DAGSTUHL
REPORTS

Volume 15, Issue 8, August 2025

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/2192-5283

Publication date
April, 2026

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `https://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.15.8.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
https://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 25341

Software Performance Engineering

Chen Ding[*1], Charles E. Leiserson[*2], Yihan Sun[*3], and Bruce Hoppe[†4]

1 **University of Rochester, US.** cding@cs.rochester.edu
2 **MIT – Cambridge, US.** cel@mit.edu
3 **University of California – Riverside, US.** syhlalala@gmail.com
4 **Connective Associates – Arlington, US.** behoppe@mit.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25341 on software performance engineering. This seminar convened researchers from diverse intellectual communities across computer science to synthesize a collective understanding of this fragmented discipline and advance software performance engineering as a rigorous and principled scientific field in its own right. With connections established by this seminar, we are creating a unique arena for computer science researchers to cross-fertilize by sharing performance-engineering tools, techniques, opportunities, challenges, and open problems in their own domains of expertise. The major activities included 29 talks on recent research, five working groups that discussed topics like tools, community-building, education, and LLMs, and three "world cafe" sessions with in-depth conversations among participants on guiding questions for advancing software performance engineering.

Seminar August 17–22, 2025 – https://www.dagstuhl.de/25341
2012 ACM Subject Classification Computer systems organization; Computing methodologies; Software and its engineering; Theory of computation
Keywords and phrases applications, productivity tools, software performance engineering, theory and practice
Digital Object Identifier 10.4230/DagRep.15.8.1

1 Executive Summary

Chen Ding (University of Rochester, US)
Bruce Hoppe (Connective Associates – Arlington, US)
Charles E. Leiserson (MIT – Cambridge, US)
Yihan Sun (University of California – Riverside, US)

This seminar convened researchers from diverse intellectual communities across computer science who share a common interest in *software performance engineering* (SPE): making software run fast or otherwise consume few resources such as time, storage, energy, network bandwidth, etc. Seminar participants explored the role of SPE in each others' home fields of computer science and sought to synthesize a collective understanding of this fragmented discipline. The seminar aimed to coalesce a community of researchers to advance SPE as a rigorous and principled scientific field in its own right.

[*] Editor / Organizer
[†] Editorial Assistant / Collector

With the demise of Moore's Law, the time is ripe for this seminar. But despite the growing importance of SPE, research in the field is separated into tribes scattered across the traditional areas of computer science. Our major concern is how the field of computer science will fare without an SPE community that provides a common infrastructure, a shared understanding of SPE principles, and a lingua franca. We have a choice: SPE can be tedious, expensive, haphazard, and controlled by "high priests"; or it can be fun, cheap, principled, and democratically available to the average programmer.

With connections established by this Dagstuhl seminar, we are creating a unique arena for computer science researchers to cross-fertilize by sharing SPE tools, techniques, opportunities, challenges, and open problems in their own domains of expertise, with the aim of discovering commonalities that transcend individual tribes. We are promoting our arena for SPE with websites like `https://fastcode.org`, regular activities like the monthly Fastcode Seminar, and special projects like creating a SIMD tutorial with Highway.

2 Table of Contents

3 Overview of Talks

3.1 Transforming High-Performance Libraries to Domain-Specific Languages and Optimizing Compilers with BuildIt

Saman Amarashinghe (MIT – Cambridge, US)

URL https://buildit.so/

There are countless high-performance library implementations available for various domains and hardware platforms, yet Domain-Specific Languages (DSLs) and compilers remain rare. A well-designed DSL can express a far broader range of programs within a domain compared to even the most comprehensive library while also enabling domain-specific, global optimizations that go beyond hand-optimized kernels. The scarcity of high-performance DSLs stems from the complexity of building DSL compilers, which are typically large, intricate systems developed by experts.

In this talk, I will introduce BuildIt, a C++ framework designed for the rapid prototyping of high-performance DSLs. BuildIt uses a multi-stage programming approach to combine the flexibility of libraries with the performance and specialization of code generation. With BuildIt, domain experts can transform existing libraries into efficient, specialized compilers simply by modifying types of the variables. Moreover, it allows them to implement analyses and transformations without needing to write traditional compiler code. Currently, BuildIt supports code generation for multi-core CPUs and GPUs, with FPGA support coming soon. I will also showcase four DSLs created with BuildIt to highlight its power and ease of use: a reimplementation of the GraphIt graph computing language, the BREeze DSL for regular expressions, StreamIt a DSL for stream computing including PyTorch, and NetBlocks, a DSL for custom network protocol development. More information on BuildIt can be found at `https://buildit.so/`.

3.2 High-Performance Graph Analytics for Motif Finding in Neuroscience Connectome Graphs and Beyond using Arachne

David A. Bader (NJIT – Newark, US)

The growth of network-structured data across domains like neuroscience and cybersecurity demands scalable graph analytics, but complex tasks like subgraph isomorphism remain accessible only to high-performance computing (HPC) specialists. Arachne is an open-source framework that democratizes high-performance graph analytics through a Python interface while abstracting parallelism complexities. It enables advanced graph algorithms to run efficiently from laptops to supercomputers. Arachne has been adopted by Harvard researchers for the MoMo connectome visualization tool, allowing neuroscientists to draw neural motifs that are translated into attributed subgraphs and searched using our novel HiPerMotif algorithm. Key innovations include HiPerMotif, which achieves up to 66× speedups over parallel approaches.Testing on large-scale datasets including FlyWire and the H01 human brain connectome demonstrates Arachne's performance: completing complex subgraph

searches in 38 seconds versus NetworkX's 16,000+ seconds. This unified platform balances high-performance computation with accessibility, enabling researchers to extract insights from billion-scale graphs and advancing pattern matching across data-driven sciences.This research is supported in part by NSF grants CCF-2109988, OAC-2402560, and CCF-2453324.

3.3 Which RDF database is the best

Hannah Bast (Universität Freiburg, DE)

RDF is a modern variant of relational databases, with a universal schema (all data is modelled as triples) and universal identifiers (think of URLs). The query language is SPARQL, which is essentially SQL on RDF data. There are hundreds of fully-featured systems for RDF/SPARQL, from companies, open source projects, and academia. RDF and SPARQL are 100% standardized, yet the performance of these systems varies by many orders of magnitude, often contrary to the claims by the respective creators. In my talk, I will shed some light on this mystery. In a nutshell: computer scientists cannot code, open source developers slept in their algorithms class, and companies eventually stop doing research.

3.4 Lock-Free Locks

Naama Ben-David (Technion – Haifa, IL)

Locks are frequently used in concurrent systems to simplify code and ensure safe access to contended parts of memory. However, they are also known to cause bottlenecks in concurrent code, leading practitioners and theoreticians to sometimes opt for more intricate lock-free implementations. In this talk, I'll show that, despite the seeming contradiction, it is possible to design practically and theoretically efficient lock-free locks; I'll present a practical lock-free lock algorithm and our library which allows easily replacing regular locks with lock-free ones. Time permitting, I'll discuss related open problems.

3.5 Performance Engineering in a Legacy System: A Report from the Trenches

Jon Louis Bentley (Hackettstown, US)

Joint work of Jon Louis Bentley, Duffy Boyle, P. Krishnan, John Meiners

This talk describes how I spent six months embedded in an industrial-strength software team working as a consulting performance engineer. The system was big (>107 lines of code written by several hundred developers), aged (grown over several decades), important ($300M annual revenue) and slow – if our team couldn't speed it up, this cash cow and its revenue

would be catastrophically lost in a matter of months. I'll describe some of the problems that we faced, and challenge teachers of performance engineering to prepare their students to work in such a role. I'll also address how performance engineers can enhance their own performance by using LLMs. (This talk describes joint work with Duffy Boyle, P. Krishnan and John Meiners.)

3.6 SPE Expedition: Teaching Performance Engineering Through Collaborative Problem Solving

Rezaul Chowdhury (Stony Brook University, US)

This talk proposes a collaborative approach to teaching performance engineering. Building on the success of the long-standing Algorithms Reading Group at Stony Brook University, the proposed Performance Engineering (PE) meetings aim to foster a hands-on, interactive learning environment. Participants will collaboratively optimize unoptimized code through real-time coding, successive refinements, and discussions of interesting developments. The bi-weekly hybrid sessions will be open to all, recorded, and scribed, with opportunities for participants to continue optimizations between meetings and submit their code to a dedicated server for performance evaluation. An online leaderboard will track successful submissions, encouraging friendly competition and continuous improvement.

3.7 Stencil Computation with Reduced Work

Rezaul Chowdhury (Stony Brook University, US)

Joint work of Rezaul Chowdhury, Zafar Ahmad, Russell Bentley, Reilly Browne, Rathish Das, Pramod Ganapathi, Aaron Gregory, Yushen Huang, Michael Santomauro, Yimin Zhu

Stencil computations are a fundamental tool in scientific computing, modeling the evolution of physical systems on multidimensional grids over multiple timesteps. They underpin numerous applications including fluid dynamics, image processing, mechanical engineering, cellular automata, electromagnetics, and meteorology. Traditionally, stencil algorithms iterate over every grid cell for each timestep, performing Omega(TN) work for a grid of size N and T timesteps.

In this talk, I will present our recent sub-Theta(TN)-work algorithms for general linear stencil computations under aperiodic and free-space boundary conditions, achieving substantial speedups over state-of-the-art methods. Our aperiodic-grid algorithm employs an FFT-based periodic solver within a recursive divide-and-conquer framework (SPAA'21, TOPC), while our free-space algorithm uses Gaussian approximations and N-body methods (SPAA'22, ACDA'25).

More recently, we have extended our FFT-based approach to handle a broad class of problems involving multiple time-varying linear stencils applied across spatial regions with morphing boundaries, without increasing the work or span by more than a logarithmic factor in T (SPAA'25). We have also developed FFT-accelerated algorithms for a class of nonlinear stencil problems in quantitative finance, specifically for pricing American options under binomial, trinomial, and Black–Scholes–Merton models (PPoPP'24).

Implementations of most of these algorithms run significantly faster than the best existing methods.

This is joint work with current and former Stony Brook University students: Zafar Ahmad, Russell Bentley, Reilly Browne, Rathish Das, Pramod Ganapathi, Aaron Gregory, Yushen Huang, Michael Santomauro, and Yimin Zhu.

3.8 Successes and challenges in RocksDB performance at Meta

Peter C. Dillinger (Meta – Menlo Park, US)

RocksDB is a library providing a convenient key-value interface for data storage and sits on top of a local or remote filesystem. It is "ource of truth" for most of the online small-to-medium object storage at Meta. We have been involved in developing some advanced data structures such as relaxed consistency concurrent hash tables (unpublished HyperClockCache) and approximate query filters / static functions (Ribbon filter). We are looking to advance these and related areas, including other mutable concurrent data structures and other static representations of indexing or filtering data.

3.9 Relational and Complexity Theories of Locality

Chen Ding (University of Rochester, US)

Computer memory is hierarchical, making locality a fundamental concern in software performance engineering. Locality theories formalize the cost of a cache hierarchy. The formalism is necessary for optimization.

This talk overviews two recent theories of locality: the Relational Theory (Yuan et al., TACO 2019), which proves the equivalence of data movement, data reuse, and working set measures; and Data Movement Distance (Smith et al., ICS 2022), a new metric that quantifies locality by its asymptotic complexity.

3.10 Rank-polymorphism and performance: have your cake and eat it, too, with persistent asynchronous adaptive specialization

Clemens Grelck (Friedrich-Schiller-Universität Jena, DE)

Rank-polymorphic array programming systematically abstracts from structural array properties such as shape and rank. As usual, abstraction is great for software engineering, but comes at the price of performance. Specialization to rank- and shape-monomorphic intermediate code is the way out of this dilemma, but is limited in practice by the non-availability of rank and shape information at compile time.

We present a staged approach where the runtime system recompiles polymorphic intermediate code concurrently with the execution of the application itself. The resulting fast(er) code is then dynamically linked into the running application, and the runtime system, henceforth, dispatches program execution to the fast clone. Asymptotically, this yields shape-monomorphic performance for rank-polymorphic code.

We further complement asynchronous adaptive specialization with a persistence layer that stores fast clones in a repository for immediate use in further application runs that happen to use the same ranks and shapes. However, what sounds like a simple engineering solution turns out to exhibit a number of critical issues that we will briefly touch upon.

3.11 Recent Advances and Challenges in Parallel Algorithm Design

Yan Gu (University of California – Riverside, US)

With the advent of modern hardware, parallelism has been more important than ever, and top-tier conference papers reporting performance results are rarely run sequentially. Parallel algorithms have been extensively studied since the 1970s, so what's new and still needs to be explored? In this talk, I argue that there are still numerous important directions to investigate. I will briefly overview some of my recent work on graph analytics (SSSP, connectivity, k-core, etc.), data structure design (search trees, priority queues, kd-trees, etc.), and highlight ongoing challenges such as space-efficiency, synchronization costs, and the need for simplicity. If time permits, I will also discuss some future topics that may be of interest to this audience.

3.12 Compiler Technology for Multi-Scale Heterogeneity: a Data-Centric View

Mary W. Hall (University of Utah – Salt Lake City, US)

We need compilers that support the increasingly heterogeneous landscape of architectures. The end of Moore's Law and Dennard scaling has given rise to architecture specialization at multiple scales, from heterogeneous chip architectures that include chiplets and accelerators to integrated CPU+GPU nodes and heterogeneous clusters of heterogeneous nodes. This talk will focus on the role of data layout in generating code for all of these scales of heterogeneity.

3.13 Locating software performance engineering for the greater good

Bruce Hoppe (Connective Associates – Arlington, US)

The proposal for the Dagstuhl SPE seminar states: "We have a choice: SPE can be tedious, expensive, haphazard, and controlled by "high priests"; or it can be fun, cheap, principled, and democratically available to the average programmer." I call this "choosing where and how to locate SPE." In this talk, I introduce Fastcode – an open-source community for choosing where and how to locate SPE for the greater good.

3.14 LiBox: A Learned Index with Array-Like Efficiency and No Last-Mile Search

Song Jiang (University of Texas at Arlington, US)

Learned indexes can outperform traditional structures in speed and space efficiency by predicting key positions in sorted arrays. However, prediction errors necessitate costly last-mile searches, and model evaluation itself can be expensive, limiting performance gains. We present LiBox, a hierarchical, box-based learned index that eliminates these inefficiencies. LiBox partitions keys into "boxes" such that the target box can be identified with zero error using a simple linear regression function. The remaining in-box search requires only a single AVX-512 instruction, enabling highly predictable and minimal instruction counts per query. By using modest extra space to accommodate irregular key distributions, LiBox supports both read and write queries at near array-access speed. Its structure adapts reorganizations to workload patterns, sustaining high read performance while hiding update costs.

3.15 Speedcode: Software performance engineering education via the coding of didactic exercises

Timothy Kaler (MIT – Cambridge, US)

URL https://speedcode.org

This talk discusses recent work to develop structured playgrounds for learning about software performance engineering. Speedcode is an online programming platform that aims to improve the accessibility of software performance-engineering education. At its core, Speedcode provides a platform that lets users gain hands-on experience in software performance engineering and parallel programming by completing short programming exercises. Speedcode challenges users to develop fast multicore solutions for short programming problems and evaluates their code's performance and scalability in a quiesced cloud environment.

I will also discuss the development of FastCoder-Factory and Lesson which relate to synthetic data generation and collaborative learning for code-optimization tasks.

3.16 Talk: Portable Compilation in an Accelerated World

Fredrik Kjolstad (Stanford University, US)

The future of software performance engineering must include the specialized hardware that is being developed across the industry and academia. Although performance engineers play a critical role, such hardware places a large burden on the software stack and thus increases the need for compilers and programming models for productivity. I will share my thoughts on designing programming systems that permit portable compilation across disparate hardware. These programming systems must raise the level of abstraction to diverse operations on abstract collections. (I think four such collections cover the lion's share of large-scale computing.) By raising the level of abstraction and by introducing new compiler techniques, we can make programs portable across different machines and different data structures. To manage complexity, compilers should target hardware-facing abstract machines that separate the software and hardware implementations. Finally, intermediate languages can also help us describe hardware to the compiler, so that we can target it without rewriting large parts of the compiler.

3.17 Dynamic Partial Deadlock Detection and Recovery via Garbage Collection

I-Ting Angelina Lee (Washington University – St. Louis, US)

A challenge of writing concurrent message-passing pro- grams is ensuring the absence of partial deadlocks, which can cause severe memory leaks in long-running systems. The Go programming language is particularly susceptible to this problem due to its support of message passing and ease of lightweight concurrency creation.

We propose a novel dynamic technique to detect partial deadlocks by soundly approximating liveness using the garbage collector's marking phase. The approach allows systems to not only detect, but also automatically redress partial deadlocks and alleviate their impact on memory.

We implement the approach in the tool Golf, as an extension to the garbage collector of the Go runtime system and evaluate its effectiveness in a series of experiments. Preliminary results show that the approach is effective at detecting 94% and 50% of partial deadlocks in a series of microbenchmarks and the test suites of a large-scale industrial codebase, respectively. Furthermore, we deployed Golf on a real service used by Uber, and over a period of 24 hours, effectively detected 252 partial deadlocks caused by three programming errors.

3.18 Towards Zero Spawn Overhead: Work Stealing Without Deques

I-Ting Angelina Lee (Washington University – St. Louis, US)

In a randomized work-stealing scheduler, parallel speedup depends on the spawn overhead, which workers pay to allow tasks to execute in parallel, and the steal overhead, which thieves pay to start executing new work. It's important to minimize the spawn overhead, because the spawn overhead incurred by the parallel code must first be offset by parallel scalability before any speedup can be observed.

In pursuit of zero spawn overhead, this work considers a strategy that eliminates the use of deques entirely, obviating the need for a worker to perform explicit bookkeeping or set up a deque to enable parallelism. To that end, we propose DLite, a compiler and runtime ABI (Application Binary Interface) that incurs near-zero spawn overhead, empirically measured to be about 6% compared to a regular function invocation. DLite decreases the spawn overhead to almost nil, at the expense of a high steal cost. Specifically, DLite employs a backtracking strategy: When a steal attempt occurs, the victim provides its current stack and base pointers to the thief, and the thief then reconstructs the necessary state to realize the parallel execution.

We have implemented Cilk-DLite, which extends the OpenCilk platform to implement DLite. When the application has ample parallelism, Cilk-DLite exhibits similar scalability to OpenCilk with much lower spawn overhead. When the application lacks parallelism, the high steal cost in Cilk-DLite can impede scalability due to slower work distribution. We deep dive into one benchmark to analyze the performance impact of the high steal cost.

3.19 The resurgence of software performance engineering

Charles E. Leiserson (MIT – Cambridge, US)

Today, most application developers write code without much regard for how quickly it will run. Moreover, once the code is written, it is rare for it to be reengineered to run faster. Historically, gains in performance from miniaturization, codified in Moore's Law, relieved programmers from the burden of making software run fast. But with the end of Moore's Law, interest is resurging in software performance engineering: making software run fast or otherwise consume few resources such as time, storage, file IO's, network bandwidth, energy, etc. In the future, to develop innovative products and applications, programmers will need to engage in performance engineering, which is an integrative field requiring an understanding of algorithms, software, and computer architecture. Unfortunately, performance engineering is not widely taught, and most people consider it an unstructured collection of ad hoc tricks. Now is the time to establish software performance engineering as a science-based discipline in its own right alongside traditional areas of computer science.

3.20 Dataflow-Specific Algorithms for Resource-Constrained Scheduling and Memory Design

Quanquan C. Liu (Yale University – New Haven, US)

We introduce the Weighted Red-Blue Pebble Game, an extension of the classic red-blue pebble game with weighted operation costs. This weighted formulation enables constant-factor analysis of highly resource-constrained systems with bounded fast memory, unlimited slow memory, and strict energy and power constraints.

We apply our model to computational kernels in ultra-low-power brain-computer interfaces (BCIs) implanted near the brain. We express these kernels as computational directed acyclic graphs (CDAGs), enabling modular composition of operation schedules with data movement. We derive theoretically optimal schedules for a broad class of tree-structured CDAGs and apply them to on-chip memory design with circuit-level validations for power and area.

Our algorithms result in an average 63% memory area reduction and 43% static power reduction for BCI workloads–critical improvements for ensuring safe, thermally constrained operation in implantable devices. Beyond BCIs, our results underscore the broader utility of weighted pebble games in optimizing memory and I/O across resource-constrained computing environments.

3.21 HPCToolkit: A Tool for Application Performance Engineering at Scale

John Mellor-Crummey (Rice University – Houston, US)

URL https://hpctoolkit.org/

HPCToolkit is a tool for performance analysis of programs on systems ranging from desktops to GPU-accelerated supercomputers. Hardware support for instruction-level performance measurement in AMD, Intel, and NVIDIA GPUs was developed at the urging of the HPCToolkit project team. When measuring a GPU-accelerated application, HPCToolkit employs novel wait-free queues to communicate performance measurements between tool threads and application threads. To accelerate attribution of an execution's performance measurements, HPCToolkit analyzes the execution's CPU and GPU binaries to recover mappings between machine instructions and source code. To analyze terabytes of performance measurements gathered during executions at exascale, HPCToolkit employs distributed-memory parallelism, multithreading, sparse data structures, and out-of-core streaming algorithms. To support interactive exploration of profiles up to terabytes in size, HPCToolkit's hpcviewer GUI uses out-of-core methods to visualize performance data. Recently, HPCToolkit was extended with support for top-down CPU performance analysis. This talk will describe key aspects of HPCToolkit, successes analyzing applications, and some challenges ahead.

3.22 Making Waves in the Cloud: A Paradigm Shift for Scientific Computing through Compiler Technology

William S. Moses (University of Illinois – Urbana-Champaign, US)

Scientific models are today limited by compute resources, forcing approximations driven by feasibility rather than theory. They consequently miss important physical processes and decision-relevant regional details. Advances in AI-driven supercomputing – specialized tensor accelerators, AI compiler stacks, and novel distributed systems – offer unprecedented computational power. Yet, scientific applications such as ocean models, often written in Fortran, C++, or Julia and built for traditional HPC, remain largely incompatible with these technologies. This gap hampers performance portability and isolates scientific computing from rapid cloud-based innovation for AI workloads. In this work, we bridge that gap by transpiling existing programs using the MLIR compiler infrastructure. This process enables advanced optimizations, deployment on AI hardware, and automatic differentiation. In particular, we demonstrate execution of a state of the art Julia-based ocean model (Oceananigans), with >277 custom single-node CUDA kernels on thousands of distributed GPUs and Google TPUs. Our results demonstrate that cloud-based hardware and software designed for AI workloads can significantly accelerate simulations, opening a path for scientific programs to benefit from cutting-edge computational advances.

3.23 Zombie Hashing Reanimating Tombstones in a Graveyard

Prashant Pandey (Northeastern University – Boston, US)

Linear probing-based hash tables offer high data locality and are considered among the fastest in real-world applications. However, they come with an inherent tradeoff between space efficiency and speed, i.e. when the hash table approaches full capacity, its performance tends to decline considerably due to an effect known as primary clustering. As a result they are only used at low load factors.

Tombstones (markers for deleted elements) can help mitigate the effect of primary clustering in linear probing hash tables. However, tombstones require periodic redistribution, which, in turn, requires a complete halt of regular operations. This makes linear probing not suitable in practical applications where periodic halts are unacceptable.

In this talk, we present a solution to forestall primary clustering in linear probing hash tables, ensuring high data locality and consistent performance even at high load factors. Our approach redistributes tombstones within small windows, deamortizing the cost of mitigating primary clustering and eliminating the need for periodic halts. We provide theoretical guarantees that our deamortization method is asymptotically optimal in efficiency and cost. We also design an efficient implementation within dominant linear-probing hash tables and show performance improvements.

We introduce Zombie hashing in two variants: ordered (compact) and unordered (vectorized) linear probing hash tables. Both variants achieve consistent, high throughput and lowest variance in operation latency compared to other state-of-the-art hash tables across numerous

churn cycles, while maintaining 95% space efficiency without downtime. Our results show that Zombie hashing overcomes the limitations of linear probing while preserving high data locality.

3.24 Concurrent Size

Erez Petrank (Technion – Haifa, IL)

The size of a data structure (i.e., the number of elements in it) is a widely used property of a data set. However, for concurrent programs, obtaining a correct size efficiently is non-trivial. In fact, the literature does not offer a mechanism to obtain a correct (linearizable) size of a concurrent data set without resorting to inefficient solutions, such as taking a full snapshot of the data structure to count the elements, or acquiring one global lock in all update and size operations. I will talk about a methodology for adding a concurrent linearizable size operation to sets and dictionaries with a relatively low performance overhead.

3.25 Rank-Polymorphism for High-Level Performance Engineering

Sven-Bodo Scholz (Radboud University Nijmegen, NL)

URL https://www.sac-home.org/

Rank-Polymorphism relates to the ability of specifying algorithms that operate on arrays of statically undetermined dimensionality and shape. As it turns out, this capability is instrumental for enabling various performance-related aspects such as parallelism and locality to be captured in array shapes rather than explicit loop nests. At the example of SaC (www.sac-home.org), we demonstrate how this can be achieved, delivering competitive performance from very concise, easily verifiable, rank-polymorphic specifications.

3.26 Automating Performance Optimization of Data Flow Within HPC Workflows

Nathan Tallent (Pacific Northwest National Lab. – Richland, US)

Scientific workflows that require HPC resources are critical in many areas of scientific exploration. Because these workflows tend to be data intensive, severe bottlenecks emerge in storage systems and I/O networks. Although there has been much prior work on coordination of workflows, scheduling algorithms, and HPC storage systems, there are no comprehensive workflow performance diagnosis suites that can automatically identify and alleviate bottlenecks. We present DataFlowDrs, a new comprehensive suite of tools for performance optimization of HPC workflows that especially focuses on data flow and storage. Our suite

introduces (a) lightweight high-resolution tools for measurement and profiling of executions; (b) novel methods for automatically predicting data flow bottlenecks after only a 3-5 runs using automatically generated interpretable models of data flow; (c) effective performance analysis and bottleneck detection that can automatically rank order bottlenecks for different combinations of task parallelism and storage resources; (d) actionable performance optimization in the form of new schedules and resource assignments.

3.27 Memory Systems Challenges in Graph Processing

Hans Vandierendonck (Queen's University of Belfast, GB)

Graph analytics are reputed for posing significant challenges to the memory system, which is primarily due to "random" memory access patterns. In this talk, we revisit performance models and characterisation of the memory system issues. We discuss two solutions to the problem: vectorisation and data compaction and evaluate how they alleviate the bottleneck presented by the memory system.

3.28 Parallelism-corrected profiling

Jan Wassenberg (Google – Zürich, CH)

Amdahl's Law – that serial sections ultimately limit parallel speedup – is often obscured by modern profilers. In fork-join parallelism, standard tools misrepresent execution costs, allowing serial bottlenecks to hide in plain sight. We introduce Parallelism-corrected profiling, a straightforward method to adjust profiler outputs according to their "wall time" contribution. We apply this method to LLM inference on CPU, finding several bottlenecks that resulted in an easy 1.4x speedup.

3.29 Benchmarking and developing dynamic-graph data structures

Helen Xu (Georgia Institute of Technology – Atlanta, US)

This talk will cover everything you need to know about how to design efficient parallel applications that operate on dynamic graphs! It will focus on three key aspects:(1) How do you design the containers (ie, data structures) that encapsulate the dynamic graph? (2) What is the right framework (ie, interface) for interacting with a dynamic graph? (3) How do you fairly benchmark the performance of your parallel application for dynamic graphs?

To answer these questions, this talk will discuss two main results. First, I will present a new container for dynamic graphs called F-Graph. F-Graph is a multicore batch-parallel dynamic-graph system that is optimized for spatial locality. It is built on top of a batch-parallel packed-memory array, yielding fast performance for a variety of graph applications.

Next, I will present BYO, a unified framework for large-scale graph containers designed to facilitate benchmarking. BYO provides a simple and abstract container API, along with a clean interface. The evaluation uses BYO to evaluate 27 different graph containers on 10 different graph algorithms using 10 large graph datasets. The resulting data illuminates the issues and tradeoffs involved in designing parallel applications for dynamic graphs. Overall, the talk hopes to provide insight into both the theory and practice of efficient parallel computation for dynamic graphs.

4 Working groups

4.1 SPE and LLMs

Saman Amarashinghe (MIT – Cambridge, US), David A. Bader (NJIT – Newark, US), Jon Louis Bentley (Hackettstown, US), Albert Cohen (Google – Paris, FR), Timothy Kaler (MIT – Cambridge, US), Charles E. Leiserson (MIT – Cambridge, US), and Quanquan C. Liu (Yale University – New Haven, US)

4.1.1 ML Impact on Performance Engineering

Discussion of topics for further discussion

- Popular topics emerged from voting:
  - How will software performance engineering (SPE) change/improve with LLMs?
  - What are the LLM-related research topics for SPE?
- Group split into two discussion groups focusing on these areas
- Initial demonstration showed live coding with LLM generating a 400+ line app in minutes

What are the LLM-related research topics for SPE

- Role of correctness in LLM-generated code:
  - Shift from generating correct code by design to verifying correctness after generation
  - Two approaches debated:
    - Use existing tools to constrain LLMs and increase correctness probability
    - Let LLMs "go wild" with superoptimization, then verify afterward
- Tool integration questions:
  - Whether existing performance engineering tools remain relevant in LLM environment
  - Scheduling languages (like Mary Hall's morning presentation) could become LLM vocabulary
    - LLMs could generate schedules that compose into correct, high-performance code
    - Domain-specific languages (DSLs) designed for machine consumption
- Transferability research opportunities:
  - LLMs translating optimized code between languages (e.g., C/OpenMP to Rust)
  - Challenge: very few parallel code bases exist for training
  - Transpiler research evolution from syntactic to semantic approaches
- Expanded scope beyond LLMs to general ML approaches

Common thread: correctness

- Central challenge: making it easier for ML to generate correct code
  - LLMs can generate code faster than verification tools can check it
  - Much faster generation than humans can comprehend or analyze
- Consensus that LLMs cannot guarantee correctness in foreseeable future
  - Verification will likely remain harder than code generation
  - Need for dynamic correctness checking beyond static analysis
- Bridge approach proposed: use current tools to guide LLMs without complete restriction

LLM can generate code

- Speed advantages:
  - Faster than correctness tools can verify
  - Much faster than humans can comprehend
- Agreement that trusting LLM correctness remains problematic
- Resources for staying current:
  - Martin Maas blog (though may not be recently updated)
  - ML for systems conferences and SysML conferences

What are LLMs good for

- Performance diagnosis and analysis:
  - Processing complex performance traces and providing optimization suggestions
  - Identifying serialization bottlenecks in large-scale traces
  - Acting as intelligent performance analysis assistant
- Autotuning replacement potential
- Research assistance:
  - "DeepResearch" mode for literature review
  - AI co-scientist frameworks for hypothesis generation
  - Example: Google DeepMind's multi-agent system identifying research directions
- Hypothesis generation and literature synthesis:
  - Processing 200+ papers quickly for initial analysis
  - Generating ranked research direction recommendations

What SPE can do for LLM?

- Jan Wassenberg identified as expert resource for this perspective
- Topic noted but not extensively discussed in this session

Tools to help us

- Research tools mentioned:
  - Research Rabbit: citation graph tracing and clustering
  - Notebook LM: high-confidence summarization with papers loaded in context
  - Consensus AI: survey and research assistance
- Limitations noted:
  - Research Rabbit includes tangentially related papers
  - Need for careful curation before feeding into analysis tools

Questions

- Most interesting/productive use of LLMs for SPE research priorities
- Research topics to explore in ML for systems community
- Correctness approaches:
  - Design-time correctness vs. post-generation verification
  - Integrating theorem provers with LLM workflows
- Promising applications:
  - LLMs for performance data analysis
  - Compiler error message interpretation and debugging assistance
  - Survey paper generation as starting point for human refinement

4.1.2 Roundtable Discussion Notes

How would you use LLMs for coding; how do you get started?

1. Tim: use $10 credit on OpenAI playground and play around with prompts.
   a. Try Cursor. To study the models, get the API for OpenAI.
2. David: all models have strengths and weaknesses. Pay for the more advanced models, not just use the free version; the paid version will be much better.
   a. Mileage may vary with each model and version number.
   b. There may be a meta-model that calls all other models.
   c. I'm sold on Anthropic Claude, does better than ChatGPT.
   d. Deepseek is quite good as well.
3. Jon: What was the nature of the prompts? Context offers a great deal.
4. Saman: What's the difference between using interfaces (on the web) and agents?
   a. Tim: Can create custom workflows with agents. Now I'm using Cursor and I don't even need to make a custom workflow. For unstructured things like analyzing data or asking a topic I don't have any expertise in, I use OpenAI Playground. The interesting workflow right now is just to use Cursor. Iterate through generating and testing.
   b. David: Copy and paste from emacs.

What do you usually do using LLMs?

1. Saman: Can you write me an app for voting for topics to discuss for this discussion session? It did two things wrong. Then I put screenshots stating the problems. Then, it didn't start. Then, I said make these three changes. Then, it worked.
2. David: how do we benchmark the LLM-generated code? What is a systematic way for benchmarking the LLMs to each other? Is there a way to systematically check all combinations of optimizations? Is there a way to check all possible combinations of optimizations?
3. Jon: I've talked to LLMs to make summaries and transfer the summaries from one context to another. Can refer to details about everything.
4. Saman: We were trying to write a full Apple app. At some point, we ran out of context and then we couldn't get past it.
5. Jon: Gemini has much more context than ChatGPT; 1 million tokens versus 50k which makes a big difference.
6. David: We've used Rag to deal with context issues.
7. Jon: Use case: read a book and asked what are the lessons for software engineers from this book?

8. Tim: LLMs are particularly good at acting as experts. They are particularly good at being experts at AI.
   a. Use cases: using LLMs for productivity tools. You can use LLMs to evaluate LLMs using these tools. You can use them for user evaluations. It's an objective evaluation of productivity tools.
   b. It's good to have formal verification but people don't want to write these formal verification languages.
9. Saman: It'll be great to write a formal verification proof for transforms. Give a proof that the software is correct.
   a. David: We still need a verifier for verifying the formal verification proof for transforms.
10. Saman: can we use LLM fine-tuning to train on machine code and have it output machine code?
11. Jon: good blog posts: a) use cases for SPE b) what can you use LLMs in real life.
    a. Saman: 5 use cases for LLMs that people in our community have used it for.
12. Jon: give a technical paper to an LLM and let it give critique for the paper.
13. Saman: how do I improve my chances for acceptance? They gave a lot of good ideas.
14. Long discussion on using LLMs for recommendations and bios.
15. Tim: if there's something that's verification, you can use reinforcement learning to fine-tune the LLM.

Looking into the future

16. Jon: in the future will there still be books? Or will LLMs supercede books?
17. Saman: using LLMs, you don't have to go through the middle part about understanding why an answer is correct. All previous generations have to go through this training. Every generation, you lose some abilities.
18. Charles: but what do you get back?
19. David: I think in 2 or three years, all coding will be done in English.

4.2 SPE Education (Curriculum and Teaching)

Naama Ben-David (Technion – Haifa, IL), Saman Amarashinghe (MIT – Cambridge, US), Rezaul Chowdhury (Stony Brook University, US), Bruce Hoppe (Connective Associates – Arlington, US), Song Jiang (University of Texas at Arlington, US), Timothy Kaler (MIT – Cambridge, US), Fredrik Kjolstad (Stanford University, US), Charles E. Leiserson (MIT – Cambridge, US), Nikolai Maas (KIT – Karlsruher Institut für Technologie, DE), Manuel Penschuck (Goethe University – Frankfurt am Main, DE), Hans Vandierendonck (Queen's University of Belfast, GB), Marvin Williams (KIT – Karlsruher Institut für Technologie, DE), and Helen Xu (Georgia Institute of Technology – Atlanta, US)

4.2.1 SPE Curriculum

Context: MIT 6.106, course on SPE, is restricted to multi-core (parallelism). It is not: concurrency; distributed systems; databases; networking; etc. However, students are trained in general ideas, not specifics of each of the different areas, which they can acquire on the job.

Observation: multi-core course sufficient as initial course for undergraduates.

Summary of discussion

- Consensus on core set of SPE curriculum: experimental skills on performance measurement, tools enabling detective-work, examples of relevant optimisations
- Parallel programming may have to take precedence over single-core optimisation to make memory system optimisations relevant on many-core processors
- Advanced courses can branch out in different directions and are largely independent, e.g., accelerators, file systems, data bases, distributed systems

4.2.2 Teaching SPE

Who we are and why we are here

We all teach something closely related to SPE:
- Teaching Algorithm Engineering & related courses
- Wanting to offer new course on SPE
- Teaching DB systems & scalable algorithms course; starting course on SPE
- Already teaching course related to SPE

The fastcode SPE instructors community

`https://fastcode.org/get-involved/instructors/`

- offers helpful resources and a "meta-class" on teaching SPE
- Plan: discuss how to use the materials
- Valuable conversations about teaching are hard because of private political restrictions within each institution.

Opinions on MPI course

- in the past, no test harness or benchmark framework was provided
  - Try to use educational cluster, but has high variability -> e.g. due to other workloads
  - AWS is expensive and a lot of work
  - Fastcode has some infrastructure resources
    - speedcode coding platform
    - project on providing a script system is on the way
  - Is it possible to use a simulator or instruction counts? -> problematic if parallelism is involved
- It is valuable to offer resources which are immediately usable
- Students sometimes prefer to run code online instead of locally
  - It is possible to teach them how to debug locally by providing an introduction / instructions

Should we teach profilers / how to best do it

- start with a debugging assignment
- it is not really possible to force students to use a profiler, but you can encourage them
- competition on optimizing code (online leaderboard) helps

Curriculum: how does the basic course look like

- there is a page on fastcode.org with public slides
- there is a lot of different things that should be included in such a course
- using simulators (e.g. simulate cache misses in Python) can help to make concepts more accessible, especially for more theory-leaning students
- it can be valuable to show differences between theoretical expectations and practice

Experience from talking with practitioners

- systems they work on are severely restricted
- joy of SPE comes from finding things that work within such an environment
- how to share this with students?

It is important to teach about concepts / guiding principles

- instead of only showing how things work in practice (course on low-level languages vs. OS course)
- differentiate programming as a tool from the abstract concepts
- decoupling is really important in all fields in CS
  - but often loses performance -> SPE tends to break down some abstractions

4.3 Ten Questions of SPE

Jon Louis Bentley (Hackettstown, US)

Q1: What, precisely, is Software Performance Engineering (SPE)?

Please give a definition that is useful for deciding what work falls into this area. Various tests for any proposed definition.

- Is it broad enough to include all of the people at this workshop? If not, should it?
- Does it include the items in Q2 and Q3 below?

Do we need both a regular definition and a "Big Tent" definition? What is the relationship of SPE to other fields, including

- Compiler Optimization – is this a subfield of SPE?
- Algorithm Engineering – Is this identical to SPE?
- Software Engineering – Is SPE a subfield of SE?

What CS topics clearly are *not* part of SPE?

- For any field X, the efficient implementation of X objects falls within SPE.
- Excluding that connection, what CS fields are not part of SPE?
- A theory-only field, such as Axiomatic Complexity Theory

What is research in SPE? Is there research in SPE, or does SPE build on the research done in other fields? Potential definitions include the following components:

- *SPE is Engineering*: A branch of science and technology concerned with the design, building and use of computing systems.

- *SPE is about Performance*: Performance engineering is a systematic approach to optimizing the efficiency of software and systems throughout their entire lifecycle.
- SPE is about Performance Engineering *applied to software systems*.

One LLM Generated Definition: Software Performance Engineering (SPE) is a systematic, data-driven discipline focused on the design, measurement, and optimization of computing systems to meet defined performance goals throughout their entire lifecycle.

Q2: What are some classic readings in SPE?

Books

- Abbott & Fisher, The Art of Scalability
- Bentley, Writing Efficient Programs
- Kounev, Systems Benchmarking
- McGeoch, A Guide to Experimental Algorithmics
- Sites, Understanding Software Dynamics

Articles

- Knuth, An Empirical Study of Fortran Programs
- Bentley & McIlroy, Engineering a Sort Function

Q3: Give an overview, a map, of the field of SPE today

Questions Q1, Q2 and Q3 are intimately related. The responses to Q2 and Q3 will depend on the definition in Q1. The definition of Q1 can be tested by whether it includes the responses to Q2 and Q3.

Q4: What are the Fundamental Principles of SPE?

- Measure first, and measure often.
- Reality is the only judge.
- Understand the entire stack.
- Performance is a continuous process, not a one-time event.

Q5: What are the Essential Techniques of SPE?

(This is distinct but perhaps related to the Essential Tools of SPE)
Benchmarking as a technique...

Q6: What are the Ten Big Problems of SPE today?

What is a Big Problem? Do we mean "grand challenges" or do we mean "looming risks"? The two can be closely related: any obvious looming risk implies the grand challenge of mitigating that risk. Correctness is one such looming risk that implies a grand challenge

Q7: How should an SPE determine when to stop optimizing?

What are different ways of setting performance goals? How to choose how much effort to expend in increasing performance?

- How to estimate what future performance requirements will be?
- How to choose safety factors?
- How to balance human productivity and machine performance?

List different criteria for stopping the performance improvement process.

Q8: What is the value of SPE? Please illustrate with examples.

This is easy to see when things go wrong, but how to illustrate the value when everything is going right? (The benefits of keeping healthy rather than curing illnesses.) What is there to lose if we do not invest in SPE?

Q9: What is the role of asymptotic analyses in SPE?

Probably not a goal in itself, but one of many useful tools.

QUESTIONS COVERED BY OTHER GROUPS

- What are the essential tools for an SPE?
- What roles can LLMs and other ML tools play in SPE? What should be the structure of an SPE undergraduate curriculum?
- Should there be a graduate SPE curriculum? If so, what should its structure be?

4.4 SPE Tools

John Mellor-Crummey (Rice University – Houston, US) and Jan Wassenberg (Google – Zürich, CH)

4.4.1 Diverse set of architectures

- Cerebras
  - Tiled architecture with 48KB memory per tile
  - CSL programming model supports HPC applications beyond deep learning
  - Student worked on targeting with custom applications, required vendor engagement to extend communication capabilities
  - G42 from Abu Dhabi made major investment to keep company viable
- Neuromorphic architectures
  - Most are actually just ML-targeted accelerators, not true neuromorphic
- Sambanova
  - Dataflow graph compilation model
  - Tooling would require compiler passes that decorate graphs for instrumentation
  - Programming model not publicly exposed
- GROQ
  - Intel expressed interest in targeting this architecture
  - Availability unclear for purchase
- Risk-V based accelerators
  - STX accelerator from Fraunhofer
    - Stencil Tensor accelerator with data management cores and compute engines
    - Two-level offloading model using bastardized OpenMP
  - LLVM extensions being developed

4.4.2 What kinds of tools

- Programming systems and compilers
  - Most vendors building proprietary MLIR compilers
  - Limited software accessibility beyond deep learning frameworks
- Runtime systems
  - Message passing and vector streaming communication models
  - Fine-grain partitioning required for tile architectures
- Performance tools
  - Simulators provide timing and memory contention information
  - Analysis tools need architecture-specific development
  - 80% of specialized hardware expected to disappear within couple years

4.4.3 Challenges

- Proprietary hardware dominance
  - Vendors unwilling to open source competitive advantages
  - Secret sauce optimizations kept internal
- F64 support disappearing
  - Requires analysis to determine when reduced precision acceptable
  - F64 emulation with F16 requires dozens of passes
  - Need tools to verify correctness of reduced precision results
- Most programming systems focused on machine learning
  - PyTorch, TensorFlow primarily ML-oriented
  - HPC applications struggle to find suitable tooling
- Industry vs open source divide
  - AI development unlike Internet era – no unified direction
  - Industry investment outpacing federal funding dramatically
  - Department of Energy focused on GPU-accelerated supercomputers, not novel architectures

4.4.4 Challenge: Open source tools

- Programming models seeking commonality
  - DSLs as vehicle for targeting multiple systems
  - Abstraction layers closer to application level needed
  - Rust interest for safe concurrent software development
- Promising approaches
  - **Triton**: Tile-level loop fusion programming model for GPUs
  - **PALLAS**
    - Embedded in JAX (Python)
    - References to array slices with explicit placement control
    - Can generate Triton code
    - Open source status confirmed
    - Higher level abstractions than Triton
  - **ONNX**: Common graph representation
    - Powerful enough to express non-ML computations
    - Still requires sophisticated partitioning for tile architectures
    - Need higher-level models to generate ONNX representations

- **JAX**:
  - JIT compilation for NumPy interface
  - Supports CPU, GPU, TPU platforms
  - Used at Google for weather simulation and other scientific computing
  - Allows custom code integration while bypassing autograd
- **Single Assignment C (SAC)**:
  - Generates C/CUDA code from high-level descriptions
  - Targets multiple architectures including potential new ones
- IR target options: LLVM + MLIR
  - LLVM becoming de facto standard (Intel, IBM, Google, Microsoft support Clang)
  - MLIR offers flexibility but fragmented into many dialects
  - Convergence toward MHLO dialect for high-level operations
  - MHLO used for non-ML applications (compute sin, cos operations)
  - Challenge: MLIR not designed for heterogeneous programming
  - Need explicit parallelism mapping from implicit dataflow graphs

(Note: StableHLO is apparently the preferred replacement for/derivative of MHLO.)

4.5 "Highlights of SPE" Workshop

Yihan Sun (University of California – Riverside, US), Yan Gu (University of California – Riverside, US), Rob Johnson (Broadcom – San Jose, US), and Prashant Pandey (Northeastern University – Boston, US)

Below are very brief notes of the discussions:

- **Name?** Highlights of Performance Engineering (HOPE) or Highlights of Performance Engineering on Software (HOPES)?
- **When/Where?** Possible plan: 2027 PPoPP/HPCA/CGO/CC/.. Co-located conferences on multiple relevant areas; good location (Salt Lake City); flexible timeline. It doesn't need to be always at PPoPP, though. Maybe we can reconsider the venue every year based on time/location/...
- **How to submit?**
  - Submissions will be based on "highlights", which are papers published elsewhere but highly relevant to SPE. Accepted papers will be invited to give talks.
  - We can make it submission+invitation based. The PC (or an even smaller committee) may recommend papers, but submissions are open to the public.
  - Papers will be evaluated by a 2-4 pages abstract. Full papers should be provided. Accepted papers will be invited to give talks; more papers may be accepted as posters. The quality of the write-up is important in the review process as it is an indicator of the talk quality.
  - A submission can contain multiple papers or a series of work. In this case the abstract should provide a nice overview of them.
  - Accepted manuscripts can be published as abstract/workshop short papers on ACM DL.

- **Other discussions**
  - Tracks? Expected number of submissions/acceptance rate? Balance topics?
  - Possible funding for student traveling? (would make it more attractive to get more submissions)
  - For invited papers, should we always accept or they still compete with other papers normally?

Participants

- Saman Amarashinghe
MIT – Cambridge, US
- David A. Bader
NJIT – Newark, US
- Hannah Bast
Universität Freiburg, DE
- Naama Ben-David
Technion – Haifa, IL
- Jon Louis Bentley
Hackettstown, US
- Rezaul Chowdhury
Stony Brook University, US
- Florina M. Ciorba
Universität Basel, CH
- Albert Cohen
Google – Paris, FR
- Alexander Conway
Cornell Tech – New York, US
- Peter C. Dillinger
Meta – Menlo Park, US
- Chen Ding
University of Rochester, US
- Clemens Grelck
Friedrich-Schiller-Universität Jena, DE
- Yan Gu
University of California – Riverside, US
- Mary W. Hall
University of Utah – Salt Lake City, US
- Bruce Hoppe
Connective Associates – Arlington, US
- Song Jiang
University of Texas at Arlington, US
- Rob Johnson
Broadcom – San Jose, US
- Timothy Kaler
MIT – Cambridge, US
- Fredrik Kjolstad
Stanford University, US
- I-Ting Angelina Lee
Washington University – St. Louis, US
- Charles E. Leiserson
MIT – Cambridge, US
- Quanquan C. Liu
Yale University – New Haven, US
- Nikolai Maas
KIT – Karlsruher Institut für Technologie, DE
- John Mellor-Crummey
Rice University – Houston, US
- William S. Moses
University of Illinois – Urbana-Champaign, US
- Prashant Pandey
Northeastern University – Boston, US
- Manuel Penschuck
Goethe University – Frankfurt am Main, DE
- Erez Petrank
Technion – Haifa, IL
- Sven-Bodo Scholz
Radboud University Nijmegen, NL
- Diane Souvaine
Tufts University – Medford, US
- Yihan Sun
University of California – Riverside, US
- Nathan Tallent
Pacific Northwest National Lab. – Richland, US
- Hans Vandierendonck
Queen's University of Belfast, GB
- David Wajc
Technion – Haifa, IL
- Jan Wassenberg
Google – Zürich, CH
- Marvin Williams
KIT – Karlsruher Institut für Technologie, DE
- Helen Xu
Georgia Institute of Technology – Atlanta, US

Report from Dagstuhl Seminar 25342

Frontiers of Parameterized Algorithmics of Matching under Preferences

Jiehua Chen[*1], Christine Cheng[*2], David Manlove[*3], Ildikó Schlotter[*4], and Manuel Sorge[†5]

1 **TU Wien, AT.** jiehua.chen@tuwien.ac.at
2 **University of Wisconsin – Milwaukee, US.** ccheng@uwm.edu
3 **University of Glasgow, GB.** david.manlove@glasgow.ac.uk
4 **ELTE KRTK – Budapest, HU.** ildiko@krtk.elte.hu
5 **TU Wien, AT.** manuel.sorge@ac.tuwien.ac.at

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25342 "Frontiers of Parameterized Algorithmics of Matching under Preferences", held from August 17–22, 2025. The seminar brought together researchers from the Matching Under Preferences (MATCH-UP) and Parameterized Complexity Theory (PCT) communities to systematically apply parameterized techniques to computationally hard matching problems. The program included tutorials on parameterized algorithmics, surveys on MATCH-UP complexity and structure of stable matchings, contributed talks, and intensive working group sessions that explored fundamental open problems. This seminar represents the first focused effort to comprehensively map the parameterized complexity landscape of matching markets, establishing frameworks for ongoing collaboration between these communities. The report presents abstracts of talks, tutorials, working groups, and open problems in alphabetical order by speaker.

Seminar August 17–22, 2025 – https://www.dagstuhl.de/25342
2012 ACM Subject Classification Theory of computation → Algorithmic game theory and mechanism design; Theory of computation → Computational complexity and cryptography; Theory of computation → Design and analysis of algorithms; Theory of computation → Parameterized complexity and exact algorithms; Theory of computation
Keywords and phrases Algorithmic design and complexity analysis, Matching markets, Matching theory, Parameterizec complexity analysis
Digital Object Identifier 10.4230/DagRep.15.8.29

1 Executive Summary

Jiehua Chen (TU Wien, AT)
Christine Cheng (University of Wisconsin – Milwaukee, US)
David Manlove (University of Glasgow, GB)
Ildikó Schlotter (ELTE KRTK – Budapest, HU)

Matching Under Preferences (MATCH-UP) is a research field which investigates the complexities and algorithms of matching markets, where agents or entities are paired based on individual preferences to meet certain criteria such as stability or fairness. Although matching

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Frontiers of Parameterized Algorithmics of Matching under Preferences, *Dagstuhl Reports*, Vol. 15, Issue 8, pp. 29–45
Editors: Jiehua Chen, Christine Cheng, David Manlove, and Ildikó Schlotter
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

markets have broad real-world applications – including labor allocation, organ exchanges, and educational placements – the computational landscape of the problems therein often presents as NP-complete or beyond.

Parameterized Complexity Theory (PCT), established by Downey and Fellows in 1980s, has emerged as a robust framework for dissecting the computational intricacies of NP-hard problems. However, the primary application of PCT has largely been confined to graph-theoretical challenges.

In the last five to ten years, there has been a surge in parameterized investigations of problems that fall within the MATCH-UP area. However, such research only constitutes a small part of the literature, and there are numerous topics where our understanding of the computational complexity of the problems involved, and the parameters influencing them, is insufficient.

This seminar addressed a fundamental gap: the lack of comprehensive parameterized analysis across the MATCH-UP landscape. By convening experts from the MATCH-UP and PCT communities, we aimed to identify which structural restrictions render hard matching problems tractable, and which parameters provably cannot help under standard complexity assumptions.

Seminar Composition and Structure

The seminar brought together 23 researchers from the USA, UK, Japan, India, and across Europe. The balanced composition of PCT and MATCH-UP experts enabled genuine bidirectional knowledge transfer between communities that rarely interact at traditional venues.

The week-long program comprised:

- Two tutorials providing crash courses on parameterized algorithmics, covering fixed-parameter tractability, kernelization techniques, and treewidth-based algorithm design
- Two survey talks on recent parameterized complexity results for MATCH-UP problems and structural properties of stable matchings
- One application-focused talk highlighting real-world deployment challenges in matching markets
- Eight contributed talks presenting current research on MATCH-UP and PCT
- Two rump sessions for open problem proposals and challenge identification
- Seven working group sessions for intensive collaborative investigation
- Multiple plenary discussions where working groups reported progress and input was obtained from the wider audience

Notably, the assignment of working groups was computed using an optimal matching tool[1] developed by one of the organizers (David Manlove) to produce a popular matching respecting participant preferences.

Outcomes and Impact

The seminar enabled productive exchange between the PCT and MATCH-UP communities, establishing shared vocabulary and research frameworks. The seven working groups explored fundamental questions including:

- Identifying structural parameters specific to MATCH-UP instances, such as agent types
- Designing dynamic algorithms for stable matching

[1] Freely available at `https://matwa.optimalmatching.com`

- Parameterized complexity of weighted envy-free matchings
- Computing competitive equilibrium in house allocation
- Parameterized approximation for stable matching variants

Conclusion

Dagstuhl Seminar 25342 was the first seminar focusing on recent advancement of parameterized complexity research in the practically motivated field of matching markets. The organizers believe it was a successful meeting, bringing together researchers from the PCT community to work on MATCH-UP problems and enabling productive cross-disciplinary exchange.

We hope that this seminar will act as a springboard to future synergies between the PCT and MATCH-UP communities, and that this will be reflected at events such as the MATCH-UP series of workshops in the future

Acknowledgments

The organizers thank the Dagstuhl staff for their outstanding professional support, all participants for their engaging contributions to tutorials, talks, and working groups, and Manuel Sorge for collecting abstracts of contributed talks and working group results.

Christine Cheng, Jiehua Chen, David Manlove, and Ildikó Schlotter

2 Table of Contents

3 Overview of Talks

3.1 Complexity of optimal and stable exchange problems

Péter Biró (HUN-REN KRTK – Budapest, HU)

In this talk we survey some recent developments on the CS/OR aspects of kidney exchange programmes (KEPs). These programmes have been established in most of the Western countries in the last two decades to facilitate the exchange of living donors for those recipients who have willing, but immunologically incompatible donors. In the first part we describe the European practices including the hierarchical optimisation used in the matching runs of these KEPs for computing optimal solutions, and the first algorithm implemented in the UK, which was an FPT algorithm with a special parameter. In the second part we describe an alternative solution concept based on the individual fairness notion of stability. Depending on the nature of blocking coalition, we study the core, the competitive equilibrium, and the strong core of the corresponding game.

3.2 Matchings under preferences: Strength of stability and trade-offs

Jiehua Chen (TU Wien, AT)

Joint work of Jiehua Chen, Piotr Skowron, Manuel Sorge

We propose two solution concepts for matchings under preferences: *robustness* and *near stability*. The former strengthens while the latter relaxes the classical definition of stability by Gale and Shapley [1]. Informally speaking, robustness requires that a matching must be stable in the classical sense, even if the agents slightly change their preferences. Near stability, on the other hand, imposes that a matching must become stable (again, in the classical sense) provided the agents are willing to adjust their preferences a bit. Both of our concepts are quantitative; together they provide means for a fine-grained analysis of the stability of matchings. Moreover, our concepts allow the exploration of trade-offs between stability and other criteria of social optimality, such as the egalitarian cost and the number of unmatched agents. We investigate the computational complexity of finding matchings that implement certain predefined trade-offs. We provide a polynomial-time algorithm that, given agent preferences, returns a socially optimal robust matching (if it exists), and we prove that finding a socially optimal and nearly stable matching is computationally hard.

References

1 David Gale and Lloyd S. Shapley. *College Admissions and the Stability of Marriage*. The American Mathematical Monthly 120 (5), p. 386–391, 1962.

3.3 The structure of stable matchings

Christine Cheng (University of Wisconsin – Milwaukee, US)

The set of stable matchings has a very interesting structure in both the Stable Marriage (SM) and Stable Roommates (SR) settings. For SM, I describe the role of posets and distributive lattices and how they are used to design polynomial-time algorithms or prove hardness results. For SR, I show how the closed subsets of posets are generalized to the complete closed subsets of mirror posets and how distributive lattices are generalized to median graphs. I end the talk by noting that structural results on stable matchings can be exported to other fields that study distributive lattices and median graphs.

References

1 D. Gusfield and R.W. Irving *The Stable Marriage Problem: Structure and Algorithms*, The MIT Press, 1989.
2 C. Cheng. *Understanding the Generalized Median Stable Matchings*, Algorithmica 58:1 (2010) pp. 34-51.
3 C. Cheng and A. Lin. *Stable Roommates Matchings, Mirror Posets, Median Graphs and the Local/Global Median Phenomenon in Stable Matchings*, SIAM Journal on Discrete Mathematics 25:1 (2011) pp. 72-94.
4 C. Cheng *A Poset-based Approach to Embedding Median Graphs in Hypercubes and Lattices*, Order 29 (2012), pp. 147-163.
5 C. Cheng, E. McDermid and I. Suzuki. *Eccentricity, Center and Radius Computations on the Cover Graphs of Distributive Lattices with Applications to Stable Matchings*, Discrete Applied Mathematics 205 (2016) pp. 27-34.

3.4 Diversity constraints in stable many-one matching

Thekla Hamm (TU Eindhoven, NL)

We consider bipartite stable many-one matching in combination with so called *diversity constraints*: entities on the "many"-side have certain (possibly intersecting) types and entities on the "one"-side have lower and upper bounds constraining how many entities of each type must be and are allowed to be matched to them respectively. This can for example be used to model college admission with affirmative action towards certain groups of students. In joint work with Jiehua Chen and Robert Ganian [1] we showed that this problem is complete for the second level of the polynomial hierarchy and carried out an extensive analysis of its paraNP-hardness versus XP-time solvability with respect to all combinations of a set of natural parameters. In this talk, I gave an overview of these results.

References

1 J. Chen, R. Ganian and T. Hamm. *Stable Matchings with Diversity Constraints: Affirmative Action is beyond NP*. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, 2020

3.5 Strongly stable matchings and non-wastefulness

Naoyuki Kamiyama (Kyushu University – Fukuoka, JP)

Main reference Naoyuki Kamiyama: "The Strongly Stable Matching Problem with Closures", CoRR, Vol. abs/2401.02666, 2024.
URL https://doi.org//10.48550/ARXIV.2401.02666

In this talk, we consider the problem of finding a matching between two disjoint groups D, H of agents. Each agent has a preference over a subset of the agents in the other group, and the preferences may contain ties. Strong stability is one of the well-studied properties of a matching in this setting. In this talk, we consider the following variant of strong stability. We are given a subset S of H. If an agent in S is not matched to any partner in the current matching, then this agent cannot become a member of a blocking pair. We prove that the problem of checking the existence of a strongly stable matching in this setting is generally hard, and give some polynomial-time solvable cases. Interestingly, one positive result gives a unified approach to the strongly stable matching problem in the ordinary setting and the envy-free matching problem.

3.6 Adapting stable matchings to evolving preferences

Dušan Knop (Czech Technical University – Prague, CZ)

We continue and extend previous work on the parameterized complexity analysis of the NP-hard Stable Roommates with Ties and Incomplete Lists problem, thereby strengthening earlier results both on the side of parameterized hardness as well as on the side of fixed-parameter tractability. Other than for its famous sister problem Stable Marriage which focuses on a bipartite scenario, Stable Roommates with Incomplete Lists allows for arbitrary acceptability graphs whose edges specify the possible matchings of each two agents (agents are represented by graph vertices). Herein, incomplete lists and ties reflect the fact that in realistic application scenarios the agents cannot bring all other agents into a linear order. Among our main contributions is to show that it is W[1]-hard to compute a maximum-cardinality stable matching for acceptability graphs of bounded treedepth, bounded tree-cut width, and bounded disjoint paths modulator number (these are each time the respective parameters). However, if we 'only' ask for perfect stable matchings or the mere existence of a stable matching, then we obtain fixed-parameter tractability with respect to tree-cut width but not with respect to treedepth. On the positive side, we also provide fixed-parameter tractability results for the parameter feedback edge set number.

3.7 From applications to problems in matching under preferences

David Manlove (University of Glasgow, GB)

Matching problems involving ordinal preferences and cardinal utilities arise in many practical applications. In this talk I will outline four applications that I have been involved in over the last 25 years, namely course allocation, project allocation, resident doctor allocation and kidney exchange. In each case I will describe the application and define the underlying matching model that arises from it. I will then give an overview of the known algorithmic results and suggest some directions for future research, especially from the perspective of parameterised complexity.

3.8 FPT tutorial, part I

Dániel Marx (CISPA – Saarbrücken, DE)

In my talk, I will give a brief overview of the motivation and main definition of parameterized algorithms. I introduce some of the standard techniques that are used to show parameterized problems to be fixed-parameter tractable. In particular, I show how matroid-based techniques are relevant for matching problems.

3.9 An application of matching under preferences in team formation

Matthias Mnich (TU Hamburg, DE)

We describe complex matching scheme for assigning teams of human topic experts to projects, in such a way that several hard and soft criteria for the assignment are satisfied. Moreover, there are preferences of experts over projects, and over other experts, who they would like to be in a team with and who they prefer not having as team mates. Structural insights into the nature of the data lets us decompose highly intractable problem into simpler phases, each of which can be solved separately and the solutions to which can be amalgamated to a full assignment. The theoretical findings for the algorithm design are then turned into an efficient implementation and evaluated on real data set of experts and projects, leading to significant savings in computation time over previously used manual approaches.

3.10 FPT tutorial, part II

Marcin Pilipczuk (University of Warsaw, PL)

In the second part, we will explore the rich world of structural parameters. Then, we will discuss tractable parameterizations of Integer Programming and examples of IP-based parameterized algorithms.

3.11 Stable matchings with groups of similar agents

Baharak Rastegari (University of Southampton, GB)

Many important stable matching problems are known to be NP-hard, even when strong restrictions are placed on the input. In this work we seek to identify structural properties of instances of stable matching problems which will allow us to design efficient algorithms using elementary techniques. We focus on the setting in which all agents involved in some matching problem can be partitioned into k different types, where the type of an agent determines his or her preferences, and agents have preferences over types (which may be refined by more detailed preferences within a single type). This situation would arise in practice if agents form preferences solely based on some small collection of agents' attributes. We also consider a generalisation in which each agent may consider some small collection of other agents to be exceptional, and rank these in a way that is not consistent with their types; this could happen in practice if agents have prior contact with a small number of candidates. We show that (for the case without exceptions), several well-studied NP-hard stable matching problems including Max SMTI (that of finding the maximum cardinality stable matching in an instance of stable marriage with ties and incomplete lists) belong to the parameterised complexity class FPT when parameterised by the number of different types of agents needed to describe the instance. For Max SMTI this tractability result can be extended to the setting in which each agent promotes at most one "exceptional" candidate to the top of his/her list (when preferences within types are not refined), but the problem remains NP-hard if preference lists can contain two or more exceptions and the exceptional candidates can be placed anywhere in the preference lists, even if the number of types is bounded by a constant.

3.12 Stable matchings with restricted preferences

Will Rosenbaum (University of Liverpool, GB)

We consider the stable marriage problem when agents' preferences are restricted by (1) bounded preference lists, (2) number of attributes, and (3) the "range" of the preference lists. We show that models 1 and 2 realize arbitrary rotation posets for bound parameters.

Consequently, many problems that are hard in general, e.g., counting stable matchings, are hard in these restricted models. On the other hand, k-range instance have rotation posets with pathwidth $O(k^2)$. Consequently, the following problems admit FPT algorithms parameterized by the instance's range:

1. counting stable matchings,
2. sampling stable matchings uniformly, and
3. finding balanced, median, and sex-equal stable matchings.

3.13 Recent advances in parameterized matching markets

Ildikó Schlotter (ELTE KRTK – Budapest, HU)

This talk presents recent advances in parameterized algorithmics for matching markets. We discuss parameterized complexity results for hard problems that emerge from the classical stable matching framework when incorporating constraints such as fairness or diversity, as well as from extensions that address aspects of control, dynamic or multidimensional settings.

4 Working groups

4.1 Computing maximum size competitive equilibrium solutions in Shapley-Scarf housing markets

Péter Biró (HUN-REN KRTK – Budapest, HU), Jiehua Chen (TU Wien, AT), Gergely Csáji (Eötvös Lorand University – Budapest, HU), Simon Mauras (INRIA Saclay – Île-de-France, FR), and Ildikó Schlotter (ELTE KRTK – Budapest, HU)

In the Shapley-Scarf housing markets the agents have houses and preferences over the others' houses. The market solution is an exchange with no payment allowed. When preferences are strict then the strong core coincides with the set of competitive equilibrium (CE) allocations and it can be obtained by the Top Trading Cycles (TTC) algorithm of Gale. However, when the preferences are weak then the strong core can be empty, and the set of CE allocations can be large, obtained by the TTC algorithm with tie-breakings. Motivated by the kidney exchange programmes, we ask the complexity of the problem of finding a CE allocation of maximum size. If the problem is NP-hard, are there FPT-algorithm with parameters representing the length and structure of the ties?

4.2 Complexity and algorithms for partial w-envy-free many-one matchings

Robert Bredereck (TU Clausthal, DE), Tamás Fleiner (Budapest University of Technology & Economics, HU), Naoyuki Kamiyama (Kyushu University – Fukuoka, JP), and Dušan Knop (Czech Technical University – Prague, CZ)

Motivated by the open questions on partial envy-free allocations of indivisible goods [1], where envy-free matchings serve as a key algorithmic tool, our working group focused on advancing the understanding of w-envy-free many-one matching problems.

Interestingly, the literature (notably Aigner-Horev and Segal-Halevi [2]) had left open whether polynomial-time algorithms exist even for simpler envy-free 1-to-1 matchings with arbitrary value functions. On the positive side, by adapting an algorithmic approach from Gan, Suksompong, and Voudouris [3], our working group answered one of the open questions posed by Aigner-Horev and Segal-Halevi [2]. Moreover, our investigation revealed inherent computational hardness: we prove NP-hardness already for 2-to-1 envy-free matchings with very restricted weight functions, which unfortunately implies that the open questions from [1] cannot be fully resolved by relying solely on envy-free matchings as an algorithmic tool.

Our results illuminate a nuanced picture where some natural classes of envy-free matching problems remain tractable while others become computationally intractable even under simple restrictions. Our ongoing work aims to further delineate this boundary, finding new algorithmic methods and complexity results to better understand fair division under envy-freeness constraints.

References

1 R. Bredereck, A. Kaczmarczyk, J. Luo, and B. Sun. *Computing Efficient Envy-Free Partial Allocations of Indivisible Goods*. In: Sanmay Das, Ann Nowé, and Yevgeniy Vorobeychik (eds.), Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, MI, USA, May 19–23, 2025, pp. 390–398. Int. Foundation for Autonomous Agents and Multiagent Systems / ACM, 2025. DOI: 10.5555/3709347.3743553.

2 E. Aigner-Horev and E. Segal-Halevi. *Envy-free matchings in bipartite graphs and their applications to fair division*. Information Sciences 587 (2022): 164–187. DOI: 10.1016/j.ins.2021.11.059.

3 J. Gan, W. Suksompong, and A. A. Voudouris. *Envy-freeness in house allocation problems*. Mathematical Social Sciences 101 (2019): 104–106. DOI: 10.1016/j.mathsocsci.2019.07.005.

4.3 FPT-Approximability of stable matching problems

Jiehua Chen (TU Wien, AT), Shuichi Miyazaki (University of Hyogo – Kobe, JP), and Ildikó Schlotter (ELTE KRTK – Budapest, HU)

Main reference Jiehua Chen, Sanjukta Roy, Sofia Simola: FPT-Approximability of Stable Matching Problems. CoRR abs/2508.10129 (2025)
URL https://arxiv.org/abs/2508.10129

Many optimization problems regarding stable matchings are NP-hard to approximate and their decision variants remain parameterized intractable. Recently, Chen, Roy, and Simola [1] initiated the study of parameterized approximation on three NP-hard stable matching variants:

1. Min-BP-SMI: Given a stable marriage instance and a number k, find a size-at-least-k matching that minimizes the number b of blocking pairs;
2. Min-BP-SRI: Given a stable roommates instance, find a matching that minimizes the number b of blocking pairs;
3. Max Size-SMTI: Given a stable marriage instance with preferences containing ties, find a maximum-size stable matching.

In this working group, we aim to advance the research on parameterized approximation of further stable matching problems such as Min Egalitarian SRI which aims for finding a stable matching with minimum egalitarian cost and Incremental SR which aims for finding a stable matching that is closest to a given one.

We mainly focus on Min Egalitarian SRI parameterized by the maximum preference list length d. For instance, we obtain that the problem remains NP-hard to approximate even if $d = 3$. This excludes parameterized approximation algorithms with respect to d.

References

1 Jiehua Chen and Sanjukta Roy and Sofia Simola. *FPT-Approximability of Stable Matching Problems*. Technical Report. Available on-line at http://https://arxiv.org/abs/2508.10129 ACM Computing Research Repository (CoRR) (2025)

4.4 Complexity of proportionally diverse stable matching

Thekla Hamm (TU Eindhoven, NL), Péter Biró (HUN-REN KRTK – Budapest, HU), Henning Fernau (Universität Trier, DE), David Manlove (University of Glasgow, GB), and Danielius Sukys (University of Glasgow, GB)

In the classic hospital-residents problem with capacities a set of doctors should be assigned to a set of hospitals such that no hospital h should still have space for a doctor d that would prefer to be assigned to h than whichever hospital (if any) that d is assigned to or be able to make space for d by removing a doctor assigned to h which is less preferred by h than d. To model more detailed constraints on which "diverse" kinds of doctors a hospital needs to be assigned, e.g. a hospital needs between 5 and 8 heart surgeons or should hire at least a tenth of its doctors locally, the problem was more recently generalized to include upper and lower quotas over so called *types* which are possibly overlapping attributes of doctors. In many situations, it is reasonable to assume that these quotas are given as absolute values

but sometimes, e.g. when affirmative action is supposed to be imposed, it is more natural to formulate these quotas proportionally to the number of doctors that each hospital is actually assigned. The former generalization has already been studied in detail from a complexity theoretic and parameterized complexity theoretic angle but with proportional quotas we have no understanding of the problem's complexity.

In this working group we first established the general hardness of the problem, even in very restricted settings. Specifically, we can show hardness, even if there are only two completely disjoint types which are not hospital-specific and all hospitals only have only constant capacity and no upper quotas on the types. Future work will target the more specific questions such as the impact of the number of hospitals on the problems complexity and a more precise understanding of the subtle differences in the problems difficulty when considering proportional versus absolute quotas.

4.5 Dynamic algorithms for dynamic matching markets

Matthias Mnich (TU Hamburg, DE), Viktória Nemkin (Budapest University of Technology & Economics, HU), Marcin Pilipczuk (University of Warsaw, PL), and Manuel Sorge (TU Wien, AT)

Dynamic matching markets are commonly modeled as a set of agents with preferences over each other where these preferences change over time, new agents may arrive or present agents may leave. Such markets have received tremendous attention in terms of their economics. The algorithms and complexity of computing stable matchings in dynamic matching markets have also been studied extensively in the temporal setting (where all of the changes are known in advance) or the online setting (where none of the changes are known). However, the algorithmic dynamic setting has, to our knowledge, not been studied: Here the changes arrive one by one and in each step we need to efficiently maintain a stable matching. This working group set out to close that gap.

In our discussions, we were looking for a data structure that has some polynomial-time initialization time for a Stable Marriage instance with a fixed set of $2n$ agents, and supports efficiently a query operation that reports a stable matching for the current set of preferences, and an update operation that receives a change made to the current instance such as a shift of an alternative in a preference list.

In terms of lower bounds we found that, even when we allow only a swaps of two adjacent agents in preference lists, $O(n^{2-\epsilon})$ query time and $O(n^{1-\epsilon})$ update time for some $\epsilon > 0$ would contradict the so-called online matrix-vector multiplication conjecture and seems therefore unlikely.

In terms of upper bounds, it seems challenging to improve on the trivial $O(n^2)$ query-time upper-bound. We could achieve $O(n)$ query time and almost linear update time for the case where all the agents of one side have the same preference list. But in the more general setting where all agents' preference lists follow an order over all pairs of agents we could achieve $O(n)$ query time only with quadratic updates so far. Hence, next in particular we want to look into stronger lower bounds that show superlinear required update time for subquadratic query time.

4.6 Stable marriage with groups of similar agents

Baharak Rastegari (University of Southampton, GB), Tamás Fleiner (Budapest University of Technology & Economics, HU), and Shuichi Miyazaki (University of Hyogo – Kobe, JP)

Many important stable matching problems are known to be NP-hard, even when strong restrictions are placed on the input. In [1] we aimed to identify structural properties of instances of stable matching problems which will allow us to design efficient algorithms using elementary techniques. We focused on the setting in which all agents involved in some matching problem can be partitioned into k different types, where the type of an agent determines their preferences, and agents have preferences over types (which may be refined by more detailed preferences within a single type). This situation would arise in practice if agents form preferences solely based on some small collection of agents' attributes. We also considered a generalisation in which each agent may consider some small collection of other agents to be exceptional, and rank these in a way that is not consistent with their types; this could happen in practice if agents have prior contact with a small number of candidates. We showed that (for the case without exceptions), several well-studied NP-hard stable matching problems including Max SMTI (that of finding the maximum cardinality stable matching in an instance of stable marriage with ties and incomplete lists) belong to the parameterised complexity class FPT when parameterised by the number of different types of agents needed to describe the instance. For Max SMTI this tractability result can be extended to the setting in which each agent promotes at most one "exceptional" candidate to the top of their list (when preferences within types are not refined), but the problem remains NP-hard if preference lists can contain two or more exceptions and the exceptional candidates can be placed anywhere in the preference lists, even if the number of types is bounded by a constant.

In this working group we explored open problems involving exceptional candidates, such as the case when the exceptional candidate can be moved to the bottom of the preference list, and the case where the exceptional candidate can be moved to the top or bottom of their respective type. Other avenues, such as considering other relaxation on types, inspired by real life instances, were briefly considered but not explored given the limited time available.

References

1 Kitty Meeks and Baharak Rastegari *Solving hard stable matching problems involving groups of similar agents*, Theoretical Computer Science, Volume 844, Pages 171–194, Dec 2020.

4.7 Stable roommates parameterized by range

Will Rosenbaum (University of Liverpool, GB), Christine Cheng (University of Wisconsin – Milwaukee, US), Sushmita Gupta (The Institute of Mathematical Sciences – Chennai, IN), David Manlove (University of Glasgow, GB), and Dániel Marx (CISPA – Saarbrücken, DE)

Given a stable roommates instance, we define the range of the *range* of an agent a to be the difference between a's maximum and minimum rank in all other agents' preference lists. The range of the instance is the maximum range of any agent in the instances. We consider parameterized algorithms for stable matching problems parameterized by the range of the instance.

We showed that the following problems are FPT with respect to range:

- Counting stable matchings
- Sampling stable matchings uniformly
- Finding Optimal stable matchings
- Nearly stable matchings.

We have not yet determined if the following problems admit FPT algorithms:

- Finding egalitarian stable matchings
- Finding a matching that minimizes the number of blocking pairs.

5 Open problems

5.1 Stable matching in the semi-streaming model

Sushmita Gupta (The Institute of Mathematical Sciences – Chennai, IN)

We discussed the challenge of computing *stable matchings* in the (semi-)streaming model, motivated by massive bipartite graphs where storing the full input is infeasible. Classical Gale-Shapley style algorithms require quadratic space, while streaming models allow only near-linear space. Our focus was on defining and approximating stability when edges and preferences arrive as a stream. The core questions include: *What is the "right" relaxation of stability for this setting-minimizing the number of blocking pairs or blocking agents? Can we design semi-streaming algorithms that output perfect matchings with provably few blocking pairs? What trade-offs exist between space complexity and approximate stability guarantees?* We also asked whether kernelization or sketching techniques can yield efficient semi-streaming approximations, and whether known lower bounds for maximum matching extend to stability notions. These problems lie at the intersection of matching theory, streaming algorithms, and parameterized complexity, and aim to bridge the gap between large-scale preference data and theoretical guarantees.

Participants

- Péter Biró
HUN-REN KRTK – Budapest, HU
- Robert Bredereck
TU Clausthal, DE
- Jiehua Chen
TU Wien, AT
- Christine Cheng
University of Wisconsin – Milwaukee, US
- Gergely Csáji
Eötvös Lorand University – Budapest, HU
- Henning Fernau
Universität Trier, DE
- Tamás Fleiner
Budapest University of Technology & Economics, HU
- Sushmita Gupta
The Institute of Mathematical Sciences – Chennai, IN
- Thekla Hamm
TU Eindhoven, NL
- Naoyuki Kamiyama
Kyushu University – Fukuoka, JP
- Dušan Knop
Czech Technical University – Prague, CZ
- David Manlove
University of Glasgow, GB
- Dániel Marx
CISPA – Saarbrücken, DE
- Simon Mauras
INRIA Saclay – Île-de-France, FR
- Shuichi Miyazaki
University of Hyogo – Kobe, JP
- Matthias Mnich
TU Hamburg, DE
- Viktória Nemkin
Budapest University of Technology & Economics, HU
- Marcin Pilipczuk
University of Warsaw, PL
- Baharak Rastegari
University of Southampton, GB
- Will Rosenbaum
University of Liverpool, GB
- Ildikó Schlotter
ELTE KRTK – Budapest, HU
- Manuel Sorge
TU Wien, AT
- Danielius Sukys
University of Glasgow, GB

Report from Dagstuhl Seminar 25351

Computational Proteomics

Rebekah Gundry*[1], Magnus Palmblad*[2], and Mathias Wilhelm*[3]

1 **University of Nebraska – Omaha, US.** rebekah.gundry@unmc.edu
2 **Leiden University Medical Center, NL.** n.m.palmblad@lumc.nl
3 **TU München – Freising, DE.** mathias.wilhelm@tum.de

Abstract

In 2025 the Dagstuhl Seminar "Computational Proteomics" (25351), part of a series of Dagstuhl Seminars with the same name, brought together experts from proteomics, glycomics and machine learning to address key challenges in the field. Discussions emphasized the need for scalable and interoperable data infrastructures, a new initiative to generate large, AI-ready proteomics datasets, and community standards for reproducible and interpretable machine learning and harmonized glycomics workflows. Participants identified several barriers in clinical translation, multi-omics integration, and quantitative glyco-proteomics, highlighting limited data interoperability, heterogeneous experimental designs, and insufficient statistical and reporting frameworks. The seminar concluded with concrete action plans toward new standards, best practices, and collaborative initiatives to advance reproducible, sustainable and clinically relevant proteomics.

Seminar August 24–29, 2025 – https://www.dagstuhl.de/25351
2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains; Computing methodologies → Machine learning; Applied computing → Life and medical sciences
Keywords and phrases proteomics, glycomics, glycoproteomics, machine learning, mass spectrometry
Digital Object Identifier 10.4230/DagRep.15.8.46

1 Executive Summary

Rebekah Gundry (University of Nebraska – Omaha, US)
Magnus Palmblad (Leiden University Medical Center, NL)
Mathias Wilhelm (TU München – Freising, DE)

In 2025 the Dagstuhl Seminar "Computational Proteomics" (25351), part of a series of Dagstuhl Seminars with the same name, brought together researchers in proteomics, glycomics, computational biology, translational biomarker research, mass spectrometry, statistics, and machine learning for a week of intense discussions and collaboration. Building on the Dagstuhl Seminar "Computational Proteomics" (23301) in 2023, we extended the agenda in four directions that reflect where our field is now pushing hardest:

Translational Proteomics

The translational proteomics group defined translational proteomics as a continuum from discovery to clinical implementation, spanning basic model systems (cell lines, mouse), human biospecimens, clinical decision support, and ultimately population health. The group

* Editor / Organizer

emphasized that translation is not just "applying proteomics in the clinic," but instead structuring the entire value chain: standardized sample handling, acquisition, annotation, processing, interpretation, and delivery of actionable outputs (e.g. patient stratification, tumor board support). Major barriers identified include a lack of interoperable and well-annotated datasets, underpowered cohorts (especially in rare diseases), weak incentives for repetitive but clinically necessary assays, and difficulty converting molecular readouts into clinical recommendations. The group proposed ENIGMA, a staged, global-scale effort to generate and harmonize >100,000 proteomics datasets, starting in controlled mouse models and extending to human samples, as an AI-ready foundation for translation.

Machine Learning (in Proteomics and Glycomics)

The machine learning (ML) group concluded that the current culture of "incremental performance improvements" is unsustainable and often scientifically marginal. Instead, the group argued for community standards around software quality, reproducibility, interpretability, and dataset (ML/AI) readiness. Discussions focused on updating and extending existing recommendations (e.g. DOME, FAIR4RS) to address maintainability, testability and bias, and on defining what actually constitutes a publishable ML contribution in proteomics or glycomics. The group also highlighted the need for well-annotated, uncertainty-aware training and benchmarking datasets, including glycopeptide data, and began drafting two manuscripts – one on software quality and reporting expectations for ML in proteomics, and one on explainable and interpretable AI in MS-based proteomics.

Glycomics and Glycoproteomics

The "glyco" group focused on two tightly linked goals: improving confidence and comparability in glycan/glycopeptide identification and quantification, and lowering the barrier of entry for new researchers. First, the group outlined a plan for harmonizing glycan search spaces and reporting. A key recommendation is that outputs should carry standardized GlyTouCan identifiers and clearly encode the level of structural specificity (composition-only, topology, full linkage) so that results from different software tools can be compared on a common specificity level. The group also emphasized the need for explicit false-discovery rate (FDR) frameworks for glycan assignments, including topology- and isomer-sensitive scoring. Second, the working group substantially advanced two manuscripts: a best-practices/tutorial document for new glycosylation researchers (terminology, pitfalls, reporting standards), and a focused manuscript on how glycan structure affects glycopeptide signal intensity and the downstream challenges for quantification and biological interpretation. Writing responsibilities, timelines, and revision plans were agreed, and a first integrated draft was produced on-site.

Cross-cutting Topics: Federated Learning, Data Sharing and Credit, Multi-Omics Integration, and Quantitative Glyco-Proteomics

In the second half of the seminar week, people rotated between groups to discuss a number of cross-cutting topics and common challenges: Federated learning and controlled-access clinical data: While federated learning is still rarely used in proteomics due to widespread centralized data deposition, this is expected to change as clinical data are increasingly held locally for regulatory and privacy reasons. The group concluded that now is the time to define incentives, governance, and credit mechanisms for data generators so that high-value but unpublished datasets can still drive model development without leaving institutional boundaries. Multi-omics integration: Participants discussed how to integrate proteomics with transcriptomics,

phosphoproteomics, glycoproteomics, metabolomics/lipidomics, immunopeptidomics, and mass spectrometry imaging (and other spatially resolved data). The consensus was that most current "integration" is actually late-stage comparison of separate analyses. True multi-omics fusion is blocked by non-synchronous sampling, heterogeneous sample preparation, lack of common quality control (standards), differing biological timescales, and underdeveloped statistical control across layers. Guidance is needed on realistic experimental design and on levels of integration (early, mid-level latent, late/pathway). Quantitative glyco-proteomics statistics: The group outlined a plan to benchmark methods for glycoform quantification, normalization, missing value handling, and site occupancy estimation, leveraging tools such as MSstats/MSstats-PTM and experimentally perturbed datasets.

The 2025 Dagstuhl Seminar crystallized a shift in the field: from chasing marginal analytical improvements toward building scalable, interpretable, clinically relevant, and socially sustainable proteomics. The seminar concluded with concrete action items: manuscripts on ML standards and explainability; a best-practices/tutorial manuscript and a quantitative glycoform analysis manuscript from the glyco group; an ENIGMA proposal for large-scale, AI-ready translational proteomics data; a position statement on data sharing, incentive structures, and federated learning; and guidance on realistic multi-omics integration and QC. Together, these efforts define a forward-looking agenda for computational and translational proteomics in the coming years.

The discussions on trustworthiness and quality control in proteomics also fed directly into the planning of a Lorentz Center Workshop on "Trustworthiness in Proteomics", successfully co-organized by Mathias Wilhelm and Magnus Palmblad in Leiden, in February 2026.

2 Table of Contents

3 Working groups

3.1 Working Group Report: Translational Proteomics

Isabell Bludau (Unviversitätsklinikum Heidelberg, DE), Tine Claeys (Ghent University, BE), Stephanie Cologna (University of Illinois – Chicago, US), Melanie Föll (Universitätsklinikum Freiburg, DE), Wassim Gabriel (TU München – Freising, DE), Paula González Menéndez (University of Adelaide, AU), Zhiwei Liu (Westlake University – Hangzhou, CN), Tobias Schmidt (MSAID – Garching, DE), Nicola Ternette (University of Dundee, GB), Hans Wessels (Radboud University Nijmegen, NL), Mathias Wilhelm (TU München – Freising, DE), and Bernd Wollscheid (ETH Zürich, CH)

3.1.1 Scope and Motivation

The translational proteomics working group defined translational proteomics as a stepwise process linking basic proteomics research to clinical and population-level impact. This includes:

- Experiments in model systems (cell lines, mouse models).
- Studies on human biospecimens (tissues, fluids).
- Clinical application (diagnostics, patient stratification, therapy guidance).
- Integration into routine decision-making (e.g. tumor boards).
- Extension to population health and precision medicine at scale.

The group agreed that translational proteomics is not only about "measuring patient samples," but about structuring and standardizing the entire value chain: sample acquisition, metadata capture, MS acquisition, data processing, integration with other omics, interpretation, and delivery of clinically actionable output.

3.1.2 Key Barriers

Data quality and interoperability: Most proteomics datasets are not captured with standardized metadata, ontologies, or controlled vocabularies suitable for clinical use. Heterogeneous assay formats, missing metadata, and inconsistent annotation impede reuse and prevent robust comparisons across sites.

Underpowered cohorts: Many studies operate at N too small for reliable biomarker development, particularly in rare diseases. Cohort fragmentation across institutions and7 slow/complex material transfer agreements impede the assembly of sufficiently powered datasets.

Workflow robustness and reproducibility: Clinical deployment requires durable, validated, auditable workflows. Current research pipelines are often bespoke, sensitive to batch effects, and not easily transferable to regulated environments.

Motivation and incentives: Clinically relevant work often requires repetitive, "boring" targeted assays and rigorous QC. Such work is rarely rewarded academically and is often unfunded. As a result, essential confirmatory and longitudinal measurements are underperformed.

Interpretability and clinical usability: Even when high-quality measurements exist (e.g. phosphoproteomics, glycoproteomics, immunopeptidomics), delivering a clear, defensible recommendation to a tumor board remains difficult. Clinicians want interpretable, validated

evidence (what pathway is active? which drug class is likely to work?), not just feature lists. Regulatory and legal constraints: Clinical proteomics data raise privacy, compliance, and regulatory concerns analogous to those in genomics. Institutions are cautious about releasing raw data, which complicates model development and validation.

3.1.3 Computational / Infrastructural Priorities

Standardized ontologies and metadata schemas (including disease context, acquisition parameters, and sample processing information) are required to make datasets reusable across institutions. Scalable, transparent pipelines are needed to analyze imperfect, real-world study designs (non-ideal controls, heterogeneous sampling times, varying platforms). Multi-omics integration must move beyond naive overlap of "differential features" and instead support mechanistic interpretation and phenotype prediction. Cloud-based or federated frameworks may be required to enable collaborative analysis when data cannot leave clinical boundaries. Quantitative modeling should address proteoforms, phosphorylation, glycosylation, immunopeptidomics, etc., rather than collapsing information to a single "protein ID."

3.1.4 ENIGMA Initiative

To address the chronic shortage of large, harmonized, clinically relevant datasets, the group proposed ENIGMA: an international, multi-site initiative to create an AI-ready proteomics resource on the order of >100,000 samples.
The initiative is structured in phases:
Phase 1 (Pilot): A few thousand well-controlled samples (initially including defined mouse lines and select human material) are processed to establish baseline protocols, metadata standards, benchmarking pipelines, and QC criteria.
Phase 2 (Scale-up): Tens of thousands of samples are acquired across multiple mass spectrometry laboratories worldwide under harmonized DIA-style acquisition. Heterogeneity (instrument type, sample preparation differences) is explicitly captured rather than avoided, to allow development of normalization, batch correction, and cross-lab integration strategies.
Phase 3 (Translation): The trained models and pipelines are applied to human cohorts, including rare disease and clinically complex cases, with the goal of informing stratification, therapeutic targeting, and mechanism-of-action hypotheses.
Critical design features:

- Use of leftover tissues and archived material to reduce ethical/consent burden and accelerate scale.
- Systematic metadata capture (including tissue, condition, preparation, instrument, acquisition settings), with machine learning support for harmonization and outlier detection.
- Community governance around data ownership, licensing, authorship, and credit.
- Defined benchmark subsets for public release, similar in spirit to CPTAC and AlphaFold-style "challenge" datasets.

3.1.5 Outcomes and Next Steps

The working group concluded that translational proteomics requires a coordinated village: clinicians to articulate clinical questions; experimentalists to generate high-quality material; computational groups to build robust, interpretable pipelines; and funding/industry partners to sustain longitudinal measurement. The ENIGMA plan will serve both as a rallying point for funding proposals and as an anchor for future community standards in translational proteomics.

3.2 Working Group Report: Cross-cutting Topics

Frédérique Lisacek (Swiss Institute of Bioinformatics – Geneva, CH), Charlotte Adams (University of Antwerp, BE), Kiyoko Aoki-Kinoshita (Soka University – Tokyo, JP), Gad Armony (Bruker Nederland – Leiderdorp, NL), Wout Bittremieux (University of Antwerp, BE), Isabell Bludau (Unviversitätsklinikum Heidelberg, DE), Robert Chalkley (University of California – San Francisco, US), Tine Claeys (Ghent University, BE), Stephanie Cologna (University of Illinois – Chicago, US), Eric Deutsch (Institute for Systems Biology – Seattle, US), Patrick Emery (Matrix Science Ltd. – London, GB), Melanie Föll (Universitätsklinikum Freiburg, DE), Wassim Gabriel (TU München – Freising, DE), Paula González Menéndez (University of Adelaide, AU), Rebekah Gundry (University of Nebraska – Omaha, US), Devon Kohler (Northeastern University – Boston, US), Lev Levitskiy (University of Southern Denmark – Odense, DK), Klaus Lindpaintner (Bruker – Concord, US), Zhiwei Liu (Westlake University – Hangzhou, CN), Sriram Neelamegham (University at Buffalo – SUNY, US), Magnus Palmblad (Leiden University Medical Center, NL), Daniel Polasky (University of Michigan – Ann Arbor, US), Rene Ranzinger (University of Georgia, US), Tobias Schmidt (MSAID – Garching, DE), Nicola Ternette (University of Dundee, GB), Sergey Vakhrushev (University of Copenhagen, DK), Hans Wessels (Radboud University Nijmegen, NL), Mathias Wilhelm (TU München – Freising, DE), and Dirk Winkelhardt (Ruhr-Universität-Bochum, DE)

The Thursday cross-over sessions brought together participants from translational proteomics, ML, glycomics, clinical proteomics, and multi-omics integration to address challenges that cut across all domains.

3.2.1 Federated Learning, Privacy, and Incentives for Data Sharing

Federated learning (FL) is widely discussed in biomedical AI but is not yet common in proteomics practice. The group identified why: For most discovery proteomics, data are still deposited centrally (e.g. PRIDE, MassIVE). Central pooling is simpler than setting up FL. FL infrastructure is non-trivial: institutions must maintain local compute, synchronize software/hardware, address batch effects and fairness in distributed model updating, and handle versioning and auditing. Repositories themselves would bear significant cost to orchestrate FL at scale. However, participants agreed that this will change as proteomics becomes clinically embedded. Clinical and translational datasets (especially targeted assays, patient-derived longitudinal data, immunopeptidomics, and phospho-signaling panels) are increasingly held locally for regulatory and privacy reasons. In that emerging setting, FL – or at least controlled-access, privacy-aware model training – becomes essential.

The group emphasized that the blockers are not just technical but social:

- Privacy/compliance anxiety: When rules are unclear, many investigators choose not to share data at all.
- Fear of being scooped: High-value clinical datasets are expensive to generate, and groups are reluctant to expose them prior to publication without formal recognition.
- Fear of policing: Investigators worry about reputational risk if preliminary or messy data are scrutinized out of context.

Participants noted that vast quantities of clinically interesting proteomics data remain unpublished and effectively invisible. Releasing even partial access to these datasets (or enabling them to participate in FL-like analysis) would massively accelerate method development and translational validation.
Proposed actions:

- Draft a community position / opinion piece that (a) documents current barriers to sharing unpublished proteomics data, (b) argues for explicit dataset credit and citation, (c) proposes governance and recognition mechanisms for dataset contributors, and (d) frames FL as one of several approaches (not the only one) for responsibly leveraging sensitive data.
- Promote dataset-level DOIs, ORCID linkage, citation tracking, and usage metrics as first-class research outputs.
- Encourage journals and funders to recognize data notes / dataset publications and not penalize manuscripts that build on previously described datasets.

3.2.2 Multi-Omics Integration

The group critically assessed the state of "multi-omics integration" in proteomics-driven biology and translation. The conclusion was that true integration is still rare. Core challenges:

- Different omics layers (genomics, transcriptomics, proteomics, phosphoproteomics, glycoproteomics, metabolomics, lipidomics, immunopeptidomics, spatial MS imaging) are often collected on different material, at different time points, using incompatible extraction chemistries.
- Biological timescales differ: DNA is static, RNA is dynamic on hours, protein abundance and PTMs reflect turnover and signaling, metabolites respond on minutes. Naively correlating steady-state measurements across these layers can be misleading.
- Statistical control is underdeveloped. Integrating multiple high-dimensional datasets multiplies the number of hypotheses tested, but FDR control is often applied separately per layer, if at all.
- QC practices are inconsistent. Many studies still lack standardized spike-ins, technical controls, and batch assessment across all omics layers.

The group distinguished three levels of integration:

- Late integration: Interpret each omics layer independently, then compare or combine conclusions (e.g. pathway enrichment or overlapping differentially regulated features). This is the most common today.
- Mid-level / latent integration: Learn joint low-dimensional representations or network structures across layers (e.g. linking phosphoproteomics to transcriptomics through signaling pathway models).
- Early fusion: Combine raw or minimally processed quantitative data across layers into a single model. This is the most ambitious but also most fragile with respect to batch effects, missingness, and sampling asynchrony.

Use cases discussed:

- Proteogenomics to detect neoantigens, fusion proteins, and tumor-specific sequence variants for immunopeptidomics.
- Phosphoproteomics + total proteomics to distinguish signaling changes from protein abundance changes.
- Glycoproteomics to refine clinically relevant biomarkers (e.g. glycoform-specific PSA outperforms total PSA).
- Spatial MS (laser capture microdissection, MSI-guided microproteomics/metabolomics/glycomics) for tumor microenvironment profiling.

Actionable needs:

- Clear guidance on experimental design for clinically realistic studies, acknowledging that perfect co-sampling is often impossible in the hospital setting.
- Shared QC expectations across layers (technical controls, spike-ins).
- Agreement on how to claim "integration": studies should report whether they did late, mid-level, or early fusion, rather than labeling any multi-assay project as "multi-omics integration."
- Availability of paired, public benchmark datasets suitable for method development.

3.2.3 Quantitative Glyco-Proteomics Statistics

The cross-over discussions also addressed quantification and statistical analysis for glycomics and glycoproteomics, linking glyco specialists, statisticians, and ML practitioners.
Key points:

- Glycopeptide intensities are not directly comparable across glycoforms, because different glycans alter ionization and fragmentation efficiency.
- Missing values should not always be blindly imputed. If an entire glycoform is absent in one condition, treating that as "low abundance" can generate false positives.
- Normalization strategies must reflect biology. Global scaling can erase true global glycosylation shifts; site-focused normalization may be more appropriate.
- Site occupancy and glycoform distribution are biologically meaningful readouts but require modeling analogous to PTM-centric statistical frameworks.

Planned work:

- Evaluate MSstats/MSstats-PTM–style approaches where glycoforms are treated as modified analytes, using controlled datasets (e.g. exoglycosidase-treated samples, known congenital glycosylation defects) to benchmark differential analysis, normalization, and missing value handling.
- Define minimal reporting requirements for glyco-quantitative studies, to make them interpretable and reusable in translational pipelines.

3.3 Working Group Report: Glycomics and Glycoproteomics

Rene Ranzinger (University of Georgia, US), Kiyoko Aoki-Kinoshita (Soka University – Tokyo, JP), Gad Armony (Bruker Nederland – Leiderdorp, NL), Robert Chalkley (University of California – San Francisco, US), Wassim Gabriel (TU München – Freising, DE), Rebekah Gundry (University of Nebraska – Omaha, US), Klaus Lindpaintner (Bruker – Concord, US), Frédérique Lisacek (Swiss Institute of Bioinformatics – Geneva, CH), Sriram Neelamegham (University at Buffalo – SUNY, US), Daniel Polasky (University of Michigan – Ann Arbor, US), Sergey Vakhrushev (University of Copenhagen, DK), and Hans Wessels (Radboud University Nijmegen, NL)

3.3.1 Motivation

Glycomics and glycoproteomics remain computationally and analytically challenging due to structural complexity (branching, linkage, isomerism), diverse acquisition strategies, non-uniform software support, and historically limited standardization. The glyco working group

concentrated on two urgent needs: (i) harmonizing identification and reporting so that results are comparable across software and labs, and (ii) establishing reliable quantitative and statistical practices, especially for glycopeptide-level measurements.

3.3.2 Harmonizing Glycan/Glycopeptide Identification

A core outcome was agreement that glycan and glycopeptide search results must become machine-readable, comparable, and traceable across tools. The group recommends:

- Use of GlyTouCan IDs in all reported results, at minimum at the composition level, with increasing specificity (topology, linkage) where supported by evidence.
- Explicit declaration of structural specificity: Results should indicate whether the assignment is composition-only, topology (branching pattern without full linkage certainty), or fully linkage-resolved.
- Subsumption / hierarchical comparison: Since different tools resolve glycans to different specificity levels, there must be a way to "collapse" structures to a shared lower-specificity representation for comparison and benchmarking.
- Shared glycan reference sets: The group began assembling a curated "human adult reference glycan list," analogous to a reviewed FASTA for proteins. Each participating lab/software group will contribute the glycan lists they currently search (e.g. as used in Byonic, MSFragger-glyco, pGlyco, Protein Prospector, GlycanFinder, etc.). These will be merged, deduplicated, and annotated with GlyTouCan ID, species/context, linkage class (N- vs O-linked), biological source (plasma, tissue, cell line), and whether the glycan is free or peptide-conjugated.
- The aim is to generate a community-supported glycan search space that is biologically grounded, not "everything in GlyTouCan," and thus suitable for routine glycoproteomics searches and cross-tool benchmarking.

3.3.3 Confidence Scoring and FDR for Glycans/Glycopeptides

The group addressed statistical confidence, which currently lags behind peptide-centric proteomics:

- Glycan/glycopeptide identification often lacks robust FDR control, especially at the level of glycan topology or isomer resolution.
- Scoring schemes must consider diagnostic fragment ions, instrument- and method-specific fragmentation behavior (e.g. stepped HCD vs ETD/EThcD), and the ability to distinguish core vs antenna fucosylation, high-mannose vs complex glycans, etc.
- Structural-level FDR (e.g. topology-level "localization" of glycan features on the peptide) was viewed as analogous to PTM localization scoring in phosphoproteomics.

Participants presented ongoing work on topology-aware scoring frameworks that progressively reduce candidate lists using diagnostic ions and instrument-aware fragmentation rules, then apply multi-level FDR (PSM-level, composition-level, topology-level). The consensus was pragmatic: "some FDR is better than none," provided the confidence level is clearly stated. Work remains to prevent overconfidence in cases where only one plausible glycan remains in the database but the MS/MS evidence is insufficient to distinguish isomers.

3.3.4 Quantification, Normalization, and Biological Interpretation

A second major focus was quantification. Glycopeptide signals are strongly influenced by the attached glycan, so intensity is not a straightforward proxy for occupancy or abundance.

The group discussed:

- Handling missing values: imputing partially missing precursors may be acceptable, but imputing entire missing glycoforms is risky and can create false biological conclusions.
- Normalization: global median-centering or total-signal normalization can be misleading if overall glycosylation shifts biologically. Alternative strategies include normalizing within a site to the dominant glycoform, or modeling site occupancy relative to the unmodified protein level.
- Statistical modeling: Treating glycoforms as site-specific PTMs suggests that tools such as MSstats/MSstats-PTM could be adapted by treating each glycoform as the "modified species."

The group proposed benchmarking known perturbation datasets (e.g. exoglycosidase-treated samples, congenital disorders of glycosylation, cell-surface enrichment studies) to evaluate differential analysis, missing value handling, and normalization strategies. This work connects directly to translational aims (e.g. leveraging glycosylation patterns in diagnostics and tumor board discussions) and to ML aims (e.g. training glyco-aware spectral/RT predictors, or "glyco-Prosit" models).

3.3.5 Best Practices and Onboarding for New Researchers

The working group advanced two manuscript efforts that originated in a previous Dagstuhl meeting:

- Best-practices/tutorial manuscript
- Introduces newcomers to protein glycosylation analysis, including released glycans and intact glycopeptides.
- Defines key terminology (composition, topology, linkage-defined structure).
- Recommends that reported results always include GlyTouCan IDs and explicitly state confidence and specificity level.
- Summarizes common pitfalls in identification, quantification, and interpretation.
- Aligns with MIRAGE-style reporting expectations.

During the seminar, section leads were assigned, timelines were agreed upon, prior contributors were re-engaged, and a coordinated writing sprint produced 21 pages of draft text.

Glycoform quantification manuscript

- Examines how glycan structure affects glycopeptide fragmentation and intensity.
- Explains why naïve fold-change analysis across glycoforms can mislead biological interpretation.
- Outlines statistical approaches for site-specific glycoform analysis and occupancy estimation.

3.3.6 Summary

The glycomics/glycoproteomics group delivered concrete progress toward harmonized reporting, FDR-aware scoring, and quantitative interpretation, and established a clear writing and dissemination plan. The group also linked their agenda to ML (glyco-aware prediction models, benchmark datasets) and translational proteomics (inclusion of glycan biology in clinically oriented pipelines).

3.4 Working Group Report: Machine Learning in Proteomics

Tobias Schmidt (MSAID – Garching, DE), Charlotte Adams (University of Antwerp, BE), Wout Bittremieux (University of Antwerp, BE), Tine Claeys (Ghent University, BE), Eric Deutsch (Institute for Systems Biology – Seattle, US), Patrick Emery (Matrix Science Ltd. – London, GB), Wassim Gabriel (TU München – Freising, DE), Devon Kohler (Northeastern University – Boston, US), Lev Levitskiy (University of Southern Denmark – Odense, DK), Zhiwei Liu (Westlake University – Hangzhou, CN), Magnus Palmblad (Leiden University Medical Center, NL), and Dirk Winkelhardt (Ruhr-Universität-Bochum, DE)

3.4.1 Motivation and Problem Statement

The ML working group took a critical view: the field currently rewards incremental improvements (e.g. 2% gain in spectral prediction accuracy) without necessarily delivering interpretability, robustness, reproducibility, or biological/clinical insight. The group argued for a rebalancing of values toward sustainability and impact.

Four core themes structured the discussions:

(i) Software Quality, Standards, and Publication Expectations

Participants assessed the adequacy of existing frameworks such as the DOME recommendations for reporting ML in proteomics/metabolomics and FAIR4RS for FAIR research software. They concluded that these frameworks are necessary but incomplete. Key gaps:

- Maintainability and testability: ML code should be modular, documented, versioned, licensed, and accompanied by retraining instructions and tests. These expectations are rarely enforced.
- Adherence to community standards: ML tools must input and output established HUPO-PSI formats (mzML, mzIdentML, mzTab, SDRF, ProForma, etc.) where applicable, to ensure interoperability.
- Bias and applicability domain: Authors should explicitly analyze dataset composition (organism, tissue, modification class, charge states, instrument type), identify biases (e.g. human and HLA bias in immunopeptidomics, charge state bias, peptide length bias), and articulate where the model should and should not be used.
- Novelty and publication worthiness: DOME specifies how to describe an ML method, but not whether the method represents a meaningful advance. The group argued that journals and reviewers should demand either conceptual novelty, new interpretability, improved robustness/maintainability, or concrete biological/clinical utility – not just a small numeric gain.

Outcome: The group began drafting a manuscript that extends previous Dagstuhl-driven work (including "Interpretation of the DOME Recommendations for Machine Learning in Proteomics and Metabolomics" and recent guidance on FAIR/open-source research software in proteomics) by incorporating the EVERSE research software quality dimensions (modularity, reusability, analysability, modifiability, testability, sustainability). The goal is a reviewer/author checklist that goes beyond reporting and into quality, longevity, and responsible reuse.

(ii) Training Data, Metadata, Uncertainty, and Benchmarking

Robust ML depends on training data that is complete, well-annotated, and accompanied by uncertainty estimates. The group emphasized that current proteomics and glycoproteomics datasets often lack even basic metadata (sample type, organism, sex, preparation method, instrument settings), which blocks fair benchmarking and leads to hidden biases. Priorities:

- Centralized metadata capture: Multiple ongoing efforts (curated SDRF files, reprocessing pipelines, LLM-based annotation of manuscripts and raw mzML headers, large-scale reanalyses such as MassIVE, PRIDE, PeptideAtlas, and MassNet) should not remain siloed. A shared portal under an existing infrastructure (e.g. ProteomeXchange) could collect experimental design metadata, derived identifications/quantifications, QC metrics, and completeness scores.
- Gold standard datasets: For each ML task (fragmentation prediction, retention time prediction, identification, quantification, modification localization, glycopeptide assignment, etc.), the community should define benchmark datasets that include raw data, identifications, quantitative outputs, and experimental design metadata in standard formats.
- These benchmarks must reflect biological and technical diversity (different tissues, instruments, acquisition modes) rather than a single "easy" mixture.
- Uncertainty propagation: Both PSM-level confidence and metadata confidence should be carried through into training. Models should not be trained exclusively on idealized "perfect" identifications; probabilistic or Bayesian treatment of noisy/ambiguous cases is encouraged.
- Synthetic data: Simulated/synthetic data can support benchmarking, experimental design, rare-event modeling, and privacy-preserving analysis. But the group raised integrity and governance concerns: synthetic vendor-format raw files must be clearly watermarked and traceable, to prevent fraud and to allow repositories and journals to distinguish simulated from measured data. Synthetic data are not a substitute for high-quality real data in final model evaluation.

(iii) Interpretability and Explainable AI (XAI)
The group emphasized that interpretable ML is essential if models are to influence biological discovery, clinical triage, or mechanistic reasoning. Two levels were distinguished:

- Explainability: Tools such as SHAP values, saliency maps, causal/graphical models, and feature attribution methods can reveal which features drive predictions in a given model or for a given sample.
- Interpretability: The biological and biochemical meaning of those features must then be contextualized. For example, a classifier that distinguishes tissue-of-origin should identify pathways and protein/glycoform signatures plausibly linked to that tissue, rather than artifacts such as systematic missingness or batch effects.

Planned output: The group outlined a manuscript arguing for explainable and interpretable AI in MS-based proteomics and glycoproteomics. This manuscript will catalog use cases (e.g. tissue-of-origin prediction, phospho-signaling interpretation, glycoform-driven biomarker hypotheses, failure analysis in de novo sequencing), pitfalls (spurious correlations, hidden batch effects), and recommendations for reporting.
(iv) Education and Community Infrastructure
Interest in ML for proteomics and glycomics is accelerating, but training materials remain fragmented. Rather than generating entirely new courses, the group agreed to curate and maintain a community-driven "Awesome Proteomics/ML" style resource, linked to ProteomicsML and ELIXIR TeSS. This resource will collect:

- Introductory material for ML researchers entering proteomics (instrument basics, data structure, pitfalls).
- Practical material for experimentalists entering ML (basic statistics/ML literacy, bias analysis, model evaluation).

- Glyco-specific onboarding resources, which are currently underrepresented compared to proteomics.

The group discussed adding vetted content to TeSS and ProteomicsML, tagging material by audience and task, and exploring lightweight retrieval-augmented assistants based on this curated corpus.

Summary

The ML working group shifted attention from "How do we get slightly better predictions?" to "How do we build models, datasets, software, and training practices that the community can trust, reuse, and interpret – and that actually matter biologically and clinically?" Deliverables in progress include:

- A standards/checklist manuscript integrating DOME, FAIR4RS, and EVERSE for ML in proteomics and glycomics.
- A perspective on explainable and interpretable AI in MS-based proteomics.
- A shared educational and onboarding resource, to be maintained in community infrastructure.

4 Open problems

4.1 Outlook

Rebekah Gundry (University of Nebraska – Omaha, US), Magnus Palmblad (Leiden University Medical Center, NL), and Mathias Wilhelm (TU München – Freising, DE)

The 2025 Dagstuhl Seminar crystallized a shift in the field: from chasing marginal analytical improvements toward building scalable, interpretable, clinically relevant, and socially sustainable proteomics.

Across all working groups, several unifying priorities emerged:

- Standards and reproducibility before novelty.
- Participants called for enforceable expectations around metadata, software quality, model explainability, and statistical rigor – in translational workflows, ML models, glycan/glycopeptide assignment, and multi-omics integration.
- AI-ready, creditable datasets at scale.
- Whether through ENIGMA (to generate >100,000 harmonized datasets across organisms and clinical samples) or through curated benchmark datasets with explicit uncertainty and metadata, the community is moving toward data infrastructure as a collective asset. This shift demands new incentive structures for data contributors, including dataset DOIs, citation tracking, dataset papers, and recognition in funding and hiring.
- Interpretability and clinical relevance.
- The community emphasized not only accurate predictions, but predictions that can be explained, trusted, and acted upon in biological and clinical contexts – from pathway-level interpretation to tumor board recommendations.
- Inclusion of glycosylation, PTMs, and spatial context.

- Participants repeatedly stressed that clinically useful proteomics must incorporate proteoforms, phosphorylation, glycosylation, immunopeptidomics, and spatial heterogeneity, rather than collapsing biology to "protein ID + abundance."
- Bridging to clinical reality.

Translational proteomics was framed as a pipeline problem, not a single breakthrough: standardized acquisition, continuous data ingestion, interpretable computation, and clinically grounded reporting. Federated and privacy-aware learning will likely become central as proteomics enters regulated clinical environments. The seminar concluded with concrete action items: manuscripts on ML standards and explainability; a best-practices/tutorial manuscript and a quantitative glycoform analysis manuscript from the glyco group; an ENIGMA proposal for large-scale, AI-ready translational proteomics data; a position statement on data sharing, incentive structures, and federated learning; and guidance on realistic multi-omics integration and QC. Together, these efforts define a forward-looking agenda for computational and translational proteomics in the coming years.

Last but not least, the participants discussed potential future topics for a next Dagstuhl meeting on computational proteomics. Such a meeting could be structured around a coherent progression from data generation to biological insight. It could begin with the current state of proteomics technologies, focusing on challenges and open gaps in state-of-the-art mass spectrometry, including ion mobility MS, alongside the role of non-MS approaches such as affinity proteomics and single-molecule protein sequencing. From there, the meeting could address fundamental questions of protein identification and characterization, including the status and future of de novo sequencing methods and the ongoing debate around the existence and relevance of the dark proteome. Building on these measurements, structural proteomics – encompassing cross-linking, HDX, FPOP, PTM analysis (including glycans), AlphaFold-based models, and integrated structural modeling – connects molecular structure to function and modification. This in turn motivates deeper consideration of data integration challenges, including N- and P-integration and links between proteomics, cryo-EM, and other structural modalities. At a higher level of abstraction, the integration of multi-omics data with machine learning, foundational models, and the concept of virtual or digital cells offers a unifying computational framework. Finally, the meeting could broaden its scope to challenging and underrepresented application areas such as meta-omics, plant proteomics, paleoproteomics, non-model organisms, and non-human proteomics relevant to food and nutrition, with data integration serving as a cross-cutting theme throughout all topics.

Participants

- Charlotte Adams
University of Antwerp, BE
- Kiyoko Aoki-Kinoshita
Soka University – Tokyo, JP
- Gad Armony
Bruker Nederland – Leiderdorp, NL
- Wout Bittremieux
University of Antwerp, BE
- Isabell Bludau
Unviversitätsklinikum Heidelberg, DE
- Robert Chalkley
University of California – San Francisco, US
- Tine Claeys
Ghent University, BE
- Stephanie Cologna
University of Illinois – Chicago, US
- Eric Deutsch
Institute for Systems Biology – Seattle, US
- Patrick Emery
Matrix Science Ltd. – London, GB
- Melanie Föll
Universitätsklinikum Freiburg, DE
- Wassim Gabriel
TU München – Freising, DE
- Paula González Menéndez
University of Adelaide, AU
- Rebekah Gundry
University of Nebraska – Omaha, US
- Devon Kohler
Northeastern University – Boston, US
- Lev Levitskiy
University of Southern Denmark – Odense, DK
- Klaus Lindpaintner
Bruker – Concord, US
- Frédérique Lisacek
Swiss Institute of Bioinformatics – Geneva, CH
- Zhiwei Liu
Westlake University – Hangzhou, CN
- Sriram Neelamegham
University at Buffalo – SUNY, US
- Magnus Palmblad
Leiden University Medical Center, NL
- Daniel Polasky
University of Michigan – Ann Arbor, US
- Rene Ranzinger
University of Georgia – Athens, US
- Tobias Schmidt
MSAID – Garching, DE
- Nicola Ternette
University of Dundee, GB
- Sergey Vakhrushev
University of Copenhagen, DK
- Hans Wessels
Radboud University Nijmegen, NL
- Mathias Wilhelm
TU München – Freising, DE
- Dirk Winkelhardt
Ruhr-Universität-Bochum, DE
- Bernd Wollscheid
ETH Zürich, CH

Report from Dagstuhl Seminar 25361

Natural Language Processing for Mental Health

Dana Atzil-Slonim*[1], Iryna Gurevych*[2], Dirk Hovy*[3], and Diyi Yang*[4]

1 **Bar-Ilan University – Ramat-Gan, IL.** dana.slonim@gmail.com
2 **Department of Computer Science, TU Darmstadt, DE.** iryna.gurevych@tu-darmstadt.de
3 **Bocconi University – Milan, IT.** dirk.hovy@unibocconi.it
4 **Stanford University, US.** diyi@stanford.edu

Abstract

NLP has made remarkable progress in recent years, driven by breakthroughs in large language models (LLMs) and the availability of large-scale datasets such as social media posts, online forums, and patient records. These advances have made NLP models highly capable of extracting valuable insights from text data related to mental health. This development raises two natural questions: (1) How well, if at all, can NLP enable early detection, diagnosis, and intervention – not only for patients or support seekers but also for therapists or support providers? (2) Can NLP-driven solutions help bridge the gap between the escalating demand for mental health resources and the limited availability of mental health professionals, providing scalable and immediate support through chatbots, virtual therapists, and data-driven interventions? Both questions address the technical feasibility and the ethical concerns about using a developing technology in a sensitive application. This Dagstuhl Seminar brought together researchers across NLP, clinical science, human–computer interaction, and digital mental health to reflect on how NLP and AI can support mental health outcomes. Over the course of the week, we looked at key areas where NLP has the potential to transform mental health: understanding how mental states change and how therapeutic change occurs; exploring how NLP can help therapist training and feedback; identifying technological gaps and multilingual challenges in building reliable mental health models; and addressing pressing concerns around evaluation, validation, privacy, and ethics. Through vision talks, lightning sessions, and breakout groups, participants explored both the opportunities and limitations of deploying NLP for mental health, laying the groundwork for responsible, interdisciplinary research in this vital direction.

Seminar August 31 – September 5, 2025 – https://www.dagstuhl.de/25361

2012 ACM Subject Classification Computing methodologies → Natural language processing; Applied computing → Health care information systems; Human-centered computing → Empirical studies in HCI

Keywords and phrases Mental Health, NLP, Human-Centered AI, Large Language Models

Digital Object Identifier 10.4230/DagRep.15.8.62

* Editor / Organizer

Natural Language Processing for Mental Health, *Dagstuhl Reports*, Vol. 15, Issue 8, pp. 62–79
Editors: Dana Atzil-Slonim, Iryna Gurevych, Dirk Hovy, and Diyi Yang

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Dana Atzil-Slonim (Bar-Ilan University – Ramat-Gan, IL, dana.slonim@gmail.com)
Iryna Gurevych (TU Darmstadt, DE, iryna.gurevych@tu-darmstadt.de)
Dirk Hovy (Bocconi University – Milan, IT, dirk.hovy@unibocconi.it)
Diyi Yang (Stanford University, US, diyi@stanford.edu)

This document summarizes the outcomes of our Dagstuhl Seminar on "Natural Language Processing for Mental Health" (25361). The seminar was motivated by an urgent and growing global need: mental health issues are affecting millions of people worldwide, yet the majority of individuals in need of care do not receive any treatment, especially those from marginalized, low-income, or rural populations. While many mental health conditions are treatable, or even preventable with early detection and intervention, people often don't receive support until their concerns have escalated. In parallel, NLP has made remarkable progress in recent years, largely due to advances in large language models (LLMs) and the availability of large-scale textual data from sources like social media, online forums, and clinical records. These technologies therefore offer a variety of opportunities in addressing the problems outlined above, both for patients and in therapist training. Indeed, NLP models are already being deployed in mental health applications. However, both the existing and envisioned use cases raise critical questions around feasibility, scalability, and responsibility.

To address these, the seminar brought together an interdisciplinary group of researchers from NLP, clinical psychology, HCI, and digital health to assess where, how, and under what conditions NLP can be responsibly used to support mental health needs. Concretely, the seminar focused on:

- Understanding the potential of NLP, and in particular LLMs, to support mental health diagnosis, intervention, and therapeutic processes;
- Identifying critical gaps in technical feasibility, evaluation, and deployment of these tools in real-world, high-stakes clinical settings;
- Exploring responsible and interdisciplinary solutions that bridge NLP, psychology, human–computer interaction, and ethics.

As a major result from the seminar, we identified the following problems and future directions:

1. The need for systematic evaluation frameworks for NLP models used in psychotherapy and mental health support, including benchmarks, as well as longitudinal, multilingual, and real-world evaluation protocols.
2. The challenge of simulating clients or patients, and therapists using LLMs to support therapist training, research, and practice, particularly concerning modeling therapeutic authenticity, diversity, and change processes.
3. The persistent knowledge and communication gap between NLP and clinical psychology communities, and the urgent need to bridge this divide through interdisciplinary collaboration to ensure clinical relevance and real-world utility.
4. The lack of visibility and impact pathways for research in NLP and mental health across both technical and applied domains, and how to promote this work across venues, funding agencies, and policy-making spaces.

The seminar's structure was designed to support both critical discussion and creative collaboration. Through a program of vision talks, lightning presentations, breakout groups, and informal exchanges, the seminar was organized around four thematic clusters: (1)

understanding how mental states change and how therapeutic change occurs; (2) how NLP can support therapist training and real-time feedback; (3) identifying technological, privacy, and multilingual challenges; and (4) addressing evaluation, validation, and ethical concerns.

This seminar has laid a solid foundation for a crucial research area. **Three perspective papers** are in development: one focused on simulating patients using LLMs, one on evaluation challenges and opportunities in psychotherapy applications, and one on strengthening interdisciplinary collaboration between NLP and mental health communities. Additionally, **a blog post** is being prepared to reflect on how to promote the broader impact of this work, and **a workshop submission** motivated by the seminar is currently under review for SIGCHI 2026.

These outcomes align with our initial goals: (1) to produce joint research publications and collaboration opportunities, such as position papers that map out the challenges and opportunities in building responsible and robust NLP systems for mental health; and (2) to form a cross-field community that continues to connect NLP and mental health researchers – both in technical venues and clinical practice contexts.

2 Table of Contents

3 Overview of Talks

3.1 Integrating Innovations in Clinical Science and Artificial Intelligence to Study the Dynamics of Therapeutic Change

Dana Atzil-Slonim (Bar-Ilan University – Ramat-Gan, IL)

In psychological therapies, understanding what works for whom and when remains one of the most enduring challenges in mental health research and practice. Today, however, we are better equipped than ever to address this question – thanks to transformative advances in both clinical science and AI. In this talk, I demonstrate how our interdisciplinary team of clinicians and AI researchers is bridging top-down, theory-driven approaches with bottom-up, data-driven AI methods to uncover the dynamics that drive therapeutic change.

I begin by outlining key challenges in conceptualizing, treating, and researching mental health. I then discuss significant theoretical shifts in clinical science – such as the transition from general treatment models to transtheoretical processes and from one-person to two-person psychology – that have paved the way for addressing these challenges. Building on these foundations, I illustrate how our team has leveraged theoretical innovations and advancements in multimodal analysis and AI to explore the dynamic processes within clients (intrapersonal dynamics) and between clients and therapists (interpersonal dynamics) that are linked to improved treatment outcomes.

Central to this effort is the development of temporally aware, multimodal AI methods designed to address key limitations in modeling temporality, complex reasoning, situational awareness, personalization, and the integration of verbal and non-verbal data. I conclude by discussing how this integrative approach can enhance diagnostic precision, support personalized interventions, and improve the overall effectiveness of mental health treatments.

3.2 A (Humane) Vision for Digital Mental Health in a Post-AI World

Munmun De Choudhury (Georgia Institute of Technology – Atlanta, US)

Digital mental health has undergone three paradigm shifts: from the clinical gaze, to the quantified self, to predictive models built on social media and pervasive data. Today, we are on the cusp of a fourth shift – one defined by the rapid rise of large language models (LLMs) and generative AI. These technologies hold promise to expand access to support, reframe cognition, and facilitate empathic conversations. Yet, they also risk cultural misalignment, shallow validation, and the erasure of lived experience. Drawing on empirical studies spanning algorithmic prediction, online support, therapeutic alliance with AI, and cross-lingual evaluations of LLMs, this talk surfaces the tensions between correctness and care, agency and automation, identity and institutional power. I argue that the prevailing focus on efficiency and scale often neglects the ecological realities of people's lives and the invisible labor they contribute when their data fuel AI systems. A humane vision for digital mental health in a post-AI world requires moving from inference to interventions that center authenticity, inclusivity, and responsibility. I conclude by outlining design principles for an AI that connects without co-opting, empowers rather than replaces, and offers truth with care – not just comfort.

3.3 The Construct Mining Pipeline – a Computational Method to Reveal Psychological Constructs from Text Data

Jana Lasser (University of Graz, AT)

When starting to formalize psychological constructs, researchers traditionally rely on two distinct approaches: the quantitative approach, which defines constructs as part of a testable theory based on prior research and domain knowledge, often using self-report questionnaires, or the qualitative approach, which gathers data mainly in the form of text and bases construct definitions on exploratory analyses. We present a new computational method that combines the comprehensiveness of qualitative research and the scalability of quantitative analyses to define psychological constructs from semi-structured text data. Using structured questions, participants are prompted to generate sentences that reflect instances of the construct of interest. We apply computational methods to calculate embeddings as numerical representations of the sentences, which we then run through a clustering algorithm to arrive at groupings of sentences as psychologically relevant classes. The method includes steps for measuring and correcting bias introduced by data generation and for assessing cluster validity through human judgment. We demonstrate the applicability of our method on an example from emotion regulation. Based on short descriptions of emotion regulation attempts collected through an open-ended situational judgment test, we use our method to derive classes of emotion regulation strategies. Our approach shows how machine learning and psychology can be combined to provide new perspectives on the conceptualization of psychological processes.

3.4 Can Language-based Assessments Outperform the Instruments on which They are Trained?

H. Andrew Schwartz (Stony Brook University, US)

Joint work of H. Andrew Schwartz, Brenda Curtis, Salvatore Giorgi, Lyle Ungar, Huy Vu, David Yaden, Tingting Liu, Kenna Yadeta, Gregory Park, Johannes Eichstaedt, Evelyn Bromet, Benjamin Luft, Roman Kotov, Sean Clouston, Youngseo Son, Martin Seligman, Varadarajan Varadarajan

This two-part talk begins with a vision for HLAB focusing on (a) a more accurate representation of people, (b) LLMs for safe mental health therapy, and (c) eudaimonix content recommendation from AI. It continues with a series of experiments addressing the assumption that models trained on standard instruments are bounded by their predictive validity. We demonstrate cases where language-based assessments (LBAs) predict psychological outcomes better than the instruments they were trained on. Evidence includes: a historical review, a simulated experiment, an experiment with pseudo-observed data, and another with fully observed data predicting external criteria. These findings provide theoretical and empirical evidence that language-based assessments can more closely approximate true psychological states. Mechanisms by which these assessments outperform traditional tools are explored, highlighting the potential for AI-based language analysis to reshape psychological measurement.

3.5 Progress and challenges in NLP for mental health: Personalised longitudinal monitoring and beyond

Maria Liakata (Queen Mary University of London, GB)

The first part of the talk provides an overview of the NLP landscape for mental health, discussing the range of applications and the challenges faced by systems based on large language models (LLMs), with examples from the literature, especially regarding how these factors impact their suitability and applicability to mental health. There are many unresolved challenges, among others, regarding appropriate generation, temporal robustness, temporal and other forms of reasoning, and privacy concerns, especially when working with sensitive content such as mental health data. The programme of work I have been leading in the past five years consists of three core research directions: (1) data representation and generation, (2) methods for personalised longitudinal models and temporal understanding, (3) evaluation in real-world settings, with use cases in mental health. I will give an overview of the work in my group on these topics and conclude with a presentation of the evaluation platform for LLM-based systems that we have been developing within the AdSoLve project.

3.6 Using AI in Measurement-Based and Data-Informed Psychological Therapy

Wolfgang Lutz (Trier University, DE)

This presentation discusses a dynamic network model and research program supporting the delivery of personalized feedback to therapists at the start and throughout psychological therapy. The approach focuses on identifying the core individual elements of psychological distress and resources, while also enabling the quantification of multiple dimensions of distress (e.g., cognitive-behavioral, emotional, motivational, and interpersonal stressors) through the application of large language models. It addresses questions such as: How can psychological distress and resources be extracted, both qualitatively and quantitatively, from session transcripts, and how can they be integrated into clinical support tools and adaptive treatment planning for therapists? How can such models support the identification of patients at risk of treatment failure? And how can these models support clinical training? What are the technical and practical challenges that need to be addressed?

3.7 Using Large Language Models to Create Personalized Networks From Therapy Sessions

Hiba Arnaout (TU Darmstadt, DE)

Personalizing psychotherapy often relies on individual-level networks, but estimating these networks typically requires intensive longitudinal data, limiting scalability. Large Language Models (LLMs) offer a potential alternative by analyzing therapy transcripts directly. We introduce a pipeline that automatically generates client networks to support case conceptualization and treatment planning.

3.8 Monitoring Patient Emotions at Scale to Assess Psychotherapy Outcomes Using Language Models

Matteo Malgaroli (NYU Grossman School of Medicine – New York, US)

Traditional methods of psychiatric evaluation face ongoing challenges regarding reliability, objectivity, and scalability. The integration of language models offers a potential solution for assessing psychiatric symptoms and informing theory through digital biomarkers. In this talk, I will discuss findings on using linguistic markers to capture mental health symptoms from clinical conversations. In particular, I will introduce VISTA, a scalable method for capturing temporal flows, and apply it to emotions expressed by a sample of over 10,000 patients receiving digital mental-health treatment. I show how the resulting clusters relate to clinical outcomes. These findings highlight the opportunity to monitor patient outcomes using only linguistically captured emotions, especially when direct measurement of mental-health outcomes is infeasible.

3.9 Insights from Human-centered & HCI approaches for NLP/AI for Mental Health: Some Provocations

Stevie Chancellor (University of Minnesota – Minneapolis, US)

I will discuss three provocations for the field of AI/NLP for mental health, guided by insights from empirical work in HCI and human-centered AI. My approach to this problem combines my disciplinary training in Media Studies and Computer Science. To unpack the transformative potential of human-centered AI, we'll look at my group's work in mental illness and online social systems (social media and generative AI) as examples. The goal of this is to inspire conversation among participants, encourage solutions, and spark interdisciplinary reflection.

3.10 From Physics to Psychiatry: Dynamical Systems, Language, and Control

Hamidreza Jamalabadi (Philipps-Universität Marburg, DE)

Psychiatric interventions often focus on symptom alleviation, lacking the ability to fully capture the intricate dynamics of cognitive-affective states or tailor treatments to individual needs. From a dynamical systems perspective, though, the intervention can be observed in terms of an optimal control problem, where the ability to recover the dynamics based on continuous observation is key, as is the flexible implementation of control signals aimed at influencing cognitive-affective states – mirroring the severity and dynamics of mental disorders such as depression. Advances in natural language processing (NLP) and large language models (LLMs) offer transformative potential to address these limitations by enabling continuous, high-resolution monitoring of mental states through language, and further optimizing the processes of language-based therapies. This talk presents an integrative framework that combines physics-inspired dynamical systems theory with NLP. Language serves as a rich, temporally dynamic proxy for mental states, with LLMs facilitating the extraction of latent states and their temporal dynamics for predictive and interventional purposes. Optimized sampling rates, informed by ecological momentary assessment (EMA) and nonlinear modeling, enhance the detection of short- and long-term mood fluctuations. Furthermore, NLP-driven interventions, such as optimized psychotherapy and affective priming, can act as control inputs to reshape these trajectories, paralleling advances in neurostimulation that target pathological neural attractors. Early studies in our group show promising results in these directions, including those based on social media data (400,000 texts from more than 1,600 individuals over 6 years) and further experimental studies on optimizing interventions such as affective priming. This AI-informed neurocognitive dynamical systems framework paves the way for personalized interventions that stabilize resilient cognitive-affective trajectories, offering a paradigm shift from current approaches.

3.11 Supporting Mental-Health Communication: Towards a Proactive AI Support for (Human) Therapists

Cristian Danescu-Niculescu-Mizil (Cornell University – Ithaca, US)

URL https://convokit.cornell.edu

Recent years have seen a gold rush to replace people with AI agents in communication: they can serve as your therapist, your tutor, your financial advisor, and your interviewer. In this talk, I propose a contrasting vision: one in which AI supports humans in their communication while preserving their agency. Achieving this vision requires moving beyond the current transactional paradigm embodied by current generative AI systems, which are designed to fulfill the immediate goals of a single person, such as answering a question, solving a math problem, booking a flight, or (repeatedly) replying in character. To meaningfully support human–human communication without disrupting or supplanting it, an AI system must instead follow a proactive paradigm: it needs to decide when to intervene to offer support

as the interaction unfolds, rather than wait to explicitly be prompted as AI agents and chatbots do today. In this talk, I present initial progress on AI technologies that enable such a proactive mode of operation and demonstrate them in the context of mental health crisis counseling. Data and code are available through ConvoKit.

3.12 Catching Disengagement Early: Development and Validation of an LLM Rating Scale for Client Engagement in Psychotherapy

Steffen T. Eberhardt (Trier University, DE)

Main reference Steffen T. Eberhardt, Antonia Vehlen, Jana Schaffrath, Brian Schwartz, Wolfgang Lutz, Tobias Baur, Dominik Schiller, Tobias Hallmen, Elisabeth André: "Development and validation oflarge language model rating scalesfor automatically transcribedpsychological therapy sessions", Sci Rep 15, 29541 (2025)
URL https://doi.org/10.1038/s41598-025-14923-y

Rating scales have shaped psychological research, but are resource-intensive and can burden participants. Large Language Models (LLMs) provide a tool for assessing latent constructs in text. This study introduces *LLM rating scales* that use LLM responses rather than human ratings. We demonstrate this approach using an LLM rating scale to measure patient engagement in therapy transcripts. Automatically transcribed videos of 1,131 sessions from 155 patients were analyzed using DISCOVER, a software framework for local multimodal human behavior analysis. Llama 3.1 8B rated 120 engagement items, averaging the top eight into a total score. Psychometric evaluation showed a normal distribution, strong reliability ($\omega = 0.953$), and acceptable fit (CFI = 0.968, SRMR = 0.022), except RMSEA = 0.108. Validity was supported by significant correlations with engagement determinants (e.g., motivation, $r = .413$), processes (e.g., between-session efforts, $r = .390$), and outcomes (e.g., symptoms, $r = -.304$). Results remained robust across bootstrap resampling and cross-validation, accounting for the nested data structure. The LLM rating scale exhibited strong psychometric properties, demonstrating the potential of the approach as an assessment tool. Importantly, this automated approach uses interpretable items, ensuring a clear understanding of measured constructs, while supporting local implementation and protecting confidential data.

3.13 CARE: Training Counselors via LLMs

Ryan Louie (Stanford University, US)

The global mental health crisis demands innovative approaches to scale high-quality care. While AI chatbots for therapeutic use are on the rise, I argue that research should develop scalable solutions for contexts where human support remains essential. I will present our work developing CARE, a Large Language Model (LLM)-based training system that empowers human counselors through practice with AI-simulated patients and feedback from AI mentors. In this talk, I highlight two research thrusts: technical challenges in training counselors with LLMs and evaluating the impact of LLM-based training with novice counselors. Advancing

these research thrusts requires answering interdisciplinary questions at the intersection of natural language processing, mental health, and human AI interaction. How might we develop realistic LLM simulations of patients when privacy concerns restrict data access and domain-expert feedback is expensive? How can we mitigate the chance of generic or clinically-inappropriate LLM feedback? How can we conduct stage-appropriate user studies of LLM training systems that yield actionable insights to improve design? Our work advancing CARE contributes to responsible AI in mental health by developing new tools, algorithms, and empirical evidence for scaling counselor education via LLMs, thereby increasing the supply of well-trained human therapists to meet society's growing demands.

3.14 Helping the Helper: How AI Can Support Training of Peers in Delivery of Behavioral Health Care

Daniel Blonigen (VA Palo Alto Health Care System, US & Stanford University, US)

Globally, there is a demand–capacity problem in mental health care. One in four individuals has a behavioral health disorder (substance use and/or mental illness), yet more than half of all behavioral health providers report no openings for new patients, and many report burnout with existing caseloads. Peer Recovery Workers (PRWs) have been cited as critical to improving access and engagement in behavioral healthcare and mitigating staffing shortages. However, high rates of burnout and turnover are well-documented challenges to implementing PRWs, often stemming from inadequate training and supervision. Artificial Intelligence (AI) could be an ideal solution for ensuring high-quality training and ongoing supervision of PRWs in a strained healthcare system with limited resources. In this talk, we review how peer competencies include many of the interpersonal skills foundational to effective psychotherapy (e.g., rapport-building, active listening, reflecting, validating, empathizing), and the value of scaling peer training in these competencies using AI. CARE is an AI-powered web-based platform to train and empower individuals in basic counseling skills. In CARE, individuals select and/or create AI patients to roleplay emotional support conversations. Using a Large Language Model trained by senior psychotherapy supervisors, CARE provides counselors with feedback on their responses to the AI patients using a multi-level structure that mirrors how supervisors provide feedback to trainees. Although lab experiments have established CARE's proof-of-concept to improve the feedback quality provided to novice counselors, the program was not designed for PRWs whose scope of practice and training needs differ from those of traditional counselors (e.g., sharing lived experiences). Consequently, there may be a need to customize and pilot the use of CARE with PRWs. Mixed-method designs are suggested to collect systematic feedback from relevant stakeholders to guide customization of CARE for peers and then evaluate the feasibility and acceptability of the customized version with novice peers working in real-world clinical settings. This research program has the potential to address the demand–capacity problem by increasing the adoption of the PRW workforce in behavioral health settings and improving the quality of care they provide to patients.

3.15 Designing Technologies for Digital Mental Health: an HCI Perspective

Gavin Doherty (Trinity College Dublin, IE)

In this talk, I present a Human–Computer Interaction perspective on the design of digital health interventions for mental health. Drawing on experience developing and evaluating a range of novel systems to support the delivery of mental health care, the talk considers the critical yet often ill-defined concepts of acceptability and engagement, before examining in more detail the HCI issues surrounding the use of machine learning technologies in this context, looking at a specific example relating to outcome prediction. The talk concludes with a brief consideration of more personalised interventions, including the use of LLM-based capabilities within a broader design space for Ecological Momentary Interventions, along with the associated design issues.

3.16 Practices of NLP for Mental Health in China

Minlie Huang (Tsinghua University – Beijing, China)

URL http://coai.cs.tsinghua.edu.cn/hml/

In this talk, the speaker discusses his NLP practices for mental health applications in China. Specifically, he developed tools with LLMs for mental state assessment, built chatbots to provide effective emotional support, and developed LLM models for simulating clients and therapists for the purpose of training and evaluation. Some of these tools have been deployed in real-world applications.

3.17 Culture, Personalization and Mental Health

Monojit Choudhury (Mohamed Bin Zayed University of Artificial Intelligence – Abu Dhabi, AE)

Mental health conditions exhibit extreme behavioral variability. A stunning example is Autism Spectrum Disorder (ASD), a catchall term covering a wide range of symptoms and varying degrees of communication difficulties. Only recently have we begun to understand the types and causes of ASD, and specialized interventions that work for every individual are still a far-off dream. AI, particularly large language models (LLMs), offers a unique opportunity to continuously learn from a user's behavior, along with demographic information, patient history, and assessment results, which could serve as sandboxes for testing different (potentially novel) interventions before delivering the optimal one. In my talk, I discuss recent work that simulates user behavior in terms of what a user knows and does not know, and how to explain unfamiliar concepts in a lucid and personalized way from a cross-cultural communication perspective. These studies reveal that LLMs can learn to personalize, but

they also tend to produce stereotypical, less variable responses than real users. LLMs often agree with responses from other LLMs, but not as much with human users, who align more strongly with each other. This insight suggests that culture can serve as an excellent prior for personalization, but systems must also continuously learn from unfolding behavior to achieve the best results. Such methods provide a promising direction toward individual mental health–relevant behaviors for intervention sandboxing and innovation.

3.18 Artificial Intelligence, Affective Computing, and Health: Opportunities and Ethical Considerations in Real-World Data Collections

Emily Mower Provost (University of Michigan – Ann Arbor, US)

Emotions provide critical cues into our health and well-being. They are particularly important in the context of mental health, where changes in emotion may signify changes in symptom severity. However, information about emotion and its temporal variation is often accessible only through survey methodologies (e.g., ecological momentary assessment, EMA), which can become burdensome over time. Affective computing technologies, such as automated speech emotion recognition systems, could provide an alternative, namely offering quantitative measures of emotion using acoustic data captured passively from a consented individual's environment. However, these technologies are not without risk and can pose a potential for harm. There are critical ethical issues that must be thoughtfully considered. In this talk, I discuss our journey in affective computing for health modeling, presenting the design of these technologies alongside the ethical considerations that have shaped their development.

3.19 Dear ChatGPT, Can You Keep My Secret? Privacy and Security in the Era of LLMs

Anmol Goel (TU Darmstadt, DE & University of Copenhagen, DK)

In this talk, I explore the critical privacy and security challenges of Large Language Models (LLMs). Users are increasingly turning to LLMs for sensitive and personal interactions, including advice, companionship, and counseling. This trend creates a "Personalization–Privacy Paradox," in which the utility of personalized AI is in direct conflict with the need to protect user data.

The talk outlines concrete privacy threats, including membership inference, data leakage, and prompt extraction attacks. To combat these risks, a dual approach is proposed: Proactive Privacy (privacy by design) and Reactive Privacy (unlearning). I discuss recent results showing that differential privacy can reliably work for steering vectors, and that current unlearning evaluations are suboptimal – providing only a false sense of privacy. Finally, I present ongoing work on data poisoning and attribution methods for language models.

3.20 Computational Paralinguistics – In a Nutshell

Andreas Triantafyllopoulos (Technical University of Munich, DE)

Speech is about more than content – it's about what you say, when, and how. The latter two facets of spoken language – the "when" and the "how" – fall under the purview of computational paralinguistics. The use of *para* emphasizes contrast to classic linguistics. Paralinguistics concerns phenomena often neglected by traditional computational linguistics. In this talk, I provide a brief overview of the field, starting with motivational examples and a taxonomy of the relevant phenomena. I outline methodological approaches used in computational paralinguistics, from traditional feature-based analysis to modern foundation model–based methods that integrate speech and language processing. I also present a recent case study on using speech in the mental health context.

4 Working Groups

In addition to vision talks and breakout discussions, the seminar gave rise to three ongoing working groups. Each group brings together participants from NLP and clinical psychology to develop joint position or perspective papers that extend the conversations initiated at Dagstuhl into concrete research agendas and community guidelines. Together, these efforts reflect the seminar's overarching themes: understanding dynamic mental health processes, supporting therapist training, and ensuring that AI systems for mental health are evaluated and deployed responsibly.

4.1 Responsible Evaluation of AI for Mental Health Systems

This working group, led by Iryna Gurevych (Technical University of Darmstadt, DE), focuses on how AI systems for mental health should be evaluated before they can be trusted in sensitive clinical and community settings. Building on analysis of recent NLP work on mental health, the group identified recurring gaps in current practice, including over-reliance on automated metrics, limited involvement of clinicians or people with lived experience, and insufficient attention to safety, equity, and context.

In response, the group is developing a taxonomy that distinguishes assessment, intervention, and support systems and links each aspect to appropriate evaluation dimensions. The resulting position paper will provide a shared, clinically grounded framework that allows AI researchers, clinicians, implementation scientists, and other stakeholders to speak a common evaluative language and move toward more reliable, interpretable, and socially responsible AI for mental health.

4.2 Simulating AI Patients for Psychotherapy: Challenges and Opportunities

A second working group, led by Diyi Yang (Stanford University, US), examines the emerging use of AI-simulated patients as training tools for psychotherapy and counseling. The group starts from the long tradition of actor-based roleplay in clinical training and asks what it

would take to extend these practices into scalable, AI-driven simulation environments. Key questions include how to balance conversational realism with pedagogical value and safety; how to represent long-term therapeutic trajectories versus short, skills-focused exercises; and how to ensure cultural, linguistic, and demographic diversity, including rare and edge-case presentations.

The group also considers data governance, risks of "overfitting" trainees or models to simulators, and methods for benchmarking educational utility across therapeutic orientations. The planned perspective paper will articulate design principles, evaluation standards, and a research roadmap for AI-simulated patients that are not only technically sophisticated but also ethically sound and transferable to real-world clinical practice.

4.3 Identifying Intrapersonal and Interpersonal Dynamics Predictive of Change in Mental Health

The third working group, led by Dana Atzil-Slonim (Bar-Ilan University, IL), focuses on intrapersonal and interpersonal dynamics as central mechanisms of change in mental health. The group brings together clinical and AI researchers to ask which aspects of within-person experience and between-person interaction – such as emotional trajectories, synchrony, co-regulation, or conversational redirection – are most predictive of positive (or negative) outcomes, and how these phenomena can be modeled using modern AI methods.

The planned position paper will first synthesize clinical and methodological work on verbal and non-verbal dynamics at multiple temporal scales (from utterances to sessions and longer-term courses of care), and then review AI techniques for capturing temporal and multimodal patterns. Building on this, the group will identify promising intersections between the two literatures, highlighting opportunities to leverage advances in NLP and multimodal models to analyze longitudinal mental health datasets to identify intrapersonal and interpersonal dynamics predictive of positive outcomes.

5 Conclusion and Outlook

This Dagstuhl Seminar on "Natural Language Processing for Mental Health" brought together researchers from NLP, clinical psychology, human–computer interaction, and digital mental health to examine how language technologies can responsibly support mental health needs. Across vision talks, lightning presentations, and breakout discussions, participants highlighted both the transformative potential of NLP and large language models for understanding mental states, supporting therapeutic processes, and scaling access to care, as well as the substantial challenges around evaluation, temporality, multilinguality, privacy, and ethics that must be addressed before these systems can be trusted in real-world, high-stakes settings.

The seminar has laid the necessary groundwork for tackling these challenges through concrete, interdisciplinary collaborations. The three working groups initiated at Dagstuhl are developing perspective and position papers on responsible evaluation of AI for mental health systems, the design and use of AI-simulated patients for psychotherapy training, and the identification of intrapersonal and interpersonal dynamics predictive of change in mental health. Alongside planned community-building activities such as blogs, workshops, and joint projects, these efforts aim to articulate shared frameworks, datasets, and research

agendas that bridge technical and clinical expertise. In doing so, the seminar advances a growing cross-field community committed to developing NLP and AI for mental health that is clinically meaningful, empirically grounded, and ethically responsible.

Participants

- Tim Althoff
University of Washington – Seattle, US
- Hiba Arnaout
TU Darmstadt, DE
- Dana Atzil-Slonim
Bar-Ilan University – Ramat-Gan, IL
- Daniel Blonigen
Stanford University, US
- Stevie Chancellor
University of Minnesota – Minneapolis, US
- Monojit Choudhury
MBZUAI – Abu Dhabi, AE
- Torrey Creed
University of Pennsylvania, US
- Cristian Danescu-Niculescu-Mizil
Cornell University – Ithaca, US
- Munmun De Choudhury
Georgia Institute of Technology – Atlanta, US
- Gavin Doherty
Trinity College Dublin, IE
- Steffen Eberhardt
Universität Trier, DE
- Anmol Goel
TU Darmstadt, DE
- Philipp Graffe
Universität Stuttgart, DE
- Iryna Gurevych
TU Darmstadt, DE
- Nick Haber
Stanford University, US
- Dirk Hovy
Bocconi University – Milan, IT
- Darya Hryhoryeva
Charles University – Prague, CZ
- Minlie Huang
Tsinghua University – Beijing, CN
- Zac Imel
University of Utah, US
- Hamidreza Jamalabadi
Universität Marburg, DE
- Christopher Landau
Universitätsklinikum Frankfurt, DE
- Jana Lasser
Universität Graz, AT
- Maria Liakata
Queen Mary University of London, GB
- Ryan Louie
Stanford University, US
- Wolfgang Lutz
Universität Trier, DE
- Matteo Malgaroli
NYU School of Medicine – New York, US
- Emily Mower Provost
University of Michigan – Ann Arbor, US
- Clarissa Ong
University of Louisville, US
- Flor Miriam Plaza del Arco
Leiden University, NL
- Julia R Pozuelo
Harvard Medical School, US
- Alla Rozovskaya
City University of New York, US
- Brian Schwartz
Universität Trier, DE
- H. Andrew Schwartz
Vanderbilt University – Nashville, US
- Raj Sanjay Shah
Georgia Institute of Technology – Atlanta, US
- Bhavyajeet Singh
TU Darmstadt, DE
- Thamar Solorio
MBZUAI – Abu Dhabi, AE
- Aseem Srivastava
MBZUAI – Abu Dhabi, AE
- Jina Suh
Microsoft Research – Redmond, US
- Andreas Triantafyllopoulos
Klinikum rechts der Isar der TU München, DE
- Diyi Yang
Stanford University, US

Report from Dagstuhl Seminar 25362

Optimization and Automated Reasoning for Designing Future Space Missions

Max Bannach[*1], Johannes Klaus Fichte[*2], Dario Izzo[*3], Inês Lynce[*4], and Giacomo Acciarini[†5]

1 **ESA / ESTEC – Noordwijk, NL.** max.bannach@esa.int
2 **Linköping University, SE.** johannes.fichte@liu.se
3 **ESA / ESTEC – Noordwijk, NL.** dario.izzo@esa.int
4 **INESC-ID – Lisbon, PT.** ines.lynce@tecnico.ulisboa.pt
5 **University of Surrey – Guildford, GB.** g.acciarini@surrey.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25362 *Optimization and Automated Reasoning for Designing Future Space Missions,* which explored fundamental optimization and reasoning tasks that arise in early stages of designing complex space missions. Such tasks include *selecting* and *scheduling* the bodies that should be encountered, *routing* a spacecraft across multiple bodies optimally, or strategically *placing facilities* to support future missions. Many of these problems are still solved by hand, as current missions only contain a few celestial objects. However, with larger and increasingly complex missions, these problems become more relevant and, thus, there is an increasing need to solve space-related optimization, scheduling, and planning problems automatically.

Despite the promising opportunities for collaboration, the entry barrier to many of these problems remains high for those without a background in celestial mechanics. Conversely, modern tools and techniques from constraint reasoning and optimization are still largely unfamiliar to many aerospace researchers. The Dagstuhl Seminar 25362 successfully established a bridge between computer scientists working in automated reasoning and experts from the space domain focused on mission analysis and operations. This Dagstuhl Seminar brought together researchers from academia, industry, and space agencies, fostering interdisciplinary dialogue. Problems and tools from both communities were presented in a language accessible to the other, laying the groundwork for future joint research and development.

Seminar August 31 – September 3, 2025 – https://www.dagstuhl.de/25362

2012 ACM Subject Classification Applied computing → Aerospace; Applied computing → Operations research; Astrodynamics; Computing methodologies → Artificial intelligence; Theory of computation → Automated reasoning; Theory of computation → Constraint and logic programming

Keywords and phrases Automated Reasoning, Satellite Constellation Design, Space Logistics, Trajectory Optimization, Astrodynamics, Global Trajectory Optimization

Digital Object Identifier 10.4230/DagRep.15.8.80

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Optimization and Automated Reasoning for Designing Future Space Missions, *Dagstuhl Reports*, Vol. 15, Issue 8, pp. 80–94
Editors: Max Bannach, Johannes Klaus Fichte, Dario Izzo, Inês Lynce, and Giacomo Acciarini

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Max Bannach (ESA / ESTEC – Noordwijk, NL)
Johannes Klaus Fichte (Linköping University, SE)
Dario Izzo (ESA / ESTEC – Noordwijk, NL)
Inês Lynce (INESC-ID – Lisbon, PT)

Many tasks in early-stage mission design are still solved *manually,* as mission profiles tend to be small and subject to numerous constraints. However, the rise of the *new space* movement has significantly reduced mission costs and increased their frequency, creating a growing demand for *automation* in early design phases. This shift brings traditional computer science problems into focus, including *route planning* (e.g., traveling salesperson problems), *reliability analysis* (e.g., model counting), *scheduling* (e.g., graph coloring), and *facility location* (e.g., dominating set problems). As a result, automating mission design requires close collaboration between mission analysts and experts in automated reasoning. Yet, many of the modern tools developed in cost-optimal reasoning (e.g., maximum satisfiability), probabilistic reasoning (e.g., model counting), and constraint reasoning remain largely unfamiliar to the aerospace research community. Historically, this community has focused more on *local optimization,* e.g., computing optimal trajectories between celestial bodies, rather than on global optimization, like identifying optimal sequences across multiple targets. The goal of this Dagstuhl Seminar 25362 *Optimization and Automated Reasoning for Designing Future Space Missions* was to establish a bridge between these two communities to enable and activate future collaborations.

Computational Competitions

Early in the seminar, a shared passion quickly emerged as common ground between both communities: *computational competitions.* These are deeply rooted in the automated reasoning field, with flagship events such as the annual SAT competition [10] and the MAXSAT evaluation [6]. Computer scientists showed strong interest in the efforts of mission analysts to establish similar competitions within the space domain, e.g., the bi-annual Global Trajectory Optimisation Competition (GTOC) [1] and ESA's Space Optimization Competition (SPOC) [8]. Conversely, aerospace researchers were keen to learn from the SAT community's long-standing experience in organizing such events, particularly in the development of standardized interfaces, file formats, validators, and benchmark sets.

Future Space Logistics

The vast majority of hypothetical space missions discussed during the seminar are from the *space logistics* domain. According to the AIAA Space Logistics Technical Committee, space logistics is "the theory and practice of driving space system design for operability and supportability, and of managing the flow of materiel, services, and information needed throughout a space system lifecycle" [3]. Yuri Shimane provided a detailed tutorial on the topic, highlighting in particular the rise of mega-constellations such as Starlink or OneWeb due to massively reduced launch costs. The design, construction, and maintenance of such structures involves various problems that can naturally be solved with tools from the constraint programming toolbox, which was presented to the participants of the seminar in a tutorial by Laurent Perron.

Routing Problems under Keplerian Dynamics

One of the actively discussed topics during the seminar was a variant of the traveling salesperson problem with moving targets, where the targets follow Keplerian dynamics [2, 4, 12]. This formulation naturally arises in applications such as in-space servicing [7], active debris removal [11], in-orbit refueling [13], and asteroid mining [5]. In contrast to these multi-rendezvous missions (the spacecraft must match position and velocity with the target), some versions only require flybys (matching only the position). A representative example discussed during the seminar was asteroid observation missions, for which Naoya Ozaki presented results on the design of flyby cycler trajectories – a promising approach for repeated asteroid visits. Participants explored how techniques like *dynamic discretization discovery* and *branch-and-price* could help to address the time-dependent nature of these problems. Additionally, discussions focused on the potential advantages of leveraging technologies such as MAX-SAT, given the highly dynamic and multi-objective characteristics inherent to these routing problems.

Orbital Facility Location Problems

Another actively discussed topic during the seminar was facility location problems in orbital environments [14]. A classical example involves placing fuel depots in orbit to support sustainable in-orbit refueling missions, where a servicing spacecraft retrieves propellant from a depot and delivers it to a client. Participants explored how such problems can be discretized to make them amenable to automated reasoning techniques. It turned out that in some cases, these problems can be treated as static – for instance, when servicing times significantly exceed orbital periods (e.g., weeks or months versus hours). However, when such assumptions cannot be made, the problem becomes highly dynamic and time-dependent, requiring more sophisticated modeling and solution approaches similar to the routing problems.

Scheduling and Packing Problems

Additional relevant problem domains were proposed during the seminar, including satellite, constellation, and fleet scheduling problems [9], as well as 3D packing problems under physical constraints such as the system's center of mass (e.g., for cargo vessel loading). While these challenges appear to be natural candidates for techniques from the constraint optimization community, they were not explored within the scope of this seminar due to lack of time.

Artificial Intelligence

The design and operation of in-space infrastructure involves constraints driven by the system's time-varying properties (e.g., transfer costs), which are often non-linear. Two approaches discussed during the seminar were: (1) *full discretization* through pre-computation, which is conceptually straightforward but computationally expensive; and (2) the use of *surrogate models,* typically neural-network-based approximators, which can be integrated into optimization frameworks. In discussions with industry experts such as Robert Luce, participants explored how such integrations could be realized and what kinds of interfaces commercial solvers should support to facilitate this interaction.

Verification and Validation of Neural Networks

Although not a central theme of the seminar, attention was drawn to the stringent safety and reliability standards that neural networks must meet to be certified for on-board use. The automated reasoning community, with its expertise in formal verification, offers promising tools to certify neural network reliability automatically. As a result, future collaborations in this domain were initiated.

Seminar Agenda

Given that this Dagstuhl Seminar brought together two distinct communities, each day began with two tutorial talks: one focused on a computer science topic and the other on a space-related topic. On the first day, Harry Holt presented a tutorial on the fundamental building blocks of (multi-)rendezvous missions, while Matti Järvisalo introduced the concept of maximum satisfiability. The second day featured the tutorials discussed in the previous section, and on the final day, Abdin Adam provided a compelling bridge between optimization techniques and space logistics.

To foster collaboration and interaction, the seminar contained a *problem session* on the first day. Zhong Zhang introduced the *Global Trajectory Optimization Competition,* while Giacomo Acciarini and Manuel López-Ibáñez presented various formulations of the *traveling salesperson problem under Keplerian dynamics.* Following this session, participants engaged in breakout groups to explore the proposed challenges in more depth. These sessions focused on three main topics: (1) the use of MAX-SAT and *dynamic discretization discovery* for solving time-dependent routing problems (chaired by Max Bannach), (2) the integration and support of *non-linear constraints in modern solvers* (chaired by Robert Luce), and (3) the computational aspects of a *future mission to the Saturn system,* including how a spacecraft might leverage its moons for gravitational braking (chaired by Laurent Beauregard).

Additionally, two sessions of *inspiring talks* were organized, giving young researchers the opportunity to share ideas from their current work and spark new discussions. Robyn Natherson spoke about challenges in *low-thrust trajectory design,* Chit Hong Yam addressed issues in *sustainable lunar logistics,* and Thorsten Ehlers presented on *trajectory optimization at DLR.* These space-focused insights were complemented by contributions from the computer science community: Anna Latour discussed *reasoning under uncertainty,* Alexandra Lassota analyzed *structural properties of integer programs,* and Stefan Szeider explored *synergies between language models and constraint reasoning.* As is tradition at Dagstuhl, some of the most engaging conversations took place during the Tuesday hike, which provided an informal yet productive setting for deeper interdisciplinary exchanges.

Future Work

As the primary objective of this seminar was to raise awareness of the tools and challenges developed within the computer science and space communities in recent years, much of the time was dedicated to presenting these resources rather than solving specific problems. A natural next step is a more solution-oriented workshop, focused on developing algorithms for concrete applications using techniques from the automated reasoning community. To facilitate this collaboration, participants expressed a clear desire for standardized interfaces, file formats, and benchmark sets.

Moreover, due to the limited duration of the seminar, many important topics could only be touched upon briefly or not at all. These include cargo packing, reliability analysis of constellations under uncertainty, sustainability aspects, applications to planetary defense, and satellite traffic management. These areas present promising directions for future interdisciplinary exploration.

References

1 Global Trajectory Optimisation Competition. https://sophia.estec.esa.int/gtoc_portal/, 2024. Accessed: 11.04.2024.

2 Adam Abdin. Strategic Management of On-Orbit Servicing: Leveraging Operations Research Methods for Enhanced Mission Planning and Scheduling. In *18th International Conference on Space Operations*, 2025.

3 AIAA Space Logistics Technical Committee. Definition of Space Logistics. https://www.aiaa-sltc.org/, 2024. Accessed: 06.09.2025.

4 Max Bannach, Giacomo Acciarini, and Dario Izzo. On the Keplerian TSP and VRP: Benchmarks and Encoding Techniques. In *International Astronautical Congress*, 2024.

5 A. Bellome, J.P. Sánchez, J.C. García Mateas, L. Felicetti, and S. Kemble. Modified Dynamic Programming for Asteroids Belt Exploration. *Acta Astronautica*, 215:142–155, 2024.

6 Jeremias Berg, Matti Järvisalo, Ruben Martins, Andreas Niskanen, and Tobias Paxian. MaxSAT Evaluation 2024: Solver and Benchmark Descriptions. 2024.

7 Alec J Cavaciuti, Joseph H Heying, and Joshua Davis. In-space Servicing, Assembly, and Manufacturing for the New Space Economy. *Aerospace Center for Space Policy and Strategy*, pages 2022–07, 2022.

8 ESA. SpOC. https://www.esa.int/Enabling_Support/Space_Engineering_Technology/Help_make_an_orbital_megastructure_with_genetic_computation, 2024. Accessed: 08.04.2024.

9 Benedetta Ferrari, Jean-François Cordeau, Maxence Delorme, Manuel Iori, and Roberto Orosei. Satellite Scheduling Problems: A Survey of Applications in Earth and Outer Space Observation. *Comput. Oper. Res.*, 173:106875, 2025.

10 Marijn JH Heule, Markus Iser, Matti Järvisalo, and Martin Suda. Proceedings of SAT Competition 2024: Solver, Benchmark and Proof Checker Descriptions. 2024.

11 Dario Izzo, Ingmar Getzner, Daniel Hennes, and Luís Felismino Simões. Evolving Solutions to TSP Variants for Active Space Debris Removal. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2015, Madrid, Spain, July 11-15, 2015*, pages 1207–1214, 2015.

12 Manuel López-Ibáñez, Francisco Chicano, and Rodrigo Gil-Merino. The Asteroid Routing Problem: A Benchmark for Expensive Black-Box Permutation Optimization. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pages 124–140. Springer, 2022.

13 Daria Malyh, Sergey Vaulin, Victor Fedorov, Ruslan Peshkov, and Mikhail Shalashov. A Brief Review on in-orbit Refueling Projects and Critical Techniques. *Aerospace Systems*, 5(2):185–196, 2022.

14 Yuri Shimane, Nicholas Gollins, and Koki Ho. Orbital Facility Location Problem for Satellite Constellation Servicing Depots. *Journal of Spacecraft and Rockets*, 61(3):808–825, 2024.

2 Table of Contents

3 Overview of Talks

3.1 The Keplerian Traveling Salesperson Problem

Giacomo Acciarini (ESA / ESTEC – Noordwijk, NL)

Joint work of Max Bannach, Giacomo Acciarini, Dario Izzo
Main reference Max Bannach, Giacomo Acciarini, Dario Izzo: "On the Keplerian TSP and VRP: Benchmarks and Encoding Techniques", International Astronautical Congress, Milan, 2024.
URL https://openresearch.surrey.ac.uk/esploro/outputs/99927464502346

This talk addresses a central challenge in space mission design and logistics: planning interplanetary trajectories for missions that must rendezvous with multiple bodies, such as in active debris removal, in-orbit servicing, or asteroid belt exploration. The problem is captured by the Keplerian Traveling Salesperson Problem (KTSP), an extension of the classical TSP that incorporates the orbital motion of targets. In contrast to the standard TSP, the KTSP features time-dependent and asymmetric transfer costs.

The talk presents a rigorous formalization of the KTSP together with a benchmark suite that includes globally optimal solutions, providing a basis for comparison with heuristic methods. A time-unfolding technique reformulates the continuous orbital dynamics as a discrete optimization problem in a time-expanded network, making the benchmark accessible to the discrete optimization community without prior expertise in celestial mechanics. An alternative formulation as an integer linear program is also introduced, using Interval-based Dynamic Discretization Discovery to capture the time-dependent structure of orbital transfers.

To enable practical comparisons and foster research in the field, exact methods are complemented with initial solution heuristics, improvement strategies, and preprocessing routines, and compared to heuristics like beam search and variants of the 2-opt algorithm. The overall framework demonstrates how complex multi-body rendezvous problems, ranging from asteroid exploration to on-orbit servicing and debris removal, can be systematically addressed within a rigorous optimization setting.

3.2 Optimization in Future Space Logistics

Abdin Adam (CentraleSupélec – Gif sur Yvette, FR)

The continued expansion of space activities is giving rise to a new class of logistical challenges, where the effective management of orbital assets is critical to the resilience, adaptability, and sustainability of space infrastructure. Missions such as satellite refueling, repair, debris removal, and life-extension will increasingly rely on carefully optimized planning and coordination across multiple timescales. These missions introduce distinctive challenges in modeling, optimization, and decision-making, as they must reconcile resource limitations, orbital dynamics, operational feasibility, and uncertainty inherent to the space environment. Operations research (OR) provides a rigorous methodological basis for addressing such problems, combining mathematical modeling with analytical and computational techniques. We discuss OR frameworks relevant to space logistics and future space missions management, including deterministic optimization, stochastic programming, robust optimization, sequential decision-making, and associated decision analysis methods. We further illustrate how these approaches can be integrated with automated reasoning and intelligent algorithms to support mission planning across different temporal and operational scales.

References

1 Ho, K. (2024). Space logistics modeling and optimization: Review of the state of the art. *Journal of Spacecraft and Rockets*, 61(5), 1417–1427. American Institute of Aeronautics and Astronautics.
2 Abdin, A. (2025). Strategic Management of On-Orbit Servicing: Leveraging Operations Research Methods for Enhanced Mission Planning and Scheduling. In *Proceedings of the 18th International Conference on Space Operations*.
3 Bannach, M., Acciarini, G., & Izzo, D. (2024). On the Keplerian TSP and VRP: Benchmarks and Encoding Techniques. In *Proceedings of the International Astronautical Congress*.

3.3 Trajectory Optimization at DLR

Thorsten Ehlers (DLR – Hamburg, DE)

In this talk I will show some optimization problems that were solved in the institute for air transport at DLR. While the main focus of our institute is on the aerospace side, it is highly relevant for us to use the right optimization algorithms, e.g. in evaluating operational concepts for new aircraft.

3.4 Tutorial Talk on Multi-Rendezvous Missions

Harry Holt (ESA / ESTEC – Noordwijk, NL)

The objective of this talk was to introduce the field of multi-rendezvous/encounter missions in astrodynamics to a non-expert audience. After covering some of the building blocks in orbital mechanics, such as propagation, orbital elements and reference frames, we discussed the constraints imposed by different encounters and propulsion systems. Lambert's problem was introduced for solving two-impulse transfers, and direct encodings were presented for deep-space manoeuvres, multi-impulse trajectories, low-thrust trajectories, and gravity assists. Finally, space-specific methods for solving the outer combinatorial part were presented, including pruning approaches and approximating the cost function.

3.5 Tutorial on Maximum Satisfiability

Matti Järvisalo (University of Helsinki, FI)

Main reference Fahiem Bacchus, Matti Järvisalo, Ruben Martins: "Maximum Satisfiability", pp. 929–991, IOS Press, 2021.
URL http://dx.doi.org/10.3233/FAIA201008

We provide a high-level overview of maximum satisfiability (MaxSAT), covering basics on encoding problems as MaxSAT, the algorithmic approaches implemented in modern MaxSAT solvers, and pointers to further recent developments in MaxSAT solving.

3.6 What do we do with Integer Programs in Theory?

Alexandra Lassota (TU Eindhoven, NL)

This talk will give you a little snippet on the TCS side of integer programs: What are some of us actually doing? What are our challenges? What are our accomplishments?

3.7 Which Variables Matter? Structure-based Sensitivity Analysis for Reasoning Under Uncertainty

Anna Latour (TU Delft, NL)

Real-world problems typically contain a lot of structure that we can exploit to solve them fast in practice. For example: in many problems that reason about uncertainty, we can capture the problem in a decision diagram, and use that DD's structure to not only prune the search space of strategies, but also efficiently integrate over scenarios to evaluate the quality of a strategy.

For a satellite constellation for Earth observation design problem, we recently showed that we can improve upon the state of the art by framing the original problem as a computationally harder problem, but with an exponentially smaller encoding. Thanks to the speed of solvers for high-complexity problems in practice, we can make orders of magnitude improvement in the size of the problems that we can solve. This is all thanks to smartly leveraging the structure of the problem. We are currently working on how to apply this trick to different (and in some cases more realistic) variants of the problem.

The next step is to increase our capabilities of optimisation under uncertainty by applying structure-based approaches to sensitivity analysis. The classical approach to identifying how sensitive your decisions are to the exact value of certain input parameters relies on simulations. These are computationally expensive, might miss important sensitivities due to their probabilistic nature (and hence provide statistical guarantees at best), and can typically only handle one parameter at a time.

I propose a new method that is based on the logical structure of the problem, instead. Advantages include that you get the interaction of variables for free and that you get access to formal verification technology.

3.8 An Exact Framework for Solving the Space-Time Dependent TSP

Manuel López-Ibáñez (University of Manchester, GB)

Main reference Isaac Rudich, Quentin Cappart, Manuel López-Ibáñez, Michael Römer, Louis-Martin Rousseau: "An Exact Framework for Solving the Space-Time Dependent TSP", 2024
URL https://doi.org/10.48550/arXiv.2312.01404

Many real-world scenarios involve solving bi-level optimization problems in which there is an outer discrete optimization problem, and an inner problem involving expensive or black-box computation. This arises in space-time dependent variants of the Traveling Salesman Problem, such as when planning space missions that visit multiple astronomical objects. Planning these missions presents significant challenges due to the constant relative motion of the objects involved. There is an outer combinatorial problem of finding the optimal order to visit the objects and an inner optimization problem that requires finding the optimal departure time and trajectory to travel between each pair of objects. The constant motion of the objects complicates the inner problem, making it computationally expensive. This paper introduces a novel framework utilizing decision diagrams (DDs) and a DD-based branch-and-bound technique, Peel-and-Bound, to achieve exact solutions for such bi-level optimization problems, assuming sufficient inner problem optimizer quality. The framework leverages problem-specific knowledge to expedite search processes and minimize the number of expensive evaluations required. As a case study, we apply this framework to the Asteroid Routing Problem (ARP), a benchmark problem in global trajectory optimization. Experimental results demonstrate the framework's scalability and ability to generate robust heuristic solutions for ARP instances. Many of these solutions are exact, contingent on the assumed quality of the inner problem's optimizer.

3.9 The Asteroid Routing Problem: a Benchmark for Expensive Black-Box Permutation Optimization

Manuel López-Ibáñez (University of Manchester, GB)

Main reference Manuel López-Ibáñez, Francisco Chicano, Rodrigo Gil-Merino: "The Asteroid Routing Problem: A Benchmark for Expensive Black-Box Permutation Optimization", in Proc. of the Applications of Evolutionary Computation, pp. 124–140, Springer International Publishing, 2022.
URL https://doi.org/10.1007/978-3-031-02462-7_9

Inspired by the recent 11th Global Trajectory Optimisation Competition, this paper presents the asteroid routing problem (ARP) as a realistic benchmark of algorithms for expensive bound-constrained black-box optimization in permutation space. Given a set of asteroids' orbits and a departure epoch, the goal of the ARP is to find the optimal sequence for visiting the asteroids, starting from Earth's orbit, in order to minimize both the cost, measured as the sum of the magnitude of velocity changes required to complete the trip, and the time, measured as the time elapsed from the departure epoch until visiting the last asteroid. We provide open-source code for generating instances of arbitrary sizes and evaluating solutions to the problem. As a preliminary analysis, we compare the results of two methods for expensive black-box optimization in permutation spaces, namely, Combinatorial Efficient Global Optimization (CEGO), a Bayesian optimizer based on Gaussian processes,

and Unbalanced Mallows Model (UMM), an estimation-of-distribution algorithm based on probabilistic Mallows models. We investigate the best permutation representation for each algorithm, either rank-based or order-based. Moreover, we analyze the effect of providing a good initial solution, generated by a greedy nearest neighbor heuristic, on the performance of the algorithms. The results suggest directions for improvements in the algorithms being compared.

3.10 Reachability-Informed Low-Thrust Trajectory Design: Progress and Challenges

Robyn Natherson (University of Colorado Boulder, US)

Joint work of Robyn Natherson, Daniel J Scheeres
Main reference Robyn Natherson, Daniel J Scheeres: "Reachability-informed missed thrust design". In AAS/AIAA Astrodynamics Specialist Conference, number AAS 25-609, 2025.
URL https://www.researchgate.net/publication/395012227_Reachability-Informed_Missed_Thrust_Design_AAS_25-609

Low-thrust propulsion enables deep space exploration at a fraction of the fuel cost. However, these systems require significantly longer thrust arcs compared to those with conventional chemical propulsion systems. Burn durations can span days, weeks, or even months. The extended thrusting periods of low-thrust systems increase the likelihood that a spacecraft anomaly will overlap with a planned burn, causing the spacecraft to deviate from its nominal trajectory. This is referred to as the missed thrust problem. Without robust trajectory design, losses from missed thrust jeopardize mission objectives, especially when flight paths include encounters with celestial bodies.

My doctoral research focuses on how to leverage reachability results to design robust transfers accounting for missed thrust. Reachability results can be formulated in two ways, forwards and backwards. My work specifically utilizes backwards reachable sets, or controllable sets. Controllable sets characterize the region of full-state initial conditions which can reach a target within a finite time. Typically, designing for missed thrust involves an iterative process where candidate transfers are tested for robustness and updated until design constraints are satisfied. However, applying reachability theory to study the missed thrust problem would allow the entire solution space to be studied at once, avoiding this tedious process. I will discuss the use of a reachability framework to compute a robustness metric, the missed thrust recovery margin (MTRM). The MTRM is the amount of time a spacecraft can coast away from a nominal trajectory before the target becomes inaccessible. By sampling points on a controllable set, we map the robustness of regions in phase-space visualized with a heatmap. The goal is to leverage this reachability information to design flight paths through regions that have inherently high MTRM.

Challenges associated with this research include:

- Reachability computation – efficiency, accuracy, and sampling of full-state sets
- Representing reachability point-cloud data in a usable form – especially for non-convex reachability results
- Testing inside/outside of a controllable or reachable set
- Visualization of 4D or 6D full-state reachability data

Due to the missed thrust problem, robust design strategies are integral for promoting mission assurance for flight projects employing low-thrust systems. Our novel reachability-informed trajectory design approach has the potential to change how to view the missed thrust problem.

3.11 Asteroid Flyby Cycler Trajectory Design Using Deep Neural Networks

Naoya Ozaki (JAXA – Sagamihara, JP)

Asteroid exploration has been attracting more attention in recent years. Nevertheless, we have just visited tens of asteroids while we have discovered more than one million bodies. As our current observation and knowledge should be biased, it is essential to explore multiple asteroids directly to better understand the remains of planetary building materials. One of the mission design solutions is utilizing asteroid flyby cycler trajectories with multiple Earth gravity assists.

An asteroid flyby cycler trajectory design problem is a subclass of global trajectory optimization problems with multiple flybys, involving a trajectory optimization problem for a given flyby sequence and a combinatorial optimization problem to decide the sequence of the flybys. As the number of flyby bodies grows, the computation time of this optimization problem expands maliciously.

This paper presents a new method to design asteroid flyby cycler trajectories utilizing a surrogate model constructed by deep neural networks approximating trajectory optimization results. Since one of the bottlenecks of machine learning approaches is the computation time to generate massive trajectory databases, we propose an efficient database generation strategy by introducing pseudo-asteroids satisfying the Karush-Kuhn-Tucker conditions.

The numerical result applied to JAXA's DESTINY+mission shows that the proposed method is practically applicable to space mission design and can significantly reduce the computational time for searching asteroid flyby sequences.

3.12 Tutorial Talk on CSP

Laurent Perron (Google – Paris, FR)

We present OR-Tools, the suite of Operations Research tools build at Google and exported to github.

We give a more detailed overview of the CP-SAT solver, an award winning hybrid solver using MaxSAT, CP, MIP, and Meta-Heuristics techniques.

3.13 Tutorial Talk on Future Space Logistics

Yuri Shimane (Georgia Institute of Technology, US)

Main reference Yuri Shimane, Kento Tomita, Koki Ho: "Cislunar Space Situational Awareness Constellation Design and Planning with Facility Location Problem", Journal of Spacecraft and Rockets, Vol. 62(6), pp. 1938–1960, 2025.
URL http://dx.doi.org/10.2514/1.A36361

Space Logistics, which studies the design, operation, and maintenance of in-space infrastructures, requires reconciling combinatorial aspects of the problem with the underlying orbital mechanics. In this light, we begin by exploring the history of Space Logistics, from the Space Shuttle era to the present decade. We then explore two space-based applications of the facility location problem (FLP) in closer detail. The first example is for the placement of in-orbit servicing depots and their allocation to GNSS constellations. In this example, due to the periodicity of the depots' candidate orbits and the difference in time scale between the servicing allocation and the orbital periods, we assume a time-independent formulation. In contrast, the second example studying the placement and sensor-tasking of a cislunar space situational awareness constellation is time-dependent, requiring a time-expanded FLP. To tackle the problem size growth with the inclusion of discretized time, we develop a Lagrangian relaxation scheme that features an analytical relaxed solution, customized heuristics for the feasible solution, and that scales linearly with the number of time-steps. We conclude by looking towards the future – areas such as coordinated space traffic management, data logistics, and planetary defense, present exciting research avenues with real-life implications for the space economy.

3.14 Neural Meets Symbolic: Synergies Between Language Models and Constraint Reasoning

Stefan Szeider (TU Wien, AT)

Integrating Large Language Models (LLMs) with traditional solving techniques creates new synergies in automated reasoning. This talk explores both (i) how LLMs can enhance SAT and constraint solving through structural analysis and search guidance and (ii) how formal reasoning can help LLMs tackle hard reasoning and optimization problems. We will present case studies exploring the practical advances and future potential of combining neural and symbolic approaches in computational reasoning.

3.15 Challenges of Sustainable Lunar Logistics

Chit Hong Yam (ispace – Tokyo, JP)

Future lunar exploration will depend on building a sustainable logistics framework that connects data, mission components, and technologies into a coherent system. Yet critical uncertainties remain: how much and what type of lunar data is sufficient for planning? How

should interdependent mission elements – landers, habitats, power systems, and resource utilization – be prioritized? And who will coordinate logistics standards across multiple players? This talk highlights the challenges and trade-offs of sustainable lunar logistics, emphasizing the tension between incomplete knowledge and long-term commitments. Rather than presenting solutions, it frames the open questions that must be addressed before a truly sustainable lunar supply chain can emerge.

3.16 Global Trajectory Optimization Competition (GTOC) – GTOC12 Asteroid Mining

Zhong Zhang (Tsinghua University – Beijing, CN)

Main reference Zhong Zhang, Nan Zhang, Xiang Guo, Di Wu, Xuan Xie, Jia Yang, Fanghua Jiang, Hexi Baoyin: "Sustainable Asteroid Mining: On the design of GTOC12 problem and summary of results". Astrodyn 9, 3–17 (2025).
URL https://doi.org/10.1007/s42064-024-0199-3

The 12th Global Trajectory Optimization Competition (GTOC12) focused on the challenge of asteroid mining. Teams were tasked with designing trajectories for multiple mining spacecraft departing from Earth, visiting asteroids, extracting resources, and returning the mined material to Earth. The primary objective was to maximize the total returned mass of minerals. The problem combined optimal control and combinatorial optimization. From the control perspective, spacecraft needed to rendezvous with asteroids, requiring precise position and velocity matching under a two-body low-thrust dynamical model. From the combinatorial side, teams had to decide how many spacecraft to use, which asteroids to target, in what order, and at what times. The competition highlighted the intersection of astrodynamics, optimization, and mission design. Ultimately, innovative strategies enabled the top teams to achieve impressive results. The next competition, GTOC13, will be hosted by NASA JPL.

Participants

- Giacomo Acciarini
University of Surrey – Guildford, GB
- Abdin Adam
CentraleSupélec – Gif sur Yvette, FR
- Carlos Ansotegui
University of Lleida, ES
- Max Bannach
ESA / ESTEC – Noordwijk, NL
- Laurent Beauregard
Telespazio – Darmstadt, DE
- Thorsten Ehlers
DLR – Hamburg, DE
- Johannes Klaus Fichte
Linköping University, SE
- Harry Holt
ESA / ESTEC – Noordwijk, NL
- Dario Izzo
ESA / ESTEC – Noordwijk, NL
- Matti Järvisalo
University of Helsinki, FI
- Alfons Laarman
Leiden University, NL
- Alexandra Lassota
TU Eindhoven, NL
- Anna Latour
TU Delft, NL
- Manuel López-Ibáñez
University of Manchester, GB
- Robert Luce
Gurobi Optimization – Berlin, DE
- Inês Lynce
INESC-ID – Lisbon, PT
- Robyn Natherson
University of Colorado Boulder, US
- Naoya Ozaki
JAXA – Sagamihara, JP
- Laurent Perron
Google – Paris, FR
- Yuri Shimane
Georgia Institute of Technology – Atlanta, US
- Stefan Szeider
TU Wien, AT
- Polina Verkhovodova
Georgia Institute of Technology – Atlanta, US
- Felix Winter
TU Wien, AT
- Chit Hong Yam
ispace – Tokyo, JP
- Zhong Zhang
Tsinghua University – Beijing, CN