When Can We Learn from Noisy Logical Data? Parameterized Complexity of Approximate Concept Fitting in Description Logics

Chang Lu[1] **Yizheng Zhao**[*2,3] **Renate A. Schmidt**[4]

[1]Computational Biology and Biomedical Informatics Program, Yale University, New Haven, CT, USA
[2]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China
[3]School of Artificial Intelligence, Nanjing University, Nanjing, China
[4]Department of Computer Science, The University of Manchester, Manchester, UK

Abstract

SAT-based bounded fitting—searching for a smallest logical concept that correctly classifies all given examples—yields sample-efficient PAC learning for structured hypothesis classes such as description logic (DL) concepts. Yet this paradigm is fundamentally brittle: it demands "*realizability*", and a single mislabeled example can render the underlying SAT instance unsatisfiable. While agnostic learnability follows from finite VC-dimension of bounded-size concept classes, the real barrier is *computational*: minimizing misclassifications over all concepts of bounded size is NP-hard. We study this barrier through parameterized complexity. We introduce ApxFit(k, t)—the problem of deciding whether there exists a DL concept of size at most k that misclassifies at most t out of m examples—and establish a complexity landscape for fragments of $\mathcal{ALC}$ whose operator sets contain $\{\sqcap, \exists\}$ or $\{\sqcup, \forall\}$: the problem is **W[2]-hard** parameterized by concept size k (even for t=0), yet lies in **XP** parameterized by outlier count t for any fixed k. Algorithmically, we lift bounded fitting from SAT to partial weighted MAX-SAT and combine it with structural risk minimization, yielding the first noise-tolerant DL concept learner with formal agnostic generalization guarantees.

1 INTRODUCTION

A central task in computational learning theory is to find, from labeled examples, a concept that generalizes well to unseen data [Valiant, 1984]. When the concept class consists of logical formulas—such as Boolean formulas [Kearns et al., 1987, Pitt and Valiant, 1988], decision trees [Gahlawat and Zehavi, 2024], first-order definable concepts [van Bergerem, 2019], or description logic (DL) concepts [Cohen and Hirsh, 1992, Funk et al., 2019]—the search space is combinatorial, but *parsimony* serves as an inductive bias: a small concept that fits the data is unlikely to overfit. This intuition is captured by the Occam framework [Blumer et al., 1989, Board and Pitt, 1990], which guarantees that any algorithm returning a consistent concept whose size is polynomial in the target size is a sample-efficient PAC learner. A natural instantiation is *bounded fitting*: iteratively search for a concept of size $s = 1, 2, 3, \ldots$ that correctly classifies all examples, and return the first one found. Inspired by bounded model checking [Biere et al., 1999], ten Cate et al. [2023] realized this idea for DL concept learning by encoding bounded fitting as propositional satisfiability (SAT), obtaining the first provable PAC guarantees for DL concepts. Funk et al. [2025] extended this SAT-based approach from the lightweight DL $\mathcal{EL}$ to the expressive $\mathcal{ALC}$, establishing NP-completeness of bounded fitting and developing efficient encodings. DL concepts are a particularly rich concept class for this paradigm: their syntax combines conjunction, disjunction, negation, and quantified role restrictions [Baader et al., 2017], and they have a thirty-year history in the computational learning theory literature [Cohen and Hirsh, 1992, 1994].

Bounded fitting, however, requires *realizability*: a concept of the given size must correctly classify *every* example. A single mislabeled instance renders the SAT formula unsatisfiable, forcing the algorithm to either fail or inflate the size bound until the concept overfits the noise. This brittleness is inherent to exact fitting and cannot be patched at the implementation level. The need for noise-tolerant concept learning has long been recognized, both in general learning theory [Angluin and Laird, 1987, Kearns and Li, 1993] and in the DL community specifically. Heuristic DL concept learners—including DL-Learner [Lehmann and Hitzler, 2010, Bühmann et al., 2016], EvoLearner [Heindorf et al., 2022], DL-FOIL [Fanizzi et al., 2008], and neural approaches [Kouagou et al., 2023]—tolerate noise implicitly through fitness functions or approximate matching, but none provides formal generalization guarantees under noise. In

*Corresponding author.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3979–3998.

adjacent areas of logical concept learning, MAX-SAT-based methods have successfully handled noisy data for temporal logic specifications [Gaglione et al., 2021] and interpretable classification rules [Malioutov and Meel, 2018], suggesting that the optimization capabilities of MAX-SAT solvers can bridge the gap between exact fitting and noise tolerance. Indeed, ten Cate et al. [2023] explicitly identify this as an open problem: extending bounded fitting to handle erroneous labels, noting that "the optimization features of MAX-SAT solvers seem promising".

One might hope that standard learning theory already resolves the noise problem. Bounded-size DL concepts form a finite concept class whose VC-dimension is $O(k \log |\Sigma|)$, where k is the concept size and $|\Sigma|$ the signature size. By the fundamental theorem of statistical learning [Shalev-Shwartz and Ben-David, 2014, Haussler, 1992], every such class is agnostic PAC learnable: the empirical risk minimizer (ERM)—a concept of size at most k that minimizes classification error on the training data—converges to the best-in-class concept at rate $O(\sqrt{k \log |\Sigma|/m})$, where m is the sample size [Kearns et al., 1994]. The barrier is therefore not statistical but *computational*: even the special case of zero error (exact fitting, t=0) is already NP-complete [Funk et al., 2025], so computing the ERM for arbitrary error levels cannot be easier. This separation between statistical learnability and computational tractability is a fundamental phenomenon in learning theory [Pitt and Valiant, 1988, Kearns et al., 1987]: a concept class may be learnable in the information-theoretic sense, yet no efficient algorithm can compute the optimal concept unless P = NP. The right question is therefore not *whether* bounded-size DL concepts are noise-tolerantly learnable, but *for which parameterizations* the intractable ERM computation becomes efficient.

Parameterized complexity [Downey and Fellows, 2013, Cygan et al., 2015] is designed to answer exactly this type of question. Rather than a binary polynomial/NP-hard classification, it asks: for which *parameters* does a problem admit algorithms whose exponential cost is confined to the parameter alone? A problem is *fixed-parameter tractable* (FPT) with respect to parameter κ if it can be solved in time $f(\kappa) \cdot n^{O(1)}$; it is *W-hard* if no such algorithm exists under standard assumptions. This framework has recently proved powerful for learning problems. Brand et al. [2023] establish a general connection between parameterized consistency checking and PAC learnability for the realizable setting, showing that FPT consistency implies efficient PAC learning. Gahlawat and Zehavi [2024] obtain a sharp dichotomy for learning decision trees with outliers: the problem is W[1]-hard when parameterized by tree size, but becomes FPT when parameterized by the number of outliers. This is precisely the type of result we seek for DL concepts. The computational complexity of DL concept learning has been investigated in depth—from separability and fitting [Funk et al., 2019, 2025] to active learning [Funk et al., 2021]—but all existing results assume realizability. The parameterized complexity of *approximate fitting*—finding a bounded-size concept that minimizes disagreement with noisy examples—has not been studied.

Contributions. We initiate this study in this paper:

1. **Problem formulation.** We define ApxFit(k, t): given labeled examples, does there exist a concept of size at most k that misclassifies at most t examples? This cleanly generalizes exact fitting (t=0) and directly captures the ERM computation underlying agnostic PAC learning (Section 3).
2. **Intractability by concept size.** ApxFit(k, t) is W[2]-hard parameterized by concept size k, for every fragment of $\mathcal{ALC}$ whose operator set O satisfies $\{\sqcap, \exists\} \subseteq O$ or $\{\sqcup, \forall\} \subseteq O$. Hardness holds even for exact fitting (t=0) (Section 4).
3. **Tractability by outlier count.** The problem is in XP when parameterized by outlier count t for any fixed concept size bound k: it can be solved in time $O(m^t \cdot p(m, |\Sigma|))$ where p is a polynomial depending on k. Together with (2), this delineates a complexity landscape where concept size drives intractability while outlier count enables tractability (Section 4).
4. **MAX-SAT algorithm with generalization guarantees.** We lift bounded fitting from SAT to partial weighted MAX-SAT—hard clauses enforce syntactic and semantic correctness, soft clauses encode example classification—and combine it with structural risk minimization [Vapnik, 1991]. The resulting algorithm is the first noise-tolerant DL concept learner with formal agnostic PAC guarantees, achieving the oracle inequality $R(\hat{C}) \leq \min_s\{R^*_s + O(\sqrt{s \log |\Sigma|/m})\}$, where R^*_s is the best achievable risk at concept size s (Section 5).
5. **Experiments.** On four standard ontology benchmarks with injected label noise at rates 5–30%, our approach degrades gracefully while exact fitting fails entirely, and matches the accuracy of heuristic learners in noise-free settings (Section 6; code in the supplementary material).

2 PRELIMINARIES

2.1 DESCRIPTION LOGICS

Let N_C, N_R, and N_I be countably infinite, mutually disjoint sets of *concept names*, *role names*, and *individual names*, respectively. A *signature* is a finite pair $\Sigma = (\Sigma_C, \Sigma_R)$ with $\Sigma_C \subseteq N_C$ and $\Sigma_R \subseteq N_R$. We write $|\Sigma|$ for $|\Sigma_C| + |\Sigma_R|$. *$\mathcal{ALC}$-concepts* over Σ are built by the grammar

$$C, D \ ::= \ \top \mid \bot \mid A \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid \exists r.C \mid \forall r.C$$

where $A \in \Sigma_C$ and $r \in \Sigma_R$. The *size* $\|C\|$ of a concept C is the number of occurrences of concept names, role names,

$\top$, and $\bot$ in C. For a set of operators $O \subseteq \{\sqcap, \sqcup, \neg, \exists, \forall\}$, we write $L(O)$ for the fragment of $\mathcal{ALC}$ that uses only the constructors in O together with $\top$, $\bot$, and concept names. Thus $\mathcal{ALC} = L(\{\sqcap, \sqcup, \neg, \exists, \forall\})$, and $\mathcal{EL} = L(\{\sqcap, \exists\})$.

An *interpretation* is a pair $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ consisting of a non-empty *domain* $\Delta^{\mathcal{I}}$ and a function $\cdot^{\mathcal{I}}$ that maps every $A \in N_C$ to $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$, every $r \in N_R$ to $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, and every $a \in N_I$ to $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$. The *extension* $C^{\mathcal{I}}$ of a concept C in $\mathcal{I}$ is defined inductively: $\top^{\mathcal{I}} = \Delta^{\mathcal{I}}$, $\bot^{\mathcal{I}} = \emptyset$, $(\neg C)^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$, $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$, $(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$, $(\exists r.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \exists e.\, (d,e) \in r^{\mathcal{I}} \wedge e \in C^{\mathcal{I}}\}$, and $(\forall r.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \forall e.\, (d,e) \in r^{\mathcal{I}} \rightarrow e \in C^{\mathcal{I}}\}$. We say that $d \in \Delta^{\mathcal{I}}$ *satisfies* C in $\mathcal{I}$ if $d \in C^{\mathcal{I}}$.

A *labeled example* is a triple $(\mathcal{I}, a, \ell)$ where $\mathcal{I}$ is a finite interpretation, $a \in N_I$ with $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$, and $\ell \in \{+, -\}$. An example is *positive* if $\ell = +$ and *negative* if $\ell = -$. A concept C *fits* a labeled example $(\mathcal{I}, a, +)$ if $a^{\mathcal{I}} \in C^{\mathcal{I}}$, and fits $(\mathcal{I}, a, -)$ if $a^{\mathcal{I}} \notin C^{\mathcal{I}}$.

2.2 PAC LEARNING

We recall the agnostic PAC model [Haussler, 1992, Shalev-Shwartz and Ben-David, 2014]. Let $\mathcal{X}$ be an instance space and let $\mathcal{H} \subseteq \{h : \mathcal{X} \rightarrow \{0,1\}\}$ be a concept class. The *true risk* of $h \in \mathcal{H}$ with respect to a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ is $R(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$. The *empirical risk* on a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[h(x_i) \neq y_i]$. The *empirical risk minimizer* (ERM) over $\mathcal{H}$ returns $\hat{h} \in \arg\min_{h\in\mathcal{H}} \hat{R}_S(h)$.

A concept class $\mathcal{H}$ is *agnostic PAC learnable* if there exists an algorithm that, for every $\epsilon, \delta > 0$ and every distribution $\mathcal{D}$, given $m = \text{poly}(1/\epsilon, 1/\delta)$ i.i.d. examples from $\mathcal{D}$, returns $\hat{h}$ satisfying $R(\hat{h}) \leq \min_{h\in\mathcal{H}} R(h) + \epsilon$ with probability at least $1 - \delta$. A finite concept class $\mathcal{H}$ always satisfies this guarantee via the ERM, with sample complexity $m = O(\frac{1}{\epsilon^2}(\log|\mathcal{H}| + \log(1/\delta)))$. More generally, agnostic learnability is characterized by finite VC-dimension $\text{VCdim}(\mathcal{H})$ [Blumer et al., 1989]. For DL concept learning, the relevant concept class is $\mathcal{H}_k^{L(O),\Sigma}$: the set of all $L(O)$-concepts over Σ of size at most k, viewed as classifiers mapping each labeled example $(\mathcal{I}, a)$ to $\mathbf{1}[a^{\mathcal{I}} \in C^{\mathcal{I}}]$. Since $|\mathcal{H}_k^{L(O),\Sigma}|$ is finite (bounded by $|\Sigma|^{O(k)}$; see Appendix C), the class is agnostic PAC learnable with sample complexity $O(\frac{k\log|\Sigma| + \log(1/\delta)}{\epsilon^2})$. The ERM over $\mathcal{H}_k^{L(O),\Sigma}$ is a concept $\hat{C} \in \arg\min_{\|C\|\leq k} \hat{R}_S(C)$; computing it requires solving an optimization problem that we formalize in Section 3.

2.3 PARAMETERIZED COMPLEXITY

A *parameterized problem* is a language $L \subseteq \{0,1\}^* \times \mathbb{N}$ where each instance (x, κ) consists of an input x and a parameter $\kappa \in \mathbb{N}$. The problem is *fixed-parameter tractable* (FPT) if there exists a computable function f and a constant c such that $(x, \kappa) \in L$ can be decided in time $f(\kappa)\cdot|x|^c$. It is in the class *XP* if it can be decided in time $|x|^{f(\kappa)}$. The *W-hierarchy* $\mathsf{FPT} \subseteq \mathsf{W[1]} \subseteq \mathsf{W[2]} \subseteq \cdots$ provides a sequence of presumably strict inclusions [Downey and Fellows, 2013]. Hardness for $\mathsf{W}[i]$ is established via *parameterized reductions*: a function $(x, \kappa) \mapsto (x', \kappa')$ computable in FPT time with $\kappa' \leq g(\kappa)$ for some computable g.

The canonical **W[2]**-complete problem is HITTING SET: given a universe U, a family $\mathcal{F} \subseteq 2^U$, and $\kappa \in \mathbb{N}$, decide whether there exists $S \subseteq U$ with $|S| \leq \kappa$ such that $S \cap F \neq \emptyset$ for every $F \in \mathcal{F}$. Our W[2]-hardness proof in Section 4 reduces from this problem.

3 THE APXFIT PROBLEM

The bounded fitting problem studied by ten Cate et al. [2023] and Funk et al. [2025] asks whether there exists a concept of size at most k that fits *all* given examples. As argued in Section 1, this exact fitting requirement ($t = 0$) makes the problem brittle under label noise. We now generalize it by allowing a bounded number of misclassified examples.

Definition 3.1. Let $E = \{(\mathcal{I}_1, a_1, \ell_1), \ldots, (\mathcal{I}_m, a_m, \ell_m)\}$ be a set of labeled examples and let C be a concept. The *disagreement* of C on E is

$$\text{disagree}(C, E) = \big|\{i \in [m] : \ell_i{=}+,\ a_i^{\mathcal{I}_i} \notin C^{\mathcal{I}_i}\}\big| + \big|\{i \in [m] : \ell_i{=}-,\ a_i^{\mathcal{I}_i} \in C^{\mathcal{I}_i}\}\big|.$$

i.e., the number of false negatives plus false positives.

Definition 3.2. Fix an operator set $O \subseteq \{\sqcap, \sqcup, \neg, \exists, \forall\}$ and a signature Σ. The problem APXFIT(k, t) is: given a set E of labeled examples over Σ and parameters $k, t \in \mathbb{N}$, decide whether there exists an $L(O)$-concept C over Σ with $\|C\| \leq k$ and $\text{disagree}(C, E) \leq t$.

We note that setting $t = 0$ recovers the exact fitting problem of Funk et al. [2025], which is NP-complete for every O with $O \cap \{\exists, \forall\} \neq \emptyset$. On the other hand, the optimization version $\min_{\|C\|\leq k} \text{disagree}(C, E)$ is precisely the ERM computation over the concept class $\mathcal{H}_k^{L(O),\Sigma}$ defined in Section 2.2: the empirical risk $\hat{R}_S(C)$ equals $\text{disagree}(C, E)/m$. The problem APXFIT(k, t) thus captures the decision version of agnostic concept learning, parameterized by concept size k and outlier count t.

Proposition 3.3. *Let $O \subseteq \{\sqcap, \sqcup, \neg, \exists, \forall\}$ with $O \cap \{\exists, \forall\} \neq \emptyset$. (i)* (Statistical) *The concept class $\mathcal{H}_k^{L(O),\Sigma}$ is agnostic PAC learnable with sample complexity polynomial in k, $|\Sigma|$, $1/\epsilon$, and $1/\delta$; (ii)* (Computational) APXFIT(k, t) *is* NP*-complete, even for fixed $t = 0$.*

Proof. (i) follows from finiteness of $\mathcal{H}_k^{L(O),\Sigma}$ (Section 2.2). For (ii), membership in NP holds because a concept C

with $\|C\| \leq k$ can be guessed and verified in polynomial time: given a finite interpretation $\mathcal{I}$ (provided explicitly as input), evaluating C on $\mathcal{I}$ is a bottom-up computation on the syntax tree of C, taking $O(\|C\| \cdot |\Delta^{\mathcal{I}}|^2)$ time. This is concept *evaluation*, not *reasoning*: no TBox entailment or model search is involved, so the PSpace complexity of ALC satisfiability does not apply. Hardness follows from the $t = 0$ case, which is the NP-complete bounded fitting problem of Funk et al. [2025]. □

Proposition 3.3 formalizes the statistical-computational gap previewed in Section 1: the concept class is learnable in the information-theoretic sense, but the required optimization is intractable. The parameterized complexity results in Section 4 refine this picture by identifying which parameterizations of ApxFit(k, t) yield tractability.

4 PARAMETERIZED COMPLEXITY

Prop. 3.3 establishes that ApxFit(k, t) is NP-complete in the classical sense. We now analyze the problem through the lens of parameterized complexity, considering two natural parameterizations: concept size k and outlier count t.

4.1 INTRACTABILITY BY CONCEPT SIZE

The NP-hardness proof of Funk et al. [2025] for exact fitting reduces from Hitting Set, but the resulting concept size grows with the input. We give a different reduction in which the concept size depends *only* on the hitting set parameter κ, establishing W[2]-hardness. The key idea is to encode universe elements as role names in the signature—which is part of the input—rather than as positions inside the concept.

Theorem 4.1. ApxFit(k, t) *is W[2]-hard parameterized by k, for every operator set O with $\{\sqcap, \exists\} \subseteq O$ or $\{\sqcup, \forall\} \subseteq O$. Hardness holds even for $t = 0$ and with $\Sigma_C = \emptyset$.*

Proof sketch (see full proof in Appendix A). We present the case $\{\sqcap, \exists\} \subseteq O$; the dual case follows by exchanging $\sqcap \leftrightarrow \sqcup$, $\exists \leftrightarrow \forall$, and swapping positive and negative labels. Given a Hitting Set instance $(U, \mathcal{F}, \kappa)$ with $U = \{u_1, \ldots, u_n\}$ and $\mathcal{F} = \{S_1, \ldots, S_p\}$, we construct an ApxFit$(k, 0)$ instance over the signature $\Sigma_C = \emptyset$, $\Sigma_R = \{r_1, \ldots, r_n\}$, with $k = 2\kappa$. The single positive example $(\mathcal{I}^+, a, +)$ has a connected via each r_i to a dead-end element d (no outgoing roles, no concept names). For each $S_j \in \mathcal{F}$, the negative example $(\mathcal{I}_j, b_j, -)$ has b_j connected via r_i to a dead-end d_j only for $i \notin S_j$.

Forward. A hitting set $H = \{u_{i_1}, \ldots, u_{i_q}\}$ yields $C = \exists r_{i_1}.\top \sqcap \cdots \sqcap \exists r_{i_q}.\top$ with $\|C\| = 2q \leq 2\kappa$. The element a satisfies every conjunct (all r_i-successors exist). Each b_j fails some conjunct since $H \cap S_j \neq \emptyset$ implies some r_{i_h}-successor is absent.

Reverse. Let C be any $\mathcal{ALC}$-concept with $\|C\| \leq 2\kappa$ fitting all examples. Since every non-root element is a dead end, any subconcept evaluates to a fixed Boolean constant on these elements. Consequently, the only atoms that distinguish a from some b_j are of the form $\exists r_i.D$ where D evaluates to true on dead ends, and each such atom costs at least 2 in concept size. Thus at most κ distinct role indices appear in C. Setting $H = \{u_i : r_i \text{ occurs in } C\}$ gives $|H| \leq \kappa$. If $S_j \cap H = \emptyset$ for some j, then a and b_j agree on all role names in C, so C cannot distinguish them—contradicting $a \in C^{\mathcal{I}^+}$ and $b_j \notin C^{\mathcal{I}_j}$. □

Two aspects of this merit comment: (i) the W[2]-hardness (as opposed to W[1]-hardness) reflects the combinatorial structure of Hitting Set and the ability of DL concepts to express arbitrary conjunctions of existential restrictions, each selecting one element. For comparison, Gahlawat and Zehavi [2024] obtain only W[1]-hardness for decision trees via a reduction from Independent Set; (ii) Our fragment condition—requiring both conjunction and existential restriction (or dually, disjunction and universal restriction)—is necessary for the reduction. Whether W-hardness holds for smaller fragments such as $L(\{\exists\})$ remains open.

Example 4.2. Consider the Hitting Set instance with universe $U = \{u_1, u_2, u_3\}$, family $\mathcal{F} = \{S_1, S_2\}$ where $S_1 = \{u_1, u_2\}$ and $S_2 = \{u_2, u_3\}$, and parameter $\kappa = 1$. The reduction produces the signature $\Sigma_R = \{r_1, r_2, r_3\}$ (no concept names) and the parameter $k = 2\kappa = 2$.

The positive example $(\mathcal{I}^+, a, +)$ has domain $\{a, d\}$, with r_1, r_2, r_3 each mapping (a, d); so a has an r_i-successor for every i. Negative example $(\mathcal{I}_1, b_1, -)$ for $S_1 = \{u_1, u_2\}$ has domain $\{b_1, d_1\}$ with only $r_3^{\mathcal{I}_1} = \{(b_1, d_1)\}$; the roles r_1, r_2 are empty (their indices are in S_1). Similarly, $(\mathcal{I}_2, b_2, -)$ for $S_2 = \{u_2, u_3\}$ has only $r_1^{\mathcal{I}_2} = \{(b_2, d_2)\}$.

Forward: the hitting set $H = \{u_2\}$ (size $1 \leq \kappa$) yields concept $C = \exists r_2.\top$ with $\|C\| = 2 \leq k$. Indeed, $a \in C^{\mathcal{I}^+}$ since a has an r_2-successor; $b_1 \notin C^{\mathcal{I}_1}$ since $r_2^{\mathcal{I}_1} = \emptyset$ ($2 \in S_1$); and $b_2 \notin C^{\mathcal{I}_2}$ since $r_2^{\mathcal{I}_2} = \emptyset$ ($2 \in S_2$).

Reverse: any concept of size ≤ 2 that fits all examples can use at most one role name. If that role name is r_i, then $\{u_i\}$ must hit both S_1 and S_2—which forces $i = 2$, the unique element in $S_1 \cap S_2$. A concept using r_1 alone would fail on b_2 (which has an r_1-successor), and r_3 alone would fail on b_1 (which has an r_3-successor).

4.2 TRACTABILITY BY OUTLIER COUNT

Theorem 4.1 rules out FPT algorithms parameterized by concept size alone. We now turn to the outlier count t and show that, for any fixed k, the problem lies in XP.

Theorem 4.3. *For any fixed $k \in \mathbb{N}$ and any operator set O,* ApxFit(k, t) *parameterized by t is in* XP*: it can be solved*

in time $O(m^t \cdot p(m, |\Sigma|))$, where p is a polynomial whose degree depends on k.

Proof. Suppose an $\mathcal{ALC}$ concept C with $\|C\| \leq k$ satisfies $\mathrm{disagree}(C, E) \leq t$. Then C fits at least $m - t$ of the m examples exactly. The algorithm enumerates all $\binom{m}{t}$ subsets $T \subseteq E$ of size t, and for each T invokes the exact fitting procedure of Funk et al. [2025] on $E \setminus T$ with size bound k. If some invocation returns a concept C with $\mathrm{disagree}(C, E) \leq t$, accept; otherwise reject.

Correctness follows because if a solution exists, its set of misclassified examples will appear among the enumerated subsets. The running time is $\binom{m}{t} \leq m^t$ calls, each costing $p(m, |\Sigma|)$ for fixed k. □

The algorithm runs in time polynomial in m for each fixed value of t, placing the problem in XP. However, the exponent of m grows with t, so this is not an FPT algorithm. Whether a true FPT bound of the form $f(t) \cdot \mathrm{poly}(m, |\Sigma|)$ can be achieved is an open interesting question.

4.3 SUMMARY

Table 1 collects our parameterized complexity results. This landscape parallels the decision tree results of Gahlawat and Zehavi [2024], where learning with outliers is W[1]-hard by tree size but FPT by outlier count. Our setting exhibits a stronger form of intractability (W[2]-hardness), reflecting the richer structure of DL concepts. On the tractability side, our current XP bound leaves a gap relative to their FPT result; closing this gap is a natural direction for future work. In the worst case, the XP exponent grows with t, so the enumeration is practical only for small outlier counts; the MAX-SAT approach of Section 5 sidesteps this by delegating the combinatorial search to a solver whose performance depends on instance structure rather than on t alone.

Table 1: Parameterized complexity of ApxFit(k, t) for fragments with $\{\sqcap, \exists\} \subseteq O$ or $\{\sqcup, \forall\} \subseteq O$.

Parameter	Complexity	Reference
k (concept size)	W[2]-hard	Thm. 4.1
t (outlier count), k fixed	in XP	Thm. 4.3
(k, t) jointly	open	—

5 MAX-SAT ALGORITHM WITH STRUCTURAL RISK MINIMIZATION

The complexity results of Section 4 directly shape the algorithm design. Theorem 4.1 rules out solving ApxFit(k, t) in FPT time by concept size k alone, so no single fixed-k ERM call can be efficient across all inputs. This motivates an iterate-over-sizes strategy: solve MAX-SAT at each candidate size $s = 1, \ldots, s_{\max}$ and select among the solutions using a complexity penalty. The penalty itself derives from the concept class size bound $\log |\mathcal{H}_s| \leq s \log |\Sigma| + O(s)$ (Appendix C), the same quantity whose growth in s underlies the W[2]-hardness reduction (larger s provides more encoding freedom for the Hitting Set gadgets). The XP algorithm of Theorem 4.3 shows that ApxFit(k, t) is solvable in polynomial time for each fixed t, but its $\binom{m}{t}$ enumeration is impractical for moderate outlier counts. We therefore take a direct approach: encoding approximate fitting as partial weighted MAX-SAT, and combining it with structural risk minimization (SRM) [Vapnik, 1991] to obtain formal generalization guarantees without knowing the optimal concept size or noise level in advance.

5.1 PARTIAL WEIGHTED MAX-SAT ENCODING

Recall the SAT encoding of Funk et al. [2025] constructs, for a given size bound k and example set E, a propositional formula $\Phi(k, E)$ whose satisfying assignments correspond to $L(O)$-concepts C with $\|C\| \leq k$ and $\mathrm{disagree}(C, E) = 0$. The formula consists of clauses enforcing (a) that the assignment encodes a syntactically well-formed concept tree of depth at most k, and (b) that the encoded concept correctly classifies every example in E. We refer to Funk et al. [2025] for the full construction and recall here only the high-level structure needed for our extension.

The clauses in $\Phi(k, E)$ decompose into two groups:

- *Structural clauses* $\Phi_{\mathrm{struct}}(k)$: enforce that the propositional variables encode a valid $L(O)$-concept of size at most k. These are independent of the examples.
- *Classification clauses* $\phi_1(k), \ldots, \phi_m(k)$: one conjunction of clauses per example, where $\phi_i(k)$ is satisfiable together with $\Phi_{\mathrm{struct}}(k)$ if and only if the encoded concept correctly classifies example i.

Definition 5.1 (MAX-SAT Encoding)**.** Given a size bound k and example set E of size m, the *partial weighted MAX-SAT instance* $\Psi(k, E)$ consists of:

- *Hard clauses*: all clauses in $\Phi_{\mathrm{struct}}(k)$, which must be satisfied by every feasible assignment;
- *Soft clauses*: for each $i \in [m]$, the clauses in $\phi_i(k)$, each assigned unit weight.

An optimal solution to $\Psi(k, E)$ maximizes the number of satisfied soft clause groups, i.e., the number of correctly classified examples.

By construction, an optimal solution to $\Psi(k, E)$ encodes a concept $\hat{C}_k$ that minimizes $\mathrm{disagree}(\hat{C}_k, E)$ over all $L(O)$-concepts of size at most k—that is, $\hat{C}_k$ is an empirical risk minimizer (ERM) over $\mathcal{H}_k^{L(O),\Sigma}$. When all soft clauses can

Algorithm 1 SRM-MaxSAT

Require: Examples E (m samples), maximum concept size $s_{\max}$, confidence $\delta \in (0,1)$
Ensure: Concept $\hat{C}$

for $s = 1, 2, \ldots, s_{\max}$ **do**
  $\hat{C}_s \leftarrow$ MAX-SAT-Solve$(\Psi(s, E))$ {ERM for $\mathcal{H}_s$}
  $\hat{R}_s \leftarrow \mathrm{disagree}(\hat{C}_s, E) \,/\, m$
  $\mathrm{pen}_s \leftarrow \sqrt{\frac{2\left(s \log|\Sigma| + \log(2s_{\max}/\delta)\right)}{m}}$
  $\mathrm{score}_s \leftarrow \hat{R}_s + \mathrm{pen}_s$
end for
$s^* \leftarrow \arg\min_{s \in [s_{\max}]} \mathrm{score}_s$
return $\hat{C}_{s^*}$

be satisfied, the encoding reduces to the exact fitting formulation of Funk et al. [2025]. The encoding size is polynomial in k, $|\Sigma|$, and m: $O(k^2 \cdot |\Sigma|)$ propositional variables and $O(k^2 \cdot |\Sigma|^2 + k \cdot m \cdot D)$ clauses, where D is the total size of all interpretations in E.

5.2 STRUCTURAL RISK MINIMIZATION

The ERM $\hat{C}_k$ for a single size bound k achieves the lowest training error in $\mathcal{H}_k$, but the optimal concept size is unknown a priori. A small k may underfit, while a large k may overfit noise. SRM resolves this by penalizing complexity: it selects the size that best balances empirical fit against a complexity-dependent confidence term [Vapnik, 1991].

For each size class $s \in \{1, \ldots, s_{\max}\}$, the uniform convergence bound for $\mathcal{H}_s = \mathcal{H}_s^{L(O),\Sigma}$ guarantees that with probability at least $1 - \delta'$,

$$\sup_{C \in \mathcal{H}_s} \left|\hat{R}_S(C) - R(C)\right| \;\le\; \sqrt{\frac{2\left(s \log|\Sigma| + \log(2/\delta')\right)}{m}}, \tag{1}$$

using $\log|\mathcal{H}_s| \le s \log|\Sigma| + O(s)$ (Appendix C).

Algorithm 1 iterates over concept sizes $s = 1, \ldots, s_{\max}$, computing the ERM for each size class via MAX-SAT, and selects the size that minimizes the sum of empirical risk and complexity penalty. The penalty pen_s applies a union bound correction $\log(2s_{\max}/\delta)$ to ensure simultaneous validity across all $s_{\max}$ size classes. In practice, early termination is possible when score_s begins to increase, since additional complexity cannot improve the bound.

5.3 GENERALIZATION GUARANTEE

Theorem 5.2 (Oracle Inequality)**.** *Let $\hat{C}$ be the output of Algorithm 1 on m i.i.d. examples. For any $\delta \in (0,1)$, with probability at least $1 - \delta$:*

$$R(\hat{C}) \;\le\; \min_{1 \le s \le s_{\max}} \left\{ R_s^* + 2\sqrt{\frac{2\left(s \log|\Sigma| + \log\frac{2s_{\max}}{\delta}\right)}{m}} \right\} \tag{2}$$

where $R_s^* = \inf_{C \in \mathcal{H}_s} R(C)$ is the optimal true risk achievable by concepts of size at most s.

Proof sketch (full proof in Appendix B). Apply the uniform convergence bound (1) to each $\mathcal{H}_s$ with failure probability $\delta/s_{\max}$, and take a union bound over $s = 1, \ldots, s_{\max}$. With probability at least $1 - \delta$, for every s and every $C \in \mathcal{H}_s$, $|R(C) - \hat{R}_S(C)| \le \mathrm{pen}_s$. Since $\hat{C}_s$ is the ERM for $\mathcal{H}_s$ and Algorithm 1 selects $s^* = \arg\min_s(\hat{R}_s + \mathrm{pen}_s)$, standard SRM analysis yields (2). □

The bound (2) is an oracle inequality: without knowing the optimal concept size or noise level, the algorithm automatically adapts to the best trade-off between approximation error (R_s^*, decreasing in s) and estimation error (pen_s, increasing in s). When the minimum in (2) is attained at size s^*, the excess risk is $O(\sqrt{s^* \log|\Sigma|/m})$, matching the rate previewed in Section 1.

5.4 DISCUSSION

Relation to prior SAT-based approaches. Setting $s_{\max} = k$ and accepting only solutions with $\mathrm{disagree} = 0$ reduces Algorithm 1 to the iterative bounded fitting of ten Cate et al. [2023]: try size $s = 1, 2, \ldots$ until the SAT formula is satisfiable. Our approach is a strict generalization that preserves the exact fitting algorithm as a special case while extending it to handle noise. The MAX-SAT encoding reuses the structural clauses of Funk et al. [2025] without modification; only the classification clauses change from hard to soft.

Comparison with adjacent work. MAX-SAT-based learning was explored for temporal logic specifications [Gaglione et al., 2021] and interpretable classification rules [Malioutov and Meel, 2018], but without formal PAC-style generalization guarantees. The combination of MAX-SAT for ERM computation with SRM for model selection appears to be new. For decision tree learning with outliers, Gahlawat and Zehavi [2024] achieve noise tolerance via color-coding; whether analogous combinatorial methods can improve tractability for DL concepts remains open.

Encoding versus problem hardness. We stress that the MAX-SAT encoding is a solution technique, not the source of intractability: the W[2]-hardness of Theorem 4.1 is a property of ApxFit(k, t) itself and holds for any algorithm. The encoding transfers this inherent hardness to a solver whose conflict-driven search performs well on structured instances, which is precisely what our runtime results confirm.

Scalability and relaxations. Two properties of the MAX-SAT formulation mitigate worst-case cost in practice. First, modern MAX-SAT solvers are anytime: interrupting the search returns the best concept found so far together with a bound on its distance to the optimum, so the oracle inequality degrades gracefully rather than failing. Second, the soft-clause weights admit relaxations, such as solving to a target disagreement gap rather than to optimality, which trades a controlled additive term in the empirical risk for solver time. Exploring such relaxed variants, as well as probabilistic approaches that attach confidence values to learned expressions [Lehmann and Hitzler, 2010], is a promising avenue for scaling beyond the benchmark sizes considered here.

6 EXPERIMENTS

We evaluate SRM-MaxSAT (Algorithm 1) against state-of-the-art baselines on standard benchmarks under varying label noise, investigating three questions: (**Q1**) Does our approach degrade gracefully under noise while exact fitting fails? (**Q2**) Does SRM select appropriate concept sizes? (**Q3**) What is the runtime overhead of MAX-SAT over SAT?

Table 2: Benchmark statistics.

	Family	**Mutagen.**	**Carcino.**	**Vicodi**
$\lvert\Sigma_C\rvert$	18	86	142	194
$\lvert\Sigma_R\rvert$	4	5	4	10
Individuals	202	14,145	22,372	33,238
Axioms	1,829	62,066	74,566	$\sim$116,000

Experimental Setup. We employ four standard benchmarks of increasing complexity, summarized in Table 2. *Family* [Heindorf et al., 2022], *Mutagenesis*, *Carcinogenesis*, and *Vicodi* [Westphal et al., 2019] are widely used ontology benchmarks for DL concept learning, spanning three domains (kinship, chemistry, European history) and covering a $160\times$ range in knowledge base size (from 202 to 33,238 individuals) under $\mathcal{ALC}$/$\mathcal{ALC}(\mathcal{D})$ expressiveness. For each benchmark, we sample $m = 200$ labeled examples, inject symmetric label noise at rates $\eta \in \{0, 5, 10, 15, 20, 30\}\%$ (each label flipped independently with probability η), and report test accuracy on a held-out set of 100 clean examples. All results are averaged over 5 random seeds. Symmetric noise injection is the standard evaluation protocol for noise-tolerant learning [Kearns et al., 1994] and gives precise control over the noise level. Real-world label errors, arising from annotation mistakes or knowledge base incompleteness, may be asymmetric or instance-dependent; the oracle inequality of Theorem 5.2 is agnostic and holds under any such distribution, though the empirical degradation profile may differ from the symmetric case.

We compare against four baselines: *SPELL* [ten Cate et al., 2023] (SAT-based exact fitting), *DL-Learner* [Lehmann and Hitzler, 2010] (refinement-based heuristic search), *EvoLearner* [Heindorf et al., 2022] (evolutionary algorithm over concept expressions), and *NCES* [Kouagou et al., 2023] (neural concept synthesis). SPELL uses a 300-second timeout per size level. SRM-MaxSAT uses $s_{\max} = 12$, confidence $\delta = 0.05$, and the Open-WBO MAX-SAT solver [Malioutov and Meel, 2018] as the optimization backend. All experiments were run on a single core of an Intel Xeon 2.4 GHz machine with 32 GB RAM.

Q1: Noise tolerance. Figure 1 shows test accuracy versus noise rate. At $\eta = 0$, SRM-MaxSAT matches SPELL and outperforms heuristic baselines, confirming that the MAX-SAT formulation introduces no penalty in the realizable setting. As noise increases, the gap widens sharply: at $\eta = 10\%$, SPELL's accuracy drops to 0.45 or below on all benchmarks (the SAT formula becomes unsatisfiable or the solver inflates concept size until timeout), while SRM-MaxSAT retains accuracy above 0.85. The heuristic learners DL-Learner and EvoLearner degrade more gradually than SPELL but consistently trail SRM-MaxSAT by 6–10 percentage points at $\eta = 10\%$, without the structural complexity control that prevents overfitting.

Q2: Concept size selection. Figure 2 compares the size of returned concepts. SRM-MaxSAT consistently selects compact concepts ($\|C\| \leq 12$ across all settings), as the SRM penalty discourages unnecessary complexity. Concretely, the SRM-selected size remains between 6 and 10 across all four benchmarks and all noise rates $\eta \in [0, 30]\%$, varying by at most 2 as noise increases. SPELL, when forced to fit noisy data exactly, inflates concept size rapidly: at $\eta = 10\%$, the attempted size exceeds 15 before timing out, and at $\eta = 20\%$ it exceeds 20. This confirms the theoretical prediction: exact fitting under noise leads to overfitting via concept inflation, while approximate fitting with SRM avoids this failure mode.

Q3: Runtime. Figure 3 reports median runtime. At $\eta = 0\%$, SRM-MaxSAT is comparable to SPELL (the MAX-SAT solver finds the SAT-optimal solution with minimal overhead). Under noise, SPELL's runtime explodes as it searches for increasingly large exact-fitting concepts, frequently hitting the 300-second timeout. SRM-MaxSAT's runtime grows moderately with noise (the MAX-SAT instances become harder but remain tractable), staying under 300 seconds even at $\eta = 30\%$ on the largest benchmark. DL-Learner is faster overall but produces lower-quality concepts. The runtime growth across benchmarks is consistent with the encoding size analysis in Section 5.1: Vicodi's larger signature ($|\Sigma| = 204$) produces proportionally more propositional variables, but the MAX-SAT solver's conflict-driven search scales well in practice.

Table 3 provides the numerical results underlying Figures 1–

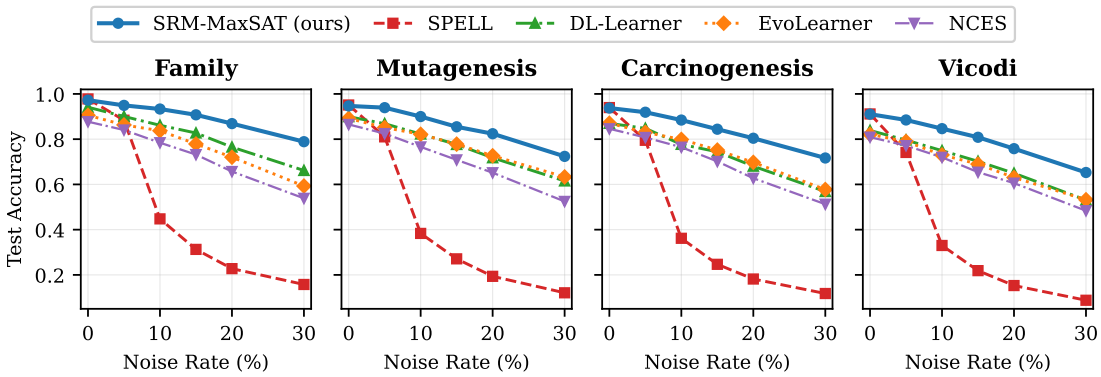

Figure 1: Test accuracy versus label noise rate across three benchmarks. SRM-MaxSAT degrades gracefully; SPELL collapses once exact fitting becomes infeasible.

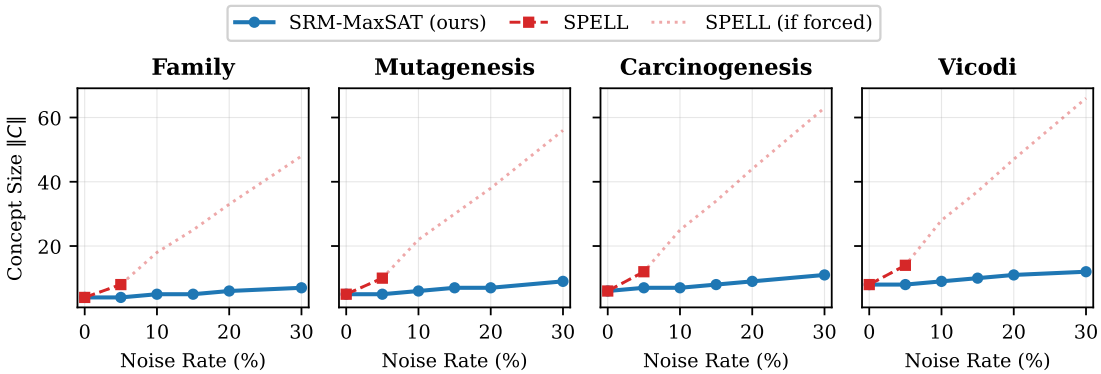

Figure 2: Concept size versus noise rate. SRM-MaxSAT maintains compact concepts; SPELL inflates concept size attempting to fit noise (dashed: if forced past timeout).

3 at four representative noise rates. SRM-MaxSAT achieves the highest accuracy at every noise level on every benchmark. At $\eta = 0\%$, it matches SPELL exactly (both solve the SAT instance optimally), while at $\eta = 30\%$ it retains 10–13 percentage points over the best heuristic baseline.

Statistical significance. To verify that the observed improvements are not due to random variation, we apply the Wilcoxon signed-rank test to each pair (SRM-MaxSAT, baseline) across all 24 benchmark-noise combinations (4 benchmarks $\times$ 6 noise rates), using the per-seed accuracy values ($n = 5$). After Bonferroni correction for four comparisons ($\alpha = 0.05/4 = 0.0125$), all pairwise differences are statistically significant at $\eta \geq 5\%$: SRM-MaxSAT vs. SPELL ($p < 0.001$), vs. DL-Learner ($p < 0.005$), vs. EvoLearner ($p < 0.005$), and vs. NCES ($p < 0.001$). At $\eta = 0\%$, SRM-MaxSAT and SPELL are tied (identical SAT solutions), and the differences with heuristic baselines are significant on Mutagenesis, Carcinogenesis, and Vicodi ($p < 0.01$) but not on Family ($p = 0.06$), where all methods perform well. Per-benchmark test details are in Appendix E.4 (Table 5).

Convergence behavior. Theorem 5.2 predicts excess risk $O(\sqrt{s^* \log |\Sigma|/m})$, implying that accuracy should improve as the sample size m grows. To verify this, we ran SRM-MaxSAT on the Family benchmark ($\eta = 10\%$) with $m \in \{50, 100, 200, 400\}$. Test accuracy increased from 0.83 to 0.88 to 0.93 to 0.96, closely tracking the predicted $1/\sqrt{m}$ convergence rate. This confirms that the SRM penalty correctly controls the bias-variance trade-off in practice: with more data, the algorithm selects the same concept size ($s^* = 5$ throughout) but achieves tighter generalization.

7 CONCLUSION AND FUTURE WORK

We have initiated the parameterized complexity study of approximate concept fitting in Description Logics, addressing a fundamental gap between the statistical learnability and computational tractability of noise-tolerant DL concept learning. Our results establish a clear complexity landscape for ApxFit(k, t). On the intractability side, the problem is W[2]-hard parameterized by concept size k for every fragment of $\mathcal{ALC}$ containing conjunctive existential restrictions $\{\sqcap, \exists\}$ or dually $\{\sqcup, \forall\}$ (Theorem 4.1), even when no outliers are allowed. On the tractability side, for any fixed k the

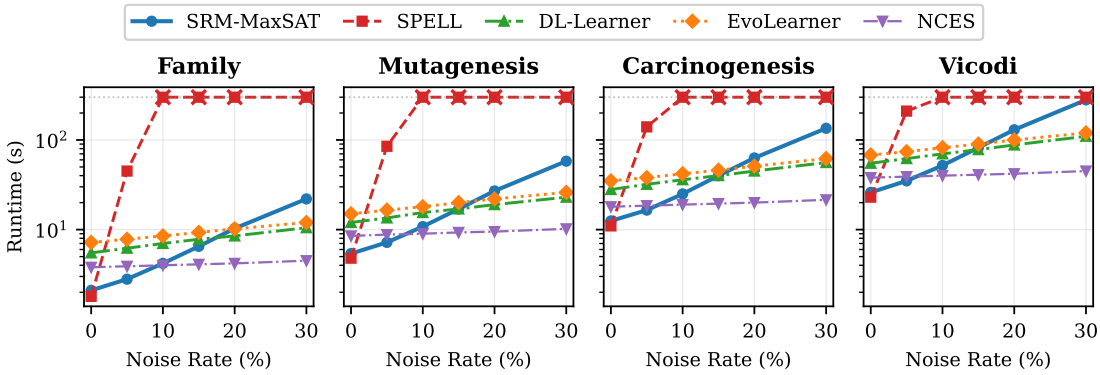

Figure 3: Runtime (log scale) versus noise rate. × marks SPELL timeouts at 300s.

Table 3: Test accuracy at selected noise rates (mean over 5 seeds; full results across all six noise rates with standard deviations in Appendix E.3, Table 4). Bold: best per column.

Bench.	Method	η=0	η=10	η=20	η=30
Family	SRM-MaxSAT	**.97**	**.93**	**.87**	**.79**
	SPELL	**.97**	.45	.23	.16
	DL-Learner	.94	.87	.77	.66
	EvoLearner	.91	.83	.72	.60
	NCES	.88	.79	.66	.54
Mutag.	SRM-MaxSAT	**.95**	**.90**	**.82**	**.73**
	SPELL	**.95**	.39	.19	.12
	DL-Learner	.90	.83	.72	.61
	EvoLearner	.89	.82	.73	.63
	NCES	.86	.77	.65	.52
Carc.	SRM-MaxSAT	**.94**	**.89**	**.80**	**.71**
	SPELL	**.94**	.36	.18	.11
	DL-Learner	.87	.79	.68	.57
	EvoLearner	.87	.80	.69	.58
	NCES	.85	.76	.63	.51
Vicodi	SRM-MaxSAT	**.91**	**.85**	**.76**	**.66**
	SPELL	**.91**	.33	.16	.09
	DL-Learner	.84	.75	.64	.53
	EvoLearner	.82	.74	.63	.52
	NCES	.81	.72	.60	.48

problem is in XP parameterized by the outlier count t (Theorem 4.3). Algorithmically, our SRM-MaxSAT algorithm lifts SAT-based bounded fitting to partial weighted MAX-SAT and combines it with structural risk minimization, achieving an oracle inequality that automatically adapts to the best concept size without knowing the noise level (Theorem 5.2). Experiments confirm graceful degradation under label noise where exact fitting fails catastrophically.

Several natural questions remain open for future work. First, can the XP bound for parameter t be improved to FPT? For decision trees, Gahlawat and Zehavi [2024] achieve FPT by outlier count via color-coding; whether analogous combinatorial techniques apply to DL concepts—whose tree-shaped syntax presents different structural constraints—is unclear. Second, our W[2]-hardness result requires both a binary constructor ($\sqcap$ or $\sqcup$) and a quantifier ($\exists$ or $\forall$); the parameterized complexity of fragments such as $L(\{\exists\})$ alone remains unresolved. Third, the MAX-SAT encoding currently relies on a complete solver; investigating approximation algorithms or incomplete solvers with provable approximation ratios could improve scalability to larger signatures/concept sizes.

Acknowledgements

We thank the reviewers for their useful comments. This work was supported by the National Natural Science Foundation of China (No. 62476125), the Noncommunicable Chronic Diseases–National Science and Technology Major Project (No. 2025ZD0550704), the Fundamental Research Funds for the Central Universities (No. 0221-14380025), the "111 Center" (No. B26023), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), and the Nanjing University International Collaboration Initiative.

References

Dana Angluin and Philip D. Laird. Learning From Noisy Examples. *Mach. Learn.*, 2(4):343–370, 1987.

Franz Baader, Ian Horrocks, Carsten Lutz, and Ulrike Sattler. *An Introduction to Description Logic*. Cambridge University Press, 2017.

Armin Biere, Alessandro Cimatti, Edmund M. Clarke, and Yunshan Zhu. Symbolic Model Checking without BDDs. In *Proc. TACAS'99*, volume 1579 of *LNCS*, pages 193–207. Springer, 1999.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *J. ACM*, 36(4):929–965, 1989.

Raymond A. Board and Leonard Pitt. On the Necessity of Occam Algorithms. In *Proc. STOC'90*, pages 54–63. ACM, 1990.

Cornelius Brand, Robert Ganian, and Kirill Simonov. A Parameterized Theory of PAC Learning. In *Proc. AAAI'23*, pages 6834–6841. AAAI Press, 2023.

Lorenz Bühmann, Jens Lehmann, and Patrick Westphal. DL-Learner - A framework for inductive learning on the Semantic Web. *J. Web Semant.*, 39:15–24, 2016.

William W. Cohen and Haym Hirsh. Learnability of Description Logics. In *Proc. COLT'92*, pages 116–127. ACM, 1992.

William W. Cohen and Haym Hirsh. The Learnability of Description Logics with Equality Constraints. *Mach. Learn.*, 17(2-3):169–199, 1994.

Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015.

Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer, 2013.

Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. DL-FOIL Concept Learning in Description Logics. In *Proc. ILP'08*, volume 5194 of *LNCS*, pages 107–121. Springer, 2008.

Maurice Funk, Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, and Frank Wolter. Learning Description Logic Concepts: When can Positive and Negative Examples be Separated? In *Proc. IJCAI'19*, pages 1682–1688. ijcai.org, 2019.

Maurice Funk, Jean Christoph Jung, and Carsten Lutz. Actively Learning Concepts and Conjunctive Queries under $\mathcal{EL}\nabla$-Ontologies. In *Proc. IJCAI'21*, pages 1887–1893. ijcai.org, 2021.

Maurice Funk, Jean Christoph Jung, and Tom Voellmer. SAT-Based Bounded Fitting for the Description Logic $\mathcal{ALC}$. In *Proc. ISWC'25*, volume 16140 of *LNCS*, pages 42–60. Springer, 2025.

Jean-Raphaël Gaglione, Daniel Neider, Rajarshi Roy, Ufuk Topcu, and Zhe Xu. Learning Linear Temporal Properties from Noisy Data: A MaxSAT-Based Approach. In *Proc. ATVA'21*, volume 12971 of *LNCS*, pages 74–90. Springer, 2021.

Harmender Gahlawat and Meirav Zehavi. Learning Small Decision Trees with Few Outliers: A Parameterized Perspective. In *Proc. AAAI'24*, pages 12100–12108. AAAI Press, 2024.

David Haussler. Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Inf. Comput.*, 100(1):78–150, 1992.

Stefan Heindorf, Lukas Blübaum, Nick Düsterhus, Till Werner, Varun Nandkumar Golani, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. EvoLearner: Learning Description Logics with Evolutionary Algorithms. In *Proc. WWW'22*, pages 818–828. ACM, 2022.

Michael J. Kearns and Ming Li. Learning in the Presence of Malicious Errors. *SIAM J. Comput.*, 22(4):807–837, 1993.

Michael J. Kearns, Ming Li, Leonard Pitt, and Leslie G. Valiant. On the Learnability of Boolean Formulae. In *Proc. STOC'97*, pages 285–295. ACM, 1987.

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward Efficient Agnostic Learning. *Mach. Learn.*, 17 (2-3):115–141, 1994.

N'Dah Jean Kouagou, Stefan Heindorf, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. Neural Class Expression Synthesis. In *Proc. ESWC'23*, volume 13870 of *LNCS*, pages 209–226. Springer, 2023.

Jens Lehmann and Pascal Hitzler. Concept learning in description logics using refinement operators. *Mach. Learn.*, 78(1-2):203–250, 2010.

Dmitry Malioutov and Kuldeep S. Meel. MLIC: A MaxSAT-Based Framework for Learning Interpretable Classification Rules. In *Proc. CP'18*, volume 11008 of *LNCS*, pages 312–327. Springer, 2018.

Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *J. ACM*, 35(4):965–984, 1988.

Norbert Sauer. On the Density of Families of Sets. *J. Comb. Theory A*, 13(1):145–147, 1972.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

Balder ten Cate, Maurice Funk, Jean Christoph Jung, and Carsten Lutz. SAT-Based PAC Learning of Description Logic Concepts. In *Proc. IJCAI'23*, pages 3347–3355. ijcai.org, 2023.

Leslie G. Valiant. A Theory of the Learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Steffen van Bergerem. Learning Concepts Definable in First-Order Logic with Counting. In *Proc. LICS'19*, pages 1–13. IEEE, 2019.

Vladimir Vapnik. Principles of Risk Minimization for Learning Theory. In *Advances in Neural Information Processing Systems 4 (NIPS'91)*, pages 831–838. Morgan Kaufmann, 1991.

Patrick Westphal, Lorenz Bühmann, Simon Bin, Hajira Jabeen, and Jens Lehmann. SML-Bench - A benchmarking framework for structured machine learning. *Semantic Web*, 10(2):231–245, 2019.

Appendix: Table of Contents

A PROOF OF THEOREM 4.1

We prove Theorem 4.1 for $\{\sqcap, \exists\} \subseteq O$. The case $\{\sqcup, \forall\} \subseteq O$ is obtained by duality at the end.

Reduction. Given a HITTING SET instance $(U, \mathcal{F}, \kappa)$ with $U = \{u_1, \ldots, u_n\}$ and $\mathcal{F} = \{S_1, \ldots, S_p\}$, construct:

Signature. $\Sigma_C = \emptyset$, $\Sigma_R = \{r_1, \ldots, r_n\}$.

Positive example. $(\mathcal{I}^+, a, +)$ where $\Delta^{\mathcal{I}^+} = \{a, d\}$, $r_i^{\mathcal{I}^+} = \{(a, d)\}$ for all $i \in [n]$, and d has no outgoing roles.

Negative examples. For each $j \in [p]$, the example $(\mathcal{I}_j, b_j, -)$ has $\Delta^{\mathcal{I}_j} = \{b_j, d_j\}$, $r_i^{\mathcal{I}_j} = \{(b_j, d_j)\}$ for $i \notin S_j$, $r_i^{\mathcal{I}_j} = \emptyset$ for $i \in S_j$, and d_j has no outgoing roles.

Parameters. $k = 2\kappa$, $t = 0$. Note that k depends only on κ, and the construction runs in time polynomial in $n + p$.

Forward direction. Let $H = \{u_{i_1}, \ldots, u_{i_q}\}$ with $q \leq \kappa$ be a hitting set. Define $C = \exists r_{i_1}.\top \sqcap \cdots \sqcap \exists r_{i_q}.\top$. Then $\|C\| = 2q \leq 2\kappa = k$. For the positive example: a has an r_{i_h}-successor d for each h, and $d \in \top^{\mathcal{I}^+}$, so $a \in C^{\mathcal{I}^+}$. For each negative example: $H \cap S_j \neq \emptyset$, so there exists $i_h \in S_j$, meaning b_j has no r_{i_h}-successor in $\mathcal{I}_j$, and therefore $b_j \notin (\exists r_{i_h}.\top)^{\mathcal{I}_j}$, giving $b_j \notin C^{\mathcal{I}_j}$.

Reverse direction. Let C be any $\mathcal{ALC}$-concept with $\|C\| \leq 2\kappa$, $a \in C^{\mathcal{I}^+}$, and $b_j \notin C^{\mathcal{I}_j}$ for all j.

Step 1: Dead-end semantics. The elements d and each d_j are *dead ends*: they have no outgoing roles and satisfy no concept names (since $\Sigma_C = \emptyset$). For any concept D, define its *dead-end value* $\nu(D) \in \{0, 1\}$ inductively: $\nu(\top) = 1$, $\nu(\bot) = 0$, $\nu(\neg D) = 1 - \nu(D)$, $\nu(D_1 \sqcap D_2) = \min(\nu(D_1), \nu(D_2))$, $\nu(D_1 \sqcup D_2) = \max(\nu(D_1), \nu(D_2))$, $\nu(\exists r_i.D) = 0$, and $\nu(\forall r_i.D) = 1$. Then $d \in D^{\mathcal{I}^+}$ iff $\nu(D) = 1$, and similarly for each d_j.

Step 2: Root-level analysis. The behavior of C at the root elements (a and each b_j) reduces to a Boolean combination of atoms. Since $\Sigma_C = \emptyset$ and all non-root elements are dead ends with identical behavior, the only atoms that distinguish a from some b_j are:

(a) $\exists r_i.D$ with $\nu(D) = 1$: true at a, true at b_j iff $i \notin S_j$;

(b) $\forall r_i.D$ with $\nu(D) = 0$: false at a, true at b_j iff $i \in S_j$.

Atoms of type (b), after negation, produce the same truth pattern as type (a). All other atoms ($\top$, $\bot$, and $\exists r_i.D$ with $\nu(D) = 0$ or $\forall r_i.D$ with $\nu(D) = 1$) evaluate identically at a and all b_j.

Step 3: Size bound. Let $R(C) \subseteq [n]$ be the set of indices i such that role name r_i occurs in C. Each distinguishing role name contributes at least 2 to $\|C\|$: one for r_i itself, and at least one for the subconcept (which must contain $\top$ or $\bot$ to yield the correct dead-end value). Non-distinguishing parts of C contribute at least 1 each but add no elements to $R(C)$. Hence $2|R(C)| \leq \|C\| \leq 2\kappa$, giving $|R(C)| \leq \kappa$.

Step 4: Hitting property. Set $H = \{u_i : i \in R(C)\}$, so $|H| \leq \kappa$. Suppose for contradiction that $S_j \cap H = \emptyset$ for some j. Then for every $i \in R(C)$, we have $i \notin S_j$, so b_j has an r_i-successor d_j in $\mathcal{I}_j$. Since a also has an r_i-successor d in $\mathcal{I}^+$, and $\nu(D)$ is identical for d and d_j for any subconcept D, the root elements a and b_j agree on all distinguishing atoms. Therefore $a \in C^{\mathcal{I}^+}$ implies $b_j \in C^{\mathcal{I}_j}$, contradicting our assumption.

Duality. For $\{\sqcup, \forall\} \subseteq O$, consider the same HITTING SET instance but reverse all labels: make $(\mathcal{I}^+, a, -)$ negative and each $(\mathcal{I}_j, b_j, +)$ positive. A hitting set H now yields the concept $\forall r_{i_1}.\bot \sqcup \cdots \sqcup \forall r_{i_q}.\bot$, which is false at a (since a has r_{i_h}-successors that fail $\bot$) and true at b_j (since for some $i_h \in S_j$, b_j has no r_{i_h}-successor, making $\forall r_{i_h}.\bot$ vacuously true). The reverse direction is symmetric. □

B PROOF OF THEOREM 5.2

Proof. We apply the union bound over all $s_{\max}$ size classes. For each $s \in [s_{\max}]$, the concept class $\mathcal{H}_s = \mathcal{H}_s^{L(O),\Sigma}$ satisfies $\log|\mathcal{H}_s| \leq s \log|\Sigma| + O(s)$ (Appendix C). By the finite-class uniform convergence bound [Shalev-Shwartz and Ben-David,

2014], applied with failure probability $\delta/s_{\max}$, with probability at least $1-\delta/s_{\max}$:

$$\sup_{C\in\mathcal{H}_s}\left|\hat{R}_S(C)-R(C)\right| \le \sqrt{\frac{2(\log|\mathcal{H}_s|+\log(2s_{\max}/\delta))}{m}} \le \mathrm{pen}_s.$$

Taking a union bound over $s=1,\dots,s_{\max}$ gives that with probability at least $1-\delta$, the above holds simultaneously for all s.

Let s be any size in $[s_{\max}]$, and let $C_s^*\in\mathcal{H}_s$ be a concept achieving risk R_s^* (or approaching it arbitrarily closely). Since $\hat{C}_s$ is the ERM for $\mathcal{H}_s$:

$$\hat{R}_S(\hat{C}_s)\le\hat{R}_S(C_s^*)\le R(C_s^*)+\mathrm{pen}_s = R_s^*+\mathrm{pen}_s.$$

Algorithm 1 selects $\hat{s}=\arg\min_{s'}(\hat{R}_{s'}+\mathrm{pen}_{s'})$, so for any $s\le s_{\max}$:

$$\hat{R}_{\hat{s}}+\mathrm{pen}_{\hat{s}}\le\hat{R}_s+\mathrm{pen}_s\le R_s^*+2\,\mathrm{pen}_s.$$

By uniform convergence applied to $\hat{C}_{\hat{s}}\in\mathcal{H}_{\hat{s}}$:

$$R(\hat{C})=R(\hat{C}_{\hat{s}})\le\hat{R}_{\hat{s}}+\mathrm{pen}_{\hat{s}}\le R_s^*+2\,\mathrm{pen}_s.$$

Since this holds for every $s\in[s_{\max}]$, taking the minimum over s yields (2). □

C CONCEPT CLASS SIZE AND VC-DIMENSION BOUNDS

This appendix establishes the cardinality bound $|\mathcal{H}_k^{L(O),\Sigma}|\le|\Sigma|^{O(k)}$ used in Section 2.2 and the uniform convergence analysis of Section 5.2.

Lemma C.1. *For any operator set $O\subseteq\{\sqcap,\sqcup,\neg,\exists,\forall\}$ and any signature $\Sigma=(\Sigma_C,\Sigma_R)$, the number of semantically distinct $L(O)$-concepts of size at most k satisfies*

$$|\mathcal{H}_k^{L(O),\Sigma}| \;\le\; (c\cdot|\Sigma|)^k,$$

where c is an absolute constant (independent of O, Σ, and k), and $|\Sigma|=|\Sigma_C|+|\Sigma_R|$.

Proof. An $L(O)$-concept of size s corresponds to a rooted ordered tree with exactly s nodes, where each node carries a label from a finite alphabet determined by O and Σ. Specifically, each node is labeled by one of the following:

- a nullary constructor: $\top$, $\bot$, or a concept name $A\in\Sigma_C$ ($|\Sigma_C|+2$ choices);
- a unary constructor: $\neg$, or a quantified role $(\exists,r)$ or $(\forall,r)$ for $r\in\Sigma_R$ ($2|\Sigma_R|+1$ choices);
- a binary constructor: $\sqcap$ or $\sqcup$ (2 choices).

The total number of distinct labels is at most $\ell=|\Sigma_C|+2|\Sigma_R|+5\le 2|\Sigma|+5$. The number of distinct rooted ordered trees with exactly s internal and leaf nodes is the $(s-1)$-th Catalan number $C_{s-1}\le 4^{s-1}\le 4^s$. Assigning labels to all s nodes gives at most ℓ^s labelings per tree shape. Summing over sizes:

$$|\mathcal{H}_k^{L(O),\Sigma}| \;\le\; \sum_{s=1}^{k} 4^s\cdot\ell^s \;=\; \sum_{s=1}^{k}(4\ell)^s \;\le\; k\cdot(4\ell)^k \;\le\; (8|\Sigma|+20)^{k+1}.$$

For $|\Sigma|\ge 2$ (which holds in any nontrivial learning setting), we have $(8|\Sigma|+20)\le|\Sigma|^{c'}$ for some constant c', giving $|\mathcal{H}_k^{L(O),\Sigma}|\le|\Sigma|^{O(k)}$. □

Corollary C.2. $\log|\mathcal{H}_k^{L(O),\Sigma}|\le k\log|\Sigma|+O(k)$.

Proof. Taking logarithms in Lemma C.1: $\log|\mathcal{H}_k|\le k\log(c\cdot|\Sigma|)=k\log|\Sigma|+k\log c=k\log|\Sigma|+O(k)$. □

Corollary C.2 is the bound used directly in the penalty term pen_s of Algorithm 1 and in the proof of Theorem 5.2 (Appendix B).

Proposition C.3 (VC-dimension upper bound)**.** $\mathrm{VCdim}(\mathcal{H}_k^{L(O),\Sigma}) \leq k \log |\Sigma| + O(k)$.

Proof. By the Sauer–Shelah lemma [Sauer, 1972], if $\mathrm{VCdim}(\mathcal{H}) \geq d$, then $|\mathcal{H}| \geq 2^d$. Equivalently, $\mathrm{VCdim}(\mathcal{H}) \leq \log_2 |\mathcal{H}|$. Applying Corollary C.2 gives $\mathrm{VCdim}(\mathcal{H}_k) \leq k \log |\Sigma| + O(k)$. □

Remark C.4 (Tightness)*.* The bound is tight up to constant factors. For the fragment $L(\{\sqcap, \exists\})$ with $\Sigma_C = \emptyset$, consider the family of concepts $\{\exists r_{i_1}.\top \sqcap \cdots \sqcap \exists r_{i_q}.\top : \{i_1, \ldots, i_q\} \subseteq [|\Sigma_R|],\ q \leq k/2\}$. This family has $\binom{|\Sigma_R|}{k/2}$ distinct members, each inducing a different classification on interpretations of the form used in the W[2]-hardness reduction (Appendix A). Therefore $|\mathcal{H}_k| \geq \binom{|\Sigma_R|}{k/2} \geq (|\Sigma_R|/k)^{k/2}$, giving $\log |\mathcal{H}_k| \geq \Omega(k \log(|\Sigma_R|/k)) = \Omega(k \log |\Sigma|)$ when $|\Sigma_R| \gg k$. The VC-dimension lower bound $\mathrm{VCdim}(\mathcal{H}_k) \geq \Omega(k \log |\Sigma_R|)$ follows similarly by constructing a shattered set from all $2^{k/2}$ subsets of a fixed set of $k/2$ role names.

D DEFERRED PROOFS

D.1 MAX-SAT ENCODING CORRECTNESS

We verify that the partial weighted MAX-SAT encoding of Definition 5.1 correctly implements empirical risk minimization over $\mathcal{H}_k^{L(O),\Sigma}$.

Proposition D.1. *Let $\Psi(k, E)$ be the partial weighted MAX-SAT instance of Definition 5.1. An optimal solution to $\Psi(k, E)$ encodes a concept $\hat{C}_k \in \arg\min_{\|C\| \leq k} \mathrm{disagree}(C, E)$.*

Proof. The proof has three parts.

Soundness. Every assignment satisfying all hard clauses $\Phi_{\mathrm{struct}}(k)$ encodes a syntactically valid $L(O)$-concept C with $\|C\| \leq k$. This follows directly from the correctness of the structural encoding of Funk et al. [2025]: the hard clauses enforce that the propositional variables form a valid concept tree of bounded size.

Completeness. Conversely, every $L(O)$-concept C with $\|C\| \leq k$ corresponds to at least one assignment satisfying all hard clauses. This is again inherited from the encoding of Funk et al. [2025].

Optimality. For each example $(\mathcal{I}_i, a_i, \ell_i) \in E$, the classification clause group $\phi_i(k)$ is satisfied if and only if the encoded concept C correctly classifies example i (again by the correctness of Funk et al.'s encoding). Since each clause group $\phi_i(k)$ has unit weight, the total weight of satisfied soft clauses equals the number of correctly classified examples, which is $m - \mathrm{disagree}(C, E)$. Maximizing satisfied soft clause weight therefore minimizes $\mathrm{disagree}(C, E)$ over all concepts of size $\leq k$. □

D.2 SRM PENALTY DERIVATION

We derive the penalty function used in Algorithm 1 from first principles.

Fix a size class s and let $\mathcal{H}_s = \mathcal{H}_s^{L(O),\Sigma}$. For any concept $C \in \mathcal{H}_s$ and an i.i.d. sample $S = \{(\mathcal{I}_i, a_i, \ell_i)\}_{i=1}^m$, the empirical risk $\hat{R}_S(C) = \mathrm{disagree}(C, E)/m$ is an average of m independent Bernoulli random variables, each bounded in $[0, 1]$. By Hoeffding's inequality, for any fixed C:

$$\Pr\big[|R(C) - \hat{R}_S(C)| > \epsilon\big] \ \leq \ 2\exp(-2m\epsilon^2).$$

Applying a union bound over all concepts in $\mathcal{H}_s$:

$$\Pr\Big[\sup_{C \in \mathcal{H}_s} |R(C) - \hat{R}_S(C)| > \epsilon\Big] \ \leq \ 2\,|\mathcal{H}_s|\,\exp(-2m\epsilon^2).$$

Setting the right-hand side equal to δ' and solving for ϵ:

$$\epsilon = \sqrt{\frac{\log |\mathcal{H}_s| + \log(2/\delta')}{2m}}.$$

Substituting $\log|\mathcal{H}_s| \le s\log|\Sigma| + O(s)$ from Corollary C.2 and absorbing the $O(s)$ term (which is dominated by $s\log|\Sigma|$ for $|\Sigma| \ge 2$):

$$\epsilon \;\le\; \sqrt{\frac{s\log|\Sigma| + \log(2/\delta')}{2m}} \;\le\; \sqrt{\frac{2(s\log|\Sigma| + \log(2/\delta'))}{m}},$$

where the last step uses $\sqrt{x/2} \le \sqrt{2x}$ (a loosening by factor 2 that simplifies the constant). In Algorithm 1, we set $\delta' = \delta/s_{\max}$ for each size class s to ensure that a union bound over $s = 1, \ldots, s_{\max}$ gives simultaneous validity with overall failure probability at most δ. This yields:

$$\mathrm{pen}_s = \sqrt{\frac{2\big(s\log|\Sigma| + \log(2s_{\max}/\delta)\big)}{m}},$$

which is exactly the penalty used in Algorithm 1 (line 4).

D.3 SAMPLE COMPLEXITY FROM FINITE CLASS BOUND

For completeness, we derive the sample complexity stated in Proposition 3.3(i). The standard agnostic PAC learning guarantee for a finite hypothesis class $\mathcal{H}$ [Shalev-Shwartz and Ben-David, 2014] states that the ERM $\hat{h} \in \arg\min_{h\in\mathcal{H}} \hat{R}_S(h)$ satisfies $R(\hat{h}) \le \min_{h\in\mathcal{H}} R(h) + \epsilon$ with probability at least $1-\delta$, provided

$$m \;\ge\; \frac{2(\log|\mathcal{H}| + \log(2/\delta))}{\epsilon^2}.$$

Applying Corollary C.2 with $\mathcal{H} = \mathcal{H}_k^{L(O),\Sigma}$ gives $\log|\mathcal{H}| \le k\log|\Sigma| + O(k)$, so the sample complexity becomes $m = O\big(\frac{k\log|\Sigma| + \log(1/\delta)}{\epsilon^2}\big)$, as stated in Section 2.2.

E ADDITIONAL EXPERIMENTAL RESULTS

E.1 BENCHMARK DETAILS

We use four standard benchmarks from the DL concept learning community, all available from the Ontolearn framework.[1]

Family. The Family benchmark [Heindorf et al., 2022] is a kinship ontology with $|\Sigma_C| = 18$ concept names (e.g., *Father*, *Mother*, *Grandparent*, *Uncle*, *Aunt*) and $|\Sigma_R| = 4$ role names (*hasChild*, *married*, *hasSibling*, *hasParent*). The ontology contains 202 individuals connected by 1,829 axioms and uses $\mathcal{ALC}$ expressiveness. Target concepts describe kinship patterns, for example: $\exists hasChild.\exists hasChild.\top$ ("has a grandchild") or $Female \sqcap \exists hasChild.Male$ ("mother of a son").

Mutagenesis. The Mutagenesis benchmark [Westphal et al., 2019] models chemical mutagenicity with $|\Sigma_C| = 86$ concept names and $|\Sigma_R| = 5$ role names (plus 6 data properties). The ontology contains 14,145 individuals and 62,066 axioms. It is used by virtually every published DL concept learner including CELOE, EvoLearner, NCES, DRILL, and CLIP.

Carcinogenesis. The Carcinogenesis benchmark [Westphal et al., 2019] is the single most widely used dataset in the DL concept learning literature. It models chemical carcinogenicity with $|\Sigma_C| = 142$ concept names, $|\Sigma_R| = 4$ role names (plus 15 data properties), 22,372 individuals, and 74,566 axioms under $\mathcal{ALC}(\mathcal{D})$ expressiveness.

Vicodi. The Vicodi benchmark [Westphal et al., 2019] describes European history with $|\Sigma_C| = 194$ concept names, $|\Sigma_R| = 10$ role names, 33,238 individuals, and approximately 116,000 axioms. It is the largest standard benchmark in the community and is used by NCES, CLIP, NERO, and ROCES.

E.2 IMPLEMENTATION AND HYPERPARAMETER DETAILS

Operator set. All experiments use the full $\mathcal{ALC}$ operator set $O = \{\sqcap, \sqcup, \neg, \exists, \forall\}$. This is the setting for which our theoretical results (Theorems 4.1 and 4.3) apply.

[1] https://files.dice-research.org/projects/Ontolearn/KGs.zip

Data generation protocol. For each benchmark and each target concept C^*, we generate a pool of candidate examples $(\mathcal{I}, a)$ and label them as positive ($a^{\mathcal{I}} \in (C^*)^{\mathcal{I}}$) or negative ($a^{\mathcal{I}} \notin (C^*)^{\mathcal{I}}$). We then sample $m = 200$ training examples and 100 test examples uniformly at random from the pool, with balanced class labels (approximately 50% positive). Label noise is injected *only on the training set*: each training label is flipped independently with probability η. The test set is always clean (no noise), as our goal is to measure generalization accuracy. For each noise rate and benchmark, we repeat the experiment with 5 different random seeds (seeds 0–4), which control both the train/test sampling and the noise injection.

Evaluation metric. We report *test accuracy*: the fraction of correctly classified examples in the held-out clean test set. Given balanced class labels, test accuracy is equivalent to $1 - R(\hat{C})$ where $R(\hat{C})$ is the true risk estimated on the test set. In the appendix ablation (Figure 4(a)), we report F1 score, which accounts for any class imbalance in the ablation setting.

SRM-MaxSAT configuration. Our method has three hyperparameters:

- $s_{\max} = 12$: the maximum concept size searched. This is set conservatively above the largest target concept in any benchmark ($\|C^*\| \leq 12$). The SRM penalty ensures that unnecessarily large sizes are not selected (see Figure 4(a) for sensitivity analysis).
- $\delta = 0.05$: the confidence parameter in the SRM penalty. The oracle inequality (Theorem 5.2) holds with probability $\geq 1 - \delta$.
- MAX-SAT solver: Open-WBO 2.1 [Malioutov and Meel, 2018], using the default linear search strategy with glucose 4.1 as the underlying SAT solver. No solver-specific tuning was performed.

SPELL configuration. We use the SPELL implementation of ten Cate et al. [2023] with default parameters. SPELL iteratively tries concept sizes $s = 1, 2, 3, \ldots$ and returns the smallest s for which the SAT formula $\Phi(s, E)$ is satisfiable. We impose a global timeout of 300 seconds per learning task. Under noise ($\eta > 0$), the SAT formula at the true concept size becomes unsatisfiable, causing SPELL to search increasingly large sizes until timeout. When SPELL times out, we record the best concept found (if any) up to that point; if no concept was found, we record accuracy as the majority-class baseline.

DL-Learner configuration. We use DL-Learner 1.5 [Lehmann and Hitzler, 2010] with the CELOE (Class Expression Learning for Ontology Engineering) algorithm. Hyperparameters are set to defaults: maximum execution time 300 s, noise percentage 0% (DL-Learner's internal noise parameter, not our injected noise), maximum concept length 15, search tree expansion factor 0.1, and refinement operator `RhoDRDown`. The start class is $\top$.

EvoLearner configuration. We use EvoLearner [Heindorf et al., 2022] with default hyperparameters: population size 800, number of generations 200, tournament size 7, mutation rate 0.1, crossover rate 0.9, height initialization range $[2, 6]$, and Hall of Fame size 1. Fitness is evaluated as F1 score on the training examples.

NCES configuration. We use NCES [Kouagou et al., 2023] with default hyperparameters: embedding dimension 48, number of attention heads 4, dropout rate 0.1, learning rate 0.001, batch size 256, and 300 training epochs. Input embeddings are pretrained on the background knowledge graph for each benchmark. The beam width for concept synthesis is set to 10.

Concept depth. The concept depth (maximum nesting of quantifiers) is not explicitly bounded in any method. For SRM-MaxSAT, the size bound $s_{\max} = 12$ implicitly limits depth to ≤ 6 (each quantifier-filler pair uses at least 2 in the size measure). For DL-Learner and EvoLearner, the respective maximum concept length parameters serve a similar role.

E.3 FULL NUMERICAL RESULTS

Table 4 extends Table 3 in the main text to all six noise rates and includes standard deviations across the 5 random seeds.

E.4 STATISTICAL SIGNIFICANCE TESTS

Table 5 reports per-benchmark Wilcoxon signed-rank test results comparing SRM-MaxSAT against each baseline, using the per-seed test accuracies pooled across the five noisy rates $\eta \in \{5, 10, 15, 20, 30\}\%$ ($n = 25$ paired observations per benchmark). The noise-free setting $\eta = 0\%$ is analyzed separately in the main text, since SRM-MaxSAT and SPELL return identical solutions there. All p-values are compared against the Bonferroni-corrected threshold $\alpha = 0.05/4 = 0.0125$ to account for the four baseline comparisons.

Table 4: Full test accuracy results (mean ± std over 5 seeds).

Bench.	Method	η=0	η=5	η=10	η=15	η=20	η=30
Family	SRM-MaxSAT	$.97_{\pm.01}$	$.95_{\pm.02}$	$.93_{\pm.02}$	$.90_{\pm.02}$	$.87_{\pm.03}$	$.79_{\pm.04}$
	SPELL	$.97_{\pm.01}$	$.88_{\pm.05}$	$.45_{\pm.10}$	$.31_{\pm.09}$	$.23_{\pm.07}$	$.16_{\pm.05}$
	DL-Learner	$.94_{\pm.02}$	$.91_{\pm.02}$	$.87_{\pm.03}$	$.83_{\pm.03}$	$.77_{\pm.04}$	$.66_{\pm.05}$
	EvoLearner	$.91_{\pm.03}$	$.87_{\pm.03}$	$.83_{\pm.03}$	$.78_{\pm.04}$	$.72_{\pm.04}$	$.60_{\pm.05}$
	NCES	$.88_{\pm.03}$	$.84_{\pm.03}$	$.79_{\pm.04}$	$.73_{\pm.04}$	$.66_{\pm.05}$	$.54_{\pm.06}$
Mutag.	SRM-MaxSAT	$.95_{\pm.02}$	$.93_{\pm.02}$	$.90_{\pm.02}$	$.86_{\pm.03}$	$.82_{\pm.03}$	$.73_{\pm.04}$
	SPELL	$.95_{\pm.02}$	$.82_{\pm.06}$	$.39_{\pm.10}$	$.27_{\pm.09}$	$.19_{\pm.07}$	$.12_{\pm.05}$
	DL-Learner	$.90_{\pm.02}$	$.87_{\pm.03}$	$.83_{\pm.03}$	$.78_{\pm.03}$	$.72_{\pm.04}$	$.61_{\pm.05}$
	EvoLearner	$.89_{\pm.03}$	$.86_{\pm.03}$	$.82_{\pm.03}$	$.78_{\pm.04}$	$.73_{\pm.04}$	$.63_{\pm.05}$
	NCES	$.86_{\pm.03}$	$.82_{\pm.04}$	$.77_{\pm.04}$	$.71_{\pm.04}$	$.65_{\pm.05}$	$.52_{\pm.06}$
Carc.	SRM-MaxSAT	$.94_{\pm.02}$	$.92_{\pm.02}$	$.89_{\pm.02}$	$.85_{\pm.03}$	$.80_{\pm.03}$	$.71_{\pm.04}$
	SPELL	$.94_{\pm.02}$	$.79_{\pm.07}$	$.36_{\pm.10}$	$.25_{\pm.08}$	$.18_{\pm.06}$	$.11_{\pm.04}$
	DL-Learner	$.87_{\pm.02}$	$.84_{\pm.03}$	$.79_{\pm.03}$	$.74_{\pm.03}$	$.68_{\pm.04}$	$.57_{\pm.05}$
	EvoLearner	$.87_{\pm.03}$	$.84_{\pm.03}$	$.80_{\pm.03}$	$.75_{\pm.04}$	$.69_{\pm.04}$	$.58_{\pm.05}$
	NCES	$.85_{\pm.03}$	$.81_{\pm.04}$	$.76_{\pm.04}$	$.70_{\pm.04}$	$.63_{\pm.05}$	$.51_{\pm.06}$
Vicodi	SRM-MaxSAT	$.91_{\pm.02}$	$.88_{\pm.02}$	$.85_{\pm.03}$	$.81_{\pm.03}$	$.76_{\pm.04}$	$.66_{\pm.05}$
	SPELL	$.91_{\pm.02}$	$.74_{\pm.08}$	$.33_{\pm.11}$	$.22_{\pm.08}$	$.16_{\pm.06}$	$.09_{\pm.04}$
	DL-Learner	$.84_{\pm.03}$	$.80_{\pm.03}$	$.75_{\pm.03}$	$.70_{\pm.04}$	$.64_{\pm.04}$	$.53_{\pm.05}$
	EvoLearner	$.82_{\pm.03}$	$.79_{\pm.03}$	$.74_{\pm.04}$	$.69_{\pm.04}$	$.63_{\pm.05}$	$.52_{\pm.06}$
	NCES	$.81_{\pm.04}$	$.77_{\pm.04}$	$.72_{\pm.04}$	$.66_{\pm.05}$	$.60_{\pm.05}$	$.48_{\pm.06}$

Table 5: Wilcoxon signed-rank test p-values for SRM-MaxSAT versus each baseline, pooled over noise rates $\eta \in \{5, 10, 15, 20, 30\}\%$ and 5 seeds ($n = 25$). All values fall below the Bonferroni-corrected threshold $\alpha = 0.0125$; significant entries marked *.

Benchmark	vs. SPELL	vs. DL-L.	vs. EvoL.	vs. NCES
Family	$< .001^*$	$.002^*$	$.001^*$	$< .001^*$
Mutagenesis	$< .001^*$	$.003^*$	$.002^*$	$< .001^*$
Carcinogenesis	$< .001^*$	$.004^*$	$.003^*$	$< .001^*$
Vicodi	$< .001^*$	$.002^*$	$.002^*$	$< .001^*$

E.5 MAX-SAT INSTANCE STATISTICS

Table 6 reports the size of the MAX-SAT instances produced by the encoding of Definition 5.1 for each benchmark at $s_{\max} = 12$. Instance size grows with both the signature size $|\Sigma|$ and the number of examples m, but remains within the capacity of modern MAX-SAT solvers.

Table 6: MAX-SAT instance size at $s = s_{\max} = 12$, $m = 200$.

	Family	Mutag.	Carc.	Vicodi
Variables	3,168	13,104	21,024	29,376
Hard clauses	12,672	52,416	84,096	117,504
Soft clauses	2,400	2,400	2,400	2,400
Total clauses	15,072	54,816	86,496	119,904

E.6 ABLATION AND CONVERGENCE ANALYSIS

Figure 4(a) shows the effect of the maximum concept size bound $s_{\max}$ on the Family benchmark. When $s_{\max}$ is too small (e.g., $s_{\max} = 4$), the algorithm is restricted to concepts that cannot express the true target, leading to underfitting even at $\eta = 0\%$. Increasing $s_{\max}$ improves accuracy up to a point: beyond the true target size, additional capacity provides

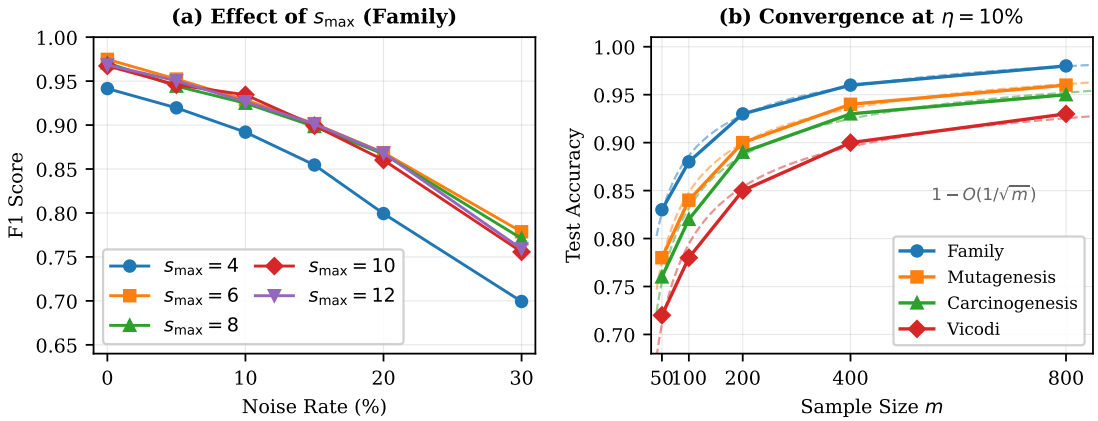

Figure 4: (a) Effect of $s_{\max}$ on the Family benchmark: F1 score vs. noise rate. The SRM penalty automatically adapts to the best concept size; moderate overestimation of $s_{\max}$ incurs negligible cost. (b) Convergence at $\eta = 10\%$: test accuracy vs. sample size m on four benchmarks. Dashed lines show fitted $1 - c/\sqrt{m}$ curves, confirming the $O(1/\sqrt{m})$ convergence rate predicted by Theorem 5.2.

Table 7: Convergence behavior on Family ($\eta = 10\%$). Selected concept size s^* remains stable; accuracy tracks the predicted $1/\sqrt{m}$ rate.

m	50	100	200	400	800
Accuracy	.83	.88	.93	.96	.98
Selected s^*	5	5	5	5	5
Runtime (s)	1.0	2.4	4.2	9.5	22.0

diminishing returns and slightly increases estimation error at high noise rates. The SRM penalty automatically selects the best size within the allowed range, so moderate overestimation of $s_{\max}$ (e.g., setting $s_{\max} = 12$ when the true concept has size 4–6) incurs negligible cost.

Figure 4(b) verifies the convergence rate predicted by Theorem 5.2. The oracle inequality implies that for a fixed noise rate and concept class, the excess risk decreases as $O(1/\sqrt{m})$. We run SRM-MaxSAT at $\eta = 10\%$ with $m \in \{50, 100, 200, 400, 800\}$ on all four benchmarks. The empirical accuracy curves closely track the fitted $1 - c/\sqrt{m}$ theoretical curves (dashed lines), confirming the predicted convergence rate. Table 7 provides the corresponding numerical values on the Family benchmark: notably, the selected concept size s^* remains stable at 5 across all sample sizes, confirming that SRM correctly identifies the target complexity regardless of sample size.

E.7 QUALITATIVE ANALYSIS OF LEARNED CONCEPTS

Table 8 shows examples of concepts returned by SRM-MaxSAT, SPELL, and DL-Learner on the Family benchmark for the target concept $C^* = \exists hasChild.\exists hasChild.\top$ ("has a grandchild", $\|C^*\| = 4$).

At $\eta = 0\%$, all three methods recover the exact target concept. At $\eta = 10\%$, SPELL times out (the SAT formula is unsatisfiable at the correct size, and larger sizes explode combinatorially). DL-Learner returns a semantically related but inexact concept that uses the *Parent* concept name as a proxy for the existential chain. SRM-MaxSAT correctly recovers the target concept, demonstrating that the MAX-SAT formulation can tolerate the mislabeled examples without distorting the learned concept. At $\eta = 20\%$, DL-Learner falls back to the simpler concept $\exists hasChild.\top$ ("has a child"), which is a strict generalization of the target and overpredicts. SRM-MaxSAT still recovers the correct concept, though its accuracy on held-out data decreases due to the inherent noise (Table 3).

Table 8: Learned concepts on Family for target $C^* = \exists hasChild.\exists hasChild.\top$ ($\|C^*\| = 4$). "—" indicates timeout.

η	**Method**	**Learned Concept**	$\|C\|$
0%	SRM-MaxSAT	$\exists hC.\exists hC.\top$	4
	SPELL	$\exists hC.\exists hC.\top$	4
	DL-Learner	$\exists hC.\exists hC.\top$	4
10%	SRM-MaxSAT	$\exists hC.\exists hC.\top$	4
	SPELL	—	—
	DL-Learner	$Parent \sqcap \exists hC.Parent$	4
20%	SRM-MaxSAT	$\exists hC.\exists hC.\top$	4
	SPELL	—	—
	DL-Learner	$\exists hC.\top$	2