Simulation-based inference of yeast centromeres

Eloïse Touron
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
eloise.touron@inria.fr

Pedro L. C. Rodrigues
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
pedro.rodrigues@inria.fr

Julyan Arbel
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
julyan.arbel@inria.fr

Nelle Varoquaux
TIMC, Univ. Grenoble Alpes
CNRS, Grenoble INP, France
nelle.varoquaux@univ-grenoble-alpes.fr

Michael Arbel
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France
michael.arbel@inria.fr

Abstract

The chromatin folding and the spatial arrangement of chromosomes in the cell play a crucial role in DNA replication and genes expression. An improper chromatin folding could lead to malfunctions and, over time, diseases. For eukaryotes, centromeres are essential for proper chromosome segregation and folding. Despite extensive research using *de novo* sequencing of genomes and annotation analysis, centromere locations in yeasts remain difficult to infer and are still unknown in most species. Recently, genome-wide chromosome conformation capture coupled with next-generation sequencing (Hi-C) has become one of the leading methods to investigate chromosome structures. Some recent studies have used Hi-C data to give a point estimate of each centromere, but those approaches highly rely on a good pre-localization. Here, we present a novel approach that infers in a stochastic manner the locations of all centromeres in budding yeast based on both the experimental Hi-C map and simulated contact maps.

1 Introduction

Hi-C maps have become one of the main assets in understanding DNA folding, notably through the study of chromatin loops or topologically associated domains (TADs) in mammalian cells [6, 20]. These maps capture the contact counts between fragments of chromosomes among a population of DNA into a 2D squared and symmetric matrix made of cis- and trans- blocks of interactions.

Beside chromatin loops or TADs, centromeres are also of great interest to genome structure investigation due to their essential role in many biological processes: they facilitate chromosome segregation through the formation of the kinetochore [2] during mitosis and meiosis, and act as key regulators of genome stability via the prevention of chromosome breakage. In yeasts, centromeres are highly compact regions spanning about 125 base pairs (bp) [4] and tend to cluster near the spindle pole body within the nucleus. This clustering results in a distinct peak in the trans-contact counts Hi-C matrices, centered at the position of each centromere pair. Many studies have attempted to annotate yeast centromeres, usually through Fluorescent In Situ Hybridization (FISH) [14] or chromatin

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

immunoprecipitation (ChIP) [12]. However, these approaches often remain imprecise and may even fail to infer centromeres for some species [7]. To bypass these limitations, newer methods have been proposed, which use Hi-C contact maps [13, 18]. Working directly with a Hi-C contact map or with the corresponding Pearson correlation matrix, they fit a Gaussian to each interaction peak to precisely infer centromere locations. Beyond the optimization of a non-convex function, these methods highly rely on good pre-localization of the centromeres to be precise and output only a point estimate of each centromere whereas it is actually a whole segment of chromosome.

We propose a novel approach to infer centromere positions that differs from existing methods in two key aspects. Firstly, we adopt a stochastic approach by quantifying the uncertainties about the centromere candidates we infer. Secondly, the inference processes are not only based on a reference Hi-C matrix (denoted C_{ref}) but also on simulated contacts maps (denoted C). We adopt a Bayesian approach where centromere positions (denoted θ) are sampled from a prior distribution and the contact maps C are generated from a custom-designed simplified simulator. We estimate the posterior distribution $p(\theta|C_{\text{ref}})$, which amounts to solving the following inverse problem:

Given contact map C_{ref}, what are the most probable centromere positions θ to have generated it?

2 Methods

2.1 Framework

Setting. We work with the budding yeast *Saccharomyces cerevisiae* for which the positions of its 16 centromeres and length of each chromosome (in base pairs) are known. We used data from Duan et al. [5] to construct the reference Hi-C contact map. Centromere candidates $\theta = (\theta_1, \ldots, \theta_{16})$ are sampled from a prior distribution $p(\theta)$ that is poorly informative, e.g. a multivariate uniform distribution where the interval is the length of each chromosome in each dimension. To simulate contact maps, we designed a simplified simulator (described below) that takes θ as input and directly outputs realistic C without simulating any DNA folding. Other studies have used simulators to do Bayesian inference of biological elements [1], but those biological simulators try to simulate the 3D folding of the chromatin in the nucleus before computing the corresponding contact map. This renders them too slow and unnecessarily complex if we want to have many contact maps in a reasonable time.

Contact maps and data normalization. The contact map C projects the information contained in a population of 3D chromatin foldings into a 2D squared and symmetric matrix made of cis- (or intra-chromosomal) and trans- (or inter-chromosomal) blocks of interactions between pairs of chromosomes. To construct it, we cut each chromosome into genomic windows of a given length (called resolution), e.g. 32 kilobases (kb). Each matrix entry is then a non-negative number, called the contact count, representing the number of times a given window was in contact with another one over the population (see Appendix A and Figure 3 for more details).

During inference, we use a reference Hi-C map C_{ref} and simulate synthetic contact maps C. Hi-C contact maps have many biases due to sequencing and mapping errors or to the inherent structure of the chromatin [10]. Therefore, C_{ref} is actually a normalized Hi-C map, where the normalization corrects those biases, iteratively forcing all rows and columns to sum up to one [10]. The quality of the contact map depends on the chosen resolution and the signal-to-noise ratio gets smaller if we work at higher resolution. We thus choose to set the contact maps at resolution 32 kb. In yeasts, the main informative part about centromeres rely on the upper trans-contact blocks: the matrix of contacts between chromosomes i and j contains an enrichment of interactions at the location of both centromeres (θ_i, θ_j).

Simulator. We exploit the structure of yeast contact maps to design a very efficient simulator that directly creates the upper trans-contact blocks given its centromeres positions θ. At the centromere positions, the chromatin has a brush-like organization: chromosomal regions near the centromeres often enter in contact over the population whereas the further we move away from the centromeres, the rarer the contacts become. To mimic this effect, we simulate a Gaussian spot at the position (θ_i, θ_j) for each trans-contact block. Between chromosomes, we also observe rare interactions over the population that we reproduce by adding Gaussian noise to all the trans-contacts blocks up to 10% of the maximal contact count (see Appendix B with Algorithm 1 and Figure 4 for more details).

2.2 Simulation-based inference

Our goal is to infer θ from C_{ref} using a probabilistic framework based on simulations. The usual way for doing so would be to search for the most appropriate θ for a given C_{ref} by maximizing the likelihood:

$$\hat{\theta} = \underset{\theta \in \Omega}{\text{argmax}} \ \log p(C_{\text{ref}}|\theta) \ .$$

However, as the simulator is often very complex (e.g. biological simulators that try to mimic a 3D folding of the DNA given a set of constraints), the likelihood $p(C|\theta)$ may be intractable. As such, we directly target the posterior density $p(\theta|C_{\text{ref}})$ using data from the joint model $(\theta, C) \sim p(\theta)p(C|\theta)$, either via approximate Bayesian computation (Sequential Monte-Carlo ABC: SMC-ABC) or by estimating the posterior density with a conditional normalizing flow (Sequential neural posterior estimation: SNPE) [15, 8] .

SMC-ABC. We use a variant of ABC coupled with sequential Monte-Carlo (SMC) [17]. It consists of multiple rounds of ABC where, at each round, relevant $\{\theta^{k,*}\}_k$ are selected from the training set $\{(\theta^n, C^n)\}_n$ depending on a closeness criterion between C and C_{ref}. We then associate weights $\{w^k\}_k$ to those selected $\{\theta^{k,*}\}_k$, and use the set $\{(\theta^{k,*}, w^k)\}_k$ to create the next population of $\{\theta^n\}_n$ for the next round of ABC. This sequential approach enables us to refine the relevant θ at each round. However, we need to define a metric for discriminating (θ^n, θ^m) based on their associated observations (C^n, C^m).

<u>Metric 1</u> : Pearson correlation – ABC-Pearson. To measure the closeness between C and C_{ref}, the Pearson correlation is commonly used [16, 13, 18]. We find that the vector-based Pearson correlation averaged over all trans-contacts blocks is the most discriminative metric: each trans-contacts block of C and C_{ref} is vectorized and the Pearson correlation is computed between both. We then average all the correlations over the trans-contacts blocks (see Algorithm 2 in Appendix C.1). However, this metric is fine-tuned to this specific inference task.

<u>Metric 2</u> : Data-driven summary statistic – ABC-CNN. Instead of looking for a specific metric to compare C to C_{ref}, we choose to use the classical l^2-norm. For this, we need a summary statistic S that will extract the main features of C and project it into a low-dimensional vector. One relevant candidate for a summary statistic is $\mathbb{E}[\theta|C]$ because with this one:

$$\mathbb{E}\left[\theta \mid \|S(C) - S(C_{\text{ref}})\| \leq \epsilon\right] \underset{\epsilon \to 0}{\to} \mathbb{E}\left[\theta|C_{\text{ref}}\right] \ ,$$

where ϵ is the ABC-threshold.
When $\epsilon \to 0$, we don't lose any first-order information when summarizing C [11]. Moreover, $\mathbb{E}[\theta|C]$ is the analytical solution of the regression of θ on C i.e.

$$\mathbb{E}\left[\theta|C\right] = \underset{S \in \mathcal{F}}{\text{argmin}} \ \mathbb{E}\left[\|S(C) - \theta\|_2^2\right] \ , \tag{1}$$

where $\mathcal{F}$ is the set of square integrable functions. As this statistic is unavailable, we learn it via a (deep) neural network (DNN) S_ϕ with parameters ϕ [11]. The DNN encoding S_ϕ is composed of a convolutional neural network (CNN) followed by a multi-layer perceptron (MLP). Using Monte Carlo estimator of (1) with N samples $(\theta^n, C^n) \sim p(\theta)p(C|\theta)$, the DNN loss to be minimized in ϕ is then

$$\widehat{\mathcal{L}}_{\text{DNN}}(\phi) = \frac{1}{N} \sum_{1 \leq n \leq N} \|S_\phi(C^n) - \theta^n\|_2^2 \ .$$

For large N, we expect $S_\phi(C) \approx \mathbb{E}[\theta|C]$. This approach has two phases: first learn S_ϕ by minimizing $\widehat{\mathcal{L}}_{\text{DNN}}(\phi)$, then run sequential ABC with this summary statistic and the l^2-norm as discriminative criterion (see Algorithm 3 in Appendix C.2).

SNPE – SBI-CNN. SMC-ABC yields only samples from the target posterior distribution $p(\theta|C_{\text{ref}})$, but evaluating log-probabilities can be useful for downstream tasks. In contrast, a conditional normalizing flow $p_\psi(.|.)$ [15, 8] used to estimate the posterior distribution can both easily sample from the posterior and return the values of its log-probabilities. To ensure that $p_\psi(\theta|C_{\text{ref}})$ is close to $p(\theta|C_{\text{ref}})$, we minimize their Kullback–Leibler divergence (D_{KL}) averaged over the observations C as per

$$\mathbb{E}_C\left[D_{\text{KL}}\big(p(\cdot|C) \| p_\psi(\cdot|C)\big)\right].$$

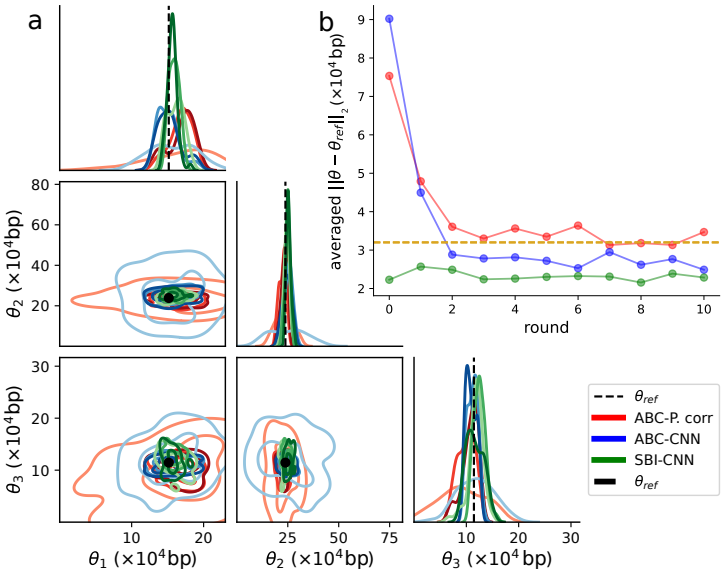

Figure 1: Inference using **ABC-Pearson**, **ABC-CNN**, and **SBI-CNN** (**a**). Color shades increase from lightest to darkest across rounds. Densities are estimated with the 5% best θ according to the ABC criterion or sampled from the flow. We also report the mean Euclidean distance between θ and θ_{ref}, computed over the 5% best-performing samples in the top right corner (**b**). The horizontal dashed line stands for the resolution of the contact map C_{ref} (in bp) in the top right figure. Results with **SBI-CNN** are uniformly better and both approaches based on data-driven summary statistics have errors smaller than the resolution of the contact maps.

After simplifications and using a Monte Carlo estimator, the flow is trained to minimize

$$\widehat{\mathcal{L}}_{\text{NPE}}(\psi) = -\frac{1}{N}\sum_n \log(p_\psi(\theta^n|C^n))\,,\ (\theta^n, C^n) \sim p(\theta)p(C|\theta).$$

Once trained, we obtain an amortized estimator of the posterior densities $p(\theta|C)$ valid for any C. We just have to plug in C_{ref} to get the estimated posterior density $p_\psi(.|C_{\text{ref}})$ (see Algorithm 4 in Appendix D). Since we are actually interested in the posterior at C_{ref}, parameters θ with very low posterior density may not be useful for learning ψ. Thus, we consider a sequential approach with several rounds of NPE to get an iterative refinement of the posterior estimate [8]. From the second round, θ^n are sampled from the latest estimated posterior found instead of the prior. This way, training samples are more informative about C_{ref}, gradually improving the learning of ψ. When the observations C are high-dimensional (e.g. 2D-matrices), we encode them in a summary statistic S using a convolutional neural network.

3 Numerical experiments

We showcase our methodology on two settings involving the genome of the yeast *S. cerevisiae*, for which we have access to the true position of all its centromeres: firstly, we run our inference pipeline on only *S. cerevisiae*'s first three chromosomes, secondly on its whole genome (16 chromosomes). We assess the performance of different inference methods by comparing their approximate posterior distributions to a ground-truth distribution consisting of Diracs on each dimension located at the true centromere positions. All experiments can be run on the CPU of a laptop, requiring ~ 1 h for the small genome and ~ 5 h for the entire genome.

3.1 Study case – small genome (3 chromosomes)

In low-dimensional settings, we can jointly infer θ given the entire contact map C_{ref}. For each inference method, we consider 11 rounds, each with a training dataset $\{(\theta^n, C^n)\}_n$ of size 10^3. The summary statistic S_ϕ is pre-trained using a different training set of size 5×10^3 with the optimizer `Adam` and a fixed learning rate 5×10^{-4}. For SNPE, we use a masked autoregressive flow (MAF) and the version SNPE-C [8] from the Python package `sbi` [3] (see Appendix E for more details).

3.2 High dimensional problem – whole genome (16 chromosomes)

When extending the analysis to the entire genome, we end up facing the curse of dimensionality: the space of parameters θ becomes too large to cover with few simulations and the contact maps

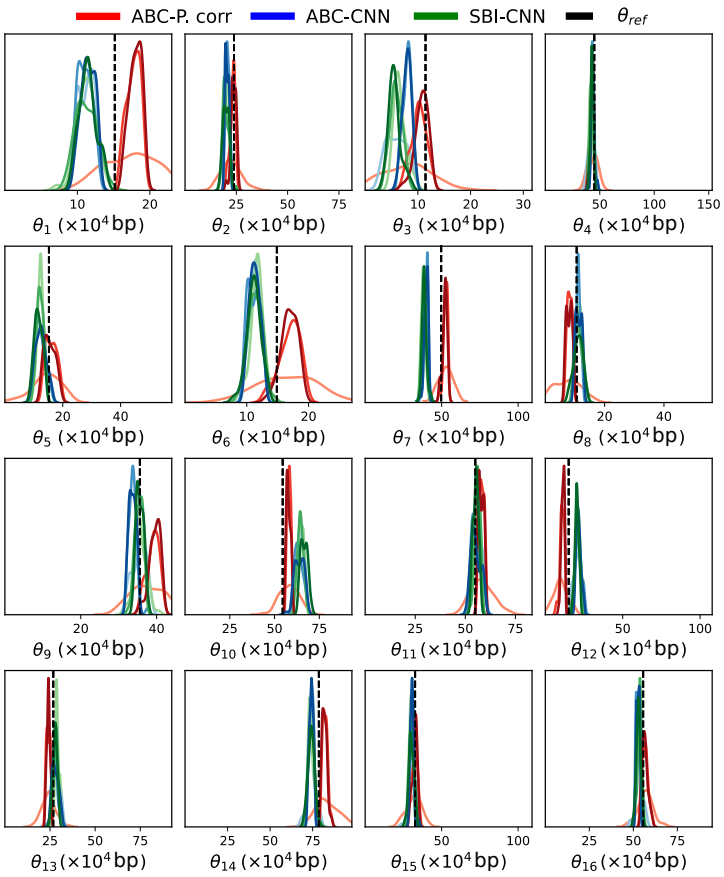

Figure 2: Inference using **ABC-Pearson**, **ABC-CNN**, and **SBI-CNN**. Color shades increase from lightest to darkest across rounds. Densities are estimated with the 5% best θ according to the ABC criterion or sampled from the flow. In some dimensions, the densities are very peaky and centered around θ_i (e.g. chromosome 4, 13, 15) but in others, the inference is not precise (e.g. chromosome 1, 6, 10). Data-driven summary statistics approaches do not outperform Pearson correlation-based method.

C are too big. As such, the resulting neural network encoding S_ϕ has too many parameters to be optimized. To reduce the dimension of the problem, we run 16 parallel inferences (one per dimension of θ) extracting each time only the informative part of the contact maps. With this approach, the space of θ is cut into several chromosome-length 1D intervals reducing the train set size. Let C_i and $C_{\text{ref},i}$ be the ith lines of blocks of matrices C and C_{ref}, respectively. To infer θ_i with **ABC-Pearson**, we compute the vector-based Pearson correlation averaged over all blocks between C_i and $C_{\text{ref},i}$. Concerning the data-driven summary statistic approaches **ABC-CNN** and **SBI-CNN**, the summary statistic S_{ϕ_i} also tries to project C_i to θ_i.

Using the redundancy of the data between rows of blocks, and to minimize the number of parameters of $\{S_{\phi_i}\}_i$, we consider a shared architecture where the CNN parameters are shared between chromosomes and the MLP ones are chromosome-specific. For **SBI-CNN**, we learn 16 normalizing flows $p_{\psi_i}(\theta_i | S_{\phi_i}(C_{\text{ref},i}))$ (see Appendix F for more details and results).

4 Discussion

We present a novel methodology to infer the positions of the centromeres of the yeast *S. cerevisiae* using Hi-C contact maps. The probabilistic framework that we use allows us to quantify the uncertainty about the centromere candidates. Our entire inference pipeline is based on a large number of simulations relating centromere positions and contact maps. To mitigate computing bottlenecks, we have designed a simplified but efficient simulator that yields very convincing results when coupled with inferences on real experimental data.

In the case of a small genome, we obtained accurate inference of the centromere positions (Figure 1). The estimated densities for the summary statistic-based methods (**ABC-CNN** and **SBI-CNN**) are not very biased and peaky around the ground truth. In each dimension, θ is estimated at a precision under the resolution of the contact map C_{ref} (Figure 6a) and the Euclidean distance to θ_{ref} is also under the resolution (Figure 1b). Moreover, **SBI-CNN** outperforms **ABC-CNN** that itself outperforms **ABC-Pearson**, reinforcing the use of a summary statistic and the flexibility of the normalizing flows.

In the case of the whole genome, our approaches are not as accurate and could be improved. In some dimensions, θ is estimated at a precision under the resolution (Figure 7b), and we obtain peaky densities but in others the inference is not precise (Figure 2).

An advantage of our method is that we do not rely on any initialization or pre-localization: instead, we use an uninformative prior, setting each centromere randomly in the range of its corresponding chromosome. Also, our approach is naturally scalable: the pre-trained summary statistic could be reused for inference on centromeres of others yeasts without any re-training. To improve our approach, we will focus our efforts in developing a summary statistic independent of the size of the genome via notably the use of transformer architectures [19].

Acknowledgments and Disclosure of Funding

This work was supported by the ANR project BONSAI (grant ANR-23-CE23-0012-01) and by the ANR project Bayes-Duality (grant ANR-21-JSTM-0001).

References

[1] Jean-Michel Arbona, Sébastien Herbert, Emmanuelle Fabre, and Christophe Zimmer. Inferring the physical properties of yeast chromatin through Bayesian analysis of whole nucleus simulations. In *Genome Biology*, 2017.

[2] Kerry S. Bloom. Centromeric heterochromatin: the primordial segregation machine. *Annu. Rev. Genet.*, 48:457–484, 2014.

[3] Jan Boelts, Michael Deistler, Manuel Gloeckler, Álvaro Tejero-Cantero, Jan-Matthis Lueckmann, Guy Moss, Peter Steinbach, Thomas Moreau, Fabio Muratore, Julia Linhart, Conor Durkan, Julius Vetter, Benjamin Kurt Miller, Maternus Herold, Abolfazl Ziaeemehr, Matthijs Pals, Theo Gruner, Sebastian Bischoff, Nastya Krouglova, Richard Gao, Janne K. Lappalainen, Bálint Mucsányi, Felix Pei, Auguste Schulz, Zinovia Stefanidi, Pedro Rodrigues, Cornelius Schröder, Faried Abu Zaid, Jonas Beck, Jaivardhan Kapoor, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. sbi reloaded: a toolkit for simulation-based inference workflows, 2024.

[4] Guillaume Cottarel, James H. Shero, Philip Hieter, and Johannes H. Hegemann. A 125-base-pair CEN6 DNA fragment is sufficient for complete meiotic and mitotic centromere functions in Saccharomyces cerevisiae. *Mol Cell Biol*, 9(8):3342–3349, 1989.

[5] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465:363–367, 2010.

[6] Kyle P Eagen. Principles of chromosome architecture revealed by Hi-C. In *Trends Biochem Sci.*, 2018.

[7] Jonathan L. Gordon, Kevin P. Byme, and Kenneth H. Wolfe. Mechanisms of Chromosome Number Evolution in Yeast. *PLoS Genet*, 7, 2011.

[8] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. In *ICML*, 2019.

[9] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. In *Journal of Machine Learning Research*, 2012.

[10] Maxim Imakaev, Geoffrey Fudenberg, Rachel P. McCord, Natalia Naumova, Anton Goloborodko, Bryan Lajoie, Job Dekker, and Leonid Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9:999–1003, 2012.

[11] Bai Jiang, Tung yu Wu, Charles Zheng, and Wing H. Wong. Learning summary statistic for approximate Bayesian computation via deep neural network. In *Statistica Sinica*, 2018.

[12] Philippe Lefrançois, Ghia M. Euskirchen, Raymond K. Auerbach, Joel Rozowsky, Theodore Gibson, Christopher M. Yellman, Mark Gerstein, and Michael Snyder. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, 10(37), 2009.

[13] Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Gianni Liti, Gilles Fischer, Christophe Zimmer, and Romain Koszul. Filling annotation gaps in yeast genomes using genome-wide contact maps. In *Bioinformatics*, 2014.

[14] Angela Nietzel, Mariano Rocchi, Heike Starke, Anita Heller, Wolfgang Fiedler, Iwona Wlodarska, Ivan Loncarevic, Volkmar Beensen, Uwe Claussen, and Thomas Liehr. A new multicolor-FISH approach for the characterization of marker chromosomes: centromere-specific multicolor-FISH (cenM-FISH). *Human Genetics*, 2001.

[15] George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In *NeurIPS*, 2016.

[16] Harianto Tjong, Ke Gong, Lin Chen, and Frank Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. In *Genome Research*, 2012.

[17] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. In *Journal of The Royal Society Interface*, 2008.

[18] Nelle Varoquaux, Ivan Liachko, Ferhat Ay, Joshua N Burton, Jay Shendure, Maitreya J Dunham, Jean-Philippe Vert, and William S Noble. Accurate identification of centromere locations in yeast genomes using Hi-C. In *Nucleic Acids Research*, 2015.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[20] Joachim Wolff, Rolf Backofen, and Björn Grüning. Loop detection using Hi-C data with HiCExplorer. In *GigaScience*, volume 11, 2022.

A Contact maps

A contact map summarizes all the chromatin contacts observed over a population of DNA configurations. To construct it, we define the resolution of the map (the length of the chromosome fragment that will represent one pixel in the map). Each chromosome is then cut into fragments and each entry of the map represents the contact counts of any fragment with another over the population of DNA. This creates a matrix by blocks of interactions between chromosomes. Usually, we represent them by a heatmap.

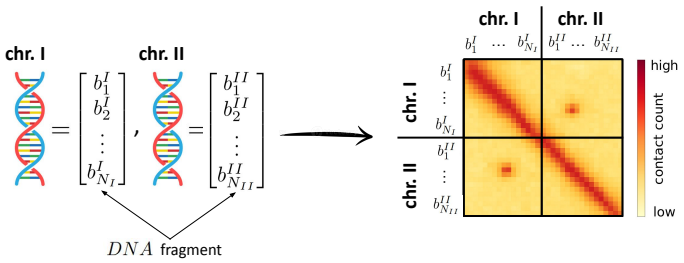

Figure 3: Process to construct a contact map in the case of 2 chromosomes.

B The simulator

The goal of the simulator is to create the upper trans-contact blocks of a contact map C rapidly given the centromere positions θ. We want to mimic the peak of interaction that appears in those blocks, as well as some rare interactions that can occur among the population of DNA.
Given the L chromosome lengths in bp $\{l_i\}_{1 \leq i \leq L}$, the centromere positions θ are sampled from the prior $\mathcal{U}(\prod_{1 \leq i \leq L} [1, l_i - 1])$. To create each contact map C, the process is described in Algorithm 1.

Algorithm 1 Simulator of contact maps

Input: L chromosome lengths in bp $\{l_i\}_{1 \leq i \leq L}$, resolution of the contact map in bp r (e.g. $r = 32$ kb), centromere positions θ
Return: the upper trans-contact blocks of a simulated contact map C at the resolution r bp.

choose the size of the peaks of interaction: sample σ^2 from $\mathcal{U}(0.1, 10)$
choose the intensity of interaction α to simulate the DNA population size: sample α from $\mathcal{U}([\![1, 1000]\!])$
for each chromosome pair (i, j) **do**
  define a block of interaction C_{ij} of size $(\frac{l_i}{r}, \frac{l_j}{r})$
  define the center of the peak (θ_i, θ_j)
  apply Gaussian density $\mathcal{N}((\theta_i/r, \theta_j/r), \sigma^2)$ to the pixels of the block C_{ij}
  multiply each pixel of C_{ij} by the intensity factor α
  add Gaussian noise up to 10% of the maximal value of C_{ij} to mimic the rare contacts:
  construct a random matrix M_{ij} of size $(\frac{l_i}{r}, \frac{l_j}{r})$ where each pixel is sampled from $\mathcal{N}(\max(C_{ij}) \times 0.05, (\max(C_{ij}) \times 0.05)^2)$, then add M_{ij} to C_{ij}
end for
return a simulated contact map C at resolution r bp

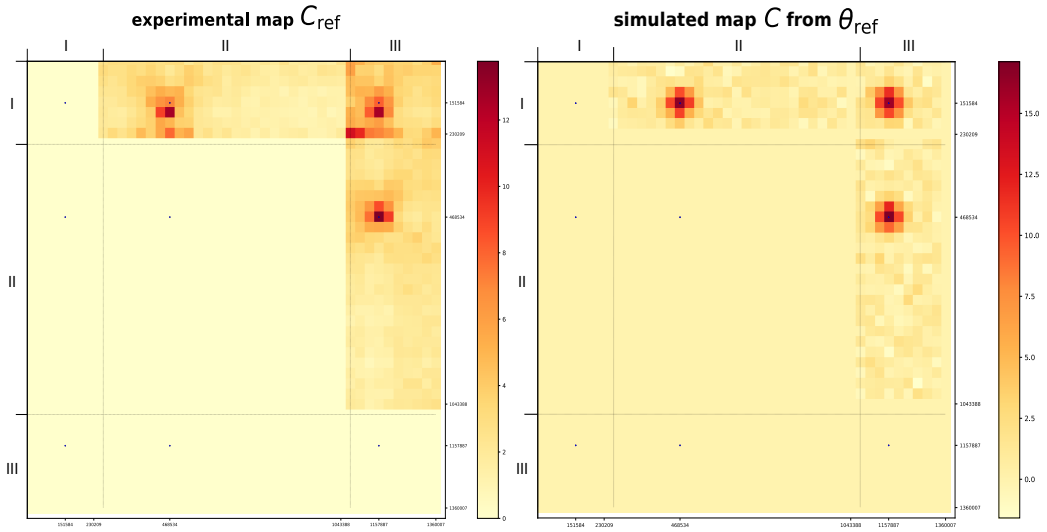

Figure 4: Hi-C map and our simulated map in the case of a small genome (resolution 32 kb).

Our simulator outputs contacts maps that present some dissimilarities with Hi-C maps. If we compute the row-based averaged Pearson correlation between C_{ref} and C simulated from θ_{ref} as in [16], we get a correlation of 0.18 in the case of 3 chromosomes and 0.12 in the case of 16 chromosomes, which is quite low.
However, concerning the inference task, our simulator estimates the centromere positions θ nearly as well on synthetic data (Figure 5) as on Hi-C data (Figure 1).

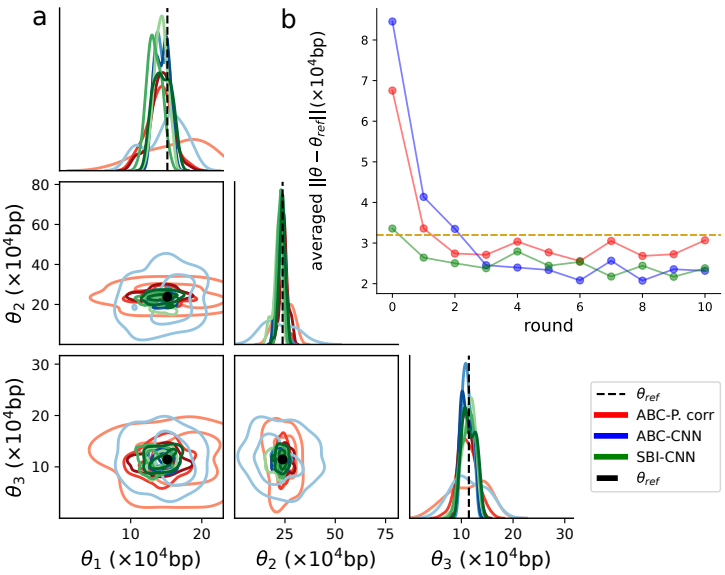

Figure 5: Inference using **ABC-Pearson**, **ABC-CNN**, and **SBI-CNN** from synthetic data (**a**). Color shades increase from lightest to darkest across rounds. Densities are estimated with the 5% best θ according to the ABC criterion or sampled from the flow. We also report the mean Euclidean distance between θ and θ_{ref}, computed over the 5% best-performing samples in the top right corner (**b**). The horizontal dashed line stands for the resolution of the contact map C (in bp) in the top right figure. Results with data-driven summary statistics approaches are uniformly better even if all approaches have errors smaller than the resolution of the contact maps.

C SMC-ABC

C.1 With the metric Pearson correlation – ABC-Pearson

One of the inference methods used is sequential ABC with the metric vector-based Pearson correlation averaged over all trans-contacts blocks.

Algorithm 2 SMC-ABC based on Pearson correlation inspired from [17]

Input: T rounds, prior π, train set of size N, acceptance size M, perturbation kernel $K = \mathcal{N}(., \sigma^2\text{Id})$ ($\sigma =$ resolution (bp))
Return: $\theta \sim p(\theta|\text{corr}(C, C_{\text{ref}}) \geq \epsilon_{\text{corr}})$
round $t = 0$
- sample $\theta^n \sim \pi$, and $C^n \sim p(.|\theta^n), n \in [\![1, N]\!]$
- compute $\text{corr}(C^n, C_{\text{ref}})$ and keep the top 5% of $\{\theta^n\}_n$ in terms of the highest correlation: $\{\theta^{m,0}, m \in [\![1, M]\!]\}$
- compute weights $\{w^{m,0} = \frac{1}{M}, m \in [\![1, M]\!]\}$
output round $t = 0$: $\{(\theta^{m,0}, w^{m,0})\}_{m \in [\![1,M]\!]}$
for $0 < t < T$ **do**
round t
- from the previous accepted $\{\theta^{m,t-1}\}_{m \in [\![1,M]\!]}$, sample $\{\bar{\theta}^k, k \in [\![1, M]\!]\}$ from multinomial $\mathcal{M}(\{\theta^{m,t-1}\}_m, \{w^{m,t-1}\}_m)$ with replacement
- perturb $\frac{N}{M}$ times the M samples $\bar{\theta}^k$ to have N samples θ^n

$$\theta^n \leftarrow \bar{\theta}^k + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2\text{Id}) \text{ for } k = n \text{ mod } M \text{ and } n = 1, ..., N$$

- check that θ^n is in the prior bound otherwise, set $\theta^n \leftarrow \bar{\theta}^k$
- from this set $\{\theta^n\}_{n \in [\![1,N]\!]}$, sample $C^n \sim p(.|\theta^n), n \in [\![1, N]\!]$
- compute $\text{corr}(C^n, C_{\text{ref}})$ and keep the top 5% of $\{\theta^n\}_n$ in terms of the highest correlation: $\{\theta^{m,t}, m \in [\![1, M]\!]\}$
- compute corresponding weights

$$w^{m,t} = \frac{\pi(\theta^{m,t})}{\sum_{k=1}^{M} w^{k,t-1} K(\theta^{m,t}; \theta^{k,t-1})}$$

output round t: $\{(\theta^{m,t}, w^{m,t})\}_{m \in [\![1,M]\!]}$
end for
return accepted samples $\theta^n \sim p(\theta|\text{corr}(C^n, C_{\text{ref}}) \geq \epsilon_{\text{corr}})$

When $\epsilon_{\text{corr}} \to 1$, $p(\theta|\text{corr}(C, C_{\text{ref}}) \geq \epsilon_{\text{corr}}) \to p(\theta|C_{\text{ref}})$.

C.2 With a summary statistic and the classical l^2-norm – ABC-CNN

The other ABC approach uses a pre-learned summary statistic S_ϕ.

Algorithm 3 ABC with learned summary statistic inspired from [11]

Input: (deep) neural network (DNN) S_ϕ, threshold ϵ, Euclidean norm in $\mathbb{R}^n$, simulator, prior p
Return: Samples θ from the estimated posterior density $p(. \mid \|S_\phi(C) - S_\phi(C_{\text{ref}})\| \leq \epsilon)$

Stage 1: learn the summary statistic $S_\phi(.)$ s.t. $S_\phi(C) \approx \mathbb{E}[\theta|C]$
generate a train set (θ^n, C^n) from $p(\theta)p(C|\theta)$
train a DNN S_ϕ on this train set with the loss to minimize in ϕ

$$\widehat{\mathcal{L}}_{\text{DNN}}(\phi) = \frac{1}{N} \sum_{1 \leq n \leq N} \|S_\phi(C^n) - \theta^n\|_2^2$$

output $S_\phi(.)$ s.t. $S_\phi(C) \approx \mathbb{E}[\theta|C]$
Stage 2: run ABC with the learned summary statistic S_ϕ **and the criterion** $\|S_\phi(C) - S_\phi(C_{\text{ref}})\| \leq \epsilon$
return accepted samples $\theta^n \sim p(. \mid \|S_\phi(C^n) - S_\phi(C_{\text{ref}})\| \leq \epsilon)$

For S_ϕ informative enough, and when $\epsilon \to 0$,

$$p(\theta \mid \|S_\phi(C) - S_\phi(C_{\text{ref}})\| \leq \epsilon) \to p(\theta|S_\phi(C_{\text{ref}})) \approx p(\theta|C_{\text{ref}}).$$

D SNPE – SBI-CNN

The last inference approach is SNPE based on normalizing flows and the pre-learned summary statistic S_ϕ.
It is a sequential method: in the first round, θ is drawn from an uninformative prior. From the next rounds, θ is drawn from a proposal: the posterior density estimated from the previous round. This way, θ is more informative about C_{ref} and the inference is expected to be refined across rounds.

Algorithm 4 SNPE inspired from [15] and [8]

Input: T rounds, posterior density estimator p_ψ, simulator, prior p, simulation budget N, observation C_{ref}, pre-learned summary statistic S_ϕ
Return: The estimated posterior density $p_\psi(.|S_\phi(C_{\text{ref}}))$

for round $t = 1, ..., T$ **do**
  if $t = 1$ **then** $p_t = p$
  end if
  for $n = 1, ..., N$ **do**
    sample $\theta^n \sim p_t$
    sample $C^n \sim p(.|\theta^n)$
  end for
  train the posterior estimator p_ψ on $\mathcal{D} = \{(\theta^n, C^n)\}_n$ with the loss to minimize in ψ

$$\widehat{\mathcal{L}}_{\text{NPE}}(\psi) = -\frac{1}{N} \sum_{1 \leq n \leq N} \log p_\psi(\theta^n|S_\phi(C^n))$$

  use p_ψ to construct the estimated posterior : $p_\psi(.|S_\phi(C_{\text{ref}}))$.
  define the proposal for the next round : $p_t(\theta) = p_\psi(\theta|S_\phi(C_{\text{ref}}))$
end for
return samples $\theta^n \sim p_\psi(\theta|S_\phi(C_{\text{ref}}))$

E Small genome inference

We work with the *S. cerevisiae*'s first three chromosomes. θ is directly inferred from the entire contact map C_{ref}. We present a benchmark of metrics to assess the performance of the different inference methods : they evaluate both the proximity of the samples to θ_{ref} and the closeness of the densities to the 'true' posterior $\delta_{\theta_{\text{ref}}}$.

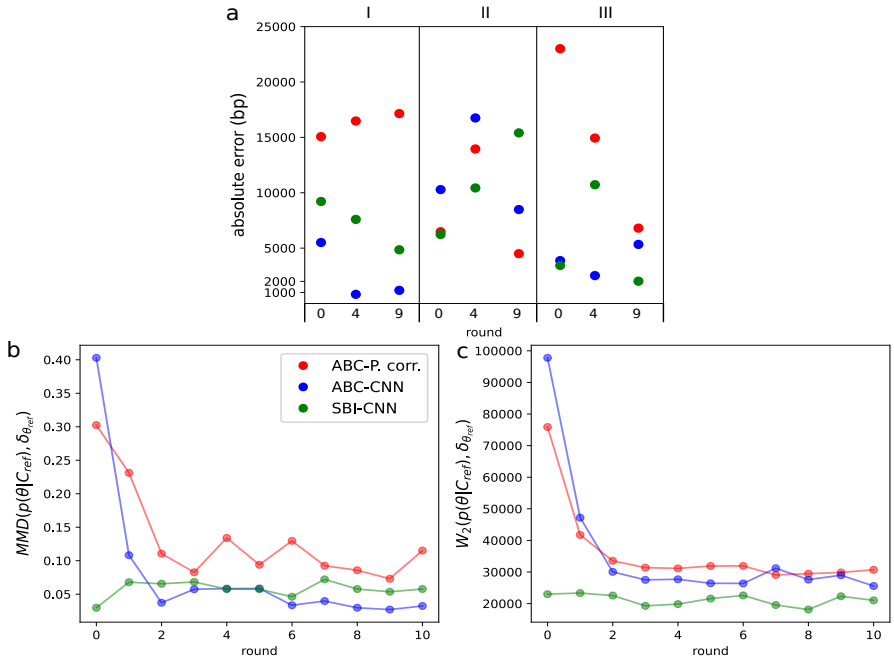

Figure 6: We report the absolute error per dimension of θ between the mean computed over the 5% best-performing samples and θ_{ref} (**a**) as well as the Maximum Mean Discrepancy (MMD) (**b**) and the Wasserstein-2 distance (**c**) between $p(\theta|C_{\text{ref}})$ and $\delta_{\theta_{\text{ref}}}$ [9].

F Whole genome inference

To reduce the dimension of the problem, we carry 16 parallel inferences: one per dimension of θ. Thus, we have 16 $1D$ inference problems where the parameter θ_i is drawn from a Uniform prior whose range is the size of the chromosome i in bp. The simulator creates the ith row of trans-contact blocks of a contact map C (denoted C_i). All the inference methods target the posterior $p(\theta_i|C_{\text{ref},i})$. We need also to learn 16 summary statistics $\{S_{\phi_i}\}_i$ to project each row of trans-contact blocks C_i to θ_i.

S_{ϕ_i} is a CNN to capture the information of C_i followed by an MLP to project this information into θ_i. On the one hand, as the rows of trans-contact blocks C_i are quite similar, we choose a shared architecture for the CNN between chromosomes. On the other hand, each MLP depends on the size of each chromosome so a chromosome-specific architecture is thus needed for this part of the DNN. For the SBI method, we also need to learn 16 normalizing flows. As for the 3-chromosomes case, we choose a MAF as well as SNPE-C for the experiments.

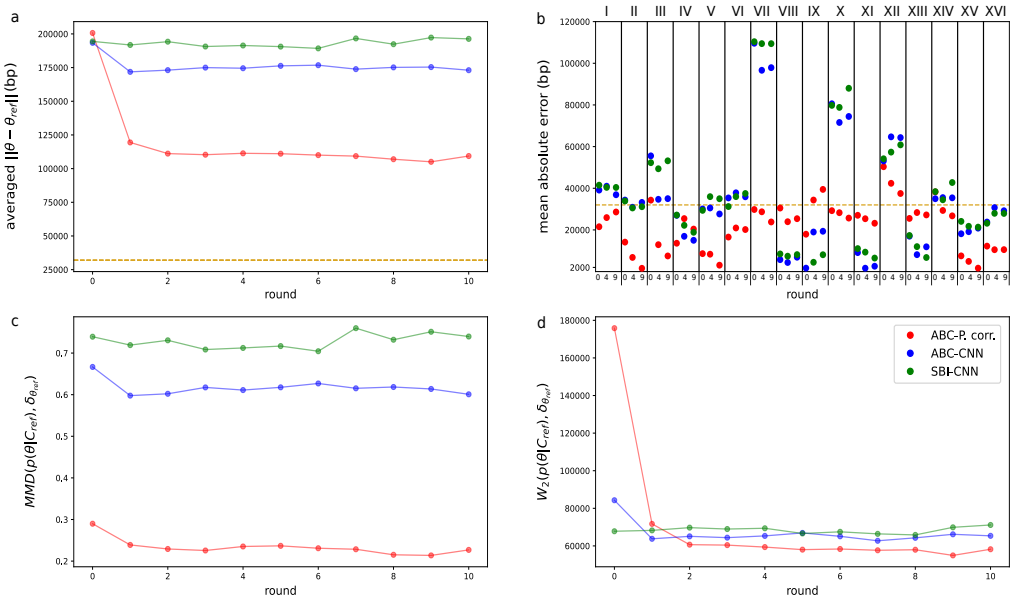

Figure 7: We report the mean Euclidean distance between θ and θ_{ref} (**a**), computed over the 5% best-performing samples, the absolute error per dimension of θ between the mean θ computed over the 5% best-performing samples and θ_{ref} (**b**) as well as the MMD (**c**) and the Wasserstein-2 distance (**d**) between $p(\theta|C_{\text{ref}})$ and $\delta_{\theta_{\text{ref}}}$. The horizontal dotted line stands for the resolution of the contact map C_{ref} (in bp) in the top figures.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present in abstract an inference method whose strengths and weaknesses are detailed all along the paper through real case examples.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We point out the limitations of our method that is not as precise as the other existing ones. We also propose ideas for further improvements.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There are no theoretical results in the paper, all theoretical results are taken from other papers (cited in the text) and applied in practice in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the algorithms as well as the simulator are described in the main text and detailed in appendices. Experimental details (hyper parameters, network architecture,...) are also provided for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: No open source code is available for the moment as it is still a work in progress.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The train set size as well as training details (optimizer, learning rate, acceptance rate,...) are provided either in the core of the paper or in appendices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide a bench of metrics (based on the samples and on the distributions) to measure the performance of the inference methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details about computer resources to run the inference methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We would like to render our inference approach more flexible, thus being adaptable to other species. This way, our work could have greater impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate

deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not manipulate models with high risk. We designed a simplified simulator that we use and the biological data used as reference are open access.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use our custom-designed simplified simulator, and the algorithm for the inference are either implemented by ourself or open source.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We detail the experiments done as well as the data used.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We only use open source or hand-written data/code.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not need participants for our experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: We do not use LLMs to find ideas or methods for our research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.