

Supplementary Material: PiFlow: Principle-aware Scientific Discovery with Multi-Agent Collaboration

Anonymous submission

Contents

1	Justification for Benchmark Selection	2
1.1	Motivation and Challenges	2
1.2	Selection Principles	2
1.3	Domain Task Selection and Justification	2
1.3.1	Nanohelix Design	2
1.3.2	Bio-molecule Optimization	2
1.3.3	Superconductor Optimization	2
1.4	Alternative Domain Considerations	3
2	Establishing the “Last Record” in Task Description	4
2.1	Domain-Specific Baseline Configurations	4
2.1.1	Nanohelix Optimization	4
2.1.2	Bio-molecules Optimization	5
2.1.3	Superconductor Optimization	5

This document provides supplementary information to accompany the main manuscript titled “*PiFlow: Principle-aware Scientific Discovery with Multi-Agent Collaboration*”. (Note: Detailed implementation code, surrogate model architectures, training procedures, and performance statistics, e.g., r^2 values, are provided in the accompanying code directories and main manuscript Appendix.)

1 Justification for Benchmark Selection

1.1 Motivation and Challenges

Large language model-based multi-agent systems address scientific discovery through iterative experiment design and validation loops. Evaluating these frameworks requires balancing computational efficiency with the complexity of scientific research, particularly in domains with substantial accumulated knowledge depth. The challenge lies in selecting benchmark scenarios that are both (a) **computationally tractable** and **scientifically meaningful**, while providing (c) **sufficient diversity** to evaluate system versatility across different methods.

1.2 Selection Principles

Our benchmark selection follows five fundamental principles that collectively ensure comprehensive evaluation of computational discovery frameworks:

1. **Scientific Impact and Contemporary Relevance.** Each domain represents an active area of cutting-edge research with transformative potential. Recent developments underscore their continued importance: chiral materials have recently converged with deep learning to ignite innovative breakthroughs [8, 12, 7], AI in drug development continues to see strong growth [6, 5, 9, 2, 1, 3], and room temperature superconductivity [10, 4, 11, 13] could introduce more efficient power grids, better magnetic resonance imaging, faster Magnetic Levitation trains, and new motors and scientific instruments.
2. **Complexity and Diversity.** The scenarios span different types of optimization landscapes. The continuous parameter spaces of nanohelices, discrete molecular spaces of bio-molecules, and compositional/structural spaces of superconductors, provide a comprehensive test of PiFlow’s adaptability across varied mathematical formulations and constraint structures.
3. **Principle Potential.** Each domain offers distinct opportunities for incorporating scientific principles into the discovery process, from electromagnetic theory in nanophotonics to structure-activity relationships in drug discovery and quantum mechanical principles in superconductivity.
4. **Surrogate Model Feasibility.** All three scenarios permit the development of reliable surrogate models using established computational methods (electromagnetic simulations, molecular property prediction, and materials property modeling), enabling controlled and reproducible benchmarking without prohibitive experimental costs.
5. **Measurable Optimization Objectives:** Each scenario provides quantifiable target properties (g-factor, pChEMBL value, critical temperature T_c) that enable direct performance comparison and statistical analysis of different discovery strategies.

1.3 Domain Task Selection and Justification

Based on these principles, we selected three complementary domains that collectively address the full spectrum of computational discovery challenges, as shown in Table 1.

1.3.1 Nanohelix Design

Chiral nanomaterials represent a convergence of nanotechnology and quantum optics with applications in biosensing, asymmetric catalysis, and quantum information processing. Recent developments in deep learning-enabled optical design have demonstrated significant breakthroughs in this field.

1.3.2 Bio-molecule Optimization

AI-driven drug discovery continues to show exponential growth with significant industrial and therapeutic impact. The structure-activity relationship paradigm provides a well-established framework for principle-guided molecular optimization.

1.3.3 Superconductor Optimization

Room-temperature superconductivity represents one of the most significant unsolved challenges in condensed matter physics, with potential to revolutionize energy transmission, magnetic applications, and quantum computing.

Table 1: Methodological characteristics of benchmark scenarios

Characteristic	Nanohelix	Optimization	Bio-molecule	Optimization	Superconductor	Optimization
Optimization Landscape	Continuous space with multimodal functions	parameter smooth but objective	Discrete graph space with complex structure-property relationships	molecular with complex structure-property relationships	Compositional and structural space with complex many-body quantum effects	
Constraint Structure	Geometric constraints with electromagnetic effects	feasibility with electromagnetic coupling	Chemical validity constraints with ADMET property boundaries		Thermodynamic stability and electronic structure constraints	
Principle Integration	Maxwell equations, electromagnetic field theory, chirality principles		Structure-activity relationships, pharmacophore principles, medicinal chemistry rules		BCS theory, band structure principles, materials design rules	
Surrogate Model Basis	Finite-difference time-domain (FDTD) electromagnetic simulations		Graph neural networks trained on experimental bioactivity databases		Density functional theory calculations and materials property databases	
Parameter Space	4D continuous (fiber radius, helix radius, turns, pitch)		High-dimensional discrete (molecular graphs)		Variable dimensional compositional (elemental ratios, crystal structures)	
Target Property	g-factor (chirality measure, 0.0–2.0)		pChEMBL value (bioactivity, target > 6.5)		Critical temperature T_c (target: 298.15 K)	

1.4 Alternative Domain Considerations

Protein Design. While protein folding and design represent highly active research areas, the multi-scale physics spanning quantum mechanical, molecular dynamics, and thermodynamic effects create substantial challenges for reliable surrogate model development. Current computational approaches require significant resources and often lack the prediction accuracy needed for systematic benchmarking studies.

Metal-Organic Framework (MOF) Synthesis. MOF design involves complex synthesis-structure-property relationships that are difficult to capture in surrogate models due to limited high-quality training data and the strong dependence on synthesis conditions. The gap between computational predictions and experimental realizability remains a significant barrier for benchmarking applications.

Battery Materials Design. While lithium-ion battery optimization is a critical area for energy storage, the complex interplay between electrochemical properties, ion transport dynamics, and material stability across multiple length scales makes surrogate modeling challenging. The strong coupling between electrode materials, electrolyte composition, and interface phenomena requires computationally intensive multi-physics simulations that are difficult to approximate reliably.

Catalyst Discovery. Heterogeneous catalysis represents a major industrial application area, but the reaction mechanisms involve surface chemistry, adsorption energetics, and kinetic pathways that are highly sensitive to atomic-scale details. The computational cost of density functional theory calculations for reaction pathway analysis, combined with the vast combinatorial space of multi-metallic compositions and support materials, limits the feasibility of rapid surrogate model evaluation.

Polymer Design. While polymer materials have diverse applications, the relationship between monomer sequence, processing conditions, and macroscopic properties involves multiple scales from molecular to continuum. The configurational entropy and kinetic effects during polymerization create a complex landscape that current surrogate models struggle to capture with sufficient fidelity for benchmarking purposes.

Due to the complexity and meaningfulness, we seek the optimal balance between scientific relevance and computational tractability. Our selected benchmark scenarios serve as typical examples for advancing scientific discovery with language model systems.

2 Establishing the “Last Record” in Task Description

To ensure rigorous and unbiased evaluation, we established standardized starting points for each benchmark scenario through carefully selected “Last Record” baselines. This approach addresses two critical evaluation challenges: eliminating algorithmic bias from arbitrary initialization and providing consistent comparative benchmarks across different discovery strategies. The Last Record selection follows a principled approach based on:

1. Selection of well-characterized examples from established research.
2. Ensuring sufficient room for improvement to demonstrate discovery capabilities.
3. Choosing configurations that reflect typical starting points in each domain.
4. Using publicly available or easily reproducible initial conditions.

2.1 Domain-Specific Baseline Configurations

2.1.1 Nanohelix Optimization

Task

This team work MUST follow the standard scientific research.

This research task aims to discover a nanohelix with ‘structure parameter’

- expression exhibiting optimal chirality, quantified by maximizing its
- ‘g-factor’. All members operate within a **Hypothesis-Validation** mechanism.

1. The ‘g-factor’ (a value, known ranges from 0.0 to 2.0) is a critical
 - descriptor, representing the chirality of a nanohelix, with higher values
 - indicating stronger chiral effect.
2. The nanohelix discovery process involves exploring diverse chemical space
 - through strategic structural modifications.
3. For each proposed nanohelix structure, Experiment Agent should check its
 - ‘g-factor’.
4. The core objective is to identify the nanohelix with the **highest** predicted
 - ‘g-factor’, effectively pinpointing a candidate with potentially superior
 - chirality.

The outcome will be the identification of the nanohelix structure parameter with

- the best predicted ‘g-factor’.

Definitions

- * **fiber-radius (nm):** Radius of the actual fiber/wire that forms the helix
 - structure. The values for this parameter will range from 20 nm to 60 nm,
 - with **10** evenly spaced values.
- * **helix-radius (nm):** Radius of the helix (distance from the central axis to
 - the center of the helical path). The values for this parameter will range
 - from 20 nm to 90 nm, with **10** evenly spaced values.
- * **n-turns (float):** Number of complete turns in the helix. The values for this
 - parameter will range from 3 to 10, with **10** evenly spaced values.
- * **pitch (nm):** Axial distance between adjacent turns. The values for this
 - parameter will range from 60 nm to 200 nm, with **10** evenly spaced values.

[Important Warning] The Hypothesis scope MUSTN’t out of the definitions below.

- There are many principles/mechanisms from correlation perspectives to be
- tested. When suggest experiments, any slight changes will be strongly
- rejected!

Last record of the experiment

- Parameters: fiber_radius=20.0, helix_radius=20.0, n_turns=10.0 and pitch=60.0
- Property: g-factor=0.5213

The expected outcome

- Property: g-factor>1.8
- Parameters: ? (should be grounded in physicochemical principles)

Rationale. The selected geometric parameters represent a “minimal” helix configuration with small radii and high turn density. This configuration provides multiple optimization vectors: increasing fiber radius can enhance electromagnetic coupling, expanding helix radius affects the chiral volume, reducing turn number may optimize the length-to-chirality ratio, and modifying pitch controls the helical periodicity. The baseline g-factor of 0.5213 indicates measurable chirality while leaving substantial room for enhancement through geometric optimization.

2.1.2 Bio-molecules Optimization

Task

This team work MUST follow the standard scientific research.

This research task aims to discover a molecule with SMILES expression exhibiting

- optimal bio-activity, quantified by maximizing its ‘pChEMBL’ value. All
- members operate within a **Hypothesis-Validation** mechanism.

1. The ‘pChEMBL’ value (our objective maximum is 6.5) is a critical descriptor,
 - representing the biological activity of a molecule, with higher values
 - indicating stronger biological effect.
2. The molecular discovery (optimization) process involves exploring diverse
 - chemical space with rationals.
3. For each proposed molecule structure, Experiment Agent should check its
 - ‘pChEMBL’ value.
4. The core objective is to identify the molecule with the **highest** predicted
 - ‘pChEMBL’ value, effectively pinpointing a candidate with potentially
 - superior bio-activity.

The outcome will be the identification of the nanohelix structure parameter with

- the best predicted ‘pChEMBL’.

Last record of the experiment

- molecule: ‘n1cccc2ccccc12’
- Property: 2.6910

Rationale. The quinoline core “1cccc2ccccc12”, as mentioned in the last record, represents a pharmaceutically relevant heterocyclic system with established bioactivity profiles. This scaffold offers multiple sites for structural elaboration: the benzene ring can accommodate various substituents, the pyridine nitrogen provides hydrogen bonding opportunities, and the fused ring system allows for extension into larger pharmacophores. The pChEMBL value of 2.691 corresponds to moderate bioactivity, providing a realistic starting point for structure-activity relationship exploration.

2.1.3 Superconductor Optimization

Task

This team work MUST follow the standard scientific research.

This research task aims to discover a superconductor with larger critical

- temperature of superconducting materials, quantified by maximizing its ‘Tc’
- value. All members operate within a **Hypothesis-Validation** mechanism.

1. The ‘Tc’ (a value, our objective is 298.15) is a critical descriptor,
 - representing the critical temperature of a superconductor, with higher
 - values indicating stronger chiral effect.
2. The superconductor discovery process involves exploring diverse material space
 - through rationals.
3. For each proposed superconductor structure, Experiment Agent should check its
 - ‘Tc’.
4. The core objective is to identify the superconductor with the **highest**
 - predicted ‘Tc’, effectively pinpointing a candidate with potentially
 - superior Room Temperature Superconductivity.

The outcome will be the identification of the superconductor expression with the

- largest predicted ‘Tc’.

Last record of the experiment

- Candidate: ‘Nb₃Sn’
- Property (T_c): 17.6

Rationale. “Nb₃Sn” serves as an ideal baseline due to its historical significance as one of the first practical high-field superconductors and its well-understood A15 crystal structure. With $T_c = 17.6$ K, it represents established superconducting technology while being far from room temperature targets.

References

- [1] M. K. G. Abbas, A. Rassam, F. Karamshahi, R. Abunora, and M. Abouseada. The role of ai in drug discovery. *ChemBioChem*, 25, 2024.
- [2] Y. Chen and P. Esmaeilzadeh. Generative ai in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26, 2024.
- [3] V. Gallego, R. Naveiro, C. Roca, D. R. Insua, and N. E. Campillo. Ai in drug development: a multidisciplinary perspective. *Molecular Diversity*, 25:1461 – 1479, 2021.
- [4] J. B. Gibson, A. C. Hire, P. M. Dee, O. Barrera, B. Geisler, P. J. Hirschfeld, and R. G. Hennig. Accelerating superconductor discovery through tempered deep learning of the electron-phonon spectral function. *ArXiv*, abs/2401.16611, 2024.
- [5] W. Lan, Z. Tang, M. Liu, Q. Chen, W. Peng, Y.-P. P. Chen, and Y. Pan. The large language models on biomedical data analysis: A survey. *IEEE journal of biomedical and health informatics*, PP, 2025.
- [6] S. Liu, Y. Lu, S. Chen, X. Hu, J. Zhao, T. Fu, and Y. Zhao. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *ArXiv*, abs/2411.15692, 2024.
- [7] C. Luo, T. Sang, Z. Ge, J. Lu, and Y. Wang. Flexible design of chiroptical response of planar chiral metamaterials using deep learning. *Optics express*, 32 8:13978–13985, 2024.
- [8] W. Ma, F. Cheng, and Y. Liu. Deep-learning-enabled on-demand design of chiral metamaterials. *ACS nano*, 12 6:6326–6334, 2018.
- [9] M. K. Tripathi, A. Nath, T. P. Singh, A. S. Ethayathulla, and P. Kaur. Evolving scenario of big data and artificial intelligence (ai) in drug discovery. *Molecular Diversity*, 25:1439 – 1460, 2021.
- [10] D. Viatkin, B. Garcia-Zapirain, A. Méndez-Zorrilla, and M. A. Zakharov. Deep learning approach for prediction of critical temperature of superconductor materials described by chemical formulas. In *Frontiers in Materials*, 2021.
- [11] D. Wines and K. Choudhary. Data-driven design of high pressure hydride superconductors using dft and deep learning. *Materials Futures*, 3, 2023.
- [12] W. Wu, W. Hu, G. Qian, H. Liao, X. Xu, and F. Berto. Mechanical design and multifunctional applications of chiral mechanical metamaterials: A review. *Materials & Design*, 2019.
- [13] M. Yazdani-Asrami. Artificial intelligence, machine learning, deep learning, and big data techniques for the advancements of superconducting technology: a road to smarter and intelligent superconductivity. *Superconductor Science and Technology*, 36, 2023.