Fitting Ontologies and Constraints to Relational Structures

Simon Hosemann[1], **Jean Christoph Jung**[2], **Carsten Lutz**[1,4], **Sebastian Rudolph**[3,4]

[1]Leipzig University [2]TU Dortmund University [3]TU Dresden
[4]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)

simon.hosemann@uni-leipzig.de, jean.jung@tu-dortmund.de,
carsten.lutz@uni-leipzig.de, sebastian.rudolph@tu-dresden.de

Abstract

We study the problem of fitting ontologies and constraints to positive and negative examples that take the form of a finite relational structure. As ontology and constraint languages, we consider the description logics $\mathcal{EL}$ and $\mathcal{ELI}$ as well as several classes of tuple-generating dependencies (TGDs): full, guarded, frontier-guarded, frontier-one, and unrestricted TGDs as well as inclusion dependencies. We pinpoint the exact computational complexity, design algorithms, and analyze the size of fitting ontologies and TGDs. We also investigate the related problem of constructing a finite basis of concept inclusions / TGDs for a given set of finite structures. While finite bases exist for $\mathcal{EL}$, $\mathcal{ELI}$, guarded TGDs, and inclusion dependencies, they in general do not exist for full, frontier-guarded and frontier-one TGDs.

1 Introduction

In a *fitting problem* as studied in this article, one is given as input a finite set of positive and negative examples, each taking the form of a logical structure, and the task is to produce a logical formula that satisfies all examples. Problems of this form play a fundamental role in several applications. A prime example is the query by example paradigm in the field of data management, also known as query reverse engineering (Li *et al.* 2015; Barceló and Romero 2017; ten Cate *et al.* 2023a). In that case, the positive and negative examples are database instances and the formula to be constructed is a database query. In concept learning in description logics (DLs) (Lehmann and Hitzler 2010; Jung *et al.* 2021; ten Cate *et al.* 2023c), the examples are ABoxes and the formula is a DL concept; note that fitting problems are connected to PAC learning by the fundamental theorem of computational learning theory. Another example is entity comparison (Petrova *et al.* 2017; Petrova *et al.* 2019) where the examples are knowledge graphs and the formula is a SPARQL query.

In this article, we study fitting problems that aim to support the construction of two kinds of artefacts: (i) ontologies formulated in a description logic or an existential rule language and (ii) constraints on databases that take the form of tuple-generating dependencies (TGDs). Since 'existential rule' and 'TGD' are two names for the same thing, from now on we shall only speak of TGDs. We will consider both unrestricted TGDs and restricted classes of TGDs: full, guarded, frontier-guarded, and frontier-one, as well as inclusion dependencies, which also form a restricted class of TGDs. In the DLs $\mathcal{EL}$ and $\mathcal{ELI}$ that we consider in this article, an ontology is a set of concept inclusions that can be translated into an equivalent TGD. For uniformity, we shall thus also refer to concept inclusions as TGDs.

Let us be more precise about the fitting problems that we study. Examples are finite relational structures that, as in data management, we refer to as instances. In the DL case, such structures only admit unary and binary relations and are usually called interpretations. The formulas that we seek to construct fall into two classes. We may either be interested in a single TGD, to be used as a building block in an ontology or as a database constraint, or we may want to construct a finite set of TGDs to be used as an ontology or as a collection of constraints. From our perspective, there is in fact no difference between an ontology and a set of constraints: any set of TGDs can be used as an ontology when an open world semantics is adopted and as a set of constraints under a closed world semantics. We are interested both in the construction of fitting TGDs and ontologies, and in deciding their existence.

The problem of fitting an ontology to a given set of examples turns out to be closely related to a problem that has been studied in the area of description logic and is known as finite basis construction (Distel 2011; Guimarães *et al.* 2023; Kriegel 2024). One fixes an ontology language $\mathcal{L}$ and is given as input a finite instance I and the task is to produce an $\mathcal{L}$-ontology $\mathcal{O}$ such that $I \models \rho$ if and only if $\mathcal{O} \models \rho$, for all $\mathcal{L}$-TGDs ρ. We generalize this problem to a finite set H of input instances. It turns out that a finite basis $\mathcal{O}$ for the set of positive examples is a canonical fitting ontology in the sense that if any $\mathcal{L}$-ontology fits the given (positive and negative) examples, then $\mathcal{O}$ does. Thus, constructing finite bases provides an approach to constructing fitting ontologies and to deciding their existence. This approach in fact often yields decidability and tight upper complexity bounds ('tight' meaning that we prove matching lower and upper bounds).

We first consider the DLs $\mathcal{EL}$ and $\mathcal{ELI}$ as well as their extensions with the $\bot$ concept. We reprove the existence of finite bases for $\mathcal{EL}$, already known from (Baader and Distel 2008; Distel 2011), and simultaneously prove that finite bases exist also for $\mathcal{ELI}$, which, to the best of our knowledge, is a new result. In contrast to the proofs from (Baader and

Distel 2008; Distel 2011), our proofs are direct in that they do not rely on the machinery of formal concept analysis. The constructed bases are of double exponential size, but can be succinctly represented in single exponential size by structure sharing. We also show that these size bounds are tight, both for $\mathcal{EL}$ and for $\mathcal{ELI}$. We obtain from this an EXPTIME upper bound for the fitting existence problem for $\mathcal{EL}(\mathcal{I})$-ontologies.

We then provide a semantic characterization of fitting $\mathcal{EL}(\mathcal{I})$-TGD existence in terms of simulations and direct products. This characterization gives rise to an algorithm for fitting $\mathcal{EL}(\mathcal{I})$-TGD existence and opens up an alternative path to algorithms for fitting $\mathcal{EL}(\mathcal{I})$-ontology existence. It also enables us to prove lower complexity bounds and we in fact show that all four problems are EXPTIME-complete. We also prove tight bounds on the size of fitting TGDs and fitting ontologies, which are identical to the size bounds on finite bases described above.

We next turn to TGDs. For guarded TGDs (GTGDs), we implement exactly the same program described above for $\mathcal{EL}(\mathcal{I})$, but obtain different complexities. We show that finite GTGD-bases always exist and establish a tight single exponential bound on their size. Succinct representation does not help to reduce the size. We give a characterization of fitting GTGD existence and fitting GTGD-ontology existence in terms of products and homomorphisms, show that fitting GTGD existence and fitting GTGD-ontology existence are CONEXPTIME-complete, and give a tight single exponential bound on the size of fitting GTGDs and GTGD-ontologies. The CONEXPTIME upper bound may be obtained either via finite bases or via the semantic characterization.

For all other considered classes of TGDs, the approach via finite bases fails: for the frontier-guarded, frontier-one, and full cases, we prove that finite bases need not exist. For inclusion dependencies, finite bases trivially exist but approaching fitting via this route does not result in an optimal upper complexity bound. For unrestricted TGDs, the existence of finite bases is left open.

We may, however, still approach fitting existence in a direct way or via a semantic characterization. For inclusion dependencies (INDs), we use direct arguments to show that fitting IND existence and fitting IND-ontology existence are NP-complete, and that the size of fitting IND-ontologies is polynomial. For all other remaining cases, we establish semantic characterizations and then use them to approach fitting existence. In this way, we prove the following. Fitting ontology existence and fitting TGD existence are CONEXPTIME-complete for TGDs that are frontier-guarded or frontier-one. For full TGDs, fitting TGD existence is CONEXPTIME-complete and fitting ontology existence is in Σ_2^p and DP-hard. In the case of unrestricted TGDs, both problems are CONEXPTIME-hard and we prove a CO2NEXPTIME upper bound for fitting ontology existence and a CO3NEXPTIME upper bound for fitting TGD existence. We also show tight single exponential size bounds for fitting TGDs and ontologies in the case of frontier-guarded and frontier-one TGDs. We do the same for fitting full TGDs, whereas if a fitting ontology of full TGDs exists, there always exists one of polynomial size. For unrestricted TGDs we give a single exponential lower bound and a triple (for TGDs) and double (for ontologies) exponential upper bound on the size of fittings.

Proofs are provided in the Appendix.

Related Work. There are several related lines of work. One is (ten Cate and Dalmau 2015; Alexe *et al.* 2011) where the fitting problem is studied for schema mappings that take the form of a set of TGDs. While the setup in (Alexe *et al.* 2011) uses universal examples and diverges from ours, the fitting problems considered in (ten Cate and Dalmau 2015) are closely related to the ones studied here. For schema mappings, separate source and target instances (in possibly different schemas) are used in every example while we only have a single instance per example. Upper bounds carry over from the schema mapping setting to our setting because we can choose the source and target instance to be identical, while there seems no easy way to carry over lower bounds. We remark that (ten Cate and Dalmau 2015) consider unrestricted TGDs and so-called GAV constraints, which correspond to single-head full TGDs. They assume a fixed schema (which we do not) and only study sets of TGDs/schema mappings (corresponding to ontologies), but not single TGDs.

Also related is inductive logic programming (ILP), especially in its 'learning from interpretations' incarnation. There, one question of interest is fitting existence (often called the 'consistency problem') for definite first-order clauses, which correspond to single-head full TGDs (Muggleton and Feng 1990; Kietz and Džeroski 1994; Cohen 1994; Gottlob *et al.* 1997). However, the ILP literature typically admits background knowledge as an additional input to the fitting problem and, moreover, adopts additional biases such as a constant number of variables, determinacy conditions that pertain to the functionality of relations, or variable depth. We are not aware that results have been obtained for the unrestricted case studied in this article.

There are also more loosely related lines of work concerning fitting problems for conjunctive queries (ten Cate and Dalmau 2015; ten Cate *et al.* 2023a; ten Cate *et al.* 2023b) and for description logic concepts (Funk *et al.* 2019; Jung *et al.* 2020; Jung *et al.* 2022; ten Cate *et al.* 2023c). We also mention (De Raedt *et al.* 2018), which is concerned with fitting propositional logic constraints to data. Our work on finite bases is related to (Rudolph 2006; Baader and Distel 2008; Distel 2011; Guimarães *et al.* 2023; Kriegel 2024). All those works employ formal concept analysis to construct finite bases while we use a direct approach.

2 Preliminaries

Schema, (Pointed) Instance, Homomorphism. A *schema* $\mathcal{S}$ is a non-empty finite set of *relation symbols* R, each with an associated *arity* $\mathsf{ar}(R) \geq 1$. An $\mathcal{S}$*-fact* is an expression $R(a_1, \ldots, a_m)$, where $a_1, \ldots, a_m$ are *values*, $R \in \mathcal{S}$, and $\mathsf{ar}(R) = m$. An $\mathcal{S}$*-instance* is a (possibly infinite) set I of $\mathcal{S}$-facts. The *active domain* of I, denoted $\mathsf{adom}(I)$, is the set of all values that occur in facts of I. A *pointed* $\mathcal{S}$*-instance* is a pair $(I, \bar{a})$ where I is an $\mathcal{S}$-instance and $\bar{a}$ is a tuple of values. We refer to $\bar{a}$ as the *distinguished values* of $(I, \bar{a})$ and

define its *arity* to be $|\bar{a}|$. The distinguished values need not belong to $\mathsf{adom}(I)$, though most of the time they do.

Given two instances I, J over the same schema, a *homomorphism* is a map $h : \mathsf{adom}(I) \to \mathsf{adom}(J)$ such that $R(\bar{a}) \in I$ implies $R(h(\bar{a})) \in J$ where $h(\bar{a})$ means the component-wise application of h to $\bar{a}$. For pointed instances $(I, \bar{a})$ and $(J, \bar{b})$, we additionally demand that $h(\bar{a}) = \bar{b}$. To indicate that there exists a homomorphism from I to J, we write $I \to J$ and likewise for pointed instances.

Conjunctive Query. A *conjunctive query (CQ)* over a schema $\mathcal{S}$ is a formula $q(\bar{x})$ of the form $\exists \bar{y}\, \varphi(\bar{x}, \bar{y})$ where $\bar{x}$ and $\bar{y}$ are tuples of variables and φ is a conjunction of relational atoms that uses only variables that occur in $\bar{x}$ or $\bar{y}$ and only relation symbols from $\mathcal{S}$. Every variable from $\bar{x}$ must occur in one of the atoms in φ, a condition known as the *safety* of the CQ. The *arity* of q is the number of variables in $\bar{x}$, also called *answer variables*. We use $\mathsf{var}(q)$ to denote the set of all variables that occur in q. We say that q is *guarded* if φ contains an atom that mentions all variables from $\bar{x}$ and $\bar{y}$.

There is a natural correspondence between CQs over a schema $\mathcal{S}$ and finite pointed $\mathcal{S}$-instances of the same arity. First, the *canonical instance* of a CQ $q(\bar{x})$ is $(I_q, \bar{x})$, where $\mathsf{adom}(I_q) = \mathsf{var}(q)$ and the facts of I_q are the atoms in q. Conversely, if $(I, \bar{a})$ satisfies $\bar{a} \subseteq \mathsf{adom}(I) \neq \emptyset$ and no value appears more than once in $\bar{a}$, then the *canonical CQ* of $(I, \bar{a})$ is $q_{(I,\bar{a})}(\bar{a})$ where $\mathsf{var}(q_{(I,\bar{a})}) = \mathsf{adom}(I)$ and the atoms of $q_{(I,\bar{a})}$ are the facts of I.

With a homomorphism from $q(\bar{x})$ to $(I, \bar{a})$, we mean a homomorphism from $(I_q, \bar{x})$ to $(I, \bar{a})$. For a CQ $q(\bar{x})$ over some schema $\mathcal{S}$ and an $\mathcal{S}$-instance I, we use $q(I)$ to denote the set of all *answers* $\bar{a}$ to q on I, that is, all tuples $\bar{a} \subseteq \mathsf{adom}(I)$ such that there is a homomorphism from q to $(I, \bar{a})$.

TGDs, Ontologies. A *tuple-generating dependency* (TGD) over a schema $\mathcal{S}$ is a formula of the form

$$\rho = \forall \bar{x} \forall \bar{y} \big(\varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z}) \big)$$

where $\varphi(\bar{x}, \bar{y})$ and $\exists \bar{z}\, \psi(\bar{x}, \bar{z})$ are conjunctions of relational atoms over $\mathcal{S}$ called the *body* and *head* of ρ. The variables $\bar{x}$, which occur both in the body and in the head, are called *frontier variables*. A TGD is *full* or a FullTGD if the head contains no existentially quantified variables, *guarded* or a GTGD if the body contains an atom that mentions all variables that occur in the body, *frontier-guarded* or an FGTGD if the body contains an atom that mentions all frontier variables, and *frontier-one* or an F1TGD if it has at most one frontier variable. A TGD is an *inclusion dependency* or an IND if both the body and head contain only one atom. The repeated use of variables is admitted. When writing TGDs, the universal quantifiers are usually omitted.

An instance I satisfies a TGD $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, denoted $I \models \rho$, if $\exists \bar{y}\, \varphi(\bar{x}, \bar{y})(I) \subseteq \exists \bar{z}\, \psi(\bar{x}, \bar{z})(I)$, where we view the body and head each as a CQ. An *ontology* is a finite set of TGDs. We speak of a GTGD-ontology if all TGDs are guarded, and likewise for FGTGD and F1TGD. An instance I is a *model* of an ontology $\mathcal{O}$ if it satisfies all TGDs in $\mathcal{O}$. For an ontology $\mathcal{O}$ and a TGD ρ, we write $\mathcal{O} \models \rho$ if every model of $\mathcal{O}$ satisfies ρ. For ontologies $\mathcal{O}, \mathcal{O}'$, we write $\mathcal{O} \models \mathcal{O}'$ if $\mathcal{O} \models \rho$ for every $\rho \in \mathcal{O}'$.

Description Logics $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, $\mathcal{ELI}_\bot$. In the context of description logics, one uses schemas $\mathcal{S}$ that contain only unary and binary relation symbols, also referred to as *concept names* and *role names*. We let A range over concept names of $\mathcal{S}$ and R over the role names of $\mathcal{S}$. The set of $\mathcal{ELI}_\bot$-concepts over $\mathcal{S}$ is given by the grammar

$$C ::= \top \mid \bot \mid A \mid C \sqcap C \mid \exists R.C \mid \exists R^-.C.$$

$\mathcal{EL}_\bot$-concepts are obtained by the same grammar without *inverse roles* R^-, and $\mathcal{ELI}$- and $\mathcal{EL}$-concepts are defined likewise, but do not permit the use of '$\bot$'.

The semantics of DLs is given in terms of interpretations. Every non-empty instance I may be viewed as an interpretation $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ with $\Delta^\mathcal{I} = \mathsf{adom}(I)$, $A^\mathcal{I} = \{d \mid A(d) \in I\}$ for all concept names A, and $R^\mathcal{I} = \{(d, e) \mid R(d, e) \in I\}$ for all role names R. The extension $C^\mathcal{I}$ of an $\mathcal{ELI}_\bot$-concept C is then defined as usual (Baader *et al.* 2017). For easier reference we set $C^I = C^\mathcal{I}$ and, for the empty instance J, $C^J = \emptyset$ for all $\mathcal{ELI}_\bot$-concepts C. We also write Δ^I in place of $\mathsf{adom}(I)$.

Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$. An *$\mathcal{L}$-concept inclusion* over $\mathcal{S}$ takes the form $C \sqsubseteq D$ where C and D are $\mathcal{L}$-concepts over $\mathcal{S}$. An *$\mathcal{L}$-ontology* is a finite set of $\mathcal{L}$-concept inclusions. An instance I satisfies a concept inclusion $C \sqsubseteq D$, written $I \models C \sqsubseteq D$, if $C^I \subseteq D^I$. An instance I is a *model* of an ontology $\mathcal{O}$ if it satisfies all concept inclusions in $\mathcal{O}$. Every $\mathcal{ELI}$-concept inclusion can be expressed as an F1TGD in a straightforward way. From now on we may thus view $\mathcal{ELI}$-ontologies as F1TGD-ontologies and concept inclusions as TGDs.[1] We also remark that by a straightforward normalization which removes syntactic nesting in rule bodies at the cost of introducing additional unary predicates, every $\mathcal{ELI}$-ontology $\mathcal{O}$ can be converted into a GTGD-ontology that is a conservative extension of $\mathcal{O}$. In our context, however, the change of schema matters and results for GTGD-ontologies do not automatically transfer to $\mathcal{ELI}$.

For any syntactic object O such as a conjunctive query, a TGD, or an ontology, we use $||O||$ to denote the *size* of O, meaning the length of O when encoded as a word over a suitable finite alphabet. We next explain what we mean by the succinct representation of $\mathcal{ELI}_\bot$-concepts and ontologies. An $\mathcal{ELI}_\bot$-concept C is a *subconcept* of an $\mathcal{ELI}_\bot$-concept D (ontology $\mathcal{O}$) if C occurs in D (in $\mathcal{O}$) as a syntactic subexpression. An ontology $\mathcal{O}$ (or concept C) can be represented succinctly by using structure sharing, that is, representing every subconcept C of $\mathcal{O}$ only once and using pointers to share that representation among different occurrences of C in $\mathcal{O}$. This representation can sometimes make the size of the representation exponentially smaller. It is easy to see that computationally, succinct representation of concepts and ontologies is almost always harmless. In particular, the following holds.

[1] For uniformity, we will also take the freedom to refer to an $\mathcal{ELI}_\bot$-concept inclusion as a TGD, despite the fact that TGDs do not admit $\bot$ as the rule head.

Lemma 1. *Given a finite instance I and an $\mathcal{ELI}_\bot$-ontology $\mathcal{O}$ in succinct representation, it can be decided in polynomial time whether $I \models \mathcal{O}$.*

Simulation. The expressive power of $\mathcal{EL}$ and $\mathcal{ELI}$ is closely linked to simulations. With an *$\mathcal{EL}$-role* or *$\mathcal{EL}_\bot$-role*, we mean a role name. An *$\mathcal{ELI}$-role* or *$\mathcal{ELI}_\bot$-role* is a role name or an inverse role. Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. For instances I, J, we call a relation $Z \subseteq \Delta^I \times \Delta^J$ an *$\mathcal{L}$-simulation* from I to J if it satisfies the following conditions:

1. If $(d, d') \in Z$, then $d \in A^I$ implies $d' \in A^J$.
2. If $(d, d') \in Z$ and $(d, e) \in R^I$ with R an $\mathcal{L}$-role, then there is a $(d', e') \in R^J$ with $(e, e') \in Z$.

For unary pointed instances $(I, d), (J, e)$ we write $(I, d) \preceq_{\mathcal{L}} (J, e)$ to denote the existence of an $\mathcal{L}$-simulation Z from I to J with $(d, e) \in Z$. As a special case, we also write $(I, d) \preceq_{\mathcal{L}} (J, e)$ if $d \notin \Delta^I$.

With the *role depth* of an $\mathcal{ELI}_\bot$-concept C, we mean its quantifier depth, defined in the standard way. The *outdegree* of C is the maximum number of existential restrictions in any conjunction in C. The following lemma is based on the classic elimination procedure for computing the maximal $\mathcal{L}$-simulation between two interpretations. Since other proofs refer back to it, we recall the proof details in the appendix.

Lemma 2. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. There is a polynomial time algorithm that, given finite pointed instances (I, d) and (J, e), decides whether $(I, d) \preceq_{\mathcal{L}} (J, e)$ and, if this is not the case, outputs the succinct representation of an $\mathcal{L}$-concept C of role depth at most $|\Delta^I| \cdot |\Delta^J|$ and outdegree at most $|\Delta^J|$ such that $d \in C^I$ and $e \notin C^J$.*

Disjoint Union, Direct Product. Let H be a non-empty finite set of instances with pairwise disjoint active domains. Then the *disjoint union* of the instances in H, denoted $\biguplus\mathsf{H}$, is the instance $\bigcup \mathsf{H}$. When the domains of the instances in H are not pairwise disjoint, we assume that renaming is used to achieve disjointness before forming $\biguplus\mathsf{H}$. The *direct product* of two pointed instances $(I, \bar{a})$ and $(J, \bar{b})$, with $\bar{a} = \langle a_1, \ldots a_k \rangle$ and $\bar{b} = \langle b_1, \ldots, b_k \rangle$ is the pointed instance $(I \times J, \bar{a} \times \bar{b})$ where $I \times J$ consists of all facts $R((c_1, d_1), \ldots, (c_n, d_n))$ such that $R(c_1, \ldots, c_n)$ is a fact of I and $R(d_1, \ldots, d_n)$ is a fact of J, and $\bar{a} \times \bar{b} = \langle (a_1, b_1), \ldots, (a_k, b_k) \rangle$. Note that the elements of $\bar{a} \times \bar{b}$ are not necessarily in $\mathsf{adom}(I \times J)$, and in fact this is precisely why we do not require all values of $\bar{a}$ to lie in $\mathsf{adom}(I)$ for a pointed instance $(I, \bar{a})$. The product construction extends naturally to non-empty finite sets S of pointed instances. We then use $\prod S$ to denote its product.

Fitting Problems. Let $\mathcal{L}$ be one of the ontology languages introduced above, such as $\mathcal{ELI}$ and GTGD. A *fitting instance* is a pair (P, N) with P and N finite and non-empty sets of finite instances. Let $\mathcal{S}$ be the set of symbols used in P and N. An $\mathcal{L}$-ontology $\mathcal{O}$ over $\mathcal{S}$ *fits* (P, N) if $I \models \mathcal{O}$ for all $I \in \mathsf{P}$ and $J \not\models \mathcal{O}$ for all $J \in \mathsf{N}$. For a single $\mathcal{L}$-TGD ρ over $\mathcal{S}$, fitting (P, N) is defined in exactly the same way. Note that an ontology $\mathcal{O}$ fitting (P, N) does not imply that every TGD $\rho \in \mathcal{O}$ fits (P, N).

We consider the following computational problems:

PROBLEM:	Fitting $\mathcal{L}$-ontology existence
INPUT:	Fitting instance (P, N).
QUESTION:	Does (P, N) have a fitting $\mathcal{L}$-ontology?

PROBLEM:	Fitting $\mathcal{L}$-ontology construction
INPUT:	Fitting instance (P, N).
OUTPUT:	$\mathcal{L}$-ontology that fits (P, N) if existent, "no fitting" otherwise.

Analogous problems for fitting $\mathcal{L}$-TGDs in place of $\mathcal{L}$-ontologies are defined in the expected way.

Example 1. *Consider the instances*

$$P = \{R(a, b), R(b, a)\}$$
$$N = \{R(a, b), R(b, c), R(c, a)\}.$$

Then $(\{P\}, \{N\})$ has no fitting $\mathcal{ELI}$-TGD or -ontology, but it has fitting GTGDs such as $R(x, y) \to R(y, x)$. Now let

$$N' = N \cup \{R(b, a), R(c, b), R(a, c)\}.$$

Then $(\{P\}, \{N'\})$ has no fitting GTGD or GTGD-ontology. But it has fitting F1TGDs such as

$$R(x, y) \wedge R(y, z) \wedge R(z, x) \to R(x, x).$$

All negative claims in this example are a consequence of the semantic characterizations established below.

How are fitting ontologies and fitting TGDs related? The following is an immediate consequence of the definition of fitting and the semantics of ontologies and TGDs.

Lemma 3. *Let (P, N) be a fitting instance. Then there is an $\mathcal{L}$-ontology that fits (P, N) if and only if for every $N \in \mathsf{N}$, there is an $\mathcal{L}$-TGD that fits $(\mathsf{P}, \{N\})$.*

This clearly implies the following.

Lemma 4. *Let (P, N) be a fitting instance. If there is an $\mathcal{L}$-ontology that fits (P, N) then there is an $\mathcal{L}$-ontology $\mathcal{O}$ that fits (P, N) and contains at most $|\mathsf{N}|$ TGDs.*

We also make an observation regarding full TGDs and the number of head atoms. If there is a full TGD ρ that fits an instance $(\mathsf{P}, \{N\})$ with a single negative example, then there is a fitting FullTGD with a single head atom. In fact, there must be a head atom falsified in the single negative example, and we may drop all other head atoms. This is not the case, however, if more than one negative example is present.

Example 2. *Let $\mathsf{P} = \{P\}$ with $P = \{A(a), B_1(a), B_2(a)\}$ and $\mathsf{N} = \{N_1, N_2\}$ with $N_i = \{A(a), B_i(a)\}$ for $i \in \{1, 2\}$. Then the full TGD $A(x) \to B_1(x) \wedge B_2(x)$ fits (P, N), but there is no fitting FullTGD with a single head atom.*

By Lemma 3, this implies that if there is a FullTGD-ontology that fits an instance (P, N), with N containing any number of examples, then there is a FullTGD-ontology that fits (P, N) and in which every TGD has a single head atom.

Finite Bases. Let H be a non-empty finite set of finite $\mathcal{S}$-instances, for some schema $\mathcal{S}$, and let $\mathcal{L}$ be one of the ontology languages introduced above. An $\mathcal{L}$-ontology $\mathcal{O}$ over $\mathcal{S}$ is a *finite $\mathcal{L}$-basis* of H if for all $\mathcal{L}$-TGDs ρ, the following holds:

$$\mathcal{O} \models \rho \text{ iff } I \models \rho \text{ for all } I \in \mathsf{H}.$$

If $\mathsf{H} = \{I\}$ is a singleton, we also say that $\mathcal{O}$ is a *finite $\mathcal{L}$-basis* of I. We study the following computational problem:

PROBLEM:	$\mathcal{L}$-basis construction
INPUT:	A finite set $\mathsf{H} \neq \emptyset$ of finite instances.
OUTPUT:	A finite $\mathcal{L}$-basis of H.

We remind the reader that the entailment $\mathcal{O} \models \rho$ is defined relative to all instances, including infinite ones. An alternative definition of finite bases would use entailment only over finite models. However, in almost all of the cases studied in this paper, finite and unrestricted entailment coincide. The only exception is the case of unrestricted TGDs (Beeri and Vardi 1981).

The following lemma connects finite basis construction with fitting $\mathcal{L}$-ontology existence. Informally, it states that a finite basis of the positive examples is a canonical candidate for a fitting $\mathcal{L}$-ontology.

Lemma 5. *Let (P, N) be a fitting instance and let $\mathcal{O}_\mathsf{P}$ be a finite $\mathcal{L}$-basis of P. Then $\mathcal{O}_\mathsf{P}$ fits (P, N) if and only if (P, N) has a fitting $\mathcal{L}$-ontology.*

For some choices of $\mathcal{L}$, finite bases of a set H of instances coincide with finite bases of the single instance obtained by taking the disjoint union of all instances in H. It in fact follows directly from the definition of finite bases that this is the case if $\mathcal{L}$-ontologies $\mathcal{O}$ are *invariant under disjoint union*, that is, for all non-empty finite sets of instances H: $\biguplus \mathsf{H} \models \mathcal{O}$ if and only if $I \models \mathcal{O}$ for all $I \in \mathsf{H}$. It is well known that $\mathcal{ELI}_\bot$-ontologies are invariant under disjoint union while this is not the case for any of the classes of TGDs that we consider.

3 Finite Bases for $\mathcal{EL}$ and $\mathcal{ELI}$

We reprove that finite bases always exist in $\mathcal{EL}$ and prove that the same is true for $\mathcal{ELI}$, both with and without '$\bot$'. We also show that, in both cases, finite bases can be constructed in double exponential time, and in single exponential time if represented succinctly. This also establishes upper bounds on the size of finite bases and we prove matching lower bounds. Our main result is the following.

Theorem 1. *In $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, finite bases can be constructed in double exponential time, and in single exponential time if ontologies are represented succinctly.*

By Lemmas 1 and 5, we immediately get the following.

Corollary 1. *In $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, fitting ontology existence is decidable in* EXPTIME *and fitting ontologies can be constructed in double exponential time, and in exponential time if ontologies are represented succinctly.*

We now prove Theorem 1, treating all four cases simultaneously. Thus, for the remainder of this section, let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$. Since $\mathcal{L}$-ontologies are invariant under disjoint union, we may without loss of generality concentrate on finite bases of single instances.

Let I be a finite instance given as an input to the $\mathcal{L}$-basis construction problem. A set $X \subseteq \Delta^I$ is *$\mathcal{L}$-definable* if there is an $\mathcal{L}$-concept C such that $X = C^I$. For every $\mathcal{L}$-definable $X \subseteq \Delta^I$, choose an $\mathcal{L}$-concept E^*_X such that $(E^*_X)^I = X$. Note that if $\bot$ is an $\mathcal{L}$-concept, then the empty set is $\mathcal{L}$-definable. Otherwise, this is the case if and only if Δ^I contains no value d that is *$\mathcal{L}$-total* in I, meaning that $d \in C^I$ for all $\mathcal{L}$-concepts C over the schema $\mathcal{S}$ of I. We generally assume that $E^*_{\Delta^I} = \top$ and, provided that $\bot$ is an $\mathcal{L}$-concept, $E^*_\emptyset = \bot$.

Example 3. *Let $\mathcal{S} = \{R\}$, R a binary relation symbol, and let $I_1 = \{R(a,b)\}$ and $I_2 = \{R(a,b), R(c,c)\}$. Then the empty set is $\mathcal{EL}$-definable in I_1 by the concept $\exists R.\exists R.\top$ while the empty set is not $\mathcal{ELI}$-definable in I_2. It is, however, trivially $\mathcal{EL}_\bot$-definable by $\bot$.*

In the following, we sometimes consider $\mathcal{L}$-concepts that contain subsets $X \subseteq \Delta^I$ as subconcepts. We interpret those in I by setting $X^I = X$.

Define the ontology $\mathcal{O}_I$ to contain the following concept inclusions, for all $\mathcal{L}$-definable sets X, X', Y:

1. $E^*_X \sqsubseteq A$ for all concept names A with $I \models X \sqsubseteq A$;
2. $E^*_X \sqsubseteq \exists R.E^*_Y$ for all $\mathcal{L}$-roles R with $I \models X \sqsubseteq \exists R.Y$;
3. $A \sqsubseteq E^*_X$ for all concept names A with $I \models A \sqsubseteq X$;
4. $\exists R.E^*_X \sqsubseteq E^*_Y$ for all $\mathcal{L}$-roles R with $I \models \exists R.X \sqsubseteq Y$;
5. $E^*_X \sqcap E^*_{X'} \sqsubseteq E^*_Y$ if $I \models X \sqcap X' \sqsubseteq Y$;

We show the following in the appendix.

Lemma 6. *$\mathcal{O}_I$ is a finite $\mathcal{L}$-basis of I.*

The construction of $\mathcal{O}_I$ hinges on effectively identifying the $\mathcal{L}$-definable sets X and constructing the corresponding $\mathcal{L}$-concepts E^*_X. We use brute force enumeration over all sets $X \subseteq \Delta^I$ and then, for each X, apply the algorithm that achieves the following, based on Lemma 2.

Lemma 7. *There is an algorithm that, given as input an instance I and a set $X \subseteq \Delta^I$, in double exponential time outputs an $\mathcal{L}$-concept C with $C^I = X$ if such a concept exists, and reports 'undefinable' otherwise. The algorithm can be modified to run in single exponential time and output a succinct representation of C.*

Results in formal concept analysis imply a single exponential lower bound on the size of finite $\mathcal{EL}$-bases, even when no role names are present (Kuznetsov 2004; Kriegel 2024). The same is also known when the schema contains a single role name and no concept names (Guimarães *et al.* 2023). We establish a tight double exponential lower bound that also applies to $\mathcal{ELI}$ and a tight single exponential lower bound when the basis is represented succinctly. In fact, the following is a consequence of Lemma 5 and Theorem 6 below.

Theorem 2. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$. For every $n \geq 1$, there is a finite instance I_n such that*

1. *the size of I_n is bounded by $p(n)$, p a polynomial;*
2. *the smallest finite $\mathcal{L}$-basis $\mathcal{O}$ of I_n is of size 2^{2^n}, and of size 2^n when represented succinctly.*

4 Fitting Existence for $\mathcal{EL}$ and $\mathcal{ELI}$

We consider fitting TGD existence and fitting ontology existence in $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, proceeding in parallel. We establish semantic characterizations in terms of products and simulations which suggest an alternative algorithm for fitting existence and construction with optimal complexity, reproving Corollary 1. They also support proving lower complexity bounds.

Example 4. *Having $\bot$ or not makes a difference. Let $\mathcal{S} = \{R\}$ with R binary, $P = \{R(a,b)\}$, and $N = \{R(a,a)\}$. Then $\exists R.\exists R.\top \sqsubseteq \bot$ fits $(\{P\},\{N\})$, but it follows from the characterizations below that there is no fitting $\mathcal{ELI}$-ontology.*

We next give a model-theoretic characterization for fitting TGD existence. Via Lemma 3, it also provides a characterization for fitting ontology existence.

Theorem 3. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. Let (P,N) be a fitting instance where $\mathsf{N} = \{N_1,\dots,N_k\}$ and let $P = \biguplus\mathsf{P}$. Then no $\mathcal{L}$-TGD fits (P,N) if and only if for every $\bar{d} = (d_1,\dots,d_k) \in \Delta^{\prod \mathsf{N}}$ such that d_i is non-$\mathcal{L}$-total in N_i for all $i \in [k]$, the following conditions are satisfied:*

1. *the set $S_{\bar{d}} = \{(P,e) \mid (\prod \mathsf{N}, \bar{d}) \preceq_{\mathcal{L}} (P,e)\}$ is non-empty;*
2. *$\prod S_{\bar{d}} \preceq_{\mathcal{L}} (N_i, d_i)$ for some $i \in [k]$.*

The same is true for $\mathcal{L}_\bot$ when the condition 'd_i non-$\mathcal{L}$-total in N_i for all $i \in [k]$' is dropped.

Theorem 3 provides us with an alternative algorithm for fitting TGD existence and fitting ontology existence, by checking the conditions given there. This reproves the upper bounds in fitting ontology existence from Corollary 1 and gives the same upper bounds for fitting TGD existence. The proofs are constructive in the sense that they also yield algorithms for fitting TGD and ontology construction. Unlike in the approach via finite bases, the constructed ontologies contain only one TGD per negative example.

Theorem 4. *In $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, fitting TGD existence is decidable in* ExpTime *and fitting TGD construction is possible in double exponential time, and in exponential time if TGDs are represented succinctly. The same is true for fitting ontology existence and construction.*

We next show that the obtained upper bounds for fitting existence are optimal. This is done by a non-trivial reduction from the product simulation problem, which is known to be ExpTime-hard (Harel *et al.* 2002).

Theorem 5. *In $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, fitting TGD existence and fitting ontology existence are* ExpTime*-hard.*

We finally establish tight lower bounds on the size of fitting TGDs and ontologies. This builds on lower bounds on the size of fitting $\mathcal{EL}$-concepts in concept learning, obtained in (Funk 2019).

Theorem 6. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$. For every $n \geq 1$, there is a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ such that*

1. *the size of $(\mathsf{P}_n, \mathsf{N}_n)$ is bounded by $p(n)$, p a polynomial;*
2. *$(\mathsf{P}_n, \mathsf{N}_n)$ admits a fitting $\mathcal{L}$-TGD and a fitting $\mathcal{L}$-ontology;*
3. *the smallest $\mathcal{L}$-TGD that fits $(\mathsf{P}_n, \mathsf{N}_n)$ is of size at least 2^{2^n} and of size 2^n when represented succinctly, and the same is true for the smallest fitting $\mathcal{L}$-ontology.*

5 Finite Bases for TGDs

We show that finite bases for guarded TGDs always exist and that basis construction is possible in single exponential time. This also yields a single exponential upper bound on the size of bases, and we establish a matching lower bound as well. For inclusion dependencies, the existence of finite bases is trivial as there are only finitely many IDs over a fixed schema $\mathcal{S}$. For frontier-one TGDs, frontier-guarded TGDs, and full TGDs, we prove that finite bases do not always exist. Finite basis existence remains open for unrestricted TGDs.

We remark that we have defined TGDs so that '$\bot$' is not admitted as a head. The results in this section and the subsequent one can be adapted to that case, in analogy with the differences between $\mathcal{ELI}$ and $\mathcal{ELI}_\bot$ in the previous sections. We start with the following positive result.

Theorem 7. *GTGD-basis construction is possible in single exponential time.*

By Lemma 5, we immediately get the following.

Corollary 2. *Fitting GTGD-ontology existence is decidable in* coNExpTime *and fitting GTGD-ontology construction is possible in single exponential time.*

Since guarded TGDs are not invariant under disjoint union, we directly construct finite bases of *sets* of instances. We start by introducing a preliminary notion. Let $(I,\bar{a})$ be a pointed instance with $\bar{a} = \langle a_1,\dots,a_k\rangle$. We use $(I^*,\bar{a}^*)$ to denote the *diversification* of $(I,\bar{a})$, that is, the pointed instance obtained from $(I,\bar{a})$ by introducing fresh and distinct values $\bar{a}^* = \langle a_1^*,\dots,a_k^*\rangle$ and adding to I each a_i^* as a 'clone' of a_i. More precisely, I^* consists of all facts that can be obtained from a fact in I by replacing, for every a_i, zero or more occurrences of a_i with a_i^*. Note that $(I^*,\bar{a}^*) \to (I,\bar{a})$ while the converse does not hold in general, since $\bar{a}$ may contain repeated values whereas $\bar{a}^*$ does not (which is, in fact, the aim of diversification).

Let H be a non-empty finite set of $\mathcal{S}$-instances, for some schema $\mathcal{S}$. We construct an ontology $\mathcal{O}_\mathsf{H}$ that contains one guarded TGD for each of the (finitely many) guarded CQs over $\mathcal{S}$. Let $q(\bar{x}) = \exists\bar{y}\,\varphi(\bar{x},\bar{y})$ be such a CQ. We distinguish two cases:

1. $q(I) = \emptyset$ for all $I \in \mathsf{H}$.
   Then $\mathcal{O}_\mathsf{H}$ contains the rule $\varphi \to \psi$ where ψ is the conjunction of all possible atoms that use only the variables from $\bar{x}$ and the relation symbols from $\mathcal{S}$;
2. $q(I) \neq \emptyset$ for some $I \in \mathsf{H}$. Set
   $$S = \{(I,\bar{a}) \mid I \in \mathsf{H}, \bar{a} \in q(I)\},$$
   let $(P,\bar{b}) = \prod S$ and let $q_{(P^*,\bar{b}^*)} = \exists\bar{z}\,\psi$ be the canonical CQ of $(P^*,\bar{b}^*)$. Then $\mathcal{O}_\mathsf{H}$ contains the rule $\varphi \to \exists\bar{z}\,\psi'$ where ψ' is obtained from ψ by renaming the answer variables of $q_{(P^*,\bar{b}^*)}$ to $\bar{x}$. Note that the renaming relies on $\bar{b}^*$ not containing repeated values.

Example 5. *Let H consist of the single $\mathcal{S}$-instance $I = \{R(a,a)\}$ where $\mathcal{S} = \{R\}$ with R a binary relation. The guarded CQ $q(x) = \exists y\,R(x,y)$ induces in $\mathcal{O}_\mathsf{H}$ the TGD*

$$R(x,y) \to \exists z\,\big(R(x,x) \wedge R(x,z) \wedge R(z,x) \wedge R(z,z)\big).$$

Now consider the guarded CQ $q(x,y) = R(x,y)$. It induces in $\mathcal{O}_\mathsf{H}$ the TGD

$$\begin{aligned} R(x,y) \to \exists z\,\big(&R(x,x) \wedge R(x,y) \wedge R(x,z) \wedge \\ &R(y,x) \wedge R(y,y) \wedge R(y,z) \wedge \\ &R(z,x) \wedge R(z,y) \wedge R(z,z)\big). \end{aligned}$$

Note that this may be viewed as a weakening of the equality-generating dependency $R(x,y) \to x = y$, which is true in I, but not expressible as a guarded TGD.

We show the following in the appendix.

Lemma 8. *$\mathcal{O}_{\mathsf{H}}$ is a finite GTGD-basis of H.*

The basis $\mathcal{O}_{\mathsf{H}}$ constructed above is of double exponential size: there are double exponentially many possible guarded CQs which are all used as the body of a TGD in $\mathcal{O}_{\mathsf{H}}$. Each single TGD, however, is only of single exponential size. To reduce the size of $\mathcal{O}_{\mathsf{H}}$ to single exponential, it is thus enough to show that we can select single exponentially many TGDs from $\mathcal{O}_{\mathsf{H}}$ and still obtain a finite basis of H. To achieve this, we take inspiration from the case of $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$ where we were concerned only with subsets of instances that are $\mathcal{L}$-definable. Analogously, in Point 2 of the definition of $\mathcal{O}_{\mathsf{H}}$, we should only be interested in sets S that are definable by a guarded CQ q. In contrast to the case of $\mathcal{EL}$ and $\mathcal{ELI}$, however, it does not suffice to choose one guarded CQ for every definable set S. Instead, we need to consider all guarded CQs q that define S and are minimal in the sense that any CQ obtained from q by dropping an atom no longer defines S. This gives rise to the following.

Theorem 8. *Let $\mathcal{O}'_{\mathsf{H}}$ be the subset of $\mathcal{O}_{\mathsf{H}}$ that contains all TGDs from $\mathcal{O}_{\mathsf{H}}$ whose body has at most $n := ||\mathsf{H}|| + 1$ atoms. Then $\mathcal{O}'_{\mathsf{H}}$ is a GTGD-basis of H with $O(S^n \cdot \ell^{\ell n})$ TGDs where S is the number of relation symbols in the schema $\mathcal{S}$ of H, and ℓ is the maximal arity of a symbol in $\mathcal{S}$.*

We now consider lower bounds on the size of GTGD-bases. As a consequence of Lemma 5 and Theorem 21 below, we may obtain a single exponential lower bound. That bound, however, does not apply to singleton sets H. Here, we show the same result also for this case.

Theorem 9. *For every $n \geq 1$, there is an instance I_n (over a schema with only unary and binary relations) such that*

1. *the size of I_n is bounded by $p(n)$, p a polynomial;*
2. *the smallest finite GTGD-basis $\mathcal{O}$ of I_n is of size 2^n.*

Proof. Let P and R be unary and binary symbols, respectively. For $m \geq 1$, let L_m denote the "lasso" instance, with values $a_0^m, \ldots, a_{2m-1}^m$ and facts

$$\{R(a_i^m, a_{i+1}^m) \mid i < 2m-1\} \cup \{R(a_{2m-1}^m, a_m^m), P(a_m^m)\}$$

Then define, for $n \geq 1$, $I_n = \bigcup_{i=1}^{n} (L_{p_i} \cup \{A(a_0^{p_i})\})$ where p_i is the i-th prime number. We show in the appendix that I_n is as required. ❑

Let us now discuss finite bases for inclusion dependencies. Over any schema $\mathcal{S}$ with relations of maximum arity k, there are at most $|\mathcal{S}|^2(2k)^{2k}$ many inclusion dependencies. Therefore, a finite basis for a non-empty set H of finite instances always exists: we can simply use the set of all INDs that are true in H. If the arity of relation symbols is bounded by a constant, then this basis is only of polynomial size. Otherwise, exponential size cannot be avoided.

Theorem 10. *There are finite instances $I_1, I_2, \ldots$ such that for all $n \geq 1$, the size of I_n is polynomial in n, but all finite IND-bases of I_n contain at least 2^n INDs.*

Via Lemma 5 and the fact that INDs can be evaluated in polynomial time, we also obtain decidability of fitting IND-ontology existence in ExpTime. However, we will see in the subsequent section that this complexity is not optimal and thus do not state the result here as a formal theorem. We also obtain the following.

Theorem 11. *Fitting IND-ontology construction is possible in single exponential time.*

We do not know whether the bound in Theorem 11 is optimal.

We next prove that for frontier-guarded TGDs and for frontier-one TGDs, finite bases do not always exist.

Theorem 12. *There are instances I that have no finite FGTGD-basis and no finite F1TGD-basis.*

To prove Theorem 12, we use the rather simple instance $I = \{R(a,b), R(b,a)\}$. For every $n \geq 1$, consider the frontier-one TGD

$$\rho_n = \bigwedge_{i \in [n-1]} R(x_i, x_{i+1}) \wedge R(x_n, x_1) \to R(x_1, x_1).$$

The TGD expresses that if x_1 lies on a cycle of length n, then x_1 has a reflexive loop. We have $I \models \rho_n$ for all odd n because no cycle of odd length homomorphically maps to I. Note that the rule bodies of the TGDs ρ_n with n odd get larger with increasing n. Intuitively, this means that also the rule bodies of any finite FGTGD-basis of I must be of unbounded size, which means that there is no finite FGTGD-basis. In the appendix, we formally prove that I indeed neither has a finite FGTGD-basis nor a finite F1TGD-basis.

We now consider full TGDs. The instance I used in the proof of Theorem 12 is not suitable here because the set

$$\begin{aligned} \mathcal{O}_I = \{ \quad & R(x,y) \to R(y,x) \\ & R(x,y) \wedge R(y,z) \wedge R(z,u) \to R(x,u) \\ & R(x,x) \wedge \mathsf{true}(y) \wedge \mathsf{true}(z) \to R(y,z) \ \} \end{aligned}$$

is a finite FullTGD-basis of I. The third TGD above represents four TGDs, as $\mathsf{true}(v)$ is a placeholder for either $R(u,v)$ or $R(v,u)$ with u a fresh variable and $v \in \{y,z\}$. In fact, we even have the following.

Lemma 9. *$\mathcal{O}_I$ is a finite FullTGD-basis of I and also a finite TGD-basis.*

Let $J = \{R(u,v) \mid u,v \in \{a,b,c\},\ u \neq v\}$. It is well known and easy to see that an undirected graph G has a homomorphism to the undirected 3-clique if and only if G is 3-colorable. Note that J is essentially an undirected 3-clique except that undirected edges are replaced with bidirectionally directed edges. We may modify any undirected graph G in the same way, making it bidirectional. For every graph G, consider the full TGD

$$\rho_G = \varphi_G \to R(x,x)$$

where φ_G is the conjunction of the edges of G viewed as R-atoms and x is a vertex in G chosen arbitrarily. By what was said above, it is clear that $J \models \rho_G$ if and only if G is not 3-colorable. It is known that for any $m \geq 0$, there are non-3-colorable graphs of girth exceeding m (Erdős 1959). We

remind the reader that the girth of a graph is the length of a shortest cycle in it (and ∞ if the graph is acyclic). Intuitively, this means that also the rule bodies of any finite FullTGD-basis of J must be of unbounded size, and thus there is no such basis.

Theorem 13. *There are instances J that have no finite FullTGD-basis.*

We remark in passing that there appears to be a loose connection between finite FullTGD-bases and constraint satisfaction problems. Indeed, it is not very difficult to see that any finite FullTGD-basis $\mathcal{O}$ of a finite instance I gives rise to a datalog-rewriting of the complement of the (non-uniform) constraint satisfaction problem CSP(I) obtained by using I as a template. For more background on these notions, we refer to (Feder and Vardi 1998).

6 Fitting Existence for TGDs

We study fitting TGD existence and fitting ontology existence for various classes of TGDs. As in Section 4, we start with semantic characterizations that are the basis for developing decision procedures and determining the computational complexity. In the case of GTGDs, the characterization allows us to reprove the upper complexity bound from Corollary 2. For frontier-guarded TGDs, frontier-one TGDs, and full TGDs, where the approach via finite bases is precluded, we may nevertheless use our characterizations to obtain algorithms for fitting TGD existence and fitting ontology existence. The same is true for unrestricted TGDs, for which the existence of finite bases remains open. We identify tight complexity bounds for all studied existence problems, with the exception of full TGDs and unrestricted TGDs where a gap remains, and also tight bounds on the size of fitting TGDs and ontologies.

We start with the case of inclusion dependencies. Here we skip a characterization because, due to the simplicity of INDs, they are rather convenient to deal with directly. Since an IND that fits a given fitting instance must be of size linear in the size of that instance, fitting IND existence is clearly in NP. By Lemma 3, the same is true for fitting IND-ontologies. We prove a matching lower bound by reduction from 3SAT. A core idea is to use two relation symbols R and S that provide one position for each literal in the 3SAT formula given as an input, and to represent truth values in an IND by distinguishing whether a position in the head atom holds the same variable as the corresponding position in the body atom or an existentially quantified variable.

Theorem 14. *Fitting IND existence and fitting IND-ontology existence are* NP*-complete.*

Turning to other classes of TGDs, we first give some preliminaries. Let I be an $\mathcal{S}$-instance. A k-tuple $\bar{a} \subseteq \mathsf{adom}(I)$ is *total* in I if $\bar{a} \in q(I)$ for all k-ary CQs $q(\bar{x})$ over $\mathcal{S}$. Note that this is the case if and only if I contains every possible fact built from a relation symbol in $\mathcal{S}$ and values from $\bar{a}$. A set $M \subseteq \mathsf{adom}(I)$ is *guarded* if there is a fact $R(\bar{a}) \in I$ such that $M \subseteq \bar{a}$, and M is *maximally guarded* if there is no guarded set $N \subseteq \mathsf{adom}(I)$ with $M \subsetneq N$. Fix an arbitrary order on $\mathsf{adom}(I)$. For a set $M \subseteq \mathsf{adom}(I)$, we write $\bar{M}$ to denote the tuple that contains each element of M exactly once, adhering to the fixed order. By $I|_M$ we denote the subset of I that consists of precisely those facts that use only values from M. Consider instances $I_1, \ldots, I_k$ and an n-tuple

$$\bar{b} \in \mathsf{adom}(\prod_{i=1}^{k} I_i)^n.$$

We may write $\bar{b}$ as $\bar{a}_1 \times \cdots \times \bar{a}_k$ where $\bar{a}_i \in \mathsf{adom}(I_i)^n$ for all $i \in [k]$. We then use $\bar{b}[i]$ to denote $\bar{a}_i$. We next present a characterization of fitting GTGD existence. Via Lemma 3, it also applies to fitting ontology existence.

Theorem 15. *Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \ldots, N_k\}$. Then no GTGD fits (P, N) if and only if for every non-empty maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, the following conditions are satisfied:*

1. *the following set is non-empty:*

$$S_M = \{(J, \bar{b}) \mid J \in \mathsf{P} \text{ and } \bar{b} \in \mathsf{adom}(J)^{|M|} \text{ such that } (\textstyle\prod \mathsf{N}|_M, \bar{M}) \to (J, \bar{b})\}$$

2. *$\exists i \in [k]$: $(K^*, \bar{c}^*) \to (N_i, \bar{M}[i])$ where $(K, \bar{c}) = \prod S_M$.*

Characterizations for other classes of TGDs can be obtained by varying the characterization in Theorem 15. For brevity, we only give the differences. For the convenience of the reader, all three characterizations are given in an explicit form in the appendix.

Theorem 16. *Consider Theorem 15 modified so that in the definition of the set S_M, '$(\prod \mathsf{N}|_M, \bar{M})$' is replaced with '$(\prod \mathsf{N}, \bar{M})$'. The resulting theorem holds for*

1. *FGTGD;*
2. *F1TGD when, additionally, singleton sets are considered for M instead of maximally guarded sets;*
3. *TGD when, additionally, unrestricted sets are considered for M instead of maximally guarded sets.*

The characterization for full TGDs requires some more changes, we state a self-contained version.

Theorem 17. *Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \ldots, N_n\}$. Then no FullTGD fits (P, N) if and only if, for all relation symbols $R_1, \ldots, R_n$ and tuples $\bar{a}_1, \ldots, \bar{a}_n$ such that*

$$\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)} \text{ and } R_i(\bar{a}_i[i]) \notin N_i \quad \text{for } i \in [n],$$

there is a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P such that $R_j(h(\bar{a}_j)) \notin P$ for some $j \in [n]$.

Moreover, if (P, N) admits a fitting FullTGD, then it admits one in which the number of head atoms is bounded by the number of examples in N.

The above characterizations give rise to algorithms for fitting TGD existence and fitting ontology existence.

Theorem 18.

1. *For $\mathcal{L} \in \{$GTGD, FGTGD, F1TGD$\}$, fitting $\mathcal{L}$-ontology existence and fitting $\mathcal{L}$-TGD existence are in* coNExpTime.

2. *For full TGDs, fitting ontology existence is in Σ_2^p (and in* coNP *if the arities of relation symbols are bounded by a constant), and fitting TGD existence is in* coNExpTime.
3. *For unrestricted TGDs, fitting ontology existence is in* co2NExpTime *and fitting TGD existence is in* co3NExpTime.

To prove Point 1 of Theorem 18, the characterizations for GTGDs can be implemented straightforwardly. Note that the number of sets M is linear and thus the sets S_M can be computed deterministically in single exponential time, checking $(\prod \mathsf{N}|_M, \bar{M}) \rightarrow (J, \bar{b})$ by brute force enumeration. For frontier-guarded and frontier-one TGDs, in contrast, there is no obvious way to compute the sets S_M in coNExpTime. This is because $(\prod \mathsf{N}|_M, \bar{M})$ is replaced by $(\prod \mathsf{N}, \bar{M})$, which has single exponentially many values rather than linearly many, and thus brute force enumeration no longer works. In the appendix, we show how to get around this problem.

For Point 2, the characterization can be implemented straightforwardly. The difference in complexity between fitting ontology existence and fitting TGD existence is due to the fact that, by Lemma 3, fitting ontology existence corresponds to fitting TGD existence with a single negative example. But in this case the product $\prod \mathsf{N}$ is of course of linear size only, rather than of single exponential size.

For Point 3, there are double exponentially many sets M to be considered (as these no longer need to be guarded) and the sets S_M are of double exponential size and the products $\prod S_M$ are of triple exponential size. This explains the co3NExpTime upper bound for fitting TGD existence. If there is only a single negative example as in ontology fitting via Lemma 3, the number of sets S_M and the size of the products $\prod S_M$ reduce by one exponential, and the complexity drops to co2NExpTime.

In the proofs of Theorems 15, 16, and 17, concrete fitting TGDs and ontologies are constructed. A straightforward analysis of their size yields the following.

Theorem 19.

1. *Let $\mathcal{L} \in \{GTGD, FGTGD, F1TGD\}$ and let (P, N) be a fitting instance. If there is a fitting $\mathcal{L}$-TGD or a fitting $\mathcal{L}$-ontology for (P, N), then there is one of single exponential size that can be constructed in double exponential time.*
2. *The same is true for full TGDs where, in the case of fitting ontologies, there is even a fitting ontology of polynomial size that can be constructed in single exponential time.*
3. *If there is a fitting TGD (fitting TGD-ontology) for (P, N), then there is one of triple (double) exponential size that can be constructed in quadruple (triple) exponential time.*

We now establish lower complexity bounds.

Theorem 20.

1. *Let $\mathcal{L} \in \{GTGD, FGTGD, F1TGD, TGD\}$. Then fitting $\mathcal{L}$-TGD existence and fitting $\mathcal{L}$-ontology existence are* coNExpTime*-hard.*
2. *For full TGDs, fitting TGD existence is* coNExpTime*-hard and fitting ontology existence is DP-hard (and* coNP*-hard if the arities of relation symbols are bounded by a constant).*

We thus obtain coNExpTime-completeness for GTGD, FGTGD, and F1TGD. All coNExpTime lower bounds in Theorem 20 are proved by reduction from the product homomorphism problem (ten Cate and Dalmau 2015) and apply already to schemas that contain only binary relation symbols. The coNExpTime lower bounds in Points 1 and 2 are based on different constructions. This is because the former exploit the unrestricted heads of the TGD classes considered there while the latter exploits the unrestricted bodies of full TGDs.

We finally turn to lower bounds on the size of fitting TGDs and ontologies. These can be derived from results implicit in (ten Cate and Dalmau 2015; Willard 2010). The following theorem yields lower bounds for the size of fitting TGDs and ontologies for the cases of GTGD, FGTGD, F1TGD, and unrestricted TGDs.

Theorem 21. *For every $n \geq 1$, there is a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ such that*

1. *the size of $(\mathsf{P}_n, \mathsf{N}_n)$ is bounded by $p(n)$, p a polynomial;*
2. *$(\mathsf{P}_n, \mathsf{N}_n)$ admits a fitting guarded and frontier-one TGD;*
3. *the smallest TGD fitting $(\mathsf{P}_n, \mathsf{N}_n)$ has size 2^n.*

The same is true for ontologies instead of single TGDs.

Note that for GTGD, FGTGD, and F1TGD, these bounds are tight. We have the same lower bounds also for full TGDs (where they are not tight).

Theorem 22. *For every $n \geq 1$, there is a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ such that*

1. *the size of $(\mathsf{P}_n, \mathsf{N}_n)$ is bounded by $p(n)$, p a polynomial;*
2. *$(\mathsf{P}_n, \mathsf{N}_n)$ admits a fitting full TGD;*
3. *the smallest full TGD fitting $(\mathsf{P}_n, \mathsf{N}_n)$ has size 2^n.*

For inclusion dependencies, it is an easy consequence of Lemma 3 that if a fitting IND-ontology exists, then there is one of polynomial size.

7 Conclusion

We have studied finite bases as well as TGD and ontology fittings for several description logics and classes of TGDs. We mention some interesting open questions. We would like to know the exact complexity of deciding fitting TGD existence and fitting ontology existence for unrestricted TGDs. Note that the same gap left open here also shows up in (ten Cate and Dalmau 2015). We would also like to know whether finite bases always exist for unrestricted TGDs. We conjecture that this is not the case and that the instance J used in Section 5 to show that finite FullTGD-bases need not exist can be used to show that. For the time being, however, we do not have a proof. It would also be interesting to characterize precisely the sets of instances H that have a finite basis of frontier-guarded TGDs, and likewise for frontier-one TGDs and full TGDs.

Acknowledgments

This work is partly supported by BMFTR in DAAD project 57616814 (SECAI). The second and third authors were supported by DFG project JU 3197/1-1.

References

Bogdan Alexe, Balder ten Cate, Phokion Kolaitis, and Wang-Chiew Tan. Designing and refining schema mappings via data examples. *Proc. of SIGMOD*, pages 133–144, 2011.

Franz Baader and Felix Distel. A finite basis for the set of EL-implications holding in a finite model. *Proc. of ICFCA*, 4933:46–61, 2008.

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

Pablo Barceló and Miguel Romero. The complexity of reverse engineering problems for conjunctive queries. *Proc. of ICDT*, 68:7:1–7:17, 2017.

Catriel Beeri and Moshe Y. Vardi. The implication problem for data dependencies. *Proc. of ICALP*, 115:73–85, 1981.

Balder ten Cate and Víctor Dalmau. The product homomorphism problem and applications. *Proc. of ICDT*, 31:161–176, 2015.

Balder ten Cate, Víctor Dalmau, Maurice Funk, and Carsten Lutz. Extremal fitting problems for conjunctive queries. *Proc. of PODS*, pages 89–98, 2023.

Balder ten Cate, Maurice Funk, Jean Christoph Jung, and Carsten Lutz. Fitting algorithms for conjunctive queries. *SIGMOD Rec.*, 52(4):6–18, 2023.

Balder ten Cate, Maurice Funk, Jean Christoph Jung, and Carsten Lutz. SAT-based PAC learning of description logic concepts. *Proc. of IJCAI*, pages 3347–3355, 2023.

Balder ten Cate, Maurice Funk, Jean Christoph Jung, and Carsten Lutz. On the non-efficient PAC learnability of conjunctive queries. *Inf. Process. Lett.*, 183:106431, 2024.

William W. Cohen. PAC-learning nondeterminate clauses. *Proc. of AAAI*, pages 676–681, 1994.

Alin Deutsch, Alan Nash, and Jeffrey B. Remmel. The chase revisited. *Proc. of PODS*, pages 149–158, 2008.

Felix Distel. *Learning description logic knowledge bases from data using methods from formal concept analysis*. PhD thesis, Dresden University of Technology, 2011.

Paul Erdős. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

Tomás Feder and Moshe Y. Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory. *SIAM J. Comput.*, 28(1):57–104, 1998.

Maurice Funk. Concept-by-Example in $\mathcal{EL}$ knowledge bases. Master's thesis, University of Bremen, 2019.

Maurice Funk, Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, and Frank Wolter. Learning description logic concepts: When can positive and negative examples be separated? *Proc. of IJCAI*, pages 1682–1688, 2019.

Georg Gottlob, Nicola Leone, and Francesco Scarcello. On the complexity of some inductive logic programming problems. *Proc. of ILP*, 1297:17–32, 1997.

Ricardo Guimarães, Ana Ozaki, Cosimo Persia, and Baris Sertkaya. Mining $\mathscr{E}\mathscr{L}\bot$ bases with adaptable role depth. *J. Artif. Intell. Res.*, 76:883–924, 2023.

David Harel, Orna Kupferman, and Moshe Y. Vardi. On the complexity of verifying concurrent transition systems. *Inf. Comput.*, 173(2):143–161, 2002.

Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, and Frank Wolter. Logical separability of incomplete data under ontologies. *Proc. of KR*, pages 517–528, 2020.

Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, and Frank Wolter. Separating data examples by description logic concepts with restricted signatures. *Proc. of KR*, pages 390–399, 2021.

Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, and Frank Wolter. Logical separability of labeled data examples under ontologies. *Artif. Intell.*, 313:103785, 2022.

Jörg-Uwe Kietz and Sašo Džeroski. Inductive logic programming and learnability. *SIGART Bull.*, 5(1):22–32, 1994.

Francesco Kriegel. Efficient axiomatization of OWL 2 EL ontologies from data by means of formal concept analysis. *Proc. of AAAI*, 38(9):10597–10606, 2024.

Sergei O. Kuznetsov. On the intractability of computing the Duquenne-Guigues base. *JUCS - Journal of Universal Computer Science*, 10(8):927–933, 2004.

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

Hao Li, Chee-Yong Chan, and David Maier. Query from examples: An iterative, data-driven approach to query construction. *Proc. VLDB Endow.*, 8(13):2158–2169, 2015.

Stephen H. Muggleton and Cao Feng. Efficient induction of logic programs. *Proc. of ALT*, pages 368–381, 1990.

Alina Petrova, Egor V. Kostylev, Bernardo Cuenca Grau, and Ian Horrocks. Query-based entity comparison in knowledge graphs revisited. *Proc. of ISWC*, 11778:558–575, 2019.

Alina Petrova, Evgeny Sherkhonov, Bernardo Cuenca Grau, and Ian Horrocks. Entity comparison in RDF graphs. *Proc. of ISWC*, 10587:526–541, 2017.

Luc De Raedt, Andrea Passerini, and Stefano Teso. Learning constraints from examples. *Proc. of AAAI*, 32(1):7965–7970, 2018.

Sebastian Rudolph. *Relational exploration: combining description logics and formal concept analysis for knowledge specification*. PhD thesis, Dresden University of Technology, Germany, 2006.

Ross Willard. Testing expressibility is hard. *Proc. of CP*, 6308:9–23, 2010.

A Some Basic Lemmas

Lemma 10. *Let (I, d) and (J, e) be pointed instances and $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. Then $(I, d) \preceq_{\mathcal{L}} (J, e)$ implies that for every $\mathcal{L}_\bot$ concept C, $d \in C^I$ implies $e \in C^J$.*

Lemma 11. *Let S be a non-empty finite set of k-ary pointed instances with $(I, \bar{a}) \in S$ and let $(J, \bar{b}) = \prod S$. Then $(J, \bar{b}) \to (I, \bar{a})$.*

Lemma 12. *Let S be a non-empty finite set of k-ary pointed instances and $(J, \bar{b}) = \prod S$. Then for every k-ary CQ $q(\bar{x})$:*

$$\bar{b} \in q(J) \text{ if and only if } \bar{a} \in q(I) \text{ for all } (I, \bar{a}) \in S.$$

Lemma 13. *Let $(I_1, a_1), \ldots, (I_n, a_n)$ be pointed instances, $(P, \bar{a}) = \prod_{1 \leq i \leq n} (I_i, a_i)$, and C an $\mathcal{ELI}_\bot$-concept. Then $\bar{a} \in C^P$ if and only if $a_i \in C^{I_i}$ for $1 \leq i \leq n$.*

B Proofs for Section 2

Lemma 2. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. There is a polynomial time algorithm that, given finite pointed instances (I, d) and (J, e), decides whether $(I, d) \preceq_{\mathcal{L}} (J, e)$ and, if this is not the case, outputs the succinct representation of an $\mathcal{L}$-concept C of role depth at most $|\Delta^I| \cdot |\Delta^J|$ and outdegree at most $|\Delta^J|$ such that $d \in C^I$ and $e \notin C^J$.*

Proof. Let I and J be finite interpretations. We

1. start with the relation Z_0 that consists of all pairs $(d, e) \in \Delta^I \times \Delta^J$ such that for all concept names A, $d \in A^I$ implies $e \in A^J$ and then
2. construct a sequence of relations $Z_0, Z_1, \ldots$ where we obtain Z_{i+1} from Z_i by starting with $Z_{i+1} = Z_i$ and then deleting as follows: if $(d, e) \in Z_i$ and $(d, d') \in R^I$, with R an $\mathcal{L}$-role, and there is no $(e, e') \in R^J$ with $(d', e') \in Z_i$, then remove (d, e) from Z_{i+1}.

The process stabilizes after $m \leq |\Delta^I| \cdot |\Delta^J|$ rounds because that is the maximum number of pairs in Z_0. Each round clearly also only needs polynomial time. It is well known and easy to prove that the resulting relation Z_m is the maximal $\mathcal{L}$-simulation (w.r.t. set inclusion) from I to J. Thus $(I, d) \preceq_{\mathcal{L}} (J, e)$ iff $(d, e) \in Z_m$.

The procedure also allows us to identify for every pair $(d, e) \in (\Delta^I \times \Delta^J) \setminus Z_m$ a concept $C_{d,e}$ with $d \in C^I_{d,e}$ and $e \notin C^J_{d,e}$:

- If $(d, e) \notin Z_0$, then we may choose as $C_{d,e}$ any concept name A with $d \in A^I$ and $e \notin A^J$.
- If a pair (d, e) is deleted in Step 2 because of some $(d, d') \in R^I$, then we set $C_{d,e}$ to $\exists R.D$ with D the conjunction of $C_{d',e'}$, for all $(e, e') \in R^J$ (the empty conjunction being $\top$).

Let us analyze the size of the constructed concepts $C_{d,e}$. An easy analysis reveals that the role depth of each concept $C_{d,e}$ is bounded by $|\Delta^I| \cdot |\Delta^J|$, the maximum number of rounds of the algorithm, and that the outdegree is bounded by $|\Delta^J|$. Consequently, $C_{d,e}$ is of size at most single exponential. Let us now consider succinct representation. Clearly, every concept $C_{d,e}$ constructed by the algorithm is a concept name or of the form $\exists R.D$ with D a conjunction of concepts $C_{d',e'}$. Moreover, there are at most $|\Delta^I| \cdot |\Delta^J|$ concepts. It follows that the number of subconcepts of each concept $C_{d,e}$ is $O((|\Delta^I| \cdot |\Delta^J|)^2)$, and thus the succinct representation of $C_{d,e}$ is of polynomial size. Moreover, it is easy to see that the succinct representations of the concepts $C_{d,e}$ can be constructed in polynomial time, in parallel to executing the elimination procedure. ❏

Lemma 5. *Let (P, N) be a fitting instance and let $\mathcal{O}_\mathsf{P}$ be a finite $\mathcal{L}$-basis of P. Then $\mathcal{O}_\mathsf{P}$ fits (P, N) if and only if (P, N) has a fitting $\mathcal{L}$-ontology.*

Proof. Since "$\Rightarrow$" is trivial, we consider "$\Leftarrow$". Assume that $\mathcal{O}_\mathsf{P}$ does not fit (P, N) and take any $\mathcal{L}$-ontology $\mathcal{O}$. We have $I \models \mathcal{O}_\mathsf{P}$ for all $I \in \mathsf{P}$ and hence there must be a $J \in \mathsf{N}$ such that $J \models \mathcal{O}_\mathsf{P}$. We have to show that $\mathcal{O}$ does not fit (P, N). If $I \not\models \mathcal{O}$ for some $I \in \mathsf{P}$, then we are done. Otherwise every TGD in $\mathcal{O}$ is satisfied in all instances in P and thus $\mathcal{O}_\mathsf{P} \models \mathcal{O}$ by definition of finite bases. But then $J \models \mathcal{O}$ and thus $\mathcal{O}$ does not fit (P, N). ❏

C Proofs for Section 3

We want to show that the ontology $\mathcal{O}_I$ constructed in the main part of the paper is a finite $\mathcal{L}$-basis of I. A central step is to prove the following lemma, where $\mathcal{O} \models C \equiv D$ is an abbreviation for $\mathcal{O} \models C \sqsubseteq D$ and $\mathcal{O} \models D \sqsubseteq C$.

Lemma 14. *$\mathcal{O}_I \models C \equiv E^*_{C^I}$ for all $\mathcal{L}$-concepts C.*

Proof. The proof is by induction on the structure of C. We show both directions simultaneously. In the induction start, C is either a concept name, $\top$ or $\bot$, if $\bot$ is an $\mathcal{L}$-concept. First assume that $C = A$ is a concept name. The semantics yields $I \models A \equiv A^I$ and thus $\mathcal{O}_I \models A \equiv E^*_{A^I}$ by Points 1 and 3 of the construction of $\mathcal{O}_I$. For $\top$ and $\bot$, $\mathcal{O}_I \models E^*_{\top^I} \sqsubseteq \top$ and $\mathcal{O}_I \models \bot \sqsubseteq E^*_{\bot^I}$ hold trivially by the semantics. Regarding the other directions, Point 6 yields the desired result for $\top$ and Point 7 for $\bot$.

For the induction step, we begin with $C = D_1 \sqcap D_2$. The semantics yields $I \models D_1^I \sqcap D_2^I \sqsubseteq C^I$ and $I \models C^I \sqcap C^I \sqsubseteq D_i^I$ for $i \in \{1, 2\}$, and thus from Point 5 of the construction of $\mathcal{O}_I$ we obtain

$$\mathcal{O}_I \models E^*_{D_1^I} \sqcap E^*_{D_2^I} \sqsubseteq E^*_{C^I} \text{ and}$$
$$\mathcal{O}_I \models E^*_{C^I} \sqsubseteq E^*_{D_i^I} \text{ for } i \in \{1, 2\}.$$

Therefore

$$\mathcal{O}_I \models E^*_{C^I} \equiv E^*_{D_1^I} \sqcap E^*_{D_2^I}.$$

Applying the induction hypothesis yields $\mathcal{O}_I \models D_i \equiv E^*_{D_i^I}$ for $i \in \{1, 2\}$ and thus we may conclude from the above that

$$\mathcal{O}_I \models D_1 \sqcap D_2 \equiv E^*_{C^I}.$$

Now consider $C = \exists R.D$ with R an $\mathcal{L}$-role. By the semantics, $I \models \exists R.D^I \equiv (\exists R.D)^I$. By Points 2 and 4 of the construction of $\mathcal{O}_I$, we obtain

$$\mathcal{O}_I \models \exists R.E^*_{D^I} \equiv E^*_{(\exists R.D)^I}.$$

The induction hypothesis yields $\mathcal{O}_I \models D \equiv E^*_{D^I}$, which, combined with the above gives $\mathcal{O}_I \models \exists R.D \equiv E^*_{(\exists R.D)^I}$, as required. ❏

Based on Lemma 14, the following is easy to show.

Lemma 6. *$\mathcal{O}_I$ is a finite $\mathcal{L}$-basis of I.*

Proof. We have to show that $\mathcal{O}_I \models C \sqsubseteq D$ iff $I \models C \sqsubseteq D$, for all $\mathcal{L}$-concept inclusions $C \sqsubseteq D$.

"$\Rightarrow$". It suffices to observe that $I \models \mathcal{O}_I$, which is immediate from the construction of $\mathcal{O}_I$.

"$\Leftarrow$". Assume that $I \models C \sqsubseteq D$. Then $I \models E^*_{C^I} \sqsubseteq E^*_{D^I}$ by definition of $E^*_{C^I}$ and $E^*_{D^I}$. By Point 5 of the construction of $\mathcal{O}_I$, this implies $E^*_{C^I} \sqcap E^*_{C^I} \sqsubseteq E^*_{D^I} \in \mathcal{O}_I$, which of course gives $\mathcal{O}_I \models E^*_{C^I} \sqsubseteq E^*_{D^I}$. Lemma 14 yields $\mathcal{O}_I \models C \sqsubseteq D$, as required. ❑

Lemma 7. *There is an algorithm that, given as input an instance I and a set $X \subseteq \Delta^I$, in double exponential time outputs an $\mathcal{L}$-concept C with $C^I = X$ if such a concept exists, and reports 'undefinable' otherwise. The algorithm can be modified to run in single exponential time and output a succinct representation of C.*

Proof. Given an instance I and a set $X \subseteq \Delta^I$, the algorithm claimed to exist by Lemma 7 returns $\bot$ if $X = \emptyset$ and $\bot$ is an $\mathcal{L}$-concept. If this special case does not apply, then it first constructs $(P, \bar{d}) = \prod_{d \in X}(I, d)$ (in single exponential time) and then checks that $(P, \bar{d}) \not\preceq_{\mathcal{L}} (I, e)$ for every $e \in \Delta^I \setminus X$, using the algorithm from Lemma 2. If the check fails for some e, it outputs 'undefinable'. Otherwise, for each e it identifies an $\mathcal{L}$-concept $C_{\bar{d},e}$ with $X \subseteq C^I_{\bar{d},e}$ and $e \notin C^I_{\bar{d},e}$ and returns the conjunction of all these concepts. This is clearly an $\mathcal{L}$-concept that defines X. The desired concepts $C_{\bar{d},e}$ are in fact the concepts obtained from Lemma 2. They satisfy $\bar{d} \in C^P_{\bar{d},e}$ and $e \notin C^I_{\bar{d},e}$, and by Lemma 13 the former implies $X \subseteq C^I_{\bar{d},e}$, as required. ❑

D Proofs for Section 4

We first define characteristic concepts of bounded depth. Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$ and let (I, d) be a finite pointed instance. For every $i \in \mathbb{N}$, the concept $C^I_{\mathcal{L},i}(d)$ is defined by:

- $C^I_{\mathcal{L},0}(d) = \top \sqcap \bigsqcap \{A \mid d \in A^I\}$
- $C^I_{\mathcal{L},i+1}(d) = C^I_{\mathcal{L},0}(d) \sqcap \bigsqcap_{\substack{R \mathcal{L}\text{-role} \\ (d,e) \in R^I}} \exists R.C^I_{\mathcal{L},i}(e)$.

Note that the size of $C^I_{\mathcal{L},i}(d)$ is $O(|I| \cdot |\Delta^I|^i)$ and the number of subconcepts in $C^I_{\mathcal{L},i}(d)$ is $O(|I| \cdot |\Delta^I|^2 \cdot i)$.

Lemma 15. *Let (I, d) and (J, e) be finite pointed instances. If $e \in C^I_{\mathcal{L},n}(d)^J$, with $n = |\Delta^I| \cdot |\Delta^J|$, then $(I, d) \preceq_{\mathcal{L}} (J, e)$.*

Proof. Towards proving the contrapositive, assume that $(I, d) \not\preceq_{\mathcal{L}} (J, e)$. Then Lemma 2 yields an $\mathcal{L}$-concept C of role depth at most n such that $d \in C^I$ and $e \notin C^J$. However, the former clearly implies $\emptyset \models C^I_{\mathcal{L},n}(d) \sqsubseteq C$ by definition of characteristic concepts, and thus $e \notin C^J$ implies $e \notin C^I_{\mathcal{L},n}(d)^J$, which is what we have to show. ❑

Theorem 3. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$. Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \dots, N_k\}$ and let $P = \biguplus \mathsf{P}$. Then no $\mathcal{L}$-TGD fits (P, N) if and only if for every $\bar{d} = (d_1, \dots, d_k) \in \Delta^{\prod \mathsf{N}}$ such that d_i is non-$\mathcal{L}$-total in N_i for all $i \in [k]$, the following conditions are satisfied:*

1. *the set $S_{\bar{d}} = \{(P, e) \mid (\prod \mathsf{N}, \bar{d}) \preceq_{\mathcal{L}} (P, e)\}$ is non-empty;*
2. *$\prod S_{\bar{d}} \preceq_{\mathcal{L}} (N_i, d_i)$ for some $i \in [k]$.*

The same is true for $\mathcal{L}_\bot$ when the condition 'd_i non-$\mathcal{L}$-total in N_i for all $i \in [k]$' is dropped.

Proof. "$\Leftarrow$". Assume that for all $\bar{d} = (d_1, \dots, d_k) \in \Delta^{\prod \mathsf{N}}$ with d_i non-$\mathcal{L}$-total in N_i for all $i \in [k]$, Conditions 1 and 2 are satisfied. Further assume to the contrary of what we have to show that there exists an $\mathcal{L}$-concept inclusion $C \sqsubseteq D$ that fits (P, N). Then for every $i \in [k]$ there is a $d_i \in \Delta^{N_i}$ such that $d_i \in C^{N_i} \setminus D^{N_i}$. Consider the tuple $\bar{d} = (d_1, \dots, d_k)$. First note that, by Lemma 13, $\bar{d} \in C^{\prod \mathsf{N}}$. Also note that d_i is not $\mathcal{L}$-total in N_i for all $i \in [k]$ and therefore $\bar{d}$ must satisfy Conditions 1 and 2. By Condition 1, the product $(J, \bar{f}) = \prod S_{\bar{d}}$ is defined. Condition 2 yields $(J, \bar{f}) \preceq_{\mathcal{L}} (N_i, d_i)$ for some $i \in [k]$ and since $d_i \notin D^{N_i}$, it follows that $\bar{f} \notin D^J$. By Lemma 13 this implies $e \notin D^P$ for some $(P, e) \in S_{\bar{d}}$. But by definition of $S_{\bar{d}}$ we have $(\prod \mathsf{N}, \bar{d}) \preceq_{\mathcal{L}} (P, e)$ and thus from $\bar{d} \in C^{\prod \mathsf{N}}$ we obtain $e \in C^P$. It follows that $P \not\models C \sqsubseteq D$, and since $\mathcal{L}$ is invariant under disjoint union, this contradicts the assumption that $C \sqsubseteq D$ fits (P, N).

"$\Rightarrow$". We show the contrapositive. Assume that there is a non-$\mathcal{L}$-total $\bar{d} \in \Delta^{\prod \mathsf{N}}$ such that Conditions 1 and 2 do not both hold. Let $\prod S_{\bar{d}} = (J, \bar{f})$ and $n = |\Delta^{\prod \mathsf{N}}| \cdot |\Delta^P|$.

First assume that d satisfies Condition 1 (and thus $\prod S_{\bar{d}}$ is defined), but violates Condition 2. Consider the concept inclusion

$$C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}) \sqsubseteq C^J_{\mathcal{L},m}(\bar{f})$$

where $m = |\Delta^J| \cdot \max_{I \in \mathsf{N}} |\Delta^I|$. We show that every example in P satisfies this inclusion while all examples in N violate it. For the former, it suffices to show that the disjoint union P satisfies the inclusion. Let $e \in (C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}))^P$. Then Lemma 15 yields $(\prod \mathsf{N}, \bar{d}) \preceq_{\mathcal{L}} (P, e)$ and thus $(P, e) \in S_{\bar{d}}$ is one of the instances in the product $(J, \bar{f})$. By Lemma 13, this implies $e \in (C^J_{\mathcal{L},m}(\bar{f}))^P$, as required. To show that every $N \in \mathsf{N}$ violates $C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}) \sqsubseteq C^J_{\mathcal{L},m}(\bar{f})$, first note that $\bar{d} \in (C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}))^{\prod \mathsf{N}}$ and thus Lemma 13 yields $d_i \in (C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}))^{N_i}$ for all $i \in [k]$. Moreover, $(J, \bar{f}) \not\preceq_{\mathcal{L}} (N_i, d_i)$ for all $i \in [k]$ because Condition 2 is violated. From Lemma 15 we obtain $d_i \notin (C^J_{\mathcal{L},\ell}(\bar{f}))^{N_i}$ for $\ell = |\Delta^{N_i}| \cdot |\Delta^J|$. This clearly implies $d_i \notin (C^J_{\mathcal{L},m}(\bar{f}))^{N_i}$ since $m \geq \ell$, and thus we are done.

Now assume that Condition 1 is violated. Let (K, t) be the pointed interpretation with $\Delta^K = \{t\}$ and such that t satisfies all concept names in the schema $\mathcal{S}$ of (P, N) and there is a reflexive loop on t for every role name in $\mathcal{S}$. Then clearly t is $\mathcal{L}$-total. Consider the concept inclusion

$$C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}) \sqsubseteq C^K_{\mathcal{L},m}(t)$$

where $m = \max_{I \in \mathsf{N}} |\Delta^I|$. We again show that every example in P satisfies this inclusion while no $N \in \mathsf{N}$ does. Concerning the latter, it is again clear that $d_i \in (C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}))^{N_i}$ for

all $i \in [k]$ and thus it suffices to show that $d_i \notin (C^K_{\mathcal{L},m}(t))^{N_i}$. Assume to the contrary. Then by Lemma 15 $(\mathcal{K}, t) \preceq_{\mathcal{L}} (N_i, d_i)$. This implies that every concept satisfied by $(\mathcal{K}, t)$ is also satisfied by (N_i, d_i). As t is $\mathcal{L}$-total in K, d_i must be $\mathcal{L}$-total in N_i too, a contradiction. To show that every example in P satisfies $C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}) \sqsubseteq C^K_{\mathcal{L},m}(t)$, it suffices to show that the disjoint union P satisfies the inclusion, and in particular that $(C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}))^P = \emptyset$. This, however, follows from Lemma 15 and the fact that Condition 1 is violated.

The proof of the claim for $\mathcal{L}_\bot$ can be obtained from the proof for $\mathcal{L}$ above by minor changes. For the direction "$\Leftarrow$" and the case of violation of Condition 2 in the "$\Rightarrow$" direction, it suffices to lift the restriction to non-$\mathcal{L}$-total values in the assumptions, which has no effect on the arguments. When it comes to the violation of Condition 1 in the "$\Rightarrow$" direction, the proof becomes simpler. The inclusion $C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}) \sqsubseteq C^K_{\mathcal{L},m}(t)$ can then be replaced by $C^{\prod \mathsf{N}}_{\mathcal{L},n}(\bar{d}) \sqsubseteq \bot$. ❑

Theorem 4. *In $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, fitting TGD existence is decidable in* ExpTime *and fitting TGD construction is possible in double exponential time, and in exponential time if TGDs are represented succinctly. The same is true for fitting ontology existence and construction.*

Proof. Let $\mathcal{L}$ be any of the DLs mentioned in the theorem. To decide whether a given fitting instance (P, N) admits a fitting $\mathcal{L}$-TGD, we have to check whether there is a $\bar{d} = (d_1, \dots, d_k) \in \Delta^{\prod \mathsf{N}}$, with d_i non-$\mathcal{L}$-total in N_i for all $i \in [k]$ if $\mathcal{L}$ does not admit $\bot$, such that Conditions 1 and 2 of Theorem 3 are not both satisfied. Making use of the facts that the existence of a simulation can be decided in polynomial time and the involved products are of single exponential size and can be computed in output polynomial time, it is easy to implement the required checks to obtain an ExpTime upper bound.[2] The fitting TGDs obtained if the check fails, as in the proof of Theorem 3, are at most double exponential in size and at most single exponential if represented succinctly.

Regarding fitting ontology existence and construction, Lemma 3 yields a simple reduction to the TGD fitting case that gives the desired results. We remark that, unlike the fitting ontologies obtained from finite bases, the resulting ontologies contain at most one concept inclusion per instance in N. ❑

Theorem 5. *In $\mathcal{EL}$, $\mathcal{ELI}$, $\mathcal{EL}_\bot$, and $\mathcal{ELI}_\bot$, fitting TGD existence and fitting ontology existence are* ExpTime*-hard.*

Proof. Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$. The *product $\mathcal{EL}$-simulation problem* means to decide, given pointed instances $(I_1, a_1), \dots, (I_n, a_n), (J, b)$, whether $\prod_{i=1}^n (I_i, a_i) \preceq_{\mathcal{EL}} (J, b)$. This problem is known to be ExpTime-hard (Funk *et al.* 2019). We provide a polynomial time reduction from the product $\mathcal{EL}$-simulation problem to fitting $\mathcal{L}$-ontology existence and fitting $\mathcal{L}$-TGD existence. In fact, we shall use fitting instances with a single negative example for which, by Lemma 3, the existence of fitting ontologies and of fitting TGDs coincides.

Let the input to the product $\mathcal{EL}$-simulation problem consist of the pointed instances $(I_1, a_1), \dots, (I_n, a_n), (J, b)$ over some schema $\mathcal{S}$. We may assume without loss of generality that all these instances are pairwise disjoint.

We construct a fitting instance (P, N) over an extended schema

$$\mathcal{S}' = \mathcal{S} \cup \{A_c \mid c \in \Delta^J \cup \{u, v\}\} \cup \{R\}$$

where the A_c are fresh concept names and R is a fresh role name. We also introduce fresh values u, v. Set $\mathsf{N} = \{N\}$ where

$$\begin{aligned} N \;=\; & J \cup \{A_c(c) \mid c \in \Delta^J \cup \{u, v\}\} \cup \{R(u, b)\} \\ & \cup \{S(v, c) \mid c \in \Delta^J \cup \{u, v\} \text{ and } S \in \mathcal{S}'\} \\ & \cup \{A(v) \mid A \in \mathcal{S}'\}. \end{aligned}$$

To define P, let $N_1, \dots, N_n$ be pairwise disjoint copies of N. For every $d \in \Delta^N$ we use $\langle d, i\rangle$ to denote the copy of d in N_i. Moreover, set $\langle d, i\rangle^\downarrow = d$. Now define $\mathsf{P} = \{P\}$ with P the disjoint union of the instances $I_1, \dots, I_n$ and $N_1, \dots, N_n$, extended with the facts $R(\langle u, i\rangle, a_i)$, $i \in [n]$.

We show that $\prod_{i=1}^n (I_i, a_i) \preceq_{\mathcal{EL}} (J, b)$ if and only if (P, N) has no fitting $\mathcal{L}$-TGD, making use of the characterization provided by Theorem 3. Note that because we only have a single negative example N, the product $\prod \mathsf{N}$ in Theorem 3 is simply N, and the second condition simplifies to $\prod S_{\bar{d}} \preceq_{\mathcal{L}} (N, d)$. Take any $d \in \Delta^N$ and consider the set $S_d = \{(P, e) \mid (N, d) \preceq_{\mathcal{EL}} (P, e)\}$. We must have

$$\begin{aligned} S_d \;=\; & \{(P, \langle d, 1\rangle), \dots, (P, \langle d, n\rangle)\} \cup \\ & \{(P, \langle v, 1\rangle), \dots, (P, \langle v, n\rangle)\}. \end{aligned}$$

In fact, the "$\supseteq$" direction is immediate and the "$\subseteq$" follows from the use of the fresh concept names A_d. In particular, S_d is non-empty and thus Condition 1 of Theorem 3 is satisfied for d.

As a consequence of this initial consideration and of Theorem 3, it suffices to show the following.

Claim 1. $\prod_{i=1}^n (I_i, a_i) \preceq_{\mathcal{EL}} (J, b)$ if and only if $\prod S_d \preceq_{\mathcal{L}} (N, d)$ for every $d \in \Delta^N$.

This should be clear in the case $\mathcal{L} \in \{\mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$ where Conditions 1 and 2 of Theorem 3 have to be satisfied for all $d \in \Delta^N$, not just for the non-$\mathcal{L}$-total ones. Note that, due to the use of the fresh concept names A_c, none of the values $d \in \Delta^N \setminus \{v\}$ is $\mathcal{L}$-total while v is $\mathcal{L}$-total. But showing the claim also suffices in the case $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}\}$: for all $\mathcal{L}$-total d in N, we trivially have $\prod S_d \preceq_{\mathcal{L}} (N, d)$ if S_d is non-empty, which we have argued to be the case.

To simplify matters, we consider the set

$$T_d = \{(P, \langle d, 1\rangle), \dots, (P, \langle d, n\rangle)\}$$

and observe the following.

Claim 2. $\prod S_d \preceq_{\mathcal{L}} (N, d)$ if and only if $\prod T_d \preceq_{\mathcal{L}} (N, d)$ for every $d \in \Delta^N$.

[2] Note that d_i is $\mathcal{L}$-total in N_i iff $(I_\bot, d) \preceq_{\mathcal{EL}} (N_i, d_i)$ where $I_\bot$ is the instance that contains $A(d)$ for every concept name A in the schema if (P, N) and $R(d, d)$ for every role name R in this schema. As a consequence, totality can be decided in polynomial time.

To prove Claim 2, it suffices to observe that from any $\mathcal{L}$-simulation Z that witnesses $\prod S_d \preceq_{\mathcal{L}} (N, d)$, we can obtain an $\mathcal{L}$-simulation Z' that witnesses $\prod T_d \preceq_{\mathcal{L}} (N, d)$ by using

$$\{((b_1, \ldots, b_n), c) \mid ((b_1, \ldots, b_n, \langle v, 1\rangle, \ldots, \langle v, n\rangle), c) \in Z\}.$$

This is indeed an $\mathcal{L}$-simulation because Z is and each $\langle v, i\rangle$ satisfies all concept names and has a reflexive loop for all role names. Conversely, from any $\mathcal{L}$-simulation Z that witnesses $\prod T_d \preceq_{\mathcal{L}} (N, d)$, we can obtain an $\mathcal{L}$-simulation Z' that witnesses $\prod S_d \preceq_{\mathcal{L}} (N, d)$ by using

$$\{((b_1, \ldots, b_n, c_1, \ldots, c_n), c) \mid ((b_1, \ldots, b_n), c) \in Z \text{ and } c_1, \ldots, c_n \in \Delta^P\}.$$

The virtue of Claim 2 is that it suffices to prove Claim 1 with $\prod T_d$ in place of $\prod S_d$. This is what we do in what follows. As a preparation, note that for distinct $d, d' \in \Delta^N$, the products $\prod T_d$ and $\prod T_{d'}$ are based on the same instance, which is the n-fold product of P with itself, denoted P^n, and differ only in the distinguished tuple, which is $(\langle d, 1\rangle, \ldots, \langle d, n\rangle)$ and $(\langle d', 1\rangle, \ldots, \langle d', n\rangle)$, respectively. We continue to show Claim 1 with $\prod T_d$ in place of $\prod S_d$.

"$\Rightarrow$". Let $\prod_{i=1}^n (I_i, a_i) \preceq_{\mathcal{EL}} (J, b)$ and let Z be a witnessing $\mathcal{EL}$-simulation. We call a value $(d_1, \ldots, d_n) \in \Delta^{P^n}$ *well-sliced* if $d_i \in \Delta^{I_i} \cup \Delta^{N_i}$ for $1 \leq i \leq n$. Extend Z to a relation Z', as follows:

1. add, for every well-sliced $(d_1, \ldots, d_n) \in \Delta^{P^n}$ that contains at least one copy of a value in N, the pair
$$((d_1, \ldots, d_n), d_j^{\downarrow})$$
where d_j is the left-most value in $(d_1, \ldots, d_n)$ that is a copy of a value in N (this choice is arbitrary);
2. add, for every $(d_1, \ldots, d_n) \in \Delta^{P^n}$, the pair
$$((d_1, \ldots, d_n), v).$$

We show that Z' is an $\mathcal{ELI}$-simulation from $\prod T_d$ to (N, d), for every $d \in \Delta^N$. This suffices also when $\mathcal{L} \in \{\mathcal{EL}, \mathcal{EL}_\bot\}$ since every $\mathcal{ELI}$-simulation is also an $\mathcal{EL}$-simulation. By Point 1 of the extension, Z' contains the pair $((\langle d, 1\rangle, \ldots, \langle d, n\rangle), d)$ for every $d \in \Delta^N$. It thus remains to show that Z' is an $\mathcal{ELI}$-simulation from P^n to N.

Let $(\bar{e}, e) \in Z \subseteq Z'$. We have to show that the two conditions of $\mathcal{ELI}$-simulations are satisfied. Condition 1 is satisfied since Z is an $\mathcal{EL}$-simulation. For Condition 2, let $(\bar{e}, e) \in Z$ and $(\bar{e}, \bar{f}) \in S^{P^n}$. If S is a role name, then it suffices to observe that Z satisfies Condition 2 of $\mathcal{EL}$-simulations. If S is an inverse role, it suffices to observe that $(e, v) \in S^N$ and $(\bar{f}, v) \in Z'$ by Point 2 of the extension.

Now assume that $(\bar{e}, e) \in Z' \setminus Z$ was added in Point 1 of the extension, with $\bar{e} = (e_1, \ldots, e_n)$. Then there is some e_j that is a copy of a value of N and for the leftmost such e_j we have $e_j^{\downarrow} = e$. The two conditions of $\mathcal{ELI}$-simulations are again satisfied:

1. Assume that $\bar{e} \in A^{P^n}$ for some concept name A. Then $e \in A^N$, by definition of products and because $\bar{e} \in A^{P^n}$ and e_j is a copy of e.
2. Assume that $(\bar{e}, \bar{f}) \in S^{P^n}$, for some $\mathcal{ELI}$-role S. Let $\bar{f} = (f_1, \ldots, f_n)$. We distinguish two cases.

   Case 1. $S \notin \{R, R^-\}$. Then $(\bar{e}, \bar{f}) \in S^{P^n}$ implies that f_ℓ is a copy of a value in N if and only if e_ℓ is a copy of a value in N, for $1 \leq \ell \leq n$. In particular, this means that f_j is the left-most value in $\bar{f}$ that is a copy of a value of N. Point 1 of the definition of Z' yields $(\bar{f}, f_j) \in Z'$. By definition of products, $(\bar{e}, \bar{f}) \in S^{P^n}$ implies that $(e_j^{\downarrow}, f_j^{\downarrow}) \in S^N$ and we are done.

   Case 2. $S = R$. First assume that $\bar{f}$ does not contain a copy of a value in N. Since $\bar{e}$ is well-sliced and $S = R$, we must then have $e_i = \langle u, i\rangle$ and $f_i = a_i$ for $1 \leq i \leq n$. Note that $(u, b) \in R^N$. Moreover, $Z \subseteq Z'$ contains the pair $(\bar{f}, b)$ and thus we are done.

   Now assume that $\bar{f}$ contains a copy of a value in N. Let f_k be the left-most such copy. It follows together with $(\bar{e}, \bar{f}) \in R^{P^n}$ that e_k is also a copy of a value in N and hence by definition of product we have $(e_k^{\downarrow}, f_k^{\downarrow}) \in R^N$. Moreover, Point 1 of the definition of Z' yields $(\bar{f}, f_k^{\downarrow}) \in Z'$ as required.

   Case 3. $S = R^-$. Recall $e_j^{\downarrow} = e$. By construction we have $(e_j^{\downarrow}, v) \in R^{-N}$. From Point 2 of the definition of Z' we obtain $(\bar{f}, v) \in Z'$ and thus we are done.

Finally assume that $(\bar{e}, e) \in Z' \setminus Z$ was added in Point 2 of the extension. Then $e = v$. Recall that v satisfies all concept names from $\mathcal{S}'$ and has a reflexive loop for every $\mathcal{S}'$-role name. Together with Point 2 of the definition of Z', this clearly implies that Conditions 1 and 2 of $\mathcal{ELI}$-simulations are satisfied.

"$\Leftarrow$". Assume that $\prod T_d \preceq_{\mathcal{L}} (N, d)$ for every $d \in \Delta^N$. Then in particular $\prod T_u \preceq_{\mathcal{L}} (N, u)$. Let Z be a witnessing simulation. Then specifically $(\bar{u}, u) \in Z$ where $\bar{u} = (\langle u, 1\rangle, \ldots, \langle u, n\rangle)$. Define $Z' = Z \cap (\Delta^{\prod_{i=1}^n I_i} \times \Delta^J)$. We show that Z' is an $\mathcal{EL}$-simulation that witnesses $\prod_{i=1}^n (I_i, a_i) \preceq_{\mathcal{EL}} (J, b)$.

Note that $R(\bar{u}, \bar{a}) \in P^n$ where $\bar{a} = (a_1, \ldots, a_n)$. Since b is the only R-successor of u in N, it follows that $(\bar{a}, b) \in Z$. Hence, $(\bar{a}, b) \in Z'$. It remains to show that Z' is an $\mathcal{EL}$-simulation. This, however, follows from the fact that Z is an $\mathcal{EL}$-simulation and that all values $\Delta^N \setminus \Delta^J$ that are reachable in N from some value in Δ^J are reachable only via an inverse role, but not via a 'forward' role. ❏

For the subsequent proof, we first recall a result from (Funk 2019).

Theorem 23. *Let $n \geq 1$. There are pointed instances $(I_1, d_1), \ldots, (I_n, d_n), (J, e)$ such that*

1. *the size of the instances is bounded by $p(n)$, p a polynomial;*
2. *$\prod_i (I_i, d_i) \not\preceq_{\mathcal{EL}} (J, e)$;*
3. *the smallest $\mathcal{EL}$-concept C such that $d_i \in C^{I_i}$ for all $i \in \{1, \ldots, n\}$ but $e \notin C^J$ has size 2^{2^n} and is of role depth 2^n.*[3]

[3] Reference (Funk 2019) does not explicitly mention a lower

We use Theorem 23 to prove the following result.

Theorem 6. *Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$. For every $n \geq 1$, there is a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ such that*

1. *the size of $(\mathsf{P}_n, \mathsf{N}_n)$ is bounded by $p(n)$, p a polynomial;*
2. *$(\mathsf{P}_n, \mathsf{N}_n)$ admits a fitting $\mathcal{L}$-TGD and a fitting $\mathcal{L}$-ontology;*
3. *the smallest $\mathcal{L}$-TGD that fits $(\mathsf{P}_n, \mathsf{N}_n)$ is of size at least 2^{2^n} and of size 2^n when represented succinctly, and the same is true for the smallest fitting $\mathcal{L}$-ontology.*

Proof. Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{ELI}, \mathcal{EL}_\bot, \mathcal{ELI}_\bot\}$ and $n \geq 1$. Let $(I_1, d_1), \ldots, (I_n, d_n), (J, e)$ be the instances from Theorem 23, and let their schema be $\mathcal{S}$. Without loss of generality, we can assume that the domains of these instances are disjoint. By Condition 2 of Theorem 23 and Lemma 10, there is an $\mathcal{EL}$-concept C_0 such that $d_i \in C_0^{I_i}$ for all $i \in \{1, \ldots, n\}$ but $e \notin C_0^J$.

We construct sets of instances $\mathsf{P}_n = \{I_1', \ldots, I_n'\}$ and $\mathsf{N}_n = \{J'\}$ over an extended schema

$$\mathcal{S}' = \mathcal{S} \cup \{A, S\}$$

where A is a fresh concept name and S a fresh role name, as follows:

- J' is obtained from J by adding two fresh values e', f as well as
  - assertions $A(e')$ and $S(e', e)$ and
  - assertions $R(f, d), B(f)$ for all role names $R \in \mathcal{S}'$, concept names $B \in \mathcal{S}'$, and values $d \in \mathsf{adom}(J) \cup \{f\}$.
- I_i' is obtained from the union of I_i and J' by adding a fresh value d_i' and the statements $S(d_i', d_i)$, $S(d_i', e)$, and $A(d_i')$.

Clearly, the sizes of $I_1', \ldots, I_n', J'$ are bounded by some fixed polynomial because the sizes of $I_1, \ldots, I_n, J$ are. It remains to verify Conditions 2 and 3.

For Condition 2, observe that the TGD

$$A \sqsubseteq \exists S.C_0$$

is satisfied in every I_i' but not in J'. Thus, this TGD fits $(\mathsf{P}_n, \mathsf{N}_n)$, and so does the ontology that contains exactly this TGD.

Before verifying Condition 3, we observe two properties of the construction.

Claim. For every $d \in \mathsf{adom}(J) \cup \{f\}$ and $i \in \{1, \ldots, n\}$, we have (i) $(J', d) \preceq_{\mathcal{ELI}} (I_i', d)$ and (ii) $(I_i', d) \preceq_{\mathcal{ELI}} (J', d)$.

Proof of the claim. Point (i) is clear since $J' \subseteq I_i$, for every i, and hence the identity is a witnessing $\mathcal{ELI}$-simulation. For Point (ii), observe that the relation S_i defined by

$$S_i = \{(d, d) \mid d \in \mathsf{adom}(J')\} \cup \{(d, f) \mid d \in \mathsf{adom}(I_i) \setminus \mathsf{adom}(J')\}$$

is the witnessing $\mathcal{ELI}$-simulation, for every i. This finishes the proof of the claim.

To show Condition 3, let $X \sqsubseteq Y$ be a smallest TGD that fits $(\mathsf{P}_n, \mathsf{N}_n)$. Then $I_i' \models X \sqsubseteq Y$ for all I_i', but $J' \not\models X \sqsubseteq Y$. By the latter, we know that there is some $d \in \Delta^{J'}$ such that $d \in X^{J'}$ but $d \notin Y^{J'}$. Suppose first that $d \in \mathsf{adom}(J) \cup \{f\}$. Then the Claim together with Lemma 10 implies that d satisfies the same $\mathcal{ELI}_\bot$-concepts in J' and I_i', contradicting $I_i' \models X \sqsubseteq Y$. Thus, $d = e'$ is the fresh value introduced in the construction of J'. Observe that by construction $(J', e') \preceq_{\mathcal{ELI}} (I_i', d_i')$ for every $i \in \{1, \ldots, n\}$. Since $e' \in X^{J'}$, Lemma 10 implies $d_i' \in X^{I_i'}$, for all i. Since all I_i' satisfy $X \sqsubseteq Y$, also $d_i' \in Y^{I_i'}$.

By the structure of I_i' and J', Y is of shape $\exists S.Z$. Note that $e' \notin Y^{J'}$ implies $e \notin Z^{J'}$. By Point (ii) of the claim and Lemma 10, it follows that $e \notin Z^{I_i'}$, for every i. Hence, $d_i \in Z^{I_i}$ for all i. If Z is an $\mathcal{EL}$-concept, then by the choice of $I_1, \ldots, I_n, J$ and Condition 3 of Theorem 23, this Z has to have size at least 2^{2^n} and depth 2^n. Thus, the TGD $X \sqsubseteq Y$ has size at least 2^{2^n} and 2^n when represented succinctly.

It remains to consider the case when Z is an $\mathcal{ELI}$-concept. Let $U = \exists R^-.Z'$ be a subconcept of Z and let $\bar{Z}$ be the concept obtained from Z by replacing $\exists R^-.Z'$ with $\top$. Notice that we still have $e \notin \bar{Z}^{J'}$ since every value in $\mathsf{adom}(J)$ satisfies U in J', due to the presence of f. Moreover, $d_i \in \bar{Z}^{I_i'}$ for all i and hence $X \sqsubseteq \exists S.\bar{Z}$ is a smaller fitting TGD than $X \sqsubseteq Y$, in contradiction to $X \sqsubseteq Y$ being of smallest possible size.

The argument for smallest fitting ontologies $\mathcal{O}$ is almost identical. We start with choosing from $\mathcal{O}$ some TGD $X \sqsubseteq Y \in \mathcal{O}$ such that $I_i' \models X \sqsubseteq Y$ for all I_i', but $J' \not\models X \sqsubseteq Y$ and then continue as above, in the end obtaining a smaller fitting ontology $\{X \sqsubseteq \exists S.\bar{Z}\}$. ❏

bound on the role depth, but it is not hard to derive it using the analysis given there.

E Proofs for Section 5

Lemma 8. *$\mathcal{O}_\mathsf{H}$ is a finite GTGD-basis of H.*

Proof. We have to show that for every guarded TGD ρ:

$$\mathcal{O}_\mathsf{H} \models \rho \text{ if and only if } I \models \rho \text{ for all } I \in \mathsf{H}.$$

"⇒". It suffices to show that if $\rho \in \mathcal{O}_\mathsf{H}$, then $I \models \rho$ for all $I \in \mathsf{H}$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. For brevity, let $q(\bar{x}) = \exists \bar{y}\, \varphi$ and $p(\bar{x}) = \exists \bar{z}\, \psi$. Consider any $I \in \mathsf{H}$. If $q(I) = \emptyset$, then $I \models \rho$ holds trivially. Otherwise take any $\bar{a} \in q(I)$. We have to show that $\bar{a} \in p(I)$. By construction of the rules in $\mathcal{O}_\mathsf{H}$, the right-hand side p of ρ is obtained from the CQ $q_{(P^*, \bar{b}^*)}$ by renaming the free variables, where

$$(P, \bar{b}) = \prod S \quad \text{and} \quad S = \{(J, \bar{c}) \mid J \in \mathsf{H}, \bar{c} \in q(J)\}.$$

Clearly, $(I, \bar{a})$ is in S and thus by Lemma 11 there exists a homomorphism h from $(P, \bar{b})$ to $(I, \bar{a})$. We can easily extend h to a homomorphism h' from $(P^*, \bar{b}^*)$ to $(I, \bar{a})$: if value a^* is a clone of value a, then set $h'(a^*) = h(a)$. It is now simple to convert h' into a homomorphism g from p to I with $g(\bar{x}) = \bar{a}$ by renaming the objects in the domain of h'. Thus $\bar{a} \in p(I)$ as required.

"⇐". Assume that $I \models \rho$ for all $I \in \mathsf{H}$. Moreover, let I be some instance with $I \models \mathcal{O}_\mathsf{H}$. We have to show that $I \models \rho$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$ and, for brevity, let $q(\bar{x}) = \exists \bar{y}\, \varphi$ and $p(\bar{x}) = \exists \bar{z}\, \psi$. If $q(I) = \emptyset$, then $I \models \rho$

holds trivially. Thus assume that $q(I) \neq \emptyset$ and take any $\bar{a} \in q(I)$. We want to show that $\bar{a} \in p(I)$. Consider two cases.

First, let $q(J) = \emptyset$ for all $J \in \mathsf{H}$. Then $\mathcal{O}_\mathsf{H}$ includes the rule $\varphi \to \psi'$ with ψ' the conjunction of all atoms over $\mathcal{S}$ that use only the variables from $\bar{x}$. Since $I \models \mathcal{O}_\mathsf{H}$ and $\bar{a} \in q(I)$, it follows that I contains all facts over $\mathcal{S}$ that use only values from $\bar{a}$. But then we find a homomorphism h from p to I with $h(\bar{x}) = \bar{a}$, no matter what p is. Thus $\bar{a} \in p(I)$, as required.

Second, assume that $q(J) \neq \emptyset$ for some $J \in \mathsf{H}$. Then $\mathcal{O}_\mathsf{H}$ contains the rule $\varphi \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$ where $\exists \bar{z}\, \psi(\bar{x}, \bar{z})$ is obtained from the CQ $q_{(P^*, \bar{b}^*)}$ by renaming the free variables, where

$$(P, \bar{b}) = \prod S \quad \text{and} \quad S = \{(J, \bar{c}) \mid J \in \mathsf{H}, \bar{c} \in q(J)\}.$$

For brevity, let $p_\Pi(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Since $I \models \mathcal{O}_\mathsf{H}$, $\bar{a} \in p_\Pi(I)$ and thus there is a homomorphism h from p_Π to I with $h(\bar{x}) = \bar{a}$. Since $J \models \rho$ for all $J \in \mathsf{H}$, we have $\bar{c} \in p(J)$ for all $(J, \bar{c}) \in S$. By Lemma 12, this implies $\bar{b} \in p(P)$, that is, there is a homomorphism g from p to P with $g(\bar{x}) = \bar{b}$. Note that the variables in $\bar{x} = x_1 \cdots x_k$ are all distinct as $\bar{x}$ is simply the list of frontier variables of ρ. By definition of diversification, we obtain from g a homomorphism g' from p to P^* with $g'(\bar{x}) = \bar{b}^*$ by defining $g'(x_i)$ to be the fresh value a_i^* introduced by diversification as a clone of the i-th value in $\bar{b}$. The composition $h \circ g'$ is a homomorphism from p to I and satisfies $h \circ g'(\bar{x}) = \bar{a}$. Consequently, $\bar{a} \in p(I)$, as required. ❏

Theorem 8. *Let $\mathcal{O}'_\mathsf{H}$ be the subset of $\mathcal{O}_\mathsf{H}$ that contains all TGDs from $\mathcal{O}_\mathsf{H}$ whose body has at most $n := ||\mathsf{H}|| + 1$ atoms. Then $\mathcal{O}'_\mathsf{H}$ is a GTGD-basis of H with $O(S^n \cdot \ell^{\ell n})$ TGDs where S is the number of relation symbols in the schema $\mathcal{S}$ of H, and ℓ is the maximal arity of a symbol in $\mathcal{S}$.*

Proof. We start with analyzing the size. Due to guardedness, we can assume w.l.o.g. that only variables $x_1, \ldots, x_\ell$ are used in TGD bodies in $\mathcal{O}_\mathsf{H}$. Then the number of possible atoms is

$$\sum_{P \in \mathcal{S}} \ell^{\mathsf{ar}(P)} \leq s \cdot \ell^\ell.$$

The number of possible TGD bodies built from at most $n = ||H|| + 1$ of the possible atoms is

$$\binom{s \cdot \ell^\ell}{n} \in O\left((s \cdot \ell^\ell)^n\right),$$

as required.

To show that $\mathcal{O}'_\mathsf{H}$ is indeed a basis, it suffices to verify that $\mathcal{O}'_\mathsf{H} \models \mathcal{O}_\mathsf{H}$. Let $q(\bar{x}) \to p(\bar{x}) \in \mathcal{O}_\mathsf{H}$. We claim that there is a TGD $q'(\bar{x}) \to p(\bar{x}) \in \mathcal{O}'_\mathsf{H}$ such that all atoms in q' are atoms in q. Clearly, this implies $q'(\bar{x}) \to p(\bar{x}) \models q(\bar{x}) \to p(\bar{x})$, which closes the proof.

Let $q(\bar{x}) = \exists \bar{y}\, \varphi(\bar{x}, \bar{y})$ and let $R(\bar{z})$ be the guard atom in $\varphi(\bar{x}, \bar{y})$ where $\bar{z} = \bar{x} \cup \bar{y}$. For simplicity, suppose that the guard atom takes the shape $R(z_1, \ldots, z_n)$ and that $\bar{x} = z_1, \ldots, z_k$ and $\bar{y} = z_{k+1}, \ldots, z_n$ for some k. The other cases can be treated similarly.

We make two observations on $q(I)$ for each $I \in \mathsf{H}$:

1. $q(I) \subseteq \{(a_1, \ldots, a_k) \mid R(a_1, \ldots, a_n) \in I\}$, and
2. for every $R(\bar{b}) \in I$, $\bar{b} = (b_1, \ldots, b_n)$ such that $(b_1, \ldots, b_k) \notin q(I)$, we find an atom $\alpha_{I,\bar{b}} = P(z_{i_1}, \ldots, z_{i_m})$ in q such that $P(b_{i_1}, \ldots, b_{i_m}) \notin I$.

Let $q' = \exists \bar{y}\, \varphi'(\bar{x}, \bar{y})$ where φ' is the conjunction of $R(z_1, \ldots, z_n)$ and all atoms $\alpha_{I,\bar{b}}$ identified in Item 2 above. By definition, all atoms in q' are atoms in q, as intended. By construction of q', we additionally have

$$\{(I, \bar{a}) \mid I \in \mathsf{H}, \bar{a} \in q(I)\} = \{(I, \bar{a}) \mid I \in \mathsf{H}, \bar{a} \in q'(I)\}.$$

Moreover, q' has at most $||\mathsf{H}|| + 1$ atoms. Hence, $q'(x) \to p(x) \in \mathcal{O}'_\mathsf{H}$ as claimed. ❏

The proof of Theorem 9 relies on a standard characterization of entailments of TGD-ontologies via the chase, which we briefly recall next. Let J be an instance, $\rho = p(\bar{x}) \to q(\bar{x})$ be a TGD, and $\bar{a}$ a tuple over $\mathsf{adom}(J)$. We say that ρ is *applicable to* $J, \bar{a}$ if there is a homomorphism from $p(\bar{x})$ to $(J, \bar{a})$. The *result of applying* ρ *to* $J, \bar{a}$ is obtained as the union of J and $I_q[\bar{x}/\bar{a}]$, which is the canonical instance I_q of q with $\bar{x}$ renamed to $\bar{a}$. A sequence of instances $J_0, J_1, \ldots$ is called a *chase sequence for J and $\mathcal{O}$* if $J = J_0$ and each J_{i+1} is the result of applying some $\rho \in \mathcal{O}$ to $J_i, \bar{a}$ for some $\bar{a}$. The sequence $J_0, J_1, \ldots$ is *fair* if for each $i \geq 0$ and each tuple $\bar{a}$ over $\mathsf{adom}(J_i)$ such that some $\rho \in \mathcal{O}$ is applicable to $J_i, \bar{a}$, this ρ is applied to $J_j, \bar{a}$ for some j. The *result* of the chase sequence $J_0, J_1, \ldots$ is the instance $\bigcup_{i \geq 0} J_i$. The following characterization is well known, see for example (Deutsch *et al.* 2008, Proposition 3).

Lemma 16. *Let $\mathcal{O}$ be some TGD-ontology and $\rho = p(\bar{x}) \to q(\bar{x})$ some TGD. Then*

$$\mathcal{O} \models \rho \quad \textit{iff} \quad \widehat{J}, \bar{a} \models q(\bar{a})$$

for some (equivalently: all) result $\widehat{J}$ of a fair chase sequence for I_p and $\mathcal{O}$.

Theorem 9. *For every $n \geq 1$, there is an instance I_n (over a schema with only unary and binary relations) such that*

1. *the size of I_n is bounded by $p(n)$, p a polynomial;*
2. *the smallest finite GTGD-basis $\mathcal{O}$ of I_n is of size 2^n.*

Proof. Let P and R be unary and binary symbols, respectively. For $m \geq 1$, let L_m denote the "lasso" instance, with values $a_0^m, \ldots, a_{2m-1}^m$ and facts

$$L_m = \{R(a_i^m, a_{i+1}^m) \mid i < 2m - 1\} \cup \{R(a_{2m-1}^m, a_m^m), P(a_m^m)\}$$

We now define, for $n \geq 1$, I_n as

$$I_n = \bigcup_{i=1}^{n} \left(L_{p_i} \cup \{A(a_0^{p_i})\}\right),$$

where p_i denotes the i-th prime number (that is, $p_1 = 2, p_2 = 3, \ldots$).

We claim that the constructed instances are as required. For Point 1 observe that, by the prime number theorem, $p_n = O(n \log n)$, and thus the size of I_n is bounded by some polynomial in n.

For Point 2, we rely on the following fact from (ten Cate *et al.* 2024):

(†) Any unary CQ $q(x)$ which contains a P-atom connected to x and such that the TGD $A(x) \to q(x)$ is satisfied in I_n of size at least 2^n.

Here, *connected* is defined as usual in terms of the undirected graph $(\mathsf{var}(q), \{\{x, y\} \mid x, y \text{ co-occur in some atom in } q\})$ induced by q. The fact (†) can be proved along the lines of the proof of Theorem 3.2 in (ten Cate *et al.* 2024). Since it is rather close to that proof we omit the details. The intuition is that P-atoms occur only in prime distance to A atoms, so the assumption that $A(x) \to q(x)$ is satisfied in I_n means that the distance of any P-atom from x is the product of all the used primes which is at least 2^n.

We proceed with the proof of Point 2. Let $\mathcal{O}$ be a smallest GTGD-basis for I_n, and consider the TGD $\rho_n = A(x) \to q_1(x_1)$ with

$$q_1(x_1) = \exists x_2 \dots x_k . R(x_1, x_2) \wedge \dots R(x_{k-1}, x_k) \wedge P(x_k)$$

where $k = \Pi_{i=1}^{n} p_i$. Since $I_n \models \rho_n$, we must have $\mathcal{O} \models \rho_n$. By Lemma 16, we have $\widehat{J} \models q_1(a)$ for the result $\widehat{J}$ of some fair chase sequence for $\{A(a)\}$ and $\mathcal{O}$. Since $\mathcal{O}$ is a GTGD-ontology, the left-hand sides of TGDs in $\mathcal{O}$ have a very restricted shape. Essentially, they can be only one of the following CQs:

$$A(x) \qquad R(x,y) \qquad A(x), R(x,y) \tag{1}$$

$$P(x) \qquad R(x,y), P(y) \qquad P(x), R(x,y) \tag{2}$$

(There are more guarded CQs but those will have no match into $\widehat{J}$, e.g. $A(x), P(x)$.)

For the analysis, let $J_0, J_1, \dots$ be the fair chase sequence that leads to $\widehat{J}$ and let i be the first index such that J_i contains some P-fact connected to the value a in the initial instance J_0. Since we can consider any fair chase sequence, it is without loss of generality to assume that before that step i only rules with rule heads of shape (1) have been applied. We distinguish now cases on the head of the rule applied in the last step:

- Consider first rule head $A(x)$, that is, a rule of shape
$$\rho = A(x) \to q(x)$$
for some CQ $q(x)$. By assumption, q contains some atom $P(y)$ with y connected to x. Since $\rho \in \mathcal{O}$ and $\mathcal{O}$ is a GTGD-basis for I_n, we have $I_n \models \rho$. Fact (†) implies that $q(x)$ and thus $\mathcal{O}$ is of size at least 2^n.
- Consider next rule head $R(x, y)$, that is, a rule of shape
$$\rho = R(x,y) \to q(x,y)$$
for some CQ $q(x, y)$. As in the previous case, we can conclude that $q(x, y)$ contains an atom $P(z)$ for some z connected to x. Let $q'(x, y)$ be connected component of x in $q(x, y)$. Since $\rho \in \mathcal{O}$, we know that there is a homomorphism from $q'(x, y)$ to $(I_n, (b_1, b_2))$ for every atom $R(b_1, b_2) \in I_n$. A straightforward analysis shows that q' cannot contain a P atom, a contradiction. (Informally, for each distance ℓ we can find an atom $R(b_1, b_2)$ so that there is no P-atom in distance ℓ from b_1 in I_n.)
- The case of rule head $A(x), R(x, y)$ can be treated as the first case.

This finishes the proof of Theorem 9. ❑

Theorem 10. *There are finite instances $I_1, I_2, \dots$ such that for all $n \geq 1$, the size of I_n is polynomial in n, but all finite IND-bases of I_n contain at least 2^n INDs.*

Proof. The instance I_n contains, for $1 \leq i \leq n$, the facts

$$S(\bar{a}_i) \quad R(\bar{b}_i) \quad R(\bar{c}_i)$$

with $\bar{a}_i$ a tuple of $2n$ pairwise distinct values, $\bar{b}_i$ identical to $\bar{a}_i$ except that a fresh value is in position $2i-1$, and $\bar{c}_i$ identical to $\bar{a}_i$ except that a fresh value is in position $2i$. Then among the INDs true in I_i we find all 2^n INDs of the form

$$S(x_1, y_1, \dots, x_n, y_n) \to \exists z_1 \cdots \exists z_n \, R(u_1, v_1, \dots, u_n, v_n)$$

where for $1 \leq i \leq n$, either $(u_i, v_i) = (x_i, z_i)$ or $(u_i, v_i) = (z_i, y_i)$. Moreover, it is not difficult to see that (i) each of these INDs ρ is logically strongest among the INDs in I_n, that is, there is no IND ρ' true in I_n such that $\rho' \models \rho$; and (ii) for a set of INDs $\mathcal{O}$ and an IND ρ, $\mathcal{O} \models \rho$ if and only if there is a $\rho' \in \mathcal{O}$ such that $\rho' \models \rho$. As a consequence, any finite IND-basis of I_n must contain all of the 2^n INDs above. ❑

Lemma 17. *The instance I has no finite FGTGD-basis and no finite F1TGD-basis.*

Proof. Assume to the contrary that $\mathcal{O}$ is a finite FGTGD-basis or F1TGD-basis of I, and let n be the maximal number of variables contained in the body of any TGD in $\mathcal{O}$. Set $m = 2n + 1$ and let J denote the instance that is a bi-directional R-cycle of length m, that is, $R(d_i, d_{i+1}) \in J$ and $R(d_{i+1}, d_i) \in J$ for all $i \in [m]$ (indices taken modulo m).

Consider ρ_m. Since m is odd and I satisfies ρ_k for odd k, $I \models \rho_m$. Therefore, $\mathcal{O} \models \rho_m$. Regarding J, by construction every d_i lies on a cycle of length m but $R(d_i, d_i) \notin J$, so $J \not\models \rho_m$. We proceed by showing that $J \models \mathcal{O}$, which implies $\mathcal{O} \not\models \rho_m$, yielding the desired contradiction.

Take any TGD $\rho \in \mathcal{O}$ and let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z} \, \psi(\bar{x}, \bar{z})$. For brevity, define $q(\bar{x}) = \exists \bar{y} \, \varphi$ and $p(\bar{x}) = \exists \bar{z} \, \psi$. Let $h_{q \to J}$ be a homomorphism from q to J. Define J' as the maximal subinstance of J whose active domain is the image of $h_{q \to J}$. Clearly, J' has at most n values. This allows us to define a homomorphism $h_{J' \to I}$ from J' to I: the elements of any maximal contiguous sequence $d_i, \dots, d_{i+k} \in \mathsf{adom}(J')$, are mapped alternately to a and b.

The composition $h_{q \to I} = h_{J' \to I} \circ h_{q \to J}$ defines a homomorphism from q to I. Since I satisfies ρ, there is thus also a homomorphism $h_{p \to I}$ from p to I with $h_{p \to I}(\bar{x}) = h_{q \to I}(\bar{x})$.

Next observe that the restriction g of $h_{J' \to I}$ to the set of elements $S = \{h_{q \to J}(x) \mid x \text{ frontier-variable in } \rho\}$ is injective. This is trivially the case if ρ contains at most one frontier variable. It is also the case if ρ contains two frontier variables since ρ is frontier-guarded and I contains no reflexive R-loop. Therefore, the inverse g^- of g is a well-defined mapping from $\mathsf{adom}(I)$ to $\mathsf{adom}(J)$.

Moreover, it is easy to verify that g^- is a homomorphism from I to J because the cycle that we had added in the construction of J is bi-directional and thus for every edge $R(d, e) \in J'$ there is also the edge $R(e, d)$. It follows that

$h_{p\to J} = h_{p\to I} \circ g^-$ is a homomorphism from p to J with $h_{p\to J}(\bar{x}) = h_{q\to J}(\bar{x})$, and therefore $J \models \rho$. We conclude $J \models \mathcal{O}$, as required. ❑

$$\begin{aligned}\mathcal{O}_I = \{\quad & R(x,y) \to R(y,x) \\ & R(x,y) \wedge R(y,z) \wedge R(z,u) \to R(x,u) \\ & E(x,x) \wedge \mathsf{true}(x) \wedge \mathsf{true}(y) \to R(y,z)\ \}\end{aligned}$$

Lemma 9. *$\mathcal{O}_I$ is a finite FullTGD-basis of I and also a finite TGD-basis.*

Proof. Let J be a model of $\mathcal{O}_I$ and take any full TGD $\rho = q \to R(x_1, x_1') \wedge \cdots \wedge R(x_k, x_k')$ with $I \models \rho$. Let h be a homomorphism from q to J. We have to show that $R(h(x_i), h(x_i')) \in J$ for $1 \le i \le k$.

First assume that q does not admit a homomorphism to I. By choice of I, the canonical instance I_q of q then contains a cycle of odd length, and so does J. But since the first and second TGD in $\mathcal{O}_I$ are satisfied in J, this implies that J must contain a reflexive loop. By the third TGD, J is total, that is, it contains $R(a,b)$ for all $a, b \in \mathsf{ind}(J)$. Therefore, we clearly have $R(h(x_i), h(x_i')) \in J$ for $1 \le i \le k$, as required.

Next assume that q admits a homomorphism g to I. Then the canonical instance I_q of q is bipartite with partition P_1, P_2 where P_1 contains all variables $x \in \mathsf{var}(q)$ with $g(x) = a$, and likewise for P_2 and $g(x) = b$.

Take any head atom $R(x_i, x_i')$. Then x_i' are neither both in P_1 nor in P_2 because otherwise g would witness that I does not satisfy ρ. What is more, x_i and x_i' must belong to the same maximally connected component of I_q. For if they belonged to different connected components, we could find a homomorphism from q to I that maps both x_i and x_i' to the same value a or b, but I has no reflexive loops which contradicts satisfaction of ρ in I.

Let Q_1 be the set of all values $h(x) \in \mathsf{adom}(J)$ such that $x \in P_1$ and x_1 is in the same maximally connected component as x_i and x_i', and likewise for Q_2. Since J satisfies the first and second TGD in $\mathcal{O}_I$, there must be an edge $R(c_1, c_2) \in J$ for all $c_1 \in Q_1$ and $c_2 \in Q_2$; note that this relies on our careful use of maximally connected components. But since x_i and x_i' are neither both in P_1 nor in P_2, this implies $R(h(x_i), h(x_i')) \in J$, as required. ❑

Theorem 13. *There are instances J that have no finite FullTGD-basis.*

Proof. Assume to the contrary that $\mathcal{O}$ is a finite FullTGD-basis of J, and let n be the maximal number of variables contained in the body of any TGD in $\mathcal{O}$.

Take any graph G that is not 3-colorable and has girth exceeding n and view it as an instance, that is, make it bidirectional. Consider the TGD ρ_G. Since G is not 3-colorable and J satisfies ρ_G for non-3-colorable G, we have $J \models \rho_G$. Therefore, $\mathcal{O} \models \rho_G$. We proceed to show that $G \models \mathcal{O}$, which implies $\mathcal{O} \not\models \rho_G$ since clearly $G \not\models \rho_G$, yielding the desired contradiction.

Take any TGD $\rho \in \mathcal{O}$, with $\rho = q \to R(x_1, x_1') \wedge \cdots \wedge R(x_k, x_k')$. Let h be a homomorphism from q to G. Define G' as the subinstance/graph of G induced by the values/vertices in the domain of h. Clearly, G' has at most n values. Since the girth of G exceeds n, G' is acyclic and thus 3-colorable. Moreover, for any two vertices u, v in G', we find a 3-coloring of G' that colors u and v with different colors. Choosing $u = h(x_i)$ and $v = h(x_i')$ for some $i \in [n]$, this implies that there is a homomorphism g from G' to J such that $R(g(h(x_i)), g(h(x_i'))) \notin J$. The composition $g \circ h$ is a homomorphism from q to J. Since J satisfies ρ, we must thus have $R(g \circ h(x_i), g \circ h(x_i')) \in J$, a contradiction. ❑

F Proofs for Section 6

Theorem 14. *Fitting IND existence and fitting IND-ontology existence are* NP*-complete.*

Proof. *3SAT* is the problem to decide the satisfiability of propositional logic formulas that are in conjunctive normal form with exactly three distinct literals per clause. It is famously known to be NP-hard. Without loss of generality, we additionally require that no clause contains both a variable and its negation. We provide a polynomial time reduction from *3SAT* to fitting IND-ontology existence and fitting IND existence. The fitting instances used in the reduction contain a single negative example and thus, by Lemma 3, the existence of a fitting ontology and a fitting IND coincide.

Let the input to *3SAT* be the propositional formula $\varphi = C_1 \wedge \cdots \wedge C_m$ using the variables $p_1, \ldots, p_n$. We construct a fitting instance $(\mathsf{P} = \{I_1, \ldots, I_m\}, \mathsf{N} = \{J\})$ over the schema $\{R, S\}$, where $\mathsf{ar}(R) = \mathsf{ar}(S) = 2n$. With every position $i \in [2n]$, we associate the literal p_i, if $i \le n$, and $\neg p_{i-n}$, if $i > n$. For brevity, we speak about *position* p_i to mean position i and *position* $\neg p_i$ to mean position $n + i$.

Fix a tuple $\bar{a} = (a_1, \ldots, a_{2n})$ of values. We will consider tuples of values obtained from $\bar{a}$ by replacing some values with fresh values. Each clause C_i gives rise to a set M_i of exactly 8 such tuples, obtained from $\bar{a}$ by replacing, for ℓ_1, ℓ_2, ℓ_3 the literals in C_i, either the value in position ℓ_i or the value in position $\neg\ell_i$ with a fresh value, for all $i \in \{1, 2, 3\}$. The tuple in M_i that has fresh values in all three positions ℓ_1, ℓ_2, ℓ_3 is referred to as the *falsifying tuple*. We now define the instances I_i and J as follows:

- I_i contains $R(\bar{a})$ and all facts $S(\bar{b})$ with $\bar{b} \in M_i$ non-falsifying;
- J contains $R(\bar{a})$ and, for $1 \le i \le m$, the fact $S(\bar{b})$ where $\bar{b} \in M_i$ is falsifying.

We need to show that φ is satisfiable if and only if (P, N) has a fitting IND. Whenever we consider tuples of variables $\bar{x}$ of length $2n$ and a literal ℓ, then we use x_ℓ to denote the variable in $\bar{x}$ in position ℓ, and likewise for $\bar{y}$ etc.

"⇒" Assume φ is satisfiable and let V be a satisfying variable assignment. Consider the IND

$$\rho = R(\bar{x}) \to \exists \bar{z}\, S(\bar{y}),$$

where $\bar{x}$, $\bar{y}$, $\bar{z}$ are tuples of variables of length $2n$. The variables in $\bar{x}$ and $\bar{z}$ are all distinct. If $V(p) = 1$, then the variable in position p of $\bar{y}$ is x_p and the variable in position $\neg p$ is $z_{\neg p}$. If $V(p) = 0$, then the variable in position p of $\bar{y}$ is z_p and the variable in position $\neg p$ is $x_{\neg p}$. We show that ρ fits (P, N), starting with the positive examples.

Let $I_i \in \mathsf{P}$ and let $C_i = \ell_1 \vee \ell_2 \vee \ell_3$ be the corresponding clause. There is exactly one homomorphism h from the body of ρ to I_i, mapping the atom $R(\bar{x})$ to the fact $R(\bar{a})$. Consider the fact $S(\bar{b})$ where $\bar{b}$ is obtained from $\bar{a}$ by replacing, for all $i \in \{1,2,3\}$, the value in position $\neg\ell_i$ by a fresh value if $V \models \ell_i$ and the value in position ℓ_i by a fresh value if $V \not\models \ell_i$. Since V satisfies C_i, the tuple $\bar{b}$ is not the falsifying tuple. By construction, it follows that I_i contains $S(\bar{b})$. Using the construction of ρ, it is now easy to verify that the mapping g defined by $g(\bar{z}) = \bar{b}$ is a homomorphism from the head of ρ to I_i that maps the atom $S(\bar{y})$ to $S(\bar{b})$. This homomorphism witnesses satisfaction of the IND.

Now consider the negative example. Again, there is exactly one homomorphism h from the body of ρ to J, mapping the atom $R(\bar{x})$ to the fact $R(\bar{a})$. Consider any fact $S(\bar{b})$ in J. We show that $S(\bar{b})$ does witness satisfaction of the IND. By construction of J, $S(\bar{b})$ is a falsifying tuple, say for clause $C_i = \ell_1 \vee \ell_2 \vee \ell_3$. We have $V(\ell_i) = 1$ for some $i \in \{1,2,3\}$. But then the tuple $\bar{y}$ has variable x_{ℓ_i} in position ℓ_i whereas the tuple $\bar{b}$ has a fresh value (rather than the value from position ℓ_i of $\bar{a}$) in position ℓ_i. Thus, $S(\bar{y})$ cannot be mapped to $S(\bar{b})$ by a homomorphism that agrees with h on the variables that occur both in $\bar{y}$ and $\bar{x}$.

"$\Leftarrow$" Assume that (P, N) has a fitting IND ρ. We first prove that ρ satisfies the following properties:

1. there are no repeated variables in the body atom, and neither in the head atom;
2. every frontier variable of ρ has the same position in the head atom as it has in the body atom;
3. the relation symbol in the body of ρ is R, the relation symbol in the head is S;
4. for every variable p, not more than one of the variables in positions p and $\neg p$ in the body atom occurs in the frontier of ρ.

Property 1. It is easy to verify that for every instance I in (P, N) and every atom α over schema $\{R, S\}$ that contains variables $\bar{x}$, we have $I \models \exists\bar{x}\,\alpha$ if and only if α does not contain any variable more than once. Thus, if an atom with repeated variables occurs in the body of ρ, then $J \models \rho$; if no such atom occurs in the body, but some such atom occurs in the head, then $I \not\models \rho$ for all $I \in \mathsf{P}$. Either case contradicts that ρ is a fitting, hence ρ satisfies Property 1.

Property 2. Since Property 1 is satisfied, the body atom of ρ contains every variable at most once. Every such atom admits a homomorphism to any positive example I_i. But in I_i, every value occurs only in exactly one position across all facts. Since $I_i \models \rho$, this implies that Property 2 must be satisfied.

Property 3. We first observe that ρ must use both R and S as, otherwise, Properties 1 and 2 imply that ρ is a tautology, in contradiction to $J \not\models \rho$. The remaining case to be ruled out is thus that ρ uses S in the body and R in the head. The frontier must contain some variable as otherwise $J \models \rho$. Let i be the position in which x occurs in the head atom $R(\bar{y})$. By construction, there is a positive example I_j that contains a fact $S(\bar{b})$ with a fresh value in position i. But the only R-fact in I_j is $R(\bar{a})$, which contains a non-fresh (and thus distinct) value in position i. It follows that $I_j \not\models \rho$, a contradiction.

Property 4. Let C_i be a clause in which variable p occurs and consider the positive example I_i. There is only one homomorphism from the body $R(\bar{x})$ of ρ to I_i, mapping $R(\bar{x})$ to $R(\bar{a})$. By construction of I_i, every S-fact in I_i has a fresh value on position p or $\neg p$. Consequently, if the head atom $S(\bar{y})$ in ρ contains a frontier variable in both position p and $\neg p$, then $I_i \not\models \rho$, a contradiction.

By Properties 1–3, we may assume that

$$\rho = R(\bar{x}) \to \exists\bar{z}\, S(\bar{y})$$

where $\bar{x}$ and $\bar{y}$ contain no repeated variables and if some variable x from $\bar{x}$ occurs in $\bar{y}$, then this is in the same place as the occurrence of x in $\bar{x}$.

We construct an assignment V that φ: set $V(p) = 1$ if the variable in position p of $\bar{y}$ is from $\bar{x}$ and $V(p) = 0$ if it is from $\bar{z}$. We have to argue the V satisfies all clauses in φ.

Take any such clause C. Then J contains the fact $S(\bar{b})$ where $\bar{b}$ is the falsifying tuple for C. Since ρ fits the negative example, the homomorphism h that maps $R(\bar{x})$ to $R(\bar{a})$ cannot be extended to a homomorphism that maps $S(\bar{y})$ to $S(\bar{b})$. This can only be the case if there is a position i in $\bar{y}$ that contain a variable from $\bar{x}$, but such that position i of $\bar{b}$ contain a fresh value (rather than a value from $\bar{a}$). First assume that position i is associated with a variable p. Since $\bar{y}$ contains a variable from $\bar{x}$ in position p, we have $V(p) = 1$. Since position p of $\bar{b}$ contains a fresh value and b is the falsifying tuple, p is one of the literals in C. The case that position i is associated with a negated variable is similar, additionally using Property 4 above. ❏

Theorem 15. *Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \ldots, N_k\}$. Then no GTGD fits (P, N) if and only if for every non-empty maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, the following conditions are satisfied:*

1. *the following set is non-empty:*

$$S_M = \{(J, \bar{b}) \mid J \in \mathsf{P} \text{ and } \bar{b} \in \mathsf{adom}(J)^{|M|} \text{ such that } (\textstyle\prod \mathsf{N}|_M, \bar{M}) \to (J, \bar{b})\}$$

2. *$\exists i \in [k]$: $(K^*, \bar{c}^*) \to (N_i, \bar{M}[i])$ where $(K, \bar{c}) = \prod S_M$.*

Proof. "$\Rightarrow$". We show the contrapositive. Assume that there is some non-empty maximally guarded $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$ and at least one of the Conditions 1 and 2 is violated. We have to show that (P, N) has a fitting guarded TGD.

To this end we will construct a GTGD ρ such that $N \not\models \rho$ for all $N \in \mathsf{N}$, and $J \models \rho$ for all $J \in \mathsf{P}$. Consider $(\prod \mathsf{N}|_M, \bar{M})$. Note that since M is non-empty and maximally guarded, $\prod \mathsf{N}|_M$ cannot be empty, and thus the canonical CQ $q_{(\prod \mathsf{N}|_M, \bar{M})}$ is defined. Then the body of ρ is $\varphi(\bar{x})$, where the CQ $q_M(\bar{x}) = \varphi(\bar{x})$ is obtained from $q_{(\prod \mathsf{N}|_M, \bar{M})}$ by renaming the free variables. Clearly $\bar{M} \in q_M(\prod \mathsf{N})$, and thus Lemma 12 yields $\bar{M}[i] \in q_M(N_i)$ for every $i \in [n]$. Since M is guarded, there is a fact in $\prod \mathsf{N}|_M$ that contains

every value from $\mathsf{adom}(\prod \mathsf{N}|_M)$. Then by definition there exists an atom in q_M that mentions every variable from $\mathsf{var}(q_M)$, and thus q_M is guarded, ensuring that any TGD with $\varphi(\bar{x})$ as body is guarded. Concerning the head of ρ, we make a case distinction.

First assume the violation of Condition 1, so $S_M = \emptyset$. Since for every $i \in [k]$, the tuple $\bar{M}[i]$ is non-total in N_i, there exists a CQ $q_i(\bar{x}) = \exists \bar{z}_i\, \psi_i(\bar{x})$, such that $\bar{M}[i] \notin q_i(N_i)$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, where $\psi = \bigwedge_{i\in[k]} \psi_i$, and $\bar{z} = \bigcup_{i\in[k]} \bar{z}_i$. For brevity define $q_H(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Clearly $\bar{M}[i] \notin q_H(N_i)$, and thus $N_i \not\models \rho$ for $i \in [n]$. Now let $J \in \mathsf{P}$. To show that $J \models \rho$, it suffices to argue that $q_M(J) = \emptyset$. Suppose to the contrary that there is a tuple $\bar{a}$ with $\bar{a} \in q_M(J)$. Then there is a homomorphism h from $q_M(\bar{x})$ to $(J, \bar{a})$. By appropriately renaming the elements in the domain of h to values from $\mathsf{adom}(\prod \mathsf{N}|_M)$, we obtain a homomorphism witnessing $(\prod \mathsf{N}|_M, \bar{M}) \to (J, \bar{a})$. Hence $(J, \bar{a}) \in S_M$, which contradicts the assumption that $S_M = \emptyset$.

Now assume that Condition 1 is satisfied (and thus S_M is non-empty), but Condition 2 is violated. Let $(K, \bar{c}) = \prod S_M$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, where $\exists \bar{z}\, \psi(\bar{x}, \bar{z})$ is obtained from the CQ $q_{(K^*, \bar{c}^*)}$ by renaming the free variables. For brevity, let $q_\Pi(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Then the violation of Condition 2 by $\bar{M}$ implies $\bar{M}[i] \notin q_\Pi(N_i)$, and thus $N_i \not\models \rho$ for $i \in [k]$. To see that all positive examples satisfy ρ, let $J \in \mathsf{P}$ and $\bar{a} \in q_M(J)$. Then $(J, \bar{a}) \in S_M$, which by Lemma 11 implies $(K, \bar{c}) \to (J, \bar{a})$. Clearly $(K^*, \bar{c}^*) \to (K, \bar{c})$ and thus $\bar{a} \in q_\Pi(J)$, so $J \models \rho$.

"$\Leftarrow$". Assume that for all non-empty maximally guarded $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, such that $\bar{M}[i]$ is non-total in N_i for $i \in [k]$, Conditions 1 and 2 are satisfied. Further assume, contrary to what we aim to show, that there exists a GTGD that fits (P, N).

Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$ be such a TGD. For brevity, define $q(\bar{x}) = \exists \bar{y}\, \varphi$ and $p(\bar{x}) = \exists \bar{z}\, \psi$. As ρ fits (P, N), it follows that for every $i \in [k]$ there is some $\bar{a}_i \in \mathsf{adom}(N_i)^{|\bar{x}|}$ with $\bar{a}_i \in q(N_i)$ and $\bar{a}_i \notin p(N_i)$, witnessed by a homomorphism $h_{q\to N_i}$ from $q(\bar{x})$ to $(N_i, \bar{a}_i)$. Our goal is to show that for some i, we have $\bar{a}_i \in p(N_i)$, thereby obtaining the desired contradiction. To utilize both conditions, we first need to find an appropriate non-empty maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$.

Consider the instance $(\prod \mathsf{N}, \bar{b})$, where $\bar{b} = \bar{a}_1 \times \cdots \times \bar{a}_k$. Define $h_{q\to\Pi} : \mathsf{adom}(I_q) \to \mathsf{adom}(\prod \mathsf{N})$ by $h_{q\to\Pi}(y) = (h_{q\to N_1}(y), \ldots, h_{q\to N_k}(y))$. Since ρ is guarded, so is q, and thus, there exists a maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $h_{q\to\Pi}(\mathsf{adom}(I_q)) \subseteq M$. The non-emptiness of M follows from the fact that I_q, as the canonical instance of q, trivially has a non-empty active domain. By the definition of products, $h_{q\to\Pi}$ is a homomorphism witnessing $\bar{b} \in q(\prod \mathsf{N})$, and since $\bar{b} \subseteq M$, this implies that $h_{q\to\Pi}$ also witnesses $\bar{b} \in q(\prod \mathsf{N}|_M)$.

To ensure that M satisfies Conditions 1 and 2, it remains to show that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$. Let $i \in [k]$ and observe that $\bar{b}[i] = \bar{a}_i$, and thus $\bar{b}[i]$ is non-total in N_i because $\bar{a}_i \notin p(N_i)$. As an immediate consequence of the definition of totality, every $M' \subseteq \mathsf{adom}(N_i)$ with $\bar{b}[i] \subseteq M'$ is also non-total in N_i. Note that $(d_1, \ldots, d_k) \in \bar{M}$ implies $d_i \in \bar{M}[i]$ by definition of products. Therefore $\bar{b} \subseteq M$ implies $\bar{b}[i] \subseteq \bar{M}[i]$ and thus $\bar{M}[i]$ is non-total in N_i. Hence, M satisfies Conditions 1 and 2.

This allows us to invoke Condition 1, which states that the set S_M is non-empty, so let $(P_1, \bar{d}_1), \ldots, (P_m, \bar{d}_m)$ be an enumeration of the pointed instances in S_M, with $m > 0$. Consider $(P_j, \bar{d}_j)$ for some $j \in [m]$. The definition of S_M yields $(\prod \mathsf{N}|_M, \bar{M}) \to (P_j, \bar{d}_j)$. Composing $h_{q\to\Pi}$ with the witnessing homomorphism yields a homomorphism $h_{q\to P_j}$ from q to P_j such that $h_{q\to P_j}(\bar{x}) \subseteq \bar{d}_j$. Recall that ρ is a fitting TGD and therefore $P_j \in \mathsf{P}$ implies $h_{q\to P_j}(\bar{x}) \in p(P_j)$. Hence, there exists a homomorphism $g_{p\to P_j}$ from $p(\bar{x})$ to $(P_j, h_{q\to P_j}(\bar{x}))$.

Let $(K, \bar{c}) = \prod S_M$ and define $g_{p\to K} : \mathsf{adom}(I_p) \to \mathsf{adom}(K)$ by setting $g_{p\to K}(y) = (g_{p\to P_1}(y), \ldots, g_{p\to P_m}(y))$. By construction $g_{p\to K}(\bar{x}) \subseteq \bar{c}$ and by definition of products $g_{p\to K}$ is a homomorphism. Define a homomorphism $g_{p\to K^*}$ from $p(\bar{x})$ to $(K^*, \bar{c}^*)$ by $g_{p\to K^*}(y) = c_i^*$ if $g_{p\to K}(y) = c_i$ and $g_{p\to K^*}(y) = g_{p\to K}(y)$ otherwise. Clearly $g_{p\to K^*}(\bar{x}) \subseteq \bar{c}^*$ and since $g_{p\to K}$ is a homomorphism, the definition of diversification guarantees that $g_{p\to K^*}$ is also a homomorphism.

By Condition 2 we have $(K^*, c^*) \to (N_i, \bar{M}[i])$ for some $i \in [k]$. Composing $g_{p\to K^*}$ with the witnessing homomorphism yields a homomorphism $g_{p\to N_i}$ from p to N_i such that $g_{p\to N_i}(\bar{x}) \subseteq \bar{M}[i]$. We next establish that in fact $g_{p\to N_i}(\bar{x}) = \bar{a}_i$, thus showing $\bar{a}_i \in p(N_i)$ and thereby obtaining a contradiction, as desired. Let $x \in \bar{x}$ and let $\bar{M} = (M_1, \ldots, M_n)$. Observe that if $h_{q\to\Pi}(x) = M_\ell$ for $\ell \in [n]$, then $h_{q\to N_i}(x)$ is the ℓ-th element of $\bar{M}[i]$ due to construction of $h_{q\to\Pi}$. Furthermore the construction of the homomorphisms $h_{q\to P_j}$, $g_{p\to P_j}$, $g_{p\to K}$, $g_{p\to K^*}$ and $g_{p\to N_i}$ all preserve the order of the original mapping of $\bar{x}$ to $\bar{M}$ induced by $h_{q\to\Pi}$. Thus $g_{p\to N_i}(x) = h_{q\to N_i}(x)$, which entails $g_{p\to N_i}(\bar{x}) = \bar{a}_i$, as required. ❏

For easier reference, we spell out the characterizations for F1TGD, FGTGD and TGD explicitly below. In order to prove them, we need the following lemma.

Lemma 18. *Let ρ be a TGD over $\mathcal{S}$ such that ρ has no frontier variables. Then, there exists a TGD ρ' over $\mathcal{S}$ with exactly one frontier variable such that, for all finite $\mathcal{S}$-instances I, the following equivalence holds:*

$$I \models \rho \text{ if and only if } I \models \rho'.$$

Proof. Let $\rho = \varphi(\bar{y}) \to \exists \bar{z}\, \psi(\bar{z})$ be a TGD over $\mathcal{S}$ with an empty frontier and let I be a finite $\mathcal{S}$-instance. Consider an arbitrary atom $R(\bar{x})$ of $\varphi(\bar{y})$ and some $x \in \bar{x}$. Let $R(\bar{v})$ be the result of replacing every x_i, except x, by a variable v_i not occurring in ρ. Consider the TGD $\rho' = \varphi(\bar{y}) \to \exists \bar{w} \exists \bar{z}\, (\psi(\bar{z}) \wedge R(\bar{v}))$, where $\bar{w} = \bar{v} \setminus x$. Since the frontier of ρ' is x, ρ' is frontier-one and it is easy to see that for every instance I the equivalence $I \models \rho \Leftrightarrow I \models \rho'$ holds. ❏

Before we attend to the characterization of fitting FGTGDs, we show another lemma. The proof of the FGTGD characterization relies on guarded sets rather than maximally guarded sets. However, a characterization in terms of maximally guarded sets is desirable. As we shall see later, it

enables a complexity analysis that yields a tight upper bound for fitting FGTGD and ontology existence. The following lemma will be used to bridge the gap between guarded sets and maximally guarded sets.

Lemma 19. *Let* (P, N) *be a fitting instance where* $\mathsf{N} = \{N_1, \ldots, N_k\}$. *Let* $\emptyset \subsetneq M \subseteq M' \subseteq \mathsf{adom}(\prod \mathsf{N})$ *be such that* M' *satisfies the following conditions:*

1. *the following set is non-empty:*

$$S_{M'} = \{(J, \bar{b}) \mid J \in \mathsf{P} \text{ and } \bar{b} \in \mathsf{adom}(J)^{|M'|} \text{ such that } (\prod \mathsf{N}, \bar{M'}) \to (J, \bar{b})\}$$

2. $(K^*, \bar{c}^*) \to (N_i, \bar{M'}[i])$ *for some* $i \in [k]$ *where* $(K, \bar{c}) = \prod S_{M'}$.

Then M *also satisfies Conditions 1 and 2.*

Proof. Let $\emptyset \subsetneq M \subseteq M' \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that M' satisfies Conditions 1 and 2. Let π be the projection mapping a tuple of length $|M'|$ to a subtuple of length $|M|$, induced by $\bar{M'}$ and $\bar{M}$, that is $\pi(\bar{M'}) = \bar{M}$. Note that π is well-defined because $M \subseteq M'$ and $\bar{M'}$ contains no value more than once. Observe that

1. every homomorphism witnessing $(\prod \mathsf{N}, \bar{M'}) \to (I, \bar{a})$ also witnesses $(\prod \mathsf{N}, \bar{M}) \to (I, \pi(\bar{a}))$, and
2. every homomorphism witnessing $(\prod \mathsf{N}, \bar{M}) \to (I, \bar{a})$ also witnesses $(\prod \mathsf{N}, \bar{M'}) \to (I, \bar{a}')$ for some $\bar{a}'$ with $\pi(\bar{a}') = \bar{a}$.

Since M' satisfies Condition 1, $S_{M'}$ is non-empty. Then it is an immediate consequence of Observation 1 that S_M is also non-empty.

Define $(K, \bar{c}) = \prod S_M$ and $(K', \bar{c}') = \prod S_{M'}$. In order to prove that M satisfies Condition 2, we first show that $(K^*, \bar{c}^*) \to (K'^*, \pi(\bar{c}'^*))$. Let $|S_M| = n$ and $|S_{M'}| = m$. It is not hard to see from Observations 1 and 2 that $n \leq m$ and that any $I \in \mathsf{P}$ occurring in some pointed instance of S_M, must also occur in some pointed instance of S'_M and vice versa. Define $f : [m] \to [n]$ such that for $j \in [m]$, if $(I, \bar{a})$ is the j-th component of $(K', \bar{c}')$, then $f(j) = i$, where i is the position of $(I, \pi(\bar{a}))$ as a component of $(K, \bar{c})$.

This allows us to define a homomorphism from $(K^*, \bar{c}^*)$ to $(K'^*, \pi(\bar{c}'^*))$. Let $h : \mathsf{adom}(K^*) \to \mathsf{adom}(K'^*)$, such that $h(\bar{c}^*) = \pi(\bar{c}'^*)$ and for every $a = (a_1, \ldots, a_n) \notin \bar{c}^*$, $h(a) = (a_{f(1)}, \ldots, a_{f(m)})$. Intuitively, every a that is not a clone is sent to an $h(a)$ that has only components that already occur in a. To see that h indeed is a homomorphism, let $R(\bar{a}) \in K^*$. If none of the values of the fact $R(\bar{a})$ is from $\bar{c}^*$, by the definition of h and f, for every $j \in [m]$, $h(\bar{a})[j] = \bar{a}[i]$ for some $i \in [n]$. Observe that by the definition of products, $R(\bar{a}) \in K^*$ implies that $R(\bar{a}[i])$ is a fact of the i-th component of S_M for every $i \in [n]$. Thus by the definition of products $R(h(\bar{a})) \in K' \subseteq K'^*$. If some values from $\bar{c}^*$ do occur in $\bar{a}$, diversification guarantees that there is a fact $R(\bar{a}') \in K$, where in place of any cloned c^*, the corresponding c occurs. Since $R(\bar{a}')$ contains no cloned values, the previous argument applies and thus $R(h(\bar{a}')) \in K'$. Observe that f ensures that h preserves the order induced by π, that is, if c is the i-th element of $\bar{c}$, then $h(c)$ is the i-th element of $\pi(\bar{c}')$. Since by definition $h(\bar{c}^*) = \pi(\bar{c}'^*)$, the fact $R(h(\bar{a}))$ is obtained from $R(h(\bar{a}'))$ by diversification, and thus $R(h(\bar{a})) \in K'^*$.

We will now show that $(K^*, \bar{c}^*) \to (K'^*, \pi(\bar{c}'^*))$ implies that M satisfies Condition 2. Since M' satisfies Condition 2, there is an $i \in [k]$ such that $(K'^*, (\bar{c}'^*)) \to (N_i, \bar{M'}[i])$, which clearly implies $(K'^*, \pi(\bar{c}'^*)) \to (N_i, \pi(\bar{M'}[i]))$. The composition yields a homomorphism witnessing $(K^*, \bar{c}^*) \to (N_i, \pi(\bar{M'}[i]))$. By definition $\pi(\bar{M'})[i] = M[i]$, and it can be easily verified that $\pi(\bar{M'})[i] = \pi(\bar{M'}[i])$. Thus $(K^*, \bar{c}^*) \to (N_i, M[i])$, which concludes the proof. ❏

Theorem 24. *Let* (P, N) *be a fitting instance where* $\mathsf{N} = \{N_1, \ldots, N_k\}$. *Then no FGTGD fits* (P, N) *if and only if for every non-empty maximally guarded set* $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ *such that* $\bar{M}[i]$ *is non-total in* N_i *for all* $i \in [k]$, *the following conditions are satisfied:*

1. *the following set is non-empty:*

$$S_M = \{(J, \bar{b}) \mid J \in \mathsf{P} \text{ and } \bar{b} \in \mathsf{adom}(J)^{|M|} \text{ such that } (\prod \mathsf{N}, \bar{M}) \to (J, \bar{b})\}$$

2. $(K^*, \bar{c}^*) \to (N_i, \bar{M}[i])$ *for some* $i \in [k]$ *where* $(K, \bar{c}) = \prod S_M$.

Proof. Instead of showing the above statement, we prove an equivalent one, where instead of maximally guarded sets, guarded sets are considered. First we show that both statements are indeed equivalent. Let (P, N) be a fitting instance. It is sufficient to show that every non-empty maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, satisfies Conditions 1 and 2 if and only if the same holds for every non-empty guarded set $M' \subseteq \mathsf{adom}(\prod \mathsf{N})$. Since every maximally guarded set is guarded, the direction from guarded to maximally guarded sets holds trivially. Regarding the other direction, assume that every maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, satisfies Conditions 1 and 2. Consider a non-empty guarded set $M' \subseteq \mathsf{adom}(\prod \mathsf{N})$, such that $\bar{M'}[i]$ is non-total in N_i for all $i \in [k]$. Then there is a non-empty maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ with $M' \subseteq M$. It is not hard to see from the definitions of totality and products, that $\bar{M'}[i]$ being non-total in N_i for all $i \in [k]$ implies that $\bar{M}[i]$ must also be non-total in N_i for all $i \in [k]$, and thus M satisfies Conditions 1 and 2. By Lemma 19 this implies that M' satisfies Conditions 1 and 2, establishing the desired equivalence.

We will now continue to show the statement for guarded sets, rather than maximally guarded sets. The rest of the proof follows the same structure as the proof of Theorem 15, with significant changes in the reasoning explicitly marked by the symbol (†).

"⇒". We show the contrapositive. Assume that there is some non-empty guarded $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, and at least one of the Conditions 1 and 2 is violated. We have to show that (P, N) has a fitting frontier-guarded TGD.

To this end we will construct an FGTGD ρ such that $N \not\models \rho$ for all $N \in \mathsf{N}$, and $J \models \rho$ for all $J \in \mathsf{P}$. Consider

$(\prod \mathsf{N}, \bar{M})$. Note that since M is non-empty, $\prod \mathsf{N}$ cannot be empty, and thus the canonical CQ $q_{(\prod \mathsf{N}, \bar{M})}$ is defined. Then the body of ρ is $\varphi(\bar{x}, \bar{y})$, where the CQ $q_M(\bar{x}) = \exists \bar{y}\, \varphi(\bar{x}, \bar{y})$ is obtained from $q_{(\prod \mathsf{N}, \bar{M})}$ by renaming the free variables. Clearly $\bar{M} \in q_M(\prod \mathsf{N})$, and thus Lemma 12 yields $\bar{M}[i] \in q_M(N_i)$ for every $i \in [n]$.

(†) Since M is guarded, there is a fact in $\prod \mathsf{N}$ that contains every value from $\bar{M}$. Then by definition there exists an atom in q_M that mentions every variable from $\bar{x}$, ensuring that any TGD with body $\varphi(\bar{x}, \bar{y})$ and frontier $\bar{x}$ is frontier-guarded. Concerning the head of ρ, we make a case distinction.

First assume the violation of Condition 1, so $S_M = \emptyset$. Since for every $i \in [k]$, the tuple $\bar{M}[i]$ is non-total in N_i, there exists a CQ $q_i(\bar{x}) = \exists \bar{z}_i\, \psi_i(\bar{x})$, such that $\bar{M}[i] \notin q_i(N_i)$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, where $\psi = \bigwedge_{i \in [k]} \psi_i$, and $\bar{z} = \bigcup_{i \in [k]} \bar{z}_i$. For brevity define $q_H(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Clearly $\bar{M}[i] \notin q_H(N_i)$, and thus $N_i \not\models \rho$ for $i \in [n]$. Now let $J \in \mathsf{P}$. To show that $J \models \rho$, it suffices to argue that $q_M(J) = \emptyset$. Suppose to the contrary that there is a tuple $\bar{a}$ with $\bar{a} \in q_M(J)$. Then there is a homomorphism h from $q_M(\bar{x})$ to $(J, \bar{a})$. By appropriately renaming the elements in the domain of h to values from $\mathsf{adom}(\prod \mathsf{N})$, we obtain a homomorphism witnessing $(\prod \mathsf{N}, \bar{M}) \to (J, \bar{a})$. Hence $(J, \bar{a}) \in S_M$, which contradicts the assumption that $S_M = \emptyset$.

Now assume that Condition 1 is satisfied (and thus S_M is non-empty), but Condition 2 is violated. Let $(K, \bar{c}) = \prod S_M$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, where $\exists \bar{z}\, \psi(\bar{x}, \bar{z})$ is obtained from the CQ $q_{(K^*, \bar{c}^*)}$ by renaming the free variables. For brevity, let $q_\Pi(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Then the violation of Condition 2 by M implies $\bar{M}[i] \notin q_\Pi(N_i)$, and thus $N_i \not\models \rho$ for $i \in [k]$. To see that all positive examples satisfy ρ, let $J \in \mathsf{P}$ and $\bar{a} \in q_M(J)$. Then $(J, \bar{a}) \in S_M$, which by Lemma 11 implies $(K, \bar{c}) \to (J, \bar{a})$. Clearly $(K^*, \bar{c}^*) \to (K, \bar{c})$ and thus $\bar{a} \in q_\Pi(J)$, so $J \models \rho$.

"⇐". Assume that for all non-empty guarded $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, such that $\bar{M}[i]$ is non-total in N_i for $i \in [k]$, Conditions 1 and 2 are satisfied. Further assume, contrary to what we aim to show, that there exists an FGTGD that fits (P, N).

(†) By Lemma 18, this implies the existence of a fitting FGTGD with a non-empty frontier.

Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$ be such a TGD. For brevity, define $q(\bar{x}) = \exists \bar{y}\, \varphi$ and $p(\bar{x}) = \exists \bar{z}\, \psi$. As ρ fits (P, N), it follows that for every $i \in [k]$ there is some $\bar{a}_i \in \mathsf{adom}(N_i)^{|\bar{x}|}$ with $\bar{a}_i \in q(N_i)$ and $\bar{a}_i \notin p(N_i)$, witnessed by a homomorphism $h_{q \to N_i}$ from $q(\bar{x})$ to $(N_i, \bar{a}_i)$. Our goal is to show that for some i, we have $\bar{a}_i \in p(N_i)$, thereby obtaining the desired contradiction. To utilize both conditions, we first need to find an appropriate non-empty guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$.

Consider the instance $(\prod \mathsf{N}, \bar{b})$, where $\bar{b} = \bar{a}_1 \times \cdots \times \bar{a}_k$. Define $h_{q \to \Pi} : \mathsf{adom}(I_q) \to \mathsf{adom}(\prod \mathsf{N})$ by $h_{q \to \Pi}(y) = (h_{q \to N_1}(y), \ldots, h_{q \to N_k}(y))$.

(†) By the definition of products, $h_{q \to \Pi}$ is a homomorphism witnessing that $\bar{b} \in q(\prod \mathsf{N})$. Let $M = \{b \in \mathsf{adom}(\prod \mathsf{N}) \mid b \in \bar{b}\}$. Since ρ has a non-empty frontier and is frontier-guarded, and $h_{q \to \Pi}$ maps $\bar{x}$ to $\bar{b}$, M is non-empty and guarded.

To ensure that M satisfies Conditions 1 and 2, it remains to show that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$. Let $i \in [k]$ and observe that $\bar{b}[i] = \bar{a}_i$, and thus $\bar{b}[i]$ is non-total in N_i because $\bar{a}_i \notin p(N_i)$. As an immediate consequence of the definition of totality, every $M' \subseteq \mathsf{adom}(N_i)$ with $\bar{b}[i] \subseteq M'$ is also non-total in N_i. Note that $(d_1, \ldots, d_k) \in M$ implies $d_i \in \bar{M}[i]$ by definition of products. Therefore $\bar{b} \subseteq M$ implies $\bar{b}[i] \subseteq \bar{M}[i]$ and thus $\bar{M}[i]$ is non-total in N_i. Hence, M satisfies Conditions 1 and 2.

This allows us to invoke Condition 1, which states that the set S_M is non-empty, so let $(P_1, \bar{d}_1), \ldots, (P_m, \bar{d}_m)$ be an enumeration of the pointed instances in S_M, with $m > 0$. Consider $(P_j, \bar{d}_j)$ for some $j \in [m]$. The definition of S_M yields $(\prod \mathsf{N}, \bar{M}) \to (P_j, \bar{d}_j)$. Composing $h_{q \to \Pi}$ with the witnessing homomorphism yields a homomorphism $h_{q \to P_j}$ from q to P_j such that $h_{q \to P_j}(\bar{x}) \subseteq \bar{d}_j$. Recall that ρ is a fitting TGD and therefore $P_j \in \mathsf{P}$ implies $h_{q \to P_j}(\bar{x}) \in p(P_j)$. Hence, there exists a homomorphism $g_{p \to P_j}$ from $p(\bar{x})$ to $(P_j, h_{q \to P_j}(\bar{x}))$.

Let $(K, \bar{c}) = \prod S_M$ and define $g_{p \to K} : \mathsf{adom}(I_p) \to \mathsf{adom}(K)$ by setting $g_{p \to K}(y) = (g_{p \to P_1}(y), \ldots, g_{p \to P_m}(y))$. By construction $g_{p \to K}(\bar{x}) \subseteq \bar{c}$ and by definition of products $g_{p \to K}$ is a homomorphism. Define a homomorphism $g_{p \to K^*}$ from $p(\bar{x})$ to $(K^*, \bar{c}^*)$ by $g_{p \to K^*}(y) = c_i^*$ if $g_{p \to K}(y) = c_i$ and $g_{p \to K^*}(y) = g_{p \to K}(y)$ otherwise. Clearly $g_{p \to K^*}(\bar{x}) \subseteq \bar{c}^*$ and since $g_{p \to K}$ is a homomorphism, the definition of diversification guarantees that $g_{p \to K^*}$ is also a homomorphism.

By Condition 2 we have $(K^*, c^*) \to (N_i, \bar{M}[i])$ for some $i \in [k]$. Composing $g_{p \to K^*}$ with the witnessing homomorphism yields a homomorphism $g_{p \to N_i}$ from p to N_i such that $g_{p \to N_i}(\bar{x}) \subseteq \bar{M}[i]$. We next establish that in fact $g_{p \to N_i}(\bar{x}) = \bar{a}_i$, thus showing $\bar{a}_i \in p(N_i)$ and thereby obtaining a contradiction, as desired. Let $x \in \bar{x}$ and let $\bar{M} = (M_1, \ldots, M_n)$. Observe that if $h_{q \to \Pi}(x) = M_\ell$ for $\ell \in [n]$, then $h_{q \to N_i}(x)$ is the ℓ-th element of $\bar{M}[i]$ due to construction of $h_{q \to \Pi}$. Furthermore the construction of the homomorphisms $h_{q \to P_j}$, $g_{p \to P_j}$, $g_{p \to K}$, $g_{p \to K^*}$ and $g_{p \to N_i}$ all preserve the order of the original mapping of $\bar{x}$ to $\bar{M}$ induced by $h_{q \to \Pi}$. Thus $g_{p \to N_i}(x) = h_{q \to N_i}(x)$, which entails $g_{p \to N_i}(\bar{x}) = \bar{a}_i$, as required. ❏

For F1TGD, we provide a characterization that slightly differs from the one in Theorem 16. Instead of singleton sets $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, it considers values $\bar{a} \in \mathsf{adom}(\prod \mathsf{N})$. Moreover, it uses the interpretation $\prod S_{\bar{a}}$ in the second condition, rather than its diversification, as the latter is redundant for F1TGD.

Theorem 25. *Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \ldots, N_k\}$. Then no F1TGD fits (P, N) if and only if for every $\bar{a} = (a_1, \ldots, a_k) \in \mathsf{adom}(\prod \mathsf{N})$ such that a_i is non-total in N_i for all $i \in [k]$, the following conditions are satisfied:*

1. *the following set is non-empty:*

$$S_{\bar{a}} = \{(J, b) \mid J \in \mathsf{P} \text{ and } b \in \mathsf{adom}(J) \text{ such that } (\prod \mathsf{N}, \bar{a}) \to (J, b)\}$$

2. $\prod S_{\bar{a}} \to (N_i, a_i)$ *for some* $i \in [k]$.

Proof. The proof follows the same structure as the proof of Theorem 15, with significant changes in the reasoning explicitly marked by the symbol (†).

"⇒". We show the contrapositive. Assume that there is some $\bar{a} = (a_1, \ldots, a_k) \in \mathsf{adom}(\prod \mathsf{N})$ such that a_i is non-total in N_i for all $i \in [k]$, and at least one of the Conditions 1 and 2 is violated. We have to show that (P, N) has a fitting frontier-one TGD.

To this end we will construct an F1TGD ρ such that $N \not\models \rho$ for all $N \in \mathsf{N}$, and $J \models \rho$ for all $J \in \mathsf{P}$. Consider $(\prod \mathsf{N}, \bar{a})$.

(†) Note that since $\bar{a} \in \mathsf{adom}(\prod \mathsf{N})$, the canonical CQ $q_{(\prod \mathsf{N}, \bar{a})}$ is defined and has exactly one free variable.

Then the body of ρ is $\varphi(x, \bar{y})$, where the CQ $q_{\bar{a}}(x) = \exists \bar{y}\, \varphi(x, \bar{y})$ is obtained from $q_{(\prod \mathsf{N}, \bar{a})}$ by renaming the free variable. Clearly $\bar{a} \in q_{\bar{a}}(\prod \mathsf{N})$, and thus Lemma 12 yields $a_i \in q_{\bar{a}}(N_i)$ for every $i \in [n]$. Concerning the head of ρ, we make a case distinction.

First assume the violation of Condition 1, so $S_{\bar{a}} = \emptyset$. Since for every $i \in [k]$, the value a_i is non-total in N_i, there exists a CQ $q_i(x) = \exists \bar{z}_i\, \psi_i(x)$, such that $a_i \notin q_i(N_i)$. Let $\rho = \varphi(x, \bar{y}) \to \exists \bar{z}\, \psi(x, \bar{z})$, where $\psi = \bigwedge_{i \in [k]} \psi_i$, and $\bar{z} = \bigcup_{i \in [k]} \bar{z}_i$. For brevity define $q_H(x) = \exists \bar{z}\, \psi(x, \bar{z})$. Clearly $a_i \notin q_H(N_i)$, and thus $N_i \not\models \rho$ for $i \in [n]$. Now let $J \in \mathsf{P}$. To show that $J \models \rho$, it suffices to argue that $q_{\bar{a}}(J) = \emptyset$. Suppose to the contrary that there is a value b with $b \in q_{\bar{a}}(J)$. Then there is a homomorphism h from $q_{\bar{a}}(x)$ to (J, b). By appropriately renaming the elements in the domain of h to values from $\mathsf{adom}(\prod \mathsf{N})$, we obtain a homomorphism witnessing $(\prod \mathsf{N}, \bar{a}) \to (J, b)$. Hence $(J, b) \in S_{\bar{a}}$, which contradicts the assumption that $S_{\bar{a}} = \emptyset$.

Now assume that Condition 1 is satisfied (and thus $S_{\bar{a}}$ is non-empty), but Condition 2 is violated. Let $(K, c) = \prod S_{\bar{a}}$. Let $\rho = \varphi(x, \bar{y}) \to \exists \bar{z}\, \psi(x, \bar{z})$, where $\exists \bar{z}\, \psi(x, \bar{z})$ is obtained from the CQ $q_{(K,c)}$ by renaming the free variable. For brevity, let $q_\Pi(x) = \exists \bar{z}\, \psi(x, \bar{z})$. Then the violation of Condition 2 by M implies $a_i \notin q_\Pi(N_i)$, and thus $N_i \not\models \rho$ for $i \in [k]$. To see that all positive examples satisfy ρ, let $J \in \mathsf{P}$ and $b \in q_{\bar{a}}(J)$. Then $(J, b) \in S_{\bar{a}}$, which by Lemma 11 implies $(K, c) \to (J, b)$. It follows that $b \in q_\Pi(J)$, and thus $J \models \rho$.

"⇐". Assume that for all $\bar{a} = (a_1, \ldots, a_k) \in \mathsf{adom}(\prod \mathsf{N})$, such that a_i is non-total in N_i for $i \in [k]$, Conditions 1 and 2 are satisfied. Further assume, contrary to what we aim to show, that there exists an F1TGD that fits (P, N).

(†) By Lemma 18, this implies the existence of a fitting F1TGD with a frontier of size exactly 1.

Let $\rho = \varphi(x, \bar{y}) \to \exists \bar{z}\, \psi(x, \bar{z})$ be such a TGD. For brevity, define $q(x) = \exists \bar{y}\, \varphi$ and $p(x) = \exists \bar{z}\, \psi$. As ρ fits (P, N), it follows that for every $i \in [k]$ there is some $a_i \in \mathsf{adom}(N_i)$ with $a_i \in q(N_i)$ and $a_i \notin p(N_i)$, witnessed by a homomorphism $h_{q \to N_i}$ from $q(x)$ to (N_i, a_i). Our goal is to show that for some i, we have $a_i \in p(N_i)$, thereby obtaining the desired contradiction. To utilize both conditions, we first need to find an appropriate $\bar{a} \in \mathsf{adom}(\prod \mathsf{N})$.

Consider the instance $(\prod \mathsf{N}, \bar{a})$, where $\bar{a} = (a_1, \ldots, a_k)$. Define $h_{q \to \Pi} : \mathsf{adom}(I_q) \to \mathsf{adom}(\prod \mathsf{N})$ by $h_{q \to \Pi}(y) = (h_{q \to N_1}(y), \ldots, h_{q \to N_k}(y))$. By the definition of products, $h_{q \to \Pi}$ is a homomorphism witnessing that $\bar{a} \in q(\prod \mathsf{N})$.

To ensure that $\bar{a}$ satisfies Conditions 1 and 2, it remains to show that a_i is non-total in N_i for all $i \in [k]$. However, for every $i \in [k]$, this is an immediate consequence of $a_i \notin p(N_i)$, and thus $\bar{a}$ indeed satisfies Conditions 1 and 2.

This allows us to invoke Condition 1, which states that the set $S_{\bar{a}}$ is non-empty, so let $(P_1, d_1), \ldots, (P_m, d_m)$ be an enumeration of the pointed instances in $S_{\bar{a}}$, with $m > 0$. Consider (P_j, d_j) for some $j \in [m]$. The definition of $S_{\bar{a}}$ yields $(\prod \mathsf{N}, \bar{a}) \to (P_j, d_j)$. Composing $h_{q \to \Pi}$ with the witnessing homomorphism yields a homomorphism $h_{q \to P_j}$ from q to P_j such that $h_{q \to P_j}(x) = d_j$. Recall that ρ is a fitting TGD and therefore $P_j \in \mathsf{P}$ implies $d_j \in p(P_j)$. Hence, there exists a homomorphism $g_{p \to P_j}$ from $p(x)$ to (P_j, d_j).

Let $(K, c) = \prod S_{\bar{a}}$ and define $g_{p \to K} : \mathsf{adom}(I_p) \to \mathsf{adom}(K)$ by setting $g_{p \to K}(y) = (g_{p \to P_1}(y), \ldots, g_{p \to P_m}(y))$. By construction $g_{p \to K}(x) = c$ and by definition of products $g_{p \to K}$ is a homomorphism.

By Condition 2 we have $(K, c) \to (N_i, a_i)$ for some $i \in [k]$. Composing $g_{p \to K}$ with the witnessing homomorphism yields a homomorphism $g_{p \to N_i}$ from p to N_i such that $g_{p \to N_i}(x) = a_i$, thus showing $a_i \in p(N_i)$, a contradiction. ❑

Theorem 26. *Let* (P, N) *be a fitting instance where* $\mathsf{N} = \{N_1, \ldots, N_k\}$. *Then no TGD fits* (P, N) *if and only if for every non-empty set* $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ *such that* $\bar{M}[i]$ *is non-total in* N_i *for all* $i \in [k]$, *the following conditions are satisfied:*

1. *the following set is non-empty:*

$$S_M = \{(J, \bar{b}) \mid J \in \mathsf{P} \text{ and } \bar{b} \in \mathsf{adom}(J)^{|M|} \text{ such that } (\prod \mathsf{N}, \bar{M}) \to (J, \bar{b})\}$$

2. $(K^*, \bar{c}^*) \to (N_i, \bar{M}[i])$ *for some* $i \in [k]$ *where* $(K, \bar{c}) = \prod S_M$.

Proof. The proof follows the same structure as the proof of Theorem 15, with significant changes in the reasoning explicitly marked by the symbol (†).

"⇒". We show the contrapositive. Assume that there is some non-empty $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, and at least one of the Conditions 1 and 2 is violated. We have to show that (P, N) has a fitting TGD.

To this end we will construct a TGD ρ such that $N \not\models \rho$ for all $N \in \mathsf{N}$, and $J \models \rho$ for all $J \in \mathsf{P}$. Consider $(\prod \mathsf{N}, \bar{M})$. Note that since M is non-empty, $\prod \mathsf{N}$ cannot be empty, and thus the canonical CQ $q_{(\prod \mathsf{N}, \bar{M})}$ is defined. Then the body of ρ is $\varphi(\bar{x}, \bar{y})$, where the CQ $q_M(\bar{x}) = \exists \bar{y}\, \varphi(\bar{x}, \bar{y})$ is obtained from $q_{(\prod \mathsf{N}, \bar{M})}$ by renaming the free variables. Clearly $\bar{M} \in q_M(\prod \mathsf{N})$, and thus Lemma 12 yields $\bar{M}[i] \in q_M(N_i)$ for

every $i \in [n]$. Concerning the head of ρ, we make a case distinction.

First assume the violation of Condition 1, so $S_M = \emptyset$. Since for every $i \in [k]$, the tuple $\bar{M}[i]$ is non-total in N_i, there exists a CQ $q_i(\bar{x}) = \exists \bar{z}_i\, \psi_i(\bar{x})$, such that $\bar{M}[i] \notin q_i(N_i)$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, where $\psi = \bigwedge_{i\in[k]} \psi_i$, and $\bar{z} = \bigcup_{i\in[k]} \bar{z}_i$. For brevity define $q_H(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Clearly $\bar{M}[i] \notin q_H(N_i)$, and thus $N_i \not\models \rho$ for $i \in [n]$. Now let $J \in \mathsf{P}$. To show that $J \models \rho$, it suffices to argue that $q_M(J) = \emptyset$. Suppose to the contrary that there is a tuple $\bar{a}$ with $\bar{a} \in q_M(J)$. Then there is a homomorphism h from $q_M(\bar{x})$ to $(J, \bar{a})$. By appropriately renaming the elements in the domain of h to values from $\mathsf{adom}(\prod \mathsf{N})$, we obtain a homomorphism witnessing $(\prod \mathsf{N}, \bar{M}) \to (J, \bar{a})$. Hence $(J, \bar{a}) \in S_M$, which contradicts the assumption that $S_M = \emptyset$.

Now assume that Condition 1 is satisfied (and thus S_M is non-empty), but Condition 2 is violated. Let $(K, \bar{c}) = \prod S_M$. Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$, where $\exists \bar{z}\, \psi(\bar{x}, \bar{z})$ is obtained from the CQ $q_{(K^*, \bar{c}^*)}$ by renaming the free variables. For brevity, let $q_\Pi(\bar{x}) = \exists \bar{z}\, \psi(\bar{x}, \bar{z})$. Then the violation of Condition 2 by M implies $\bar{M}[i] \notin q_\Pi(N_i)$, and thus $N_i \not\models \rho$ for $i \in [k]$. To see that all positive examples satisfy ρ, let $J \in \mathsf{P}$ and $\bar{a} \in q_M(J)$. Then $(J, \bar{a}) \in S_M$, which by Lemma 11 implies $(K, \bar{c}) \to (J, \bar{a})$. Clearly $(K^*, \bar{c}^*) \to (K, \bar{c})$ and thus $\bar{a} \in q_\Pi(J)$, so $J \models \rho$.

"$\Leftarrow$". Assume that for all $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, such that $\bar{M}[i]$ is non-total in N_i for $i \in [k]$, Conditions 1 and 2 are satisfied. Further assume, contrary to what we aim to show, that there exists a TGD that fits (P, N).

(†) By Lemma 18, this implies the existence of a fitting TGD with a non-empty frontier.

Let $\rho = \varphi(\bar{x}, \bar{y}) \to \exists \bar{z}\, \psi(\bar{x}, \bar{z})$ be such a TGD. For brevity, define $q(\bar{x}) = \exists \bar{y}\, \varphi$ and $p(\bar{x}) = \exists \bar{z}\, \psi$. As ρ fits (P, N), it follows that for every $i \in [k]$ there is some $\bar{a}_i \in \mathsf{adom}(N_i)^{|\bar{x}|}$ with $\bar{a}_i \in q(N_i)$ and $\bar{a}_i \notin p(N_i)$, witnessed by a homomorphism $h_{q\to N_i}$ from $q(\bar{x})$ to $(N_i, \bar{a}_i)$. Our goal is to show that for some i, we have $\bar{a}_i \in p(N_i)$, thereby obtaining the desired contradiction. To utilize both conditions, we first need to find an appropriate non-empty set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$.

Consider the instance $(\prod \mathsf{N}, \bar{b})$, where $\bar{b} = \bar{a}_1 \times \cdots \times \bar{a}_k$. Define $h_{q\to\Pi} : \mathsf{adom}(I_q) \to \mathsf{adom}(\prod \mathsf{N})$ by $h_{q\to\Pi}(y) = (h_{q\to N_1}(y), \ldots, h_{q\to N_k}(y))$.

(†) By the definition of products, $h_{q\to\Pi}$ is a homomorphism witnessing that $\bar{b} \in q(\prod \mathsf{N})$. Let $M = \{b \in \mathsf{adom}(\prod \mathsf{N}) \mid b \in \bar{b}\}$. Since ρ has a non-empty frontier, M is non-empty.

To ensure that M satisfies Conditions 1 and 2, it remains to show that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$. Let $i \in [k]$ and observe that $\bar{b}[i] = \bar{a}_i$, and thus $\bar{b}[i]$ is non-total in N_i because $\bar{a}_i \notin p(N_i)$. As an immediate consequence of the definition of totality, every $M' \subseteq \mathsf{adom}(N_i)$ with $\bar{b}[i] \subseteq M'$ is also non-total in N_i. Note that $(d_1, \ldots, d_k) \in M$ implies $d_i \in \bar{M}[i]$ by definition of products. Therefore $\bar{b} \subseteq M$ implies $\bar{b}[i] \subseteq \bar{M}[i]$ and thus $\bar{M}[i]$ is non-total in N_i. Hence, M satisfies Conditions 1 and 2.

This allows us to invoke Condition 1, which states that the set S_M is non-empty, so let $(P_1, \bar{d}_1), \ldots, (P_m, \bar{d}_m)$ be an enumeration of the pointed instances in S_M, with $m > 0$. Consider $(P_j, \bar{d}_j)$ for some $j \in [m]$. The definition of S_M yields $(\prod \mathsf{N}, \bar{M}) \to (P_j, \bar{d}_j)$. Composing $h_{q\to\Pi}$ with the witnessing homomorphism yields a homomorphism $h_{q\to P_j}$ from q to P_j such that $h_{q\to P_j}(\bar{x}) \subseteq \bar{d}_j$. Recall that ρ is a fitting TGD and therefore $P_j \in \mathsf{P}$ implies $h_{q\to P_j}(\bar{x}) \in p(P_j)$. Hence, there exists a homomorphism $g_{p\to P_j}$ from $p(\bar{x})$ to $(P_j, h_{q\to P_j}(\bar{x}))$.

Let $(K, \bar{c}) = \prod S_M$ and define $g_{p\to K} : \mathsf{adom}(I_p) \to \mathsf{adom}(K)$ by setting $g_{p\to K}(y) = (g_{p\to P_1}(y), \ldots, g_{p\to P_m}(y))$. By construction $g_{p\to K}(\bar{x}) \subseteq \bar{c}$ and by definition of products $g_{p\to K}$ is a homomorphism. Define a homomorphism $g_{p\to K^*}$ from $p(\bar{x})$ to $(K^*, \bar{c}^*)$ by $g_{p\to K^*}(y) = c_i^*$ if $g_{p\to K}(y) = c_i$ and $g_{p\to K^*}(y) = g_{p\to K}(y)$ otherwise. Clearly $g_{p\to K^*}(\bar{x}) \subseteq \bar{c}^*$ and since $g_{p\to K}$ is a homomorphism, the definition of diversification guarantees that $g_{p\to K^*}$ is also a homomorphism.

By Condition 2 we have $(K^*, c^*) \to (N_i, \bar{M}[i])$ for some $i \in [k]$. Composing $g_{p\to K^*}$ with the witnessing homomorphism yields a homomorphism $g_{p\to N_i}$ from p to N_i such that $g_{p\to N_i}(\bar{x}) \subseteq \bar{M}[i]$. We next establish that in fact $g_{p\to N_i}(\bar{x}) = \bar{a}_i$, thus showing $\bar{a}_i \in p(N_i)$ and thereby obtaining a contradiction, as desired. Let $x \in \bar{x}$ and let $\bar{M} = (M_1, \ldots, M_n)$. Observe that if $h_{q\to\Pi}(x) = M_\ell$ for $\ell \in [n]$, then $h_{q\to N_i}(x)$ is the ℓ-th element of $\bar{M}[i]$ due to construction of $h_{q\to\Pi}$. Furthermore the construction of the homomorphisms $h_{q\to P_j}$, $g_{p\to P_j}$, $g_{p\to K}$, $g_{p\to K^*}$ and $g_{p\to N_i}$ all preserve the order of the original mapping of $\bar{x}$ to $\bar{M}$ induced by $h_{q\to\Pi}$. Thus $g_{p\to N_i}(x) = h_{q\to N_i}(x)$, which entails $g_{p\to N_i}(\bar{x}) = \bar{a}_i$, as required. ❏

Theorem 17. *Let (P, N) be a fitting instance where $\mathsf{N} = \{N_1, \ldots, N_n\}$. Then no FullTGD fits (P, N) if and only if, for all relation symbols $R_1, \ldots, R_n$ and tuples $\bar{a}_1, \ldots, \bar{a}_n$ such that*

$$\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)} \text{ and } R_i(\bar{a}_i[i]) \notin N_i \quad \text{for } i \in [n],$$

there is a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P such that $R_j(h(\bar{a}_j)) \notin P$ for some $j \in [n]$.

Moreover, if (P, N) admits a fitting FullTGD, then it admits one in which the number of head atoms is bounded by the number of examples in N.

Proof. "$\Rightarrow$". We show the contrapositive. Suppose there exist $R_1, \ldots, R_n$ and $\bar{a}_1, \ldots, \bar{a}_n$ with $\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)}$ and $R_i(\bar{a}_i[i]) \notin N_i$ for $i \in [n]$, and such that for all $P \in \mathsf{P}$ and h witnessing $\prod \mathsf{N} \to P$, we have $R_1(h(\bar{a}_1)), \ldots, R_n(h(\bar{a}_n)) \in P$. We need to show that (P, N) has a fitting FullTGD.

We will construct a FullTGD ρ such that $N \not\models \rho$ for all $N \in \mathsf{N}$ and $P \models \rho$ for all $P \in \mathsf{P}$. Define $\bar{a} = \bigcup_{i\in[n]} \bar{a}_i$ and

$$\rho = \varphi(\bar{x}, \bar{y}) \to R_1(\bar{x}_1) \wedge \cdots \wedge R_n(\bar{x}_n) \text{ with } \bar{x} = \bigcup_{i\in[n]} \bar{x}_i,$$

where $q_{\bar{a}}(\bar{x}) = \exists \bar{y}\, \varphi(\bar{x}, \bar{y})$ is obtained from the canonical CQ of $(\prod \mathsf{N}, \bar{a})$ by renaming the free variables. Then $\bar{a} \in q_{\bar{a}}(\prod \mathsf{N})$, which yields $\bar{a}[i] \in q_{\bar{a}}(N_i)$ for all $i \in [n]$ by Lemma 12. Since by assumption $R_i(\bar{a}_i[i]) \notin N_i$, it follows that $N_i \not\models \varphi(\bar{x}, \bar{y}) \to R_i(\bar{x}_i)$, which implies $N_i \not\models \rho$ for

$i \in [n]$. It remains to show that $P \models \rho$ for every $P \in \mathsf{P}$. Let $P \in \mathsf{P}$.

Case 1. Assume $\prod \mathsf{N} \to P$, witnessed by h with $h(\bar{a}) = \bar{b}$. This entails $R_1(h(\bar{a}_1)), \ldots, R_n(h(\bar{a}_n)) \in P$ by assumption, hence $P \models \rho$.

Case 2. No homomorphism from $\prod \mathsf{N}$ to P exists. Then $P \not\models q_{\bar{a}}$, and therefore $P \models \rho$ holds vacuously. Thus, ρ is a fitting FullTGD for (P, N).

"$\Leftarrow$". Assume that for all $R_1, \ldots, R_n$ and $\bar{a}_1, \ldots, \bar{a}_n$ with $\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)}$ and $R_i(\bar{a}_i[i]) \notin N_i$ for $i \in [n]$, there exists a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P, such that $R_j(h(\bar{a}_j)) \notin P$ for some $j \in [n]$. We show that (P, N) has no fitting FullTGD. Towards a contradiction, suppose there is a fitting FullTGD

$$\rho = \varphi(\bar{x}, \bar{y}) \to \psi(\bar{x}).$$

For brevity, set $q(\bar{x}) = \exists \bar{y}\, \varphi$. Since ρ is a fitting FullTGD, there is a homomorphism h_i witnessing $(I_q, \bar{x}) \to (N_i, h_i(\bar{x}))$ and an atom $R_i(\bar{x}_i)$ in ψ such that $R_i(h_i(\bar{x}_i)) \notin N_i$ for every $i \in [n]$.

From the atoms $R_i(\bar{x}_i)$ and homomorphisms h_i we construct facts $\bar{a}_1, \ldots, \bar{a}_n$ with $\bar{a}_j \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_j)}$ and $R_j(\bar{a}_j[j]) \notin N_j$ for $j \in [n]$. Let

$$\bar{a}_i = h_1(\bar{x}_i) \times \cdots \times h_n(\bar{x}_i) \quad \text{for } i \in [n].$$

Fix an $i \in [n]$. To see that indeed $\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)}$, let $\bar{a}_i = (\bar{b}_1, \ldots, \bar{b}_{|\bar{a}_i|})$ and let $k \in [|\bar{a}_i|]$. We will show that $\bar{b}_k \in \mathsf{adom}(\prod \mathsf{N})$. Observe that for x in position k of $\bar{x}_i$, $q(\bar{x})$ must contain an atom $S(\bar{z})$ with $\bar{z} \subseteq \bar{x}$, and $x \in \bar{z}$, because $\bar{x}_i \subseteq \bar{x}$. Then $h_j(\bar{x}) \in q(N_j)$ implies $S(h_j(\bar{z})) \in N_j$ for $j \in [n]$. Thus, by definition of products, $\prod \mathsf{N}$ contains the fact $S(h_1(\bar{z}) \times \cdots \times h_n(\bar{z}))$, which has $(h_1(x), \ldots, h_n(x)) = \bar{b}_k$ as one of its values. Therefore $\bar{b}_k \in \mathsf{adom}(\prod N)$, as required.

Since $\bar{a}_i[i] = h_i(\bar{x}_i)$, we immediately obtain $R_i(\bar{a}_i[i]) \notin N_i$ for $i \in [n]$. Therefore, we may apply the assumption to $R_1, \ldots, R_n$ and $\bar{a}_1, \ldots, \bar{a}_n$, which yields a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P, such that $R_j(h(\bar{a}_j)) \notin P$ for some $j \in [n]$. Define

$$\bar{a} = h_1(\bar{x}) \times \cdots \times h_n(\bar{x}).$$

For every $i \in [n]$, $(I_q, \bar{x}) \to (N_i, h_i(\bar{x}))$ implies $h_i(\bar{x}) \in q(N_i)$. We immediately obtain $\bar{a} \in q(\prod \mathsf{N})$ by Lemma 12. Then there exists a homomorphism g witnessing $(I_q, \bar{x}) \to (\prod \mathsf{N}, \bar{a})$. Thus, $g \circ h$ is a witness to $(I_q, \bar{x}) \to (P, h(\bar{a}))$, and therefore $h(\bar{a}) \in q(P)$. By definition $h(\bar{a}_j) \subseteq h(\bar{a})$, and since R_j is an atom in ψ, $R_j(h(\bar{a}_j)) \notin P$ implies $P \not\models \rho$, the desired contradiction.

The claimed bound on the number of atoms in the head of the fitting FullTGD by the number of examples in N follows directly from the constructed FullTGD in the "$\Rightarrow$" direction of the proof. ❑

Theorem 18.

1. *For $\mathcal{L} \in \{GTGD, FGTGD, F1TGD\}$, fitting $\mathcal{L}$-ontology existence and fitting $\mathcal{L}$-TGD existence are in* coNExpTime.
2. *For full TGDs, fitting ontology existence is in Σ_2^p (and in* coNP *if the arities of relation symbols are bounded by a constant), and fitting TGD existence is in* coNExpTime.
3. *For unrestricted TGDs, fitting ontology existence is in* co2NExpTime *and fitting TGD existence is in* co3NExpTime.

Proof. We start with $\mathcal{L} \in \{\text{GTGD}, \text{FGTGD}, \text{F1TGD}\}$ and concentrate on fitting TGD existence since Lemma 3 yields a simple reduction from fitting ontology existence to fitting TGD existence that gives the desired results.

We start with fitting GTGD existence, using the characterization from Theorem 15 to argue that fitting non-existence is in NExpTime.

To decide whether a given fitting instance (P, N) with $\mathsf{N} = \{N_1, \ldots, N_k\}$ admits no fitting GTGD, we have to check whether for every maximally guarded set $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ such that $\bar{M}[i]$ is non-total in N_i for all $i \in [k]$, Conditions 1 and 2 of Theorem 15 are satisfied.

We first analyze the cardinality of the sets S_M defined in Condition 1 and of the products $\prod S_M$ used in Condition 2. By definition, the cardinality of the sets S_M is bounded by $\sum_{P \in \mathsf{P}} |P|$. This is because M is a maximally guarded set in $\prod \mathsf{N}|_M$ and thus $(J, \bar{b}) \in S_M$ implies that J contains a fact that contains all values from $\bar{b}$ and no other values. Consequently, the product $\prod S_M$ is of at most single exponential size.

We also observe that each set S_M can be computed in single exponential time. To see this first note that each instance $\prod \mathsf{N}|_M$ contains at most linearly many values simply because M is a maximally guarded set. We can thus decide in single exponential time whether $(\prod \mathsf{N}|_M, \bar{M}) \to (J, \bar{b})$ by a brute force algorithm.

Making use of the above observations and of the fact that the existence of a homomorphism can be decided in NP, it is easy to derive a NExpTime algorithm. We iterate over all relevant sets $M \subseteq \mathsf{adom}(\prod \mathsf{N})$, of which there are only single exponentially many because these sets must be maximally guarded, and then check Conditions 1 and 2. Condition 1 can be checked deterministically in single exponential time. For Condition 2, we can guess and check the required homomorphism. We remark that, despite the use of diversification, the structure $(K^*, \bar{c}^*)$ is still of single exponential size.

The frontier-guarded case is very similar. A noteworthy difference is that in Condition 1, we now need to decide the existence of a homomorphism from $(\prod \mathsf{N}, \bar{M})$ to $(J, \bar{b})$, rather than from $(\prod \mathsf{N}|_M, \bar{M})$. However, the structure $\prod \mathsf{N}$ has single exponentially many values rather than linearly many, and thus a brute force approach no longer works. We solve this problem as follows:

1. we observe that the characterization still holds if Condition 2 is rephrased as follows:

   2′. $(K^*, \bar{c}^*) \to (N_i, \bar{M}[i])$ for some $i \in [k]$ and with $(K, \bar{c}) = \prod S'_M$ for some non-empty $S'_M \subseteq S_M$.

   In fact, Conditions 2 and 2′ are easily seen to be equivalent.
2. to verify Condition 1, guess $J \in \mathsf{P}$ and $\bar{b} \in \mathsf{adom}(J)^{|M|}$ and a homomorphism from $(\prod \mathsf{N}, \bar{M})$ to $(J, \bar{b})$

3. to verify Condition 2, guess a set S'_M of pairs $(J, \bar{b})$ with $J \in \mathsf{P}$ and $\bar{b} \in \mathsf{adom}(J)^{|M|}$, verify that $S'_M \subseteq S_M$ by guessing a homomorphism from $(\prod \mathsf{N}, \bar{M})$ to $(J, \bar{b})$ for each $(J, \bar{b}) \in S'_M$, and finally guess an $i \in [k]$ and a homomorphism from $(K^*, \bar{c}^*) \to (N_i, \bar{M}[i])$ where $(K, \bar{c}) = \prod S'_M$.

The case of frontier-one TGDs is identical to that of frontier-guarded TGDs. The only difference is that the cardinality of the sets S_M is bounded by $\sum_{P \in \mathsf{P}} |\mathsf{adom}(P)|$, since only singleton sets M need to be considered.

For unrestricted TGDs, the algorithm exactly parallels the one for GTGDs, that is, the modifications for frontier-guarded TGDs and frontier-one TGDs are not needed. Let us verify that this algorithm yields a CO3NEXPTIME upper bound. The sets $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ that we iterate over in the outermost loop are no longer maximally guarded. There are therefore double exponentially many such sets. The cardinality of S_M is at most double exponential rather than polynomial as in the case of GTGDs, the reason also being that the sets M are no longer maximally guarded. The product $\prod S_M$ in Condition 2 is thus of triple exponential size, which explains the CO3NEXPTIME upper bound.

For fitting ontology existence in place of fitting TGD existence, the same algorithm yields a CO2NEXPTIME upper bound when combined with Lemma 3, which allows us to reduce fitting ontology existence to fitting TGD existence with a single negative example. In that case, there are only single exponentially many sets $M \subseteq \mathsf{adom}(\prod \mathsf{N})$ to iterate over in the outermost loop. Moreover, the cardinality of S_M and the size of $\prod S_M$ drops by one exponential. As a consequence, we can compute $(K^*, \bar{c}^*)$ in double exponential time and guess the homomorphism in Condition 2.

We now turn to fitting FullTGD existence and show that it is in CONEXPTIME. We use the characterization from Theorem 17 for fitting FullTGD non-existence.

The theorem states that no FullTGD fits (P, N) with $\mathsf{N} = \{N_1, \ldots, N_n\}$ if and only if for all choices of relation symbols $R_1, \ldots, R_n$ and tuples $\bar{a}_1, \ldots, \bar{a}_n$ such that $\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)}$ and $R_i(\bar{a}_i[i]) \notin N_i$ for all $i \in [n]$, there exists a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P such that $R_j(h(\bar{a}_j)) \notin P$ for some $j \in [n]$.

First, we analyze the size of the objects involved. The product instance $\prod \mathsf{N}$ can be of exponential size. The arity of the occurring relation symbols can grow with the input, so for a relation symbol R, the number of possible tuples $\bar{a} \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R)}$ is exponential in the worst case. Since $n = |\mathsf{N}|$, the number of combinations of relation symbols $R_1, \ldots, R_n$ and tuples $\bar{a}_1, \ldots, \bar{a}_n$ that need to be considered, is bounded exponentially. This allows us to iterate over all combinations in single exponential time.

For each combination, we need to verify that there is a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P, such that $R_i(h(\bar{a}_i)) \notin P$ for some R_i and $\bar{a}_i$ with $i \in [n]$. We can guess an R_i, a $P \in \mathsf{P}$, and a homomorphism h witnessing $\prod \mathsf{N} \to P$. This puts us in NEXPTIME because $\prod \mathsf{N}$ is of exponential size and guessing a homomorphism is in NP. Verification of $R_i(h(\bar{a}_i)) \notin P$ clearly is possible in polynomial time. Exponentially many such guesses can be done in NEXPTIME, which is therefore the overall complexity. Since this analysis covers the complement of FullTGD existence, the latter is decidable in CONEXPTIME.

We show that fitting FullTGD-ontology existence is decidable in Σ^p_2. By Lemma 3 it suffices to analyze fitting FullTGD existence for (P, N) with $|\mathsf{N}| = 1$. We again use the negation of the characterization. It states that an FullTGD fits (P, N) with $\mathsf{N} = \{N\}$ if and only if there is a relation symbol R and a tuple $\bar{a} \in \mathsf{adom}(N)^{\mathsf{ar}(R)}$ with $R(\bar{a}) \notin N$, and such that for all $P \in \mathsf{P}$ and $\bar{b} \in \mathsf{adom}(P)^{\mathsf{ar}(R)}$, if $(N, \bar{a}) \to (P, \bar{b})$, then $R(\bar{b}) \in P$. We can non-deterministically guess in polynomial time an R and $\bar{a} \in \mathsf{adom}(N)^{\mathsf{ar}(R)}$ with $R(\bar{a}) \notin N$, and then use a CONP oracle to verify that all $P \in \mathsf{P}$ and $\bar{b} \in \mathsf{adom}(P)^{\mathsf{ar}(R)}$ with $(N, \bar{a}) \to (P, \bar{b})$ imply $R(\bar{b}) \in P$. This puts fitting FullTGD-ontology existence in Σ^p_2.

If the arity of the relation symbols occurring in (P, N) is bounded by a constant, fitting FullTGD-ontology existence can be decided in CONP. To see that, observe that the number of tuples $\bar{a}$ with $R(\bar{a}) \notin N$ is polynomial in this case. This implies that fitting non-existence is in NP, as a polynomial-sized certificate can specify a witnessing homomorphism for each of the polynomially many $(R, \bar{a})$. Thus, fitting FullTGD-ontology existence with bounded arity is in CONP. ❑

Theorem 20.

1. *Let $\mathcal{L} \in \{GTGD, FGTGD, F1TGD, TGD\}$. Then fitting $\mathcal{L}$-TGD existence and fitting $\mathcal{L}$-ontology existence are* CONEXPTIME*-hard.*
2. *For full TGDs, fitting TGD existence is* CONEXPTIME*-hard and fitting ontology existence is DP-hard (and* CONP*-hard if the arities of relation symbols are bounded by a constant).*

Proof. We prove all CONEXPTIME-hardness-results by reducing the *product homomorphism problem (PHP)*. The product homomorphism problem takes as input finite instances $I_1, \ldots, I_n, J$ over the same schema and asks, whether $\prod_{i \in [n]} I_i \to J$. This problem is known to be NEXPTIME-hard, even over a fixed schema consisting only of a binary relation symbol E (ten Cate and Dalmau 2015).

We start with Point 1. For $\mathcal{L} \in \{\mathrm{GTGD}, \mathrm{FGTGD}, \mathrm{F1TGD}, \mathrm{TGD}\}$, we provide a polynomial time reduction from the product homomorphism problem to the complement of fitting $\mathcal{L}$-ontology existence and to the complement of fitting $\mathcal{L}$-TGD existence. We will use fitting instances with a single negative example, where by Lemma 3, the existence of fitting ontologies and fitting TGDs coincides.

Let $I_1, \ldots, I_n, J$ be an input to the product homomorphism problem. We assume w.l.o.g. that J has a domain that is disjoint from those of $I_1, \ldots, I_n$. We construct a fitting instance (P, N) as follows. Take a fresh value a and for every $b \in \mathsf{adom}(J) \cup \{a\}$ a fresh unary relation symbol R_b. Define instances

$$J' = J \cup \{R_b(b) \mid b \in \mathsf{adom}(J) \cup \{a\}\}$$

and

$$I'_i = I_i \cup J' \text{ for every } i \in [n].$$

Then set $\mathsf{P} = \{I'_i \mid i \in [n]\}$ and $\mathsf{N} = \{J'\}$. We show that $\prod_{i\in[n]} I_i \to J$ if and only if (P, N) has no fitting $\mathcal{L}$-TGD, based on the characterization given for fitting $\mathcal{L}$-TGD existence.

First, let $\mathcal{L} =$ GTGD. Then by Theorem 15 (P, N) has no fitting GTGD if and only if for every non-empty maximally guarded set $M \subseteq \mathsf{adom}(J')$ such that $\bar{M}$ is non-total in J', the two conditions in Theorem 15 are satisfied.

Observe that $\bar{M}$ is non-total in J' for every non-empty maximally guarded $M \subseteq \mathsf{adom}(J')$, as a result of the newly introduced relation symbols R_b and the fresh value a.[4] We now show that the first condition of Theorem 15 is satisfied for every such M. Let $M \subseteq \mathsf{adom}(J')$ be non-empty and maximally guarded. We have to argue that

$$S_M = \{(I', \bar{b}) \mid I' \in \mathsf{P} \text{ and } \bar{b} \in \mathsf{adom}(I')^{|M|} \text{ such that } (J'|_M, \bar{M}) \to (I', \bar{b})\}$$

is non-empty. In fact, we have that

$$S_M = \{(I'_i, \bar{M}) \mid i \in [n]\}.$$

The "$\supseteq$" direction holds since $J' \subseteq I'_i$ and the "$\subseteq$" direction is due to the use of the fresh relation symbols R_b.

Therefore it suffices to show that $\prod_{i\in[n]} I_i \to J$ if and only if the second condition of Theorem 15 is satisfied by (P, N), for every non-empty maximally guarded $M \subseteq \mathsf{adom}(J')$. That is:

Claim. $\prod_{i\in[n]} I_i \to J$ if and only if $(K^*, \bar{c}^*) \to (J', \bar{M})$, where $(K, \bar{c}) = \prod S_M$, for every non-empty maximally guarded $M \subseteq \mathsf{adom}(J')$.

"$\Rightarrow$". Assume that there exists a homomorphism h from $\prod_{i\in[n]} I_i$ to J and let $M \subseteq \mathsf{adom}(J')$ be non-empty and maximally guarded and $\prod S_M = (K, \bar{c})$. We have $(K^*, \bar{c}^*) \to (K, \bar{c})$. Therefore it suffices to show that h can be extended to a homomorphism h' that witnesses $(K, \bar{c}) \to (J', \bar{M})$. To this end, let h' be the extension of h with

$$h'((b_1, \ldots, b_n)) = b_i$$

for every $(b_1, \ldots, b_n) \in \mathsf{adom}(K) \setminus \mathsf{adom}(\prod_{i\in[n]} I_i)$, where $i \in [n]$ is smallest with $b_i \in \mathsf{adom}(J')$. Note that such an i must exist, since otherwise $(b_1, \ldots, b_n) \in \mathsf{adom}(\prod_{i\in[n]} I_i)$.

It remains to show that h' is a homomorphism. Thus let $R(\bar{d}_1, \ldots, \bar{d}_m) \in K$. If $\bar{d}_1, \ldots, \bar{d}_m \in \mathsf{adom}(\prod_{i\in[n]} I_i)$, then $R(h'(\bar{d}_1), \ldots, h'(\bar{d}_m)) = R(h(\bar{d}_1), \ldots, h(\bar{d}_m)) \in J \subseteq J'$. Otherwise, there is some $\bar{d}_i$ that contains a value from $\mathsf{adom}(J')$. Let k be smallest such that $\bar{d}_i$ contains such a value in its k-th component. Now let $d^k_1, \ldots, d^k_m$ be the k-th components of $\bar{d}_1, \ldots, \bar{d}_m$. By definition of products, $R(\bar{d}_1, \ldots, \bar{d}_m) \in K$ implies that $R(d^k_1, \ldots, d^k_m) \in I'_i$. But I'_i is the disjoint union of I_i and J', and thus $d^k_i \in \mathsf{adom}(J')$ implies that $d^k_1, \ldots, d^k_m \in \mathsf{adom}(J')$ and thus we have $R(d^k_1, \ldots, d^k_m) \in J'$. It remains to observe that, by definition of h', $R(h'(\bar{d}_1), \ldots, h'(\bar{d}_m)) = R(d^k_1, \ldots, d^k_m)$.

"$\Leftarrow$". Assume that $(K^*, \bar{c}^*) \to (J', \bar{M})$, where $(K, \bar{c}) = \prod S_M$, for every non-empty maximally guarded $M \subseteq \mathsf{adom}(J')$. Take any such M and a homomorphism h that witnesses $(K^*, \bar{c}^*) \to (I, \bar{M})$. Let h' denote the restriction of h to $\mathsf{adom}(\prod_{i\in[n]} I_i)$. Then clearly h' is a homomorphism from $\prod_{i\in[n]} I_i$ to J'. Since every value from $\prod_{i\in[n]} I_i$ occurs in some fact using a relation symbol different from R_a, h' cannot map any value from $\prod_{i\in[n]} I_i$ to a, the fresh value. Thus h' is also a homomorphism from $\prod_{i\in[n]} I_i$ to J, as required. This concludes the proof for GTGD.

For $\mathcal{L} \in \{\text{FGTGD}, \text{F1TGD}, \text{TGD}\}$, we follow the structure of the proof for GTGD and establish analogous results. Instead of Theorem 15, we invoke the following:

- Theorem 24, when $\mathcal{L} =$ FGTGD;
- Theorem 25, using values from $\mathsf{adom}(J')$ rather than maximally guarded sets, when $\mathcal{L} =$ F1TGD;
- Theorem 26, using unrestricted sets rather than maximally guarded sets, when $\mathcal{L} =$ TGD.

With respect to the non-totality of $\bar{M}$, for FGTGD we also consider maximally guarded sets, as for GTGD. For F1TGD and TGD, note that for any non-empty $M \subseteq \mathsf{adom}(J')$, whether M is a singleton or unrestricted, $\bar{M}$ is non-total in J', following the same argument as in the proof for GTGD. The non-totality of values is implied by the non-totality of $\bar{M}$, where M is a singleton.

Regarding Condition 1 of the respective characterizations for fitting $\mathcal{L}$-TGD, consider the sets defined in Condition 1 for each $\mathcal{L}$. Note that for FGTGD, even though maximally guarded sets are used, as in the case of GTGD, the definition of the sets S_M uses homomorphisms from $(J', \bar{M})$, as opposed to $(J'|_M, \bar{M})$. However, this has no influence on the argument. Thus, from here on, we can follow the exact same proof as for GTGD, since in both cases maximally guarded sets are considered. For $\mathcal{L} \in \{\text{F1TGD}, \text{GTGD}\}$, by the same reasoning as in the proof for GTGD, we observe:

- $S_M = \{(I'_i, \bar{M}) \mid i \in [n]\}$ holds for unrestricted M,
- $S_a = \{(I'_i, a) \mid i \in [n]\}$ holds for values a.

This confirms the non-emptiness of the sets defined in Condition 1 of each respective characterization.

Thus, for each $\mathcal{L}$, the reduction depends solely on Condition 2 of the theorem characterizing fitting $\mathcal{L}$-TGD existence. This allows us to state a claim, analogous to the one in the proof for GTGD: $\prod_{i\in[n]} I_i \to J$ if and only if

- $(K^*, \bar{c}^*) \to (J', \bar{M})$, where $(K, \bar{c}) = \prod S_M$, for unrestricted M, when $\mathcal{L} =$ TGD,
- $\prod S_a \to (J', a)$, for $a \in \mathsf{adom}(J')$, when $\mathcal{L} =$ F1TGD.

Observe that the proof of the claim for GTGD does not fundamentally depend on M being maximally guarded. Thus, the respective claim for TGD can be shown in the exact same way, assuming M is unrestricted. For $\mathcal{L} =$ F1TGD, values take the place of maximally guarded sets. This does not alter the overall argument but simplifies the proof slightly,

[4] The presence of a guarantees non-totality even if $\mathsf{adom}(J)$ is a singleton.

as $\prod S_a$ is considered directly, rather than its diversification. This finishes the proof of Point 1 of the theorem.

We now turn to the proof of Point 2 of Theorem 20, starting with the coNEXPTIME-hardness. As announced, we reduce the PHP. Recall that the PHP is already NEXPTIME-hard over instances over a single binary relation symbol E (ten Cate and Dalmau 2015), so we assume such instances from now on. Let $I_1, \ldots, I_n, J$ be an input to the PHP. We assume without loss of generality that $I_1, \ldots, I_n, J$ have pairwise disjoint active domains and that c is a fresh value (that does not occur in any of the domains). We construct a fitting instance (P, N) as follows:

- $\mathsf{P} = \{P_1, P_2\}$ where $P_1 = J \cup \{B(c)\}$ and $P_2 = I_1 \cup \{A(c), B(c)\}$;
- $\mathsf{N} = \{I'_1, \ldots, I'_n\}$ where I'_i is obtained from I_i by adding $A(a)$ for every $a \in \mathsf{adom}(I_i)$ and $B(c)$.

The following claim establishes correctness of the reduction.

Claim. $\prod_i I_i \to J$ iff (P, N) has no fitting FullTGD.

Proof of the Claim. For "⇒" suppose $\prod_i I_i \to J$ and let h be a witness for this. Let $R_1, \ldots, R_n$ be relation symbols and $\bar{a}_1, \ldots, \bar{a}_n$ be tuples such that $\bar{a}_i \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R_i)}$ and $R_i(\bar{a}_i[i]) \notin I'_i$ for $i \in [n]$. We have to show that there is a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P such that $R_j(h(\bar{a}_j)) \notin P$ for some $j \in [n]$. We distinguish cases on the relations symbol R_1:

- If $R_1 = A$, then $\bar{a}_i[1] = c$ for all $i \in [n]$ and h extended with $h(c, \ldots, c) = c$ is the required homomorphism to P_1;
- If $R_1 = B$, then $a_i[1] \neq c$ for all $i \in [n]$ and the map $g : \mathsf{adom}(\prod \mathsf{N}) \to \mathsf{adom}(P_2)$ defined by

$$g(b_1, \ldots, b_n) = \begin{cases} b_1 & \text{if } c \notin \{b_1, \ldots, b_n\} \\ c & \text{otherwise} \end{cases}$$

  is the required homomorphism to P_2;
- If $R_1 = E$, then the map g from the previous point is the required homomorphism to P_2;

For the other direction "⇐", suppose for every relation symbol R and tuple $\bar{a} \in \mathsf{adom}(\prod \mathsf{N})^{\mathsf{ar}(R)}$ with $R(\bar{a}[i]) \notin I'_i$ for all $i \in [n]$, there is a $P \in \mathsf{P}$ and a homomorphism h from $\prod \mathsf{N}$ to P such that $R(h(\bar{a})) \notin P$ for some $j \in [n]$. Consider relation symbol A and tuple $\bar{a} = c, \ldots, c$. Since $B(\bar{a}) \in \prod \mathsf{N}$, the only potential target of $\bar{a}$ by such homomorphism h is the point c in P_1. But then h has to map $\prod_i I_i$ into J, hence $\prod_i I_i \to J$ as required. ⊣

It remains to prove the DP-hardness result. We do this by reduction from 3-colorability/non-homomorphism, that is, we are given a triple (G, I, J) with $G = (V, E)$ an undirected graph and I and J instances over a single binary relation E, and we want to decide whether G is 3-colorable and $I \not\to J$. It is easy to prove that this problem is DP-complete.

Let (G, I, J) be a triple of the described form. Further let $G = (V, E)$ and $V = \{v_1, \ldots, v_n\}$. We use the following relation symbols:

- unary relation symbols $V_1, \ldots, V_n$ and $C_1, \ldots, C_n$ (for identifying the vertices of G and colors for each vertex in G);
- a relation symbol W of arity 4 (for *well-colored*), representing possible colorings of pairs of nodes in G;
- a relation symbol R of arity $2n$ for choosing a 3-coloring of G;
- a binary relation symbol E for edges of I.

We introduce the following examples:

- one negative example N that contains the following facts:
  - $V_i(v_i)$ for $1 \leq i \leq n$;
  - $C_i(r_i), C_i(g_i), C_i(b_i)$ for $1 \leq i \leq n$;
  - for $1 \leq i < j \leq n$ and all $c, d \in \{r, g, b\}$, the fact $W(v_i, c_i, v_j, d_j)$ if $\{v_i, v_j\} \notin E$ or $c \neq d$;
  - all facts from the instance I.
- a positive example P_0 which ensures that if we choose a non-R-tuple $\bar{a}$ in N that is of the wrong 'type', then we find a homomorphism from N to P_0 such that $R(h(\bar{a})) \notin P_0$. It contains the following facts:
  - $V_1(v_1), \ldots, V_n(v_n)$ and $C_1(c_1), \ldots, C_n(c_n)$;
  - all facts $W(v_i, c_i, v_j, c_j)$ with $1 \leq i < j \leq n$;
  - $R(v_1, c_1, \ldots, v_n, c_n)$;
  - $E(\bot, \bot)$.
- a positive example P_1 that contains the following facts:
  - $V_j(u_i)$ and $C_j(u_i)$ for $1 \leq j \leq n$ and all $i \in \{1, 2\}$;
  - $C_j(\bot_i)$ for $1 \leq j \leq n$ and $i \in \{1, 2\}$;
  - $W(a_1, b_1, a_2, b_2)$ for all $a_1, a_2 \in \{u_1, u_2\}$ and $b_1, b_2 \in D$, where $D = \{u_1, u_2, \bot_1, \bot_2\}$, such that
  - (i) $a_1 = a_2$ or $(a_1, a_2) = (u_1, u_2)$;
  - (ii) $a_i = b_i$ or $(a_i, b_i) = (u_1, \bot_1)$ or $(a_i, b_i) = (u_2, \bot_2)$ for all $i \in \{1, 2\}$;
  - (iii) $(a_1, b_1, a_2, b_2) \neq (u_1, u_1, u_2, u_2)$.
  - $R(\bar{a})$ for all $\bar{a} = (a_1, b_1, \ldots, a_n, b_n) \in D^{2n}$ such that
  - (iv) $a_i = a_{i+1}$ or $(a_i, a_{i+1}) = (u_1, u_2)$ for $1 \leq i < n$;
  - (v) $a_i = b_i$ or $(a_i, b_i) = (u_1, \bot_1)$ or $(a_i, b_i) = (u_2, \bot_2)$;
  - (vi) $\{(u_1, u_1), (u_2, u_2)\} \not\subseteq \{(a_1, b_1), \ldots, (a_n, b_n)\}$;
  - $E(\bot, \bot)$.
- a positive example P_2 that contains the following facts:
  - all facts on the relation symbols $V_1, \ldots, V_n, C_1, \ldots, C_n, W$ that use the single value $\bot$;
  - all facts from the instance J.
- a positive example P_3 that contains the following facts:
  - all facts on the relation symbols $V_1, \ldots, V_n, C_1, \ldots, C_n, W, R$ that use the single value $\bot$;
  - all facts from I.

It is not hard to verify that the size of our examples is only polynomial in the size of G. By Theorem 17, it suffices to show the following.

Claim 1. G is 3-colorable and $I \not\to J$ if and only if there is

a relation symbol X of some arity k and an $\bar{a} \in \mathsf{adom}(N)^k$ with $X(\bar{a}) \notin N$ such that for all homomorphisms h from N to a positive example P, we have $X(h(\bar{a})) \in P$.

We first argue that the only interesting choice for the relation symbol X is R, because for all $X \in \{V_1, \ldots, V_n, C_1, \ldots, C_n, W, E\}$ and all $\bar{a} \in \mathsf{adom}(N)^k$ with $X(\bar{a}) \notin N$, there is a homomorphism h from N to a positive example P with $X(h(\bar{a})) \notin P$. Whenever $X \in \{V_1, \ldots, V_n, C_1, \ldots, C_n\}$, we may in fact choose $P = P_0$ and the homomorphism h with

- $h(v_i) = v_i$ for $1 \leq i \leq n$;
- $h(r_i) = h(g_i) = h(b_i) = c_i$ for $1 \leq i \leq n$;
- $h(a) = \bot$ for all $a \in \mathsf{adom}(I)$.

If $X = W$ and $\bar{a}$ is not of the form (v_k, c, v_ℓ, d) with $c \in \{r_k, g_k, b_k\}$ and $d \in \{r_\ell, g_\ell, b_\ell\}$, then we may again use $P = P_0$ and the homomorphism h just given. Thus the remaining case is that $X = W$ and $\bar{a}$ is of the form (v_k, c, v_ℓ, d) with $c \in \{r_k, g_k, b_k\}$ and $d \in \{r_\ell, g_\ell, b_\ell\}$. We then choose $P = P_1$ and the following homomorphism h:

- $h(v_i) = u_1$ for $1 \leq i \leq \ell$;
- $h(r_i) = h(g_i) = h(b_i) = u_1$ for $1 \leq i \leq k$;
- $h(c_k) = u_1$ and $h(d) = \bot_1$ for all $d \in \{r_k, g_k, b_k\} \setminus \{c_k\}$;
- $h(r_i) = h(g_i) = h(b_i) = u_1$ for $k < i \leq \ell$;
- $h(v_i) = u_2$ for $\ell \leq i \leq n$;
- $h(c_\ell) = \bot_2$ and $h(d) = u_2$ for all $d \in \{r_\ell, g_\ell, b_\ell\} \setminus \{c_\ell\}$;
- $h(r_i) = h(g_i) = h(b_i) = u_2$ for $\ell < i \leq n$;
- $h(a) = \bot$ for all $a \in \mathsf{adom}(I)$.

Finally, if $X = E$, then we may choose $P = P_3$ and the following homomorphism:

- h maps all values v_i, r_i, g_i, b_i to $\bot$;
- h is the identity on all values from I.

By what was said above, to show correctness of the reduction it suffices to prove the following claim.

Claim 2. G is 3-colorable and $I \not\to J$ if and only if there is an $\bar{a} \in \mathsf{adom}(N)^{2n}$ with $R(\bar{a}) \notin N$ such that for all homomorphisms h from N to a positive example P, we have $R(h(\bar{a})) \in P$.

Proof of the claim. "$\Rightarrow$". Assume that G is 3-colorable with $\chi : V \to \{r, g, b\}$ a proper 3-coloring, and that $I \not\to J$. We show that Point 1 of Claim 2 is satisfied. Set $\bar{a} = (v_1, \chi(v_1)_1, \ldots, v_n, \chi(v_n)_n)$. We have $R(\bar{a}) \notin N$, simply because N contains no R-tuples. Let h be a homomorphism into a positive example P. We have to show that $R(h(\bar{a})) \in P$. We distinguish cases:

- $P = P_0$. Then because of the use of the V_i and C_i relations, we must have $h(\bar{a}) = (v_1, c_1, \ldots, v_n, c_n)$. But then $R(h(\bar{a})) \in P_0$, as required.
- $P = P_1$. Let $h(\bar{a}) = (a_1, b_1, \ldots, a_n, b_n)$. Because $\bar{a}$ is derived from a proper 3-coloring and by definition of N, we have $W(v_i, \chi(v_i)_i, v_j, \chi(v_j)_j) \in N$ for $1 \leq i < j \leq n$. Due to the definition of the W-relation in P_1, this implies that Properties (iv) and (v) from the definition of P_1 are satisfied. It also implies that Property (vi) is satisfied since there is no fact $W(u_1, u_1, u_2, u_2) \in P_1$. But then $R(h(\bar{a})) \in P_1$, as required
- $P = P_2$. Then h witnesses that $I \to J$, in contradiction to $I \not\to J$.

"$\Leftarrow$". First assume that there is an $\bar{a} \in \mathsf{adom}(N)^{2n}$ with $R(\bar{a}) \notin N$ such that for all homomorphisms h from N to a positive example P, we have $R(h(\bar{a})) \in P$.

Considering the positive example $P = P_0$, this means that $\bar{a}$ takes the form $(v_1, c_1, \ldots, v_n, c_n)$ where $c_i \in \{r_i, g_i, b_i\}$ for $1 \leq i \leq n$. In fact, consider the homomorphism h from N to P_0 defined as follows:

- $h(v_i) = v_i$ for $1 \leq i \leq n$;
- $h(r_i) = h(g_i) = h(b_i) = c_i$ for $1 \leq i \leq n$.

If $\bar{a}$ does not take the stated form, then $R(h(\bar{a})) \notin P_0$, in contradiction to our choice of $\bar{a}$.

Note that $\bar{a}$ thus represents a coloring χ of G defined by setting $\chi(v_i) = r$ if $c_i = r_i$, $\chi(v_i) = g$ if $c_i = g_i$, and $\chi(v_i) = b$ if $c_i = b_i$. Considering the positive example $P = P_1$, we now argue that χ is proper.

Assume to the contrary that χ is not proper. Then by definition of N there are k, ℓ with $1 \leq k < \ell \leq n$ such that $W(v_k, c_k, v_\ell, c_\ell) \notin N$. As a consequence, the following map h is a homomorphism from N to P_1:

- $h(v_i) = u_1$ for $1 \leq i \leq \ell$;
- $h(r_i) = h(g_i) = h(b_i) = u_1$ for $1 \leq i \leq k$;
- $h(c_k) = u_1$ and $h(d) = \bot_1$ for all $d \in \{r_k, g_k, b_k\} \setminus \{c_k\}$;
- $h(r_i) = h(g_i) = h(b_i) = u_1$ for $k < i \leq \ell$;
- $h(v_i) = u_2$ for $\ell \leq i \leq n$;
- $h(c_\ell) = \bot_2$ and $h(d) = u_2$ for all $d \in \{r_\ell, g_\ell, b_\ell\} \setminus \{c_\ell\}$;
- $h(r_i) = h(g_i) = h(b_i) = u_2$ for $\ell < i \leq n$;
- $h(a) = \bot$ for all $a \in \mathsf{adom}(I)$.

But $R(h(\bar{a})) \notin P_1$, contradicting the choice of $\bar{a}$.

Assume to the contrary of what is left to be shown that there is a homomorphism h from I to J. Then there is clearly also a homomorphism h from N to P_2. So we have $R(h(\bar{a})) \in P_2$. This, however, contradicts the fact that P_2 does not contain any R-facts. ❑

We recall a known result that we are going to use as a black box.

Theorem 27. *(ten Cate and Dalmau 2015; Willard 2010) Let $n \geq 1$. There are instances $I_1, \ldots, I_n, J$ such that*

1. *the size of the instances is bounded by $p(n)$, p a polynomial;*
2. *$\prod_i I_i \not\to J$, that is, there is a Boolean CQ q with $\prod_i I_i \models q$ and $J \not\models q$;*
3. *the smallest such CQ q has size 2^n.*

We can now show the claimed lower bounds based on Theorem 27 and a construction similar to the one in the complexity hardness proof in Theorem 15.

Theorem 21. *For every $n \geq 1$, there is a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ such that*

1. *the size of $(\mathsf{P}_n, \mathsf{N}_n)$ is bounded by $p(n)$, p a polynomial;*
2. *$(\mathsf{P}_n, \mathsf{N}_n)$ admits a fitting guarded and frontier-one TGD;*
3. *the smallest TGD fitting $(\mathsf{P}_n, \mathsf{N}_n)$ has size 2^n.*

The same is true for ontologies instead of single TGDs.

Proof. Let $n \geq 1$ and $I_1, \ldots, I_n, J$ be the instances that exist by Theorem 27, and let q_0 be the Boolean CQ witnessing Point 2. We assume without loss of generality that the active domains of all these instances are pairwise disjoint. We construct $\mathsf{P}_n, \mathsf{N}_n$ by taking:

- $\mathsf{N}_n = \{J\}$;
- $\mathsf{P}_n = \{I'_1, \ldots, I'_n\}$ where $I'_i = I_i \cup J$, for all i.

Clearly, Point 1 is satisfied. For Point 2, observe that $\top \to q_0$ witnesses that $(\mathsf{P}_n, \mathsf{N}_n)$ admit a fitting guarded and frontier-one TGD. For Point 3, suppose $\rho = p(\bar{x}) \to q(\bar{x})$ is a smallest TGD fitting $(\mathsf{P}_n, \mathsf{N}_n)$. We make a couple of observations.

Observe first that we can assume that $q(\bar{x})$ is connected. Here, we call $q(\bar{x})$ connected if the undirected graph $(\mathsf{var}(q), \{\{x, y\} \mid x, y \text{ co-occur in some atom in } q\})$ is connected. In fact, since J is a negative example for ρ, there is a tuple $\bar{c}$ such that there is a homomorphism h from $p(\bar{x})$ to $(J, \bar{c})$ but not a homomorphism from $q(\bar{x})$ to $(J, \bar{c})$. Hence, we can select a connected component $q'(\bar{x})$ of $q(\bar{x})$ that does not have a homomorphism to $(J, \bar{c})$.

Suppose now that $\bar{x}$ is not the empty tuple. Since $J \subseteq I_i$ for all i, h from the previous paragraph is also a homomorphism from $p(\bar{x})$ to all $(I'_i, \bar{c})$. Since all I'_i are positive examples for ρ, there is also a homomorphism from $q(\bar{x})$ to $(I'_i, \bar{c})$. As $q(\bar{x})$ is connected, this is actually a homomorphism from $q(\bar{x})$ to $(J, \bar{c})$, a contradiction. We conclude that $\bar{x}$ is the empty tuple and q is in fact a Boolean CQ.

It remains to note that q fits $I'_1, \ldots, I'_n$ and J that is, $I'_i \models q$, but $J \not\models q$. Since q is connected and J is contained in each I'_i, we conclude that $I_i \models q$ for all i. By Point 3 of Theorem 27, q and thus ρ is of size 2^n.

To see that the same also holds for ontologies, recall that by Lemma 3 a smallest ontology fitting $(\mathsf{P}_n, \mathsf{N}_n)$ contains only one TGD since there is only one negative example involved. ❑

Theorem 22. *For every $n \geq 1$, there is a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ such that*

1. *the size of $(\mathsf{P}_n, \mathsf{N}_n)$ is bounded by $p(n)$, p a polynomial;*
2. *$(\mathsf{P}_n, \mathsf{N}_n)$ admits a fitting full TGD;*
3. *the smallest full TGD fitting $(\mathsf{P}_n, \mathsf{N}_n)$ has size 2^n.*

Proof. Let $n \geq 1$ and $I_1, \ldots, I_n, J$ be the instances that exist by Theorem 27. We assume without loss of generality that the active domains of all these instances are pairwise disjoint. Let A be a fresh unary relation symbol and let I_i^A denote the extension of I_i with all facts $A(d)$ for $d \in \text{adom}(I_i)$;

We construct a fitting instance $(\mathsf{P}_n, \mathsf{N}_n)$ by taking:

- $\mathsf{P}_n = \{J'\}$ for $J' = J \cup \bigcup_i I_i^A$;
- $\mathsf{N}_n = \{I_1, \ldots, I_n\}$.

Clearly, Point 1 of the theorem is satisfied. For Point 2, let q be any Boolean CQ with $\prod_i I_i \models q$, but $J \not\models q$, which exists due to Point 2 of Theorem 27. Let $q'(x)$ be the variant of q in which one (any) of the quantified variables of q is made an answer variable. Then $q'(x) \to A(x)$ is a fitting full TGD for $(\mathsf{P}_n, \mathsf{N}_n)$.

For Point 3, let $\rho = \varphi(\bar{x}, \bar{y}) \to \psi(\bar{y})$ be a fitting full TGD. Since all I_i are negative examples, there are, for $i \leq n$, homomorphisms h_i from $q(\bar{y}) = \exists \bar{x}\, \varphi(\bar{x}, \bar{y})$ to I_i such that $I_i \not\models \psi(h_i(\bar{y}))$. Suppose first that ψ does not mention A. But then such h_i is also a homomorphism from $q(\bar{y})$ to I_i^A and by construction $I_i^A \not\models \psi(h_i(\bar{y}))$. Hence, J' is not a positive example for ρ, a contradiction. Thus, $\psi(\bar{y})$ contains an atom $A(z)$. Since none of the $I_i \in \mathsf{N}$ contains an A-fact, we can actually assume that $\psi(\bar{y})$ is a single fact $A(z)$, and that ρ takes the shape $q(z) \to A(z)$. Now, this $q(z)$ cannot have a homomorphism to J since otherwise J' would not be a positive example for ρ. Consider the Boolean CQ $q' = \exists z\, q(z)$. By what was said above q' has a homomorphism into every I_i, but not into J. By Point 3 of Theorem 27, q' is of size at least 2^n. ❑