Agentic AI for Cross-Disciplinary Next-Generation Gravitational-Wave Forecasting

Zihan Zhou
Department of Physics,
Princeton University,
Princeton, NJ 08540, USA
zihanz@princeton.edu

Digvijay Wadekar
Center for Gravitational Physics,
University of Texas at Austin,
Austin, TX 78712, USA

Abstract

We demonstrate the modular use of LLM agentic systems for cross-disciplinary scientific forecasts spanning gravitational-wave (GW) astronomy, cosmology, and renormalization theory. We create a framework comprising agentic skills for astrophysical compact-object population synthesis, GW detector-network construction, Fisher forecasting, and analytic GW waveform construction. Our modular framework can propagate astrophysical or fundamental-physics models to GW detector-level observables. We demonstrate the system on two next-generation detector science cases that have not previously been studied: (i) distinguishing reionization-era source-history scenarios using future high-redshift binary black hole detections, and (ii) forecasting constraints on the renormalization-group evolution of multipole moments from GW observations. We also show the importance of our agentic skills in a domain-specific benchmark test and show that they improve the accuracy of state-of-the-art agentic frameworks (Claude Code and GPT/Codex) by a factor of $10-100$.

1 Introduction

Fisher forecasts play a central role in assessing the scientific reach of future experiments across particle phenomenology, gravitational-wave (GW) astronomy, and cosmology. Their appeal is that they provide a fast and interpretable approximation to the information content of prospective measurements, making them especially useful for detector design studies, theory target assessments, and science-priority planning.

At the same time, many scientifically interesting forecasting problems now require combining ideas and software from several subfields. This makes them a natural target for LLM agentic systems, especially when the agents are equipped with domain-specific skills that constrain planning, code generation, validation, and scientific interpretation (see e.g., [20, 21, 17]).

We develop the GW forecast agentic framework for Fisher-matrix-based forecasting with next-generation GW detectors. The framework supports two independent GW domain-specific codes (GWFAST [10, 11] and GWFish [5]), two population-synthesis backends spanning analytic models and numerical codes (RAPSTER and COSMIC [2, 14]), and multiple third-generation detector networks [15]. Because next-generation GW forecasts require multi-stage workflows, source population generation, detector network specification, parameter-estimation surrogates, and science-specific post-processing, they provide a stringent test of agentic scientific infrastructure. We therefore apply the framework to two challenging science cases that, to our knowledge, have not previously been studied in the literature: distinguishing reionization-era source-history scenarios with high-redshift BBH detections, and forecasting constraints on a renormalization-group evolution term in the gravitational-wave phase.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

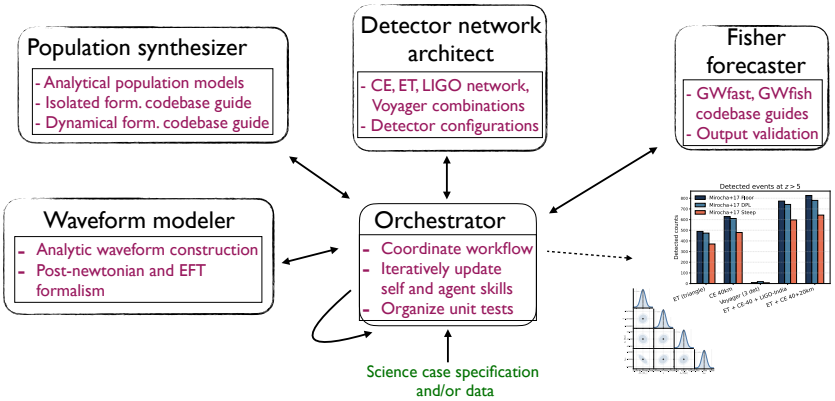

Figure 1: Agentic workflow used in this work. A science-case specification and/or dataset is routed through an orchestrator that coordinates four specialized agents: a population synthesizer, waveform modeler, detector network architect, and Fisher forecaster. Each agent is paired with domain-specific skills grounded in public scientific codebases for reionization modeling, compact-binary population synthesis, and Fisher forecasting [16, 2, 14, 10, 11, 5].

2 Agentic Framework

Our framework takes a science-case specification and routes it through an orchestrator that coordinates four specialized agents with domain-specific skills. Within this workflow, we interchangeably use the Claude Code and Codex agentic coding frameworks as the execution layer for planning, code modification, validation, and orchestration [1, 19]. The modularity of our framework makes it easier to check LLM hallucinations.

1. **Orchestrator**: coordinates the workflow, iteratively updates agent skills, and organizes validation tests.
2. **Population synthesizer**: generates compact-binary populations using analytical models together with isolated and dynamical formation codebase guides. In the current framework we explicitly use `COSMIC` for isolated binary evolution and `RAPSTER` for dynamical formation channels [2, 14], since these codebases are fast to deploy in an agentic workflow; future extensions could incorporate additional population-synthesis tools such as `COMPASS` or `CMC`.
3. **Waveform modeler**: constructs analytic waveforms and incorporates post-Newtonian and effective field theory inspired ingredients.
4. **Detector network architect**: specifies detector configurations and network combinations spanning CE, ET, LIGO network, and Voyager concepts.
5. **Fisher forecaster**: evaluates detectability and parameter-information metrics using `GWfast` and `GWFish`-based forecasting and output validation [10, 11, 5]. We also encode scientific guardrails and validation rules. For example, we regularize or truncate Fisher matrices with large condition numbers before inversion, as can occur in the presence of parameter degeneracies.

3 Case Study I: Reionization-Era Binary Black Holes

One of the most compelling science advantages of 3G detectors is their ability to probe star formation and compact binary synthesis at high redshifts which are inaccessible to electromagnetic observations. Earlier work has therefore proposed connecting future GW observations to star-formation history, metallicity evolution, and high-redshift compact binary populations (see e.g., [22, 18, 6, 7, 3]). We ask whether 3G detector networks can distinguish alternative reionization-era source histories through the detected high-redshift BBH population. We use the `ARES` reionization star formation history models [16] referenced in Fig. 2. We combine them with isolated binaries from `COSMIC` [2] and a sparse dynamical channel from `RAPSTER`, and construct high-redshift conditioned catalogs with observed merger redshift $z > 7$. One practical challenge of this study is that it requires coordinating

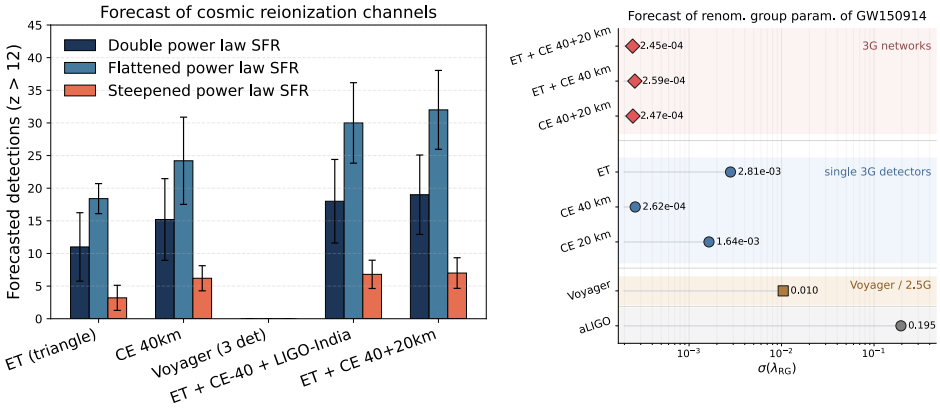

Figure 2: We show forecast results from our agentic workflow for two novel cases from different scientific domains. **Left:** Mean detected counts of binary black holes in the early universe ($z > 12$) for the three cosmic reionization scenarios. The strongest 3G networks preserve a clear hierarchy among the source-history models, which is correlated with the steepness of the star formation efficiency function. **Right:** Forecasted uncertainty on the GW renormalization-group evolution parameter for the detector networks considered in this work.

four domain-specific public codebases (`ARES`, `COSMIC`, `RAPSTER`, and `GWFAST`). However, the native strength of coding agents in reading, adapting, and connecting scientific software helps make this integration tractable.

We refer the reader to [16] for details on the reionization models. As an example, the double power-law model is given by $f_{\rm dpl}(M_{\rm h}) = \frac{f_{*,0}}{\left(\frac{M_{\rm h}}{M_{\rm p}}\right)^{-\gamma_{\rm lo}} + \left(\frac{M_{\rm h}}{M_{\rm p}}\right)^{\gamma_{\rm hi}}}$ where $f \equiv \dot{M}_*/\dot{M}_b$ is the star formation efficiency, $M_{\rm h}$ is the halo mass, $f_{*,0}, M_{\rm p}$ correspond to the peak efficiency and peak halo mass, and $\gamma_{\rm lo}, \gamma_{\rm hi}$ are the low and high mass power-law indices.

Figure 2 shows the mean detected counts above $z > 12$. The error bars are estimated from multiple iterations and show the variance due to the finite number of detections. The separation is clearly visible at the count level for the stronger networks, while the weaker Voyager network contributes essentially no detections.

4 Case Study II: Renormalization-Group Evolution in the GW Phase

This RG evolution forecast redirects the same agentic framework from population astrophysics to fundamental physics. Unlike the reionization example, there is no pre-existing workflow to reuse: the agent should read a recent theory paper, extract a useful approximation, propagate the new parameter through the Fisher pipeline, and return a meaningful physics result.

Waveform model and basic setup We build the waveform model based on Section IV of Ref. [4]. The skill keeps the dominant mode and introduces a single deformation parameter, $\lambda_{\rm RG}$, which modifies the radiative tail sector of Eq. (129) in Ref. [4]:

$$\hat{\hat{\ell}} = \ell + \lambda_{\rm RG}\gamma_{\ell m}^{\rm univ} \ . \tag{1}$$

$\lambda_{\rm RG} = 1$ corresponds to the value in general relativity. The Fisher matrix is evaluated on the parameter set $\{\mathcal{M}_c, \eta, d_L, \lambda_{\rm RG}\}$, where $\mathcal{M}_c$ is the chirp mass, η the symmetric mass-ratio, d_L the luminosity distance. We choose the fiducial source to be GW150914-like binary black holes ($\mathcal{M}_c \simeq 28.3\, M_\odot$, $\eta = 0.247$, $d_L = 410\,{\rm Mpc}$) at $\lambda_{\rm RG} = 1$ (GR). We compute a single-source Fisher forecast for eight detector configurations as shown in Fig. 2.

Agent workflow.

The RG skill first converts Ref. [4] into a forecast target and records the key conventions. The skill then provides instructions to implement the waveform model and to check its numerical stability. The Fisher forecasting skill evaluates the model across detector configurations and turns the outputs into

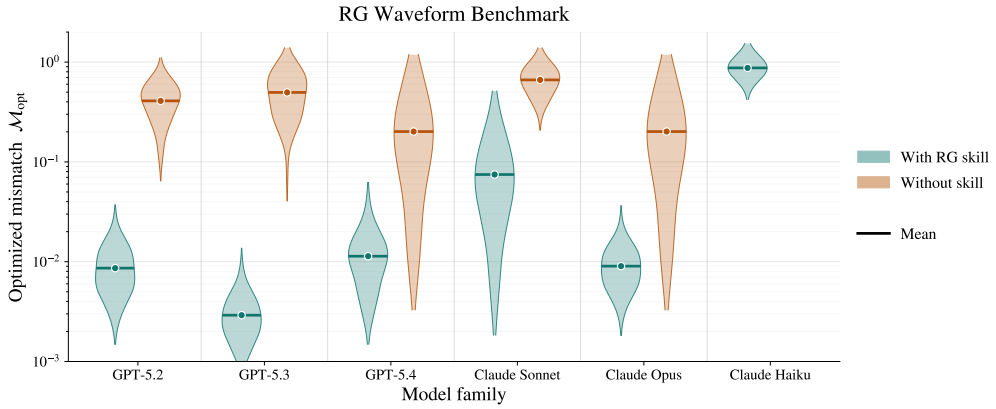

Figure 3: Benchmark comparison for the renormalization-group (RG) evolution waveform model task. The RG skill improves waveform reproduction accuracy by roughly one to two orders of magnitude relative to the no-skill baseline. Claude Opus and Claude Sonnet refer to version 4.6. Among the tested models, GPT-5.3 with the RG skill attains the lowest mismatch.

Fig. 2. In this workflow, we primarily use and refine the analytic waveform creation skill, leaving the downstream forecasting pipeline untouched. For the skill benchmarking, both the with-skill and without-skill conditions are given the same source paper, so the two differ not in the information available but in how it is presented: the without-skill condition supplies only the paper together with the project summary, whereas the with-skill condition additionally provides step-by-step guidance in the waveform construction that points to the specific equations to consult and explains their physical meaning. The benchmark therefore isolates the value of structured procedural guidance rather than reflecting differential access to source information.

Results. The forecasted 1σ uncertainty on $\lambda_{\rm RG}$ ranges from $\sigma(\lambda_{\rm RG}) = 1.95 \times 10^{-1}$ for aLIGO to $\sigma(\lambda_{\rm RG}) = 2.45 \times 10^{-4}$ for the ET+CE 40+20 km reduced network, an improvement of roughly 8×10^2. Voyager already substantially improves on the aLIGO baseline, while larger gains come from 3G sensitivity. Among single detectors, CE 40 km provides the most powerful constraints.

5 Benchmarking the Skills

We show that the agentic skills created in this project improve not only usability but also scientific reliability. We therefore use the RG-tail waveform from Section 4 as a controlled implementation benchmark, which is also part of the follow-up GWBENCHMARKS suite [12]: frontier LLM agents are asked to generate a standalone waveform module in a fresh workspace, and the resulting code is scored against the checked waveform implementation in Section 4. We use the noise-weighted mismatch, optimized over reference time and phase difference, as a metric to quantify performance:

$$\mathcal{M}_{\rm opt} = 1 - \max_{\Delta t, \Delta\phi} \frac{\langle h_{\rm true}, h_{\rm cand} e^{i(2\pi f \Delta t + \Delta\phi)} \rangle}{\|h_{\rm true}\| \, \|h_{\rm cand}\|} \, . \tag{2}$$

We test six LLM agent families, GPT-5.2, GPT-5.3, GPT-5.4, Claude Sonnet, Claude Opus, and Claude Haiku, under two conditions. In the *with-skill* condition, the model receives the project's RG skill documentation, which encodes the waveform conventions, key equations from Ref. [4], and physical interpretations. In the *without-skill* condition, the model receives only a natural-language project summary plus the paper references, without any RG-specific skill layer.

We show the benchmark outcome in Fig. 3. With the skill enabled, GPT-5.3, GPT-5.2, Claude Opus, and GPT-5.4 achieve $\langle \mathcal{M}_{\rm opt} \rangle \simeq 2.9 \times 10^{-3}$, 8.6×10^{-3}, 9.0×10^{-3}, and 1.1×10^{-2}, respectively, while Claude Sonnet reaches 7.5×10^{-2}. Without the skill layer, the corresponding performance degrades by an order of magnitude. As an additional single-agent baseline, we tried the same no-skill prompt with Claude Opus 4.6 and found a mismatch of roughly 70%.

6 Discussion and Conclusion

We have outlined an agentic AI framework for gravitational-wave forecasting and shown how it can be used to address two novel science cases: high-redshift population inference during reionization and a fundamental-physics phase-correction forecast.

One limitation of the present study is that the baseline comparison is restricted to the waveform-skill benchmark rather than a full end-to-end comparison against a single-agent forecasting workflow or an independently produced human analysis. Such comparisons would be useful for separating the benefits of multi-agent decomposition from the benefits of the domain-specific skill layer itself. Important framework-level failure modes include public-codebase API drift, which can be mitigated by agentic skill updates, and Fisher-matrix inversion failures when parameters are strongly correlated, for which we use a pseudoinverse with a tolerance threshold.

We build domain-specific skills for GW forecasting that can be reused in other aspects of GW data analysis, such as parameter inference, population inference, and waveform modeling. We plan to make the domain-specific skills used in this framework publicly available with the camera-ready release.

A GW Background Knowledge and Failure Modes

The New Physics Implementation Bench evaluates whether LLM agents can translate compact theoretical physics derivations into numerically correct scientific software modules under realistic implementation constraints. The benchmark is based on the radiative section in the inspiral phase discussed in Refs. [4, 13, 8, 9], adapted into a standalone frequency-domain waveform implementation task. Agents are provided only with a restricted prompt and a compact formula packet containing the dominant nonspinning $(2, 2)$ RG-tail ingredients. They are not given executable reference code or access to external implementations. The benchmark uses a factorized correction to the dominant Fourier-domain inspiral mode,

$$\tilde{h}_{22}(f) = \tilde{h}_{22}^{\mathrm{Newt}}(f)\, \hat{h}_{22}(x), \tag{3}$$

where the dimensionless correction factor is

$$\hat{h}_{22}(x) = H_{\mathrm{eff}}(x)\, T_{22}(x)\, \rho_{22}^{2}(x)\, e^{i\delta_{22}(x)}. \tag{4}$$

Here H_{eff} is the conservative effective source term, T_{22} is the radiative tail factor, and ρ_{22}, δ_{22} are residual PN amplitude and phase corrections. The benchmark deformation parameter λ_{RG} modifies the running contribution inside the tail sector through

$$\hat{\ell}_{22} = 2 + \lambda_{\mathrm{RG}}\, \gamma_{22}^{\mathrm{univ}}, \tag{5}$$

with $\lambda_{\mathrm{RG}} = 1$ corresponding to the GR limit. The universal anomalous dimension depends on the dimensionless radiative frequency $\hat{k}$ and angular-momentum combination J,

$$\gamma_{22}^{\mathrm{univ}} = -\frac{214}{105}\hat{k}^2 + \frac{2mJ\hat{k}^3}{3} - \frac{3390466}{1157625}\hat{k}^4 + \frac{381863\, mJ\, \hat{k}^5}{99225}. \tag{6}$$

Agents must construct a frequency-domain waveform $h(f; \mathcal{M}_c, \eta, d_L, t_c, \phi_c, \lambda_{\mathrm{RG}})$ using detector-frame quantities and geometric units,

$$M_{\mathrm{sec}} = \mathcal{M}_c\, M_{\odot}\, \eta^{-3/5}, \qquad x = (\pi M_{\mathrm{sec}} f)^{2/3}, \tag{7}$$

where M_{sec} is the detector-frame total mass in geometric seconds and x is the PN frequency variable. The waveform is truncated near the innermost stable circular orbit (ISCO),

$$f_{\mathrm{ISCO}} = \frac{1}{\pi 6^{3/2} M_{\mathrm{sec}}}, \tag{8}$$

and smoothly tapered using the Fermi window

$$W(f) = \frac{1}{1 + \exp[(f - f_{\mathrm{ISCO}})/\sigma]}, \qquad \sigma = \sigma_{\mathrm{taper}}\, f_{\mathrm{ISCO}}, \tag{9}$$

where σ controls the taper width near the cutoff frequency. The benchmark intentionally leaves several implementation details implicit, including Fourier-phase conventions, stationary-phase amplitude

normalization, numerical stabilization choices, and complex tail evaluation. As a result, qualitatively reasonable waveforms can still fail quantitatively due to subtle scientific implementation errors such as inconsistent phase conventions, omitted tail phases, incorrect geometric-unit conversions, or unstable exponential evaluations.

Submitted implementations are evaluated against a hidden reference waveform over a 144-case test grid spanning chirp mass, symmetric mass ratio, luminosity distance, and λ_{RG}, using a detector-noise-weighted frequency-domain mismatch defined as follows. For two frequency-domain waveforms $a(f)$ and $b(f)$, the noise-weighted inner product is

$$\langle a, b\rangle = 4\,\mathrm{Re}\int_{f_{\min}}^{f_{\max}} \frac{\tilde{a}(f)\,\tilde{b}^{*}(f)}{S_n(f)}\,\mathrm{d}f, \tag{10}$$

where $S_n(f)$ is the one-sided power spectral density—here `aLIGOZeroDetHighPower`—and the analysis band $[f_{\min}, f_{\max}]$ is bounded below by the PSD low-frequency cutoff and above by the tapered ISCO frequency.

References

[1] Anthropic. Claude code: Agentic coding in the terminal. `https://docs.anthropic.com/en/docs/claude-code/overview`, 2026. Accessed: 2026-04-16.

[2] Katelyn Breivik, Scott Coughlin, Michael Zevin, Carl L. Rodriguez, Kyle Kremer, Claire S. Ye, Jeff J. Andrews, Michael Kurkowski, Matthew C. Digman, Shane L. Larson, and Frederic A. Rasio. COSMIC Variance in Binary Population Synthesis. ApJ, 898(1):71, July 2020. doi: 10.3847/1538-4357/ab9d85.

[3] Aditya Chatterjee, Tirthankar Roy Choudhury, et al. Gravitational wave informed inference of 21-cm global signal parameters. *arXiv e-prints*, 2026. Preprint.

[4] Andrea Cipriani, Alessandro Nagar, Francesco Fucito, and José Francisco Morales. From the confluent Heun equation to a new factorized and resummed gravitational waveform for circularized, nonspinning, compact binaries. 2 2026.

[5] Ulyana Dupletsa, Jan Harms, Biswajit Banerjee, Marica Branchesi, Boris Goncharov, Andrea Maselli, Ana Carolina Silva Oliveira, Samuele Ronchini, and Jacopo Tissino. gwfish: A simulation software to evaluate parameter-estimation capabilities of gravitational-wave detector networks. *Astron. Comput.*, 42:100671, 2023. doi: 10.1016/j.ascom.2022.100671.

[6] Irina Dvorkin, Elisabeth Vangioni, Joseph Silk, Jean-Philippe Uzan, and Keith A. Olive. Metallicity-constrained merger rates of binary black holes and the stochastic gravitational wave background. *Monthly Notices of the Royal Astronomical Society*, 461(4):3877–3885, 2016. doi: 10.1093/mnras/stw1477.

[7] Maya Fishbach. Probing cosmic chemical enrichment with next-generation gravitational-wave observatories. *Class. Quant. Grav.*, 42(5):055009, 2025. doi: 10.1088/1361-6382/adaf70.

[8] Francesco Fucito, Jose Francisco Morales, and Rodolfo Russo. Gravitational wave forms for extreme mass ratio collisions from supersymmetric gauge theories. *Phys. Rev. D*, 111(4):044054, 2025. doi: 10.1103/PhysRevD.111.044054.

[9] Walter D. Goldberger and Andreas Ross. Gravitational radiative corrections from effective field theory. *Phys. Rev. D*, 81:124015, 2010. doi: 10.1103/PhysRevD.81.124015.

[10] Francesco Iacovelli, Michele Mancarella, Stefano Foffa, and Michele Maggiore. Forecasting the Detection Capabilities of Third-generation Gravitational-wave Detectors Using GWFAST. *Astrophys. J.*, 941(2):208, 2022. doi: 10.3847/1538-4357/ac9cd4.

[11] Francesco Iacovelli, Michele Mancarella, Stefano Foffa, and Michele Maggiore. GWFAST: A Fisher Information Matrix Python Code for Third-generation Gravitational-wave Detectors. *Astrophys. J. Supp.*, 263(1):2, 2022. doi: 10.3847/1538-4365/ac9129.

[12] Tousif Islam, Digvijay Wadekar, and Zihan Zhou. gwBenchmarks: Stress-Testing LLM Agents on High-Precision Gravitational Wave Astronomy. 5 2026.

[13] Mikhail M. Ivanov, Yue-Zhou Li, Julio Parra-Martinez, and Zihan Zhou. Resummation of Universal Tails in Gravitational Waveforms. *Phys. Rev. Lett.*, 135(14):141401, 2025. doi: 10.1103/jzd1-qzkt.

[14] Konstantinos Kritos, Vladimir Strokov, Vishal Baibhav, and Emanuele Berti. Dynamical formation of black hole binaries in dense star clusters: Rapid cluster evolution code. *Phys. Rev. D*, 110(4):043023, 2024. doi: 10.1103/PhysRevD.110.043023.

[15] Michele Maggiore, Chris Van Den Broeck, Nicola Bartolo, Enis Belgacem, Daniele Bertacca, Marie Anne Bizouard, Marica Branchesi, Sebastien Clesse, Stefano Foffa, Juan García-Bellido, Stefan Grimm, Jan Harms, Mark Hindmarsh, Manuel Hohmann, Mick Hubner, Alexander C. Jenkins, Gerard Johansson, Ramunas Juska, Stavros Katsanevas, Erik Katsavounidis, Keiko Kokeyama, Mikhail Korobko, Danny Laghi, Brahim Lamine, Jan Langendorff, Annalisa Lartaux-Vollard, Andreas Leins, Tingxuan Li, Vincent Loriette, Harald Lück, Salvatore Maggiore, Dragoljub Marković, Suvodip Mukherjee, Alexander H. Nitz, Alessio Notari, Maria Okounkova, Cristiano Palomba, Daniela Paoletti, Tania Regimbau, Antonio Riotto, Mairi Sakellariadou, Bangalore S. Sathyaprakash, Beate Schulte, Alberto Sesana, Lijing Shao, Xavier Siemens, Magdalena Sieniawska, Kentaro Somiya, Lorenzo Speri, Christian Stephan-Otto, Nikolaos Stergioulas, Nicola Tamanini, Alexandre Toubiana, Daniyar Tuyenbayev, Artemiy Utina, Gabriele Vajente, Alberto Vecchio, Justin Vines, Will Vousden, Matias Zaldarriaga, Wen Zhao, and Tao Zhu. Science case for the Einstein Telescope. *Journal of Cosmology and Astroparticle Physics*, 2020(03):050, 2020. doi: 10.1088/1475-7516/2020/03/050.

[16] Jordan Mirocha, Steven R. Furlanetto, and Guochao Sun. The global 21-cm signal in the context of the high- z galaxy luminosity function. MNRAS, 464(2):1365–1379, January 2017. doi: 10.1093/mnras/stw2412.

[17] Nikhil Mukund, Yifang Luo, Fan Zhang, Lisa Barsotti, and Erik Katsavounidis. MARVEL: A Multi Agent-based Research Validator and Enabler using Large Language Models. *arXiv e-prints*, art. arXiv:2601.03436, January 2026. doi: 10.48550/arXiv.2601.03436.

[18] Ken K. Y. Ng, Salvatore Vitale, Will M. Farr, and Carl L. Rodriguez. Probing multiple populations of compact binaries with third-generation gravitational-wave detectors. *The Astrophysical Journal Letters*, 913(1):L5, 2021. doi: 10.3847/2041-8213/abf8be.

[19] OpenAI. Openai codex. `https://platform.openai.com/docs/guides/code`, 2025. Accessed: 2026-04-16.

[20] Tilman Plehn, Daniel Schiller, and Nikita Schmal. MadAgents. 1 2026.

[21] Francisco Villaescusa-Navarro, Boris Bolliet, Pablo Villanueva-Domingo, Adrian E. Bayer, Aidan Acquah, Chetana Amancharla, Almog Barzilay-Siegal, Pablo Bermejo, Camille Bilodeau, Pablo Cárdenas Ramírez, Miles Cranmer, Urbano L. França, ChangHoon Hahn, Yan-Fei Jiang, Raul Jimenez, Jun-Young Lee, Antonio Lerario, Osman Mamun, Thomas Meier, Anupam A. Ojha, Pavlos Protopapas, Shimanto Roy, David N. Spergel, Pedro Tarancón-Álvarez, Ujjwal Tiwari, Matteo Viel, Digvijay Wadekar, Chi Wang, Bonny Y. Wang, Licong Xu, Yossi Yovel, Shuwen Yue, Wen-Han Zhou, Qiyao Zhu, Jiajun Zou, and Íñigo Zubeldia. The denario project: Deep knowledge ai agents for scientific discovery, 2025. URL `https://arxiv.org/abs/2510.26887`.

[22] Salvatore Vitale, Will M. Farr, Ken K. Y. Ng, and Carl L. Rodriguez. Measuring the star formation rate with gravitational waves from binary black holes. *The Astrophysical Journal Letters*, 886(1):L1, 2019. doi: 10.3847/2041-8213/ab50c0.