Amortized Simulation-Based Inference of Colliding-Wind Binaries from Short, Noisy Image Time Series

Niklas Knöll [1 2] **Tobias Buck** [1 2] **Lorenzo Branca** [1 2] **Giuseppe Viterbo** [1 2]

Abstract

Colliding-wind binaries (CWBs), which are systems of two massive stars whose supersonic winds collide into bow shocks, encode rich information about stellar wind properties in their multi-frequency emission, e.g. images in the Hα, X-ray, and radio wavelengths. Inferring physical parameters (mass-loss rates, terminal wind velocities, orbital elements) from short time-series observations is a compelling but challenging inverse problem, because the forward hydrodynamic simulator is computationally expensive and the likelihood is intractable. We adopt a factorized spatio-temporal architecture for amortized posterior inference that separates spatial encoding from temporal aggregation. This design aligns with the structure of the underlying physical process of local morphology and global dynamical evolution, induces time-translation equivariance in the learned representation, and improves identifiability in low-signal regimes. Coupled with a neural spline flow conditioned on these spatio-temporal embeddings of 10-frame Hα photon-count time series, we present a complete simulation-based inference pipeline for CWBs. Our method jointly infers seven physical parameters from synthetic observations under realistic detector noise, with posteriors verified as well-calibrated via TARP and SBC diagnostics. The approach naturally expands posterior width in information-poor regimes (low photon counts) and robustly recovers orbital parameters and mass-loss rates, demonstrating the feasibility of amortized likelihood-free inference for this challenging astrophysical inverse problem.

[1]Interdisciplinary Center for Scientific Computing, Heidelberg, Germany [2]University Heidelberg, Heidelberg, Germany. Correspondence to: Tobias Buck <tobias.buck@iwr.uni-heidelberg.de>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Stellar winds are continuous outflows of mass, momentum, and energy from a star into its surrounding medium, and the dominant driving mechanism depends strongly on stellar type. In low-mass, solar-type stars, the wind is primarily powered by thermal pressure from a hot corona, together with magnetohydrodynamic processes providing additional acceleration. In hot, luminous massive stars however, the wind is launched by the transfer of radiative momentum to the gas via absorption and scattering in spectral lines. Because spectral lines have very large cross sections compared to continuum processes, a small fraction of the stellar luminosity suffices to accelerate substantial mass, producing terminal speeds of order a few times the stellar escape speed in massive stars. These stars therefore lose a large fraction of their mass through line-driven supersonic winds, and when two such stars form a binary their winds collide into a bow-shaped shocked region bounded by two contact discontinuities (Cantó et al., 1996). The resulting colliding-wind binary (CWB) radiates in Hα, thermal X-rays and radio, and its observational appearance encodes information about the mass-loss rates $\dot{M}_{1,2}$, the terminal wind velocities $v_{\infty,1/2}$, and the orbital dynamics of the system (Pejcha et al., 2022). Constraining these quantities is of astrophysical interest, as mass-loss shapes the evolutionary tracks and chemical yields of the most massive stars in the Universe.

In practice, inferring wind and orbital parameters from observations is challenging for two main reasons. First, the forward mapping from parameters to observables involves a three-dimensional compressible hydrodynamics simulation coupled with an N-body integrator, followed by emissivity calculations, projection, and a detector-noise model, which makes it computationally expensive. Second, the induced likelihood is implicit and cannot be evaluated pointwise, so traditional Bayesian workflows are unavailable. These two properties make the problem a natural fit for *Simulation-Based Inference* (SBI; Papamakarios & Murray, 2018; Greenberg et al., 2019; Cranmer et al., 2020; Boelts et al., 2025), where a neural conditional density estimator is trained on a precomputed dataset of simulation–parameter pairs and, once trained, provides amortized posterior estimates for new observations.

In this work, we demonstrate that SBI recovers the physical parameters of CWBs from short time series of synthetic Hα photon-count maps. The core ingredients are (i) a GPU-accelerated, JAX-based fluid solver (astronomix; Storcks, 2025), coupled with an RK4 N-body integrator for the forward model, (ii) a factorized spatio-temporal (2+1)D CNN architecture that is commonly used in video understanding (Donahue et al., 2015; Tran et al., 2018) that separates spatial encoding from temporal aggregation and compresses a 10-frame 64×64 time series into a low-dimensional summary, and (iii) a neural spline flow (Durkan et al., 2019) trained via amortized neural posterior estimation (Papamakarios & Murray, 2018). A critical observation motivates the need for time-series data in our study. The orbital elements of a CWB parametrize the binary's temporal evolution by construction, so a single Hα frame, which captures only one instant of that evolution, is structurally insufficient to constrain them. In projection, several combinations of inclination, orbital phase, eccentricity, and argument of periastron produce essentially indistinguishable 2D morphologies. The temporal evolution carries the information needed to resolve these degeneracies, including changing separation, shock-cone reorientation, and phase-dependent brightening and dimming. Spatial features simultaneously encode the wind-density structure, which motivates our spatio-temporal approach. Additionally, the classical likelihood-free alternative, Approximate Bayesian Computation (Beaumont et al., 2002, ABC), using hand-crafted summary statistics, is in principle applicable to this problem. Nevertheless, even a carefully designed physics-motivated summary fails to constrain all but the mass-loss rates, and at substantially broader posterior widths than the neural estimator, underlining the need for a learned embedding.

The goal of this study consists of three parts. **First**, we develop a complete SBI pipeline for CWBs that jointly infers seven physical parameters from short Hα time series in the presence of realistic detector noise (Fig. 1), with posteriors verified as well-calibrated via TARP and SBC. **Second**, we show that the factorized two-stage spatio-temporal encoder design is well-matched to the physics of CWB observations, and that the temporal information resolves fundamental identifiability degeneracies that are provably unsolvable from static observations. **Third**, we demonstrate that mass-loss rates are recovered across the full prior range, wind velocities and orbital parameters are recovered when wind luminosity is sufficient, and posterior width expands appropriately in information-poor regimes.

2. Related work

SBI has developed rapidly in the last decade, with neural posterior, neural likelihood and neural ratio estimation now available in a common software ecosystem (Boelts et al., 2025; Papamakarios & Murray, 2018; Greenberg et al., 2019; Durkan et al., 2019). Applications in physics include galactic chemical evolution (Gunes et al., 2025; Buck et al., 2025; Viterbo & Buck, 2024), gravitational-wave astronomy (Wildberger et al., 2023), cosmology (e.g. Lemos et al., 2023a), galactic dynamics (Viterbo & Buck, 2026) or supernova inference (Boyd et al., 2026). Within stellar-wind modeling, analytic thin-shell solutions provide semi-quantitative geometric insight (Cantó et al., 1996), but a data-driven, likelihood-free inference pipeline that ingests *images* of CWBs has not been reported before to our knowledge.

3. Problem Setup and Forward Model

We consider a CWB system parameterized by

$$\boldsymbol{\theta} \;=\; \left(\dot{M}_1, \dot{M}_2, v_{\infty,1}, v_{\infty,2}, e, i, \eta\right) \tag{1}$$

where the first four entries are the individual mass-loss rates and terminal wind velocities, e and i the orbital eccentricity and inclination, and $\eta = \dot{E}_{\mathrm{turb}} \,/\, \langle L_{\mathrm{wind}} \rangle$ is the fraction of additional forced random turbulence (scheme adopted and modified from Seo & Ryu, 2023) relative to the mean inserted wind luminosity. The priors, summarized in Table 1, span ranges relevant for massive binaries and realistic geometries. The mass-loss rates and terminal wind velocities are chosen to span the parameter space of observed massive binaries.

Table 1. Prior parameter distributions for our model. $\dot{M}$ is sampled log-uniformly, v_∞, e and η are sampled linear-uniformly and i is sampled from the standard isotropic-inclination prior.

Parameter	Lower	Upper
$\dot{M}_{1,2}$ [$\mathrm{M_\odot\,yr^{-1}}$]	$8{\cdot}10^{-9}$	$1{\cdot}10^{-5}$
$v_{\infty,1/2}$ [$\mathrm{km\,s^{-1}}$]	1200	3200
e	0	0.85
i [deg]	0	90
η	0	0.025

3.1. Physical simulator

The hydrodynamics are solved with a modified version of `astronomix`, a JAX-based, conservative finite-volume fluid solver (Storcks, 2025; Storcks & Buck, 2024). The two stars are represented as point sources with an energy-and-mass wind injection scheme. The mass is deposited at a rate $\dot{M}_s/V$ and the kinetic energy at a rate $\frac{1}{2}v_{\infty,s}^2\,\dot{M}_s/V$ over a spherical injection region of volume V. The momentum source then arises self-consistently through the induced pressure gradient. Gravity is handled in a two-step scheme: the stars are integrated with an explicit fourth-order Runge–Kutta N-body integrator (treating the gas self-gravity as negligible compared to the stellar masses), while the gas

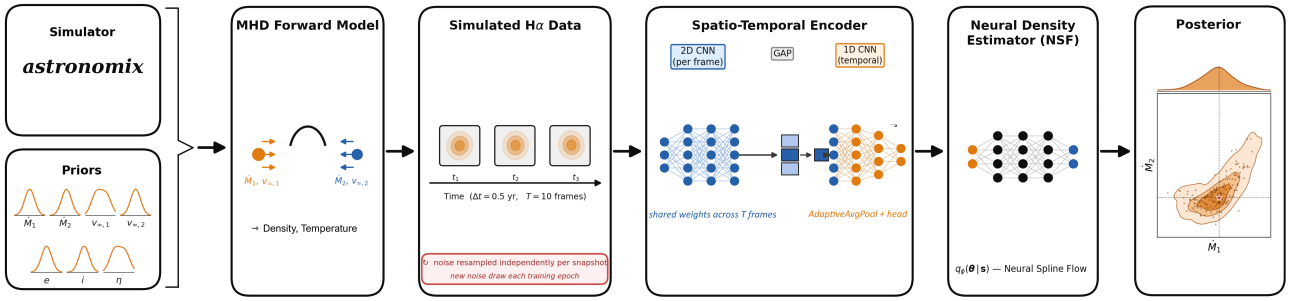


Figure 1. **SBI flow chart for the CWB problem**. From a prior $p(\boldsymbol{\theta})$ over seven physical parameters, we run 3D hydrodynamic simulations with a modified version of the JAX-based fluid solver `astronomix` (Storcks, 2025) together with an orbit integrator and an Hα emissivity + Poisson noise pipeline to produce 10-frame photon-count time series. The resulting $\{(\boldsymbol{\theta}_i, \mathbf{x}_i)\}$ pairs train a neural spline flow with a spatio-temporal CNN embedding as the conditional density estimator, while resampling noise per simulation and per snapshot. Given a new observation, the trained estimator returns an approximate posterior $p(\boldsymbol{\theta} \mid \mathbf{x})$, which allows inference of stellar and orbital parameters.

feels the stars via a Poisson solve in which the stellar masses are deposited onto the grid with a nearest-grid-point kernel. Each simulation is run on a three-dimensional $N = 64^3$ Cartesian grid for a physical time of $T_{\rm end} = 5\,{\rm yr}$, and 10 equally spaced snapshots are stored. The simulations are initialized with a uniform ambient density of $\rho_0 = 2\,m_p\,{\rm cm}^{-3}$ and temperature $T_0 = 15\,000\,{\rm K}$, representing the warm ionized interstellar medium surrounding the binary. In this case, wind ram pressure vastly exceeds ambient thermal pressure, so shock geometry is determined by wind-wind interactions rather than external confinement. The forward problem is scale-free in $(a, \dot{M}, v_\infty)$ up to small orbital-motion asymmetries, so we fix the semi-major axis $a = 10\,{\rm AU}$ throughout. Additionally, we couple the sampled mass-loss rate $\dot{M}$ to a stellar mass via an interpolated empirical mass / mass-loss relation $M_0(\dot{M})$ with a 5% Gaussian spread,

$$M \sim \mathcal{N}\Big(M_0(\dot{M}),\, 0.05\, M_0(\dot{M})\Big)\,. \tag{2}$$

The mapping $M_0(\dot{M})$ is based on stellar evolution models from Ekström et al. (2012), from which stellar wind parameters were calculated following the procedure in Haid et al. (2018). The resulting distribution of orbital periods lies between $P_{\rm min} \approx 2.04\,{\rm yr}$ and $P_{\rm max} \approx 5.85\,{\rm yr}$, so that $\sim 94\%$ of sampled systems complete at least one orbit within the observational window. This range covers the massive eccentric CWB η Carinae ($P \approx 5.54\,{\rm yr}$; Damineli et al., 2008). Inference on longer-period systems is accessible to the same pipeline by retraining at an adjusted a with a correspondingly extended $T_{\rm end}$, at the cost of additional simulation budget. Appendix A details the empirical $M_0(\dot{M})$ relation (Fig. 6) and the derived prior for the orbital period.

3.2. From the forward simulation to synthetic observations

Starting from the simulated density and pressure fields, the temperature is obtained from the ideal gas law, $T = \mu\, m_H P/(k_B \rho)$. Assuming a fully photoionized H II region around the hot, massive stars (following the approach in Green, Samuel et al., 2019) we compute the Hα emissivity from an interpolated table in Osterbrock (1989):

$$j_{H\alpha} \;=\; 2.63{\cdot}10^{-33}\,\frac{n_e n_H}{T^{0.9}}\;{\rm erg\,cm^{-3}\,s^{-1}\,arcsec^{-2}}, \tag{3}$$

with $n_e = 0.86\,\rho/m_p$ and $n_H = 0.71\,\rho/m_p$. In the low-density CWB environment we adopt the optically thin limit of the radiative transfer equation, so that the specific intensity along the line of sight reduces to $J = \int j_{H\alpha}\,{\rm d}l$. Due to the quadratic dependence of $j_{H\alpha}$ on the density via $n_e n_H$, the emission is dominated by the bow-shock apex and the dense wind-collision region between the two contact discontinuities (see e.g. Fig. 2). The intensity is then converted into an expected observed photon count by adopting representative instrument parameters (telescope diameter $D = 2.4\,{\rm m}$, angular solid angle $A_{\rm ap} = 0.04\,{\rm arcsec}^2$, exposure time $t_{\rm exp} = 600\,{\rm s}$, throughput $\eta_{\rm tel} = 0.11$). Finally, a per-pixel noise model is applied. The most important contributions are given by the shot noise (Poisson photon-counting statistics) with variance $\sigma_{\rm Poisson} = \sqrt{N}$, which dominates the low/medium count regime ($N \lessapprox 10^4$), and a flat-field calibration uncertainty with fractional amplitude $\epsilon_{\rm flat} = 0.01$, which dominates the high count regime. The resulting dependence between physical density, emissivity and intensity for a representative simulation (without applied noise) is shown in Fig. 2.

During training, each of the 10 snapshots in a simulation receives an independent noise realisation that is resampled between epochs. The network is therefore exposed to many possible observational outcomes of the same underlying physical system, which both acts as augmentation and prevents the network from memorizing a specific noise pattern.

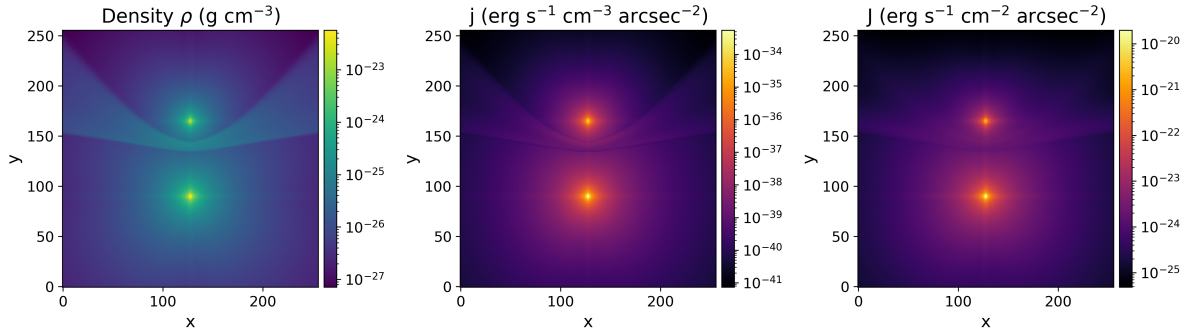


Figure 2. Forward-model pipeline for a representative CWB ($N = 256^3$; x and y in grid units). Left to right: mass-density slice in the orbital plane; Hα emissivity $j_{H\alpha}$ (orbital plane slice); and line-of-sight-integrated Hα intensity J. The intensity traces the high-density bow-shock apex and the wind-collision region bounded by the two contact discontinuities.

4. Simulation-Based Inference

We target the posterior $p(\boldsymbol{\theta} \mid \mathbf{x})$ over the seven physical parameters of Eq. 1 given an observation $\mathbf{x} \in \mathbb{R}^{T\times H\times W}$ consisting of $T = 10$ noisy Hα photon-count maps of size $H = W = 64$. Following the neural-posterior-estimation (NPE) paradigm (Papamakarios & Murray, 2018), we train a conditional density estimator $q_{\boldsymbol{\phi}}(\boldsymbol{\theta} \mid \mathbf{s})$ with parameters $\boldsymbol{\phi}$ by minimizing the expected negative log-likelihood of training parameters given the corresponding simulation summaries,

$$\mathcal{L}(\boldsymbol{\phi}) \;=\; \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{x})}\big[-\log q_{\boldsymbol{\phi}}(\boldsymbol{\theta} \mid f_\psi(\mathbf{x}))\big], \tag{4}$$

where $\mathbf{s} = f_\psi(\mathbf{x})$ is a neural embedding of the raw time-series cube and is trained jointly with $q_{\boldsymbol{\phi}}$. We parameterize $q_{\boldsymbol{\phi}}$ as a neural spline flow (Durkan et al., 2019), which we found to outperform masked autoregressive flows and mixture density networks in preliminary experiments.[1] The spline's local curvature accommodates the curved, non-linear degeneracies typical of CWB posteriors more efficiently than affine coupling blocks. All training is performed with the `sbi` python package (Boelts et al., 2025).

4.1. Spatio-temporal embedding

The embedding f_ψ is a two-stage factorized spatio-temporal encoder, following approaches in video understanding (Donahue et al., 2015; Tran et al., 2018). The architecture has several discrete hyperparameters (written below as the sets searched over), which are tuned jointly with the flow in Sec. 4.2. The selected values are reported in Table 2. In the first stage, a shared 2D CNN $\mathcal{E}$ of depth $L \in \{2, 3\}$ with 3×3 strided convolutions and a doubling channel schedule $C_\ell = \min(2C_{\ell-1}, 512)$ processes each frame independently, followed by global average pooling over the spatial axes to yield a per-frame embedding $\mathbf{e}_t$. In the second stage, the sequence $\{\mathbf{e}_t\}_{t=1}^{T}$ is passed through $L_t \in \{1, 2, 3\}$ 1D convolutions of kernel size $k_t = 3$, then adaptively average-pooled to $T_{\text{out}} \in \{1, 2, 4\}$ temporal bins and mapped to the final summary by a single fully-connected head of width d_{fc}. The overall map can be written compactly as

$$f_\psi(\mathbf{x}) \;=\; h \circ \mathcal{A} \circ \text{gap} \circ \mathcal{E}, \tag{5}$$

where $\mathcal{E}$ is the shared per-frame 2D encoder, gap denotes global average pooling over the spatial axes, $\mathcal{A}$ is the temporal aggregator, and h is the fully-connected output head.

The choice of this embedding, rather than a single 3D CNN over the full $T \times H \times W$ cube, is motivated by how the physical parameters are encoded in the observations. First, the morphology of each individual Hα frame already carries most of the information about wind momentum ratio, bow-shock opening angle and projected inclination (see Fig. 7 for an illustration), which is a purely spatial inference task that a shared 2D encoder solves equally effectively at every time step. Second, the temporal dimension is short ($T = 10$) and carries coherent global changes (orbital phase, periastron-driven brightening in eccentric orbits), not local spatio-temporal patterns.

Sharing $\mathcal{E}$ across frames makes the per-frame feature-extraction strictly time-translation equivariant. An observation shifted in orbital phase produces a correspondingly shifted sequence of frame embeddings, which the temporal aggregator $\mathcal{A}$ can then process consistently. *This matches the physical symmetry that the orbital dynamics are invariant under a shift of the time origin*, and ensures that the network does not have to learn this invariance separately at every spatial layer.

4.2. Hyperparameter search

Both the embedding and the flow contain free hyperparameters. We search jointly over nine integer/categorical variables (five for the embedding CNN, two for the temporal ag-

[1] The most important parts of our pipeline, including forward simulator, noise formalism, hyperparameter tuning and calibration routines, are available publicly in this GitHub repository: https://github.com/n-knoell/SBI-for-CWBs

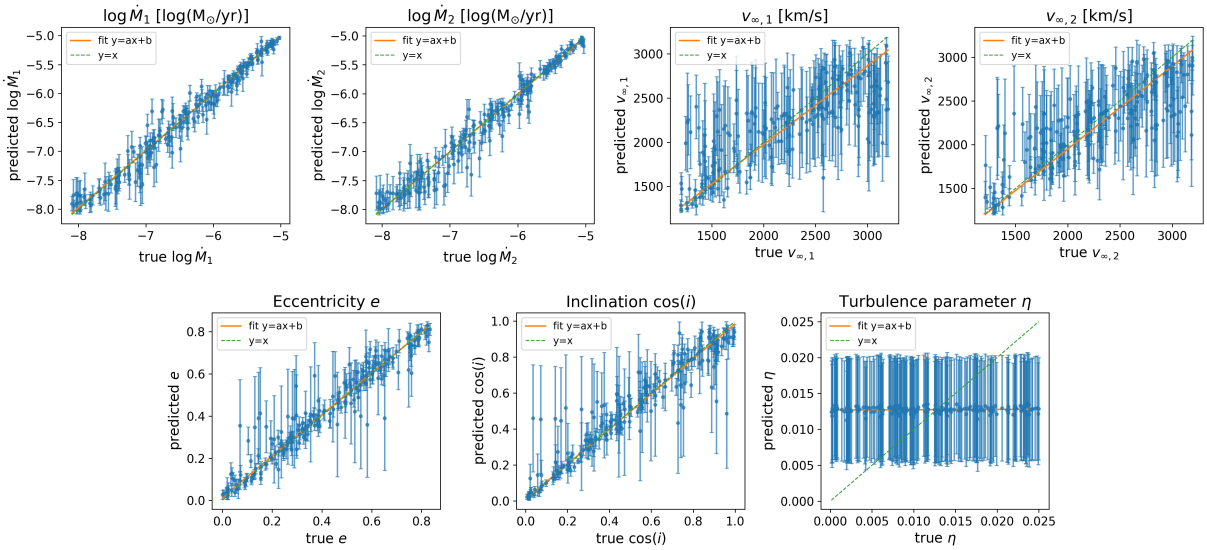


Figure 3. True vs. inferred values for a $N = 200$ subset of the test simulations. Error bars are 1σ intervals estimated from 2000 posterior samples. Mass-loss rates are recovered accurately over the 4-decade range, while terminal wind velocities slightly compress to the mean. The eccentricity e and the inclination i show phase-dependent recoverability. The turbulence parameter η remains at the prior.

gregator, and two for the flow: hidden width $H_{\text{flow}} \in [4, 30]$ and number of transforms $N_{\text{tr}} \in [3, 25]$) using Optuna's NSGA-II multi-objective sampler (Akiba et al., 2019; Deb et al., 2002). Two objectives are minimized simultaneously: the validation NLL and the mean TARP deviation from the ideal coverage curve (Sec. 5). The 128-trial Pareto front optimization plot is shown in the appendix (Fig. 8).

The final-selected configuration has $L = 3$ convolutional blocks with base width $C_1 = 32$, a final embedding dimension $d_{\text{fc}} = 128$, $L_t = 2$ temporal layers with stride $s_t = 1$, temporal output $T_{\text{out}} = 2$, and a neural spline flow with $N_{\text{tr}} = 20$ transforms of width $H_{\text{flow}} = 7$ (Table 2). The learning rate for the Adam optimizer is $5 \cdot 10^{-4}$. The flow's narrow per-transform width is a consequence of the problem's low intrinsic parameter dimensionality, which favours deeper rather than wider flows.

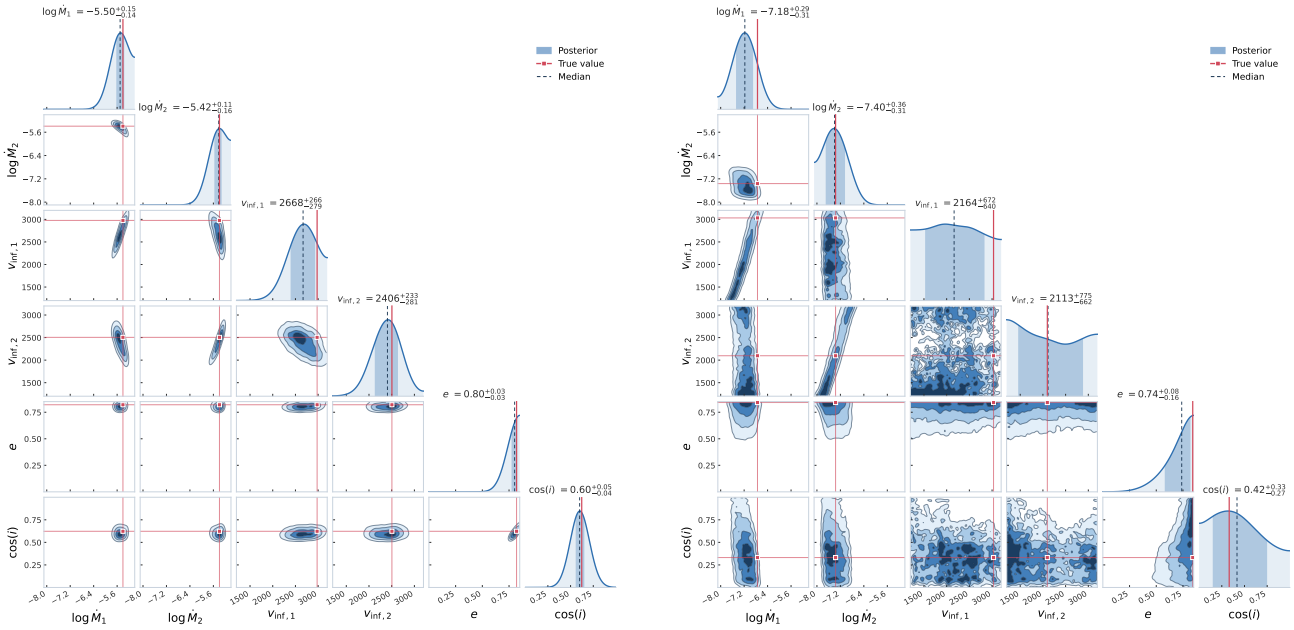
Table 2. Best hyperparameters from the Optuna search

Hyperparameter	Value
Conv blocks L	3
Base channels C_1	32
Final FC width d_{fc}	128
Temporal layers L_t	2
Temporal first-layer stride s_t	1
Temporal pool bins T_{out}	2
Flow transforms N_{tr}	20
Flow hidden features H_{flow}	7

5. Experiments and Results

5.1. Calibration

We generate $N = 40,000$ simulations for training and $1,000$ additional simulations for held-out testing. On a single NVIDIA H200 GPU, we were able to run approximately $4,100$ simulations per day. The embedding and the flow are optimized jointly via amortized neural posterior estimation (Papamakarios & Murray, 2018). We assess posterior calibration along two complementary axes. Tests of Accuracy with Random Points (TARP; Lemos et al., 2023b) probes the *joint* 7-D credibility coverage by constructing posterior balls around random reference points and measuring empirical containment probabilities, and perfect calibration produces a diagonal coverage curve. Simulation-based calibration (SBC; Talts et al., 2018) probes the *marginal* calibration by comparing the distribution of true-parameter ranks among posterior samples against the expected uniform distribution. Both diagnostics are shown in Fig. 10 of the appendix. The TARP result is computed on the full $N = 1000$ held-out test set (with 1000 posterior samples per test case) with minimal coverage deviation from the optimum (Average Test Coverage ATC $= -0.061$), which indicates that the full 7-D posterior is well calibrated. On the marginal level the flow shows mild overconfidence in eccentricity and mild underconfidence in $\dot{M}_2$. All other parameters are within statistical tolerance of the expected distribution under uniformity.



(a) Case 1 (high wind luminosity, low relative noise).

(b) Case 2 (low wind luminosity, high relative noise).

Figure 4. Inferred posterior distributions and parameter correlations for two reference CWBs. Diagonal panels show the one-dimensional marginal posteriors with the shaded band marking the central 68% credible interval and the dashed line the posterior median. Off-diagonal panels show the two-dimensional joint posteriors as filled contours enclosing the 0.5σ, 1σ, 1.5σ and 2σ credible regions. Solid red lines indicate the true parameter values. In the high-luminosity regime (a) all parameters (except η) are tightly constrained. In the low-luminosity regime (b) the shock features are washed out by Poisson noise, essentially all information about the terminal wind velocities is lost, and the corresponding marginals approximately return the prior.

5.2. Point estimates and posteriors

Figure 3 compares the posterior mean and 1σ interval against the true parameters for 200 randomly chosen test simulations. The mass-loss rates are recovered accurately over the four decades of prior range. Terminal wind velocities are systematically recovered but with larger per-simulation error bars and a mild tendency to under-predict at the high end of the prior. Eccentricity and inclination show a bimodal error pattern, with many simulations pinned to narrow posteriors and others retaining significant prior width, reflecting whether the 5-year observational window happens to capture dynamically informative portions of the orbit (particularly the periastron passage). The turbulence parameter η is essentially unconstrained, as the posterior is statistically indistinguishable from the prior across the full range. We believe this is caused by a combination of (i) the artificial detector noise dominating the much weaker turbulent signal and (ii) strided pooling in the embedding filtering out small-scale turbulent texture. We will return to this point in Sec. 6.

Reference cases. Table 3 lists three representative held-out simulations spanning different wind-luminosity regimes. Corner plots of the inferred posteriors for Case 1 (high luminosity) and Case 2 (low luminosity) are shown in Fig. 4.

Table 3. Three reference cases spanning wind-luminosity regimes.

#	$\dot{M}_1$	$\dot{M}_2$	$v_{\infty,1}$	$v_{\infty,2}$	e	$\cos i$	η
1	$4{\cdot}10^{-6}$	$4{\cdot}10^{-6}$	2977	2505	0.82	0.62	0.016
2	$1.8{\cdot}10^{-7}$	$4.3{\cdot}10^{-8}$	3033	2097	0.84	0.33	0.011
3	$8.4{\cdot}10^{-6}$	$1.2{\cdot}10^{-8}$	3080	1740	0.09	0.89	0.007

Case 3 (highly asymmetric wind strengths) is in the appendix (Fig. 12) together with the full time series of Case 1 (Fig. 11). It should be noted that the turbulence parameter η is omitted here, because it resembles the prior in all cases, as already mentioned throughout this work.

In Case 1 both stars generate strong winds and the shocked collision region is dense and bright. As a result, the fractional Poisson noise per-pixel is very low, allowing all parameters except η to be tightly constrained around their true values (see left panel of Fig. 4). In Case 2 the wind luminosities of both stars are very low (the full time series for Case 2 is shown in Fig. 5). Consequently, noise dominates in most snapshots and the stagnation point along with much of the shock structure is only intermittently visible. Mass-loss rates and orbital parameters e and i are still reasonably well estimated (see right panel of Fig. 4), because they are encoded in the overall brightness and large-scale geometry.

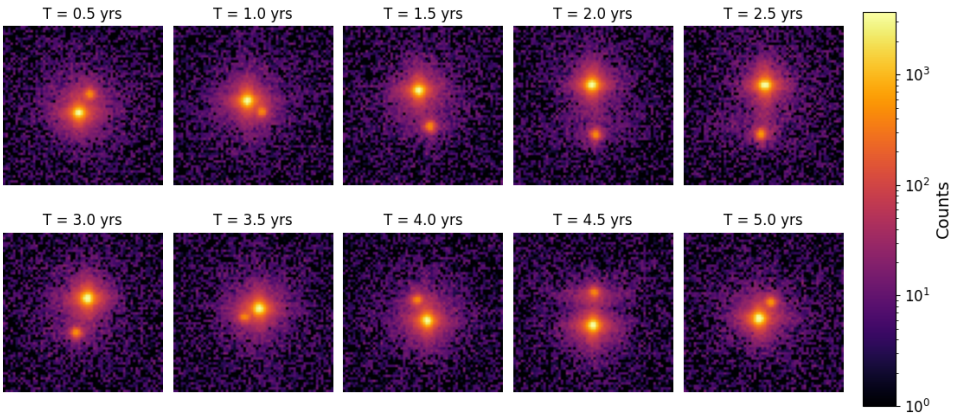


Figure 5. Complete 10-frame Hα photon-count time series for Case 2 (low wind luminosity, high relative noise). Noise dominates in most snapshots and shock features and the stagnation point are only intermittently visible.

However, the terminal-velocity posteriors approximately coincide with the prior, because the distinctive features that encode v_∞ (opening angle, width of the shocked region, and periastron asymmetry) are lost in the noise and most of the shock structure is washed out (see Fig. 5). Case 3 is a highly asymmetric case in which one strong wind dominates. Now, the recovery pattern matches expectations, namely that $v_{\infty,1}$ is well constrained while $v_{\infty,2}$ returns approximately the prior. In addition, eccentricity and inclination are still very well constrained (see Fig. 12 in the appendix).

Effect of training-set size. We retrain the selected architecture on seven subsets $N \in \{10^2, 5{\cdot}10^2, 10^3, 5{\cdot}10^3, 10^4, 2{\cdot}10^4, 4{\cdot}10^4\}$ of the full set of training simulations to map out how posterior quality scales with the available simulation budget. The resulting RMSE and posterior-width curves are provided in Appendix D (Fig. 9). Both metrics remain essentially flat until $N \approx 10^4$ and then improve rapidly until $N \approx 2 \cdot 10^4$. Further small improvement is still observed out to our largest size $N = 4 \cdot 10^4$ without fully plateauing, suggesting that additional training data would still slightly improve RMSE and tighten the posteriors. However, pushing N further is costly, as the $\sim$4,100 sims/day throughput reported above means that an additional $4{\cdot}10^4$ simulations would consume roughly 10 GPU-days on a single NVIDIA H200, and we therefore report results at the largest size afforded by our computational budget. The turbulence parameter η is the only exception to this scaling pattern, remaining uninformative at every value of N tested.

Baseline comparison (ABC). To assess whether the learned spatio-temporal embedding is actually needed, or whether a carefully designed low-dimensional summary would already suffice, we compare against an Approximate Bayesian Computation (ABC) rejection sampler operating on a 62-dimensional hand-crafted summary of the same 10-frame Hα time series. The summary stacks per-frame log-total flux and log-peak flux (to target $\dot{M}$ through $n_e n_H$), brightness-weighted centroids (to capture orbital motion) and second moments (bow-shock extent and, therefore, v_∞), together with two global temporal descriptors that test for phase-driven variability (due to the eccentricity). The full definitions and results of this process are given in Appendix G. Figure 13 shows the resulting ABC posteriors for the same reference high- and low-luminosity cases as in Fig. 4. In both cases ABC recovers only the mass-loss rates, and at substantially broader widths than the neural estimator. Additionally, $v_{\infty,1,2}$, e, $\cos i$ and η effectively return the prior. This is consistent with $\dot{M}$ dominating the integrated Hα flux, while the more subtle morphological and temporal signatures that encode the orbit and wind kinematics are destroyed when $T{\times}H{\times}W = 40{,}960$ pixels are collapsed to 62 scalars. In other words, classical summary-statistic ABC does not scale to the dimensionality of the CWB problem, and the NPE approach is a necessary rather than optional component of our pipeline.

6. Discussion and Limitations

Three empirical patterns emerge from our experiments. First, mass-loss rates are by far the easiest parameter group, because they modulate the overall brightness of the bow-shock quadratically through $n_e n_H$, and this dominant signal survives even in the low-luminosity regime. Second, terminal wind velocities are encoded in higher-order morphological features (bow-shock opening angle, the width of the shocked layer, and the left–right asymmetry induced by orbital motion), which are precisely the features that are first erased by Poisson noise. Third, the turbulence parameter η is never informative. A possible explanation is that this results from a combination of the artificial per-pixel noise with amplitude comparable to the turbulent signal, and

the information-bottleneck character of the spatial-pooling layers, which average out the small-scale turbulent texture that would be needed to determine η. A plausible mitigation would be an embedding that explicitly retains high-frequency spatial content, for example through a parallel branch operating on spectral or wavelet features (e.g. Rost et al., 2025). We note that η was introduced primarily to inject small-scale turbulence into the observations, not as a target of scientific interest. Its non-recovery therefore does not affect the parameters that matter physically (mass-loss rates, wind velocities, and orbital elements).

The ABC baseline shows that even a deliberately rich 62-D hand-crafted summary, built to explicitly target the physical signals in the data, recovers only mass-loss rates (with substantially broader posteriors) and leaves orbital and kinematic parameters at the prior. The learned spatio-temporal embedding therefore enables recovery of parameters that classical methods leave essentially unconstrained.

Model misspecification. All experiments use synthetic data generated by the same simulator that trains the network. A real-world deployment would require addressing the inevitable simulator-observation gap: stellar radiation fields, dust absorption, instrumental point-spread functions, and intrinsic stellar variability are all absent from our forward model. Recent work on robust NPE and misspecified simulators has introduced a range of strategies to identify model misspecification and perform amortized model selection (e.g. Bürkner et al., 2022; Radev et al., 2023; Gunes et al., 2025).

7. Conclusion

We have presented an amortized simulation-based inference framework that recovers the physical parameters of colliding-wind binaries from short, noisy Hα time series, yielding well-calibrated posteriors in a challenging seven-dimensional inverse problem. Our findings point to a broader principle, namely that in many simulator-based settings non-identifiability arises not from model limitations but from insufficient observations. We show that even short time series can break these degeneracies, while single-frame observations remain fundamentally ambiguous. Crucially, this requires architectures that reflect the structure of the data. By adopting a factorized spatial-then-temporal encoder, our model aligns with the underlying physics and improves identifiability compared to unstructured approaches. We further observe that the learned posterior adapts to the information content of the data, tightening in high-signal regimes and reverting to the prior when observations are dominated by noise. This behavior is essential for reliable uncertainty quantification in likelihood-free inference and highlights a key advantage of amortized neural methods over classical approaches. Looking forward, bridging the gap to real observations and integrating multi-modal data will be critical next steps. More generally, our results suggest that structured amortization, which combines inductive biases with flexible density estimators, offers a scalable path toward reliable inference in complex, high-dimensional scientific models.

Impact Statement

This paper presents work that aims to advance the field of machine learning for the physical sciences. We do not foresee direct societal risks or applications beyond the scientific community, and the methods developed here are generic tools for likelihood-free inference on simulator-based forward models.

Acknowledgements

This work is funded by the Carl-Zeiss-Stiftung through the NEXUS program. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). We acknowledge the usage of the AI-clusters Tom and Jerry funded by the Field of Focus 2 of Heidelberg University.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 12 2002. ISSN 1943-2631. doi: 10.1093/genetics/162.4.2025. URL https://doi.org/10.1093/genetics/162.4.2025.

Boelts, J., Deistler, M., Gloeckler, M., Álvaro Tejero-Cantero, Lueckmann, J.-M., Moss, G., Steinbach, P., Moreau, T., Muratore, F., Linhart, J., Durkan, C., Vetter, J., Miller, B. K., Herold, M., Ziaeemehr, A., Pals, M., Gruner, T., Bischoff, S., Krouglova, N., Gao, R., Lappalainen, J. K., Mucsányi, B., Pei, F., Schulz, A., Stefanidi, Z., Rodrigues, P., Schröder, C., Zaid, F. A., Beck, J., Kapoor, J., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. sbi reloaded: a toolkit for simulation-based inference workflows. *Journal of Open Source Software*, 10(108):7754, 2025. doi: 10.21105/joss.07754. URL https://doi.org/10.21105/joss.07754.

Boyd, B. M., Mandel, K. S., Grayling, M., Mitra, A., Kessler, R., Autenrieth, M., Do, A., Ginolin, M., Kelsey, L., Narayan, G., O'Callaghan, M., Sarin, N., and Thorp, S. FlowSN: Neural Simulation-Based Inference under Realistic Selection Effects applied to Supernova Cosmology. *arXiv e-prints*, art. arXiv:2603.11165, March 2026. doi: 10.48550/arXiv.2603.11165.

Buck, T., Günes, B., Viterbo, G., Oliver, W. H., and Buder, S. Inferring Galactic parameters from chemical abundances with simulation-based inference. *A&A*, 702:A184, October 2025. doi: 10.1051/0004-6361/202554306.

Bürkner, P.-C., Scholz, M., and Radev, S. T. Some models are useful, but how do we know which ones? Towards a unified Bayesian model taxonomy. *arXiv e-prints*, art. arXiv:2209.02439, September 2022. doi: 10.48550/arXiv.2209.02439.

Cantó, J., Raga, A. C., and Wilkin, F. P. Exact, Algebraic Solutions of the Thin-Shell Two-Wind Interaction Problem. *ApJ*, 469:729, October 1996. doi: 10.1086/177820.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117.

Damineli, A., Hillier, D. J., Corcoran, M. F., Stahl, O., Levenhagen, R. S., Leister, N. V., Groh, J. H., Teodoro, M., Colombo, J. F. A., Gonzalez, F., Arias, J., Levato, H., Grosso, M., Morrell, N., Gamen, R., Wallerstein, G., and Niemela, V. The periodicity of the η carinae events. *Monthly Notices of the Royal Astronomical Society*, 384(4):1649–1656, March 2008. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2007.12815.x. URL http://dx.doi.org/10.1111/j.1365-2966.2007.12815.x.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197, 2002. doi: 10.1109/4235.996017.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., and Saenko, K. Long-term recurrent convolutional networks for visual recognition and description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2625–2634, 2015. doi: 10.1109/CVPR.2015.7298878.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows, 2019. URL https://arxiv.org/abs/1906.04032.

Ekström, S., Georgy, C., Eggenberger, P., Meynet, G., Mowlavi, N., Wyttenbach, A., Granada, A., Decressin, T., Hirschi, R., Frischknecht, U., Charbonnel, C., and Maeder, A. Grids of stellar models with rotation - i. models from 0.8 to 120 m solar metallicity (z = 0.014). *A&A*, 537:A146, 2012. doi: 10.1051/0004-6361/201117751. URL https://doi.org/10.1051/0004-6361/201117751.

Green, Samuel, Mackey, Jonathan, Haworth, Thomas J., Gvaramadze, Vasilii V., and Duffy, Peter. Thermal emission from bow shocks - i. 2d hydrodynamic models of the bubble nebula. *A&A*, 625:A4, 2019. doi: 10.1051/0004-6361/201834832. URL https://doi.org/10.1051/0004-6361/201834832.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2404–2414. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/greenberg19a.html.

Gunes, B., Buder, S., and Buck, T. A COMPASS to Model Comparison and Simulation-Based Inference in Galactic Chemical Evolution. *arXiv e-prints*, art. arXiv:2507.05060, July 2025. doi: 10.48550/arXiv.2507.05060.

Haid, S., Walch, S., Seifried, D., Wünsch, R., Dinnbier, F., and Naab, T. The relative impact of photoionizing radiation and stellar winds on different environments. *Monthly Notices of the Royal Astronomical Society*, 478 (4):4799–4815, May 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty1315. URL http://dx.doi.org/10.1093/mnras/sty1315.

Lemos, P., Cranmer, M., Abidi, M., Hahn, C., Eickenberg, M., Massara, E., Yallup, D., and Ho, S. Robust simulation-based inference in cosmology with Bayesian neural networks. *Machine Learning: Science and Technology*, 4(1):01LT01, March 2023a. doi: 10.1088/2632-2153/acbb53.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-based accuracy testing of posterior estimators for general inference, 2023b. URL https://arxiv.org/abs/2302.03026.

Osterbrock, D. E. *Astrophysics of gaseous nebulae and active galactic nuclei*. 1989.

Papamakarios, G. and Murray, I. Fast ϵ-free inference of simulation models with bayesian conditional density estimation, 2018. URL https://arxiv.org/abs/1605.06376.

Pejcha, O., Calderón, D., and Kurfürst, P. Supernovae in colliding-wind binaries: observational signatures in the first year. *Monthly Notices of the Royal Astronomical Society*, 510(3):3276–3290, 03 2022. ISSN 0035-8711. doi: 10.1093/mnras/stab3729. URL https://doi.org/10.1093/mnras/stab3729.

Radev, S. T., Schmitt, M., Pratz, V., Picchini, U., Köthe, U., and Bürkner, P.-C. JANA: Jointly Amortized Neural Approximation of Complex Bayesian Models. *arXiv e-prints*, art. arXiv:2302.09125, February 2023. doi: 10.48550/arXiv.2302.09125.

Rost, R., Branca, L., and Buck, T. Emulating Radiative Transfer in Astrophysical Environments. *arXiv e-prints*, art. arXiv:2511.08219, November 2025. doi: 10.48550/arXiv.2511.08219.

Seo, J. and Ryu, D. How-mhd: A high-order weno-based magnetohydrodynamic code with a high-order constrained transport algorithm for astrophysical applications. *The Astrophysical Journal*, 953(1):39, August 2023. ISSN 1538-4357. doi: 10.3847/1538-4357/acdf4b. URL http://dx.doi.org/10.3847/1538-4357/acdf4b.

Storcks, L. astronomix - differentiable mhd in jax, 2025. URL https://zenodo.org/doi/10.5281/zenodo.17782162.

Storcks, L. and Buck, T. Differentiable conservative radially symmetric fluid simulations and stellar winds – jf1uids, 2024. URL https://arxiv.org/abs/2410.23093.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018. doi: 10.1109/CVPR.2018.00675.

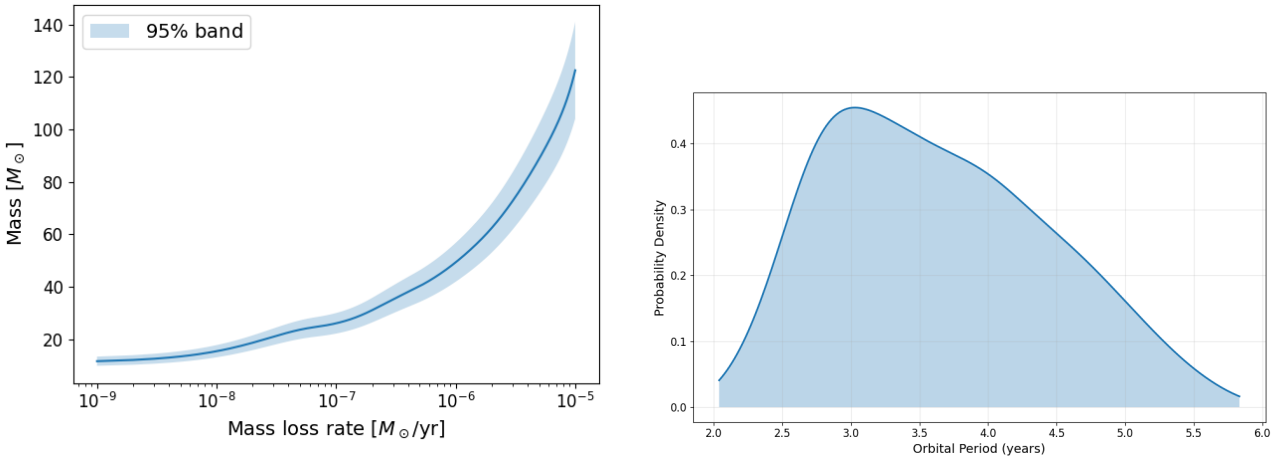
Viterbo, G. and Buck, T. CASBI – Chemical Abundance Simulation-Based Inference for Galactic Archeology. *arXiv e-prints*, art. arXiv:2411.17269, November 2024. doi: 10.48550/arXiv.2411.17269.

Viterbo, G. and Buck, T. The dynamical memory of tidal stellar streams: Joint inference of the Galactic potential and the progenitor of GD-1 with flow matching. *A&A*, 707:A363, March 2026. doi: 10.1051/0004-6361/202558358.

Wildberger, J. B., Dax, M., Buchholz, S., Green, S. R., Macke, J., and Schölkopf, B. Flow Matching for Scalable Simulation-Based Inference. In *Machine Learning for Astrophysics*, pp. 34, July 2023. doi: 10.48550/arXiv.2305.17161.

A. Prior distributions

Figure 6 shows the empirical mass / mass-loss relation $M_0(\dot{M})$ used to assign a stellar mass to each sampled $\dot{M}$ (with 5% Gaussian scatter), together with the prior-induced distribution of orbital periods for a binary with semi-major axis $a = 10\,\mathrm{AU}$.



(a) $M_0(\dot{M})$ with 5% Gaussian scatter.

(b) Prior-induced period distribution.

Figure 6. Derived quantities from the mass-loss prior. (a) the empirical mass / mass-loss mapping with Gaussian scatter used to draw stellar masses, and (b) the resulting prior distribution of orbital periods.

B. Geometry of CWBs

For physical intuition, it is helpful to introduce the analytical thin-shell solution of Cantó et al. (1996) for two interacting, axisymmetric, hypersonic, non-accelerating winds with efficient radiative cooling, launched from a stationary binary (orbital motion neglected). The resulting geometric properties and surface-density distributions provide useful insight, even when these assumptions are only approximately satisfied. For two stars with masses M_1, M_2 and wind velocities $v_{\infty,1}, v_{\infty,2}$ it is useful to start by defining the ratios of momenta β and terminal wind velocities α as

$$\alpha \equiv \frac{v_{\infty,1}}{v_{\infty,2}}, \qquad \beta \equiv \alpha \frac{\dot{M}_1}{\dot{M}_2} \tag{6}$$

Under the above listed assumptions of constant-velocity, isotropic winds, mass conservation yields the individual density profiles

$$\rho_1(r) = \frac{\dot{M}_1}{4\pi r^2 v_{\infty,1}}, \quad \rho_2(r_2) = \frac{\dot{M}_2}{4\pi r_2^2 v_{\infty,2}}. \tag{7}$$

where r and r_2 denote the distances from star 1 and star 2, respectively. In this case, the stand-off distance (stagnation radius), which is the point along the connection line of both stars where the ram pressures $P = \rho v_\infty^2$ are in balance, is given by

$$R_0 = a \frac{\sqrt{\beta}}{1+\sqrt{\beta}} \tag{8}$$

with orbital separation a. For $\beta > 1$ we have $R_0 > a/2$, i.e. the stagnation point lies closer to the weaker-wind star, as expected. By combining mass and momentum conservation with geometric arguments, one obtains an expression for the asymptotic opening angle θ_∞ of the bow-shock cone, measured at star 1 from the inter-stellar axis to the asymptotic shell direction, in the limit $r \to \infty$:

$$\theta_\infty - \tan\theta_\infty = \frac{\pi}{1-\beta}. \tag{9}$$

For an illustration of both quantities, see Fig. 7.

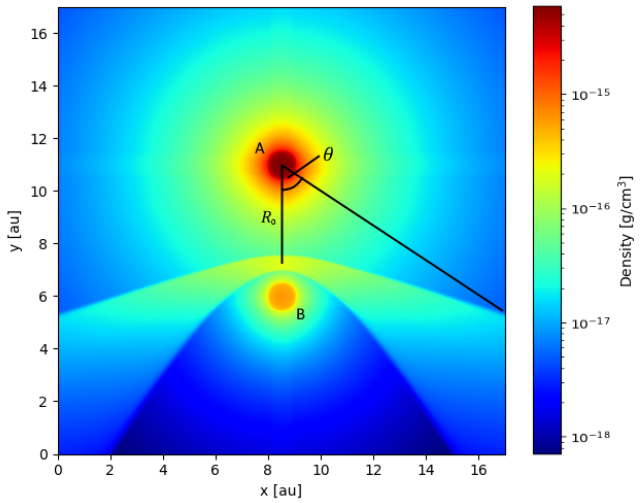


Figure 7. Mass-density slice in the orbital plane of a stationary CWB, showing the stand-off distance R_0 measured from the stronger-wind star A (Eq. 8) and the opening angle θ of the bow-shock cone (becomes θ_∞ for $r \to \infty$; Eq. 9). Example parameters: $\dot{M}_1 = 5.2 \cdot 10^{-6}$ M$_\odot$/yr, $\dot{M}_2 = 5.2 \cdot 10^{-7}$ M$_\odot$/yr, $v_{\infty,1} = 2200$ km/s, $v_{\infty,2} = 2000$ km/s, $a = 5$ AU

C. Hyperparameter tuning

Fig. 8 shows the 128-trial Pareto front optimization plot, in which the validation NLL and the mean TARP deviation from the ideal coverage curve are optimized simultaneously. The green marked and annotated trial 20 giving the best final calibration among the chosen best trials for a training on the complete test simulation set ($N = 40,000$).

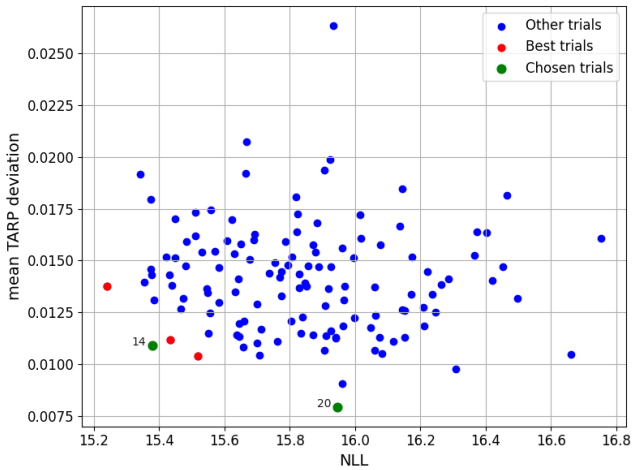
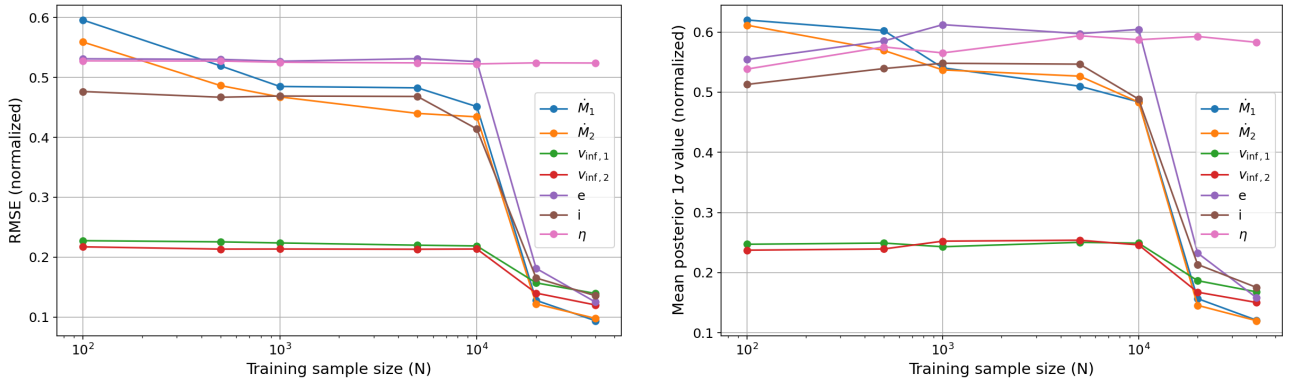


Figure 8. Pareto front over 128 Optuna trials for the joint minimization of validation NLL and mean TARP deviation. Red points mark the Pareto-optimal set. Trials 14 and 20 (green) were retrained at full batch, trial 20 giving the best final calibration.

D. Effect of training-set size

Figure 9 shows how the quality of inferred posteriors depends on the number of training simulations. We retrain the selected architecture on the subsets $N \in \{10^2, 5{\cdot}10^2, 10^3, 5{\cdot}10^3, 10^4, 2{\cdot}10^4, 4{\cdot}10^4\}$ of the training simulation set and evaluate on the held-out test set. Both the RMSE of the posterior mean against the ground truth and the 1σ Gaussian-approximated posterior width are approximately flat up to $N \approx 10^4$, then improve rapidly until $N \approx 2 \cdot 10^4$, and slowly flatten with only a small residual gain out to the largest training size. This suggests that additional simulations would further reduce bias and sharpen posteriors, but with diminishing returns. The turbulence parameter η is an exception, as it remains uninformative

across all training-set sizes, consistent with the discussion in Sec. 6. Note that the hyperparameters were held fixed at the values of Table 2 (tuned at the largest N) for all training-set sizes. Per-N retuning might improve the smaller-N results slightly, which makes the curves a conservative estimate of the achievable performance at small sample sizes.



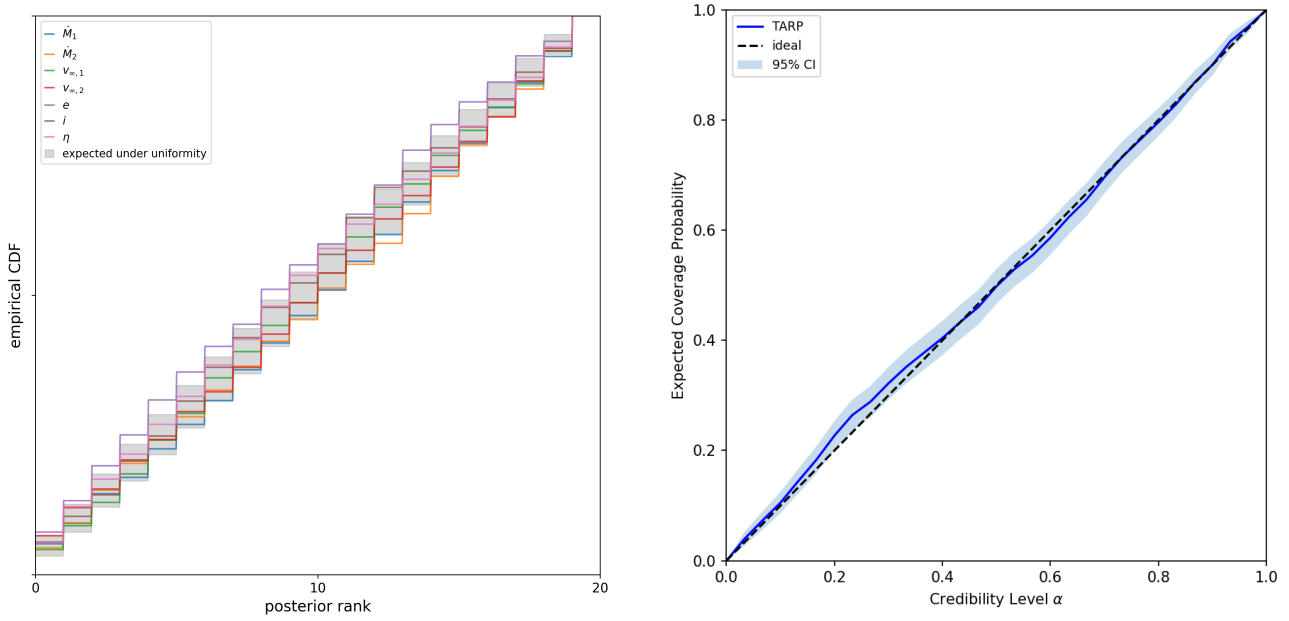
(a) RMSE vs. true values.

(b) Posterior 1σ width.

Figure 9. RMSE and posterior width as a function of training-set size. Both metrics saturate slowly past $N \approx 2 \cdot 10^4$ simulations.

E. Calibration diagnostics

Figure 10 shows the joint TARP coverage curve and the per-parameter SBC plot for the trained model. The TARP curve tracks the ideal diagonal across all credibility levels, indicating that the full 7-dimensional posterior is jointly well-calibrated (ATC $= -0.061$). The SBC as an empirical CDF vs. posterior rank plot confirms an approximately uniform rank distribution for mass-loss rate 1, both terminal wind velocities, inclination, and the turbulence parameter. Slight deviations from the optimum band indicate mild overconfidence for eccentricity (-8.8% too narrow posteriors) and mild underconfidence (5.1% too wide posteriors) for $\dot{M}_2$.



(a) SBC rank plot, per parameter.

(b) TARP joint coverage.

Figure 10. Posterior calibration diagnostics for the 7-parameter model. (a) simulation-based calibration rank plot are within uniform. (b) the TARP coverage curve tracks the diagonal across all credibility levels α.

F. Additional reference cases

Figure 11 shows the full 10-frame time series for Case 1 of Table 3, the high-luminosity counterpart of Case 2 above. The bow shock, the two contact discontinuities, and the periastron-driven brightening are all visible at a per-pixel relative noise of a few percent.

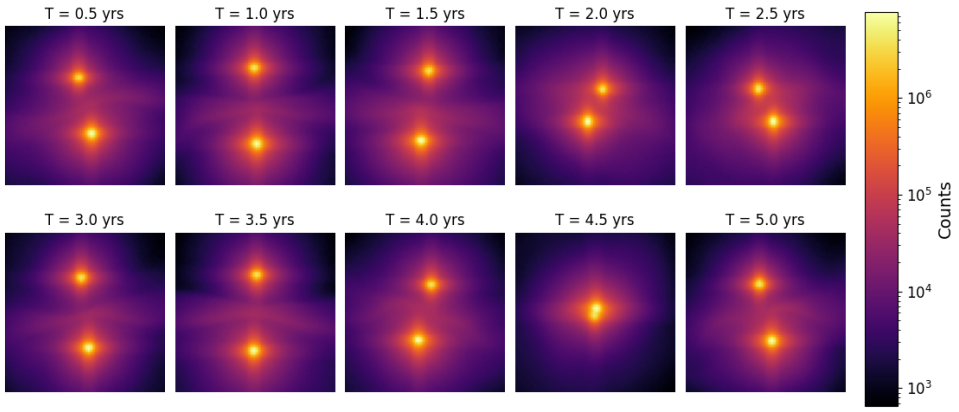


Figure 11. Complete 10-frame Hα photon-count time series for Case 1 (high wind luminosity, low relative noise, highly eccentric orbit). The periastron passage produces a clear brightening and reshaping of the bow shock.

Figure 12 shows the full posterior corner plot for Case 3 of Table 3: a strongly asymmetric CWB with $\dot{M}_1/\dot{M}_2 \sim 7{\cdot}10^2$. As anticipated, the dominant wind's velocity $v_{\infty,1}$ is well constrained, while $v_{\infty,2}$ returns a slightly skewed prior, because the weaker wind's shock features are no longer above the noise floor. Both mass-loss rates and orbital parameters e and i remain robustly recovered.

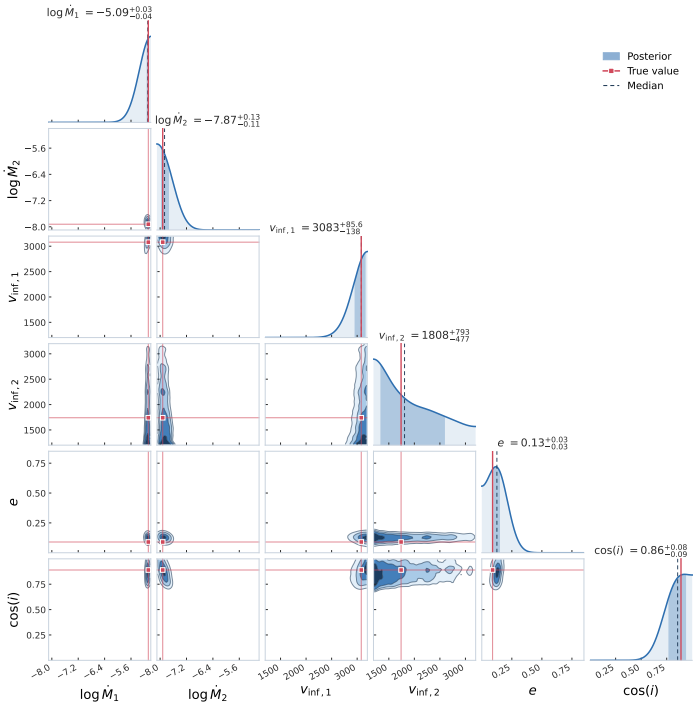


Figure 12. Inferred posterior for Case 3, a highly asymmetric CWB with one dominant wind. The dominant star's v_∞ is well constrained, the secondary's v_∞ returns a slightly distorted prior.

G. ABC baseline: summary statistics and posteriors

As a classical likelihood-free baseline we implement an Approximate Bayesian Computation (ABC) rejection sampler on a physics-motivated summary of the same 10-frame Hα photon-count time series that our neural pipeline takes in. This appendix defines the summary statistic, the distance, the acceptance threshold, and shows the resulting posteriors for the two reference cases of Sec. 5.

G.1. Summary statistic

For an observation I of the Hα photon count with $T = 10$ time frames and $H = W = 64$ pixels per spatial dimension we define a $(6T + 2) = 62$-dimensional summary vector

$$s(\mathbf{I}) \;=\; \left\{ \log_{10} F_t,\; \log_{10} P_t,\; c_{x,t},\; c_{y,t},\; \sigma_{x,t},\; \sigma_{y,t} \right\}_{t=1}^{T} \;\cup\; \{\phi,\; \pi\} \tag{10}$$

with per-frame components (and using p, q as pixel row/column indices). F_t is the total flux and P_t the peak flux, $c_{x,t}$ and $c_{y,t}$ are brightness-weighted centroids and second moments $\sigma^2_{x/y,t}$,

$$F_t = \sum_{p,q} I_{t,p,q}, \qquad P_t = \max_{p,q} I_{t,p,q}, \tag{11}$$

$$c_{x,t} = \frac{1}{F_t} \sum_{p,q} p\, I_{t,p,q}, \qquad c_{y,t} = \frac{1}{F_t} \sum_{p,q} q\, I_{t,p,q}, \tag{12}$$

$$\sigma^2_{x,t} = \frac{1}{F_t} \sum_{p,q} (p - c_{x,t})^2\, I_{t,p,q}, \qquad \sigma^2_{y,t} = \frac{1}{F_t} \sum_{p,q} (q - c_{y,t})^2\, I_{t,p,q}, \tag{13}$$

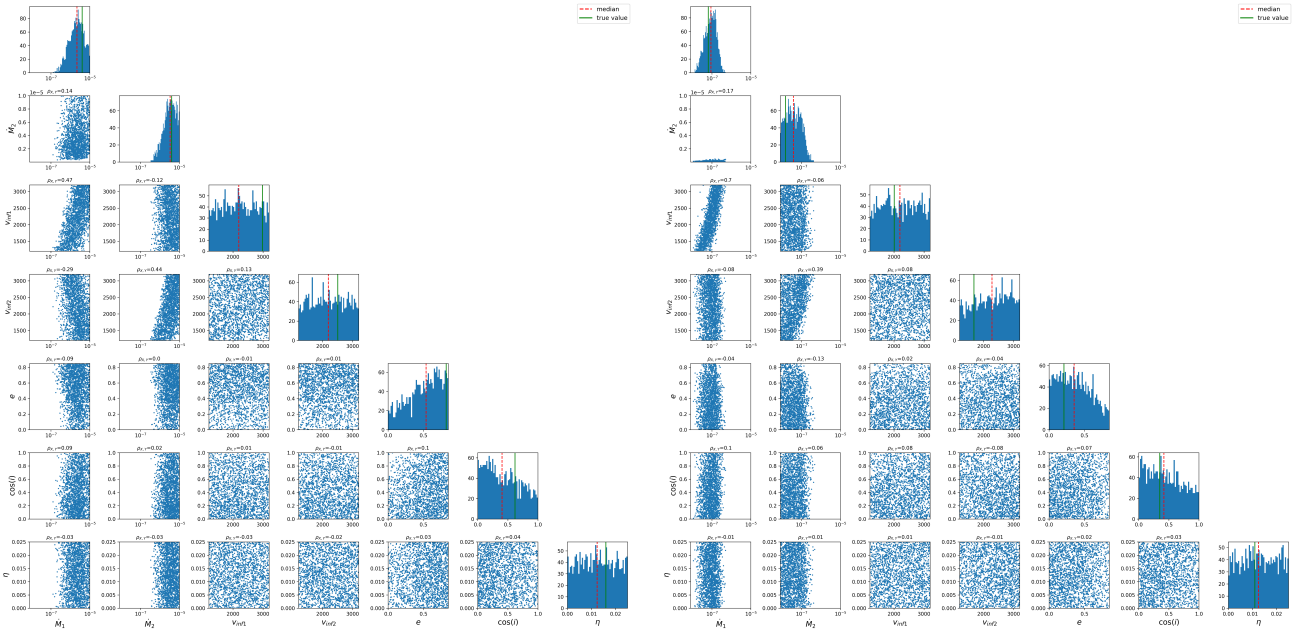
together with two global temporal descriptors

$$\phi \;=\; \frac{\mathrm{std}_t(F_t)}{\mathrm{mean}_t(F_t)}, \qquad \pi \;=\; \frac{\max_t P_t}{\min_t P_t}. \tag{14}$$

Per-frame F_t, P_t trace $\dot{M}$, while the centroids $(c_{x,t}, c_{y,t})$ trace orbital motion. The second moments $(\sigma_{x,t}, \sigma_{y,t})$ trace the bow-shock extent and thus v_∞, and ϕ, π try to pick up periastron-driven temporal variability due to the eccentricity.

We standardize every summary component by its empirical mean $\bar{s}_k$ and standard deviation $\hat{\sigma}_k$ estimated from a batch of 1000 test simulations, and use the standardized ℓ_2 distance. The acceptance threshold ϵ is set adaptively as the 5% quantile of distances measured in a pilot run of 500 previous simulations. We then sweep the full training pool of $N = 40{,}000$ pre-computed $(\boldsymbol{\theta}_i, \mathbf{I}_i)$ pairs and accept $\boldsymbol{\theta}_i$ whenever the condition $d(\mathbf{s}(\mathbf{I}_i), \mathbf{s}_{\mathrm{obs}}) \leq \epsilon$ is met.

Reference-case posteriors



(a) Case 1 (high wind luminosity). *(b)* Case 2 (low wind luminosity).

Figure 13. ABC-rejection posteriors obtained from the 62-D hand-crafted summary of Eq. 10 for the reference cases 1 and 2 (see Tab. 3). Only the mass-loss rates are constrained, with widths noticeably broader (1σ CI spans around a full decade for $\dot{M}_{1/2}$ in case 1) than those of the neural estimator. All other posterior distributions closely resemble their respective prior distributions.

Structural limitations of ABC rejection. Beyond the poor recovery of kinematic and orbital parameters, the rejection sampler carries three structural drawbacks that motivate the neural approach. First, the rejection step is inherently simulation-wasteful: at a 5% acceptance rate, 95% of the $N = 40{,}000$ forward simulations are discarded, so that the effective posterior is built from at most ~ 2000 samples regardless of total compute invested. Second, ABC rejection is not amortized, which means that the full sweep over $N = 40{,}000$ simulations must be repeated independently for every new observation, consuming the entire simulation budget per target. Third, global calibration diagnostics such as TARP and SBC (Appendix E) require evaluating posteriors over many held-out observations simultaneously, which is computationally prohibitive for rejection sampling and no population-level calibration statement can therefore be made for the ABC estimates.