

Systematic selection of surrogate models for nonequilibrium chemistry

Robin Janssen^{1,2,*}, Lorenzo Branca¹, and Tobias Buck¹

¹ Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany

² Institute of Computer Engineering, Heidelberg University, Heidelberg, Germany

Received 12 January 2026 / Accepted 18 February 2026

ABSTRACT

Context. Nonequilibrium chemistry is central to many astrophysical environments, but remains a major computational bottleneck in simulations because solving the associated stiff, coupled, ordinary differential equation systems is expensive. Neural surrogates promise substantial increases in speed, yet most existing studies are limited to proof-of-concept demonstrations and lack rigorous, dataset-grounded comparisons of architectures or systematic optimization toward accuracy and efficiency.

Aims. We aim to establish a principled procedure for optimizing and selecting surrogate models for astrochemistry that would enable representative and quantitative comparisons across architectures. This requires joint optimization for accuracy and efficiency, suitable metrics for performance assessment, and evaluation of surrogate reliability under practical constraints such as uncertainty quantification (UQ) and iterative prediction.

Methods. To this end, we employed CODES, a benchmarking framework that performs multi-objective hyperparameter tuning, trains optimized configurations, and evaluates their behavior across multiple dimensions. We compared four surrogate families, two fully connected models and two latent-evolution models. These surrogates were optimized and trained on four KROME-generated datasets spanning primordial and molecular-cloud chemistry, with up to 287 reactions across 37 species, including parametric variations in radiation field and metallicity. Each model predicted chemical abundances and temperature over a 10 kyr interval for user-specified output times.

Results. Dual-objective optimization reveals pronounced accuracy–efficiency trade-offs for all architectures and enables substantial efficiency gains with minimal accuracy loss. Across datasets, architectures group naturally by inductive bias: fully connected models, which impose minimal structural assumptions, achieve the highest accuracy and the most reliable UQ, but show the characteristic long-term error growth associated with low-bias models. Latent-evolution models – though less accurate – exhibit reduced error accumulation under iterative rollouts.

Conclusions. Our results underscore the importance of systematic optimization and comprehensive architectural comparison to make trade-offs explicit. The datasets, architectures, metrics, and benchmarking procedure are publicly bundled in CODES to support representative and reproducible comparisons.

Key words. astrochemistry – methods: numerical – ISM: abundances – evolution – ISM: molecules

1. Introduction

Nonequilibrium chemistry plays a fundamental role in regulating a wide range of astrophysical processes, from the early Universe (Galli & Palla 1998; Glover & Abel 2008), to protoplanetary disks (Caselli & Ceccarelli 2012), the evolution of galaxies (Pallottini et al. 2017; Lupi 2019), and the star formation process (Decataldo et al. 2019; Kim et al. 2018). In particular, accurately solving the chemical evolution in numerical simulations is crucial for modeling realistic thermodynamics. Directly solving the chemical evolution avoids the need for approximate treatments, such as precomputed analytical cooling functions, which may fail in dynamically evolving environments (Bovino et al. 2019).

On the one hand, several codes implement extensive photochemical networks and achieve high physical fidelity, but at significant computational cost. Examples include CLOUDY (Ferland et al. 2017), UCLCHEM (Holdship et al. 2017), and MAIHEM (Gray et al. 2019). This type of approach (see Olsen et al. 2018 for a review) is typically used to post-process astrophysical simulations to obtain emission lines (e.g.,

Vallini et al. 2018). On-the-fly coupling of such detailed chemistry is therefore commonly limited to one-dimensional applications, such as photochemical evolution during shock processes (e.g., Danehkar et al. 2022).

On the other hand, some software packages are designed for coupling to full three-dimensional hydrodynamic simulations. ASTROCHEM (Kumar & Fisher 2013) and KROME (Grassi et al. 2014) generate solver code for user-defined chemical networks, while NIRVANA (Ziegler 2016) and GRACKLE (Smith et al. 2017) provide subroutines to solve selected, astrophysically motivated networks. More recently, CARBOX (Vermariën et al. 2025b), the first fully differentiable astrochemical code written entirely in JAX (Bradbury et al. 2021), was introduced. It resembles previous approaches but exposes the full computational graph (reaction rates, Jacobians, and state updates), enabling direct access to model internals and making fast, gradient-based inference on external parameters feasible.

However, solving the full system of ordinary differential equations (ODEs) describing chemical networks on the fly remains one of the dominant computational bottlenecks in hydrodynamical simulations. This occurs for several reasons: (i) astrochemical networks contain reactions with rate

* Corresponding author: sy192@uni-heidelberg.de

coefficients that span many orders of magnitude. Consequently, the associated chemical relaxation times can be much shorter than, comparable to, or longer than the local hydrodynamical or advection time depending on density, temperature, and irradiation. Thus, while some channels drive very fast transients (often far faster than the concurrent fluid evolution), equilibrium cannot be assumed globally or at all times; (ii) the governing ODEs are stiff and strongly coupled (Branca & Pallottini 2023; Nejad 2005), which necessitates costly implicit integrators; (iii) the number of reactions grows super-linearly with the number of chemical species, inflating the required floating-point operations; and (iv) in massively parallel simulations, load imbalance and communication or synchronization overheads reduce effective throughput.

Recent advances in deep learning have enabled the development of surrogate models that aim to mitigate these bottlenecks by replacing numerical solvers with computationally cheaper, typically data-driven architectures. A variety of surrogate approaches have been proposed, including latent ODEs with autoencoders (Grassi et al. 2022; Holdship et al. 2021), physics-informed neural networks (Branca & Pallottini 2023), neural ODEs with autoencoders (Sulzer & Buck 2023), neural fields (Asensio Ramos et al. 2024), parametric neural ODEs (Vermar- ien et al. 2025a), and neural operators (Branca & Pallottini 2024; van de Bor et al. 2025; Ono & Sugimura 2026).

Compared to numerical solvers, neural surrogates are often more efficient because they transform the computational workload into forms that map to modern accelerator hardware well. Inference in neural networks is dominated by dense linear algebra operations, which are highly parallel and can be executed efficiently on GPUs, often at reduced numerical precision. Depending on the architecture, surrogate inference additionally has a fixed computational cost per evaluation, in contrast to the adaptive behavior of ODE solvers. This predictability simplifies load balancing and improves throughput in large-scale simulations. However, these efficiency gains come at the cost of approximation error. It may be for this reason that up to now surrogate models have not been widely adopted in current state-of-the-art astrophysical simulations. Developing surrogate models for this application context is challenging; chemical solvers are invoked at every hydrodynamic time step, which means that even small inaccuracies in surrogate predictions can lead to significant errors through numerical instability and error accumulation.

To avoid this error build-up, surrogate models generally require a high level of accuracy, but there are additional factors to be taken into account when determining the optimal surrogate architecture and configuration for a task; a good average accuracy may not be sufficient if the surrogate occasionally produces catastrophic errors. Errors may be more consequential for some predictions than for others, the prediction interval must often be flexible due to adaptive hydrodynamic time stepping, and a reliable uncertainty quantification (UQ) mechanism is desirable to enable fallback to the numerical solver when predictions are untrustworthy. Given these (and possibly additional) desirables, we argue that advancing from proof-of-concept models to simulation-ready tools requires a paradigm shift in how surrogate models for nonequilibrium chemistry are developed and evaluated. To facilitate this shift, we developed the CODES benchmark (Janssen et al. 2024).

The remainder of this paper is structured as follows. In Sect. 2, we describe the astrochemical datasets used for training and evaluation. Section 3 introduces the surrogate architectures and their variants, while Sect. 4 details the benchmarking,

optimization, and evaluation framework. The results are presented in Sect. 5, followed by a discussion in Sect. 6 and conclusions in Sect. 7.

2. Chemistry datasets

The synthetic data used for training, validation and testing of the surrogate models was generated with KROME (Grassi et al. 2014), a Python preprocessor that produces Fortran code to solve a given chemical network. The system of ODEs that describes the time evolution of the chemical species is

$$\dot{n}_k = \sum_{i,j} A_k^{ij} n_i n_j + \sum_i B_k^i n_i, \quad (1)$$

where the coefficients $A_k^{ij} = A_k^{ij}(T, \mathbf{n})$ and $B_k^i = B_k^i(\mathbf{F})^1$ denote net reaction contributions to species k . In particular, while the underlying rate coefficients are non-negative, A_k^{ij} and B_k^i include the appropriate stoichiometric numbers (with their signs) so that they account for both the production and destruction of n_k . Here T is the gas temperature, $\mathbf{n}$ is the vector of species abundances, and $\mathbf{F}$ denotes any ionizing or dissociation flux. The system is closed by simultaneously evolving the gas temperature (T):

$$\dot{T} = \frac{(\gamma - 1)}{k_b \sum_i n_i} (\Gamma - \Lambda), \quad (2)$$

where k_b is the Boltzmann constant, γ is the gas adiabatic index, $\Gamma = \Gamma(T, \mathbf{n}, \mathbf{F})$ and $\Lambda = \Lambda(T, \mathbf{n}, \mathbf{F})$ are the heating and cooling functions, respectively². To solve Eqs. (1) and (2), KROME uses a backward differentiation formula solver of DLSODES (Hindmarsh 2019), an implicit multistep solver that exploits the sparsity of the Jacobian matrix constructed with the chemical fluxes³.

To demonstrate the performance of our models, we constructed four datasets with distinct characteristics: (i) primordial, (ii) primordial parametric, (iii) cloud, and (iv) cloud parametric. In this context, parametric refers to datasets where the sampling varies not only over the initial conditions but also over certain external parameters. This distinction is important because these parameters constitute additional, structurally distinct inputs that may require architectural modifications (see Sect. 3). Specifically, we varied two physical parameters: the radiation field intensity (G) and the metallicity (Z). For the primordial and primordial parametric datasets, we considered 9 chemical species: e^- , H^- , H , H^+ , He , He^+ , He^{++} , H_2 , and H_2^+ . These species evolve through 46 reactions⁴, including H_2 formation on dust grains (Jura 1975; Wakelam et al. 2017), photo-chemistry, and cosmic-ray ionization. For the cloud and cloud parametric datasets, we extended the primordial network with heavier atoms and molecules: C^- , O^- , C , Si , O , OH , CO , CH , CH_2 , C_2 , HCO , H_2O , O_2 , C^+ , Si^+ , O^+ , HOC^+ , HCO^+ , H_3^+ , CH^+ , CH_2^+ , CO^+ , CH_3^+ ,

¹ Although the coefficients B_k^i can generally depend on gas temperature, they do not for the reaction networks considered here.

² Equations (1) and (2) may, in principle, also depend on additional external parameters, such as the metallicity (Z) or the rate of molecule formation on dust grains; however, for clarity of notation, these dependencies have been omitted.

³ We adopt the default errors for KROME, i.e., relative and absolute tolerances are fixed at 10^{-6} and 10^{-20} respectively.

⁴ The reaction rates are taken from Bovino et al. (2016): reactions (1)–(31), (53), (54), and from (58)–(61) in their Tables B.1 and B.2, photo-reactions P1 to P9 in their Table 2.

OH^+ , H_2O^+ , H_3O^+ , O_2^+ , and Si^{2+} , resulting in a network of 37 species and 287 reactions⁵.

In the case of the nonparametric datasets, the ionizing radiation intensity is fixed at $G = G_0$, where $G_0 = 1.6 \times 10^{-3} \text{ erg cm}^{-2} \text{ s}^{-1}$, and the metallicity is set to the solar value, $Z = Z_\odot$. For the parametric datasets, the radiation intensity varies over the range $G \in [0.1 G_0, 10 G_0]$, and the metallicity spans the interval $Z \in [10^{-3} Z_\odot, Z_\odot]$ ⁶. In all datasets, we assume a cosmic-ray ionization rate $\zeta_{\text{cr}} = 3 \times 10^{-17} \text{ s}^{-1}$ and a dust to gas ratio in solar units $f_d = 0.3$.

A further critical aspect in constructing representative datasets for training surrogate models is the choice of how to sample the initial parameter space. For surrogate training, the primary objective is to obtain a space-filling design that broadly covers the admissible domain for abundances and potential additional parameters, rather than to reproduce an assumed astrophysical prior distribution. The dimensionality of this domain ranges from ten parameters in the primordial case to 39 in the cloud parametric case. We therefore employed Sobol sampling (Sobol 1967), a deterministic quasi-random sampling strategy that provides near-uniform coverage of bounded parameter domains. Compared to standard random sampling, Sobol sequences reduce clustering and uncovered regions at fixed sample size, improving representation of rare parameter combinations while remaining reproducible. Sobol sampling is a global, non-adaptive strategy and therefore does not focus additional samples on regions where the surrogate model exhibits larger prediction errors—addressing such regions would require an active learning or adaptive sampling strategy.

After drawing a point in parameter space, we construct the initial abundances $\mathbf{n}_0$ by enforcing a helium mass fraction consistent with a 3:7 He-to-(H+He) ratio, and we compute the electron abundance a posteriori to ensure charge neutrality. In the cloud parametric case, all metal-bearing species are initialized by scaling a reference abundance pattern linearly with the sampled metallicity Z (i.e. $n_X \propto Z$ for heavy elements), while H/He species follow the helium constraint above. The thermochemical evolution is then computed for 10 kyr using KROME. The resulting trajectories are sampled at 100 logarithmically spaced time points starting just below 0.1 yr to resolve the stiff early-time dynamics. The ranges of initial conditions and parameters are summarized in Table 1.

3. Surrogate architectures

The surrogate architectures investigated in this work are implemented within our benchmarking framework CODES, which is described further in Sect. 4. To ensure fair comparison, reproducibility, and extensibility, all architectures follow a mandatory code structure. For example, all surrogates implement a common data preprocessing and training interface and expose a unified prediction method. Design choices were further guided by the

Table 1. Parameter ranges used to generate the datasets in this work.

Quantity	Variable	Min	Max
Gas density	$\log(n/\text{cm}^{-3})$	−2	3.5
Abundance (H and He)	$\log(n_i/n)$	−6	0
Abundance (metals)	$\log(n_i/n)$	−20	−6
Temperature	$\log(T/\text{K})$	$\log(20)$	5.5
Radiation	$\log(G/G_0)$	−15	−5
Metallicity	$\log(Z/Z_\odot)$	−3	0
Time	$\log(t/\text{yr})$	−1	4

Notes. All datasets vary the initial gas density (n), species abundances (n_i/n), and temperature (T), while parametric datasets additionally vary the radiation intensity (G) and metallicity (Z) (see Sect. 2). The datasets contain 2048 (primordial), 4096 (primordial parametric and cloud), and 8192 (cloud parametric) initial conditions, each evaluated at 100 time steps.

target application: replacing the numerical solver used to evolve chemical abundances across a hydrodynamical time step. Since the hydrodynamical time step can vary during a simulation, all surrogates are trained to predict the system state at user-specified times within the temporal range covered by the training data, given an initial condition. During training, trajectories are sampled at fixed, logarithmically spaced time points, so that the time variable takes only discrete values in the training set; however, time is treated as a continuous input by the architectures, allowing predictions at intermediate times within the trained interval.

CODES currently includes four surrogate architectures:

- FullyConnected (FCNN). A standard fully connected neural network. Initial abundances and the desired output time are concatenated and provided jointly as input to the model.
- MultiONet (MON). An adaptation of DeepONet (Lu et al. 2021) for multiple outputs as proposed in Lu et al. (2022). The inputs to the branch network are the initial abundances, while the trunk network receives the desired output time.
- LatentNeuralODE (LNODE). Following Sulzer & Buck (2023), an autoencoder is combined with a latent space neural ODE (Chen et al. 2018). The encoder receives the initial abundances as input. In contrast to FCNN and MON, this model features an explicit time evolution mechanism, hence the time input is used as the endpoint of the numerical integration, not as an explicit network input.
- LatentPoly (LP). An autoencoder with a learnable latent space polynomial, also following Sulzer & Buck (2023). Similarly to LNODE, the initial state is input to the encoder network, and the time input controls the evaluation point of the learnable polynomial.

Schematics of all architectures are displayed in Fig. 1. Generally, the surrogate architectures investigated in this work are not autoregressive: the desired output time t is provided explicitly, and the models learn to predict the system state at arbitrary times for a given initial condition. This design enables flexible prediction intervals required for simulations with adaptive hydrodynamical time stepping. All architectures accept a continuous time input t , allowing predictions at arbitrary times within the training horizon (10 kyr in this work).

As mentioned in Sect. 2, the ODE governing the evolution of a chemical species may depend not only on the abundances of other species, but also on additional physical quantities. One important such quantity is the temperature (T), which coevolves with chemical abundances. Due to this explicit time-dependence,

⁵ The reactions rate are taken from Glover et al. (2010), Glover & Mac Low (2011), Kingdon & Ferland (1996), O’Connor et al. (2015), Glover & Savin (2009) and KIDA (Wakelam et al. 2012), OSU (Prasad & Huntress 1980), and UMIST (Millar 1991) databases.

⁶ Because the chemical networks differ between the datasets, the interpretation of “metallicity” and the corresponding cooling treatment also differs. In the primordial case, metallicity denotes the total abundance of all elements heavier than helium and is implemented following the prescription of Shen et al. (2013). In the cloud case, by contrast, metallicity is applied by rescaling the relative abundances of heavy atoms and molecules with respect to hydrogen and helium.

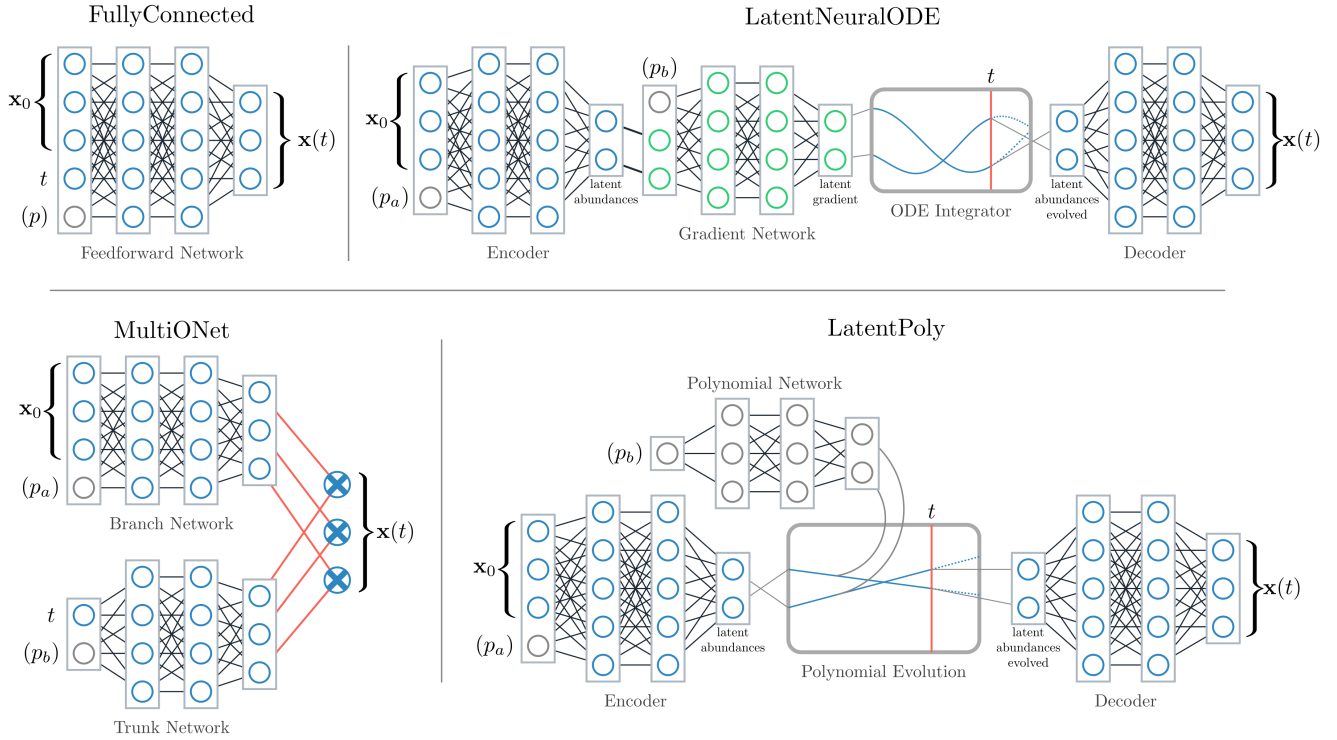


Fig. 1. Schematic overview of the four surrogate architectures investigated in this work. *Left*: fully connected surrogates (FCNN, MON). *Right*: latent-evolution surrogates (LNODE, LP). All models take the initial state (x_0), the desired output time (t), and, for parametric datasets, additional physical parameters (p), and output the predicted state ($x(t)$). Optional components are shown in gray, and parameter-handling options (p_a and p_b) denote alternative ways of incorporating the parameters.

T is treated as an additional species in the chemical reaction network. Other physical quantities, such as the radiation field intensity (G) and the metallicity (Z), are updated independently of the chemistry and therefore remain constant over a hydrodynamical time step. In principle, these quantities could also be treated as additional species that stay constant across one set of trajectories. In this view, we simply concatenate the parameters to the abundances and hand the combined vector to the architecture as input. However, since these parameters are structurally and physically distinct from chemical abundances, we propose architectural variations that better reflect this difference for all architectures except FCNN⁷.

For MON, the parameters can be handed to either the branch or the trunk network. The latter approach is in line with the original motivation to have two separate networks, namely the fact that the inputs to the branch and trunk network have different domains. For LNODE, the alternative to handing parameters to the encoder network (concatenated with the abundances) is to hand them to the gradient network (concatenated to the encoded latent-space abundances). The intuitive reading of this would be that while the abundances should be encoded to evolve on a lower-dimensional latent manifold, the parameters might be more aptly viewed as influencing the gradients in this latent space, rather than changing the projection into this space. Finally, for LP, we propose as alternative to the joint encoding of abundances and parameters that the parameters should influence

the learnable polynomial coefficients directly. This is achieved by introducing an additional fully connected neural network (the polynomial network), which predicts the polynomial coefficients based on the parameter values. These architectural variations are treated as optional hyperparameters and are only used when selected during hyperparameter tuning.

4. Methodology: Benchmarking and optimization framework

Initially developed to compare surrogate models on a given dataset, the CODES (Coupled ODE Surrogates) benchmark (Janssen et al. 2024) has evolved into a general framework for optimizing and evaluating surrogate architectures for coupled ODE systems⁸. Optimization is achieved through extensive hyperparameter tuning, while evaluation consists of training multiple instances of each optimized architecture on varying subsets of the dataset and assessing their performance across multiple dimensions. The result of this evaluation step is a collection of metrics and diagnostic plots aimed at both quantifying surrogate performance and comparing architectures. These metrics and plots extend beyond standard measures of accuracy to address the above-mentioned desirables, yielding detailed insights into the computational demand, error distribution, generalisation capabilities, and the reliability of UQ.

CODES is implemented in PYTORCH (Paszke et al. 2019) and builds on established tools including OPTUNA (Akiba et al. 2019) for hyperparameter optimization (HPO), SCHEDULEFREE

⁷ The motivation to include FCNN as a surrogate architecture is to simply hand all available information to the model without any additional inductive bias. Due to this, it did not seem apt to derive an architectural variation for the parameters for this model. The variations for the other architectures are structurally well motivated, but for FCNN, we feel that this would not be the case.

⁸ In terms of data structure, the only assumption made in CODES is that trajectories evolve with respect to an independent variable, but some architectures' inductive bias is aimed specifically at ODE data.

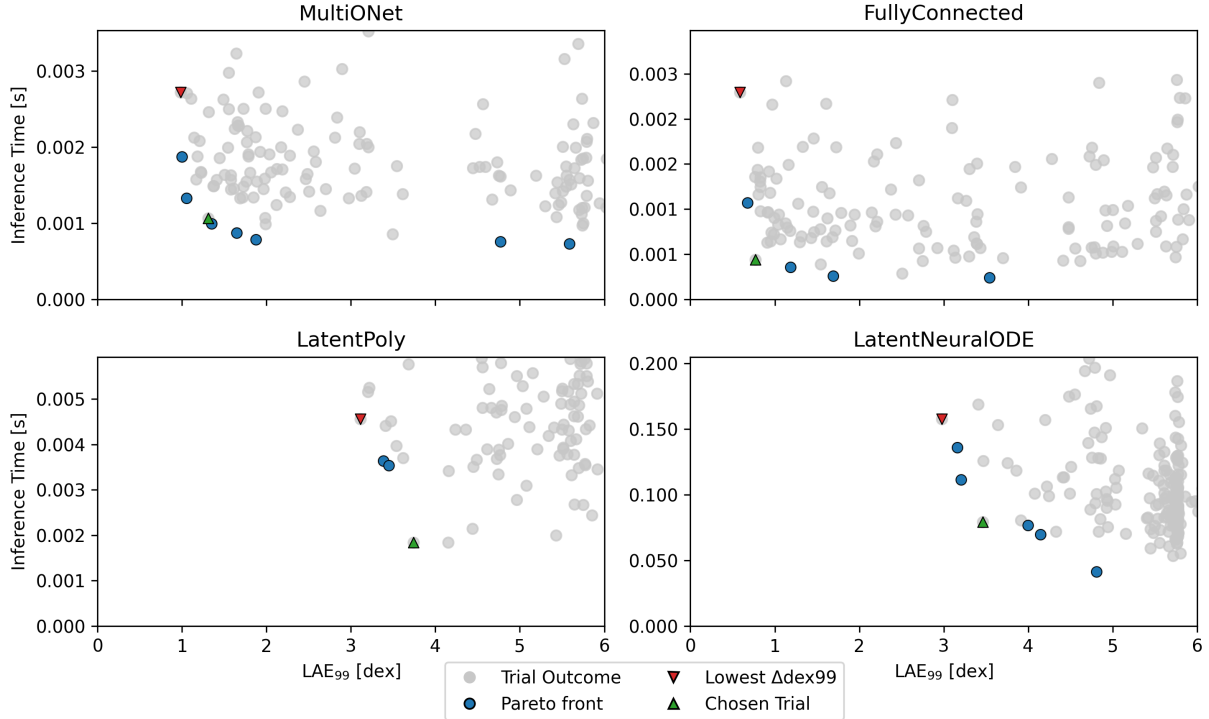


Fig. 2. Pareto fronts obtained from dual-objective hyperparameter tuning for the primordial dataset. The colored points indicate Pareto-optimal configurations, while the gray points are dominated solutions. The red marker denotes the lowest-error configuration, and the green marker indicates the selected trade-off configuration balancing accuracy and inference time.

(Defazio et al. 2024) for learning-rate scheduling, and the TORCHODE (Lienen & Günnemann 2022) package for differentiable GPU-based ODE solvers. The code is available on GitHub⁹. CODES follows standard software best practices, including automated testing, documentation, and reproducibility via controlled seeding and stored configuration files. In the following, we introduce the three-step process of hyperparameter tuning, training and evaluation, and elaborate on metrics and UQ.

4.1. Multi-objective hyperparameter tuning

Hyperparameter optimization is a necessary first step in the benchmarking pipeline, as meaningful comparisons require well-performing configurations and the performance of neural networks strongly depends on the chosen hyperparameters. Compared to its initial formulation in Janssen et al. (2024), CODES was extended to replace the assumption of known optimal hyperparameters with an explicit HPO pipeline. To clarify terminology, a trial denotes a single training run within HPO. In each trial, the model is trained on the training set using the selected hyperparameters, and its performance is subsequently evaluated on the validation set. A study corresponds to the optimization of one surrogate architecture on one dataset and thus comprises multiple trials. A tuning refers to the joint optimization of all architectures on a given dataset. In this work, each tuning consisted of four studies, hence across the four datasets considered, a total of sixteen studies were performed. Given the high computational cost of the hyperparameter tuning runs, we rely on OPTUNA to efficiently explore the hyperparameter space. CODES leverages OPTUNA’s adaptive sampling strategies, early stopping of unpromising trials, and parallel execution to reduce computational overhead while maintaining robust optimization

performance. The specific samplers and pruning strategies used in this work are described in Appendix A.1.

An optimization procedure requires the definition of an objective function. For the results presented in Sect. 5, we performed dual-objective optimization, simultaneously targeting prediction accuracy and computational efficiency. Accuracy is quantified using the LAE₉₉, while efficiency is measured via inference time, defined as the wall-clock duration required for a complete forward pass over all samples¹⁰. The choice and interpretation of these metrics are discussed in detail in Sect. 4.3. Each trial is thus associated with two objective values, LAE₉₉ and inference time. To handle this dual-objective setting, CODES employs OPTUNA’s NSGA-II evolutionary algorithm (Deb et al. 2002), which is designed to identify a set of Pareto-optimal solutions. A configuration is Pareto-optimal if no other trial achieves strictly better performance in both accuracy and efficiency simultaneously, and the collection of such configurations forms a Pareto front in the two-dimensional objective space. NSGA-II is well suited for this setting, as it directly supports multi-objective optimization in heterogeneous search spaces and promotes broad exploration.

Examples of the resulting Pareto fronts for the primordial dataset are shown in Fig. 2. From each Pareto front, we manually selected a configuration near the knee of the curve, balancing accuracy and efficiency¹¹. The number of trials per study is set heuristically to $15 \times N_h$, where N_h (ranging from 11 to 17)

¹⁰ Inference time is measured on the validation set during HPO and on the test set in the final evaluation.

¹¹ In principle, this step could be automated, but in practice such automation offers little benefit. Manual selection is more robust given the stochastic nature of the tuning. Further details are given in Appendix A.

⁹ <https://github.com/AstroAI-Lab/CODES-Benchmark>.

denotes the number of tuned hyperparameters. The hyperparameters optimized this way fall into two categories: architecture-specific and shared hyperparameters. The architecture-specific parameters differ between models and include latent dimensionalities, hidden-layer widths, or choices governing how external parameters are incorporated. The shared hyperparameters govern aspects of the training procedure and are common across architectures. The resulting search space is heterogeneous and includes real-valued, categorical, and conditionally activated parameters. We refer to Appendix A for further details on the tuning procedure and a full list of optimized hyperparameters.

4.2. Training and evaluation protocol

Following HPO, all surrogate architectures were trained under standardized conditions using the optimized hyperparameters to enable fair comparison across datasets. For most evaluations, it is sufficient to train one instance of the surrogate architecture on the full training dataset (the main model). For additional characterization, CODES provides five modalities, each of which can be toggled via a central configuration file and results in training extra models on modified datasets:

- Interpolation. Tests a surrogate’s ability to interpolate between training time steps by omitting intermediate points.
- Extrapolation. Probes how well a model generalizes beyond the training time range by truncating trajectories.
- Sparsity. Assesses robustness when the number of training samples is reduced.
- Batch size. Examines the effect of training with different batch sizes.
- Uncertainty quantification. Estimates predictive uncertainty with deep ensembles (DEs; [Lakshminarayanan et al. 2017](#)).

All modalities except UQ support setting multiple parameter settings¹², enabling the construction of scaling curves that show how performance changes as interpolation gaps widen, trajectories are truncated earlier, training data grows sparse, or batch size decreases. These modalities serve as diagnostic tools for probing surrogate robustness under controlled perturbations of the training setup. In this work, we focused on UQ, as it is most directly relevant to assessing surrogate reliability in simulation settings.

Upon completion of the training step, architectures are evaluated on the test set. The evaluation procedure reloads the trained weights and computes a core set of quantitative performance metrics for the main model and all additional models trained under the enabled modalities. Depending on the configuration, further diagnostics can be enabled, including loss trajectories, correlations between prediction errors and true gradients, inference times, and the computational footprint of each architecture (trainable parameters and memory usage). The results of these evaluations form the basis for the comparative plots and tables shown in Sect. 5.

4.3. Accuracy metrics

Many standard accuracy metrics are unsuitable to assess model performance in the context of astrochemistry, where chemical abundances spread over many orders of magnitude (in our datasets up to 30 dex). On such datasets, absolute error metrics such as the mean absolute error or the mean squared error do not account for errors on scarce quantities. In principle, relative error metrics such as the mean relative error mitigate this

problem by dividing the absolute error by the actual magnitude of the quantity in question. However, when aggregating relative error distributions, a small number of disproportionately large errors can strongly skew summary statistics, especially the mean. Furthermore, relative errors are not symmetric – significantly overestimating an abundance produces large relative errors, whereas significantly underestimating it produces at most a relative error of 100% – and division by very small values can be numerically unstable.

These problems with relative errors could be mitigated by introducing a relative error threshold value, which would serve as a stabilizing factor in the division, preventing explosive error values and numerical instabilities. This, however, implicitly introduces a weighting of abundances, as errors below the chosen threshold are effectively discounted. While weighting errors is a viable approach, this weighting should be motivated by the actual importance of the abundance to the evolution of the system, which is hard to quantify and presumably not captured adequately by a blanket threshold value. In the present work, all species and time steps are treated equally when aggregating error statistics, although CODES may provide custom weighting schemes in future extensions.

For these reasons, we propose to focus on log-space metrics when measuring accuracy of surrogate models in astrochemistry (i.e., across HPO, training and evaluation). All models in this work are trained on log-transformed abundances, hence it is natural to compute aggregate values, losses and optimization objectives in log-space as well. Concretely, we report the models’ log-space mean absolute error (mLAE) as well as the LAE₉₉ in Table 2, both computed on the log₁₀-transformed abundances. As mentioned in Sect. 4.1, the latter metric was also used as optimization objective during HPO¹³. We deliberately employed a percentile-based tail metric rather than the maximum error. Optimizing directly for low maximum errors directs computational effort toward eliminating rare failure cases rather than achieving good performance in the majority of cases. Percentile-based metrics accept rare failures while still capturing overall predictive performance, which is consistent with the use of a UQ-enabled solver fallback. The choice of the 99th percentile is pragmatic – ideally, this threshold should be motivated by the requirements of the target application.

4.4. Uncertainty quantification

Even with a highly representative dataset, a surrogate model will still make significant errors for some fraction of its predictions. A well-calibrated UQ mechanism serves to detect these unreliable predictions and can enable a fallback to the numerical solver for such cases. To investigate UQ and its potential for error mitigation, CODES optionally trains a DE ([Lakshminarayanan et al. 2017](#)) with a configurable number of models. DE rests on the idea that introducing some stochasticity¹⁴ into the training of multiple, architecturally identical models will cause them to converge to different, yet equally representative points on the loss landscape. The ensemble’s prediction is obtained as the mean across ensemble member predictions, while the standard deviation of these predictions can be used as a measure of uncertainty. In the

¹³ Due to the usage of log₁₀, the metrics indicate errors in orders of magnitude and hence have the unit “dex”.

¹⁴ In CODES, this stochasticity is implemented by assigning each model a different seed for the random number generators, which influences the weight initialization and dataset shuffling during training.

¹² Deep ensembles require multiple models by definition, hence there is only a single setting for UQ: the number of models in the DE.

Table 2. Performance metrics across datasets and models.

Dataset	Arch	mLAE	LAE ₉₉	mRE	Inference time	Params	mLAE _{iter}	mLU _{DE}	mLAE _{DE}	UQ
		(dex)	(dex)	(%)	(μ s)	(k)	(dex)	(dex)	(dex)	PCC
Primordial	MON	0.343	2.47	38.5	703 $\pm$ 77	498	0.499	0.225	0.174	0.531
	FCNN	0.094	0.693	11.8	307 $\pm$ 38	896	0.273	0.154	0.101	0.581
	LNODE	0.661	3.04	80.6	18 190 $\pm$ 960	1418	1.26	0.251	0.686	0.267
	LP	0.843	3.35	90.8	619 $\pm$ 78	461	1.24	0.301	0.833	0.263
Primordial Parametric	MON	0.613	5.96	41.6	2030 $\pm$ 90	86	0.667	0.431	0.251	0.695
	FCNN	0.259	1.74	33.4	642 $\pm$ 42	454	0.719	0.172	0.121	0.512
	LNODE	0.693	3.32	81.7	41 520 $\pm$ 2360	409	1.58	0.821	0.783	0.278
	LP	0.743	3.43	85.4	1040 $\pm$ 120	27	1.60	0.488	0.805	0.425
Cloud	MON	0.034	0.326	4.1	13 010 $\pm$ 10 910	459	0.187	0.031	0.024	0.654
	FCNN	0.035	0.384	3.5	814 $\pm$ 98	24	0.246	0.111	0.054	0.796
	LNODE	1.92	4.55	100.0	43 430 $\pm$ 610	286	1.92	0.008	1.91	0.122
	LP	1.64	4.70	99.9	1910 $\pm$ 190	193	1.78	0.306	1.62	0.023
Cloud Parametric	MON	0.121	0.831	16.3	8470 $\pm$ 4090	372	0.938	0.077	0.066	0.559
	FCNN	0.244	2.80	15.0	1510 $\pm$ 130	361	0.441	0.144	0.083	0.82
	LNODE	1.51	4.80	99.7	49 070 $\pm$ 520	234	1.85	0.197	1.49	-0.017
	LP	1.37	4.51	99.7	1820 $\pm$ 120	17	2.04	0.679	1.40	0.051

following, we refer to this standard deviation as the predicted uncertainty.

For the evaluations in this work, the ensembles always consist of $M = 5$ models. Employing a DE comes at additional computational cost, because predictions require M forward passes rather than one. Most surrogates investigated in this work are computationally lightweight, hence this overhead remains minor. Nevertheless, the additional expenditure must be justified by performance gains. The DE could deliver two main benefits: improving accuracy over the single-model setting, and providing a reliable identification mechanism for significant errors. In Sect. 5.3, we investigate whether these benefits are achieved in practice.

5. Results

In the following, we present the results of applying CODES to the four datasets primordial, primordial parametric, cloud, and cloud parametric, described in Sect. 2. We begin with an analysis of the hyperparameter tuning runs in Sect. 5.1. The benchmark results following HPO are then examined from several perspectives: prediction accuracy (Sect. 5.2), UQ in application-relevant settings (Sect. 5.3), and error propagation under iterative prompting (Sect. 5.4). Key metrics are shown in Table 2, although this is only an excerpt of the full list of metrics generated by CODES.

5.1. Hyperparameter tuning

Following the order of the benchmarking pipeline, we begin by evaluating the dual-objective optimization procedure. The goal of this evaluation is not to analyze individual hyperparameters, but to investigate whether the hyperparameter tuning procedure proposed in Sect. 4.1 yields meaningful accuracy–efficiency trade-offs and how these compare to accuracy-only optimization.

For each dataset and architecture, the optimization using the NSGA-II sampler yields a well-defined Pareto front. While the

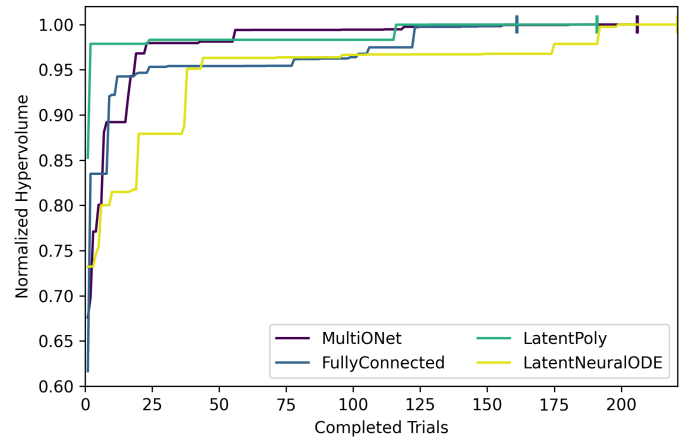


Fig. 3. Evolution of the normalized hypervolume spanned by the Pareto front during HPO for the primordial dataset. Vertical dashes indicate the final trial of each architecture-specific study. Most gains occur early in the optimization, followed by gradual saturation, suggesting diminishing returns from additional trials.

stochastic nature of the sampler does not allow us to make definitive statements about the state of convergence, we can use the development of the area covered by the Pareto front over the course of the tuning as a post-hoc heuristic to judge the progress of the optimization procedure. These areas were computed with respect to a reference point 10% worse than the worst values observed across all trials in both objectives. Figure 3 shows the evolution of this area, normalized to the final Pareto-front area, over the course of the studies performed for the primordial dataset. Corresponding plots for the other datasets show similar behavior (see Appendix A.3). We observe a saturation effect toward the end of the run, with most gains made early in the optimization procedure. This does not prove convergence, but indicates that substantial further gains are unlikely if the optimization were extended.

The resulting Pareto fronts exhibit substantial, architecture-dependent variability in both accuracy and efficiency, typically spanning factors of two to five with occasional extremes around an order of magnitude for both LAE₉₉ and inference time. The FCNN architecture displays the broadest accuracy spread, reflecting a strong sensitivity to hyperparameter choice, whereas the latent-evolution surrogates (LP, LNODE) show comparatively flat accuracy fronts but several-fold differences in runtime.

Figure 2 exemplifies the resulting Pareto fronts for the primordial dataset, highlighting both the selected configuration (green) and the configuration with the lowest LAE₉₉ (red). These plots qualitatively illustrate that selecting the most accurate configuration without regard for efficiency often yields only a modest error reduction at the cost of a substantial increase in inference time. A quantitative analysis confirms this: in five out of the sixteen studies, the Pareto-selected configuration coincides with the lowest-error configuration. In the remaining eleven cases, prioritizing accuracy alone yields only modest error reductions at substantially increased inference cost. Averaged across these studies, LAE₉₉ decreases by ~15% at the expense of a ~170% increase in inference time. Examined on a per-study basis, this trade-off is highly asymmetric: reducing the error by 1% typically requires an ~20% increase in inference time, with costs ranging from a few percent to more than 75% depending on dataset and architecture. These results demonstrate that systematic, dual-objective tuning exposes a broad accuracy – efficiency trade-off surface that would remain hidden under single-objective or manual adjustment. Even within the limited search budget used here, the optimizer consistently identifies configurations offering substantial efficiency gains for minimal accuracy loss, highlighting the practical value of explicit multi-objective optimization in surrogate selection.

5.2. Surrogate performance

An overview of selected performance metrics is provided in Table 2. Metrics directly related to prediction accuracy include the mLAE, the LAE₉₉, and the median relative error (mRE)¹⁵. For completeness, additional diagnostics on worst-case errors and temperature-only accuracy are reported in Appendix B. Computational demand is quantified by the inference time, measured as the mean and standard deviation over five measurements of the time required to process the full test set on an NVIDIA TITAN Xp¹⁶. These values should be interpreted only as a relative measure of computational demand, as absolute inference times depend strongly on hardware and on the size of the test set.

Overall, FCNN dominates these performance metrics. It consistently exhibits the lowest inference time and lowest mRE across all datasets, often by a large margin, and achieves the lowest mLAE and LAE₉₉ on both primordial datasets. It is also nearly on par with MON on the cloud dataset, with cloud parametric being the only case where MON clearly outperforms it in mLAE and LAE₉₉. While the low inference time of FCNN is expected given its simple structure and efficient GPU execution, its consistently superior accuracy is noteworthy.

More generally, we observe a clear divide between two classes of architectures: the fully connected models (FCNN, MON) outperform the latent-evolution models (LNODE, LP) across all datasets, often by a wide margin. The former class consists of plain feed-forward networks that differ mainly in how they handle structural differences between inputs, and make few assumptions about the underlying data-generating process (low inductive bias). In contrast, the latter class features autoencoder-like bottlenecks and differs primarily in the choice of latent evolution mechanism, thereby imposing explicit assumptions about the dimensionality and structure of the latent dynamics (high inductive bias). The divide between these two classes is more pronounced on the cloud dataset variants than on the primordial dataset variants. Interestingly, the fully connected architectures generally perform better on the cloud variants than on the primordial variants across all three accuracy metrics¹⁷, whereas the opposite trend is observed for the latent-evolution architectures. This suggests that the encoder–decoder structure or the assumed latent evolution may be ill suited to the cloud datasets. Key differences between the datasets that may contribute to this discrepancy are the larger number of species (38 versus 10), smaller average gradients, and the wider dynamic range of abundances in the cloud datasets.

A similar pattern emerges when comparing parametric and nonparametric variants: fully connected architectures achieve markedly higher accuracy on nonparametric datasets, whereas latent-evolution architectures perform similarly on both variants, occasionally even favoring the parametric case. The error distributions shown in Fig. 4 complement the scalar metrics in Table 2 by illustrating variability in model predictions across chemical states. They largely mirror the ranking observed in the summary statistics: fully connected architectures exhibit lower typical errors and tighter distributions, while latent-evolution architectures show broader spreads extending to higher error ranges, consistent with their larger LAE₉₉ values.

With respect to the second optimization objective, inference time, we observe clear trends, albeit not as clearly grouped by architecture type. Within each dataset, FCNN is consistently the fastest model, whereas LNODE is by far the slowest, incurring nearly two orders of magnitude higher inference time in extreme cases. This reflects the high computational demand of the numerical integration required to evolve latent trajectories. In contrast, LP consistently outperforms MON, ranking second across datasets, indicating that an encoder–decoder structure does not necessarily incur high runtime overhead when the latent evolution is expressed in closed form rather than via numerical integration.

Interestingly, computational efficiency shows no direct correspondence to model size. For example, FCNN contains up to an order of magnitude more parameters than LP in some cases, yet remains faster by a factor of two to three. Even the architecturally much more similar MON is slower than FCNN at lower parameter count. This suggests that increased computational cost arises from non-standard operations – the scalar product in MON, polynomial evaluation in LP, and, most prominently, numerical integration in LNODE – whose forward passes cannot fully exploit highly optimized compute kernels.

5.3. Uncertainty quantification

We first investigate whether a DE achieves improved accuracy compared to a single model. This is assessed by comparing the

¹⁵ While we deem relative error metrics to be suboptimal, we still want to provide at least one relative error metric for reference and choose the median RE because it is less affected by a few strong outliers.

¹⁶ Inference is primarily compute-bound rather than memory-bound for the surrogate models considered here. Peak GPU memory usage remained below 431 MB across all experiments, making the NVIDIA TITAN Xp a convenient reference platform rather than a requirement.

¹⁷ Comparisons are made separately for parametric and nonparametric variants.

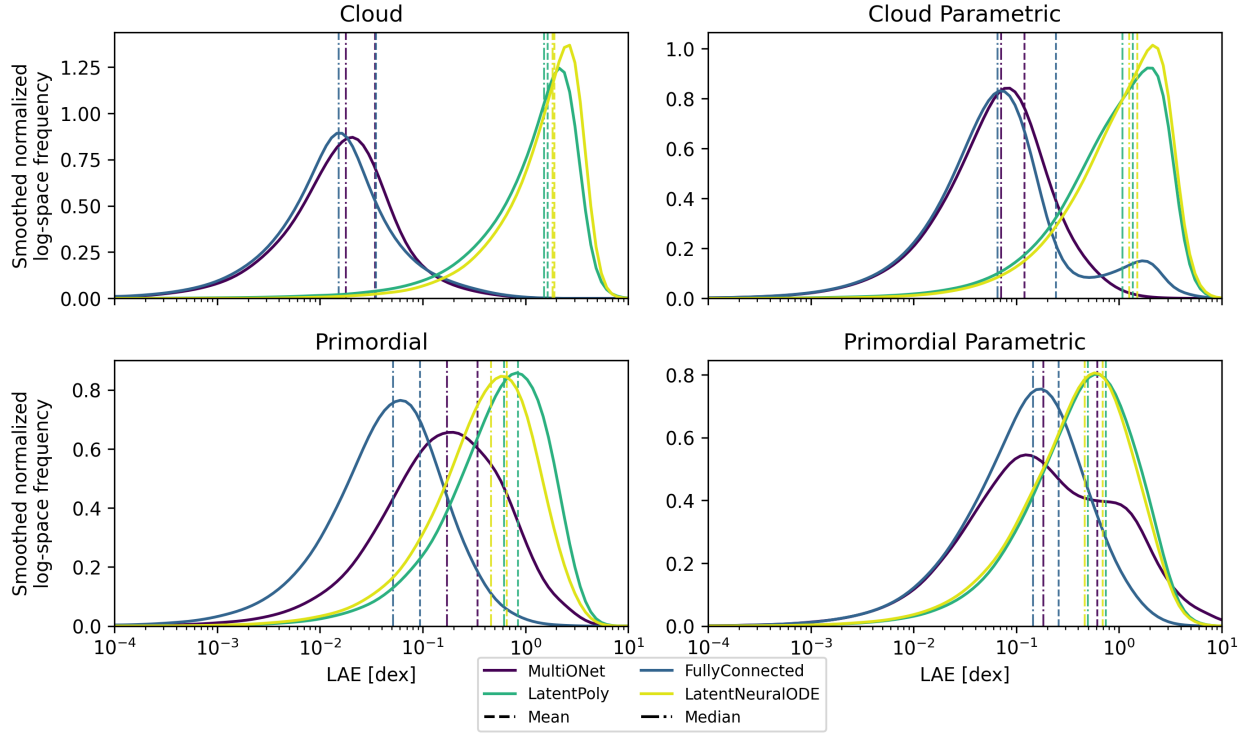


Fig. 4. Smoothed histograms of the LAE on the test set across surrogates and datasets, constructed in log-space and shown alongside the corresponding mean and median values. Latent-evolution surrogates exhibit systematically higher LAE values, with distribution peaks, means, and medians typically closely aligned. In contrast, for fully connected models the mLAE is frequently shifted toward higher values than the distribution peak, indicating a stronger influence of comparatively rare high-error predictions despite lower typical errors.

mLAE of the single model with mLAE (DE) in Table 2, which is the mLAE obtained using the ensemble mean as prediction, rather than the prediction of the main model. For clarity, we note that CODES forms the DE by training $M - 1$ additional models (with $M = 5$ in our case), incorporating the main model in the DE. This comparison reveals substantial improvements in mLAE over the single-model case for MON across all datasets, as well as for FCNN on the parametric dataset variants. On the nonparametric variants, the FCNN DE exhibits slightly higher mLAE than the single model. For the latent-evolution surrogates, only minor changes in mLAE are observed, without a consistent trend in either direction. One possible explanation is that models within latent-evolution ensembles tend to agree more closely with each other than those in fully connected ensembles. This is reflected in the mean log uncertainty (mLU), computed as the mean standard deviation of ensemble predictions across the test set. The mLU is of similar magnitude across surrogate classes, but it must be interpreted relative to the corresponding ensemble mLAE values. After all, the size of an error bar is only representative in relation to the magnitude of the value it refers to. Thus, on average, predictions within latent-evolution ensembles exhibit a much smaller spread relative to the true error magnitude.

To disentangle uncertainty calibration from absolute scale, we additionally report the UQ Pearson correlation coefficient (UQ PCC) between the ensemble standard deviation and the corresponding LAE across all test-set predictions, i.e., between predicted uncertainty and actual error. By nature, the PCC only measures linear correlation, remaining agnostic to the scaling of the quantities. The PCC differs systematically between surrogate classes: it exceeds 0.5 in all cases for fully connected models, often by a large margin, but remains below 0.5 for all latent-evolution models. The smallest correlations are observed

for the cloud dataset variants, where the correlation is close to 0 for the latter surrogate class. In summary, these metrics indicate that DEs based on fully connected models more accurately estimate the magnitude of their prediction error than those based on latent-evolution models.

Finally, we investigate whether DEs can realistically trigger a fallback to the numerical solver in cases of catastrophic prediction failure. For the purposes of this investigation, we define errors that exceed the 99th percentile as catastrophic errors. This means that for each surrogate-dataset pair, catastrophic errors are those above LAE_{99} in Table 2. Using this definition, Fig. 5 shows how reliably catastrophic errors can be identified based on the predicted uncertainty. Predictions are ordered by their mLU, and increasing fractions of predictions are flagged as unreliable (from most to least uncertain). For each such fraction, we report the corresponding recall (true positive rate) of catastrophic errors, i.e., the fraction of predictions with errors exceeding LAE_{99} that are correctly flagged. Plotting recall against the fraction of flagged predictions makes the trade-off between detection performance and fallback cost explicit, without relying on a specific uncertainty threshold.

In an application setting, the approach would be inverted: we would set a threshold value based on the desired recall or the affordable loss of efficiency. This efficiency loss is an inherent consequence of UQ-based fallback mechanisms, since the numerical solver may be orders of magnitude slower. In Fig. 5 we observe that the fully connected ensembles capture their prediction errors much more reliably than latent-evolution ensembles. While the former detect more than 80% of prediction errors at a threshold that flags less than 20% of predictions in all cases, the latter only reach this recall at much higher fractions. This aligns with the observations about the UQ PCC – fully connected

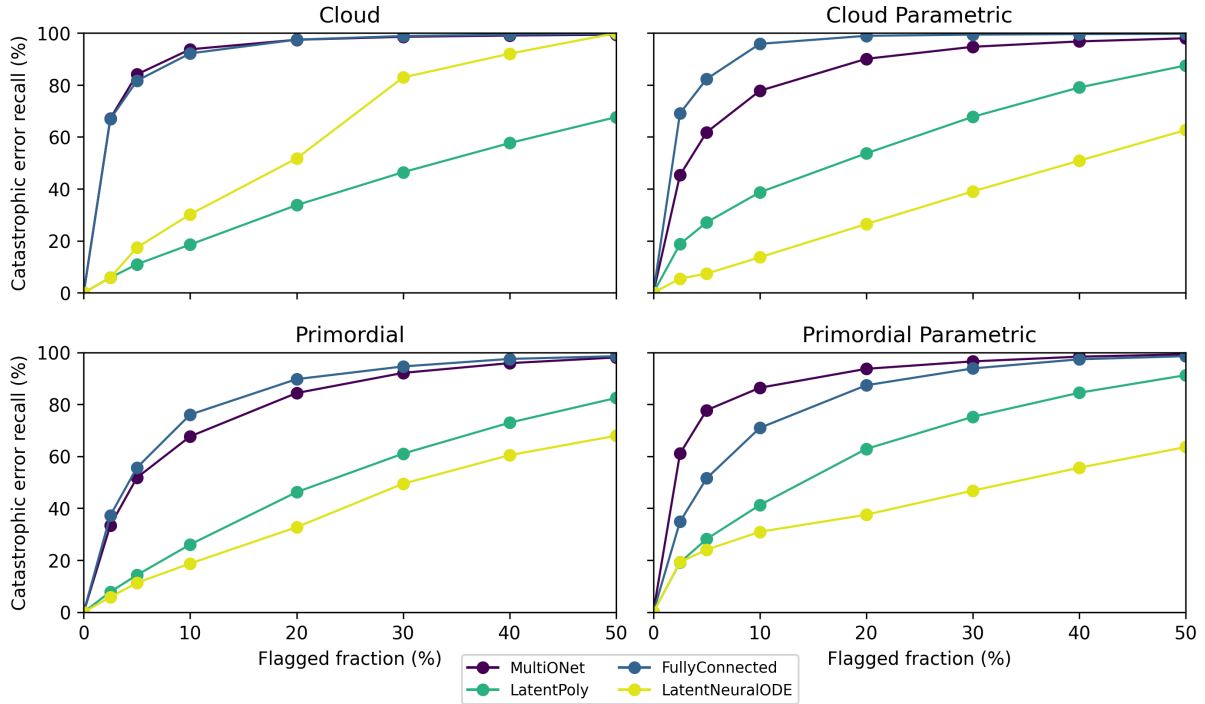


Fig. 5. Recall of catastrophic errors on the test set as a function of the fraction of flagged predictions. Predictions are ranked by the predicted uncertainty (mLU) estimated from a DE ($M = 5$), and increasing fractions of the most uncertain predictions are flagged. Catastrophic errors are defined as predictions whose log-space error exceeds the 99th percentile (LAE₉₉ in Table 2). Fully-connected surrogates achieve higher recall at lower flagged fractions than latent-evolution models, indicating more effective uncertainty-based error detection.

ensembles predict the magnitude of their prediction error more reliably, translating directly to higher catastrophic error detection efficiency. This behavior is robust to the precise definition of catastrophic error: repeating the analysis with a lower threshold yields qualitatively identical recall trends across architectures (see Appendix C).

5.4. Error propagation

Since surrogate models are intended to replace costly numerical solver calls at every hydrodynamical time step, their stability under repeated application is a key property. In a full simulation, abundances would be modified by hydrodynamics between surrogate calls. As a proxy, we instead reuse the surrogate’s prediction over one part of the time domain as the initial state for the next, following van de Bor et al. (2025). In CODES, the time domain is divided into segments of i time steps, and predictions are generated iteratively by using the final state of each segment as the initial condition for the next. Figure 6 compares the LAEs across datasets and architectures for $i = 10$ resulting from both iterative prediction and the standard one-shot prediction across the entire time domain (i.e., predicting all time steps directly from the initial state). Additionally, Table 2 reports the mean iterative log-space absolute error (mLAE_{iter}), corresponding to the mean of the trajectories shown in Fig. 6. In all cases, the mean iterative error exceeds the one-shot error, indicating that repeated surrogate application degrades accuracy. Additional tests show that the degree of error accumulation depends on the choice of i , with more frequent surrogate calls leading to larger accumulated errors (see Appendix D for $i = 3$).

The two architectural classes again exhibit distinct behavior in the iterative setting, with higher-inductive-bias models showing greater robustness. Fully connected surrogates, despite

their superior one-shot accuracy across most datasets, accumulate errors more rapidly under iterative application than latent-evolution architectures. This is most apparent on the cloud dataset, where the latter class of surrogates incurs almost no loss in accuracy when evaluated iteratively. Although latent-evolution models typically start with higher errors at early times, the effect is strong enough that MON, despite achieving the lowest initial error on the cloud parametric dataset, exhibits the highest error at the end of the time domain. These findings suggest that architectures with stronger inductive bias, while underperforming in standard accuracy metrics, capture more of the governing dynamics and are therefore more reliable when making predictions for previously unseen initial conditions.

6. Discussion

The results presented in Sect. 5 offer a coherent picture of how the proposed surrogate architectures behave when applied to realistic, challenging astrochemical datasets. We identify multiple key takeaways from these investigations.

Hyperparameter tuning is crucial. As evidenced by the substantial span of the Pareto fronts across both optimization objectives, chosen hyperparameters strongly influence surrogate performance across architectures. The interdependence of hyperparameters and the complexity of the search space render manual optimization impractical, necessitating systematic tuning. Most importantly, however, our findings in Sect. 5.1 emphasize the importance of making the desired surrogate properties explicit optimization objectives. The observed trade-off – where in some cases a fivefold increase in inference time yielded only minimal accuracy gains – would not be visible if efficiency were

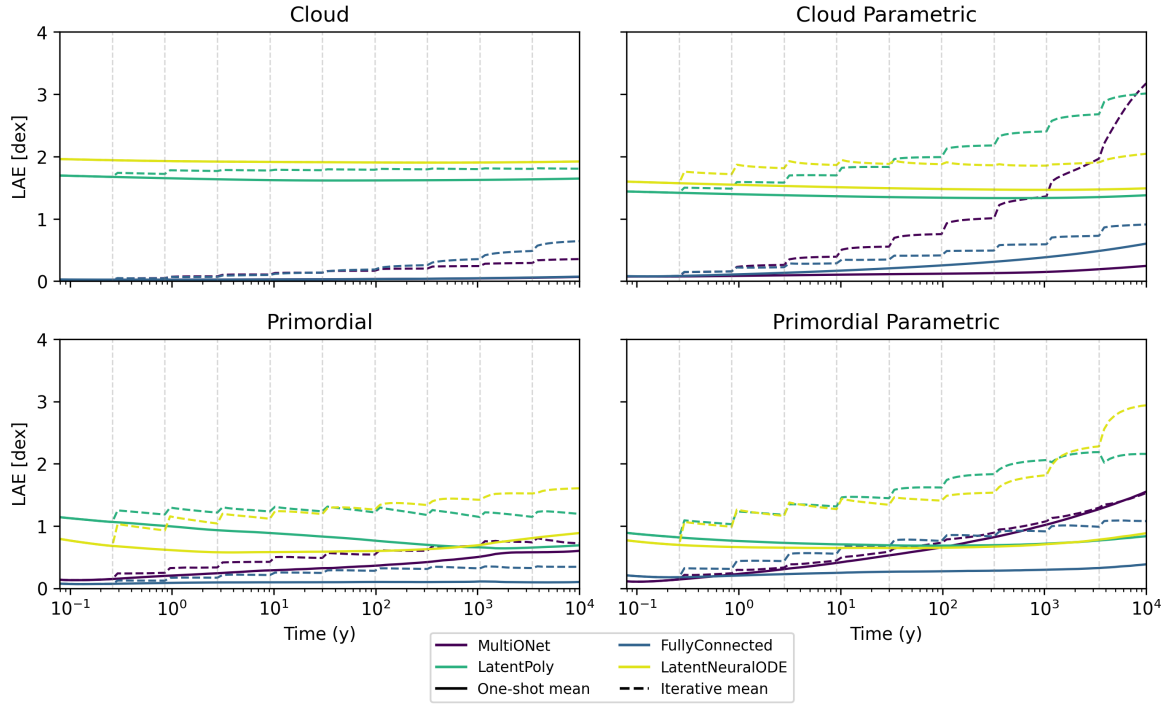


Fig. 6. Evolution of the mLAE on the test set over time across surrogates and datasets. To investigate error accumulation over time, we compared the one-shot mean, where predictions are obtained by providing the initial state and the desired time step, with the iterative mean. For the latter, we obtained predictions by dividing the domain into intervals of $i = 10$ time steps (indicated by the vertical dashed lines), and using the predicted abundances at the end of one interval as initial state for the next interval. In this iterative setting, fully connected surrogates generally exhibit a more pronounced accrual of error over time, while latent-evolution models prove more robust.

excluded from the optimization procedure. Moreover, this trade-off emerges using a sampling algorithm explicitly designed to explore the Pareto front and identify configurations that balance both objectives. A single-objective optimization using a corresponding algorithm (e.g., the Tree-Parzen Estimator) may optimize more aggressively toward accuracy, sacrificing even more efficiency in the process.

Inductive bias influences performance. Across Sect. 5, we observe pronounced differences between two classes of architectures: fully connected surrogates and latent-evolution surrogates. These groups differ systematically in the amount of inductive bias they introduce. The former make few assumptions about the underlying problem, whereas the latter are explicitly designed for trajectories arising from coupled ODEs. On accuracy alone, the fully connected architectures are superior in many cases, especially FCNN. While we do not claim a definitive explanation, we hypothesize that for the data considered here – complex coupled ODEs with additional external parameters – strong inductive biases, such as enforced low-dimensional latent trajectories, may restrict model flexibility more than they help regularize learning. This interpretation is consistent with observations reported by Grassi et al. (2022), where a latent-evolution model was found to effectively time-average latent dynamics, thereby attenuating high-frequency components associated with stiff behavior. For LNODE in particular, an additional contributing factor may be vanishing gradients, which pose a fundamental challenge for neural differential equations acting on stiff problems (Fronk & Petzold 2025). The picture is more nuanced for inference time, which largely depends on the computational cost of the non-standard operations involved. However, similar distinctions between the two groups emerge in UQ, where fully connected architectures are more reliable, and in iterative

evaluation, where latent-evolution models prove more robust. Both observations are consistent with differences in inductive bias: stronger bias reduces variability among ensemble members, degrading uncertainty estimates, while simultaneously increasing robustness under challenging conditions and limiting performance degradation under iterative predictions.

Deep ensembles provide reliable uncertainty estimates. The investigations in Sect. 5.3 show that DEs with $M = 5$ members provide a reliable mechanism for detecting catastrophic prediction errors. We observe pronounced differences in detection efficiency between architectures: latent-evolution models achieve comparable recall only when flagging a much larger fraction of predictions. The fully connected surrogates, however, detect catastrophic errors with high efficiency. This makes a confidence-triggered fallback mechanism to the numerical solver, which we expect to improve overall accuracy and significantly enhance trajectory stability, a realistic option.

Performance depends on representative data. Directly investigating trajectory stability via iterative surrogate calls in Sect. 5.4, we again find pronounced differences between architectures. The latent-evolution architectures prove more robust, whereas fully connected models exhibit pronounced error accrual. While strategies such as adding noise during training can mitigate error accumulation (Holdship et al. 2021), we conjecture that the dominant cause is initial-condition drift: repeated surrogate evaluations expose the model to states outside its training domain. A perfect surrogate that fully represents the underlying ODE would not be affected by this drift. However, for data-driven surrogates, leaving the training domain degrades accuracy, as evidenced by the jumps in mean error at the start of each iteration interval. This hypothesis is consistent

with the observed robustness of latent-evolution architectures – their higher inductive bias promotes learning the underlying mathematical structure rather than merely reproducing observed trajectories.

Regardless of the architecture of choice, this observation emphasizes the importance of training data being representative of the later application. For surrogate deployment in large-scale astrophysical simulations, this implies that the distribution of initial abundances in the training data should closely match that encountered in solver calls during the simulation. While this distribution is a priori unknown, a practical approach is to approximate it using small-scale simulations driven by the accurate but expensive numerical solver.

Limitations. In this work, we compare surrogates across multiple performance dimensions, four datasets, and four architectures. While CODES provides a rigorous framework for this comparison, ensuring that evaluations are fully balanced and representative remains challenging due to architectural differences and application-specific implementation details. One example is the measurement of inference time. We measure the time required to predict all trajectories in the test set using the same batched setup as during training, which favors architectures that benefit from batching. Additionally, for LNODE, we opt for evaluating entire trajectories at once, allowing a single numerical integration, rather than letting the solver solve anew for each time step. Whether these choices fully reflect realistic applications remains open.

A natural question is how these surrogate runtimes compare to those of full chemical solvers. We do not attempt such a comparison here for multiple reasons. First, establishing a fair baseline is non-trivial: astrochemical solvers vary widely in implementation details, may run on CPU or GPU, and performance is tightly coupled to integrator settings. Additionally, as described in Sect. 1, neural surrogates map naturally to modern GPU hardware, where dense tensor operations dominate performance. Their runtime characteristics therefore differ fundamentally from those of ODE solvers, making direct one-to-one comparisons difficult to interpret. Furthermore, the literature already provides multiple demonstrations that neural surrogates outperform traditional solvers by orders of magnitude. [Maes et al. \(2024\)](#) find a speed-up by a factor of 24–28, while [Branca & Pallottini \(2024\)](#) measure a factor of 128, and [Sulzer & Buck \(2023\)](#) observed a speed-up of factor 55 for their original implementation of LNODE and an impressive factor 4270 for the original LP.

While we performed thorough relative comparisons, assessing surrogate accuracy in absolute terms remains challenging, primarily because the impact of a given error on a downstream simulation is difficult to determine. In the future, integrating surrogates into fully differentiable magnetohydrodynamic frameworks such as *astronomix* ([Storcks 2025](#)) may enable new forms of sensitivity analysis, allowing errors to be traced back to individual species and predictions, which is not possible in a benchmark decoupled from the simulation. We also do not explicitly enforce mass or charge conservation, and leave conservation-aware training or renormalization strategies for future work. In addition to this, we did not compare our accuracy results to other works for two reasons: we use different metrics to measure accuracy (see Sect. 4.3), and we work on different datasets.

To facilitate future comparisons, our datasets are publicly available and CODES automates optimization, training, and evaluation on these and other datasets. Further work is needed to

quantify dataset difficulty, potentially via measures such as stiffness (as a proxy for dynamical complexity) or the hypervolume of the initial-condition space.

Another limitation concerns details of the HPO setup. While we consider our choices well motivated, aspects such as accuracy metrics, trial budgets, and optimizer selection remain largely heuristic. We identify the systematic exploration of these HPO design choices in the context of astrochemical surrogate modeling as a key opportunity for future research.

A limitation of our UQ analysis is that it relies on the test split of the same Sobol-sampled dataset used for training and validation. While the initial conditions in this split are unseen during training and optimization, they stem from the same underlying distribution. Although our results suggest that ensemble spread should increase for out-of-distribution inputs, further investigation is required to rule out failure modes such as mode collapse, where ensemble uncertainty may not reliably reflect prediction error.

Finally, we would like to emphasize that this work does not claim to present optimal or best-performing surrogates. While to our knowledge, the architectures presented here were optimized extensively for ideal performance on each dataset, we are aware that, as ever in machine learning, performance depends substantially on finding the optimal configuration of an array of training tools and architectural details. We do not expect substantial gains from retuning hyperparameters alone, as this is precisely what the HPO procedure addresses. But it is entirely possible that changes to architectures or training, for example adding a new loss term, altering the optimizer, or modifying regularization, could significantly improve performance.

Similarly, we do not claim that the architectures we present here are ideal. Our aim is to encourage architecture research by providing a systematic and rigorous approach to measuring and comparing performance across architectures. To facilitate this, we proposed challenging datasets that capture the complexities of the application context, a dual-objective optimization procedure to ensure surrogates perform as intended, log-space metrics to accurately evaluate surrogate performance, and additional investigations of surrogate behavior. These proposals are bundled in the open-source benchmarking suite CODES, which is well-documented and easily extensible.

Surrogates are viable for realistic astrochemical dynamics. Despite these limitations, the observed accuracy gains, substantial improvements in computational efficiency, and the additional safety provided by DE uncertainty estimation suggest that replacing numerical solvers with surrogates in complex astrophysical simulations is generally feasible. Our analysis indicates that, for the datasets considered here, fully connected models – most notably FCNN – are better suited to this task: they are more accurate, faster, and provide more reliable uncertainty estimates. It is, however, plausible that other datasets may benefit from the additional regularization and physical structure imposed by latent-evolution architectures. CODES facilitates informed, data-driven choices of surrogate architecture and configuration for a given task.

7. Conclusions

In this work, we proposed a systematic approach to identifying well-performing surrogate models for a given task, as represented by a dataset. This was achieved by separately optimizing four architectures for accuracy and inference time, followed by a

multidimensional comparison of their behavior. For this comparison, we introduced log-space accuracy metrics tailored to the large dynamic range characteristic of astrochemical data. The entire pipeline – tuning, training, and evaluation – is implemented in the CODES benchmarking framework.

We applied this procedure to four different datasets that represent realistic application scenarios, characterized by large dynamic range, relative stiffness, and additional challenges such as external ODE parameters. Our findings across surrogates and datasets are multifaceted: joint optimization for accuracy and inference time reveals an intrinsic trade-off between these objectives. We found that across accuracy, UQ, and error propagation measurements, the architectures fall into two groups. This grouping aligns closely with the degree of inductive bias incorporated by each architecture. For the datasets studied here, higher inductive bias – in the form of explicit latent-space time evolution – improves robustness in challenging settings but comes at the cost of reduced accuracy and less reliable DE uncertainty estimates.

A natural next step is to deploy the optimized surrogates within full-scale astrophysical simulations. This will reveal whether the performance characteristics observed in benchmark studies – accuracy, UQ with fallback mechanisms, and error propagation – translate to practical application settings. We expect this transition to a real application setting to introduce new challenges, while also enabling deeper insight into practical surrogate behavior. To support and encourage such efforts, CODES is made available as an open framework, and we welcome its use and extension by the community to explore additional datasets, architectures, and application scenarios.

Data availability

Our benchmarking framework CODES is publicly available on GitHub at <https://github.com/AstroAI-Lab/CODES-Benchmark>. A dedicated documentation website including a tutorial and auto-generated API documentation is provided at <https://astroai-lab.de/CODES-Benchmark>. The datasets used in this work are publicly available on Zenodo <https://doi.org/10.5281/zenodo.18018572>. Additional datasets included in CODES are also publicly available via Zenodo and are automatically downloaded on demand.

Acknowledgements. This work was funded by the Carl-Zeiss-Stiftung through the NEXUS program and supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1–390900948 (the Heidelberg STRUCTURES Excellence Cluster). The authors acknowledge support from the IWR Collaborate Program at Heidelberg University. We also acknowledge the usage of the AI clusters Tom and Jerry, funded by Field of Focus 2 of Heidelberg University. This work benefited from discussions at the Lorentz Center workshop “Machine Learning in Astrochemistry”. In particular, we thank Immanuel Sulzer, Simon Glover, Felix Priestley, and Tommaso Grassi for valuable exchanges.

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, 2623
- Asensio Ramos, A., Westendorp Plaza, C., Navarro-Almida, D., et al. 2024, *MNRAS*, **531**, 4930
- Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. 2011, in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2546
- Bovino, S., Grassi, T., Capelo, P. R., Schleicher, D. R. G., & Banerjee, R. 2016, *A&A*, **590**, A15
- Bovino, S., Schleicher, D. R. G., & Grassi, T. 2019, *Bol. Asoc. Argentina Astron. La Plata Argentina*, **61**, 274
- Bradbury, J., Frostig, R., Hawkins, P., et al. 2021, JAX: Autograd and XLA, Astrophysics Source Code Library [[record ascl:2111.002](https://ui.adsabs.org/abs/2021ASCL..2111..002)]
- Branca, L., & Pallottini, A. 2023, *MNRAS*, **518**, 5718
- Branca, L., & Pallottini, A. 2024, *A&A*, **684**, A203
- Caselli, P., & Ceccarelli, C. 2012, *A&A Rev.*, **20**, 56
- Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. 2018, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6572
- Danehar, A., Oey, M. S., & Gray, W. J. 2022, *ApJ*, **937**, 68
- Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. 2002, *IEEE Trans. Evol. Comput.*, **6**, 182
- Decataldo, D., Pallottini, A., Ferrara, A., Vallini, L., & Gallerani, S. 2019, *MNRAS*, **487**, 3377
- Defazio, A., Yang, X. A., Mehta, H., et al. 2024, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 37, 9974
- Ferland, G. J., Chatzikos, M., Guzmán, F., et al. 2017, *Rev. Mexicana Astron. Astrofis.*, **53**, 385
- Fronk, C., & Petzold, L. 2025, *Chaos*, **35**, 113113
- Galli, D., & Palla, F. 1998, *A&A*, **335**, 403
- Glover, S. C. O., & Abel, T. 2008, *MNRAS*, **388**, 1627
- Glover, S. C. O., & Mac Low, M. M. 2011, *MNRAS*, **412**, 337
- Glover, S. C. O., & Savin, D. W. 2009, *MNRAS*, **393**, 911
- Glover, S. C. O., Federrath, C., Mac Low, M. M., & Klessen, R. S. 2010, *MNRAS*, **404**, 2
- Grassi, T., Bovino, S., Schleicher, D. R. G., et al. 2014, *MNRAS*, **439**, 2386
- Grassi, T., Nauman, F., Ramsey, J. P., et al. 2022, *A&A*, **668**, A139
- Gray, W. J., Oey, M. S., Silich, S., & Scannapieco, E. 2019, *ApJ*, **887**, 161
- Hindmarsh, A. C. 2019, ODEPACK: Ordinary differential equation solver library, Astrophysics Source Code Library [[record ascl:1905.0211](https://ui.adsabs.org/abs/1905.0211)]
- Holdship, J., Viti, S., Jiménez-Serra, I., Makrymallis, A., & Priestley, F. 2017, *AJ*, **154**, 38
- Holdship, J., Viti, S., Haworth, T. J., & Ilee, J. D. 2021, *A&A*, **653**, A76
- Janssen, R., Sulzer, I., & Buck, T. 2024, arXiv e-prints [[arXiv:2410.20886](https://arxiv.org/abs/2410.20886)]
- Jura, M. 1975, *ApJ*, **197**, 575
- Kim, J.-G., Kim, W.-T., & Ostriker, E. C. 2018, *ApJ*, **859**, 68
- Kingdon, J. B., & Ferland, G. J. 1996, *ApJS*, **106**, 205
- Kumar, A., & Fisher, R. T. 2013, *MNRAS*, **431**, 455
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6405
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. 2018, *J. Mach. Learn. Res.*, **18**, 1
- Lienen, M., & Günnemann, S. 2022, arXiv e-prints [[arXiv:2210.12375](https://arxiv.org/abs/2210.12375)]
- Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. 2021, *Nat. Mach. Intell.*, **3**, 218
- Lu, L., Meng, X., Cai, S., et al. 2022, *Comput. Methods Appl. Mech. Eng.*, **393**, 114778
- Lupi, A. 2019, *MNRAS*, **484**, 1687
- Maes, S., De Ceuster, F., Van de Sande, M., & Decin, L. 2024, *ApJ*, **969**, 79
- Millar, T. J. 1991, *A&A*, **242**, 241
- Nejad, L. A. M. 2005, *Ap&SS*, **299**, 1
- O'Connor, A. P., Urbain, X., Stützel, J., et al. 2015, *ApJS*, **219**, 6
- Olsen, K. P., Pallottini, A., Wofford, A., et al. 2018, *Galaxies*, **6**, 100
- Ono, S., & Sugimura, K. 2026, *ApJ*, **996**, 9
- Pallottini, A., Ferrara, A., Bovino, S., et al. 2017, *MNRAS*, **471**, 4128
- Paszke, A., Gross, S., Massa, F., et al. 2019, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8026
- Prasad, S. S., & Huntress, Jr., W. T. 1980, *ApJS*, **43**, 1
- Shen, S., Madau, P., Guedes, J., et al. 2013, *ApJ*, **765**, 89
- Smith, B. D., Bryan, G. L., Glover, S. C. O., et al. 2017, *MNRAS*, **466**, 2217
- Sobol, I. M. 1967, *U.S.S.R. Comput. Math. Math. Phys.*, **7**, 86
- Storcks, L. 2025, astronomix – differentiable MHD in JAX, zenodo, <https://doi.org/10.5281/zenodo.17782162>
- Sulzer, I., & Buck, T. 2023, arXiv e-prints [[arXiv:2312.06015](https://arxiv.org/abs/2312.06015)]
- Vallini, L., Pallottini, A., Ferrara, A., et al. 2018, *MNRAS*, **473**, 271
- van de Bor, P., Brennan, J., Regan, J. A., & Mackey, J. 2025, *Open J. Astrophys.*, **8**, 96
- Vermariën, G., Bisbas, T. G., Viti, S., et al. 2025a, *Mach. Learn. Sci. Technol.*, **6**, 025069
- Vermariën, G., Grassi, T., Van de Sande, M., et al. 2025b, arXiv e-prints [[arXiv:2511.10558](https://arxiv.org/abs/2511.10558)]
- Wakelam, V., Herbst, E., Loison, J. C., et al. 2012, *ApJS*, **199**, 21
- Wakelam, V., Bron, E., Cazaux, S., et al. 2017, *Mol. Astrophys.*, **9**, 1
- Ziegler, U. 2016, *A&A*, **586**, A82

Appendix A: Hyperparameter optimization details and results

This appendix provides supplementary details on the HPO procedure used in CODES, including the optimization setup, the selection of Pareto-optimal configurations, and additional results and tables supporting Sects. 4.1 and 5.1.

A.1. Optimization setup and algorithms

Given the high computational cost of the tuning runs, CODES employs the OPTUNA framework (Akiba et al. 2019) to efficiently explore heterogeneous hyperparameter search spaces. In particular, CODES leverages OPTUNA’s adaptive sampling strategies, pruning of unpromising trials, and asynchronous parallel execution to reduce computational overhead while maintaining robust optimization performance.

For adaptive sampling, CODES uses Optuna samplers that propose new configurations based on outcomes of previous trials. For single-objective optimization, CODES employs the TPESampler, based on the Tree-Parzen estimator approach (Bergstra et al. 2011). For the dual-objective optimization used throughout this work, CODES employs NSGAIIISampler, implementing the NSGA-II genetic algorithm (Deb et al. 2002). We used a fixed population size of 50 for NSGA-II, while all other sampler parameters (including crossover and mutation settings) are left at OPTUNA’s default values. All reported Pareto fronts and selected configurations in this paper are obtained using NSGAIIISampler.

To reduce wasted compute on poor configurations, CODES uses OPTUNA pruners to terminate unpromising trials early. By default, CODES employs the HyperbandPruner (Li et al. 2018), which dynamically allocates computational resources across trials and can substantially reduce overall cost compared to running all trials to completion. However, OPTUNA’s built-in pruning mechanisms are only directly compatible with single-objective optimization. To nevertheless retain some of the benefits of early stopping in our dual-objective setup, we implemented a lightweight heuristic: during tuning, we tracked the running mean and standard deviation of trial durations and pruned trials exceeding the mean duration by more than one standard deviation after an initial warmup phase. This strategy is motivated by the observation that, within a given architecture and dataset, unusually long training times typically arise from configurations with increased model complexity or solver overhead, which also tend to incur higher inference cost. Pruning only the slowest trials therefore yields a disproportionate reduction in computational cost, as these configurations dominate the overall resource consumption by definition. We emphasize that this heuristic is not guaranteed to preserve all Pareto-optimal configurations, but serves as a lightweight mechanism to eliminate clearly inefficient trials in the absence of native multi-objective pruning support.

Finally, OPTUNA supports asynchronous parallel execution of trials, allowing multiple configurations to be evaluated concurrently. While this does not reduce total compute requirements, it substantially reduces wall-clock time.

The tuning budget for each study was set heuristically to $15 \times N_h$, where N_h denotes the number of tuned hyperparameters. Across studies, N_h ranges from 11 to 17 depending on the architecture and on whether the dataset includes additional physical parameters (see Sect. 3), resulting in 165 to 255 trials per study. To control cost, we used a trial length of 2000 epochs during the main tuning stage. After selecting Pareto-optimal configurations, we additionally performed a fine-tuning stage with 8000 epochs,

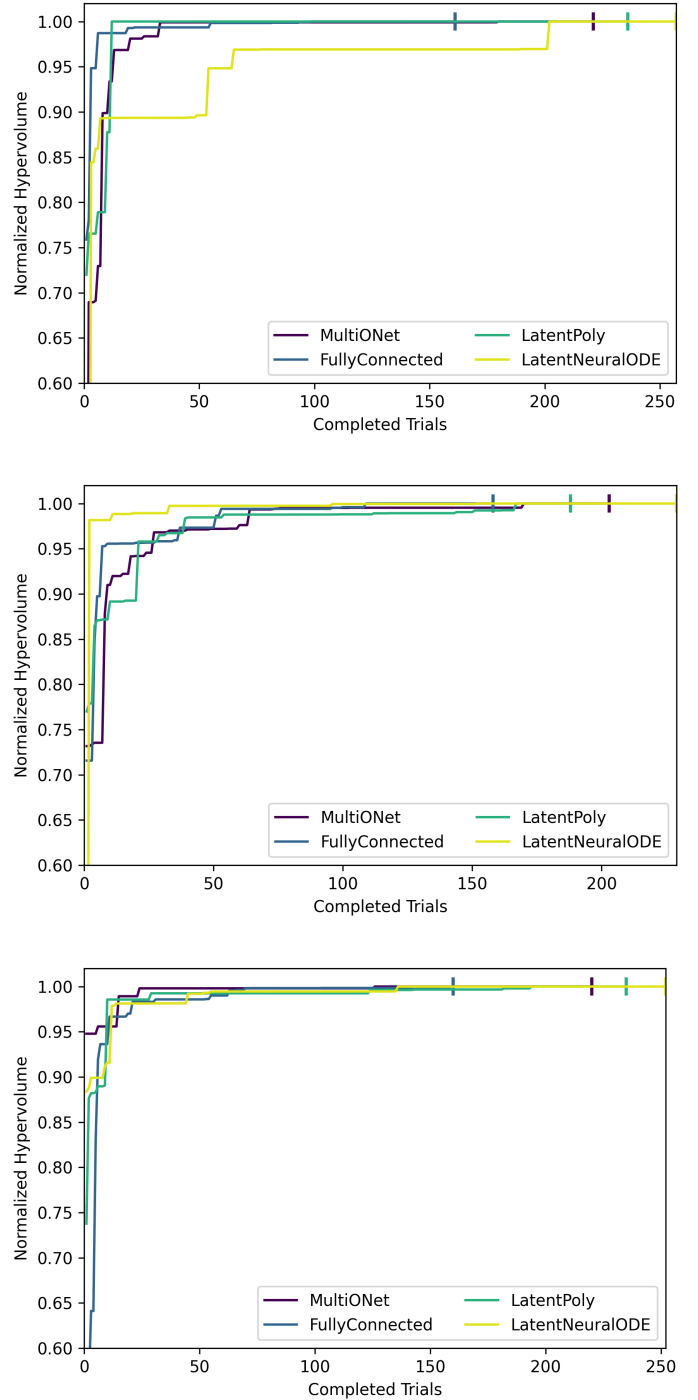


Fig. A.1. Normalized hypervolume evolution during HPO for the primordial parametric (top), cloud (middle), and cloud parametric (bottom) datasets.

restricting the search to hyperparameters that directly couple to training duration (learning rate, weight decay, and, where applicable, schedule or optimizer coefficients). This fine-tuning stage accounts for the fact that the computational budget of individual trials is typically lower than for the final training run, which can bias the optimal values of hyperparameters whose effect depends on training duration. The number of fine-tuning trials was set to $10 \times N_h$ (typically 30–40 trials per study).

A.2. Selection of Pareto-optimal configurations

For each dataset–architecture pair, dual-objective optimization yields a set of Pareto-optimal trials in the two-dimensional objective space spanned by LAE₉₉ and inference time (see Sect. 4.1). From each Pareto front, we selected one configuration for downstream training and evaluation. The selection was performed manually by choosing a configuration near the knee of the Pareto curve, prioritizing substantial efficiency gains while avoiding disproportionate losses in accuracy.

In principle, this step could be automated via knee-finding heuristics or scalarization schemes. In practice, such automation offers limited benefit in our setting for two reasons. First, the curvature and scaling of Pareto fronts differ between architectures and datasets, making a single, scale-invariant selection rule difficult to define. Second, Pareto fronts are obtained from stochastic optimization with finite budgets. Automated knee-finding can be sensitive to sampling noise and small perturbations of the front. We therefore adopted manual knee selection as a robust and transparent procedure that aligns with the practical surrogate-selection question of interest: selecting one of the most accurate configurations that does not sacrifice efficiency disproportionately.

Table A.1. Accuracy–efficiency trade-off between the lowest-error and Pareto-selected configurations (see Fig. 2).

Dataset	Arch	$\Delta\text{Error} \downarrow$ [%]	$\Delta\text{Time} \uparrow$ [%]	Ratio R
Primordial	MON	24.9	155	6.2
	FCNN	24.0	417	17.4
	LP	16.7	148	8.9
	LNODE	13.9	98	7.1
Primordial Parametric	MON	29.6	109	3.7
	FCNN	37.3	265	7.1
Cloud	LP	9.5	143	15.1
	LNODE	1.7	70	41.2
Cloud Parametric	MON	4.6	36	7.8
	LP	5.4	411	76.1
	LNODE	1.6	28	17.5
Averages		15.4	171	18.9

Notes. Architectures where both configurations coincide are omitted. Reported values indicate relative percentage changes of the accuracy-only configuration with respect to the Pareto-selected one. The final column reports the ratio $R = \Delta\text{Time}/\Delta\text{Error}$, i.e., the percentage increase in inference time required to achieve a 1% reduction in error.

A.3. Pareto-front area evolution across datasets

To provide a post-hoc indication of optimization progress within the allocated budgets, we monitor the evolution of the Pareto-front area (hypervolume) over the course of each study, computed with respect to a reference point slightly worse than the worst observed configuration in both objectives (see Sect. 5.1). Figure A.1 shows the normalized area evolution for the datasets primordial parametric, cloud and cloud parametric datasets, respectively.

Table A.2. Chosen hyperparameters for MultiONet.

Hyperparameter	Cloud	Cloud Parametric	Primordial	Primordial Parametric
Scheduler	poly	poly	schedulefree	poly
Optimizer	AdamW	AdamW	AdamW	AdamW
Loss	MSE	MSE	SmoothL1	MSE
Activation	Mish	ELU	LeakyReLU	PReLU
Hidden layers B	8	4	3	5
Hidden size	100	100	160	50
Output factor	50	40	130	74
Hidden layers T	1	4	2	1
Params branch	–	True	–	True
Poly power	0.647	8.20	–	2.57
Learning rate	3.7e-03	3.2e-03	2.7e-03	5.2e-03
Reg. factor	4.6e-05	6.5e-03	0.117	1.4e-03
Beta	–	–	0.732	–

Table A.3. Chosen hyperparameters for LatentNeuralODE.

Hyperparameter	Cloud	Cloud Parametric	Primordial	Primordial Parametric
Scheduler	poly	schedulefree	schedulefree	cosine
Optimizer	SGD	SGD	SGD	SGD
Loss	MSE	SmoothL1	MSE	MSE
Activation	SiLU	GELU	SiLU	SiLU
Coder layers	8	2	1	1
Coder width	80	180	580	360
Latent features	1	10	9	10
ODE layers	3	3	6	8
ODE width	250	220	480	220
ODE tanh reg	False	False	True	True
Encode params	–	False	–	False
Poly power	3.43	–	–	–
Momentum	0.143	0.075	0.274	0.331
Learning rate	3.9e-05	4.5e-03	5.9e-04	2.8e-06
Reg. factor	3.41	3.6e-04	9.2e-03	6.4e-05
Eta min	–	–	–	5.3e-04
Beta	–	14.4	–	–

Table A.4. Chosen hyperparameters for FullyConnected.

Hyperparameter	Cloud	Cloud Parametric	Primordial	Primordial Parametric
Scheduler	cosine	poly	poly	poly
Optimizer	AdamW	AdamW	AdamW	AdamW
Loss	SmoothL1	SmoothL1	SmoothL1	SmoothL1
Activation	PReLU	ELU	ELU	Mish
Hidden size	310	290	470	470
Hidden layers	1	5	5	3
Poly power	–	8.16	3.59	0.163
Eta min	2.7e-03	–	–	–
Beta	8.39	2.66	0.863	39.5
Learning rate	2.2e-03	0.0122	1.8e-03	1.3e-04
Reg. factor	3.4e-04	0.114	5.3e-05	0.0333

Table A.5. Chosen hyperparameters for LatentPoly.

Hyperparameter	Cloud	Cloud Parametric	Primordial	Primordial Parametric
Scheduler	poly	poly	schedulefree	poly
Optimizer	AdamW	AdamW	SGD	AdamW
Loss	MSE	SmoothL1	MSE	MSE
Activation	GELU	ReLU	ReLU	ELU
Coder layers	2	1	2	1
Coder width	290	170	470	700
Latent features	2	10	8	7
Degree	6	9	4	3
Coeff network	–	False	–	False
Poly power	5.18	2.46	–	0.264
Momentum	–	–	0.0161	–
Learning rate	9.3e-05	8.2e-03	2.4e-04	6.0e-06
Reg. factor	2.1e-04	0.0567	1.1e-03	0.0510

A.4. Accuracy–efficiency trade-off statistics

Table A.1 quantifies the accuracy–efficiency trade-off between the Pareto-selected and accuracy-only (lowest-error) configurations for all studies where these differ. It reports relative percentage changes in LAE₉₉ and inference time, as well as the ratio $R = \Delta\text{Time}/\Delta\text{Error}$, interpreted as the percentage increase in inference time required to achieve a 1% reduction in error.

A.5. Optimized hyperparameters

Tables A.2, A.3, A.4, and A.5 report the hyperparameter values selected from the Pareto fronts for each architecture and dataset. Hyperparameter names follow the notation used in the CODES configuration files. Further implementation details are provided in the CODES documentation¹⁸. The optimized hyperparameters fall into two categories: shared training hyperparameters, which are common across architectures and architecture-specific hyperparameters, which control structural choices such as layer widths, depths, latent dimensionality, or parameter handling. Entries marked as “–” indicate that a given hyperparameter is not applicable to the corresponding architecture.

Appendix B: Additional accuracy diagnostics

The additional diagnostics in Table B.1 provide context for the tail behavior of surrogate errors and for temperature-specific accuracy. Across all datasets and architectures, the maximum log-space absolute error (LAE₁₀₀) is substantially larger than LAE₉₉, typically exceeding it by factors of two to more than an order of magnitude. These values correspond to extreme outliers, with errors spanning several to tens of orders of magnitude, and confirm that worst-case failures are indeed catastrophic in absolute terms. While models with lower LAE₉₉ generally also exhibit lower LAE₁₀₀, this relationship is loose and non-monotonic, underscoring the statistical instability of the maximum error as a performance metric. The magnitude of LAE₁₀₀ appears to be primarily determined by the dataset rather than the architecture, with more complex datasets exhibiting higher worst-case errors, consistent with their larger dynamic ranges.

In contrast, the temperature-only error (Temp. LAE) is consistently small across all architectures and datasets, and typically far below the corresponding mLAE values. This reflects the comparatively limited dynamic range of temperature relative to chemical abundances and indicates that large chemical outliers do not necessarily translate into large temperature errors. From an application perspective, this is encouraging, as temperature evolution is often the primary quantity of interest in coupled hydrodynamical–chemical simulations.

Appendix C: Catastrophic error detection

In Sect. 5.3, catastrophic prediction errors are defined as those exceeding the LAE₉₉, and detection efficiency is evaluated relative to this threshold. To assess whether the observed detection behavior depends strongly on this specific definition, Fig. C.1 repeats the analysis using a less extreme threshold, defining catastrophic errors as those exceeding the 90th percentile (LAE₉₀).

The resulting recall curves closely resemble those obtained for LAE₉₉: fully connected ensembles consistently achieve high

Table B.1. Additional accuracy diagnostics.

Dataset	Arch	mLAE (dex)	LAE ₉₉ (dex)	LAE ₁₀₀ (dex)	Temp. LAE (dex)
Primordial	MON	0.343	2.47	6.24	0.126
	FCNN	0.094	0.693	3.47	0.030
	LNODE	0.661	3.04	6.88	0.203
	LP	0.843	3.35	7.07	0.27
Primordial Parametric	MON	0.613	5.96	18.5	0.126
	FCNN	0.259	1.74	14.5	0.060
	LNODE	0.693	3.32	14.8	0.149
	LP	0.743	3.43	14.2	0.17
Cloud	MON	0.034	0.326	6.36	0.008
	FCNN	0.035	0.384	4.33	0.005
	LNODE	1.92	4.55	8.64	0.993
	LP	1.64	4.70	8.36	0.497
Cloud Parametric	MON	0.121	0.831	9.43	0.017
	FCNN	0.244	2.80	16.9	0.019
	LNODE	1.51	4.80	14.7	0.253
	LP	1.37	4.51	13.5	0.274

Notes. In addition to mLAE and LAE₉₉, we report the maximum log-space absolute error (LAE₁₀₀) and the temperature-only absolute error (Temp. LAE). Best values per dataset are shown in bold.

recall at low fractions of flagged predictions, whereas latent-evolution ensembles require substantially higher flagged fractions to reach comparable recall. This demonstrates that the qualitative conclusions regarding uncertainty-based detection efficiency are robust to the precise choice of catastrophic-error threshold.

Appendix D: Iterative evaluation

In Sect. 5.4 we investigate error accumulation under iterative surrogate application using an interval length of $i = 10$ time steps. To assess the sensitivity of these findings to the choice of interval length, Fig. D.1 shows the same analysis performed with a shorter interval of $i = 3$ time steps, corresponding to more frequent reuse of surrogate predictions as initial conditions.

As expected, reducing the interval length leads to stronger error accumulation across all architectures, reflecting the increased exposure to prediction drift under more frequent iterative updates. Importantly, however, the qualitative trends observed in the main text remain unchanged: fully connected surrogates exhibit a more pronounced growth of error over time in relative terms, while latent-evolution architectures retain comparatively stable performance despite their higher initial errors. This confirms that the observed differences in iterative robustness are not specific to a particular choice of interval length, but reflect systematic architectural properties.

¹⁸ <https://astroai-lab.de/CODES-Benchmark/>.

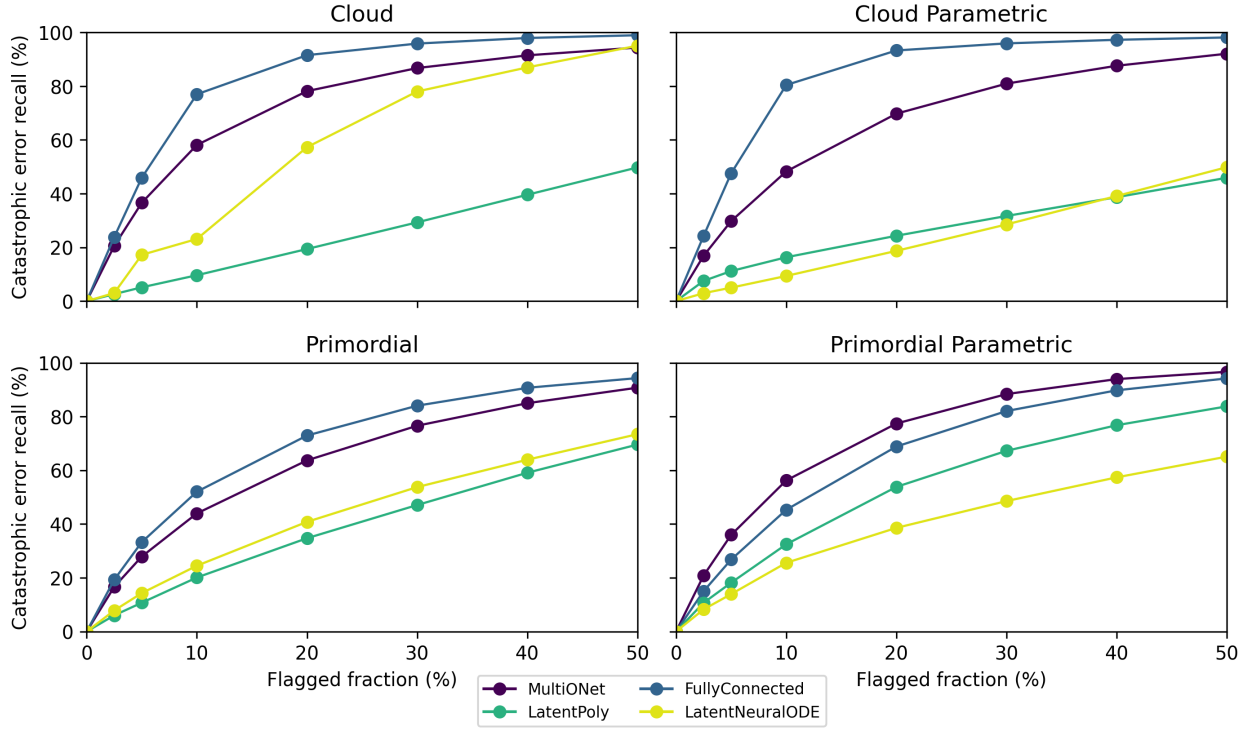


Fig. C.1. Same as Fig. 5 but defining catastrophic errors as those exceeding the 90th percentile log-space error, illustrating the robustness of the detection trends to the chosen threshold.

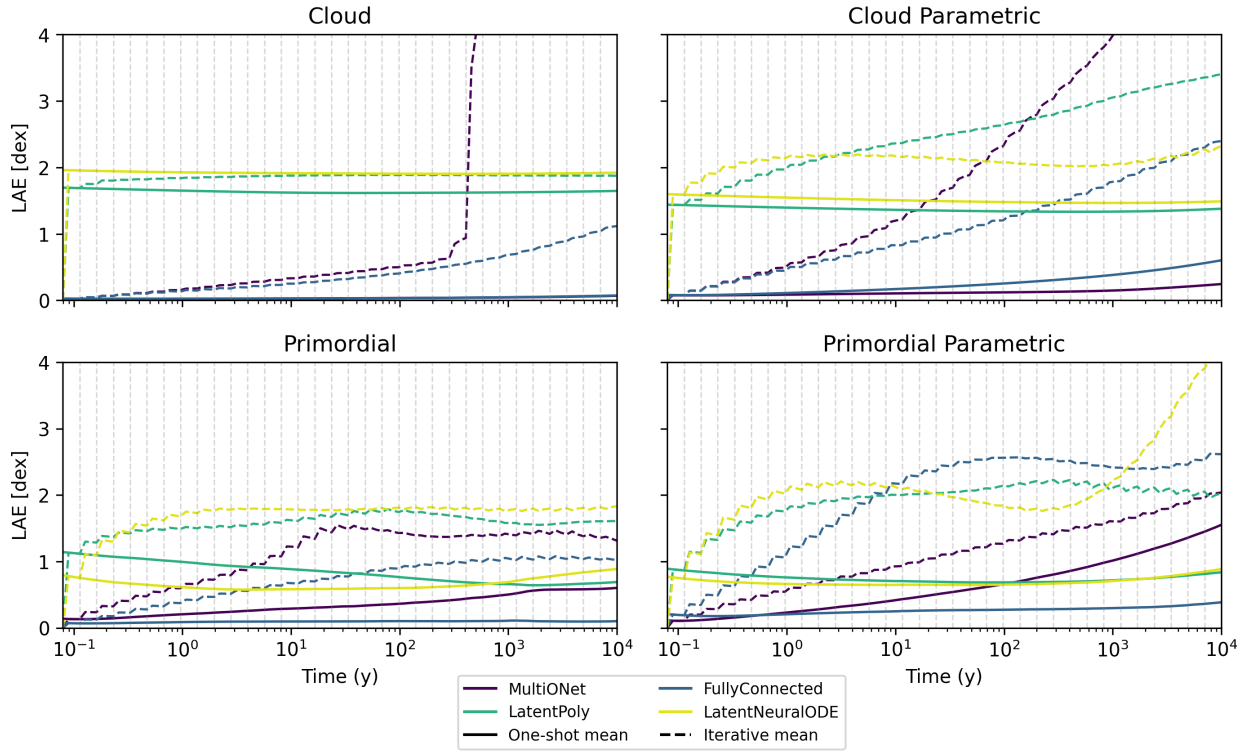


Fig. D.1. Same as Fig. 6 but for an interval length of $i = 3$ time steps, illustrating stronger sensitivity of error accumulation to more frequent iterative updates.