Postdoc in synthetic data simulation for deep learning in malaria genomic epidemiology

Institut Pasteur, Paris, France

This 2-year, full-time postdoc, starting January 2027 (or as soon as possible thereafter) is part of a French National Research Agency (ANR) funded project, UniGEM (Unified Inference for Genomic Epidemiology of Malaria), led by Aimee Taylor. Using simulation-based inference (SBI) with deep learning, UniGEM aims to build a neural network to estimate epidemiological parameters of *P. falciparum,* the deadliest malaria parasite species. Within UniGEM, the postdoc will be responsible for the simulation of synthetic data used to train, develop, and test the neural network.

Scientific context: Malaria parasites are transmitted between humans by mosquitoes. In the mosquito host, they undergo an obligate stage of sexual recombination. Transmission-dependent rates of out-crossing, inbreeding, and self-fertilization between parasites are governed by superinfection and co-transmission on the mosquito-to-human level. These overlapping, non-linear processes obfuscate estimation of malaria epidemiological parameters from parasite genetic data; cf., viral genomic epidemiology, where bifurcating trees capture the ancestry of DNA sequences, and human population genetics, where the ancestral recombination graph (ARG) needs no dynamic rate of self-fertilisation or hierarchical structure over processes on a host-to-host level [1]. To estimate malaria epidemiological parameters using simulation-based inference, UniGEM requires a simulator that captures malaria-specific recombination and transmission processes, generating realistic synthetic datasets while retaining sufficient computational efficiency for large-scale simulation.

Postdoc work programme: The postdoc will review the literature on epidemiological parameters and existing simulators, then benchmark a subset of simulators against key criteria. They will likely iterate twice over data simulation: once towards the end of their first year, using the most appropriate simulator available at that time. A second year provides time to develop an enhanced simulator and to simulate a refined batch of synthetic data. For each batch, the postdoc will estimate Bayes optimal error, an important guide for realistic goals for deep learning. The first batch of synthetic data will be used by a PhD student, likely starting October 2027, to initiate development of a deep-learning architecture. The refined batch of synthetic data will be used by the PhD student to finalise the deep-learning model. Expected outputs include parameter priors, benchmarking results, a simulation pipeline, curated synthetic datasets, and Bayes optimal errors.

Publication / conference: Depending on progress, there is potential for one or two first-author publications: a benchmarking study comparing existing simulators and a methodological study detailing the development of an enhanced simulator. In addition, the postdoc will be invited to co-author all publications on UniGEM activities that make use of the postdoc results and encouraged to disseminate findings at the annual American Society of Tropical Medicine and Hygiene conference.

Candidate profile: An ideal candidate will likely have a PhD in statistical population genetics (e.g., prior experience working with ARGs / coalescent-with-recombination), computational biology, bioinformatics, infectious-disease epidemiology, evolutionary or ecological biology, or a closely related field.

- Strong programming skills (R, Python, and/or other)
- Familiarity with version control (Git / GitHub)
- Proficient written and spoken English (B2 level or higher)
- Ability to communicate methodological assumptions and limitations clearly and succinctly
- Experience with reproducible computational workflows and high-performance computing
- Experience handling large population-genetic simulations (e.g., SLiM [2], msprime [3], tskit [4])

Application: Interested applicants should email Aimee Taylor at ataylor@pasteur.fr, describing their motivation and relevant experience, providing a CV, and providing links to relevant software, code repositories, preprints, or publications. The position will remain open until filled, with applications reviewed on a rolling basis.

All critical information is included on this page. Interested candidates can find more details on location, supervision, the work programme, and references below.

Location: Infectious Disease Epidemiology and Analytics (IDEA) Unit, Institut Pasteur, Paris, France. Institut Pasteur is a world-leading centre for biomedical research. Postdocs are supported by a variety of dedicated resources including a postdoc office, a careers office (offering a career programme, career events, access to an alumni network), and an office of Diversity, Equity, and Inclusion. Financial support for childcare, public transport, and other daily expenditures is also provided.

Within Pasteur, IDEA is a highly international and multidisciplinary unit, with expertise in mathematical modelling and machine learning, wet-lab expertise in multiplex serological and genetic assays, expertise in diagnostic development and production, and expertise coordinating field studies and large cluster-randomised clinical trials, especially with collaborators across Africa.

IDEA is located within the Department of Global Health (another highly international and multidisciplinary structure) and has close ties with the Department of Parasites and Insect Vectors, which includes various groups working on malaria, and the Department of Computational Biology, which includes the Bioinformatics and Biostatistics Hub, providing support in computational biology to all research groups across Institut Pasteur. In addition to all the resources of Institut Pasteur, the postdoc will have access to a large network of world-leading experts in the Paris Artificial Intelligence Research Institute (PRAIRIE).

Supervision: Aimee Taylor (permanent researcher, IDEA, and junior chair, PRAIRIE) is the principal investigator of UniGEM and will lead the postdoc project, providing critical guidance on the specificities of malaria genomic epidemiology. Michael White (Unit director, IDEA) will provide guidance on the identification of epidemiological parameters of interest, and on how to link them to simulator inputs.

Detailed work programme

Identify priors: Epidemiological parameters of interest, their plausible ranges, and their notable features, will be identified using a literature review (an example of a malaria epidemiological parameter of interest is the entomological inoculation rate; it ranges from < 1 infective mosquito bites per person per year in very low transmission settings to > 800 in some very high transmission settings [5]). Information gathered in the literature review will then be used to construct epidemiologically realistic prior distributions, from which a wide range of values linked to parameters of interest can be drawn, and thus used to simulate synthetic data. The literature on malaria epidemiological parameters is rich and can contain conflicting values; meanwhile, some parameters (e.g., superinfection prevalence) lack empirical measures. The postdoc will consult with transmission modelling expert, Michael White, to resolve literature discrepancies. Where domain knowledge is entirely lacking or conflicting, broad and non-informative prior distributions will be used.

Identify a simulator: Candidate simulators will be identified using a literature review. Examples of malaria-specific simulators include EMOD malaria [6], forward-dream [7], Magenta [8], SimpleGEN (unpublished R package, mrc-ide.github.io/SIMPLEGEN/), which are all forward-time simulators, and the genomic transmission graph of Dominic Kwiatkowski [9], which is the only malaria-specific coalescent-with-recombination (i.e., backward-time) simulator. These specialised simulators feature various malaria-specific processes (e.g., person-to-person mosquito transmission, using modifications of the Ross-Macdonald model [10], superinfection, and co-transmission). Examples of general-purpose simulators include the forward-time simulator SLiM [2], and the backward-time coalescent-with-recombination simulator msprime [3]. These two approaches have been combined via the tskit library [4]. This hybrid approach, which is extremely powerful, was used in a recent study of the impact of selection on malaria-parasite population-genetic inferences [11].

Benchmark simulators: Simulators will be scored in terms of their realism regarding malaria-specific processes, their range of user-specifiable input variables that can be linked to epidemiological parameters, and their computational efficiency. High-scoring simulators will be benchmarked against theoretical expectations and by prior predictive checks, i.e., data will be simulated given parameter values drawn from the priors identified above, and summary statistics of the synthetic data will be compared to those computed using real malaria parasite genetic (an intern working alongside the postdoc in the first six months will generate a pipeline to compute these summary statistics and use it to compute values for the real malaria parasite data). Almost certainly, no simulator will have input variables that link to all the desirable parameters. Preference will be given to those that enable inference on parameters whose values are operationally informative. If no simulator performs well across the range of summary statistics used for prior predictive checking, the simulator that performs best on average (possibly using a weighted average to preference summary statistics that we believe are most informative) will be chosen, and the postdoc will be encouraged to develop an enhanced simulator in their second year (e.g., by modifying and/or extending an existing one). For any edge cases where simulators out-

perform the chosen / enhanced simulator, out-of-sample development and test sets can be used to assess the impact of the chosen / enhanced simulator.

Simulate and summarise synthetic data: Using a wide range of parameter values drawn from the priors identified above, millions of well-documented synthetic data sets will be simulated on Institut Pasteur's High-Performance Computing (HPC) cluster Maestro. The postdoc will have priority use of a single CPU node plus a GPU workstation. Each simulated dataset will include genotyping calls on approximately 500,000 across 20 samples. For each set of 20 samples, a per-population vector of summary statistics will be generated using the pipeline developed by the intern.

Approximate Bayes optimal error: Different epidemiological processes can sometimes generate identical genetic observations. In this setting, epidemiological parameters cannot be estimated with zero error using genetic data alone. The postdoc will help the PhD student set realistic goals for optimal deep learning, by estimating simulator non-identifiability. For example, for a random subset of synthetic dataset pairs, one could compute the Euclidean distance between the vectors of parameter values used to simulate the first and second dataset within each pair; compute the Euclidean distance between the vectors of summary statistics (or the Wasserstein distance between the empirical distributions over simulated sequences) for first and second dataset within each pair; and compute the proportion of distinct parametrisations that are observationally equivalent (e.g., non-zero distances in parameter space associated with near-zero distances in summary-statistic space).

References

[1] F. Camponovo *et al.*, "Measurably recombining malaria parasites," *Trends Parasitol.*, vol. 39, no. 1, Jan. 2023.

[2] B. C. Haller *et al.*, "Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes," *Mol. Ecol. Resour.*, vol. 19, no. 2, Mar. 2019.

[3] J. Kelleher *et al.*, "Coalescent simulation with msprime," *Methods Mol. Biol.*, vol. 2090, 2020.

[4] J. Kelleher *et al.*, "Efficient pedigree recording for fast population genetics simulation," *PLoS Comput. Biol.*, vol. 14, no. 11, Nov. 2018.

[5] T. Degefa *et al.*, "Methods of sampling malaria vectors and their reliability in estimating entomological indices in Africa," *J. Med. Entomol.*, vol. 61, no. 3, May 2024.

[6] A. Bershteyn *et al.*, "Implementation and applications of EMOD, an individual-based multi-disease modeling platform," *Pathog. Dis.*, vol. 76, no. 5, July 2018, doi: 10.1093/femspd/fty059.

[7] J. A. Hendry *et al.*, "Elucidating relationships between P.falciparum prevalence and measures of genetic diversity with a combined genetic-epidemiological model of malaria," *PLoS Comput. Biol.*, vol. 17, no. 8, Aug. 2021.

[8] O. J. Watson *et al.*, "Evaluating the performance of malaria genetics for inferring changes in transmission intensity using transmission modeling," *Mol. Biol. Evol.*, vol. 38, no. 1, Jan. 2021.

[9] D. Kwiatkowski, "Modelling transmission dynamics and genomic diversity in a recombining parasite population," *Wellcome Open Res.*, vol. 9, Apr. 2024.

[10] D. L. Smith *et al.*, "Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens," *PLoS Pathog.*, vol. 8, no. 4, Apr. 2012.

[11] B. Guo *et al.*, "Strong positive selection biases identity-by-descent-based inferences of recent demography and population structure in Plasmodium falciparum," *Nat. Commun.*, vol. 15, no. 1, Mar. 2024.