BIOINFERENCE

Book of Abstracts

Posters (A-Z)

Lorenzo Alonso Requena (Imperial College London)

Longitudinal Multi-Omic Data Integration for the Study Relapse in an MCF7 model of ER+ Breast Cancer

Relapse is a major problem in oestrogen receptor positive breast cancer. The rate of relapse remains constant for at least 15 years after endocrine treatment discontinuation. This points towards a subpopulation of cancer cells entering dormancy, avoiding treatment, and then stochastically awakening. Genomic, transcriptomic and epigenomic profiles of bulk samples from an MCF7 cell line model of this temporal process have been generated and analysed separately. Our aim in this work is to integrate the datasets in a multi-omic analysis accounting for their temporal structure to gain a holistic understanding of the underlying biological mechanisms. To address the issue of heterogeneous sampling regimes between datasets, we applied data imputation strategies for harmonisation. We performed regression using linear splines with mixed effects and gaussian processes due to their ability to adaptively fit non-linear trajectories. The former approached has shown to positively affect visualisation in our dataset, and to capture features that change with time. We are currently exploring computational tools for gaussian process regression. Using the modelled temporal trajectories, we performed data visualisation and clustering. Initial k-means clustering revealed similarities and differences between datasets, which will be followed up by exploring multi-omic network analysis. A potential co-localisation of H3K4me3 and H4K20me3 histone modifications during awakening was detected.

Tarek Alrefae (University of Oxford)

Inferring the Force of Infection via Serocatalytic Modelling

Serological data and related seroepidemiological models have been used to uncover complex, dynamic trends in the force of infection (FoI) of various diseases. These have in turn been used to better understand the burden of infection(s) in populations and to inform public health policymaking. We will present the progress we have made toward the development and implementation of a serocatalytic model to help infer the FoI of Enteroviruses D68 (EV68) and A71, as well as Coxsackievirus A6 in England using a set of three cross-sectional titration assays. The talk or poster will focus on the methodological aspects of our investigation and their application to the aforementioned viruses. This extends previous work which investigated the changes in transmission of EV68 in England using seroprevalence data via a set of mechanistic models which estimated individual antibody concentration levels with (considerable and likely important) uncertainty. The incorporation of this uncertainty into our

serocatalytic modelling framework then allows us to infer the FoI and serodynamics with reasonable uncertainty. This framework represents a more robust inference framework that can be generally applied to many other viruses and serological data sets.

Inga Huld Ármann (Imperial College London)

Investigating the role of metabolites in prostate cancer

Prostate cancer (PrCa) is the most common cancer in men in the UK with 52,000 cases and 12,000 deaths per year. Leveraging data obtained from accessible screening methods, such as blood tests, would enhance early diagnosis and improve patient care. This work analyses clinical and metabolomics data from two case-control studies on PrCa to identify key biomarkers and uncover underlying data structures. High correlation and complex dependencies pose challenges, addressed through penalised regression methods for variable selection. Sparse group methods further enhance performance by incorporating known metabolomics group structures. Additionally, penalised regression integrates clinical risk factors, such as BRCA mutations. Clustering techniques are applied to explore potential PrCa subtypes, particularly those linked to disease aggressiveness. Finally, dimensionality reduction techniques facilitate data visualisation, providing deeper insights into metabolomics patterns in PrCa.

Nadav Ben Nun (Tel Aviv University)

Inference from Highly Variable Experimental Replicates with the Collective Posterior Distribution

Simulation-based Bayesian inference methods estimate a posterior distribution over model parameters, representing the probability of parameters θ given an observation x, p(θ|x), when the likelihood function p(x|θ) is intractable. However, biological experiments often yield collections of observations assumed to originate from the same underlying distribution, offering an opportunity for collective inference. To address this, we propose the collective posterior distribution: p(θ|x_1,...,x_n), a formulation that conditions model parameters θ on an entire collection of observations x_1,...,x_n using an axiomatic approach. We implemented the collective posterior distribution and demonstrated its application for inference from the Wright-Fisher model using approximate Bayesian computation (ABC) and neural network simulation-based inference. The collective posterior framework is adaptable to any simulation-based Bayesian inference pipeline, enabling the transition from individual posteriors to a unified posterior conditioned on all observations. Our

results demonstrate that inference with the collective posterior achieves higher accuracy than method-specific alternatives, particularly for evolutionary models, datasets governed by hierarchical processes, and empirical observations. These findings highlight its potential to improve inference quality in highly variable biological experiments.

Lucie Brolon (Université Paris Cité)

Modeling T-cell and dendritic cell communication

When an infectious agent enters the body, it is detected by dendritic cells (DCs), which emit biochemical signals to initiate an immune response. These signals induce the differentiation of T cells (TCs), which subsequently release signals to combat the infectious agent.

Modeling the relationship between the signals emitted by DCs (input) and those of TCs (output) is challenging. DCs can produce diverse signals, multiple DCs may emit identical or distinct signals simultaneously, and TCs process combinations of these signals. A single DC signal can have varying effects on TC signals depending on the simultaneous signals emitted by other DCs—referred to as the context. Each signal must therefore be analyzed within its context.

To address this challenge, we selected two algorithms: a regression tree (PILOT[1]) and a random forest (SIRUS[2]). Each algorithm has its own specific strengths in order to address our challenge. PILOT detects contexts in which there are linear relationships between inputs and outputs, while SIRUS generates a set of stable contexts. We therefore combined these two algorithms to develop an algorithm that is both stable and relevant. In this presentation, I will introduce this hybrid algorithm and compare it to other approaches (MOB[3], MARS[4], M5[5]) using a simulation study.

References:

[1] J. Raymaekers et al. “Fast linear model trees by PILOT”. In:Mach Learn113 (2024),6561–6610.

[2] C. Bénard et al. “SIRUS: Stable and Interpretable RUle Set for classification”. In:Electron. J. Statist.15(1) (2021), pp. 427–505.

[3] A. Zeileis, T. Hothorn, and K. Hornik. “Journal of Computational and Graphical Statistics”. In:Cell17(2) (2008), pp. 492–514.

[4] J. Friedman. “Multivariate Adaptive Regression Splines”. In:Ann. Statist.19 (1991),pp. 1–67.

[5] J.R Quinlan. “Learning with continuous classes”. In:Proceedings Australian Joint Conference on Artificial Intelligence(1992), pp. 343–348.

Giulia Capitoli (University of Milano-Bicocca)

Spatially informed sparse Gaussian Graphical Mixture Model to detect latent patterns in mass spectrometry imaging

Mass spectrometry is a class of imaging techniques that measures molecular abundance in tissue samples at cellular resolution, while preserving the spatial structure of the tissue. In particular, mass spectrometry imaging (MALDI-MSI) has the capability to differentiate regions that are indistinguishable to pathologists at the microscopic level.

A central goal in mass spectrometry data analysis is to identify molecules with similar functions within the analyzed biological system, enabling a better understanding of abnormal molecular mechanisms. Our aim is to identify relevant biomolecules associated with cancer cells and the tumor microenvironment, thereby expanding biological knowledge.

In this study, we propose a Gaussian Graphical Mixture Model (GGMM) to address unobserved heterogeneity and segment tissue sections into regions based on distinct molecular profiles. Specifically, we aim to identify groups of molecules with similar activation patterns, investigate their spatial mapping within renal cancer tissue samples, and discover clusters of molecules whose activation is linked to specific biological mechanisms.

To model this heterogeneity, we reconstruct underlying molecular graphs from the data using sparsity constraints and incorporate spatial dependencies between neighboring pixels. To account for the spatial nature of the dataset, we utilize Hidden Markov Random Fields, ensuring that the spatial structure is effectively captured.

Arianna Ceccarelli (University of Oxford)

A Bayesian inference framework to calibrate one-dimensional velocity-jump models for single-agent motion using discrete-time noisy data

Advances in experimental techniques allow the collection of high-resolution spatio-temporal data that track individual motile entities over time. These tracking data motivate the use of mathematical models to characterise the motion observed. In this paper, we aim to describe the solutions of velocity-jump models for single-agent motion in one spatial dimension, characterised by successive Markovian transitions within a finite network of n states, each with a specified velocity and a fixed rate of switching to every other state. In particular, we focus on obtaining the

solutions of the model subject to discrete-time noisy observations, with no direct access to the agent state. The lack of direct observation of the hidden state makes the problem of finding the exact distributions generally intractable. Therefore, we derive a series of approximations for the data distributions. We verify the accuracy of these approximations by comparing them to the empirical distributions generated through simulations of four example model structures. These comparisons confirm that the approximations are accurate given sufficiently infrequent state switching relative to the imaging frequency. The approximate distributions computed can be used to obtain fast forwards predictions, to give guidelines on experimental design, and as likelihoods for inference and model selection.

Alex Cecchetto (Università degli studi di Padova)

Generalized Matrix Factorization: A Flexible Framework for Spatially-Resolved Omics Data

As omics technologies continue to evolve, the demand for statistical methods to characterize gene expression at high resolution is increasing. Single-cell data, with their high dimensionality and complexity, offer opportunities to uncover new biological insights.

Matrix factorization techniques are a powerful tool for analyzing high-dimensional omics data, which often present complex noise structures. Generalized Matrix Factorization (GMF) models, which extend classical approaches to support a broader range of data types (e.g., those from the exponential family), are particularly well-suited for high-dimensional datasets and have been successfully applied to single-cell data.

The advent of spatially-resolved transcriptomics technologies offers new opportunities to explore the structure of complex, intact tissues. This presents an opportunity to integrate spatial information into transcriptomic analyses, allowing for deeper insights into the structure and function of complex tissues. Moreover, the integration of multi-sample data remains a critical area for extending these approaches to support more comprehensive and comparative studies.

Here we explore the application of GMF models to high-dimensional omics data, emphasizing their flexibility in addressing noise and their potential to incorporate spatial and multi-sample components. By leveraging these advancements, this work aims to contribute to the development of robust tools for extracting meaningful biological insights from complex datasets.

Holly Chambers (Imperial College London)

Benchmarking Causal Discovery Methods for Partially Observed Biochemical Kinetics

Systems of intracellular biochemical reactions are complex, often involving components that cannot be directly measured. Representing these systems as networks, with nodes representing biochemical species and edges their reactions, helps quantitatively characterize their function and effects of dysregulation. Causal discovery methods can uncover functional interactions within these networks from purely observational data, detecting hidden effects from partial observations. These effects appear as common causes of observed variables, or through time-lagged effects from intermediate causes.

We benchmark the causal discovery method temporal Multivariate Information-based Inductive Causation (tMIIC) alongside other state-of-the-art tools, for time series data from biochemical kinetic models. Our results demonstrate tMIIC's high recall in identifying interactions within toy reaction networks. By selectively omitting data, we consider both latent confounders (the standard choice for benchmarking these methods) and unobserved species participating in reactions. tMIIC detects latent confounders using bidirected edges, and unobserved species through time-delayed edges, locating hidden effects and estimating their typical timescales. Finally, we extend these benchmarks to reconstruct an experimentally calibrated model of the epidermal growth factor receptor signalling network – a system frequently dysregulated in cancer.

Altogether, our work showcases the feasibility and usefulness of causal discovery methods like tMIIC for data-driven mathematical modelling of biochemical reactions.

Valerii Chirkov (Humboldt University of Berlin)

Dynamics of Spatio-Temporal Clusters in Social Foraging

Human foraging often occurs in social contexts, yet the factors driving aggregation versus solitary behavior remain unclear. This study introduces a novel application of spatio-temporal clustering methods to analyze human foraging dynamics, using data from ice-fishing competitions held over two years (2022–2023) on ten lakes in Eastern Finland. Competitors, tracked via GPS and head cameras, aimed to catch the highest amount of fish, frequently forming large clusters that persisted for extended periods.

Clusters were identified using spatio-temporal DBSCAN applied to participants' GPS tracks. Next, we compared the success of individuals within clusters to those outside, and studied how cluster density, catch rate, centrality and size impacted the rise and collapse of clusters. Finally, we examined the spatio-temporal dynamics of individuals' movements within clusters, focusing on the relationship between their position within the cluster and their foraging success.

Denser clusters with higher catch rates persisted longer and shrank more slowly, while larger clusters were less stable. Cluster formation appears to be driven by random joining but shaped by adaptive, success-driven leaving behavior. Within clusters, less successful individuals tended to move toward areas of higher expected success, tracking their neighbors' performance. These findings clarify the role of individual behavior in shaping collective foraging patterns and provide a basis for exploring similar dynamics in other contexts.

Emma Davis (University of Warwick)

Modelling the role of human migration in disease elimination and resurgence

The Neglected Tropical Disease lymphatic filariasis (LF) in Togo has been classified as eliminated as a public health problem - meaning disease levels have been brought sufficiently low that the disease is expected to die out without further intervention. However, complex human movement patterns, including high levels of cross-border migration, pose a risk of disease resurgence where neighbouring countries still have high disease prevalence. We developed a meta-population model representing Northern Togo, which shares borders with Burkina Faso, Benin and Ghana, using a combination of a radiation model to represent short-term migration and travelling nomadic populations, to investigate the role of human movement at this critical stage of elimination. We demonstrate that, under certain conditions, cross-border movement has the potential to undermine elimination efforts, resulting in resurgence back to pre-intervention disease levels. This highlights a need for continued surveillance, especially in migratory groups, to allow for targeted interventions.

Bowen Fang (University of Warwick)

Numerical Splitting Methods for SDEs: Simulation and Inference

Many real-world biological phenomena, such as population dynamics, neuronal activity, and ecological systems, are modeled using stochastic differential equations (SDEs) with multiplicative noise. Important examples include the Jacobi (Wright-Fisher) processes for genetic drift and neuronal models, as well as the broader Pearson diffusion class, the stochastic Ginzburg-Landau equation, and the stochastic Verhulst equation. %, commonly used in genetics, neuroscience, and ecology.

However, exact simulation schemes for these models are often unavailable or computationally prohibitive.

In this work, we propose a general numerical splitting method for SDEs with locally Lipschitz drift and Hölder continuous diffusion coefficients. Specifically, we decompose the original equation into tractable subequations and apply Lie-Trotter and Strang compositions to recover the full solution. Our approach outperforms traditional stochastic Taylor expansion methods, such as Euler-Maruyama, in both order of convergence and property preservation. The proposed method ensures boundary preservation for SDEs with constrained state spaces (e.g., Wright-Fisher diffusion) and improves empirical distribution convergence to invariant measures, allowing for more accurate and robust simulations.

Beyond simulation, these splitting schemes admit tractable transition densities, enabling parameter inference via pseudo-maximum likelihood estimation and Bayesian approaches, providing a practical framework for learning parameters of interest in complex biological systems.

Alicia Gill (University of Oxford)

Bayesian Inference of Reproduction Number from Epidemic and Genomic Data using Particle MCMC Methods

Typically, the time-varying reproduction number is inferred using only epidemic data, such as prevalence per day. However, prevalence data is often noisy and partially observed, and it can be difficult to identify whether you have observed many cases of a small epidemic or few cases of a large epidemic. Genomic data is therefore increasingly being used to understand infectious disease epidemiology, and inference methods incorporating both genomic and epidemiological information are an active area of research. We use particle Markov chain Monte Carlo methods to infer parameters of the epidemic using both a dated phylogeny and partial prevalence data to improve inference compared with using only one source of information. To do this, we have implemented a sequential Monte Carlo algorithm to infer the latent unobserved epidemic, which is then used to infer the reproduction number as it varies through time. We

then analyse the performance of this approach using simulated data. Finally we present case studies applying the method to real datasets.

Isaac Hayden (Imperial College London)

Estimating Hypertension Incidence in Uganda: A Bayesian Framework for Incomplete Blood Pressure Data

We present an extended Bayesian modeling framework for estimating disease incidence when binary disease status is incompletely observed for some subjects, leveraging information across individuals to improve inference. Using four rounds of health survey data (2016–2024) from the Rakai Community Cohort Study in Uganda, we apply this approach to identify risk factors for incident hypertension. Historic blood pressure data in the study exhibit both a terminal digit rounding effect due to manual measurement, and a limited number of readings for many individuals. To address these challenges, we model joint probability distributions for systolic and diastolic blood pressure, allowing estimation of hypertension probability for rounded values. A linear regression model further estimates an error correction term for individuals with sparse data. By allowing the incorporation of multiple crude, correlated measurements rather than restricting to complete data cases, this approach maximises data utilisation and enables broader trend analysis, particularly in resource-limited settings. We calculate incidence rate ratios, stratified by patient covariates, to identify key risk factors and compare the strength of association between hypertension and anthropometric indicators of obesity. This framework is widely applicable to many non-communicable diseases diagnosed through routine healthcare measurements, and provides critical insights to inform hypertension surveillance and management in Sub-Saharan Africa and beyond.

Shu Huang (University of Warwick)

Inference for Diffusion Processes via Controlled Sequential Monte Carlo and Splitting Schemes

We introduce an inferential framework for a wide class of semi-linear stochastic differential equations (SDEs) with constant diffusion coefficient, and drift satisfying a global one-sided Lipschitz condition with at most polynomial growth. Recent work has shown that mean-square converging numerical splitting schemes can preserve critical properties of such types of SDEs, give rise to explicit pseudolikelihoods, and hence allow for parameter inference for fully observed processes. Here, under several observation regimes such as fully observed processes and partially observed processes with/without noise, we represent the implied

pseudolikelihood as the normalising constant of a Feynman--Kac flow, allowing its efficient estimation by the controlled sequential Monte Carlo algorithm. We then adapt likelihood-based methods to exploit this pseudolikelihood for inference. The self-contained strategy developed herein allows us to obtain good results across a range of problems without recourse to more complex time discretisation schemes which typically require considerable application-specific efforts.

Josef Janák (University of Pavia)

Parameter estimation in an SPDE model for cell repolarisation

As a concrete setting where stochastic partial differential equations (SPDEs) are able to model real phenomena, we propose a stochastic Meinhardt model for cell repolarization and study how parameter estimation techniques developed for simple linear SPDE models apply in this situation. We establish the existence of mild SPDE solutions, and we investigate the impact of the driving noise process on pattern formation in the solution. We then pursue estimation of the diffusion term and show asymptotic normality for our estimator as the space resolution becomes finer. We demonstrate the performance of the model in numerical and real data experiments.

Petar Jovanovski (Chalmers University of Technology)

Simulation-based inference using splitting integrators for partially observed diffusions in chemical reaction networks

We address the problem of parameter inference for chemical reaction networks described by the chemical Langevin equation, a stochastic differential equation (SDE) representation of the dynamics of the chemical species. This is particularly challenging for two main reasons. First, the (multi-dimensional) SDEs cannot be explicitly solved, requiring the development of advanced numerical schemes for their approximation and simulation. Second, not all components of the SDEs are observed, as the available discrete-time data are typically incomplete and subject to error. Our first contribution is developing a novel mean-squared numerical method for conditionally Cox–Ingersoll–Ross-type SDEs (SDEs with linear drift and square root diffusion component) that preserves structural properties such as oscillations, state space and invariant distributions, differently from the Euler-Maruyama scheme. This enables accurate inference with larger integration time steps. Our second contribution is an extension of an Approximate Bayesian Computation Sequential Monte Carlo algorithm to multidimensional SDEs, incorporating data-conditional simulation and sequential learning of summary statistics. We validate our

approach on models such as the stochastic Repressilator, Lotka-Volterra, and Two-pool systems, demonstrating its effectiveness for Bayesian inference in chemical reaction networks.

Endri Mjeku (Imperial College London)

Stochastic Modelling of mtDNA Replication Dynamics

Our cells contain up to thousands of copies of mitochondrial DNA (mtDNA), which are all dynamically replicating and degrading, even in post-mitotic cells. This dynamical process plays a central role in mutation spread; a cell with one mutant mtDNA can quickly become a cell with many mutant mtDNAs, simply through multiple mutant mtDNA replication events. This, in turn, can have significant health impacts. Despite this importance, the specific birth-death dynamics of mtDNA remains poorly understood. Using novel data collected from human fibroblast cells, we build a collection of stochastic birth-death models, using ABC (Approximate Bayesian Computation) to fit and select between them. In doing so, we find strong evidence of the existence of a small proportion of replicative mtDNA driving copy number, and we compute posterior estimates of multiple biological quantities of interest, including the mtDNA replication rate, proportion of replicative mtDNA in a cell, and dwell time in each subpopulation. Moreover, using neural granger causality, a nonlinear generalisation of granger causality, on time series data, we uncover the causal relationship between three fundamental dynamical quantities of interest - mtDNA number, total mitochondrial volume, and cell volume. We confirm this uncovered causal graph with linear granger causality, as well as conditional independence testing on static data.

Tom Morrish (University of Warwick)

Model-Informed Classification with Path Signatures: Predicting Bacterial Health

Inspired by collaborative work with a biomedical company, we explore theoretical and applied aspects of a model-informed classification framework. This approach leverages path signatures as linear features and employs a surrogate model to approximate the effects of a discretely observed, hidden real-world system. Building on previous work in Stratford et al. (2019), our objective is to create a classification scheme that is efficient, robust, and physically interpretable, while also accommodating complex, multivariate, and rough signals, such as those generated by fractional Brownian motion. A significant aspect of this work involves the study of two key distances: between the classification boundaries and the simulated path space, and between the simulation

model and the real system. Understanding these distances allow us to control and communicate the method's robustness and accuracy, key factors for biomedical applications.

Ella Orme (Imperial College London)

Non-negative matrix tri-factorisation for multi-view biclustering

Due to advancements in technology, the collection of large volumes of data is now feasible both practically and fiscally. Researchers conducting clinical studies can now obtain multiple omics dataset, such as genomics, proteomics, etc, on participating individuals. Integrating these omics, rather than utilising them individually, allows for a reduction of the effect of noise and the strengthening of signals seen weakly in the views.

Studying these datasets can be invaluable in the identification of disease subtypes. However, not all features are relevant to describe the patterns for all individuals. Multi-omics biclustering aims to simultaneously cluster both rows and columns, discovering clusters of individuals as well as their identifying omic-specific features.

This work will present the problem of multi-view biclustering and a novel approach based on matrix factorisation, Restrictive Multi-View Non-Negative Matrix Tri-Factorisation (ResNMTF). Demonstrated through extensive experiments on both synthetic and real datasets, including a single-cell dataset, ResNMTF successfully identifies both overlapping and non-exhaustive biclusters, without pre-existing knowledge of the number of biclusters present, and is able to incorporate any combination of shared dimensions across views. Further, to address the lack of a suitable bicluster-specific intrinsic measure, the popular silhouette score is extended to the bisilhouette score. The bisilhouette score is demonstrated to align well with known extrinsic measures, and proves useful as a tool for hyper-parameter tuning as well as visualisation.

Berk Tan Perçin (CNR)

A stochastic model to simulate pest dynamics

Population dynamics models are a useful tool for optimizing pest control strategies. Unlike overly simplistic deterministic models, stochastic models can better capture the reality of the dynamics because they are able to account for variability due to various factors. In this project, a system of Kolmogorov equations will be presented to model the dynamics of the pest Cacopsylla pyricola. It is a stage-structured phenological model

consisting of immature and adult stages. Each stage is characterized by a development rate function and the adult reproductive stage also by a fecundity rate function. The dynamics are simulated with numerical methods and compared with the abundance data of each stage collected in some pear crops in the Emilia Romagna region. The main objective of this project is to obtain a prediction of the dynamics of the pest to enable farmers to intervene in advance to control the invasion.

Finlay Plumb (University of Warwick)

Signature Methods for Sleep Stage Classification

In sleep research, there exists great interest in automating sleep stage classification as a means of objectively assessing sleep quality and diagnosing sleep disorders. This problem is typically addressed with deep learning models, which, despite their strong predictive performance, are computationally expensive and lack interpretability. To address these limitations, we propose a novel approach that combines random projections with iterated integrals of a path (signature transform) to construct expressive and structured feature representations of Polysomnography (PSG) time series. As we are considering multivariate biomedical signals, the truncated signature provides a principled mathematical framework for capturing the temporal evolution of these signals, while random projections have shown an ability to enhance the discriminative power of the extracted features. These representations are then used within an ensemble learning model, providing a robust and computationally efficient alternative to deep learning. Evaluating our method on the Sleep-EDF dataset, we demonstrate that it outperforms deep learning models while maintaining a greater sense of interpretability.

Alicia Quirós (Universidad de León)

Bayesian variable selection with missing data: an application to cardiology

Variable selection in regression models with missing data is crucial in medical research. Existing methods often discard incomplete cases, leading to biased estimates and reduced power. We propose a novel approach combining objective Bayesian variable selection with multiple imputation using readily available R packages. This approach utilizes all available data, avoids subjective user input, and provides a comprehensive assessment of variable importance. We demonstrate its effectiveness in identifying myocardial damage factors in a real dataset with missing data in 63% of the cases and show results in simulation studies. Compared to existing methods, our approach shows a superior

performance in terms of false discovery rate and scalability to high-dimensional problems. This innovative methodology provides a valuable tool for researchers dealing with missing data in medical research.

Dario Righelli (University of Naples "Federico II")

SpaceTrooper, an R package for the preprocessing and quality control of imaging-based spatial transcriptomics data

Several computational pipelines exist for preprocessing and quality control of spatially resolved omics data, but these often adapt single-cell RNA-seq methods with limited consideration of geospatial features. While sequencing-based platforms may suffice with this approach, imaging-based spatial transcriptomics require bespoke methods, particularly for segmentation. Segmentation defines cell boundaries as spatial entities for summarizing transcripts into cell-level count matrices, yet segmentation errors are predominantly evaluated through visual inspection. Simple, quantitative metrics for assessing cell boundary quality remain scarce.

To address these challenges, we present SpaceTrooper, an R package designed for the preprocessing and quality control of imaging-based spatial transcriptomic data. SpaceTrooper leverages Bioconductor data structures and geospatial R packages (e.g., sf, terra) to generate cell geometries from image and shape files in formats such as TIFF, HDF5, and parquet. It integrates a generalized linear model to evaluate probe expression alongside morphological features like cell area, aspect ratio, and unrealistic cell polygons. By combining these metrics, SpaceTrooper effectively flags or removes low-quality cells, addressing segmentation errors and boundary artifacts inherent in spatial transcriptomics.

Megan Ruffle (University of Bristol)

Benchmarking Probabilistic Programming Languages: Stan vs. NumPyro in Infectious Disease Modelling

Infectious disease modelling plays a crucial role in understanding epidemic and pandemic dynamics, guiding public health interventions, and informing policy decisions. Probabilistic Programming Languages (PPLs) provide a flexible framework for specifying complex statistical models, enabling Bayesian inference to estimate key epidemiological parameters and effectively integrate uncertainty into predictions of disease spread. Epidemiological research that utilises PPLs is growing,

demonstrating their capabilities, but systematic comparisons of model performance across different PPLs remain limited.

In this study, we evaluate the performance of two widely used PPLs, Stan and NumPyro, across three common models: a normal distribution model, an SIR model, and a Hawkes process model. We assess their computational efficiency and statistical accuracy using metrics such as runtime, posterior diagnostics (e.g., R-hat and effective sample size), and model-specific predictive performance measures.

When working with simpler models, preliminary results suggest that Stan and NumPyro perform similarly. However, as model complexity increases, their differences become more apparent. Stan offers a more mature and well-established ecosystem with extensive documentation, while NumPyro benefits from running on both CPUs and GPUs and appears to run faster.

Sviatoslav Rybnikov (The Hebrew University of Jerusalem)

Integrating gene drives and conventional pest control methods to minimize spillover risk

Gene drives are artificial genetic constructs that violate standard Mendelian inheritance laws. This emerging technology is considered promising for combating disease vectors and pests. However, while gene drives can potentially eradicate target populations within a few generations, they raise concerns regarding spillover into non-target populations, which could drastically affect ecosystems. Using mathematical models and optimal control theory, we investigate how combining gene drives with conventional control methods, such as pesticides and sterile males, can reduce spillover risk. Using dynamic programming, we identify the optimal timing and intensity of supplemental conventional controls to minimize the total costs of eradication, accounting for the damage of spillovers. Our findings show that integrating conventional control methods together with gene drive can significantly reduce spillover risk. Specifically, higher relative spillover costs require stronger and earlier interventions, sometimes before gene drives are released. For pesticides, treatment intensity increases gradually, while for sterile males, a sharp transition occurs when the relative spillover cost exceeds a critical threshold. Combining gene drives with conventional control methods also changes the optimal design of gene drive configurations, favoring gene drives with lower fitness costs.

Ivan Sciascia (Università di Torino)

An archive of bioimages from confocal microscopy

An archive of cellular images from confocal microscopy is presented here and summarizes the characteristics of a database that can be queried with search keys that allow the recovery of images or series of images in time lapse in confocal microscopy. The work aims to build bioimage identifiers that can be useful for building a relational bioimage database.

Shigeru Shinomoto (Kyoto University)

Nondifferentiable activity in the brain

Rapid advances in measurement technology have made it possible to record spike signals from numerous neurons in the brain. Accurate analysis of these data may reveal how signals are transmitted between neurons and, ultimately, the circuit structure of the brain. Recently, we have developed an analysis technique for estimating neural connections [1]. As we began to analyze large amounts of data, we newly discovered that the neural activity in awake animals has large steep coherent fluctuations, which may have caused errors in the estimation of neural connections. By analyzing the phenomena, we further improved the connection estimation algorithm [2]. We are now beginning to analyze the latest big data using these new analysis techniques and discovering new aspects of neural circuits.

References:

[1] Kobayashi et al., Nature Communications 10:4468 (2019)

[2] Tsubo and Shinomoto, PNAS Nexus 3:261 (2024)

Andrey Shternshis (Uppsala University)

Predicting allergy and postpartum depression from incomplete compositional microbiome

Data presented as compositional vectors obtained at several time points are considered. Data from some time points are missing, which reduces the size of the complete dataset. We propose a method for binary classification that includes imputation for missing values and logarithmic transformation of compositional data. Imputation approaches entail models that incorporate artificial data alongside collected time points, thereby supplementing the dataset. We consider two datasets with associated target labels, aiming to improve prediction accuracy. We

predict infants' food allergy from their gut microbiome with a balanced accuracy of 0.72. We forecast postpartum depression based on gut microbiome data collected during pregnancy, achieving a balanced accuracy of 0.62. Features extracted from the microbiome time series, specifically ratios of bacterial abundance, are statistically significant indicators of depression.

Astrid Sierens (Hasselt University, Vrije Universiteit Brussel)

Modelling contact and airborne transmission in an individual-based model with venues for COVID-19

The transmission of COVID-19 occurs through contact and airborne pathways, yet most individual-based models (IBMs) focus on only one of these. This study extends the STRIDE IBM, which originally accounted for contact transmission, by incorporating a venue-based structure, which modulates the contact network and facilitates airborne transmission. Instead of a general community pool for leisure contacts, individuals are now assigned to specific venues (e.g., restaurants, shops) based on time-use data.

This venue-based structure enables the modelling of airborne transmission by incorporating environmental factors, such as ventilation and the number of individuals per air mass, which affect viral spread through aerosols. It also introduces heterogeneity in contact behavior, reflecting that some individuals visit more venues than others, increasing variability in contact patterns. Contact transmission is further refined by including venue characteristics, such as respiratory activity and contact duration, which vary by venue.

Integrating the venue-specific framework and both transmission routes within the STRIDE IBM presents several challenges, including determining their relative contributions and introducing assumptions to address data availability limitations (e.g., for droplet emission and environmental factors). Despite these challenges, our model extensions provide a more detailed representation of transmission heterogeneity, offering insights into how both individual behavior and venue characteristics shape disease spread and contribute to superspreading. This helps the development of public health strategies.

Lucas Siu (Imperial College London)

Multi-trait signed linkage disequilibrium profile regression

Genome-wide association studies (GWAS) have been successful in identifying genomic regions linked to a trait. Fine-mapping, a follow-up analysis, seeks to pinpoint the specific variants within these regions that are directly causal with the tested traits. Fine-mapping from a statistics perspective can be regarded as a variable selection problem with two key challenges – sparsity and high correlation among variants. These challenges have driven the development of a wide array of specialised methods. There is growing interest in leveraging functional information data to improve the power of fine-mapping in identifying the causal genetic variants. We have explored the possibility of incorporating functional information as part of the prior distribution of the variant effects. As part of our work we have explored different approaches for computing the posterior credible sets of causal variants using both Markov chain Monte Carlo (MCMC) and variational inference (VI) methods. As part of our work we will present simulation scenarios that illustrate the effectiveness of incorporating functional information in fine-mapping analysis.

References:

[1] Reshef, Y.A., Finucane, H.K., Kelley, D.R. et al. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. Nat Genet 50, 1483–1493 (2018). https://doi.org/10.1038/s41588-018-0196-7 "

Jack Soulsby (Imperial College London)

High resolution Markov transition probabilities for scRNA-seq analysis

Single-cell RNA sequencing (scRNA-seq) analysis frequently employs transition probabilities in its initial stages, as they provide an elegant representation of differentiation by viewing differentiation through the lens of Markov transitions on a similarity graph. However, current approaches often bias transition probabilities towards more differentiated states in an oversimplified manner, potentially misrepresenting the true biological dynamics. Here, we introduce a novel computational method that generates transition probabilities with reduced entropy, better capturing the underlying diffusion processes in cellular state transitions. Through rigorous mathematical analysis, we prove that our method converges asymptotically to a true diffusion process in the limit of large data. Additionally, we demonstrate that by carefully selecting an appropriate temporal scale for the dynamics, our approach achieves comparable performance to existing state-of-the-art methods across various downstream analyses. Importantly, our method generates more biologically plausible trajectory samples, providing a more accurate representation of cellular differentiation pathways. This advancement offers researchers a more reliable framework for modeling cell state

transitions while maintaining computational efficiency. Most significantly, by generating biologically plausible transition probabilities, our method enables the conversion of static scRNA-seq snapshots into pseudo-temporal trajectories, opening new avenues for analyzing differentiation through established time series methodologies.

Luca Vedovelli (Università Degli Studi di Padova)

spatialBenchR: a Simulation Framework for Single-Cell Spatial Data

Current single-cell spatial analysis methods are rapidly evolving, but their validation remains challenging due to the lack of standardized benchmarking datasets. We introduce a simulation framework designed to generate realistic single-cell spatial data, enabling systematic testing and validation of analytical methods. A key feature of our approach is the use of spatial intensity patterns from real images to define cell clusters based on grayscale intensity and spatial proximity. Transcriptomic profiles are modeled with controlled variability both within and between clusters, incorporating spatial correlations to reflect realistic biological scenarios. To demonstrate the framework's applicability, we tested multiple single-cell analysis pipelines. Specifically, we modified the Seurat pipeline to enhance its precision in distinguishing different cell types by replacing Principal Component Analysis (PCA) with Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and integrating HDBSCAN as the clustering method. These modifications resulted in similar clustering precision compared to the original Seurat pipeline. In each case, the simulations demonstrated biological plausibility, captured relevant spatial relationships, and were produced within acceptable runtimes. By providing a standardized, flexible, and reproducible benchmarking platform, our simulation framework has the potential to support method development, improve comparative evaluations, and contribute to advancements in single-cell spatial research.

Edwin Weinholtz (TU Dresden)

Quantifying the Collective Behaviour of Cancer Cells: A Novel Framework for Analyzing Cell Clustering and Therapy Resistance

Invasion into surrounding tissues and metastasis are hallmarks of cancer (Hanahan and Weinberg, 2011). Notably, collective clusters of cancer cells have been shown to exhibit increased resistance to radiotherapy and DNA damage, contributing to therapy resistance (Haeger et al., 2017). Here, we introduce a novel framework to quantify the collective behavior of cancer

cells and assess how different conditions influence their clustering using image data of nuclei and actin filaments at fixed time points.

Our approach constructs a cell graph based considering nuclei as nodes that are linked using the actin filament data. Additionally, we introduce connection strengths on the graph's edges, depending on the distance of the nuclei and diameter of the actin signal between the cells using a principal component analysis. This enables our method to capture subtle differences in cellular interactions under different experimental conditions.

The overarching aim of our study is to uncover heterogeneities of the cell graph's observables that promote therapy resistence. By identifying these patterns, our findings aspire to facilitate the development of targeted interventions to prevent their appearance in the tumor, thereby augmenting the efficacy of cancer therapies.

Adriana Zanca (University of Melbourne)

Inferring chemical reaction networks with Catalyst.jl

Chemical reaction networks are ubiquitous, relevant in subjects ranging from material science to gene transcription. The reactions in a chemical reaction network can be described mathematically using ordinary differential equations derived from the law of mass action. Given some data, it is often not clear what the underlying network structure is that produced the data. In this work, I will explore how to recover the structure of a chemical reaction network — and the reaction rates within the network — from data computationally. To perform structural inference, we use the package Catalyst.jl in the Julia language to systematically generate all possible networks, then use optimisation techniques to determine which network(s) were most likely to have produced the data. While the applications presented are inference on chemical reaction networks, the methods shown can generalise to other kinds of compartmental models.