Journal of Machine Learning Research 27 (2026) 1-40 Submitted 5/25; Revised 6/26; Published 6/26

Leakage and Interpretability in Concept-Based Models

Enrico Parisini* ENRICO.PARISINI@KCL.AC.UK
The Alan Turing Institute
London, UK
PharosAI, Kings College London
London, UK

Tapabrata Chakraborti TCHAKRABORTY@TURING.AC.UK
The Alan Turing Institute
London, UK
University College London
London, UK

Chris Harbron CHRIS.HARBRON@ROCHE.COM
Roche Pharmaceuticals
Welwyn Garden City, UK

Ben D. MacArthur* BDM@SOTON.AC.UK
Mathematical Sciences, University of Southampton
Southampton, UK

Christopher R.S. Banerji* CBANERJI@TURING.AC.UK
The Alan Turing Institute
London, UK
Kings College London
London, UK

*EP, BDM and CRSB should be considered joint corresponding authors. BDM and CRSB are joint senior authors.

Editor: Manuel Gomez-Rodriguez

Abstract

Concept-based Models aim to improve interpretability by predicting high-level intermediate concepts, representing a promising approach for deployment in high-risk scenarios. However, they are known to suffer from information leakage, whereby models exploit unintended information encoded within the learned concepts. We introduce an information-theoretic framework to rigorously characterise and quantify leakage, and define two complementary measures: the concepts-task leakage (CTL) and interconcept leakage (ICL) scores. We show that these measures are strongly predictive of model behaviour under interventions and outperform existing alternatives. Using this framework, we identify the primary causes of leakage and, as a case study, analyse how it manifests in Concept Embedding Models, revealing interconcept and alignment leakage in addition to the concepts-task leakage present by design. Finally, we present a set of practical guidelines for designing concept-based models to reduce leakage and ensure interpretability.

Keywords: Concept Learning, Explainable AI (XAI), Interpretability metrics, Concept Bottleneck Models, Information leakage

1. Introduction

Explainability and transparency are essential aspects of model design, especially in high-risk scenarios such as biomedical applications. Concept Bottleneck Models (CBMs) (Kumar et al., 2009; Lampert et al., 2009; Koh et al., 2020) stand out in the landscape of interpretable models as they learn a set of high-level intermediate concepts associated with an outcome, in a manner which facilitates human oversight (Banerji et al., 2025). CBMs first predict the level of activation of each concept for a given input data point, then use these activations to predict a task. Under the assumption that the set of annotated concepts are appropriately learned by the model to encode the same amount of information as the ground-truth concepts, a CBM is thus *interpretable*, in the sense that its output can be deterministically and uniquely traced to the level of activation of semantically meaningful concepts (e.g. "presence vs. absence" in the binary case). Their structure allows for direct human supervision and intervention, as opposed to *post hoc* explainability approaches.[1]

In earlier CBMs, concept activations are real numbers (either probabilities or logits); more recently there have been efforts to generalise this approach to concept vector representations. Concept Embedding Models (CEMs, Espinosa Zarlenga et al., 2022) are often regarded as state of the art in this respect, as they are capable of achieving higher task performance than CBMs, frequently matching that of end-to-end models.

A trained concept-based model, however, may be significantly less interpretable than expected, despite presenting the illusion of being so. A key reason for this is the phenomenon of *information leakage* (Kazhdan et al., 2021; Margeloiu et al., 2021; Mahinpei et al., 2021; Havasi et al., 2022; Makonnen et al., 2025). Leakage occurs when a poorly designed concept-based model uses additional information to attain a higher task accuracy than a well-designed concept-based model, at the cost of undermining concept interpretability. Although distinct from shortcut learning (Enouen and Galhotra, 2025), leakage can enable shortcuts when the additional information encoded in the learnt concepts is exploited by the final task predictor. Storing such additional information in the learnt concepts is beneficial for the model with respect to loss function minimisation, as it typically allows the final head to leverage this information, attaining higher task performance while preserving high accuracy in concept predictions. This ultimately results in a model that is not interpretable in the sense defined above: interpretability requires not only that the output be traceable to the concept bottleneck, but also that the bottleneck activations remain semantically faithful to the annotated concepts they are intended to represent.

As an example, consider a concept-based model trained to make treatment recommendations for cancer patients from local biopsy data, using concepts routinely extracted from such data, such as tumor grade (i.e. how aggressive the tumor appears under the microscope). However, when metastases are present, treatment decisions depend on body-level information that cannot be inferred from local biopsy concepts alone. A concept-based model trained only on local biopsies may therefore be misspecified and, during training, can learn to encode task-relevant information (such as the presence of metastases) within its concept activations. In such a setting, a clinician can neither meaningfully supervise the model's predictions (since decisions rely on hidden information not captured by the exposed

1. Other model classes that offer a degree of interpretability, beyond concept-based models, include prototype-based networks (Chen et al., 2019) and B-cos networks (Böhle et al., 2022).

concepts) nor safely intervene on an incorrectly predicted concept, because modifying that concept would change the output in a way that is causally unrelated to its clinical meaning. A concept-based model which displays leakage may therefore be unsafe for critical decision-making applications, such as clinical scenarios.

Beyond leakage, several additional risks and limitations affect current concept-based architectures (Sinha and Zhang, 2025). An important line of work in this context concerns robustness and security against adversarial attacks (Sinha et al., 2023; Baniecki and Biecek, 2025), and recent studies have also shown that leakage itself can facilitate backdoor attacks that influence model predictions (Lai et al., 2025). The practical human-in-the-loop integration of interpretable models also raises several important challenges, including interventions, user interfaces, and interactive debugging (Ramaswamy et al., 2022; Bontempelli et al., 2023).

Contributions. We propose and test a framework based on information theory to assess the interpretability of concept-based models. In particular, we provide the first precise definition of information leakage, and identify two main ways it can manifest - as information leaking into each learnt concept from either the class label, or the other concepts (Section 3). After summarising the limitations of current leakage measures (Section 4), we accordingly define two measures, the concepts-task leakage (CTL) and the interconcept leakage (ICL) scores (Section 5), and in Section 6 we show that they

- are highly predictive of differences in performance upon intervention across models (the only robust indicator of interpretability in a CBM);
- substantially outperform previously proposed measures (including the OIS and NIS defined in Espinosa Zarlenga et al. (2023a)) across models and datasets.

We then employ these measures to identify and assess the primary causes of leakage in CBMs (Section 7), including: over-expressive concept representations; an incomplete set of annotated concepts; a misspecified final head; and insufficient concept supervision. We then employ these measures to systematically quantify and compare several known or suspected causes of leakage in CBMs (Section 7), including: over-expressive concept representations; an incomplete set of annotated concepts; a misspecified final head; and insufficient concept supervision. While some of these effects have been previously observed in specific settings, our framework provides a unified way to measure their contribution to leakage and their impact on intervention behaviour. Building on this novel information-theoretic framework, in Section 8 we confirm that concepts-task leakage is structurally embedded in CEMs to boost task performance, consistent with their original architectural design and prior understanding (Espinosa Zarlenga et al., 2022, 2023b, 2025). We further establish the presence of interconcept leakage in CEMs and identify alignment leakage, a novel form of concepts-task leakage that arises in architectures trained under concept-level interventions such as CEMs. Finally, in Section 9 we propose a set of guidelines for the design of concept-based models based on preventing leakage, and we argue that evaluating leakage is a crucial step of model development to ensure interpretability.

The complete codebase for computing leakage scores, evaluating concept-based models, and reproducing all experimental results is publicly available at `https://github.com/enricoparisini/xai-concept-leakage`.

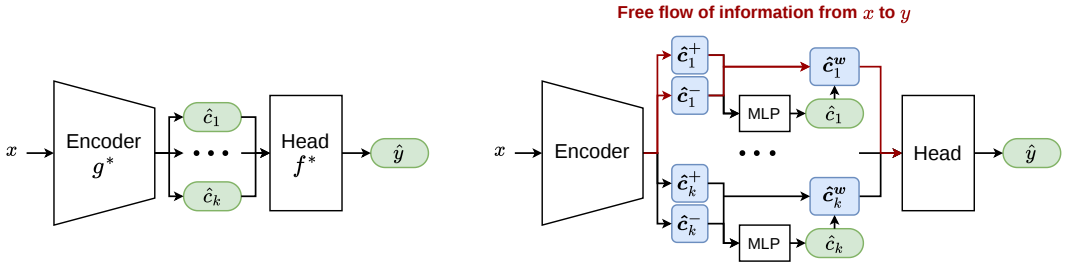

Figure 1: Scheme of CBM (*left*) and CEM (*right*) architectures. Quantities with green and blue backgrounds are predicted scalars and vectors respectively.

2. Concept bottleneck models

Consider a dataset $\left\{\boldsymbol{x}^{(n)}, c^{(n)}, y^{(n)}\right\}_{n=1}^{N}$ with N observations, where $\boldsymbol{x}^{(n)} \in \mathbb{R}^m$ is the input, $c^{(n)} \in \{0,1\}^k$ represents a set of k annotated concepts, and $y^{(n)} \in \{1,\dots,\ell\}$ is the target label. A CBM (Figure 1) is the composition of a concept encoder g^* mapping the input to the predicted concept activations $\hat{c} = g^*(x)$, and a classifier head f^* which maps the concepts to the predicted label $\hat{y} = f^*(\hat{c})$. This framework easily extends to regression tasks y and to categorical or continuous concepts.

The presence of a concept bottleneck displaying the predicted concept activations for each input allows a human overseer to directly supervise the model reasoning without resorting to post-hoc approaches to explainability such as CAM (Zhou et al., 2016), GradCAM (Selvaraju et al., 2017), LIME (Ribeiro et al., 2016) or SHAP scores (Lundberg and Lee, 2017). Additionally, this setup enables human interventions: if one or more predicted concepts are identified as incorrect for a given input $\boldsymbol{x}^{(n)}$, a human overseer can update their values according to their expertise, resulting in an improved task prediction. Recent work argues that CBM interventions can be interpreted through a causal lens as approximate interventions on concept variables in a concept-to-label causal model (Shin et al., 2022). Counterfactual CBMs (Dominici et al., 2025) make this connection to causality more explicit by learning to generate concept-level counterfactuals.

There are three possible strategies to train CBMs as introduced by Koh et al. (2020):

- *independent training*, where the concept encoder and the final head are trained independently on ground-truth data. In this case,

$$g^* = \text{argmin}_g \, \mathcal{L}_c(g(x), c)\,, \qquad f^* = \text{argmin}_f \, \mathcal{L}_y(f(c), y)\,, \tag{1}$$

  where $\mathcal{L}_c$ and $\mathcal{L}_y$ are suitable loss functions for the concepts and the label respectively;

- *sequential training*, where the concept encoder is trained on ground-truth data, while the final head is trained on the predicted concepts. In this case,

$$g^* = \text{argmin}_g \, \mathcal{L}_c(g(x), c)\,, \qquad f^* = \text{argmin}_f \, \mathcal{L}_y(f(g^*(x)), y)\,; \tag{2}$$

- *joint training*, where the model is trained end-to-end. In this case,

$$g^*, f^* = \text{argmin}_{g,f} \left[\lambda \, \mathcal{L}_c(g(x), c) + \mathcal{L}_y(f(g(x)), y)\right]\,, \tag{3}$$

where $\lambda \geq 0$ is a hyperparameter that quantifies the relative importance of concept and task learning. $\lambda = 0$ corresponds to an end-to-end model.

Common choices for concept encoding include binary probabilities, soft probabilities or logits. We will refer to *hard* CBMs as independently trained models with binary concept encoding, while *soft* CBMs denote jointly trained models with soft probability-based concept activations. Additionally, we will refer to jointly trained models using logit-based concept encoding as *logit* CBMs.

More recently, CEMs (Figure 1) were proposed in Espinosa Zarlenga et al. (2022). Here a pair of vectors $\left(\hat{\boldsymbol{c}}_i^+, \hat{\boldsymbol{c}}_i^-\right) \in \mathbb{R}^{2d}$ is predicted for each input and concept c_i, $i = 1 \dots k$ (where the hyperparameter d indicates the embedding dimension). This pair of vectors is subsequently used to predict the activation $\hat{c}_i$ as a soft probability. The weighted vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}} = \hat{c}_i\, \hat{\boldsymbol{c}}_i^+ + (1 - \hat{c}_i)\, \hat{\boldsymbol{c}}_i^-$ are then constructed and concatenated before being fed into the final head to predict the label. In this setup $\hat{\boldsymbol{c}}_i^+$ and $\hat{\boldsymbol{c}}_i^-$ are meant to represent concept i being active or inactive respectively. Learning a two-fold representation for each concept enables interventions, by modifying the concept activations $\hat{c}_i$ and thus selecting either $\hat{\boldsymbol{c}}_i^+$ or $\hat{\boldsymbol{c}}_i^-$ as an input for the final head.

Training in CEMs is performed end-to-end. Similarly to joint CBMs, the loss function consists of reconstruction losses for both the concept activations and the class label, weighted by a parameter $\lambda \geq 0$ as in (3). To make them more receptive to interventions (i.e. to realize any improvements in task performance upon intervention), CEMs are exposed to interventions during training: each activation $\hat{c}_i$ is set to its ground-truth value c_i with probability defined by a parameter $p_{int} \in (0, 1)$.

Reasoning in CEMs takes place at the level of the vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$, as opposed to CBMs where it is based on concept activations $\hat{c}_i$. As high-dimensional objects trained only to be predictive of concepts $c_i \in \{0, 1\}$, the vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}} \in \mathbb{R}^d$ are expressive representations capable of encoding large amounts of information about the task label. We demonstrate in Section 8 that this additional task-relevant information is the main driver of task performance. While Espinosa Zarlenga et al. (2022) present CEMs as interpretable models that resolve the accuracy-interpretability trade-off, we argue that they are still affected by this trade-off: indeed, they favour accuracy over interpretability. Although requiring a case-by-case evaluation, other CEM-based architectures, such as Espinosa Zarlenga et al. (2023b); Ridzuan et al. (2024); Gao et al. (2024); Hu et al. (2024); Srivastava et al. (2024), are likely to face similar interpretability challenges.

3. Definition of leakage

The interpretability of CBMs and CEMs assumes that learnt concepts are aligned with ground-truth human-interpretable concepts. However, on general grounds that is not the case. During training, concept-based models often encode additional input information into the learned concepts to enhance task accuracy. When this happens *information leakage* occurs (Kazhdan et al., 2021; Margeloiu et al., 2021; Mahinpei et al., 2021; Havasi et al., 2022; Lockhart et al., 2022; Ragkousis and Parbhoo, 2024), and the model does not rely solely on the value of concepts for its task prediction, but also on leaked information, which

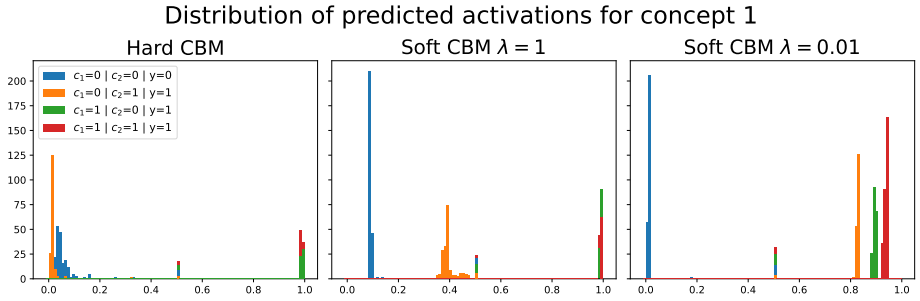

Figure 2: Distributions of predicted activations for concept 1 on the test set in CBMs with increasing leakage from left to right. In a faithful concept representation, the distribution should reflect the ground-truth value of the binary concept 1. Leakage appears as additional structure within these concept groups, separating samples according to the task label or other concepts. Colours are based on the ground-truth values of concepts and task label; to improve visibility, bars with lower counts are rendered in front of those with higher counts.

hinders interpretability. The reasoning of the model thus becomes effectively obscure to human overseers, to a degree specified by the amount of leakage.

In practice, information leakage manifests as additional structure arising in the learnt concept distributions, as shown in Figure 2. In this synthetic example, models must learn two binary concepts c_1, c_2 and the binary task is to predict the function $y = c_1 \texttt{ OR } c_2$ (see Appendix A.1 for more details). We compare the distributions of activations for the first concept by a hard CBM and two soft CBMs, with either intermediate ($\lambda = 1$) or low ($\lambda = 0.01$) concept supervision. By definition no leakage can be present in hard CBMs, since concept learning is independent of task learning, and concepts are converted to binary values before being passed to the final head, preventing leaked information. Soft CBMs with lower concept supervision are instead more prone to leakage as accurate concept learning is not enforced during training (see Koh et al., 2020; Kazhdan et al., 2021, and our results in Section 7).

Neglecting the peak close to 0.5, which corresponds to observations the model is most uncertain about, the concept distribution in the hard CBM is very close to the binary ground-truth distribution for concept 1. In the soft CBM with $\lambda = 1$ additional structure is learnt based on the value of the other concept and of the class label, while for $\lambda = 0.01$ the concept itself is a very strong predictor of the value of both the class label and the other concept. In particular, the distribution is more similar to the ground-truth binary task distribution than to the concept distribution.

Based on these observations, we identify two main types of leakage that concept-based models suffer from, and propose the following classification:

1) *Concepts-task leakage:* additional information about the task is stored into the learnt concepts. For training to be successful, the annotated concepts must be predictive of the task label as measured by the ground-truth mutual information (MI, see more below) between each concept and the task label. If concepts-task leakage is present, then the learned concepts-task MI will be higher than ground-truth concepts-task MI. This is the predom-

inant type of leakage and the main cause of non-interpretability, as it directly allows the model to achieve a better task performance (see experiments in Section 6).

2) *Interconcept leakage:* additional information about the other concepts is stored into each learned concept. In a given dataset, concepts are predictive of the value of other concepts as determined by their ground-truth pair-wise MI. This type of leakage manifests as learned concepts being more predictive of the value of the other learned concepts, resulting in a learned interconcept MI being higher than the ground-truth. This effect is usually a secondary factor of non-interpretability (see experiments in Section 6), but it may be beneficial for models as an internal error-correcting tool: interconcept leakage provides redundant pathways for the model to attain high task performance even when concept predictions are poor (Heidemann et al., 2023).

This classification, highlighting the information-theoretic nature of leakage, forms the foundation of our quantitative approach to detect and measure it.

4. Existing measures of interpretability and their shortcomings

Previous works (Koh et al., 2020; Kazhdan et al., 2021; Margeloiu et al., 2021; Mahinpei et al., 2021; Havasi et al., 2022; Espinosa Zarlenga et al., 2023a) have proposed several measures of leakage, however none take into account its information-theoretic nature outlined in Section 3. As we highlight in this section, these existing measures consequently exhibit inherent limitations in sensitivity and robustness, limiting their usefulness in most practical scenarios.

Concept performance metrics. Although performance metrics such as concept accuracy, F1 score and AUC are general quality indicators of concept learning, they are not sensitive to more subtle effects undermining interpretability, such as leakage. In particular, two models may exhibit the same evaluation scores while considerably differing in terms of leakage. This is apparent from the example in Figure 2, where a classification threshold of 0.5 yields essentially the same concept accuracy and F1 score for the hard and $\lambda = 1$ soft CBM, while the latter encodes a non-trivial amount of additional structure. Concept AUC is a finer measure of interpretability than accuracy and F1 score as it captures more information about the concept distribution, and in this simple example it is able to discriminate between those two models. However it is not generally sensitive enough to quantify leakage in more complex setups as we illustrate in the following experiments.

OIS and NIS. These scores were defined in Espinosa Zarlenga et al. (2023a) to capture leakage in concept-based models. The Oracle Impurity Score (OIS) is obtained by training $k(k-1)$ additional neural networks $\psi_{i,j}$ (typically 2-layer perceptrons of modest size), whose objective is to predict the value of the ground-truth concept j from the learnt representation of concept i. The $k \times k$ impurity matrix defined as $\pi_{ij}(\hat{c}, c) = \text{AUC}\left(\psi_{i,j}(\hat{c}_i), c_j\right)$ is meant to estimate the predictivity of each learnt concept representation with respect to any other across the dataset. The OIS is the average difference between the predictivities of learnt and ground-truth concepts over pairs of concepts,

$$\text{OIS}(\hat{c}, c) = \frac{2}{k} \left|\left|\pi(\hat{c}, c) - \pi(c, c)\right|\right|_F, \tag{4}$$

where $||\cdot||_F$ indicates the Frobenius norm, and the normalisation ensures $0 \leq \text{OIS} \leq 1$. The OIS is thus intended to provide an estimate of interconcept leakage – although, it is unable to account for concepts-task leakage, which is the prevalent effect as we illustrate in Section 6. Although demonstrated to be superior to other concept-quality metrics in Espinosa Zarlenga et al. (2023a), the OIS is subject to additional limitations that hinder its applicability. In particular, in Section 6 we demonstrate that:

1. its value is typically non-vanishing and relatively high for hard CBMs, indicating bias and a lack of sensitivity;

2. due to the intrinsic stochasticity in its definition which involves training neural networks, it is extremely variable when repeatedly evaluated for a given model at test time. The resulting broad confidence intervals typically prevent drawing any statistically significant conclusion when comparing models with different amounts of leakage, ultimately limiting the utility of this score for model design.

The Niche Impurity Score (NIS) was defined to capture leakage manifesting as additional information about concept j stored into a joint subset of the learnt representations $\{i_1, \ldots, i_p\}$ not containing concept j (see Espinosa Zarlenga et al., 2023a, for details). This score is meant to go beyond pair-wise interconcept leakage and estimate (generally subordinate) higher-order effects. As illustrated in Appendix B, the NIS is generally non-vanishing and very high for hard CBMs, and moreover it appears to be anticorrelated with intervention performance and leakage, casting doubt on its suitability as a measure of leakage.

Performance upon intervention. Interventions can be used to evaluate the interpretability of a model. They are performed by substituting the ground-truth values of concepts to the corresponding predicted activation, with a given policy defining the order of the concepts to intervene on. The behaviour of the task accuracy after each intervention measures the extent to which the learnt and ground-truth concepts are aligned. A task accuracy y_{acc} that decreases is a coarse indication of leakage since it suggests that the final head expects additional information that is not present in the ground-truth concept distribution. Figure 3 shows the intervention performance of the same three models as in Figure 2. As expected, models with higher leakage have worse intervention performance, while the hard CBM displays a monotonically increasing task accuracy.

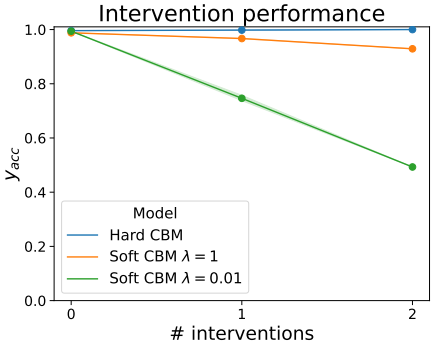

Figure 3: Performance upon random intervention of the three models analysed in Figure 2. See Appendix A.3 for more details.

As a metric of leakage, however, intervention performance has several limitations:

1. Because it is based on evaluation scores, it does not address the information-theoretical nature of leakage, resulting in a limited resolution. It is unable to distinguish between

interconcept and concepts-task leakage, and it does not directly quantify interpretability at the level of individual concepts. It is therefore often insufficiently informative to guide model design or hyperparameter selection.

2. Its evaluation is computationally expensive, especially in real-world applications in which models have a large number of concepts and are evaluated using state-of-the-art policies with a high overhead such as CooP (Chauhan et al., 2023).

3. In more complex architectures such as CEMs, it is not a measure of leakage, and models with higher leakage may display better performance upon interventions (see Appendix G).

Collectively these issues indicate that poor intervention performance is a measure of leakage with perfect specificity (leakage has always occurred when intervention performance is poor), but with poor sensitivity (when intervention performance is good, leakage is not guaranteed to be absent). Although the leakage scores we propose in Section 5 address these issues, intervention performance generally still remains a coarse but useful indicator of leakage in CBMs.

Let us denote by $y_{acc}^{*(k)}$ the task accuracy after all the k concepts have been intervened on in the model under evaluation, which coincides with the accuracy of the final head on the ground-truth concepts. As a benchmark of the proposed leakage scores we will use the intervention score

$$\mathrm{S}_{int} = y_{acc}^{(k)} - y_{acc}^{*(k)}\,, \tag{5}$$

where $y_{acc}^{(k)}$ is the reference accuracy of the final head when separately trained on the ground-truth concepts. It represents the maximal task accuracy attainable by the final head and it measures the fundamental misspecification of the final head with respect to the ground-truth functional dependence $y = f(c)$. S_{int} thus quantifies the further decrease in task accuracy caused by leakage accounting for final head misspecification. Its value is independent of the intervention policy adopted. Note that $y_{acc}^{(k)}$ coincides with $y_{acc}^{*(k)}$ for a successfully trained hard model, thus $\mathrm{S}_{int} = 0$ for hard models by construction.

5. Concepts-task and interconcept leakage scores

Information leakage is information-theoretic in nature and, as such, it is most appropriately captured by metrics based on quantities from information theory. As discussed in Section 3, the leaked information and the corresponding additional structure in the learnt concept distributions can be measured as modifications in the ground-truth interconcept and concepts-task MIs respectively. This motivates the definition of the following set of scores based on information theory to quantify leakage and the interpretability of concept-based models:

Concepts-task leakage scores. Denoting by $H(z)$ the entropy of the random variable z and by $I(z, w)$ the MI between the random variables z and w, we define the concepts-task leakage score (CTL) for concept i as the difference between predicted and ground-truth MIs between concept i and the task label,

$$\mathrm{CTL}_i(\hat{c}_i, c_i, y) = \max\left(0, \frac{I(\hat{c}_i, y)}{H(y)} - \frac{I(c_i, y)}{H(y)}\right). \tag{6}$$

This score quantifies the additional information that the learnt concept i encodes about the task label with respect to the ground-truth. MIs are appropriately normalised by the entropy of the class label to ensure $0 \leq \mathrm{CTL}_i \leq 1$. When the difference inside the brackets is negative, the learnt concepts are less predictive of the task than expected from the ground truth, indicating poor concept learning rather than leakage. We further define the average concepts-task leakage score as

$$\mathrm{CTL}(\hat{c}, c, y) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{CTL}_i(\hat{c}_i, c_i, y)\,, \tag{7}$$

as the average extra information about the task stored into the learnt concepts. Note that a weighted average over concepts can be adopted in use cases where concept importances are non-uniform.

Interconcept leakage scores. In a similar fashion, we define the pairwise interconcept leakage score (ICL) between concepts i and j as

$$\mathrm{ICL}_{ij}(\hat{c}, c) = \max\left(0, \frac{I(\hat{c}_i, \hat{c}_j)}{\sqrt{H(\hat{c}_i)\,H(\hat{c}_j)}} - \frac{I(c_i, c_j)}{\sqrt{H(c_i)\,H(c_j)}}\right), \tag{8}$$

which measures the additional predictivity of the learnt concept i for concept j with respect to ground-truth. Again, normalisation by the entropies ensures that $0 \leq \mathrm{ICL}_{ij} \leq 1$. Note that $\mathrm{ICL}_{ii} = 0$ since $I(z, z) = H(z)$. We also define the per-concept and average interconcept leakage scores as

$$\mathrm{ICL}_i(\hat{c}, c) = \frac{1}{k-1} \sum_{j=1}^{k} \mathrm{ICL}_{ij}(\hat{c}, c)\,, \qquad \mathrm{ICL}(\hat{c}, c) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{ICL}_i(\hat{c}, c)\,, \tag{9}$$

to quantify the extra information that each concept encodes about the other concepts, and the average additional interconcept predictivity respectively.

The CTL and ICL scores are meant to summarise complementary aspects of the leakage present in a given concept-based model, as well as its overall interpretability. Based on these measures, we define the following criterion to detect leakage:

Leakage Criterion. *When comparing two models (or two model classes) A and B, a sufficient condition for A having higher leakage than B is that either of the following conditions are satisfied,*

- *both the* CTL *and* ICL *scores are higher in A than in B with high statistical confidence,*
- *either the* CTL *or the* ICL *score is higher in A with high statistical confidence, while the other score takes statistically compatible values in A and B.*

Note that this is a conservative criterion covering the majority of cases, while it cannot be used in the cases where one measure is higher in A and the other is lower in B (or vice versa) with high statistical confidence.

These measures also convey more detailed information about each type of leakage at the level of single concepts. This is particularly useful for model design and when performing interventions (as they essentially indicate the risk of intervening on each concept). The framework above can be applied to concept-based models with a range of concept encodings, including binaries, soft probabilities, logits and vector representations. Alternative normalisations of MI can be used, although we found those adopted in (6) and (8) to be particularly transparent from an information theory perspective, and robust in the experiments we carried out. We estimate MI and entropy using the Kraskov-Stögbauer-Grassberger (KSG) estimator (Kraskov et al., 2004), based on k-nearest neighbour statistics.

6. Robustness of leakage scores

We now demonstrate the robustness of the CTL and ICL scores and how they overcome the issues of the existing measures of leakage.

Datasets. For our experiments we considered three synthetic datasets, TabularToy(δ), dSprites(γ) and 3dshapes(γ) (presented in Espinosa Zarlenga et al. (2023a)), as well as two real-world datasets, the Caltech-UCSD Birds-200-2011 dataset (CUB, Wah et al., 2025) and HAM10K (Tschandl et al., 2018). TabularToy(δ) is a binary-class tabular dataset based on the dataset from Mahinpei et al. (2021), while dSprites(γ) and 3dshapes(γ) are multi-class image-based datasets building upon dSprites (Matthey et al., 2017) and 3dshapes (Burgess and Kim, 2018) respectively. The CUB dataset contains bird images annotated with 112 binary attributes capturing phenotypic traits, and the task is to classify each image into one of 200 species. HAM10K consists of dermatoscopic lesion images labelled with 18 morphological concepts, with the task of classifying each lesion as benign or malignant. The use of synthetic and real-world datasets enables us to (i) modify the functional dependence $y = f(c)$ and hence tune the ground-truth concepts-task MI, and (ii) tune the ground-truth interconcept MI via the parameters $\delta \in (0, 1)$ and γ ($\in \{0, \dots, 4\}$ in dSprites and $\in \{0, \dots, 5\}$ in 3dshapes), an essential aspect to assess interpretability (Heidemann et al., 2023). See Appendix A for more details of these experiments.

CTL and ICL are highly sensitive to leakage. We consider pairs of models encoding different levels of leakage as assessed by intervention performance (Figure 4). The models in each pair were chosen to have essentially identical concepts and task evaluation scores (Table 1), evidencing the limitations of these metrics in capturing leakage. In all cases the CTL and ICL scores correctly detect the different amounts of leakage according to the Leakage Criterion, with minimal uncertainty. In contrast, the OIS has large confidence intervals and so does not allow statistically significant conclusion on leakage to be drawn. Appendix C presents the concept-wise leakage scores for these models and illustrates how they provide more fine-grained information at the level of individual concepts.

CTL and ICL are vanishing for hard CBMs. Figure 5 shows the leakage scores for hard models trained on a number of datasets in both low and high regimes of ground-truth interconcept MI. By construction hard CBMs have vanishing leakage (Koh et al., 2020; Kazhdan et al., 2021; Margeloiu et al., 2021). In all cases, the CTL and ICL scores do not deviate significantly from zero, regardless of the interconcept MI. In contrast, the OIS is non-vanishing in all cases, indicating that it does not meet a fundamental requirement as a

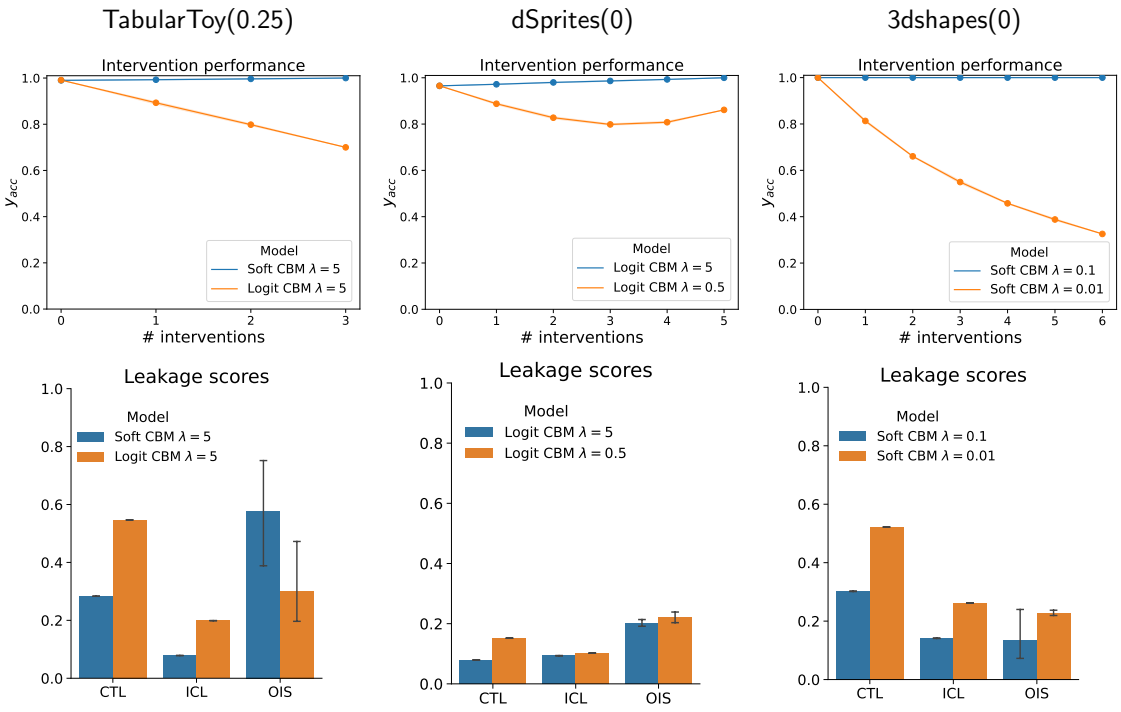

Figure 4: Intervention performance and leakage scores computed for pairs of soft and logit CBMs with different amounts of leakage but similar evaluation scores (see Table 1) trained on TabularToy(0.25), dSprites(0) and 3dshapes(0). Metrics are evaluated on a 5-fold basis for each model.

		c_{acc}	c_{F1}	c_{AUC}	y_{acc}	y_{F1}	y_{AUC}	$\text{S}_{int}(\downarrow)$
TabularToy(0.25)	Soft CBM $\lambda = 5$	0.993	0.992	0.992	0.990	0.990	0.990	0.000
	Logit CBM $\lambda = 5$	0.995	0.995	0.995	0.991	0.991	0.991	0.301
dSprites(0)	Logit CBM $\lambda = 5$	0.993	0.993	0.993	0.965	0.976	0.698	0.000
	Logit CBM $\lambda = 0.5$	0.992	0.992	0.992	0.965	0.975	0.689	0.139
3dshapes(0)	Soft CBM $\lambda = 0.1$	1.000	1.000	1.000	1.000	1.000	0.201	0.000
	Soft CBM $\lambda = 0.01$	1.000	1.000	1.000	1.000	1.000	0.207	0.674

Table 1: Concept and task evaluation scores, and intervention scores for pairs of soft and logit CBMs with different levels of leakage (see Figure 4).

measure of leakage. We further note that for each dataset the estimated OIS for hard CBMs are comparable to those of the models shown in Figure 4, which exhibit a significant amount of leakage. Thus, in these cases OIS does not distinguish interpretable from uninterpretable concept-based models.

CTL and ICL strongly correlate with intervention performance. We trained a number of soft and logit CBMs on each dataset in both high and low regimes of ground-truth interconcept MI and with different levels of concept supervision (see Appendix D for details

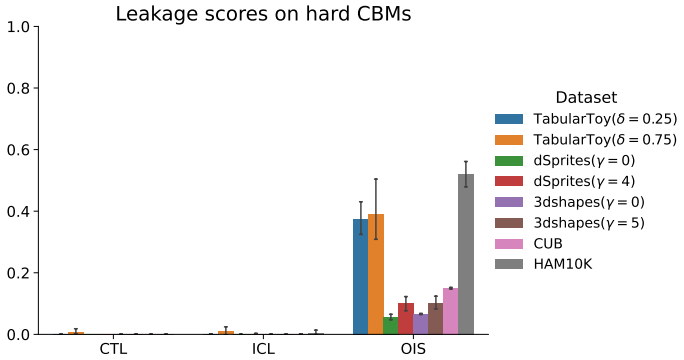

Figure 5: Leakage scores evaluated on hard CBMs on datasets with high and low ground-truth interconcept MI. The 95% confidence intervals are obtained from 5-fold training on each dataset.

	CTL		ICL		OIS	
	Pearson r	p-value	Pearson r	p-value	Pearson r	p-value
TT(0.25)	**0.75**	$\mathbf{3.4 \times 10^{-3}}$	0.69	1.4×10^{-2}	0.49	1.2×10^{-1}
dS(0)	**0.83**	$\mathbf{1.1 \times 10^{-3}}$	0.68	1.5×10^{-2}	0.70	1.1×10^{-2}
3ds(0)	0.84	4.2×10^{-5}	**0.86**	$\mathbf{2.4 \times 10^{-5}}$	0.82	2.0×10^{-4}
TT(0.75)	**0.54**	$\mathbf{4.9 \times 10^{-2}}$	0.51	1.1×10^{-1}	0.27	4.1×10^{-1}
dS(4)	**0.67**	$\mathbf{1.9 \times 10^{-2}}$	0.48	1.2×10^{-1}	0.55	7.4×10^{-2}
3ds(5)	**0.79**	$\mathbf{3.9 \times 10^{-4}}$	0.69	3.5×10^{-2}	0.70	6.5×10^{-3}
CUB	**0.72**	$\mathbf{9.0 \times 10^{-3}}$	0.67	2.1×10^{-2}	0.44	1.1×10^{-1}
HAM10K	**0.49**	$\mathbf{1.8 \times 10^{-2}}$	0.41	5.4×10^{-2}	-0.06	7.7×10^{-1}

Table 2: Pearson r coefficients and corresponding p-values measuring the correlations of CTL, ICL and OIS metrics against S_{int}.

on this experiment). The evaluation of the OIS, ICL and CTL scores was repeated 5 times for each individual model to provide estimates of uncertainty. We then took 10,000 Monte-Carlo (MC) samples from the resulting distributions of each score (assuming normality), and for each sample we computed the Pearson r coefficient between each score and the intervention score S_{int}. Pooling using Rubin's rule (see Rubin, 2004; Barnard and Rubin, 1999), we obtain a single best estimate for the Pearson r and its associated p-value for each leakage score and dataset, which are reported in Table 2.

The proposed CTL and ICL scores strongly correlate with intervention performance, they systematically outperform OIS across all datasets and correspond robustly to ground-truth interconcept MI. This analysis also quantifies the importance of concepts-task and interconcept leakage in each case. CTL generally appears as the prominent form of leakage over ICL. Its correlation with intervention performance is typically stronger than ICL, and is, moreover, significant in all datasets. This behaviour suggests that optimization favours the storage of additional task-relevant information in the learnt concepts rather than information about the other concepts. Nevertheless, these results also provide evidence that interconcept leakage systematically arises during model training; indeed, correlations

of comparable magnitude to concept-task leakage were observed in the experiments on 4 out of 8 datasets.

7. Causes of leakage

The previous results indicate that our proposed information-theoretic leakage scores are robust and sensitive and can therefore be used to identify the most common causes of leakage and assess their impact on interpretability.

A first effect that directly impacts the interpretability of a concept-based model is the concept encoder being misspecified with respect to the target concept distribution. When such a misspecification is significant, the model is essentially not capable of learning concepts and their relations. This may result in additional issues, such as non-locality (Raman et al., 2023; Huang et al., 2024). To address this issue, concept-based architectures that enforce accurate learning of interconcept relations have been recently proposed (Kim et al., 2023; Vandenhirtz et al., 2024; Xu et al., 2024; Kim et al., 2025). We recognise however that this problem is strongly use-case specific and it is usually sufficient to opt for a concept encoder with an improved architecture to solve it. In the following experiments we will thus consider concept encoders that are sufficiently well-specified, as relevant for real-world high-risk scenarios where the aim is both high concept performance and the absence of leakage.

Insufficient concept supervision. When λ is set to be small in the loss (3) during training, the model will achieve better task performance at the cost of poor concept learning and high leakage. This issue was originally observed in Koh et al. (2020), where poor intervention performance was found when training models with low λ. To precisely quantify this effect, we trained soft and logit CBMs with low, intermediate and high values of λ on datasets with both low and high interconcept MI. The resulting leakage scores, displayed in Figures 6 and 20, confirm that leakage decreases as concept supervision is increased.

However, as detailed in Appendix E, leakage does not fall to zero as λ is increased – instead, the leakage scores reach a non-vanishing minimal value, specific to each model class. At higher values (here, $\lambda \gtrsim 10$), models focus excessively on concept learning, and the task objective may be poorly learnt. This results in a drop in task performance or an increase in leakage. Each model class thus has an associated optimal range of $\lambda \in (\lambda_{min}, \lambda_{max})$ where the minimal amount of leakage is attained, below which models exhibit proportionally higher leakage and above which training often fails. As evidenced by our results, such an optimal interval can be identified using the CTL and ICL scores.

Over-expressive concept encoding. When the chosen concept representation is significantly more expressive than the annotated concept representation, the model will generally misuse the surplus expressivity to store additional information in the concepts to aid task prediction. This phenomenon is illustrated in Figures 6 and 20 where annotated concepts are binary, while learnt concepts are soft probabilities and logits, and as such over-expressive for the setup. We note that raising λ is less effective at decreasing leakage when concepts representations are more expressive, as in the logit models, and furthermore the minimal attainable leakage is significantly higher for such models over soft CBMs. CEMs vector representations with binary ground-truth concepts exhibit a similar behaviour, as we discuss

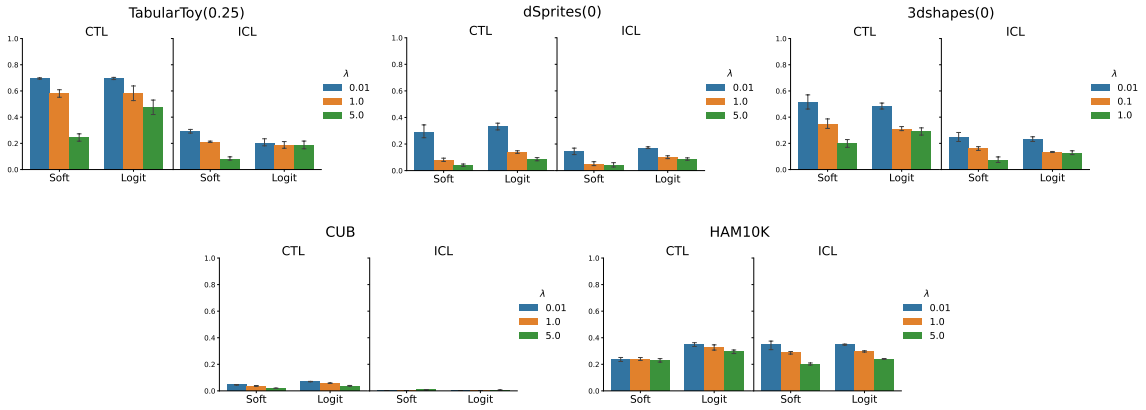

Figure 6: Leakage scores evaluated for soft and logit models with different levels of concept supervision across datasets.

in Section 8. Finally, we observe that models with low λ generally have a comparable level of high leakage regardless of the concept encoding.

Incomplete set of concepts. When the set of annotated concepts is incomplete for the task – meaning that it is not sufficiently predictive – models tend to encode information about the missing concepts (or, more generally, any additional information that may be required) into the learnt concepts to nonetheless achieve a high task performance. This typically translates into sizeable leakage and a loss of interpretability. This issue has been discussed in previous work (Kazhdan et al., 2021; Margeloiu et al., 2021; Mahinpei et al., 2021; Havasi et al., 2022; Marconato et al., 2022).

To accurately assess this phenomenon using our information-theoretic scores, we trained soft CBMs with low, intermediate and high levels of concept supervision on a range of datasets with a significant fraction of concepts removed (see Appendix A.1 for more details on which concepts are removed for each dataset). This resulted in a sizeable decrease of the reference task accuracy $y^{(k)}_{acc}$ of the final head trained on the ground-truth concepts introduced in equation (5) (Table 3, left). The leakage scores of the trained models are displayed in Figure 7, along with those of models with complete sets of concepts for comparison. The relevant concepts and task accuracies are presented in Appendix F.

According to the Leakage Criterion and as expected, we detected more leakage in 14/15 model classes trained on incomplete sets of concepts. At low and intermediate values of λ, there is significantly more leakage in models trained on an incomplete set.

Misspecified final head. If the final head is not flexible enough to learn the ground-truth $y = f(c)$ dependence, during training then the model tends to store the necessary dependences into the learnt concepts, which are thereby deformed to encode additional information about the task and the other concepts. In many practical applications a common choice for the final head of a CBM is a linear layer which in principle allows for full model explainability. In reality, many tasks are best described by a non-linear function of the concepts, and a linear layer can be misspecified in such cases.

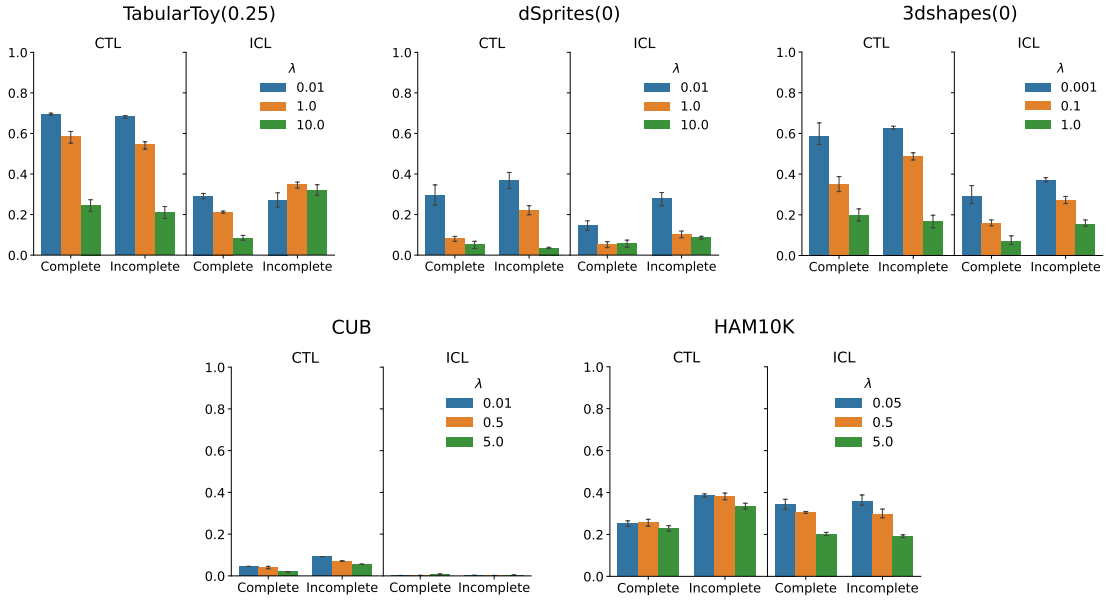

Figure 7: Leakage scores evaluated for soft CBMs at different levels of concept supervision, on datasets with complete and incomplete sets of concepts.

$y_{acc}^{(k)}$	complete	incomplete	well-specified	misspecified
TabularToy(0.25)	1.000 ± 0.000	0.786 ± 0.000	1.000 ± 0.000	0.687 ± 0.000
dSprites(0)	1.000 ± 0.000	0.749 ± 0.004	1.000 ± 0.000	0.807 ± 0.000
3dshapes(0)	1.000 ± 0.000	0.655 ± 0.002	1.000 ± 0.000	0.859 ± 0.008
CUB	1.000 ± 0.000	0.836 ± 0.012	1.000 ± 0.000	–
HAM10K	0.999 ± 0.001	0.840 ± 0.009	0.999 ± 0.001	–

Table 3: First two columns: baseline $y_{acc}^{(k)}$ task accuracies of final heads trained on datasets with complete and incomplete sets of concepts. Second two columns: baseline $y_{acc}^{(k)}$ task accuracies of a linear head on the linear (*well-specified*) and non-linear (*misspecified*) ground-truth task labels. We report 95% confidence intervals over 5-fold training.

To explore the general effects of final head misspecification we considered soft CBMs with a linear head trained on datasets where the ground-truth task is either a linear or non-linear function of the concepts. To do so we modified the functional dependences in TabularToy(0.25), dSprites(0) and 3dshapes(0) by adding non-linear terms to the original tasks (see Appendix A.1 for details). As a baseline measuring the final head misspecification, for each dataset we trained a separate linear head on the ground-truth concepts. Table 3 shows the resulting decrease in task performance in terms of the reference task accuracy $y_{acc}^{(k)}$. The corresponding leakage scores are displayed in Figure 8 and their concepts and task accuracies are presented in Appendix F. According to the Leakage Criterion, a misspecified final head causes a higher leakage in 7/9 model classes. Moreover, high concept supervision

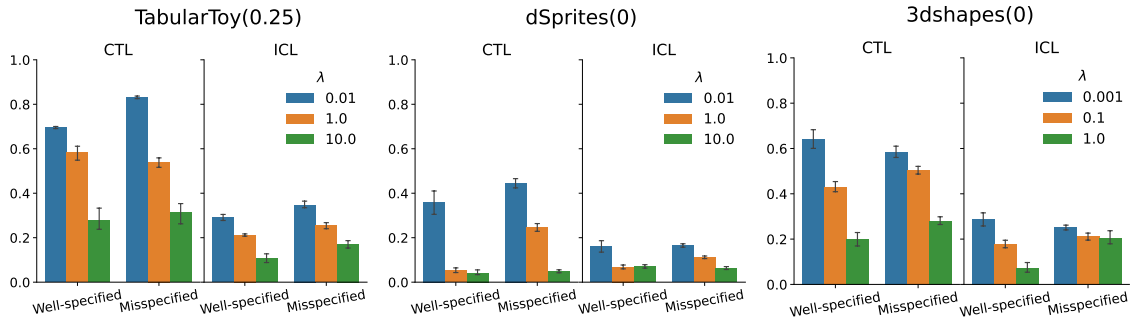

Figure 8: Leakage scores evaluated for soft CBMs with a linear final head trained on datasets where the task is either a linear (*Well-specified*) or a non-linear (*Misspecified*) function of the concepts, at low, intermediate and high levels of concept supervision.

is generally not sufficient to compensate for the additional leakage induced by the final head misspecification.

8. Leakage in CEMs

CEMs (Espinosa Zarlenga et al., 2022) are widely recognized for achieving higher task accuracy than CBMs, particularly when the concept set is incomplete, and can often approach the performance of end-to-end models without concept supervision. CEMs reasoning is based on data point-dependent concept vector representations, and not directly on concept activations as in CBMs. As noted in Espinosa Zarlenga et al. (2022, 2023b, 2025), this allows them to encode additional task-related information into concept vectors boosting task performance. In the following, we show that our information-theoretic framework confirms this picture, and that CEM's design results in concepts-task leakage for any value of the hyperparameters. Moreover, we find that CEMs exhibit significant interconcept leakage which, in contrast to CBMs, typically increases with higher concept supervision. Finally, we identify *alignment leakage*, a novel subtype of concepts-task leakage that can arise in models exposed to interventions during training such as CEMs. Taken together, these observations indicate that CEMs may be ill-suited for high-risk applications due to their limited interpretability.

Information-theoretic measures of CEM leakage. CEMs incorporate several concept representations – the concept probabilities $\hat{c}_i$ just as CBMs, as well as the vector encodings $\hat{\boldsymbol{c}}_i^+, \hat{\boldsymbol{c}}_i^-$ and $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$. Reasoning happens at the level of the weighted vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$, making them critical objects where leakage can hinder interpretability. Entropy and MI estimators are generally affected by biases depending on the dimension of variables, as a consequence of the curse of dimensionality (Lord et al., 2018; Carrara and Ernst, 2019; Czyż et al., 2023; Gowri et al., 2024). Thus, one cannot compare quantities estimated for concept representations of different dimensions as required by the definitions of the CTL and ICL scores in equations (6)-(8), since this would involve subtracting normalised MIs estimated on the ground-truth concepts c_i from those on the vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$. To avoid this bias

we will therefore analyse the behaviour of the normalised MIs on $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ across different models without referring to ground-truth normalised MIs.

We define the concepts-task normalised MI between each vector $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ and the task label, averaged across the k concepts,

$$\widetilde{I}^{(CT)}\left(\hat{\boldsymbol{c}}^{\boldsymbol{w}}, y\right) = \frac{1}{k}\sum_{i=1}^{k}\frac{I\left(\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}, y\right)}{H(y)}. \tag{10}$$

This score captures how informative the weighted vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ are, on average, of the task. This score can be considered as a simpler version of the CTL score we introduced for CBMs, that does not use the ground-truth concepts-task MI as a reference.

In a similar spirit to the ICL scores, we can also define a simplified metric capturing the amount of information that a concept representation $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ encodes about the other concepts. Averaging over the $k(k-1)/2$ non-trivial concept pairs, we get:

$$\widetilde{I}^{(IC)}(\hat{\boldsymbol{c}}^{\boldsymbol{w}}, c) = \frac{2}{k(k-1)}\sum_{i=1}^{k}\sum_{j<i}\frac{I(\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}, c_j)}{H(c_j)}. \tag{11}$$

Note that in contrast to the ICL scores, $\widetilde{I}^{(IC)}$ does not measure how much a learnt concept representation is predictive of another. This approach, based on annotated concepts instead, has the benefit of limiting the overall dimensionality of the space used to estimate MIs (in this case $d+1$ instead of $2d$), mitigating the effects of dimension-associated bias.

An associated quantity that is relevant whenever higher-dimensional representations contain non-trivial information is the normalised MI of a concept representation $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ relative to its ground-truth concept c_i. Its average over the k concepts is:

$$\widetilde{I}^{(\text{self})}(\hat{\boldsymbol{c}}^{\boldsymbol{w}}, c) = \frac{1}{k}\sum_{i=1}^{k}\frac{I(\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}, c_i)}{H(c_i)}, \tag{12}$$

which provides a measure of how predictive each concept representation is of its own concept, averaged over concepts.

Our framework detects CEM's concepts-task leakage. By design, the concept representations $\hat{\boldsymbol{c}}_i^{+}$, $\hat{\boldsymbol{c}}_i^{-}$ and $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ in CEMs are encouraged to encode additional task-relevant information besides the concept state itself, with the aim of improving task performance. This is a consequence of the fact that no direct supervision is performed on the inputs $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ to the final head, but only on the predicted concept activations $\hat{c}_i$ via the concept reconstruction loss. This means that in the forward pass, an arbitrary amount of information is allowed to flow from the input x to the task prediction through the concept vectors, as long as the embeddings $\hat{\boldsymbol{c}}_i^{+}$ and $\hat{\boldsymbol{c}}_i^{-}$ generated from x are predictive of the scalar c_i.

An effect of this task-relevant information is the clear clustering of the embeddings based on the value of the task label, as shown in Figure 9 (see also Figure 23 for similar results at non-vanishing p_{int}, as well as Espinosa Zarlenga et al. (2022); Santis et al. (2025)). This parallels the structure that appears in the concept distributions of CBMs affected by leakage (in this case, accounting for higher dimensional concept representations).

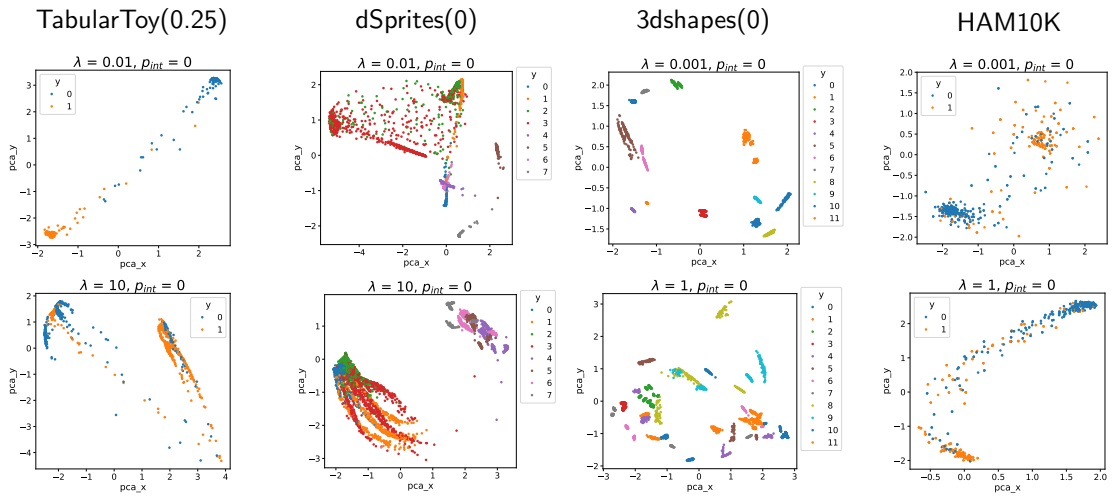

Figure 9: 2-dimensional PCA projections of the weighted embeddings $\hat{\boldsymbol{c}}_1^{\boldsymbol{w}}$ for the first concept across datasets and for different values of λ at $p_{int} = 0$. The colouring indicates the value of the ground-truth task label. See Espinosa Zarlenga et al. (2022) for similar representations in CUB.

Information theory offers additional insight into this phenomenon, in particular via the $\widetilde{I}^{(CT)}(\hat{\boldsymbol{c}}^{\boldsymbol{w}}, y)$ score defined in (10). As shown in Figure 10 this score is representative of the degree of structure in the learnt embeddings given in Figures 9 and 23, and is closely associated with task accuracy. On the other hand, concept accuracy does not correlate with task accuracy or $\widetilde{I}^{(CT)}$. This not only indicates the presence of concepts-task leakage, but also suggests that CEMs rely heavily on concepts-task leakage for their success: regardless of concepts, task accuracy is higher in models where the vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ encode more information about y. This picture agrees with the previous evidence that task accuracy is insensitive to the number of used concepts (see Espinosa Zarlenga et al. (2022), Appendix 8). In particular, when training CEMs on the CUB dataset with up to 90% of the annotated concepts removed, the task accuracy remains essentially the same as that of models trained on all the concepts.

Interconcept leakage increases with concept supervision. On closer inspection of the structure of the learned embeddings $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$, strong clustering is also observed based on the ground-truth values of the other concepts $j \neq i$. For example, Figure 11 shows PCA projections of $\hat{\boldsymbol{c}}_1^{\boldsymbol{w}}$ in terms of the values of c_2 and c_3 (see Figure 24 for similar plots for models with non-vanishing p_{int}). This phenomenon is particularly apparent in the TabularToy(0.25) data, where the task is binary and there are three concepts: in this case, the fine structure present in Figures 9 and 23 which is not explained by the y labels is clearly associated with the value of the other concepts.

This effect can be quantified using the information-theoretic measures $\widetilde{I}^{(IC)}(\hat{\boldsymbol{c}}^{\boldsymbol{w}}, c)$ and $\widetilde{I}^{(\text{self})}(\hat{\boldsymbol{c}}^{\boldsymbol{w}}, c)$ defined in equations (11)-(12). Each vector $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ becomes more and more predictive of the corresponding c_i as one increases λ and p_{int}, as indicated in Figure 12 by higher c_{acc} and $\widetilde{I}^{(\text{self})}$. This is the expected and desired behaviour as it corresponds to increasing concept supervision. However, $\widetilde{I}^{(IC)}$ undergoes a very similar growth, indicating

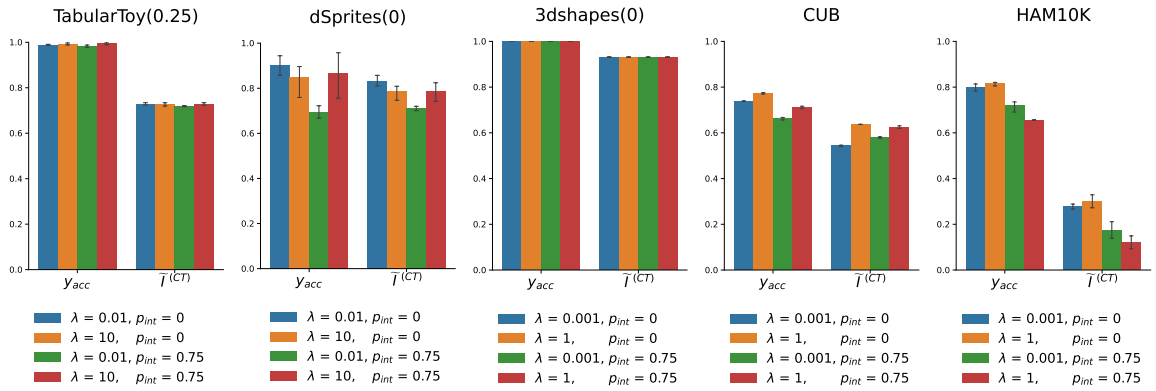

Figure 10: Task accuracy and $\widetilde{I}^{(CT)}$ score on a range of CEMs and datasets.

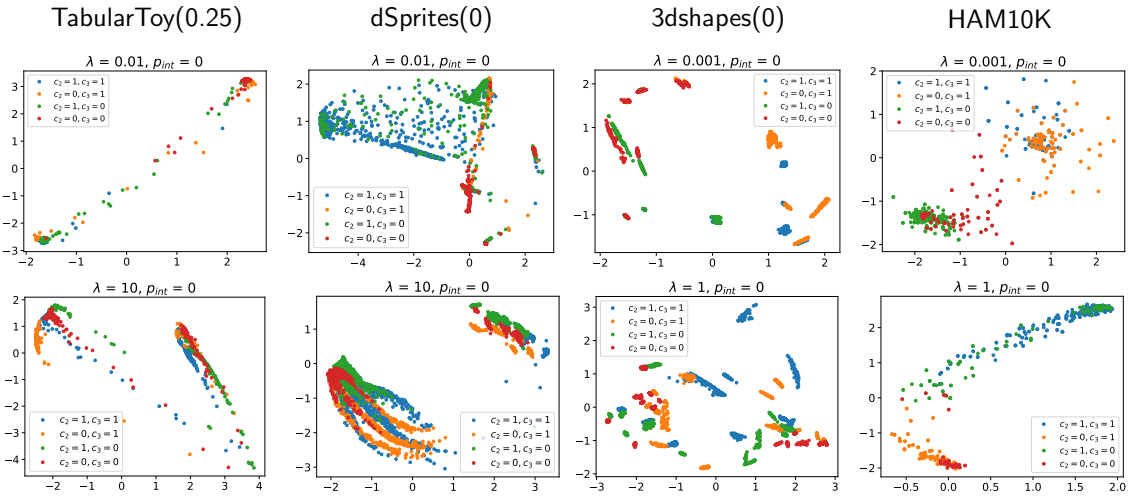

Figure 11: 2-dimensional PCA projections of the weighted embeddings $\hat{\boldsymbol{c}}_1^{\boldsymbol{w}}$ for the first concept across datasets and for different values of λ at $p_{int} = 0$. The colouring indicates the ground-truth value of concepts 2 and 3.

that the vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ contain increasing information about the other concepts, which, by definition, is interconcept leakage. Thus, and in contrast to CBMs, there is no choice of the hyperparameters λ and p_{int} that improves concept learning while decreasing interconcept leakage.

To gain deeper insight into interconcept leakage in CEMs, we evaluated the CTL and ICL scores on the predicted concept probabilities $\hat{c}_i$ (Figure 13). Although not indicative of model interpretability – since reasoning occurs at the level of the weighted vectors for CEMs – it is worth noting that they decrease in a manner similar to CBMs as concept supervision increases. The picture that emerges is thus that as one raises λ and p_{int}, the leaked information vanishes from the concept probabilities, and is effectively transferred into the vector representations as interconcept leakage. The vector representations that are

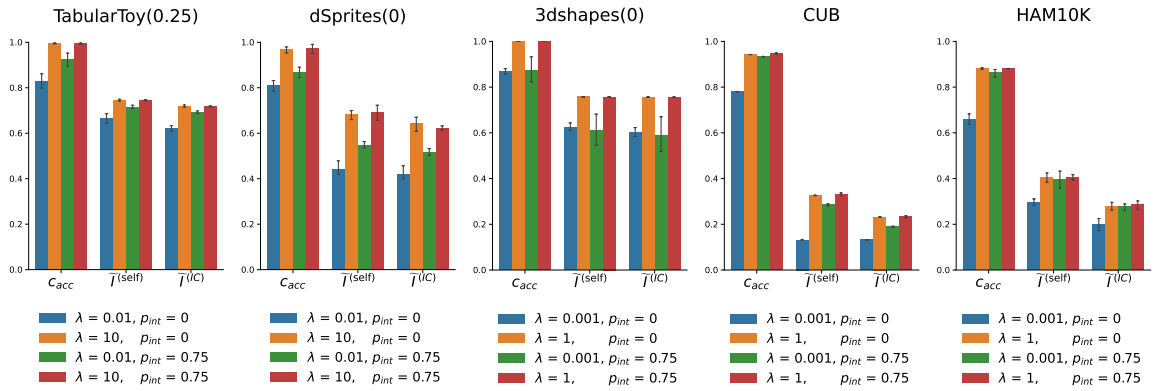

Figure 12: Concept accuracy, $\widetilde{I}^{(IC)}$ and $\widetilde{I}^{(\text{self})}$ scores for CEMs with low and high λ and p_{int}.

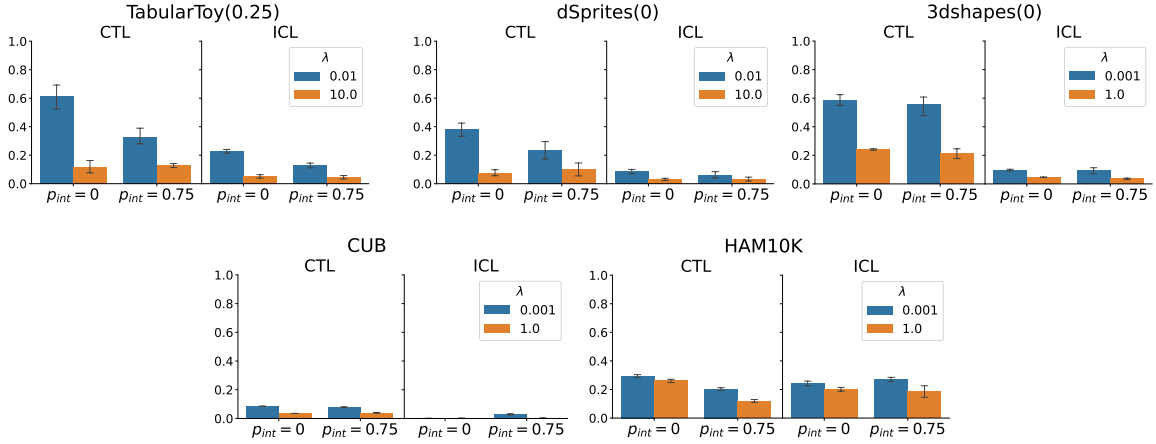

Figure 13: CTL and ICL scores evaluated on the predicted concept probabilities $\hat{c}_i$ in CEMs with different values of λ and p_{int}.

used in CEMs hence provide a mechanism for interconcept leakage to persist even at high concept supervision.

Alignment leakage in CEMs. A subtle leakage effect may in principle arise when exposing a model like CEMs to interventions during training. For each data point $n \in \{1, \ldots, N\}$ with ground-truth value $c_i^{(n)}$ for concept i, we label $\hat{c}_i^{+(n)}$ as aligned if $c_i^{(n)} = 1$, and unaligned otherwise, while we label $\hat{c}_i^{-(n)}$ as aligned if $c_i^{(n)} = 0$, and unaligned otherwise. The meaning of this distinction is that after intervening on the i-th concept at training or test time, the weighted vector $\hat{c}_i^{w(n)}$ is equal to the aligned vector. A model exposed to interventions may thus learn to improve performance upon interventions by generating aligned vectors that are more predictive of the task than the unaligned ones. This form of leakage results in embeddings that have inhomogeneous reliability depending on the alignment with ground-truth concepts, leading to an overall reduction in model interpretability.

This effect, which we label *alignment leakage*, is a specific manifestation of concepts-task leakage. To quantify it, we define the following information-theoretic score, which indicates the extent to which aligned vectors are more predictive of the task than unaligned vectors across both positive and negative embeddings, as an excess in the concepts-task normalised MIs defined in (10). Specifically, we define:

$$\begin{aligned}\widetilde{I}^{(\text{align})}\left(\hat{\boldsymbol{c}}^{+}, \hat{\boldsymbol{c}}^{-}, c, y\right) = \widetilde{I}^{(CT)}\left(\hat{c}^{+(\text{aligned})}, y\right) - \widetilde{I}^{(CT)}\left(\hat{c}^{+(\text{unaligned})}, y\right) \\ + \widetilde{I}^{(CT)}\left(\hat{c}^{-(\text{aligned})}, y\right) - \widetilde{I}^{(CT)}\left(\hat{c}^{-(\text{unaligned})}, y\right),\end{aligned} \tag{13}$$

where $\hat{c}^{\pm(\text{aligned})}$ denote the set of positive/negative vectors aligned with the corresponding ground-truth concept, and analogously for $\hat{c}^{\pm(\text{unaligned})}$. $\widetilde{I}^{(\text{align})}$ takes values in (-2,2) in units of normalised MI. Positive values indicate that the aligned vectors are more predictive of the task than the unaligned ones, while a value of zero corresponds to no alignment leakage.

We assessed this form of leakage in CEMs trained at different values of λ and p_{int}. While in models trained on dSprites(0), 3dshapes(0) and HAM10K it does not appear, we find significantly non-zero scores for models trained on TabularToy(0.25) and CUB (Figure 14). Alignment leakage is stronger in models at high values of λ and p_{int}, supporting the intuition that this effect is associated with exposure to interventions during training.

Mechanisms analogous to alignment leakage may help explain why CEMs trained with $p_{int} > 0$ often perform well under interventions (see Appendix G and Espinosa Zarlenga et al. (2022)). Importantly, this is not in contradiction with the claim that CEM embeddings may fail to be faithful concept representations. Successful interventions indicate that the embeddings encode some information about the corresponding concept state, but not that this is the only, or even the dominant, task-relevant information they encode. In particular, exposure to interventions during training may encourage the model to generate aligned embeddings that are better predictive of the task, so that interventions improve performance partly by selecting more task-informative embeddings rather than only by correcting the concept value. This interpretation is also consistent with the observation that CEMs trained with $p_{int} = 0$ are typically only mildly responsive to interventions (see Figure 6 and A.8 in Espinosa Zarlenga et al. (2022)). Taken together, these findings suggest that the positive and negative embeddings may encode the concept state, but in a way that is weaker, less isolated, or more entangled with task information than in standard CBMs.

9. Lessons for model design

Our framework and results on leakage enable us to outline the key steps for designing and training a concept-based model to ensure its interpretability. The first step consists of assessing whether the annotated concepts are sufficiently predictive of the task. To that aim, one should train several classifier heads on the ground-truth concepts, starting from a linear baseline. Depending on the use-case, while the choice of an interpretable head (such as a linear classifier) may be desirable, it may also result in severe misspecification. In this case, one should expect the gain in functional interpretability to be compensated to a certain extent by higher leakage, as per the results of Section 7.

If a sufficiently high task accuracy is not achieved with arbitrarily complex architectures, this is generally an indication that the set of concepts is not complete for the task. This

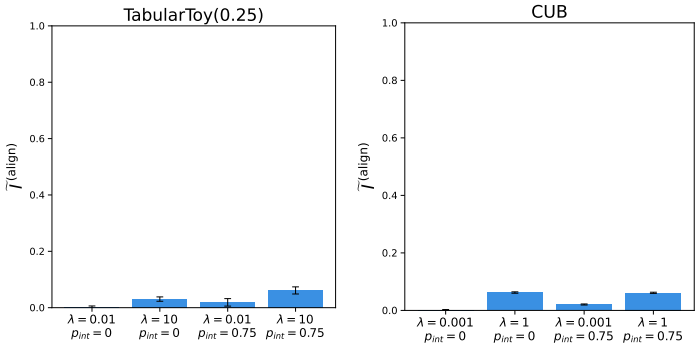

Figure 14: Alignment leakage score for CEMs trained on the TabularToy(0.25) and CUB datasets with low and high values of λ and p_{int}.

either causes a drop in task performance in the case of hard architectures, or introduces a significant amount of leakage in the case of soft architectures. Furthermore, the experiments from Section 7 suggest that higher concept supervision is not sufficient in general to prevent leakage in such cases. If additional annotated concepts are not available, one should thus conclude that the task is not amenable to a concept-based approach given the dataset at hand.

Possible solutions to concept set incompleteness have been studied. These include models that allow leakage into concept representations to encode the additional required information; CEMs and related architectures are one such example (Espinosa Zarlenga et al., 2022). Another approach consists of adding unsupervised concepts or channels, meant to account for the missing information while in principle preserving the interpretability from the annotated concepts (see Mahinpei et al., 2021; Havasi et al., 2022; Sawada and Nakamura, 2022; Yuksekgonul et al., 2023; Sheth and Kahou, 2023; Ismail et al., 2024, 2025). However, when adopting this approach, one has to ensure the model does not over-rely on such unsupervised channels, effectively bypassing the concept bottleneck when making a task prediction, and resulting equivalent to an end-to-end model without concept supervision. Especially for high-risk scenarios, a general framework to interpret unknown concepts and safely include them in the concept refinement process is still missing.

Another approach to dealing with an incomplete (or missing) set of annotated concepts that has recently gained popularity is the use of CBMs with label-free concept representations (Yuksekgonul et al., 2023; Oikarinen et al., 2023; Yang et al., 2023). A common modelling strategy in this line of work is to define task-relevant concepts using a large language model and encode them as concept prototypes with a vision-language model such as CLIP (Radford et al., 2021) The projection of an input image embedding onto these prototypes yields soft concept activations, on top of which a final head is trained sequentially to predict the task label. From the perspective of leakage, these models behave as standard sequential CBMs, which are known to exhibit leakage similar to jointly trained CBMs (Mahinpei et al., 2021). In particular, in these models the structure and geometry of the concept embeddings from the multimodal encoder may carry additional task-relevant information beyond the ground-truth interconcept and concepts-task relationships, which can then leak into the concept activations. Our framework could help clarify these effects

using analyses similar to those applied to CEMs in Section 8, addressing questions such as whether task performance is truly driven by the activation of the selected concepts, or rather by task-relevant information encoded in the concept and input representations (using the analogue of the $\widetilde{I}^{(CT)}$ score defined in (10)). Since little or no concept annotation is available in these setups, it would also be valuable to accurately benchmark leakage using our framework on datasets where annotated concepts are available and explicitly imposed as the bottleneck of an otherwise label-free model.

It is also worth noting that an incomplete concept set and a misspecified classifier head are in principle degenerate issues, both leading to a drop in task performance and similarly higher leakage scores. Quantifying the completeness of a concept set for a given task is by itself an information-theoretic problem that has been extensively discussed in the context of feature selection, and some measures have been presented, for example in Yeh et al. (2020); Havasi et al. (2022), which may provide indications of ways to avoid such degeneracy.

If such issues do not arise, the next design choice concerns the form of concept encoding. The results of Section 7 show that this step is critical: the encoding should not be overly expressive relative to the annotated concepts, otherwise leakage may be encouraged. Logit encoding is over-expressive for binary concepts, while it is arguably a suitable choice for continuous concepts. Similarly, vector encoding may be generally over-expressive for one-dimensional variables.

For binary annotated concepts, two common setups are soft and hard CBMs. To reduce leakage and achieve accurate concept learning, soft CBMs should be trained at relatively high $\lambda \in (\lambda_{min}, \lambda_{max})$, according to the results of Section 7. The CTL and ICL scores enable tuning λ to values that minimise leakage. Throughout our experiments they also indicate that leakage is often unavoidable in soft CBMs. This may be acceptable depending on the use case, particularly since at high λ the model's reliance on leaked information may be secondary to concept activations. Note that when designing a concept-based model, it can also be beneficial to train an end-to-end model without concept supervision, in order to establish an upper bound on the overall task performance, and assess the intrinsic suitability of the input data for the task.

More broadly, when developing concept-based architectures beyond CBMs, our metrics can assess how different design choices promote leakage within the concept representations. Interconcept leakage may be significantly reduced by encouraging the concept encoder to learn the ground-truth interconcept correlation structure, for instance by adopting a probabilistic perspective as in Stochastic (Vandenhirtz et al., 2024; Hoffmann et al., 2025) and Autoregressive CBMs (Mahinpei et al., 2021). However, per se a probabilistic approach does not prevent concepts-task leakage: if the concept encoder $p(c|x)$ is learnt jointly or sequentially with the final head and no hard threshold is imposed on the sampled concepts, $p(c|x)$ and the sampled concepts will generally encode task-relevant information.

In summary we have proposed an information-theoretic framework to assessing the interpretability of concept-based models. The leakage measures we introduce are robust and sensitive, and we propose their use as standard tools for ensuring interpretability when designing concept-based models.

Acknowledgments

We thank Aya Abdelsalam Ismail, Anthony Baptista and Gregory Verghese for insightful discussions on related topics. All authors except CH are supported by the Turing-Roche partnership. EP is funded by the Department of Science Innovation and Technology under PharosAI. CRSB is supported by a King's College London AI+ Fellowship. TC is supported by a principal research fellowship from University College London. CH is an employee of Roche.

Appendix A. Details on the experimental setup

A.1 Datasets

For additional details on these datasets, we refer the reader to Espinosa Zarlenga et al. (2023a).

TabularToy(δ). This dataset is constructed by sampling a 3-dimensional latent variable $z \sim \mathcal{N}(0, \Sigma(\delta))$, with correlation matrix $\Sigma(\delta)_{ij} = \delta_{ij} + \delta(1 - \delta_{ij})$ with $i, j = 1, 2, 3$, where δ_{ij} is the Kronecker symbol and $\delta \in (-1, 1)$ indicating the correlation between components. The 7-dimensional input x is a trigonometric function of the latent variables (see Mahinpei et al., 2021), while the three binary concepts are defined as the sign of the latent variables, $c_i = (z_i > 0)$. The binary task label is the following linear function of concepts,

$$y_{\text{TT}}^{(orig)} = (c_1 + c_2 + c_3 \geq 2)\,. \tag{14}$$

For our experiments we generate a dataset of size 10,000, with a 0.7/0.2/0.1 train / validation / test ratio. **For Figures 2 and 3**, we consider a simplified version of TabularToy(δ) with only the first two concepts c_1 and c_2 for each input, and binary task $y = (c_1 + c_2 \geq 1)$.

dSprites(γ). For this family of datasets, each input $x \in \{0, 1\}^{64\times64\times1}$ is an image generated deterministically from five latent variables $z = (\text{shape} \in \{0, 1, 2\}$, scale $\in \{0, \ldots, 5\}$, $\theta \in \{0, \ldots, 39\}$, $X \in \{0, \ldots, 31\}$, $Y \in \{0, \ldots, 31\})$, where shape corresponds to either a square, ellipse or a heart, while θ, X and Y indicate the rotation and the position of the shape in the plane respectively. Following Espinosa Zarlenga et al. (2023a), we define dSprites(0) as the datasets with no interconcept correlations. Here the five binary concepts are based on the value of the latent variables, $c = (z_1 < 2,\ z_2 < 3,\ z_3 < 20,\ z_4 < 16,\ z_5 < 16)$, and the 8-class task label is

$$y_{\text{dS}}^{(orig)} = (1 - c_1)(2c_4 + c_5 + 4) + c_1(2c_2 + c_3)\,. \tag{15}$$

To generate dSprites(γ) with increasing interconcept correlations for $\gamma = 1, \ldots, 4$, we adopt the procedure detailed in Espinosa Zarlenga et al. (2023a). We use datasets with approximatively 25K and 14K samples for dSprites(0) and dSprites(4) respectively, with a 0.7/0.1/0.2 train / validation / test ratio.

3dshapes(γ). Each input is an image $x \in \{0, 1\}^{64\times64\times3}$ generated from six latent variables $z = (\text{floor\_hue} \in \{0, \ldots, 9\}$, wall_hue $\in \{0, \ldots, 9\}$, object_hue $\in \{0, \ldots, 9\}$, scale $\in \{0, \ldots, 7\}$, shape $\in \{0, 1, 2, 3\}$, orientation $\in \{0, \ldots, 14\})$, where shape denotes either a sphere, a cube, a capsule or a cylinder. Following Espinosa Zarlenga et al. (2023a), in

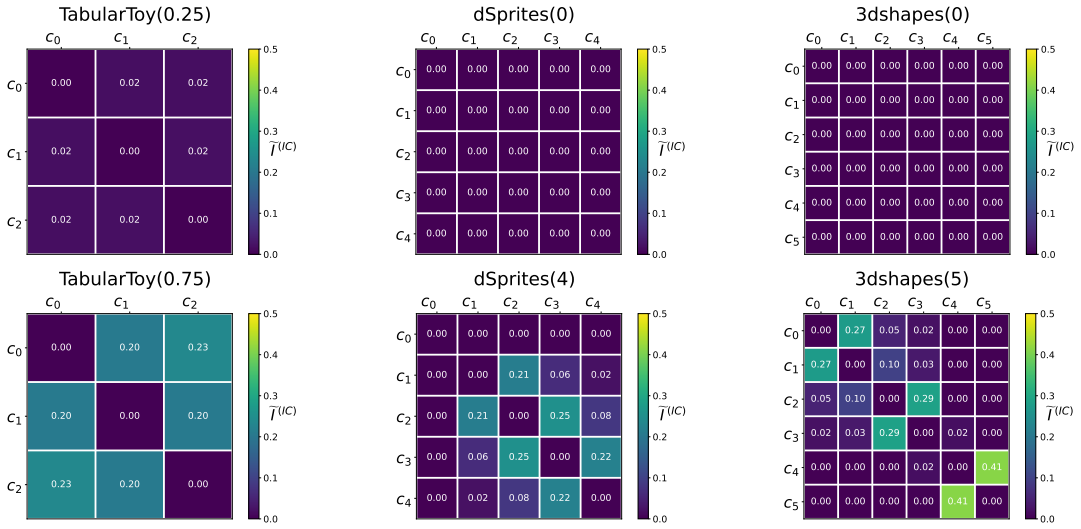

Figure 15: Ground-truth interconcept normalised MI in the synthetic datasets.

3dshapes(0) concepts are defined as $c = (z_1 < 5,\ z_2 < 5,\ z_3 < 5,\ z_4 < 4,\ z_5 < 2,\ z_6 < 7)$, and the 12-class labels are obtained as

$$y_{\text{3ds}}^{(orig)} = (1 - c_5)(2c_1 + c_2) + c_5(4c_3 + 2c_4 + c_6 + 4) \tag{16}$$

Increasing interconcept correlations are induced for $\gamma = 1, \ldots, 5$ in a similar fashion to dSprites(γ). For our experiments we consider 3dshapes(0) and 3dshapes(5) with size 16K and 14K respectively, and a 0.7/0.1/0.2 train / validation / test ratio.

CUB. This dataset consists of 11,788 images of birds labelled with 112 binary phenotype attributes, with the objective of predicting the correct species among 200 classes (Wah et al., 2025). It provides rich fine-grained concept annotations that collectively offer a near-complete description of the target classification task. We preprocessed it following Koh et al. (2020), and consider a balanced 0.4/0.1/0.5 train / validation / test ratio.

HAM10K. In this dataset the objective is to determine whether a skin lesion is benign or malignant, using dermatoscopic images annotated with 18 morphological attributes. We use the annotations from Chanda et al. (2024) and the preprocessing from Patrício et al. (2025), yielding 6498 images in total, split with a 0.8/0.1/0.1 train / validation / test ratio.

Ground-truth interconcept correlations. In Figure 15 we display as a reference the ground-truth interconcept normalised MI for the synthetic training datasets used in the experiments.

Experiments with an incomplete concept set. In TabularToy(0.25), we removed c_3 from the three concepts. In dSprites(0), we removed c_4 and c_5 from the five concepts. In 3dshapes(0), we removed c_3 and c_6 from the six concepts. For CUB we use its concept-incomplete version defined in Espinosa Zarlenga et al. (2025) where only 22 out of the 112 annotated concepts are kept. In HAM10K we remove 8 out of the 18 concepts, with the resulting concept set corresponding to the annotated features `ESA, PRL, PES, PIF, OPC, SPC, MVP, PRLC, PLF, PDES`.

Experiments with a misspecified head. We provide here more details on the experiments to assess the effects of a misspecified final head discussed in Section 7. To generate non-linear concepts-task dependences in TabularToy(δ), we added a non-linear term to the original linear task in (14), resulting in the binary classification problem,

$$y_{\text{TT}}^{(new)} = (c_1 + c_2 + c_3 - c_1 c_2 \geq 2)\,.$$

The original tasks in (15) and (16) for dSprites(γ) and 3dshapes(γ) are non-linear, however a linear classifier is able to reach essentially perfect task accuracy (Table 3). To induce a sizeable misspecification, we thus added further higher-order non-linear terms. The tasks implemented for our experiments on dSprites(0) and 3dshapes(0) are:

$$\begin{aligned} y_{\text{dS}}^{(new)} &= y_{\text{dS}}^{(orig)} - (1 - c_1)c_3 c_4 - c_1(c_2 c_5 + c_3 c_5 + c_2 c_4)\,, \\ y_{\text{3ds}}^{(new)} &= y_{\text{3ds}}^{(orig)} - 3c_1 c_2 c_3 - c_4 c_5 c_6 - c_1 c_3 c_5 + c_2 c_4 c_6\,. \end{aligned}$$

We cannot run a similar experiment on CUB and HAM10K, since a linear final head already achieves near-perfect task performance.

A.2 Model architectures and training

On TabularToy(δ) we used as concept encoder a 4-layer leaky-ReLU MLP with activations $\{7, 64, 64, 3\}$, and a linear classifier head. On dSprites(γ), 3dshapes(γ) and CUB we used a ResNet-18 encoder, while for HAM10K a ViT-B/16 encoder pretrained on ImageNet. For these datasets we used a 4-layer ReLU MLP with activations $\{k, 64, 64, \ell\}$ as a final head. We trained CEMs with 16-dimensional embeddings.

We trained all models for 200 epochs, using the Adam optimiser with momentum 0.9 and learning rate 10^{-3} for TabularToy(δ), dSprites(γ) and 3dshapes(γ); 10^{-2} for CUB; 10^{-4} for HAM10K. For hard models, we trained the encoder and the classifier head for 200 and 20 epochs respectively. We trained with batch sizes 512 for TabularToy(δ); 32 for dSprites(γ), 3dshapes(γ) and HAM10K; 64 for CUB.

A.3 Score evaluations

Except for Figure 15, all the displayed scores are evaluated on test sets. The reported intervention performances were obtained using a random policy: for each data point, the next concept to intervene on was selected uniformly at random from the set of concepts that had not yet been intervened on. In Figures 3, 4 and 17 we show results for single models, and the displayed means and 95% confidence intervals refer to a 5-fold evaluation of each score, including intervention performance. In the rest of the paper, we repeated the evaluation of each score 5 times for individual models, and we display means and 95% confidence intervals representing the distribution of the mean upon 5-fold training for each model class. MIs and entropies were estimated via the KSG estimator (Kraskov et al., 2004) with 3 nearest neighbours. The OIS and NIS were evaluated using their original implementations and default settings from Espinosa Zarlenga et al. (2023a).

Appendix B. Undesired features of the Niche Impurity Score

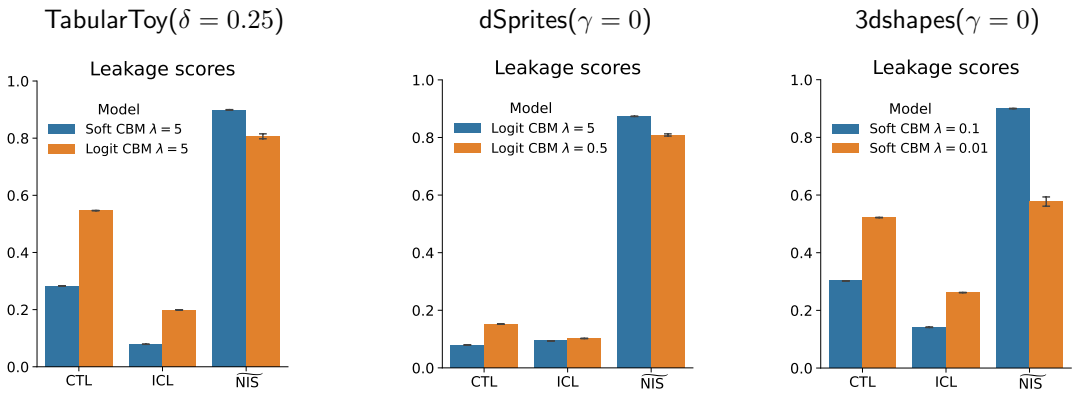

Figure 17: $\widetilde{\text{NIS}}$ and leakage scores for the pairs of models assessed in Figure 4 and Table 1.

Because it is computed as an AUC, NIS $= 1/2$ indicates that concept representation subsets are not predictive of the value of concepts, while NIS $= 1$ indicates maximal higher-order interconcept leakage. For a fair comparison of our scores with NIS, we plotted $\widetilde{\text{NIS}} = 2(\text{NIS} - 1/2)$, so that scores of 0 and 1 correspond to no and maximal leakage respectively.

Evaluating NIS on the pairs of models considered in Figure 4, this score appears to be anticorrelated with intervention performance and leakage (Figure 17). Furthermore, it takes proportionally high values regardless of the model and dataset. This is confirmed by its behaviour on hard CBMs (Figure 16): the NIS is non-vanishing and high in all cases. These undesirable features indicate that NIS is not a suitable measure of leakage.

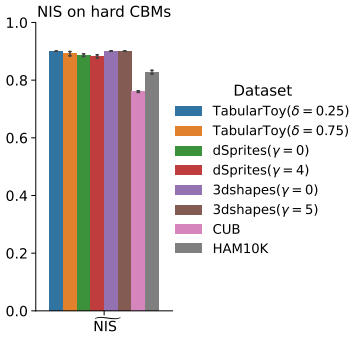

Figure 16: $\widetilde{\text{NIS}}$ evaluated on hard CBMs across datasets.

Appendix C. Concept-wise leakage scores

The concept-wise leakage scores defined in (6) and (9) provide more granular information on the leakage encoded in each concept. In Figure 18 we display the CTL_i and ICL_i scores for the pairs of models analysed in Figure 4 and Table 1. Note in particular that (i) leakage is not necessarily learnt homogeneously across concepts, and (ii) comparing two models with different levels of interpretability, certain concepts may exhibit comparable amounts of leakage (such as c_4 in the dSprites(0) example), while for others leakage may be significantly different. Concept-wise scores are sensitive to both phenomena, making them valuable indicators of per-concept leakage risk upon intervention and therefore useful measures for model design.

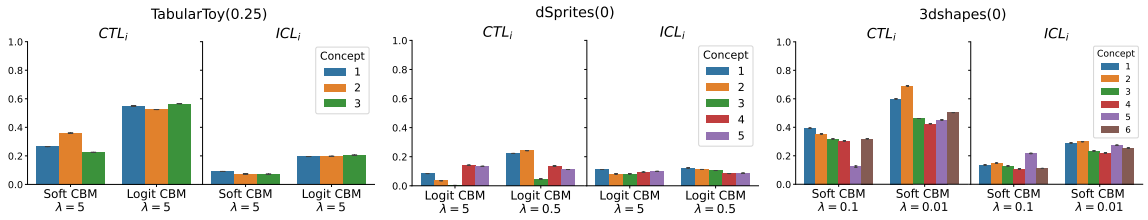

Figure 18: Concept-wise leakage scores evaluated on the pairs of models in Figure 4.

Appendix D. Details of the correlation between leakage scores and intervention performance

The models considered for this computation were

- for the TabularToy(0.25), TabularToy(0.75), dSprites(0), dSprites(4) datasets: soft and logit CBMs with $\lambda = 0.01, 0.1, 0.5, 1, 5, 10$;
- for the 3dshapes(0), 3dshapes(5) datasets: soft and logit CBMs with $\lambda = 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1$.
- for the CUB, HAM10K datasets: soft and logit CBMs with $\lambda = 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10$.

Each class of models was trained over 5 folds, amounting to a total of 60 evaluated models for each of the first set of four datasets, and to a total of 70 models for each of the second set of two datasets. We use Rubin's rule (see Rubin, 2004; Barnard and Rubin, 1999) to combine within-imputation and between-imputation variances and obtain the best-estimate Pearson r and p-value.

Appendix E. Behaviour of CBMs at high λ

In soft and logit CBMs leakage is minimal for values of $\lambda > \lambda_{min}$, where λ_{min} is specific to the considered model class (Figure 19). As λ is increased leakage scores remain essentially constant to start. However, at higher values of $\lambda \gtrsim \lambda_{max}$ concept supervision may become too strong, at the expense of task learning. In such cases,

1. task performance may experience a drop. For instance, models trained on TabularToy(0.25) reach a task accuracy above 90% in only 3 out of 5 folds for each $\lambda = 10, 20, 50$;
2. leakage may raise again, to well above the minimal value. The soft models trained on 3dshapes(0) in Figure 19 are a compelling example of this.

Finally, note that the minimal attainable leakage in logit CBMs is higher than in soft CBMs for any given dataset and choice of λ, in line with the expectations.

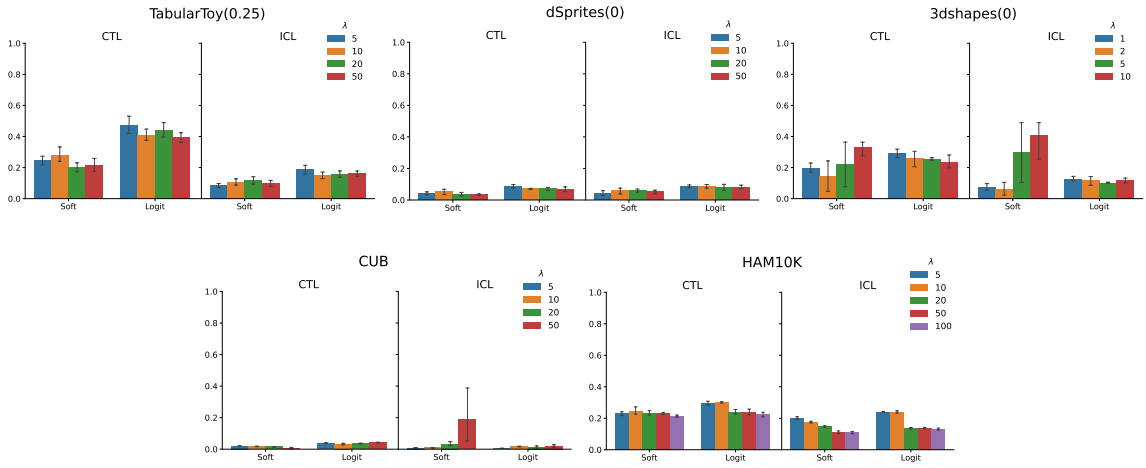

Figure 19: Leakage scores evaluated on soft and logit CBMs at high λ.

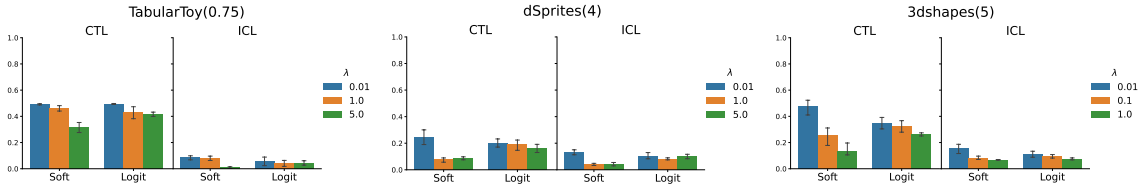

Figure 20: Leakage scores evaluated for soft and logit models with different levels of concept supervision across datasets with high ground-truth interconcept MIs.

Appendix F. Performance and scores of models with over-expressive concept representations, incomplete concept sets and misspecified heads

In Figure 20 we present the leakage scores across datasets with high ground-truth interconcept MI. In Figure 21 we show the task and concept accuracies of the soft CBMs trained on the complete and incomplete sets of concepts discussed in Section 7. Note in particular the decrease in task accuracy at high values of λ that occurs in the case of incomplete concept sets. At low λ, leakage is sufficient to ensure task accuracy remains essentially constant and approximately equivalent to that of the case of complete concept sets. Figure 22 shows the task and concept accuracies of the well-specified and misspecified soft CBMs evaluated in Figure 8 and discussed in Section 7.

Appendix G. Further results on CEMs

Embedding structure at non-vanishing p_{int}. In Figures 23 and 24 we display the PCA projections of the weighted vectors $\hat{\boldsymbol{c}}_1^{\boldsymbol{w}}$ with colouring based on the ground-truth value of the task label and of concepts 2 and 3 respectively, at non-vanishing values of p_{int} and for low and high values of λ.

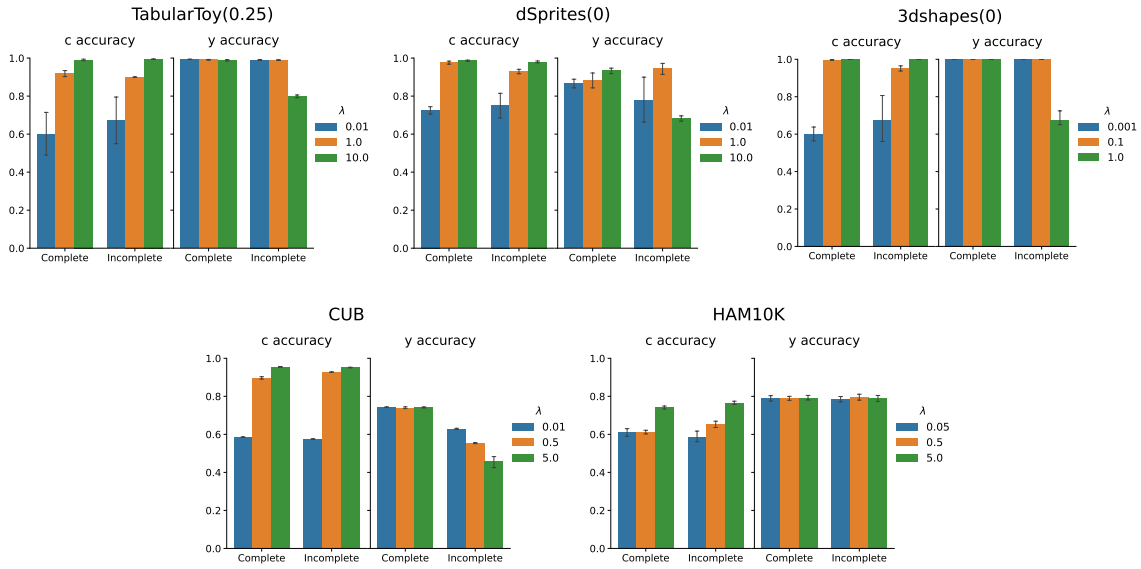

Figure 21: Concepts and task accuracies for the models analysed in Figure 7 on datasets with complete and incomplete sets of concepts.

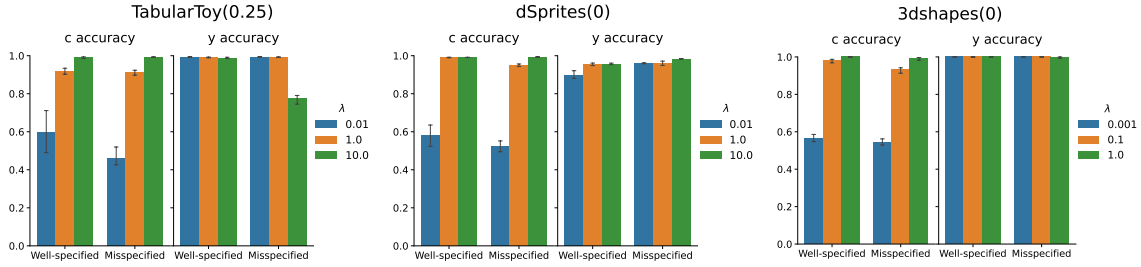

Figure 22: Concepts and task accuracies evaluated for models with a linear head analysed in Figure 8 trained on datasets with linear and non-linear tasks as functions of concepts.

Concept accuracy and intervention performance are not measures of leakage in CEMs. As demonstrated in Section 8, CEMs generally encode high amounts of concepts-task leakage, and from this perspective, concept accuracy is only a measure of how predictive the vectors $\hat{\boldsymbol{c}}_i^{\boldsymbol{w}}$ are of concept i, regardless of the leaked information about the task in the weighted embeddings. Furthermore, as shown in Figure 12, interconcept leakage increases with concept supervision. Thus c_{acc} typically correlates with interconcept leakage in CEMs.

Intervention performance is not an indicator of leakage either; indeed it can be high in models with severe leakage (see Figure 25). In particular, CEMs with high concepts-task leakage in both $\hat{\boldsymbol{c}}_i^+$ and $\hat{\boldsymbol{c}}_i^-$, or where alignment leakage is present are able to achieve a vanishing S_{int}. This motivates the definition and adoption of more sensitive information-theoretic measures for CEMs, such as (10), (11) and (13).

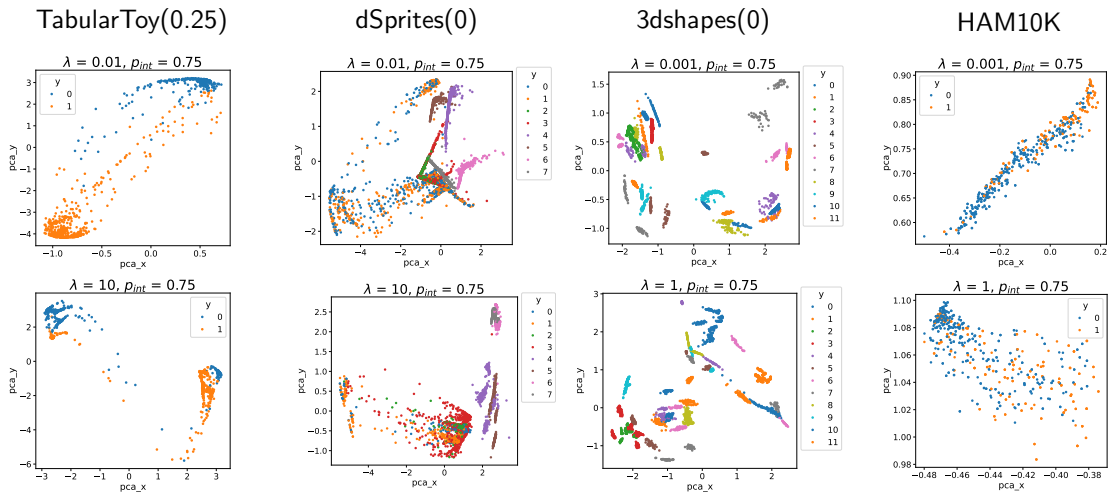

Figure 23: 2-dimensional PCA projections of the weighted embeddings $\hat{\boldsymbol{c}}_1^{\boldsymbol{w}}$ for the first concept across datasets and for different values of λ at high $p_{int} = 0.75$. The colouring indicates the value of the ground-truth task label.

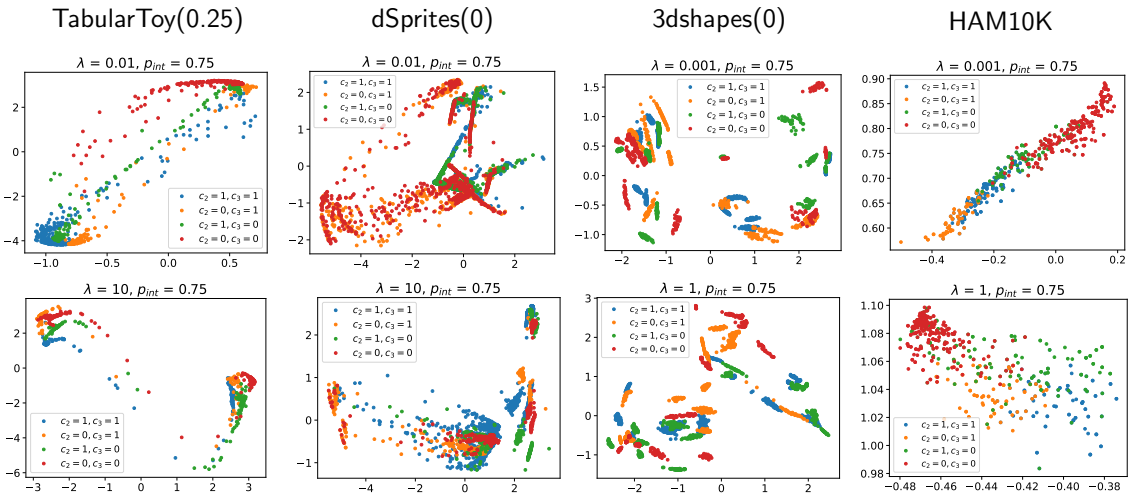

Figure 24: 2-dimensional PCA projections of the weighted embeddings $\hat{\boldsymbol{c}}_1^{\boldsymbol{w}}$ for the first concept across datasets and for different values of λ at high $p_{int} = 0.75$. The colouring indicates the ground-truth value of concepts 2 and 3.

References

Christopher RS Banerji, Tapabrata Chakraborti, Aya Abdelsalam Ismail, Florian Ostmann, and Ben D MacArthur. Train clinical ai to reason like a team of doctors. *Nature*, 639 (8053):32–34, 2025.

Hubert Baniecki and Przemyslaw Biecek. Birds look like cars: adversarial analysis of intrinsically interpretable deep learning. *Machine Learning*, 114(12):284, 2025.

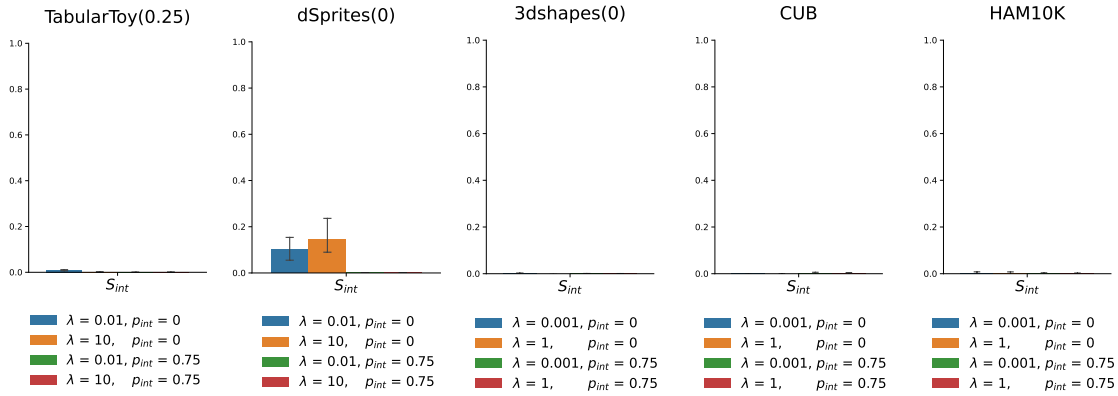

Figure 25: The S_{int} score for CEMs with low and high values of λ and p_{int}.

ISSN 1573-0565. doi: 10.1007/s10994-025-06896-w. URL https://doi.org/10.1007/s10994-025-06896-w.

John Barnard and Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999. ISSN 00063444, 14643510.

Andrea Bontempelli, Stefano Teso, Katya Tentori, Fausto Giunchiglia, and Andrea Passerini. Concept-level debugging of part-prototype networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=oiwXWPDTyNk.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10319–10328, 2022. doi: 10.1109/CVPR52688.2022.01008.

Nicholas Carrara and Jesse Ernst. On the estimation of mutual information. *Proceedings*, 33(1), 2019. ISSN 2504-3900. doi: 10.3390/proceedings2019033031. URL https://www.mdpi.com/2504-3900/33/1/31.

Tirtha Chanda, Katja Hauser, Sarah Hobelsberger, Tabea-Clara Bucher, Carina Nogueira Garcia, Christoph Wies, Harald Kittler, Philipp Tschandl, Cristian Navarrete-Dechent, Sebastian Podlipnik, Emmanouil Chousakos, Iva Crnaric, Jovana Majstorovic, Linda Alhajwan, Tanya Foreman, Sandra Peternel, Sergei Sarap, İrem Özdemir, Raymond L. Barnhill, Mar Llamas-Velasco, Gabriela Poch, Sören Korsing, Wiebke Sondermann, Frank Friedrich Gellrich, Markus V. Heppt, Michael Erdmann, Sebastian Haferkamp, Konstantin Drexler, Matthias Goebeler, Bastian Schilling, Jochen S. Utikal, Kamran Ghoreschi, Stefan Fröhling, Eva Krieghoff-Henning, Alexander Salava, Alexander Thiem, Dimitrios Alexandris, Amr Mohammad Ammar, Ana Sanader Vučemilović, Andrea Miyuki Yoshimura, Andzelka Ilieva, Anja Gesierich, Antonia Reimer-Taschenbrecker, Antonios G. A. Kolios, Arturs Kalva, Arzu Ferhatosmanoğlu, Aude Beyens, Claudia

Pföhler, Dilara Ilhan Erdil, Dobrila Jovanovic, Emoke Racz, Falk G. Bechara, Federico Vaccaro, Florentia Dimitriou, Gunel Rasulova, Hulya Cenk, Irem Yanatma, Isabel Kolm, Isabelle Hoorens, Iskra Petrovska Sheshova, Ivana Jocic, Jana Knuever, Janik Fleißner, Janis Raphael Thamm, Johan Dahlberg, Juan José Lluch-Galcerá, Juan Sebastián Andreani Figueroa, Julia Holzgruber, Julia Welzel, Katerina Damevska, Kristine Elisabeth Mayer, Lara Valeska Maul, Laura Garzona-Navas, Laura Isabell Bley, Laurenz Schmitt, Lena Reipen, Lidia Shafik, Lidija Petrovska, Linda Golle, Luise Jopen, Magda Gogilidze, Maria Rosa Burg, Martha Alejandra Morales-Sánchez, Martyna Sławińska, Miriam Mengoni, Miroslav Dragolov, Nicolás Iglesias-Pena, Nina Booken, Nkechi Anne Enechukwu, Oana-Diana Persa, Olumayowa Abimbola Oninla, Panagiota Theofilogiannakou, Paula Kage, Roque Rafael Oliveira Neto, Rosario Peralta, Rym Afiouni, Sandra Schuh, Saskia Schnabl-Scheu, Seçil Vural, Sharon Hudson, Sonia Rodriguez Saa, Sören Hartmann, Stefana Damevska, Stefanie Finck, Stephan Alexander Braun, Tim Hartmann, Tobias Welponer, Tomica Sotirovski, Vanda Bondare-Ansberga, Verena Ahlgrimm-Siess, Verena Gerlinde Frings, Viktor Simeonovski, Zorica Zafirovik, Julia-Tatjana Maul, Saskia Lehr, Marion Wobser, Dirk Debus, Hassan Riad, Manuel P. Pereira, Zsuzsanna Lengyel, Alise Balcere, Amalia Tsakiri, Ralph P. Braun, Titus J. Brinker, and Reader Study Consortium. Dermatologist-like explainable ai enhances trust and confidence in diagnosing melanoma. *Nature Communications*, 15(1):524, 2024. ISSN 2041-1723. doi: 10.1038/s41467-023-43095-4. URL `https://doi.org/10.1038/s41467-023-43095-4`.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i5.25736. URL `https://doi.org/10.1609/aaai.v37i5.25736`.

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. *This looks like that: deep learning for interpretable image recognition*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Paweł Czyż, Frederic Grabowski, Julia E Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=25vRtG56YH`.

Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Counterfactual concept bottleneck models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=w7pMjyjsKN`.

Eric Enouen and Sainyam Galhotra. Debugging concept bottlenecks through intervention: Shortcut removal + retraining. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025. URL `https://openreview.net/forum?id=KuHArH2mMQ`.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: beyond the accuracy-explainability trade-off. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Mateo Espinosa Zarlenga, Pietro Barbiero, Zohreh Shams, Dmitry Kazhdan, Umang Bhatt, Adrian Weller, and Mateja Jamnik. Towards robust metrics for concept representation evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10): 11791–11799, June 2023a. ISSN 2159-5399. doi: 10.1609/aaai.v37i10.26392. URL http://dx.doi.org/10.1609/aaai.v37i10.26392.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 37849–37875. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/770cabd044c4eacb6dc5924d9a686dce-Paper-Conference.pdf.

Mateo Espinosa Zarlenga, Gabriele Dominici, Pietro Barbiero, Zohreh Shams, and Mateja Jamnik. Avoiding leakage poisoning: Concept interventions under distribution shifts. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=7mxDGiF01U.

Yibo Gao, Zheyao Gao, Xin Gao, Yuanye Liu, Bomin Wang, and Xiahai Zhuang. Evidential concept embedding models: Towards reliable concept explanations for skin disease diagnosis. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 308–317, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72117-5.

Gokul Gowri, Xiaokang Lun, Allon M Klein, and Peng Yin. Approximating mutual information of high-dimensional variables using learned representations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HN05DQxyLl.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23386–23397. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/944ecf65a46feb578a43abfd5cddd960-Paper-Conference.pdf.

Lena Heidemann, Maureen Monnet, and Karsten Roscher. Concept correlation and its effects on concept-based models. In *2023 IEEE/CVF Winter Conference on Applications*

of Computer Vision (WACV), pages 4769–4777, 2023. doi: 10.1109/WACV56688.2023.00476.

Wiktor Jan Hoffmann, Sonia Laguna, Moritz Vandenhirtz, Emanuele Palumbo, and Julia E. Vogt. Post-hoc stochastic concept bottleneck models. *ArXiv*, abs/2510.08219, 2025. URL https://api.semanticscholar.org/CorpusID:281952428.

Lijie Hu, Tianhao Huang, Huanyi Xie, Chenyang Ren, Zhengyu Hu, Lu Yu, and Di Wang. Semi-supervised concept bottleneck models. *ArXiv*, abs/2406.18992, 2024. URL https://api.semanticscholar.org/CorpusID:270765040.

Qihan Huang, Jie Song, Jingwen Hu, Haofei Zhang, Yong Wang, and Mingli Song. On the concept trustworthiness in concept bottleneck models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30109. URL https://doi.org/10.1609/aaai.v38i19.30109.

Aya Abdelsalam Ismail, Julius Adebayo, Hector Corrada Bravo, Stephen Ra, and Kyunghyun Cho. Concept bottleneck generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=L9U5MJJleF.

Aya Abdelsalam Ismail, Tuomas Oikarinen, Amy Wang, Julius Adebayo, Samuel Don Stanton, Hector Corrada Bravo, Kyunghyun Cho, and Nathan C. Frey. Concept bottleneck language models for protein design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Yt9CFhOOFe.

Dmitry Kazhdan, B. Dimanov, Helena Andrés-Terré, Mateja Jamnik, Pietro Lio', and Adrian Weller. Is disentanglement all you need? comparing concept-based & disentanglement approaches. *CoRR*, abs/2104.06917, 2021.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Sangwon Kim, Dasom Ahn, Byoung Chul Ko, In-su Jang, and Kwang-Ju Kim. Eq-cbm: A probabilistic concept bottleneck with energy-based models and quantized vectors. In Minsu Cho, Ivan Laptev, Du Tran, Angela Yao, and Hongbin Zha, editors, *Computer Vision – ACCV 2024*, pages 270–286, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-96-0963-5.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372, 2009. doi: 10.1109/ICCV.2009.5459250.

Songning Lai, Jiayu Yang, Yu Huang, Lijie Hu, Tianlang Xue, Zhangyi Hu, Jiaxu Li, Haicheng Liao, and Yutao Yue. Cat: Concept-level backdoor attacks for concept bottleneck models, 2025. URL https://arxiv.org/abs/2410.04823.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. doi: 10.1109/CVPR.2009.5206594.

Joshua Lockhart, Nicolas Marchesotti, Daniele Magazzeni, and Manuela Veloso. Towards learning to explain with concept bottleneck models: mitigating information leakage. *ArXiv*, abs/2211.03656, 2022. URL https://api.semanticscholar.org/CorpusID:253384003.

Warren M. Lord, Jie Sun, and Erik M. Bollt. Geometric k-nearest neighbor estimation of entropy and mutual information. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3):033114, 03 2018. ISSN 1054-1500. doi: 10.1063/1.5011683. URL https://doi.org/10.1063/1.5011683.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems (NeurIPS)*, pages 4765–4774, 2017.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *CoRR*, abs/2106.13314, 2021. URL https://api.semanticscholar.org/CorpusID:235652059.

Mikael Makonnen, Moritz Vandenhirtz, Sonia Laguna, and Julia E Vogt. Measuring leakage in concept-based methods: An information theoretic approach, 2025. URL https://arxiv.org/abs/2504.09459.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: interpretable, leak-proof concept-based models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended?, 2021. URL https://arxiv.org/abs/2105.04289.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Cristiano Patrício, Luís F. Teixeira, and João C. Neves. A two-step concept-based approach for enhanced interpretability and trust in skin lesion diagnosis. *Computational and Structural Biotechnology Journal*, 28:71–79, 2025. ISSN 2001-0370. doi: 10.1016/j.csbj.2025.02.013. URL https://doi.org/10.1016/j.csbj.2025.02.013.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Angelos Ragkousis and Sonali Parbhoo. Tree-based leakage inspection and control in concept bottleneck models, 2024. URL https://arxiv.org/abs/2410.06352.

Naveen Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. Do concept bottleneck models obey locality? In *XAI in Action: Past, Present, and Future Applications*, 2023. URL https://openreview.net/forum?id=F6RPYDUIZr.

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth C. Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10932–10941, 2022. URL https://api.semanticscholar.org/CorpusID:258676658.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pages 1135–1144, 2016.

Muhammad Ridzuan, Mai A. Shaaban, Numan Saeed, Ikboljon Sobirov, and Mohammad Yaqub. Hulp: Human-in-the-loop for prognosis. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 328–338, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72086-4.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley, 2004. ISBN 9780471655749.

Francesco De Santis, Gabriele Ciravegna, Philippe Bich, Danilo Giordano, and Tania Cerquitelli. V-cem: Bridging performance and intervenability in concept-based models. *ArXiv*, abs/2504.03978, 2025. URL https://api.semanticscholar.org/CorpusID:277621728.

Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:41758–41765, 2022. URL https://api.semanticscholar.org/CorpusID:246485723.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 618–626, 2017.

Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022. URL https://openreview.net/forum?id=PUspzfGsgY.

Sanchit Sinha and Aidong Zhang. A comprehensive survey on the risks and limitations of concept-based models, 2025. URL https://arxiv.org/abs/2506.04237.

Sanchit Sinha, Mengdi Huai, Jianhui Sun, and Aidong Zhang. Understanding and enhancing robustness of concept-based models. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i12.26765. URL https://doi.org/10.1609/aaai.v37i12.26765.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Jm2aK3sDJD.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL https://doi.org/10.1038/sdata.2018.161.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia E Vogt. Stochastic concept bottleneck models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=iSjqTQ5S1f.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-ucsd birds-200-2011, 2025. URL https://dx.doi.org/10.21227/fg6w-vh29.

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I1quoTXZzc.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pages 19187–19197, 2023. doi: 10.1109/CVPR52729.2023.01839.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2921–2929, 2016.