capstract:2611.00030v1 [cs.LG] 17 Nov 2026

Predicting Proof Search Cost from Care Labels on Stitched Premise Graphs

Ailsa Verdon-Kato∗
School of Computing
Aldren Institute of Technology
avk@aldren.example

Matías Orellac
Department of Textile Conservation
Instituto de Estudios de Ribera Alta

Irena Solvic-Daher
School of Computing
Aldren Institute of Technology
Faculty of Mathematics
Lestava College of Science

Abstract

A static estimate of proof cost is useful when a collection of goals competes for a fixed search budget. We study the information retained by a garment care label assigned to the premises of a goal. Each premise set is sewn onto linen, with a free span of thread for each occurrence of an antecedent. Three textile conservators examine the reverse of the panel and independently specify how it may be washed, dried and pressed. They receive no logical notation. We fit a linear predictor of capped log expansion count to the resulting labels. On 144 goals from 24 held-out rule families it has a mean absolute error of 0.81 bits, compared with 1.08 for a predictor using syntax counts and 0.68 for a graph kernel with access to the original premises. The paired reduction relative to syntax counts is 0.27 bits (95% family bootstrap interval 0.12 to 0.41). Removing the drying instruction costs 0.14 bits. On a paired set of 48 goals, securing the reverse threads changes 31 drying instructions and leaves the proof search itself invariant. The experiment localizes part of the signal to the unsupported span of an antecedent. Several of the most expensive proofs nevertheless admit unrestricted tumble drying.

1 Introduction

Before a search begins, the solver has a list of rules and a goal. From these it is often possible to tell which jobs are worth admitting to a queue with a short time limit. Counts of constants and rules cost little to obtain. A learned embedding can distinguish more instances but makes it difficult to say which part of the input supplied the estimate. Our interest is in the middle of this range. We want a small description with coordinates whose meanings were settled before anyone measured search cost.

A garment care label meets this last condition. Its instructions summarize the consequences of moving a textile when wet, supporting it while drying, and pressing its surface. These judgments depend in part on what is attached to what. A long thread secured only at its ends is treated differently from the same length distributed over several seams. We encode the dependency of one proposition on another as a span on the reverse of a linen panel. The inspector sees the panel in the condition in which it leaves the embroidery frame. We keep the care label and discard the inspector’s notes for purposes of prediction.

∗Corresponding author: avk@aldren.example

Preprint.

The panel contains only information available in the premise list. A conservator, however, must issue one drying instruction for the whole piece. Several spans that each tolerate hanging can require flat drying when they share one anchor. We ask how much a decision at this scale can reveal about searches in which several paths terminate in a common obstruction.

Finite propositional systems fit on a single frame. As a control we can fasten a premise differently without changing its proofs (Section 4). We repeated inspection after this alteration on a paired sample from the held-out families (Section 6).

2 Previous work

Algorithm selection starts with a description of the instance available before the solver runs (Rice, 1976). Work on empirical hardness models treats the running time of a specified algorithm as the response (Hutter et al., 2014). We use this definition with a count in place of elapsed time. Learned theorem provers can exploit premise and proof structure directly (Bansal et al., 2019; Yang and Deng, 2019). Their output may change the course of search. Our search policy has no input from the cost estimator. This restriction allows us to assign all predictors exactly the same response vector, including the runs that exhaust the budget.

The closest representation to ours in machine learning is a concept bottleneck (Koh et al., 2020). A concept in this experiment has an observer, a physical specimen and an explicit set of allowed answers. The observer can make a mistake about the care of the specimen without making a mistake about the solution of the underlying goal. The effect of substituting measured concepts for inferred ones depends on how the concept measurements were obtained (Shin et al., 2023). We report predictions from measured instructions only. In particular, we do not train a model to imitate the conservator from a photograph and count such an imitation as an independent assessment.

Graph kernels compare neighborhoods in the original structure (Shervashidze et al., 2011). Structural descriptors of logical instances can remain informative where the names of symbols carry little information (Norrel and Oltavi, 2020). Some textile recording systems represent discontinuities and floating threads in similarly local terms (Bervenne, 2014; Ridellan and Kaulvik, 2018). An inventory of floats gives little indication of where the cloth will cease to carry its own weight after wetting. The present inspection follows work in which a conservator must name an acceptable method for the assembled piece before treatment (Carreloux and Aveldi, 2019). Each method is recorded as a separate field. The professional-cleaning field was present in our pilot, but all twelve pilot pieces received the same entry. We removed it before drawing the study families.

3 Proof problems and loss

A problem consists of facts, a set of propositional Horn clauses with one or two antecedents, and a distinguished goal atom. Each atom occupies one of eight depth levels. A clause head belongs to a higher level than either antecedent. A family specifies the number of atoms at each level, the number of rules per head and the available facts at level zero. Within a family we redraw which lower-level atoms supply the rules. All six instances have the same counts of atoms, rules and antecedent slots. Some share a chain of consequences but no two have the same identified rule graph. We reject repeated graphs within a family by canonical comparison with the goal and the facts marked. Families with the same count specification are also rejected.

We draw a chain of support for the goal first and fill the remaining slots from admissible atoms. The support chain can branch at a two-antecedent rule. We require that every leaf of this construction be a fact. This gives every instance a proof and prevents the fabric supplier from being asked to make a separate type of panel for an unprovable goal. The search program receives the resulting clauses in canonical order. It has no record of which ones were present when the chain was drawn. Our population consists of provable problems made by this procedure. The restriction matters since exhausting a finite failed branch can be as expensive as finding the first successful one.

The solver works backward from the first outstanding goal. A state is a sequence of goals still to discharge. Expansion either drops an initial fact or replaces the first goal by a rule body in canonical order. Equal successor sequences are merged, and a queue visits the remaining sequences in order of length, breaking ties by arrival order. An empty successor is a proof. A chain guaranteed by

the construction can be found before the visit limit or can still be in the queue when the limit is reached. We count one expansion per popped nonempty sequence, including sequences with no successors. A cap of 4,096 expansions applies to every search. All costs in the paper, including those recorded for unfinished runs, use this cap.

Let N_i be the number of expansions required for the first proof in an unrestricted run on instance i, and set $B = 4096$. Our loss on a set S is mean absolute error in the log of capped expansions,

$$y_i = \log_2 \min(N_i, B), \qquad L_S(f) = \frac{1}{|S|} \sum_{i \in S} |y_i - f(x_i)|. \tag{1}$$

In Eq. (1) a one-bit absolute residual corresponds to a factor of two in the capped count. One may compute the response without running the uncapped search. A model that predicts the ceiling exactly on an unfinished problem receives zero error there. This treatment estimates resource use under a fixed allowance. Estimating the cost of an uncapped prover from these runs would require a separate model for the upper tail.

We sampled 96 families after checking dimensional limits on the loom. A public integer permutation of family indices with initializer 1729 selected the partitions in Table 1. Each family contributes all six of its instances to one partition. The twelve pilot problems used to choose the form and frame are outside these counts. We chose the test size in units of panels before assigning inspection work. Twenty-four families exhaust the capacity of one test ledger.

Table 1: Allocation and number of searches that used the full allowance. Inspection counts omit repeated assessments of modified panels.

Partition	Families	Panels	Inspections	At cap
Training	60	360	1080	26
Validation	12	72	216	5
Test	24	144	432	11
Total	96	576	1728	42

4 Specimen preparation

The compiler assigns each atom to a hole in a strip of evenweave linen. Columns indicate depth. Within a column the compiler orders atoms by their canonical indices. We designate a small block of neighboring holes at each atom for its incident threads. The knot at the first hole fixes the block’s position on the reverse. The head of each rule receives a thread, taken across the back to the first antecedent, returned along the face in short stitches, then taken across the back to the second antecedent if present. No travel stitch on the face is longer than two grid intervals. A rear thread with no intermediate fastening is a float. A two-antecedent rule therefore has two floats leaving the head. Both are anchored at their remote ends. There is no free tail that an inspector could use to identify the last rule embroidered.

A small buttonhole ring marks the goal. Facts receive a single overcast stitch, recognizable by touch. A border of eight plain stitches on all four sides keeps the area available for gripping independent of the number of atoms. We cut every panel to 18 by 18 cm, including this border. The mapping uses the same 3 mm grid pitch for a short proof as for a long one. Enlarging the shortest graphs to fill the piece would make some short antecedent spans artificially fragile. A head may have at most three rules in the family generator. This limit keeps the attachment points distinct under a single ring. The layout convention for cases at the edge of the grid is given in Appendix A.

For material control we bought four bolts from one production lot and assigned each panel to a bolt in block-random order before sewing. Six machines ran the same stitch file interpreter. Border trimming and removal from the hoop followed the same bench procedure across partitions. The machine operators received files under temporary accession numbers. They could see a graph while embroidering it, but could not look up its expansions. Eight panels failed inspection for a skipped anchor. We made their replacements from the existing files, before any observation of care labels.

A skipped anchor is a missing dependency in the physical graph and cannot be corrected by dropping the problem from the evaluation set.

Inspectors could hold the panel by the edge, turn it over, or run a finger under a float to check whether a lining was needed. Only one piece was on the bench at a time. We mixed partitions within sessions and recoded the specimens for each inspector. Identifiers have no place on the care form.

4.1 Care form and feature vector

Three conservators assessed each piece independently in a room with the dry panels. Each chose one entry per field. Washing ranged from normal machine action through reduced action and hand washing to no immersion. For drying, the alternatives were tumble permitted, line drying and dry flat. For pressing they were direct press, press through a cloth and do not press. The last distinction concerns relief on the face. Drying and pressing restrictions need not agree. The labels use words on the form. The corresponding symbols were stamped onto a duplicate kept with the textile. In what follows, a label for a problem means the three inspectors’ answers, except where we explicitly give the median instruction.

For each field and each possible entry we take the fraction of inspectors choosing that entry. The ten fractions sum to one separately within each field. Unlike a majority vote they retain a split verdict on a piece that will tolerate hanging if supported at two corners. Before fitting the model we omit one reference entry per field, giving seven columns and an intercept. Shrinkage toward zero then favors the reference entries. Centering and scaling the seven columns on the training partition keeps this dependence small. In the repeat experiment the inspectors received the same form, printed from the original supply. They were asked to look at the condition of the specimen before them, irrespective of any earlier disposition.

5 Estimators and inference

For the care estimator we fit ridge regression on the seven fractions (Hoerl and Kennard, 1970). Predictions are clipped to the response range $[0, 12]$. The syntax baseline uses the counts of facts, atoms, rules and occupied depth levels, the number of two-antecedent clauses and the maximum number of clauses at a head. It has the same linear fitting rule and clipping. Both estimators choose a penalty from $\{0.1, 1, 10, 100\}$ by validation absolute error. The optimization target on training data is squared error. We use absolute error for evaluation to limit the influence of a small number of deep searches on model choice. Penalties multiply the squared Euclidean norm of the scaled coefficient vector in a sum of squared residuals. The intercept has no penalty.

The graph estimator takes one vertex per rule and one per atom, retaining the flags for facts and the goal. Labeled edges record the role of each atom in the rule. Repeated neighborhood refinement produces a vector of color counts, and the kernel is the normalized inner product of these vectors accumulated over all rounds (Shervashidze et al., 2011). We fit kernel ridge regression, selecting depth from one through four and penalty from the same four-element grid on a unit-diagonal Gram matrix. Unseen colors contribute to the test vector norm and meet zero training counts. For comparison we fit a random forest to the syntax inputs (Breiman, 2001). This separates a limitation of a linear fit from one of repeatedly supplying six different theorems with the same summary of their rules.

Validation selected a penalty of 1 for care and 10 for syntax. The kernel used three rounds and a penalty of 0.1. All reported test predictions use the fits obtained on training data. Records from a family can resemble one another closely enough to make an item-level split uninformative (Sulcan and Tessier-Vaud, 2021; Vargelin, 2022). We therefore resample 24 families with replacement for 10,000 paired bootstrap replicates of the difference in losses (Efron, 1979; Davison and Hinkley, 1997). We use the 2.5th and 97.5th percentiles. These intervals condition on the training set, the three inspectors and the selected fits. A second sample of professionals is outside the design. Our one prespecified contrast was care against linear syntax. We give descriptive results for the other estimates without interpreting individual intervals as tests.

6 Held-out families

The average difference $L_{\text{test}}(\text{syntax}) - L_{\text{test}}(\text{care})$ was 0.27 bits, with a paired family interval of 0.12 to 0.41. Thus the three instructions together account for a measurable part of the variation that the syntax counts miss. Inspection supplies information on which antecedents occupy the available slots. The full graph still gives the smallest loss in Table 2. A forest with the same syntax input improves on linear syntax by only 0.03 bits. Any estimator that uses just these six inputs must tie all six goals within each family.

Table 2: Prediction on 144 test problems. Rank correlation is pooled Spearman correlation with midranks. All losses use capped expansion counts.

Input and estimator	MAE (bits)	Rank correlation
Training median	1.49	undefined
Syntax, linear	1.08	0.62
Syntax, forest	1.05	0.65
Care form, linear	0.81	0.77
Premise graph, kernel	0.68	0.84

A field-by-field analysis gives most of the credit to the instructions on drying. In Figure 1 we removed each entire field, including its reference entry, before fitting and tuning the smaller model. The largest change was 0.14 bits when drying was omitted. Some of the lost accuracy can be recovered from the washing instruction. On a loose panel the inspector often requests both reduced action and flat drying. A full model can distinguish this case from a well-attached panel whose bulk makes machine action unsuitable. Pressing supplied little extra information on these specimens, which were all made of the same thread.

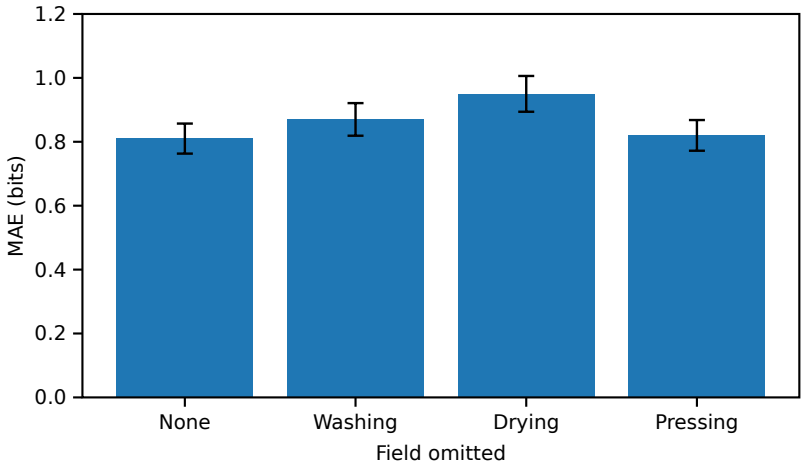


Figure 1: Held-out loss after removing one care instruction. Each point uses a separately fitted model. Error bars are bootstrap standard errors over test families.

The table of verdicts in Figure 2 retains all 144 test cases, including 11 at the cap. The row classification is the median drying verdict across the three inspectors. The four columns divide capped expansions at 128, 512 and 2,048. The order follows from the instructions on the form; it was fixed before we made a cross-tabulation. Four pieces approved for the tumble dryer required at least 2,048 expansions. On two of them the useful rule for the goal occurred late among alternatives, a computational distinction that the order of stitches within a block does not preserve particularly well. All four passed reinspection of their anchors.

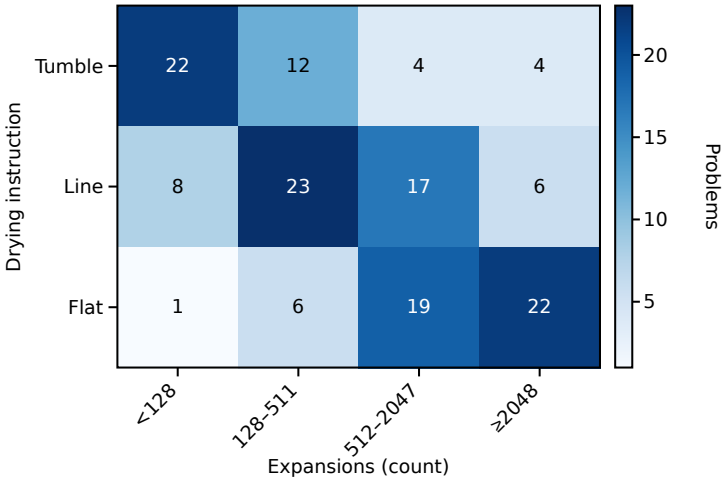


Figure 2: Test problems by median drying instruction and expansion count. Cells contain numbers of problems.

The unit of comparison partly determines the advantage of a sewn representation. The kernel has access to the identity of each shared antecedent. The care form can respond to a concentration of attachments without distinguishing two goals with similar physical disposition. At the other extreme, syntax costs very little to compute and is occasionally adequate for choosing a search budget for a whole family. We calculated the log error in predicting the family mean by averaging the six predictions and the six log outcomes separately. Linear syntax then gave 0.63 bits and the care form gave 0.55. Some of the gap seen in Table 2 is specific to choosing among close relatives.

6.1 Effect of securing a float

We marked two randomly selected positions out of the six in each test family before unsealing the cost ledger. For each of these 48 problems we made a second panel in which long floats were caught down every two grid intervals with a fine thread. The endpoints, thread color and forward stitches were the same as in the original file. The additions have no counterpart in the clauses. They distribute the load from the center of a span to the backing. A clean copy of the original 48 panels was also produced to measure the effects of a second manufacturing and inspection pass. We submitted the resulting 96 pieces as a single mixed batch, with fresh numbers, six weeks after the first assessment. The conservators knew the pieces would be kept as reference material. They did not know which specimens formed pairs.

Relative to the first inspection, 31 secured pieces changed median drying verdict, compared with seven clean copies. All 31 changes on secured pieces allowed a less restrictive drying method. The factory alteration also affected washing instructions, a consequence we could have missed by simply editing one coordinate of the model input. For alteration a and original drying group G, we computed

$$D_{a,G} = |G|^{-1} \sum_{i \in G} \big(f(x_{ia}) - f(x_{i0}) \big), \tag{2}$$

where 0 denotes the first inspection. All fields enter the predictions in Eq. (2). Figure 3 gives the shifts for the three original groups. Subtracting within problem cancels the target in Eq. (1) when calculating signed error. It does not cancel it for absolute error, which we recomputed against the original search ledger.

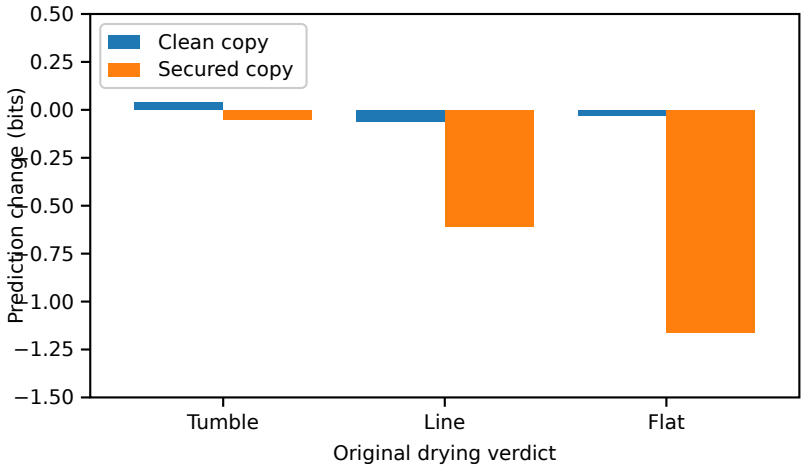


Figure 3: Mean change in predicted log cost after a second manufacture, by original median drying verdict. Each group contains the same proof problems under both treatments.

On these 48 cases the original model had an absolute error of 0.78 bits. The clean copies gave 0.80 and the secured copies 1.10. The mean prediction fell little for the group already permitted to tumble. The largest change was among the panels originally approved for flat drying alone. A change of fastening therefore shifts the numerical prognosis even when the mathematical object is held fixed. It would be possible to retrain on secured panels. Such a fit would answer a different question about which information survives when much of the original span structure can no longer be felt.

7 Resources and limits

We logged 38.4 machine-hours for the first 576 acceptable panels and 6.5 for the two sets of copies. These times run from first to last needle movement and omit trimming, preparation and rejected pieces. Search and model fitting, including the validation grids and resampling, occupied 71 minutes on eight general-purpose CPU cores. No GPU was used. Inspection occupied 27.3 person-hours across both passes, measured between opening and closing each form. An operator interested solely in minimizing computing time could have obtained every target in the study before the first day’s inspection ended. The resource comparison becomes less unequal only where a collection has already been sewn and accessioned.

The three conservators consented to the use of their care decisions and were paid at their ordinary assessment rates. They could return an incomplete form or stop at the end of any session. All forms were complete. We kept accession numbers for manufactured specimens but did not retain video or individual timing data. Agreement estimates pertain to people who handle constructed textiles professionally. A public rating study would introduce a different source of uncertainty. Whether training can reproduce the effect is particularly uncertain for drying, where two members of the panel sometimes recorded different acceptable positions for the same piece.

The specimen alphabet is narrow. Every panel has the same backing, and its longest untreated span is shorter than the side of the frame. We have no estimate for clauses with variables, for cyclic dependency, or for material that contracts at a different rate from the embroidery. A clause whose two antecedents are interchanged gets the same care treatment on most pieces, although it may send a backward search into a long failed branch before a short one. This places a limit on fitting more expressive models to the existing labels. For search policies that reorder antecedents at every step, some of this loss might cease to matter. The cap is a further restriction. Fitting a runtime model to the eleven test problems beyond the allowance would require following them through to completion.

8 Discussion

The smallest useful indication on the form is a horizontal line telling the custodian to lay the piece down to dry. It distinguishes two panels made with the same number of stitches at their heads and the same number of propositions at each depth. It loses much of this ability when the threads are caught to the backing. An archive that enters only the final care disposition of a repaired specimen in its catalog will therefore be mixing two measurement protocols. For the present collection we kept the first label in the proof record and attached a dated second label to the secured copy. A search-cost estimate can be reproduced from the first record even after the corresponding theorem has been made suitable for hanging.

A Construction and inspection record

Within each depth column, a head occupies a rectangular block of holes. The top row holds the ring, if any, and the next three rows hold the heads of candidate rules in canonical body order. We sort antecedents by their level and index before stitching and pass them to the prover in the same order. At a boundary we reserve the entire unused row for the neighboring block. A jump across an empty row is counted at its physical length. A repeated antecedent in the same clause is excluded by the generator. Two different clauses can still require the same pair of antecedents under different heads. This case produces parallel spans close together on the reverse, with separate anchor points identifiable at both ends.

A fact has a stitch even if unreachable from the goal. Removing these facts would require graph analysis before sewing. We kept every fact and every failed branch in its original position.

The verdict data in Figure 2 sum to 42 pieces permitted to tumble, 54 to line dry and 48 to dry flat. The 48 selected positions for the repeat experiment happened to contain 14, 17 and 17 of these verdicts respectively. The individual changes on the secured specimens are given in Table 3. Two specimens moved directly from flat to tumble. They shared an exposed return span across nearly the full width of the occupied columns. Catching down this one span was enough for two inspectors on each piece to approve a tumble cycle. We did not alter which branch the solver tried first.

Table 3: Median drying instruction on 48 specimens before and after the addition of intermediate fastenings. Counts are paired by proof problem.

Original	Secured tumble	Secured line	Secured flat
Tumble	14	0	0
Line	15	2	0
Flat	2	14	1

To assess agreement we calculated the fraction of the 432 inspector pairs on test pieces that gave exactly the same drying answer. Each of the 144 pieces contributes three pairs. The observed agreement was 0.76. A chance-corrected coefficient would depend on the prevalence of flat drying among these specifically constructed pieces (Fleiss, 1971). For the prediction model the fractions matter directly. For example, unanimous flat drying supplies the flat column with one. A flat majority with a single line verdict supplies it with two thirds. Replacement of the fractions by the three field medians gave 0.86 bits after the reduced input was refit. Finite collections of assessors can yield misleadingly precise comparisons if each pair is treated as a new sample (Solanis and Dentrel, 2023). Our bootstrap unit remains the rule family, including all ratings and copies for the selected problem.

B Cost accounting and fit details

The family generator has no global random state shared with the stitch compiler. It draws the choice of first support, followed by all distractor rules, before asking for the first stitch. The accession scheduler reads a completed list of instance identifiers and allocates bolts without access to

targets. During evaluation the row key joining the run ledger to the forms is a canonical graph identifier. We discarded inspection room and operator codes when forming the learning table. Keys derived from the loom schedule have been shown to recover production batches from nominally anonymous disposition records (Tervula and Amendros, 2021). A batch key could improve apparent fit if all the variants of one family were made in sequence.

The counter detects an empty sequence on generation. Success on the last allowed expansion records a proof at the ceiling. The visited set includes sequences already expanded and uses the order of outstanding goals. Sorting those goals could move a fact ahead of a failing branch.

Forest fits used 400 trees, bootstrap sampling of training cases and all six candidate inputs at each split. Validation chose a minimum leaf size of four from the values two, four and eight. Random initialization was fixed at 614 for all forest fits. Other regressors are deterministic after the split. For tied validation scores we select the larger penalty, then the smaller refinement depth. Clipping is applied before computing a validation score, as on test. For the family-mean analysis we first clip each of the six predictions and then take the arithmetic mean on the logarithmic scale. These conventions are sufficient to distinguish a question about one piece from a disposition made for an entire drawer of theorems.

Acknowledgments

We thank the collection technicians for keeping the first and second care records under separate accession numbers.

References

Kshitij Bansal, Sarah Loos, Markus Rabe, Christian Szegedy, and Stewart Wilcox. HOList: An Environment for Machine Learning of Higher Order Logic Theorem Proving. In *International Conference on Machine Learning*, 2019.

Corine Bervenne. Recording discontinuities in reverse-worked panels. *Textile Register Quarterly*, 2014.

Leo Breiman. Random Forests. *Machine Learning*, 45, 2001.

Denis Carreloux and Satine Aveldi. Agreement in the assessment of flat storage and drying support. *Bulletin of Applied Textile Records*, 2019.

Anthony C. Davison and David V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7, 1979.

Joseph L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76, 1971.

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12, 1970.

Frank Hutter, Lin Xu, Holger H. Hoos, and Kevin Leyton-Brown. Algorithm Runtime Prediction: Methods and Evaluation. *Artificial Intelligence*, 206, 2014.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *International Conference on Machine Learning*, 2020.

Hede Norrel and Renato Oltavi. Symbol renaming in learned estimates of clause-set difficulty. *Transactions on Empirical Deduction*, 2020.

John R. Rice. The Algorithm Selection Problem. In *Advances in Computers*, volume 15, 1976.

Nerea Ridellan and Ester Kaulvik. Float inventories for compound machine embroidery. *Studies of Structural Needlework*, 2018.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-Lehman Graph Kernels. *Journal of Machine Learning Research*, 12, 2011.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A Closer Look at the Intervention Procedure of Concept Bottleneck Models. In *International Conference on Machine Learning*, 2023.

Veer Solanis and Imogen Dentrel. Repeated experts in the evaluation of judgment-based predictors. *Annals of Computational Measurement*, 2023.

Mirja Sulcan and Benoit Tessier-Vaud. Holding out generator families in small reasoning corpora. *Transactions on Empirical Deduction*, 2021.

Siiri Tervula and Luc Amendros. Batch effects in the indexing of machine-worked sample books. *Textile Register Quarterly*, 2021.

Dana Vargelin. Dependence among relabeled items in algorithm tests. *Annals of Computational Measurement*, 2022.

Kaiyu Yang and Jia Deng. Learning to Prove Theorems via Interacting with Proof Assistants. In *International Conference on Machine Learning*, 2019.