BIOINFERENCE

Schedule, Useful Information & Talk Abstracts

BioInference 2025: Sessions of Talks

Wednesday 28th - Morning

09:00 – 10:00 Registration

10:00 – 11:40 *Chair – Enrico Bibbona, Politecnico di Torino*

10:00 - 10:30		Welcome
10:30 – 11:10	Jennifer Asimit	Improved genetic discovery and fine-mapping resolution through multivariate latent factor analysis of high-dimensional traits
11:10 – 11:40	Jaromir Sant	Inference of genome-wide genealogical relationships between ancient and modern individuals

12:00 – 13:10 *Chair – Ioana Bouros, University of Oxford*

12:00 – 12:40	Antonietta Mira	On the notion of data intrinsic dimension and its biomedical applications
12:40 – 13:10	Florian Schunck	Modeling the Bigger Picture: Hierarchical Approaches for Integrating Diverse Toxicological Data in Mechanistic Models

Wednesday 28th - Afternoon

14:10 – 15:10 *Chair – Julia Brettschneider, University of Warwick*

14:10 – 14:40	Hetvi Jethwani & Kamal Sukandar	The role of rare somatic mtDNA mutations in proliferative tissue aging
14:40 – 15:10	Margherita Bruno	Inferring the acquisition age of JAK2-V617F mutation in MPN patients from phylogenetic data and an ABC-SMC model selection procedure

15:30 – 17:10 *Chair – Marina Evangelou, Imperial College London*

15:30 – 16:00	Hélène Ruffieux	A Bayesian functional factor model for high-dimensional molecular curves
16:00 – 16:30	Andrea Corbetta	Genetics helps determine differences in response to statin treatment: analysis on short and long-term low-density lipoprotein cholesterol trajectories
16:30 – 17:10	Zhana Kuncheva	Can causal inference complete the multi-omics game

17:10 – 19:00 Poster Session

Thursday 29th - Morning

08:30 – 10:40 *Chair – Giulia Capitoli, University of Milano-Bicocca*

08:30 – 09:00	Yoav Ram	Quantifying Cultural Evolution Using Population Genetic Methods
09:00 – 09:30	Ferdinando Insalata	Mutational signatures of deterministic and noise-induced evolutionary mechanisms
09:30 – 10:00	Meritxell Brunet Guasch	Quantifying the evolution of metastati potential across cancer tissues
10:00 – 10:40	Jonasz Słomka	How encounters at the microscale prime microbial interactions

11:00 – 12:40 *Chair – Massimiliano Tamborrino, University of Warwick*

11:00 – 11:30	Andrew Golightly	Accelerating Bayesian inference for stochastic epidemic models using incidence data
11:30 – 12:00	Henrik Häggström	Simulation-based inference for stochastic nonlinear mixed-effects models with applications in systems biology
12:00 – 12:40	Diego di Bernardo	The role of rare somatic mtDNA mutations in proliferative tissue aging

Friday 30th - Morning

09:00 – 10:40 *Chair – Elena Sabbioni, University of Oxford*

09:00 – 09:30	Richard Kettle	Quantifying Immune Cell Subtypes and States under Cytokine Perturbations States
09:30 – 10:00	Alberto Cassese	Bison - Bi-clustering of spatial omics data with feature selection
10:00 – 10:40	Davide Risso	Scalable and interpretable analysis of single-cell omics data

11:20 – 13:00 *Chair – Alicia Gill, University of Oxford*

11:20 – 12:00	Ben Swallow	Spatio-temporal Gaussian processes on GPUs
12:00 – 12:30	Nicholas Steyn	SMC methods for epidemic renewal models
12:30 - 13:00	Anastasia Mantziou	Bayesian model-based clustering for populations of network data

Friday 30th - Afternoon

14:00 – 15:30 *Chair – Martina Amongero, University of Turin*

14:00 - 14:30	Veronica Biancacci	Physics-Informed model-based reinforcement learning for pandemic mitigation
14:30 – 15:00	Simone Tiberi	IsoBayes: a Bayesian approach for single-isoform proteomics inference
15:00 – 15:30	Matthew Adeoye	Bayesian Spatio-temporal Modelling for Infectious Disease Outbreak Detection

15:30 – 16:00 **Prize Awards + Closing Remarks**

BioInference 2025: Useful Information

Conference Venue and covered meals/refreshments:

The talks and the poster session of the conference will be held at Palazzo delle Feste in Bardonecchia. The paid registration fee covers the three lunches, four coffee breaks, the conference dinner (see below) and the aperitivo that will be offered during the poster session on Wednesday.

Address: Piazza Valle Stretta, 1, 10052 Bardonecchia TO, Italy

Registration:

Registration will start from 9am on Wednesday 28th of May at Palazzo delle Feste.

Hiking / Networking Activity:

On Thursday afternoon, after the talks, a hiking activity in the Bardonecchia area will take place. This is an informal networking event. Packed lunch will be provided for the hiking activity. More details will be announced during the conference.

Suitable hiking clothes (e.g. a fleece jacket/jumper) and hiking shoes are recommended. Alternative routes based on difficulty/length will be proposed, and individuals will be able to choose between them. Please do note that there may be the possibility of encountering snow on some of the paths at higher altitudes, so keep that in mind.

Conference Dinner:

The social dinner will be on Thursday evening from 8pm at Harald's Ski Restaurant Bar, and it is included in the registration fee.

Prizes:

We will be awarding three prizes, each worth £200. The first prize is the *Reproducibility Prize* for the best reproducible code, to be submitted to the BioInference team (bioinference@gmail.com) by the **18th May 2025**. To be considered for this prize, the code must be housed in an open-source location (e.g. a public GitHub repository).

The other two prizes are meant for Early Career Researchers (ECRs), and consist of the *Best ECR Talk* and the *Best ECR Poster*. These will be voted by the BioInference participants (that is you!) during the conference, with more details shared at the conference.

Early Career Researchers Travel Claims:

We are happy to announce that due to the generous sponsorship received, we are able to reimburse some travel expenses of ECRs that are unable to fund through other means. If you are an ECR (i.e. if you do not hold a permanent academic position), you are invited to submit a travel claim for attending the BioInference conference. Email any requests alongside receipts to bioinference@gmail.com after the conference. We will consider any requests made by the **30th June**. The amount reimbursed will depend on the number of requests and available funds, as we aim to reimburse as many ECRs as possible.

Sponsors:

This year's conference is generously sponsored by:

- Politecnico di Torino
- Elsevier
- Society for Mathematical Biology
- European Society for Mathematical and Theoretical Biology (ESMTB)

We also thank IBS Italy for the patronage of BioInference2025.

BioInference 2025 Committee:

Conference Organisers present at the conference:

Enrico Bibbona

(Politecnico di Torino)

Marina Evangelou

(Imperial College London)

Massimiliano Tamborrino

(University of Warwick)

Martina Amongero

(University of Turin)

Elena Sabbioni

(University of Oxford)

Julia Brettschneider

(University of Warwick)

Ioana Bouros

(University of Oxford)

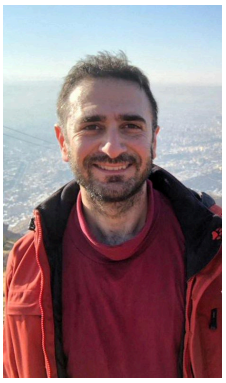
Alicia Gill

(University of Oxford)

Giulia Capitoli

(University of Milano-Bicocca)

Other Committee Members:

 Ben Lambert (University of Oxford)	 Constandina Koki (University of Warwick)	 Aden Forrow (Maine)
Andrii Krutsylo (Institute of Computer Science Polish Academy of Science)	 Hamid Rahkooy (University of Oxford)	

BioInference 2025: Talk Abstracts

Matthew Adeoye (University of Warwick)

Bayesian Spatio-temporal Modelling for Infectious Disease Outbreak Detection

The Bayesian analysis of infectious disease surveillance data from multiple locations typically involves building and fitting a spatio-temporal model of how the disease spreads in the structured population. Here we present new generally applicable methodology to perform this task. We introduce a parsimonious representation of seasonality and a biologically informed specification of the outbreak component to avoid parameter identifiability issues. We develop a computationally efficient Bayesian inference methodology for the proposed models, including techniques to detect outbreaks by computing marginal posterior probabilities at each spatial location and timepoint. We show that it is possible to efficiently integrate out the discrete outbreak parameters, enabling the use of dynamic Hamiltonian Monte Carlo as a complementary alternative to a hybrid Markov Chain Monte Carlo algorithm. We introduce a robust Bayesian model comparison framework based on importance sampling to approximate model evidence in high-dimensional space. The performance of our methodology is validated through systematic simulation studies, where simulated outbreaks were successfully detected, and model comparison strategy demonstrates strong reliability. We apply the methodology to monthly incidence data on invasive meningococcal disease from 28 European countries. The results highlight outbreaks across multiple countries and months, with model comparison analysis favouring the new specification over others.

Jennifer Asimit (University of Cambridge MRC Biostatistics Unit)

Improved genetic discovery and fine-mapping resolution through multivariate latent factor analysis of high-dimensional traits

Genome-wide association studies (GWAS) of high-dimensional traits, such as blood cell or metabolic traits, often use univariate approaches, ignoring trait relationships. Biological mechanisms generating variation in high-dimensional traits can be captured parsimoniously through GWAS of

latent factors. FlashfmZero is a latent-factor based multi-trait fine-mapping approach to identify shared and distinct causal variants underlying genetic associations amongst multiple traits. Applying flashfmZero to 25 latent factors derived from 99 blood cell traits in the INTERVAL cohort, we show that this approach uncovers genetic signals missed by standard univariate methods and refines credible sets in 87% of our comparisons. These latent factor approaches can be applied to GWAS summary statistics and will enhance power for the discovery and fine-mapping of associations for many traits.

Veronica Biancacci (Vrije Universiteit Brussel)

Physics-Informed model-based reinforcement learning for pandemic mitigation

Mitigating infectious diseases is a critical public health challenge. In recent work, the use of reinforcement learning (RL) in combination with epidemiological models was demonstrated to learn optimal mitigation strategies. RL techniques nonetheless suffer from sample inefficiency, which impedes their use for real-time decision support.

Model-Based RL (MBRL) offers a promising alternative, by learning a transition model from observed data. However, standard MBRL approaches typically struggle with capturing complex epidemiological dynamics, requiring extensive interactions with the environment to learn accurate models. To remedy this, integrating expert knowledge into the learning process can enhance model reliability and sample efficiency. Moreover, such an approach is necessary to move towards real-time decision support, where there is no room for trial-and-error.

Physics-Informed Neural Networks (PINNs) facilitate the integration of epidemic expertise into learning algorithms by embedding differential equations that govern disease spread. This ensures consistency with established epidemiological principles while providing flexibility to handle real-world uncertainties. Additionally, PINNs facilitate the estimation of unknown epidemic parameters, further increasing the model's interpretability and reliability.

This work explores a Physics-Informed MBRL framework that integrates prior epidemiological knowledge into policy optimization. By leveraging PINNs for both model learning and parameter estimation, we aim to enhance decision-making for pandemic mitigation, leading to more effective and robust intervention strategies. We conduct our initial

experiments in a SIHR compartment model where we consider social contact reductions as a preventive action.

Meritxell Brunet Guasch (University of Edinburgh)

Quantifying the evolution of metastati potential across cancer tissues

Cancers of different origins exhibit remarkable variation in the incidence of metastatic disease, yet the underlying causes of this heterogeneity remain largely unexplored. In particular, the evolution of traits that enable metastasis seeding is poorly understood. Here, we integrate multi-region tumor biopsies from paired primary and metastatic samples across different cancer types (n=32 colorectal, n=17 breast and n=15 pancreatic cancers) and SEER epidemiological data with a mathematical model of metastasis evolution. The model parameters include the rate at which the primary tissue evolves metastatic ability, and the rate at which metastatically-able cells disseminate to form metastases, which are learned from the data via Approximate Bayesian Computation (ABC). Our findings reveal differences across both primary tissues and metastases organs. In colorectal cancers, only a fraction of cancers will seed metastasis to the lymph nodes, but those that do acquire this ability early in tumor progression. In both colorectal and breast cancer, distant metastases are seeded by one or only few metastatically competent clones, suggesting the evolution of a 'special' trait necessary for metastasis. A similar pattern is observed in breast adenocarcinoma. Conversely, in pancreatic adenocarcinoma – a far more aggressive disease – most primary tumor regions are capable of distant metastasis, indicating an inherent metastatic proclivity of the founder cell. This study provides a unifying framework for understanding the evolution of metastasis, which may guide future, more systematic investigations into the potential molecular drivers of metastatic progression.

Margherita Bruno (CentraleSupélec-Université Paris-Saclay)

Inferring the acquisition age of JAK2-V617F mutation in MPN patients from phylogenetic data and an ABC-SMC model selection procedure

Myeloproliferative neoplasms (MPN) mainly arise from acquiring the JAK2-V617F mutation in hematopoietic stem cells years before symptoms appear. Estimating the acquisition age of this mutation is essential to understand MPN oncogenesis and enable early diagnosis.

Previous approaches by Van Egeren et al. (Cell Stem Cell, 2021) and Williams et al. (Nature, 2022) have been developed by collecting patient data structured as phylogenetic trees, establishing mathematical stochastic models of clonal development, and estimating model parameters using ABC. However, analyzing patients separately limits model generalizability.

Our objective was to propose a robust ABC framework for model selection and hierarchical Bayesian inference to estimate patient-specific acquisition ages and common biological parameters. Due to computational costs, we first optimized parameter estimation for each model and patient separately. We implemented an ABC-SMC strategy with Gaussian perturbation kernels, reducing variance at each step to improve convergence.

Then, we extended our method to model selection across multiple patients simultaneously. Only parameters that generate the closest simulations to the full dataset, measured by lineage-through-time distance, are retained in the posterior distributions while individual model-specific characteristics, like prior probability, are still preserved.

Our first individual-specific procedure revealed model-dependent variations in the estimated acquisition ages. Our hierarchical approach, performed with two patients, already gives information about the confidence of each model and more robust individual estimations. As a perspective, we will extend the number of models and patients for a more cohesive inference.

Alberto Cassese (University of Florence)

Bison - Bi-clustering of spatial omics data with feature selection

The advent of next-generation sequencing-based spatially resolved transcriptomics (SRT) techniques has reshaped genomic studies by enabling high-throughput gene expression profiling while preserving spatial and morphological context. Comprehending gene functions and interactions in different spatial domains is crucial, as it can enhance our comprehension of biological mechanisms, such as cancer-immune

interactions and cell differentiation in various regions. To achieve this, it is necessary to cluster tissue regions into distinct domains and identify genes with similar expression patterns within each domain, referred to as spatial-domain-marker genes. Existing methods for identifying these genes typically rely on a two-stage approach, which can lead to the phenomenon known as double-dipping. We propose a unified Bayesian latent block model that simultaneously detects a list of informative genes contributing to spatial domain identification while clustering these informative genes and spatial locations. The efficacy of our proposed method is validated through a series of simulation experiments, and its capability to identify spatial-domain-marker genes is demonstrated through applications to benchmark SRT datasets, including the mouse olfactory bulb ST dataset and breast cancer Visium dataset.

Andrea Corbetta (Human Technopole)

Genetics helps determine differences in response to statin treatment: analysis on short and long-term low-density lipoprotein cholesterol trajectories

Background: Understanding the genetic basis of lipid-lowering responses to statin therapy may provide valuable insights into personalized cardiovascular therapies. This study examines how genetic predisposition, as captured by polygenic scores (PGS) for low-density lipoprotein cholesterol (LDL-C), influences short and long-term changes in LDL-C levels after statin initiation.

Materials and Methods: From the FinnGen cohort, we extracted 11,343 individuals with LDL-C measurements within one year before and after statin treatment initiation (short-term) and 15,864 individuals with at least five years of statin treatment and LDL-C measurements (long-term), from which we defined trajectories and disentangled independent patterns using functional principal components. We tested the effect of polygenic scores (PGS) for LDL on absolute and relative reduction of LDL (short-term) and the constant and reduction pattern (long-term) and performed genome-wide association studies.

Results: We find that individuals in the top tertile of PGS have a greater absolute reduction (8.12[6.93-9.57] mg/dl) of LDL-C in the first year but a smaller relative reduction (1.81[0.06-2.99] %) compared with the bottom tertile. While individuals with high PGS are associated with higher five-year LDL-C measurements, PGS is not associated with long-term changes in LDL-C. The GWAS detects significant genome-wide signals for

relative reduction and longitudinal mean LDL-C, previously reported for LDL-C levels.

Conclusion: Short-term LDL-C reduction after statin initiation appears to have a genetic basis closely linked to LDL-C regulation. Long-term changes in LDL-C levels do not appear to be regulated by genetic factors, yet individuals with high PGS for LDL have higher LDL-C in the long term.

Diego di Bernardo (Telethon Institute of Genetics and Medicine)

Investigation of dynamic regulation of TFEB nuclear shuttling by microfluidics and quantitative modelling

Transcription Factor EB (TFEB) controls lysosomal biogenesis and autophagy in response to nutritional status and other stress factors. Although its regulation by nuclear translocation is known to involve a complex network of well-studied regulatory processes, the precise contribution of each of these mechanisms is unclear. Using microfluidics technology and real-time imaging coupled with mathematical modelling, we explored the dynamic regulation of TFEB under different conditions. Our model integrates literature-based and experimental observations and explains how different mechanisms interact to regulate TFEB activation. Furthermore, our work shows the power of quantitative modelling in helping to elucidate complex biological systems.

Andrew Golightly (University of Durham)

Accelerating Bayesian inference for stochastic epidemic models using incidence data

This work considers the case of performing Bayesian inference for stochastic epidemic compartment models, using incomplete time course data consisting of incidence counts that are either the number of new infections or removals in time intervals of fixed length. The most natural Markov jump process representation of the model is eschewed for reasons of computational efficiency, and replaced by a stochastic differential equation representation. This is further approximated to give a tractable Gaussian process, that is, the linear noise approximation (LNA). Unless the observation model linking the LNA to data is both linear and Gaussian, the observed data likelihood remains intractable. Unlike

previous approaches that use the LNA in this setting, two approaches for marginalising over the latent process are considered: a correlated pseudo-marginal method and analytic marginalisation via a Gaussian approximation of the noise model. These approaches are compared using synthetic data with the best performing method applied to real data consisting of removal incidence of Oak Processionary moth nests in Richmond Park, London.

Henrik Häggström (Chalmers University of Technology and University of Gothenburg)

Simulation-based inference for stochastic nonlinear mixed-effects models with applications in systems biology

We propose a novel methodology for Bayesian inference in hierarchical mixed-effects models. By building on our work [1], we construct a simulation-based inference (SBI) framework that is highly scalable, where amortized approximations to the likelihood and the parameters posterior are first obtained, and these are rapidly refined for each individual dataset, to ultimately approximate the parameters posterior across many individuals. Unlike the current state-of-art SBI methods, which use neural networks, our approximations are expressed via Gaussian mixture models, leading to easily trainable, parsimonious yet expressive surrogate models of both the likelihood function and the posterior distribution. The methodology is exemplified via stochastic differential equation mixed-effects models to describe translation kinetics after mRNA transfection, however the methodology is general and can accommodate other types of stochastic and deterministic models. We compare our approximate inference with exact pseudomarginal inference and show that our methodology is fast and competitive.

References:

[1] H. Häggström, P. Rodrigues, G. Oudoumanessah, F. Forbes and U. Picchini (2024). Fast, accurate and lightweight sequential simulation-based inference using Gaussian locally linear mappings, Transactions on Machine Learning Research, https://openreview.net/forum?id=Q0nzpRcwWn.

Ferdinando Insalata (Imperial College London)

Mutational signatures of deterministic and noise-induced evolutionary mechanisms

When two or more species compete, a typical problem is understanding what is the mechanism that could lead to one prevailing over the others.

While it is intuitive to posit that the prevailing species has a higher overall growth rate, for example a higher replication rate, noise-induced selection mechanisms have attracted increasing attention in recent years. Models showing stochastic selection are often counterintuitive, as they can present identical growth rates for all species, but are widely relevant, given the inherent randomness in biological systems. A crucial point is whether we can distinguish between these two classes of mechanisms in a typical experiment.

Here we compare the frequency distributions of randomly occurring neutral mutations in a spatially extended system where a species is expanding in a wave-like fashion. We find qualitative and quantitative signatures of these frequency distributions that discriminate between selection based on a replicative advantage (RA) and noise-induced selection, driven by differences in carrying capacity (SSD) or in baseline turnover rates. We find that standard statistical tests are able to detect these differences with practically feasible sample sizes.

Our findings are applicable to current debates in the field of evolutionary biology, as noise-induced selection has been repeatedly implied in the spread of altruistic traits.

Hetvi Jethwani and Kamal Sukandar (Imperial College London)

The role of rare somatic mtDNA mutations in proliferative tissue aging

Mitochondrial dysfunction is a hallmark of aging. Somatic mitochondrial DNA mutations contribute to mitochondrial dysfunction, and can manifest in time spans similar to human life. Previous work discovered a class of mtDNA mutations unique to a single cell in a sample of post-mitotic tissues, showing that these rare somatic mutations accumulate to high levels by mid/late life, and covary with genotypic hallmarks of aging. We build upon previous work and study the role played by rare somatic

mtDNA mutations (or low prevalence/LP mutations) within proliferative tissues by analyzing single-cell RNA sequencing data from 300+ samples across 10+ datasets. We discover that these LP mutations are the most abundant class of distinct mutations forming ~80% of the repertoire. We show that LP mutations accumulate with age across proliferative tissue types including PBMCs, lung, skin, and muscle; we model the accumulation by using a Kingman coalescence model. We use a Beta-Binomial model and infer that the mutation rate of synonymous and non-synonymous mutations is similar, suggesting that these LP mutations (otherwise invisible in bulk) evade selection. We show the presence of LP mutations robustly covaries with aging markers, highlighting that LP mutations may play a crucial role in aging.

Richard Kettle (University of Edinburgh)

Quantifying Immune Cell Subtypes and States under Cytokine Perturbations States

Characterizing cell heterogeneity is critical for medicine, as it informs on how different cells behave, interact and respond to treatments. Current approaches for inferring heterogeneity rely on projecting cells into a reduced dimensional transcriptomic space where they are clustered into discrete categories. This approach disallows cells to be labelled multiply, resulting in the definition of arbitrary boundaries, and thus cannot capture states occupied by cells of different types.

We introduce Stator (https://doi.org/10.1038/s44320-024-00074-1), a novel method that overcomes these limitations. Stator first estimates gene interactions at higher than pairwise ($3 \leq n \leq 7$) order, before performing hierarchical clustering of these gene interactions (not cells) based on their co-occurrence in single cells. This hierarchy of Stator cell states is then used to multiply label cells at fine resolution with regard to cell type, sub-type, and state (e.g., cell cycle phase, activation etc).

Here, I have applied Stator to The Immune Dictionary: a set of highly heterogeneous mouse cell populations that were perturbed, in vivo, by a raft of signalling protein treatments (cytokines) (https://doi.org/10.1038/s41586-023-06816-9). I first used Stator to evaluate cellular heterogeneity in a fully data driven manner, at multiple resolutions, without relying on traditional and potentially spurious cell labels. Second, I tested for associations between protein treatments and cell states not previously identified using whole-transcriptome clustering approaches.

Zhana Kuncheva (Bioxcelerate AI)

Can causal inference complete the multi-omics game?

The path from data to drugs is increasingly shaped by our ability to extract causal insights from vast, complex multi-omics datasets. In early-phase discovery, identifying and validating successful drug targets remains one of the highest-risk stages in the value chain. To improve the success rate of therapeutic development, we need tools that can integrate diverse data sources to generate interpretable and actionable evidence of causality.

Causal inference methods—such as fine-mapping, colocalisation, and Mendelian randomisation—are key to this effort, allowing the integration of genomic, transcriptomic, proteomic and many other omics data into structured, biologically meaningful models of disease aetiology. Our bioX Discovery Platform builds on top of these causal inference tools and data integration capabilities to meet the challenges faced by pharma and biotech. It combines in-house machine learning algorithms, scalable cloud engineering, and proprietary software to deliver actionable insights.

In this talk, we present a use case in Parkinson's disease, showcasing how our integrative platform—bioX Discovery—applies advanced causal inference methods to large-scale GWAS and molecular QTL datasets. Using our proprietary fine-mapping and summary statistic imputation pipelines, we integrate brain eQTLs and plasma pQTLs to identify molecular traits associated to Parkinson's risk. These relationships are embedded into a disease-focused knowledge graph, enriched through multi-trait colocalisation and gene-level phenotype mapping, ultimately revealing mechanistic sub-networks within a broader Parkinson's disease phenotype network.

Our approach illustrates the transformative potential of causal inference in multi-omics integration. By providing scalable, data-driven frameworks for hypothesis generation and validation, we aim to shift the paradigm of early-stage drug discovery—turning data complexity into biological clarity.

Anastasia Mantziou (University of Warwick)

Bayesian model-based clustering for populations of network data

There is increasing appetite for analysing populations of network data due to the fast-growing body of applications demanding such methods. While methods exist to provide readily interpretable summaries of heterogeneous network populations, these are often descriptive or ad hoc, lacking any formal justification. In contrast, principled analysis methods often provide results difficult to relate back to the applied problem of interest. Motivated by two complementary applied examples, we develop a Bayesian framework to appropriately model complex heterogeneous network populations, while also allowing analysts to gain insights from the data and make inferences most relevant to their needs. The first application involves a study in computer science measuring human movements across a university. The second analyses data from neuroscience investigating relationships between different regions of the brain. While both applications entail analysis of a heterogeneous population of networks, network sizes vary considerably. We focus on the problem of clustering the elements of a network population, where each cluster is characterised by a network representative. We take advantage of the Bayesian machinery to simultaneously infer the cluster membership, the representatives, and the community structure of the representatives, thus allowing intuitive inferences to be made. The implementation of our method on the human movement study reveals interesting movement patterns of individuals in clusters, readily characterised by their network representative. For the brain networks application, our model reveals a cluster of individuals with different network properties of particular interest in neuroscience. The performance of our method is additionally validated in extensive simulation studies.

Antonietta Mira (Universita' della Svizzera Italiana)

On the notion of data intrinsic dimension and its biomedical applications

Real-world datasets often exhibit a high degree of (possibly) non-linear correlations and constraints among their features. Consequently, despite residing in a high-dimensional embedding space, the data typically lie on a manifold with a much lower intrinsic dimension (ID), which—under the presence of noise—may depend on the scale at which the data are analyzed. This situation raises interesting questions: How many variables

or combinations thereof are necessary to describe a real-world dataset without significant information loss? What is the appropriate scale at which one should analyze and visualize the data? Although these two issues are often considered unrelated, they are in fact strongly entangled and can be addressed within a unified framework.

We introduce an approach in which the optimal number of variables and the optimal scale are determined self-consistently, recognizing and bypassing the scale at which the data are affected by noise. To this end, we estimate the data ID in an adaptive manner. Sometimes, within the same dataset, it is possible to identify more than one ID, meaning that different subsets of data points lie on manifolds with different IDs. Identifying these manifolds provides a clustering of the data; in many real-world applications, a simple topological feature such as the ID allows us to uncover a rich data structure and improves our insight into subsequent statistical analyses.

Examples of biomedical applications range from gene expression to protein folding, pandemic evolution, and all the way to fMRI data.

Yoav Ram (Tel Aviv University)

Quantifying Cultural Evolution Using Population Genetic Methods

Cultural evolution, much like genetic evolution, is influenced by both vertical transmission from parents to offspring and non-vertical processes, such as exchange, amalgamation, and innovation. We therefore adapted population-genetic computational methods to examine cultural transmission and variation among the Austronesian societies in the Pacific Ocean.

Due to the often noisy and incomplete nature of cultural data, we utilized Bayesian PCA to both impute missing values and denoise the data. We then used the posterior distributions from the PCA to develop a new method for estimating the number of significant principal components across four categories of cultural features: subsistence, social organization, religion, and cultural interaction.

Next, we applied archetypal analysis to each cultural class, to reduce dimensionality and cluster the data around extreme "archetypes," which we named based on their correlations with the original features. We then we constructed cultural phylogenetic networks and evaluated their "tree-likeness" to assess the relative verticality of cultural transmission

among the different cultural classes. Lastly, we developed a method for identifying cultural outliers and compared these outliers with known linguistic outliers.

This study demonstrates how modified population genetic methods can be effectively utilized to analyze cultural and anthropological data. It emphasizes the adaptability of mathematical and statistical tools from evolutionary biology in uncovering the mechanisms that contribute to cultural diversity.

Davide Risso (University of Padova, Italy)

Scalable and interpretable analysis of single-cell omics data

Single-cell and spatial omics techniques allow the quantification of gene and/or protein expression at the individual cell level, enabling the study of tissue cellular heterogeneity and expression dynamics. From an analytical point of view, dimensionality reduction, supervised classification, and clustering are the basis for identifying biological signals, cell types, and spatial domains. However, these steps are challenging because of the size and complexity of the data. Here, I will present a generalized matrix factorization (GMF) framework for scalable single-cell data analysis. I will show that many of the proposed approaches in the single-cell dimensionality reduction literature can be seen as special cases of this model.

Hélène Ruffieux (University of Cambridge)

A Bayesian functional factor model for high-dimensional molecular curves

The increasing availability of longitudinal measurements on large panels of gene products is set to improve our understanding of the molecular processes underlying disease risk and progression. While functional data analysis is an active area of research, methods for modelling complex multivariate functional dependencies remain limited.

Motivated by a COVID-19 study conducted in Cambridge, we propose a Bayesian approach for representing high-dimensional curves, combining latent factor modelling and functional principal component analysis (FPCA). This approach captures correlations across variables (e.g., biomarkers) and time, by positing that subsets of variables contribute to a

small number of FPCA expansions (e.g., representing latent disease processes) through variable-specific loadings. Subject variability is modelled using a small number of functional principal components, each characterised by a smoothly varying temporal function. We develop a variational inference algorithm, with analytical updates, that couples efficiency and principled parameter uncertainty quantification, and we propose strategies to learn the number of factors and FPCA components involved in each factor's expansion.

Extensive numerical experiments illustrate the ability of the approach to (i) accurately estimate variable-specific loadings, FPCA latent functions and subject-specific component scores, and (ii) scale to high-dimensional datasets (e.g., with panels of 20,000 genes measured longitudinally for a few hundred subjects). In the COVID-19 study, our framework helps disentangle disease heterogeneity and clarifies which biomarkers coordinate over time, pointing to key biological pathways, towards targeted interventions and personalised treatments.

Jaromir Sant (Universita di Torino)

Inference of genome-wide genealogical relationships between ancient and modern individuals

Recent advances in sequencing technologies have greatly increased the number of available ancient DNA (aDNA) human genomes. Combining aDNA with large modern biobanks can provide new insights into the relationships between ancient and modern individuals, and reveal details of human evolutionary history. We developed a method, ThreaDNA, to study such genome-wide genealogical relationships by threading unphased ancient diploid genomes into an ancestral recombination graph (ARG) inferred from modern data. Our approach relies on an efficient haplotype matching algorithm based on the Li-Stephens model and the positional Burrows-Wheeler transform, coupled with approximate likelihood-based inference of coalescence times. We show that ThreaDNA is accurate when applied to simulated unphased aDNA samples, and that it scales linearly in the number of ancient individuals and sub-linearly in the number of modern individuals. We then apply our method to the diverse imputed ancient individuals present within the Allen Ancient DNA Resource dataset together with a previously inferred genome-wide ARG comprising 487,409 modern UK Biobank samples. By jointly modelling both modern and ancient individuals, we are able to recover several well-documented demographic shifts within the UK in the past ~2000

years. These findings underscore the utility of constructing large-scale genealogies that include both ancient and modern individuals to gain insights into historical migrations, population dynamics, and recent natural selection within the UK.

Florian Schunck (Osnabrück University)

Modeling the Bigger Picture: Hierarchical Approaches for Integrating Diverse Toxicological Data in Mechanistic Models

Over 50,000 chemicals in commerce today, presenting significant challenges for human and environmental risk assessment that cannot be addressed solely through experimental testing, which also burdens animal subjects. But alternatives exist, molecular measurements are increasingly available, enhancing the diverse toxicological data. Integrating these data into mechanistic models can elucidate the processes linking exposure to effects. However, significant obstacles remain in integrating qualitatively and quantitatively different datasets from different experiments into mechanistic models. 1) Data sparsity due to incomplete datasets and 2) Data variability arising from experimental and biological noise.

This study addresses these challenges using Bayesian hierarchical modeling, on a diverse dataset, applying a molecular TKTD model to time-resolved internal concentration, gene expression, and survival data from exposure to three chemicals across 42 zebrafish embryo experiments. We estimate model parameters and uncertainty, coupling a auto-differentiable ODE solver with gradient based stochastic variational inference and hierarchically modeling the true external concentration as a function of the nominal external concentration adjusted by an experiment-specific deviation factor.

Early results show that, on average, the modeled external concentrations deviated by a factor of 2.1 (94% HDI = [1.9, 2.3]) with a log standard deviation of 0.81 from the nominal concentrations. The hierarchical approach allowed for the inclusion of integration of all experimental data regardless of experimental discrepancies between nominal concentration and measured effects, demonstrating its potential in enhancing predictive risk assessment through improved integration of existing and new data.

Jonasz Słomka (ETH Zurich)

How encounters at the microscale prime microbial interactions

Microbial interactions often critically depend on the rate of physical cell–cell or cell–resource encounters. In a liquid environment, many prominent examples include encounters among phytoplankton in the ocean that lead to the formation of marine snow, the formation of living aggregates by cyanobacteria, bacterial chemotaxis towards leaky phytoplankton, and horizontal gene transfer between bacteria. Microscale encounters are nearly always quantified as encounters between inanimate spheres, borrowing from the physics of gases, coagulating colloids, and rain formation. However, these classical approaches often fail to account for important traits of microorganisms, such as cell elongation, motility, or gradient sensing. Even more importantly, experimental assays typically do not control cell–cell encounters. In my talk, I will outline how more realistic models of encounters at the microscale can contribute to our understanding of fundamental ecological processes controlled by microbes, from active aggregation through chemotaxis to gene exchanges. I will close by presenting our recent experimental evidence that encounters driven by fluid shear strongly control the rates of horizontal gene transfer between bacteria.

Nicholas Steyn (University of Oxford)

SMC methods for epidemic renewal models

Renewal models are widely employed in statistical epidemiology as semi-mechanistic models of disease transmission. While primarily used for estimating the instantaneous reproduction number, they can also be used for generating projections, estimating elimination probabilities, modelling the effect of interventions, and more. Many methods for fitting these models exist, oftentimes specifically constructed to account for biases in the epidemiological data the models are fit to.

We present the use of sequential Monte Carlo (SMC) methods as a simple framework for fitting a wide range of epidemic renewal models. We focus on the flexibility of these methods and their ability to fit models that account for multiple biases simultaneously, thus unifying many existing approaches in the literature. In this talk, we demonstrate how these methods can be used for both inference and forecasting and highlight how this approach can unify many existing models.

While we focus on epidemiological applications, these methods work with a wide range of sequential hidden-state models.

A companion website SMC and epidemic renewal models contains many worked examples and self-contained code. We emphasise that, other than a Julia installation, these methods require no external software packages, enabling researchers to set-up their own model in minutes.

Ben Swallow (University of St Andrew's)

Spatio-temporal Gaussian processes on GPUs

Gaussian process are a widely-used statistical tool for conducting non-parametric inference, particularly with spatio-temporally indexed data. Whilst there are many computational packages available to train and predict on observed data, the inherent challenge with GPs is their computational demand for realistic data sizes. In this talk I will discuss the embedding of spatio-temporal GPs within a Bayesian Markov chain Monte Carlo framework, using software packages that can interact with tensorflow probability. The approach enables a relatively straightforward implementation on GPUs with significant speed increase over standard CPU calculations, whilst still allowing asymptotically exact inference and uncertainty quantification. The inference pipeline is applied to weekly UKHSA data on tuberculosis in the East and West Midlands regions of England over a period of two years.

Simone Tiberi (University of Bologna)

IsoBayes: a Bayesian approach for single-isoform proteomics inference

Background:

Inferring protein isoforms is a crucial step in biomedical research. At present, proteins are indirectly measured via peptides. However, this processes is noisy because most peptides are shared across multiple proteins; furthermore, peptides may also be erroneously detected. As a consequence, inference is typically abstracted at the gene-level.

Results:

Here, we describe IsoBayes, a novel Bayesian statistical method for protein inference at the isoform level. To this aim, we designed a two-layer latent variable approach where: i) we sample if a peptide has been correctly detected, and, ii) we allocate the abundance of such selected peptides across the protein isoform(s) they are compatible with. Furthermore, in order to enhance the information available, we integrate proteomics and transcriptomics data. This allows us to: i) infer the presence/absence of each protein isoform (via a posterior probability), ii) estimate its abundance (and credible interval).

In order to validate our approach, we designed comprehensive benchmarks, based on simulated and real datasets, where IsoBayes displays good sensitivity and specificity when detecting proteins (Figure attached), and where its estimated abundances highly correlate with the ground truth.

Availability:

Our method is flexible and works with peptide identifications obtained by any proteomics tool, is distributed open-access as a Bioconductor R package and is accompanied an example usage vignette.