Stochastic Concept Bottleneck Models

Moritz Vandenhirtz [*1] **Sonia Laguna** [*1] **Ričards Marcinkevičs** [1] **Julia E. Vogt** [1]

Abstract

Concept Bottleneck Models (CBMs) have emerged as a promising interpretable method whose final prediction is based on intermediate, human-understandable concepts rather than the raw input. Through time-consuming manual interventions, a user can correct wrongly predicted concept values to enhance the model's downstream performance. We propose *Stochastic Concept Bottleneck Models* (SCBMs), a novel approach that models concept dependencies. In SCBMs, a single-concept intervention affects all correlated concepts. Leveraging the parameterization, we derive an effective intervention strategy based on the confidence region. We show empirically on synthetic tabular and natural image datasets that our approach improves intervention effectiveness significantly. Notably, we showcase the versatility and usability of SCBMs by examining a setting with CLIP-inferred concepts, alleviating the need for manual concept annotations.

1. Introduction

In today's world, machine learning plays a crucial role in making important decisions, from healthcare to finance and law. Recent studies have focused on Concept Bottleneck Models (CBMs) (Koh et al., 2020; Havasi et al., 2022; Shin et al., 2023), a class of models that predict human-understandable concepts upon which the final target prediction is based. If a user disagrees with a concept prediction, they can intervene by adjusting it to the right value, which in turn affects the target prediction. For example, consider the yellow warbler in Figure 1 (a), where a user might notice that the binary concept 'yellow primary color' is mispredicted. Upon this realization, they can intervene on the CBM by setting its value to 1, which increases the probability of the class yellow warbler. In this work, we propose to extend the concept predictions with the modeling of their dependencies, such that interventions also affect correlated concepts, as depicted in Figure 1 (a,c).

The proposed approach captures the concept dependencies by modeling the concept logits with a learnable non-diagonal normal distribution, which enables efficient, scalable computing of the effect of interventions on other concepts. By integrating concept correlations, we reduce the time and effort of having to laboriously intervene on many correlated variables and increase the efficacy of interventions on the downstream prediction. Lastly, based on the distributional concept parameterization, we propose a novel approach for computing dependency-aware interventions through the likelihood-based confidence region.

Contributions This work contributes to the line of research on concept bottleneck models in several ways. (*i*) We propose to capture and model concept dependencies with a multivariate normal distribution. (*ii*) We derive a novel intervention strategy based on the confidence region of the normal distribution that incorporates concept correlations. Using the learned concept dependencies during the intervention procedure allows for stronger interventional effectiveness. (*iii*) We provide a thorough empirical assessment of the proposed method on synthetic tabular and natural image data. Additionally, we combine our method with concept discovery where we alleviate the need for annotations by using CLIP-inferred concepts. In particular, we show the proposed method (a) discovers meaningful, interpretable patterns in the form of concept dependencies, (b) allows for fast, scalable inference, and (c) outperforms related work with respect to intervention effectiveness thanks to the proposed concept modeling and intervention strategy.

2. Background

Concept bottleneck models (Koh et al., 2020; Lampert et al., 2009; Kumar et al., 2009) are typically trained on data points $(\boldsymbol{x}, \boldsymbol{c}, y)$, comprising the covariates $\boldsymbol{x} \in \mathcal{X}$, target $y \in \mathcal{Y}$, and C annotated binary concepts $\boldsymbol{c} \in \mathcal{C}$. Consider a neural network $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ and a slice $\langle g_{\boldsymbol{\psi}}, h_{\boldsymbol{\phi}} \rangle$ (Leino et al., 2018) s.t. $\hat{y} := f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\psi}}(h_{\boldsymbol{\phi}}(\boldsymbol{x}))$. CBMs enforce a concept bottleneck $\hat{\boldsymbol{c}} := h_{\boldsymbol{\phi}}(\boldsymbol{x})$ such that the model's final output depends on the covariates $\boldsymbol{x}$ solely through the predicted concepts $\hat{\boldsymbol{c}}$.

[*]Equal contribution [1]Department of Computer Science, ETH Zürich, Switzerland. Correspondence to: Moritz Vandenhirtz <moritz.vandenhirtz@inf.ethz.ch>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment.

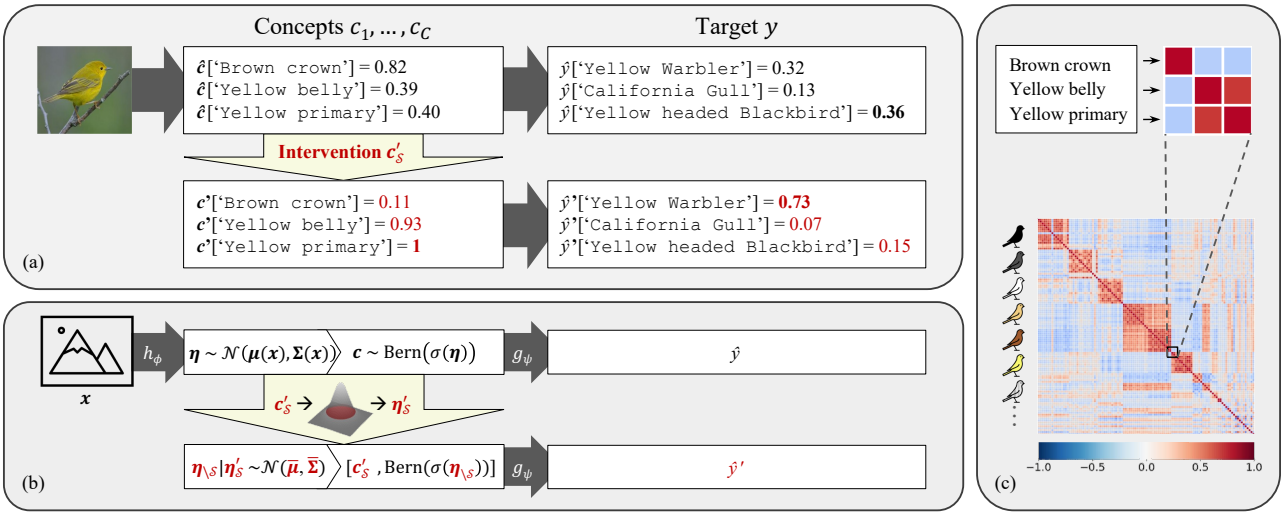


Figure 1. Overview of the proposed method for the CUB dataset (Wah et al., 2011). (a) A user intervenes on the concept of 'primary color: yellow'. Unlike CBMs, our method then uses this information to adjust the predicted probability of correlated concepts, thereby affecting the target prediction. (b) Schematic overview of the intervention procedure. A user's intervention $\boldsymbol{c}'_S$ is used to infer the logits $\boldsymbol{\eta}_{\setminus S}$ of the remaining concepts. (c) Visualization of the learned global dependency structure as a correlation matrix for the 112 concepts of CUB. Characterization of concepts on the left. The anonymized code is available here: https://anonymous.4open.science/r/scbm-A1AA/.

3. Methods

We propose Stochastic Concept Bottleneck Models (SCBM), a novel concept-based method that relaxes the implicit CBM assumption of independent concepts. SCBM captures the concept dependencies by learning their multivariate distribution. As a result, interventions become more effective and scalable, as a single intervention can influence multiple correlated concepts. A schematic overview of the proposed method is depicted in Figure 1 (b).

3.1. Model Formulation

To capture the concept dependencies, we model the concept logits $\boldsymbol{\eta}$ with a learned multivariate normal distribution. A neural network is trained to predict the distribution's parameters $\boldsymbol{\eta} \mid \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x})), \boldsymbol{\Sigma}(\boldsymbol{x}))$, where $\boldsymbol{\mu}(\boldsymbol{x}) \in \mathbb{R}^C$, and $\boldsymbol{\Sigma}(\boldsymbol{x}) \in \mathbb{R}^{C \times C}$. To learn the distribution, we minimize the negative log-likelihood $-\log p(\boldsymbol{c} \mid \boldsymbol{x}) = -\log \int p(\boldsymbol{c} \mid \boldsymbol{\eta}) p_{\boldsymbol{\phi}}(\boldsymbol{\eta} \mid \boldsymbol{x}) d\boldsymbol{\eta}$. Due to its intractability, the integral is approximated by M Monte-Carlo samples

$$-\log \int p(\boldsymbol{c} \mid \boldsymbol{\eta}) p_{\boldsymbol{\phi}}(\boldsymbol{\eta} \mid \boldsymbol{x}) d\boldsymbol{\eta} \approx -\log \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{c} \mid \boldsymbol{\eta}^{(m)}),$$

where we employ the reparameterization trick $\boldsymbol{\eta}^{(m)} \mid \boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{x}) + \mathbf{L}(\boldsymbol{x})\boldsymbol{\epsilon}^{(m)}, \mathbf{L}(\boldsymbol{x})\mathbf{L}(\boldsymbol{x})^T = \boldsymbol{\Sigma}(\boldsymbol{x}), \boldsymbol{\epsilon}^{(m)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to compute gradients. Each concept c_i then depends on their corresponding sigmoid-transformed logit $\sigma(\eta_i)$ such that $\log p(\boldsymbol{c} \mid \boldsymbol{\eta}) = \sum_{i=1}^{C} \log p(c_i \mid \eta_i)$, where $p(c_i \mid \eta_i)$ describes a Bernoulli distribution. Combining the above considerations results in the following reformulation of the negative log-likelihood:

$$\mathcal{L}_{\boldsymbol{c}} = -\log \sum_{m=1}^{M} \exp \sum_{i=1}^{C} \left[-\text{BCE}(c_i, \sigma(\eta_i^{(m)})) \right], \quad (1)$$

where BCE stands for Binary Cross Entropy, and the log-sumexp trick is used for numerical stability.

The distribution-based modeling procedure allows for efficient sampling, thus, enabling SCBMs to train concept and target predictors jointly, sequentially, or independently. To prevent leakage, we follow Havasi et al. (2022) and train the model with the hard $\{0, 1\}$ concept values as bottleneck rather than the logits used in the original CBM (Koh et al., 2020). To retain differentiability, we employ the straight-through Gumbel-Softmax trick (Jang et al., 2017; Maddison et al., 2017). The target predictor $g_{\boldsymbol{\psi}}$ is then learned by minimizing the negative log-likelihood

$$\mathcal{L}_y = -\log \frac{1}{M} \sum_{m=1}^{M} p_{\boldsymbol{\psi}}(y \mid \boldsymbol{c}^{(m)}), \boldsymbol{c}^{(m)} \sim p(\boldsymbol{c} \mid \boldsymbol{x}). \quad (2)$$

Lastly, the learned dependencies are regularized by following Occam's razor and to prevent overfitting. We take inspiration from the Graphical Lasso (Friedman et al., 2008) and penalize the off-diagonal elements of the precision matrix $\mathcal{L}_{\boldsymbol{\Sigma}} = \sum_{i \neq j} \boldsymbol{\Sigma}(\boldsymbol{x})^{-1}_{i,j}$.

By combining concept, target, and precision loss with weighting factors λ_1 and λ_2, we arrive at the final loss function

$$\mathcal{L} = \mathcal{L}_{\boldsymbol{c}} + \lambda_1 \mathcal{L}_y + \lambda_2 \mathcal{L}_{\boldsymbol{\Sigma}}. \quad (3)$$

3.2. Covariance Learning

The introduced amortized covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{x})$ provides the flexibility to tailor its predicted concept dependencies to each data point. However, an amortized covariance matrix comes at the price of not being able to visualize and interpret a unified concept structure on a dataset level. Thus, we propose a variation of SCBM, where the covariance matrix is not *amortized* ($\boldsymbol{\Sigma}(\boldsymbol{x})$), but learned *globally* ($\boldsymbol{\Sigma}$). An example of the global concept structure learned on CUB is shown in Figure 1 (c). We recommend using the more flexible, amortized version by default and only utilizing a global covariance if the strong assumption of fixed dependencies is reasonable. We will explore this empirically in more detail in Section 5.

3.3. Interventions

A distinguishing property of CBM-like methods is the user's capacity to correct wrongly predicted concepts, which in turn affects the target prediction (Marcinkevičs et al., 2024). For a big concept set, this intervention procedure can become quite laborious as a user has to inspect and manually intervene on each concept separately. SCBMs are designed to alleviate this need by utilizing the learned concept dependencies such that a single intervention affects all related concepts as modeled by the multivariate normal distribution.

The parameterization as a multivariate normal distribution allows for a quick, scalable intervention procedure. Given a set $\mathcal{S} \subset \{1, \ldots, C\}$ of concept interventions, the effect on the remaining concepts $\boldsymbol{c}_{\setminus\mathcal{S}}$ is computed via their logits $\boldsymbol{\eta}_{\setminus\mathcal{S}}$ by conditioning on the intervention logits $\boldsymbol{\eta}'_{\mathcal{S}}$, utilizing the known properties of the normal distribution

$$\begin{aligned} \boldsymbol{\eta}_{\setminus\mathcal{S}} \mid \boldsymbol{x}, \boldsymbol{\eta}'_{\mathcal{S}} &\sim \mathcal{N}\left(\bar{\boldsymbol{\mu}}(\boldsymbol{x}), \overline{\boldsymbol{\Sigma}}(\boldsymbol{x})\right), \\ \bar{\boldsymbol{\mu}} &= \boldsymbol{\mu}_{\setminus\mathcal{S}} + \boldsymbol{\Sigma}_{\setminus\mathcal{S},\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1}(\boldsymbol{\eta}'_{\mathcal{S}} - \boldsymbol{\mu}_{\mathcal{S}}), \\ \overline{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{\setminus\mathcal{S},\setminus\mathcal{S}} - \boldsymbol{\Sigma}_{\setminus\mathcal{S},\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S},\setminus\mathcal{S}}. \end{aligned} \tag{4}$$

For a standard CBM (Koh et al., 2020), η'_i are set to the 5th (if $c_i = 0$) or 95th (if $c_i = 1$) percentile of the training distribution. Although this strategy is effective for SCBMs, see Appendix E.3, the explicit parameterization of our method enables us to take concept dependencies into account. To this end, we utilize the likelihood-based confidence region[1] that provides a natural way of capturing the area of possible $\boldsymbol{\eta}'_{\mathcal{S}}$ while taking into account concept dependencies. To determine the specific point within this region, we search for the values $\boldsymbol{\eta}'_{\mathcal{S}}$, which maximize the log-likelihood of the known, intervened-on concepts $\boldsymbol{c}_{\mathcal{S}}$, implicitly focusing on concepts that the model predicts poorly. We provide the explicit optimization problem in Appendix D.

[1] A confidence region is the multivariate generalization of a confidence interval.

4. Experimental Setup

Inspired by Marcinkevičs et al. (2024), we introduce a synthetic tabular dataset with a generating mechanism that contains fixed concept dependencies we can regulate. As natural image benchmark, we evaluate the Caltech-UCSD Birds-200-2011 dataset (Wah et al., 2011), with 200 bird classes and 112 concepts, as proposed in (Koh et al., 2020). Additionally, we explore CIFAR-10 (Krizhevsky et al., 2009) to mitigate the concept annotations requirement. 143 concepts classes are generated via GPT-3 (Brown et al., 2020) adopted from prior work (Oikarinen et al., 2023), and we obtain their binary labels using CLIP (Radford et al., 2021). Appendix B contains further details about all datasets.

To compare methods, we evaluate the concept and target accuracy before and after intervening on an increasing number of concepts. The order of intervened concepts is determined by an uncertainty-based policy (Shin et al., 2023). We also show results for a random policy in Appendix E.1. Additionally, we evaluate the calibration of the concept uncertainties with the Brier score (Brier, 1950) and Expected Calibration Error (Naeini et al., 2015; Kumar et al., 2019) in Appendix E.6. We compare our method with state-of-the-art models, detailed in Appendices A and C. Namely, we focus on the vanilla concept bottleneck model (CBM) by Koh et al. (2020) in its *hard* version (Havasi et al., 2022), the concept embedding model (CEM) by Espinosa Zarlenga et al. (2022), and the autoregressive CBM (AR CBM) (Havasi et al., 2022), with concept dependencies modelled in an autoregressive structure.

5. Results

Test performance In Table 1, we report the results of the concept and target accuracy prior to interventions. Overall, SCBM performs on par with the baseline methods, with no clear outperforming or underperforming technique throughout the datasets. This shows that the additional overhead of learning the concept dependencies does not negatively affect the predictive performance. We note that the amortized covariance consistently surpasses the globally learned variant due to its ability to adjust the predicted concept dependency structure and uncertainty on an instance level. On the other hand, the global variant offers a unified understanding of the concept correlations, as the example presented in Figure 1 (c). In Appendices E.5 and E.6, we further discuss time complexity and performance of proposed baselines.

Interventions In Figure 2, we show the intervention performances across ten seeds based on the concept and target accuracy. The first row shows that SCBMs are superior in modeling the concept dependencies, as evidenced by their steeper intervention curves, and the second row that the strong concept modeling translates to an improvement in

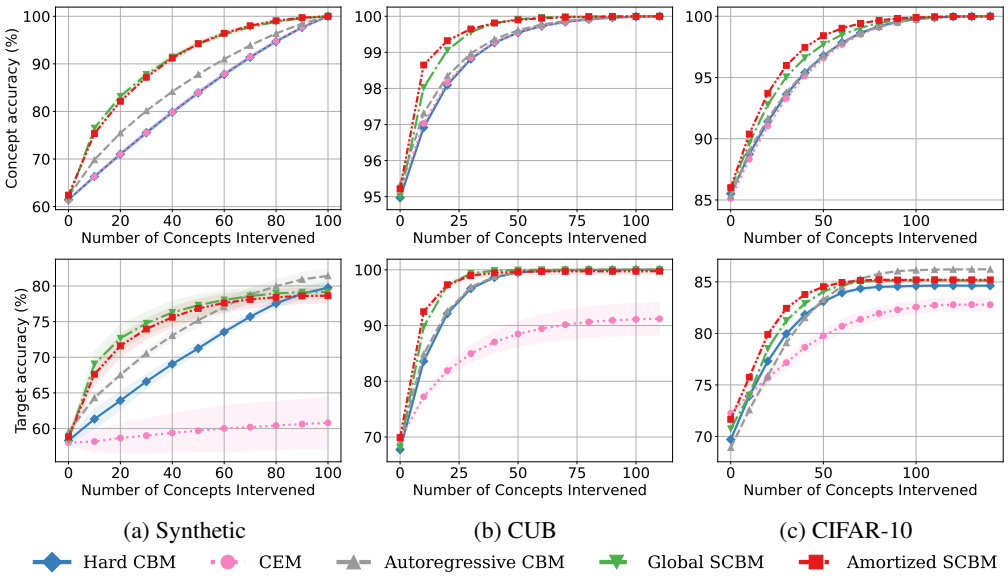


Figure 2. Performance after intervening on concepts in the order of highest predicted uncertainty. Concept and target accuracy (%) are shown in the first and second rows, respectively. Results are reported as averages and standard deviations across ten seeds.

Table 1. Test-set concept (c) and target (y) accuracy (%) prior to interventions reported as averages and standard deviations of performance across ten seeds. For each dataset and metric, the best-performing method is **bolded** and the runner-up is underlined.

Dataset	Method	c Accuracy	y Accuracy
Synthetic	Hard CBM	61.42 ± 0.07	58.38 ± 0.39
	CEM	61.42 ± 0.12	58.01 ± 0.49
	AR CBM	$\underline{62.17} \pm 0.11$	$\mathbf{59.60} \pm 0.62$
	Global SCBM	61.57 ± 0.05	58.39 ± 0.53
	Amortized SCBM	$\mathbf{62.41} \pm 0.20$	$\underline{58.96} \pm 0.38$
CUB	Hard CBM	94.97 ± 0.07	67.72 ± 0.57
	CEM	95.12 ± 0.07	$\underline{69.60} \pm 0.30$
	AR CBM	$\mathbf{95.33} \pm 0.07$	69.24 ± 0.44
	Global SCBM	94.99 ± 0.09	68.19 ± 0.63
	Amortized SCBM	$\underline{95.22} \pm 0.09$	$\mathbf{69.87} \pm 0.56$
CIFAR-10	Hard CBM	85.51 ± 0.04	69.73 ± 0.29
	CEM	85.12 ± 0.14	$\mathbf{72.24} \pm 0.33$
	AR CBM	85.31 ± 0.06	68.88 ± 0.47
	Global SCBM	$\underline{85.86} \pm 0.04$	70.74 ± 0.29
	Amortized SCBM	$\mathbf{86.00} \pm 0.03$	$\underline{71.66} \pm 0.25$

downstream performance, partly thanks to the intervention strategy introduced in Section 3.3. We note that especially for the most practical scenario of only a small number of interventions, SCBMs outperform their counterparts. The success of SCBMs on CIFAR-10, with CLIP-based concepts, shows it can work without human-annotated concepts.

Analyzing the AR CBM, which also captures concept dependencies, but not to a full extent, we observe a better intervention performance than the hard vanilla CBM, which does not take correlations into account but still lower than SCBMs. This shows in the target accuracy, where they only match or outperform SCBMs towards the full set of intervened concepts. We attribute this improvement to the independent training procedure utilized by AR CBMs, which comes at the cost of lower test performance in CIFAR-10. Finally, the CEM shows reduced intervention performance as the expressive concept embeddings, prone to information leakage, suboptimally adapt to the injected concept information.

6. Conclusion

In this paper, we introduced SCBMs, a new concept-based method that models concept dependencies with a multivariate normal distribution. We proposed a novel, effective intervention strategy that takes concept correlations into account and is based on the confidence region inferred from the distributional parameterization. We showed that our modeling approach retains CBMs' training and inference speed, thus, being able to harness the benefits of end-to-end concept and target training. Additionally, the explicit parameterization offers the user a clearer understanding of the learned concept dependencies. Empirically, we demonstrated that SCBMs offer a substantial improvement in intervention effectiveness while retaining test performance prior to interventions. We showed that our method excels when iteratively intervening on the most uncertain concept predictions, sparing users from having to manually search through the concept set to identify necessary interventions. Finally, the versatility of SCBMs is highlighted through their success on CIFAR-10, with CLIP-based rather than human annotations. Limitations and future work are discussed in Appendix F.

Acknowledgements

The authors would like to thank Dr. Alexander Marx, Department of Statistics, TU Dortmund, for insightful discussions. SL and MV are supported by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number MB22.00047. RM is supported by the SNSF grant #320038189096.

References

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 929–947, 2024.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. URL https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., and Dvijotham, K. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5948–5955, 2023.

Collins, K. M., Barker, M., Zarlenga, M. E., Raman, N., Bhatt, U., Jamnik, M., Sucholutsky, I., Weller, A., and Dvijotham, K. Human uncertainty in concept-based AI systems. In Rossi, F., Das, S., Davis, J., Firth-Butterfield, K., and John, A. (eds.), *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, pp. 869–889. ACM, 2023.

Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21400–21413, 2022.

Espinosa Zarlenga, M., Collins, K., Dvijotham, K., Weller, A., Shams, Z., and Jamnik, M. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36, 2024.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Havasi, M., Parbhoo, S., and Doshi-Velez, F. Addressing leakage in concept bottleneck models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=tglniD_fn9.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heidemann, L., Monnet, M., and Roscher, K. Concept correlation and its effects on concept-based models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4780–4788, 2023.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Kim, E., Jung, D., Park, S., Kim, S., and Yoon, S. Probabilistic concept bottleneck models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16521–16540. PMLR, 2023. URL https://proceedings.mlr.press/v202/kim23g.html.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference*

Track Proceedings, 2014. URL http://arxiv.org/abs/1312.6114.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, Virtual, 2020. PMLR. URL https://proceedings.mlr.press/v119/koh20a.html.

Kraft, D. A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372, Kyoto, Japan, 2009. IEEE. URL https://doi.org/10.1109/ICCV.2009.5459250.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009. IEEE. URL https://doi.org/10.1109/CVPR.2009.5206594.

Leino, K., Sen, S., Datta, A., Fredrikson, M., and Li, L. Influence-directed explanations for deep convolutional networks. In *2018 IEEE International Test Conference (ITC)*. IEEE, 2018. URL https://doi.org/10.1109/test.2018.8624792.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., and Pan, W. Promises and pitfalls of black-box concept learning models, 2021. URL https://doi.org/10.48550/arXiv.2106.13314. *arXiv:2106.13314*.

Marcinkevičs, R., Laguna, S., Vandenhirtz, M., and Vogt, J. E. Beyond concept bottleneck models: How to make black boxes intervenable? *arXiv preprint arXiv:2401.13544*, 2024.

Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Chin-Cheong, K., Paschke, A., Zerres, J., Denzinger, M., Niederberger, D., Wellmann, S., Ozkan, E., Knorr, C., and Vogt, J. E. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis*, 91:103042, 2024. URL https://www.sciencedirect.com/science/article/pii/S136184152300302X.

Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., and Weller, A. Do concept bottleneck models learn as intended?, 2021. URL https://doi.org/10.48550/arXiv.2105.04289. *arXiv:2105.04289*.

Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., and Glocker, B. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In *Advances in neural information processing systems*, volume 33, pp. 12756–12767, 2020.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Neal, R. M. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, Canada, 1995. URL https://librarysearch.library.utoronto.ca/permalink/01UTORONTO_INST/14bjeso/alma991106438365706196.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sheth, I., Rahman, A. A., Sevyeri, L. R., Havaei, M., and Kahou, S. E. Learning from uncertain concepts via test time interventions. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022. URL https://openreview.net/forum?id=WVe3vok8Cc3.

Shin, S., Jo, Y., Ahn, S., and Lee, N. A closer look at the intervention procedure of concept bottleneck models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31504–31520. PMLR, 2023. URL https://proceedings.mlr.press/v202/shin23a.html.

Silvey, S. *Statistical Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1975. ISBN 9780412138201. URL https://books.google.ch/books?id=qIKLejbVMf4C.

Steinmann, D., Stammer, W., Friedrich, F., and Kersting, K. Learning to intervene on concept bottlenecks, 2023. URL https://doi.org/10.48550/arXiv.2308.13453. *arXiv:2308.13453*.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

A. Related Work

We review relevant research on concept bottleneck models to highlight current approaches and improve understanding of the associated challenges and opportunities. While Koh et al. (2020) propose the *soft* vanilla CBM, where the concept logits parameterize the bottleneck, Havasi et al. (2022) argue that such a representation leads to leakage, where additional unwanted information in the concept representation is used to predict the target (Margeloiu et al., 2021; Mahinpei et al., 2021). Thus, they parameterize the bottleneck by binarized concept predictions and call it the *hard* CBM. Then, Havasi et al. (2022) equip the hard CBM with an autoregressive structure of the form $c_i|\boldsymbol{x}, \boldsymbol{c}_{<i}$, which is supposed to learn the concept dependencies. As such, the implicit autoregressive modeling of concept dependencies by Havasi et al. (2022) is the most related to the current work. Complementary to our work, Heidemann et al. (2023) analyze how a CBM's performance is affected by concept correlations. Unlike approaches that restrict the bottleneck to prevent leakage, Concept Embedding Models (CEM) (Espinosa Zarlenga et al., 2022) represent each concept with a predicted embedding vector from which the concept probabilities can be inferred, treating the problem akin to a multi-task setting. Kim et al. (2023) model the embedding with a normal distribution, assuming a diagonal covariance matrix, which prevents them from capturing concept dependencies. Recent works explored how a CBM-like structure can be enforced even without a concept-annotated training set. Yuksekgonul et al. (2023) transform a pre-trained model into a CBM via a concept bank from concept activation vectors and multimodal models (Kim et al., 2018), while Oikarinen et al. (2023) query GPT-3 (Brown et al., 2020) for the concept set $\mathcal{C}$ and assign the values of the concept activations to each datapoint $\boldsymbol{x}$ with CLIP (Radford et al., 2021) similarities. Marcinkevičs et al. (2024) instead relax the need for a concept labeled training set to a smaller validation set by fine-tuning a pre-trained model.

Intervenability (Marcinkevičs et al., 2024) is a crucial element of CBMs as it allows the user to correct wrongly predicted concepts $\hat{\boldsymbol{c}}$ to $\boldsymbol{c}'$, which in turn affects the target prediction of the model $\hat{y}'$. If multiple concepts are intervened on, then the order of interventions is important. To this end, Sheth et al. (2022) and Shin et al. (2023) explore multiple policies according to which the order of concepts is determined. Chauhan et al. (2023) propose to combine predefined policies with learnable weighting parameters, while Espinosa Zarlenga et al. (2024) learn the policy itself. Steinmann et al. (2023) argue that instance-specific interventions are costly and store previous interventions in a memory to automatically reapply them for similar data points. Lastly, Collins et al. (2023) explore the advantages of including uncertainty rather than treating humans as oracles.

Our work models concept dependencies by parameterizing the bottleneck with a distribution. In a similar vein, Variational Autoencoders (Kingma & Welling, 2014) parameterize the bottleneck with a normal distribution to model and generate new data. Stochastic Segmentation Networks (Monteiro et al., 2020) parameterize the logits of a segmentation map with a non-diagonal normal distribution to capture the spatial correlations of pixels and model the aleatoric uncertainty. The modeling of uncertainty with a distribution is also explored by Bayesian Neural Networks (Neal, 1995) that learn a probability distribution over the neurons of a neural network.

B. Dataset Details

In this section, we provide additional details on the datasets that are being used in the experiments.

B.1. Synthetic Data-Generating Mechanism

Here, we describe the data-generating mechanism of the synthetic dataset in more detail. In particular, the concept logits $\boldsymbol{\eta}$ are sampled from a randomly initialized positive definite covariance matrix and generate $\boldsymbol{x}$. Binary concept values $\boldsymbol{c}$ are inferred from $\boldsymbol{\eta}$ and generate the target y. Let N, p, and C denote the number of independent data points $\{(\boldsymbol{x}_n, \boldsymbol{c}_n, y_n)\}_{n=1}^N$, covariates, and concepts, respectively. We set $N = 50{,}000$, $p = 1{,}500$, and $C = 100$, with a 60%-20%-20% train-validation-test split. The generative process is as follows:

1. Randomly sample $\boldsymbol{W} \in \mathbb{R}^{C \times 10}$ s.t. $w_{i,j} \sim \mathcal{N}(0, 1)$ for $1 \leq i \leq C$ and $1 \leq j \leq 10$.
2. Generate a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{C \times C}$ s.t. $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{D}$. Let $\boldsymbol{D} \in \mathbb{R}^{C \times C}$ s.t. $\boldsymbol{D} = \boldsymbol{\delta}\boldsymbol{I}$, where $\delta_i \sim \mathcal{U}_{[0,1]}$ for $1 \leq i \leq C$.
3. Randomly sample logits $\boldsymbol{H} \in \mathbb{R}^{N \times C}$ s.t. $\boldsymbol{\eta}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $1 \leq n \leq N$.
4. Let $c_{n,i} = \mathbb{1}_{\{\eta_{n,i} \geq 0\}}$ for $1 \leq n \leq N$ and $1 \leq i \leq C$.

5. Let $h : \mathbb{R}^C \to \mathbb{R}^p$ be a randomly initialised multilayer perceptron with ReLU nonlinearities.
6. Let $\boldsymbol{x}_n = h(\boldsymbol{\eta}_n) + \boldsymbol{\epsilon}_n$ s.t. $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ for $1 \leq n \leq N$.
7. Let $g : \mathbb{R}^C \to \mathbb{R}$ be a randomly initialized linear perceptron.
8. Let $y_n = \mathbb{1}_{\{(g(\boldsymbol{c}_n) \geq y_{med})\}}$ for $1 \leq n \leq N$, where y_{med} denotes the median of $g(\boldsymbol{c}_n)$.

B.2. Natural Image Datasets

Caltech-UCSD Birds-200-2011 We evaluate on the Caltech-UCSD Birds-200-2011 (CUB)[2] dataset (Wah et al., 2011). It comprises 11,788 photographs from 200 distinct bird species annotated with 312 concepts, such as belly color and pattern. In this manuscript, we follow the original train-test split and revised the proposed dataset in the initial CBM work (Koh et al., 2020). Here, only the 112 most widespread binary attributes are included in the final dataset, and concepts are shared across samples in identical classes. The images were resized to a resolution of 224 × 224 pixels. Finally, following the original proposed augmentations, we applied random horizontal flips, modified the brightness and saturation, and applied normalization during training.

CIFAR-10 CIFAR-10[3] (Krizhevsky et al., 2009) is a natural image benchmark with 60,000 32x32 colour images and 10 classes. We kept the original train-test split, with 50,000 samples in the train set and a balanced total of 6,000 images per class. We generated 143 concept labels as described in Section 4 using large language and vision models. In particular, we compute the similarity between each instance of an image with the concept text embedding and compare it to the similarity of its negative counterpart, i.e. *not* the concept. At training time, as for CUB, we applied augmentations including modifications to brightness and saturation, random horizontal flips and normalisation. Images were rescaled to a size of 224 × 224 pixels.

C. Experimental Details

C.1. Baselines

We evaluate the performance of our method in comparison with state-of-the-art models. Namely, we focus on the vanilla concept bottleneck model (CBM) by Koh et al. (2020) in its *hard* version (Havasi et al., 2022), trained jointly using the straight-through Gumbel-Softmax trick (Jang et al., 2017; Maddison et al., 2017), as a sensical baseline to our binary modeling of concepts. Additionally, we explore the concept embedding model (CEM) by Espinosa Zarlenga et al. (2022) that learns two concept embeddings, $\hat{\boldsymbol{c}}_i^+$ and $\hat{\boldsymbol{c}}_i^-$. These representations are used to predict the final concept probability with a learnable scoring function $\hat{p}_i = s(\hat{\boldsymbol{c}}_i^+, \hat{\boldsymbol{c}}_i^-) = \sigma(\mathbf{W}_s[\hat{\boldsymbol{c}}_i^+, \hat{\boldsymbol{c}}_i^-]^T + \mathbf{b}_s)$ and are then combined on a final concept embedding $\hat{\boldsymbol{c}}_i = (\hat{p}_i \hat{\boldsymbol{c}}_i^+ + (1 - \hat{p}_i)\hat{\boldsymbol{c}}_i^-)$ that is passed to the target predictor. Interventions are modeled by altering the concept probabilities $\hat{p}_i$. Finally, we evaluate the autoregressive CBM structure proposed by Havasi et al. (2022), where concept dependencies are learned with an autoregressive structure. Here, each concept c_i is predicted with a separate MLP that takes as input a shared latent representation of the input $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and all previous concepts $c_1, ..., c_{i-1}$. To obtain a good initialization of the autoregressive structure, it is pretrained for 50 epochs. As the Monte-Carlo sampling from the autoregressive structure is time-consuming, the target predictor $g_{\boldsymbol{\psi}}$ is trained independently using the ground-truth concepts as input. At intervention time, a normalized importance sampling algorithm is used to estimate the concept distribution.

C.2. Implementation Details

This section provides the implementation details of SCBM and the evaluated baselines. All methods were implemented using PyTorch (v 2.1.1) (Ansel et al., 2024). All models are trained for 150 epochs for the synthetic and 300 epochs for the natural image datasets with the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-4} and a batch size of 64. For the independently trained autoregressive model, we split the training epochs into $2/3$ for the concept predictor and $1/3$ for the target predictor. For the methods requiring sampling, the number of Monte-Carlo samples is set to $M = 100$. The model architectures are comprised of a backbone for concept prediction followed by a linear layer as head for an interpretable target prediction. For the synthetic tabular data, we use a fully connected neural network as backbone, with 3 non-linear layers, batch normalization, and dropout. For the CUB dataset, we use a pretrained ResNet-18 (He et al., 2016),

[2] https://www.vision.caltech.edu/datasets/cub_200_2011/, no license available
[3] https://www.cs.toronto.edu/~kriz/cifar.html, no license available

and for the lower-resolution CIFAR-10 a simple convolutional neural network with 2 convolutional layers followed by ReLU, Dropout, and a fully connected layer. For fairness in the comparisons, all baselines have the same model architecture choices and all experiments are performed over 10 random seeds.

To ensure the positive definiteness of the concept covariance matrix $\boldsymbol{\Sigma}$, we parameterize it via its Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$. Thus, we solely predict the lower triangular Cholesky matrix $\boldsymbol{L}$. We will evaluate two options for SCBMs: using a *global* ($\boldsymbol{\Sigma}$) or an *amortized* covariance matrix ($\boldsymbol{\Sigma}(\boldsymbol{x})$). For the amortized version, we set the weighting terms λ_1 and λ_2 of Equation 3 to 1. For the global version, we initialize it with the estimated empirical covariance matrix and set $\lambda_2 = 0$, as we did not observe big differences when varying λ_2. In Appendix E.2, we provide an ablation study, demonstrating that SCBMs are not very sensitive to the choice of λ_2. At intervention time, we solve the optimization problem based on the 99%-confidence region with the SLSQP algorithm (Kraft, 1988). In Appendix E.4, we provide an ablation with different confidence levels.

Resource Usage For the experiments of the main paper, we used a cluster of mostly GeForce RTX 2080's with 2 CPU workers. Over all methods, we estimate an average runtime of 8h per experiment. This amounts to 5 methods $\times$ 3 datasets $\times$ 10 seeds $\times$ 8 hours $=$ 1200 hours. Adding to that, the Ablation Figures required another 40 runs, amounting to a full total of 1520 hours of compute. Please note that we only report the numbers to generate the final results but not the development time, which we roughly estimate to be around 10 times bigger.

D. Intervention Strategy

For a standard CBM (Koh et al., 2020), intervention logits η_i' are set to the 5th (if $c_i = 0$) or 95th (if $c_i = 1$) percentile of the training distribution. This strategy presents certain limitations that result in a suboptimal intervention performance when interventions affect other concepts. For example, if the initially predicted μ_i was more extreme than the selected training percentile, the interventional shift guided by $\eta_i' - \mu_i$ would point in the wrong direction. This, in turn, would cause $\boldsymbol{\eta}_{\setminus\mathcal{S}}$ to shift incorrectly. Thus, we pose the desideratum that an appropriate intervention strategy should determine η_i' such that $\eta_i' - \mu_i \geq 0$ if $c_i = 1$, and $\eta_i' - \mu_i \leq 0$ if $c_i = 0$. Additionally, $\eta_i' - \mu_i$ should not be "too large" as to avoid that the intervention completely disregards the predicted $\boldsymbol{\mu}_{\setminus\mathcal{S}}$.

Here manifests an additional benefit of the explicit distributional representation: the likelihood-based confidence region provides a natural way of capturing the region of possible $\boldsymbol{\eta}_{\mathcal{S}}'$ that fulfill our desiderata. Note that the confidence region takes concept dependencies into account when describing the area of possible $\boldsymbol{\eta}_{\mathcal{S}}'$. To pinpoint the specific location within this region, we search for the values $\boldsymbol{\eta}_{\mathcal{S}}'$ that maximize the log-likelihood of the known intervened concepts $\boldsymbol{c}_{\mathcal{S}}$, thereby focusing on poorly predicted concepts.

$$
\begin{aligned}
\boldsymbol{\eta}_{\mathcal{S}}' = \arg\max_{\boldsymbol{\eta}_{\mathcal{S}}} & \log p(\boldsymbol{c}_{\mathcal{S}} \mid \boldsymbol{\eta}_{\mathcal{S}}) \\
\text{s.t.} & -2\left(\log p(\boldsymbol{\eta}_{\mathcal{S}} \mid \boldsymbol{\mu}_{\mathcal{S}}, \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}) - \log p(\boldsymbol{\mu}_{\mathcal{S}} \mid \boldsymbol{\mu}_{\mathcal{S}}, \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}})\right) \leq \chi^2_{d,1-\alpha} \\
& \eta_i' - \mu_i \geq 0 \text{ if } c_i = 1, \quad \forall i \in \mathcal{S} \\
& \eta_i' - \mu_i \leq 0 \text{ if } c_i = 0, \quad \forall i \in \mathcal{S},
\end{aligned}
\tag{5}
$$

where $d = |\mathcal{S}|$. The first inequality describes the confidence region. It is based on the logarithm of the likelihood ratio, which, after multiplying with -2, asymptotically follows a χ^2 distribution (Silvey, 1975). The last two inequalities restrict the region to the desired direction. Note that $\boldsymbol{\eta}_{\mathcal{S}}'$ is computed to determine the conditional effect of the interventions on $\boldsymbol{\eta}_{\setminus\mathcal{S}}$ using Equation 4. When predicting $\hat{y}'$ under interventions, the logits $\boldsymbol{\eta}_{\setminus\mathcal{S}}$ are then used for sampling the binary concept values $\boldsymbol{c}_{\setminus\mathcal{S}}$ while the intervened-on concepts $\boldsymbol{c}_{\mathcal{S}}'$ are directly set to their known, binary value.

E. Further Experiments

In this section, we show additional experiments to provide a more in-depth understanding of SCBM's effectiveness. We ablate multiple hyperparameters to provide an understanding of how they influence the model performance.

E.1. Random Intervention Policy

In Figure 3, we present the intervention performance of SCBM and baseline methods. Compared to the uncertainty-based intervention policy of Figure 2, the intervention curves of all methods are less steep, confirming the usefulness of Shin

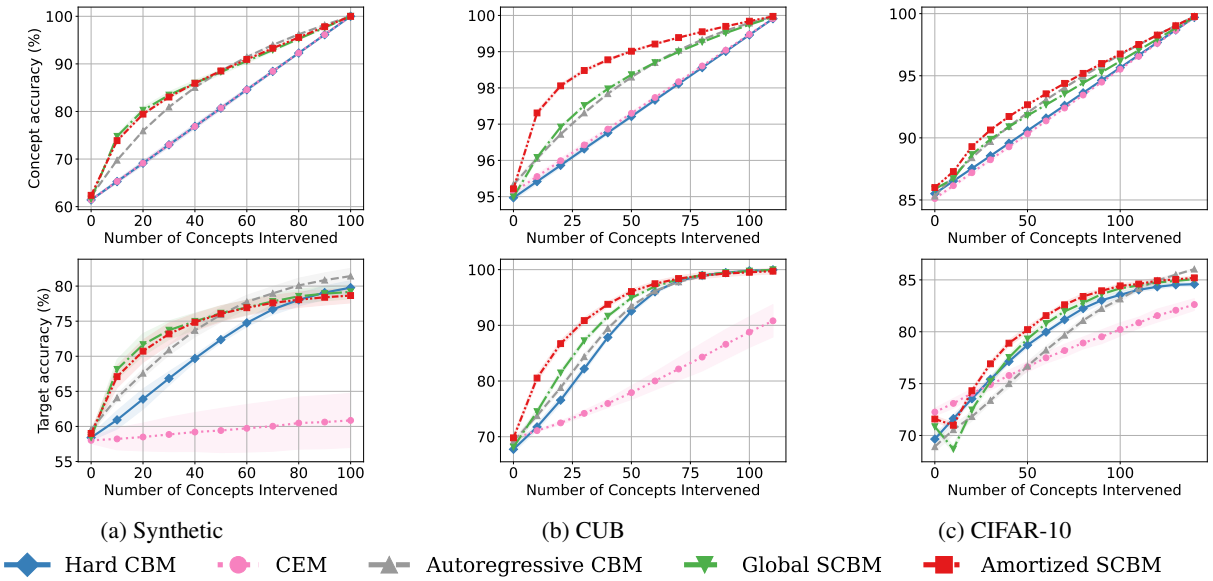


Figure 3. Performance after intervening on concepts in random order. Concept and target accuracy (%) are shown in the first and second rows, respectively. Results are reported as averages and standard deviations of model performance across ten seeds.

et al. (2023)'s proposed policy. Following the previous statements, SCBMs still outperform baseline methods with the amortized beating the global variant for real-world datasets. We observe that in CIFAR-10 for the first interventions, an improvement in concept accuracy is not directly reflected in improved target prediction for SCBMs, which is likely due to the low signal-to-noise ratio of the CLIP-inferred concepts.

E.2. Regularization Strength

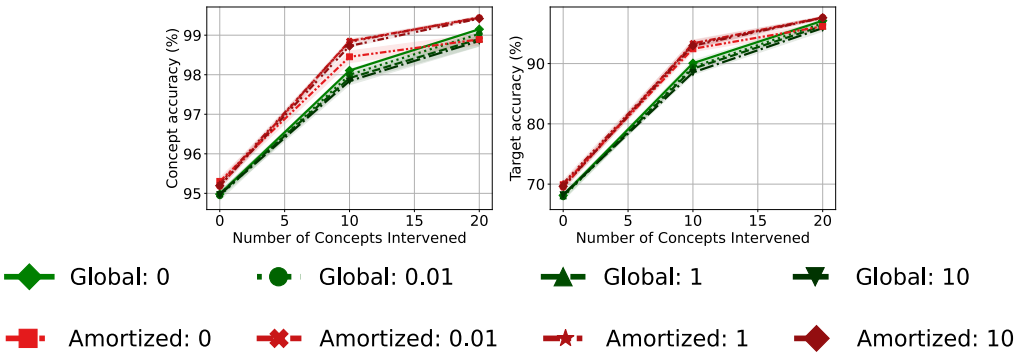


Figure 4. Performance on CUB after intervening on concepts in the order of highest predicted uncertainty with differing regularization strengths. Concept and target accuracy (%) are shown in the first and second columns, respectively. Results are reported as averages and standard deviations of model performance across five seeds. For each SCBM variant, we choose a darker color, the higher the regularization strength of λ_2.

In Figure 4, we analyze the impact of the strength of λ_2 from Equation 3. Due to environmental considerations, we conducted experiments using only 5 seeds and limited the number of interventions to 20. Our findings indicate that SCBMs are not sensitive to the choice of λ_2, except that the unregularized amortized variant exhibits slight patterns of overfitting.

E.3. Intervention Strategy

In Figure 5, we analyze the effect of the intervention strategy. Our findings indicate that while SCBMs are still effective with the proposed strategy from Koh et al. (2020), that sets the logits to the 5th (if $c_i = 0$) or 95th (if $c_i = 1$) percentile of the training distribution, our proposed strategy based on the confidence region results in stronger intervenability.

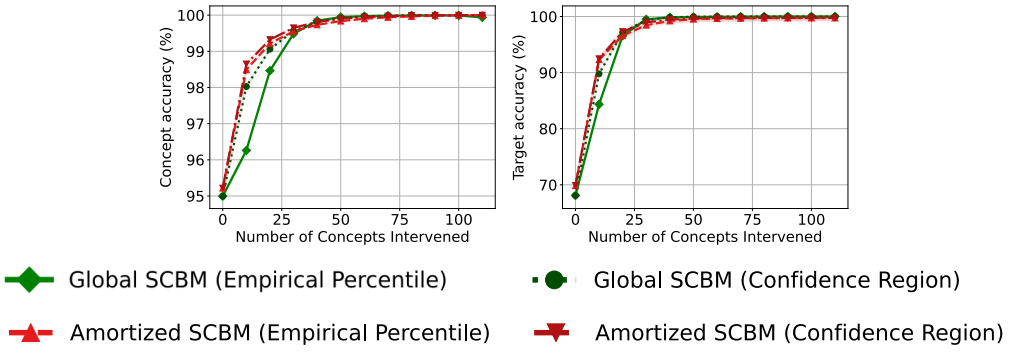


Figure 5. Performance on CUB after intervening on concepts in the order of highest predicted uncertainty, comparing the proposed intervention strategy to Koh et al. (2020)'s intervention of setting the logits to the 5th or 95th empirical percentile of the training distribution. Concept and target accuracy (%) are shown in the first and second columns, respectively. Results are reported as averages and standard deviations of model performance across five seeds.

E.4. Confidence Region Level

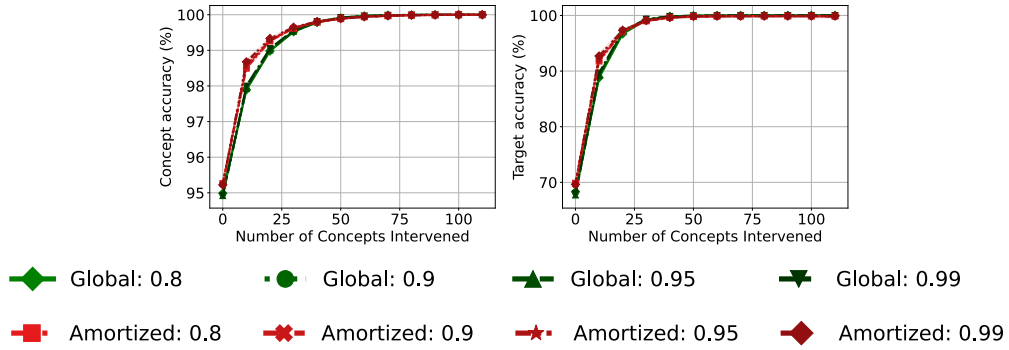


Figure 6. Performance on CUB after intervening on concepts in the order of highest predicted uncertainty with differing levels $1 - \alpha$ of the confidence region. Concept and target accuracy (%) are shown in the first and second columns, respectively. Results are reported as averages and standard deviations of model performance across three seeds.

In Figure 6, we analyze the effect of the level $1 - \alpha$ of the likelihood-based confidence region. Our findings indicate that the SCBMs are not sensitive to the choice of $1 - \alpha$, with higher levels being slightly better in performance.

E.5. Comparative performance

In this section, we further discuss the comparison between baseline methods without interventions, as introduced in Section 5 and introduce the time it takes for training and testing of the methods. Notably, from Table 1 in CIFAR-10, even though the concept performance of CEM is the worst of all methods, it has the best target performance. This might suggest the presence of leakage in CEM's embeddings, as in CIFAR-10, the concept set alone is not sufficient to predict the target, and learning additional information might be useful. In Table 2, it is evident that the autoregressive CBM of Havasi et al. (2022) suffers from a slow sampling process due to its autoregressive structure, while SCBMs retain the efficiency of CBMs and CEMs.

Table 2. Relative time it takes for one epoch in the CUB dataset when training on the training set, or evaluating on the test set, respectively.

Method	Training	Inference
Hard CBM	5x	1x
CEM	5x	1x
Autoregressive CBM	5x	14x
Global SCBM	5x	1x
Amortized SCBM	5x	1x

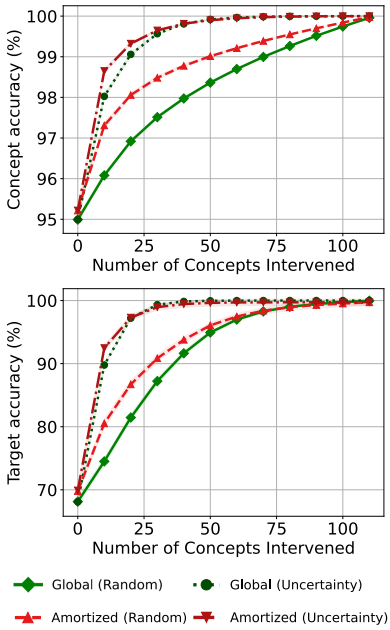
E.6. Modeling the concept distribution

A cornerstone of SCBMs is the explicit, distributional parameterization of concepts. This helps in understanding the data correlations and allows for visualization, as the example seen in Figure 1 (c). The explicit probabilistic modeling results in improved concept uncertainty estimates compared to the baseline CBM counterparts, as shown in Table 3, where lower metrics imply better estimates. This proves useful for interventions, where the uncertainty estimates can be leveraged for the choice of concept to intervene on, improving the target prediction more effectively and reducing the need for manual user inspection. In Figure 7, we compare the performance of randomly intervening versus intervening based on the predicted uncertainty. We observe that there is a big gap between the two policies, indicating the usefulness of the estimated probabilities. Nevertheless, note that intervening at random remains successful and supports the observations made in the previous paragraph, as shown in Appendix E.1.

Table 3. Test-set calibration (%) of concept predictions. Results are reported as averages and standard deviations of model performance across ten seeds. For each dataset and metric, the best-performing method is **bolded** and the runner-up is underlined. Lower is better.

Dataset	Method	Brier	ECE
Synthetic	Hard CBM	28.79 ± 0.09	22.38 ± 0.15
	CEM	29.32 ± 0.08	23.55 ± 0.09
	Autoregressive CBM	$\mathbf{24.84} \pm 0.32$	$\mathbf{13.54} \pm 0.49$
	Global SCBM	27.73 ± 0.09	20.10 ± 0.14
	Amortized SCBM	$\underline{25.58} \pm 0.20$	$\underline{15.57} \pm 0.55$
CUB	Hard CBM	3.93 ± 0.05	2.44 ± 0.06
	CEM	4.04 ± 0.05	3.25 ± 0.07
	Autoregressive CBM	$\underline{3.75} \pm 0.05$	2.73 ± 0.05
	Global SCBM	3.87 ± 0.06	$\underline{2.33} \pm 0.09$
	Amortized SCBM	$\mathbf{3.64} \pm 0.07$	$\mathbf{1.85} \pm 0.08$
CIFAR-10	Hard CBM	10.42 ± 0.05	4.93 ± 0.17
	CEM	11.06 ± 0.16	7.11 ± 0.39
	Autoregressive CBM	10.70 ± 0.05	6.07 ± 0.10
	Global SCBM	$\underline{9.95} \pm 0.02$	$\underline{2.88} \pm 0.11$
	Amortized SCBM	$\mathbf{9.84} \pm 0.02$	$\mathbf{2.22} \pm 0.12$

Figure 7. Intervention performance of SCBMs measured in concept and target accuracy (%) on CUB for random and uncertainty-based policy.



F. Limitations & Future Work

In this section, we discuss the limitations that our proposed method poses together with the multiple new research avenues that it opens. A natural extension is to go beyond binary concepts, such as continuous domains with their corresponding adaptations of modeling the concept distribution. Additionally, addressing the quadratic memory complexity of the covariance matrix is essential for scaling to larger concept sets. Current interventions focus on editing the concept values. However, this work allows the editing of the learned dependency structure by adjusting the entries of the predicted covariance matrix, which could be explored. Lastly, to model additional information and reduce leakage, Koh et al. (2020); Havasi et al. (2022) propose the adoption of a side channel. The complementary effectiveness of incorporating the side channel in the covariance structure could be explored in the context of SCBMs.