MadEvolve: Evolutionary Optimization of Cosmological Algorithms with Large Language Models

Shihui Zang
Department of Physics
University of Wisconsin–Madison
szang4@wisc.edu

Tianyi Li
Department of Physics
University of Wisconsin–Madison
tianyi.li@wisc.edu

Moritz Münchmeyer
Department of Physics
University of Wisconsin–Madison
muenchmeyer@wisc.edu

Abstract

We present MadEvolve, a framework for discovering scientific algorithms by integrating Large Language Models (LLMs) into evolutionary search. In this work, we implement a nested optimization scheme consisting of an outer evolutionary loop that explores algorithmic structures and an inner auto-differentiable loop that tunes continuous parameters via gradient descent. We apply MadEvolve to three problems in computational cosmology—baryon acoustic oscillation (BAO) reconstruction, 21cm foreground wedge reconstruction, and baryonic physics prediction from N-body simulations—achieving improvements of 22.8%, 30.7%, and 63%, respectively. The discovered algorithms remain interpretable and provide useful heuristics for future sky surveys.

1 Introduction

As modern sky surveys produce an unprecedented volume of data for cosmology [e.g. 1, 2], the associated data analysis pipelines are becoming increasingly complex, requiring substantial effort from researchers to keep designing and testing analysis algorithms for optimal usage of survey data.

Recently, the rapid development of LLM-based automated algorithm discovery has shown promising potential in algorithm search. Examples include FunSearch [3] and AlphaEvolve [4], which combine large language models with evolutionary search to discover algorithms for mathematical and engineering problems. Inspired by these approaches, we adapt this paradigm to computational cosmology, where LLMs act as intelligent mutation operators while physics-based evaluators guide the evolutionary search.

Contributions We introduce *MadEvolve*, a framework that combines LLM-driven evolutionary search with physics-based evaluation and automatic parameter optimization. Our approach features a nested optimization scheme that tunes parameters of LLM-generated algorithms via gradient-based methods, and automated report generation for interpreting the discovered algorithms. Specifically, we demonstrate MadEvolve on three computational cosmology tasks, achieving substantial improvements over baselines while maintaining interpretability.

Related Work Automated algorithm discovery has progressed from evolutionary methods like AutoML-Zero [5] and learned search systems such as AlphaTensor [6] and AlphaDev [7] to LLM-guided evolution, where language models serve as intelligent mutation operators: FunSearch [3] pioneered this for mathematics, AlphaEvolve [4] extended it to full codebases, and Eureka [8] applied

Preprint.

it to reward design. End-to-end "AI-Scientist" systems [9–11] and domain-specific AI agents for astrophysics [12] attempt fully automated scientific discovery. In the physical sciences, symbolic regression has been used to discover conservation laws [13] and rediscover physics equations [14].

2 The MadEvolve Framework

Evolutionary Loop MadEvolve treats scientific algorithms as programs iteratively improved by an LLM ensemble (Fig. 2). Each iteration: (1) selects a parent from a diverse population; (2) retrieves exemplar programs as inspiration; (3) constructs a prompt with code, metrics, and exemplars; (4) queries the LLM for code modifications; (5) optimizes continuous parameters via an inner loop before fitness evaluation; and (6) updates the program population. Diversity is maintained through a MAP-Elites quality-diversity grid [15], an island model with 5 subpopulations and periodic migration, and a global elite archive.

Nested Parameter Optimization When an LLM proposes a new algorithm with tunable parameters, evaluating it with arbitrary values may reject a potential good idea. MadEvolve automatically optimizes parameters before evaluation: for differentiable algorithms using JAX [16], gradients are computed via reverse-mode autodiff and optimized with Adam optimizer [17] (η=0.1, 30–50 iterations), enabling efficient tuning of 10+ parameters. Greedy search is used for non-differentiable algorithms. This ensures fitness comparisons reflect each algorithm's best performance.

LLM Ensemble Selection Seven frontier LLMs are managed via a UCB bandit [18]: $\mathrm{UCB}_i = (\bar{r}_i + c_t\sqrt{\ln N/n_i})\, w_i$, where $\bar{r}_i$ is model M_i's mean fitness improvement, n_i its query count, N total queries, and c_t a decaying exploration coefficient. This dynamically routes queries to the most productive model. MadEvolve also automatically generates readable reports comparing evolved algorithms against baselines.

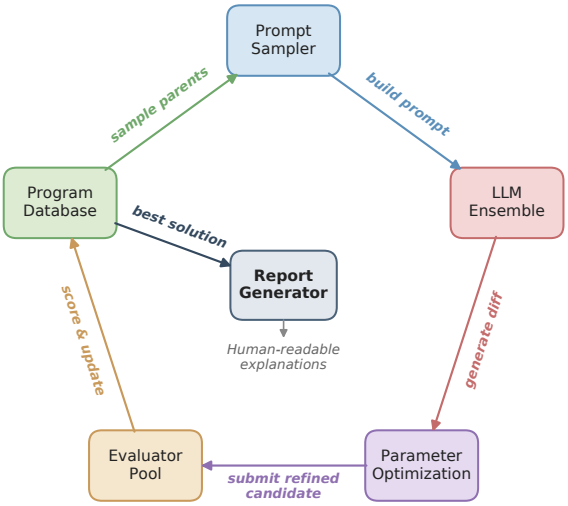

Figure 1: MadEvolve pipeline. The Prompt Sampler retrieves parent algorithms from the Program DataBase to query the LLM Ensemble. Generated code undergoes Parameter Optimization before fitness evaluation. The scores are subsequently updated in the database.

3 Applications and Results

We applied MadEvolve to three cosmological problems, using Quijote [19] and CAMELS [20] simulations. In all cases, algorithms are trained on a single simulation and tested on held-out realizations. Table 1 summarizes the results for each application.

Application I: BAO Reconstruction. Baryon acoustic oscillations (BAO) are a key cosmological standard ruler whose precision is degraded by nonlinear gravitational evolution at low redshifts. Conventionally, standard reconstruction algorithm [21] based on linear theory is used to remove bulk displacements thus sharpening the BAO signal. Since then, a variety of reconstruction methods have been developed to extract more information from BAO wiggles [e.g. 22–26]. We apply MadEvolve to evolve the reconstruction algorithm by maximizing the cross-correlation $r(k)$ between the reconstructed and the true initial density fields using the Quijote N-body simulation suite [19]. In the evaluation step, $r(k)$ is averaged over $k \in [0.01, 0.5]\, h\,\mathrm{Mpc}^{-1}$. We denote this evaluation metric by $\bar{r}$.

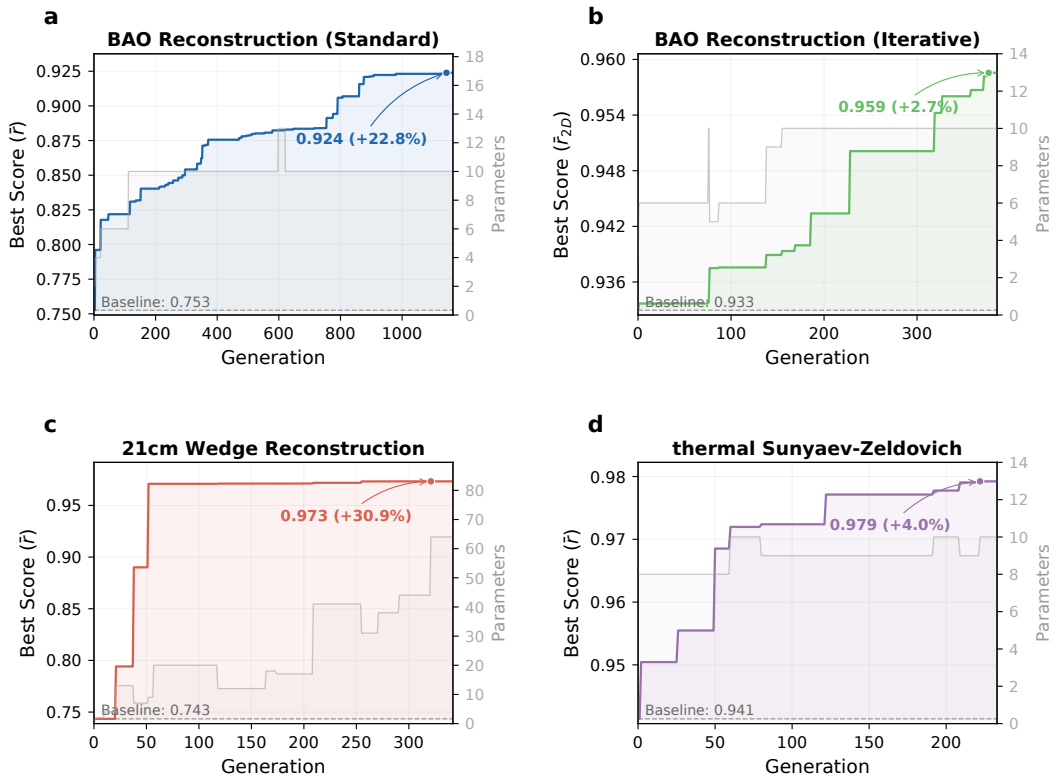

Figure 2: Evolution dynamics of four evolutionary runs. Colored curves (left axis) show best fitness score till current generation. Grey curves (right axis) refer to number of tunable parameters. Dashed lines are the baseline performance. (a) BAO reconstruction (start from standard reconstruction), (b) BAO reconstruction (start from iterative reconstruction), (c) 21 cm wedge reconstruction, (d) tSZ prediction from N-body simulations.

Application	Baseline	Evolved	Improv.	Gens
BAO Recon. (standard)	0.752	0.924	22.8%	1,165
BAO Recon. (iterative SotA)	0.933	0.959	2.8%	386
21 cm Tidal Recon.	0.743	0.971	30.7%	342
N-body baryonic (test loss)	0.613	0.230	63%	233

Table 1: Performance improvements across four evolutionary runs. BAO reconstruction and tidal reconstruction metrics are cross-correlation coefficients with the underlying true fields $\bar{r}$ (higher is better). For tSZ field modeling, performance is quantified by the L1 loss between the predicted and the reference field (lower is better). All results represent the best algorithm discovered during evolution.

Starting from the standard reconstruction ($\bar{r}$=0.752), the evolution discovers higher-order perturbation theory corrections—tensor invariant augmentation using shear tensor invariants (I_1, I_2, I_3) of the displacement field, spectral refinement via non-Gaussian residual subtraction, and multi-tracer fusion with scale-dependent weights—reaching $\bar{r}$=0.924 after 1,165 generations. In a second run starting from the iterative reconstruction baseline [26] ($\bar{r}$=0.933), MadEvolve achieves even higher $\bar{r}$=0.959 by applying differentiable post-processing corrections including anisotropic diffusion guided by the tidal tensor and residual advection along large-scale velocity flows. Notably, evolution from standard reconstruction did not independently rediscover the iterative algorithm, suggesting the framework is better at refining existing methods than retrieving known solutions. Moreover, the discovered algorithms generalize to higher-resolution mesh grids beyond the resolution used during evolution.

Application II: 21 cm Wedge Reconstruction. Observations of the 21 cm hydrogen line at cosmic dawn are contaminated by foreground emission which is orders of magnitude brighter than the signal. This causes mode mixing in a characteristic "wedge" in cylindrical $(k_\perp, k_\parallel)$ Fourier space [27]. Tidal reconstruction [28] exploits the fact that large-scale tidal field modulates the evolution of small-scale structures, enabling statistical recovery of contaminated modes. We apply MadEvolve to evolve the reconstruction algorithm by maximizing the two-dimensional cross-correlation $r_{2D}(k_\perp, k_\parallel)$ between the reconstructed and the underlying true matter fields, averaged over the wedge region defined by $k_\perp \in [0.08, 0.65]\ h$ Mpc^{-1} and $k_\parallel < 0.08\ h$ Mpc^{-1}, evaluated on the Quijote N-body simulation suite [19]. We denote this evaluation metric by $\bar{r}_{2D}$.

Starting from the original tidal reconstruction method ($\bar{r}_{2D}$=0.743), MadEvolve discovers a refined anisotropic tensor reconstruction algorithm. The key innovations of the evolved algorithm include anisotropic filtering in replace of the isotropic Gaussian filter, a generalized potential with tunable spectral index and anisotropy factor $f_{\rm ani}$, and independently weighted transverse, longitudinal, and cross tidal components. The 9-parameter algorithm at generation 52 achieves $\bar{r}_{2D}$=0.971, capturing 99.8% of the gains from the final 64-parameter variant. This 30.7% improvement reduces reconstruction error $(1-\bar{r}_{2D}^2)$ by a factor of eight, and the evolved algorithm outperforms both the baseline tidal method and the cluster fossil quadratic estimator [29–33].

Application III: Baryonic Physics from N-body Simulations. Full hydrodynamic simulations incorporating baryonic physics are prohibitively expensive for large-volume surveys. Lagrangian Deep Learning (LDL) [34] offers an alternative by mapping gravity-only dark matter simulations to baryonic observables. Here we apply MadEvolve to improve LDL specifically for predicting the thermal Sunyaev-Zeldovich (tSZ) effect, which is proportional to electron pressure $P_e = n_e k_B T_e$. The baseline LDL uses a simple power-law transformation from dark matter density to pressure with 18 trainable parameters. The evolution is guided by maximizing the cross-correlation $r(k)$ between the predicted field and the tSZ y-parameter field from hydrodynamic simulations in the CAMELS suite [20], averaged over $k \in [0, 10]\, h\,\mathrm{Mpc}^{-1}$.

The evolved algorithm discovers a physically motivated multiplicative decomposition $y \propto n_e \cdot T_{\rm eff}$ that mirrors the physical definition, factorizing the prediction into an electron density component and an effective temperature. The temperature term combines a screened gravitational potential and a gated shock-heating term that mixes isotropic gradient and anisotropic tidal shear contributions, modulated by local density to amplify heating in collapsing regions. After 233 generations on CAMELS simulations [20], test loss decreases by 63% (from 0.613 to 0.230) and cross-correlation improves from 0.943 to 0.969.

4 Discussion and Conclusion

In this work, we implemented a framework for iteratively improving scientific algorithms by combining LLM with evolutionary search. We show the effectiveness of this framework with three computational cosmology tasks, obtaining significant improvements without human intervention. Besides, we find that the quality of the initial baseline strongly influences the final outcome in the BAO reconstruction task. This suggests that LLM-driven search is most effective as a complement to human expertise, refining and extending existing methods rather than replacing them.

More broadly, this work highlights a new paradigm for scientific discovery with machine learning: rather than treating ML as a black-box predictor, it can be used to explore the space of algorithmic and conceptual possibilities. This approach is particularly promising for cosmology, where modern surveys generate increasingly large and complex datasets. We expect this framework to find wide applications in cosmology and beyond.

Despite these promising results, several limitations still remain. The current framework is computationally expensive, with the main bottleneck arising from repeated evaluations rather than LLM generation. In addition, while the evolved algorithms achieve strong performance, they often lack the clarity and elegance of human-designed solutions. Furthermore, the evolved algorithms have so far been developed and validated exclusively on simulations. Assessment on real observational data is essential, as additional challenges such as survey geometry, selection effects, and astrophysical systematics may limit performance gains. Further work is needed to address these limitations and to better interpret the discovered algorithms.

References

[1] Amir Aghamousa et al. The DESI Experiment Part I: Science,Targeting, and Survey Design. 10 2016.

[2] M. Amiri et al. The CHIME Fast Radio Burst Project: System Overview. 3 2018. doi: 10.3847/1538-4357/aad188.

[3] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, Alhussein Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

[4] Alexander Novikov, Ngan Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131, 2025.

[5] Esteban Real, Chen Liang, David R. So, and Quoc V. Le. Automl-zero: Evolving machine learning algorithms from scratch. *Proc. of ICML*, 119:8007–8019, 2020.

[6] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, and et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610:47–53, 2022. doi: 10.1038/s41586-022-05172-4.

[7] Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, and et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618:257–263, 2023. doi: 10.1038/s41586-023-06004-9.

[8] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *Proc. of ICLR*, 2024. arXiv:2310.12931 [cs.RO].

[9] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292, 2024. URL `https://arxiv.org/abs/2408.06292`.

[10] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv preprint arXiv:2504.08066, 2025. URL `https://arxiv.org/abs/2504.08066`.

[11] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, et al. Towards an ai co-scientist. arXiv preprint arXiv:2502.18864, 2025. URL `https://arxiv.org/abs/2502.18864`.

[12] Francisco Villaescusa-Navarro, Boris Bolliet, Pablo Villanueva-Domingo, et al. The denario project: Deep knowledge ai agents for scientific discovery. arXiv preprint arXiv:2510.26887, 2025. URL `https://arxiv.org/abs/2510.26887`.

[13] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

[15] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

[16] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 12 2014.

[18] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[19] Francisco Villaescusa-Navarro et al. The Quijote simulations. *Astrophys. J. Suppl.*, 250:2, 2020. doi: 10.3847/1538-4365/ab9d82.

[20] Francisco Villaescusa-Navarro et al. The CAMELS project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, 2021.

[21] Daniel J. Eisenstein, Hee-Jong Seo, Edwin Sirko, and David N. Spergel. Improving cosmological distance measurements by reconstruction of the baryon acoustic peak. *Astrophys. J.*, 664:675–679, 2007. doi: 10.1086/518712.

[22] Svetlin Tassev and Matias Zaldarriaga. Towards an Optimal Reconstruction of Baryon Oscillations. *JCAP*, 10:006, 2012. doi: 10.1088/1475-7516/2012/10/006.

[23] Ixandra Achitouv and Chris Blake. Improving reconstruction of the baryon acoustic peak : the effect of local environment. *Phys. Rev. D*, 92(8):083523, 2015. doi: 10.1103/PhysRevD.92.083523.

[24] Hong-Ming Zhu, Yu Yu, Ue-Li Pen, Xuelei Chen, and Hao-Ran Yu. Nonlinear reconstruction. *Phys. Rev. D*, 96(12):123502, December 2017. doi: 10.1103/PhysRevD.96.123502.

[25] Yanlong Shi, Marius Cautun, and Baojiu Li. New method for initial density reconstruction. *Phys. Rev. D*, 97(2):023505, 2018. doi: 10.1103/PhysRevD.97.023505.

[26] Marcel Schmittfull, Tobias Baldauf, and Matias Zaldarriaga. Iterative initial condition reconstruction. *Phys. Rev. D*, 96:023505, 2017. doi: 10.1103/PhysRevD.96.023505.

[27] Miguel F. Morales, Bryna Hazelton, Ian Sullivan, and Adam Beardsley. Four Fundamental Foreground Power Spectrum Shapes for 21 cm Cosmology Observations. *Astrophys. J.*, 752:137, 2012. doi: 10.1088/0004-637X/752/2/137.

[28] Hong-Ming Zhu, Ue-Li Pen, Yu Yu, and Xuelei Chen. Recovering lost 21 cm radial modes via cosmic tidal reconstruction. *Physical Review D*, 98(4):043511, 2018. doi: 10.1103/PhysRevD.98.043511.

[29] Donghui Jeong and Marc Kamionkowski. Clustering Fossils from the Early Universe. *Phys. Rev. Lett.*, 108:251301, 2012. doi: 10.1103/PhysRevLett.108.251301.

[30] Peikai Li, Scott Dodelson, and Rupert A. C. Croft. Large Scale Structure Reconstruction with Short-Wavelength Modes. *Phys. Rev. D*, 101(8):083510, 2020. doi: 10.1103/PhysRevD.101.083510.

[31] Peikai Li, Rupert A. C. Croft, and Scott Dodelson. New Probes of Large Scale Structure. 7 2020.

[32] Omar Darwish, Simon Foreman, Muntazir M. Abidi, Tobias Baldauf, Blake D. Sherwin, and P. Daniel Meerburg. Density reconstruction from biased tracers and its application to primordial non-Gaussianity. *Phys. Rev. D*, 104(12):123520, 2021. doi: 10.1103/PhysRevD.104.123520.

[33] Zhenyuan Wang and Donghui Jeong. Reconstructing the long-wavelength matter density fluctuation modes from the scalar-type clustering fossils. *JCAP*, 07:020, 2024. doi: 10.1088/1475-7516/2024/07/020.

[34] Biwei Dai and Uroš Seljak. Learning effective physical laws for generating cosmological hydrodynamics with Lagrangian deep learning. *Proceedings of the National Academy of Sciences*, 118(16):e2020324118, 2021.