Hierarchical Causal Representation Learning

Angelos Nalmpantis[1]*, **Phillip Lippe**[1], **Sara Magliacane**[2,3]
[1]QUVA Lab, University of Amsterdam
[2]AMLab, University of Amsterdam
[3]MIT-IBM Watson AI Lab

Abstract

Learning causal representations is a crucial step toward understanding and reasoning about an agent's actions in embodied AI and reinforcement learning. In many scenarios, an intelligent agent starts learning to interact with an environment by initially performing coarse actions with multiple simultaneous effects. During the learning process, the agent starts acquiring more fine-grained skills that can now affect only some of the factors in the environment. This setting is currently underexplored in current causal representation learning methods that typically learn a single causal representation and do not reuse or refine previously learned representations. In this paper, we introduce the problem of *hierarchical causal representation learning*, which leverages causal representations learned with coarse interactions and progressively refines them, as more fine-grained interactions become available. We propose HERCULES, a method that builds a hierarchical structure where at each level it gradually identifies more fine-grained causal variables by leveraging increasingly refined interventions. In experiments on two benchmarks of sequences of images with intervened causal factors, we demonstrate that HERCULES successfully recovers the causal factors of the underlying system and outperforms current state-of-the-art methods in scenarios with limited fine-grained data. At the same time, the acquired representations of HERCULES exhibit great adaptation capabilities under local transformations of the causal factors.

1 Introduction

Causal representation learning (CRL) [48] aims at identifying the causal variables of an underlying system along their relations given high-dimensional observations, e.g., images. Learning causal representations is a crucial step toward understanding and reasoning about an agent's actions in embodied AI and reinforcement learning. Many previous studies focused on obtaining causal representations by using intervention targets [30, 32, 33], counterfactual observations [5, 35, 56], environment interactions [29, 31] or nonstationarity [58, 59]. These methods learn a single causal representation, instead of refining previously learned representations when there are changes in the data or environment that allow for a more fine-grained identifiability.

In many practical scenarios, we might have an intelligent agent, which starts learning to interact with an environment by initially performing coarse actions that have multiple simultaneous effects, e.g. perturbing many causal variables in the environment. During the learning process, the agent starts acquiring more fine-grained skills that can now perturb only some of the causal variables. While the fine-grained actions might help completely identify the causal factors, we still want to be able to efficiently reuse the coarse causal representations that we learned in the first phase. As we experimentally demonstrate, these coarse causal representations reduce the required amount of fine-grained interactions needed to fully identify the causal factors. Moreover, maintaining different

[*]Correspondence to: angelosnalm@gmail.com

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

levels of coarseness of causal representations allows us to generalize across environments with the same underlying system, but different fine-grained representations.

In this paper, we introduce the *hierarchical causal representation learning* (HCRL) setting, which considers hierarchies of progressively more fine-grained causal representations. We propose an HCRL method, HERCULES, that learns causal representations in a hierarchical manner, extending CITRIS [33]. In the initial phase, HERCULES captures high-level causal variables, inferred from the coarse interventions. Based on this initial disentanglement, HERCULES redefines the representations and further disentangles each group of causal variables into low-level causal variables, using more fine-grained interventions. Overall, this iterative process leads to a hierarchical structure adapting to the available skills (interventions) at each time. Through this hierarchical framework, increasingly fine-grained interactions are leveraged, ultimately identifying each individual causal variable.

Overall, our contributions are summarized as follows:

- We introduce the hierarchical causal representation learning (HCRL) setting, which aims at learning intermediate representations that *block-identify* [56] groups of causal factors and are further refined across hierarchical levels.
- We propose the first method for HCRL, HERCULES, which leverages interactions at different levels of granularity to build a versatile hierarchical causal framework. Its hierarchical structure can be expanded to accommodate different numbers of hierarchical levels.
- We modify two current benchmark datasets for the HCRL setting and demonstrate that HERCULES outperforms current state-of-the-art methods in scenarios with limited fine-grained interactions. At the same time, we show that in some initial experiments, the acquired representations can efficiently adapt to environments where the effects of causal variables differ in the observational space.

2 Related Work

Learning hierarchical representations and CRL are two fields that were previously explored independently. In this section, we provide a comprehensive overview of their literature.

Hierarchical Representations. The concept of hierarchical representations has been broadly studied, with many approaches ranging from attention-based models such as Swin Transformer [34] to probabilistic generative models like hierarchical variational models (HVMs) [43]. In these models, the early layers commonly capture local, low-level features while deeper layers focus on high-level concepts. Other works learn a hierarchical structure by reformulating the variational autoencoder's (VAE) [24] objective as the Lagrangian of a constrained optimization problem [27, 46], utilizing the nested Chinese Restaurant Process [13] or employing a ladder-based approach for VAEs [50] with NVAE [52] narrowing the sizeable gap between VAEs and other generative models.

In another line of work, a growing body of research has explored the use of alternative geometries, modeling embeddings in the *hyperbolic* space. Despite the success of Euclidean embeddings [8, 54], they still fall short when data exhibit latent hierarchical structures, since they require a prohibitive amount of dimensions to capture complex relations [40]. On the other hand, hyperbolic spaces can capture hierarchical relations with few dimensions due to their exponentially increasing volume [47]. Thereby, a surge of works exploited the capabilities of hyperbolic embeddings in tasks entailing symbolic data [10, 11, 41, 55], as well as in a wide variety of computer vision problems [1, 3, 22, 37, 53], showing their benefits for problems with underlying hierarchical structures.

Causal Representation Learning. One of the precursors of causal representation learning (CRL) is Independent Component Analysis (ICA) [6, 17], which is a method that seeks to recover independent variables that were measured together through a linear invertible transformation. The impossibility of identifying the sources in the general non-linear case [19] prompted research in using additional auxiliary variables [15, 16, 18]. Non-linear ICA has been further extended to deep learning architectures [20, 25]. Several studies drew connections between ICA and causality, demonstrating how practices can be transferred from one field to another [14, 38, 44, 49].

Recently, there have been significant advances in CRL, which lies in the intersection of causality and representation learning. CITRIS [32, 33], which our method builds upon, uses observed intervention targets to identify scalar as well as multidimensional causal factors from temporal high-dimensional data. BISCUIT [31] replaces the intervention targets, which require substantial supervision, with unobserved binary interaction variables. Also, using counterfactual observations where an unknown subset of variables has been intervened is a common practice for learning causal representations [2, 5, 35, 56]. Other works on temporal data provide identifiability results by employing mechanism sparsity regularization and modeling interventions as external actions [28, 29], or enforcing the independent noise condition and modeling soft-interventions as non-stationary noise [58, 59].

3 Preliminaries

We adopt the TempoRal Intervened Sequences (TRIS) setup that is used in CITRIS [33]. We assume that the underlying causal process that generates the data is modeled by a Dynamic Bayesian Network (DBN) [7, 39] with K causal factors instantiated at each time step t as $C^t = (C_1^t, C_2^t, ..., C_K^t)$. Each causal factor $C_i \in D_i^{M_i}$ can be potentially multidimensional with D_i denoting the domain, $D_i^{M_i} \subseteq \mathbb{R}^{M_i}$ for continuous variables, $D_i^{M_i} \subseteq \mathbb{Z}^{M_i}$ for discrete variables, and M_i indicating its dimensionality. Thereby, the causal factors' space is defined as $\mathcal{C} = D_1^{M_1} \times D_2^{M_2} \times ...D_K^{M_K}$. We assume there are no instantaneous effects, i.e., causal relations between variables of the same time-step, and that the DBN is first-order Markov and stationary, so the only causal relations can happen between two adjacent time steps and repeat across all pairs of timesteps.

We furthermore consider that there can be *soft interventions* on each causal factor C_i in the system at each timestep t, which can perturb its relation to its parents. We also assume the knowledge of the intervention target vector at each step $I^t \in \{0, 1\}^K$, where $I_i^t = 1$ indicates that C_i^t was intervened, while $I_i^t = 0$ that was not. We model potential dependencies between the interventions through a latent variable R^t. At each time step, we observe a high-dimensional observation $X^t = h(C_1^t, C_2^t, ..., C_K^t, E^t)$, where E^t being an exogenous variable modeling any noise. The observation function $h : \mathcal{C} \times \mathcal{E} \to \mathcal{X}$ that maps the causal factor space $\mathcal{C}$ and the noise space $\mathcal{E}$ to the observation space $\mathcal{X}$ is required to be bijective, allowing us to uniquely identify each causal factor from an observation. In this setting, CITRIS [33] is able to identify the *minimal causal variables*, which are the parts of each (potentially multidimensional) causal variable C_i that are *manipulable*, i.e., affected by the corresponding intervention variable I_i given the values of the causal variables in the previous timestep, under the assumption that each intervention target I_i is not a deterministic function of any other intervention target $I_j, i \neq j$.

We will assume for simplicity that all of the causal variables that we will consider at different levels are completely affected by their corresponding intervention variable, so in this case, minimal causal variables and causal variables will be the same. Additionally, we will leverage the potentially multidimensional causal variables to model different granularities of causal representations, allowing us to refine them as more fine-grained interventions are available in the environment.

4 Hierarchical Causal Representation Learning

For learning hierarchical causal representations, we propose HERCULES (HiERarchical CaUsaL reprESentations), a framework that gradually identifies more fine-grained causal variables by leveraging increasingly refined interactions. In this paper, we base HERCULES on the recent CRL method CITRIS [33], but the approach is general and it can be applied to other CRL methods as well.

Autoencoder. Initially, we consider a pre-trained AE that models the invertible function $g_\theta : \mathcal{X} \to \mathcal{Z}$ with $\mathcal{Z} \subseteq \mathbb{R}^M$ denoting the latent space and M its dimensionality. Specifically, the encoder g_θ transforms an image x into a vector z of fewer dimensions:

$$g_\theta(x) = z \tag{1}$$

and then tries to reconstruct the original image x using the low-dimensional vector z and a decoder that models the inverse of the encoder. Once the AE model is trained, its parameters remain frozen, and we employ a series of normalizing flows (NFs) [45] to hierarchically identify the causal factors, leading to a hierarchical causal structure of L levels.

Joint Interventions. We consider at each time point t that a causal variable C_i^t may have been intervened. To enable hierarchical causal representation learning, we assume that certain causal

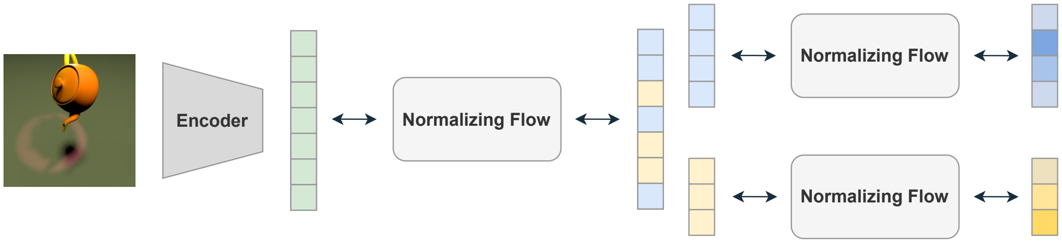


Figure 1: Overview of HERCULES. In the first step, we encode a high observational sample, e.g. an image, to obtain an entangled representation z (green). We then learn an NF to identify causal variables that are jointly intervened. In this example, two groups were considered (blue and yellow). Finally, for each block of variables, a distinct normalizing flow further disentangles the group, identifying the causal variables at an atomic level (shades of blue and yellow).

variables are consistently jointly intervened at the hierarchical level l, leading to the formation of $K_l \leq K$ groups representing the jointly intervened variables. The intervention targets at time step t and level l are denoted as $I_l^t \in \{0,1\}^{K_l}$, where $I_{l,i}^t = 1$ indicates that the group $i \in [\![1..K_l]\!]$ has been intervened at step t, while $I_{l,i}^t = 0$ that it has not been intervened upon. To allow HERCULES to ultimately attain a fully disentangled representation, progressing within the hierarchical levels, we leverage more fine-grained interventions. Note that when a group pertains to a single causal variable, HERCULES is able to individually identify it, as in CITRIS. In the case of consistently joint interventions, HERCULES identifies an entangled version of the respective causal variables.

Dataset. For each level l within the hierarchical structure, we consider a distinct training dataset D_l comprising tuples $\{x^t, x^{t+1}, I_l^{t+1}\}$ with x^t and x^{t+1} denoting the high-dimensional observations at time step t and $t+1$ obtained by the observation function. I_l^{t+1} denotes the intervention targets on the groups of causal variables at level l. Each subsequent dataset D_{l+1} contains more fine-grained intervention targets, enforcing some disentanglement in the next stage. The final dataset D_{L-1} contains independent interventions of the remaining individually unidentified causal factors, enabling their identification in the last hierarchical level.

HERCULES. HERCULES models a hierarchical causal structure through a series of L levels, with each latent space denoted as $\{\mathcal{Z}_l\}_{l=0}^{L-1}$. At each level l, a distinct invertible function $f_l : \mathcal{Z}_{l-1} \to \mathcal{Z}_l$ maps the representation $z_{l-1} \in \mathcal{Z}_{l-1}$ of level $l-1$ with the next level's representation $z_l \in \mathcal{Z}_l$, i.e.,

$$f_l(z_{l-1}) = z_l \tag{2}$$

Each f_l comprises of K_{l-1} models, denoted as $f_{l,i}$, disentangling each group i within level $l-1$:

$$f_l(z_{l-1}) = \begin{cases} f_{l,i}(z_{l-1,i}) & \text{if group } i \text{ has more than one causal variable} \\ z_{l-1,i} & \text{if group } i \text{ has exactly one causal variable} \end{cases} \tag{3}$$

with $z_{l,i} = \{(z_l)_j \mid j \in [\![1..M]\!], \psi_l(j) = i\}$ referring to the set of latent variables at level l that an assignment function $\psi_l : [\![1..M]\!] \to [\![0..K_l]\!]$ assigned to group i. In addition, we use $\psi_l(j) = 0$ to denote any latent dimension that does not belong to a causal variable or group. In cases where $z_{l-1,i}$ pertains to a single causal variable, $f_{l,i}$ is considered to be the identity function, as the causal variable has already been disentangled from the rest. As the initial level, we consider the entangled representation provided by the AE, $\mathcal{Z}_0 = \mathcal{Z}$. The invertible functions f_l require the same input and output dimensionality, thereby each space in the hierarchical structure follows the same dimensions, $\forall l \in [\![0..L-1]\!] \ : \mathcal{Z}_l \subseteq \mathbb{R}^M$. Each subsequent level within the hierarchical structure further disentangles the previous representation, gradually identifying more fine-grained causal variables.

Objective Function. For each hierarchical level l, HERCULES learns:

1. an invertible function $f_l : \mathcal{Z}_{l-1} \to \mathcal{Z}_l$, mapping a hierarchy's latent space to the next, identifying K_l blocks of causal variables;
2. an assignment function $\psi_l : [\![1..M]\!] \to [\![0..K_l]\!]$, mapping the dimensions of the latent space $\mathcal{Z}_l$ to one of the causal factors or groups present in the hierarchical level.

The hierarchical levels have to be attained in a sequential order, with the preceding levels remaining frozen throughout the remainder of the training process. This is crucial as the employed models require fixed input sizes, while at the same time it enhances the overall training stability.

The disentanglement is enforced by the *transition prior*, which conditions each latent variable precisely on one of the intervention targets. Let the invertible composite function $g_{\theta_l} : \mathcal{X} \to \mathcal{Z}_l$ map a high-dimensional image to a hierarchical level:

$$g_{\theta_l} = f_l \circ f_{l-1} \circ ... \circ f_1 \circ g_\theta \tag{4}$$

For each hierarchical level l, we model the prior $p_{\phi_l}(z_l^{t+1} \mid z_l^t, I_l^{t+1})$ with $z_l^t, z_l^{t+1} \in \mathcal{Z}_l, z_l^t = g_{\theta_l}(x^t)$ and $z_l^{t+1} = g_{\theta_l}(x^{t+1})$. To impose the disentanglement of the latent space, the prior is factorized as:

$$p_{\phi_l}\left(z_l^{t+1} \mid z_l^t, I_l^{t+1}\right) = \prod_{i=0}^{K_l} p_{\phi_l}\left(z_{l,i}^{t+1} \mid z_l^t, I_{l,i}^{t+1}\right) \tag{5}$$

with $I_{l,i}^{t+1}$ referring to whether the group i at level l was intervened or not, and $I_{l,0}^{t+1} = 0$ since an extra variable was used to capture any noise. Finally, the objective for each hierarchical level is to maximize the likelihood:

$$p_{\phi_l,\theta_l}(x^{t+1}|x^t, I_l^{t+1}) = \left|\frac{\partial g_{\theta_l}(x^{t+1})}{\partial x^{t+1}}\right| p_{\phi_l}(z_l^{t+1} \mid z_l^t, I_l^{t+1}) \tag{6}$$

In the case of a single hierarchical level and infinite fine-grained interventional data, HERCULES reverts to CITRIS and thus shares the same identifiability results. We conjecture that HERCULES provides the same identifiability results as CITRIS also for multiple hierarchies.

5 Experiments

To carry out the experimental analysis and evaluate the effectiveness of HERCULES, we use the Temporal Causal3DIdent [56, 60] and the Voronoi [33] dataset. The former provides a challenging benchmark due to the presence of various causal relations and the high-dimensionality of the observational space. The latter offers great flexibility in modeling different relation types, facilitating a thorough and rigorous assessment of the method. To quantify HERCULES' performance, we use two correlation metrics, the R^2 coefficient of determination [57] and the Spearman's rank coefficient [51], measuring the *correlation* between the learned latent variables assigned to a causal factor and the corresponding ground truth value of it. Also, we quantify the reconstruction fidelity by combining causal factors of different images, using *triplet evaluation* [33]. As our baselines, we use the VAE models iVAE [21], SlowVAE [26], DMSVAE [29], and LEAP [59], as well as the NF variant of CITRIS [33]. The same intervention targets are provided as the auxiliary variable for all methods. For HERCULES, we mainly focus on two hierarchical levels. In the first, we block identify the groups [pos_x, pos_y, pos_z], [rot_α, rot_β], [rot_s, hue_s, hue_b, hue_o] and [obj_s] for the Temporal Causal3DIdent dataset as well as [c_0, c_1, c_2] and [c_3, c_4, c_5] for the Voronoi dataset. In the second level, we individually identify each causal factor for both datasets.

Temporal Causal 3DIdent. For the Temporal Causal 3DIdent, we consider two variations, one with only the teapot shape and one entailing all shapes. In Table 1a, we provide the experimental results with both variants. First, we observe the NF-based models, i.e., CITRIS and HERCULES, significantly outperform the VAEs. This highlights the difference in the reconstruction fidelity between the AE that had negligible reconstruction error and the VAEs that were constrained by the strong priors embedded in the KL divergence term. Moreover, HERCULES provides slightly better correlation metrics, especially on the off-diagonals.

Data Efficiency. To investigate the data efficiency of the best-performing methods, we limit the fine-grained interventional data to 50% and 10% and trained CITRIS and the last hierarchical level of HERCULES solely on them. As noted in Table 1a, HERCULES significantly outperforms CITRIS, indicating that the previous hierarchical levels efficiently guided the disentanglement process. As another baseline, we train CITRIS on the coarse interventions and then finetune it on the same fine-grained ones, denoted as CITRIS-FT. Although we notice increased performance compared to CITRIS, CITRIS-FT is still unable to surpass HERCULES, strengthening our previous results that a hierarchical causal structure compensates in the case of limited fine-grained interactions.

Table 1: Experimental results with (a) the Temporal Causal3DIdent and (b) the Voronoi dataset. The results are averaged over 5 different seeds (3 for the 7 shapes variant). The triplet evaluation distance is in the range of [0, 1] and refers to the average distance of all causal variables (optimal value 0). The diag and sep in correlation metrics denote the average of the diagonals (optimal value 1) and off-diagonals (optimal value 0) respectively in the correlation matrices.

(a) Temporal Causal 3DIdent

	Dist. ↓	R^2 diag ↑	R^2 sep ↓	Sp. diag ↑	Sp. sep ↓
Baselines					
Temporal Causal3DIdent 7 Shapes					
SlowVAE	0.44	0.22	0.10	0.30	0.22
iVAE	0.12	0.82	0.08	0.83	0.12
DMSVAE	0.14	0.78	0.09	0.79	0.18
LEAP	0.14	0.78	0.12	0.80	0.21
CITRIS	**0.10**	**0.95**	0.07	0.95	0.13
HERCULES	0.11	0.94	**0.04**	**0.96**	**0.10**
Temporal Causal3DIdent Teapot					
SlowVAE	0.42	0.26	0.14	0.33	0.23
iVAE	0.22	0.70	0.06	0.74	**0.11**
DMSVAE	0.20	0.70	0.08	0.74	0.16
LEAP	0.21	0.70	0.09	0.74	0.19
CITRIS	**0.08**	0.93	0.08	0.94	0.13
HERCULES	**0.08**	**0.94**	**0.04**	**0.96**	**0.11**
Data Efficiency					
Temporal-Causal3DIdent Teapot (50%)					
CITRIS	0.10	0.91	0.10	0.91	0.15
HERCULES	**0.09**	**0.93**	**0.04**	**0.95**	**0.11**
Temporal-Causal3DIdent Teapot (10%)					
CITRIS	0.41	0.36	0.24	0.44	0.38
CITRIS-FT	0.22	0.76	0.15	0.78	0.24
HERCULES	**0.15**	**0.86**	**0.10**	**0.87**	**0.14**
Environment Generalization					
Temporal-Causal3DIdent Teapot (rotating xy-plane by 20 degrees)					
HERCULES	0.06	0.97	0.04	0.98	0.10
Temporal-Causal3DIdent Teapot (applying shear on the xy-plane)					
HERCULES	0.08	0.94	0.04	0.95	0.11

(b) Voronoi benchmark

	Dist. ↓	R^2 diag ↑	R^2 sep ↓	Sp. diag ↑	Sp. sep ↓
Baselines					
Voronoi - 6 variables (Graph 1)					
CITRIS	**0.10**	**0.91**	0.10	**0.96**	0.26
HERCULES	0.11	0.86	**0.09**	0.93	**0.25**
Voronoi - 6 variables (Graph 2)					
CITRIS	0.15	0.83	0.17	0.91	0.36
HERCULES	**0.09**	**0.92**	**0.04**	**0.96**	**0.19**
Voronoi - 6 variables (Graph 3)					
CITRIS	0.12	0.86	0.14	0.91	0.32
HERCULES	**0.09**	**0.87**	**0.09**	**0.92**	**0.24**
False Positive Interventions					
Voronoi - 6 variables (Graph 1, 20% noisy targets)					
CITRIS	0.13	0.85	0.14	0.92	0.33
HERCULES	**0.11**	**0.86**	**0.09**	**0.93**	**0.24**
Voronoi - 6 variables (Graph 1, 50% noisy targets)					
CITRIS	0.15	0.78	0.12	0.88	0.31
HERCULES	**0.11**	**0.86**	**0.08**	**0.93**	**0.24**
Voronoi - 6 variables (Graph 1, 99% noisy targets)					
CITRIS	0.17	0.67	0.14	0.80	0.33
HERCULES	**0.11**	**0.86**	**0.08**	**0.93**	**0.23**
Ablation Study: 3 Hierarchical Levels					
Voronoi - 6 variables					
HERCULES (Graph 1)	0.08	0.89	0.08	0.94	0.24
HERCULES (Graph 2)	0.06	0.93	0.04	0.96	0.16
HERCULES (Graph 3)	0.06	0.90	0.08	0.95	0.23
Ablation Study: Grouping					
Voronoi - 6 variables (Graph 1)					
HERCULES 3-3	0.11	0.86	0.09	0.93	0.25
HERCULES Interleaving	0.09	0.90	0.07	0.95	0.21
HERCULES 5-1	**0.07**	**0.92**	**0.06**	**0.96**	**0.20**

Environment Generalization. We evaluate the adaptability of the first level's representations, which capture the abstract variables of position, rotation and color, for identifying similar, albeit not identical, causal factors in the next level. We employ the first level which is trained on the regular xy-plane and subsequently train the second level on a transformed plane. The applied transformation influences how the causal factors `pos_x` and `pos_y` affect the object's position in the observational space $\mathcal{X}$. The results in Table 1a show that under a rotation of 20 degrees and a shear transformation with $\lambda = 0.1$, HERCULES is still able to attain the same performance as in the regular plane, indicating its adaptation capabilities.

Voronoi. In Table 1b, we report the experimental results with 3 different Voronoi datasets, generated using distinct graphs with 6 variables. HERCULES consistently exhibits lower off-diagonal values compared to its non-hierarchical counterpart, CITRIS. Moreover, it shows robustness when a percentage of the intervention targets have *one* false positive intervention element, i.e., converting an entry from 0 to 1. HERCULES accommodates two or potentially more hierarchical levels without compromising its performance. Finally, we evaluate HERCULES when different joint interventions are used for training the initial hierarchical level and obtain similar levels of performance.

6 Conclusion

In this paper, we introduced HERCULES, a hierarchical CRL approach that leverages coarse interventions to build a hierarchical causal structure. HERCULES is able to leverage coarse-grained representations and perform well even when fine-grained interactions are scarce, outperforming SOTA methods. In the early levels of the hierarchy, it identifies groups of causal variables and then refines their representations to further disentangle them. HERCULES exhibits considerable flexibility in its structure, accommodating varying numbers of levels and capturing diverse causal variables.

Hierarchical causal structures hold promise in the field of reinforcement learning, where agents learn to interact with a system. In this context, a coarse disentanglement could facilitate in identifying actions for individual variables, essentially guiding the skill-learning process. Moreover, HERCULES' framework can be extended to various CRL methods, such as BISCUIT [31], while the results on the representations' adaptability pave the way for the employment of pre-trained causal models. Finally, since hyperbolic spaces befit capturing hierarchical structures [37], incorporating hyperbolic geometry could potentially enhance the overall outcomes of our method.

Acknowledgements

We thank SURF for the support in using the National Supercomputer Snellius. This work is financially supported by Qualcomm Technologies Inc., the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy. SM was supported by the Air Force Office of Scientific Research under award number FA8655-22-1-7155. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

[1] O. Ahmad and F. Lecue. FisheyeHDK: Hyperbolic deformable kernel learning for ultra-wide field-of-view image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5968–5975, 2022.

[2] K. Ahuja, J. S. Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022.

[3] M. G. Atigh, J. Schoep, E. Acar, N. van Noord, and P. Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4453–4462, June 2022.

[4] L. J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[5] J. Brehmer, P. De Haan, P. Lippe, and T. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

[6] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[7] P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In *Uncertainty in artificial intelligence*, pages 41–48. Elsevier, 1992.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[9] W. Falcon and The PyTorch Lightning team. PyTorch Lightning, Mar. 2019.

[10] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018.

[11] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

[12] M. Germain, K. Gregor, I. Murray, and H. Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015.

[13] P. Goyal, Z. Hu, X. Liang, C. Wang, and E. P. Xing. Nonparametric variational auto-encoders for hierarchical representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5094–5102, 2017.

[14] L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

[15] A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[16] A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[17] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[18] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[19] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[20] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[21] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[22] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6418–6428, 2020.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[24] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[25] D. A. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

[26] D. A. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

[27] A. Klushyn, N. Chen, R. Kurle, B. Cseke, and P. van der Smagt. Learning hierarchical priors in vaes. *Advances in neural information processing systems*, 32, 2019.

[28] S. Lachapelle and S. Lacoste-Julien. Partial Disentanglement via Mechanism Sparsity. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[29] S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

[30] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Intervention Design for Causal Representation Learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[31] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. BISCUIT: Causal representation learning from binary interactions. In R. J. Evans and I. Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1263–1273. PMLR, 31 Jul–04 Aug 2023.

[32] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In *International Conference on Learning Representations*, 2023.

[33] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and S. Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[35] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[36] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[37] P. Mettes, M. G. Atigh, M. Keller-Ressel, J. Gu, and S. Yeung. Hyperbolic deep learning in computer vision: A survey, 2023.

[38] R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

[39] K. P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

[40] M. Nickel, X. Jiang, and V. Tresp. Reducing the rank in relational factorization models by including observable patterns. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[41] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[43] R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International conference on machine learning*, pages 324–333. PMLR, 2016.

[44] P. Reizinger, L. Gresele, J. Brady, J. von Kügelgen, D. Zietlow, B. Schölkopf, G. Martius, W. Brendel, and M. Besserve. Embrace the Gap: VAEs Perform Independent Mechanism Analysis. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022.

[45] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

[46] D. J. Rezende and F. Viola. Taming vaes. *CoRR*, 2018.

[47] F. Sala, C. De Sa, A. Gu, and C. Re. Representation tradeoffs for hyperbolic embeddings. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4460–4469. PMLR, 10–15 Jul 2018.

[48] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[49] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[50] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

[51] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[52] A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[53] M. van Spengler, E. Berkhout, and P. Mettes. Poincaré resnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5419–5428, October 2023.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[55] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[56] J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[57] S. Wright. Correlation and causation. *Journal of agricultural research*, 20:557–580, 1921.

[58] W. Yao, G. Chen, and K. Zhang. Temporally Disentangled Representation Learning. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022.

[59] W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022.

[60] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

Appendix

A Datasets

In this section, we describe the datasets used for the experimental evaluation. For the Temporal-Causal3DIdent with the teapot and the Voronoi datasets, we rendered 100,000 images for training, while the test dataset comprised 8,000 images which were also used to produce 8,000 triplet test samples. For the variant with all 7 shapes, we increased the training size to 250,000 images.

A.1 Temporal-Causal3DIdent

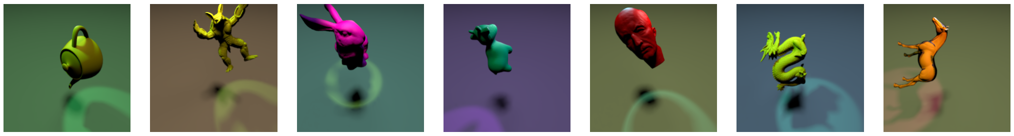
The Temporal Causal3DIdent dataset poses significant challenges because of the high-dimensional observations and the variety of relations between the causal factors. The dataset consists of 3D objects projected in 2D under varying conditions. The underlying system generating the images consists of ten causal factors: three for the object's spatial position $[\texttt{pos\_x, pos\_y, pos\_z}] \in [-2, 2]^3$, two for the object's rotation $[\texttt{rot\_}\alpha\texttt{, rot\_}\beta] \in [0, 2\pi)^2$, the hue of the object, background and spotlight $[\texttt{hue\_o, hue\_b, hue\_s}] \in [0, 2\pi)^3$, the spotlight's rotation $\texttt{rot\_s} \in [0, 2\pi)$ and finally, the object's shape, denoted as `obj_s` with seven possible categorical values. All continuous variables follow a Gaussian distribution with the mean value provided by a non-linear combination of their parents. An instance of each available shape combined with different values of the other variables is illustrated in Figure 2. The dataset contains a diverse range of causal relations between these variables including chains of causal relations and variables with up to four causal parents. Figure 3 provides an extensive visual depiction of the relations.

The dataset generation process starts by initially assigning a random value to each of the causal factors and then rendering an image. Following the causal relations depicted in Figure 3, for each subsequent time step, we sample new values and create a new image. Additionally, during each step, we consider that an intervention may be performed to a causal factor, replacing its value using a uniform distribution. Particularly, for the rotation and hue variables, the distribution $U(0, 2\pi)$ is used to sample the intervention values, while for the position variables we use the distribution $U(-2, 2)$. For the shape, one of the categorical values is uniformly sampled. At each level of the hierarchy, the decision to perform an intervention is determined by sampling from a discrete uniform distribution, which selects one of the groups within the level or no group at all. All the causal variables in the selected group are then intervened. For the rest of the baselines, the intervention targets for each causal factor are sampled from a Bernoulli distribution, Bernoulli$(0, 1)$. The generated images have a resolution of 64×64, however, for visualization purposes, the images reported have a higher resolution of 256×256. Sequential samples from the dataset are presented in Figure 4.

A.2 Voronoi

Voronoi is a synthetic dataset generated by a flexible system that can support any causal graph with an arbitrary number of variables. For each generated graph, an edge is created for every pair of variables in C^t and C^{t+1} with a probability of 0.25. Then, a randomly initialized neural network is employed to model these relations. Specifically, the model takes as input a subset of the causal factors C^t, which are the parents of C_i according to the causal graph, to parameterize a Gaussian distribution for C_i. Similarly to the Temporal Causal3DIdent, at each time step, we may perform an intervention, replacing the value of the intervened causal variable using a Gaussian distribution. The intervention targets follow a uniform distribution, sampling one of the grouped causal variables or none at all. The causal variables of the sampled group are then intervened. For the baselines, we perform an intervention upon each causal variable with a probability of 0.15.

After computing the causal variables for each time step, we proceed to render them into a high observational space. The mapping process, initially, involves the employment of a two-layer NF to entangle the causal factors. Then, an image is generated with a resolution of 32×32 where the causal variables are depicted as colors in a fixed Voronoi diagram. Figure 5 illustrates samples from the dataset, showcasing the diagrams for 6 causal variables. The 3 temporal causal graphs used to generate the data are depicted in Figure 6. For visualization purposes, the displayed images are at a higher resolution of 320×320.



(a) Teapot (b) Armadillo (c) Bunny (d) Cow (e) Head (f) Dragon (g) Horse

Figure 2: An example image for all 7 object shapes used in the Temporal-Causal3DIdent dataset.

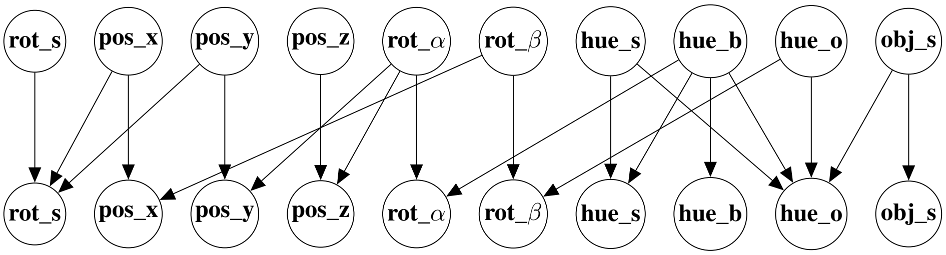


Figure 3: The relations between the causal variables in the Temporal Causal3DIdent dataset. The arrows indicate the temporal relations between the causal variable C_i^t and C_j^{t+1}. [33]

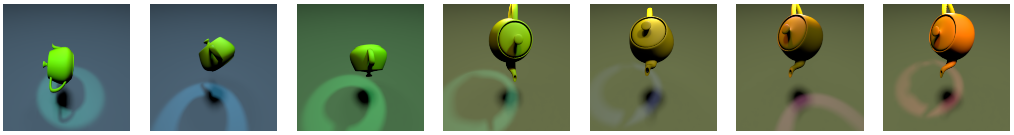


Figure 4: Sequential samples from the Temporal Causal3DIdent dataset, where at each time-step each causal variable may have been intervened.

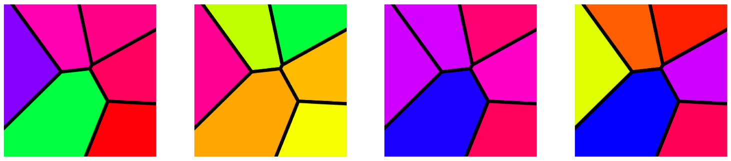
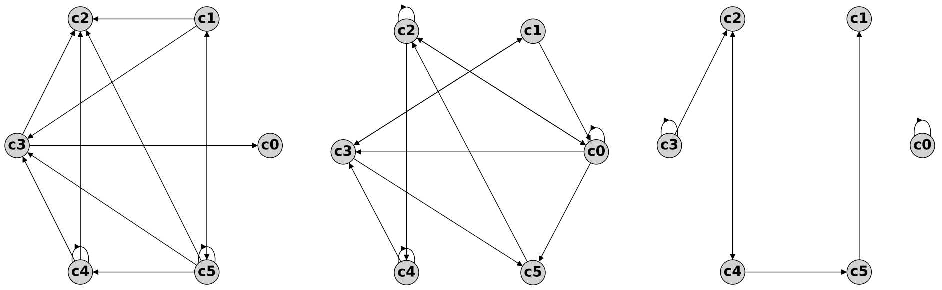


Figure 5: Sequential samples from the Voronoi dataset with 6 causal variables.



(a) Temporal Graph 1 (b) Temporal Graph 2 (c) Temporal Graph 3

Figure 6: The summary graphs modeling the temporal causal relations in the Voronoi datasets. An arrow denotes a relation between C_i^t and C_j^{t+1}.

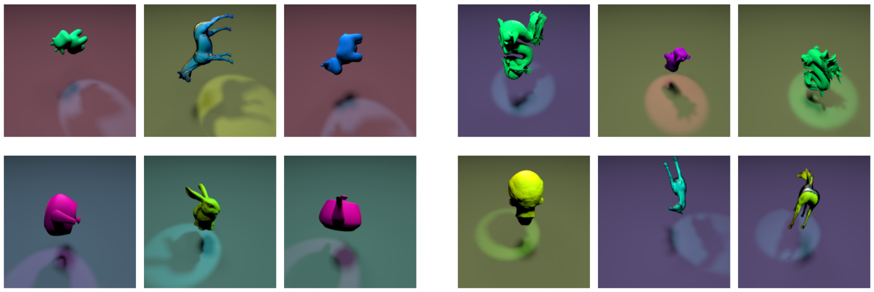


Figure 7: Triplet examples used for triplet evaluation. In each triplet, a combination of causal variables of the first two images is used to generate the third image. For example, in the top left triplet, we use the shape (cow) from the first image and the object's color (blue) from the second one.

B Metrics

In this section, we describe the evaluation metrics used for HERCULES and the baselines.

Correlation Metrics To evaluate the proposed method we measure the correlation between the learned latent variables and the ground truth causal variables. Since the latent variables describing causal variables can be multidimensional, we learn a mapping between the two by employing a Multilayer Perceptron (MLP). For each set of latent variables that are assigned to the same causal variable, a distinct MLP is trained. The evaluation phase starts after the completion of HERCULES' training process, with its parameters remaining frozen. In this way, no error is backpropagated and gradients are not calculated to further update the model's weights. The MLPs are trained using the mean squared error (MSE) loss for continuous variables, apart from circular values. Predicting the value $2\pi - \epsilon$ with the ground truth being 0, MSE would falsely yield a high error. Therefore, to address this issue, for such variables, we predict a vector instead and calculate the cosine distance from the ground truth angle projected onto the unit circle. Finally, for categorical causal variables, the cross entropy loss is used. Once the MLPs are trained, their predictions are utilized to measure their correlation with the corresponding causal variables. We use two correlation metrics, the R^2 coefficient of determination [57] and the Spearman's rank coefficient [51].

Triplet Evaluation The evaluation of HERCULES also incorporates the use of *triplet evaluation*, which quantifies the fidelity of the reconstruction based on a random combination of causal variables from two different images. Specifically, given two randomly sampled images from the test set, a third image is generated by combining different causal variables originating from the two original images. Initially, the two images are independently encoded, and then the latent variable i is extracted from the first image, if it has been assigned to a causal variable linked to that image. Otherwise, the variable i is obtained from the second image. This process is repeated for all latent variables, resulting in the formation of a new latent space. Then, the inverse of the NFs is applied, followed by the decoder to generate a new image. This concept is visually depicted in Figure 7 using the Temporal Causal3DIdent dataset where the ground truth image is the newly generated one. In the top-left triplet, it can be observed that the shape component, specifically the depiction of a cow, was extracted from the first image, while the object's color, which is blue, was derived from the second one. Consequently, the outcome of merging these two components yields a blue cow.

Naively relying on the measurement of the reconstruction error fails to offer a comprehensive analysis of the method's performance, as potential inaccuracies in certain causal variables may yield high errors, while inaccuracies in others may relatively be overlooked. For example, in the Temporal-Causal3DIdent dataset, the background color would have a substantial impact on the reconstruction loss, while the rotation of specific shapes may have negligible influence. Instead, a more descriptive metric is employed, involving the use of an additional CNN that predicts the values of the causal variables based on the input images. The distance between the predictions and the ground truth values can be subsequently quantified for each individual causal variable, enabling a more precise analysis. The employed CNN model had overall minimal prediction errors with both datasets.

Table 2: The architecture of the encoder and the decoder. The same architecture was used for all the VAE models. The residual blocks in the decoder consist of 2 convolutions with 64 channels, kernel of size 3 and stride 1. The upsampling denotes bilinear upsampling with a scaling factor of 2.

	Layer	Feature Dimension (H × W × C)	Kernel	Stride	Activation Function
f_{enc}	Conv	$32 \times 32 \times 64$	3	2	BatchNorm+SiLU
	Conv	$32 \times 32 \times 64$	3	1	BatchNorm+SiLU
	Conv	$16 \times 16 \times 64$	3	2	BatchNorm+SiLU
	Conv	$16 \times 16 \times 64$	3	1	BatchNorm+SiLU
	Conv	$8 \times 8 \times 64$	3	2	BatchNorm+SiLU
	Conv	$8 \times 8 \times 64$	3	1	BatchNorm+SiLU
	Conv	$4 \times 4 \times 64$	3	2	BatchNorm+SiLU
	Conv	$4 \times 4 \times 64$	3	1	BatchNorm+SiLU
	Reshape	$1 \times 1 \times 1024$	-	-	-
	Linear	$1 \times 1 \times 256$	-	-	LayerNorm+SiLU
	Linear	$1 \times 1 \times 2$·num_latents	-	-	-
f_{dec}	Linear	$1 \times 1 \times 256$	-	-	LayerNorm+SiLU
	Linear	$1 \times 1 \times 1024$	-	-	-
	Reshape	$4 \times 4 \times 64$	-	-	-
	Upsample	$8 \times 8 \times 64$	-	-	-
	ResidualBlock	$8 \times 8 \times 64$	3	1	-
	Upsample	$16 \times 16 \times 64$	-	-	-
	ResidualBlock	$16 \times 16 \times 64$	3	1	-
	Upsample	$32 \times 32 \times 64$	-	-	-
	ResidualBlock	$32 \times 32 \times 64$	3	1	-
	Upsample	$64 \times 64 \times 64$	-	-	-
	ResidualBlock	$64 \times 64 \times 64$	3	1	-
	Pre-Activation	$64 \times 64 \times 64$	-	-	BatchNorm+SiLU
	Conv	$64 \times 64 \times 64$	1	1	BatchNorm+SiLU
	Conv	$64 \times 64 \times 3$	1	1	Tanh

C Architectures

In Table 2, we report the architecture of the encoder and the decoder used for the experiments. The VAE models shared the same design to ensure a fair comparison. For CITRIS and HERCULES, we used the same models for the Voronoi dataset while we increased the residual blocks by a factor of 2 per resolution for the experiment conducted with the Temporal Causal3DIdent dataset. Increasing the complexity of the VAE's decoder did not have a substantial influence. The prior distribution was modeled by an autoregressive model, following a MADE architecture [12] with 2 layers and SiLU activation functions. For each latent variable, 16 neurons were assigned per layer. For the normalizing flows, an affine autoregressive flow was employed. The autoregressive model followed the same MADE architecture. Between each coupling layer, activation normalization and invertive 1×1 convolutions were used. For the target classifier, we used an MLP consisting of a single layer and 128 neurons with Layer Normalization [4] and SiLU activation functions. All models were developed using the deep learning framework PyTorch [42] and PyTorch Lightning [9].

Table 3: An overview of the hyperparameter used for the VAE models. This configuration resulted in approximately 20 minutes, 1 hour and 5 hours of training on an NVIDIA A100-SXM4-40GB with the Voronoi, the teapot and the 7 shapes variant of the Temporal Causal3DIdent dataset respectively.

Hyperparameter	**Value**
Batch size	512
Optimizer	Adam [23]
Learning rate	1e-3
Learning rate scheduler	Cosine Warmup (100 steps)
KL divergence factor β	1.0
Number of latents	32
Number of epochs	200 (Voronoi, teapot), 1000 (7 shapes)
Target classifier weight	2.0
Gumbel Softmax temperature	1.0

Table 4: An overview of the hyperparameter used for the AE model. This configuration resulted in approximately 1 hour, 2.5 hours and 30 hours of training on an NVIDIA A100-SXM4-40GB with the Voronoi, the teapot and the 7 shapes variant of the Temporal Causal3DIdent dataset respectively.

Hyperparameter	**Value**
Batch size	512
Optimizer	Adam [23]
Learning rate	1e-3
Learning rate scheduler	Cosine Warmup (100 steps)
Number of latents	32
Number of epochs	200 (Voronoi, teapot), 1000 (7 shapes)
Gaussian noise std	0.05

Table 5: An overview of the hyperparameter used for the CITRIS and HERCULES models. In the case of HERCULES, the hyperparameters apply for each hierarchical level's model. This configuration with CITRIS resulted in approximately 20 minutes, 30 minutes and 3.5 hours of training on an NVIDIA A100-SXM4-40GB with the Voronoi, the teapot and the 7 shapes variant of the Temporal Causal3DIdent dataset respectively. HERCULES required approximately double the training time, since we sequentially trained two hierarchical levels.

Hyperparameter	**Value**
Batch size	512, 64 (setting with 10%)
Optimizer	Adam [23]
Learning rate	1e-3
Learning rate scheduler	Cosine Warmup (100 steps)
Number of latents	32
Number of epochs	200 (Voronoi, teapot), 1000 (7 shapes)
Number of coupling layers	4 (Voronoi, teapot), 6 (7 shapes)
Target classifier weight	2.0
Gumbel Softmax temperature	1.0

D Training Hyperparameters

For all the experiments, we used the Adam optimizer [23] with the learning rate being set to 0.001. For the target classifier, we used the AdamW variant [36] with a learning rate of 0.004 and weight decay 0.0001. Also, a cosine warmup of 100 steps was employed. The batch size was set to 512 for all the experiments, with the exception of the setting with only 10% of the fine-grained dataset where the batch size was 64. For the Voronoi and the teapot dataset, all the models were trained for 200 epochs while for the Temporal Causal3DIdent dataset that considered all 7 shapes the epochs were increased to 1000. The latent spaces were composed of 32 latent variables. A summary of the hyperparameters is provided in Table 3, 4 and 5.

Table 6: Overview of the experimental results on the Temporal Causal3DIdent dataset considering only the teapot shape. The results are averaged over 5 different seeds.

	Triplet evaluation distances ↓										**Correlation metrics**			
	`pos_x`	`pos_y`	`pos_z`	rot_α	rot_β	`rot_s`	`hue_s`	`hue_b`	`hue_o`	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Baselines														
Temporal Causal3DIdent Teapot														
SlowVAE	0.27 ± 0.050	0.26 ± 0.040	0.31 ± 0.025	0.49 ± 0.004	0.50 ± 0.002	0.15 ± 0.029	0.76 ± 0.022	0.62 ± 0.019	0.65 ± 0.021	0.42 ± 0.024	0.26 ± 0.020	0.14 ± 0.016	0.33 ± 0.017	0.23 ± 0.026
iVAE	0.07 ± 0.002	0.04 ± 0.001	0.06 ± 0.002	0.43 ± 0.015	0.49 ± 0.002	0.03 ± 0.001	0.60 ± 0.041	0.02 ± 0.010	0.11 ± 0.011	0.22 ± 0.009	0.70 ± 0.009	0.06 ± 0.007	0.74 ± 0.020	**0.11** ± 0.023
DMSVAE	0.09 ± 0.004	0.05 ± 0.007	0.13 ± 0.004	0.47 ± 0.012	0.49 ± 0.003	0.05 ± 0.003	0.33 ± 0.224	0.00 ± 0.000	0.08 ± 0.017	0.20 ± 0.032	0.70 ± 0.037	0.08 ± 0.022	0.74 ± 0.043	0.16 ± 0.024
LEAP	0.12 ± 0.033	0.05 ± 0.006	0.14 ± 0.006	0.45 ± 0.027	0.49 ± 0.003	0.05 ± 0.002	0.40 ± 0.115	0.00 ± 0.001	0.10 ± 0.034	0.21 ± 0.024	0.70 ± 0.040	0.09 ± 0.021	0.74 ± 0.037	0.19 ± 0.029
CITRIS	0.04 ± 0.001	0.03 ± 0.001	0.04 ± 0.001	0.20 ± 0.014	0.28 ± 0.027	0.03 ± 0.002	0.06 ± 0.011	0.01 ± 0.006	0.03 ± 0.008	**0.08** ± 0.008	0.93 ± 0.005	0.08 ± 0.014	0.94 ± 0.006	0.13 ± 0.015
HERCULES	0.04 ± 0.002	0.03 ± 0.001	0.04 ± 0.002	0.15 ± 0.004	0.22 ± 0.018	0.04 ± 0.003	0.14 ± 0.020	0.03 ± 0.014	0.05 ± 0.005	**0.08** ± 0.008	**0.94** ± 0.004	**0.04** ± 0.004	**0.96** ± 0.003	**0.11** ± 0.008
Data Efficiency														
Temporal-Causal3DIdent Teapot (50%)														
CITRIS	0.05 ± 0.001	0.04 ± 0.001	0.05 ± 0.002	0.22 ± 0.009	0.30 ± 0.016	0.05 ± 0.037	0.11 ± 0.023	0.02 ± 0.016	0.06 ± 0.014	0.10 ± 0.013	0.91 ± 0.005	0.10 ± 0.031	0.91 ± 0.006	0.15 ± 0.013
HERCULES	0.05 ± 0.002	0.03 ± 0.002	0.04 ± 0.002	0.16 ± 0.005	0.22 ± 0.019	0.05 ± 0.004	0.19 ± 0.048	0.05 ± 0.023	0.06 ± 0.004	**0.09** ± 0.013	**0.93** ± 0.005	**0.04** ± 0.005	**0.95** ± 0.004	**0.11** ± 0.006
Temporal-Causal3DIdent Teapot (10%)														
CITRIS	0.38 ± 0.055	0.31 ± 0.041	0.32 ± 0.052	0.36 ± 0.012	0.45 ± 0.009	0.23 ± 0.033	0.65 ± 0.024	0.44 ± 0.060	0.55 ± 0.044	0.41 ± 0.036	0.36 ± 0.049	0.24 ± 0.029	0.44 ± 0.050	0.38 ± 0.026
CITRIS-FT	0.12 ± 0.023	0.10 ± 0.055	0.12 ± 0.055	0.25 ± 0.008	0.35 ± 0.026	0.08 ± 0.046	0.46 ± 0.052	0.32 ± 0.131	0.18 ± 0.079	0.22 ± 0.050	0.76 ± 0.067	0.15 ± 0.040	0.78 ± 0.066	0.24 ± 0.046
HERCULES	0.05 ± 0.002	0.03 ± 0.002	0.04 ± 0.002	0.21 ± 0.011	0.27 ± 0.024	0.07 ± 0.013	0.36 ± 0.130	0.18 ± 0.065	0.13 ± 0.023	**0.15** ± 0.031	**0.86** ± 0.020	**0.10** ± 0.022	**0.87** ± 0.015	**0.14** ± 0.012
Environment Generalization														
Temporal-Causal3DIdent Teapot (rotating `xy`-plane by 20 degrees)														
HERCULES	0.03 ± 0.002	0.02 ± 0.001	0.03 ± 0.002	0.11 ± 0.006	0.16 ± 0.021	0.03 ± 0.002	0.08 ± 0.019	0.02 ± 0.010	0.02 ± 0.003	0.06 ± 0.008	0.97 ± 0.003	0.04 ± 0.005	0.98 ± 0.003	0.10 ± 0.003
Temporal-Causal3DIdent Teapot (applying shear on the `xy`-plane)														
HERCULES	0.06 ± 0.001	0.03 ± 0.001	0.04 ± 0.002	0.14 ± 0.005	0.22 ± 0.016	0.04 ± 0.004	0.14 ± 0.018	0.03 ± 0.013	0.05 ± 0.007	0.08 ± 0.007	0.94 ± 0.004	0.04 ± 0.005	0.95 ± 0.004	0.11 ± 0.005

E Experimental Details

In this section, we provide additional details about some of the experiments. Also, in Table 6, 7 and 8, we report a more comprehensive analysis of our results, including the standard deviations and the individual distances for each causal variable.

False Positive Noise To evaluate the robustness of HERCULES to false positive noise, we converted one of the intervention targets' entries from 0 into 1. This modification was applied for a random subset comprising 20%, 50%, and 99% of the last level's intervention targets.

Grouping In the experiments with the Voronoi dataset, for the first hierarchical level, we mainly considered the 3-3 grouping where $[\texttt{c\_0}, \texttt{c\_1}, \texttt{c\_2}]$ and $[\texttt{c\_3}, \texttt{c\_4}, \texttt{c\_5}]$ comprised the groups containing the jointly intervened causal variables. To evaluate the performance of HERCULES when other joint interventions are present, we instead trained the first hierarchical level of HERCULES on the interleaving grouping, which is $[\texttt{c\_0}, \texttt{c\_2}, \texttt{c\_4}]$ and $[\texttt{c\_1}, \texttt{c\_3}, \texttt{c\_5}]$, and the 5-1 grouping comprising $[\texttt{c\_0}, \texttt{c\_1}, \texttt{c\_2}, \texttt{c\_3}, \texttt{c\_4}]$ and $[\texttt{c\_5}]$. In each case, in the first hierarchical level, the causal factors within the same group were jointly intervened while in the second level, we had atomic interventions.

3 Hierarchical Levels In our experiments, we focused on two hierarchical levels. However, in many environments, agents do not transition directly from coarse actions to highly precise interactions but instead perform actions of an intermediate granularity. Therefore, it is important for HERCULES to be able to leverage interactions of all levels of granularity by incorporating the respective number of hierarchical levels. We evaluated HERCULES with three hierarchical levels. The groups formed in the first stage were $[\texttt{c\_0}, \texttt{c\_1}, \texttt{c\_2}, \texttt{c\_3}, \texttt{c\_4}]$ and $[\texttt{c\_5}]$ while, in the second level, $\texttt{c\_4}$ was disentagled from the rest, resulting in the groups $[\texttt{c\_0}, \texttt{c\_1}, \texttt{c\_2}, \texttt{c\_3}]$ and $[\texttt{c\_4}]$. Finally, in the last level, every causal variable was individually identified.

Table 7: Overview of the experimental results on the Temporal Causal3DIdent dataset considering all 7 shapes. The results are averaged over 3 different seeds.

	Triplet evaluation distances ↓											**Correlation metrics**			
	pos_x	pos_y	pos_z	rot_α	rot_β	rot_s	hue_s	hue_b	hue_o	obj_s	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Baselines															
Temporal Causal3DIdent 7 Shapes															
SlowVAE	0.28	0.18	0.31	0.47	0.46	0.18	0.80	0.61	0.64	0.62	0.44	0.22	0.10	0.30	0.22
	± 0.022	± 0.015	± 0.039	± 0.001	± 0.003	± 0.019	± 0.012	± 0.049	± 0.013	± 0.001	± 0.019	± 0.002	± 0.004	± 0.013	± 0.015
iVAE	0.07	0.04	0.09	0.39	0.35	0.03	0.04	0.00	0.06	0.22	0.12	0.82	0.08	0.83	0.12
	± 0.002	± 0.002	± 0.004	± 0.021	± 0.013	± 0.001	± 0.008	± 0.000	± 0.030	± 0.046	± 0.009	± 0.017	± 0.011	± 0.007	± 0.011
DMSVAE	0.11	0.06	0.21	0.40	0.41	0.04	0.02	0.00	0.03	0.25	0.14	0.78	0.09	0.79	0.18
	± 0.004	± 0.004	± 0.011	± 0.025	± 0.007	± 0.001	± 0.006	± 0.000	± 0.002	± 0.036	± 0.007	± 0.025	± 0.008	± 0.019	± 0.011
LEAP	0.10	0.06	0.20	0.39	0.38	0.05	0.02	0.00	0.10	0.24	0.14	0.78	0.12	0.80	0.21
	± 0.007	± 0.005	± 0.001	± 0.007	± 0.011	± 0.002	± 0.003	± 0.000	± 0.139	± 0.041	± 0.019	± 0.022	± 0.013	± 0.018	± 0.006
CITRIS	0.07	0.04	0.07	0.20	0.26	0.04	0.12	0.03	0.09	0.07	**0.10**	**0.95**	0.07	0.95	0.13
	± 0.003	± 0.001	± 0.005	± 0.039	± 0.064	± 0.003	± 0.017	± 0.020	± 0.007	± 0.030	± 0.018	± 0.014	± 0.017	± 0.018	± 0.016
HERCULES	0.07	0.04	0.07	0.16	0.18	0.06	0.19	0.10	0.13	**0.04**	0.11	0.94	**0.04**	**0.96**	**0.10**
	± 0.001	± 0.002	± 0.002	± 0.029	± 0.041	± 0.005	± 0.036	± 0.033	± 0.012	± 0.001	± 0.018	± 0.009	± 0.005	± 0.013	± 0.006

Table 8: Experimental results with three different Voronoi datasets. The results are averaged over 5 different seeds.

	Triplet evaluation distances ↓							**Correlation metrics**			
	c_0	c_1	c_2	c_3	c_4	c_5	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Baselines											
Voronoi - 6 variables (Graph 1)											
CITRIS	0.11	0.07	0.13	0.11	0.11	0.09	**0.10**	**0.91**	0.10	**0.96**	0.26
	± 0.056	± 0.035	± 0.044	± 0.042	± 0.019	± 0.036	± 0.039	± 0.031	± 0.032	± 0.016	± 0.056
HERCULES	0.09	0.09	0.16	0.10	0.09	0.10	0.11	0.86	**0.09**	0.93	**0.25**
	± 0.038	± 0.022	± 0.051	± 0.013	± 0.020	± 0.030	± 0.029	± 0.070	± 0.042	± 0.040	± 0.058
Voronoi - 6 variables (Graph 2)											
CITRIS	0.14	0.12	0.16	0.14	0.20	0.15	0.15	0.83	0.17	0.91	0.36
	± 0.038	± 0.023	± 0.018	± 0.018	± 0.064	± 0.022	± 0.032	± 0.057	± 0.050	± 0.039	± 0.049
HERCULES	0.08	0.07	0.09	0.07	0.12	0.08	**0.09**	**0.92**	**0.04**	**0.96**	**0.19**
	± 0.016	± 0.015	± 0.038	± 0.023	± 0.045	± 0.014	± 0.027	± 0.031	± 0.019	± 0.017	± 0.037
Voronoi - 6 variables (Graph 3)											
CITRIS	0.11	0.17	0.09	0.10	0.13	0.13	0.12	0.86	0.14	0.91	0.32
	± 0.044	± 0.075	± 0.028	± 0.030	± 0.082	± 0.059	± 0.052	± 0.120	± 0.076	± 0.084	± 0.100
HERCULES	0.08	0.15	0.08	0.07	0.09	0.12	**0.09**	**0.87**	**0.09**	**0.92**	**0.24**
	± 0.029	± 0.049	± 0.016	± 0.032	± 0.046	± 0.025	± 0.034	± 0.049	± 0.039	± 0.035	± 0.045
False Positive Interventions											
Voronoi - 6 variables (Graph 1, 20%)											
CITRIS	0.14	0.08	0.17	0.12	0.15	0.11	0.13	0.85	0.14	0.92	0.33
	± 0.056	± 0.034	± 0.064	± 0.040	± 0.031	± 0.037	± 0.045	± 0.075	± 0.059	± 0.046	± 0.077
HERCULES	0.09	0.09	0.17	0.10	0.09	0.10	**0.11**	**0.86**	**0.09**	**0.93**	**0.24**
	± 0.038	± 0.022	± 0.051	± 0.013	± 0.020	± 0.030	± 0.029	± 0.073	± 0.044	± 0.041	± 0.060
Voronoi - 6 variables (Graph 1, 50%)											
CITRIS	0.16	0.10	0.18	0.15	0.16	0.13	0.15	0.78	0.12	0.88	0.31
	± 0.057	± 0.037	± 0.052	± 0.045	± 0.038	± 0.048	± 0.046	± 0.106	± 0.027	± 0.061	± 0.050
HERCULES	0.09	0.09	0.17	0.10	0.10	0.10	**0.11**	**0.86**	**0.08**	**0.93**	**0.24**
	± 0.039	± 0.023	± 0.053	± 0.013	± 0.021	± 0.032	± 0.03	± 0.073	± 0.042	± 0.041	± 0.064
Voronoi - 6 variables (Graph 1, 99%)											
CITRIS	0.16	0.13	0.20	0.18	0.19	0.16	0.17	0.67	0.14	0.80	0.33
	± 0.057	± 0.049	± 0.074	± 0.032	± 0.032	± 0.070	± 0.049	± 0.149	± 0.028	± 0.111	± 0.029
HERCULES	0.09	0.09	0.17	0.11	0.10	0.10	**0.11**	**0.86**	**0.08**	**0.93**	**0.23**
	± 0.040	± 0.023	± 0.053	± 0.014	± 0.021	± 0.033	± 0.03	± 0.072	± 0.038	± 0.041	± 0.063
Ablation Study: Grouping											
Voronoi - 6 variables (Graph 1)											
HERCULES 3-3	0.09	0.09	0.16	0.10	0.09	0.10	0.11	0.86	0.09	0.93	0.25
	± 0.038	± 0.022	± 0.051	± 0.013	± 0.020	± 0.030	± 0.029	± 0.070	± 0.042	± 0.040	± 0.058
HERCULES Interleaving	0.08	0.06	0.13	0.09	0.09	0.08	0.09	0.90	0.07	0.95	0.21
	± 0.044	± 0.016	± 0.060	± 0.027	± 0.027	± 0.033	± 0.035	± 0.066	± 0.036	± 0.036	± 0.063
HERCULES 5-1	0.07	0.05	0.11	0.07	0.06	0.10	**0.07**	**0.92**	**0.06**	**0.96**	**0.20**
	± 0.019	± 0.011	± 0.032	± 0.015	± 0.012	± 0.037	± 0.018	± 0.021	± 0.035	± 0.012	± 0.050
Voronoi - 6 variables (Graph 2)											
HERCULES 3-3	0.08	0.07	0.09	0.07	0.12	0.08	0.09	0.92	**0.04**	0.96	**0.19**
	± 0.016	± 0.015	± 0.038	± 0.023	± 0.045	± 0.014	± 0.027	± 0.031	± 0.019	± 0.017	± 0.037
HERCULES Interleaving	0.10	0.10	0.10	0.09	0.13	0.10	0.11	0.90	0.06	0.95	0.21
	± 0.044	± 0.051	± 0.024	± 0.020	± 0.045	± 0.012	± 0.037	± 0.034	± 0.026	± 0.018	± 0.049
HERCULES 5-1	0.06	0.04	0.07	0.06	0.06	0.13	**0.06**	**0.93**	0.06	**0.97**	0.20
	± 0.007	± 0.003	± 0.014	± 0.011	± 0.008	± 0.056	± 0.009	± 0.021	± 0.045	± 0.012	± 0.061
Voronoi - 6 variables (Graph 3)											
HERCULES 3-3	0.08	0.15	0.08	0.07	0.09	0.12	0.09	0.87	**0.09**	0.92	**0.24**
	± 0.029	± 0.049	± 0.016	± 0.032	± 0.046	± 0.025	± 0.034	± 0.049	± 0.039	± 0.035	± 0.045
HERCULES Interleaving	0.14	0.14	0.09	0.09	0.14	0.11	0.12	0.82	0.10	0.89	0.26
	± 0.073	± 0.046	± 0.020	± 0.026	± 0.054	± 0.014	± 0.044	± 0.095	± 0.043	± 0.066	± 0.033
HERCULES 5-1	0.05	0.09	0.06	0.04	0.05	0.12	**0.06**	**0.91**	0.10	**0.95**	0.25
	± 0.023	± 0.043	± 0.006	± 0.008	± 0.009	± 0.034	± 0.018	± 0.044	± 0.028	± 0.024	± 0.039
Ablation Study: 3 Hierarchical Levels											
Voronoi - 6 variables											
HERCULES (Graph 1)	0.08	0.05	0.14	0.07	0.06	0.10	0.08	0.89	0.08	0.94	0.24
	± 0.018	± 0.013	± 0.054	± 0.017	± 0.012	± 0.039	± 0.023	± 0.041	± 0.044	± 0.036	± 0.033
HERCULES (Graph 2)	0.07	0.05	0.07	0.06	0.07	0.13	0.06	0.93	0.04	0.96	0.16
	± 0.008	± 0.006	± 0.020	± 0.014	± 0.009	± 0.058	± 0.011	± 0.026	± 0.038	± 0.014	± 0.059
HERCULES (Graph 3)	0.06	0.10	0.06	0.04	0.05	0.12	0.06	0.90	0.08	0.95	0.23
	± 0.022	± 0.042	± 0.006	± 0.009	± 0.009	± 0.033	± 0.018	± 0.043	± 0.034	± 0.024	± 0.039