Formalizing Coherence and Consistency Applied to Transfer Learning in Neuro-Symbolic Autoencoders

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the study of reasoning in neural networks, recent efforts have sought to improve coherence and consistency of neural sequence models. This is an important development in the study of neuro-symbolic systems. In symbolic AI, however, the concepts of consistency and coherence are defined formally. The provision of such formal definitions is needed to offer a common basis for the quantitative evaluation and systematic comparison of connectionist, neuro-symbolic and transfer learning approaches. In this paper we introduce formal definitions for coherence and consistency of neural systems. To illustrate the usefulness of the definitions, we propose a new dynamic relation-decoder model built around the principles of consistency and coherence. By comparing several existing relation-decoders on a partial relation transfer learning task and novel data set introduced in this paper, our experiments show that relation-decoders that can maintain consistency over unobserved regions of representation space, retain coherence across domains and achieve better transfer learning performance.

1 Introduction

Humans are capable of learning concepts that can be applied to many different scenarios [17, 33, 22]. An important principle is that human-like concepts remain *coherent* across contexts [30]. As an example, consider the concept of *ordinality*, *e.g*. "A is larger than B", which allows comparisons to be made between ordered sets. Ordinality should apply equally whether A and B are digits or a tower of blocks. It is said that a concept may pertain to a multitude of properties: position, volume, reach, *etc*. As long as one of these properties can be attributed to an object, a set of objects can be compared on that basis. All in all, if the concept of ordinality was to be learned in its most general form, its use should be consistent across objects and coherent across object properties.

In [30], empirical results on story generation and instruction-following have shown that an intuitive use of consistency and coherence can increase the accuracy of neural networks. It is argued in [30] that *System 1* approaches, fast and capable of learning patterns efficiently from data, "are often inconsistent and incoherent", and that "adding *System 2*-inspired logical reasoning" as a logically-consistency, training-free module allows for an improved selection of candidate stories generated by *System 1*. While [30] makes an important contribution by exploring several variations on the theme, in this paper we offer a formal definition for consistency and coherence in the context of neural networks, in particular autoencoders. We also take one step further and apply and evaluate consistency and coherence to transfer learning, where we believe that the theme will have its most practical impact.

We argue that for a concept to be useful during transfer learning, the system of relations that define the concept in the source domain must be coherent with the target domain, whereby logical consistency achieved in the source is retained in the target domain. This is to say that the concept-specific relations

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

learned in the source ought to be consistent with a logical theory that defines their semantics, and that such consistency must extend beyond the representations learned in the source domain and, in particular, hold for the embeddings learned in the target domain.

In this paper, we offer a formal definition for consistency and coherence of sub-symbolic learners, inspired by analogous definitions from symbolic AI. This is expected to define the conditions that make a learned concept transfer well across properties and objects. We propose a simple neural-symbolic autoencoder architecture consisting of a neural encoder for objects coupled with consistent and modular relation-decoders, and we show in comparison with alternative popular approaches that this simple architecture is capable of achieving an improved transfer learning performance by being coherent across object properties [37, 12, 3, 44, 28, 11].

Specifically, consistency and coherence metrics are shown to offer a more fine-grained measure for transfer learning than accuracy alone. The proposed architecture is evaluated on a new Partial Relation Transfer (PRT) task and data set introduced in this paper. The application of a set of logical relations to a domain is specified as a model-theoretic structure with an analogous (soft-)structure for non-symbolic learners. Consistency and coherence of soft-structures is then shown to provide a practical score calculation to the evaluation of autoencoders. The benchmark PRT learning task uses a new BlockStacks data set derived from the CLEVR data set rendering agent. This is compared with several existing relation-decoder models on transfer learning tasks from BlockStacks to the MNIST handwritten digits data set, on relations such as isGreater, isEqual..., such that the learning of ordinality among the digits is evaluated against the learning of the relative position of a block in the stack. 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. In summary, the contributions of this paper are:

- A formal definition of consistency and coherence for sub-symbolic learners offering a practical evaluation score for concept coherence;
- A derived model implementation and partial relation transfer experimental setup used to evaluate the interplay between concept coherence and concept transfer;
- A comprehensive critical evaluation of results and comparison of multiple relation-decoder models with varied model capacity, showing that regularisation via model capacity or β-induced disentanglement pressure improves concept coherence.

In Section 2 we provide the required logic background. Section 3 introduces soft-structures and formally defines coherence and consistency. Section 4 describes the neuro-symbolic architecture and its associated practical consistency loss. After detailing the PRT task and data set in Section 5, comparative experimental results are discussed in Section 6. Section 7 concludes the paper with a discussion, including limitations and future work. We discuss related work, experimental setup and data set characteristics, model details and parameterization, and we make the code and additional experimental results available in the Supplementary Material.

2 Preliminaries

Notation: 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), $\wedge$ (conjunction), $\rightarrow$ (implication), and universal quantification $\forall$ (for all) with their conventional meaning (see [38]). Relations express knowledge over the 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 a relation is defined by an *interpretation* $I_{\mathcal{S}_\sigma}$ over elements of an non-empty *domain* $\mathcal{S}$.

Definition 2.1 (Signature, Interpretation, Structure)**.** The *signature* of a language $\mathcal{L}$ is a set of relations $\sigma = \{r \in \mathcal{L}\}$ whose elements have *arity* given by $\mathsf{ar} : \sigma \to \mathcal{N}$, where $\mathcal{N}$ is the set of natural numbers. 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 example, $r(s_1, s_2)$, 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 2.2 (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 hand-written digits and σ the signature of binary relations $\sigma = \{\mathsf{isGreater}, \mathsf{isEqual}, \mathsf{isLess}, \mathsf{isSuccessor}, \mathsf{isPredecessor}\}$, 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 $\forall i, j.\, \mathsf{G}(i, j) \to \neg \mathsf{E}(i, j)$ (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 A Formalization of Consistency and Coherence

In this section we turn our attention to the challenge of learning a model of a theory over a real-world domain given a signature. 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 example, the handwritten digits in the MNIST dataset contain stylistic details such as line thickness and digit skew that are irrelevant to the notion of ordinality, which makes learning the structure from Example 1 non-trivial.

Following the convention from the disentanglement literature [4, 20, 16, 15], we make 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:

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

tasked with approximating the conditional 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 $\mathcal{Z}$ [32, 31], and refer to these as *relation-decoders*:

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

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 3.1 (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 $\psi_{\mathcal{S}} : \mathcal{S} \to \mathcal{Z}$, $\tilde{\mathcal{S}}_\sigma = (\psi_{\mathcal{S}}(\mathcal{S}), \phi)$ 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 (logical) 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:

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

with $f(\phi_r, \psi_\mathcal{S}, O, \gamma^r_{O,\mathcal{S}_\sigma}) = (\phi_r(\psi_\mathcal{S}(O)))^{\gamma^r_{O,\mathcal{S}_\sigma}} \cdot (1-\phi_r(\psi_\mathcal{S}(O)))^{1-\gamma^r_{O,\mathcal{S}_\sigma}}$, 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 a relation 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 $\mathcal{M}^\mathcal{T}_\mathcal{S}$ 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 3.2 (ϵ-Consistency of Soft-Structure)**.** Given a finite soft-structure $\tilde{\mathcal{S}_\sigma}$, if $1 - \Gamma^{\tilde{\mathcal{S}_\sigma}}_\mathcal{T} \leq \epsilon$ 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 [30]. A consistent soft-structure $\tilde{\mathcal{S}_\sigma}$ ensures that ϕ gives high belief only to interpretations that satisfy, and therefore are logically consistent with, $\mathcal{T}$. As expected, consistency pertains to the domain encodings of $\tilde{\mathcal{S}_\sigma}$, *i.e.* $\psi_\mathcal{S}(\mathcal{S})$. 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 with their corresponding soft-structures, which gives our definition of coherence between soft-structures, as follows. Consider a situation where a deep network has already learned a soft-structure that has high model weight given the relations $\{\mathsf{G}, \mathsf{E}, \mathsf{L}, \mathsf{S}, \mathsf{P}\}$ from Example 1. Now suppose that we are given a new domain of images, $\mathcal{Y}$, showing single block stacks of different heights, and we wish to re-use the signature of ordinal relations and $\mathcal{T}$ from Example 1. Lastly, let $I_{\mathcal{Y}_\sigma}$ be a interpretation in the new domain that orders images according to block stack height and is a model of $\mathcal{T}$. We can summarise this with the following two structures:

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

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 learned 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 $\psi_\mathcal{Y}(\mathcal{Y})$? Concretely, we are interested in when the following *coherence* condition holds.

Definition 3.3 (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 $\tilde{\mathcal{X}_\sigma}$ is ϵ_1-consistent with $\mathcal{T}$, $\tilde{\mathcal{Y}_\sigma}$ is ϵ_2-consistent with $\mathcal{T}$, $\epsilon_1 \leq \epsilon$, and $\epsilon_2 \leq \epsilon$.

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].

[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.

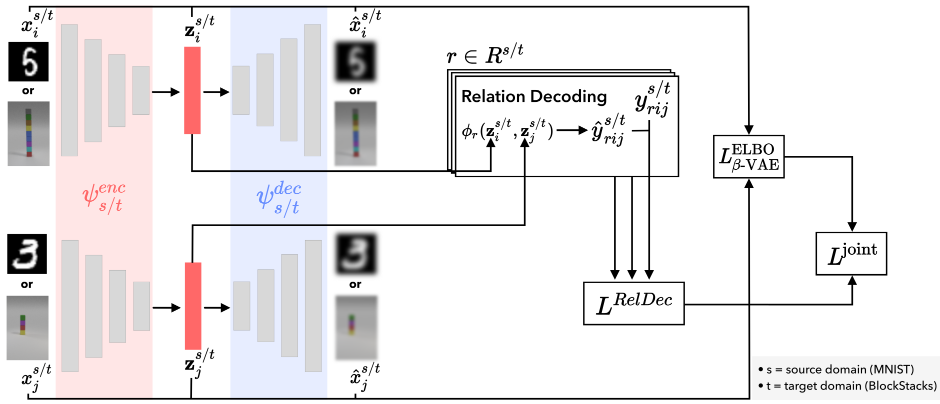

Figure 1: Network architecture used for PRT task.Relational learning is performed on the source MNIST data set (to learn e.g. that digit 5 is greater than 3). Moving to the target data set (to learn that a stack of blocks is greater than another) involves training a new encoder-decoder together with a subset of the relation-decoders (with fixed parameters) from MNIST. The remaining relations are held-out to evaluate zero-shot transfer learning performance.

4 A Consistent and Coherent Neuro-Symbolic Autoencoder

In order to ground our definitions of consistency (3.2) and coherence (3.3) into a real system and evaluate their practical value, in this section we propose a simple autoencoder neuro-symbolic architecture intended to satisfy our definitions. To derive an efficient loss function, we introduce an estimate measure for a soft-structure's ϵ-consistency and coherence with a given theory when access to every logical model is not available or computationally feasible.[2]

Suppose there is a fixed domain $\mathcal{S}$ and theory $\mathcal{T}$ whose sentences use relations from a signature σ. Let $k \in \{1, ..., K_0\}$ denote the index associated with each unique ground instance of the relations in $\mathcal{T}$. Take $B_\mathcal{T}$ to be a Boolean random variable. The probability of $\mathcal{T}$ being satisfied under a soft-structure $\tilde{\mathcal{S}_\sigma}$ is expressed as $p(b_\mathcal{T}|\tilde{\mathcal{S}_\sigma}, k)$, where $b_\mathcal{T} = 1$ if $\mathcal{T}$ is satisfied (i.e. $true$), or 0 otherwise (denoting $false$). By definition, $p(b_\mathcal{T} = 1|\mathcal{S}_\sigma, k) = 1$ if $\mathcal{S}_\sigma \in \mathcal{M}^\mathcal{T}_\mathcal{S}$, where $\mathcal{M}^\mathcal{T}_\mathcal{S}$ denotes the set of models of $\mathcal{T}$. When $\tilde{\mathcal{S}_\sigma}$ is consistent with $\mathcal{T}$ then we should also find that $p(b_\mathcal{T} = 1|\tilde{\mathcal{S}_\sigma}, k) \approx 1$. Hence, we define a loss function as the 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 simplifies to the expected negative log-likelihood of satisfying $\mathcal{T}$ under a random sampling from the set of ground instances:

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

where $p(k) = \frac{1}{K_0}$ is taken to be uniform distribution over the set of unique groundings. A measure based on this loss is required to enable the practical evaluation of 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 the already defined $\Gamma^{\tilde{\mathcal{S}_\sigma}}_\mathcal{T}$ to obtain a bound on the loss function:

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

where $\bar{\epsilon} \geq 1 - \bar{\Gamma}^{\tilde{\mathcal{S}_\sigma}}_\mathcal{T}$. We take the coherence to be the upper value of $\ln \frac{1}{1-\bar{\epsilon}}$ between domains.[3]

Figure 1 outlines the main components of our autoencoder: a domain-encoder $\psi_\mathcal{S}$ and modular relation-decoders ϕ form an autoencoding architecture that, given a domain of images $\mathcal{S} \subset \mathbb{R}^{W\times H}$ and a d-dimensional latent space $\mathcal{Z} = \mathbb{R}^d$, converts sub-symbolic encodings from $\psi_\mathcal{S}$ into a modular relational representation via decoding 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, a domain-decoder produces domain reconstructions $\hat{\mathcal{S}}$. In Figure 1, we use $\psi^{\text{enc}}_\mathcal{S}$ to refer to the domain-encoder and $\psi^{\text{dec}}_\mathcal{S}$ for the domain-decoder. To train the model, ground-truth interpretations $I_{\mathcal{S}_\sigma}$ are given, allowing us to directly

[2]Calculating Eqn. 4 can become intractable as it involves computing ϕ beliefs for every grounding.
[3]The complete derivation of loss function and bounds is presented in the Supplementary Material.

maximise Eqn. 3 via the 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, known as disentanglement (although disentanglement is not seen here as a requirement for consistency and coherence) [16, 6, 20]. We therefore combine the ELBO objective with an additional β scalar hyperparameter that seeks to achieve disentanglement ($L^{\text{ELBO}}_{\beta\text{-VAE}}$) with $L(\mathcal{T}, \tilde{\mathcal{S}_\sigma})$ over each component of the autoencoder architecture to obtain the following aggregate objective (we provide the full ELBO derivation with a detailed explanation in the Supplementary):

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

where λ is a scalar weighting parameter.

Together with the $L^{\text{ELBO}}_{\beta\text{-VAE}}$, the choice of relation-decoder can shape the domain-encodings [14]. In our evaluation, the following choices are made. We propose a Dynamic Comparator (DC) model composed of two modes, a distance-based measure, $\phi^{\dagger}_r$, to measures the distance between two inputs relative to a reference point, and a step-function, $\phi^{\ddagger}_r$, that determines the sign of the difference between two points, optionally with an offset. Although any function can be used that has the required characteristics for $\phi^{\dagger}$ and $\phi^{\ddagger}$, in this paper we use the following implementation:

$$\phi^{DC}_r(\boldsymbol{z}_i, \boldsymbol{z}_j) = a_{r,0} \cdot \phi^{\dagger}_r + a_{r,1} \cdot \phi^{\ddagger}_r \tag{11}$$

where,

$$\phi^{\dagger}_r = f_0\big(-\eta_{r,0}(\|\boldsymbol{u}_r \odot (\boldsymbol{z}_i - \boldsymbol{z}_j + \boldsymbol{b}^{\dagger}_r)\|_2)\big) \tag{12}$$

$$\phi^{\ddagger}_r = f_1\big(\eta_{r,1} \cdot \boldsymbol{u}_r^\top(\boldsymbol{z}_i - \boldsymbol{z}_j + \boldsymbol{b}^{\ddagger}_r)\big). \tag{13}$$

Here $\boldsymbol{a}_r = \texttt{Softmax}(\boldsymbol{A}_r) \in (0,1)^2$ is an attention weighting between the two modes, $\phi^{\dagger}_r$ and $\phi^{\ddagger}_r$; f_0 and f_1 are an $\exp$ 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. Further details are provided in the Supplementary Material.

5 Relational Transfer Learning Experiment Design

We now describe an experimental design to compare coherence of different relation-decoders.

Partial Relation Transfer (PRT): We evaluate a novel PRT task across two soft-structures $\tilde{\mathcal{X}}_\sigma$ and $\tilde{\mathcal{Y}}_\sigma$. They share a common signature σ and relation-decoders ϕ but have disjoint domains $\mathcal{X}$ and $\mathcal{Y}$, respectively. The experimental design 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 are expected to help guide training of $\psi^{\text{enc}}_{\mathcal{Y}}$. Held-out relation-decoders are then evaluated in the new domain on zero-shot transfer learning performance. For domain $\mathcal{X}$ we use the MNIST handwritten digits data set [23], and for domain $\mathcal{Y}$ we use a proposed BlockStacks data set, which includes a single stack of multi-colored cubes of differing heights, each containing one randomly positioned red cube (see Supplementary Material for details and examples). The shared signature includes the ordinal relations $\sigma = \{\mathsf{G}, \mathsf{E}, \mathsf{L}, \mathsf{S}, \mathsf{P}\}$ 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 the theory in the Supplementary Material. When transferring relations from $\psi^{\text{enc}}_{\mathcal{X}}$ to $\psi^{\text{enc}}_{\mathcal{Y}}$, one could use the full set ϕ of relation-decoders. However, this is not necessary from a logical standpoint because the entire system of relations can be expressed in terms of isSuccessor (e.g. the successor of a number is larger than that number). We therefore only employ the isSuccessor relation-decoder as

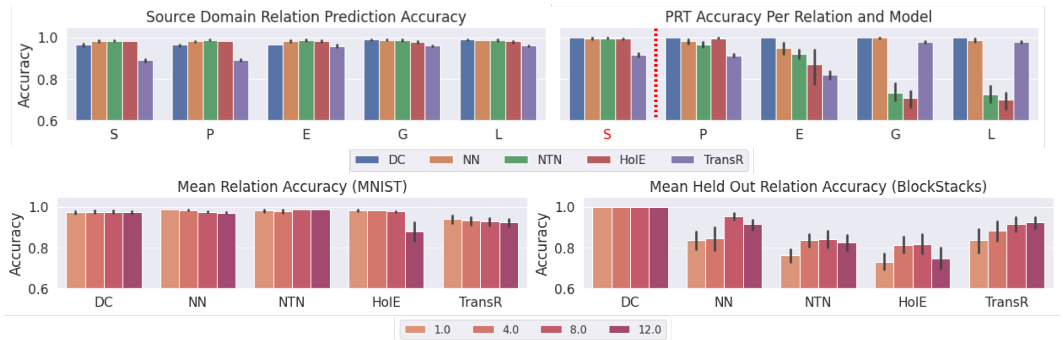

Figure 2: **[Top]** Relation-decoder prediction accuracy per model (DC, NN, NTN, HoIE, TransR) and relation (abbreviated on the x-axis by {S: isSuccessor, P: isPredecessor, E: isEqual, G: isGreater, L: isLess}), in the source domain (MNIST, left) and target domain (BlockStack, right). A red highlighted S and dotted line (top right) indicates that relation isSuccessor is included in training the target domain autoencoder, but none of the other relations are. Both DC and NN retain a good performance while all other models show a decrease of accuracy in the target domain for one or more of the relations not included in training. **[Bottom]** Impact of different values of $\beta \in \{1, 4, 8, 12\}$ for each relation-decoder averaged across all relations in the source domain (left) and held-out relations $\{P, E, G, L\}$ in the target domain (right). It can be seen that DC is not impacted by changes in β and it maintains performance in the target domain. All other models show a decrease of accuracy for the held-out relations in the target domain.

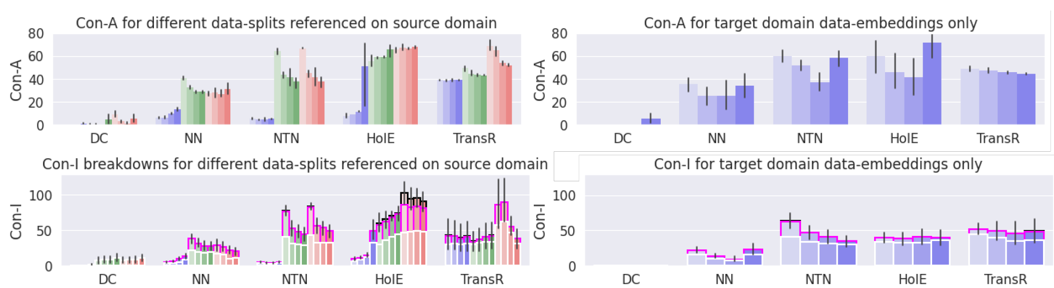

Figure 3: Consistency losses (lower values are better) for the models (DC, NN, NTN, HoIE, TransR) using the MNIST data set (source domain $\mathcal{X}$) **[left]** and BlockStacks (target domain $\mathcal{Y}$) **[right]**. The blue bars show the consistency loss of the data embeddings, with darker shades corresponding to models trained with higher β (disentanglement pressure). Two additional data splits are shown: interpolation (in green) with samples coming from the MNIST data-embedding cluster, and extrapolation (in red) with samples drawn from outside the cluster. Results are further divided into consistency across relations (Con-A) **[top]** and consistency of individual relations (Con-I) **[bottom]**. The following relations are used (see stacked bars at the bottom graphs): transitivity (in white), asymmetry (in magenta) and reflexivity (in black). Notice the large difference in MNIST between data-embedding Con-A vs. interpolated and extrapolated Con-A results, wherein BlockStacks data-embedding Con-A results are similar to the MNIST interpolated/extrapolated Con-A results.

a fixed-parameter guide for $\psi_{\mathcal{Y}}^{\text{enc}}$. If coherence, as defined in this paper, is carried across domains, we would expect the transferring of isSuccessor to be sufficient to produce an improved performance.

Neural model components and Hyperparameters: Together with DC, existing relation-decoder models evaluated here are: TransR [24], HolE [29], NTN [39]. We additionally include a basic feedforward neural-network baseline, NN. To produce domain-encodings, all experiments use a β-VAE. We provide further details for all models, including details about training regimen and implementation in the Supplementary Material. 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 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 $\mathcal{Z} = \mathbb{R}^{10}$.

Table 1: Coherence comparison with respect to source and target data-embeddings. Results are reported with the corresponding $\beta = \beta^*$ value (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 consistencies. DC outperforms all other approaches in coherence scores.

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)

6 Main Experimental Results and Comparative Evaluation

In this section, experimental results demonstrate the relevance of a model-theoretic perspective on the learning of concepts with neural networks. Results show that transfer learning performance is positively correlated with measures for consistency within and consistency across domains, i.e. coherence. This holds particularly true for embeddings that are close by but different from source domain embeddings. As we have argued, for a neural model to perform well on concept transfer, its representations must maintain high probability of consistency with a logical theory that can provide a semantics for the concept. We further argue that the most robust way of doing this is to maintain consistency across regions of embedding space, rather than relying exclusively on the specific data-points observed at training time in the source domain.

Figure 2 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 offers a direct coherence comparison between relation-decoders, using the derived coherence measure (Eqn. 8)[4].

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 the impact of adjusting β (disentanglement pressure) 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.

To gain deeper insight as to which underlying characteristics can explain the observed PRT accuracy profiles, Figure 3-top presents consistency losses against formulae that constrain truth-value assignments across relations under a theory of ordinality, referred to as consistency-across (Con-A).[5]. Results refer to both source (left) and target domain embeddings (right). We note that DC shows excellent Con-A in the target domain 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. DC sustains strong Con-A results for target domain data-embeddings (right). Results for all other models are notably worse with respect to their source data-embedding performances and are instead comparable with their interpolation or extrapolation results in the source domain. Together, these results paint a picture wherein it may be possible to antic-

[4]We take ϕ_r prediction values above 0.5 to signify a truth prediction and those below 0.5 to signify falsity. An alternative, left as future work, would be to sample the space of ϕ values to produce a confidence measure

[5]Truth-tables for each consistency formula are given in the Supplementary Material

ipate poor transfer performance by evaluating interpolation and extrapolation consistency in the source domain. This would indeed be expected, since source and target domain data-embeddings are unlikely to perfectly overlap, and so retained consistency on regions outside the source data-embeddings should increase the probability of consistency over target domain data-embeddings.

Next, Figure 3-bottom presents consistency values for each individual relation-decoder model (Con-I). Stacked bars show the results for logical sentences defining: transitivity (white), asymmetry (magenta) and reflexivity (black). Results are averaged over individual relations and are grouped under label Con-I w.r.t. 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.

Lastly, Table 1 provides a comparison between optimal coherences achieved for each relation-decoder model, as defined in Section 4. Results are partitioned according to each consistency type (transitivity, asymmetry and reflexivity) and an aggregate value. DC clearly outperforms all other models on coherence. NN achieves strong aggregate coherence compared with NTN, HolE and TransR. Although NTN and HolE have similar aggregate coherence, TransR performs generally worse. This may be caused by TransR producing weaker belief scores 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 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 coherence are in more direct agreement - as TransR achieves its best at $\beta = 12$.

Our results indicate that increasing regularisation over relation-decoder models, either in the form of disentanglement pressure or relation-decoder model capacity, improves 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, i.e. coherence, is key.

7 Conclusion and Future Work

This paper introduced formal definitions of consistency and coherence for neuro-symbolic systems. As a result, a sub-symbolic model can have consistency and coherence measured with respect to a logical theory. We defined a neural model based on domain-encoders coupled with modular relation-decoders and experimental procedure that together allowed the investigation of how concept coherence differs for various implementations of relation-decoders applied to transfer learning. Consistency results and a comparison of coherence scores showed that the models that can achieve excellent coherence also achieve high accuracy at partial relational transfer learning tasks. The empirical evaluations in this paper only considered binary relations and a fixed signature which is learned "all at once" in a source domain. In practical applications, however, it should be possible to discover concepts gradually, e.g. as part of a curriculum or through gradual refinement of pre-learned relations after progressive exposure to different contexts. This necessitates an adaptation of the approach presented here and further evaluations as part of future work. Additionally, we only explored a signature for ordinality, whereas other fundamental properties should be investigated such as periodic (*e.g.* rotation) and unordered categorical (*e.g.* shape) properties. Further evaluations of the formalization introduced here should consider the use of different models, theories and scenarios/data sets in the evaluation of consistency and coherence metrics.

References

[1] Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. Boxe: A box embedding model for knowledge base completion. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

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

[3] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *CoRR*, abs/2012.13635, 2020.

[4] 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.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems*, pages 2787–2795. Curran Associates, Inc., Lake Tahoe, USA, 2013.

[6] 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.

[7] Junxiang Chen and Kayhan Batmanghelich. Robust ordinal VAE: employing noisy pairwise comparisons for disentanglement. *CoRR*, abs/1910.05898, 2019.

[8] 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.

[9] 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*, pages 2615—-2625, Montreal, Quebec, Canada, 2018.

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2172–2180, 2016.

[11] 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.

[12] 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*, pages 1596—-1602, 2017.

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

[14] Víctor Gutiérrez-Basulto and Steven Schockaert. From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018*, pages 379–388. AAAI Press, 2018.

[15] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018.

[16] 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*, Toulon, France, 2017.

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

[18] 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,*, pages 1–16, San Juan, Puerto Rico, 2016.

[19] 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.

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

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

[22] 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.

[23] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[24] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2181–2187. AAAI Press, 2015.

[25] 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}*, pages 4114—-4124, Long Beach, California, USA, 2019.

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

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

[28] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proc. IEEE*, 104(1):11–33, 2016.

[29] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 1955–1961. AAAI Press, 2016.

[30] 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. *In Proc. NeurIPS 2021*, abs/2107.02794, 2021.

[31] J. B. Paris. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge University Press, 1994.

[32] Jeffrey Paris and Alena Vencovská. *Pure Inductive Logic*. Perspectives in Logic. Cambridge University Press, 2015.

[33] Jean Piaget. *The Psychology of Intelligence*. Routledge and Kegan Paul, 2005.

[34] Ievgen Redko, Amaury Habrard, Emilie Morvant, Marc Sebban, and Younès Bennani. *Advances in Domain Adaptation Theory*. Elsevier, 2019.

[35] 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*, pages 185—-194, Montreal, Quebec, Canada, 2018.

[36] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. 10843:593–607, 2018.

[37] 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.

[38] Stewart Shapiro and Teresa Kouri Kissel. Classical Logic. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

[39] 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*, pages 926–934, 2013.

[40] 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.

[41] 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.

[42] 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*, pages 2071–2080, New York, NY, USA, 2016.

[43] 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*, pages 14222—-14235, Vancouver, BC, Canada, 2019.

[44] 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.

A Societal Impact Statement

This work does not have a negative societal impact, specifically it 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 proposed definition of coherent concepts and corresponding analysis provides more depth in the assessment of deep learning methods, which are typically otherwise opaque, and this can have a positive societal impact. Currently, we cannot provide interpretable descriptions regarding *how* a standard deep learning method produces its inferences, making it difficult to fully trust a model in critical applications. An important failure case is that biases are not easy to uncover from a trained deep learning model. The benefit of learning a coherent concept is that inferences uphold logical consistency, which can be formally expressed and tested. This can provide more trust in the model as practitioners can have confidence that the model should not obtain inputs that lead to incoherent inferences, wherein errors are certain. Further, if the logic does not include biases, the inferences of a coherent set of relation-decoders should not be biased. A caveat to these points is that unless the relation-decoder functional form allows us to analytically make comments/assertions about the model's performances for arbitrary regions of latent space, as with DC (see E.1), it is intractable to fully examine model coherence, as it requires a full extrapolation/interpolation evaluation. Nonetheless, a practical evaluation of coherence is an important step forward.

B 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 [39, 42, 41, 5, 28, 44, 11, 19, 1]. Although these typically do not use a shared autoencoder as we do in this paper, **(author?)** [36] 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 [4] , and has been explored using a variety of methods including both Generative Adversarial Networks [10] and VAEs [6, 16, 9, 35, 13, 21, 25]. Disentangled representations have been evaluated in terms of there transferability in [43, 40, 26]. A bridge between these two fields, wherein relation-decoders are employed as a semi-supervision to VAEs can be found in [18, 8, 7], where [18] use multiple relation-decoders but compute a triplet comparison based query and [8, 7] 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 [34]. 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.

C 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 [2]. The dataset is divided into 9000:1500:1500 train, validation and test splits.

D 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 [20]. The VAE maximises the log-probability by maximising

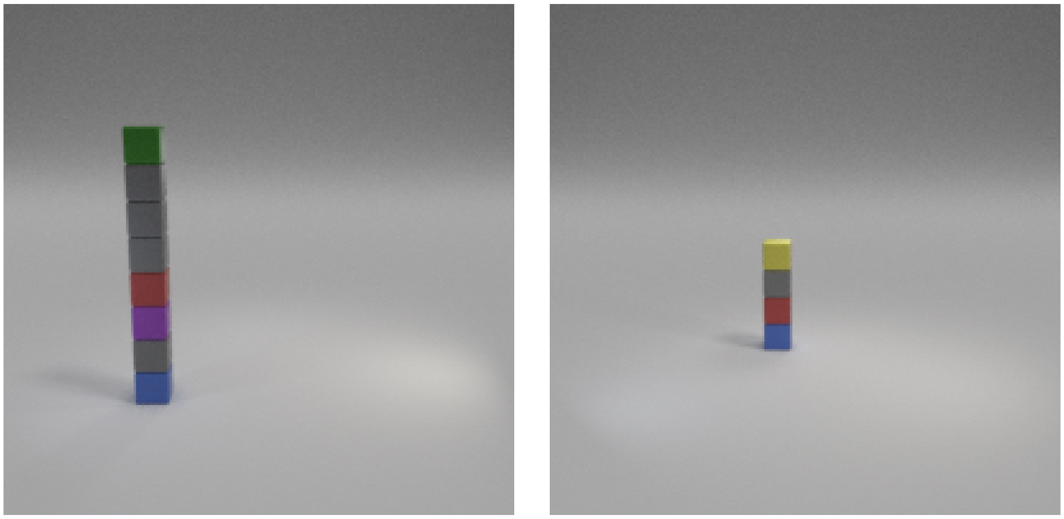

Figure 4: Example of two BlockStacks data set images. Each instance consists of a single red block varying in position within the block stack. On the left the red block is at height 3 (using a zero index) and on the right it is at height 1.

this lower bound, given by:

$$L_{\beta\text{-VAE}}^{\text{ELBO}} = \mathbb{E}_{q_\alpha(\boldsymbol{Z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{Z})] - \beta D_{KL}(q_\alpha(\boldsymbol{Z}|\boldsymbol{X}) \| p_\theta(\boldsymbol{Z})), \tag{14}$$

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 [20]). In the β-VAE [16, 6], 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 $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$). When $\beta = 1$ we obtain the standard VAE objective [20].

E 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.

E.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

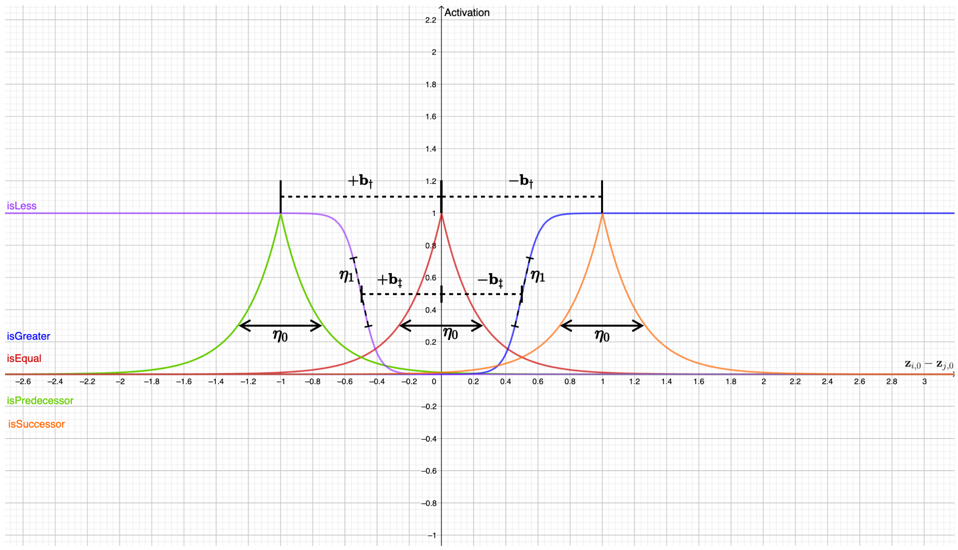

Figure 5: Depiction of a set of DC relation-decoders for binary relations isGreater, isLess, isEqual, isSuccessor and isPredecessor. Each DC relation-decoder (for each relation) has a one-hot mask, $\boldsymbol{u}_r$ (that is in this example the same across relations), which ensures only the zeroth dimensions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

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.

E.2 Relation-Decoder implementations

TransR [24]:

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

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 [39] from [12, 37]):

$$\phi_r(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \sigma\big(\boldsymbol{u}_r^\top [\tanh(\boldsymbol{z}^{c\top} \boldsymbol{M}_r \boldsymbol{z}^c + \boldsymbol{V}_r \boldsymbol{z}^c + \boldsymbol{b}_r)]\big) \tag{15}$$

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 [12, 37]. In contrast, the original NTN (see [39]) is only applicable to binary relations and does not include the outer sigmoid.

HolE [29]:

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

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 [27]. The final output layer, l_{out}, is a single value passed through a sigmoid function, to bound the output within (0,1).

E.3 β-VAE configuration

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

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

E.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 [25, 40], 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.

F 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)

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

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.

G Expanded consistency loss derivation

In this section, we present the expanded justification for reporting $-\ln 1-\bar{\epsilon}$ 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 that denote 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) *i.e.* $t^+ = \{\ t\ |\boldsymbol{T}_{k,t,K_2} = 1 \wedge t \in \{1,\ldots,K_1\}\}$, we can then rewrite $\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma}$ in terms of samples from $\boldsymbol{T}$:

$$
\begin{aligned}
\Gamma_{\mathcal{T}}^{\tilde{\mathcal{S}}_\sigma} &= \sum_{\mathcal{S}_\sigma \in \mathcal{M}_{\mathcal{S}}^{\mathcal{T}}} \prod_{r \in \sigma} \prod_{O \in \mathcal{S}^{\mathrm{ar}(r)}} \phi_r(O)^{\gamma^r_{O,\mathcal{S}_\sigma}} (1-\phi_r(O))^{1-\gamma^r_{O,\mathcal{S}_\sigma}} \qquad \text{(Eqn. 3)} \\
&= \sum_{\mathcal{S}_\sigma \in \mathcal{M}_{\mathcal{S}}^{\mathcal{T}}} \prod_{k=1}^{K_0} \sum_{t \in t^+} \mathbf{1}_{t_k^{\mathcal{S}_\sigma}}(t) \prod_{m=1}^{l} f(\phi_{r^m}, O_{km}, c_{ktm})^{N(\phi_{r^m}, O_{km}, c_{ktm}, \mathcal{S}_\sigma)^{-1}} \qquad (19)
\end{aligned}
$$

with

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

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 $\Gamma^{\tilde{\mathcal{S}_\sigma}}_{\mathcal{T}}$ 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,

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

where $\mathcal{S}_\sigma \in \mathcal{M}^{\mathcal{T}}_{\mathcal{S}}$. Given $\mathcal{S}_\sigma \in \mathcal{M}^{\mathcal{T}}_{\mathcal{S}}$, 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 $\Gamma^{\tilde{\mathcal{S}_\sigma}}_{\mathcal{T}}$ by looking at the likelihood $p(b_{\mathcal{T}} = 1|\tilde{\mathcal{S}_\sigma}, k)$. First, we define $\bar{\Gamma}^{\tilde{\mathcal{S}_\sigma}}_{\mathcal{T}}$ by isolating the likelihood:

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

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

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

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:

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

$p(\boldsymbol{c}_{kt}|\tilde{\mathcal{S}_\sigma}, k)$ 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:

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] These have been included in the Supplementary - the main societal impact of coherent concept learning is that inferences will uphold logical consistency. If the logic does not include biases, the inferences themselves should not be biased, providing that the feature extraction has been properly disentangled. All in all, coherent concepts will be easier to trust.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3 and Section 4
   (b) Did you include complete proofs of all theoretical results? [Yes] In particular, a rigorous proof for the consistency loss (Eqn. 8) is provided in the Supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] A current zip of the code used for this paper will be included in the paper's supplementary. A URL to public facing (refined and minimised) code will be provided in the camera ready version of the paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and we provide further model details in the Supplementary.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We provide error bars in bar plots. However we did not provide errors in the tabular results. These are obtained from the bar plots, so can be evaluated, but to improve clarity we will include the numeric errors in the tabular results for the camera ready version.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] This is a difficult number to extract

as experiments were run on different GPU models and with early stopping. We will endeavour to provide an estimate for the camera ready.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The MNIST data set is cited properly.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The necessary code to generate the BlockStacks data set is included in the Supplementary. We will include the actual data set providing there is space.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]