# Table of contents

# Exploring the space of self-reproducing RNA using generative models

Martin Weigt *† 1

1 Sorbonne University – Sorbonne Université UPMC Paris VI – France

Self-reproducing ribozymes represent a key challenge for understanding how complex RNA function can emerge spontaneously from less complex molecules and persist over time, yet only a handful of carefully designed examples are known. In this work, we combine generative probabilistic models with high-throughput functional screening to explore the sequence space of the Azoarcus ribozyme and related group I introns. By generating and testing over 20,000 artificial sequences, we uncover extensive mutational robustness and identify a vast number of viable self-reproducing variants far beyond engineered cases. In this context, we are able to substitute up to 30% of the wildtype nucleotides, and estimate the RNA sequence space to contain more than 10ˆ39 functional ribozyme sequences. We further demonstrate that integrating experimental feedback into generative models further improves predictive power, highlighting that better data, rather than more complex model architectures, is helpful in advancing RNA sequence design.

**Keywords:** RNA design, generative models, reintegration of experimental feedback, RNA sequence space

*Speaker

†Corresponding author: martin.weigt@sorbonne-universite.fr

# Exploring the archaic introgression landscape of admixed populations through joint ancestry inference

Jazeps Medina Tretmanis * , María Ávila-Arcos [1], Flora Jay [2], Emilia Huerta-Sanchez [3]

[1] LIIGH, UNAM – Mexico
[2] Laboratoire Interdisciplinaire des Sciences du Numérique – Institut National de Recherche en Informatique et en Automatique, CentraleSupélec, Université Paris-Saclay, Centre National de la Recherche Scientifique – France
[3] EEOB, Brown University – United States

Local Ancestry Inference (LAI) has been useful to study the evolution of recently admixed populations in the Americas. Similarly, detecting archaic introgressed tracts in Eurasian populations has provided insights into our interactions with archaic humans. Archaic introgression in recently admixed populations remains understudied, leaving questions about how recent admixture has altered patterns of archaic ancestry. This is partly due to a lack of methods that detect LAI and archaic ancestry simultaneously. Here, we present the first deep learning method capable of jointly inferring continental and archaic introgressed regions, even when trained on a mixture of real and synthetic data.

We show that we can train our model on a mixture of the 1KG dataset and synthetic data, and achieve high inference accuracies on the Simons Genome Diversity Project dataset (94.1% and 93.3% for continental and archaic ancestry classification). Our method is also better at detecting continental ancestry from old admixture events (> 100 generations), especially when the admixture proportion from one of the donors is small (_˜10%), where we demonstrate an 8% increase in accuracy for LAI compared to the next best available method.

Applying our method to American populations from the 1KG dataset, we find that Peruvian individuals harbor a large part of their archaic ancestry in Native American tracts (_˜50%), while other American populations like Puerto Ricans harbor most of their archaic ancestry tracts on European tracts (_˜70%). These results suggest that recent admixture has shifted patterns of archaic ancestry in different geographical regions due to local differences in their history of admixture. We also infer the joint ancestry of modern individuals from the Mexican BioBank dataset, and use these results to illustrate how our method can find candidates of adaptive introgression, through the identification of introgression hotspots where local ancestry patterns differ from global ancestry proportions.

**Keywords:** deep learning, ancestry, method, methods, human, humans, archaic, hominid, introgression, admixture, continental, local

*Speaker

# Predicting natural variation in the yeast phenotypic landscape with machine learning

Sakshi Khaiwal *† [1], Matteo De Chiara [2], Benjamin Barré [2], Inigo Barrio-Hernandez , Simon Stenberg , Pedro Beltrao , Jonas Warringer , Gianni Liti

[1] Institut de Recherche sur le Cancer et le Vieillissement – Université Nice Sophia Antipolis (1965 - 2019), Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Université Côte d'Azur – France
[2] Université Côte d'Azur – Université Côte d'Azur (UCA) – France

Most traits are encoded by complex interactions between genetics and the environment, ranging from height to disease susceptibility and response to medication. Predicting these traits is therefore a major goal of modern medicine, as it could pave the way for preventive and personalized healthcare. However, this remains a major challenge due to the complexity of the human genome and the many environmental factors that vary between individuals. To attempt such a challenge, we used Saccharomyces cerevisiae as a model organism, for which both complete genetic information and hundreds of traits are available for 1011 worldwide isolates grown in laboratory-controlled environments.

To investigate phenotype predictions for 223 traits measured across this collection, we developed a machine learning (ML) pipeline that integrates genomics (pangenome, SNPs, etc.), transcriptomics, and proteomics data. The pipeline evaluates both linear and non-linear models and benchmarks their performance across diverse input types and the full set of traits. Gradient boosting machines emerged as the best-performing model. Gene function disruption scores and gene presence/absence emerged as best predictors, suggesting a considerable contribution of the accessory genome in controlling phenotypes. The prediction accuracy broadly varied among phenotypes, with stress resistance being easier to predict compared to growth across nutrients. ML identified relevant genomic features linked to phenotypes, including high-impact variants with established relationships to phenotypes, despite these being rare in the population. Near-perfect accuracies were achieved when other phenomics data, mostly in similar conditions, were used, suggesting that useful information can be conveyed across phenotypes.

Our study presents the first large-scale comparison of machine learning methods across a broad range of traits and multi-omics data as predictors, highlighting their ability to decipher causative variants at the population level. We believe that our ML framework can be extended to other organisms, including humans, for predicting human traits and disease risks.

**Keywords:** Machine learning, phenotype prediction, genetic variants, S. cerevisiae, genotype, phenotype maps

[*] Speaker

[†] Corresponding author: Sakshi.KHAIWAL@univ-cotedazur.fr

# Phylodynamic modeling with unsupervised Bayesian neural networks

Marino Gabriele * [1,2], Ugnė Stolz [3,4], Daniele Silvestro [3,4], Cecilia Valenzuela Agüí [3,4], Tanja Stadler [3,4]

[1] ETH Zürich – Switzerland
[2] Swiss Institute of Bioinformatics – Switzerland
[3] ETH Zurich – Switzerland
[4] Swiss Institute of Bioinformatics – Switzerland

Phylodynamic modeling provides a powerful framework to quantify the parameters governing the growth of a phylogenetic tree-transmission, recovery, or migration in epidemiology, or speciation and extinction in macroevolution. Current approaches based on generalized linear models (GLMs) allow to integrate external, non-genetic covariates in the inference process (e.g., host traits or mobility data). Yet, these approaches often rely on restrictive assumptions-such as linear or additive effects of covariates -that limit their ability to capture nonlinear, context-dependent interactions driving transmission, speciation, and extinction. In epidemiology, this restricts integrating heterogeneous data (e.g., mobility patterns) into pathogen-spread models, while in macroevolution it constrains inference on how traits and paleoenvironment jointly influence diversification.

To address these challenges, we introduce a methodology leveraging Bayesian neural networks (BNN) as universal function approximators within a traditional Markov chain Monte Carlo (MCMC)-based phylodynamic inference framework. Instead of directly sampling phylodynamic parameters, our approach samples BNN weights, letting the network learn unsupervised mappings from predictors to parameters such as speciation, extinction, migration or transmission rates.

Through extensive simulations, we demonstrate that our framework reliably recovers the true relationships between predictors and phylodynamic parameters, avoids overfitting despite its large parameter space, and matches the performance of GLMs in linear settings while significantly outperforming them under nonlinear dynamics. Using tools from explainable AI, we further show that the model can identify which predictors meaningfully contribute to epidemiological or macroevolutionary dynamics, providing both accurate inference and interpretability. Finally, we validate the method on empirical datasets, estimating migration rates underlying COVID-19 early spread in Europe from viral sequence alignments and inferring speciation and extinction dynamics across living and extinct mammalian carnivores. Unsupervised BNNs unlock new avenues to model complex real-world epidemiological and macroevolutionary dynamics.

*Speaker

# Likelihood-free inference of phylogenetic tree posterior distributions

Luc Blassel * [1], Nicolas Lartillot , Bastien Boussau [2], Laurent Jacob [3]

[1] Département écologie évolutive [LBBE] – Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – France

[2] Université Claude Bernard Lyon 1 – LBBE – France

[3] Sorbonne Université – CNRS, CNRS : UMR7232, CNRS : UMR8001, CNRS : UMR8256, CNRS : UMR7203 – France

Phylogenetic inference, the task of reconstructing how related sequences evolved from common ancestors, is a central task in evolutionary genomics. The current state-of-the-art methods exploit probabilistic models of sequence evolution along phylogenetic trees, by searching for the tree maximizing the likelihood of observed sequences, or by estimating the posterior of the tree given the sequences in a Bayesian framework. Both approaches typically require to compute likelihoods, which is only feasible under simplifying assumptions such as independence of the evolution at the different positions of the sequence, and even then remains a costly operation. Here we present Phyloformer 2, a likelihood-free inference method for posterior distributions over phylogenies, trained end-to-end from sequences to trees. Phyloformer 2 exploits a novel encoding for pairs of sequences that makes it more scalable than previous approaches, and a parameterized probability distribution factorized over a succession of subtree merges. The resulting network outperforms both state-of-the-art maximum likelihood methods and a previous likelihood-free method for point estimation. It opens the way to fast and accurate phylogenetic inference under realistic models of sequence evolution.

**Keywords:** phylogenetic inference, likelihood free, end to end, neural posterior estimation

*Speaker

# Generative continuous time model reveals epistatic signatures in protein evolution

Barrat-Charlaix Pierre *[1]

[1] Laboratoire de Biologie Computationnelle Quantitative et Synthétique – Sorbonne Université – France

Protein evolution is fundamentally shaped by epistasis, where the effect of a mutation depends on the sequence context. As standard phylogenetic methods assume independently evolving sites, there is a need for more complex models based on accurate estimations of the fitness landscape. Good candidates are modern generative models – such as the Potts model – which successfully capture epistatic effects. However, recent work on generative evolutionary models usually use discrete time, making them difficult to integrate with the standard frameworks in evolutionary biology. We introduce a continuous-time sequence evolution model using the Gillespie algorithm and parameterized by a generative Potts model. This approach enables us to simulate realistic, family-specific evolutionary trajectories and allows for direct comparison with independent-site models. Surprisingly, we find that while epistasis significantly slows down evolution, it does not change the average evolutionary rates at individual sites. This is explained by the rate heterogeneity caused by context-dependence: we show that the rate at some positions varies between null to high values depending on the context, while other positions are essentially independent from the context. Finally, we show that epistasis leads to a systematic underestimation bias in the inference of evolutionary distance between sequences. Overall, our work provides a new tool for simulating realistic protein evolution and offers novel insights into the complex interplay between epistasis and evolutionary dynamics.

**Keywords:** sequence evolution models, generative models, phylogenetic inference, epistasis

*Speaker

# Neural posterior estimation for high-dimensional genomic data from complex population genetic models

Jiseon Min [1], Yuxin Ning * [2], Nathaniel Pope [1], Franz Baumdicker† [2], Andrew Kern‡ [1]

[1] University of Oregon [Eugene] – United States
[2] Eberhard Karls Universität Tübingen = University of Tübingen – Germany

Quantifying the evolutionary forces that shape natural populations is a key challenge in evolutionary ecology, where accurate inference improves our understanding of eco-evolutionary dynamics and local adaptation. Simulation-based inference has shown great potential in this field because it accommodates biologically realistic models when likelihoods are intractable. Approximate Bayesian Computation (ABC) is a widely used simulation-based inference method but comes with substantial computational cost and struggles to fit models to the high-dimensional data now routinely collected in ecological studies. In contrast, supervised machine learning (ML) techniques can handle such high-dimensional data, but they generally do not provide Bayesian uncertainty estimates for the quantities they predict.

Here, we present neural posterior estimation (NPE), a deep-learning framework that combines the strengths of ABC with the efficiency of supervised machine learning. By training neural density estimators on simulated data, NPE directly approximates the full posterior distribution of model parameters.

Benchmarking across a variety of population genetic tasks shows that neural posterior estimators yield amortized posterior distributions with high accuracy and better efficiency than ABC or parametric bootstrapping. We use the two-epoch demographic model of *Arabidopsis thaliana* to estimate the timing and magnitude of a bottleneck event, and the *Drosophila melanogaster* out-of-Africa model as an application to empirical data. Our experiments demonstrate the advantages of NPE in learning complex, non-linear posteriors directly from data across diverse inference tasks. NPE consistently produces well-calibrated, amortized posterior distributions from both manually computed summary statistics and neural representations learned from raw genomic data. We further show that incorporating linkage-based statistics alongside classical site frequency spectra improves prediction accuracy in certain settings. Additionally, we apply NPE to demographic inference for both simple and more complex models to highlight its versatility. Finally, we provide a user-friendly workflow that enables others to apply neural posterior estimation to their own genetic data.

**Keywords:** Simulation, based inference, Neural posterior estimation, population genetics

*Speaker

†Corresponding author: franz.baumdicker@uni-tuebingen.de

‡Corresponding author: adkern@uoregon.edu

# A differentiable model for detecting diversifying selection directly from alignments in large-scale bacterial datasets

Leonie Lorenz * [1], Joel Hellewell [1], John Lees [1]

[1] European Bioinformatics Institute [Hinxton] – United Kingdom

As populations adapt to changes in their environment, this can leave traces of diversifying selection in their genomes. In pathogens, an example of such an adaption is immune escape. Diversifying selection can be detected by calculating the ratio of nonsynonymous to synonymous substitutions, dN/dS. Classical methods for estimating dN/dS rely on phylogenetic trees, which limits scalability as sample sizes increase. Tree-based methods can also give misleading results for bacterial species as trees generally cannot be corrected such that they do not include recombination. We have developed a new tool called TOMBOMBADIL (Tree-free Omega Mapping By Observing Mutations of Bases and Amino acids Distributed Inside Loci), which computes dN/dS directly from codon counts from alignments by comparing to expected frequencies from the coalescent. We implemented TOMBOMBADIL as an end-to-end differentiable model in both STAN and Python, allowing us to estimate dN/dS values across genes for alignments with potentially millions of samples. TOMBOMBADIL includes different substitution (NY98, GTR) and codon evolution models (independent, similar within regions, or multimodal). We validated our method by comparing it to established, tree-based dN/dS methods and on well-known examples of bacterial genes under positive selection, including the outer membrane protein PorB of *Neisseria meningitidis*. Our next step is to scale to joint modelling over the whole genome scale. To do this, we are reimplementing TOMBOMBADIL in JAX to make use of stochastic gradient descent methods. The model is also extensible through adding regression terms between protein features and selection, which will help map sites of positive selection to predictions of protein structures to study the relationship of functional evolution and structure. In summary, TOMBOMBADIL will allow fast detection of positive selection in large-scale bacterial datasets using a tree-free dN/dS approach and machine learning methods.

**Keywords:** diversifying selection, positive selection, sequence alignments

*Speaker

# Detecting interspecific positive selection using transformers

Charlotte West * [1], Luca Nesterenko , Nick Goldman [1], Nicola Demaio [1], Bastien Boussau [2]

[1] European Bioinformatics Institute [Hinxton] – United Kingdom

[2] Laboratory of Biometry and Evolutionary Biology (LBBE) – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, LBBE, F-69100, Villeurbanne, France – France

Traditional statistical methods using maximum likelihood and Bayesian inference can detect positive selection from an interspecific phylogeny and a codon sequence alignment based on model assumptions, but they are prone to false positives due to alignment errors and can lack power. These problems are particularly pronounced when faced with high levels of indels and divergence. To address these issues, we trained and tested transformer models on simulated data and achieved higher accuracy in detecting selection across a wide range of phylogenetic scenarios and evolutionary modes. We developed models that excel in both binary classification (inference of the presence or absence of positive selection during the evolution of the sequences), as well as sitewise inference of dN/dS values. These advantages are particularly evident when performing inference on noisy data prone to misalignments. Our method shows some ability to account for these errors, where most statistical frameworks fail to do so in a tractable manner. Once trained, our transformer model is faster at test time, making it a scalable alternative to traditional statistical methods for large-scale, multigene analyses.

**Keywords:** positive selection, selection, simulation, self attention, transformer

*Speaker

# Predicting Multiple Sequence Alignment Uncertainty via Machine Learning

Lucia Martin-Fernandez *† [1], Mattis Bodynek [2], Ben Bettisworth‡ [1], Julia Haag [2], Alexandros Stamatakis§ [1,2,3]

[1] Biodiversity Computing Group, Institute of Computer Science, Foundation for Research and Technology - Hellas – Greece
[2] Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies – Germany
[3] Institute for Theoretical Informatics, Karlsruhe Institute of Technology – Germany

Computing a Multiple Sequence Alignment (MSA), constitutes an important and frequent operation in molecular sequences data analysis. Many tools exist to infer MSAs that rely on distinct algorithmic techniques and model assumptions. This diversity of available MSA tools therefore yields a diversity of inferred MSAs for the same input sequences. Further, even the same MSA tool can yield distinct MSAs when using slightly different input parameters. Consequently, this diversity induces MSA uncertainty. As numerous downstream analyses rely heavily on the reliability of the MSA, characterizing the extent of uncertainty in an MSA is crucial. To this end, we quantify this diversity via an MSA uncertainty score. We compute this uncertainty score as the average distance of a set of diverse MSAs from a given reference MSA, as measured by Blackburne and Whelan's *dpos* (a metric which matches positions and gaps). We then show that we can approximate the reference-based MSA uncertainty score via a reference-free method from unaligned sequences. This method summarizes the pairwise uncertainty scores from a set of distinct MSAs generated using various alignment tools and input parameters. Subsequently, we present a machine learning model that can reliably predict these reference-free uncertainty scores with a Root Mean Square Error of 0.04 for alignments in the test set. We validate our machine learning predictions on a diverse collection of empirical datasets from BAliBASE, TreeBASE, and published studies. We find that our method is accurate and robust to potential data source bias. We also find an inverse correlation between phylogenetic difficulty and MSA uncertainty.

**Keywords:** Multiple Sequence Alignment, Machine Learning, Uncertainty Analysis

*Speaker

†Corresponding author: luciamf@ics.forth.gr

‡Corresponding author: bbettis@ics.forth.gr

§Corresponding author: Alexandros.Stamatakis@h-its.org

# Graph Neural Networks for Likelihood-Free Inference in Diversification Models

Amélie Leroy *† [1,2], Ismaël Lajaaiti *

[3,4], Sophia Lambert *

[2], Jakub Voznica *

[2], Maximilian Pichler *

[4], Hélène Morlon *

[2], Florian Hartig *

[4], Laurent Jacob *

1

[1] Biologie Computationnelle, Quantitative et Synthétique – Sorbonne Universite – France
[2] Modélisation de la biologie – Institut de Biologie de l'École normale supérieure (IBENS), Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005 Paris, France. – France
[3] Institut des Sciences de l'Évolution de Montpellier – CNRS – France
[4] Theoretical Ecology Lab – Germany

A common approach to infer the processes that gave rise to past speciation and extinction rates across taxa, space and time is to formulate hypotheses in the form of probabilistic diversification models and estimate their parameters from extant phylogenies using Maximum Likelihood or Bayesian inference. A drawback of this approach is that likelihoods can easily become computationally intractable, limiting our ability to extend current diversification models with new hypothesized mechanisms. Neural networks have been proposed as a likelihood-free alternative for parameter inference of stochastic models, but so far there is little experience in using this method for diversification models, and the quality of the results is likely to depend on finding the right network architecture and data representation. As phylogenies are essentially graphs, graph neural networks (GNNs) appear to be the most natural architecture but previous results on their performance are conflicting, with some studies reporting poor accuracy of GNNs in practice. Here, we show that this underperformance was likely caused by optimization issues and inappropriate pooling operations that flatten the information along the phylogeny and make it harder to extract relevant information about the diversification parameters. When equipped with PhyloPool, a new time-informed pooling procedure, GNNs show similar or better

*Speaker

†Corresponding author: amelie.leroy@sorbonne-universite.fr

performance compared to all other architectures and data representations (including Maximum Likelihood Estimation) that we tested for two common diversification models, the Constant Rate Birth-Death and the Binary State Speciation and Extinction. We conclude that GNNs could serve as a generic tool for estimating diversification parameters of complex diversification models with intractable likelihoods.

**Keywords:** Cladogenesis, Deep Learning, Macroevolution, Probabilistic Biodiversity Model, Approximate Inference, Simulation, based Inference, Phylogeny

# Popformer: learning general signatures of genetic variation and natural selection with a self-supervised transformer

Leon Zong *† [1], Sara Mathieson [1]

[1] University of Pennsylvania – United States

Understanding natural selection can help to shed light on the genetics underpinning adaptation. The advent of large-scale data on human genetic variation has led to the development of data-driven methods for detecting signatures of selection, many of which are based on deep learning. However, these methods often fail to generalize well to the diversity of selection signatures across a broad range of evolutionary scenarios. We propose a novel transformer-based model, Popformer, for learning encodings of general patterns of genetic variation. Popformer includes site-wise and haplotype-wise attention, allowing us to capture global variation among both genetic positions and individuals. It additionally learns relative positional embeddings for inter-site positional distances. The model is pre-trained with an analog of the masked language modeling objective across a range of real human genomic data, similar to the task of genetic imputation. The pre-trained model learns meaningful encodings which can be utilized to classify signatures of natural selection, infer population ancestry, estimate demographic parameters, or to detect real versus simulated genetic data. Pre-trained Popformer outperforms naive baselines at imputing randomly masked genotypes. The performance of Popformer fine-tuned on selection detection rivals that of recent deep learning selection classifiers, without the need of being re-trained on task-specific simulations. Initial results of applying Popformer to 1000 Genomes data reveal its ability to generalize from diverse simulated data to real data. We also demonstrate the effectiveness of fine-tuning Popformer on separating real and simulated data, which can help identify weaknesses in simulation programs.

**Keywords:** selection detection, transformer, population genetics inference, selfsupervised learning, simulations

*Speaker

†Corresponding author: lzong@sas.upenn.edu

# PRIVET: PRIVacy metric based on Extreme value Theory

Antoine Szatkownik *† [1], Aurélien Decelle [2], Beatriz Seoane [2], Nicolas Béreux [1], Léo Planche [1], Guillaume Charpiat [3], Burak Yelmen [4], Flora Jay [1], Cyril Furtlehner [3]

[1] LISN – Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) – France
[2] Universidad Complutense de Madrid – Spain
[3] LISN – Université Paris-Saclay, CNRS, INRIA Tau team, LISN – France
[4] University of Tartu – Estonia

Deep generative models are often trained on sensitive data, such as genetic sequences, health data, or more broadly, any copyrighted, licensed or protected content. This raises critical concerns around privacy-preserving synthetic data, and more specifically around privacy leakage, an issue closely tied to overfitting. Existing methods almost exclusively rely on global criteria to estimate the risk of privacy failure associated to a model, offering only quantitative non interpretable insights. The absence of rigorous evaluation methods for data privacy at the sample-level may hinder the practical deployment of synthetic data in real-world applications. Using extreme value statistics on nearest-neighbor distances, we propose PRIVET, a generic sample-based, modality-agnostic algorithm that assigns an individual privacy leak score to each synthetic sample. We empirically demonstrate that PRIVET reliably detects instances of memorization and privacy leakage across diverse data modalities, including settings with very high dimensionality, limited sample sizes such as genetic data and even under underfitting regimes. We compare our method to existing approaches under controlled settings and show its advantage in providing both dataset level and sample level assessments through qualitative and quantitative outputs. As generative models become more prominent both in research and real-life applications, importance of privacy preservation grows substantially especially for datasets with potential to leak personal information, such as genetic datasets. To the best of our knowledge, our method is the first to offer a comprehensive assessment of both individual samples and entire datasets, while remaining theoretically grounded, scalable, and robust across low-sample regimes and diverse domains, thus helping bridge the gap between advances in generative modeling and privacy protection.

**Keywords:** Generative modeling, privacy, extreme value statistics, genetic data

*Speaker

†Corresponding author: antoine.szatkownik@universite-paris-saclay.fr

# Generative models for inferring the evolutionary history of the malaria vector Anopheles gambiae

Amelia Eneli [1], Matteo Fumagalli [1], Sara Mathieson * [2]

[1] Queen Mary University of London – United Kingdom
[2] University of Pennsylvania – United States

Malaria in sub-Saharan Africa is transmitted by mosquitoes, in particular the *Anopheles gambiae* complex. Efforts to control the spread of malaria have often focused on these vectors, but relatively little is known about the relationships between populations and species in the Anopheles complex. Here, we first quantify the genetic structure of mosquito populations in sub-Saharan Africa using unsupervised machine learning. We then adapt and apply an innovative generative deep learning algorithm to infer the joint evolutionary history of populations sampled in Guinea and Burkina Faso, West Africa. This algorithm (based on the method **pg-gan**, a Generative Adversarial Network) was adapted for the unusually high SNP density and population sizes of mosquitoes. We further developed a novel model selection approach and discovered that an evolutionary model with migration fits this pair of populations better than a model without post-split migration. For the migration model, we find that our method outperforms earlier work based on summary statistics, especially in capturing genetic differentiation between populations. We are currently modifying our framework to include transformers as the generative model, with the goal of joint inference of demography and natural selection. This modification means we can scale up our results to all sequenced populations, creating a detailed map of mosquito density across Africa. Finally, we interpret the trained models to understand what features of the real data are currently unmodeled by generative approaches. Overall our work has the ability to help us understand changes in population size, migration patterns, and adaptation in hosts, vectors, and pathogens. Ultimately this information could assist malaria control interventions, with the goal of predicting nuanced outcomes from insecticide resistance to population collapse.

**Keywords:** evolutionary inference, mosquitoes, malaria vector, generative models, GANs, transformers, interpretability

*Speaker

# Language Models Outperform Supervised-Only Approaches for Conserved Element Comprehension

Eyes Robson * [1], Nilah Ioannidis

[1] University of California [Berkeley] – United States

Supervised genomic sequence-to-function models such as Enformer have long demonstrated promise for comprehending genome sequences and interpreting noncoding variant effects. While recent research has revealed several of their limitations, such as modeling distal contexts or predicting the impact of personal genome variation, supervised-only approaches remain the dominant sequence-to-function strategy, due partly to a lack of head-to-head evaluations against alternatives on a single large-scale benchmark.

Here, we discuss findings from the updated genomic sequence-to-function benchmark, GUANinE v1.1. In addition to three newly-constructed variant effect tasks, we expand our baseline evaluations with over two dozen new models, including self-supervised genome Language Models (gLMs). Our findings suggest that while supervised-only methods maintain a competitive lead in functional tasks, they underperform on phylogenomic sequence conservation tasks relative to gLMs – their lower-context and less-supervised counterparts. We hypothesize that this is due to the sequence-introspective nature of language modeling, which requires estimating predictable (and thus conserved) regions of the genome.

GUANinE v1.1 answers questions such as: Is a model's understanding of sequence conservation sufficient to estimate variant deleteriousness? Does increased performance on functional annotation correlate with improved variant effect scores? How do gLMs behave on variant effect prediction outside of zero-shot settings? Altogether, our findings illustrate the different strengths of gLMs vs supervised approaches, and suggest a need for novel hybrid-supervision approaches which optimize their internal representations of sequences while benefitting from the information-rich nature of functional annotations.

**Keywords:** Benchmarks, sequence conservation, deleteriousness, sequence to function models, Transformers, Enformer, Borzoi, Language Models, gLMs

[*] Speaker

# Identification and Classification of Orphan Genes, Spurious Orphan Genes, and Conserved Genes from the human microbiome

Chen Chen * [1]

[1] Chen Chen – Netherlands

Orphan genes, defined as genes without detectable homologs outside a species, are widespread in microbial genomes and thought to play roles in adaptation and innovation. A challenging question is whether orphan genes actually include only entirely novel, functional coding sequences unique to each species. False positive orphan genes, or spurious orphan genes, can arise when non-existing genes are mistakenly classified as orphan genes in our study. We reason that if an orphan gene is found not to be expressed in a wide range of conditions, it may be spurious. Here, we combined large-scale metatranscriptomic profiling of the human gut microbiome with machine learning models to distinguish expressed orphan genes from spurious orphan genes and to explore how expressed orphan genes differ from conserved genes.

Using nearly 5,000 gut metatranscriptomic libraries, we identified _˜218,000 orphan genes supported by expression evidence, while over 330,000 predicted orphan genes lacked detectable expression and were classified as spurious. We extracted 154 sequence, structural, and evolutionary features for each gene, and trained XGBoost classifiers while accounting for genomic representation biases. The models achieved an area under the receiver operating characteristic curve (AUC) of 0.93 in distinguishing expressed orphan genes from conserved genes, and an AUC of 0.82 in distinguishing expressed orphan genes from spurious orphan genes. SHAP-based interpretation revealed clear distinguishing signals across sequence composition, GC3 content, and evolutionary features. Structural embeddings from ProstT5 further showed that conserved genes are characterized by stable protein structural profiles, whereas expressed orphan genes display less constrained structures, consistent with a more recent evolutionary origin.

This work advances methods for orphan gene discovery and demonstrates that expressed orphan genes differ systematically from conserved genes and spurious orphan genes in sequence composition, structural constraints, and evolutionary signals.

**Keywords:** orphan genes, spurious orphan genes, conserved genes, machine learning, human microbiome

*Speaker

# Neural Simulation-based inference of demography and selection

Francisco De Borja Campuzano Jiménez *† [1], Hannes Svardal

[1] University of Antwerp – Belgium

A central goal of evolutionary genomics is to understand how natural selection shapes genomes. Yet, many widely used approaches, such as genome scans for differentiated loci, are agnostic to demography. Model-based inference offers a principled alternative, jointly capturing the effects of selection and demographic history.
The challenge, however, is that likelihoods in population genetics are rarely tractable, even for simple models. Standard workarounds rely on composite likelihoods (fast but inexact) or Approximate Bayesian Computation (ABC) (flexible but computationally costly).

In this talk, I will highlight a promising alternative: Simulation-based Inference (SBI). The key idea is to train generative neural networks, such as normalizing flows, to directly approximate the posterior distribution \$p$(\theta \mid x)\$fromsimulateddata(i.e.pairsof\$(\theta, x)\$sampledfromthejointdistribution\$p(\theta,$
$estimates, SBIlearnsthefulldistribution, enablingefficientandflexibleinference.$

$Iwillillustratethisapproachwithapplicationstorecentdemographicchangeandpositiveselection.Specifically, I$
$(1)neurallikelihoodsaredifferentiable, whichallowsexplorationofcomplexparameterspaceswithgradient-$
$basedsamplerslikeNUTS, and(2)amortizationmakeslikelihoodevaluationsextremelyfastoncetrained, unlike$
$Bymergingmodel-basedinferencewithmachinelearning, wecanturnscientificmodelsintopracticalinference$
$specifyingamodel, generatingsimulations, andrecoveringalikelihoodfunctionthatisdifferentiableandcompu$

**Keywords:** Simulation, Based Inference, Population Genetics, Natural Selection, Neural Likelihood Estimation, Likelihood, free Inference

[*] Speaker

[†] Corresponding author: curro.campuzanojimenez@uantwerpen.be

# Species Identification and aDNA Read Mapping Using k-mer Embeddings

Filip Thor * [1], Carl Nettelblad * † [2]

[1] Uppsala Universitet [Uppsala] – Sweden
[2] Uppsala Universitet [Uppsala] – Sweden

We present a method for analyzing short-read data, with a particular focus on ancient DNA (aDNA) reads, using contrastive learning. An encoder model is trained to produce embeddings that cluster sequences from the same genomic region. The sequential nature of genomic regions is preserved in the form of trajectories through this embedding space. Trained solely to capture the structure of the genome, the resulting model provides a general representation of k-mer sequences that can be applied across a range of downstream tasks involving read data. We apply our framework to the genomes of E. coli and *S. enterica* subsp. *enterica*, and a segment of human chromosome 20. We demonstrate its use in simulated aDNA read mapping, species identification, and the detection of structural variations. Incorporating a domain-specific noise model is shown to enhance embedding robustness, and a supervised contrastive learning setting can be adopted when a linear reference genome is available, by introducing a distance-thresholding parameter. The model can also be trained fully self-supervised on read data, enabling analysis without the need to construct any complete genome assembly. Lightweight prediction models built on top of pre-trained embeddings achieve performance on par with BWA-aln, the current gold standard approach for aDNA mapping, in terms of accuracy and runtime for short genomes. Given the method's favorable scaling properties with respect to total genome size, inference using our approach shows strong potential for metagenomic applications and for mapping of reads to mammalian genomes. We conclude by discussing how we want to use these ”assembly without assembly” methods to model intraspecies genomic heterogeneity.

**Keywords:** aDNA, contrastive learning, read mapping, species identification

*Speaker

†Corresponding author: carl.nettelblad@it.uu.se

# Contrastive Learning for Population Structure and Trait Prediction

Filip Thor [1], Max Kovalenko [1], Carl Nettelblad *† [1]

[1] Uppsala Universitet [Uppsala] – Sweden

We have previously presented low-dimensional embeddings of genome variant data on populations such as 1000 Genomes and dog breed datasets, using convolutional autoencoders and contrastive learning. Here, we extend this approach to much larger cohorts with more than 100,000 samples, specifically the UK Biobank. We discuss the challenges associated with training on datasets of this size. Beyond describing broad population structure and providing visualization, we also train high-dimensional embeddings and explore their use in transfer learning for trait prediction. Starting from embeddings trained on large populations, we develop predictors for scalar phenotypes such as body height using only 1,000-20,000 individuals for training. To demonstrate the wide application of our methodology, we also predict geographic origin within the Baltic Sea herring population, formulating the task as a deep scalar regression of latitude and longitude from embeddings. The extent of genetic distinctiveness among subpopulations remains an open question, and our findings nuance the picture. Furthermore, our experience in suitable model architectures also translates to the application of diffusion model techniques in the genomic domain, including using diffusion models for imputation through inpainting.

**Keywords:** embeddings, contrastive learning, trait prediction, large populations

*Speaker

†Corresponding author: carl.nettelblad@it.uu.se

# Protein and genomic language models chart a vast landscape of antiphage defenses

Mordret Ernest * [1,2]

[1] Institut Pasteur [Paris] – Institut Pasteur de Paris – France
[2] Institut National de la Santé et de la Recherche Médicale – Institut National de la Santé et de la Recherche Médicale - INSERM – France

The bacterial pangenome harbors a vast and largely unexplored diversity of antiphage defense systems. Here, we develop machine learning approaches to systematically predict novel antiphage proteins. First, we fine-tune the protein language model ESM2 to detect distant homology with known defense proteins. Second, we train a genomic language model using the ALBERT architecture to infer defensive function from genomic context. Finally, we integrate protein sequence and context into a hybrid model, GeneCLR. Applied to a dataset of 30,000 bacterial genomes, these methods generalize effectively, expanding the diversity of the bacterial defense repertoire fivefold.

**Keywords:** Protein language model, Genomic language model, antiphage defense, Microbiology

*Speaker

# The Phylogenomics and Sparse Learning of Trait Innovations

Gaurav Diwan *† [1,2], Robert Russell [1,2]

[1] Bioquant – Germany
[2] Ruprecht-Karls Universität Heidelberg = Ruprecht-Karls University = Universität Heidelberg = Heidelberg University – Germany

A central challenge in evolutionary genomics is to identify the genomic signatures that underlie phenotypic adaptation. My research addresses this by asking: **given an environment or selective pressure, what genomic changes repeatedly enable organisms to adapt?** To answer this, I developed a phylogenomics framework that integrates comparative datasets with machine-learning approaches to uncover the gene families and amino acid variants associated with key traits.
The approach combines a **species tree**, **orthogroup-based gene family sizes**, and **one-hot encoding of multiple sequence alignments** with **Evolutionary Sparse Learning (ESL)**, which provides interpretable, minimal sets of predictive sequence features that distinguish species with and without a focal trait. Together with phylogenetic regressions, this pipeline yields ranked lists of gene families and variants repeatedly linked to trait shifts.

I will present analyses of **cancer resistance in large-bodied mammals**, a trait that has evolved independently in elephants, whales, and other mammalian lineages despite their vast cell numbers and long lifespans. Early results highlight convergence in canonical tumour suppressors such as **p53** and in transcriptional regulators (**TAF5-like RNA polymerase II p300, Mortality factor 4-like proteins, La ribonucleoprotein**). Additional candidates include stress-response and metabolic regulators (**heat shock cognate 71 kDa protein-like, adenylate kinase 6, translationally controlled tumour protein**), immune and extracellular components (**intelectins, CD99 antigen-like proteins**), and lineage-specific adaptations such as **hibernation-associated plasma protein HP-25-like**. These findings suggest that cancer resistance arises from a combination of conserved tumour suppressor pathways and clade-specific innovations in stress buffering and immune modulation.
This framework is extendable to other traits, including **climate resilience** (heat and salt tolerance in plants, insects, and microbes) and **pathogen tolerance** (in bats, amphibians, and bacteria). By integrating phylogenomics with sparse learning, the work provides a scalable method for mapping genotype to phenotype across the tree of life, revealing both general rules and lineage-specific solutions.

**Keywords:** phylogenomics, evolutionary sparse learning, genotype–phenotype mapping

*Speaker

†Corresponding author: gaurav.diwan@bioquant.uni-heidelberg.de

# Author Index