

Efficient Rule Induction by Ignoring Pointless Rules

Andrew Cropper^{1,2*}, David M. Cerna^{3,4*}

¹ELLIS Institute Finland

²University of Helsinki

³Dynatrace Research

⁴Czech Academy of Sciences Institute of Computer Science

andrew.cropper@helsinki.fi, david.cerna@dynatrace.com, dcerne@cs.cas.cz

Abstract

The goal of inductive logic programming (ILP) is to find a set of logical rules that generalises training examples and background knowledge. We introduce an ILP approach that identifies *pointless* rules. A rule is pointless if it contains a redundant literal or cannot discriminate against negative examples. We show that ignoring pointless rules allows an ILP system to soundly prune the hypothesis space. Our experiments on multiple domains, including visual reasoning and game playing, show that our approach can reduce learning times by 99% whilst maintaining predictive accuracies.

Code — <https://github.com/logicand-learning-lab/aaai26-implications>

Extended version — <https://arxiv.org/pdf/2502.01232>

1 Introduction

The goal of inductive logic programming (ILP) is to induce a hypothesis (a set of logical rules) that generalises training examples and background knowledge (BK) (Muggleton 1991; Cropper and Dumančić 2022).

To illustrate ILP, suppose we have BK with the relations *succ/2*, *lt/2*, *gt/1*, *int/1*, *even/1*, *odd/1* and the following positive (E^+) and negative (E^-) examples:

$$E^+ = \{f(5), f(7)\}$$

$$E^- = \{f(2), f(3), f(4), f(6), f(8), f(9)\}$$

Given this input, we might want to learn a rule such as:

$$r_1 = f(A) \leftarrow \text{odd}(A), \text{gt}(A,3), \text{lt}(A,8)$$

This rule says that $f(A)$ is true if A is odd, greater than 3, and less than 8.

An ILP learner tests rules on the examples and BK and uses the outcome to guide the search. For instance, suppose a learner tests the rule:

$$f(A) \leftarrow \text{even}(A)$$

This rule is too specific because it does not entail any positive example. Therefore, a learner can ignore its specialisations, such as:

$$f(A) \leftarrow \text{even}(A), \text{gt}(A,2)$$

Similarly, suppose a learner tests the rule:

$$r_2 = f(A) \leftarrow \text{odd}(A), \text{int}(A)$$

This rule is too general because it entails two negative examples ($f(3)$ and $f(9)$). Therefore, a learner can ignore its generalisations. A learner will not necessarily ignore its specialisations because one could still be helpful, such as:

$$r_3 = f(A) \leftarrow \text{odd}(A), \text{int}(A), \text{gt}(A,3), \text{lt}(A,8)$$

This rule entails all the positive and none of the negative examples.

However, r_2 and r_3 are pointless because $\text{odd}(A)$ implies $\text{int}(A)$, so we can remove $\text{int}(A)$ whilst preserving semantics. We call such rules *reducible* rules.

Now consider the rule:

$$f(A) \leftarrow \text{lt}(A,B)$$

This rule entails all the positive and all the negative examples, so a learner can ignore its generalisations. However, although it entails all the negative examples, we cannot ignore its specialisations as one might still be a good rule, such as r_1 .

Now consider the specialisation:

$$f(A) \leftarrow \text{lt}(A,10)$$

This rule entails all the positive and negative examples, so a learner can ignore its generalisations. However, unlike the previous rule, a learner can ignore its specialisations. The reason is that the literal $\text{lt}(A,10)$ implies all the negative examples, i.e. there is no negative example greater than 10. Moreover, this literal cannot be further specialised (besides grounding the variable A and thus disconnecting the head literal from the body). We call such rules *indiscriminate* rules.

The novelty of this paper is to show that ignoring reducible and indiscriminate rules allows us to efficiently and soundly prune the hypothesis space and thus improve learning performance. Although existing approaches find reasons why a rule fails on examples, such as having erroneous literals (Muggleton, Lin, and Tamaddoni-Nezhad 2015; Cropper and Hocquette 2024) or rules (Shapiro 1983; Raghathan et al. 2020), they do not find reducible or indiscriminate rules.

We demonstrate our idea in the new ILP system REDUCER. This system builds on POPPER (Cropper and Morel 2021; Hocquette et al. 2024), which can learn optimal and recursive hypotheses from noisy data. The key novelty of REDUCER is identifying pointless rules and building constraints from them to prune the hypothesis space.

*These authors contributed equally.

Novelty and Contributions The novelty of this paper is *identifying pointless rules and building constraints from them to prune the hypothesis space of an ILP system*. The impact, which we demonstrate on multiple domains, is vastly reduced learning times. Overall, we contribute the following:

- We define two types of pointless rules: *reducible* and *indiscriminate*. We show that specialisations of reducible rules are also reducible (Proposition 1). Likewise, we show that specialisations of indiscriminate rules are also indiscriminate (Proposition 3). Finally, we show that a hypothesis with a reducible or indiscriminate rule is not optimal (Propositions 2 and 4).
- We introduce REDUCER, an ILP system which identifies pointless rules in hypotheses, including recursive hypotheses, and builds constraints from them to prune the hypothesis space. We prove that REDUCER always learns an optimal hypothesis if one exists (Theorem 1).
- We experimentally show on multiple domains, including visual reasoning and game playing, that our approach can reduce learning times by 99% whilst maintaining high predictive accuracies.

2 Related Work

ILP. Many ILP systems use bottom clauses (Muggleton 1995; Srinivasan 2001) or variants (Ray 2009) to restrict the hypothesis space. Building bottom clauses can be expensive (Muggleton 1995). In the worst case, a learner needs a bottom clause for every positive example. Bottom clause approaches struggle to learn hypotheses with recursion or predicate invention. REDUCER does not use bottom clauses.

Redundancy in ILP. Fonseca et al. (2004) define *self-redundant* clauses similar to reducible rules (Definition 8) but cannot guarantee that specialisations are redundant (unlike Proposition 1) and require users to provide redundancy information. Raedt and Ramon (2004) check redundancy before testing but require *anti-monotonic* constraints (holding for generalisations, not specialisations) and do not identify implications between literals. By contrast, REDUCER finds implications to prune specialisations. Zeng, Patel, and Page (2014) prune syntactic redundancy (e.g. duplicate variables), whereas we detect semantic redundancy. Srinivasan and Kothari (2005) compress bottom clauses statistically; we avoid bottom clauses.

Rule selection. Many systems (Corapi, Russo, and Lupu 2011; Kaminski, Eiter, and Inoue 2019; Raghothaman et al. 2020; Bembenek, Greenberg, and Chong 2023; Law 2023) precompute every possible rule in the hypothesis space and then search for a subset that generalises the examples. Because they precompute all possible rules, they cannot learn rules with many literals and can build pointless rules. For instance, ILASP4 (Law 2023) will precompute all rules with *int(A)* and *even(A)* in the body. By contrast, REDUCER builds constraints from pointless rules to restrict rule generation. If REDUCER sees a rule containing *int(A)* and *even(A)*, it identifies that *even(A)* implies *int(A)* and henceforth never builds a rule with both literals.

Constraints. Many recent ILP systems frame the ILP problem as a constraint satisfaction problem (Corapi, Russo, and

Lupu 2011; Ahlgren and Yuen 2013; Schüller and Benz 2018; Kaminski, Eiter, and Inoue 2019; Bembenek, Greenberg, and Chong 2023). MUSPER (Cropper and Hocquette 2024) finds minimal sub-hypotheses that can never be true and builds constraints from them to prune the hypothesis space. For instance, given the rule $f(A) \leftarrow succ(A,B), odd(B), even(B)$, MUSPER can identify that a number cannot be both odd and even and then prunes the hypothesis space accordingly. By contrast, REDUCER finds pointless rules with redundancy.

Rule induction. ILP approaches induce rules from data, similar to rule learning methods (Fürnkranz and Kliegr 2015) such as AMIE+ (Galárraga et al. 2015) and RDFRules (Zeman, Kliegr, and Svátek 2021). Most rule-mining methods are limited to unary and binary relations, require facts as input, and operate under an open-world assumption. By contrast, REDUCER operates under a closed-world assumption, supports arbitrary-arity relations, and can learn from definite programs as background knowledge.

Redundancy in AI. Theorem proving has preprocessing techniques for eliminating redundancies (Hoder et al. 2012; Khasidashvili and Korovin 2016; Vukmirovic, Blanchette, and Heule 2023). In SAT, redundancy elimination techniques, such as blocked clause elimination (Kullmann 1999), play an integral role in modern solvers (Heule, Järvisalo, and Biere 2010; Barnett, Cerna, and Biere 2020; Biere, Järvisalo, and Kiesl 2021). Similar to this paper, SAT redundancy elimination identifies clauses containing literals that always resolve to tautologies.

3 Problem Setting

We assume familiarity with logic programming (Lloyd 2012) but have included summaries in the supplementary material. For clarity, we define some key terms. A rule r is a definite clause of the form $h \leftarrow p_1, \dots, p_n$ where h, p_1, \dots, p_n are literals, $head(r) = h$, and $body(r) = \{p_1, \dots, p_n\}$. A *definite program* is a set of definite clauses. We use the term *hypothesis* interchangeably with a definite program. We denote the set of variables in a literal l as $vars(l)$. The variables of a rule r , denoted $vars(r)$, is defined as $vars(head(r)) \cup \bigcup_{p \in body(r)} vars(p)$. Given a rule r and a set of literals C , by $r \cup C$ we denote the rule r' such that $head(r') = head(r)$ and $body(r') = body(r) \cup C$. A *specialisation* of a rule r is a rule r' that is θ -subsumptively more specific than r , where $\theta = \{v_1/t_1, \dots, v_n/t_n\}$ is a substitution simultaneously replaces each variable v_i by term t_i (See (Plotkin 1971) and the Appendix). We focus on a restricted form of θ -subsumption which only considers whether the bodies of two rules with the same head literal are contained in one another, i.e., no variable renaming. We formally define this restriction through a *subrule* relation:

Definition 1 (Subrule). Let r_1 and r_2 be rules. Then r_1 is a *subrule* of r_2 , denoted $r_1 \subseteq r_2$, if $head(r_1) = head(r_2)$ and $body(r_1) \subseteq body(r_2)$.

We generalise the subrule relation to a *sub-hypothesis* relation. Unlike the subrule relation, the sub-hypothesis relation is not a restriction of θ -subsumption:

Definition 2 (Sub-hypothesis). Let h_1 and h_2 be hypotheses and for all $r_1 \in h_1$ there exists $r_2 \in h_2$ such that $r_1 \subseteq r_2$. Then h_1 is a *sub-hypothesis* of h_2 , denoted $h_1 \subseteq h_2$.

The sub-hypothesis relation captures a particular type of hypothesis that we refer to as *basic*. These are hypotheses for which specific rules do not occur as part of a recursive predicate definition:

Definition 3 (Basic). Let h be a hypothesis, r_1 a rule in h , and for all r_2 in h , the head symbol of r_1 does not occur in a body literal of r_2 . Then r_1 is *basic* in h .

As we show in Section 3, under certain conditions we can prune hypotheses containing sub-hypotheses that are *basic* with respect to a contained rule.

Inductive Logic Programming

We formulate our approach in the ILP learning from entailment setting (De Raedt 2008). We define an ILP input:

Definition 4 (ILP input). An ILP input is a tuple (E, B, \mathcal{H}) where $E = (E^+, E^-)$ is a pair of sets of ground atoms denoting positive (E^+) and negative (E^-) examples, B is background knowledge, and \mathcal{H} is a hypothesis space, i.e., a set of possible hypotheses.

We restrict hypotheses and background knowledge to definite programs with the least Herbrand model semantics.

We define a cost function:

Definition 5 (Cost function). Given an ILP input (E, B, \mathcal{H}) , a cost function $cost_{E,B} : \mathcal{H} \mapsto \mathbb{N}$ assigns a numerical cost to each hypothesis in \mathcal{H} .

Given an ILP input and a cost function $cost_{E,B}$, we define an *optimal* hypothesis:

Definition 6 (Optimal hypothesis). Given an ILP input (E, B, \mathcal{H}) and a cost function $cost_{E,B}$, a hypothesis $h \in \mathcal{H}$ is *optimal* with respect to $cost_{E,B}$ when $\forall h' \in \mathcal{H}, cost_{E,B}(h) \leq cost_{E,B}(h')$.

We use a cost function that first minimises misclassified training examples, then minimises the number of literals in a hypothesis. False positives are negative examples entailed by $h \cup B$. False negatives are positive examples not entailed by $h \cup B$. We denote these as $fp_{E,B}(h)$ and $fn_{E,B}(h)$ respectively. We define $size : \mathcal{H} \rightarrow \mathbb{N}$ as the number of literals in $h \in \mathcal{H}$. We use the cost function $cost_{E,B}(h) = (fp_{E,B}(h) + fn_{E,B}(h), size(h))$.

Pointless Rules

We want to find rules that cannot be in an optimal hypothesis. We focus on *reducible* and *indiscriminate* rules.

A *reducible* rule contains a body literal that is implied by other body literals. For example, consider the rules:

$$\begin{aligned} r_1 &= h \leftarrow odd(A), int(A) \\ r_2 &= h \leftarrow odd(A) \end{aligned}$$

The rule r_1 is reducible because $odd(A)$ implies $int(A)$. Therefore, r_1 is logically equivalent r_2 .

As a second example, consider the rule:

$$h \leftarrow gt(A,B), gt(B,C), gt(A,C)$$

This rule is reducible because the relation $gt/2$ is transitive, i.e. $gt(A,B)$ and $gt(B,C)$ imply $gt(A,C)$.

Because it contains a redundant literal, a reducible rule cannot be in an optimal hypothesis. However, a specialisation of a reducible rule could be in an optimal hypothesis. For instance, consider the rule:

$$r_1 = h \leftarrow member(L,X), member(L,Y)$$

In this rule, $member(L,X)$ implies $member(L,Y)$ and vice-versa, so one of the literals is redundant. However, we could still specialise this rule as:

$$r_2 = h \leftarrow member(L,X), member(L,Y), gt(X,Y)$$

Rules r_1 and r_2 are not logically equivalent, and r_2 could be in an optimal hypothesis.

A key contribution of this paper is to identify reducible rules where we can prune all their specialisations. The idea is to identify a redundant *captured* literal. A captured literal is one where all of its variables appear elsewhere in the rule. For instance, consider the rule:

$$h \leftarrow succ(A,B), succ(B,C), gt(C,A), gt(C,D)$$

In this rule, the literal $gt(C,A)$ is captured because all its variables appear elsewhere in the rule. By contrast, the literal $gt(C,D)$ is not captured because the variable D does not appear elsewhere in the rule.

We define a captured literal:

Definition 7 (Captured literal). Let r be a rule, $l \in body(r)$, and $vars(l) \subseteq vars(body(r) \setminus \{l\}) \cup vars(head(r))$. Then l is *r-captured*.

If a literal is captured in a rule then it is captured in its specialisations:

Lemma 1. Let r_1 and r_2 be rules such that $r_2 \subseteq r_1$, $l \in body(r_2)$, and l be r_2 -captured. Then l is r_1 -captured.

Proof. Follows from Definition 1 as the subrule relation preserves variable occurrence. \square

We define a reducible rule:

Definition 8 (Reducible). Let r be a rule, B be BK, $l \in body(r)$ be r -captured, and $B \models (body(r) \setminus \{l\}) \rightarrow l$. Then r is *reducible*.

Some specialisations of a reducible rule are reducible:

Proposition 1 (Reducible specialisations). Let B be BK, r_1 be a reducible rule, and $r_1 \subseteq r_2$. Then r_2 is reducible.

Proof. Let l be a r_1 -captured literal and $B \models (body(r_1) \setminus \{l\}) \rightarrow l$. By Lemma 1, l is also r_2 -captured. Let $C = body(r_2) \setminus body(r_1)$. Then, $(body(r_1) \setminus \{l\}) \rightarrow l$ is subsumptively more general than $(body(r_1 \cup C) \setminus \{l\}) \rightarrow l$, where $r_2 = r_1 \cup C$, i.e., if the former holds for r_1 than the latter holds for r_2 . Thus, r_2 is reducible. \square

Certain hypotheses that contain a sub-hypothesis with reducible rules are not optimal:

Proposition 2 (Reducible soundness). Let B be BK, h_1 be a hypothesis, $h_2 \subseteq h_1$, r_1 be basic rule in h_1 , $r_2 \in h_2$, $r_2 \subseteq r_1$, and r_2 be reducible with respect to B . Then h_1 is not optimal.

Proof. By Proposition 1, r_1 is also reducible implying that there exists an $l \in \text{body}(r_1)$ and rule $r_3 \subset r_1$ such that (i) $B \models (\text{body}(r_1) \setminus \{l\}) \rightarrow l$, (ii) $r_1 = r_3 \cup \{l\}$, and (ii) $|r_3| < |r_1|$. Furthermore, r_3 is basic given that r_1 is basic. Let $h_3 = (h_1 \setminus \{r_1\}) \cup \{r_3\}$. Then $\text{cost}_{E,B}(h_3) < \text{cost}_{E,B}(h_1)$, i.e. h_1 is not optimal. \square

This proposition implies that if we find a reducible rule, we can ignore hypotheses that include this rule or its specialisations.

We introduce *indiscriminate* rules, a weakening of reducible rules. To motivate them, consider the rule:

$$f(A) \leftarrow \text{odd}(A), \text{lt}(A, 10)$$

This rule is not reducible because $\text{odd}(A)$ does not imply $\text{lt}(A, 10)$, nor does $\text{lt}(A, 10)$ imply $\text{odd}(A)$. However, suppose we have the negative examples $E^- = \{f(1), f(2), f(3)\}$. For these examples, the literal $\text{lt}(A, 10)$ implies all the negative examples. In other words, there is no negative example greater than 10. Therefore, this literal (and thus this rule) is pointless because it cannot discriminate against the negative examples. We formalise this notion of an indiscriminate rule:

Definition 9 (Indiscriminate). Let r be a rule, B be BK, E^- be negative examples with the same predicate symbol as the head of r , $l \in \text{body}(r)$ be r -captured, and for all $e \in E^-$, $B \models (\text{body}(r) \setminus \{l\})\theta_e \rightarrow l\theta_e$ where θ_e is a substitution with domain $\text{vars}(\text{head}(r))$ such that $\text{head}(r)\theta_e = e$. Then r is *indiscriminate*.

Under certain conditions, specialisations of an indiscriminate rule are indiscriminate:

Proposition 3 (Indiscriminate specialisations). Let B be BK, r_1 be an indiscriminate rule, $r_1 \subseteq r_2$, and E^- be negative examples with the same predicate symbol as the head of r_1 . Then r_2 is indiscriminate.

Proof. Let l be an r_1 -captured literal and for all $e \in E^-$, $B \models (\text{body}(r_1) \setminus \{l\})\theta_e^{r_1} \rightarrow l\theta_e^{r_1}$. By Lemma 1, l is also r_2 -captured. Let $C = \text{body}(r_2) \setminus \text{body}(r_1)$. We deduce the following for all $e \in E^-$: $(\text{body}(r_1) \setminus \{l\})\theta_e^{r_1} \rightarrow l\theta_e^{r_1}$ is subsumptively more general than $(\text{body}(r_1 \cup C) \setminus \{l\})\theta_e^{r_2} \rightarrow l\theta_e^{r_2}$ where $r_2 = (r_1 \cup C)$, i.e., if the former holds for r_1 , then the latter holds for r_2 . Thus, r_2 is indiscriminate. \square

As with reducible rules, some hypotheses with an indiscriminate rule are not optimal:

Proposition 4 (Indiscriminate soundness). Let B be BK, E^- be negative examples, h_1 be a hypothesis, r_1 be a basic rule in h_1 , $h_2 \subseteq h_1$, $r_2 \in h_2$, $r_2 \subseteq r_1$, and r_2 be indiscriminate with respect to B and E^- . Then h_1 is not optimal.

Proof. By Proposition 3, r_1 is also indiscriminate implying that there exists $l \in \text{body}(r_1)$ and $r_3 \subset r_1$ such that (i) for all $e \in E^-$, $B \models (\text{body}(r_1) \setminus \{l\})\theta_e^{r_1} \rightarrow l\theta_e^{r_1}$, (ii) $r_1 = r_3 \cup \{l\}$, and (iii) $|r_3| < |r_1|$. Furthermore, r_3 is basic given that r_1 is basic. Let $h_3 = (h_1 \setminus \{r_1\}) \cup \{r_3\}$. Then $\text{cost}_{E,B}(h_3) < \text{cost}_{E,B}(h_1)$, i.e. h_1 is not optimal. \square

A hypothesis with a reducible or indiscriminate rule is not optimal:

Definition 10 (Pointless). Let (E, B, \mathcal{H}) be ILP input. A hypothesis $h \in \mathcal{H}$ is *pointless* if there exists $r \in h$ such that r is reducible with respect to B or indiscriminate with respect to B and E^- .

Corollary 1. A pointless hypothesis is not optimal.

In Section 4, we introduce REDUCER, which identifies pointless rules and prunes them from the hypothesis space.

4 Algorithm

We now describe REDUCER (Algorithm 1), which builds on POPPER by identifying pointless rules and using them to prune the hypothesis space. REDUCER takes as input background knowledge (bk), positive (pos) and negative (neg) training examples, and a maximum hypothesis size (max_size). REDUCER follows POPPER and uses a generate, test, and constrain loop to find an optimal hypothesis (Definition 6). REDUCER starts with an answer set program \mathcal{P} . The (stable) models of \mathcal{P} correspond to hypotheses (definite programs) and represent the hypothesis space. In other words, \mathcal{P} encodes in ASP the search for a syntactically valid hypothesis that satisfies given constraints. In the generate stage (line 6), REDUCER uses an answer set programming (ASP) solver to find a model of \mathcal{P} . If there is no model, REDUCER increments the hypothesis size and loops again (lines 7-9). If there is a model, REDUCER converts it to a hypothesis h . In the test stage, it uses Prolog to test h on the training examples (line 10). If h is better (according to a cost function) than the previous best-seen hypothesis, REDUCER sets the best hypothesis to h (line 12). In the constrain stage, REDUCER builds hypothesis constraints (represented as ASP constraints) from h . This step is identical to POPPER. For instance, if h does not entail any positive examples, REDUCER builds a specialisation constraint to prune its specialisations. REDUCER saves these constraints (line 13) which it passes to the generate stage to add to \mathcal{P} to prune models and thus prune the hypothesis space.

Algorithm 1: REDUCER.

```

1 def reducer(bk, pos, neg, max_size):
2   cons = {}
3   size = 1
4   best_h, best_score = none, ∞
5   while size ≤ max_size:
6     h = generate(cons, size)
7     if h == UNSAT:
8       size += 1
9     continue
10    h_score = test(bk, pos, neg, h)
11    if h_score < best_score:
12      best_h, best_score = h, h_score
13    cons += build_cons(h, h_score)
14    if pointless(bk, neg, h):
15      cons += build_spec_basic_con(h)
16      cons += build_gen_basic_con(h)
17  return best_h

```

The novelty of REDUCER is checking whether a hypothesis has a pointless rule (line 14) by calling Algorithm 2. We describe Algorithm 2 below. If a hypothesis has a pointless

rule, REDUCER builds constraints to prune its generalisations and specialisations, in which the pointless rule is basic. REDUCER passes these constraints to the generate stage to add to \mathcal{P} to prune models and thus prune the hypothesis space of other hypotheses with pointless rules.

REDUCER continues its loop until it exhausts the models of \mathcal{P} or reaches a timeout, at which point it returns the best-seen hypothesis.

Algorithm 2: Finding pointless rules.

```

1 def pointless(h, neg, bk):
2   for rule in h:
3     if not basic(rule, h):
4       continue
5     head, body = rule
6     for literal in body:
7       body' = body - literal
8       if not captured(head, body', literal):
9         continue
10      if reducible(bk, body', literal):
11        return true
12      if indiscriminate(bk, neg, rule, head, body'):
13        return true
14    return false
15
16 def reducible(bk, neg, body', literal):
17   rule' = ( $\perp$ , body'  $\cup$  { $\neg$ literal})
18   return unsat(bk, rule')
19
20 def indiscriminate(bk, neg, rule, head, body'):
21   rule' = (head, body')
22   s1 = neg_covered(bk, neg, rule)
23   s2 = neg_covered(bk, neg, rule')
24   return s1 == s2

```

Pointless Rules

Algorithm 2 checks whether a hypothesis has a rule which (i) is basic¹ (Definition 3), (ii) has a captured literal (Definition 7, line 8), and (iii) is reducible (Definition 8, line 10) or indiscriminate (Definition 9, line 12). Algorithm 2 includes two subprocedures for checking if a rule is reducible or indiscriminate. The first procedure (lines 16-18) checks if the query $body' \cup \{\neg literal\}$ is unsatisfiable over the background knowledge. This query determines reducibility because it checks whether a literal is implied by the body. The second subprocedure (lines 20-24) checks if the subrule without the captured literal entails the same negative examples as the original rule. This query matches Definition 9. We use Prolog to perform the unsat (line 18) and coverage (lines 22-23) checks.

Correctness We show that REDUCER is correct:

Theorem 1 (REDUCER correctness). REDUCER returns an optimal hypothesis if one exists.

¹We enforce other syntactic restrictions, such as forcing literals in a rule to be connected, where they cannot be partitioned into two sets such that the variables in the literals of one set are disjoint from the variables in the literals of the other set.

Proof. Cropper and Morel (2021) show that given optimally sound constraints (an optimally sound constraint never prunes an optimal hypothesis), POPPER returns an optimal hypothesis if one exists (Theorem 1). REDUCER builds on POPPER by pruning pointless (one with a reducible or indiscriminate rule) hypotheses. By Corollary 1, a pointless hypothesis is not optimal. Therefore, REDUCER never prunes an optimal hypothesis so returns one if it exists. \square

5 Experiments

We claim that pruning pointless rules enables sound and efficient pruning of the hypothesis space, thereby improving learning performance. To test this claim, our experiments aim to answer the question:

Q1 Can pruning pointless rules reduce learning times whilst maintaining predictive accuracies?

To answer **Q1**, we compare the performance of REDUCER against POPPER. As REDUCER builds on POPPER, the only experimental difference between the systems is the ability to identify pointless rules and build constraints from them to prune the hypothesis space. Therefore, this comparison directly tests our claim.

Checking whether a rule is reducible or indiscriminate requires testing hypotheses on the BK and examples and thus incurs an overhead cost. To understand the cost of pruning pointless rules, our experiments aim to answer the question:

Q2 What is the overhead of pruning pointless rules?

Q1 explores whether the combination of both types of pointless rules, reducible (Definition 8) and indiscriminate (Definition 9), can improve learning performance. To understand the impact of each type of rule, our experiments aim to answer the question:

Q3 Can pruning reducible or indiscriminate rules alone reduce learning times?

To answer **Q3**, we compare the performance of REDUCER against POPPER but where REDUCER identifies only reducible rules or only indiscriminate rules, but not both.

Comparing REDUCER against other systems besides POPPER will not allow us to evaluate the idea of pruning pointless rules because it will not allow us to identify the source of empirical gains. However, many people expect comparisons against other systems. Therefore, we ask the question:

Q4 How does REDUCER compare to other approaches?

To answer **Q4**, we compare REDUCER against POPPER, ALEPH (Srinivasan 2001), and ASPSYNTH (Bembenek, Greenberg, and Chong 2023).

Setup Each task contains training and testing examples and background knowledge. We use the training examples to train the ILP system to learn a hypothesis. We test a hypothesis on the testing examples. Given a hypothesis h , background knowledge B , and a set of examples, a *true positive* (tp) is a positive example entailed by $h \cup B$, a *true negative* (tn) is a negative example not entailed by $h \cup B$, a *false positive* (fp) is a negative example entailed by $h \cup B$, and a *false negative* (fn) is a positive example not entailed by

$h \cup B$. We measure predictive accuracy as *balanced accuracy*: $\frac{1}{2} \left(\frac{tp(h)}{tp(h)+fn(h)} + \frac{tn(h)}{tn(h)+fp(h)} \right)$. In all experiments, we measure predictive accuracy, termination times, and the time taken to discover pointless rules, i.e. the overhead of our approach. For **Q1** and **Q2** we use a timeout of 60 minutes per task. For **Q3**, we use a timeout of 10 minutes per task. We repeat each experiment 10 times. We plot and report 95% confidence intervals (CI). We compute 95% CI via bootstrapping when data is non-normal. To determine statistical significance, we apply either a paired t-test or the Wilcoxon signed-rank test, depending on whether the differences are normally distributed. We use the Benjamini–Hochberg procedure to correct for multiple comparisons. All the systems use similar biases. However, as with other rule selection approaches (Section 2), ASPSYNTH precomputes all possible rules of a certain size and then uses an ASP solver to find a subset. It is infeasible to precompute all possible rules, so we set the maximum rule size to 4. We use an AWS m6a.16xlarge instance to run experiments where each learning task uses a single core.

Reproducibility. The code and experimental data for reproducing the experiments are available as supplementary material and will be made publicly available if the paper is accepted for publication.

Domains We use 449 tasks from several domains:

1D-ARC. This dataset (Xu et al. 2024) contains visual reasoning tasks inspired by the abstract reasoning corpus (Chollet 2019).

IGGP. In inductive general game playing (IGGP) (Cropper, Evans, and Law 2020), the task is to induce rules from game traces from the general game playing competition (Gene-sereth and Björnsson 2013).

IMDB. We use a real-world dataset which contains relations between movies, actors, and directors (Mihalkova, Huynh, and Mooney 2007).

List functions. The goal of each task in this dataset is to identify a function that maps input lists to output lists, where list elements are natural numbers (Rule et al. 2024).

Trains. The goal is to find a hypothesis that distinguishes east and west trains (Larson and Michalski 1977).

Zendo. Zendo is a multiplayer game where players must discover a secret rule by building structures.

Results

Q1. Can pruning Pointless Rules Reduce Learning Times Whilst Maintaining Predictive Accuracies? Figure 1 shows the difference in learning times of REDUCER vs POPPER. In other words, Figure 1 shows the reduction in learning times by ignoring pointless rules. Figure 1 shows that REDUCER consistently and drastically reduces learning times. Significance tests confirm ($p < 0.05$) that REDUCER reduces learning times on 96/449 (21%) tasks and increases learning times on 4/449 (1%) tasks. There is no significant difference in the other tasks. The mean decrease in learning time is 28 ± 4 minutes and the median is 24 minutes with 95% CI between 11 and 43 minutes. The mean increase in learning time is 8 ± 16 minutes and the median is 0 minutes with 95%

CI between 0 and 34 minutes. These are minimum improvements because POPPER often times out after 60 minutes. With a longer timeout, we would likely see greater improvements.

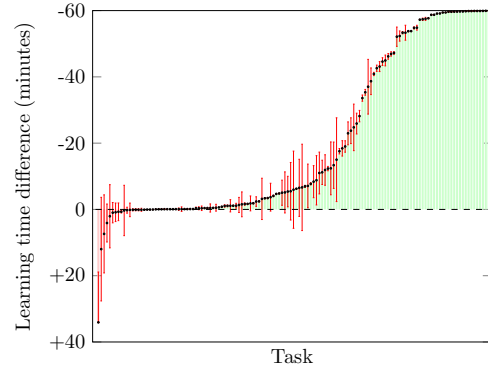


Figure 1: Learning time improvement of REDUCER over POPPER. Values above the zero line indicate that REDUCER reduces learning times. The tasks are ordered by the learning time improvement. For legibility, we only show tasks where the learning times differ by more than 1 second.

Figure 2 shows the difference in predictive accuracy of REDUCER vs POPPER. Significance tests confirm ($p < 0.05$) that REDUCER increases accuracy on 6/449 (1%) tasks. It decreases accuracy on 5/449 (1%) tasks. There is no significant difference in the other tasks. The slight accuracy discrepancy is because multiple optimal hypotheses may exist. The two systems may find different ones. Two hypotheses with the same training cost might have different test accuracy.

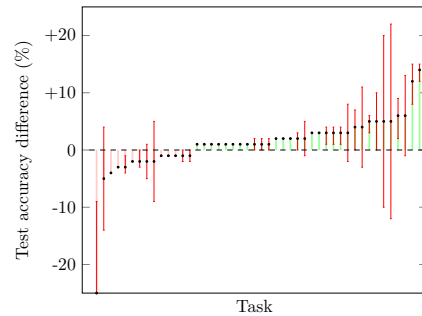


Figure 2: Predictive accuracy improvement of REDUCER compared to POPPER, i.e. when ignoring pointless rules.

REDUCER can drastically improve learning performance. For instance, in the IGGP dataset, one of the tasks is to learn a set of rules to describe a legal move in the *eight puzzle* game. For this task, both REDUCER and POPPER learn hypotheses with 100% accuracy. However, whereas POPPER does not terminate (prove optimality) within 60 minutes, REDUCER terminates (proves optimality) after only 12 ± 0 seconds, a 99% improvement. A reducible rule that REDUCER finds is:

$$\text{legal_move}(A,B,C,D) \leftarrow \text{succ}(D,E), \text{pos1}(D), \text{pos2}(E)$$

This rule is reducible because the pos_i relations denote positions in the game board, where pos_i precedes pos_{i+1} . There-

fore, if $pos_1(D)$ is true and E is the successor of D then $pos_2(E)$ must be true, and this literal is therefore redundant.

Three indiscriminate rules that REDUCER finds are:

$$\begin{aligned} legal_move(A,B,C,D) &\leftarrow role(B) \\ legal_move(A,B,C,D) &\leftarrow index(C) \\ legal_move(A,B,C,D) &\leftarrow index(D) \end{aligned}$$

These rules are indiscriminate because $role(B)$, $index(C)$, and $index(D)$ are true for every negative example and thus are redundant in the rule.

Overall, these results suggest that the answer to **Q1** is yes, pruning pointless can drastically improve learning times whilst maintaining high predictive accuracies.

Q2. What Is the Overhead of Finding Pointless Rules?

Figure 3 shows the ratio of learning time spent finding pointless rules. The mean overhead is 2%. The maximum is 80%. The overhead is less than 10% on 85% of the tasks. Overall, these results suggest that the answer to **Q2** is that the overhead of pruning pointless is typically small (<10%).

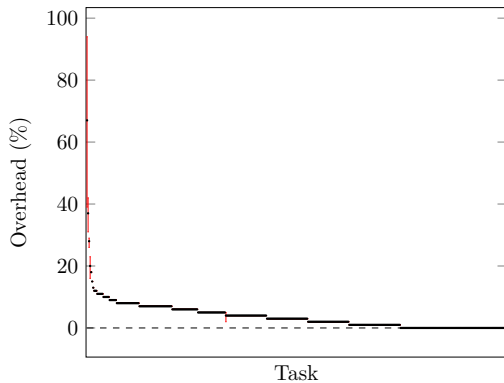


Figure 3: Overhead of finding pointless rules.

Q3. Can pruning Reducible or Indiscriminate Rules Alone Reduce Learning Times?

Figure 4 shows how the learning change depending on whether REDUCER prunes only reducible or only indiscriminate rules. Significance tests confirm ($p < 0.05$) that pruning only reducible rules reduces learning times on 93/449 (21%) tasks and increases learning times on 1/449 (0%) tasks. There is no significant difference in the other tasks. The mean decrease in learning time is 26 ± 5 minutes and the median is 17 minutes with 95% CI between 7 and 34 minutes. The mean and median increase is 1 minute. Pruning only indiscriminate rules reduces learning times on 101/449 (22%) tasks and increases learning times on 5/449 (1%) tasks. There is no significant difference in the other tasks. The mean decrease in learning time is 28 ± 4 minutes and the median is 22 minutes with 95% CI between 8 and 42 minutes. The mean increase is 6 ± 11 minutes and the median increase is 0 minutes with 95% CI between 0 and 28 minutes. Overall, these results suggest that the answer to **Q3** is that both types of pointless rules drastically reduce learning times.

Q4. How Does REDUCER Compare to Other Approaches?

Table 1 shows the predictive accuracies aggregated per do-

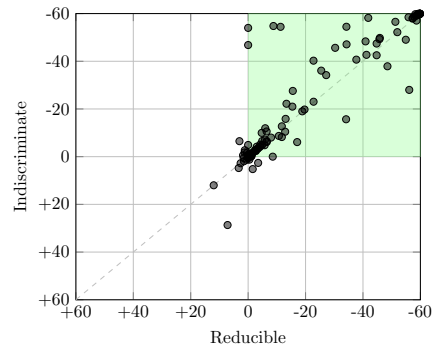


Figure 4: Learning time improvement (minutes) when pruning reducible and indiscriminate rules. Values in the upper right quadrant indicate that pruning either reducible rules or indiscriminate rules alone reduces learning time.

main of all the systems. REDUCER has higher accuracy than ALEPH and ASPSYNTH on every domain. REDUCER has higher or equal accuracy than POPPER on every domain. Overall, these results suggest that the answer to **Q4** is that REDUCER compares favourably to existing approaches in terms of predictive accuracy.

Task	ALEPH	ASPSYNTH	POPPER	REDUCER
<i>ld</i>	54 ± 4	89 ± 3	92 ± 1	92 ± 1
<i>alzheimer</i>	51 ± 1	57 ± 2	75 ± 1	75 ± 1
<i>iggp</i>	57 ± 5	68 ± 0	85 ± 0	85 ± 0
<i>imdb</i>	50 ± 0	98 ± 0	100 ± 0	100 ± 0
<i>jr</i>	64 ± 11	83 ± 2	96 ± 0	97 ± 0
<i>trains</i>	50 ± 0	76 ± 2	87 ± 9	94 ± 7
<i>zendo</i>	53 ± 6	82 ± 0	84 ± 3	84 ± 3

Table 1: Aggregated predictive accuracies (%).

6 Conclusions and Limitations

We have introduced an approach that identifies and ignores pointless (reducible and indiscriminate) rules. We have shown that ignoring pointless rules is optimally sound (Propositions 2 and 4). We implemented our approach in REDUCER, which identifies pointless rules in hypotheses and builds constraints from them to prune the hypothesis space. We have proven that REDUCER always learns an optimal hypothesis if one exists (Theorem 1). We have experimentally shown on multiple domains, including visual reasoning and game playing, that our approach can reduce learning times by 99% whilst maintaining high predictive accuracies.

Limitations. The benefits of ignoring pointless rules should generalise to other ILP approaches. For instance, the improvements should directly improve HOPPER (Pur al, Cerna, and Kaliszyk 2022), which learns higher-order programs, and PROPPER (Hillerstr om and Burghouts 2025), which uses neurosymbolic inference to learn programs from probabilistic data. Future work should empirically show how our idea benefits these systems.

Acknowledgments

Andrew Cropper was supported by his EPSRC fellowship (EP/V040340/1). David M. Cerna was supported by the Czech Science Foundation Grant 22-06414L and Cost Action CA20111 EuroProofNet.

References

- Ahlgren, J.; and Yuen, S. Y. 2013. Efficient program synthesis using constraint satisfaction in inductive logic programming. *J. Machine Learning Res.*, (1): 3649–3682.
- Barnett, L. A.; Cerna, D. M.; and Biere, A. 2020. Covered Clauses Are Not Propagation Redundant. In *Automated Reasoning - 10th International Joint Conference, IJCAR*, volume 12166, 32–47. Springer.
- Bembenek, A.; Greenberg, M.; and Chong, S. 2023. From SMT to ASP: Solver-Based Approaches to Solving Datalog Synthesis-as-Rule-Selection Problems. *Proc. ACM Program. Lang.*, 7(POPL).
- Biere, A.; Jarvisalo, M.; and Kiesl, B. 2021. Preprocessing in SAT Solving. In Biere, A.; Heule, M.; van Maaren, H.; and Walsh, T., eds., *Handbook of Satisfiability - Second Edition*, volume 336, 391–435.
- Chollet, F. 2019. On the Measure of Intelligence. *CoRR*.
- Corapi, D.; Russo, A.; and Lupu, E. 2011. Inductive Logic Programming in Answer Set Programming. In *Inductive Logic Programming - 21st International Conference*, 91–97.
- Cropper, A.; and Dumančić, S. 2022. Inductive Logic Programming At 30: A New Introduction. *Journal of Artificial Intelligence Research*, 74: 765–850.
- Cropper, A.; Evans, R.; and Law, M. 2020. Inductive general game playing. *Machine Learning*, 109(7): 1393–1434.
- Cropper, A.; and Hocquette, C. 2024. Learning Logic Programs by Finding Minimal Unsatisfiable Subprograms. In *ECAI 2024 - 27th European Conference on Artificial Intelligence*, volume 392, 4295–4302. IOS Press.
- Cropper, A.; and Morel, R. 2021. Learning programs by learning from failures. *Machine Learning*, 110(4): 801–856.
- De Raedt, L. 2008. *Logical and relational learning*. ISBN 978-3-540-20040-6.
- Fonseca, N. A.; Costa, V. S.; Silva, F. M. A.; and Camacho, R. 2004. On Avoiding Redundancy in Inductive Logic Programming. In *Inductive Logic Programming, 14th International Conference, ILP 2004, Porto, Portugal, September 6-8, 2004, Proceedings*, volume 3194 of *Lecture Notes in Computer Science*, 132–146. Springer.
- Fürnkranz, J.; and Kliegr, T. 2015. A Brief Overview of Rule Learning. In *RuleML 2015*, volume 9202 of *Lecture Notes in Computer Science*, 54–69. Springer.
- Galárraga, L.; Teflioudi, C.; Hose, K.; and Suchanek, F. M. 2015. Fast rule mining in ontological knowledge bases with AMIE+. *VLDB J.*, 24(6): 707–730.
- Genesereth, M.; and Björnsson, Y. 2013. The international general game playing competition. *AI Magazine*, 34(2): 107–107.
- Heule, M.; Jarvisalo, M.; and Biere, A. 2010. Covered Clause Elimination. In Voronkov, A.; Sutcliffe, G.; Baaz, M.; and Fermüller, C. G., eds., *Short papers for 17th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning, LPAR-17-short, Yogyakarta, Indonesia, October 10-15, 2010*, volume 13 of *EPiC Series in Computing*, 41–46. EasyChair.
- Hillerström, F.; and Burghouts, G. 2025. Towards Probabilistic Inductive Logic Programming with Neurosymbolic Inference and Relaxation. *Theory and Practice of Logic Programming*, 24: 628–643.
- Hocquette, C.; Niskanen, A.; Jarvisalo, M.; and Cropper, A. 2024. Learning MDL logic programs from noisy data. In *AAAI*.
- Hoder, K.; Khasidashvili, Z.; Korovin, K.; and Voronkov, A. 2012. Preprocessing techniques for first-order classification. In Cabodi, G.; and Singh, S., eds., *Formal Methods in Computer-Aided Design, FMCAD 2012, Cambridge, UK, October 22-25, 2012*, 44–51. IEEE.
- Kaminski, T.; Eiter, T.; and Inoue, K. 2019. Meta-Interpretive Learning Using HEX-Programs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, 6186–6190.
- Khasidashvili, Z.; and Korovin, K. 2016. Predicate Elimination for Preprocessing in First-Order Theorem Proving. In *Theory and Applications of Satisfiability Testing - SAT 2016*, volume 9710, 361–372. Springer.
- Kullmann, O. 1999. On a Generalization of Extended Resolution. *Discret. Appl. Math.*, 96-97: 149–176.
- Larson, J.; and Michalski, R. S. 1977. Inductive inference of VL decision rules. *SIGART Newsletter*, 38–44.
- Law, M. 2023. Conflict-driven inductive logic programming. *Theory and Practice of Logic Programming*, 23(2): 387–414.
- Lloyd, J. W. 2012. *Foundations of logic programming*. Springer Science & Business Media.
- Mihalkova, L.; Huynh, T. N.; and Mooney, R. J. 2007. Mapping and Revising Markov Logic Networks for Transfer Learning. In *AAAI 2007*, 608–614.
- Muggleton, S. 1991. Inductive Logic Programming. *New Generation Computing*, 8(4): 295–318.
- Muggleton, S. H. 1995. Inverse Entailment and Progol. *New Generation Computing*, 13(3&4): 245–286.
- Muggleton, S. H.; Lin, D.; and Tamaddoni-Nezhad, A. 2015. Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. *Machine Learning*, 100(1): 49–73.
- Plotkin, G. 1971. *Automatic Methods of Inductive Inference*. Ph.D. thesis, Edinburgh University.
- Purzał, S. J.; Cerna, D. M.; and Kaliszzyk, C. 2022. Learning Higher-Order Logic Programs From Failures. In *IJCAI 2022*.
- Raedt, L. D.; and Ramon, J. 2004. Condensed Representations for Inductive Logic Programming. In Dubois, D.; Welty, C. A.; and Williams, M., eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004), Whistler, Canada, June 2-5, 2004*, 438–446. AAAI Press.

Raghothaman, M.; Mendelson, J.; Zhao, D.; Naik, M.; and Scholz, B. 2020. Provenance-guided synthesis of Datalog programs. *Proc. ACM Program. Lang.*, 4(POPL): 62:1–62:27.

Ray, O. 2009. Nonmonotonic abductive inductive learning. *J. Applied Logic*, (3): 329–340.

Rule, J. S.; Piantadosi, S. T.; Cropper, A.; Ellis, K.; Nye, M.; and Tenenbaum, J. B. 2024. Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, 15(1): 6847.

Schüller, P.; and Benz, M. 2018. Best-effort inductive logic programming via fine-grained cost-based hypothesis generation - The inspire system at the inductive logic programming competition. *Machine Learning*, 107(7): 1141–1169.

Shapiro, E. Y. 1983. *Algorithmic Program DeBugging*. Cambridge, MA, USA. ISBN 0262192187.

Srinivasan, A. 2001. The ALEPH manual. *Machine Learning at the Computing Laboratory, Oxford University*.

Srinivasan, A.; and Kothari, R. 2005. A Study of Applying Dimensionality Reduction to Restrict the Size of a Hypothesis Space. In *Inductive Logic Programming, 15th International Conference, ILP 2005, Bonn, Germany, August 10-13, 2005, Proceedings*, 348–365.

Vukmirovic, P.; Blanchette, J.; and Heule, M. J. H. 2023. SAT-Inspired Eliminations for Superposition. *ACM Trans. Comput. Log.*, 24(1): 7:1–7:25.

Xu, Y.; Li, W.; Vaezipoor, P.; Sanner, S.; and Khalil, E. B. 2024. LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations. *Trans. Mach. Learn. Res.*, 2024.

Zeman, V.; Kliegr, T.; and Svátek, V. 2021. RDFRules: Making RDF rule mining easier and even more efficient. *Semantic Web*, 12(4): 569–602.

Zeng, Q.; Patel, J. M.; and Page, D. 2014. QuickFOIL: Scalable Inductive Logic Programming. *Proc. VLDB Endow.*, (3): 197–208.