Premise Selection for Higher-Order ATP Problems*

Keneni W. Tesema, Jan Jakubův, Martin Suda

Czech Technical University in Prague, Prague, Czech Republic

Automated theorem proving (ATP) in higher-order logic (HOL) is a cornerstone of modern interactive verification systems [1]. Tools like Isabelle [2], Rocq [3], Mizar [4], and Lean [5] employ a hammer tactic to dispatch proof obligations to ATPs such as Vampire [6] and E-prover [7]. A critical bottleneck in this pipeline is premise selection: identifying, from a library of potentially thousands of axioms, the small subset necessary to prove a given conjecture [8]. Presenting an ATP with an unfiltered axiom set dramatically degrades performance, as the search space explodes combinatorially.

We present and experiment with several end-to-end neural network architectures for premise selection on Typed Higher-Order Form (THF) problems. Unlike prior work predominantly targeting first-order logic or untyped formalisms, THF is significantly more expressive, featuring Church-style simple types, lambda abstractions, higher-order quantifiers, and curried function application, calling for representations that faithfully capture this full higher-order structure [9]. We propose two graph neural networks that operate on formula abstract syntax trees (ASTs) modeled as heterogeneous graphs, alongside sequence models that operate on bag-of-symbols clause representations.

Early approaches to premise selection relied on hand-crafted features and statistical ranking. Meng and Paulson [10] introduced MePo, a lightweight relevance filter based on symbol-overlap heuristics that remains a strong baseline within Isabelle. Kühlwein et al. [11] later extended this with MaSh, applying naive Bayes classifiers and k-nearest-neighbour learners that combine multiple predictors into an ensemble. While effective, these methods rely on shallow feature engineering and fail to capture the deep compositional structure of higher-order formulas. Neural networks were first applied to premise selection by Alemi et al. [12], who used convolutional and recurrent networks on character-level formula strings from the Mizar Mathematical Library. Wang et al. [9], Olšák et al. [13] shifted the paradigm toward graph-based representations, constructing deep graph embeddings where formulas are treated as directed acyclic graphs with shared symbol nodes, demonstrating that structural relationships across formulas provide a stronger inductive bias than purely sequential processing. Liu et al. [14] further refined this approach with ARCG-NN, introducing attention-based cross-graph information exchange between conjecture and axiom graphs during encoding, rather than only at the scoring stage. Hoder and Voronkov [15] introduced SInE, an axiom selection algorithm that triggers an axiom only by the least common symbols occurring in it, thereby approximating the idea that more general symbols are used to define less general ones; it is now used internally in Vampire and E-prover.

Bansal et al. [4] provided the first comprehensive deep learning environment for HOL theorem proving, targeting HOL Light with a combination of graph neural networks and sequence models. Magnushammer [16] employs a contrastive transformer trained with InfoNCE loss in a two-stage SELECT+RERANK pipeline. It operates on linearized, token-based formula representations, without explicit structural or type-dependency modeling. Concurrently, Paliwal et al. [17] developed graph representations for HOL that explicitly model term structure, types, and binding. Kogkalidis et al. [18] proposed QUILL, a structure-aware premise selection system for dependent type theory in Agda, employing structured attention and nameless variable representations via de Bruijn indices resolved to positional pointers. Their work underscores the importance of inductive biases reflecting the true grammar of formal objects. Our approach aligns with this structuralist philosophy, targeting the THF dialect used by Isabelle/Sledgehammer.

We implement two graph neural networks and two sequence models for premise selection. All problems are parsed through Vampire's internal TPTP parser.

*Supported by Renaissance Philanthropy's project DEEPER. Corresponding author: keneni.tesema@cvut.cz

Table 1: Summary of premise selection models

Model	Repr.	Symbol ID	Cross-clause	Type-aware	Loss
GNN v1	AST graph (5-node)	vocab emb.	shared sym. nodes	type edges	Focal
GNN v2	AST graph (3-node)	vocab emb.	+5-dim edge feats	type in feat	Focal
Transformer	bag-of-symbols	vocab emb.	self-attention	—	Focal
Contrastive	bag-of-symbols	vocab emb.	self-attn + cond.	BM25 feats	Focal+NCE

Transformer v1: Two-Stage Cross-Clause Attention. Sequence-based models represent each clause as a *multi-hot symbol vector* $\mathbf{f}_c \in \{0,1\}^V$, where V is the symbol vocabulary size and $\mathbf{f}_c[i] = 1$ iff symbol i appears in clause c. This bag-of-symbols representation discards formula structure but directly captures symbol overlap, the strongest known signal for premise relevance [10]. The model linearizes each clause's AST into a bag-of-symbols token sequence and applies a two-stage architecture: a per-clause formula encoder followed by a global cross-clause transformer that allows every axiom to attend jointly to the conjecture and to every other axiom. The same bilinear MLP scoring head is applied.

Transformer v2: Contrastive Transformer. We improve Transformer v1 by adopting InfoNCE contrastive loss training, which outperforms binary cross-entropy for premise selection as demonstrated in Magnushammer [16], and by integrating conjecture-conditioned encoding inspired by ARCG-NN [14].

GNN v1: Five-Node Heterogeneous Graph with SAGEConv. This model was inspired by Olšák et al. [13]'s work for first-order premise selection with shared symbol nodes. We extend this paradigm to full THF by constructing a heterogeneous graph with five node types (symbol, variable, AST operator, type, clause) and over twenty directed edge relations. Symbol nodes are shared globally across all axioms and the conjectures, creating direct cross-formula bridges that propagate symbol co-occurrence signals in a single message-passing hop. The model employs heterogeneous GraphSAGE convolutions [19] with residual connections, layer normalization, and sum aggregation across relation types.

GNN v2: Three-Node Edge-Featured Graph with GATv2Conv. GNN v1's 20+ edge types produce sparse gradient signals, as few edges per relation appear in each batch. Motivated by the observation that richer edge features can compensate for fewer relation types, GNN v2 collapses nodes to three types (term, operator, clause) and introduces five-dimensional edge feature vectors encoding argument role, syntactic direction, logical polarity, normalized AST depth, and an explicit cross-clause sharing flag. GNN v2 uses edge-feature-aware GATv2Conv attention [20], incorporating edge features directly into attention coefficients.

Initial Experiments and Results. All models were trained with Focal Loss [21] on a 26 k-problem training corpus, with a three-level caching strategy (full parse, raw graph dict, serialized HeteroData objects) that eliminates repeated parsing and graph construction across runs. We did preliminary experiments on a holdout set of 9K problems solved by E-prover. The results showed that our best model, GNN-V2 solves 16% more problems than MePo with 512 premises. The most complementary to GNN-V2 is the transformer model, which improves it to 22%.

Conclusion and Future Work. We propose a modular system for neural premise selection over typed higher-order logic, supporting both graph-based and sequence-based architectures. Our graph representations explicitly encode syntactic roles, logical polarity, and higher-order type structure, providing inductive biases that sequential approaches must learn from data. In the near future, we evaluate end-to-end proof rates within the full Vampire pipeline and compare it with SInE and MePo.

References

[1] C. Benzmüller, Higher-order automated theorem provers, All about Proofs, Proof for All, Mathematical Logic and Foundations (2015) 171–214.

[2] M. Wenzel, L. C. Paulson, T. Nipkow, The isabelle framework, in: International Conference on Theorem Proving in Higher Order Logics, Springer, 2008, pp. 33–38.

[3] G. Huet, G. Kahn, C. Paulin-Mohring, The coq proof assistant a tutorial, Rapport Technique 178 (1997) 113.

[4] K. Bansal, S. Loos, M. Rabe, C. Szegedy, S. Wilcox, Holist: An environment for machine learning of higher order logic theorem proving, in: International conference on machine learning, PMLR, 2019, pp. 454–463.

[5] L. De Moura, S. Kong, J. Avigad, F. Van Doorn, J. von Raumer, The lean theorem prover (system description), in: International Conference on Automated Deduction, Springer, 2015, pp. 378–388.

[6] A. Riazanov, A. Voronkov, System description: Vampire 1.0, in: ARW, CEUR Workshop Proceedings, CEUR-WS.org, 2000.

[7] S. Schulz, E–a brainiac theorem prover, Ai Communications 15 (2002) 111–126.

[8] Z. A. Goertzel, J. Jakubuv, C. Kaliszyk, M. Olsák, J. Piepenbrock, J. Urban, The Isabelle ENIGMA, in: 13th International Conference on Interactive Theorem Proving, ITP, 2022, pp. 16:1–16:21.

[9] M. Wang, Y. Tang, J. Wang, J. Deng, Premise selection for theorem proving by deep graph embedding, Advances in neural information processing systems 30 (2017).

[10] J. Meng, L. C. Paulson, Lightweight relevance filtering for machine-generated resolution problems, Journal of Applied Logic 7 (2009) 41–57.

[11] D. Kühlwein, J. C. Blanchette, C. Kaliszyk, J. Urban, Mash: machine learning for sledgehammer, in: International Conference on Interactive Theorem Proving, Springer, 2013, pp. 35–50.

[12] A. A. Alemi, F. Chollet, N. Een, G. Irving, C. Szegedy, J. Urban, Deepmath-deep sequence models for premise selection, arXiv preprint arXiv:1606.04442 (2016).

[13] M. Olšák, C. Kaliszyk, J. Urban, Property invariant embedding for automated reasoning, arXiv preprint arXiv:1911.12073 (2019).

[14] Q. Liu, Y. Xu, X. He, Attention recurrent cross-graph neural network for selecting premises, International Journal of Machine Learning and Cybernetics 13 (2022) 1301–1315.

[15] K. Hoder, A. Voronkov, Sine qua non for large theory reasoning, in: International Conference on Automated Deduction, Springer, 2011, pp. 299–314.

[16] M. Mikuła, S. Tworkowski, S. Antoniak, B. Piotrowski, A. Q. Jiang, J. P. Zhou, C. Szegedy, Ł. Kuciński, P. Miłoś, Y. Wu, Magnushammer: A transformer-based approach to premise selection, arXiv preprint arXiv:2303.04488 (2023).

[17] A. Paliwal, S. Loos, M. Rabe, K. Bansal, C. Szegedy, Graph representations for higher-order logic and theorem proving, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 2967–2974.

[18] K. Kogkalidis, O. Melkonian, J.-P. Bernardy, Learning structure-aware representations of dependent types, Advances in Neural Information Processing Systems 37 (2024) 65095–65118.

[19] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Advances in neural information processing systems 30 (2017).

[20] S. Brody, U. Alon, E. Yahav, How attentive are graph attention networks?, arXiv preprint arXiv:2105.14491 (2021).

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.