Dear reviewers, we want to thank you for taking the time to read through the paper and provide thorough feedback! We have adapted the paper to address the concerns that were raised and will detail them here. All changes made in the paper are shown in teal.

The main changes to the manuscript are:

- We narrowed the scope and novelty claims to theory revision for probabilistic-logic-based neurosymbolic models and clarified the propositional scope of NeTheR.
- We expanded the background and related work on theory revision and broader neurosymbolic structure learning.
- We strengthened the methodological explanation of neural concept insertion and the Sharpe-ratio-based selection heuristic, including its limitations and alternatives.
- We expanded the empirical analysis with a breakdown of the contributions of different modification types and a more detailed analysis of the settings in which NeTheR performs less well.
- We extended the discussion of interpretability, theoretical limitations, and possible generalization to first-order logic.

Review 1

Major comments

Comment 1. The paper presents itself as broadly novel in handling imperfect symbolic knowledge and performing theory revision/predicate invention in NeSy, which it isn't at all (at that level of generality). Plenty of prior NeSy work covered that also, starting from old school KBANN work and alike with the NeSy learning/revision cycle (symbolic theory $\rightarrow$ neural revision $\rightarrow$ theory extraction), d-ILP and alike with differentiable structure learning over latent symbolic theories, or LRNNs with structure learning (particularly close with its greedy, accuracy-gain-driven theory revision via gradient descent, already in a FOL/relational setting). Thus the actual, defensible novelty is much narrower: performing theory revision *within the ProbLog/arithmetic circuit paradigm* (DeepProbLog/Scallop style). That's arguably a solid contribution – just scope it respectively, for which abstract and intro call for some honest rewriting... Also related work should could mention some of the above mentioned (and alike) broader NeSy works (or scope them out by acknowledging the prob-prog focus right from the start)

Response. We agree and have rescoped the abstract (p1), introduction (p2-3), and related work (p24) accordingly: NeTheR is now positioned as theory revision for probabilistic-logic-based NeSy specifically, and we discuss KBANN, LRNNs, and differentiable ILP as related but distinct lines of work.

Comment 2. On a related note, it could be argued that the ProbLog-style NeSy paradigm (narrow compositional integration) is a more clean/auditable NeSy methodology (compared to the tighter but "fuzzy" semantics-based integrations/works). Nevertheless, NeTheR's headline move of introducing "neural concepts with no predefined semantics" actually pushes it into that fuzzy NeSy territory that ProbLog-style works like to avoid/disdain. The cleanliness/rigor/interpretability motivation is then quite circular..., and claims like "Gills = True and neural concept" is meaningfully interpretable are largely overstated (authors lists SHAP etc. as remedies, but none is evaluated)

Response. Agreed. We revised Section 2.4 (p6) and the discussion in Section 7 (p25-26) to separate interpretability of the symbolic theory from that of the invented neural concepts, and softened claims accordingly; SHAP-style methods are now presented purely as a potential future direction, not as evidence of interpretability.

Comment 3. Propositional scope not stated clearly/upfront: The paper works *entirely* with propositional theories, but this only becomes clear very subtly/slowly. It should be in the abstract, and very likely the title already! Afaik, "Predicate invention" is an ILP/FOL term – using it for propositional literal insertion will confuse readers from the community. - perhaps "Concept insertion" or "literal invention" could be more accurate, I think? - mention that instead of examples like "autonomous driving" in the abstract, which just add to the confusion/overclaiming...

Response. We now state the propositional scope explicitly in the title, abstract (p1) and introduction (p3) and consistently use "neural concept" rather than "neural predicate" throughout to avoid the ILP terminology clash. We've also clarified that a neural concept is a propositional neural predicate.

Comment 4. Experimental overclaims. The experiments are more of an ablation study than a true benchmark comparison:

- M* is generated with a structure that matches what NeTheR is designed to search over. This implicitly favors NeTheR (i.e., structural bias).
- The And Or, Or And, MLP+NN are essentially constructed foils (i.e., weak baselines). Classical theory revision systems are cited but excluded – the paper could perhaps at least explicitly say why (because they can't handle subsymbolic data? Perhaps discuss whether a hybrid pipeline would be a stronger baseline?)
- If I see correctly, Table 7 shows NeTheR actually hurts performance on some datasets (FMA/Motifs/StarLightCurves) in the uninformed setting; this is glossed over, but understanding when and why NeTheR fails matters (it's fine it does).
- the conclusion (abstract) that it "consistently achieves statistically significant improvements across all settings" is directly contradicted here, isn't it?

Overall, performance claims should thus be toned down, also...

Response. We expanded Section 5 to clarify the purpose of both settings (p17), added a new per-modification-type contribution analysis (Table 5, p22) that explains the observed degradation on the non-image datasets, and now discuss this behaviour explicitly rather than glossing over it (p20-22). We also clarified why classical theory revision systems were excluded (p19) and toned down the claims in the abstract and conclusion.

Minor comments

Comment 5. Neural training versus symbolic modification interaction lacks detail: The Sharpe ratio selection where to insert a neural concept before training it, using structural best/worst-case bounds is nice. The issue is that the greedy symbolic choice made pre-training may not be same post-training, since a trained neural concept changes the residual error landscape.

- The dynamic alpha update helps to calibrate this, which is also nice, but the interaction isn't too analyzed...
- under what conditions is greedy-before-training reliable? This could benefit from more discussion, I think. Very similarly, freezing previously inserted neural concepts is empirically shown to not hurt F1 – that's nice, but it means true end-to-end optimization across all neural concepts doesn't really happen, and the theoretical concern about suboptimality remains open.
- the ablation settles the empirical question, not the theoretical one.

Response. We expanded Section 4.2.2 (p14) to state explicitly that the heuristic is meant to identify where a neural concept is likely beneficial rather than predict its post-training performance, making NeTheR a greedy approximation. We likewise clarified in Section 4.3 (p16-17) that NeTheR follows a greedy block coordinate ascent procedure, that freezing is an empirical choice validated by ablation (Appendix C.2), and flagged the theoretical suboptimality question as open future work.

Comment 6. The Sharpe Ratio choice: The use of Sharpe ratio in this context is nice/fresh but, afaik, Sharpe uses symmetric variance as its risk measure – penalizing large deviations in both directions (including large upside gains). In theory revision, penalizing a modification because it might improve performance a lot (but variably) seems counterproductive, no? - I believe that this is a standard critique of Sharpe in finance and should at least be acknowledged (given its central role in the paper). - there is e.g. Sortino ratio, which penalizes only downside variance, could that be a better fit here?

Response. We added this justification and the Sortino alternative to Section 4.2.2 (p13-14), along with updating the ablation (Appendix C.1) showing that the dynamic Sortino ratio overestimates the value of neural revisions and underperforms relative to Sharpe.

Comment 7. If I see correctly, neural concepts seem to play a surprisingly small role. NeTheR trains on average only 1.5 neural networks during search? Given that neural predicate invention is the headline contribution, this seems somewhat weird, making it appear more like a neural theory "patching" rather than true revision... Is the greedy search algorithm immediately collapsing? Could be connected to the (medium) issue 1 above... Are neural concepts actually driving the gains, or are symbolic literal insertions doing most/all of the work? The paper could e.g., break down the contribution of each modification type.

Response. We added exactly this breakdown (Table 5, Section 5.2, p20-22): symbolic insertions and removals contribute small but reliable gains, while neural insertions contribute larger but more variable gains, particularly strong on image datasets with pretrained backbones.

Comment 8. I believe that d-DNNF compilation can be exponential in circuit size? Were any bottlenecks encountered? How does it scale with more modifications? Relevant for potential actual applicability...

Response. The main computational bottleneck was training the neural networks, which took 5.5min on average; compiling the arithmetic circuit was only a fraction of this time, as it took less than a second.

Comment 9. NeTheR / NeTHeR are used interchangeably throughout (e.g. Figure 1 vs. abstract).

Response. Thank you for noticing, this is fixed.

Review 2

Comment 10. The current framework is restricted to propositional logic circuits. Since the true strength of relational representations is central to NeSy, the discussion in Section 7 should be expanded to provide a clearer outlook on how to handle the combinatorial explosion of the search space when creating new rules in FOL.

Response. We expanded the discussion section (Section 7, p26) accordingly, addressing the combinatorial growth from rule construction and predicate invention and how classical search/pruning techniques (language biases, mode declarations, refinement operators) could be adapted.

Comment 11. To initialize the parameter α for the dynamic Sharpe ratio, Algorithm 1 requires pre-training on k randomly selected modifications. This initialization step accounts for a significant portion of the computational overhead. A discussion exploring lighter initialization strategies would be beneficial.

Response. We added this discussion to Section 5.2 (p19), covering skipping initialization when a prior α estimate is available, and using cheaper proxy models or fewer training epochs as lower-cost alternatives, with the associated calibration/cost trade-off.

Comment 12. Because the invented neural concepts are not assigned predefined semantics, the resulting logic circuit may become harder to interpret, creating a trade-off between interpretability and predictive accuracy compared to a fully symbolic initial model. While post-hoc XAI methods are mentioned, it would be valuable to discuss the potential of incorporating semantic regularization directly into the training process.

Response. We added semantic regularization (contrastive losses, sparsity constraints, concept-bottleneck regularization, disentanglement) as a promising future direction in the interpretability discussion (Section 7, p25-26).

Review 3

Comment 13. Training objective. In the "Training the model" paragraph (Section 4.3), the authors describe the training procedure but do not explicitly state the loss function that is optimized. This should be clearly specified.

Response. Section 4.3 (p16) now explicitly mentions the use of PyTorch's `BCEWithLogitsLoss`.

Comment 14. Figure 4. The performance metric used to construct the critical difference diagrams is not explicitly indicated in the figure caption or the main text. While it can be inferred that the experiments focus on F1-score, this should be stated clearly.

Response. We have updated the figures (p20).

Comment 15. Distance function d(M, M'). How much does this approximation affect the results? Is there evidence that the algorithm frequently terminates before reaching a genuinely optimal revision? This should be discussed.

Response. We have added a paragraph covering theoretical guarantees to Section 7 (p26): the search terminated when reaching the maximum number of iterations in roughly 1-third of the runs, indicating the heuristic often settles at a local optimum without exhaustively exploring the space.

Comment 16. Discussion of results. The manuscript would benefit from a substantially more detailed discussion of the empirical results. The experiments show that NeTheR consistently outperforms competing methods across most datasets, both when the initial model (M') is structurally informed and when it is not. However, the paper largely reports these findings without providing sufficient insight into why the method performs better. A deeper analysis of the factors contributing to the observed improvements would strengthen the paper considerably. For example, it would be interesting to understand whether the gains primarily arise from the flexible placement of neural concepts, the theory-revision strategy itself, the Sharpe-ratio-based selection criterion, or the interaction among these components. In addition, for some datasets and structural information, the compared methods overcome NeTheR, this must be explained to understand under which settings NeTheR is more effective. More generally, I believe the numerical results currently relegated to the appendix should be moved, at least in part, to the main paper and accompanied by a more thorough interpretation. The experimental

section demonstrates the effectiveness of the proposed approach, but the paper would be significantly stronger if it provided a deeper explanation of the observed behavior and performance trends.

Response. Section 5 (p20-22) now includes the per-modification-type analysis (Table 5, p22) attributing gains to the interaction between symbolic revision and neural concept insertion, and explicitly discusses the settings (non-image, from-scratch neural training) where NeTheR underperforms. We have also moved the numerical results from the appendix to tables 2-4 (p21-22) in the main text.

Comment 17. Regarding the bibliography, only NeSy methods based on probabilistic semantic losses are discussed/introduced. A more general overview including also methods based on fuzzy semantic losses would improve the scope of the paper.

Response. We added a paragraph covering Fuzzy Logic Methods to the related work in Section 6 (p25).

Review 4

Comment 18. Abstract and introduction

- Change “achieving the advantages” to “aiming for the advantages”, as it better reflects the paper’s intent.

- Fix the typo “their their”.

- The abstract mentions limiting the number of “high-impact modifications.” Please clarify how the number of modifications is restricted and how “high-impact” is formally measured.

- The claims made in the second paragraph of the Introduction lack references. Please provide adequate literature citations to support these assertions.

- The authors state that neural concepts are represented by symbolic literals. Please clarify this mechanism.

- In Example 1.1, the claim that existing theory revision methods cannot correct the issue related to birds without hollow bones is too strong. While image data undoubtedly makes identification easier (depending on whether the knowledge is implicit or explicit), it is inaccurate to say existing methods are fundamentally incapable of inferring this knowledge. In case the method has been presented with examples of this case, they could modify the current knowledge to achieve the representation space of that case. Please tone down or better justify this claim.

Response. All points addressed: typos fixed, abstract clarified on how modifications are ranked by estimated impact (p1), references added to the introduction (p2), the neural-concept-as-literal mechanism clarified (p3), and Example 1.1 revised to avoid overstating the limitations of purely symbolic methods (p3).

Comment 19. Background and Related Work

- A significant body of modern theory revision literature is missing and must be integrated to properly contextualize this work. For example, Cerna and Cropper (AAAI 2024) focus on a limited number of examples, negation, and predicate invention to modify logical theories, while Morel and Cropper (MLJ, 2023) learn logic programs guided by failures, which is a core principle of theory revision. Concerning structural modifications, similar to the modifications implemented in the logic circutis, Azevedo et al. (MLJ, 2020) and Guimarães et al. (MLJ, 2019) propose modifications to tree-like structures supporting relational knowledge guided by examples. This is related to the logic circuit representation adopted in this study, as generalization and specialization operators are fundamentally related to adding or removing tree branches.

- The background section currently lacks foundational definitions regarding theory revision and predicate invention.

Response. We have added the suggested citations and their relation to NeTheR to related work (Section 6, p24-25), and added a Theory Revision background subsection (Section 2.5, p7) with definitions of theory revision and predicate invention.

Comment 20. Problem Setting/Terminology Alignment: Traditional theory revision explicitly distinguishes between Background Knowledge (assumed to be correct) and the Modifiable Theory (the incorrect or incomplete part open to modification). Please clarify whether the established setting defines M' as encompassing both background knowledge and the modifiable theory, or just the latter.

Response. Clarified in Section 4.1 (p12): the classical setting is recovered by restricting admissible revision locations, treating designated parts of the theory as immutable background knowledge.

Comment 21. Methodology

- Section 4.1 needs to more heavily emphasize and reflect the specific role of neural computations. As currently written, the framework heavily resembles classical, purely symbolic (logic-based) theory revision.

- Step 3 describes "neural concept insertion" via a neural network, but the description is vague enough that it implies any mechanism could be used (e.g., symbolic concepts, alternative ML models). What makes this uniquely neural-symbolic?

- It is unclear why the Sharpe Ratio is uniquely relevant or tied to neural concepts, given that it could easily be applied to purely symbolic methods.

Response. Section 4.1 (p12) now more clearly motivates neural concept insertion as trainable neural predicates learned from subsymbolic data, distinct from purely symbolic modifications. We clarified in Section 4.2 (p13) that the Sharpe ratio is introduced specifically to address the cost of evaluating neural modifications before training, since (unlike symbolic revisions) their effect cannot be assessed without training.

Comment 22. Evaluation and Experiments

- The imperfect initial model M' is artificially generated by removing variables to intentionally lower the F1-score. How are these variables identified for removal? How does this artificial degradation realistically simulate a naturally imperfect model?

- The evaluation lacks comparisons against existing Neural-Symbolic models. Please justify why these were omitted from the baselines, as they are the most direct competitors to the proposed method.

Response. Section 5 (p17) now clarifies the M' generation procedure for both settings (structurally informed via targeted variable removal; structurally uninformed via a decision tree learned from data), and Section 5.2 (p19) clarifies why existing NeSy learning systems and classical theory revision systems are not directly comparable baselines.