A&A, 706, A323 (2026)
https://doi.org/10.1051/0004-6361/202556609

Astronomy & Astrophysics

Closing the evidence gap: reddemcee, a fast adaptive parallel tempering sampler

Next-generation ladder adaptation and evidence estimators for parallel tempering

Pablo A. Peña R.[1,2,⋆] and James S. Jenkins[1,2]

[1] Instituto de Estudios Astrofísicos, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército 441, Santiago, Chile
[2] Centro de Astrofísica y Tecnologías Afines (CATA), Casilla 36-D, Santiago, Chile

Received 25 July 2025 / Accepted 17 December 2025

ABSTRACT

Context. Markov chain Monte Carlo (MCMC) excels at sampling complex posteriors, but traditionally lags behind nested sampling in accurate evidence estimation, which is crucial for model comparison in astrophysical problems.
Aims. We introduce `reddemcee`, an adaptive parallel tempering ensemble sampler, aiming to close this gap by simultaneously presenting next-generation automated temperature-ladder adaptation techniques and robust, low-bias evidence estimators.
Methods. `reddemcee` couples an affine-invariant stretch move with five interchangeable ladder-adaptation objectives – a uniform swap-acceptance rate, swap mean distance, Gaussian area overlap, a small Gaussian gap, and equalised thermodynamic length – implemented through a common differential update rule. Three evidence estimators are provided: curvature-aware thermodynamic integration (TI+), geometric-bridge stepping stones (SS+), and a novel hybrid algorithm that blends both approaches (H+). The performance and accuracy of the sampler are benchmarked on n-dimensional Gaussian shells, Gaussian egg-box, Rosenbrock functions, and the real exoplanet radial-velocity time-series dataset of HD 20794.
Results. Across shells up to 15 dimensions, `reddemcee` achieves roughly 7 times the effective sampling speed of the best dynamic nested sampling configuration. The TI+, SS+, and H+ estimators recover estimates to within $|\Delta \ln \mathcal{Z}| \lesssim 3\%$ and supply realistic error bars with as few as six temperatures. In the HD 20794 case study, `reddemcee` reproduces literature model rankings and yields tighter yet consistent planetary parameters compared with `dynesty`, with evidence errors that track run-to-run dispersion.
Conclusions. By unifying fast ladder adaptation with reliable evidence estimators, `reddemcee` delivers strong throughput and accurate evidence estimates, often matching, and occasionally surpassing, dynamic nested sampling, while preserving the rich posterior information that makes MCMC indispensable for modern Bayesian inference.

Key words. methods: numerical – methods: statistical – planets and satellites: individual: HD 20794

1. Introduction

In scientific research, any measured data needs to be tested against a suitable hypothesis. As such, robust statistical techniques are indispensable for parameter estimation and model comparison. Science frequently faces the challenge of characterising probability distributions arising from complex, high-dimensional models ridden with nuisance parameters, uncertainties, and degeneracies. Markov chain Monte Carlo (MCMC) methods are widely recognised as a vital tool in many areas of science. They enable researchers to characterise parameter uncertainties precisely and compare models rigorously. Research areas that rely on precise inference, such as phylogenetics (Rannala & Yang 1996; Drummond et al. 2002), physiochemistry (Hansmann 1997; Sugita & Okamoto 1999), gravitational waves (van der Sluys et al. 2008; Veitch et al. 2015), and exoplanet discovery (Jenkins et al. 2020; Vines et al. 2023), benefit from the flexibility of MCMC techniques to explore difficult posterior landscapes.

Early exoplanet radial-velocity (RV) detections were often confirmed by identifying significant peaks in periodograms – in which false-alarm probabilities were calculated to establish the significance of a planetary candidate – and performing non-linear least-squares fitting for Keplerian orbits. Such was the case for 51 Pegasi b (Mayor & Queloz 1995) – the first exoplanet around a Sun-like star – and for the many Jovian planets that quickly followed this discovery (Gonzalez 1997; Marcy & Butler 2000).

However, as RV data grew and multi-planet systems became common, more sophisticated statistical tools were needed to extract the more difficult exoplanet signals, which could be subtle, degenerate, and embedded in considerable noise. This led to the introduction of MCMC methods (Ford et al. 2005), revolutionising exoplanet RV fitting with robust estimation of orbital parameters with realistic uncertainties, even for non-linear parameters (such as eccentricity and longitude of periastron). Far from perfect, MCMC still has its caveats. Poorly tuned chains could wander slowly or get trapped in local maxima.

To thoroughly explore multi-modal parameter spaces (common in multi-planet systems, for example), Gregory (2005)

⋆ Corresponding author: `astro.reddtea@gmail.com`

pioneered the use of the parallel tempering (PT) MCMC method (Swendsen & Wang 1986; Earl & Deem 2005) in relation to RV-signal detection and characterisation. This approach runs several chains in parallel to sample different powers (a temperature ladder) of the posterior distribution, each tempered to a different level. Hotter chains traverse the parameter space with more freedom, while colder chains sample the fine details. Inter-chain communication allows the sampler to explore distant high-probability nodes with ease. Furthermore, PT enables the algorithm to estimate the marginalised posterior (or evidence), a crucial quantity for model comparison. Gregory (2005) shows that this method is capable of efficiently exploring all regions of the phase space, lessening the burden of multi-modality. However, this early implementation was not capable of estimating the evidence reliably without a huge performance loss.

The affine-invariant ensemble samplers (Goodman & Weare 2010) also found their way into the exoplanet domain (Hou et al. 2012a). This MCMC variant, instead of a single chain, utilises an ensemble of 'walkers', drawing multiple samples per step, making the sampler insensitive to parameter covariances, and thus providing an enormous computational performance boost.

As an extra Keplerian in one's model can always fit noise, efforts turned to providing an accurate model comparison framework, leading to the usage of nested sampling (NS) algorithms (Skilling 2004) to calculate the Bayesian evidence for competing models (with differing number of planets or noise models), allowing one to select the most likely model (Feroz & Hobson 2008; Feroz et al. 2011). Building on NS, dynamic NS (DNS, Higson et al. 2019; Speagle 2020) introduces an adaptive allocation of 'sampling effort' to higher-probability areas of the phase space, further increasing the efficiency of reaching the target evidence precision, finding use in RV fitting (Diamond-Lowe et al. 2020).

MCMC excels in parameter posterior estimation, yielding reliable uncertainties, even for complex or correlated parameters. However, while evidence estimation is certainly possible under this algorithm, it is as a by-product of the posterior exploration. Therefore, increasing the accuracy of the evidence estimation does not necessarily translate into increased posterior accuracy, while carrying a substantial computational burden. On the other hand, NS is primarily designed for efficient evidence estimation by systematically shrinking the likelihood volume. As a by-product, it provides posterior samples, which can be used to model the parameter posteriors. Consequently, this method inherently does not provide as many samples from the parameter distribution as a well-tuned MCMC, conveying less precise parameter estimates.

This work presents `reddemcee`[1], an adaptive PT MCMC `Python` algorithm for any scientific-sampling-related endeavour that handles complex, high-dimensional, multi-modal posteriors, with several competing models. Leveraging five different automated tuning strategies for the temperature ladder – two classic ones, uniform swap acceptance, and posterior area overlap; and three new implementations based on average energy differences, the system's specific heat, and the swap mean distance (SMD) – while providing original adaptations for evidence estimation, such as the stepping stones algorithm (SS, Xie et al. 2011) with a per-stone geometric-bridge, thermodynamic integration (TI, Gelman & Meng 1998) enhanced by curvature-aware interpolation, and a novel hybrid approach.

[1] https://reddemcee.readthedocs.io/

2. Bayesian framework

A full description of Bayesian inference and MCMC methods is beyond the scope of this manuscript, and we therefore refer the reader to Gelman et al. (2004). Nonetheless, a summary of essential concepts is provided.

Bayesian inference generally consists of depicting the posterior probability distribution over the parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ of the hypothesise model, M, depicting some known measured data, D:

$$p(\boldsymbol{\theta} \mid D, M) = \frac{p(\boldsymbol{\theta} \mid M) \cdot p(D \mid \boldsymbol{\theta}, M)}{p(D \mid M)}, \tag{1}$$

where $p(\boldsymbol{\theta} \mid M)$ corresponds to the prior distribution, $p(D \mid \boldsymbol{\theta}, M)$ to the likelihood function, and $p(D \mid M)$ to the marginalised likelihood, frequently denoted as the Bayesian evidence, $\mathcal{Z}$, or simply 'evidence' (as it is in this manuscript). Throughout this work, proper priors are assumed ($\int_\Theta p(\boldsymbol{\theta} \mid M) = 1$), so the evidence is well defined and finite.

The evidence quantifies the overall support for model M and is crucial for Bayesian model comparison. However, the evidence integral generally has no closed-form solution and must be estimated numerically (Oaks et al. 2018). The MCMC methods, while excelling at drawing samples from the posterior, do not directly provide $\mathcal{Z}$.

Within MCMC methods, several evidence-estimation techniques have been developed (Maturana-Russel et al. 2019). For PT MCMC, two common approaches are TI (Gelman & Meng 1998) and stepping stones (Xie et al. 2011). How PT works and how TI and SS leverage the tempered ensemble to estimate the evidence are briefly outlined below.

2.1. Parallel tempering MCMC

Parallel tempering MCMC runs an ensemble of B parallel chains, each sampling a different power of the posterior distribution (at a different temperature, T_i; Swendsen & Wang 1986; Earl & Deem 2005; Miasojedow et al. 2013). The inverse of the temperature is $\beta = \frac{1}{T}$, with $1 = \beta_1 > \beta_2 > ... > \beta_{B-1} > \beta_B \geq 0$, and the full chain $\tilde{\boldsymbol{X}}_{\boldsymbol{t}} = (X_t^{(\beta_1)}, ..., X_t^{(\beta_B)})$. By convention, $\beta_1 = 1$ for the cold chain (sampling the original posterior) and $\beta_B \approx 0$ for the hottest chain (sampling a highly flattened landscape, approaching the prior as $\beta_B \to 0$). Each chain, $X_t^{(\beta_i)}$, samples a tempered posterior where the likelihood has been raised to the power β_i,

$$\mathcal{P}_{\beta_i}(\boldsymbol{\theta}) = \frac{\Pi \cdot \mathcal{L}^{\beta_i}}{\mathcal{Z}_{\beta_i}}, \tag{2}$$

where $\mathcal{P}_{\beta_i}(\boldsymbol{\theta})$ is the posterior distribution , Π the prior, $\mathcal{L}^{\beta_i}$ the tempered likelihood, and $\mathcal{Z}_{\beta_i}$ the evidence.

In $\mathcal{P}_\beta(\boldsymbol{\theta})$, hotter temperatures flatten and broaden peaks, reducing the risk of getting trapped in local maxima, and effectively making the posterior easier to sample. As such, hot chains are able to rapidly sample a large portion of the parameter space, whilst cold chains provide precise local sampling. By allowing chains at different temperatures to swap states, PT achieves a fast, thorough sampling of the posterior. Swaps in PT are implemented with a Metropolis-like acceptance rule. The pairwise swap (Goggans & Chi 2004), which is when two adjacent chains $(X_t^{(\beta_i)}, X_t^{(\beta_{i+1})})$ propose to switch states, is accepted with probability

$$A_{i,i+1} = \min\left(0, -\Delta \ln \mathcal{L}^{\beta_i} \cdot \Delta\beta_i\right), \tag{3}$$

where $\Delta \ln \mathcal{L}^{\beta_i}$ is the log-likelihood difference between the two chain states. This criterion (derived by simplifying the detailed balance condition for swapping two Boltzmann-distributed ensembles) shows that swaps are more probable when a high-likelihood state from a hotter chain is proposed for a colder chain. Without adjacent posteriors overlapping enough to allow regular swaps, efficiency plummets. By tuning the temperature ladder (the set of β_i values), PT aims to maintain a reasonably high swap-acceptance rate (SAR) across all adjacent pairs, ensuring a thorough exploration of the posterior landscape and providing a natural framework for estimating the evidence, $\mathcal{Z}$, via the ladder of tempered distributions.

Although our APT implementation does not provide independent draws, we next set up the classic methods for evidence estimation (which assume IID-based errors). In Sect. 3.3, we address how to handle this problem.

2.2. Thermodynamic integration

Thermodynamic integration is an indirect method of computing the evidence, $\mathcal{Z}$, by leveraging the continuous sequence of intermediate posteriors between prior and posterior (Gelman & Meng 1998; Goggans & Chi 2004; Lartillot & Philippe 2006). The basic formula arises from the identity (in statistical mechanics) that relates the derivative of the log-partition function to the average energy. In Bayesian terms:

$$\ln \mathcal{Z} = \ln \mathcal{Z}_1 - \ln \mathcal{Z}_0 = \int_0^1 \mathbb{E}_\beta \left[\ln \mathcal{L}\right] d\beta, \tag{4}$$

where $\mathbb{E}_\beta [...]$ is the expectation with respect to $\mathcal{P}_\beta(\boldsymbol{\theta})$, the tempered posterior at inverse temperature β. Since $\mathcal{Z}_0 = 1$ for proper priors (so $\ln \mathcal{Z}_0 = 0$), the integral directly yields $\ln \mathcal{Z}$. With B the number of temperatures, the integral is calculated via the trapezoidal rule:

$$\ln \widehat{\mathcal{Z}}_{\rm TI} \approx \sum_{i=1}^{B-1} \frac{\mathbb{E}_{\beta_{i+1}} \left[\ln \mathcal{L}\right] + \mathbb{E}_{\beta_i} \left[\ln \mathcal{L}\right]}{2} \cdot \Delta\beta_i. \tag{5}$$

2.3. Stepping stones

The stepping-stones algorithm (Xie et al. 2011) expresses the evidence ratio $\mathcal{Z}{=}\mathcal{Z}_1/\mathcal{Z}_0$ as a telescopic product across the temperature ladder:

$$\widehat{\mathcal{Z}}_{\rm SS} = \prod_{i=1}^{B-1} \frac{\mathcal{Z}_{\beta_i}}{\mathcal{Z}_{\beta_{i+1}}} = \prod_{i=1}^{B-1} r_i, \qquad r_i \equiv \frac{\mathcal{Z}_{\beta_i}}{\mathcal{Z}_{\beta_{i+1}}} = \mathbb{E}_{\beta_{i+1}} \left[\mathcal{L}^{\beta_i - \beta_{i+1}}\right], \tag{6}$$

with r_i the stepping-stones ratios and $\mathbb{E}_\beta [...]$ the expectation with respect to $\mathcal{P}_\beta(\boldsymbol{\theta})$. It is often convenient to work in log-space for numerical stability, yielding

$$\ln \widehat{\mathcal{Z}}_{\rm SS} = \sum_{i=1}^{B-1} \ln \left[\frac{1}{N_i} \sum_{n=1}^{N_i} \mathcal{L}^{\beta_i - \beta_{i+1}}\right], \tag{7}$$

where N_i is the number of posterior samples per chain used to approximate each expectation. SS has the advantage of typically delivering improved accuracy over TI when the number of temperatures is limited, since it more directly 'bridges' distributions to evaluate each incremental evidence ratio.

2.4. Evidence error estimation

For TI, Lartillot & Philippe (2006) identified two sources of error: one from the sampling itself, $\widehat{\sigma}_S$, calculated as the MC standard error affecting the log-likelihood averages, and another from the discretisation of the integral, $\widehat{\sigma}_D$. This quantity can be estimated by considering the worst-case contribution of the trapezoidal rule (essentially misrepresenting the whole triangular part):

$$\widehat{\sigma}_D \simeq \frac{|\mathbb{E}_{\beta_i}[\ln \mathcal{L}] - \mathbb{E}_{\beta_{i+1}}[\ln \mathcal{L}]|}{2} \cdot \Delta\beta_i. \tag{8}$$

For the SS method, the product of unbiased estimators, $\hat{\mathcal{Z}}_{\rm SS}$, is also unbiased; nevertheless, changing to log scale introduces a bias. Xie et al. (2011) estimated the error as

$$\widehat{\sigma}^2_{\mathcal{Z}_{\rm SS}} \approx \frac{1}{N^2} \sum_{i=1}^{B-1} \sum_{n=1}^{N} \left(\frac{\mathcal{L}^{\beta_i - \beta_{i+1}}}{\hat{r}_i} - 1\right)^2. \tag{9}$$

3. Temperature ladder

Extending the parallelisms with statistical thermodynamics, the specific heat, C_v, is defined with the energy $U = -\ln \mathcal{L}$:

$$C_v(\beta) = \frac{d\mathbb{E}_\beta[U]}{dT} = \beta^2(\mathbb{E}_\beta[U^2] - \mathbb{E}_\beta[U]^2) = \beta^2 \mathrm{Var}_\beta\left[U\right]. \tag{10}$$

This quantity governs the energy fluctuations at different temperatures, and is closely related to both the temperature ladder and the SAR. The mean SAR – derived from Eq. (3) – for a small temperature gap can be expressed as

$$\begin{aligned} \bar{A}_{i,i+1} &= \mathbb{E}_{\beta_i,\beta_{i+1}} \left[\exp(-\Delta U_i \cdot \Delta\beta_i)\right] \\ &\propto \mathbb{E}_{\beta_i} \left[\Delta U_i \cdot \Delta\beta_i\right] = \frac{\sqrt{C_v(\beta_i)}}{\beta_i} \Delta\beta_i. \end{aligned} \tag{11}$$

In the PT scheme, the temperature ladder must be chosen to maximise the efficiency of the cold chain. To ensure that this chain has good mixing, from the SAR (see Eq. (11)), it would suffice to minimise both $\Delta\beta_i$ and ΔU_i for every single chain. This way, samples from the hot chain can easily traverse to the cold chain. As a secondary aim, the temperature ladder serves to estimate the evidence as well, so allocating the temperatures aimed at favouring either the TI or SS method also has its benefits. Achieving a uniform SAR across all chains is generally considered a good strategy for healthy mixing (Sugita & Okamoto 1999; Predescu et al. 2004; Kone & Kofke 2005; Roberts & Rosenthal 2007; Vousden et al. 2016).

3.1. Geometric spacing

Kofke (2002) and later Predescu et al. (2004), by analysing the constant specific heat scenario, concluded that adjacent temperatures geometrically spaced should approximate to a uniform SAR. With a constant ratio, $\bar{R}$, each β would be defined by

$$\bar{R} = \frac{\beta_i}{\beta_{i+1}}. \tag{12}$$

This is the starting point for ladder design, and it illustrates why adaptive methods are needed: geometric spacing assumes constant C_v, which does not hold true in complex posteriors.

3.2. Adaptive methods

Adaptive parallel tempering (APT) MCMC dynamically updates the temperature ladder during the run to improve mixing efficiency (Gilks et al. 1998; Roberts & Rosenthal 2007; Miasojedow et al. 2013; Vousden et al. 2016). Such an algorithm inherently fails to be ergodic and may not preserve the stationarity of the cold chain. In practice, `reddemcee` uses a decaying adaptive rate (see Eq. (14)), allowing the ladder to stabilise over time. Then, the adaptation can be discontinued, after which detailed balance is preserved and standard ergodicity theorems apply. This procedure was adopted for all benchmarks that follow. Next, five different adaptation strategies are shown. The first two – uniform SAR and Gaussian area overlap – are commonly seen in the literature, while for the latter three – SMD, small Gaussian gap (SGG), and equalised thermodynamic length (ETL) – we present our implementations, as effective alternatives, especially when C_ν is non-uniform.

3.2.1. Uniform swap-acceptance rate

Vousden et al. (2016) propose that the dynamic adjustments of the temperature ladder are guided by the SAR. They define the ladder in terms of logarithmic temperature intervals, S_i, between adjacent chains. Therefore, the adaptation rate corresponds to $\frac{dS_i}{dt}$. The adaptation includes a diminishing factor, $\kappa(t)$, where $\lim_{t\to\infty}\kappa(t) = 0$, here expressed as a hyperbolic decay:

$$S_i \equiv \ln\left(T_i - T_{i+1}\right), \qquad \frac{dS_i}{dt} = \kappa(t)\left[A_i(t) - A_{i+1}(t)\right], \tag{13}$$

$$\kappa(t) = \frac{1}{\nu_0}\frac{\tau_0}{(t+\tau_0)}, \tag{14}$$

with $A_i(t)$ the measured SAR, τ_0 the decay half-life, and ν_0 the evolution timescale. This leaves both β_1 and β_B static, while the intermediate temperatures settle to achieve a uniform SAR.

3.2.2. Gaussian area overlap

Rathore et al. (2005), by studying the correlation between replicas of the SAR and the area of overlap of likelihoods in Gaussian distributions, found a relation to control the acceptance rate by spacing temperatures:

$$A_{\rm overlap} = \mathrm{erfc}\left[\frac{\Delta\ln\mathcal{L}}{2\sqrt{2}\sigma_m}\right] \quad\Rightarrow\quad \left.\frac{\Delta\ln\mathcal{L}}{\sigma_m}\right|_{\beta_i} = \left[\frac{\Delta\ln\mathcal{L}}{\sigma_m}\right]_{\rm target}, \tag{15}$$

where $\sigma_m = (\sigma_1 + \sigma_2)/2$ is the mean of the deviations of the adjacent temperature log-likelihoods. While also aiming for uniform SAR, this ladder evolution is informed by a different mechanism, which may compensate for the effects of having a non-uniform C_ν.

3.2.3. Uniform swap mean distance

Considering that a non-uniform C_ν is more realistic for complex systems (as in exoplanet discovery), Katzgraber et al. (2006) presented a feedback-optimised method that maximises the round trips of each replica between the extremal temperatures. This allows information about the likelihood landscape to travel faster from the hot chain to the cold chain. A well-tempered chain (where the posterior shape has changed significantly compared to the target posterior) should propose far-away states that, if accepted, would contribute more to healthy mixing than a swap so close that it is equivalent to the intra-chain evolution. Therefore, a ladder where adjacent swaps are done by maximising the swap distance would maximise the crossing of 'useful' information across the replicas. By normalising all dimensions in the system to unity, the mean distance of adjacent swaps for all walkers at a given temperature is

$$d|_{\beta_i} = \frac{1}{W}\sum_{w=1}^{\rm W}\sqrt{\sum_{d=1}^{\rm D}\left(\frac{\theta_{d,i} - \theta_{d,i+1}}{c_d}\right)^2}, \tag{16}$$

where W is the number of walkers, D the number of dimensions, and c_d the range of the prior for dimension d. Hereafter this method is referred to as the SMD. Intuitively, this encourages swaps that carry proposals farther across the parameter space, which may be specifically beneficial in high-dimensional problems, where local swaps might exchange nearly similar states.

3.2.4. Small Gaussian gap

This scheme (SGG) targets a uniform SAR by using a Gaussian approximation for small temperature gaps. If adjacent βs are very close, the energy difference, ΔU_i, has approximately a Gaussian distribution of $\sim\mathcal{N}(0, 2\mathrm{Var}[U_i])$. In this small-gap regime, one can show that the product $(\Delta\beta_i)^2\mathrm{Var}[U_i]$ largely controls the SAR (Predescu et al. 2004), from Eq. (11):

$$\begin{aligned}\bar{A}_{i,i+1} &= \mathbb{E}_{\beta_i,\beta_{i+1}}\left[\exp(-\Delta U_i\cdot\Delta\beta_i)\right] \\ &\approx \exp\left(\frac{1}{2}(\Delta\beta_i)^2(2\mathrm{Var}_{\beta_i}[U])\right) = \exp\left((\Delta\beta_i)^2(\mathrm{Var}_{\beta_i}[U])\right).\end{aligned} \tag{17}$$

Thus, keeping the product $(\Delta\beta_i)^2(\mathrm{Var}[U_i])$ constant implies a roughly uniform SAR.

3.2.5. Equalised thermodynamic length

The thermodynamic length is defined as

$$L(\beta) = \int_{\beta_{\rm min}}^{\beta_{\rm max}}\sqrt{\mathrm{Var}_\beta[\mathrm{U}]}d\beta = \int_{\beta_{\rm min}}^{\beta_{\rm max}}\frac{\sqrt{C_\nu(\beta)}}{\beta}d\beta. \tag{18}$$

Intuitively $L(\beta)$ corresponds to the local metric that measures how quickly (or far) the distribution changes with β. Setting neighbouring replica distributions so that they are the same 'distance' apart translates to placing them uniformly in L space, effectively populating the regions where C_ν peaks with more chains. This ensures denser sampling where the posterior changes most rapidly. Each interval of this integral can be approximated as

$$\Delta L_i \approx \frac{\Delta\beta_i}{2}\left(\sqrt{\mathrm{Var}_{\beta_i}[\mathrm{U}]} + \sqrt{\mathrm{Var}_{\beta_{i+1}}[\mathrm{U}]}\right). \tag{19}$$

Then, the discrete estimate for thermodynamic distance is

$$L(\beta_N) = \sum_{i=0}^{B-1}\Delta L_i. \tag{20}$$

Shenfeld et al. (2009) arrived at the same conclusion, further demonstrating that under constant C_ν the ETL forms a geometric progression as well.

3.2.6. Ladder adaptation in `reddemcee`

`reddemcee` implements each of the methods discussed above by setting the adaptation as $\frac{dS_i}{dt}=\kappa(t)\cdot Q_i$. This way, choosing a definition for Q_i determines the adaptive approach – be it a uniform SAR, Gaussian area overlap (GAO), uniform SMD, SGG, or ETL:

$$Q_i = \begin{cases} [A_i(t) - A_{i+1}(t)] & \text{for SAR (see Sect. 3.2.1)} \\ \frac{\Delta U_i}{\sigma_m} & \text{for GAO (see Sect. 3.2.2)} \\ d|_{\beta_i} & \text{for SMD (see Sect. 3.2.3)} \\ \exp\left((\Delta\beta_i)^2(\mathrm{Var}[U_i])\right) & \text{for SGG (see Sect. 3.2.4)} \\ \frac{\Delta L_i}{L(\beta_N)} & \text{for ETL (see Sect. 3.2.5)} \end{cases}. \quad (21)$$

3.3. Revisiting evidence

The APT implementation does not provide independent draws; each walker's proposal is based on its current position, meaning successive samples retain memory until the integrated autocorrelation time has passed. The stretch move also couples walkers, influencing each other's proposals (Goodman & Weare 2010). Furthermore, the swap move creates a cross-temperature coupling as well. To account for all these effects in correlated draws, we replaced IID-based errors with autocorrelation-adjusted MC standard errors (Flegal & Jones 2008), using overlapping batch means (OBM, Wang et al. 2018). This technique, which is consistent for MCMC under standard ergodicity conditions, also propagates cross-temperature covariance across the ladder. We used this as $\widehat{\sigma}_S$ for both methods.

Both TI and SS rely on the shape of the $\mathbb{E}_\beta[U](\beta)$ curve, for which we considered three factors: (1) the ladder size is relatively small, providing few samples for either TI or SS; (2) temperatures are not uniformly spaced; (3) $\mathbb{E}_\beta[U](\beta)$ is a monotonically increasing function. With these points in mind, two different solutions for a better $\widehat{\sigma}_D$ come to mind. One approach is to use the Richardson extrapolation, which entails computing the evidence estimate on two different grids (one coarse and one fine). The difference between these estimates is then used to extrapolate the discretisation error, leveraging the known convergence order of the trapezoidal rule. Another approach is to estimate the local curvature of the $\mathbb{E}_\beta[U](\beta)$ curve (since the leading trapezoidal-rule error term is proportional to the second derivative) to derive a tailored error estimate for each interval. In the following paragraphs, two original implementations based on these ideas are presented, leveraging the properties of $\mathbb{E}_\beta[U](\beta)$ for a more precise $\ln\mathcal{Z}$ estimation.

Piecewise interpolated thermodynamic integration. Building on the TI method, an interpolation on the curve $\mathbb{E}_\beta[U](\beta)$ was applied with the local piecewise cubic hermite interpolating polynomial (PCHIP), which preserves monotonicity (physically expected) and prevents overshooting for non-smooth data (few temperatures; Fritsch & Butland 1984). For $\widehat{\sigma}_D$, we formed a coarse ladder by dropping every other temperature, built the corresponding PCHIP, and computed another coarser evidence estimate. We used its difference with respect to the finer one as the discretisation error.

The total error is then $\widehat{\sigma}_\mathcal{Z} = \sqrt{\widehat{\sigma}_D^2 + \widehat{\sigma}_S^2}$. It is worth noting that this approach addresses the discretisation error more explicitly, reducing bias at the cost of possibly smaller (but more realistic) error bars, as is shown in Section 4. More details on the method can be found in Appendix C.2.

Geometric-bridge stepping stones. The SS evidence estimator, $\hat{\mathcal{Z}}_{\mathrm{SS}}$ (see Eq. 9), estimates the ratio, r_i, with an expectation under a single distribution, $\mathbb{E}_{\beta_i+1}[\cdot]$ (see Eq. 6). We replaced the ratios with geometric-bridge sampling (Meng & Wong 1996; Gronau et al. 2017):

$$r_i = \frac{\mathbb{E}_{\beta_i}\left[\mathcal{L}^{\Delta\beta_i/2}\right]}{\mathbb{E}_{\beta_{i+1}}\left[\mathcal{L}^{-\Delta\beta_i/2}\right]}. \quad (22)$$

By using samples from both ends, we expect to diminish the estimator variance when the temperature gap is large, as the per-stone ratio would degrade less rapidly. Using the same APT samples has no additional computational cost, and by halving the exponent we reduce the dynamic range, improving numerical stability. A more detailed explanation can be read in Appendix C.1.

Hybrid method. Consider a temperature mid-point, β_*, so TI is applied in $\beta\in[0,\beta_*]$, and stepping stones in $\beta\in[\beta_*,1]$. Provided that continuity is ensured around β_*, separating where each method is stronger leads to a fair estimation of $\ln\mathcal{Z}$. From Eq. (5) the overall log-evidence can be decomposed as

$$\ln\mathcal{Z} = \int_0^1 \mathbb{E}_\beta[U]\,d\beta = \int_0^{\beta_*} \mathbb{E}_\beta[U]\,d\beta + \int_{\beta_*}^1 \mathbb{E}_\beta[U]\,d\beta. \quad (23)$$

The transition between the 'pure prior' region (β=0) to a moderately informed distribution (small β) changes statistics more dramatically than other areas. Intuitively, any sharp change in posterior shape can be seen as a phase transition, which is also reflected as a peak in $C_v(\beta)$. Since most temperature ladders are proportional in some form to C_v, chains are denser in this region, and therefore a straightforward method such as TI works well. Near β=1, SS is more efficient at handling ratios, since the posterior is strongly peaked and $\Delta\beta$ tends to increase. By decomposing the TI trapezoidal area contribution of each interval, the mid-point is selected to be where the rectangular area is larger than twice the triangular area, ensuring that we are in a region where ΔU_i dominates over $\Delta\beta_i$, and where the energy slope is pronounced:

$$\Delta\beta_i\cdot U_i \geq 2\left(\Delta\beta_i\cdot\frac{\Delta U_i}{2}\right) \Rightarrow 2U_i \geq U_{i+1}. \quad (24)$$

4. Benchmarks

This section seeks to validate the proposed ladder adaptation algorithms as well as the evidence estimators. Taken together, these tests probe the different problems that matter most when contrasting an APT sampler with DNS. Gaussian shells, Gaussian egg-box, and the hybrid Rosenbrock function capture the three classic challenges when sampling from a multimodal distribution: mode-finding, mode-hopping, and in-mode mixing, all of which are intensified as dimensionality increases. Hereafter, two distinct qualities are compared: first, the sampler performance (comprising both the intra-chain mixing and the computational cost, measured in effective samples per second or 'kenits'); and second, the evidence estimation (comprising the estimated evidence value and its uncertainty). This was measured by both the difference found with the analytical evidence $\Delta_\mathcal{Z}\equiv\exp\left(\ln\mathcal{Z}-\ln\widehat{\mathcal{Z}}\right)-1$, validating the accuracy of the estimator, and the log-likelihood $\mathcal{L}(\widehat{\mathcal{Z}})\equiv\ln\mathcal{L}_{\widehat{\mathcal{Z}}}$, validating the credibility of the uncertainties, for $\ln\widehat{\mathcal{Z}}\sim\mathcal{N}(\ln\widehat{\mathcal{Z}},\widehat{\sigma}_{\ln\widehat{\mathcal{Z}}})$, where

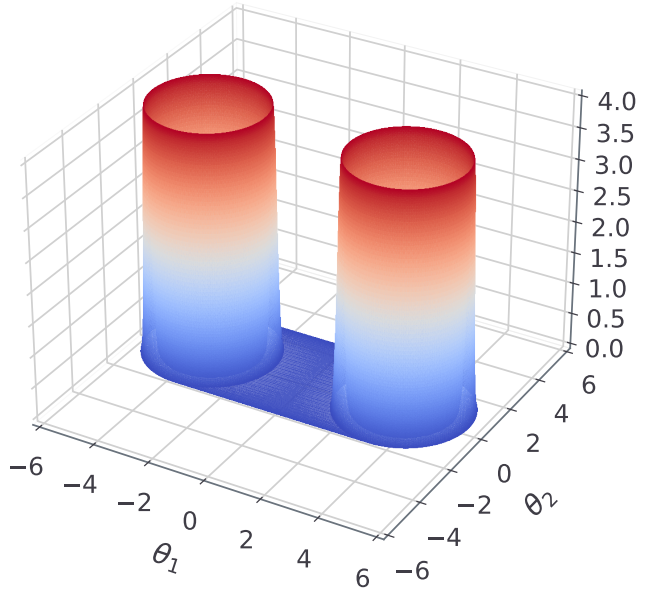

Fig. 1. 2D Gaussian shells likelihood. Radius r=2 and width w=0.1. Parameter boundaries are imposed as ±6. Likelihood values are coloured from blue (low) to red (high) to facilitate the visualisation of the figure.

$\ln \widehat{\mathcal{Z}}$ is the estimate of the evidence, and $\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}$ the estimate of the evidence uncertainty. Each benchmark was run 11 times with different known random seeds, and the reported values correspond to the mean and standard deviation of all runs.

4.1. n-d Gaussian shells

Sampling from an n-dimensional Gaussian shell is an analytically tractable problem widely used in the literature (Feroz & Hobson 2008; Speagle 2020). The likelihood contours are curved, thin, and nearly singular (see Fig. 1). Random-walk steps are inefficient, so this multimodal function stresses proposal geometry and temperature allocation. For n dimensions, radius r, width w, and peaks' centres $\boldsymbol{c}_i$, the likelihood is

$$p(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi w^2}} \exp\left(-\frac{(|\boldsymbol{\theta} - \boldsymbol{c}_i| - r)^2}{2w^2}\right). \tag{25}$$

This benchmark was used to assess the behaviour of the different samplers in efficiency and evidence estimation. `reddemcee` was set up with [16, 320, 640, 1] temperatures, walkers, sweeps, and steps, respectively, for a total of 3 276 800 samples. The adaptation hyper-parameters, τ_0, ν_0, were set as one tenth of the sweeps and one hundredth of the walkers, respectively. We adapted the temperature ladder during the first half of the sweeps, and then froze it. This entire adaptive phase was treated as burn-in: those samples were discarded before any statistical inference, but the time spent was included in the reported run-time (i.e. it affects the kenits metric). For comparison, the DNS runs were performed using `dynesty` (Speagle 2020) with its automated settings, testing three sampling modes: uniform, slice, and random-slice.

For `dynesty`, we calculated kenits using the reported number of likelihood calls, effectively measuring the sampling efficiency (fraction of independent samples) times all likelihood evaluations divided by the total run-time. For `reddemcee`, efficiency was defined as the inverse of the average autocorrelation time of the cold-chain, and kenits were computed as this efficiency multiplied by the number of likelihood evaluations (discarding burn-in samples, but not their run-time) over the full run-time. This method ensures a fair kenits comparison.

Table 1. 2D Gaussian shells performance benchmark.

Method	Time (s)	Eff (%)	Kenits
SAR	37.60±0.28	8.49±0.27	3.70±0.12
SMD	37.77±0.19	8.59±0.27	3.73±0.13
SGG	37.86±0.15	8.49±0.30	3.67±0.13
GAO	37.95±0.25	8.34±0.25	3.60±0.11
ETL	37.94±0.22	8.49±0.24	3.66±0.10
dyn-u	68.66±16.75	4.94±0.82	0.19±0.04
dyn-s	16.83±0.89	3.77±0.22	0.75±0.04
dyn-rs	13.41±0.64	4.49±0.33	0.94±0.05

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s uniform (dyn-u), slice (dyn-s), and random-slice (dyn-rs) sampling methods. From left to right: Time – total run time in seconds, Eff – sampling efficiency or percentage of independent samples, and kenits – effective samples per second in thousands.

Sampler performance. All adaptive ladder methods are statistically similar (see Table 1); variance between ladder algorithms (0.04^2) is smaller than the run-to-run variance of each method ($\approx 0.12^2$), delivering ≈3.6–3.8 kenits. The best DNS method, dyn-rs, achieves 0.94±0.05 kenits, which is ∼26% of the APT average. This method also finishes the fastest, at ∼13.4s. It is worth mentioning that once converged, the efficiency does not radically change, so even by demanding a higher total iteration count as a stop condition, the kenits (and efficiency) would not dramatically increase. The dynamic algorithm stops at target precision, which, in the dyn-u case, required a much longer run-time with many more samples, whereas the slice methods finished faster with far fewer independent samples. In other words, there is a trade-off between thorough posterior sampling and evidence-estimation convergence speed. The same reasoning can be extended to MCMC, preserving much richer posterior information.

Evidence estimation. The proposed evidence estimation methods in Sect. 2.4 were tested and compared against the analytical value for the 2D case, $\ln \mathcal{Z}$=−1.746 (see Table 2). The default TI method has $\Delta_{\mathcal{Z}}|_{\rm TI}$=8.394%, whereas the improved TI+ version achieves $\Delta_{\mathcal{Z}}|_{\rm TI+}$=0.324% (an order of magnitude closer to the true value). The likelihood improves as well, with a ratio of exp(2.80 − 0.92) = 6.58. By contrast, the default SS has the highest $\mathcal{L}(\widehat{\mathcal{Z}})|_{\rm SS}$=4.326 alongside a low $\Delta_{\mathcal{Z}}|_{\rm SS}$=0.117%. The SS+ counterpart shows slightly more conservative uncertainties (0.006 instead of 0.005) with $\mathcal{L}(\widehat{\mathcal{Z}})|_{\rm SS+}$=4.263, and a slightly higher accuracy, $\Delta_{\mathcal{Z}}|_{\rm SS+}$=0.099%. The H+ method achieves the highest accuracy of all, $\Delta_{\mathcal{Z}}|_{\rm H+}$=0.016%, along with a high likelihood, $\mathcal{L}(\widehat{\mathcal{Z}})|_{\rm H+}$=3.951. Out of the DNS methods, dyn-u is the most reliable with $\Delta_{\mathcal{Z}}|_{\rm dyn-u}$=1.751%, and a log-likelihood of $\mathcal{L}(\widehat{\mathcal{Z}})|_{\rm dyn-u}$=2.206.

4.2. Gaussian egg-box

The Gaussian egg-box, dubbed after its likelihood shape (see Fig. 2), presents a periodic landscape with several equally high

Table 2. 2D Gaussian shells SAR evidence estimation comparison.

Method	$\ln \widehat{\mathcal{Z}}$	$\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}$	$\Delta_{\mathcal{Z}}(\%)$	$\mathcal{L}(\widehat{\mathcal{Z}})$
TI	−1.833±0.008	0.066±0.005	8.394	0.917
SS	−1.744±0.008	0.005±0.001	**0.117**	**4.326**
H	−1.788±0.008	0.134±0.005	4.124	1.044
TI+	−1.749±0.008	0.024±0.005	0.324	2.801
SS+	−1.745±0.008	0.006±0.001	0.099	**4.263**
H+	−1.745±0.008	0.008±0.002	**0.016**	3.951
dyn-u	−1.763±0.054	0.040±0.001	**1.751**	**2.206**
dyn-s	−1.774±0.055	0.040±0.001	2.876	2.049
dyn-rs	−1.767±0.043	0.040±0.001	2.159	2.157

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s uniform (dyn-u), slice (dyn-s), and random-slice (dyn-rs) sampling methods. From left to right, the log-evidence estimator, the estimator uncertainty, the difference to the true value $\ln \mathcal{Z}$=−1.746 in percentage $\Delta_{\mathcal{Z}}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

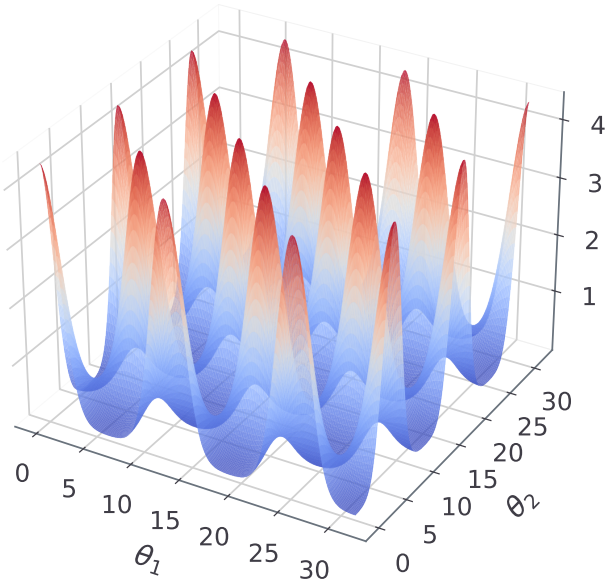

Fig. 2. Gaussian egg-box likelihood, with 16 different modes. Likelihood values are coloured from blue (low) to red (high) to facilitate visualisation.

peaks, separated by deep troughs. It is an ideal stress test for mode completeness. Missing a single peak would render catastrophically wrong evidence. The likelihood function is

$$p(\boldsymbol{\theta}) = \left(2 + \prod_{n=1}^{dim} \cos\left(\frac{\theta}{4}\right)\right)^{\bar{\beta}_e}, \tag{26}$$

where $\bar{\beta}_e = 5$ was used, defining the narrowness of the peaks. The prior volume was limited to $[0, 10\pi]$, and a fine grid integration gives $\ln \mathcal{Z}$=235.856, which was used as the true value. We used the same set-up as in Sect. 4.1.

Sampler performance. For the ladder adaptation methods, the SAR and SMD methods rank best at 5.79 ± 0.29 and 5.66 ± 0.23 kenits with overlapping uncertainties, followed closely by the other methods (see Table 3). Out of the DNS methods, dyn-rs has the lowest wall-time as well as the highest kenits count, 0.82 ± 0.07, 14.2% of the SAR kenits.

Table 3. Egg-box performance benchmark.

Method	Time (s)	Eff (%)	Kenits
SAR	22.90±0.11	8.10±0.43	5.79±0.29
SMD	23.02±0.24	7.95±0.32	5.66±0.23
SGG	23.35±0.14	7.69±0.28	5.40±0.19
GAO	23.36±0.24	7.84±0.30	5.50±0.20
ETL	23.52±0.26	7.90±0.38	5.51±0.28
dyn-u	30.41±1.51	10.71±0.88	0.45±0.02
dyn-s	19.83±0.79	3.06±0.22	0.69±0.03
dyn-rs	16.87±1.50	3.21±0.39	0.82±0.07

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s sampling methods. From left to right: Time – total run time in seconds, Eff – sampling efficiency or percentage of independent samples, and kenits – effective samples per second in thousands.

Table 4. Egg-box evidence SAR estimation comparison.

	$\ln \widehat{\mathcal{Z}}$	$\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}$	$\Delta_{\mathcal{Z}}(\%)$	$\mathcal{L}(\widehat{\mathcal{Z}})$
TI	235.733±0.013	0.748±0.008	11.532	–0.643
SS	235.843±0.012	0.008±0.001	**1.252**	**2.618**
H	235.802±0.013	0.130±0.006	5.283	1.036
TI+	235.842±0.013	0.023±0.008	1.425	2.648
SS+	235.844±0.013	0.008±0.001	1.211	2.771
H+	235.845±0.013	0.010±0.003	**1.099**	**3.082**
dyn-u	235.858±0.104	0.064±0.001	**0.201**	**1.828**
dyn-s	235.903±0.104	0.064±0.001	4.812	1.560
dyn-rs	235.953±0.118	0.064±0.001	10.186	0.679

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s sampling methods. From left to right, we show the log-evidence estimate, the estimate error, the difference to the true value $\ln \mathcal{Z}$ = 235.856 in percentage $\Delta_{\mathcal{Z}}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

Evidence estimation. As seen in Table 4, the TI algorithm misses the mark by $\Delta_{\mathcal{Z}}$=11.53%, with the worst overall likelihood: $\mathcal{L}(\widehat{\mathcal{Z}})$=−0.643. TI+ improves this to $\Delta_{\mathcal{Z}}$=1.43%, with a likelihood ratio of exp(2.658 − (−0.643)) = 26.87. The SS yields a log-likelihood of $\mathcal{L}(\widehat{\mathcal{Z}})$=2.618. Its SS+ counterpart marginally reduces $\Delta_{\mathcal{Z}}$ from 1.25% to 1.21%, with $\mathcal{L}(\widehat{\mathcal{Z}})$=2.771. The H+ method outperforms SS+ and TI+, achieving $\mathcal{L}(\widehat{\mathcal{Z}})$=1.099 and $\mathcal{L}(\widehat{\mathcal{Z}})$=3.082. Out of the DNS methods, dyn-u finds all modes (extremely accurate $\Delta_{\mathcal{Z}}$=0.201), albeit at the cost of many likelihood calls. Meanwhile, slice-based samplers are faster but slightly undersample some modes (evidence biases). dyn-u has the highest log-likelihood out of the DNS methods, at 1.828. Against H+, it gives a ratio of exp(3.082 − 1.828) = 3.5, which is almost four times as likely.

4.3. Hybrid Rosenbrock function

The Rosenbrock function is a long, thin, curved valley with strong non-linear correlations, slightly resembling a banana. The hybrid-Rosenbrock function is an extension that provides analytic evidence for any dimension (Pagani et al. 2019). Its

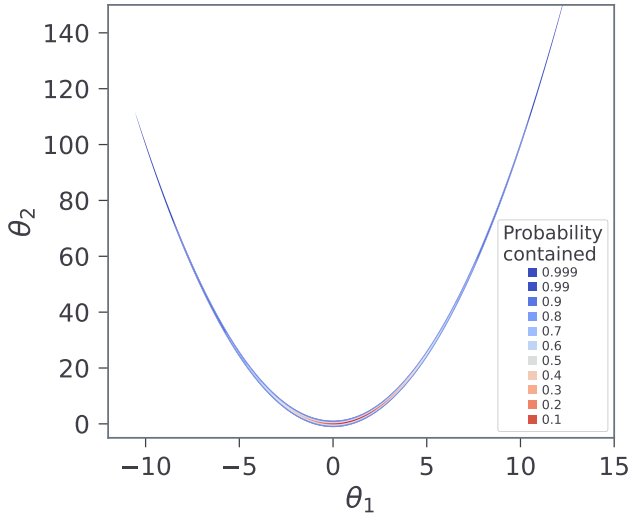

Fig. 3. Probability contour of the 2D Rosenbrock function. The inset key shows how the colours relate to the probability.

likelihood is

$$p(\boldsymbol{x}) = -a(x_1 - \mu)^2 - \sum_{j=1}^{n_2}\sum_{i=2}^{n_1} b_{j,i}(x_{j,i} - x_{j,i-1}^2)^2, \tag{27}$$

where $x_{j,i}, \mu \in \mathbb{R}$, and a, $b_{j,i} \in \mathbb{R}^+$ are arbitrary constants. Its evidence is

$$\mathcal{Z} = \sqrt{\frac{\pi^n}{a \cdot \prod b_{j,i}}}, \tag{28}$$

where $n = (n_1 - 1) \cdot n_2 + 1$, and n_1 is the number of dimensions in each of the n_2 groups. Per Goodman & Weare (2010), a scale factor of 1/20 was chosen, with $a = 1/20$ and all $b_{j,i} = 100/20$, so the distribution is shaped like a narrow ridge, providing a challenging posterior shape (see Fig. 3).

To assess temperature evolution in the 2D case, the APT set-up was changed to [24, 120, 1024, 1], increasing the number of temperatures to 24 and the sweeps to 1 024, while decreasing the walkers to 120, roughly maintaining the total iterations the same as in previous benchmarks.

Sampler performance. The SAR method ranks best with 14.65 ± 0.59 kenits, followed closely by SGG, GAO, and ETL. SMD ranks last, with 6.40 ± 0.41 kenits. For the DNS methods, dyn-rs ranks first with 0.77 ± 0.02 kenits, and dyn-u ranks last with 0.34 ± 0.02 kenits, and almost twice the wall-time (see Table 5). In this scenario, SAR presents increased kenits over dyn-rs by a factor of 19.03, almost 20 times more efficient at producing independent samples.

Evidence estimation. The H+ method relies completely on SS+, dropping TI+ segments (see Table 6). This can be attested by the poor TI+ performance compared to SS+. The slice-sampling methods present negative log-likelihoods, indicating substantially underestimated errors.

A 3D case for the Rosenbrock function was also explored (results discussed in Section 6). Generally, the trends held, with `reddemcee` outperforming in sampling efficiency, and all ladder methods producing accurate evidences (see Table 11).

Table 5. 2D hybrid Rosenbrock performance benchmark.

Method	Time (s)	Eff (%)	Kenits
SAR	19.15±0.20	9.51±0.38	**14.65±0.59**
SMD	19.14±0.14	4.15±0.28	6.40±0.41
SGG	19.44±0.26	9.03±0.53	13.70±0.69
GAO	19.41±0.13	8.89±0.48	13.51±0.74
ETL	19.59±0.18	9.04±0.50	13.61±0.83
dyn-u	38.28±2.85	8.04±0.81	0.34±0.02
dyn-s	21.74±0.93	2.18±0.06	0.59±0.03
dyn-rs	16.52±0.42	2.72±0.09	0.77±0.02

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s sampling methods with the 2D Rosenbrock function. From left to right: Time – run time in seconds, Eff – sampling efficiency or percentage of independent samples, and kenits – thousand effective samples per second.

Table 6. 2D hybrid Rosenbrock SAR evidence-estimation comparison.

Method	$\ln \widehat{\mathcal{Z}}$	$\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}$	$\Delta_{\mathcal{Z}}(\%)$	$\ln \mathcal{L}_{\widehat{\mathcal{Z}}}$
TI	1.416±0.022	0.912±0.025	34.431	–0.934
SS	1.833±0.020	0.011±0.001	**0.515**	**3.479**
H	1.833±0.020	0.011±0.001	**0.515**	**3.479**
TI+	1.781±0.022	0.227±0.017	5.565	0.531
SS+	1.833±0.022	0.012±0.001	**0.515**	**3.399**
H+	1.833±0.022	0.012±0.001	**0.515**	**3.399**
dyn-u	1.881±0.089	0.076±0.001	**4.394**	**1.501**
dyn-s	2.007±0.221	0.074±0.004	18.412	–0.936
dyn-rs	1.979±0.164	0.074±0.002	15.258	–0.144

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s uniform (dyn-u), slice (dyn-s), and random-slice (dyn-rs) sampling methods. From left to right we show the log-evidence estimate, the estimate uncertainty, the difference to the true value $\ln \mathcal{Z}$=1.838 in percentage $\Delta_{\mathcal{Z}}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

5. Exoplanet detection from radial velocities

A real-world application of the APT ladder dynamics developed in Section 3 and evidence-estimation methods introduced in Section 2.4 is presented for the nearby G-dwarf HD 20794. Hosting at least three super-Earths whose RV semi-amplitudes are lower than 1 m s^{-1}, the system combines low amplitudes, multiplicity, and a long-period planet, all features that make model selection and evidence estimation particularly demanding. The innermost signals at 18 and 90 days were first identified in the HARPS discovery paper (Pepe et al. 2011), while subsequent analyses revealed an outer candidate with a period of ∼650 d that spends a significant fraction of its eccentric orbit inside the stellar habitable zone (Nari et al. 2025).

As such, this system provides an ideal test-bed for assessing the performance of our APT dynamics and evidence-estimation algorithms on a real-world example. The main features that make exoplanet RV fitting challenging are directly reflected in the previous benchmarks: the strong phase transition in C_v is captured by the Gaussian shells, multi-modality and mode completeness in a many-alias situation (for a sparsely sampled period) in the Gaussian egg-box, and the curved, highly correlated ridges typically seen in the angular orbital parameters are emulated by

the hybrid Rosenbrock function. In this section, the work by Nari et al. (2025) was followed, with us applying `reddemcee` to the combined HARPS+ESPRESSO time series, contrasting the resulting evidence with that reported in the literature.

5.1. Model and likelihood

Each exoplanet is characterised by five Keplerian parameters, P – the period, K – the semi-amplitude, e – the eccentricity, $\bar{\omega}$ – the longitude of periastron, and M_0 – the phase of periastron passage,

$$\mathcal{K}_j(t) = K_j \cdot [\cos(\nu_j(t, P_j, e_j, M_{0j}) + \bar{\omega}_j) + e_j \cos(\bar{\omega}_j)], \tag{29}$$

where the subscript j indicates each exoplanet and ν_j is the true anomaly. In addition, each instrument is given an offset, γ (an additive constant), and a jitter, $\sigma_{\rm INS}$ (added in quadrature to the measurement error, corresponding to a white noise component). Therefore, under the assumption of Gaussian-like errors, the log-likelihood is

$$\ln \mathcal{L} = -\frac{1}{2} \sum_{\rm INS} \sum_{i}^{N_{\rm INS}} \left(\frac{\xi^2_{i,\rm INS}}{(\sigma_i^2 + \sigma^2_{\rm INS})} + \ln(\sigma_i^2 + \sigma^2_{\rm INS}) \right) - \frac{N \ln(2\pi)}{2}, \tag{30}$$

where the sub-indices i and INS correspond to the i-th RV measurement (taken at a time t_i) and to each instrument, respectively. The residuals, $\xi_{i,\rm INS}$, are defined as the difference between the data and the model. Three distinct datasets plus three Keplerian signals add up to a total of 21 parameters or dimensions.

5.2. Parameter estimation and model selection

For an in-depth methodology follow-up, refer to Appendix B. The models compared are: (1) just white noise (H_0); (2) a single sinusoid ($1S$); (3) two sinusoids ($2S$); (4) three sinusoids ($3S$); (5) three Keplerians ($3K$); and (6) four Keplerians ($4K$). The sinusoidal models (S) treat planets as having circular orbits. The $3K$ model corresponds to the three confirmed planets, and $4K$ to the inclusion of an additional candidate.

For the evidence estimation (see Table 7), dyn-rs is consistent with SAR, up to the most complex models. For the $3S$, $3K$, and $4K$ models, they have a 4.6, 6.8, and 4.6 $\ln \widehat{\mathcal{Z}}$ difference, respectively. On the other hand, the SAR log-evidence is all within the reference values' uncertainties except for the $1S$ case, which shows a moderate discrepancy (∼8.7). This could be due to random variation or differences in how noise was treated, but importantly the model ranking is unaffected.

Table 8 compares the estimated evidence errors to the actual run-to-run scatter (the ratio $\widehat{\sigma}_{\mathcal{Z},\rm estimated}/\widehat{\sigma}_{\mathcal{Z},\rm empirical}$). An ideal ratio is 1. Values $\ll$1 mean the method underestimates its true uncertainty, and $\gg$1 means overestimation. The dyn-rs ratios are all very low for complex models (0.04–0.12), severely underestimating the uncertainty with overly confident evidences, whereas `reddemcee` yields values much closer to unity (e.g., 0.63 for the $3K$). Underestimation of error in multi-planet cases could lead to overconfident claims of detection; therefore, `reddemcee` (along the H+ estimator) provides more reliable uncertainties.

Performance-wise, APT achieves a higher kenits count in the simpler models, with a turnover point at $3K$ in favour of dyn-rs. This may be attributable to the modest fixed number of walkers (256) in the APT sampler (see Appendix B for details). In practice, one could adjust the walker count for efficiency's sake, but it was kept constant for consistency between benchmarks.

For the parameter estimation, the two methods produce consistent results (see Table 9). Notably, the APT-derived parameter posteriors are slightly more asymmetrical and tighter (see K_2 and K_3). Also, the eccentricity estimates seem less averse to boundary solutions, as is seen in e_1=$0.057^{+0.045}_{-0.057}$ and e_2=$0.026^{+0.007}_{-0.026}$, both of which are consistent with $e = 0$.

Table 7. HD 20794 evidence-estimation comparison.

Model	Reference[(1)]	dyn-rs	SAR
H_0	–111.9±0.2	–111.2±0.1	–111.9±0.1
1 S	–72.9±3.0	–65.1±2.7	–64.2±0.1
2 S	–41.7±3.7	–40.2±2.4	–40.2±0.2
3 S	–12.3±2.0	–16.0±1.1	–11.5±0.1
3 K	–3.6±0.0[(2)]	–12.9±2.6	–5.8±0.5
4 K	–1.6±4.6	–4.6±3.8	0.0±1.0

Notes. (1) Nari et al. (2025); (2) unreported standard deviation. Difference between the model's mean evidence and the evidence of the best-ranking model. The models shown from top to bottom are: white noise only (H_0), increasing sinusoids (S), and increasing Keplerians (K). The reference evidences have an offset added to make the H_0 evidence equivalent to the SAR.

Table 8. HD 20794 evidence-estimation uncertainty and performance.

	$\frac{\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}}{\sqrt{\mathrm{Var}[\ln \widehat{\mathcal{Z}}]}}$		kenits (eff)	
Model	dyn-rs	SAR	dyn-rs	SAR
H_0	2.30	11.4	0.375 (2.5)	0.889 (7.8)
1 S	0.05	0.48	0.193 (1.3)	0.679 (4.7)
2 S	0.07	0.75	0.120 (1.0)	0.242 (1.9)
3 S	0.12	2.03	0.097 (0.8)	0.207 (1.7)
3 K	0.07	0.63	0.051 (0.5)	0.046 (0.4)
4 K	0.04	0.27	0.039 (0.5)	0.030 (0.3)

Notes. The left columns show the ratio of the mean of the estimated evidence uncertainty against the standard deviation of the run-to-run evidence. The right columns show the average kenits, with the efficiency in parentheses. The models are white noise only (H_0), increasing sinusoids (S), and increasing Keplerians (K).

6. Discussion

The evidence-estimation algorithms introduced in Section 2.4 have proven to be far more reliable than their unmodified counterparts. Furthermore, whether to apply the TI or the SS method is mostly problem-dependent, and the hybrid algorithm seems to be a good compromise when there is no a priori information about the problem at hand. Moreover, our proposed evidence estimators provide a solid, if not better, alternative to dedicated evidence-sampling methods. This is exemplified by the real-world case studied in Section 5.

Across the simplest problems we tested (few dimensions, nearly constant specific heat), all ladder strategies converge to an almost geometric ladder and yield a similar performance. Differences emerge as the problem's complexity grows, and since a

Table 9. HD 20794 parameter estimation.

Parameter	Reference[1]	dyn-rs	SAR
P_1 (days)	18.314±0.002	$18.314^{+0.004}_{-0.004}$	$18.313^{+0.001}_{-0.001}$
K_1 (m s^{-1})	0.614±0.048	$0.632^{+0.030}_{-0.066}$	$0.608^{+0.030}_{-0.009}$
e_1	$0.064^{+0.065}_{-0.046}$	$0.109^{+0.049}_{-0.068}$	$0.057^{+0.045}_{-0.057}$
P_2 (days)	89.68±0.10	$89.67^{+0.11}_{-0.16}$	$89.73^{+0.06}_{-0.03}$
K_2 (m s^{-1})	$0.502^{+0.048}_{-0.049}$	$0.510^{+0.038}_{-0.054}$	$0.474^{+0.041}_{-0.001}$
e_2	$0.077^{+0.084}_{-0.055}$	$0.055^{+0.053}_{-0.054}$	$0.026^{+0.007}_{-0.026}$
P_3 (days)	$647.6^{+2.5}_{-2.7}$	$652.1^{+4.5}_{-5.5}$	$648.2^{+2.6}_{-2.4}$
K_3 (m s^{-1})	$0.567^{+0.067}_{-0.064}$	$0.521^{+0.025}_{-0.091}$	$0.589^{+0.001}_{-0.055}$
e_3	$0.45^{+0.11}_{-0.10}$	$0.428^{+0.111}_{-0.412}$	$0.467^{+0.053}_{-0.024}$

Notes. (1) Nari et al. (2025). Parameter estimation results for the different algorithms for the $3K$ model. In descending order, P corresponds to the period, K to the semi-amplitude, and e to the eccentricity, where the subscript denotes a particular planet.

Table 10. Gaussian shells with increasing dimensions' performance.

Method	2D	5D	10D	15D
SAR	3.70±0.12	3.59±0.12	2.71±0.11	2.25±0.07
SMD	3.73±0.13	3.73±0.14	3.21±0.14	**2.80±0.09**
SGG	3.67±0.13	3.50±0.11	2.45±0.07	1.85±0.07
GAO	3.60±0.11	3.61±0.15	2.84±0.07	2.36±0.07
ETL	3.66±0.10	3.66±0.12	2.89±0.07	2.45±0.13
dyn-u	0.19±0.04	0.37±0.04	0.32±0.05	0.18±0.01
dyn-s	0.75±0.04	0.43±0.01	0.23±0.01	0.17±0.01
dyn-rs	0.94±0.05	0.70±0.01	0.46±0.01	0.37±0.01

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s sampling methods. The kenits – effective samples per second in thousands – are shown from left to right with increasing dimensions.

common problem astrophysicists face is that of dimensionality, we discuss it further next.

6.1. Increasing dimensionality

We took the Gaussian shells problem described in Sect. 4.1 and increased its dimensionality without modifying any hyperparameters, in order to demonstrate how the methods scale. As dimensionality grows, we see a very much expected efficiency decrease (see Table 10), as well as the SMD outperforming other methods. At 15 dimensions, it becomes the most effective, with a margin far exceeding the run-to-run scatter. In second place, GAO and ETL come tied (with overlapping confidence intervals), being around 14% slower, followed by the SAR – 19% slower – and the SGG – 34% slower. This supports the intuition that maximising swap distance is beneficial in high-dimensional spaces where local energy traps abound. On the other hand, dyn-rs, still the best DNS method, presents kenits at ∼13% of the SMD average.

For evidence estimation, with the modest amount of temperatures proposed for this benchmark, the TI method degrades considerably, with a $\Delta_\mathcal{Z}$ of ∼ 72% (see Table A.1). The TI+ soothes this, with $\Delta_\mathcal{Z}$ 11.3%. Nevertheless, the error due the discretisation remains a problem in this benchmark. For the SS method, the evidence estimate is excellent ($\Delta_\mathcal{Z}$=2.56), with the highest $\mathcal{L}(\widehat{\mathcal{Z}})$=2.1658. The SS+ method presents a slight deterioration over its classic counterpart. The H+ presents a higher likelihood (and lower $\Delta_\mathcal{Z}$) than either TI+ or SS+.

Table 11. Hybrid 3D Rosenbrock evidence estimation.

Method	$\ln \widehat{\mathcal{Z}}$	$\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}$	$\Delta_\mathcal{Z}$	$\ln \mathcal{L}_{\hat{\mathcal{Z}}}$
SAR	1.562±0.062	0.021±0.001	4.233	0.771
SMD	1.516±0.065	0.023±0.002	8.519	–4.556
SGG	1.574±0.076	0.021±0.002	3.069	1.841
GAO	1.570±0.042	0.022±0.002	3.500	1.561
ETL	1.598±0.061	0.021±0.001	**0.747**	**2.901**
dyn-u	1.923±0.207	0.109±0.002	37.366	–2.944
dyn-s	1.625±0.883	0.111±0.004	1.967	1.264
dyn-rs	1.534±0.442	0.112±0.006	6.903	1.066

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s sampling methods. From left to right we show the log-evidence estimate, the estimate error, the difference to the true value $\ln \mathcal{Z}$=1.606 in percentage in percentage $\Delta_\mathcal{Z}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

By increasing the dimensionality of the Rosenbrock function to 3D, the evidence estimation results (H+) are best for ETL with $\mathcal{L}(\widehat{\mathcal{Z}})$=2.901, followed by SGG and GAO, with $\mathcal{L}(\widehat{\mathcal{Z}})$=1.841 and 1.561, respectively (see Table 11). The kenits values present a similar trend to that of the shells, with 2.98 (SAR), 2.50 (SMD), 3.46 (SGG), 3.03 (GAO), and 2.85 (ETL), and 0.02 (dyn-u), 0.55 (dyn-s), and 0.74 (dyn-rs).

6.2. Burn-in

How fast the adaptation optimally distributes the ladder depends on both the problem at hand and the adaptation hyper-parameters ν_0 and τ_0. We stress-tested adaptation by setting the 15D shells with 640 sweeps (as in Sect. 4.2 and Sect. 6.1) to use 320, 160, 64, and 32 sweeps as burn-in (see Table A.2). H+ is the method most resilient to an uncalibrated ladder. It is the only method that retains positive $\mathcal{L}(\widehat{\mathcal{Z}})$ values after the 25% burn-in, with values of $\ln \widehat{\mathcal{Z}}|_{H+}$=–24.939±0.037, –24.980±0.031, –24.984±0.034, and –24.996±0.041 for 50%, 25%, 10%, and 5%, respectively.

6.3. Ladder adaptation

For low-dimensional targets (2D shells, egg-box), every method settles on an almost geometric ladder, due to $C_\nu(\beta)$ being almost flat. As the dimensionality grows, C_ν develops a broad peak around the phase-transition region, and the objectives react differently. For example, the SAR spreads the temperatures uniformly in swap probability, and therefore stretches the ladder on both sides of the peak, while SMD concentrates the replicas where the specific heat is large, maximising the average information transfer, as is seen by its clear lead at 15D in Table 10.

Longer runs are realised for this scenario, 15D shells, with 10 000 sweeps instead of 640, to examine the convergence of the ladder adaptation. All schemes compress the first ∼200 sweeps into a fast logistic-like relaxation of the temperature gaps, after which the diminishing adaptive factor, $\kappa(t)$ (Eq. (14)), forces a slow power-law tail. In practice, almost the whole ladder shape

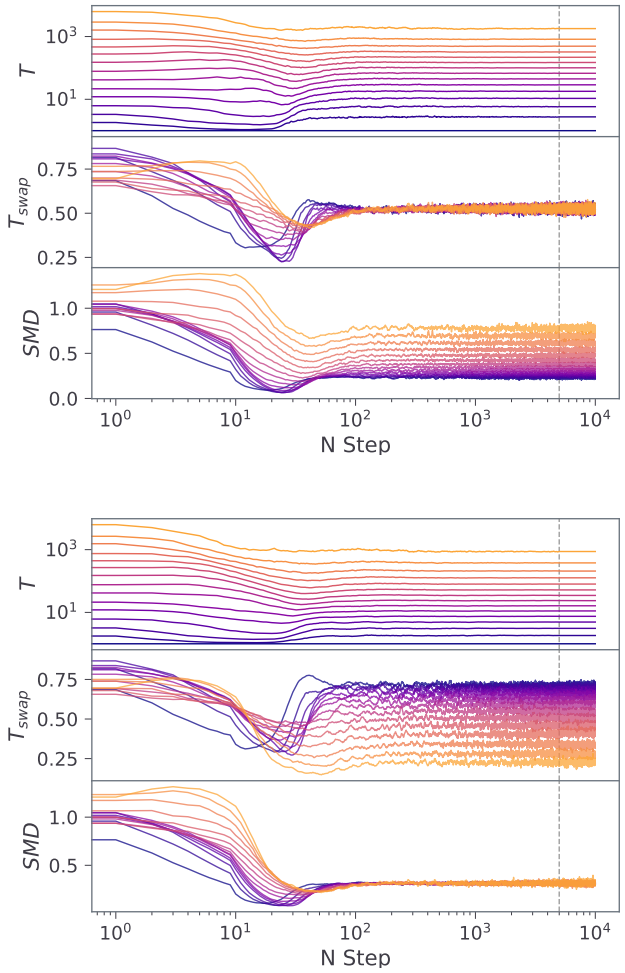

Fig. 4. Temperature ladder evolution for the 15D Gaussian shells in the SAR regime (top) and SMD regime (bottom). In the x-axis is the current iteration. From top to bottom are the temperature ladder evolution, T, the swap acceptance ratio, T_{swap}, and the SMD. Colours represent each chain, where blue is the coldest (β=1), with temperature increasing towards the red, where the hottest chain is omitted. The vertical dashed black line indicates where the adaptation stops.

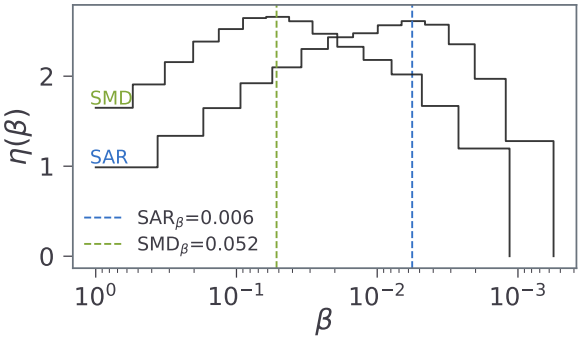

Fig. 5. Chain density per temperature range for the SAR and SMD methods in the 15D Gaussian shells. The dashed blue line denotes the maximum of the SAR, with the green one representing the maximum of the SMD. Both lines also denote the regions where the TI or SS algorithms are applied in the hybrid method.

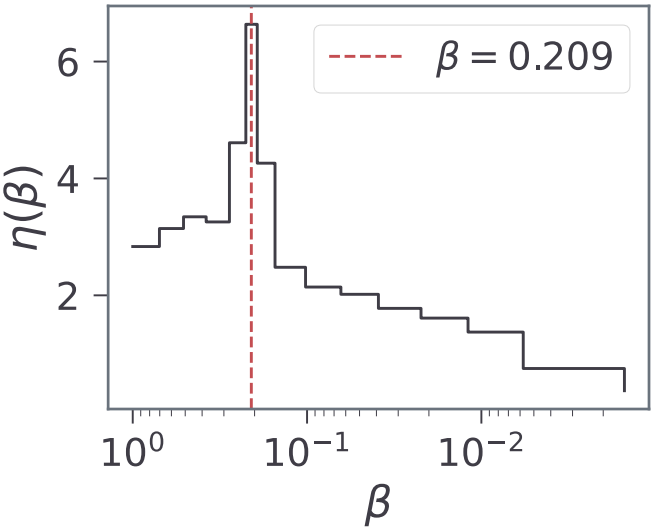

Fig. 6. Chain density per temperature range for the HD 20794 3K SAR run. The vertical dashed red line marks the distribution maximum.

is set in under 5% of the computational budget, supporting the choice of freezing the ladder after a user-defined burn-in.

Fig. 4 displays the ladder evolution of the SAR and SMD methods, respectively. A notable difference is that the SMD method has a layered swap rate, whereby the cold chain swaps frequently (0.7) and the hot chain seldom does (0.2), whereas the SAR keeps a steady swap rate of 0.5 for all temperatures.

Fig. 5 shows the final ladders as chain density (proportional to C_v) as a function of temperature. The dashed vertical lines indicate where $\Delta\beta$ is smaller, as well as where the transition between TI and SS is applied in the hybrid method.

6.4. Number of temperatures

A natural question that arises is how many temperatures are enough. The answer, as one might expect, is that it depends. When doing an APT run, there are two objectives in mind: posterior and evidence estimation. More temperatures might be redundant posterior-wise, not contributing with additional information to the cold chain. At the same time, more temperatures provide a finer grid for the evidence estimation. Nonetheless, it is possible to dabble into this question, by analysing the temperature evolution of the ladder. If the mixing is low, increasing the temperatures will result in a more efficient run. This may be measured by comparing the additional chains computational cost against the efficiency (or kenits) increase. Visualising the likelihood variance as a function of temperature (the TI) can be helpful. A smoothly increasing curve will not be affected as much as a ladder-like curve by decreasing temperatures. This can also be appreciated by the C_v plot. A ladder-like behaviour would be seen as multiple sharp peaks (as in the HD 20794 3K model, see Fig. 6 with a secondary small peak at around β=0.4), whereas a smooth curve is seen as a single smooth peak (as in the 15D shells, see Fig. 5).

Another matter is how the different evidence-estimation algorithms proposed are affected. The 15D shells (with 10 000 sweeps) were re-run with N_β=6, 8, 12, and 16 (see Table A.3). The TI method by decreasing temperatures presents a linear fall-off with $\sqrt{\mathcal{L}_{\widehat{\mathcal{Z}}}}$. This result can also be easily derived from Eq. (8), which reveals that the log-evidence error is proportional to $\frac{1}{\text{ntemps}}$. The TI+ method presents a huge improvement over

TI, although still fails to reach the mark in temperature-poor regimes. The SS methods remain extremely accurate even at low $N_\beta = 6$, with $\Delta_{\mathcal{Z}}$=1.78 and $\mathcal{L}(\widehat{\mathcal{Z}})$=2.02. The geometrical bridge in SS+ yields a small accuracy gain (lower $\Delta_{\mathcal{Z}}$ for every N_β), while bringing a small drop in $\mathcal{L}(\widehat{\mathcal{Z}})$, due to the slight increase in the error estimate.

7. Conclusions

We have introduced `reddemcee`, an APT ensemble sampler that combines three next-level techniques – flexible ladder adaptation, robust evidence estimation, and practical error quantification – into a single, easy-to-deploy package. Three of the five ladder adaptation algorithms are new implementations from this work (SMD, SGG, ETL), reliable alternatives to the most commonly used SAR method, as is demonstrated in our benchmarks (see Section 4). All of these methods converge in a few sweeps and deliver cold-chain mixing efficiencies around an order of magnitude higher than DNS across our tests. In the challenging 15D Gaussian-shell benchmark, `reddemcee` sustained > 2 kenits, about 7 times faster than the best DNS configuration, while the SMD ladder proved particularly resilient as dimensionality grew.

We devised three evidence estimators – two original implementations, TI+ and SS+, and a novel hybrid approach – that combine curvature-aware interpolation with bridge sampling. These estimators yield accurate log-evidence and maintain realistic uncertainties even when the number of temperatures is decreased beyond what is optimal.

A real-world application to the HD 20794 RV dataset shows that `reddemcee` reproduces literature model rankings, recovers planetary parameters with tighter – yet statistically consistent – credible intervals, and supplies evidence uncertainty that closely tracks run-to-run dispersion. Crucially, this performance was obtained without manual ladder tuning: the same hyperparameters were used from no-planet to four-planet models and from 6 to 21 dimensions.

Overall, `reddemcee` demonstrates that a carefully engineered APT sampler can match, and occasionally surpass, state-of-the-art DNS, both in sampling throughput and in evidence estimation, while retaining the posterior-inference strengths that make MCMC indispensable. Future work will explore on-the-fly convergence diagnostics, further widening the sampler's applicability to the increasingly complex problems faced in modern astrophysics and beyond.

Acknowledgements. PAPR and JSJ gratefully acknowledge support by FONDECYT grant 1240738, from the ANID BASAL project FB210003, and from the CASSACA China-Chile Joint Research Fund through grant CCJRF2205. For the n-d Gaussian shells, Gaussian egg-box, and Rosenbrock function benchmarks parallelisation was not used and computing was limited to single-core for all algorithms. For the exoplanet detection benchmark, parallelisation was used with 24 threads. All the benchmarks were performed on a computer with an AMD Ryzen Threadripper 3990X 64-Core Processor with 128Gb of DDR4 3200Mhz RAM. We thank the anonymous referee for insightful suggestions that enhanced the quality of this manuscript. We are grateful to Fabo Feng and Mikko Tuomi for stimulating early discussions on planet detection. We would also like to thank Dan Foreman-Mackey for the excellent library `emcee` (Foreman-Mackey et al. 2013), which opened a most exciting new world. Typesetting was carried out in `Overleaf` (Digital Science UK Ltd. 2025).

References

Diamond-Lowe, H., Charbonneau, D., Malik, M., Kempton, E. M. R., & Beletsky, Y. 2020, AJ, 160, 188
Digital Science UK Ltd. 2025, Overleaf
Drummond, A., Nicholls, G., Rodrigo, A., & Solomon, W. 2002, Genetics, 161, 1307
Earl, D. J., & Deem, M. W. 2005, Phys. Chem. Chem. Phys. (Incorp. Faraday Trans.), 7, 3910
Feroz, F., & Hobson, M. P. 2008, MNRAS, 384, 449
Feroz, F., Balan, S. T., & Hobson, M. P. 2011, MNRAS, 415, 3462
Flegal, J. M., & Jones, G. L. 2008, arXiv e-prints [arXiv:0811.1729]
Ford, E. B., Lystad, V., & Rasio, F. A. 2005, Nature, 434, 873
Foreman-Mackey, D. 2016, J. Open Source Softw., 1, 24
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Fritsch, F. N., & Butland, J. 1984, SIAM J. Sci. Statist. Comput., 5, 300
Gelman, A., & Meng, X.-L. 1998, Statist. Sci., 13, 163
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 2004, Bayesian Data Analysis, 2nd edn. (Chapman and Hall/CRC)
Gilks, W. R., Roberts, G. O., & Sahu, S. K. 1998, J. Am. Statist. Assoc., 93, 1045
Goggans, P. M., & Chi, Y. 2004, in American Institute of Physics Conference Series, 707, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, eds. G. J. Erickson, & Y. Zhai, 59
Gonzalez, G. 1997, MNRAS, 285, 403
Goodman, J., & Weare, J. 2010, Commun. App. Math. Comput. Sci., 5, 65
Gregory, P. C. 2005, ApJ, 631, 1198
Gronau, Q. F., Sarafoglou, A., Matzke, D., et al. 2017, J. Math. Psychol., 81, 80
Hansmann, U. H. E. 1997, Chem. Phys. Lett., 281, 140
Higson, E., Handley, W., Hobson, M., & Lasenby, A. 2019, Statist. Comput., 29, 891
Hou, F., Goodman, J., Hogg, D. W., Weare, J., & Schwab, C. 2012a, ApJ, 745, 198
Hou, F., Goodman, J., Hogg, D. W., Weare, J., & Schwab, C. 2012b, ApJ, 745, 198
Jenkins, J. S., Díaz, M. R., Kurtovic, N. T., et al. 2020, Nat. Astron., 4, 1148
Katzgraber, H. G., Trebst, S., Huse, D. A., & Troyer, M. 2006, J. Statist. Mech. Theory Exp., 2006, 03018
Kofke, D. 2002, jcph, 117, 6911
Kone, A., & Kofke, D. 2005, J. Chem. Phys., 122, 206101
Lartillot, N., & Philippe, H. 2006, Syst. Biol., 55, 195
Marcy, G. W., & Butler, R. P. 2000, PASP, 112, 137
Maturana-Russel, P., Meyer, R., Veitch, J., & Christensen, N. 2019, Phys. Rev. D, 99, 084006
Mayor, M., & Queloz, D. 1995, Nature, 378, 355
Meng, X.-L., & Wong, W. 1996, Statist. Sin., 6, 831
Miasojedow, B., Moulines, E., & Vihola, M. 2013, J. Computat. Graph. Statist., 22, 649
Nari, N., Dumusque, X., Hara, N. C., et al. 2025, A&A, 693, A297
Oaks, J. R., Cobb, K. A., Minin, V. N., & Leaché, A. D. 2018, arXiv e-prints [arXiv:1805.04072]
Pagani, F., Wiegand, M., & Nadarajah, S. 2019, arXiv e-prints [arXiv:1903.09556]
Pepe, F., Lovis, C., Ségransan, D., et al. 2011, A&A, 534, A58
Predescu, C., Predescu, M., & Ciobanu, C. V. 2004, J. Chem. Phys., 120, 4119
Rannala, B., & Yang, Z. 1996, J. Mol. Evol., 43, 304
Rathore, N., Chopra, M., & de Pablo, J. 2005, J. Chem. Phys., 122, 024111
Roberts, G., & Rosenthal, J. 2007, J. Appl. Probab., 44
Shenfeld, D. K., Xu, H., Eastwood, M. P., Dror, R. O., & Shaw, D. E. 2009, Phys. Rev. E, 80, 046705
Skilling, J. 2004, in American Institute of Physics Conference Series, 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, eds. R. Fischer, R. Preuss, & U. V. Toussaint, 395
Speagle, J. S. 2020, MNRAS, 493, 3132
Sugita, Y., & Okamoto, Y. 1999, Chem. Phys. Lett., 314, 141
Swendsen, R. H., & Wang, J.-S. 1986, Phys. Rev. Lett., 57, 2607
van der Sluys, M., Raymond, V., Mandel, I., et al. 2008, Class. Quant. Grav., 25, 184011
Veitch, J., Raymond, V., Farr, B., et al. 2015, Phys. Rev. D, 91, 042003
Vines, J. I., Jenkins, J. S., Berdiñas, Z., et al. 2023, MNRAS, 518, 2627
Vousden, W. D., Farr, W. M., & Mandel, I. 2016, MNRAS, 455, 1919
Wang, Y.-B., Chen, M.-H., Kuo, L., & Lewis, P. O. 2018, Bayesian Anal., 13, 311
Xie, W., Lewis, P., Fan, Y., Kuo, L., & Chen, M.-H. 2011, Syst. Biol., 60, 150

Appendix A: Gaussian shells

Table A.1. 15D Gaussian shells evidence-estimation comparison with the SAR method.

Method	$\ln \widehat{\mathcal{Z}}$	$\widehat{\sigma}_{\ln \widehat{\mathcal{Z}}}$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$
TI	-26.1818±0.0417	1.1037±0.0526	71.9266	-1.6800
SS	-24.9374±0.0410	0.0205±0.0029	**2.5617**	**2.1658**
Hybrid	-24.9523±0.0423	2.9866±0.0661	4.0070	-2.0132
TI+	-25.0308±0.0375	0.5979±0.0266	11.2557	-0.4245
SS+	-24.9411±0.0369	0.0204±0.0021	2.9299	1.9090
Hybrid+	-24.9394±0.0373	0.0242±0.0027	**2.7599**	**2.1337**
dyn-u	-24.7176±0.3144	0.1246±0.0017	21.3853	-0.0460
dyn-s	-24.8765±0.3044	0.1243±0.0015	**3.5620**	**1.1267**
dyn-rs	-24.7292±0.2163	0.1242±0.0012	19.9854	0.0908

Notes. `reddemcee`'s adaptive algorithms compared to `dynesty`'s sampling methods. From left to right, the log-evidence estimate, the estimate uncertainty, the difference to the true value $\ln \mathcal{Z}$=−24.9114 in percentage $\Delta_{\mathcal{Z}}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

Table A.2. 15D Gaussian shells $\ln \mathcal{Z}$ estimation with variable burn-in for the SAR method.

Method	$N_{\rm adapt}$=50%		$N_{\rm adapt}$=25%		$N_{\rm adapt}$=10%		$N_{\rm adapt}$=5%	
	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$
TI	71.9266	-1.6800	72.9348	-1.6919	71.0282	-1.8322	70.176	-22.5295
SS	2.5617	2.1658	7.0111	-5.9413	6.7014	-4.9155	8.3851	-8.4817
Hybrid	4.0070	-2.0132	21.8509	-1.4656	27.8067	-1.0694	17.2876	-2.0419
TI+	11.2557	-0.4245	14.1321	-0.5039	12.6212	-0.3977	12.3431	-0.3729
SS+	2.9299	1.9090	6.3493	-4.2133	6.3843	-4.7487	8.1264	-8.5345
Hybrid+	2.7599	2.1337	6.6707	1.2514	6.9972	1.0890	8.1322	0.9562

Notes. `reddemcee`'s SAR method evidence estimation with different adaptation (and burn-in) times, starting at half of the chain $N_{\rm adapt}$=50%, down to 25%, 10%, and 5% for 640 sweeps. Each sub-column shows the difference to the true value $\ln \mathcal{Z}$=−24.9114 in percentage $\Delta_{\mathcal{Z}}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

Table A.3. 15D Gaussian shells $\ln \mathcal{Z}$ estimation with variable temperatures for the SAR method.

Method	$N_\beta = 16$		$N_\beta = 12$		$N_\beta = 8$		$N_\beta = 6$	
	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$	$\Delta_{\mathcal{Z}}$	$\mathcal{L}(\widehat{\mathcal{Z}})$
TI	70.7557	-1.6760	90.4453	-2.5408	99.8465	-4.0657	>100.0	-5.2445
SS	0.5806	3.6858	0.3422	4.1314	0.1161	4.1473	1.7751	2.0183
Hybrid	15.5426	-1.8808	10.6859	0.0712	18.2425	-2.7388	43.9598	-3.0542
TI+	8.6369	-0.4128	25.0197	-1.3178	77.9445	-2.5999	99.2296	-3.5185
SS+	0.5052	3.8631	0.5405	3.7914	0.0666	4.3621	1.3469	1.8170
Hybrid+	0.6942	1.8130	0.5567	1.7850	0.8203	-0.3522	8.5362	-1.6289

Notes. `reddemcee`'s SAR method evidence estimation with N_β=16, 12, 8, and 6 temperatures. Each sub-column shows the difference to the true value $\ln \mathcal{Z}$=−24.9114 in percentage $\Delta_{\mathcal{Z}}$, and the log-likelihood of the estimator $\mathcal{L}(\widehat{\mathcal{Z}})$.

Appendix B: HD 20794

Each dataset, HARPS03, HARPS15, and E19, is nightly binned, then sigma clipped (3σ), and measurements where the error is higher than 3 times the median error are excluded, leaving out 512, 231, and 63 RVs respectively, for a grand total of 806 RVs. Priors were matched to those stated by Nari et al. (2025), with unspecified priors left to an educated guess. For offset a prior $\sim\mathcal{U}(-3, 3)$ was chosen, and for jitter $\sim\mathcal{N}(0, 5)$, truncated at [0, 3]. For eccentricity e and longitude of periastron ω, we use the change of variable

$$e_c = \sqrt{e}\cos(\omega)\,; \quad e_s = \sqrt{e}\sin(\omega)\,, \tag{B.1}$$

to linearise the circular parameter ω, improving sampler performance, as in Hou et al. (2012b). Point estimates and uncertainties for all parameters are defined as the posterior maxima and the corresponding 1-σ highest-density intervals (HDIs), rather than the more commonly used medians and corresponding 1-σ percentiles, reflecting our methodological preference for mode-based summaries.

Both methods had matching estimates for the H_0 model (with little variance). This value was chosen as offset for the reference values from the literature. We tried to follow the recipe of executing five runs and selecting the three with highest evidences. But this could not be applied to all models due to inconsistent results.

For the `dynesty` runs (random-slice sampling, 3 000 live-points), the $1S$ model consistently found $P_1 = 18.314$ d, $K = 0.61$ ms^{-1}, with $\ln\hat{\mathcal{Z}} = -1218.1 \pm 2.7$, presenting a 46.1 difference against the H_0 model $\ln\hat{\mathcal{Z}} = -1264.20 \pm 0.04$. The $2S$ run (3 000 live-points) consistently added a signal with $P_2 = 89.6$ d and $K_2 = 0.44$ ms^{-1} (signal subindex were shifted so P_1<P_2, and so on), with $\ln\hat{\mathcal{Z}} = -1193.3 \pm 2.4$, a difference of 24.9 against $1S$.

An extra sinusoid brought multiple solutions in $3S$ at $P_3 =$ 655, 1 018, and 1 426 d, the former being the solution with the highest maximum likelihood. The number of live-points was gradually increased until this solution was obtained at least three times out of five consecutive runs. At 5 000 live-points this solution was obtained consistently (5 out of 5 times), bringing an evidence difference of 24.2 with the $2S$ model.

Changing the sinusoids for Keplerians in the $3K$ model seems to eliminate this problem, since $P_3 = 655$ d appeared consistently amongst runs (3 000 live-points). For consistency, we also did the runs with increasing live-points, but when doing this the algorithm appears to get stuck in lower likelihood peaks more often. With 6 000 live-points we consistently got $P_3 = 1\,378$ d, with an evidence 15 lower than $P_3 = 655$ d, and a maximum likelihood 20 lower. The increase in evidence compared to the $3S$ model (3.1) is tinier than both the reference value and the one obtained with the SAR method.

The $4K$ model did not converge to a single solution, obtaining 85.5 d, 111 d, 1011 d, or 1420 d as the fourth signal. The best evidence was obtained by the period $P_4 = 85.54$ d, while the best likelihood by $P_4 = 111.23$ d.

By comparing the differences between each model's best solution, our dyn-rs results match the reference only for $\Delta\mathcal{Z}(H_0, S\,1)$ and $\Delta\mathcal{Z}(S\,1, S\,2)$, with diverging results in increasing models by 5.3, 2.6, and -4.2, respectively. This could be explained by the high standard deviation (e.g., for the $4K$, 4.6 in the reference and 3.8 in dyn-rs) or to sampler setup differences.

On the other hand, doing the same exercise against the SAR run provides much closer differences to the reference: 0.3, 2.2, -1.5, 1.6, and −0.2, with the highest values corresponding to $\Delta\mathcal{Z}(2S, 3S)$ and $\Delta\mathcal{Z}(3S, 3K)$, and well within variance.

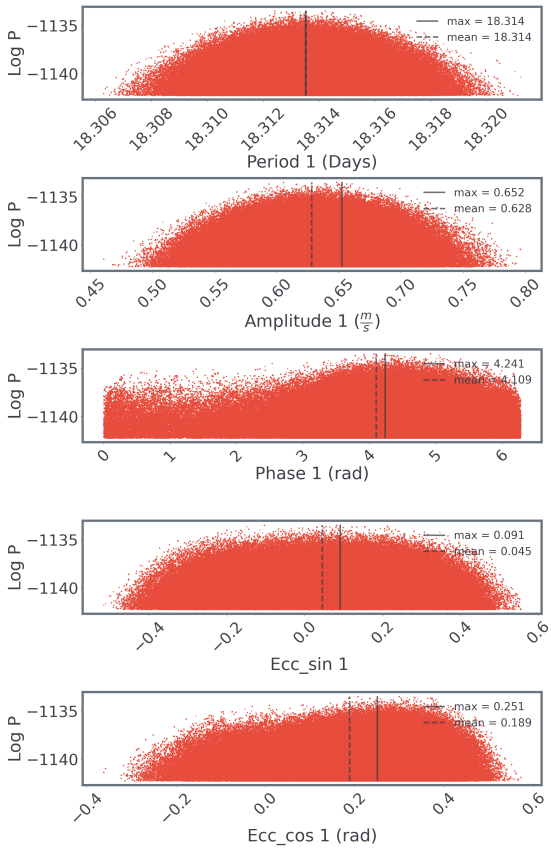

Fig. B.1. HD 20794 posteriors for the first Keplerian in the $3K$ SAR run.

The `reddemcee` runs used 16 temperatures, 256 walkers, and 5 000 sweeps, increasing by 2 500 extra sweeps for each model, with 17 500 sweeps for the $4K$, to facilitate the adaptation stationarity without 'hand-tuning' the adaptive parameters.

The $1S$ model consistently found the $P = 18.314$ d signal, with an evidence of $\ln\hat{\mathcal{Z}} = -1217.1 \pm 0.3$, and a difference of 47.9 with the H_0 model. Adding a signal found $P_2 = 89.66$ d consistently, with an evidence improvement of 23.5. The third sinusoid shifted this period slightly to 89.7 d, whilst adding one at $P_3 = 652.2$ d, and improving the evidence by 29.1. Changing the sinusoids to Keplerians brought an improvement of 5.2 evidence-wise, with eccentricities of $e_1 = 0.057$, $e_2 = 0.026$, $e_3 = 0.467$. The top of the posterior samples for the first Keplerian can be seen in Fig. B.1, with the last two parameters the aforementioned e_s and e_c in Eq. B.1. Additionally, Fig. B.2 (top) shows a corner plot (scatter plot matrix for multidimensional samples) for the same Keplerian, revealing parameter covariances: P_1 and K_1 have tight posteriors, with each off-diagonal panel being nearly circular. Both have a little correlation with eccentricity (denoted by their elongation over e_c, e_s). The centre of the e_c, e_s panel indicates a low value for eccentricity. The direction of its isotropic shape defines ω. The phase of periastron passage M_0 is coupled to the eccentricity vector, the textbook RV degeneracy. The third signal, with a higher well-defined eccentricity, does not present this degeneracy, as seen in Fig. B.2.

Finally, the $4K$ model brought both $P_4\sim$1440 d and ∼111 d, with similar evidences and likelihoods.

The step decrease in kenits for the more complex model is attributed to the low walker count. It is also worth noting the

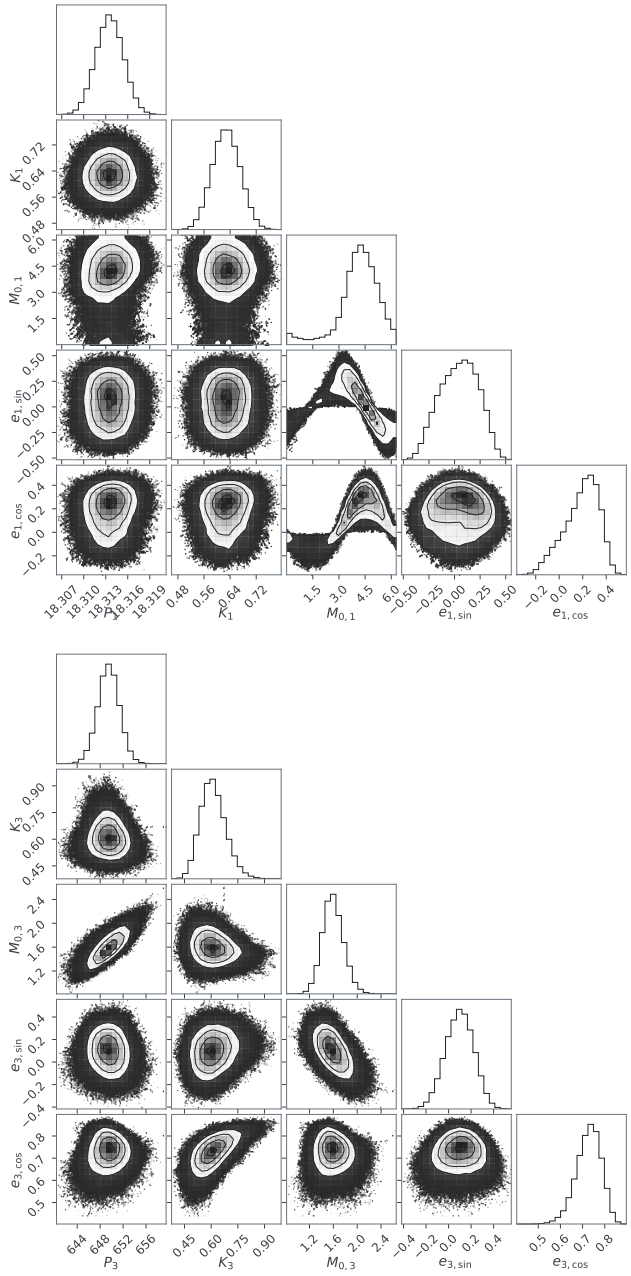

Fig. B.2. HD 20794 corner plot for the first (top) and third (bottom) signals in the $3K$ model. Figures made with the `corner` package (Foreman-Mackey 2016).

difference in both consistency and wall-time in these models. For $3K$ and $4K$ the standard deviation in dyn-rs was 2.6 and 3.8 respectively, compared to 0.4 and 0.9 in `reddemcee`. Furthermore, the average wall-time in dyn-rs was 150.26±32.46, and 112.77±5.2 minutes, whereas in `reddemcee` it was 57.68±2.57, and 77.21±2.42 minutes. Time-wise, the runs were not only significantly shorter, but also more consistent.

Appendix C: Evidence estimators

C.1. Geometric-bridge stepping stones

For the temperature ladder $1 = \beta_1 > ... > \beta_B \geq 0$, at each β_i we have an ensemble of W walkers evolving over T sweeps. For each adjacent pair (β_i, β_{i+1}) and for every sweep, we build a symmetric bridge (Meng & Wong 1996; Gronau et al. 2017) between the two ensembles

$$\begin{aligned} \text{from } \beta_i: \quad & A_{t,i} = \frac{1}{W}\sum_{w=1}^{W} \ln \mathcal{L}^{\Delta\beta_i/2}\,, \\ \text{and from } \beta_{i+1}: \quad & C_{t,i} = \frac{1}{W}\sum_{w=1}^{W} \ln \mathcal{L}^{-\Delta\beta_i/2}\,. \end{aligned} \tag{C.1}$$

If the ladder is still adapting, we use the per-sweep $\Delta\beta_{i,t}$, otherwise (as in this manuscript) a single averaged $\overline{\Delta\beta_i}$. Then we average over sweeps to get $\mu_{A,i} = \overline{A_{t,i}}$ and $\mu_{C,i} = \overline{C_{t,i}}$. Each ratio $\frac{\mu_{A,i}}{\mu_{C,i}}$ estimates r_i via the geometric bridge. Multiplying all these ratios, or adding their logs, gives the log-evidence estimate

$$\ln \widehat{\mathcal{Z}}_{\mathrm{SS+}} = \sum_{i=1}^{B-1} (\ln \mu_{A,i} - \ln \mu_{C,i})\,. \tag{C.2}$$

For the sampling error $\widehat{\sigma}_S^2$ we treat the per-sweep vectors $[A_{t,\cdot} \mid C_{t,\cdot}]$ as a correlated time series across sweeps, estimating the long-run covariance with multivariate OBM, and use the delta method for $g(\mu) = \sum_i (\ln \mu_{A,i} - \ln \mu_{C,i})$ to get $\mathrm{Var}[\hat{z}] = \widehat{\sigma}_S^2$ (Flegal & Jones 2008; Wang et al. 2018). This method effectively steps along the ladder with a geometric bridge at each step, propagating an autocorrelation-aware uncertainty.

C.2. Piecewise interpolated thermodynamic integration

For each sweep $t = 1, \ldots, T$, we take the per-temperature mean log-likelihood $\bar{\ell}_{i,t} = \frac{1}{W}\sum_{w=1}^{W} \ln \mathcal{L}^{\beta_i,t}$. Now, for each sweep we interpolate the curve $\mathbb{E}_\beta[U](\beta)$ with PCHIP (Fritsch & Butland 1984), and integrate, resulting in a series $\hat{z}_t$. The time-average of this series is the evidence estimate

$$\ln \widehat{\mathcal{Z}}_{\mathrm{TI+}} = \sum_{t=1}^{T} \hat{z}_t\,. \tag{C.3}$$

To estimate $\widehat{\sigma}_D$ we form a coarser ladder by dropping every other β, recompute the same PCHIP integral to get $\hat{z}_t^{(2)}$, time-average to $\ln \widehat{\mathcal{Z}}^{(2)}_{\mathrm{TI+}}$, and set

$$\widehat{\sigma}_D = \ln \widehat{\mathcal{Z}}^{(2)}_{\mathrm{TI+}} - \ln \widehat{\mathcal{Z}}_{\mathrm{TI+}}\,. \tag{C.4}$$

To estimate $\widehat{\sigma}_S$, if the ladder is no longer adapting, we treat $\hat{z}_t$ as a correlated univariate time series over the sweeps, estimate its long-run covariance Σ with OBM, and set $\widehat{\sigma}_S^2 = \frac{\Sigma}{T}$. Otherwise, with a ladder still adapting (under diminishing adaptation), with N_β the length of the interpolated values, we take the per-sweep trapezoid TI block as

$$\hat{z}_t = \sum_{i=1}^{N_\beta} \Delta\beta_{i,t} \cdot S_{i,t}\,, \qquad S_{i,t} = \frac{(\bar{\ell}_{i,t} + \bar{\ell}_{i+1,t})}{2}\,. \tag{C.5}$$

With the S_t series, we estimate the multivariate long-run covariance Σ_S with multivariate OBM, and propagate with mean widths $\overline{d\beta} = \overline{\Delta\beta_t}$:

$$\widehat{\sigma}_D^2 = \mathrm{Var}[\hat{z}_t] = \overline{d\beta}^{\mathsf{T}} \Sigma_S \overline{d\beta}\,. \tag{C.6}$$

Note that for the total error, since $\widehat{\sigma}_D$ and $\widehat{\sigma}_S$ are added in quadrature, they are being treated as independent. For the adaptative case, we need a stationary adaptation settled in the ladder, which should happen under diminishing adaptations after some sweeps have passed, depending on the hyper-parameters on $\kappa(t)$ (see Eq. 14).