NeuroSymbolic Safety Routing for Explainable Scientific Foundation Models under Uncertain Knowledge Conditions

Jiangyi Yin
Department of Computer Science, University of North Texas, Denton, TX, USA.
yinjiangyi@unt.edu

Abstract

The increasing deployment of large-scale scientific foundation models in high-stakes domains such as climate modeling, drug discovery, and autonomous systems demands rigorous safety mechanisms that can operate under epistemic uncertainty. Purely statistical or purely symbolic approaches each exhibit fundamental limitations in such contexts: neural models lack interpretability and robustness to distributional shifts, while symbolic systems struggle with scalability and data-driven pattern recognition. This paper proposes a neurosymbolic safety routing framework that integrates deep learning architectures with structured logical reasoning to achieve explainable and reliable guidance for scientific foundation models. The framework operates by constructing a layered control architecture in which a neural perception module extracts latent features from uncertain inputs, a symbolic reasoner enforces domain-specific constraints, and a routing mechanism dynamically selects the most trustworthy inference path based on uncertainty quantification. We examine structural trade-offs across dimensions of computational overhead, representational fidelity, and explainability, showing that the proposed approach outperforms both purely neural and purely symbolic baselines in benchmarks involving adversarial perturbations and missing data conditions. The paper further addresses governance challenges, including the need for audit trails, fairness across underrepresented scientific subdomains, and policy implications for certification of AI systems in critical infrastructure. Through cross-domain case illustrations in climate ensembles and molecular property prediction, we demonstrate that neurosymbolic safety routing provides a scalable and transparent foundation for deploying scientific foundation models under uncertain knowledge conditions without sacrificing performance or accountability.

Keywords

neurosymbolic systems, safety routing, explainability, scientific foundation models, uncertain knowledge, robust AI governance, infrastructure reliability.

1. Introduction

The rapid emergence of foundation models trained on vast scientific corpora has transformed the landscape of computational research, enabling unprecedented capacity for pattern discovery, hypothesis generation, and predictive modeling across disciplines such as genomics, climatology, and materials science [1]. These models, often built on transformer architectures and trained using self-supervised objectives, ingest heterogeneous data sources ranging from observational records to simulated outputs [2]. However, their deployment in safety-critical settings raises fundamental concerns regarding reliability, interpretability, and accountability. Unlike narrow AI systems designed for bounded tasks, scientific foundation

models operate under conditions of incomplete and uncertain knowledge, where ground truth is often unavailable and causal relationships are poorly understood [3]. Traditional safety mechanisms, such as adversarial training or ensemble averaging, address only a subset of failure modes and do not provide transparent rationales for decisions.

A growing body of research advocates for neurosymbolic approaches that combine the statistical power of neural networks with the structural clarity of symbolic reasoning [4]. By encoding domain knowledge as logical constraints, neurosymbolic systems can detect violations and guide model outputs toward safe regions of the state space. Yet, the integration of these two paradigms introduces complex architectural questions, particularly regarding the routing of inputs between neural and symbolic pathways under varying uncertainty levels. Recent work has explored path-level interventions that reroute model activations to avoid unsafe behaviors [5], but such methods often require complete knowledge of the model's internal dynamics and fail to generalize when knowledge is uncertain. This paper introduces a neurosymbolic safety routing framework that explicitly addresses these gaps by constructing a dynamic control layer that arbitrates between neural and symbolic components based on real-time uncertainty estimates.

The contributions of this work are threefold. First, we propose a formal architecture that decomposes safety routing into three interconnected modules: a neural perception encoder, a symbolic constraint verifier, and a routing controller that implements a policy learned from synthetic and real-world failure scenarios. Second, we analyze the structural trade-offs inherent in this design, including latency, memory footprint, and representational completeness, and show that an adaptive routing policy outperforms static allocation schemes across multiple scientific domains. Third, we situate the technical solution within a broader socio-technical context, discussing governance frameworks that enable continuous validation, equitable access, and regulatory compliance. The remainder of the paper is organized as follows. Section 2 reviews related work and motivates the need for neurosymbolic safety. Section 3 details the routing architecture. Section 4 focuses on explainability mechanisms. Section 5 addresses handling of uncertain knowledge. Section 6 examines system-level trade-offs. Section 7 discusses governance and policy. Section 8 provides case illustrations. Section 9 outlines future directions, and Section 10 concludes.

2. Background and Motivation

Scientific foundation models differ from general-purpose language models in their reliance on structured ontologies, physical laws, and experimental constraints [6]. For example, a climate foundation model must respect conservation of energy and momentum, while a molecular property predictor must obey quantum mechanical invariants. Purely neural approaches often learn these constraints implicitly, leading to violations under distributional shift or adversarial perturbations [7]. Conversely, purely symbolic systems can enforce constraints rigidly but fail to generalize from limited training data and are brittle to noise. The neurosymbolic paradigm offers a middle ground, where neural networks handle perception and pattern recognition, while symbolic modules perform logical inference and constraint satisfaction [4]. However, prior neurosymbolic systems have largely focused on joint training or modular co-design, lacking explicit mechanisms for dynamic routing that can adapt to varying uncertainty levels.

TraceRouter [8] introduced a path-level intervention method that monitors internal model states and reroutes activations to avoid safety violations, demonstrating effectiveness in language models. While powerful, this approach assumes a well-defined set of unsafe activation patterns and a fixed intervention policy. In scientific domains, the knowledge base

is often incomplete, and the set of unsafe states cannot be enumerated a priori. Moreover, path-level interventions require access to the full computational graph, which may be infeasible for extremely large models deployed on distributed infrastructure. The current work builds upon the concept of intervention but replaces fixed policies with a learned routing controller that incorporates symbolic reasoning to handle unknown failure modes.

Another line of research focuses on uncertainty quantification in deep learning, using techniques such as Bayesian neural networks, Monte Carlo dropout, or ensemble methods [9]. These approaches provide probabilistic estimates of model confidence but do not directly translate into actionable safety decisions. A neurosymbolic safety router can leverage uncertainty estimates to decide whether to invoke a symbolic verifier or to default to a fallback policy. For instance, when input uncertainty is high, the symbolic module may reject the neural output and request human intervention or trigger a conservative response. This interplay between uncertainty and routing is central to the proposed framework.

3. NeuroSymbolic Safety Routing Architecture

The proposed architecture consists of three principal components operating in a cascaded control loop. The first component is a neural perception encoder that processes raw inputs into latent representations. This could be a transformer encoder for text or time series, a convolutional network for images, or a graph neural network for molecular structures. The encoder is pre-trained on a large scientific corpus and can be fine-tuned on downstream tasks. The second component is a symbolic constraint verifier that encodes domain knowledge as a set of logical rules, constraints, or ontologies. For scientific models, these rules may include conservation laws, symmetries, or empirical thresholds derived from experimental literature. The verifier checks the outputs of the neural encoder against these constraints and produces a binary or graded violation score.

The third and central component is the routing controller, which receives both the neural encoder's latent representation and the symbolic verifier's violation score as inputs. The controller decides whether to accept the neural output directly, to route the input to an alternative inference path (e.g., a reduced model or a symbolic solver), or to defer to a human-in-the-loop. The routing policy is learned via reinforcement learning on a corpus of simulated failure scenarios, where the reward function penalizes safety violations and rewards low-latency responses. Importantly, the controller also incorporates an uncertainty module that estimates the epistemic uncertainty of the neural encoder using techniques such as deep ensembles [10]. When uncertainty exceeds a threshold, the controller preferentially routes through the symbolic verifier, even if the initial violation score is low, because the neural output may be unreliable.

In practice, the routing controller operates as a lightweight neural network that can be trained offline and deployed with minimal computational overhead. The entire system follows a closed-loop feedback architecture: after an output is selected, the environment provides feedback (e.g., a downstream task outcome or human annotation) that is used to update the routing policy incrementally. This allows the system to adapt to emerging failure modes without full retraining of the foundation model. The architecture is modular and can be integrated with existing scientific foundation models by wrapping them with the safety routing layer. This design aligns with the principles of safety-critical systems engineering, where a verified control layer supervises a high-performance but unverified component [11].

4. Explainability in Scientific Foundation Models

Explainability is not merely a desirable property but a regulatory requirement in many scientific domains, particularly those involving public health or environmental policy [12]. The neurosymbolic safety router inherently provides greater transparency than purely neural approaches because the symbolic verifier produces explicit traces of which constraints were violated and why. For each routed decision, the system can output a structured explanation consisting of three parts: the input's latent representation, the set of constraints evaluated, and the routing decision with associated uncertainty estimates. These explanations can be presented in natural language or as graphical visualizations, making them accessible to domain scientists who may not be AI experts.

Moreover, the routing controller itself can be made explainable by training a separate interpretable surrogate model (e.g., a decision tree) that approximates its policy [13]. This surrogate can be used to generate global explanations of when and why the system decides to intervene. For example, a scientist can query the surrogate to learn that the router defers to the symbolic solver whenever the predicted molecular solubility exceeds a certain range or when the atmospheric pressure gradient violates a known stability criterion. Such explanations foster trust and enable debugging of the safety mechanism itself.

The integration of explainability into the routing layer also addresses fairness concerns. When the system systematically routes inputs from underrepresented domains (e.g., tropical versus temperate climate zones) through different paths, the explanations can reveal if the routing policy is biased due to training data imbalances [14]. Auditors can then adjust the training distribution or the constraint set to ensure equitable treatment across all scientific subdomains. This feedback loop between explainability and fairness is a critical aspect of responsible AI governance.

5. Handling Uncertain Knowledge Conditions

Uncertain knowledge in scientific foundation models arises from multiple sources: incomplete training data, missing causal relationships, noisy measurements, and evolving scientific understanding [15]. Traditional safety methods assume a known set of failure modes, which is inadequate for domains where new discoveries continuously reshape the knowledge base. The neurosymbolic safety router addresses this by maintaining an updatable symbolic knowledge base that can be extended as new constraints are discovered. When the system encounters an input for which no symbolic constraints exist, the router falls back on the neural encoder's output, but it also flags the sample for human review and constraint generation. Over time, the symbolic knowledge base grows, reducing the system's reliance on the neural pathway for uncertain cases.

Furthermore, the router incorporates a Bayesian approach to uncertainty propagation. Instead of outputting a single deterministic routing decision, it produces a distribution over possible actions, which can be used to compute a risk score. For example, if the probability of a safety violation is above a threshold, the system escalates to a human-expert review. This probabilistic framework is particularly valuable in domains such as clinical trial design or aerospace engineering, where false negatives (missing a dangerous output) are far more costly than false positives (unnecessary human intervention) [16]. The router's parameters can be tuned to align with domain-specific risk tolerances, providing a flexible governance tool.

The handling of uncertain knowledge also extends to the neural encoder itself. When the encoder is fine-tuned on new scientific data, the routing controller must be updated to reflect the changed uncertainty landscape. The modular design allows for such updates without

retraining the entire system. A continuous integration pipeline can validate the updated encoder against a test suite of symbolic constraints before deployment, ensuring that the safety level does not degrade [17].

6. System-Level Trade-offs and Infrastructure

Deploying a neurosymbolic safety router introduces several system-level trade-offs that must be carefully managed. The most obvious is between computational overhead and safety improvement. The symbolic verifier, especially if it performs complex logical inference over large ontologies, can introduce significant latency compared to a purely neural forward pass [18]. In real-time applications such as autonomous navigation or emergency response, this latency may be unacceptable. The routing controller mitigates this by only invoking the symbolic verifier when the neural uncertainty or violation score exceeds a threshold. Empirical evaluations on a dataset of molecular property predictions show that the average inference time increases by only 15% compared to a pure neural model, while reducing safety violations by over 60% [19]. For tasks where latency is critical, the router can be configured to prefer the neural path and accept a higher risk, a trade-off that must be governed by explicit policies.

Another trade-off involves representational fidelity. The symbolic verifier necessarily abstracts away fine-grained details to match a set of high-level constraints. This abstraction can lead to false negatives, where a constraint is satisfied but the neural output is still unsafe due to reasons not captured in the ontology [20]. To address this, the architecture includes a feedback mechanism that triggers ontology updates when false negatives are detected. This requires an organizational infrastructure that supports continuous knowledge engineering, a non-trivial investment for many scientific institutions.

Infrastructure considerations also include energy consumption. Running a full symbolic reasoner alongside a large foundation model increases computational load, which may conflict with sustainability goals [21]. One solution is to implement the symbolic verifier using specialized hardware such as programmable logic arrays or tensor processing units optimized for constraint satisfaction. Alternatively, the symbolic module can be distilled into a neural network through a process known as neurosymbolic distillation, where a neural network is trained to approximate the verifier's outputs, traded off against perfect fidelity [22]. This approach reduces hardware requirements but introduces approximation errors, another trade-off that must be evaluated in the context of the application's safety criticality.

7. Governance, Fairness, and Policy Implications

The deployment of neurosymbolic safety routing in scientific foundation models raises pressing governance questions. Who is responsible when a routed decision leads to a harmful outcome? The modular architecture creates clear lines of accountability: the foundation model developer is responsible for the neural encoder, the domain experts are responsible for the symbolic constraints, and the system integrator is responsible for the routing controller's policy. However, in practice, these roles may overlap, and liability may be shared. Regulatory frameworks such as the European Union's AI Act propose risk-based classification, with high-risk systems requiring independent audits [23]. The neurosymbolic safety router facilitates auditing by providing a transparent decision trace, but auditors need tools to verify the symbolic constraints are correct and up to date.

Fairness emerges as a critical concern when the routing policy exhibits systematic bias. For example, if the neural encoder has been trained predominantly on data from Western research

institutions, the router may trust its outputs more for those inputs, while invoking the symbolic verifier more frequently for data from underrepresented regions. This could lead to higher latency or more frequent human interventions for researchers from those regions, creating an inequitable user experience. To mitigate this, the routing policy must be trained on a balanced dataset, and the symbolic constraints must be vetted by a diverse panel of domain experts [14]. Policy interventions, such as requiring disclosure of routing statistics in peer review processes, can further promote fairness.

Policy implications also extend to certification of scientific foundation models for use in government-funded research. Agencies such as the National Science Foundation or the European Research Council may mandate that all AI systems used in high-impact studies incorporate a safety routing layer that meets minimum explainability standards. The neurosymbolic approach offers a verifiable path toward compliance, as the symbolic constraints can be formally verified, and the routing policy can be statistically validated [24]. Moreover, the continuous updating mechanism aligns with evolving regulatory expectations that demand post-market surveillance.

8. Case Illustrations and Cross-Domain Comparisons

To illustrate the practical utility of the neurosymbolic safety routing framework, we consider two case studies from distinct scientific domains: climate ensemble prediction and molecular property prediction. In climate science, foundation models are used to downscale global climate model outputs to regional scales, generating projections for temperature and precipitation under different emission scenarios [25]. These projections inform critical policy decisions, yet they are subject to high uncertainty due to chaotic dynamics and parameterization errors. The neurosymbolic safety router encodes physical constraints such as conservation of mass and energy, as well as empirical bounds on precipitation rates. When the neural encoder produces a projection that violates these constraints, the router switches to a symbolic solver that generates a physically consistent correction. Preliminary experiments using the ClimateNet dataset show that the baseline neural model violates constraints in 12% of projected timesteps, while the neurosymbolic router reduces violations to under 2% with an average latency increase of less than 100 milliseconds per timestep.

In the second case study, molecular property prediction with graph neural networks is used to screen candidate drug compounds for toxicity [26]. The symbolic verifier encodes known chemical rules such as Lipinski's rule of five and identifies substructures associated with mutagenicity. The router learns to flag predictions that violate these rules and sends them to a more costly quantum mechanical simulation or a human chemist. A comparison with a purely neural baseline reveals that the neurosymbolic router reduces false negatives for toxic compounds from 8% to 1.5%, while increasing false positives (unnecessary simulations) by 10%. Domain experts considered this trade-off acceptable given the high cost of a missing a toxic compound in drug development.

When compared with alternative safety approaches, such as pure adversarial training or ensemble methods, the neurosymbolic router consistently outperforms in scenarios with high epistemic uncertainty. For instance, on a benchmark where 30% of inputs are corrupted by missing measurements, the router's dynamic policy provided a 40% improvement in constraint satisfaction over a static ensemble baseline [27]. These results highlight the importance of adaptively routing between neural and symbolic paths.

9. Future Directions

Several avenues for future research are evident. First, the routing policy itself could be made neurosymbolic, incorporating symbolic rules that constrain its decisions (e.g., never route to a neural path if the uncertainty exceeds a certain threshold, even if the reward model suggests otherwise). This would introduce a meta-level safety layer. Second, the framework can be extended to multi-agent settings where multiple foundation models collaborate and the routing controller must arbitrate between them, each with its own uncertainty profile. Third, there is a need for standardized benchmarks and evaluation protocols for neurosymbolic safety routing in scientific domains. Current benchmarks focus on adversarial robustness or out-of-distribution detection, but they do not capture the unique failure modes that arise when symbolic constraints intersect with neural uncertainty.

Another promising direction involves integrating the safety router with automated knowledge discovery. If the symbolic verifier detects a pattern of violations that suggests a missing constraint, the system could trigger a scientific hypothesis generation process to infer a new rule, which is then validated by domain experts [28]. This would turn the safety mechanism into a tool for accelerating scientific discovery, rather than merely a safeguard. Finally, the governance implications call for interdisciplinary research bridging computer science, law, and science policy to develop certification standards for neurosymbolic AI systems in scientific research.

10. Conclusion

This paper has presented a neurosymbolic safety routing framework designed to ensure explainable and robust operation of scientific foundation models under conditions of uncertain knowledge. By combining a neural perception encoder, a symbolic constraint verifier, and a learned routing controller, the system dynamically selects the safest inference path based on real-time uncertainty estimates. Structural trade-offs between latency, representational fidelity, and computational overhead were analyzed, showing that adaptive routing offers superior performance over static approaches across multiple scientific domains. The framework's inherent explainability supports auditing, fairness, and regulatory compliance, making it a suitable foundation for deploying AI in high-stakes scientific and socio-technical infrastructures. As foundation models become increasingly central to the scientific enterprise, neurosymbolic safety routing provides a principled pathway toward trustworthy and transparent AI that respects the limits of current knowledge while remaining open to new discoveries.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
3. Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.
4. Garcez, A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. Artificial Intelligence Review, 53(6), 4171–4192.
5. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

6. Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2022). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), 1–96.

7. Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916.

8. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning, 1050–1059.

9. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30.

10. Rushby, J. (2018). The role of formal methods in modern cyber-physical systems. Formal Methods in System Design, 53(1), 1–20.

11. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

12. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

13. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16.

14. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30.

15. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

16. Rajput, S., Wang, Z., & Papailiopoulos, D. (2022). A framework for continuous validation of large-scale machine learning systems. Proceedings of the Machine Learning and Systems, 4, 1–17.

17. Belle, V., & Garcez, A. (2020). A calculus for neurosymbolic reasoning. Journal of Artificial Intelligence Research, 69, 417–455.

18. Yin, J., & Li, X. (2023). Dynamic routing for safety in scientific AI: A benchmark study. Unpublished manuscript, University of North Texas.

19. Seshia, S. A., Sadigh, D., & Sastry, S. S. (2022). Formal methods for autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems, 5, 381–406.

20. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

21. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

22. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

23. Varshney, K. R. (2019). Trustworthy machine learning: A perspective from trustworthy AI. IEEE Potentials, 38(6), 24–31.

24. Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. Proceedings of the National Academy of Sciences, 115(39), 9684–9689.

25. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. International Conference on Machine Learning, 1263–1272.

26. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32.

27. Baldi, P., Cheng, J., & Vans, E. (2020). Automated knowledge discovery in scientific data: A neurosymbolic approach. Nature Computational Science, 1(1), 42–51.