PROCEEDINGS, 50th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2025
SGP-TR-229

GeoGym: a Benchmark Dataset for Evaluating Geothermal Exploration Strategies

Zhouji Liang, Sofia Cecilia Brisson, Robin Thibaut, Junjie Yu, Carl Hoiland, Ahinoam Pollack

90 S 400 W, Salt Lake City, UT, 84101

george@zanskar.us

Keywords: database, simulation, artificial intelligence, modeling

ABSTRACT

There has been increasing global attention to the growing demand for a cost-efficient and carbon free energy resource to power development of artificial intelligence (AI), as well as the application of AI for scientific challenges. AI has the potential to both transform geothermal energy exploration and greatly increase the supply of renewable energy. The development of exploration-focused AI techniques, however, requires a large dataset of geologic models of geothermal systems for training algorithms. This dataset needs to be both geologically realistic and sufficiently comprehensive in order for geoscientists to apply these algorithms effectively to real-world exploration campaigns. Unfortunately, such a dataset for geothermal exploration is currently unavailable. The scarcity of geothermal sites, compared to conventional oil and gas fields, as well as the lack of digitized and systematically formatted data from earlier exploration efforts, has contributed to this gap. In this study, we address this issue by developing a dataset of geologic models of geothermal systems that combines real-world data with advanced physical simulations. This dataset, named "GeoGym", represents a significant step forward in AI training and algorithm development for geothermal exploration. It enables a quantitative assessment of algorithm-assisted decision-making and data collection strategies.

1. INTRODUCTION

In an era of urgent climate concerns and growing clean energy demands, geothermal energy is a baseload power with a large development potential (IEA, 2024). However, despite its promise, the exploration for viable geothermal resources has historically been hampered by a high failure rate of deeper exploration wells, making project development risky and costly (Siahaan et al., 2023). This high failure rate often arises from a diverse set of challenges: how to algorithmically or manually leverage individual collected geological, geophysical and geochemical datasets for well targeting? How to integrate these datasets together? How to assess and utilize uncertainty estimates in the decision making process? What is the best drilling strategy in terms of number and depths and types of temperature gradient or slim wells? What value do individual datasets contribute to the final drilling decision? When to end a drilling campaign? Many papers address one or more of these challenges, though often on just one or a limited number of geothermal sites - limiting the ability to quantitatively evaluate the generalizability and effectiveness of exploration methods across diverse sites.

High-quality datasets are essential to address these challenges and support the rigorous testing of methods and data-types used for well targeting. The success of modern AI across diverse fields—from autonomous vehicles relying on massive driving datasets, to large language models (LLMs) trained on vast text corpora—underscores the transformative power of publicly available, large-scale data. Similarly, robust datasets and benchmarks in the geothermal exploration space, open to academia and industry, will lead to development of AI products that will enhance reservoir characterization, reduce development risks, and increase geothermal energy production.

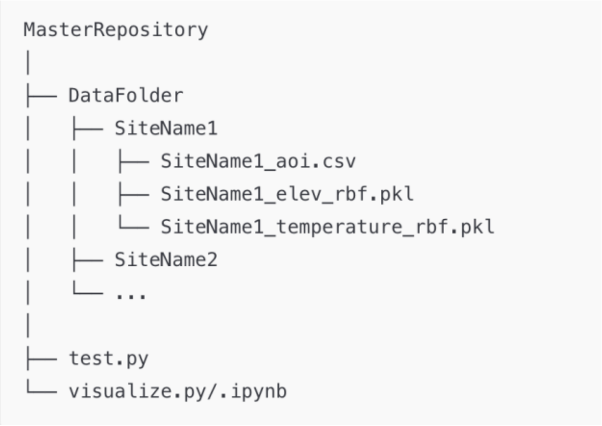
To address this gap, we assembled, for the first time, a dataset of more than 15 real-world geothermal reservoir sites, following a standardized format. This dataset aims to establish a benchmark for advancing geothermal exploration research and innovation.

2. METHOD

The presented dataset, which currently includes thermal models for over 15 geothermal sites, is publicly accessible (Liang et al., 2025), and will be expanded as new data becomes available. To construct these models, we integrated well temperature measurements, geological surveys, and other relevant subsurface information to accurately capture the complex spatial relationships among faults, layers, and lithological units. We then performed numerical simulations for each location to replicate the physical processes governing heat and fluid flow, calibrating the models against observed well temperature data. The resulting temperature distributions show good agreement with real-world measurements, highlighting the reliability of both the modeling approach and the input data.

2.1 Repository Structure

We published our benchmark dataset in a repository hosted on Kaggle. The repository is structured in the following format (Figure 1):



```
MasterRepository
│
├── DataFolder
│   ├── SiteName1
│   │   ├── SiteName1_aoi.csv
│   │   ├── SiteName1_elev_rbf.pkl
│   │   └── SiteName1_temperature_rbf.pkl
│   ├── SiteName2
│   └── ...
│
├── test.py
└── visualize.py/.ipynb
```

Figure 1: Folder structure of the dataset repository

In the following, you can find the detailed descriptions of the items in the repository.

1. Each individual site’s dataset is placed in a separate subfolder within the DataFolder directory.
2. The script test.py checks every folder to confirm that the formatting and file contents are correct.
3. The script visualize.py/.ipynb is used to quickly visualize the dataset in the PyVista package. An example of a 3D visualization is shown in Figure 2.

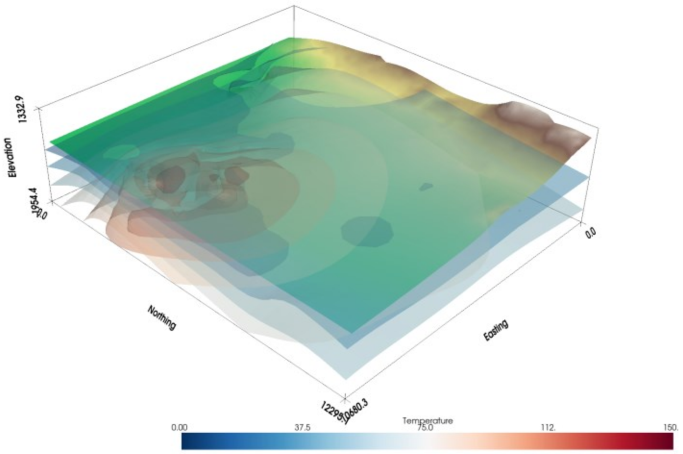


Figure 2: 3D visualization of the temperature cube shown in isosurface, with vertically exaggerated topography to enhance visual clarity.

2.2 File Description

Each site data folder contains the following files described in **Table 1**. To ensure the uniformity of dataset, the following rules are applied to the dataset:

1. All measurement units for distance are in meters and for temperature are in degrees Celsius.
2. Centered Coordinate System: All models are centered at the origin to maintain a standardized spatial reference for data integration and comparison. By removing direct geographic references, each site remains anonymous, thereby encouraging broader community contributions and more open collaboration on model improvements.

Table 1: Description of the files in the repository

File Name	Format	Description	Comments

SiteName_aoi	.csv file	Comma-separated file with six columns: Xmin, Xmax, Ymin, Ymax, Zmin, Zmax. This defines the area of interest (AOI), specifying the dataset's spatial bounds.	This file defines the scale of the dataset. All the coordinates have been shifted to the origin, i.e. Xmin = 0, Ymin=0 and Zmax =0. The reference elevation is set to the minimum elevation of the topography
SiteName_elev_rbf	.pkl (Python pickle)	Stores a SciPy RBFInterpolator providing elevation data. Given (x, y), it returns a z-value that corresponds to the local topographic elevation.	The minimum elevation of the topography of the area is shifted to Z=0 for the consistency and anonymity.
SiteName_temperature_rbf	.pkl (Python pickle)	Contains a SciPy RBFInterpolator for 3D temperature data. Given (x, y, z), it returns a corresponding temperature value based on the interpolation model.	The interpolator is constructed by a grid with 100m x 100m 100m cell size. The linear kernel is used for the RBF interpolator.

3. APPLICATIONS

The benchmark dataset presented here offers abundant opportunities for AI-driven research in the geothermal sector, enabling researchers to conduct quantitative evaluations of novel algorithms and computational methods. By providing standardized, diverse data, it also serves as a valuable resource for educational purposes, allowing students and instructors to explore real-world geothermal scenarios in a structured, consistent manner.

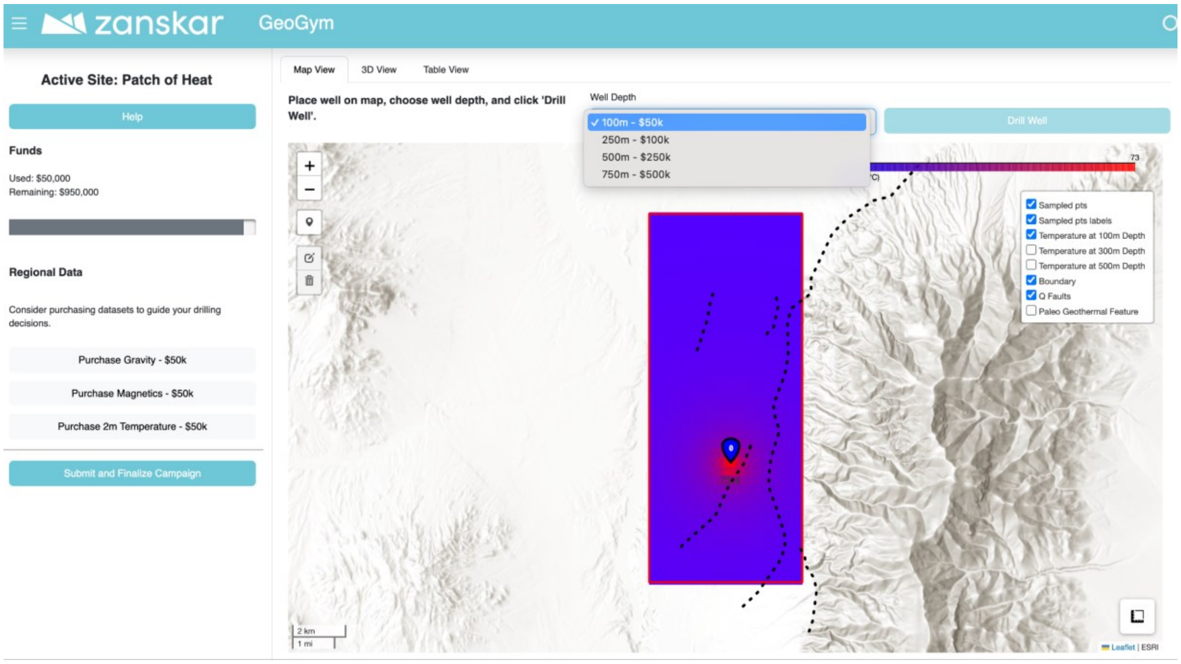
The dataset supports multiple "games" or benchmarks. One is the challenge of developing an algorithm for maximizing discovered temperature under a fixed drilling budget, across multiple drilling campaigns. Researchers can also improve such algorithms for aiming not only for the highest mean temperature outcomes across multiple sites, but also for reduced performance variance across sites. This benchmark metric is discussed in more detail in the following section. Another challenge focuses on predictive modeling: given just a few sampled wells, the task is to estimate either the complete subsurface temperature distribution or the temperature at any new well location. These complementary scenarios highlight the dataset's versatility for both cost-constrained decision-making and advanced machine-learning interpolation, reflecting the real-world complexities of geothermal exploration.

Building on the first metric of cost-contrained exploration, we developed a game environment called GeoGym—showcased at the 2024 Geothermal Rising Conference (GRC2024)—to demonstrate one practical application of this benchmark. In the next section, we delve into the details of the GeoGym setup and how it leverages the dataset for interactive geothermal exploration and learning.

3.1 GeoGym

The name GeoGym is inspired by the Gymnasium library (Towers et al., 2024) - an API standard for reinforcement learning with a diverse collection of reference environments developed by OpenAI. We designed this environment to provide a similar standard environment in geothermal for training both algorithmic and human agents. A snapshot of the graphical user interface for human players is shown in Figure 3. The game rules are set up to mimic an actual exploration process and includes the following steps:

- Each player is assigned a random site and given initial funds of $1M.
- The players have the option to purchase 3 types of datasets if available for the site: gravity data, magnetics data and 2 meter temperature data. Each dataset costs $50k.
- The players have the option to drill 4 types of wells, each of them with a different cost:
  - 100m - $50k
  - 250m - $100k
  - 500m - $250k
  - 750m - $500k
- The player can 'place a well' on a desired location within the area of interest.
- The 3D interpolation of the 'well temperature' and the 2D temperature profiles are shown.
- The player continues drilling wells until they run out of funds.
- The goal is to find the maximum temperature in the field.



Active Site: Patch of Heat

Funds

Used: $50,000
Remaining: $950,000

Regional Data

Consider purchasing datasets to guide your drilling decisions.

Place well on map, choose well depth, and click 'Drill Well'.

Figure 3: Screenshot of the GeoGym game. The left sidebar displays the used and remaining funds, along with the available data that can be purchased. On the right, three tabs allow users to toggle between the map view, 3D view, and drilled well data. A dropdown menu in the center presents different well options, each with its own associated cost. The pin point shows the location of the first drilled wells and its highest temperature. The colored map shows a slice at depth of the 3D interpolation of the temperature field based on the well data.

3.2 Performance Metrics

To account for variations in the maximum attainable temperatures across different sites, we introduced a percentage-based metric to measure a user's performance regardless of the site played. We use this metric to quantitatively evaluate the performance of different players on different sites and also later for the objective function for the algorithm.

$$f_{temp}(T) = \frac{T - T_{min}}{T_{max} - T_{min}} \times 100\% \tag{1}$$

where T_{min} and T_{max} are the minimum and maximum temperature one can achieve with the maximum depth of drilling in the game (750m in the current game).

3.3 Game Results Analysis

During the GRC 2024 conference, the GeoGym game attracted many players from both academia and industry. In total, 65 players participated, generating 235 completed game rounds across 20 different geothermal sites. One site's outcomes are shown in Figure 4, where each color-coded line represents a single player's approach. On the horizontal axis is the cumulative drilling cost, while on the vertical axis is the highest temperature found so far in that playthrough. The black dashed line near 140 °C indicates the maximum possible temperature attainable at this specific site.

Close inspection reveals contrasting strategies. The orange player (visually emphasized with round markers along the orange line) demonstrates what we call "progressive exploration," making incremental investments in shallow wells, before finally drilling a deeper, more expensive well. This approach spreads out risk and collects subsurface information along the way. In contrast, the light-blue player (marked with diamond symbols on the light-blue line) exhibits a "leap of faith" strategy, racing to a deeper well at an earlier stage to potentially secure a higher temperature faster. While this can be highly rewarding, it also risks running out of funds if the early deep well does not yield the expected results. Figure 4 underscores that even when players are given the same site and starting budget, the chosen exploration path strongly influences both the total cost spent and the peak temperature reached. The diversity of colored lines in the plot highlights the myriad ways individuals balance drilling depth, location, and data acquisitions under budget constraints. Such variability also illustrates the game's educational value—by experimenting with different strategies, players gain insights into real-world trade-offs encountered in geothermal exploration, where costly yet potentially high-payoff wells must be weighed against more systematic, stepwise investigations.

Moving beyond a single site, Figure 5 offers a broader perspective across all 20 sites, combining results under the introduced performance metric. The solid green line represents the average percentage of the maximum site temperature achieved at each cumulative cost value, while the shaded region denotes the p25–p75 range. This view highlights how some players manage to approach the site's thermal maximum quickly, whereas others remain well below the potential peak temperature, often due to premature budget exhaustion or suboptimal well placement.

Finally, in Figure 6, we categorize observed drilling sequences into "progressive exploration" (non-descending depths) versus "leap-of-faith drilling" (early direct drilling of deep wells). The box plot demonstrates that progressive explorers tended to secure higher average maximum temperatures, reflecting the benefit of gathering incremental knowledge via shallower wells before committing to deeper, more expensive ones. By contrast, those who didn't do enough exploration before drilling the deep well often achieved subpar outcomes, underscoring the risks of an unfocused strategy. These results not only emphasize the impact of informed decision-making within the GeoGym environment, but also highlight broader implications for real-world geothermal exploration, where effective sequencing and budgeting can significantly influence project success.

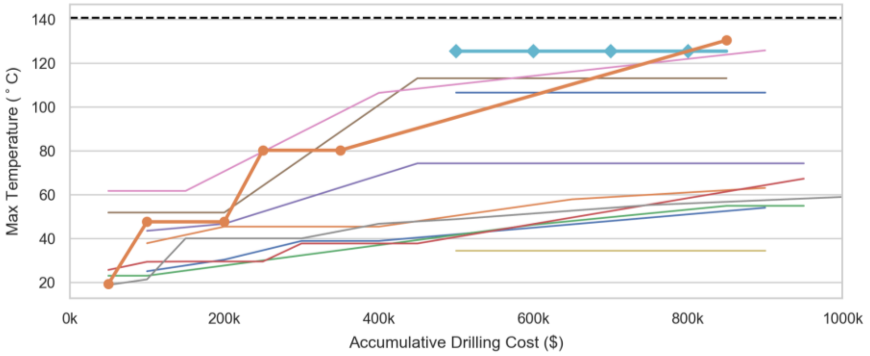


Figure 4: The game result of a representative site played during GRC2024. Each different color represents a different player. The black dashed line shows the maximum possible temperature can be reached. Markers highlight the gameplay of two example players discussed in the text.

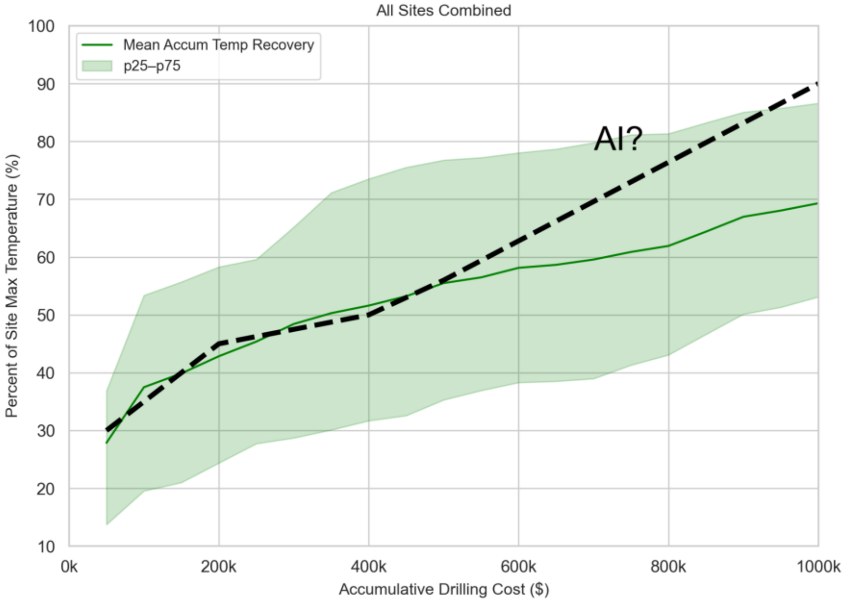


Figure 5: The game result of all sites, evaluated based on the performance metrics introduced in this paper. The black dashed line represents the anticipated AI performance benchmark.

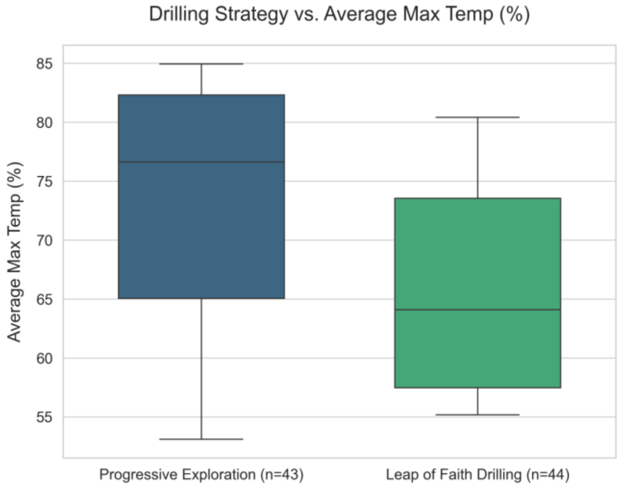


Figure 6: The performance based on different drilling strategies. The strategy is categorized based on the sequence of the drilling depth. If the depth sequence is non-descending, we categorize them as Progressive Exploration. Otherwise it will be categorized as Insufficient Exploration.

4. DISCUSSION

The results from the GeoGym game and the accompanying benchmark dataset underscore both the complexity and the promise of AI-assisted geothermal exploration. As illustrated in Figure 4 (the multi-color lines representing various players' attempts at a single site), participants adopted a wide spectrum of drilling and data-acquisition strategies. Some players opted for an aggressive "deep first" approach—investing heavily in their earliest wells to reach deeper, hotter intervals—while others followed a more conservative "ascending" strategy, starting with shallow wells and incrementally moving deeper to minimize initial expenditure. The box plot in Figure 4 comparing these two cost-sequence categories visually confirms that players who front-loaded their drilling costs often achieved higher temperatures more quickly but occasionally ran out of funds before maximizing resource potential, whereas more cautious approaches sometimes yielded lower peak temperatures but did so more consistently.

From a methodological perspective, the dataset's uniform file structure and standardized coordinate system lowered the barrier to analyzing cross-site results. While every site had different maximum attainable temperatures—necessitating a percentage-based metric to enable fair comparisons—the consistent formatting allowed rapid prototyping of AI algorithms across multiple sites. Owing to the broad geological diversity encompassed in the more than 15 real-world locations, the dataset offers unique opportunities for benchmarking algorithms on settings that vary by fault configuration, temperature gradient, lithological heterogeneity, and data availability.

The positive reception at the GRC 2024 conference supports the value of "gamifying" geothermal exploration to attract not only experts but also newcomers from academia and industry. Informal feedback revealed that participants rapidly deepened their understanding of the interplay among geological uncertainty, well costs, and diminishing financial resources. At the same time, these real-time decisions and their final outcomes—now documented in our dataset—can serve as a novel training resource for machine-learning researchers developing new methods in geothermal targeting.

Nonetheless, several limitations remain. The most critical of which is the representativeness of the training models for current exploration targets and the quality and quantity of data used for creating the models. Ongoing efforts aim to expand the dataset with additional sites, incorporate more complex geological and physical models to represent the subsurface more faithfully.

Despite these caveats, the game results and the variety of strategies exhibited underscore the importance of developing consistent and tested algorithmic approaches to exploration strategies. The benchmark dataset provides a test bed for future algorithmic innovations, bridging research and real-world application by enabling scientists to develop exploration algorithms. We look forward to future work from research teams showing performance plots similar to the black line on Figure 5.

5. CONCLUSION

In this paper, we introduced a first-of-its-kind benchmark dataset named GeoGym of over 15 geothermal sites, along with a gamified exploration environment. By blending curated real-world measurements with physically grounded simulations, the dataset provides a robust platform for testing machine-learning algorithms and for teaching the fundamentals of geothermal exploration. The analysis of GeoGym gameplay at the GRC 2024 conference demonstrates that even small differences in drilling strategy and data-purchase decisions can have an outsized effect on outcomes. Observing human decision-making "in the wild" not only validates the dataset's complexity but also opens avenues for improved AI algorithms designed to optimize well targeting under uncertainty.

This work paves the way for further advancements in geothermal exploration research. Future expansions of the benchmark will include additional sites, more varied rock-property distributions, and additional datasets. We anticipate that these data and the associated gameplay records will help researchers refine both supervised and reinforcement-learning approaches, ultimately driving down exploration risks and catalyzing wider adoption of geothermal energy. By fostering an open, collaborative environment—both on Kaggle and through interactive demos such as GeoGym—we hope to accelerate innovation and usher in the next generation of AI-enabled, cost-efficient geothermal resource discovery.

REFERENCES

International Energy Agency (IEA): The Future of Geothermal Energy, IEA, (2024). https://iea.blob.core.windows.net/assets/b5b73936-ee21-4e38-843b-8ba7430fbe92/TheFutureofGeothermal.pdf

Liang, Z., Brisson, S.C., Pollack, A., Thibaut, R., and Yu, J.: Zanskar's GeoGym [Dataset], Kaggle (2025). https://doi.org/10.34740/KAGGLE/DSV/10613518.

Siahaan, M.F., Septiani, G.A., Purba, D., and Paripurna, A.: Geothermal Exploration Project De-risking: Discussion on Various Schemes, Proceedings, 48th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA (2023).

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Tan, H., and Younis, O. G.: Gymnasium: A Standard Interface for Reinforcement Learning Environments, arXiv preprint arXiv:2407.17032 (2024).