Query-Based Entailment and Inseparability for \textit{ALC} Ontologies

Elena Botoeva,$^1$ Carsten Lutz,$^2$ Vladislav Ryzhikov,$^1$ Frank Wolter$^3$ and Michael Zakharyaschev$^4$

$^1$Faculty of Computer Science, Free University of Bozen-Bolzano
$^2$Fachbereich Informatik, University of Bremen
$^3$Department of Computer Science, University of Liverpool
$^4$Department of Computer Science, Birkbeck, University of London

Abstract

We investigate the problem whether two \textit{ALC} knowledge bases are indistinguishable by queries over a given vocabulary. We give model-theoretic criteria and prove that this problem is undecidable for conjunctive queries (CQs) but decidable in 2\textsc{Exptime} for unions of rooted CQs. We also consider the problem whether two \textit{ALC} TBoxes give the same answers to any query in a given vocabulary over all ABoxes, and show that for CQs this problem is undecidable, too, but becomes decidable and 2\textsc{Exptime}-complete in Horn-\textit{ALC}, and even \textsc{Exptime}-complete in Horn-\textit{ALC} when restricted to (unions of) rooted CQs.

1 Introduction

In recent years, data access using description logic (DL) TBoxes has become one of the most important applications of DLs [Poggi et al., 2008; Bienvenu and Ortiz, 2015], where the underlying idea is to use a TBox to specify semantics and background knowledge for the data (stored in an ABox), and thereby derive more complete query answers. A major research effort has led to the development of efficient algorithms and tools for a number of DLs ranging from DL-Lite [Calvanese et al., 2007; Rodriguez-Muro et al., 2013] via more expressive Horn DLs such as Horn-\textit{ALC} [Eiter et al., 2012; Trivela et al., 2015] to DLs with all Boolean constructors such as \textit{ALC} [Kollia and Glimm, 2013; Zhou et al., 2015].

While query answering with DLs is now well-developed, this is much less the case for reasoning services that support ontology engineering and target query answering as an application. In ontology versioning, for example, one would like to know whether two versions of an ontology give the same answers to all queries formulated over a given vocabulary of interest, which means that the newer version can safely replace the older one [Konev et al., 2012]. Similarly, if one wants to know whether an ontology can be safely replaced by a smaller subset (module), it is the answers to all queries that should be preserved [Kontchakov et al., 2010]. In this context, the fundamental relationship between ontologies is thus not whether they are logically equivalent (have the same models), but whether they give the same answers to any relevant query.

The resulting entailment problem can be formalized in two ways, with different applications. First, given a class \( Q \) of queries, knowledge bases (KBs) \( K_1 \) and \( K_2 \), and a signature \( \Sigma \) of relevant concept and role names, we say that \( K_1 \Sigma Q \text{-entails} \ K_2 \) if the answers to any \( \Sigma \)-query in \( Q \) over \( K_1 \) are contained in the answers to the same query over \( K_2 \). Further, \( K_1 \) and \( K_2 \) are \( \Sigma \text{-Q-inseparable} \) if they \( \Sigma \text{-Q-entail} \) each other. Note that a KB includes an ABox, and thus this notion of entailment is appropriate if the data is known and does not change frequently. Applications include data-oriented KB versioning and KB module extraction, KB forgetting [Wang et al., 2014], and knowledge exchange [Arenas et al., 2013].

If the data is not known or changes frequently, it is not KBs that should be compared, but TBoxes. Given a pair \( \Theta = (\Sigma_1, \Sigma_2) \) that specifies a relevant signature \( \Sigma_1 \) for ABoxes and \( \Sigma_2 \) for queries, we say that a TBox \( T_1 \Theta \text{-Q-entails} \) a TBox \( T_2 \) if, for every \( \Sigma_1 \text{-ABox} A \), the KB \( (T_1, A) \) \( \Sigma_2 \text{-Q-entails} \ (T_2, A) \). \( T_1 \) and \( T_2 \) are \( \Theta \text{-Q-inseparable} \) if they \( \Theta \text{-Q-entail} \) each other. Applications include data-oriented TBox versioning, TBox modularization and TBox forgetting [Kontchakov et al., 2010].

In this paper, we concentrate on the most important choices for \( Q \), conjunctive queries (CQs) and unions thereof (UCQs); we also consider the practically relevant classes of rooted CQs (rCQs) and UCQs (rUCQs), in which every variable is connected to an answer variable. So far, CQ-entailment has been studied for Horn DL KBs [Botoeva et al., 2014], \textit{EL} TBoxes [Lutz and Wolter, 2010; Konev et al., 2012], DL-Lite TBoxes [Kontchakov et al., 2009], and also for OBDA specifications, that is, DL-Lite TBoxes with mappings [Bienvenu and Rosati, 2015]. No results are available for non-Horn DLs (neither in the KB nor in the TBox case) and for expressive Horn DLs in the TBox case. In particular, query entailment in non-Horn DLs has had the reputation of being a technically challenging problem.

This paper makes a first breakthrough into understanding query entailment and inseparability in these cases, with the main results summarized in Figures 1 and 2 (those marked with \((*)\) are from [Botoeva et al., 2014]). Three of them came as a real surprise to us. First, it turned out that CQ- and rCQ-entailment between \textit{ALC} KBs is undecidable, even when the first KB is formulated in Horn-\textit{ALC} (in fact, \textit{EL}) and without any signature restriction. This should be contrasted with the decidability of subsumption-based entailment between \textit{ALC}
TBoxes [Ghilardi et al., 2006] and of CQ-entailment between Horn-\(\text{ALC}\) KBs [Botoeva et al., 2014]. The second surprising result is that entailment between \(\text{ALC}\) KBs becomes decidable when CQs are replaced with rUCQs. For \(\text{ALC}\) TBoxes, CQ- and rCQ-entailment are undecidable as well. We obtain decidability for Horn-\(\text{ALC}\) TBoxes (where CQ- and UCQ-entailments coincide) using the fact that non-entailment is always witnessed by tree-shaped ABoxes. As another surprise, CQ-entailment of Horn-\(\text{ALC}\) TBoxes is \(\text{ExpTime}\)-complete while rCQ-entailment is only \(\text{ExpTime}\)-complete. This should be contrasted with the \(\mathcal{E}\mathcal{L}\) case, where both problems are \(\text{ExpTime}\)-complete [Lutz and Wolter, 2010]. All upper bounds and most lower bounds hold also for inseparability in place of entailment. A model-theoretic foundation for these results is a characterization of query entailment between KBs and TBoxes in terms of (partial) homomorphisms, which, in particular, enables the use of tree automata techniques to establish the upper bounds in Figs. 1 and 2. Omitted proofs are available in the full version [Botoeva et al., 2016].

2 Preliminaries

Fix lists of individual names \(a_i\), concept names \(A_i\), and role names \(R_i\), for \(i < \omega\). \(\text{ALC}\)-concepts, \(C\), are defined by the grammar

\[
C ::= A_i \mid T \mid \neg C \mid C_1 \cap C_2 \mid \exists R_i.C.
\]

We use \(\bot\), \(C_1 \cup C_2\) and \(\forall R.C\) as abbreviations for \(\neg \top\), \(\neg(\neg C_1 \cap \neg C_2)\) and \(\exists R_\bot.C\), respectively. A concept inclusion (CI) takes the form \(C \subseteq D\), where \(C\) and \(D\) are concepts. An \(\text{ALC}\) TBox is a finite set of CIs. In a Horn-\(\text{ALC}\) TBox, no concept of the form \(\neg C\) occurs negatively and no \(\exists R_\bot.C\) occurs positively [Hustadt et al., 2005; Kazakov, 2009]. An \(\mathcal{E}\mathcal{L}\) TBox does not contain \(\neg\) at all. An ABox, \(A\), is a finite set of assertions of the form \(A_k(a_i)\) or \(R_k(a_i, a_j)\); \(\text{ind}(A)\) is the set of individual names in \(A\). Taken together, \(T\) and \(A\) form a knowledge base (KB) \(K = (T, A)\); we set \(\text{ind}(K) = \text{ind}(A)\).

The semantics is defined as usual based on interpretations \(I = (\Delta_I, \mathcal{I})\) that comply with the standard name assumption in the sense that \(a^I = a\) [Baader et al., 2003]. We write \(I \models a\) if an inclusion or assertion \(a\) is true in \(I\). If \(I \models a\), for all \(a \in T \cup \text{A}\), then we call \(I\) a model of \(K\) and write \(I \models K\). \(K\) is consistent if it has a model; we then also say that \(A\) is consistent with \(T\). \(K \models a\) means that \(I \models a\) for all \(I \models K\).

A conjunctive query (CQ) \(q(x)\) is a formula \(\exists y \varphi(x, y)\), where \(\varphi\) is a conjunction of atoms of the form \(A_k(z_1)\) or \(R_k(z_1, z_2)\) with \(z_1\) in \(x\), \(y\); the variables in \(x\) are the answer variables of \(q(x)\). We call \(q\) rooted (rCQ) if every \(y \in y\) is connected to some \(x \in x\) by a path in the graph whose nodes are the variables in \(q\) and edges are the pairs \((u, v)\) with \(R(u, v) \in q\), for some \(R\). A union of CQs (UCQ) is a disjunction \(q(x) = \bigvee_i q_i(x)\) of CQs \(q_i(x)\) with the same answer variables \(x\); it is rooted (rUCQ) if all \(q_i\) are rooted.

A tuple \(a\) in \(\text{ind}(K)\) is a certain answer to a CQ \(q(x)\) over \(K = (T, A)\) if \(I \models q(a)\) for all \(I \models K\); in this case we write \(K \models q(a)\). If \(x = \emptyset\), the answer to \(q\) is ‘yes’ if \(K \models q\) and ‘no’ otherwise. The problem of checking whether a tuple is a certain answer to a given \(U\)CQ over a given \(\text{ALC}\) KB is known to be \(\text{ExpTime}\)-complete for combined complexity [Lutz, 2008]. The \(\text{ExpTime}\) lower bound actually holds for Horn-\(\text{ALC}\) [Kroötzsch et al., 2013].

A set \(M\) of models of a KB \(K\) is called complete for \(K\) if, for every UCQ \(q(x)\), we have \(K \models q(a)\) iff \(I \models q(a)\) for all \(I \models M\). We call an interpretation \(I\) a dнтree interpretation if the directed graph \(G_T\) with nodes \(d \in \Delta_T^2\) and edges \((d, e)\) in \(R^2\), for some \(R\), is a tree and \(R^2 \cap S^2 = \emptyset\), for any distinct roles \(R, S\). \(I\) has outdegree \(n\) if \(G_T\) has outdegree \(n\). A model \(I\) of a KB \(K = (T, A)\) is forest-shaped if \(I\) is the disjoint union of dיתree interpretations \(I_a\) with root \(a\), for \(a \in \text{ind}(A)\), extended with all \(R(a, b) \in A\). The outdegree of \(I\) is the maximum outdegree of the \(I_a\). It is well known that the class \(M_K^{\text{ind}}\) of all forest-shaped models of an \(\text{ALC}\) KB \(K\) of outdegree bounded by \(|T|\) is complete for \(K\) [Lutz, 2008].

If \(K\) is a Horn-\(\text{ALC}\) KB, then a single member \(I_K\) of \(M_K^{\text{ind}}\) is complete for \(K\). \(I_K\) is constructed using the standard chase procedure and called the canonical model of \(K\).

A signature, \(\Sigma\), is a set of concept and role names. By a \(\Sigma\)-concept, \(\Sigma\)-CQ, etc. we understand any concept, CQ, etc. constructed using the names from \(\Sigma\). We say that \(\Sigma\) is full if it contains all concept and role names. A model \(I\) of a KB \(K\) is \(\Sigma\)-connected if, for any \(a \in \Delta_T^2 \setminus \text{ind}(K)\), there is a path \(R_{i1}^1(a, u_1), \ldots, R_{i_n}^n(u_n, u)\) with \(a \in \text{ind}(K)\) and \(R_i\) in \(\Sigma\).

**Definition 1.** Let \(K_1\) and \(K_2\) be consistent KBs, \(\Sigma\) a signature, and \(Q\) one of CQ, rCQ, UCQ or rUCQ. We say that \(K_1\) \(\Sigma\)-Q-entails \(K_2\) if \(K_2 \models q(a)\) implies \(a \in \text{ind}(K_1)\) and \(K_1 \models q(a)\), for all \(\Sigma\)-Q \(q(x)\) and all tuples \(a \in \text{ind}(K_2)\). \(K_1\) and \(K_2\) are \(\Sigma\)-Q inseparable if they \(\Sigma\)-Q entail each other.

As larger classes of queries separate more KBs, \(\Sigma\)-UCQ inseparability implies all other inseparabilities. The following example shows that, in general, no other implications between the different notions of inseparability hold for \(\text{ALC}\).

**Example 2.** Suppose \(T_0 = \emptyset\), \(T_0' = \{E \sqsubseteq A \cup B\}\) and \(\Sigma_0 = \{A, B, E\}\). Let \(A_0 = \{E(a)\}\), \(K_0 = (T_0, A_0)\), and \(K_0' = (T_0', A_0)\). Then \(K_0\) and \(K_0'\) are \(\Sigma_0\)-Q inseparable but not \(\Sigma_0\)-rUCQ inseparable. In fact, \(K_0\) \(\not\models q(a)\) and \(K_0'\) \(\not\models q(a)\) for \(q(x) = A(x) \lor B(x)\).

Now, let \(\Sigma_1 = \{E, B\}\), \(T_1 = \emptyset\), and \(T_1' = \{E \sqsubseteq \exists R.B\}\). Let \(A_1 = \{E(a)\}\), \(K_1 = (T_1, A_1)\), and \(K_1' = (T_1', A_1)\).
Then \( K_1 \) and \( K_2 \) are \( \Sigma_1 \)-UCQ inseparable but not \( \Sigma_1 \)-CQ inseparable. In fact, \( K'_1 \models \exists x B(x) \) but \( K_1 \not\models \exists x B(x) \).

**Definition 3.** Let \( T_1 \) and \( T_2 \) be TBoxes, \( \Theta \) one of CQ, rCQ, UCQ or rUCQ, and let \( \Theta = (\Sigma_1, \Sigma_2) \) be a pair of signatures. We say that \( T_1 \Theta \)-Q entails \( T_2 \) if, for every \( \Sigma_1 \)-ABox \( A \) that is consistent with both \( T_1 \) and \( T_2 \), the KB \( (T_1, A) \) \( \Sigma_2 \)-Q entails the KB \( (T_2, A) \). \( T_1 \) and \( T_2 \) are \( \Theta \)-Q inseparable if they \( \Theta \)-Q entail each other. If \( \Sigma_1 \) is the set of all concept and role names, we say ‘full ABox signature \( \Sigma_2 \)-Q entails’ or ‘full ABox signature \( \Sigma_2 \)-Q inseparable’.

We only consider ABoxes that are consistent with both TBoxes because the problem whether a \( \Sigma_1 \)-ABox consistent with \( T_2 \) is also consistent with \( T_1 \) is well understood: it is mutually polynomially reducible with the containment problem for ontology-mediated queries with CQs of the form \( \exists x A(x) \), which is NEXPTIME-complete for ALC and EXPTIME-complete for Horn-ALC [Bienvenu et al., 2012; 2014].

**Example 4.** Consider the TBoxes \( T_0 \) and \( T_0' \) from Example 2 and let \( \Theta = (\Sigma, \Sigma) \) for \( \Sigma = \{ R, A, B, E \} \). Then \( T_0 \) does not \( \Theta \)-rCQ entail \( T'_0 \) as \( (T'_0, A) \models \varphi(a) \) and \( (T_0, A) \not\models \varphi(a) \) for \( A \; q(x): \quad z \rightarrow R_x A \quad R_y B \quad q(y): \quad z \rightarrow R_x A \quad R_y B \).

We observe that \( \Theta \)-CQ entailment in the restricted case with \( \Theta = (\Sigma, \Sigma) \) has been investigated for E\( \mathcal{L} \) TBoxes by Lutz and Wolter [2010] and Konev et al. [2012].

As in the KB case, \( \Sigma \)-UCQ inseparability of ALC TBoxes implies all other types of inseparability, and Example 2 can be used to show that no other implications hold in general. The situation changes for Horn-ALC KBs and TBoxes. The following can be proved by observing that a Horn-ALC KB entails a UCQ iff it entails one of its disjuncts:

**Theorem 5.** Let \( K_1 \) be an ALC KB and \( K_2 \) a Horn-ALC KB. Then \( K_1 \)-\( \Sigma \)-UCQ entails \( K_2 \) iff \( K_2 \)-\( \Sigma \)-CQ entails \( K_2 \). The same holds for rUCQ and rCQ, and for TBox entailment.

### 3 Model-Theoretic Criteria for ALC KBs

We now give model-theoretic criteria for \( \Sigma \)-entailment between KBs. The product \( \prod I \) of a set \( I \) of interpretations is defined as usual in model theory [Chang and Keisler, 1990, page 405]. Note that, for any CQ \( q(x) \) and any tuple \( \alpha \) of individual names, \( \prod I \models q(\alpha) \) iff \( I \models q(\alpha) \) for each \( I \in I \).

Suppose \( I_1 \) is an interpretation for a KB \( K_1 \), \( i = 1, 2 \). A function \( h : \Delta^{i_2} \rightarrow \Delta^{i_1} \) is called a \( \Sigma \)-homomorphism if \( h u \in A^{i_2} \) implies \( h(u) \in A^{i_1} \) and \( (u,v) \in R^{i_2} \) implies \( (h(u), h(v)) \in R^{i_1} \) for all \( u, v \in \Delta^{i_2} \), \( \Sigma \)-concept names \( A \), and \( \Sigma \)-role names \( R \), and \( h(u) = a \) for all \( u \in \text{ind}(K_2) \). It is known from database theory that homomorphisms characterize CQ-entailment [Chandra and Merlin, 1977]. For KB \( \Sigma \)-query entailment, finite partial homomorphisms are required. We say that \( I_2 \) is \( n \Sigma \)-homomorphically embeddable into \( I_1 \) if, for any subinterpretation \( I'_2 \) of \( I_2 \) with \( |\Delta^{i_2}| \leq n \), there is a \( \Sigma \)-homomorphism from \( I'_2 \) to \( I_1 \). If, additionally, we require \( I'_2 \) to be \( \Sigma \)-connected then \( I_2 \) is said to be con-\( n \Sigma \)-homomorphically embeddable into \( I_1 \).

**Theorem 6.** Let \( K_1 \) and \( K_2 \) be ALC KBs, \( \Sigma \) a signature, and let \( M_i \) be complete for \( K_i \), \( i = 1, 2 \).

1. \( K_1 \)-\( \Sigma \)-UCQ entails \( K_2 \) iff, for any \( n > 0 \) and \( I_1 \in M_1 \), there exists \( I_2 \in M_2 \) that is \( n \Sigma \)-homomorphically embeddable into \( I_1 \).
2. \( K_1 \)-\( \Sigma \)-rUCQ entails \( K_2 \) iff, for any \( n > 0 \) and \( I_1 \in M_1 \), there exists \( I_2 \in M_2 \) that is con-\( n \Sigma \)-homomorphically embeddable into \( I_1 \).
3. \( K_1 \)-\( \Sigma \)-CQ entails \( K_2 \) iff \( \prod M_2 \) is \( n \Sigma \)-homomorphically embeddable into \( \prod M_1 \) for any \( n > 0 \).
4. \( K_1 \)-\( \Sigma \)-rCQ entails \( K_2 \) iff \( \prod M_2 \) is con-\( n \Sigma \)-homomorphically embeddable into \( \prod M_1 \) for any \( n > 0 \).

**Proof.** We only show (1). Suppose \( K_2 \models q \) but \( K_1 \not\models q \). Let \( n \) be the number of variables in \( q \). Take \( I_1 \in M_1 \) such that \( I_1 \not\models q \). Then no \( I_2 \in M_2 \) is \( n \Sigma \)-homomorphically embeddable into \( I_1 \). Conversely, suppose \( I_1 \in M_1 \) is such that, for some \( n \), no \( I_2 \in M_2 \) is \( n \Sigma \)-homomorphically embeddable into \( I_1 \). We can regard any subinterpretation of any \( I_2 \in M_2 \) with domain of size \( \leq n \) as a CQ (with answer variable corresponding to ABox individuals). The disjunction of all such CQs is entailed by \( K_2 \) but not by \( K_1 \).

Note that \( n \Sigma \)-homomorphic embeddability cannot be replaced by \( \Sigma \)-homomorphic embeddability. For example, in (1), let \( K_1 \) and \( K_2 \) be \( (\{ \top \iff \exists R \cdot T \}, \{ A(a) \}) \), \( M_1 = \{ I_1 \} \), where \( I_1 \) is the infinite \( R \)-chain starting with \( a \), and let \( M_2 \) contain arbitrary finite \( R \)-chains starting with \( a \) followed by an arbitrary long \( R \)-cycle. \( M_1 \) and \( M_2 \) are both complete for \( K_1 \), but there is no \( \Sigma \)-homomorphism from any \( I_2 \in M_2 \) to \( I_1 \). In Section 5, we show that in some cases we can find characterizations with full \( \Sigma \)-homomorphisms and use them to present decision procedures for entailment.

If both \( M_i \) are finite and contain only finite interpretations, then Theorem 6 provides a decision procedure for KB entailment. This applies, for example, to KBs with acyclic classical TBoxes [Baader et al., 2003], and to KBs for which the chase terminates [Grau et al., 2013].

### 4 Undecidability for ALC KBs and TBoxes

We show that CQ and rCQ-entailment and inseparability for ALC KBs are undecidable—even if the signature is full and \( K_1 \) is a Horn-ALC (in fact, E\( \mathcal{L} \)) KB. We establish the same results for TBoxes except that in the rCQ case, we leave it open whether the full ABox signature is sufficient for undecidability.

**Theorem 7.** (i) The problem whether a Horn-ALC KB \( \Sigma \)-Q entails an ALC KB is undecidable for \( \Sigma \in \{ CQ, rCQ \} \).

(ii) \( \Sigma \)-Q inseparability between Horn-ALC and ALC KBs is undecidable for \( \Sigma \in \{ CQ, rCQ \} \).

(iii) Both (i) and (ii) hold for the full signature \( \Sigma \).

**Proof.** The proof is by reduction of the undecidable \( N \times M \)-tiling problem: given a finite set \( T \) of tile types \( T \) with four colours \( \text{up}(T), \text{down}(T), \text{left}(T), \text{right}(T) \), a tile type \( I \in T \), and two colours \( W \) (for wall) and \( C \) (for ceiling), decide whether there exist \( N, M \in \mathbb{N} \) such that the \( N \times M \) grid can be tiled using \( T \) in such a way that \((1, 1) \) is covered by a tile of type \( I \); every \((N, i), \) for \( i \leq M \), is covered by a tile
of type $T$ with $\text{right}(T) = W$; and every $(i, M)$, for $i \leq N$, is covered by a tile of type $T$ with $\text{up}(T) = C$.

Given an instance of this problem, we first describe a KB $K_2 = (T_2, \{A(a)\})$ that uses (among others) 3 concept names $T_k$, $k = 0, 1, 2$, for each type $T \in \Sigma$. If a point $x$ in a model $I$ of $K_2$ is in $T_k$ and $\text{right}(T) = \text{left}(T')$, then $x$ has an $R$-successor in $T_k$. Thus, branches of $I$ define (possibly infinite) horizontal rows of tilings with $\Sigma$. If a branch contains a point $y \in T_k$ with $\text{right}(T) = W$, then this $y$ can be the last point in the row, which is indicated by an $R$-successor $z \in \text{Row}$ of $y$. In turn, $z$ has $R$-successors in all $T_{(k+1) \mod 3}$ that can be possible beginnings of the next row of tiles. To coordinate the $\text{up}$ and $\text{down}$ colours between the rows—which will be done by the CQs separating $K_1$ and $K_2$—we make every $x \in T_k$, starting from the second row, an instance of all $T_{(k+1) \mod 3}$ with $\text{down}(T) = \text{up}(T')$. The row started by $z \in \text{Row}$ can be the last one in the tiling, in which case we require that each of its tiles $T$ has $\text{up}(T) = C$. After the point in $\text{Row}$ indicating the end of the final row, we add an $R$-successor in $\text{End}$ for the end of the tiling. The beginning of the first row is indicated by a $P$-successor in $\text{Start}$ of the ABox element $a$, after which we add an $R$-successor in $I_0$ for the given initial tile type $I$; see the lowest branch in Fig. 3. To generate a tree with all possible branches described above, we only require $\mathcal{EL}$ axioms of the form $E \sqsubseteq D$ and $R \sqsubseteq \exists S.D$.

The existence of a tiling of some $N \times M$ grid for the given instance can be checked by Boolean CQs $q_n$ that require an $R$-path from $\text{Start}$ to $\text{End}$ going through $T_k$ or $\text{Row}$-points:

$$\exists x (\text{Start}(x_0) \land \bigwedge_{i=0}^{n} R(x_i, x_{i+1}) \land \bigwedge_{i=1}^{n} B_i(x_i) \land \text{End}(x_{n+1}))$$

with $B_i \in \{\text{Row}\} \cup \{T_k \mid T \in \Sigma, k = 0, 1, 2\}$; see Fig. 3. The key trick is—using an axiom of the form $D \subseteq E \cup E'$—to ensure that the $\text{Row}$-point before the final row of the tiling has two alternative continuations: one as described above, and the other one having just a single $R$-successor in $\text{End}$; see Fig. 3 where $\lor$ indicates an or-node. This or-node gives two models of $K_2$ denoted $I_l$ and $I_r$ in the picture. If $K_2 \models q_n$, then $q_n$ holds in both of them, and so there are homomorphisms $h_l : q_n \rightarrow I_l$ and $h_r : q_n \rightarrow I_r$. As $h_l(x_{n-1})$ and $h_r(x_{n-1})$ are instances of $B_{n-1}$, we have $B_{n-1} = T_{(N-M-1)}$ in the picture, and so $\text{up}(T_{(N-M-1)}) = \text{down}(T_{(N-M)})$. By repeating this argument until $x_0$, we see that the colours between horizontal rows match and the rows are of the same length. (For this trick to work, we have to make the first $\text{Row}$-point in every branch an instance of $\text{Start}$.) In fact, we have:

**Lemma 8.** An instance of the $N \times M$-tiling problem has a positive answer iff there exists $q_n$ such that $K_2 \models q_n$.

It is to be noted that to construct $T_2$ with the properties described above one needs quite a few auxiliary concept names.

Next, we define $K_1 = (\{T_i, \{A(a)\}\})$ to be the $\mathcal{EL}$ KB with the following canonical model:

$$\begin{align*}
A & \rightarrow \text{Start}, I_0, T_0^{N+1} \rightarrow \text{Start}, T_1^2 \rightarrow \text{End}, T_2^{N-1} \rightarrow \text{Start}, T_3^{N-1} \rightarrow \text{End}, \ldots, T_{\Sigma}^{N-1} \rightarrow \text{End}, \ldots
\end{align*}$$

where $\Sigma_0 = \{\text{Row}\} \cup \{T_k \mid T \in \Sigma, k = 0, 1, 2\}$. Note that the vertical $R$-successors of the $\text{Start}$-points are not instances of any concept name, and so $K_1$ does not satisfy any query $q_n$. On the other hand, $K_2 \models q$ implies $K_1 \models q$, for every $\Sigma$-CQ $q$ without a subquery of the form $q_n$ and $\Sigma = \text{sig}(K_1)$.

This proves (i) for $\Sigma$-CQ entailment. For $\Sigma$-rCQ entailment, we slightly modify the construction, in particular, by adding $R(a, a)$ and $\text{Row}(a)$ to the ABox $\{A(a)\}$, and a conjunct $R(y, x_0)$ with a free $y$ to $q_n$. (The loop $R(a, a)$ plays roughly the same role as the path between two $\text{Start}$-points in Fig. 3.) To prove (ii), we take $K'_2 = K_2 \cup K_1$ and show that $K_1$ $\Sigma$-CQ entails $K_2$ iff $K_1$ and $K_2$ are $\Sigma$-CQ inseparable. Finally, we prove (iii) by replacing non-$\Sigma$ symbols in $K_2$ with complex $\mathcal{ALC}$-concepts that cannot be used in CQs and extending the TBoxes appropriately; cf. [Lutz and Wolter, 2012, Lemma 21].

The TBoxes from the proof above can also be used to obtain

**Theorem 9.** (i) The problem whether a Horn-$\mathcal{ALC}$ TBox $\Theta$-Q entails an $\mathcal{ALC}$ TBox is undecidable for $Q \in \{\text{CQ}, r\text{CQ}\}$.

(ii) $\Theta$-Q inseparability between Horn-$\mathcal{ALC}$ and $\mathcal{ALC}$ TBoxes is undecidable for $Q \in \{\text{CQ}, r\text{CQ}\}$.

(iii) For CQs, (i) and (ii) hold for full ABox signatures and for $\Theta = (\Sigma_1, \Sigma_2)$ with $\Sigma_1 = \Sigma_2$.

Observe that our undecidability proof does not work for UCQs as the UCQ composed of the two disjunctive branches shown in Fig. 3 (for non-trivial instances) distinguishes between the KBs independently of the existence of a tiling. We now show that, at least for rUCQs, entailment is decidable.
5 rUCQ-Entailment for $\mathcal{ALC}$-KBs

Theorem 7 might seem to suggest that any reasonable notion of query inseparability is undecidable for $\mathcal{ALC}$ KBs. Interestingly, this is not the case: we show now that rUCQ-entailment is decidable. We first strengthen the characterization of Theorem 6 (2), and then develop a decision procedure based on tree automata. The first step replaces con-$\Sigma$-homomorphic embeddability with con-$\Sigma$-homomorphic embeddability, where $I_2$ is con-$\Sigma$-homomorphically embeddable into $I_1$ if the maximal $\Sigma$-connected subinterpretation of $I_2$ is $\Sigma$-homomorphically embeddable into $I_1$.

**Theorem 10.** Let $K_1$ and $K_2$ be $\mathcal{ALC}$ KBs, $\Sigma$ a signature, and let $M_1$ be complete for $K_1$. Then $K_1$-rUCQ entails $K_2$ iff for any $I_1 \in M_1$, there exists $I_2 \models K_2$ such that $I_2$ is con-$\Sigma$-homomorphically embeddable into $I_1$.

**Proof.** In view of Theorem 6 (2), it suffices to prove ($\Rightarrow$). Suppose $I_1 \models M_1$. By Theorem 6 (2), for every $n \geq 0$, we have $J \in M_{K_1}^{fo}$ and a $\Sigma$-homomorphism $h_n : J_{\leq n} \rightarrow I_1$, where $J_{\leq n}$ is the subinterpretation of $J$ whose elements are connected to ABox individuals by $\Sigma$-paths of length $\leq n$. Clearly, for any $n \geq 0$, there are only finitely many non-isomorphic pairs $(J_{\leq n}, h_n)$. It can be shown that, thus, one can construct the required $I_2 \in M_{K_2}^{fo}$ and con-$\Sigma$-homomorphism $h$ as the limits of suitable chains $J_{\leq n} \subseteq J_{\leq n+1} \subseteq \cdots$ and $h_0 \subseteq h_1 \subseteq \cdots$, respectively. □

For the second step, let $K_1$, $K_2$ be $\mathcal{ALC}$-KBs and $\Sigma$ a signature. We use two-way alternating automata on infinite trees (2ATAs) with a trivial acceptance condition (every run is accepting) and employ Theorem 10 for the class $M_{K_2}^{fo}$, encoding forest-shaped interpretations as labeled trees to make them accessible to 2ATAs. A tree is a non-empty (possibly infinite) set $T \subseteq N^*$ closed under prefixes with root $\varepsilon$. We say that $T$ is $m$-ary if, for every $x \in T$, the set $\{ i \mid x \cdot i \in T \}$ is of cardinality $m$. Let $\Gamma$ be an alphabet with symbols from the set

$$\{ \text{root}, \text{empty} \}\cup (\text{ind}(K_1) \times 2^{CN(T_1)}) \cup (\text{RN}(T_1) \times 2^{CN(T_1)})$$

where $CN(T_1)$ (resp. $RN(T_1)$) denotes the set of concept (resp. role) names in $T_1$. A $\Gamma$-labeled tree is a pair $(T, L)$ with $T$ a tree and $L : T \rightarrow \Gamma$ a node labeling function. We represent forest-shaped models of $T_1$ as $m$-ary $\Gamma$-labeled trees, with $m = \max(|T_1|, |\text{ind}(K_1)|)$. The root node labeled with root is not used in the representation. Each ABox individual is represented by a successor of the root labeled with a symbol from $\text{ind}(K_1) \times 2^{CN(T_1)}$; non-ABox elements are represented by nodes deeper in the tree labeled with a symbol from $\text{RN}(T_1) \times 2^{CN(T_1)}$. The label empty is used for padding to make sure that every tree node has exactly $m$ successors.

Now we construct three 2ATAs $A_i$, for $i = 0, 1, 2$. $A_0$ ensures that the tree is labeled in a meaningful way, e.g. that the root label only occurs at the root node; $A_1$ accepts $\Gamma$-labeled trees that represent a model of $K_1$, and $A_2$ accepts $\Gamma$-labeled trees $(T, L)$ which represent an interpretation $I_{(T, L)}$ such that some model of $K_2$ is con-$\Sigma$-homomorphically embeddable into $I_{(T, L)}$. The most interesting automaton is $A_2$, which guesses a model of $K_2$ along with a homomorphism to $I_{(T, L)}$; in fact, both can be read off from a successful run of the automaton. The number of states of the $A_i$ is exponential in $|K_1 \cup K_2|$. It then remains to combine these automata into a single 2ATA $A$ such that $L(A) = L(A_0) \cap L(A_1) \cap L(A_2)$, which is possible with only polynomial blowup, and to test (in time exponential in the number of states) whether $L(A) = \emptyset$.

**Theorem 11.** It is in $2\text{ExpTime}$ to decide whether an $\mathcal{ALC}$ KB $K_1, \Sigma$-rUCQ entails an $\mathcal{ALC}$ KB $K_2$.

The best known lower bound is ExpTime, which is easy to establish by reduction from satisfiability.

6 (r)CQ-Entailment for (Horn-)$\mathcal{ALC}$-TBoxes

We show that CQ- and rCQ-entailment between $\mathcal{ALC}$ TBoxes becomes decidable when the second TBox is given in Horn-$\mathcal{ALC}$. In this case, entailments for CQs and UCQs and, respectively, rCQs and rUCQs coincide. We start with rCQs.

Our first observation is that if a $\Sigma_1$-ABox is a witness for non-$\Theta$-rCQ entailment, then one can find a witness $\Sigma_1$-ABox that is tree-shaped and of bounded outdegree. Here, an ABox $A$ is tree-shaped if the graph with nodes ind($A$) and edges $(a, b)$ for each $R(a, b) \in A$ is a tree, and $R(a, b) \in A$ implies $S(a, b) \notin A$ for all $S \neq R$ and $S(b, a) \notin A$ for all $S$.

**Theorem 12.** Let $T_1$ be an $\mathcal{ALC}$ TBox, $T_2$ a Horn-$\mathcal{ALC}$ TBox, and $\Theta = (\Sigma_1, \Sigma_2)$. Then $T_1$ $\Theta$-rCQ-entails $T_2$ iff for all tree-shaped $\Sigma_1$-ABoxes $A$ of outdegree bounded by $|T_2|$ and consistent with $T_1$ and $T_2$, $T_2^A$ is con-$\Sigma_2$-homomorphically embeddable into any model $I_1$ of $(T_1, A)$.

**Proof.** It is known that Horn-$\mathcal{ALC}$ is unravelling tolerant, that is, $(T, A) \models C(a)$ for a Horn-$\mathcal{ALC}$ TBox $T$ and $\mathcal{EL}$-concept $C$ iff $(T, A') \models C(a)$ for a finite sub-ABox $A'$ of the tree-unravelling of $A$ at $a$ [Lutz and Wolter, 2012]. Thus, any witness ABox for non-entailment w.r.t. $\mathcal{EL}$-instance queries can be transformed into a tree-shaped witness ABox. The result follows by observing that if $T_1$ does not $\Theta$-rCQ-entail $T_2$, then this is witnessed by an $\mathcal{EL}$-instance query and by applying Theorem 10 to the KBs. The bound on the outdegree is obtained by a careful analysis of derivations. □

For the automaton construction, let $I_1$ be an $\mathcal{ALC}$ TBox, $T_2$ a Horn-$\mathcal{ALC}$ TBox, and $\Theta = (\Sigma_1, \Sigma_2)$ a pair of signatures. Though Theorem 12 provides a natural characterization that is similar in spirit to Theorem 10, we first need a further analysis of con-$\Sigma_2$-homomorphic embeddability in terms of simulations whose advantage is that they are more compositional (they can be partial and are closed under union).

Let $I_1, I_2$ be interpretations and $\Sigma$ a signature. A relation $S \subseteq \Delta^2 \times \Delta^2$ is a $\Sigma$-simulation from $I_1$ to $I_2$ if $i \in A^2$ and $(d, d') \in S$ imply $d' \in A^2$ for all $\Sigma$-concept names $A$, and (ii) if $(d, e') \in R^2$ and $(d', e') \in S$ then there is a $(d', e') \in R^2$ with $(e, e') \in S$ for all $\Sigma$-role names $R$. Let $d_i \in \Delta^2, i \in \{1, 2\}$, $(I_1, d_1)$ is $\Sigma$-simulated by $(I_2, d_2)$, in symbols $(I_1, d_1) \preceq_\Sigma (I_2, d_2)$, if there exists a $\Sigma$-simulation $S$ with $(d_1, d_2) \in S$.

**Lemma 13.** Let $A$ be a $\Sigma_1$-ABox and $I_1$ a model of $(T_1, A)$. Then $T_{2^A, A}$ is not con-$\Sigma_2$-homomorphically embeddable into $I_2$ iff there is an $a \in \text{ind}(A)$ such that one of the following holds:
(1) There is a $\Sigma_2$-concept name $A$ with $a \in A\Gamma_{\Sigma_2,A} \setminus A\Gamma_1$.
(2) There is an $R$-successor $d$ of $a$ in $A\Gamma_{\Sigma_2,A}$, for some $\Sigma_2$-role name $R$, such that $d \notin \text{ind}(A)$ and, for all $R$-successors $e$ of $a$ in $A\Gamma_1$, we have $(A\Gamma_{\Sigma_2,A}, d) \not\subseteq_2 (A\Gamma_1, e)$.

We use a mix of two-way alternating Büchi automata on finite trees (2ABTAs) and non-deterministic top-down automata on finite trees (NTAs). A finite tree $T$ is $m$-ary if, for every $x \in T$, the set $\{i \mid x \cdot i \in T\}$ is of cardinality zero or exactly $m$.

We use labeled trees to represent a tree-shaped ABox and a model $I_3$ such that, for some $a \in \text{ind}(A)$, conditions (1) and (2) from Lemma 13 are satisfied, and thus $A\Gamma_{\Sigma_2,A}$ is not con-$\Sigma_2$-homomorphically embeddable into $I_3$. To ensure that later, additional bookkeeping information is needed. Node labels are taken from the alphabet

$$
\Gamma = \Gamma_0 \times 2\text{cl}(T_1) \times 2\text{CN}(T_2) \times \{0, 1\} \times 2\text{sub}(T_3),
$$

where $\Gamma_0$ is the set of all subsets of $\Sigma_1 \cup \{R^- \mid R \in \Sigma_1\}$ that contain at most one role (a role name $R$ or its inverse $R^-$), $\text{cl}(T_1)$ is the set of subconcepts of (concepts in) $T_1$ closed under single negation, and $\text{sub}(T_3)$ is the set of subconcepts of (concepts in) $T_3$. For a $\Gamma$-labeled tree $(T, L)$ and a node $x$ from $T$, we use $L_0(x)$ to denote the first component of $L(x)$, where $\cdot_1 \in \{0, \ldots, 4\}$. Intuitively, the $L_0$-component represents the ABox $A$, the $L_1$-component represents the model $I_1$, the $L_2$-component represents $I_{\Sigma_2,A}$, and the $L_3$- and $L_4$-components help to guarantee conditions (1) and (2) from Lemma 13.

To ensure that each component $\cdot \in \{0, \ldots, 4\}$ indeed represents what it is supposed to, we impose on it an $i$-properness condition. For example, a $\Gamma$-labeled $(T, L)$ tree is $0$-proper if (i) $L_0(x)$ contains no role and (ii) for every non-root node $x$ of $T$, $L_0(x)$ contains a role. A $0$-proper $\Gamma$-labeled tree $(T, L)$ represents the following tree-shaped $\Sigma_1$-ABox:

$$
A_{(T,L)} = \{ (x, y) \mid A \in L_0(x) \} \cup \{ (x, y) \mid R \in L_0(y), y \text{ is a child of } x \} \cup \{ (x, y) \mid R^- \in L_0(y), y \text{ is a child of } x \}.
$$

Due to space limitations, we skip the remaining definitions of properness and concentrate on explaining the most interesting components $L_1$ and $L_4$ of $\Gamma$-labels. The $L_4$-component marks a single node $x$ in the tree, which is the individual $a$ from Lemma 13 that satisfies conditions (1) and (2). If (1) is satisfied, we do not need the $L_1$-component. Otherwise, we store in that component at $x$ a set of concepts $S = \{ \exists R.A \cdot B_1, \ldots, \forall R.B_\theta \}$ such that $R \in \Sigma_2$ and all concepts from $S$ are true at $x$ in $A\Gamma_{\Sigma_2,A}$. This successor set represents the $R$-successor $d$ in condition (2) of Lemma 13. We then have to make sure that, for any neighboring node $y$ of $x$ that represents an $R$-successor of $x$ in $A\Gamma_{(T,L)}$, we have $(A\Gamma_{\Sigma_2,A}, d) \not\subseteq_2 (A\Gamma_1, y)$. This can again happen via a concept name or via a successor; we are done in the former case and use the $L_1$-component of $y$ in the latter. It is important to note that we can never return to the same node in this tracing process since we only follow roles in the forward direction and the represented ABox is tree-shaped. This is crucial for achieving the ExpTime overall complexity.

We show that $T_2$ is not $\Theta$-$r$CQ-entailed by $T_1$ iff there is an $m$-ary $\Gamma$-labeled tree that is $i$-proper for any $i \in \{0, \ldots, 4\}$. It then remains to design a 2ABTA $A$ that accepts exactly those trees. We construct $A$ as the intersection of five automata $A_i$, $i < 5$, where each $A_i$ ensures $i$-properness. Some of the automata are 2ABTAs with polynomially many states while others are NTAs with exponentially many states. We mix automata models since some properness conditions (2-properness) are much easier to describe with a 2ABTA while for others (4-properness), it does not seem to be possible to construct a 2ABTA with polynomially many states. In summary, we obtain the following result.

Theorem 14. It is ExpTime-complete to decide whether an ALC TBox $T_1 (\Sigma_1, \Sigma_2)$-rCQ entails a Horn-ALC TBox $T_2$.

Note that the ExpTime lower bound holds already for entailment of $\mathcal{EL}$ TBoxes and $\Sigma_1 = \Sigma_2$ [Lutz and Wolter, 2010]. We now study the non-rooted case, starting with an analogue of Theorem 12. As expected, moving to unrestricted queries corresponds to moving to unrestricted homomorphisms.

Theorem 15. Let $T_1$ and $T_2$ be Horn-ALC TBoxes and $\Theta = (\Sigma_1, \Sigma_2)$. Then $T_1 \Theta$-CQ entails $T_2$ iff, for all tree-shaped $\Sigma_1$-ABoxes $A$ of outdegree $\leq |T_2|$ and consistent with $T_1$ and $T_2$, $A\Gamma_{\Sigma_2,A}$ is $\Sigma_2$-homomorphically embeddable into $I_{\Sigma_1,A}$.

The automata construction described above can largely be reused for this case. The main difference is that the two conditions in Lemma 13 need to be extended with a third one: there is an element $d$ in the subtree of $A\Gamma_{\Sigma_2,A}$ rooted at $a$ that has an $R$-successor $0$, $R \notin \Sigma_2$, such that, for all elements $e$ of $I_1$, we have $(I_2, d_0) \not\subseteq_2 (I_1, e)$. To deal with this condition, it becomes necessary to store multiple successor sets in the $L_4$-components instead of only a single one, which increases the overall complexity to 2ExpTime. A matching lower bound can be proved by a (non-trivial) reduction of the word problem to exponentially bounded alternating Turing machines.

Theorem 16. $\Theta$-CQ entailment for Horn-ALC TBoxes is 2ExpTime-complete. The lower bound holds for $\Theta = (\Sigma, \Sigma)$.

7 Future Work

We have made first steps towards understanding query entailment and inseparability for KBs and TBoxes in expressive DLs. Many problems remain to be addressed. From a theoretical viewpoint, it would be of interest to solve the open problems in Figures 1 and 2, and also consider other expressive DLs such as $DL-Lite^H_{\text{bdl}}$ [Artale et al., 2009] or $ACCT$. For example, if Theorem 10 could be generalized to UCQs (and $\Sigma$-homomorphisms), we would obtain a 2ExpTime upper bound for UCQ-entailment between ALC KBs using the same technique as for rUCQs. Also, our undecidability proof goes through for $DL-Lite^H_{\text{btl}}$, but the other cases remain open. From a practical viewpoint, our model-theoretic criteria for query entailment are a good starting point for developing algorithms for approximations of query entailment based on simulations. Our undecidability and complexity results also indicate that rUCQ-entailment is more amenable to practical algorithms than, say, CQ-entailment and can be used as an approximation of the latter.

Acknowledgments. This work has been supported by the EU IP project Optique, grant n. FP7-318338, DFG grant LU 1417/2-1, and EPSRC UK grants EP/M012646/1 and EP/M012670/1 (iTract).
References


