a single configuration is problematic: it can mislead conclusions about overall performance and prevents unbiased assessments of the state of the art. We address this by evaluating certified training methods via Pareto front comparisons over the natural--certified accuracy trade-off. To enable fair, method-agnostic comparisons, we perform efficient automated multi-objective hyperparameter optimisation to identify a set of Pareto-optimal configurations for each method. This approach often uncovers substantial undertuning in previously reported configurations, yielding superior performance and establishing a new state of the art. Leveraging these fronts, we present the first comprehensive multi-objective comparison of certified training approaches, showing that prior advancements are less pronounced than assumed and revealing previously unreported performance complementarities.

TabPFN-RT: Predicting Algorithm Runtime Distributions with TabPFN

Hagverdi Ibrahimli, Steven Adriaensen

Predicting algorithm runtime distributions is a central problem in meta-algorithmics, yet existing neural approaches such as DistNet rely on task-specific training, careful hyperparameter tuning, and strong parametric assumptions. In this work, we study the out-of-the-box use of TabPFN, a tabular foundation model, for algorithm runtime prediction via in-context learning. With frozen, pretrained weights, TabPFN conditions directly on observed runtime data and produces predictive distributions in a single forward pass. We present an initial comparison against DistNet-style models, highlighting three key advantages: orders-of-magnitude higher data efficiency, improved uncertainty quantification arising from the absence of restrictive parametric assumptions on runtime distributions, and operational simplicity without retraining or hyperparameter optimization. In summary, our work demonstrates the potential of tabular foundation models to power next-generation meta-algorithmic systems.

Clearing the Combinatorial Fog: Tracing the Hidden Paths of TSP Heuristics

Jonathan Heins, Darrell Whitley, Pascal Kerschke

Over decades of Traveling Salesperson Problem (TSP) research, powerful heuristics have been developed that efficiently solve many TSP instances. Among them, the local search optimizer LKH and the genetic algorithm EAX stand out as the two complementary state-of-the-art solvers. Yet, the links between instance structures and solver complementarity remain obscure, i.e., it is often unclear how instance structures affect solver performance and behavior. While visualization techniques have advanced our understanding of TSP instance structures, there is still no general method to assess the behavior of TSP heuristics directly based on the structural characteristics of the instance to be solved. Existing approaches, such as search trajectory networks, are often optimized for graph-theoretic properties that may not reflect true similarities between tours. Consequently, visually comparing solvers remains difficult, especially when few identical solutions are encountered. This paper introduces a framework for visualizing the search behavior of TSP heuristics, aiming to illuminate the still poorly understood key differences in their search behavior. To this end, we adapt the dimensionality reduction technique PaCMAP to map intermediate solutions into two-dimensional space, preserving distances that meaningfully reflect tour similarity. To support exploration, we provide an interactive dashboard that allows users to inspect individual tours and visually compare selected tour pairs. The resulting visualizations reveal fundamental differences in how LKH and EAX explore the search space, including their initialization strategies and trajectory structures. We further show that a recent EAX-inspired and performance-improved LKH variant still behaves similarly to LKH, highlighting the untapped potential for further algorithmic improvements.

Graph Instance Landscapes: When Structural Similarity Does (Not) Reflect Shortest-Path Performance

Maryam Gholami Shiri, Ivana Krminac, Marko Dzukanovic, Saso Dzeroksi, Eva Tuba, Tome Eftimov

Benchmarking shortest-path algorithms is commonly based on aggregate performance over heterogeneous graph sets, which limits insight into how different search paradigms react to instance structure. We adopt an instance-landscape view of graph benchmarking by embedding graphs into a low-cost structural feature space and clustering them into regions of similar structure. Three benchmark suites are studied: weighted Erd˝osR´enyi graphs, random geometric (wireless) graphs, and real-world road networks. We evaluate four representative shortest-path solvers spanning uninformed exact search (Dijkstra), bidi-rectional exact search (bidirectional Dijkstra), heuristic-guided exact search (A$*$), and deque-based strategies (DEQ). Clustering robustness is analyzed under multiple feature-selection schemes, and runtime distributions are compared across landscape regions using non-parametric tests. While generator parameters induce stable structural regions, we find that feature-space similarity does not necessarily imply performance similarity: significant runtime shifts are frequently observed even within the same landscape region. A merged-suite analysis further shows that different benchmark families occupy largely disjoint regions. These results highlight both the potential and the limits of structural landscapes for the structure-aware benchmarking of shortest-path algorithms.

Grasynda: Graph-based Synthetic Time Series Generation

Luis Amorim, Moisés Santos, Paulo J. Azevedo, Carlos Soares, Vitor Cerqueira

Data augmentation is a crucial tool in time series forecasting, especially for deep learning architectures that require a large training sample size to generalize effectively. However, extensive datasets are not always available in real-world scenarios. Although many data augmentation methods exist, their limitations include the use of transformations that do not adequately preserve data properties. This paper introduces Grasynda, a novel graph-based approach for synthetic time series generation that: (1) converts univariate time series into a network structure using a graph representation, where each state is a node and each tran-sition is represented as a directed edge; and (2) encodes their temporal dynamics in a transition probability matrix. We performed an extensive evaluation of Grasynda as a data augmentation method for time series forecasting. We use three neural network variations on six benchmark datasets. The results indicate that Grasynda consistently outperforms other time series data augmentation methods, including ones used in state-of-the-art time series foundation models. The method and all experiments are publicly available.

Data morphing for robustness testing in time series forecasting

Diogo Ventura

The crace package: continuous racing for automatic tuning

Yunshuang Xiao, Leslie Pérez Cáceres, Franco Ardiles, Jonas Kuckling, Pablo Contreras, Thomas Stützle

Automatic algorithm configuration procedures aim at supporting the design and application of optimization algorithms by providing specialized tools that automatically adjust their parameters to effectively apply the available computational resources. The irace configurator is a state-of-the-art method that implements an iterative racing to find an elite configuration. Although irace allows parallel execution of the target algorithm, it suffers from resource utilization inefficiency due to its sequential evaluation scheme. Inspirited by irace, the crace package performs a single continuouse race instead of iterative races and supports non-sequential model convergence as well as asynchronous executions during the race. The continuous racing procedure implemented in crace can evaluate, remove and generate new configurations asynchronously, granting a high level of flexibility in the configuration process compared to the previous iterative scheme.

Privacy-Focused Attack Detection Approach for IoT Security

Nesibe Yalçın, Semih Cakir

There has been great interest in applying machine learning-supported approaches to Internet of Things (IoT) security. However, machine learning-based intrusion detection solutions have drawbacks such as the prerequisite of having all training data on a central server, the security risks in transmitting raw data from endpoints to a central server, and especially the computational cost of training large amounts of data on a single server. Federated learning has been emerged as an effective solution for mitigating or correcting these drawbacks. Federated learning, a distributed machine learning framework, focuses on sending model updates to a central server instead of raw data, reducing communication costs, and allowing information to be shared without compromising data privacy. The heterogeneity of devices in IoT networks expands the attack surface area and increases privacy concerns. This requires the design of effective security solutions and defense strategies. Intrusion detection systems based on federated learning have been proposed as an innovative approach to improve the IoT security and reduce attack surfaces. This chapter presents the solutions offered by federated learning for the security of IoT networks, highlights the potential of federated learning-based intrusion detection systems, and discusses how the reliability of model updates can be improved.

Toward a Unified Stress Testing Framework for Large Language Models in NLP

Ons Zammel

Large Language Models (LLMs) have become foundational business tools due to their fluent text generation, yet they critically lack inherent truthfulness and often produce plausible but incorrect statements, especially when uncertain. While techniques like retrieval-augmented generation (RAG) and prompt engineering can reduce hallucinations, current evaluation benchmarks inadequately capture real-world reliability issues by rewarding confident responses over admissions of uncertainty. This work proposes a systematic framework to assess LLM truthfulness by examining whether high semantic similarity scores reflect factual correctness or merely plausibility. By combining prompt engineering, external fact-checking, and quantitative metrics, the framework deliberately probes models with realistic and ambiguous claims to expose vulnerabilities hidden by traditional evaluation methods. The results aim to provide actionable insights for developing more transparent, trustworthy, and ethically responsible language models suitable for real-world deployment.

Benchmarking and Transfer Learning for Hyperparameter Optimization of Graph Neural Networks

Marek Dědič, Michal Bělohlávek

Graph Neural Networks (GNNs) rely critically on effective hyperparameter optimization (HPO), but comprehensive HPO benchmarks for graph learning tasks remain limited. This paper addresses this gap by presenting an extensive comparison of HPO strategies for GNNs, ranging from simple random search to sophisticated Sequential Model-Based Optimization (SMBO) techniques like Bayesian Optimization (BO) and Tree-structured Parzen Estimators (TPE). Furthermore, we explore meta-learning using transfer learning, investigating its potential to accelerate HPO on new tasks by leveraging knowledge from past runs. Our results provide a practical comparative guide and demonstrate the viability of using meta-learned knowledge to significantly accelerate GNN HPO.

Automated algorithm configuration with large language models

Pablo Contreras Estrada, Thomas Stützle, Leslie Pérez Cáceres

Permutation-based combinatorial optimization problems such as the Traveling Salesman Problem (TSP) and the Quadratic Assignment Problem (QAP) are standard benchmarks; however designing efficient evolutionary algorithms for them requires substantial manual effort. We combine grammar-based automatic algorithm design with large language models (LLMs) to reduce this manual effort and improve source-code efficiency. Memetic evolutionary algorithms are represented as compositions of interchangeable components encoded by a context-free grammar and automatically configured with irace within the EMILI framework. In addition, we apply a constrained, component-wise LLM-guided optimization to accelerate individual operators while preserving algorithmic behavior. Experiments on permutation problems show significant runtime reductions and competitive or improved ARPD under fixed time budgets.

Berry Picking Across Data Landscapes: Understanding Performance Sensitivity in Collaborative Filtering

Beatriz Silva, Carlos Soares, Zafeiris Kokkinogenis

Understanding how collaborative filtering algorithms respond to variations in dataset characteristics remains an open challenge, particularly in domains where the number of available datasets is limited. We propose the use of dataset morphing as a method to generate semi-synthetic datasets that support a better understanding of algorithm behavior in domains where the number of available datasets is small. Dataset morphing has recently been proposed as a strategy that generates variations of existing datasets, with gradually varying properties, called datasetoids. We illustrate the usefulness of the methodology in the domain of nutrition-focused recommender systems and evaluate five algorithms, including baseline, matrix factorization, and neural models, and relate performance to extracted metafeatures. The results show consistent performance patterns over time, with clear differences in how each algorithm responds to changes in the data. While popularity-based methods remain comparatively stable under varying conditions, neural and factorization-based approaches exhibit stronger dependence on interaction density and distributional properties.The metafeatures analysis confirms that structural descriptors capture meaningful performance variation as the data evolve.

Comparing MechBench with BBOB via Exploratory Landscape Analysis and Performance Trajectories

Iván Olarte Rodríguez, Maria Laura Santoni, Fabian Duddeck, Carola Doerr, Thomas Bäck, Elena Raponi

Benchmarking plays a central role in the evaluation and development of optimization algorithms. The recently developed MECHBench suite introduces test problems derived from structural mechanics applications. In this work, we analyze and compare the problem instances from MechBench with those of the classical blackbox optimization benchmarking (BBOB) test suite using Exploratory Landscape Analysis (ELA) and algorithm performance profiles. Our results indicate that the two suites exhibit differences both in their landscape characteristics and in the performance of optimization algorithms. Since the MechBench problems are based on numerical simulations and thus computationally costly, we focus on Bayesian optimization methods—Vanilla-BO, TuRBO1, HEBO, and BAxUS—that are known to perform well under small budgets, complemented by CMA-ES as a population-based heuristic baseline. We find that algorithm performance and rankings vary across the two suites and dimensions (5, 10, and 20), underlining the importance of evaluating optimization methods not only on synthetic test functions but also on benchmarks that better reflect real-world scenarios.

MECHBENCH: A Set of Optimization Benchmarksinspired from Structural Mechanics V2.0

Iván Olarte-Rodríguez, Maria Laura Santoni, Fabian Duddeck, Carola Doerr, Thomas Bäck, Elena Raponi

Exploring the Parameterized Complexity of TSP Operators in Multi-Objective Optimization

Narges Tavassoli Kejani, Julien Baste, Marie-Emilie Voge, Marie-Eléonore Kessaci, Laetitia Jourdan

Many real-world problems involve several objectives that conflict with each other. In such cases, the goal is not to find a single best solution, but a set of solutions that represent different trade-offs. Although parameterized complexity is a well-established framework for studying hard optimization problems, it has rarely been applied to multi-objective optimization. In this paper, we study the Multi-Objective Traveling Salesperson Problem (MO-TSP) from a parameterized complexity perspective. We focus on two common operators, namely 2-opt and swap. We consider restricted versions of these operators using a distance parameter r, and define (k,r)-neighborhoods that allow up to k applications of the operator. For both operators, we analyze the problem of finding an improved solution with respect to Pareto dominance. We prove that this problem is fixed-parameter tractable when parameterized by k, r, and the maximum edge weight W. These results provide new insight into the complexity of multi-objective optimization and show that parameterized analysis is a useful tool in this setting.

OPL and friends: Libraries for optimisation problems, algorithms, and feature sets

Koen van der Blom, the OPL team

To benchmark, and to build algorithm selectors that represent the state of the art, we need to know what we have in terms of benchmark problems and algorithms to solve them. In black-box optimisation, this is an issue. Except for a small set of popular tools, we don't know what we have, and thus, we don't know what the state of the art in the field is. In order to get an overview of relevant optimisation problems, the optimisation problem library (OPL) was created. This library catalogues individual problems, benchmark suites, and instance generators. Each entry is annotated with high-level properties that describe the problems. This includes, for example, the number of decision variables and their types, the number of objectives, and whether there are constraints, multiple fidelities, noise, or other properties. This allows follow-up questions, such as: Which algorithm can handle these properties? A similar library that collects algorithms would allow us to get an overview of the options to solve a problem with a given set of properties. A deeper understanding of problem instances and their (dis)similarities, as well as algorithm selection systems, are facilitated by a library with feature sets, again annotated with the properties they can handle. Together, these libraries form the necessary foundations to take the next step towards thorough benchmarks and algorithm selection systems that accurately represent the state of the art and enable us to identify research gaps.

Lunch break

🕒 12:30 - 13:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Breakout session

🕒 13:30 - 15:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Coffee break

🕒 15:00 - 15:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Poster session 4

🕒 15:30 - 17:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Algorithm Recommendation for Healthcare using Meta-Learning: A Multi-Dimensional Perspective

Muhammad Asad, Mois´es Santos, Carlos Soare, Irfan Khan

The integration of machine learning (ML) in healthcare is transforming diagnostics, treatment planning, and patient management. Yet the diversity of ML algorithms makes model selection challenging. Traditional trial-and-error approaches to algorithm selection are both inefficient and often lead to sub-optimal results, which is undesirable, especially in critical healthcare situations. Meta-learning automates al- gorithm recommendation by learning from past performance and is, thus, an interesting approach to address this challenge. However, most existing meta-learning studies are based on general-purpose datasets and evaluate results with generic measures (e.g., accuracy). This becomes problematic in healthcare, because datasets are expected to have specific character- istics and model evaluation relies mostly on other measures (e.g., sen- sitivity). Therefore, existing studies do not provide information about the impact that meta-learning may have in healthcare. We introduce a multi-label, multi-dimensional meta-learning framework that is suitable for complex domains, such as healthcare. Our method balances multiple evaluation dimensions simultaneously, aligning algorithm selection with the specific needs of healthcare decision-making. The system is evaluated using 17 machine learning algorithms on 116 healthcare datasets. Our re- sults confirm that 1) healthcare datasets have characteristics that lead to different algorithm behavior; and 2) meta-learning using evaluation measures that are specific to that domain will lead to diverse recom- mendations from the ones obtained with general measures. These results show that the proposed methodology is relevant for healthcare.

Data Needs Tuning Too: Reassessing Machine Learning Performance Through a Data-Centric Lens

Sasa Mladenovic, Marius Lindauer, Carola Doerr

We show that current benchmarking practices in machine learning underemphasize the impact of data preparation on performance.

Complexity effects on synthetic data quality

Mariana Oliveira, Carlos Soares

Synthetic data can help address challenges related to privacy, fairness, and robustness in machine learning, particularly in critical domains. Evaluating the quality of synthetic data requires considering multiple dimensions, including fidelity, diversity, and utility. Data synthesizers are often sensitive to the properties of training data, such as input data quality and complexity, across these dimensions. While extensive research has analysed the relationship between complexity meta-features and predictive performance, their impact on synthetic data quality remains underexplored. In this poster, we present preliminary results from an empirical study investigating how input data complexity meta-features affect the quality of synthetic datasets produced by widely used tabular data synthesizers.

Rashomon Alignment

Carlos Soares, Peter van der Putten, Bernhard Pfahringer, and Moisés Santos

We propose Rashomon Alignment (RA), a new measure to assess functional similarity between two models. Existing functional similarity measures are distributional, quantifying differences between outputs of models applied to real-world data. However, these measures can be regarded as ecologically valid only for regions in the input space represented by the available data. We introduce a geometrical perspective on functional model similarity, which estimates it across the entire data space, offering a comprehensive view of decision boundary alignment independent of any specific data distribution. We also propose Rashomon Alignment as a measure of geometrical similarity, which is computed using data uniformly sampled from the instance space. We perform an experimental analysis on more than 90 data sets, examining critical cases where model alignment diverges from predictive accuracy. Our results show that geometrical and distributional alignment provide different and complementary perspectives on the similarity between models and algorithms. RA can be used for multiple purposes, including model selection, ensemble construction, and enhanced interpretability of machine learning models and algorithms.

Automated Rashomon Set analysis for AutoML models

Katarzyna Woźnica, Zuzanna Sieńko, Katarzyna Rogalska

AutoML frameworks frequently produce multiple models that achieve similarly high predictive performance while differing substantially in complexity, feature importance, and interpretability. This phenomenon is known as the Rashomon effect, and the collection of such models forms the Rashomon Set. At the same time, near-optimal models may produce conflicting predictions for individual samples, giving rise to the predictive multiplicity problem. Despite increasing attention in the literature, these phenomena are rarely analyzed in practical machine learning workflows and are not explicitly supported in existing AutoML systems. In this work, we investigate predictive multiplicity and the Rashomon effect in the context of AutoML-generated models. We introduce a unified framework for quantifying model diversity, disagreement, and performance similarity, enabling systematic analysis of real-world classification tasks. Furthermore, we propose the Rashomon Intersection, a novel approach to identifying sets of similarly well-performing models using multiple evaluation metrics simultaneously, rather than relying on a single performance criterion. This approach provides a more nuanced view of model equivalence and disagreement, and offers insights into robustness, uncertainty, and interpretability in automated model selection.

Learning Curve Extrapolation Across Scales

Thanika Somasundaram Thillaikumar, Aanant Chand, Huy Hoang Dang, Neeratyoy Mallik, Johannes Hog

Optimizing hyperparameters of large foundation models is prohibitively expensive, prompting practitioners to train smaller proxy models to gain insights into promising configurations. These insights are then transferred to guide optimization at larger scales. However, current general hyperparameter optimization (HPO) frameworks typically do not treat model scale as a fidelity. In contrast, freeze–thaw optimization leverages partial training information but does not directly account for cross-scale transfer. In this work, we investigate how to combine both perspectives using the ifBO surrogate model. We analyze when incorporating observations from smaller scales improves predictive performance and identify regimes in which such data can instead degrade surrogate accuracy. Our results provide practical guidance on when multi-scale information is beneficial for efficient HPO of large foundation models.

An Application of Metaheuristic Optimization Algorithms in Feature Selection for IoT Security

Nesibe Yalçın, Semih Cakir

Internet of Things (IoT) is a communication network of interconnected objects. Data obtained from devices in an IoT network can be a significant potential target for attackers. These resource-constrained devices are more susceptible to cyberattacks. Intrusion detection systems play a critical role in detecting and preventing attacks in IoT enabled networks. However, the design of an effective and rapid intrusion detection system is still a challenging research topic. To address these challenges, our study focuses on enhancing the efficiency and accuracy of the intrusion detection system. Feature selection is an important mechanism for removing irrelevant and unnecessary features from large intrusiondetection system datasets. The most meaningful/effective feature subset was selected using a meta-heuristic optimization algorithm (specifically PSO or ABC). Then this subset was fed into various machine-learning (Random Forest, XGBoost, and LSTM) models for detection of cyber-attacks with high accuracy and fewer features. The reputation of the intrusion detection system was evaluated based on the number of selected features, accuracy, F1-score, and false alarm rate metrics. The results obtained confirm the expected superior performance with optimized features and the highest accuracy.

Provable Green Hyperparameter Optimization

Leona Hennig, Jasmin Brandt, Barbara Hammer, Marius Lindauer, Marcel Wever

Large-scale hyperparameter optimization (HPO) in automated machine learning (AutoML) consumes substantial computational resources, raising growing concerns about scalability and energy efficiency. In practice, prior knowledge about promising configurations is often available and exploited, yet its impact on evaluation efficiency remains poorly understood from a theoretical perspective. In this work, we study how priors over configuration quality affect the sample complexity of HPO through the formal lens of fixed-budget best-arm identification. Modeling priors directly over arm means, we derive explicit, distribution-dependent error bounds that quantify the relationship between prior quality and evaluation budget. Our analysis shows that informative priors, those concentrating probability mass on near-optimal arms, yield exponential reductions in the number of required evaluations, while uninformative or misspecified priors incur only bounded overhead compared to the prior-free case.

Dynamic Hyperparameter Importance for Efficient Multi-Objective Optimization

Daphne Theodorakopoulos, Marcel Wever, Marius Lindauer

Choosing a suitable ML model is a complex task that can depend on several objectives, e.g., accuracy, model size, fairness, inference time, or energy consumption. In practice, this requires trading off multiple, often competing, objectives through multi-objective optimization (MOO). However, existing MOO methods typically treat all hyperparameters as equally important, overlooking that hyperparameter importance (HPI) can vary significantly depending on the trade-off between objectives. We propose a novel dynamic optimization approach that prioritizes the most influential hyperparameters based on varying objective trade-offs during the search process, which accelerates empirical convergence and leads to better solutions. Building on prior work on HPI for MOO post-analysis, we now integrate HPI, calculated with HyperSHAP, into the optimization. For this, we leverage the objective weightings naturally produced by the MOO algorithm ParEGO and adapt the configuration space by fixing the unimportant hyperparameters, allowing the search to focus on the important ones. Eventually, we validate our method with diverse tasks from PyMOO and YAHPO-Gym. Empirical results demonstrate improvements in convergence speed and Pareto front quality compared to baselines.

When More Cores Are Not Enough: Negative Results on Parallel Constraint Solving and Static Portfolios

Alessio Pellegrino

Parallelism is widely assumed to improve performance of combinatorial optimisation solvers, yet in practice, additional cores can yield limited gains or even degrade performance. In this paper, we study the effectiveness of parallel constraint solver configurations using as benchmarks the instances of the last MiniZinc Challenges, i.e., the international constraint solving competition. We evaluate the best freely available CP solvers across multiple core counts and observe that, while parallelism often reduces runtime, non-linear and instance-dependent behaviour is common because every solver exhibits cases where the sequential configuration performs better, and super-linear speedups occur only rarely. We then investigate whether simple, interpretable machine learning models can predict when allocating additional cores is beneficial and whether static portfolios of solvers can outperform the single best solver of the portfolio. Our results are largely negative: models often fail to beat a majority baseline and indicate that generalisation across unseen problem classes is difficult. Moreover, due to the existence of a dominant solver in the MiniZinc Challenge, although a static portfolio can appear advantageous in theory, shared-resource contention introduces a substantial gap that typically erases any benefit in practice. Our findings suggest that straightforward approaches to exploiting parallelism (i.e., scaling solvers blindly or running static portfolios) are insufficient, and that effective portfolio solving will likely require interference-aware dynamic selection strategies to jointly reason about solver choice and resource allocation.

Automated Black-Box Optimization of Binary Oxide Mask Generators Using CMA-ES for Realistic Nuclear Fuel Simulations

Vojtěch Bláha, Jaroslav Knotek, Jan Blažek, Martin Holeňa

Black-box optimization plays a pivotal role in automatically configuring complex simulation pipelines where gradient information is unavailable and evaluations are computationally expensive. In this work, we formulate the automatic calibration of a parametrized synthetic binary oxide mask generator as a challenging black-box optimization problem and compare the performance of CMA-ES and its advanced variant SLSQP-LQ-CMA-ES on this task. We construct a generator with 11 parameters that synthesizes binary masks of oxide distributions on nuclear fuel cross-sections through a combination of Gaussian basis functions. To assess quality, we define a loss based on matching 20 generated masks with 20 real experimental masks using 31 descriptive features (e.g., convexity, spatial extent, edge complexity). The resulting objective function is expensive to evaluate (~2.25 h per evaluation), highlighting the need for efficient black-box optimization. Our empirical evaluation shows that both CMA-ES and SLSQP-LQ-CMA-ES are capable of identifying high-quality parameter configurations for this moderately dimensional continuous optimization problem. The proposed pipeline enables fully automated hyperparameter discovery and supports adaptation across different fuel types and oxidation stages without manual tuning. This aligns the reported research with the workshop's focus on algorithm configuration, hyperparameter optimization and AutoML. At the same time, black-box optimization connects our research also to explorative landscape analysis, and through it to meta-learning.

Hyperspherical Simplex Encoding for Continuous Optimization of Discrete Search Spaces

Anna Franke, Aaron Klein, Pascal Kerschke

Many (multi-objective) optimization algorithms - such as Bayesian optimization and evolutionary approaches (EA) - are primarily designed for continuous search spaces. When applied to problems containing categorical variables, these methods require either specialized variation operators (in the case of EAs) or an appropriate encoding scheme that embeds discrete choices into a continuous domain. Common approaches include integer encoding, which introduces an artificial ordinal structure that may mislead model-based algorithms, as they can exploit non-existent relationships implied by the imposed ordering. Another widely used approach is one-hot encoding, which substantially increases the dimensionality of the search space and thus may degrade the algorithms' performance, especially in the case of evolutionary operators, due to the larger search space. Categorical search spaces take a central role in hyperparameter optimization and, in particular, in Neural Architecture Search (NAS), where architectural decisions are discrete by nature. Therefore, efficient and structure-preserving encodings are essential for scalable (multi-objective) optimization in this domain. In this work, we introduce a geometric encoding scheme for categorical variables. A categorical variable with n levels is represented as the vertices of a regular (n−1)-dimensional simplex (e.g., triangle, tetrahedron), embedded on an (n−1)-dimensional hypersphere. By exploiting the spherical structure, each category can be expressed using (n−2)-dimensional spherical coordinates, enabling continuous optimization approaches to search directly along the hypersphere. Consequently, the proposed approach yields a lower-dimensional continuous representation compared to one-hot encoding and avoids the artificial ordering imposed by integer encoding. The proposed encoding is assessed using a simple multi-objective evolutionary algorithm and compared against integer and one-hot encodings. Performance is assessed using the hypervolume indicator.

Per-class Algorithm Selection for Black-box Optimisation

Koen van der Blom, Carola Doerr

Algorithm selection can help both experts and non-experts to choose more effective algorithms for their problems. Commonly used per-instance algorithm selectors rely on instance-specific features, which require function evaluations to compute for black-box optimisation problems, and need to be defined for the domain of interest. If such features are not defined, or no compute budget is available for feature extraction, per-instance algorithm selection cannot be used. We propose a per-class automated algorithm selection framework that uses a priori known features, instead of features that require function evaluations to compute. With this approach, we can perform selection for classes of instances that have the same known feature values, instead of on a per-instance basis. The framework is general for all black-box optimisation problems, and can be implemented with any combination of known features, such as the problem dimensionality, the number of objectives, the types of decision variables, and the available computational budget. An implementation of our framework for single-objective continuous problems is compared to the hand-designed selector NGOpt from the Nevergrad platform. The results show that our automated framework is able to construct a highly effective selector, that outperforms NGOpt on the two considered test sets.

ShapIEIG: Bayesian Experimental Design for Shapley Value Estimation

David Rundel, Fabian Fumagalli, Maximilian Muschalik, Bernd Bischl, Matthias Feurer

Shapley values are a principled attribution measure widely used in interpretable machine learning, but their exact computation scales exponentially with the number of players, motivating a wide range of approximation methods based on value-function evaluations of sampled coalitions. This raises the question of whether approximation accuracy can be improved by adaptively selecting coalitions for evaluation based on previous outcomes. This is particularly relevant in settings where the value function is costly, and the number of evaluations is severely limited, such as retraining-based feature importance, data valuation, and hyperparameter importance. For this purpose, we propose ShapIEIG, a Bayesian experimental design approach that approximates the expensive value function via a Gaussian process surrogate and adaptively selects coalitions based on their expected information gain about the Shapley values. Since Shapley values are a linear function of the value function, we show that the expected information gain is available in closed form and efficiently computable. In extensive experiments across diverse costly applications, our method consistently improves estimation accuracy over state-of-the-art baselines.

Dom tour

🕒 17:30 - 19:00 📍 Johannes-Paul-II.-Straße, 52062 Aachen

Workshop dinner @ Restaurant Elisenbrunnen

🕒 19:00 - 22:00 📍 Friedrich-Wilhelm-Platz 14, 52062 Aachen

Wednesday, May 20

Arrival & Coffee

🕒 08:00 - 09:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Special Session: 50 Years of Algorithm Selection

🕒 09:00 - 11:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

09:00 - 10:00 (60 min)

Keynote

Kate Smith-Miles

10:00 - 11:00 (60 min)

Panel Discussion

Kate Smith-Miles, Pascal Kerschke, and Carlos Soares. Moderator: Lars Kotthoff

11:00 - 11:05 (3+2 min)

Dynamic Algorithm Selection: Challenges and Opportunities

Diederick Vermetten, Carola Doerr

When solving an optimization problem, it is natural to adapt the used strategy parameters based on the current state of the search. Similarly, the question of selecting the right optimizer for a given new problem is a type of adaptation before the actual optimization step. These seemingly disparate ideas can be viewed as different types of the same overall adaptation procedure. In this work, we show how dynamic algorithm selection fits into the broader context of adaptive optimization algorithms. We showcase the opportunities of dynamic algorithm selection, and highlight the challenges that need to be overcome to achieve the full potential of this method. In particular, we look at the impact of using an initial algorihtm rather than a uniform random sample (as in algorithm selection), and at the impact of warm-starting the selected optimizer based on the current search trajectory.

11:05 - 11:10 (3+2 min)

How Sequential Algorithm Portfolios can benefit Black Box Optimization

Catalin-Viorel Dinu, Diederick Vermetten, Carola Doerr

In typical black-box optimization applications, the available computational budget is often allocated to a single algorithm, typically chosen based on user preference with limited knowledge about the problem at hand or according to some expert knowledge. However, we show that splitting the budget across several algorithms yield significantly better results. This approach benefits from both algorithm complementarity across diverse problems and variance reduction within individual functions, and shows that algorithm portfolios do NOT require parallel evaluation capabilities. To demonstrate the advantage of sequential algorithm portfolios, we apply it to the COCO data archive, using over 200 algorithms evaluated on the BBOB test suite. The proposed sequential portfolios consistently outperform single-algorithm baselines, achieving relative performance gains of over 14%, and offering new insights into restart mechanisms and potential for warm-started execution strategies.

Coffee break

🕒 11:00 - 11:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

Poster session 5

🕒 11:30 - 13:00 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen

How Sequential Algorithm Portfolios can benefit Black Box Optimization

Catalin-Viorel Dinu, Diederick Vermetten, Carola Doerr

In typical black-box optimization applications, the available computational budget is often allocated to a single algorithm, typically chosen based on user preference with limited knowledge about the problem at hand or according to some expert knowledge. However, we show that splitting the budget across several algorithms yield significantly better results. This approach benefits from both algorithm complementarity across diverse problems and variance reduction within individual functions, and shows that algorithm portfolios do NOT require parallel evaluation capabilities. To demonstrate the advantage of sequential algorithm portfolios, we apply it to the COCO data archive, using over 200 algorithms evaluated on the BBOB test suite. The proposed sequential portfolios consistently outperform single-algorithm baselines, achieving relative performance gains of over 14%, and offering new insights into restart mechanisms and potential for warm-started execution strategies.

What to Monitor and How to React? Exploring Drift Detection and Reaction for Algorithm Selection under Streaming Problem Instances

Margherita Battistotti, Manuel Lopéz-Ibañez, Julia Handl, Kate Smith-Miles, Mario Andrés Muñoz.

Streams of optimization instances arise in many practical applications, yet they remain relatively underexplored in the AAS and AAC literature. While a few recent approaches address streaming scenarios, they rely on specific assumptions about the characteristics of the stream. Despite differences in methods and assumptions, a key common challenge when facing streams is the detection of and reaction to drift. Using synthetically generated instance streams designed to reflect a range of different scenarios and employing an established drift detector, we explore alternative strategies for drift detection and reaction within an AAS framework. What should be monitored to detect drift and trigger a reaction, instance feature values or algorithm selection accuracy? Once drift is detected, how should the model be updated: by retraining on all past data, by focusing only on recent instances, or by weighting instances by ‘importance’, independently of their arrival? Although these design choices depend on the characteristics of the stream, we seek strategies that are robust when such characteristics are unknown.

Dynamic Algorithm Selection: Challenges and Opportunities

Diederick Vermetten, Carola Doerr

When solving an optimization problem, it is natural to adapt the used strategy parameters based on the current state of the search. Similarly, the question of selecting the right optimizer for a given new problem is a type of adaptation before the actual optimization step. These seemingly disparate ideas can be viewed as different types of the same overall adaptation procedure. In this work, we show how dynamic algorithm selection fits into the broader context of adaptive optimization algorithms. We showcase the opportunities of dynamic algorithm selection, and highlight the challenges that need to be overcome to achieve the full potential of this method. In particular, we look at the impact of using an initial algorihtm rather than a uniform random sample (as in algorithm selection), and at the impact of warm-starting the selected optimizer based on the current search trajectory.

Graph-Attributed Meta-Features for Automated Algorithm Selection in Spatial Domains

Maamar Arbouz

Selecting the optimal machine learning algorithm for geospatial tasks often depends on the underlying spatial autocorrelation and structural heterogeneity of the data. Traditional meta-features (statistical descriptors) often fail to capture these higher-order interactions. This poster presents a novel approach to Algorithm Selection using Graph Neural Networks (GNNs) as meta-learners. By representing datasets as attributed superpixel graphs (based on GraphSAGE architectures), we extract structural embeddings that serve as "signatures" for specific problem instances. We demonstrate that these graph-based meta-features significantly improve the accuracy of performance prediction models compared to standard statistical baselines. We invite discussions on how this framework can be generalized to automated neural architecture search (NAS) within the COSEAL scope.

Algorithm Selection for Verification of Neural Networks

Guilhem Ardouin, Julien Girard-Satabin, Chokri Mraidha

The problem of Algorithm Selection has been studied for several NP-hard problems, and the same observation has been made: there exists no single algorithm that dominates all the others for solving any instance of the problem. Since the problem of verifying neural networks is also NP-hard, and since it has been demonstrated that this observation also applies to this problem [1], it would be interesting to consider Algorithm Selection to improve the verification of neural networks (VNN). Our work focuses on applying Algorithm Selection for the VNN problem. We highlight an existing gap in the literature regarding this problem, where only a single attempt [2], with no public implementation, has been made. We also present our first research tracks and early results in integrating Algorithm Selection in CAISAR [3], a platform for the VNN. [1] König et al. "Critically Assessing the State of the Art in Neural Network Verification" [2] Scott et al. "Goose: A Meta-Solver for Deep Neural Network Verification" [3] Alberti et al. "The CAISAR Platform: Extending the Reach of Machine Learning Specification and Verification"

Task-Specific vs. Task-Agnostic Text Embeddings: Which Model to Select in Each Use Case?

Ana Gjorgjevikj, Barbara Koroušić Seljak, Tome Eftimov

Selecting text embedding models that perform robustly across multiple learning tasks (e.g., classification, retrieval, clustering) remains a major challenge today, making model selection highly task-dependent. Models exhibit different abilities in different tasks, so large-scale benchmarking platforms incorporating robust performance aggregation strategies are essential for better-informed model selection. This poster presents the results from a comprehensive, transparent, and robust study of English embedding model benchmarking results, aiming to support better-informed task-specific and general-purpose model selection. We analyze model rankings sensitivity to changes in the performance aggregation strategy and dataset composition across six tasks from MTEB English v2 platform. The results demonstrate that embedding robustness is highly task-dependent, with different models having top rankings across different ranking strategies in different tasks. The task-agnostic results show that only a small number of models exhibit robust generalization, maintaining stable top rankings across tasks, ranking strategies, and dataset compositions.

Context Matters in LLM-Driven Algorithm Design

Adam Viktorin, Tomas Martinek, Zuzana Kominkova Oplatkova, Roman Senkerik, Jozef Kovac, Tomas Kadavy

Large language models (LLMs) are increasingly used to generate executable algorithms within iterative evaluation-and-feedback workflows. In this work, we study LLMs as meta-optimizers that repeatedly generate Python-based black-box optimization algorithms, receive performance feedback, and refine their solutions. Using the EASE framework, we compare twelve context strategies that differ in how previous algorithms and their performance values are reintroduced into the prompt. The results show that iterative refinement can substantially improve generated optimizers, especially in the first few iterations. However, more context does not automatically lead to better performance. Moderate and well-structured feedback, particularly combinations of recent and high-performing prior solutions, generally outperforms both no-context and context-heavy strategies.

Let Humans Select the Selector -- Investigating How and Why Algorithms Are Chosen

Markus Leyser and Pascal Kerschke

In algorithm selection (and configuration), practitioners increasingly rely on automated frameworks to identify (and configure/parametrize) high-performing algorithms for given sets of problem instances. These approaches are effective at optimizing quantitative performance measures like runtime or objective value -- including aggregations thereof, like PAR10 and ERT. However, users may care about additional aspects, including transparency, stability and explainability of the resulting algorithm (configurations), or the automated process itself. The assessment of such aspects is subjective, context-dependent, and currently receives little support in workflows of algorithm selection (and configuration). We present ongoing research on a human-centered decision support layer that, in this work, focuses specifically on the algorithm selection setting. It aims at supporting users confronted with choosing between multiple algorithm selectors and their outputs, and facing the trade-off between empirical performance and explainability. Rather than attempting to define a fixed (proxy) criteria-based explainability metric for selector-algorithm pairs, the framework treats explainability as a user-dependent concept and elicits individual selection preferences through iterative user interaction. Specifically, the framework presents users with selector-dependent explainability artifacts derived from algorithm selectors, as well as tools for analyzing the trade-off of the (automated) framework's performance and behavior. These artifacts may vary depending on both the employed algorithm selector and the selected algorithm, and can therefore differ substantially in form and content. Examples include feature importance, robustness estimates and structural properties of the features, or summaries of the selector's search behavior. Importantly, users may care both about the interpretability of the selector itself and about the explanation of the algorithm it selects. The framework targets decision scenarios in which a small number of candidate algorithms (e.g., single recommendations from different selectors) are presented to the user on a given problem instance set. Based on side-by-side comparisons of these heterogeneous artifacts, users provide pairwise explainability preference judgments between candidate selector–algorithm pairs. A preference learning approach aggregates these judgments to infer latent, user-specific explainability scores. These scores are combined with empirical performance measures in an interactive visual interface that supports users in exploring their personal performance-explainability trade-offs. This allows practitioners to reason between Pareto-optimal candidates, and ultimately choose a single selector-algorithm pair that best aligns with their personal preferences and intended use case, accounting both for how the algorithm was selected and which algorithm was selected.

Algorithm Selection of Decision-based Attacks via Racing

Henning Duwe, Holger H. Hoos

Deep neural networks remain vulnerable to adversarial attacks, which seek to find small perturbations to input images that mislead them into making incorrect predictions. Tradi- tional attacks often rely on gradient information, which is typically unavailable to attackers in real-world scenarios. To address this, black-box adversarial attacks have been developed that do not depend on full access to the network. In this study, we focused on the most challenging black-box setting, decision-based attacks, where the target model only returns output labels. We found that decision-based attacks have different strengths in terms of running time, query efficiency and performance across networks, and that it is beneficial to choose an attack for each specific scenario, i.e., which network and dataset one attacks or which restrictions one faces (time or queries). We investigate the magnitude of this performance complementarity by examining 300 scenarios, showing that five of the six investigated attacks outperform the others in specific subsets of these scenarios. To leverage this, we create a new attack, DAR, the first Decision-based Attack that uses Racing for algorithm selection. We show that DAR improves average performance on multiple datasets and networks and achieves state-of-the-art performance across a wide range of scenarios.

Neural architecture and hyperparameter selection through meta-learning on time series

Erfan Moeini, Christopher Vox, Marie Anastacio, Wadie Skaf, Mitra Baratchi, Holger H Hoos

Active research in time series classification and forecasting has led to the development of a wide range of machine learning models. For practitioners, the selection of a suitable model among these, along with their hyperparameters, remains a challenging task. While automated machine learning offers approaches for automatic selection of models for a given task, the practical efficacy of these methods is often limited, due to the computational complexity of searching over a large design space and the high dimensionality of time series datasets that poses additional challenges on generalisation quality. To fill this gap, we propose a metalearning framework that transfers past knowledge from previous searches to recommend an architecture and its hyperparameters; specifically, this framework utilises a joint representation of deep neural architectures and time series datasets, and predicts the performance of neural architectures along with their hyperparameters on time series datasets. Our computational experiments reveal that the configurations proposed by our meta-learned surrogate achieve a performance gain of up to 34% on 4 out of the 8 forecasting datasets we considered and up to 60% on 36 out of 73 of our classification datasets, whilst reducing the computational cost to 10% of that required by the hyperparameter optimisation method HEBO to tune the architectures, showcasing the effectiveness of meta-learning in the time series domain.

Specialising Algorithm Selectors Over Time for Recurring Problems

Koen van der Blom, Vanessa Volz, Carola Doerr

Black-box optimisation problems are typically solved using a single, a priori chosen algorithm. However, when faced with a stream of similar problem instances, learning how adjust the algorithm choice or configuration over time can lead to substantial performance improvements. Most existing work focuses on how optimisation methods can improve during an optimisation run (e.g., self-adaptation, dynamic algorithm selection). In this work, we look at how to improve from one run to the next. We propose a method to specialise an algorithm selector, trained on a broad set of problems, for a single recurring problem, by learning over time from a stream of problem instances that have to be optimised. This is achieved by performing problem classification with best-so-far performance trajectories, and retraining the selector for the identified problem(s). The use of best-so-far performance trajectories ensures applicability even for settings where no features can be computed. We describe this problem in detail, and implement a first end-to-end approach to serve as baseline for future work. Experimental results on the unconstrained single-objective noise-free continuous BBOB problems show clear improvements over a static selector for problems with 10 and 100 dimensions, while for 2 dimensions results are competitive, but not substantially better.

Closing Session

🕒 13:00 - 13:30 📍 Stadtpalais der Erholungsgesellschaft, Reihstraße 13, 52062 Aachen