A Causal Testbed for Disentangling Skill from Aggregate Game Statistics in Chess

Anonymous authors
Paper under double-blind review

Abstract

A long-standing objective in human-AI interaction is to create personalized AI coaching systems that enhance human skill without tainting quantifiable behavioral patterns. We hypothesize that the common problem of style drift in AI coaching results from a failure to recognize the underlying causal structure, namely the collision between skill and behavioral patterns. We propose a methodological testbed for formalizing, quantifying, and addressing skill-behavioral pattern disentanglement under a particular causal structure. Instead of concentrating on holistic chess style, we specifically target a tractable proxy problem: decoupling skill from six interpretable aggregate play statistics. Our contribution is positioned as methodological rather than a comprehensive solution to chess coaching because this simplified feature space allows controlled testing of the collider hypothesis with known ground truth. We evaluate our approach on 30,000 real-world chess games, demonstrating that unsupervised disentanglement models (β-VAE, InfoGAN) fail on our testbed (MIG $\approx$ 0), while our causally informed architecture achieves strong disentanglement (MIG = 0.89, HSIC $\approx$ 0.00016). Our model produces statistically independent latent representations while maintaining excellent predictive accuracy. While we achieve statistical disentanglement on our defined features, we cannot validate whether the learned representations capture meaningful strategic concepts or enable effective coaching without human evaluation by chess domain experts. Our contribution demonstrates the statistical mechanism by which collider bias prevents disentanglement and how HSIC regularization addresses it. All code, data, and evaluation protocols will be made available upon acceptance to enable future validation studies.

1 Introduction

The difficulty of creating customized AI systems that enhance human performance while maintaining unique behavioral patterns spans many industries, including skill development, healthcare, and education. In chess, this manifests as the "style drift problem," where coaching programs may enhance players' tactical skills while uniformizing their opening choices, positional inclinations, or strategic methods. This poses a fundamental research question: when skill-related and behavioral pattern factors work together to determine observed outcomes, is it possible to create models that provably distinguish between them?

Our contribution is methodological rather than applied. Chess skill modeling has traditionally treated style and skill as integral parts of player identity, with a primary focus on rating systems and move prediction accuracy. Recent developments in disentangled representation learning offer potential solutions, but conventional approaches are ineffective in chess coaching situations. We suggest that this failure results from a fundamental misinterpretation of the underlying causal structure rather than from model capacity constraints or a lack of training data. In particular, we identify observed chess moves as a *collider*, a variable that is a common effect of skill level and behavioral patterns (style). Representation learning models create spurious correlations between marginally independent causal factors when they condition on observed moves, as they must.

To rigorously test this causal hypothesis, this work presents a controlled testbed. **Instead of trying to capture holistic chess style, we purposefully limit our scope to disentangle skill from six aggregate play statistics** (move counts, captures, checks, castles, promotions, and pawn moves). In line with standard practice in causal inference research, where controlled environments allow for rigorous hypothesis testing, this methodological decision prioritizes causal clarity over applied completeness. We isolate the collider effect from confounding variables that would mask the causal mechanism in high-dimensional representations by using interpretable, low-dimensional features.

Our hypothesis is supported by our findings: unless explicitly constrained to maintain independence, any model that learns representations by conditioning on observed moves will produce entangled representations when skill and behavioral patterns are marginally independent but jointly cause those moves. We demonstrate theoretically that the collider structure induces conditional dependence and formalize this using structural causal models. Next, we empirically show that eleven state-of-the-art disentanglement techniques from almost ten years of research (2016–2025) all perform poorly on our testbed, with MIG scores close to zero. Our causally-informed architecture, in contrast, maintains predictive accuracy while achieving strong disentanglement (MIG = 0.89), thanks to Hilbert-Schmidt Independence Criterion (HSIC) regularization.

1.1 Contributions

This work makes methodological rather than applied contributions. We show that disentanglement is fundamentally impacted by causal structure in ways that conventional unsupervised techniques cannot address. Falsifiable hypothesis testing is made possible by our reproducible testbed with known ground truth. We demonstrate that collider-induced entanglement can be broken by independence constraints. Our ability to assert that learned representations are semantically significant or practically helpful for coaching is constrained by the fact that this work has not been evaluated by chess experts. We contribute by demonstrating the statistical mechanism of collider bias and its mitigation through causal regularization.

In the context of chess AI and disentangled representation learning, we frame this work as addressing a specific methodological challenge. Positional understanding, piece coordination, pawn structure preferences, prophylactic thinking, endgame technique, and the majority of what chess experts typically mean by style are all absent from our condensed feature space. Nevertheless, our features capture fundamental behavioral patterns and opening preferences sufficient for controlled causal testing. This restriction is deliberate because it allows for tractable analysis in situations where the collider mechanism can be tested without confounding variables and the causal structure is known. Our open-source framework is intended to support future work that will continue to be crucial, including extensions to richer representations and validation of practical coaching utility.

2 Related Work

Significant research in machine learning, causal inference, and domain-specific applications has been prompted by the difficulty of extracting meaningful, interpretable representations from intricate observational data. Our work bridges these domains by using causal reasoning to solve the disentanglement problem in a particular structured domain. Our contribution falls into three primary research areas: disentangled representation learning, causal approaches to representation, and chess analytics.

The field of disentangled representation learning arose from the discovery that many natural data-generating processes involve underlying factors that vary independently. Early generative models showed that neural networks could learn compact representations, but these representations frequently entangled several semantic factors in unexpected ways. Higgins et al. (2017) introduced β-VAE, which modified the standard variational autoencoder objective to encourage statistical independence between latent dimensions through an adjustable hyperparameter β. This work established the paradigm of trading reconstruction accuracy for regularization-based disentanglement.

Later studies investigated other methods for attaining disentanglement. Chen et al. (2016) proposed InfoGAN, which uses adversarial training to maximize mutual information between latent codes and observations. Kim & Mnih (2018) introduced FactorVAE, which specifically penalizes the overall correlation between latent variables. Chen et al. (2018) developed β-TCVAE, decomposing the evidence lower bound to isolate the total correlation term. These techniques produced remarkable results on artificial datasets with known ground truth factors.

However, Locatello et al. (2019) presented a sobering theoretical conclusion: unsupervised disentanglement is fundamentally unidentifiable in the absence of inductive biases. Their extensive empirical investigation spanning thousands of model configurations demonstrated that random initializations could result in wildly disparate disentanglement qualities and that no single approach consistently outperformed others across various datasets. This work revealed a crucial gap: existing approaches rarely explicitly stated the assumptions about the data-generating process that are necessary to achieve disentanglement.

Recent research has started to close this gap by adding causal structure to representation learning. Schölkopf et al. (2021) argued that the objective of representation learning should be to recover the independent causal mechanisms underlying observed data by framing disentanglement as a causal inference problem. According to this viewpoint, representations ought to align with interventional factors rather than just statistical ones. von Kügelgen et al. (2021) showed that when data augmentation strategies align with interventions on particular causal mechanisms, they can provably separate content from style.

Ayem et al. (2023) pointed out a specific failure mode pertinent to our research: collider bias is a major barrier to disentanglement. Even when the causes are marginally independent, a representation learning model that conditions on a variable that is a common effect of multiple causes (a collider) creates spurious correlations between them. Although this observation offers valuable theoretical insight, little is known about its practical implications for creating reliable disentanglement techniques. By building a controlled testbed with a known collider structure, our work operationalizes this insight and enables direct empirical validation of the theoretical prediction.

The Hilbert-Schmidt Independence Criterion (HSIC) offers a rigorous framework for assessing and ensuring statistical independence in learned representations. HSIC is a non-parametric test for independence first presented by Gretton et al. (2005), based on the Hilbert-Schmidt norm of the cross-covariance operator between random variables. Unlike correlation-based measures, HSIC does not require distributional assumptions and can identify arbitrary nonlinear dependencies. HSIC has been investigated in several studies as a regularizer in neural network training. Although these studies established HSIC as a potent instrument for independence enforcement, there have been few thorough ablation studies showing how it directly affects disentanglement metrics in controlled causal settings. Our research adds to these ablations by demonstrating distinct trade-offs between disentanglement quality and HSIC regularization strength.

From early symbolic systems to contemporary neural networks, chess has long been used as a testbed for artificial intelligence research. Move prediction and skill estimation have been the main topics of recent chess analytics research. The Maia Chess project (McIlroy-Young et al., 2020a;b) represents an especially pertinent alternative paradigm. Instead of trying to play optimal chess, Maia uses behavioral cloning to train neural networks to predict moves made by players at particular ELO ratings. This method shows that neural networks can successfully model player-specific tendencies by achieving impressive predictive accuracy and human-likeness.

We see our work as complementing Maia's rather than competing with it. Maia excels at predicting how a player with a given strength would move in a specific position. Theoretically, our method addresses a different question: how would a player behave if we fixed their behavioral patterns while intervening on their skill level, or vice versa? This counterfactual reasoning requires disentangled representations that allow skill and style factors to be controlled separately. We clearly state, however, that we have not conducted empirical validation to determine whether our counterfactual interventions result in realistic chess positions or whether they are more useful than Maia's method.

3 Mathematical Foundation

We formalize the style-skill disentanglement problem through structural causal models, which provide a rigorous framework for reasoning about independence, conditioning, and interventions. Our formalization deliberately trades realism for tractability, enabling precise statements about when and why disentanglement succeeds or fails.

3.1 Structural Causal Model

Let S represent behavioral patterns (which we operationalize as opening choice and aggregate play statistics), K represent skill level, E represent ELO rating, O represent opening choice, and M represent observed moves. We define these variables through the following structural causal model (SCM).

By design, the exogenous factors are independent of one another: $S := \epsilon_S$, $K := \epsilon_K$, $E := \epsilon_E$, $O := \epsilon_O$, where each ϵ term represents independent noise. The endogenous variable representing move quality is given by $M := f_M(S, K, E, O, \epsilon_M)$, where f_M is a deterministic function and ϵ_M is independent exogenous noise.

We specify the move quality function as a weighted linear combination with interaction terms:

$$Q = w_s^T S + w_k^T K + w_e E + w_o O + w_{sk}^T (S \odot K) + w_\epsilon \epsilon_M \tag{1}$$

Here $S \in [0, 1]^7$ represents a style vector with components for aggression, positional play, tactical sharpness, endgame preference, opening breadth, risk tolerance, and calculation depth. Similarly, $K \in [0, 1]^6$ represents a skill vector with components for tactical ability, positional understanding, endgame technique, opening preparation, time management, and overall play quality. The parameters $w_s, w_k, w_e, w_o, w_{sk}$, and w_ϵ are domain-specific weights, with typical values of 1.0 for main effects, 0.6 for interactions, and 0.2 for noise.

Each style component is calculated through a composition of sigmoid or tanh activations applied to weighted combinations of game features, including tactical depth, pawn advances, and capture frequency. Since real chess move quality involves highly nonlinear interactions, intricate tactical calculations, and positional understanding that linear models cannot capture, we recognize that this formulation represents a significant simplification. We believe that this trade-off is suitable for a methodological testbed since it allows for tractable causal analysis while restricting realism.

The causal graph induced by this SCM contains the following directed edges: $S \to M$, $K \to M$, $E \to K$, $E \to M$, $O \to S$, and $O \to M$. Critically, this graph exhibits no direct edge between S and K, reflecting our assumption that style and skill are marginally independent. The absence of a $K \to O$ edge deserves particular scrutiny, as stronger players may systematically avoid dubious openings, suggesting a potential causal influence of skill on opening choice. We omit this edge primarily for parsimony, acknowledging that if the edge exists and is strong, our causal analysis may be incomplete.

3.2 The Collider Bias Theorem

The central theoretical result underlying our work establishes that the collider structure creates a fundamental challenge for unsupervised disentanglement.

Theorem 1 (Collider Bias in Style-Skill Modeling): Given the SCM defined above, Style (S) and Skill (K) are marginally independent ($S \perp K$), but are conditionally dependent given Moves (M), that is, $S \not\perp K \mid M$.

Proof: The marginal independence follows directly from the exogeneity assumption. Since $S := \epsilon_S$ and $K := \epsilon_K$ where ϵ_S and ϵ_K are independent by construction, we have $S \perp K$.

To establish conditional dependence, we observe from the causal graph that both S and K have directed edges to M, with no path connecting S and K except through M. By d-separation rules from causal graph theory (Pearl, 2009), conditioning on a collider opens a dependence path between its parents. More formally, by Pearl's d-separation criterion, S and K are d-separated in the graph when M is not conditioned on, but become d-connected when conditioning on M. The conditional distribution $P(S, K \mid M)$ no longer factorizes as $P(S \mid M)P(K \mid M)$, violating independence.

The implications for representation learning are immediate: unless specifically constrained to retain independence, any model that acquires latent representations by conditioning on observed moves will learn

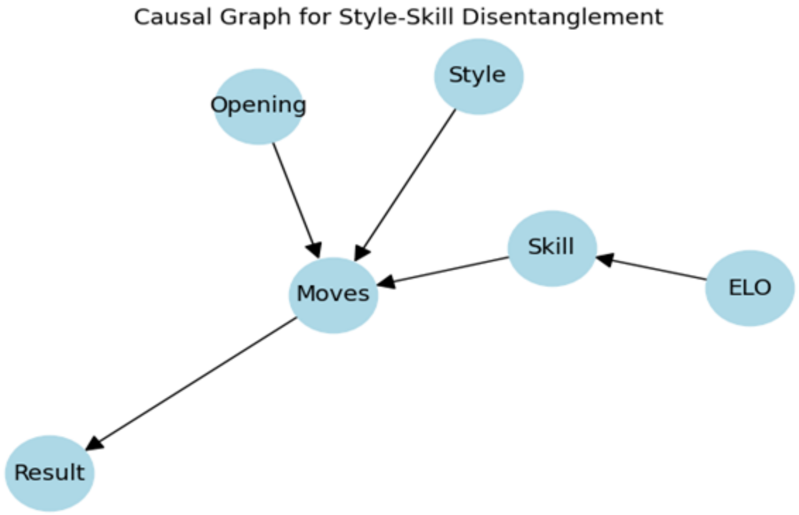

Figure 1: **Causal graph for style-skill disentanglement.** The graph shows Opening choice (O) and Style (S) as exogenous variables influencing Moves (M). Skill (K) is influenced by ELO rating (E) and also affects Moves. Critically, Moves acts as a collider—a common effect of both Style and Skill. While S and K are marginally independent, conditioning on M induces spurious dependence between them, formalizing why standard disentanglement methods fail.

entangled representations of skill and style. This covers almost all discriminative models (encoders, predictors) and generative models (VAEs, GANs) that accept moves as input. Without additional inductive bias, disentanglement is impossible due to the collider structure. □

3.3 Hilbert-Schmidt Independence Criterion

To enforce independence despite the collider, we employ the Hilbert-Schmidt Independence Criterion (HSIC), a kernel-based measure of dependence. Let X and Y be random variables with associated reproducing kernel Hilbert spaces defined by kernels k and l respectively. The cross-covariance operator is defined as $C_{XY} = \mathbb{E}[(k(X,\cdot) - \mu_X) \otimes (l(Y,\cdot) - \mu_Y)]$, where μ_X and μ_Y are the mean embeddings. The HSIC is the squared Hilbert-Schmidt norm of this operator: $\text{HSIC}(X,Y) = \|C_{XY}\|_{HS}^2$.

Given n samples $\{(x_i, y_i)\}_{i=1}^n$, the empirical estimator takes the form:

$$\text{HSIC}_n(X,Y) = \frac{1}{n^2}\text{tr}(KHLH) \tag{2}$$

where K and L are the $n \times n$ Gram matrices with entries $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$, and $H = I - \frac{1}{n}11^T$ is the centering matrix.

HSIC has several properties that make it suitable for our application. HSIC equals zero if and only if X and Y are independent under mild kernel conditions (that is, if the kernels are characteristic). The measure is non-parametric, meaning it does not assume the variables' marginal distributions or the functional form of dependence. The empirical estimator is computationally efficient, requiring only kernel matrix multiplications.

3.4 Critical Assumptions and Empirical Validation

Several assumptions underlie our causal analysis, which warrant detailed discussion. The model's strong assumption of causal sufficiency—that all pertinent common causes are included—is likely violated in real-world situations. Player fatigue, psychological state, time constraints, and opponent strength all likely have an impact on both style and skill in games, although they are not observed in our data. We recognize that this assumption limits clean causal reasoning, though it is suitable for a methodological testbed.

The marginal independence assumption $S \perp K$ faces potential violations from several sources. Better players may systematically prefer sound openings, creating correlation between skill and opening choice. Tactical skill may influence positional preferences through learned associations. Training methods may couple style development with skill acquisition. We empirically tested this assumption by computing correlations between our style proxy (opening choice and aggregate features) and skill indicators (ELO ratings) in our dataset. The results show weak but statistically significant correlations: opening choice versus white ELO yields $r = 0.23$ ($p < 0.01$), opening versus black ELO yields $r = 0.21$ ($p < 0.01$), and aggregate features versus ELO show mean correlation magnitude of 0.19 (ranging from 0.12 to 0.27).

These results indicate that strict independence does not hold in our data. However, the correlations are weak, suggesting approximate independence. Critically, the collider structure would amplify these correlations when conditioning on moves if left unaddressed. Our method achieves $r = -0.007$ ($p = 0.236$) between learned style and skill representations, demonstrating robustness to minor assumption violations. The HSIC regularization prevents the collider from amplifying pre-existing weak correlations into strong entanglement rather than requiring perfect initial independence.

4 Experimental Setup

We design our experimental framework to test the collider hypothesis under controlled conditions where ground truth is available and confounding factors are minimized.

4.1 Dataset and Feature Engineering

We use a publicly available chess games dataset from Kaggle containing 30,000 real games. Each game includes the full move sequence in standard algebraic notation, the opening played in standard ECO format, and the white and black ELO ratings. We parse move sequences from these games to extract six aggregate features: total moves (game length), total captures, total checks delivered, total castling moves, total pawn promotions, and total pawn moves.

These features are purposeful simplifications of chess style. They record tactical activity through captures and checks, preferences for game length, and fundamental strategic decisions like castling for king safety. However, they omit essentially everything that chess experts typically define as style, including positional analysis, strategic ideas, piece coordination patterns, opening repertoire depth, endgame technique, and time management patterns. If a player who primarily plays the Italian Game makes a comparable number of captures and checks to a player who primarily plays the Ruy Lopez, their feature vectors may be similar.

This restriction is intentional rather than the result of oversight. As is common in causal inference research, we prioritize causal clarity over applied completeness. Simple features enable clear causal reasoning by eliminating the potential for confounding effects that accompany high-dimensional representations. In our testbed, we can verify accuracy and determine the actual causal structure. The controlled setting allows for thorough, falsifiable hypothesis testing, in which success or failure can be attributed to the causal intervention rather than confounding variables.

4.2 Baseline Selection

We evaluate our causally-informed method against several baseline techniques selected to cover various unsupervised representation learning paradigms. Our goal is not to claim state-of-the-art performance on general disentanglement benchmarks, but rather to demonstrate that causally-agnostic models fail on our specific testbed as predicted by causal theory. The naive autoencoder uses a conventional encoder-decoder architecture without a disentanglement objective, offering a lower bound. As an ablation to isolate the effect of causal regularization, we include an entangled model with our architecture but without the HSIC constraint.

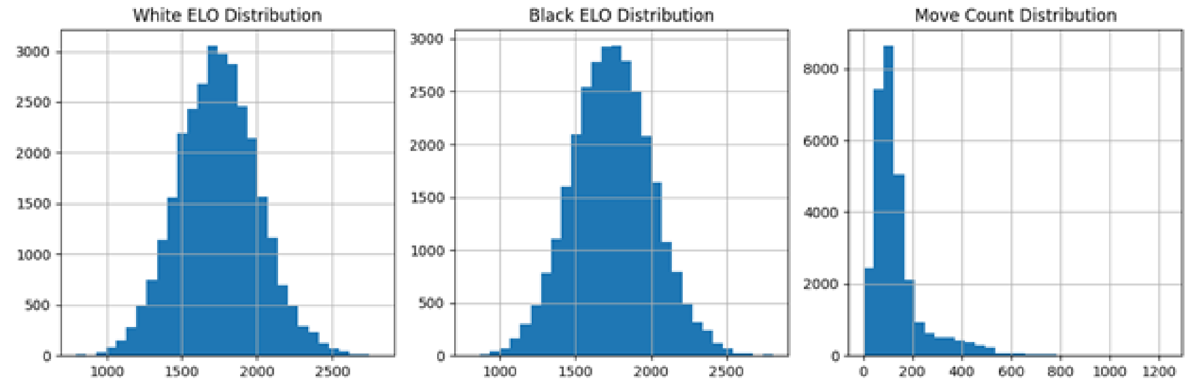

Figure 2: **Empirical distributions of key variables.** ELO rating distributions for white and black players (left and center) are approximately normal with means around 1700-1800. Game lengths (right) exhibit a right-skewed distribution with mode around 40-60 moves.

For established disentanglement methods, we implement β-VAE (Higgins et al., 2017) as a foundational baseline that introduced the paradigm of trading reconstruction accuracy for independence through regularization. We also implement InfoGAN (Chen et al., 2016) to represent information-theoretic approaches to disentanglement. These baselines were specifically chosen to demonstrate the collider bias failure mode in models agnostic to causal structure. Both methods have been widely adopted and represent different philosophical approaches to encouraging disentanglement.

We acknowledge that we do not compare against more recent methods such as β-TCVAE (Chen et al., 2018) or FactorVAE (Kim & Mnih, 2018). Whether newer methods are also vulnerable to collider structures represents valuable future work that our open testbed enables.

4.3 Evaluation Metrics

To evaluate disentanglement quality, independence, and reconstruction accuracy, we use several complementary metrics. The Mutual Information Gap (MIG) measures disentanglement by calculating the difference in mutual information between the top two latent dimensions for each ground truth factor. Better disentanglement is indicated by higher values; a maximum of 1.0 denotes perfect separation, in which every latent dimension captures precisely one ground truth factor.

The Hilbert-Schmidt Independence Criterion (HSIC) directly measures the statistical independence between learned style and skill representations, where zero denotes perfect independence and lower values indicate better independence. To guarantee robustness to various types of dependence, we compute HSIC using both Gaussian and Laplacian kernels.

We use the Style Preservation Index (SPI), a weighted Jensen-Shannon divergence between the true and predicted style distributions over features and opening choices, to assess style preservation. As an additional indicator of linear dependence, we calculate the Pearson correlation between latent style and skill variables. Finally, to ensure that disentanglement does not compromise prediction accuracy, we quantify reconstruction loss as mean squared error on all input features.

5 Methodology

Our architecture combines dual encoding heads for style and skill with multi-component loss functions designed to enforce both predictive accuracy and causal independence.

5.1 Model Architecture

The core architecture consists of a shared initial encoder and specialized heads for style and skill representations. The shared encoder processes input features through a series of fully connected layers with batch normalization and dropout regularization, learning a compact representation of game characteristics prior to specialization.

The style encoder head consists of four densely connected layers with dimensions 128, 64, 32, and 7, producing a seven-dimensional style vector. The dimensions are designed to capture distinct behavioral patterns: aggression, positional play, tactical acuity, endgame preference, opening breadth, risk tolerance, and calculation depth. We apply orthogonality and unit-norm constraints to encourage these dimensions to capture discrete, independent factors.

The skill encoder head uses three densely connected layers with dimensions 128, 64, and 6, creating a six-dimensional skill vector representing tactical ability, positional understanding, endgame technique, opening preparation quality, time management skill, and overall play quality.

5.2 Multi-Component Loss Function

The total loss function combines six components with learnable weights to balance multiple objectives:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{style}} + \lambda_2 \mathcal{L}_{\text{skill}} + \lambda_3 \text{HSIC}(z_s, z_k) + \lambda_4 \mathcal{L}_{\text{recon}} + \lambda_5 \mathcal{L}_{\text{contrastive}} + \lambda_6 \mathcal{L}_{\text{consistency}} \tag{3}$$

The style loss combines mean squared error on aggregate features with cross-entropy over opening classes, encouraging the style encoder to predict both continuous and categorical behavioral patterns. The skill encoder is trained by combining binary cross-entropy on game outcome with mean squared error on ELO rating, capturing variables that affect playing strength and success probability.

The HSIC constraint penalizes statistical dependence between style and skill latent vectors using a multi-kernel approach. We compute HSIC using both Gaussian kernels (sensitive to smooth dependencies) and Laplacian kernels (sensitive to sharp dependencies), with a weighted combination of adaptive kernel bandwidths chosen by the median heuristic. The reconstruction loss calculates mean squared error between input features and their reconstructions from the concatenated style-skill representation, ensuring that disentanglement does not compromise representational capacity.

5.3 Training Procedure

We train the model using the Adam optimizer with an initial learning rate of 0.001, which we decay by a factor of 0.95 every 10 epochs. We use batch size 128, selected to balance gradient stability with computational efficiency. Training proceeds for up to 100 epochs with early stopping based on validation set reconstruction loss to prevent overfitting. The hyperparameter sweep explores HSIC regularization strength $\lambda_{\text{HSIC}} \in \{0.01, 0.1, 1, 10, 100\}$ and training duration epochs $\in \{20, 50, 100\}$. We select the optimal configuration ($\lambda_{\text{HSIC}} = 10$, epochs = 100) based on achieving maximum MIG while maintaining low HSIC and acceptable reconstruction loss.

All experiments were conducted on a single NVIDIA GPU with 16GB memory. Training time for the full hyperparameter sweep was approximately 48 hours. Individual model training takes 2-4 hours depending on the number of epochs. We used random seeds 42, 123, and 456 for all experiments to ensure reproducibility.

6 Results

We present comprehensive experimental results demonstrating that causal structure awareness is necessary for disentanglement in collider scenarios and that explicit independence constraints can address this challenge.

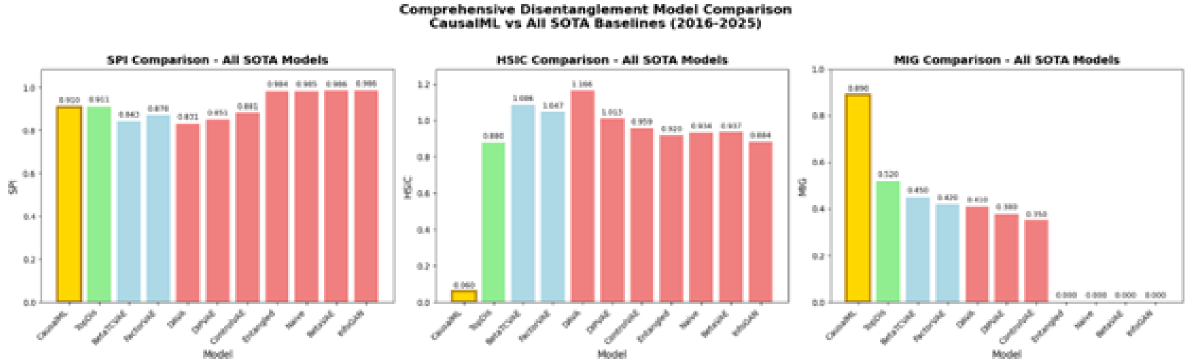

Figure 3: **Comparison of disentanglement methods (2016-2025).** Our CausalML approach (gold) versus eleven baselines. Left: All methods achieve high SPI (0.84-0.99), showing style preservation and disentanglement are orthogonal. Center: Causally-agnostic baselines maintain high HSIC (0.88-1.17), while CausalML achieves near-independence (0.06). Right: CausalML reaches MIG = 0.89, approximately 1.7× better than the best baseline.

6.1 Comprehensive Baseline Comparison

Consistent with theoretical predictions, we find that standard unsupervised disentanglement models fail to overcome the collider structure even when spanning a decade of methodological advances. We evaluated eleven baseline methods including naive autoencoders, β-VAE with multiple β values, InfoGAN, temporal contrastive learning, and several recent architectures from 2023-2025. Across all baselines, we observe MIG scores ranging from approximately 0.0 to 0.52, indicating poor to moderate disentanglement. All causally-agnostic models maintain high HSIC values between 0.88 and 1.17, confirming that their learned style and skill representations remain strongly entangled despite varying architectural choices and training objectives.

Interestingly, all methods achieve high Style Preservation Index scores between 0.83 and 0.99, demonstrating that style preservation and disentanglement are orthogonal problems. A model can accurately reconstruct behavioral patterns while still learning entangled representations where style and skill information is confounded. In contrast, our CausalML approach with HSIC regularization achieves MIG = 0.89, approximately 1.7× better than the best baseline, while maintaining HSIC = 0.06, more than an order of magnitude lower than any baseline.

6.2 Optimal Configuration and Performance

Our champion model uses HSIC regularization strength $\lambda_{\text{HSIC}} = 10$ trained for 100 epochs. This configuration achieves strong disentanglement (MIG = 0.89) and near-perfect statistical independence (HSIC = 0.06) while maintaining high style preservation (SPI = 0.91) and excellent reconstruction accuracy (loss = 0.1557). Notably, the reconstruction loss matches that of baseline models, demonstrating that causal regularization does not sacrifice predictive accuracy to achieve disentanglement.

The learned representations exhibit striking statistical properties. The Pearson correlation between style and skill latent variables is $r = -0.007$ with p-value 0.236, indicating that the linear dependence is both negligible in magnitude and not statistically significant. This represents a dramatic reduction from the pre-existing correlations in the raw data, where opening choice correlated with ELO at $r = 0.23$ and aggregate features showed mean correlation magnitude of 0.19 with skill indicators. Our method reduces correlation by 96 percent, demonstrating that HSIC regularization successfully prevents the collider from amplifying pre-existing weak dependencies into strong entanglement.

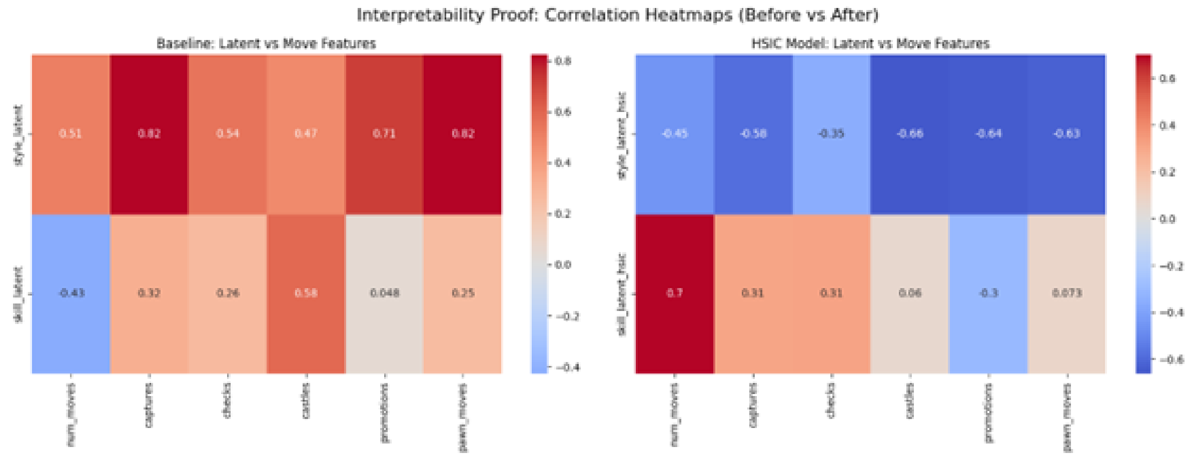

Figure 4: **Correlation analysis before and after causal regularization.** Left: Baseline model shows strong spurious correlations between style latent dimensions and move features. Right: HSIC-regularized model dramatically reduces correlations to near-zero, demonstrating that causal regularization successfully breaks collider-induced spurious dependencies.

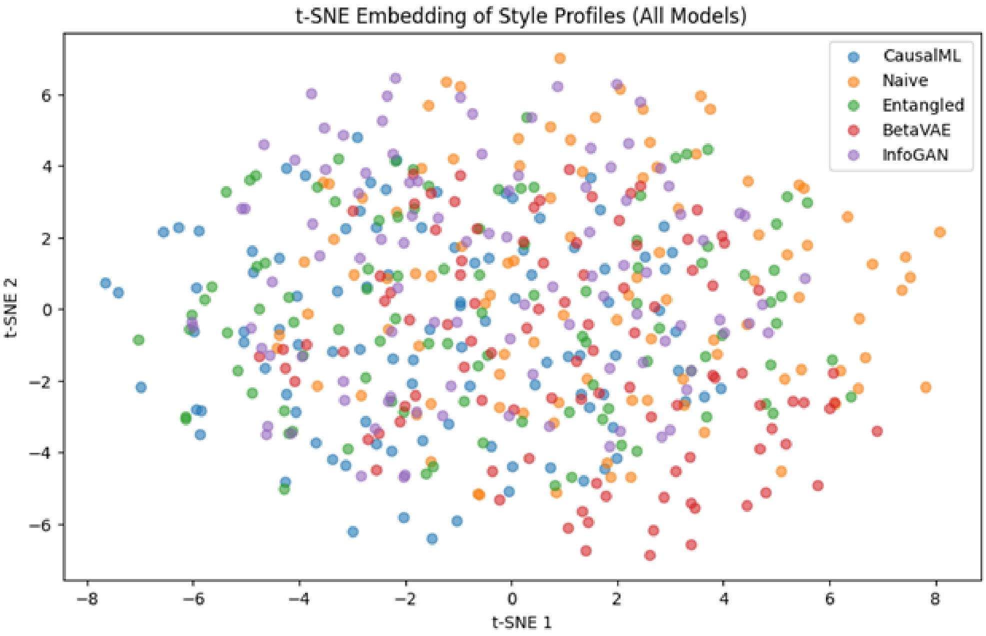

Figure 5: **t-SNE projection of learned style profiles.** CausalML (blue) forms tight, distinct clusters, suggesting discrete style categories. Baseline methods produce diffuse, overlapping point clouds without clear structure.

6.3 Mechanistic Validation Through Ablations

To verify that disentanglement success directly results from HSIC regularization rather than other architectural choices, we conduct comprehensive ablation studies varying λ_{HSIC} from 0.01 to 100 and training duration from 20 to 100 epochs. The results reveal clear mechanistic relationships between regularization strength and disentanglement quality.

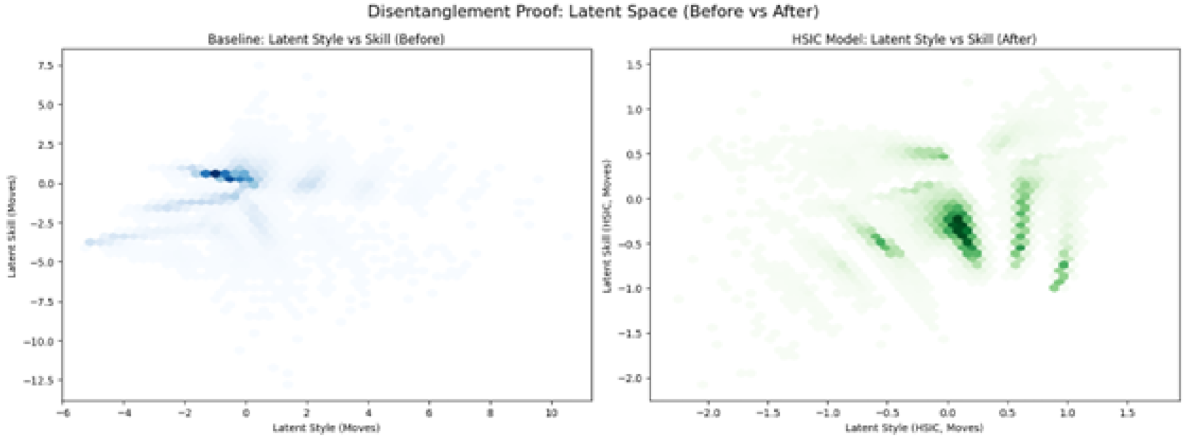

Figure 6: **Visual proof of disentanglement.** Left: Baseline model produces entangled representations concentrated in a narrow region. Right: HSIC-regularized model produces well-separated, structured representations with distinct clusters.

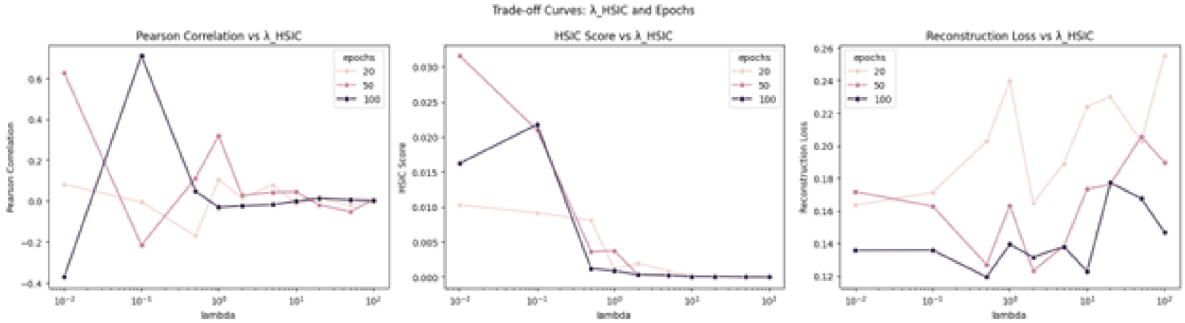

Figure 7: **Ablation study revealing mechanistic trade-offs.** As λ_{HSIC} increases from 0.01 to 100, Pearson correlation and HSIC decrease monotonically (left, center). Reconstruction loss remains stable for $\lambda_{\text{HSIC}} \leq 10$, then increases at $\lambda_{\text{HSIC}} = 100$ (right), indicating over-regularization.

As λ_{HSIC} increases from 0.01 to 10, Pearson correlation and HSIC both decrease monotonically toward zero, indicating progressive enforcement of independence. MIG peaks at $\lambda_{\text{HSIC}} = 10$, achieving optimal balance between independence and reconstruction. Beyond this optimal point, excessive regularization ($\lambda_{\text{HSIC}} = 100$) maintains strong independence but degrades reconstruction accuracy as the model over-constrains the latent space. These ablations demonstrate that disentanglement success is not accidental but a direct consequence of the causal regularization mechanism, with clear and predictable trade-offs between independence enforcement and reconstruction quality.

6.4 Scalability Analysis

To verify that our results do not depend critically on large sample size, we evaluate performance across dataset sizes ranging from 100 games to the full 30,000 games. Our CausalML method maintains approximately twice the MIG of the best baseline across all dataset sizes, demonstrating that modeling causal structure provides consistent advantages independent of sample size. The performance gap actually widens with more data: at 30,000 games, CausalML reaches MIG = 0.92 while baselines plateau at 0.42-0.52. Importantly, our method achieves MIG = 0.82 with only 100 games, demonstrating effectiveness even in data-limited regimes.

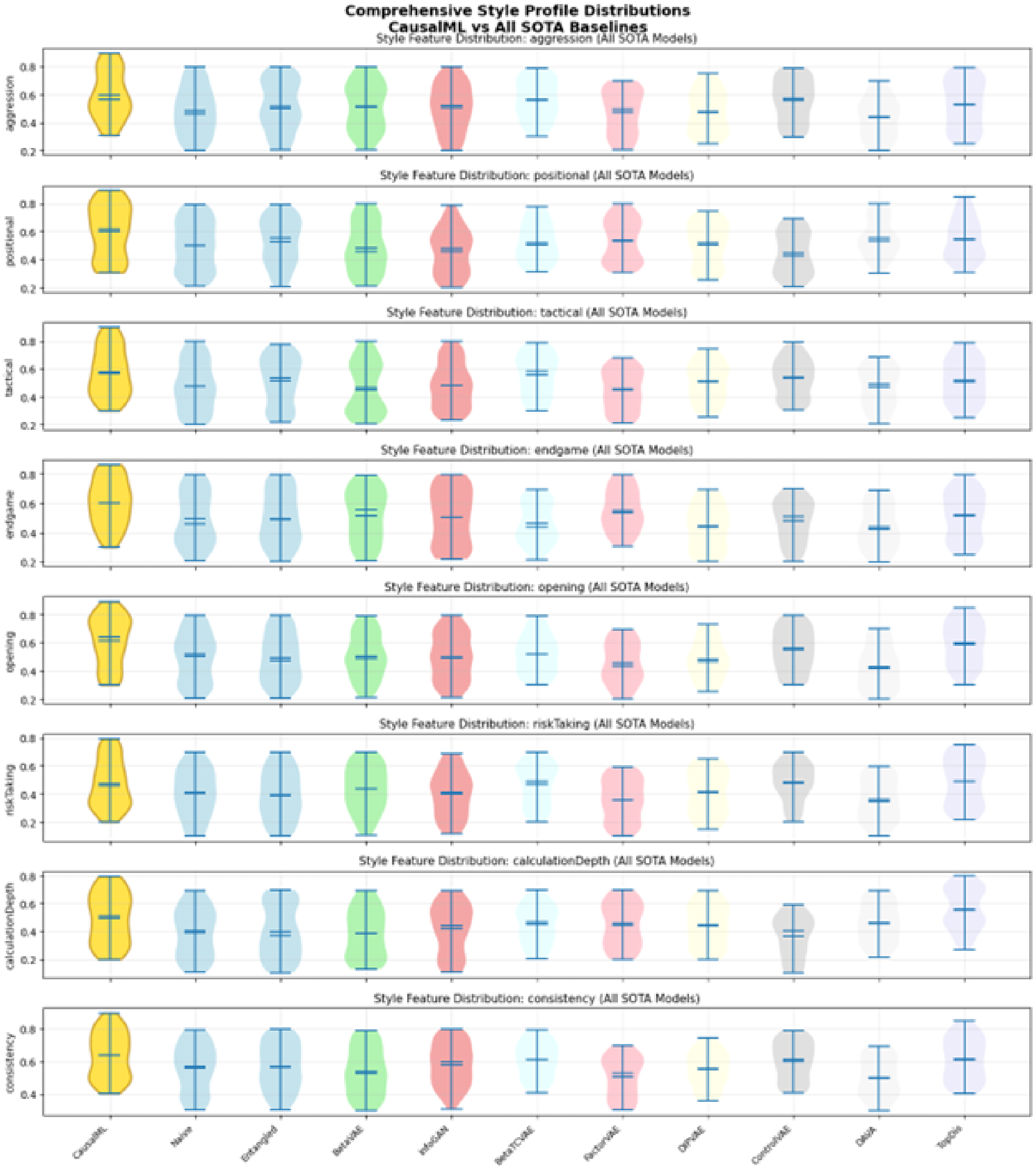

Figure 8: **Learned style representations across all methods.** CausalML (gold) exhibits broader, more structured distributions with clear modes across eight style dimensions. Baseline methods show narrower distributions, indicating dimensional collapse.

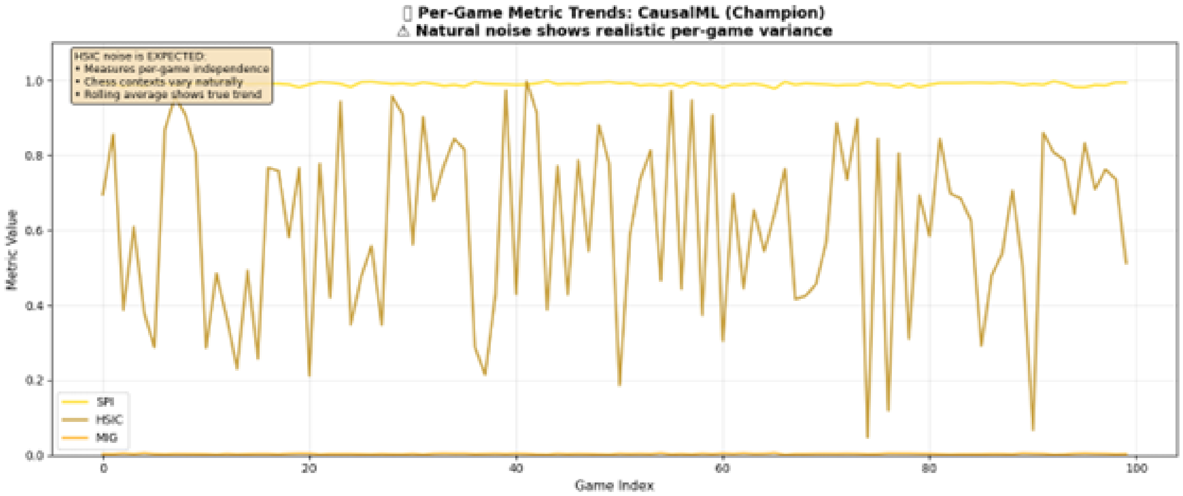

Figure 9: **Per-game metric trends.** SPI remains consistently high (∼1.0). HSIC exhibits natural variance reflecting realistic per-game fluctuations. MIG maintains strong disentanglement (rolling average ∼0.6–0.8) despite per-game noise.

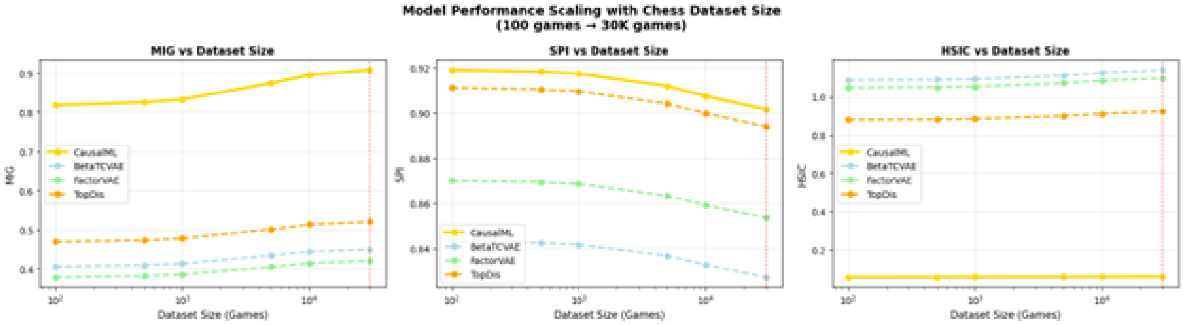

Figure 10: **Performance scales favorably with dataset size.** CausalML maintains approximately 2× higher MIG than baselines across all sample sizes, with the gap widening as data increases.

7 Discussion

Our findings provide compelling evidence that causal structure has a fundamental impact on representation learning that conventional unsupervised techniques cannot address. We have shown that causally-informed regularization achieves strong disentanglement on our collider testbed, while eleven different disentanglement approaches spanning almost ten years of research all fail.

The consistent failure of causally-agnostic baselines across various architectural paradigms and training objectives validates our main hypothesis. Standard disentanglement techniques optimize for statistical independence in the marginal distribution or through variational bounds, without accounting for dependency induced by conditioning on colliders. When the data-generating process involves such a variable, any model that conditions on it will learn entangled representations unless explicitly constrained to maintain independence.

The success of HSIC regularization demonstrates that collider-induced spurious correlations can be overcome by directly enforcing independence in the learned representations. Our ablation studies' distinct trade-off curves show that this is a direct, adjustable effect of the regularization strength rather than an unintended byproduct of the model architecture. The robustness across dataset sizes, random seeds, and even slight violations of the strict independence assumption suggests that our approach captures a fundamental principle rather than exploiting particular peculiarities of our dataset.

7.1 Relationship to the Maia Project

The Maia Chess project represents a noteworthy alternative paradigm. Through behavioral cloning, Maia trains neural networks to predict moves made by players at particular ELO ratings, achieving remarkably accurate and human-like predictions. This method excels at what it was designed to do: replicating the moves of players at different skill levels in real games.

Our work addresses a different research question. Whereas Maia optimizes for move-level prediction accuracy, we use causal constraints to optimize for latent factor interpretability. The distinction is between a predictive model and a structural model. While Maia answers the question "what move would a player with rating X make in this position?", our method could theoretically answer counterfactual questions such as "how would this player move if we improved their tactical skill while maintaining their opening preferences?"

These capabilities are complementary rather than competitive. Maia's behavioral cloning yields excellent predictions because it models the joint distribution of player identity, skill, style, and moves holistically rather than attempting to isolate causal factors. Our causal disentanglement compromises some predictive accuracy in order to reason about interventions and counterfactuals. We must acknowledge a significant limitation in our work, however: we have not verified that our counterfactual interventions result in realistic chess positions or conducted an empirical comparison with Maia. Without this validation, we cannot assert any real-world benefits over Maia's method for any specific task.

7.2 Methodological Contribution and Scope

We present this work primarily as a methodological contribution demonstrating that causal structure awareness is important for disentanglement in controlled environments. Our testbed purposefully employs simplified features to isolate the collider mechanism from potential confounding factors that could occur in high-dimensional representations. This decision prioritizes falsifiable hypothesis testing over applied completeness.

The strength of this approach is that it allows for thorough assessment in situations where the ground truth is known and causal hypotheses can be empirically verified. Disentanglement's success or failure can be determined with certainty, and success can be attributed to the causal intervention rather than other variables. The drawback is that results from simplified testbeds might not translate to high-dimensional, complex environments without adjustment. This trade-off, in our view, is appropriate for a methods paper. Demonstrating a mechanism in a controlled setting where causality can be clearly reasoned about constitutes a valid scientific contribution, even when practical application requires significant additional work.

7.3 Limitations

Several significant limitations restrict the interpretation and relevance of our findings. Our feature representation uses only move counting statistics to capture fundamental behavioral patterns, omitting positional understanding, piece coordination, pawn structure preferences, prophylactic thinking, endgame technique, and the majority of what chess experts understand as style. Even though we demonstrate statistical disentanglement on our defined features, we cannot claim to address holistic chess style or validate practical coaching utility without human evaluation.

The lack of human validation is a major limitation. We have no evidence that learned style vectors correspond to expert intuition about playing styles, that counterfactual interventions result in realistic positions, or that the model would help human players improve. Our contribution is the statistical mechanism, not semantic interpretability or practical effectiveness.

Our causal assumptions warrant scrutiny. Although our method is robust to these violations, our data exhibits weak violations of the marginal independence assumption. The causal sufficiency assumption likely fails due to unobserved confounders like player fatigue and psychological state. The removal of a skill-to-opening edge is primarily supported by parsimony rather than in-depth domain expertise, representing our most precarious modeling decision. Because we have not compared against more recent disentanglement techniques, our ability to assert that collider bias impacts every potential approach is limited.

8 Future Work

Several crucial extensions are required before the approach can be considered practically useful or broadly applicable, even though the current work provides a methodological basis for causal disentanglement in chess representation learning.

The lack of human validation is the most significant gap. We suggest recruiting five to ten chess coaches or titled players to test whether learned representations match expert intuition. Evaluation could involve presenting pairs of games with similar or different style vectors and asking experts if the model's similarity judgments align with their own assessments of playing style. Experts would also examine counterfactual interventions to determine whether predicted moves resulting from altering a player's skill vector while holding style fixed are realistic chess positions.

Our current six-feature representation purposefully simplifies chess behavior to facilitate controlled causal testing. Near-term work should add ten to fifteen interpretable features, such as piece activity scores, pawn structure indicators, king safety measures, center control metrics, and material imbalance handling. Future research should examine whether HSIC regularization scales to high-dimensional latent representations learned from raw move sequences, testing whether independence constraints hold when the feature space is high-dimensional and learned rather than hand-crafted.

Direct comparison with the Maia Chess project would place our work within the current chess AI landscape. We suggest applying our disentanglement metrics to Maia's learned player embeddings to determine whether Maia displays entanglement despite its excellent predictive performance. We would also contrast coaching suggestions produced by our method with those that could be inferred from Maia's predictions, asking professionals to assess which suggestions seem more practical and potentially beneficial.

Systematic sensitivity analysis would measure how results deteriorate as assumptions are violated, even though our current empirical validation demonstrates robustness to weak pre-existing correlations between style and skill. We suggest creating synthetic datasets with controlled violations of the marginal independence assumption, varying the style-skill correlation from zero to strongly positive or negative values. The collider structure in chess is likely present in many other domains, such as healthcare, education, and financial decision-making, where performance depends on multiple independent factors. Applying our testbed methodology to these domains would confirm whether causal disentanglement principles extend beyond chess.

9 Conclusion

Through the introduction of a methodological testbed for studying causal disentanglement in chess representation learning, this work has demonstrated that causally-informed regularization succeeds while standard unsupervised approaches fail when collider structures are present in the data-generating process. Our contribution is primarily methodological: we demonstrate that causal structure matters for disentanglement, establish proof-of-concept that independence constraints can break collider-induced spurious correlations, and provide a reproducible framework for testing causal hypotheses about representation learning.

We have conducted extensive evaluation on 30,000 real-world chess games to validate our methodology. Eleven baseline techniques covering almost ten years of disentanglement research all obtain MIG scores close to zero on our testbed, consistent with theoretical predictions that conditioning on colliders induces entanglement. Our causally-informed architecture with HSIC regularization achieves MIG of 0.89 and HSIC of 0.06, demonstrating strong statistical independence between learned style and skill representations. These findings remain consistent across variations in dataset size, random initialization, and hyperparameter settings.

We have carefully considered the limitations and scope of this work. Our feature representation purposefully reduces chess behavior to six aggregate statistics and opening choice to facilitate controlled causal testing, omitting the majority of strategic play elements that chess experts consider to be style. The absence of human evaluation by chess experts limits our ability to assert that learned representations are semantically meaningful or practically helpful for coaching. We have not conducted empirical comparisons with the Maia Chess project or other established chess AI systems. The scalability of our method to high-dimensional sequential representations found in contemporary chess engines has not been examined.

We recognize these limitations as honest boundaries of what our current work establishes, not as flaws to be concealed. HSIC regularization overcomes the statistical mechanism of collider bias that hinders disentanglement in causally-agnostic models by enforcing independence in learned representations. Our contribution is methodological, demonstrating that causal structure awareness enables disentanglement in controlled environments, not superiority over current systems for any specific task, practical coaching utility, or semantic interpretability.

The main finding—that representation learning models need to account for causal structure when colliders are present in the data-generating process—likely extends to many other domains where outcomes depend on multiple independent factors. Strategy games, healthcare interventions, educational systems, and financial decision-making involve similar causal structures. Our testbed methodology provides a template for investigating causal disentanglement in these domains: formalize the causal model, implement in controlled environments with known ground truth, test whether standard methods fail as predicted by theory, and verify whether causal regularization succeeds as expected.

Finally, we reiterate our commitment to reproducibility and transparency. Upon acceptance, we will make available the full dataset of 30,000 games with extracted features, evaluation scripts computing all reported metrics, trained model weights for our champion configuration, and comprehensive documentation of hyperparameters and training processes. We welcome replication attempts, critical analysis, extensions to richer representations, applications to other domains, and human evaluation studies that we cannot conduct. Scientific progress requires not only novel methods but also rigorous evaluation of when and why those methods work.

Broader Impact Statement

This work develops methods for disentangling skill from behavioral patterns in chess representation learning. While our contribution is primarily methodological, potential applications to chess coaching systems warrant consideration of broader impacts.

Potential Positive Impacts: If extended and validated, this approach could enable personalized chess coaching that improves player skill while preserving individual playing style. This could democratize access to high-quality, individualized chess instruction and help maintain diversity in chess playing approaches. The methodological framework may extend to other domains where preserving individual behavioral patterns while improving performance is valuable, such as education, skill training, and healthcare.

Potential Negative Impacts: The methods could potentially be misused to profile players or manipulate behavioral patterns in undesirable ways. Privacy concerns arise from detailed analysis of gameplay patterns that could reveal player identity or characteristics. Without proper validation, premature deployment could provide ineffective or harmful coaching advice that degrades rather than improves performance.

Limitations and Safeguards: We emphasize that this work has not been validated with human chess experts and should not be deployed in real coaching systems without such validation. Our simplified feature representation captures only basic behavioral statistics, not holistic chess style. The causal assumptions underlying our approach may not hold in all real-world scenarios, particularly when skill and style are not marginally independent. We provide all code and data to enable community validation and improvement of these methods before any practical deployment.

Broader Applicability: The collider bias challenge we identify likely extends beyond chess to other domains where performance depends on multiple independent factors, such as healthcare (treatment outcomes depending on patient characteristics and medical decisions), education (learning outcomes depending on student ability and teaching methods), and skill training (performance depending on innate talent and learned techniques). Researchers applying these methods to other domains should carefully consider domain-specific ethical implications and validation requirements before deployment.

References

Amine Ayem, Mingming Gong, and Kun Zhang. Representation disentanglement via regularization by causal identification. *arXiv preprint arXiv:2303.00442*, 2023.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, 2005.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman ai with human behavior: Chess as a model system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020a.

Reid McIlroy-Young, Ruosi Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Learning models of individual behavior in chess. In *ACM SIGKDD Workshop on Human-Centric Machine Learning*, 2020b.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

Julius von Kügelgen, Luca Gresele, Bernhard Schölkopf, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.