COHERENT AND CONSISTENT RELATIONAL TRANSFER LEARNING WITH AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human defined concepts are inherently transferable, but it is not clear under what conditions they can be modelled effectively by non-symbolic artificial learners. This paper argues that for a transferable concept to be learned, the system of relations that define it must be coherent across domains. This is to say that the learned concept-specific relations ought to be consistent with respect to a theory that constrains their semantics and that such consistency must extend beyond the representations encountered in the source domain. To demonstrate this, we first present formal definitions for consistency and coherence, and a proposed Dynamic Comparator relation-decoder model designed around these principles. We then perform a proposed Partial Relation Transfer learning task on a novel data set, using a neural-symbolic autoencoder architecture that joins sub-symbolic representations with modular relation-decoders. By comparing against several existing relation-decoder models, our experiments show that relation-decoders which maintain consistency over unobserved regions of representational space retain coherence across domains, whilst achieving better transfer learning performance.

1 INTRODUCTION

Humans are capable of learning concepts such that they can be applied to many different scenarios (Inhelder & Piaget, 1964; Piaget, 2005; Lake et al., 2017). An important characteristic is that human-like concepts remain *coherent* across contexts, whereby their logical consistency in one context is retained in another (Nye et al., 2021). As an example, consider the concept of ordinality which permits comparison over ordered sets, *e.g*. "A is larger than B", and pertains to a multitude of properties: position, size, volume, reach, *etc*. So long as one of these properties can be attributed to an object, a set of objects can be compared on that basis; in this sense ordinality generalises between objects. All in all, if the concept of ordinality were to be learned in its most general form, it should be coherent across properties and objects.

In this paper, we seek to define the conditions that allow a learned concept to transfer well across properties and objects in the case of sub-symbolic learners (d'Avila Garcez & Lamb, 2020; Santoro et al., 2021; Greff et al., 2020). We define consistency and coherence of sub-symbolic learners borrowing from analogous definitions from symbolic AI. We propose a neural-symbolic autoencoder architecture consisting of a neural encoder for objects coupled with modular relation-decoders (Serafini & Garcez, 2016; Donadello et al., 2017; Badreddine et al., 2020; Wang et al., 2017; Nickel et al., 2016a; Dai et al., 2020), and we show that this architecture is capable of achieving an improved transfer learning performance by being coherent across properties and objects.

We therefore claim that retaining consistency across domains dictates concept coherence, offering a more fine-grained measure of transfer learning than accuracy alone. The proposed architecture is a practical instantiation of this formalisation and is evaluated in this paper on a new Partial Relation Transfer (PRT) task and data set. We begin by expressing the symbolic application of a set of relations to some domain of interest as a model-theoretic structure, such as A is larger than B, and defining an analogous soft-structure for non-symbolic learners where relations are modelled by relation-decoders that compute beliefs. We then propose formal definitions for consistency and coherence of soft-structures which provide a practical consistency score calculation to the evaluation of autoencoders. Finally, we present a benchmark PRT learning task with the use of a new BlockStacks data set derived from the CLEVR data set rendering agent. We then compare our proposed archi-

tecture with several existing relation-decoder models on transfer learning tasks from the MNIST data set to BlockStacks. Our experiments show that relation-decoders which maintain consistency over unobserved regions of representational space retain coherence across domains whilst achieving better transfer learning performance. The contributions of this paper are:

- A formal definition of coherence and consistency for sub-symbolic learners with a practical evaluation score.
- A neural model and learning task for partial relation transfer including a new data set to evaluate concept coherence.
- A comprehensive critical evaluation of results in comparison with multiple state-of-the-art relation-decoder models.

In Section 2 we provide the required background, Section 3 introduces soft-structures and formally defines coherence and consistency, Section 4 provides a practical consistency loss and Section 5 then outlines our neural-symbolic architecture. After detailing the PRT task in Section 6, we present results in Section 7 and complete the paper in Section 8 with a discussion and concluding remarks, including limitations and future work. We provide related work in Appendix A.

2 PRELIMINARIES

Notations: We reserve uppercase calligraphic letters to denote sets and lowercase versions of the same letter to denote their elements, e.g., $\mathcal{S} = \{s_1, \ldots, s_n\}$ is a set $\mathcal{S}$ of n elements s_i. We indicate with $|\mathcal{S}| = n$ the cardinality of $\mathcal{S}$. We use uppercase roman letters to denote a random variable e.g., S, and use the uppercase calligraphic version of the same letter ($\mathcal{S}$) to denote the set from which the random variable takes values according to some corresponding probability distribution, p_{S} , over the elements of the set, such that $\sum_{i=1}^{|\mathcal{S}|} p_{\mathrm{S}}(s_i) = 1$ for a discrete $\mathcal{S}$. For brevity, we may write $p_{\mathrm{S}}(s_i)$ as $p(s_i)$, where the random variable is implied by the argument. We use bold font lowercase letters to denote vector elements, e.g., $\boldsymbol{s}_i \in \mathbb{R}^d$ is an d-dimensional vector element from the set $\mathcal{S} = \mathbb{R}^d$.

Logic and model-theoretic background: We assume a formal language $\mathcal{L}$ composed of variables, predicates (i.e. relations), logical connectives $\neg$ (negation), $\vee$ (disjunction) and $\wedge$ (conjunction), $\rightarrow$ (implication) and universal quantification $\forall$ (for all) with their conventional meaning (see Shapiro & Kouri Kissel (2021)). Relations express relational knowledge over elements of a domain. For instance, $r(s_1, s_2)$ states that elements s_1 and s_2 are related through the binary relation r. The meaning of relations is defined by an *interpretation*, $I_{\mathcal{S}_\sigma}$ over elements of an non-empty *domain* $\mathcal{S}$. Together a $\mathcal{S}$ and $I_{\mathcal{S}_\sigma}$ form a *structure* $\mathcal{S}_\sigma$.

Definition 1 (Signature, Interpretation, Structure) *The* signature *of a language $\mathcal{L}$ is $\sigma = \{r \in \mathcal{L} : r$ is a relation$\}$, whose elements have* arity *given by $\mathsf{ar} : \sigma \rightarrow \mathcal{N}$, where $\mathcal{N}$ is the set of natural numbers. For each $r \in \sigma$, $\mathsf{ar}(r)$ denotes the arity of r. Given a signature σ and a non-empty domain $\mathcal{S}$, an* interpretation *$I_{\mathcal{S}_\sigma}$ of σ over elements of $\mathcal{S}$ assigns to each relation $r \in \sigma$ a set $I_{\mathcal{S}_\sigma}(r) \subseteq \mathcal{S}^{\mathsf{ar}(r)}$. A* structure *is a tuple $\mathcal{S}_\sigma = (\mathcal{S}, I_{\mathcal{S}_\sigma})$.*

Note that for a fixed domain $\mathcal{S}$ and signature σ, different interpretations yield different structures. We construct universally quantified first-order formulae (called sentences) using the signature σ of $\mathcal{L}$, whose truth-value is defined with respect to a given structure $\mathcal{S}_\sigma$. To do so, we first consider *ground* instances of a formula. These are given by replacing all the variables in the formula with elements from the domain $\mathcal{S}$. For instance, $r(s_1, s_2)$, where s_1 and s_2 are elements of $\mathcal{S}$, is a *ground* instance of an atomic formula $r(i, j)$ where i and j are variables in $\mathcal{L}$. Given a structure $\mathcal{S}_\sigma = (\mathcal{S}, I_{\mathcal{S}_\sigma})$ a relation r and a tuple $(s_1, \ldots, s_{\mathsf{ar}(r)}) \in \mathcal{S}^{\mathsf{ar}(r)}$, a ground instance $r(s_1, \ldots, s_{\mathsf{ar}(r)})$ is true in the structure $\mathcal{S}_\sigma$ if and only if $(s_1, \ldots, s_{\mathsf{ar}(r)}) \in I_{\mathcal{S}_\sigma}(r)$. The truth value of a sentence in a given structure $\mathcal{S}_\sigma$ depends on the truth value of its respective ground instances. Specifically, a sentence is true in a structure $\mathcal{S}_\sigma$ if and only if all of its ground instances are true in $\mathcal{S}_\sigma$. When a sentence, τ, is true in a structure, $\mathcal{S}_\sigma$, we say that the structure *satisfies* τ, denoted as $\mathcal{S}_\sigma \models \tau$. A set of sentences form a *theory*, $\mathcal{T}$. A *model* of $\mathcal{T}$ is a structure that satisfies every sentence in $\mathcal{T}$.

Definition 2 (Model of a theory) *Let $\mathcal{T}$ be a theory written in a language $\mathcal{L}$ and let $\mathcal{S}_\sigma = (\mathcal{S}, I_{\mathcal{S}_\sigma})$ be a structure, where σ is the signature of $\mathcal{L}$. $\mathcal{S}_\sigma$ is a* model of *$\mathcal{T}$ if and only if $\mathcal{S}_\sigma \models \tau$ for every sentence $\tau \in \mathcal{T}$.*

Example 1 *Suppose we have the structure $\mathcal{S}_\sigma = (\mathcal{S}, I_{\mathcal{S}_\sigma})$, where $\mathcal{S}$ is a domain of images of handwritten digits and σ the signature of binary relations $\sigma = \{$isGreater, isEqual, isLess, isSuccessor, isPredecessor$\}$, or for short $\sigma = \{\mathsf{G}, \mathsf{E}, \mathsf{L}, \mathsf{S}, \mathsf{P}\}$. Let $\mathcal{T}$ be the theory that defines ordinality including, for instance, the sentence $\forall i, j.\ \mathsf{G}(i,j) \rightarrow \neg\mathsf{E}(i,j)$ (if a digit is greater than another then they are not equal). Any structure $\mathcal{S}_\sigma = (\mathcal{S}, I_{\mathcal{S}_\sigma})$ with interpretations $I_{\mathcal{S}_\sigma}$ of σ that captures a total order over the elements of $\mathcal{S}$ is a model of $\mathcal{T}$.*

3 Approximating structures that have real-world domains

In this section we turn our attention to the challenge of learning a model over a real-world domain, given a signature and theory. Here a learner must determine an appropriate interpretation over real-world data, such as images or other perceptions. This can be challenging because, firstly, we may only have a partial description of the interpretation, and secondly data may be noisy and contain information that is not relevant to the theory. For instance MNIST, a relatively simple data set by current standards, consists of stylistic details such as line thickness and digit skew (Chen & Batmanghelich, 2020), which are irrelevant to the notion of ordinality, which makes obtaining the structure from Example 1 more complicated. Nevertheless, statistical machine learning models are able to discover commonalities in data which help to infer the underlying semantics (i.e. interpretation) and disregard the noise. Following the convention in disentanglement literature (Bengio et al., 2013; Kingma & Welling, 2014; Higgins et al., 2017; 2018), we take the assumption that real-world observations S are drawn from some conditional distribution $p_{\mathrm{S}|\mathrm{Z}}$, where Z is a latent random variable, itself drawn from prior p_{Z}. It is therefore useful to define a domain *encoding* of the form,

$$\psi_{\mathcal{S}} : \mathcal{S} \rightarrow \mathcal{Z}, \tag{1}$$

tasked with approximating the conditioned expectation of the posterior, *i.e.* $\psi_{\mathcal{S}}(s) = \mathbb{E}[p_{\mathrm{Z}|\mathrm{S}}(\mathrm{Z}|s)]$. Since obtaining an interpretation from domain encodings, for a given signature, may require dealing with noise, we express the interpretation of relations over real-world data by belief functions over the space $\mathcal{Z}$ (Paris & Vencovská, 2015; Paris, 1994) , and refer to these as *relation-decoders*:

$$\phi_r : \mathcal{Z}^{\mathsf{ar}(r)} \rightarrow (0,1) \tag{2}$$

with $\phi = \{\phi_r : r \in \sigma\}$. Concretely, for a binary relation r and ordered pair $(s_i, s_j) \in \mathcal{S}^2$, $\phi_r(\psi_{\mathcal{S}}(s_i), \psi_{\mathcal{S}}(s_j))$ describes the belief that $(s_i, s_j) \in I_{\mathcal{S}_\sigma}(r)$. A belief $\phi_r(\psi_{\mathcal{S}}(s_i), \psi_{\mathcal{S}}(s_j)) \approx 1$ signifies a strong belief that $(s_i, s_j) \in I_{\mathcal{S}_\sigma}(r)$ and $\phi_r(\psi_{\mathcal{S}}(s_i), \psi_{\mathcal{S}}(s_j)) \approx 0$ signifies a strong belief that $(s_i, s_j) \notin I_{\mathcal{S}_\sigma}(r)$. Together, $\psi_{\mathcal{S}}$ and ϕ allow us to define a belief-based analogue to a structure.

Definition 3 (Soft-Structure/Soft-Substructure) *Given signature σ, a possibly infinite set $\mathcal{Z}$ and relation-decoders ϕ, a* soft-structure *is a tuple $\tilde{\mathcal{Z}}_\sigma = (\mathcal{Z}, \phi)$. For (finite) domain $\mathcal{S}$ and encoding $\psi_{\mathcal{S}} : \mathcal{S} \rightarrow \mathcal{Z}$, $\tilde{\mathcal{S}}_\sigma = (\psi_{\mathcal{S}}(\mathcal{S}), \phi)$ is a (finite)* soft-substructure *of $\tilde{\mathcal{Z}}_\sigma$, with sub-domain $\psi_{\mathcal{S}}(\mathcal{S}) = \{\psi_{\mathcal{S}}(s) | s \in \mathcal{S}\} \subseteq \mathcal{Z}$.*

A soft-structure can be used to learn a structure over a real-world domain through learning $\psi_{\mathcal{S}}$ and ϕ. Clearly, a finite soft-substructure is a soft-structure. To determine the degree to which a soft-structure supports any given structure we introduce the following measure:

$$p(\mathcal{S}_\sigma | \tilde{\mathcal{S}}_\sigma) = \prod_{r \in \sigma} \prod_{O \in \mathcal{S}^{\mathsf{ar}(r)}} (\phi_r(\psi_{\mathcal{S}}(O)))^{\gamma^r_{O,\mathcal{S}_\sigma}} (1 - \phi_r(\psi_{\mathcal{S}}(O)))^{1-\gamma^r_{O,\mathcal{S}_\sigma}} \tag{3}$$

where $\gamma^r_{O,\mathcal{S}_\sigma} = 1$ if $O \in I_{\mathcal{S}_\sigma}(r)$, and 0 otherwise; we use $\phi_r(\psi_{\mathcal{S}}(O))$ as shorthand for $\phi_r(\psi_{\mathcal{S}}(s_1), \ldots, \psi_{\mathcal{S}}(s_n))$ for $n = \mathsf{ar}(r)$. Eqn. 3 expresses the assumption that, given a finite soft-structure, the beliefs in what constitutes the (different) interpretations of (different) relations are independent of one another. It is straightforward to show that $\sum_{\mathcal{S}_\sigma} p(\mathcal{S}_\sigma | \tilde{\mathcal{S}}_\sigma) = 1$ (summed over all possible structures with domain $\mathcal{S}$ and signature σ) and so it can be treated as a probability measure, where $p(\mathcal{S}_\sigma | \tilde{\mathcal{S}}_\sigma) \approx 1$ means that there is a high probability that the interpretation sampled from $\tilde{\mathcal{S}}_\sigma$ will be $I_{\mathcal{S}_\sigma}$. If we have a theory $\mathcal{T}$ over σ then it is natural to ask with what weight $\tilde{\mathcal{S}}_\sigma$ supports any given structure that is a model of $\mathcal{T}$. In the following, we use *model weight*, $\Gamma^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}}$, to describe the support given by $\tilde{\mathcal{S}}_\sigma$ to models of $\mathcal{T}$:

$$\Gamma^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}} = \sum_{\mathcal{S}_\sigma \in \mathcal{M}^{\mathcal{T}}_{\mathcal{S}}} p(\mathcal{S}_\sigma | \tilde{\mathcal{S}}_\sigma) \tag{4}$$

where $\mathcal{M}_{\mathcal{S}}^{\mathcal{T}}$ is the set of all structures with domain $\mathcal{S}$ that are models of $\mathcal{T}$. This lets us compare soft-structures, wherein a good soft-structure will be one that has a high model weight:

Definition 4 (ϵ-Consistency of Soft-Structure) *Given a finite soft-structure $\tilde{\mathcal{S}}_\sigma$, if $1 - \Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma} \leq \epsilon$ then we say that the soft-structure is* ϵ-consistent *with theory $\mathcal{T}$.*

We propose ϵ-consistency as an appropriate quantified measure of the notion of consistency presented in (Nye et al., 2021). A consistent soft-structure $\tilde{\mathcal{S}}_\sigma$ ensures that ϕ gives high belief only to interpretations that satisfy, i.e., are *consistent* with, $\mathcal{T}$. However, this expression is limited to the domain encodings of $\tilde{\mathcal{S}}_\sigma$, *i.e.* $\psi_{\mathcal{S}}(\mathcal{S})$. Going a step further, for a concept to be learned in a manner comparable to what a human might learn, we would expect that this consistency carries over to new domains and their corresponding soft-structures, as defined in what follows.

3.1 Coherence between soft-structures

In this section we define the notion of coherence between soft-structures, which aims to characterise what it means for a concept to be learned in a human-like manner. As a motivating case, consider a situation where we have already learned a soft-structure that has high model weight with models from Example 1. Now suppose we are given a new domain of images, $\mathcal{Y}$, showing single block stacks of differing height, and let us again use the signature of ordinal relations and $\mathcal{T}$ from Example 1. Lastly, let $I_{\mathcal{Y}_\sigma}$ be a corresponding interpretation that orders images according to block stack height and is a model of $\mathcal{T}$. We can summarise this with the following two structures:

$$\mathcal{X}_\sigma = (\mathcal{X}, I_{\mathcal{X}_\sigma}) \in \mathcal{M}_{\mathcal{X}}^{\mathcal{T}} \qquad \text{and} \qquad \mathcal{Y}_\sigma = (\mathcal{Y}, I_{\mathcal{Y}_\sigma}) \in \mathcal{M}_{\mathcal{Y}}^{\mathcal{T}}, \tag{5}$$

where $\mathcal{X}_\sigma$ is the structure from Example 1 with a domain of handwritten digits and $\mathcal{Y}_\sigma$ is our new structure, with a domain of block stack images. These can be modelled by soft-structures:

$$\tilde{\mathcal{X}}_\sigma = (\psi_{\mathcal{X}}(\mathcal{X}), \phi) \qquad \text{and} \qquad \tilde{\mathcal{Y}}_\sigma = (\psi_{\mathcal{Y}}(\mathcal{Y}), \phi), \tag{6}$$

which use domain-specific encoders, $\psi_{\mathcal{X}}$ and $\psi_{\mathcal{Y}}$, but share the same relation-decoders. As we know that $\tilde{\mathcal{X}}_\sigma$ has a high model weight and since ϕ is shared with $\tilde{\mathcal{Y}}_\sigma$, a natural question to ask is: under what conditions will a ϕ that is consistent over domain-encodings $\psi_{\mathcal{X}}(\mathcal{X})$ also be consistent over $\psi_{\mathcal{Y}}(\mathcal{Y})$? If this is the case, then we know that the high model weight in $\tilde{\mathcal{X}}_\sigma$ is reciprocated for $\tilde{\mathcal{Y}}_\sigma$. Concretely, we are interested in when the following *coherence* condition holds.

Definition 5 (ϵ-Coherence across soft-structures) *Two soft-structures, $\tilde{\mathcal{X}}_\sigma$ and $\tilde{\mathcal{Y}}_\sigma$ that share relation-decoders ϕ, are said to be* ϵ-coherent *with respect to a theory $\mathcal{T}$, if they are both at least ϵ-consistent with $\mathcal{T}$.*

Coherence between $\tilde{\mathcal{X}}_\sigma$ and $\tilde{\mathcal{Y}}_\sigma$ as defined above means that the concept of ordinality that applies to digit ordering can also be applied to block stack height ordering. It is desirable that learning ordinality on the domain of digits produces a coherent concept of ordinality with respect to other ordinal properties, such as height. Since it is possible that $\psi_{\mathcal{S}}(\mathcal{X})$ and $\psi_{\mathcal{S}}(\mathcal{Y})$ produce unique encodings, coherence relies on ϕ's ability to generalise over possibly disjoint subsets of $\mathcal{Z}$[1].

4 A practical consistency loss and ϵ-proxy

Calculating Eqn. 4 can quickly become intractable as it involves computing ϕ beliefs for every grounding and comparing these with every interpretation that is a model of the theory of interest. We therefore want to derive an efficient consistency loss and a calculable ϵ-proxy, that can act as a proxy estimate for a soft-structure's ϵ-consistency/coherence with a given theory, without needing access to every model and without requiring an exhaustive calculation over every grounding. In this section, we present the ϵ-proxy derivation outline and defer the expanded derivation to Appendix H.

Suppose we have a fixed domain $\mathcal{S}$ and a theory $\mathcal{T}$, whose sentences use relations from a signature σ. A structure $\mathcal{S}_\sigma$ will be a model of $\mathcal{T}$ if and only if each ground instance of each formula of

[1] If soft-structure $\tilde{\mathcal{Z}}_\sigma$ (defined over the full space $\mathcal{Z}$) is consistent, then coherence is guaranteed between all possible soft-substructures.

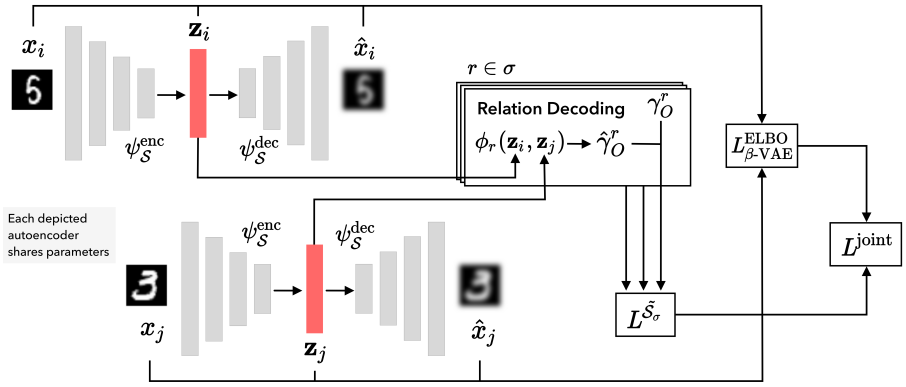

Figure 1: Network architecture used for PRT task, with $\hat{\gamma}^r_O = \phi_r(\psi^{enc}_S(O))$. The figure shows how relational learning is performed on the source MNIST data set (to learn e.g. that digit 5 is greater than 3). As discussed in Section 6, moving to the target domain (to learn that a stack of blocks is greater than another) involves training a new $\psi^{enc/dec}_{\mathcal{Y}}$ together with a subset of the ϕ_r relation-decoders (with fixed parameters), where the rest are held out as zero-shot transferred relation-decoders.

$\mathcal{T}$ is satisfied. Let $k \in \{1, ..., K_0\}$ be the index associated with each unique grounding of domain elements to the variables of $\mathcal{T}$. Further, take $B_\mathcal{T}$ to be a Boolean random variable denoting the truth-value of $\mathcal{T}$, so that the probability of the theory being either satisfied ($b_\mathcal{T} = 1$) or violated ($b_\mathcal{T} = 0$) across ground instances, under a soft-structure $\tilde{\mathcal{S}}_\sigma$ can be expressed as $p(b_\mathcal{T}|\tilde{\mathcal{S}}_\sigma, k)$. Notably, by the definition of a theory, ground instances of formulae will always hold as true for any model of $\mathcal{T}$, *i.e.* $p(b_\mathcal{T} = 1|\mathcal{S}_\sigma, k) = 1$ if $\mathcal{S}_\sigma \in \mathcal{M}^\mathcal{T}_\mathcal{S}$. Similarly, when $\tilde{\mathcal{S}}_\sigma$ is consistent with $\mathcal{T}$ then we should also find $p(b_\mathcal{T} = 1|\tilde{\mathcal{S}}_\sigma, k) \approx 1$. We thus define our consistency loss as an expectation of the binary cross entropy between $p(B_\mathcal{T}|\mathcal{S}_\sigma, k)$ and $p(B_\mathcal{T}|\tilde{\mathcal{S}}_\sigma, k)$, which, given $p(b_\mathcal{T} = 0|\mathcal{S}_\sigma, k) = 0$ for any k grounding, simplifies to the expected negative log-likelihood of satisfying $\mathcal{T}$ under a randomly sampled grounding,

$$L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma) = \mathbb{E}_{k\sim p(k)}[-\ln p(b_\mathcal{T} = 1|\tilde{\mathcal{S}}_\sigma, k)]. \tag{7}$$

where $p(k) = \frac{1}{K_0}$ is taken to be uniform over the possible unique groundings. However, we still require an ϵ-proxy measure based on this loss, to enable practical evaluation of concept coherence. To achieve this, we define $\bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_\mathcal{T} = \exp(-L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma))$ and use its relationship with $\Gamma^{\tilde{\mathcal{S}}_\sigma}_\mathcal{T}$ to define a proxy bound,

$$\ln \frac{1}{1-\bar{\epsilon}} \geq L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma) \tag{8}$$

where $\bar{\epsilon} \geq 1 - \bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_\mathcal{T}$. In our results, we take *ϵ-proxy* coherence to be the uppermost bound of $\ln \frac{1}{1-\bar{\epsilon}}$ between source and target domains.

5 Neural Model

The critical components of a soft-structure, $\tilde{\mathcal{S}}_\sigma$, are its domain-encoder $\psi_\mathcal{S}$ and modular relation-decoders ϕ. Together these form an autoencoding architecture which, given a domain of images $\mathcal{S} \subset \mathbb{R}^{W\times H}$ and with d-dimensional latent space $\mathcal{Z} = \mathbb{R}^d$, converts sub-symbolic encodings from $\psi_\mathcal{S}$ into a modular relational representation via decodings for each $\phi_r, r \in \sigma$. Additionally, to retain information in $\mathcal{Z}$ pertaining to $\mathcal{S}$ which is beyond the requirements of ϕ, we include an additional domain-decoder, which produces domain reconstructions $\hat{\mathcal{S}}$. The overall neural model is depicted by Figure 1, where we use ψ^{enc}_S to refer to the domain-encoder and ψ^{dec}_S for the domain-decoder. To train the neural model, we assume a ground truth interpretation $I_{\mathcal{S}_\sigma}$ is given, allowing us to directly maximise Eqn. 3 via negative log-likelihood loss:

$$L^{\tilde{\mathcal{S}}_\sigma} = -\log p(\mathcal{S}_\sigma|\tilde{\mathcal{S}}_\sigma), \tag{9}$$

To obtain informative latent representations for $\mathcal{S}$, we use a Variational AutoEncoder (VAE), specifically the β-VAE, given its simplicity and demonstrated ability to separate distinct factors in the latent representation (Higgins et al., 2017; Burgess et al., 2017; Kingma & Welling, 2014). The β-VAE achieves this by optimising the ELBO objective, $L^{\text{ELBO}}_{\beta\text{-VAE}}$, but with an additional β scalar hyperparameter that can be thought of as a disentanglement pressure, forcing distinct explanatory factors to align with different axes of the latent space. We provide the full ELBO loss, with a detailed explanation, in Appendix C. We combine losses over each model component to give the following aggregate objective:

$$L^{\text{joint}} = L^{\text{ELBO}}_{\beta\text{-VAE}} - \lambda L^{\tilde{\mathcal{S}}_\sigma} \tag{10}$$

where λ is a scalar weighting parameter. Together with the $L^{\text{ELBO}}_{\beta\text{-VAE}}$, the choice of relation-decoder model will shape how the domain-encodings are structured (Gutiérrez-Basulto & Schockaert, 2018). Our evaluation considers a selection of relation-decoder models that cover a range of representational flexibility. Amongst these is our proposed low-complexity, but nonetheless expressive, Dynamic Comparator (DC) model. The overall DC model is composed of two modes, a distance-based measure, $\phi^\dagger_r$, that measures the distance between two inputs relative to a reference point, and a step-like function, $\phi^\ddagger_r$, that determines the sign of the difference between two points, optionally with an offset. Although we can use any functions that have the required characteristics for $\phi^\dagger$ and $\phi^\ddagger$, in this paper we use the following implementation:

$$\phi^{DC}_r(\boldsymbol{z}_i, \boldsymbol{z}_j) = a_{r,0} \cdot \phi^\dagger_r + a_{r,1} \cdot \phi^\ddagger_r \quad \text{where} \tag{11}$$

$$\phi^\dagger_r = f_0\big(-\eta_{r,0}(\|\boldsymbol{u}_r \odot (\boldsymbol{z}_i - \boldsymbol{z}_j + \boldsymbol{b}^\dagger_r)\|_2)\big) \quad \text{and} \quad \phi^\ddagger_r = f_1\big(\eta_{r,1} \cdot \boldsymbol{u}_r^\top (\boldsymbol{z}_i - \boldsymbol{z}_j + \boldsymbol{b}^\ddagger_r)\big). \tag{12}$$

Here $\boldsymbol{a}_r = \texttt{Softmax}(\boldsymbol{A}_r) \in (0,1)^2$ is an attention weighting between the two modes, f_0 and f_1 are an exponential and sigmoid function, respectively ; $\boldsymbol{u}_r = \texttt{Softmax}(\boldsymbol{U}_r) \in (0,1)^m$ is an attention mask which is applied to m-dimensional embeddings; $\boldsymbol{b}^\dagger_r, \boldsymbol{b}^\ddagger_r \in \mathbb{R}^m$ are learnable bias terms that enables an offset to each mode; and $\eta_{r,0} \in \mathbb{R}^+$ are non-negative and $\eta_{r,1} \in \mathbb{R}$ any-valued scalar terms, respectively. Lastly, $\odot$ denotes the Hadamard product and $\|\cdot\|_2$ is the $L2$-norm. The key innovation behind DC is its ability to model each of the ordinal relations whilst encouraging generalised consistency across the full latent subspace, as defined by each $\boldsymbol{u}_r$. This is achieved without explicit weight sharing, wherein relation-decoders discover parametric relationships between relations from the data. We provide an additional depiction and analysis of DC in Appendix D.1, which further illustrates these effects.

6 Experimental Design

In this section we describe an experimental design used to compare soft-structure coherence when choosing different relation-decoder implementations.

Partial Relation Transfer (PRT): The evaluation involves a proposed PRT task across two soft-structures $\tilde{\mathcal{X}}_\sigma$ and $\tilde{\mathcal{Y}}_\sigma$. Each shares a common signature σ and relation-decoders ϕ but have disjoint domains $\mathcal{X}$ and $\mathcal{Y}$, respectively. The experimental procedure involves first learning ϕ on source domain $\mathcal{X}$, together with its domain-specific autoencoder. In the second phase, we train a new domain-specific autoencoder on the target domain, $\mathcal{Y}$, alongside a selection of the now learned ϕ relation-decoders but with fixed-parameters. The selected relation-decoders act as training guides for $\psi^{\text{enc}}_{\mathcal{Y}}$, whilst held-out relation-decoders can be evaluated against zero-shot transfer performance. For domain $\mathcal{X}$ we employ the MNIST handwritten digits data set (LeCun & Cortes, 2010), and for domain $\mathcal{Y}$ we use a proposed BlockStacks data set, which includes singular multi-colored cube stacks of differing height, each containing one randomly positioned red cube (see Appendix B for further details and examples). The shared signature includes the ordinal relations, *i.e.* $\sigma = \{\mathsf{G}, \mathsf{E}, \mathsf{L}, \mathsf{S}, \mathsf{P}\}$, and is applied to digit ordering in MNIST and red cube position ordering in BlockStacks. We provide results against a theory of ordinality, as explored in Example 1 - we provide a formal specification of this theory in Appendix G. When guiding $\psi^{\text{enc}}_{\mathcal{Y}}$ to perform a similar mapping to $\psi^{\text{enc}}_{\mathcal{X}}$, *i.e.* from domain to a similar ordinal subspace as defined by ϕ, we could use the full ϕ set of relation-decoders. However, this is not necessary from a logical standpoint, as our system of relations can all be expressed in terms of isSuccessor. We therefore only employ the isSuccessor relation-decoder as a fixed-parameter guide for $\psi^{\text{enc}}_{\mathcal{Y}}$.

Neural model components: Together with DC, existing relation-decoder models compared here are: TransR (Lin et al., 2015), HolE (Nickel et al., 2016b), NTN (Socher et al., 2013). We addition-

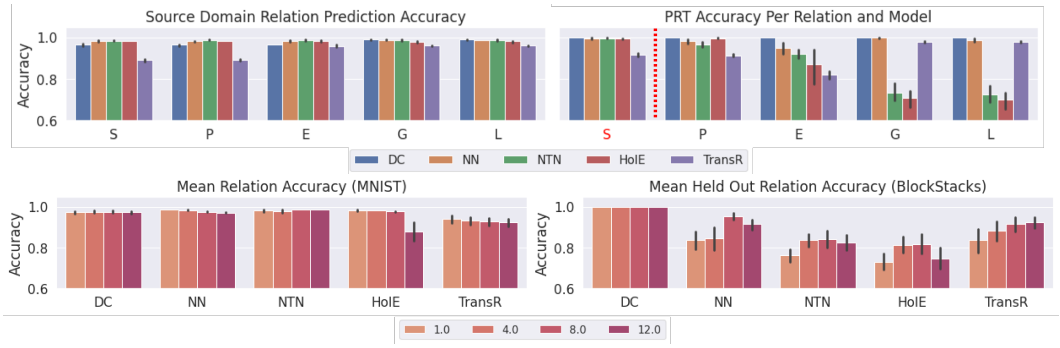

Figure 2: **[Top]** Relation-decoder prediction accuracy per relation and model, in the source (left) and target domains. Relations are abbreviated on the x-axis by {S: isSuccessor, P: isPredecessor, E: isEqual, G: isGreater, L: isLess}, with a red highlight identifying a relation included as a guide for $\psi^{enc}_{\mathcal{Y}_\sigma}$. **[Bottom]** Impact of different values of β for each relation-decoder (mean across all relations in the source domain (left) and mean for held-out relations only in the target domain (right). Notably, it can be seen that our model (DC) is not impacted while all other models show a decrease of accuracy in the target domain.

ally include a basic neural-network baseline, NN. To produce domain-encodings, all experiments use a β-VAE. We provide further details for all models in Appendix D. Due to a convergence issue when using a pretrained DC with fixed parameters, a flexible fitting procedure was necessary, in which we enable the DC parameters to train in the target domain, but with the additional loss term $\|\rho^* - \rho\|$, between pretrained ρ^* and untrained parameters ρ, respectively. In all cases we evaluated the final parameter values in the target domain and found them to be approximately equivalent to the ρ^*. We did not apply this method to the other models as they were all able to fit the isSuccessor relation in the target domain.

Hyperparameters: In the source domain we explore β values between $\{1, 4, 8, 12\}$, and set $\lambda = 10^3$ and in the target domain we first normalise losses (see Appendix D.4) and set $\beta = 10^{-4}$ and $\lambda = 10^{-2}$ as these produced good reconstructions whilst also ensuring optimisation against $L^{\tilde{\mathcal{Y}}_\sigma}$. In all experiments, we fix $\mathcal{Z} = \mathbb{R}^{10}$.

7 Key Results

In this section, Figure 2 firstly shows standard PRT prediction accuracies per relation in both the source and target domain. Figure 3 then presents consistency losses for three color-coded data splits: data-embeddings (blue), where all inputs are encodings of a domain's test data; interpolation (green), where we obtain an empirical mean and variance for the domain's data-embeddings and sample from a corresponding Gaussian distribution; and extrapolation (red), where we sample from regions strictly outside the smallest, axis aligned hyper-rectangle that encloses all data-points. Finally, Table 1 concludes with a clear ϵ-proxy coherence comparison between relation-decoders[2].

Relation-decoder PRT accuracy performance: Figure 2-top provides relation-decoder prediction accuracy in both the source MNIST (left), and target BlockStacks (right), domains. Key observations are that DC produces excellent PRT performance, whilst NN, NTN and HolE all see some degradation from their source accuracies on relations other than isSuccessor. TransR seems to maintain an target accuracy profile similar to its performance in the source domain, but this is significantly below the performance of other models in the source domain.We include β's impact on these performances in Figure 2-bottom. Barring DC which has little discernible change in either domain, PRT performance is significantly impacted by β in all models, but has little effect in the source domain. TransR shows a strong positive correlation between target domain accuracy and β, whereas the remaining models produce their best PRT performances with intermediate disentanglement pressure.

[2]We take ϕ_r prediction values above 0.5 to signify a truth prediction and those below 0.5 to signify falsity.

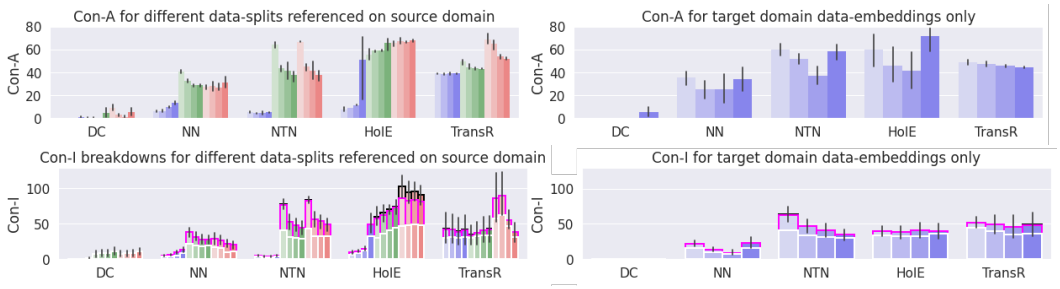

Figure 3: Modified **[Top]** Con-A values for each relation-decoder model, referenced to source (left) and target (right) domains (lower values better). **[Bottom]** Con-I values (lower values better) for each relation-decoder model referenced to source (left) and target (right) domains, where stacked bars are for formula: transitivity (white), asymmetry (magenta) and reflexivity (black). In all plots, darker color shades denote higher values of β, corresponding to greater disentanglement pressure from the β-VAE. In top-left and bottom plots, blue, green and red groups show results for data-embeddings, interpolation and extrapolation embeddings respectively (see main text for details).

Consistency Across (Con-A): To interrogate further how β affects each model, Figure 3-top presents consistency losses against formulae that constrains truth value assignments across relations, under a theory of ordinality, referred to as Con-A[3]. Results are referenced to both source (left) and target (right) domain embeddings. With reference to the source domain, we note that DC shows excellent Con-A in all regions. Most other models have worse interpolation and extrapolation consistency. Increasing β appears to improve interpolation and extrapolation performance for models NN, NTN and TransR, but there are indications that this trend does not persist into the largest $\beta = 12$ value. On the other hand, HolE shows a negative correlation between β and Con-A performance, across all data-splits. Although DC sustains impressive Con-A results for target domain data-embeddings (right), results for all other models are notably worse with respect to their source data-embeddings performances and are instead comparable with their interpolation or extrapolation results in the source domain. It may therefore be possible to anticipate/diagnose poor transfer performance by evaluating interpolation and extrapolation consistency, which suggests that DC's strong consistency generalisation indeed enables it to learn a transferable concept.

Consistency Individual (Con-I): Figure 3-bottom presents stacked consistency losses, for formula: transitivity (white), asymmetry (magenta) and reflexivity (black); results are averaged over individual relations and are together grouped under label Con-I, given that they refer to constraints on *individual* relations. Losses are again partitioned between source domain (left) and target domain (right). We firstly observe that DC and NN share the best overall Con-I performance profiles, with TransR following closely. DC and TransR both show comparable data-embedding versus interpolation/extrapolation performance, whereas NN, NTN and HolE suffer from degradation across these splits. Interestingly, these results show that: DC only suffers on transitivity, NN and TransR mainly struggle to model transitivity but show additional loss for asymmetry and HolE demonstrates difficulty in modelling each of the Con-I sub-stack. With regards to β's impact, it is not possible to determine a correlation for DC. However, NN and NTN demonstrate a negative correlation of β against overall Con-I, with comparable response for each underlying sub-stack. TransR shows a significant Con-I extrapolation improvement with increased β and HolE is for the most part adversely impacted as β is increased. Similar trends can be seen for target Con-I performance. However, notably, many models show improvements, in particular in Con-I (asymmetry). This could be due to more precise target domain data-embeddings, as a result of using a single pretrained relation-decoder.

ϵ-proxy coherence: Table 1 provides a comparison between optimal ϵ-proxy coherences achieved for each relation-decoder model, as defined in Section 4. Results are partitioned according to each consistency type and an Aggr(egate) value, which gives best summed consistency, together with best $\beta = \beta^*$ values. DC clearly outperforms all other models in ϵ-proxy coherence across all types. NN achieves strong aggregated ϵ-coherence compared with NTN, HolE and TransR outperforming across Con-A and Con-I-tr. Although NTN and HolE have similar aggregate ϵ-proxy coherence,

[3]Truth tables for each consistency formula is given in Appendix G

Table 1: ϵ-proxy coherence comparison, with respect to source and target data-embedding consistency levels. Results are reported with the corresponding $\beta = \beta^*$ setting (in parenthesis). The consistency loss abbreviations refer to: (A)cross, (tr)ansitivity, (asym)metry, (refl)exivity and (Aggr)egate, which gives the best obtained aggregate (summed) consistencies.

ϕ	Aggr.	(β^*)	Con-A	(β^*)	Con-I-tr	(β^*)	Con-I-asym	(β^*)	Con-I-refl	(β^*)
TransR	90.33	(8)	44.34	(12)	35.30	(8)	9.94	(8)	0.55	(8)
HolE	82.06	(8)	41.18	(8)	32.15	(4)	5.96	(1)	0.07	(8)
NTN	79.54	(8)	38.91	(8)	30.08	(12)	4.49	(12)	0.09	(12)
NN	34.09	(8)	24.78	(8)	7.24	(8)	3.88	(8)	**0.04**	(4)
DC	**0.34**	(1)	**0.07**	(1)	**0.18**	(1)	**0.00**	(1)	0.09	(1)

TransR performs generally worse. This may be caused by TransR producing weaker beliefs in comparison to other models, as this can result in a worse overall consistency level. Looking at β^* profiles, we see that most models achieve optimum aggregate ϵ-proxy coherence at $\beta = 8$, other than DC which performs better at $\beta = 1$. Overall, this is in agreement with the β profiles given by Figure 2-bottom (right). However, we can see that β^* profiles for Con-A based ϵ-proxy coherence are in more direct agreement - as TransR achieves its best at $\beta = 12$ - suggesting that Con-A invariance is more important to concept transfer.

8 Discussion and Concluding Remarks

In this work, we introduced the notion of a soft-structure, which can learn a structure over real-world domains, through the use of a domain-encoder coupled with modular relation-decoders. We subsequently provided formal definitions, defining what it means for a concept to be coherent in terms of domain-invariance of soft-structure consistency with respect to a theory. We then outlined a neural model and experimental procedure that together allowed us to investigate how concept coherence differs when choosing different implementations for underlying relation-decoders and its impact on concept transfer. Our results suggest that increasing regularisation over relation-decoder models, either in the form of disentanglement pressure or relation-decoder model capacity, seems to improve their ability to learn coherent concepts. Firstly, strong PRT transfer for DC and NN (given an appropriately high β setting) showed that both relation-decoder models are able to minimise Eqn. 9 in the source domain and retain good performance in the target domain. Consistency profiles over partial theories (subsets of the sentences that comprise the overall theory of ordinality), covering multiple data-splits, then further suggested that a relation-decoder's ability to retain consistency over interpolated/extrapolated regions with respect to the observed data-encodings (during training) greatly impact concept coherence. Finally, an ϵ-proxy coherence comparison showed that DC achieved excellent coherence with an aggregated ϵ-proxy of 0.34, which mirrors its strong PRT performance. NN achieved a score of 34.09, which, although significantly worse than DC, is a marked improvement over the remaining models. All in all, the empirical analysis in this work provides strong evidence towards the hypothesis that the transferability of a concept depends on its coherence, as measured by the retention of consistency across domains.

Limitations and future work: Firstly, this work only considered binary relations, and in particular unary relations, such as digit classification, are not considered. Additionally, we have only considered a fixed signature which is learned "all at once" in a source domain. In practical applications, however, it is quite possible that ordinality would be discovered gradually, either through incremental learning of the relations that form a 'complete' signature for ordinality, or through gradual refinement of pre-learned relations after being progressively exposed to different contexts in which ordinality applies. This necessitates a continual learning procedure which has not been explored in this work and would be useful as an addition in future work. Further, even in a single domain, ordinality can be applied to multiple properties (*e.g.* for BlockStacks we have: block stack height, block size, position of stacks, *etc.*) and future work can explore our framework's ability to initialize multiple instances of our signature, each applied to a different ordinal property. Lastly, we have only explored a signature for ordinality, whereas other fundamental properties are easy to find, such as periodic (*e.g.* rotation) and unordered categorical (*e.g.* shape) properties. These aspects are not explored in this work and would certainly be interesting for future work.

9 Ethics Statement

The authors declare that this work does not include any of the following: involvement of human subjects, sensitive data, harmful insights, methodologies and applications. The results, data sets and methodologies are objectively nondiscriminatory, unbiased and fair. This work does not breach any privacy or security guidelines or laws, nor any other legal restrictions. The authors declare that there are no conflicts of interest and/or external motivations through sponsorship.

10 Reproducibility Statement

To ensure reproducibility of this work, the experimental code has been open sourced, together with the proposed BlockStacks data set and rendering code;[4] references for all other employed data sets are provided in the main text. Hyperparameter configurations are specified in both Section 6 and Appendix D and the experimental procedure is detailed in Section 6.

References

Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. Boxe: A box embedding model for knowledge base completion. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6dbbe6abe5f14af882ff977fc3f35501-Abstract.html.

Masataro Asai. Photo-Realistic Blocksworld Dataset. *arXiv preprint arXiv:1812.01818*, 2018.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *CoRR*, abs/2012.13635, 2020. URL https://arxiv.org/abs/2012.13635.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2013.50.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems*, pp. 2787–2795. Curran Associates, Inc., Lake Tahoe, USA, 2013.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. In *Advances in Neural Information Processing Systems 30*, number Nips, Long Beach, CA, USA, 2017. URL http://arxiv.org/abs/1804.03599.

Junxiang Chen and Kayhan Batmanghelich. Robust ordinal VAE: employing noisy pairwise comparisons for disentanglement. *CoRR*, abs/1910.05898, 2019. URL http://arxiv.org/abs/1910.05898.

Junxiang Chen and Kayhan Batmanghelich. Weakly Supervised Disentanglement by Pairwise Similarities. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI*, New York, NY, USA, 2020.

Ricky T Q Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pp. 2615—-2625, Montreal, Quebec, Canada, 2018.

[4] Pending acceptance

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2172–2180, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html.

Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo. A Survey on Knowledge Graph Embedding: Approaches, Applications and Benchmarks. *Electronics*, 9(5):1–29, 2020. ISSN 20799292. doi: 10.3390/electronics9050750.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *CoRR*, abs/2012.05876, 2020. URL https://arxiv.org/abs/2012.05876.

Ivan Donadello, Luciano Serafini, and Artur d'Avila Garcez. Logic Tensor Networks for Semantic Image Interpretation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 1596—-1602, 2017.

Cian Eastwood and Christopher K I Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, {ICLR}*, Vancouver, BC, Canada, 2018.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *CoRR*, abs/2012.05208, 2020. URL https://arxiv.org/abs/2012.05208.

Víctor Gutiérrez-Basulto and Steven Schockaert. From Knowledge Graph Embedding to Ontology Embedding? An Analysis of the Compatibility between Vector Space Representations and Rules. 2018. doi: 1805.10461. URL http://arxiv.org/abs/1805.10461.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, {ICLR}*, Toulon, France, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv preprint arXiv:1812.02230*, 2018. doi: arXiv:1812.02230v1. URL http://arxiv.org/abs/1812.02230.

B. Inhelder and J. Piaget. *The early growth of logic in the child: classification and seriation*. Routledge and Kegan Paul, London, 1964.

Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. When crowds hold privileges: Bayesian unsupervised representation learning with oracle constraints. In *4th International Conference on Learning Representations, {ICLR}*, pp. 1–16, San Juan, Puerto Rico, 2016.

Seyed Mehran Kazemi and David Poole. Simple embedding for link prediction in knowledge graphs. *Advances in Neural Information Processing Systems*, 2018-December(Nips):4284–4295, 2018. ISSN 10495258.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Alberta, Canada, 2014. ISBN 1312.6114v10. doi: 10.1051/0004-6361/201527329.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations, {ICLR}*, Vancouver, BC, Canada, 2018.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, 40, 2017. ISSN 14691825. doi: 10.1017/S0140525X16001837.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In Blai Bonet and Sven Koenig (eds.), *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 2181–2187. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar R{\"{a}}tsch, Sylvain Gelly, Bernhard Sch{\"{o}}lkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning,{ICML}*, pp. 4114—-4124, Long Beach, California, USA, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-Supervised Disentanglement Without Compromises. *CoRR*, abs/2002.0, 2020.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning ICML*, pp. 807–814, 2010.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016a. ISSN 00189219. doi: 10.1109/JPROC.2015.2483592.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In Dale Schuurmans and Michael P. Wellman (eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 1955–1961. AAAI Press, 2016b. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484.

Maxwell I. Nye, Michael Henry Tessler, Joshua B. Tenenbaum, and Brenden M. Lake. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *CoRR*, abs/2107.02794, 2021.

J. B. Paris. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge University Press, 1994. ISBN 0-521-46089-1.

Jeffrey Paris and Alena Vencovská. *Pure Inductive Logic*. Perspectives in Logic. Cambridge University Press, 2015. doi: 10.1017/CBO9781107326194.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jean Piaget. *The Psychology of Intelligence*. Routledge and Kegan Paul, 2005. ISBN 0521781604. doi: 10.1093/acprof:oso/9780195150100.001.0001.

Ievgen Redko, Amaury Habrard, Emilie Morvant, Marc Sebban, and Younès Bennani. *Advances in Domain Adaptation Theory*. Elsevier, 2019. ISBN 978-1-78548-236-6.

Karl Ridgeway and Michael C Mozer. Learning Deep Disentangled Embeddings With the F-Statistic Loss. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pp. 185—-194, Montreal, Quebec, Canada, 2018.

Adam Santoro, Andrew K. Lampinen, Kory Mathewson, Timothy P. Lillicrap, and David Raposo. Symbolic behaviour in artificial intelligence. *CoRR*, abs/2102.03406, 2021.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling Relational Data with Graph Convolutional Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10843 LNCS(1):593–607, 2018. ISSN 16113349. doi: 10.1007/978-3-319-93417-4_38.

Luciano Serafini and Artur D.Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. In *Proceedings of the 11th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy'16) co-located with the Joint Multi-Conference on Human-Level Artificial Intelligence {(HLAI} 2016)*, New York, NY, USA, 2016.

Stewart Shapiro and Teresa Kouri Kissel. Classical Logic. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

Richard Socher, Danqi Chen, Christopher Manning, Danqi Chen, and Andrew Ng. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems*, pp. 926–934, 2013.

Xander Steenbrugge, Sam Leroux, Tim Verbelen, and Bart Dhoedt. Improving Generalization for Abstract Reasoning Tasks Using Disentangled Feature Representations. In *Neural Information Processing Systems (NeurIPS) Workshop on Relational Representation Learning*, Montreal, Canada, 2018. doi: http://arxiv.org/abs/1811.04784.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex Embeddings for Simple Link Prediction. In *Proceedings of the 33nd International Conference on Machine Learning, {ICML}*, pp. 2071–2080, New York, NY, USA, 2016. ISBN 9781510829008.

Théo Trouillon, Éric Gaussier, Christopher R. Dance, and Guillaume Bouchard. On inductive abilities of latent factor models for relational learning. *Journal of Artificial Intelligence Research*, 64: 21–53, 2019. ISSN 10769757. doi: 10.1613/jair.1.11305.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are Disentangled Representations Helpful for Abstract Visual Reasoning? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, pp. 14222—-14235, Vancouver, BC, Canada, 2019.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12): 2724—-2743, 2017. ISSN 10414347. doi: 10.1109/TKDE.2017.2754499.

A Related Work

Relational representations play a prominent role in Knowledge Graph Embedding, wherein sets of relation-decoders are jointly learned in order to obtain a semantic latent representation for data points (Socher et al., 2013; Trouillon et al., 2016; 2019; Bordes et al., 2013; Nickel et al., 2016a; Wang et al., 2017; Dai et al., 2020; Kazemi & Poole, 2018; Abboud et al., 2020). Although these typically do not use a shared autoencoder as we do in this paper, Schlichtkrull et al. (2018) did adopt an autoencoding framework, where a graph neural network is used as the encoder, however they did not work with visual data and the model was only applied to single data sets. Similarly, disentanglement is also concerned with semantic representation learning (Bengio et al., 2013) , and has been explored using a variety of methods including both Generative Adversarial Networks (Chen et al., 2016) and VAEs (Burgess et al., 2017; Higgins et al., 2017; Chen et al., 2018; Ridgeway & Mozer, 2018; Eastwood & Williams, 2018; Kumar et al., 2018; Locatello et al., 2019). Disentangled representations have been evaluated in terms of there transferability in (van Steenkiste et al., 2019; Steenbrugge et al., 2018; Locatello et al., 2020). A bridge between these two fields, wherein relation-decoders are employed as a semi-supervision to VAEs can be found in (Karaletsos et al., 2016; Chen & Batmanghelich, 2020; 2019), where (Karaletsos et al., 2016) use multiple relation-decoders but compute a triplet comparison based query and (Chen & Batmanghelich, 2020; 2019) only include a single binary relation and use function forms that are not sufficient to model the full set of relations that we include in this work. Neither presents a comprehensive analysis of resulting concept coherence. Lastly, we note that our experimental setup is most remnant of domain adaptation (Redko et al., 2019). To the best of our knowledge, no work has compared relation-decoders in their ability to learn coherent concepts, as measured by their consistency across domains.

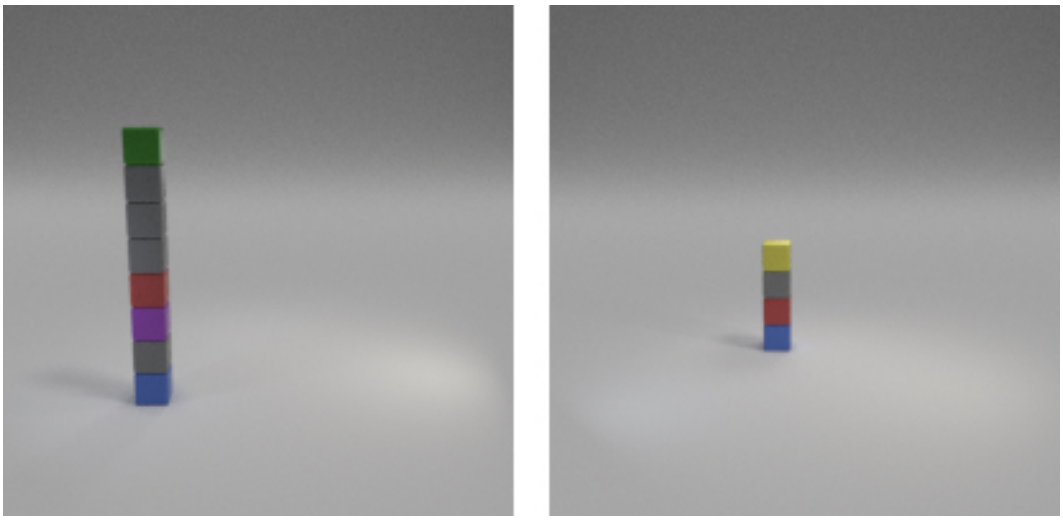

Figure 4: Example of two *BlockStacks* data set images.

B BLOCKSTACKS DATASET DESCRIPTION

The *BlockStacks* dataset consists of 12,000 images (200×200 pixels but resized in code to 128×128) of individual block stacks, of varying height (between 1-10 blocks), block colors (uniformly sampled from options: { gray, blue, green, brown, purple, cyan, yellow}) and position (uniformly sampled from x, y range (-3,-3) to (3,3)), but with the requirement that each instance consists of a single red block at a random height (see Figure 4 for example images). These were rendered using the CLEVR rendering agent with the help of code from (Asai, 2018). The dataset is divided into 9000:1500:1500 train, validation and test splits.

C EXPLANATION OF THE β-VAE

The VAE is derived by introducing an approximate posterior $q_\alpha(\boldsymbol{Z}|\boldsymbol{X})$, from which a lower bound (commonly referred to as the Evidence LOwer Bound (ELBO)) on the true marginal $\log p_\theta(\boldsymbol{X})$ can be obtained by using Jensen's inequality (Kingma & Welling, 2014). The VAE maximises the log-probability by maximising this lower bound, given by:

$$L^{\text{ELBO}}_{\beta\text{-VAE}} = \mathbb{E}_{q_\alpha(\boldsymbol{Z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{Z})] - \beta D_{KL}(q_\alpha(\boldsymbol{Z}|\boldsymbol{X})\|p_\theta(\boldsymbol{Z})), \tag{13}$$

where $q_\alpha(\boldsymbol{Z}|\boldsymbol{X})$ is typically modelled as a neural-network encoder with parameters α. Similarly $p_\theta(\boldsymbol{X}|\boldsymbol{Z})$ is often modelled as a neural-network decoder with parameters θ and is calculated as a Monte Carlo estimation. A reparameterization trick is used to enable differentiation through an otherwise undifferentiable sampling from $q_\alpha(\boldsymbol{Z}|\boldsymbol{X})$ (see (Kingma & Welling, 2014)). In the β-VAE (Higgins et al., 2017; Burgess et al., 2017), an additional β scalar hyperparameter was added as it was found to influence disentanglement through stronger distribution matching pressure with respect to the prior $p_\theta(\boldsymbol{Z})$, where this prior is typically set to an isotropic zero-mean Gaussian $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$). When $\beta = 1$ we obtain the standard VAE objective (Kingma & Welling, 2014).

D MODEL DESCRIPTIONS

In this section we firstly present an in-depth analysis of the key innovations presented by DC which provides insight into how it can learn a coherent notion of ordinality. We then provide model details for each of the compared relation-decoders in the main results and the β-VAE architecture that we employ for each data set.

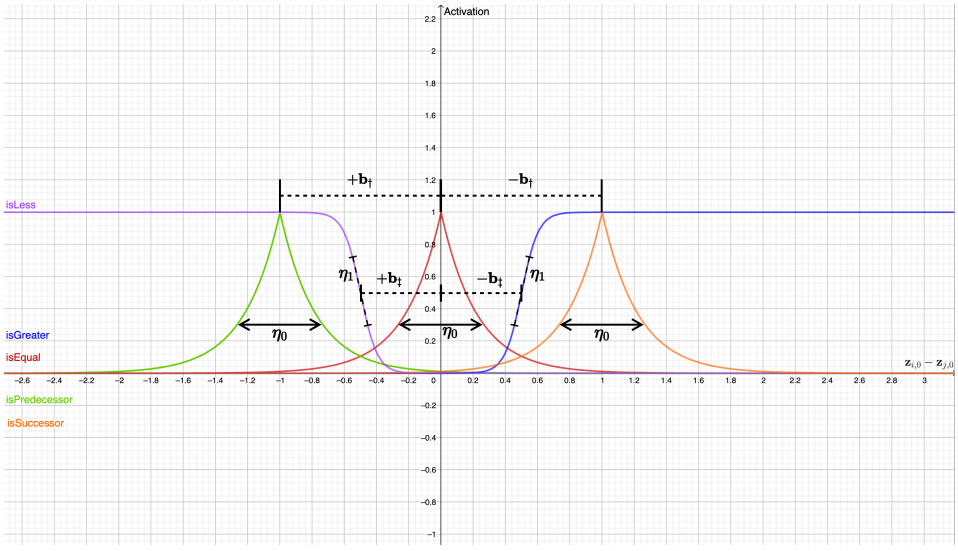

Figure 5: Depiction of a set of DC relation-decoders for binary relations isGreater, isLess, isEqual, isSuccessor and isPredecessor. Each DC relation-decoder (for each relation) has a one-hot mask, $\boldsymbol{u}_r$ (that is in this example the same across relations), which ensures only the zeroth dimensions of the embedding arguments are compared, giving $\boldsymbol{z}_{i,0}$ and $\boldsymbol{z}_{j,0}$.

D.1 Dynamic Comparator Analysis

Figure 5 depicts how DC is able to learn the isGreater, isLess, isEqual, isSuccessor and isPredecessor family of binary ordinal relations, assuming each corresponding relation-decoder has learned a common one-hot mask on the zeroth dimension *i.e.* $\boldsymbol{u}_{\mathsf{G}} = \boldsymbol{u}_{\mathsf{E}} = \ldots = \boldsymbol{u}_{\mathsf{P}} = [1, \ldots, 0]$, such that activations only depend on the $\boldsymbol{z}_{i,0} - \boldsymbol{z}_{i,1}$ difference. An important capability of DC is its ability to *select*, via $\boldsymbol{a}_r$ an appropriate functional mode, either $\phi_r^{\dagger}$ or $\phi_r^{\ddagger}$, depending on the type of relation it needs to model. As shown by Figure 5, isEqual exhibits its reflexive, symmetric and transitive characteristics, whilst isGreater and isLess both carry transitivity but are asymmetric and irreflexive. Furthermore, the use of a subtraction between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ (which, via mask $\boldsymbol{u}$ ends up only being a subtraction between their zeroth dimensions) leads to a relative comparison, not an absolute comparison, which generalises to arbitrary $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ sampled from anywhere in $\mathcal{Z}$.

Note that there is no built in parameter sharing, meaning each relation-decoder (for each individual relation r) is trained independently and has its own set of $\boldsymbol{a}_r, \boldsymbol{u}_r, \eta_{r,0}, \eta_{r,1}, \boldsymbol{b}_r^{\dagger}$ and $\boldsymbol{b}_r^{\ddagger}$ parameters. However, our experiments show that DC reliably obtains settings such that *e.g.* $\boldsymbol{u}_{\mathsf{G}} = u_{\mathsf{E}}$, or $\boldsymbol{a}_{\mathsf{G}} = \boldsymbol{a}_{\mathsf{L}} = [0, 1]$, or $\boldsymbol{b}_{\mathsf{G}}^{\ddagger} = -\boldsymbol{b}_{\mathsf{L}}^{\ddagger}$ and so on. DC is thus able to discover the interdependencies between families of relations. By learning to indirectly 'tie' together parameters in this way, whilst still being expressive enough to model each type of relation, DC can facilitate a data-driven binding between relation-decoder outputs. This helps ensure consistent generalisation across a latent subspace, as defined by the common/overlapped $\boldsymbol{u}_r$ masks.

D.2 Relation-Decoder Implementations

TransR (Lin et al., 2015):

$$\phi_r^{\text{TransR}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \|\boldsymbol{h}_r + \boldsymbol{r} - \boldsymbol{t}_r\|_2^2$$

with,

$$\boldsymbol{h}_r = \boldsymbol{M}_r \boldsymbol{z}_i \quad \text{and} \quad \boldsymbol{t}_r = \boldsymbol{M}_r \boldsymbol{z}_j.$$

where for $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathbb{R}^{d_z}$ vectors, $\boldsymbol{M}_r \in \mathbb{R}^{d_z \times d_z}$ and $\boldsymbol{r} \in \mathbb{R}^{d_z}$. As we want to obtain a [0,1] output, we modify TransR through $\phi_r^{\text{TransR}^+} = \sigma(c - \phi_r^{\text{TransR}})$, where σ is a sigmoid function and c is a scalar that ensures that at $\phi_r^{\text{TransR}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = 0$, then $\phi_r^{\text{TransR}^+}(\boldsymbol{z}_i, \boldsymbol{z}_j) \approx 1$. In all experiments we set $c = 10$.

NTN (modified version of (Socher et al., 2013) from (Donadello et al., 2017; Serafini & Garcez, 2016)):

$$\phi_r(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \sigma\big(\boldsymbol{u}_r^\top [\tanh(\boldsymbol{z}^{c\top} \boldsymbol{M}_r \boldsymbol{z}^c + \boldsymbol{V}_r \boldsymbol{z}^c + \boldsymbol{b}_r)]\big) \tag{14}$$

where $\boldsymbol{u}_r \in \mathbb{R}^k, \boldsymbol{M}_r \in \mathbb{R}^{n \cdot d_z \times n \cdot d_z \times k}, \boldsymbol{V}_r \in \mathbb{R}^{k \times n \cdot d_z)}$ and $\boldsymbol{b}_r \in \mathbb{R}^k$. The only hyperparameter to consider is k, which controls the NTN's capacity - in all experiments, we set this to 1. If $k > 1$, $\boldsymbol{z}^{c\top} \boldsymbol{M}_r \boldsymbol{z}^c$ produces a k-dimension vector by applying the bilinear operation to each of the k $\boldsymbol{M}_r$ slices. Here $\boldsymbol{z}^c \in \mathbb{R}^{n \cdot d_z}$ is a concatenation of the inputs $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, which was introduced in (Donadello et al., 2017; Serafini & Garcez, 2016). In contrast, the original NTN (see (Socher et al., 2013)) is only applicable to binary relations and does not include the outer sigmoid.

HolE (Nickel et al., 2016b):

$$\phi_r^{\text{HolE}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \sigma(\boldsymbol{r}^\top (\boldsymbol{z}_i \star \boldsymbol{z}_j))$$

where $\boldsymbol{r} \in \mathbb{R}^{d_z}$ and $\star : \mathbb{R}^{d_z} \times \mathbb{R}^{d_z} \to \mathbb{R}^d$ denotes the circular correlation operator and is given by,

$$[\boldsymbol{z}_i \star \boldsymbol{z}_j]_k = \sum_{m=0}^{d-1} z_{i,m} z_{j,(k+m) \mod d}$$

NN: a simple four-layer neural-network with layer sizes $l_{\text{in}} = 2d_z, l_1 = 2d_z$ and $l_2 = d_z$, with ReLU activations (Nair & Hinton, 2010). The final output layer, l_{out}, is a single value passed through a sigmoid function, to bound the output within (0,1).

D.3 β-VAE CONFIGURATION

The model configurations used for both *MNIST* and *BlockStacks* data sets are given in Table 2.

D.4 L^{joint} CONFIGURATION

In the source domain, we vary β values between $\{1, 4, 8, 12\}$ and fix $\lambda = 10^3$. In the target domain, we fix β to 10^{-4} and $\lambda = 10^{-2}$ and normalise the $\mathcal{L}^{ELBO}_{\beta\text{-VAE}}$ reconstruction term by dividing by a factor $\frac{1}{\sqrt{H \cdot W \cdot C}}$, for height H, width W and color channels C, and normalize the distribution matching term by a factor $\frac{1}{d_z}$, for latent representation size d_z.

To train relation-decoders over a given domain $\mathcal{S}$, it is necessary to supervise estimates of $\phi_r(\psi^{enc}_{\mathcal{S}}(O)), O \in \mathcal{S}^2$, against corresponding ground-truth labels, $\gamma^r_{O,\mathcal{S}_\sigma}$. However, doing so for every $O \in \mathcal{S}^2$ can easily become intractable and we instead only sample a subset of possible $\mathcal{S}^2$ tuples. Our sampling strategy involves first selecting a ratio $R = \frac{|\mathcal{B}|}{|\mathcal{S}|}$ where $\mathcal{B} \subset \mathcal{S}^2$ is a set of O tuples. We then sample relation-decoder specific subsets $\mathcal{B}_r$ where $|\mathcal{B}_r| = \frac{|\mathcal{B}|}{|\sigma|}$, to ensure a balanced distribution of tuples between relation-decoders. Furthermore, we ensure that each $\mathcal{B}_r$ contains a balanced ratio of $\gamma^r_{O,\mathcal{S}_\sigma} = 1$ versus $\gamma^r_{O,\mathcal{S}_\sigma} = 0$ instances. We found that each $|\mathcal{B}_r|$ set can be small without jeopardising the final relation-decoder performance level, allowing us to use $R = 1$ for MNIST experiments and $R = 3$ for BlockStacks experiments.

Finally, in all experiments we use a β-VAE trained for up to 300,000 steps, following accepted practice from (Locatello et al., 2019; Steenbrugge et al., 2018), together with any included relation-decoders. However, to ensure computation efficiency across experiments, we employ an early stopping procedure, where if the validation score does not increase over 30 and 120 training epochs for MNIST and Blockstacks experiments, respectively, we end the training early.

E SUPPLEMENTARY RESULTS

β effect on intrinsic relation-decoder characteristics: We have seen how β impacts PRT accuracy but it is not clear how this is facilitated. To understand the way in which β affects each relation-decoder we produce a *gradient-conformity* evaluation, based on the intuition that collections of relation-decoder outputs will have to shift together in order to maintain consistency, facilitated

Table 2: Specification of our β-VAE encoder and decoder model parameters, for both 28×28 (top) and 128×128 (bottom) size input data. I: Input channels, O: Output channels, K: Kernel size, S: Stride, P: Padding, A: Activation

Encoder
Input: $28 \times 28 \times N_C = 1$
Layer_ID ; I ; O ; K ; S ; P ; A
Conv2d_1 ; N_C ; 32 ; 4×4 ; 2 ; 1 ; ReLU
Conv2d_2 ; 32 ; 32 ; 4×4 ; 2 ; 1 ; ReLU
Conv2d_3 ; 32 ; 64 ; 3×3 ; 2 ; 1 ; ReLU
Conv2d_4 ; 64 ; 64 ; 2×2 ; 2 ; 1 ; ReLU
Layer_ID ; Num Nodes : In - Out ; A
FC_z ; 576 - 144 ; ReLU
FC_z_mu ; 144 - 10 ; None
FC_z_logvar ; 144 - 10 ; None

Decoder
Input: $\mathbb{R}^{10}$
Layer_ID ; Num Nodes : In - Out ; A
FC_z ; 10 - 144 ; ReLU
FC_z_mu ; 144 - 576 ; ReLU
Layer_ID ; I ; O ; K ; S ; P ; A
UpConv2d_1 ; 64 ; 64 ; 2×2 ; 2 ; 1 ; ReLU
UpConv2d_2 ; 64 ; 32 ; 3×3 ; 2 ; 1 ; ReLU
UpConv2d_3 ; 32 ; 32 ; 4×4 ; 2 ; 1 ; ReLU
UpConv2d_4 ; 32 ; N_C ; 4×4 ; 2 ; 1 ; Sigmoid

Encoder
Input: $128 \times 128 \times N_C = 3$
Layer_ID ; I ; O ; K ; S ; P ; A
Conv2d_1 ; N_C ; 32 ; 4×4 ; 2 ; 1 ; ReLU
Conv2d_2 ; 32 ; 32 ; 4×4 ; 2 ; 1 ; ReLU
Conv2d_3 ; 32 ; 64 ; 4×4 ; 2 ; 1 ; ReLU
Conv2d_4 ; 32 ; 64 ; 4×4 ; 2 ; 1 ; ReLU
Conv2d_5 ; 64 ; 64 ; 4×4 ; 2 ; 1 ; ReLU
Layer_ID ; Num Nodes : In - Out ; A
FC_z ; 1024 - 256 ; ReLU
FC_z_mu ; 256 - 10 ; None
FC_z_logvar ; 256 - 10 ; None

Decoder
Input: $\mathbb{R}^{10}$
Layer_ID ; Num Nodes : In - Out ; A
FC_z ; 10 - 256 ; ReLU
FC_z_mu ; 256 - 1024 ; ReLU
Layer_ID ; I ; O ; K ; S ; P ; A
UpConv2d_1 ; 64 ; 64 ; 4×4 ; 2 ; 1 ; ReLU
UpConv2d_2 ; 64 ; 32 ; 4×4 ; 2 ; 1 ; ReLU
UpConv2d_3 ; 32 ; 32 ; 4×4 ; 2 ; 1 ; ReLU
UpConv2d_4 ; 32 ; 32 ; 4×4 ; 2 ; 1 ; ReLU
UpConv2d_5 ; 32 ; N_C ; 4×4 ; 2 ; 1 ; Sigmoid

via a certain conformity in their gradients of input against output. For instance, suppose we have $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ such that $\phi_G(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx 0$ and $\phi_E(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx 1$. If we then take $\boldsymbol{x}_i, \boldsymbol{x}_k \in \mathcal{X}$ such that $\phi_G(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx 1$, then we must have $\phi_E(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx 0$. In fact, any time ϕ_G outputs close 1, we require ϕ_L and ϕ_E output close to 0, since only one of these three relations can be true for any common arguments. At the decision boundaries, we require that their collective outputs conform such that they are each modified appropriately, ensuring that they continue to satisfy any applicable constraints. We therefore compute a gradient based analysis for arbitrary relation-decoder inputs against the overall $\mathcal{T}$ belief state, which should be consistently ≈ 1, to see how much this varies as β is increased. Gradient-conformity (GC) is calculated as:

$$GC = \left| \frac{\boldsymbol{d}_i^T \boldsymbol{d}_j}{||\boldsymbol{d}_i||_2 \, ||\boldsymbol{d}_j||_2} \right|, \quad \text{where } \boldsymbol{d}_i = \left. \frac{d\phi_{r_i}}{d\boldsymbol{z}^c} \right|_{\boldsymbol{z}^c = \boldsymbol{z}_n^c} \text{ and } \boldsymbol{d}_j = \left. \frac{d\phi_{r_j}}{d\boldsymbol{z}^c} \right|_{\boldsymbol{z}^c = \boldsymbol{z}_n^c}, \quad \forall i \neq j \tag{15}$$

where $\boldsymbol{z}_n^c$ is the concatenation of the nth sample of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ from the latent space (and, more specifically, a particular data split). Figure 6 presents source domain referenced GC measures for each model, with the same data split schematic as in Figure 3. We see that for DC, GC is close to 1 for all β with no discernible change. All other models show a weaker GC with positive correlation between GC and β. TransR and NN achieve significantly higher GC than NTN and HolE. It appears that models that achieve a GC greater than 0.5 perform better at the overall PRT learning task.

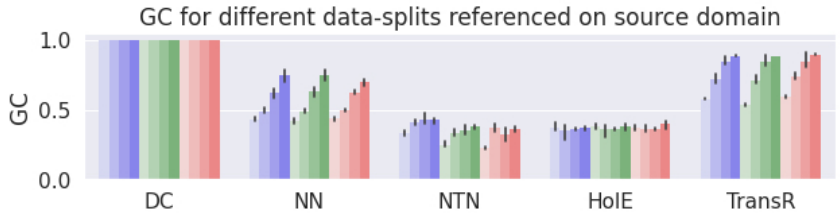

Figure 6: GC values (higher values better), for each relation-decoder model referenced to source domain. Darker color shades denote higher values of β, corresponding to greater disentanglement pressure from the β-VAE. Blue, green and red groups show results for data-embeddings, interpolation and extrapolation embeddings respectively (see main text for further details regarding the data splits).

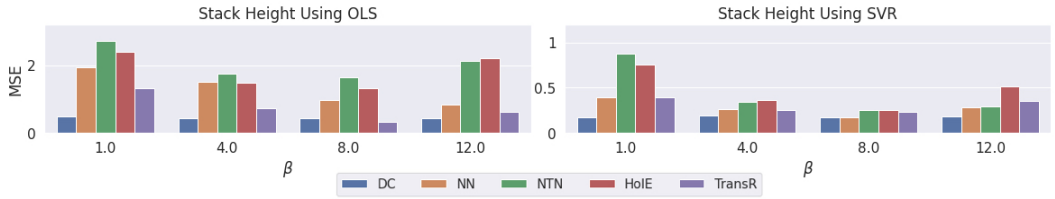

Figure 7: Analysis of domain-specific information retention by the β-VAE when using different relation-decoders for ordinality relation decoding. We attempt to predict the overall BlockStacks stack height on the final fixed embeddings obtained after isSuccessor relation-decoder alignment.

F How does each model impact the retention of domain-dependent information

Figure 7 shows results for BlockStacks overall block height prediction accuracy when training on fixed encodings of each block stack, after isSuccessor relation-decoder alignment as been applied (using a pretrained fixed-parameter relation-decoder). Note β is fixed in the target domain, so the only moving part are the choices of pretrained models which have been previously trained with varied *source* β values. Note also that DC has an unfair advantage here, as the steered fitting approach allows more flexibility to the VAE learning phase - for this reason the result is only included in the appendix. Since we are interested in capturing general representations that encode both domain-dependent and domain-independent information, we use each target encoder ψ^t_{enc} obtained from each PRT experiment and produce encodings for the full BlockStacks test set. The resulting encodings are then divided into a new train and test subset, used to train both a *Sci-Kit Learn* Linear regressor and Support Vector Machine regressor with a RBF kernel (Pedregosa et al., 2011). We present the resulting Mean Squared Errors (MSE) in Figure 7, with Ordinary Least Squares (OLS) (a) and Support Vector Regression (SVR) (b).

There are a number of noteworthy details: firstly, DC shows no dependence on β and leads to a lower MSE across all settings; second, excluding DC, for all models we observe an optimum MSE at $\beta = 8$, with TransR reaching DC MSE performance for OLS and NN doing the same for SVR. These results indicate that lower MSE can be obtained by using non-linear regression, which indicates that to some degree, the block stack height factor is not encoded linearly, regardless of selected model. Next, by contrasting with Figure 6, these results suggest that models with higher GC lead to embeddings that are more amenable to domain-specific factor prediction. However, the parabolic trend, where increasing β to 12 leads to an increase in error, is in agreement with Figure 2-bottom-right, which showed that most models do not improve at PRT for the largest β. This is perhaps due to a loss of mutual information between input and latent representation, as the distribution matching loss outweighs reconstruction in the $L^{\text{ELBO}}_{\beta\text{-VAE}}$.

G Specification for theory of ordinality

To support our claim that we can use only the isSuccessor relation as the target encoder guide due to its logical relationship the remaining relations, we include here the logical clauses:

$$\begin{aligned}
\forall i,j,k \;\; (\mathsf{isSuccessor}(i,j) \wedge \mathsf{isSuccessor}(k,j) &\rightarrow \mathsf{isEqual}(i,k)) \\
\forall i,j \;\; (\mathsf{isSuccessor}(i,j) &\rightarrow \mathsf{isGreater}(i,j)) \\
\forall i,j,k \;\; (\mathsf{isSuccessor}(i,j) \wedge \mathsf{isGreater}(j,k) &\rightarrow \mathsf{isGreater}(i,k)) \\
\forall i,j \;\; (\mathsf{isSuccessor}(i,j) &\leftrightarrow \mathsf{isPredecessor}(j,i)) \\
\forall i,j \;\; (\mathsf{isPredecessor}(i,j) &\rightarrow \mathsf{isLess}(i,j)) \\
\forall i,j,k \;\; (\mathsf{isPredecessor}(i,j) \wedge \mathsf{isLess}(j,k) &\rightarrow \mathsf{isLess}(i,k)).
\end{aligned}$$

Therefore, by knowing all of the successor relations between data instances, it should be possible to infer the remaining relationships that they share.

For completeness, we provide the truth tables for each of the sub-theories that our consistency losses evaluate against. We only include configurations that are valid under the constraints, indicated by $\subset \mathcal{T} = T$, where this notation highlights the fact each incomplete set of constraints form a subset of the overall theory $\mathcal{T}$.

Firstly, the truth-table that describes constraints shared between relation truth-values is given by the following, $\forall i, j$:

$\mathsf{G}(i,j)$	$\mathsf{E}(i,j)$	$\mathsf{L}(i,j)$	$\mathsf{S}(i,j)$	$\mathsf{P}(i,j)$	$\subset \mathcal{T}$
T	F	F	F	F	T
T	F	F	T	F	T
F	T	F	F	F	T
F	F	T	F	F	T
F	F	T	F	T	T

where we use the same relation abbreviations as in the main text results.

Next, we provide each of the three consistency individual (Con-I) truth-tables. These are referred to as being "individual" due to the fact that they describe constraints applied to the truth-state of a single relation. For transitivity, given by the rule *e.g.* $\mathsf{G}(i,j) \wedge \mathsf{G}(j,k) \rightarrow \mathsf{G}(i,k)$, we have that $\forall i, j$:

$\mathsf{G}(i,j)$	$\mathsf{G}(j,k)$	$\mathsf{G}(i,k)$	$\subset \mathcal{T}$
F	F	F	T
F	F	T	T
T	F	F	T
T	F	T	T
F	T	F	T
F	T	T	T
T	T	T	T

(16)

For asymmetry, where $\mathsf{S}(i,j) \rightarrow \neg\mathsf{S}(j,i)$, we have $\forall i, j$:

$\mathsf{S}(i,j)$	$\mathsf{S}(j,i)$	$\subset \mathcal{T}$
F	F	T
T	F	T
F	T	T

(17)

.

Finally, for reflexivity, given by $\mathsf{E}(i,i) \rightarrow \top$ (in this case describing that an object is always equal to itself) we have $\forall i$:

$\mathsf{E}(i,i)$	$\subset \mathcal{T}$
T	T

(18)

Truth-table matrices for each of the above truth-tables can be obtained by replacing T with 1 and F with 0. We provide the full set of individual constraints that are applicable to each relation covered in this paper are given by Table 3.

Table 3: Characteristic properties of ordinal relations.

Relation	asymmetric	transitive	reflexive
G	Y	Y	N
E	N	Y	Y
L	Y	Y	N
S	Y	N	N
P	Y	N	N

H EXPANDED ϵ-PROXY DERIVATION

In this section, we present the expanded justification for reporting $-\ln \bar{\epsilon}$ consistency and coherence as a proxy for ϵ-consistency/coherence as defined in Section 3. For notational clarity, in the following we omit $\psi_{\mathcal{S}}$, such that $\phi_r(\psi_{\mathcal{S}}(O))$ is abbreviated to $\phi_r(O)$.

In the following, we make no assumptions about the sizes of domain $\mathcal{S}$, signature σ and arities of each $r \in \sigma$. Further, we take $\mathcal{T}$ to be an arbitrary theory over σ consisting of universally quantified formula, and the validity of each ground instances of atomic formula with respect to $\mathcal{T}$, can be expressed by a single ground truth-table matrix, $\mathbf{T} \in \{0,1\}^{K_0 \times K_1 \times K_2}$, wherein each slice, $\mathbf{T}_{k,:,:}$ gives a unique grounding of domain objects to the variables, v, required by $\mathcal{T}$. For each grounding of the $K_0 = |\mathcal{S}|^{|v|}$ possible groundings, there are $K_1 = 2^l$ unique truth-assignments to the l atomic formulae that constitute $\mathcal{T}$, giving $K_2 = l + 1$ assignments per $\mathbf{T}_{k,t,:}$ row - one per atomic formulae and an additional value to denotes whether the particular row satisfies $\mathcal{T}$. $\boldsymbol{T}$ can be obtained by taking any truth-table from the previous section and switching true (T) for 1 and false (F) for 0, and producing K_0 copies for each assignment of domain elements to the variables. Given this truth-table matrix, notice that a structure $\mathcal{S}_\sigma$ can be composed by selecting a single row of $\boldsymbol{T}$ for each grounding (kth slice), giving a vector $\boldsymbol{c}_{kt} = \boldsymbol{T}_{k,t,1:l}$. If the structure is a model of $\mathcal{T}$, *i.e.* $\mathcal{S}_\sigma \in \mathcal{M}_{\mathcal{S}}^{\mathcal{T}}$, then only rows with $\boldsymbol{T}_{k,t,K_2} = 1$ are allowed. Taking t^+ to be the set of rows such that $\boldsymbol{T}_{k,t,K_2} = 1$ (which is identical for each k) for $t \in t^+$, we can then rewrite $\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma}$ in terms of samples from $\boldsymbol{T}$:

$$\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma} = \sum_{\mathcal{S}_\sigma \in \mathcal{M}_{\mathcal{S}}^{\mathcal{T}}} \prod_{r \in \sigma} \prod_{O \in \mathcal{S}^{\mathsf{ar}(r)}} \phi_r(O)^{\gamma^r_{O,\mathcal{S}_\sigma}} (1 - \phi_r(O))^{1-\gamma^r_{O,\mathcal{S}_\sigma}} \qquad \text{(Eqn. 3)}$$

$$= \sum_{\mathcal{S}_\sigma \in \mathcal{M}_{\mathcal{S}}^{\mathcal{T}}} \prod_{k=1}^{K_0} \sum_{t \in t^+} \mathbf{1}_{t'_{k,\mathcal{S}_\sigma}}(t) \prod_{m=1}^{l} f(\phi_{r^m}, O_{km}, c_{ktm})^{N(\phi_{r^m}, O_{km}, c_{ktm}, \mathcal{S}_\sigma)^{-1}} \qquad (19)$$

with

$$f(\phi_{r^m}, O_{km}, c_{ktm}) = \phi_{r^m}(O_{km})^{c_{ktm}} \left(1 - \phi_{r^m}(O_{km})\right)^{1-c_{ktm}}. \qquad (20)$$

In the above, $\mathbf{1}_{t'_{k,\mathcal{S}_\sigma}}(t)$ is an indicator function which equals 1 if $t = t'_{k,\mathcal{S}_\sigma}$ and 0 otherwise, for active row $t'_{k,\mathcal{S}_\sigma}$ under structure $\mathcal{S}_\sigma$ and grounding k. $\mathbf{1}_{t'_{k,\mathcal{S}_\sigma}}(t)$ has the role of only including the *single* summand where t corresponds with $t'_{k,\mathcal{S}_\sigma}$. $N(\phi_{r^m}, O_{km}, c_{ktm}, \mathcal{S}_\sigma)$ is a function that counts the number of repeat products of term $f(\phi_{r^m}, O_{km}, c_{ktm})$, such that the appropriate root can be applied. We use r^m to denote the relation for atomic formula at column m and O_{km} its corresponding arguments, under grounding k; and we use c_{ktm} to denote the truth-assignment of the atomic formula for column m, as designated by row t.

At this point, we are left with an expression for $\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma}$ in terms of truth-table matrix $\boldsymbol{T}$ entries, which is more reminiscent of $L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma)$ as defined in Section 4. However, we must go further to expose the relationship between $\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma}$ and $L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma)$ for arbitrary $\mathcal{T}$ expressed by $\boldsymbol{T}$. We will now show that the consistency loss $L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma)$ gives the negative log-likelihood of satisfying $\mathcal{T}$ given a grounding $k \in \{1, \ldots, K_0\}$, which can be further seen as a relaxation of $\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma}$ to sum over all rows $t \in t^+$ and without normalising via the $N(\phi_{r^m}, O_{km}, c_{ktm}, \mathcal{S}_\sigma)^{-1}$ exponent. With Boolean random variable $B_{\mathcal{T}}$ denoting whether $\mathcal{T}$ is ($b_{\mathcal{T}} = 1$) or is not ($b_{\mathcal{T}} = 0$) satisfied, the consistency loss for a soft-structure $\tilde{\mathcal{S}}_\sigma$ against theory $\mathcal{T}$ is given by,

$$L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma) = \mathbb{E}_{k \sim U[\{1,\ldots,K_0\}]}[H(p(B_{\mathcal{T}}|\mathcal{S}_\sigma, k), p(B_{\mathcal{T}}|\tilde{\mathcal{S}}_\sigma, k))] \qquad \text{Eqn. 7 base}$$

which can be expanded to,

$$L(\mathcal{T}, \tilde{\mathcal{S}_\sigma}) = -\sum_{k=1}^{K_0} \frac{1}{K_0} p(b_\mathcal{T} = 1|\mathcal{S}_\sigma, k) \ln p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k) \\ + (1 - p(b_\mathcal{T} = 1|\mathcal{S}_\sigma, k)) \ln 1 - p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k). \tag{21}$$

where $\mathcal{S}_\sigma \in \mathcal{M}_\mathcal{S}^\mathcal{T}$. Given $\mathcal{S}_\sigma \in \mathcal{M}_\mathcal{S}^\mathcal{T}$, then $p(b_\mathcal{T} = 1|\mathcal{S}_\sigma, k) = 1$ always holds, which means the negative case in Eqn. 21 can be ignored, yielding the following simplified form:

$$\begin{aligned} L(\mathcal{T}, \tilde{\mathcal{S}_\sigma}) &= -\sum_{k=1}^{K_0} \frac{1}{K_0} \ln p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k) \\ &= -\mathbb{E}_{k\sim U[1,\ldots,K_0]}[\ln p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k)]. \qquad \text{Eqn. 7} \end{aligned}$$

and so $L(\mathcal{T}, \tilde{\mathcal{S}_\sigma})$ is simply the negative log-likelihood of sampling a satisfied theory ($b_\mathcal{T} = 1$) from soft-structure $\tilde{\mathcal{S}_\sigma}$, for randomly sampled grounding k. Next, we show the similarities between $L(\mathcal{T}, \tilde{\mathcal{S}_\sigma})$ and $\Gamma_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ by looking at the likelihood $p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k)$. First, we define $\bar{\Gamma}_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ by isolating the likelihood:

$$\begin{aligned} \exp(-L(\mathcal{T}, \tilde{\mathcal{S}_\sigma})) &= \prod_{k=1}^{K_0} p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k)^{\frac{1}{K_0}} \\ &\doteq \bar{\Gamma}_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}} \end{aligned} \tag{22}$$

We then expand $p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k)$ to:

$$\begin{aligned} p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k) &= \sum_{t=1}^{K_1} p(b_\mathcal{T} = 1|\boldsymbol{c}_{kt}) p(\boldsymbol{c}_t|\tilde{\mathcal{S}_\sigma}, k) \\ &= \sum_{t\in t^+} p(\boldsymbol{c}_{kt}|\tilde{\mathcal{S}_\sigma}, k) \end{aligned} \tag{23}$$

where t^+ is defined as before. For all other $t \neq t^+$, $p(b_\mathcal{T} = 1|\boldsymbol{c}_{kt}) = 0$ and so this acts as a filter, yielding:

$$\bar{\Gamma}_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}} = \prod_{k=1}^{K_0} \sum_{t\in t^+} p(\boldsymbol{c}_{kt}|\tilde{\mathcal{S}_\sigma}, k)^{\frac{1}{K_0}}. \tag{24}$$

$p(\boldsymbol{c}_{kt}|\tilde{\mathcal{S}_\sigma}, k)$ is calculated by evaluating the belief of each relation-decoder against the expected truth-assignment as defined by truth-table row $\boldsymbol{c}_{kt}$:

$$\begin{aligned} p(\boldsymbol{c}_{kt}|\tilde{\mathcal{S}_\sigma}, k) &= \prod_{m=1}^{l} \phi_{r^m}(O_{km})^{c_{ktm}} (1 - \phi_{r^m}(O_{km}))^{1-c_{ktm}} \\ &= f(\phi_{r^m}, O_{km}, c_{ktm}) \end{aligned}$$

where r^m is the relation for atomic formula associated with column m (which is the same for each k slice and t row) and O_{km} is the grounding of this entry for slice k (which is the same across rows). Putting it all back together, we finally have that:

$$\bar{\Gamma}_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}} = \prod_{k=1}^{K_0} \sum_{t\in t^+} \prod_{m=1}^{l} f(\phi_{r^m}, O_{km}, c_{ktm})^{\frac{1}{K_0}}, \tag{25}$$

which makes the similarities between $\Gamma_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ and $\bar{\Gamma}_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ clear and exposes their relationship.In particular, for the special case where $|\mathcal{M}_\mathcal{S}^\mathcal{T}| = 1$, the outer sum for $\Gamma_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ can be removed, and the remaining differences between $\Gamma_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ and $\bar{\Gamma}_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ are the sum over t^+ rows and difference in exponent over $f(\phi_{r^m}, O_{km}, c_{ktm})$. For $\Gamma_\mathcal{T}^{\tilde{\mathcal{S}_\sigma}}$ to be maximised, through $p(\mathcal{S}_\sigma|\tilde{\mathcal{S}_\sigma}) \approx 1$, we would find that

$\tilde{\mathcal{S}}_\sigma$ maximally supports only the rows associated with $\mathcal{S}_\sigma$ for each k grounding. Notice that $\bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}}$ is again bound to (0,1) and achieves $\bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}} \approx 1$ when $\Gamma^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}} \approx 1$. We use the correspondence between $\Gamma^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}}$ and $\bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}}$ to define a practical ϵ-proxy consistency measure as follows. We firstly re-express ϵ-consistency/coherence but for $\bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}}$ and a different $\bar{\epsilon}$. We then trace this back to $L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma)$ so a bound in terms of the consistency loss can be reported as the overall ϵ-proxy. Together this yields the following:

$$\begin{aligned} \bar{\epsilon} &\geq 1 - \bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}} \\ \ln \frac{1}{1-\bar{\epsilon}} &\geq -\ln(\bar{\Gamma}^{\tilde{\mathcal{S}}_\sigma}_{\mathcal{T}}) \\ &\geq L(\mathcal{T}, \tilde{\mathcal{S}}_\sigma) \end{aligned} \tag{26}$$

and we arrive at an ϵ-proxy of the form $\ln \frac{1}{1-\bar{\epsilon}}$, which is reported in the main text.