Learning-Guided Higher-Order Automated Reasoning for Isabelle/HOL*

Jan Jakubův[1], Cezary Kaliszyk[2], and Martin Suda[1]

[1] Czech Technical University in Prague, Czech Republic
[2] University of Melbourne, Australia

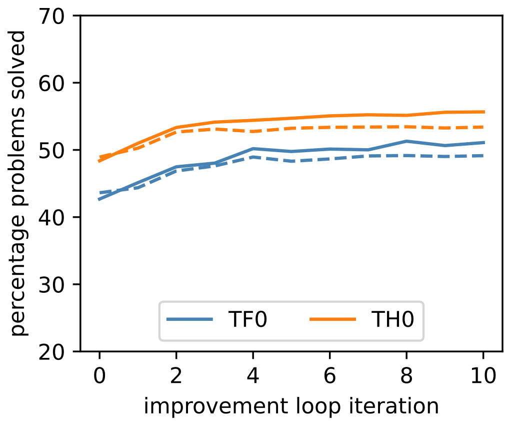
Interactive theorem provers such as Isabelle/HOL [10] rely on *hammer* systems [4] to discharge proof obligations by translating them for external automatic theorem provers (ATPs). Sledgehammer [5] is the primary such interface for Isabelle, and learning has proved effective at two complementary stages: *premise selection* (choosing library facts) and *clause guidance* (directing ATP-internal search). Systems such as ENIGMA [9] and Deepire 2.0 [12] have substantially improved saturation-based proof search on first-order (FO) Isabelle exports. Recent higher-order (HO) support in E [11] (λE) and Vampire [2, 1] opens the door to extending learned guidance to HO problems, staying closer to Isabelle's native logic and reducing the information lost in translation. We introduce λENIGMA and Deepire–2.0–HO—HO extensions of ENIGMA and Deepire 2.0—and provide a preliminary evaluation on a large Isabelle/HOL benchmark. We also improved the Sledgehammer export pipeline by fixing bugs discovered during large-scale experimentation. Updated results will be presented at AITP 2026.

Benchmark. We export proof obligations from Isabelle-2025-2 via Mirabelle and Sledgehammer (MeSh [3], 512 facts per slice) in three TPTP formats: `TF0` (monomorphic FO, for comparison with [7]), `TH0` (monomorphic HO), and `TH1` (polymorphic HO), yielding 243 856 problems per dialect. Constructs `if-then-else`, `let`, and `choice` are disabled because current λE/Vampire does not fully support them. Large-scale export exposed two Sledgehammer bugs: a name clash during η-expansion (corrupting many HOL-Analysis problems) and a name/coname clash in HOL-Nominal's `Class1`. Both are now fixed upstream, improving automation broadly.

λENIGMA. ENIGMA [9, 6] guides given-clause selection in E via gradient-boosted decision trees trained on *vertical* (top-down paths of bounded length) and *horizontal* (head symbol with heads of its arguments) features from clause syntax trees. λENIGMA extends this to λE's HO term representation, where $(\mathsf{F}\,\mathsf{X}\,\mathsf{Y})$ is encoded as $@(\mathsf{F}, \mathsf{X}, \mathsf{Y})$ and λ-abstractions as $\lambda(\mathsf{x}, \mathsf{T})$. For applications, arguments are treated as children of the functional head rather than siblings. For λ's, the bound-variable name is dropped and features record only the binder position. Monomorphic arrow types (e.g., $\mathsf{i} \to (\mathsf{i} \to \mathsf{i}) \to \mathsf{o} \mapsto \mathsf{i\text{-}(i\text{-}i)\text{-}o}$) are appended to symbol names before hashing, letting the model distinguish symbols by type. Name-independent features [8], which replace symbol names with their arities, are also supported, making models robust to symbol renaming. λE currently supports `TH0` only.

Deepire–2.0–HO. Deepire 2.0 [12, 13] extends Vampire with a three-part neural architecture: a GNN computes embeddings over the input CNF's signature and initial clauses, two RvNNs encode each derived clause's term structure and derivation history, and a final network scores each clause for selection. Deepire–2.0–HO adapts the GNN to polymorphic HO by

*Supported by Renaissance Philanthropy's project DEEPER and the Czech Science Foundation, project no. 24-12759S. Corresponding author: jakubuv@gmail.com



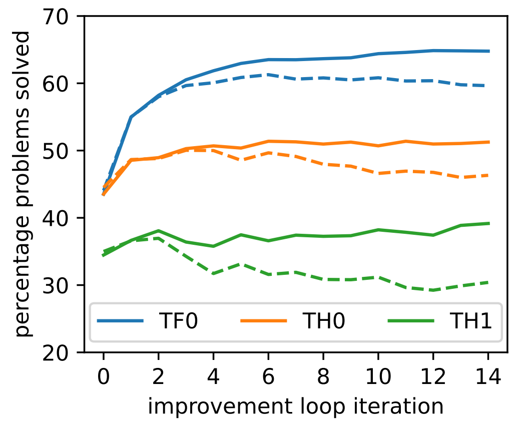


Figure 1: Preliminary performance of λENIGMA (left) and Deepire–2.0–HO (right) on TF0, TH0, and TH1 exports. Solid/dashed lines: train/test. $x = 0$: baseline.

adding type-constructor nodes, promoting sorts to compound expressions with sort variables (as required by rank-1 polymorphism), and marking universal syntactic concepts (arrow type, propositional connectives, FO quantifiers that appear explicitly in Vampire's HO CNF, `apply`, `lambda`, and de Bruijn indices) with distinct features so the GNN recognises their shared meaning across problems. Because HO terms are built from applications and constants only, the term-structure RvNN is simplified to handle a fixed number of arguments, and two HO-specific hand-crafted features (counts of applied variables and lambdas per clause) are added to the final scoring network.

Preliminary Results and Future Work. We sample 30 000 problems for training and hold out 5000 for testing (aligned across dialects), running on dual-socket AMD EPYC 7513 servers (500 GB RAM).

λENIGMA (Fig. 1, left) trains for 10 iterations (∼12 h each, 10 s per problem), selecting models by ATP performance on a 1000-problem development set. Train and test curves remain well correlated, confirming good generalisation. On the test set, λENIGMA gains 7.26 % on TH0 and 11.2 % on TF0 over the respective baseline strategies. Notably, λE alone already outperforms Vampire on TH0.

Deepire–2.0–HO (Fig. 1, right) runs 15 iterations (∼2 days per dialect) with 32 billion CPU instructions per problem. Vampire/TF0 is the strongest configuration, reaching ∼60% test coverage. The HO encodings plateau earlier, suggesting the 30 000-problem training set is too small for good HO generalisation; scaling to the full 243 856-problem dataset is a priority. That the polymorphic TH1 encoding fails to capitalise on its closer proximity to Isabelle's native logic is an open challenge we will investigate in future work.

To assess complementarity we take the best strategy per system and dialect. The two TF0 strategies cover 62.58 % of test problems; the three HO strategies cover 60.92 %; all five together reach 68.50 %, confirming a valuable non-redundant HO contribution. Future work will also focus on cross-dialect premise-selection experiments and scaling training to the full dataset.

Acknowledgments

This work was supported by the Renaissance Philanthropy grant DEEPER and by the Czech Science Foundation project no. 24-12759S.

References

[1] Filip Bártek, Ahmed Bhayat, Robin Coutelier, Márton Hajdú, Matthias Hetzenberger, Petra Hozzová, Laura Kovács, Jakob Rath, Michael Rawson, Giles Reger, Martin Suda, Johannes Schoisswohl, and Andrei Voronkov. The Vampire diary. In Ruzica Piskac and Zvonimir Rakamaric, editors, *Computer Aided Verification - 37th International Conference, CAV 2025, Zagreb, Croatia, July 23-25, 2025, Proceedings, Part III*, LNCS, pages 57–71. Springer, 2025.

[2] Ahmed Bhayat and Martin Suda. A higher-order Vampire (short paper). In Christoph Benzmüller, Marijn J. H. Heule, and Renate A. Schmidt, editors, *Automated Reasoning - 12th International Joint Conference, IJCAR 2024, Nancy, France, July 3-6, 2024, Proceedings, Part I*, LNCS, pages 75–85. Springer, 2024.

[3] Jasmin Christian Blanchette, David Greenaway, Cezary Kaliszyk, Daniel Kühlwein, and Josef Urban. A learning-based fact selector for isabelle/hol. *J. Autom. Reason.*, 57(3):219–244, 2016.

[4] Jasmin Christian Blanchette, Cezary Kaliszyk, Lawrence C. Paulson, and Josef Urban. Hammering towards QED. *J. Formaliz. Reason.*, 9(1):101–148, 2016.

[5] Sascha Böhme and Tobias Nipkow. Sledgehammer: Judgement day. In Jürgen Giesl and Reiner Hähnle, editors, *Automated Reasoning, 5th International Joint Conference, IJCAR 2010, Edinburgh, UK, July 16-19, 2010. Proceedings*, Lecture Notes in Computer Science, pages 107–121. Springer, 2010.

[6] Zarathustra Amadeus Goertzel, Karel Chvalovský, Jan Jakubův, Miroslav Olsák, and Josef Urban. Fast and slow enigmas and parental guidance. In Boris Konev and Giles Reger, editors, *Frontiers of Combining Systems - 13th International Symposium, FroCoS 2021, Birmingham, UK, September 8-10, 2021, Proceedings*, Lecture Notes in Computer Science, pages 173–191. Springer, 2021.

[7] Zarathustra Amadeus Goertzel, Jan Jakubův, Cezary Kaliszyk, Miroslav Olsák, Jelle Piepenbrock, and Josef Urban. The isabelle ENIGMA. In June Andronick and Leonardo de Moura, editors, *13th International Conference on Interactive Theorem Proving, ITP 2022, Haifa, Israel, August 7-10, 2022*, LIPIcs, pages 16:1–16:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[8] Jan Jakubův, Karel Chvalovský, Miroslav Olsák, Bartosz Piotrowski, Martin Suda, and Josef Urban. ENIGMA anonymous: Symbol-independent inference guiding machine (system description). In Nicolas Peltier and Viorica Sofronie-Stokkermans, editors, *Automated Reasoning - 10th International Joint Conference, IJCAR 2020, Paris, France, July 1-4, 2020, Proceedings, Part II*, Lecture Notes in Computer Science, pages 448–463. Springer, 2020.

[9] Jan Jakubův and Josef Urban. ENIGMA: efficient learning-based inference guiding machine. In Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke, editors, *Intelligent Computer Mathematics - 10th International Conference, CICM 2017, Edinburgh, UK, July 17-21, 2017, Proceedings*, Lecture Notes in Computer Science, pages 292–302. Springer, 2017.

[10] Lawrence C. Paulson, Tobias Nipkow, and Makarius Wenzel. From LCF to isabelle/hol. *Formal Aspects Comput.*, 31(6):675–698, 2019.

[11] Stephan Schulz, Simon Cruanes, and Petar Vukmirovic. Faster, higher, stronger: E 2.3. In Pascal Fontaine, editor, *Automated Deduction - CADE 27 - 27th International Conference on Automated Deduction, Natal, Brazil, August 27-30, 2019, Proceedings*, Lecture Notes in Computer Science, pages 495–507. Springer, 2019.

[12] Martin Suda. Deepire II = RL(GNN+2RvNN). In *10th Conference on Artificial Intelligence and Theorem Proving AITP 2025 – proceedings*, 2025. https://aitp-conference.org/2025/abstract/AITP_2025_paper_19.pdf.

[13] Martin Suda. Efficient neural clause-selection reinforcement. In Clark W. Barrett and Uwe Waldmann, editors, *Automated Deduction - CADE 30 - 30th International Conference on Automated Deduction, Stuttgart, Germany, July 28-31, 2025, Proceedings*, LNCS, pages 403–422. Springer, 2025.