

Extracting Interaction-Aware Monosemantic Concepts in Recommender Systems

Dor Arviv¹, Yehonatan Elisha¹, Oren Barkan² and Noam Koenigstein¹

¹Tel Aviv University, Israel

²The Open University, Israel

Abstract

We present a method for extracting *monosemantic* neurons, defined as latent dimensions that align with coherent and interpretable concepts, from user and item embeddings in recommender systems. Our approach employs a Sparse Autoencoder (SAE) to reveal semantic structure within pretrained representations. In contrast to work on language models, monosemanticity in recommendation must preserve the interactions between separate user and item embeddings. To achieve this, we introduce a *prediction aware* training objective that backpropagates through a frozen recommender and aligns the learned latent structure with the model’s user-item affinity predictions. The resulting neurons capture properties such as genre, popularity, and temporal trends, and support post hoc control operations including targeted filtering and content promotion without modifying the base model. Our method generalizes across different recommendation models and datasets, providing a practical tool for interpretable and controllable personalization.

Code —

<https://github.com/DeltaLabTLV/Monosemanticity4Rec>

Introduction

Modern recommender systems rely heavily on latent embeddings to achieve scalable, personalized, and accurate recommendations (He et al. 2017, 2020; Gaiger et al. 2023; Barkan et al. 2019; Barkan, Katz, and Koenigstein 2020; Barkan et al. 2021a,b,c). However, these embeddings typically lack clear semantic meaning, significantly limiting interpretability. Such opacity reduces user trust, complicates debugging, and raises concerns around fairness and accountability.

Recent advances in Sparse Autoencoders (SAEs) have shown that *monosemantic neurons*, or latent dimensions aligned with human-interpretable concepts, can be extracted from Large Language Models (LLMs) (Templeton et al. 2024; Gao et al. 2024; Pach et al. 2025). These neurons support greater transparency, robustness, and fine-grained behavioral control (Zhang et al. 2024a; Makelov, Lange, and Nanda 2024; O’Brien et al. 2024). Motivated by these findings, we explore how SAE-based monosemantic analysis can be adapted to recommender systems, which rely on modeling interactions between distinct user and item embeddings.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Crucially, recommender systems differ fundamentally from LLMs. Rather than relying on forward propagation within a single shared representation space, they derive relevance from explicit interactions between separate user and item embeddings. As a result, existing SAE methods developed for LLMs offer limited suitability for recommender systems, since they do not sufficiently preserve the user–item interaction patterns that drive recommendation quality.

To address this gap, we introduce a novel Sparse Autoencoder framework explicitly designed to extract interpretable monosemantic neurons from recommender embeddings. Our key innovation is a *prediction-aware reconstruction loss*, which backpropagates gradients through a frozen recommender model. Unlike conventional SAE setups that focus only on geometric reconstruction, our approach preserves recommendation behavior by aligning the reconstructed embeddings with the recommender’s predicted affinities. This innovative design requires a non-traditional training process in which gradients flow through the fixed user–item interaction functions, ensuring behavioral alignment. Additionally, we replace the Top-K sparsity objective commonly used in prior LLM-based approaches with KL-divergence sparsity regularization, eliminating the dead neuron issue and improving the stability of learned representations. Importantly, our method integrates seamlessly into widely deployed two-tower architectures (Covington, Adams, and Sargin 2016; Yang et al. 2020; Yu et al. 2021; Wang et al. 2025) without requiring architectural changes or additional supervision.

Beyond interpretability, our extracted monosemantic neurons enable actionable interventions: by selectively modifying specific latent dimensions, we demonstrate the ability to expose users to novel content, suppress undesired genres, or strategically promote items to targeted audiences—all post hoc, without retraining the base model. These capabilities enhance transparency, user control, and real-world utility of recommender systems.

Contributions: (1) We introduce a Sparse Autoencoder explicitly tailored to extract interpretable monosemantic neurons from recommender embeddings. (2) We propose a novel prediction-aware reconstruction loss, preserving user-item interaction semantics. (3) We demonstrate practical applications including content filtering, user preference steering, and targeted item promotion. (4) We present the first comprehensive evaluation of monosemanticity across multiple

recommender models and diverse datasets.

Related Work

Monosemanticity via Sparse Autoencoders. Monosemanticity extraction is an emerging research area that seeks to identify latent dimensions aligned with coherent and interpretable concepts. Recent work in Large Language Models (LLMs) has demonstrated that Sparse Autoencoders (SAEs) can uncover such *monosemantic neurons*, supporting greater transparency, robustness, and fine grained control (Templeton et al. 2024; Gao et al. 2024; Paulo et al. 2024; Dang et al. 2024; Makelov, Lange, and Nanda 2024; O’Brien et al. 2024; Zhang et al. 2024a; Härle et al. 2025). To address challenges such as *feature absorption*, where individual neurons mix several unrelated concepts, Matryoshka SAEs (Bussmann et al. 2025) introduce hierarchical dictionaries that encourage coarse to fine specialization. At the same time, very large dictionaries can lead to *feature hedging*, a phenomenon in which multiple neurons redundantly capture the same factor (Chanin, Dulka, and Garriga-Alonso 2025). (Pach et al. 2025) proposed a monosemanticity metric based on the similarity of each neuron’s top activating items.

Prior SAE based monosemanticity research has largely focused on latent spaces that arise from forward propagation in transformer based LLMs. Although multi modal extensions exist (for example, CLIP (Zaigrajew, Baniecki, and Biecek 2025)), they do not address the problem of extracting semantics from explicit interactions between separate embeddings. Recommender systems differ in this respect, since they determine relevance through user–item embedding interactions rather than through a single evolving representation. This interaction based structure introduces challenges that have not been explored thus far.

Interpretability in Two-Tower Recommender Architectures. Two-tower recommenders independently encode users and items, combining them via interaction functions such as dot products in Matrix Factorization (Koren, Bell, and Volinsky 2009) or neural composition layers in NeuMF (He et al. 2017). Their scalability and modularity have made them the foundation of modern personalization systems (Covington, Adams, and Sargin 2016; Yang et al. 2020; Yu et al. 2021; Wang et al. 2025), yet their interpretability remains limited because relevance arises only through embedding interactions. Existing approaches either design *model-specific* interpretable factors or attention weights (Melchiorre et al. 2022; Abdollahi and Nasraoui 2017; Barkan et al. 2023; Zhang et al. 2024b) or apply *post-hoc* explanation techniques based on feature attribution and perturbation analysis (Ghazimatin 2020; Barkan et al. 2024; Guo et al. 2023; Qin et al. 2024). These methods illuminate outputs but leave the semantics of the latent embedding space opaque.

Method

We introduce an SAE framework specifically designed to extract monosemantic concepts from user and item embeddings in two tower architectures. Our approach incorporates a novel *prediction level* loss that directly accounts for the interaction based nature of recommender embeddings, yielding improved semantic consistency with the affinity patterns

learned by the underlying recommender.

General Recommender System Formulation

We consider a generalized two-tower architecture, common to recommender systems, where user and item inputs $\mathbf{x}_u \in \mathbb{R}^{d_u}$ and $\mathbf{x}_i \in \mathbb{R}^{d_i}$ are independently encoded:

$$\mathbf{e}_u = f_u(\mathbf{x}_u; \theta_u), \quad \mathbf{e}_i = f_i(\mathbf{x}_i; \theta_i), \quad (1)$$

with a scoring function:

$$\hat{y}_{u,i} = g(\mathbf{e}_u, \mathbf{e}_i; \theta_g), \quad (2)$$

predicting the user-item affinity score $\hat{y}_{u,i} \in [0, 1]$.

SAE Architecture

Given an embedding $\mathbf{e} \in \{\mathbf{e}_u, \mathbf{e}_i\}$, our SAE encodes it into a sparse latent vector $\mathbf{z} \in \mathbb{R}^m$ and reconstructs it via:

$$\mathbf{z} = h_{enc}(\mathbf{e}; \phi_{enc}) \quad (3)$$

$$\tilde{\mathbf{e}} = h_{dec}(\mathbf{z}; \phi_{dec}). \quad (4)$$

We enhanced our SAE with a *Matryoshka SAE* structure (Bussmann et al. 2025), a recently proposed variant that simultaneously trains multiple nested SAEs with increasing dictionary sizes. Formally, given a maximum dictionary size m , Matryoshka SAEs define nested dictionary sizes $\mathcal{M} = \{m_1, m_2, \dots, m_n\}$ with $m_1 < m_2 < \dots < m_n = m$. Each nested autoencoder independently reconstructs the input using only its subset of latent neurons, inducing a hierarchy in which early latents capture general features, while later latents specialize in finer-grained concepts.

Loss Functions

Our SAE is trained with a total loss combining reconstruction and sparsity objectives:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{sparse}, \quad (5)$$

Reconstruction Loss (\mathcal{L}_{rec}) The reconstruction loss comprises two components, embedding-level and prediction-level, with the latter being a key novelty tailored specifically for recommender systems:

1. **Embedding-level loss:** ensures geometric fidelity between original and reconstructed embeddings:

$$\mathcal{L}_{emb} = \|\mathbf{e} - \tilde{\mathbf{e}}\|_2^2. \quad (6)$$

2. **Prediction-level loss (novelty):** Designed specifically for recommender systems, this term aligns reconstructed embeddings with the recommender’s predicted affinities rather than with embedding geometry alone. Let $\hat{y}_{u,i} = g(\mathbf{e}_u, \mathbf{e}_i; \theta_g)$ be the original user–item affinity and $\tilde{y}_{u,i} = g(\tilde{\mathbf{e}}_u, \tilde{\mathbf{e}}_i; \theta_g)$ its reconstruction, computed through the frozen scoring function g . We define the prediction-level loss as the mean squared difference between the original and reconstructed affinities over a random sample of user–item pairs \mathcal{S} :

$$\mathcal{L}_{pred} = \frac{1}{|\mathcal{S}|} \sum_{(u,i) \in \mathcal{S}} (\hat{y}_{u,i} - \tilde{y}_{u,i})^2. \quad (7)$$

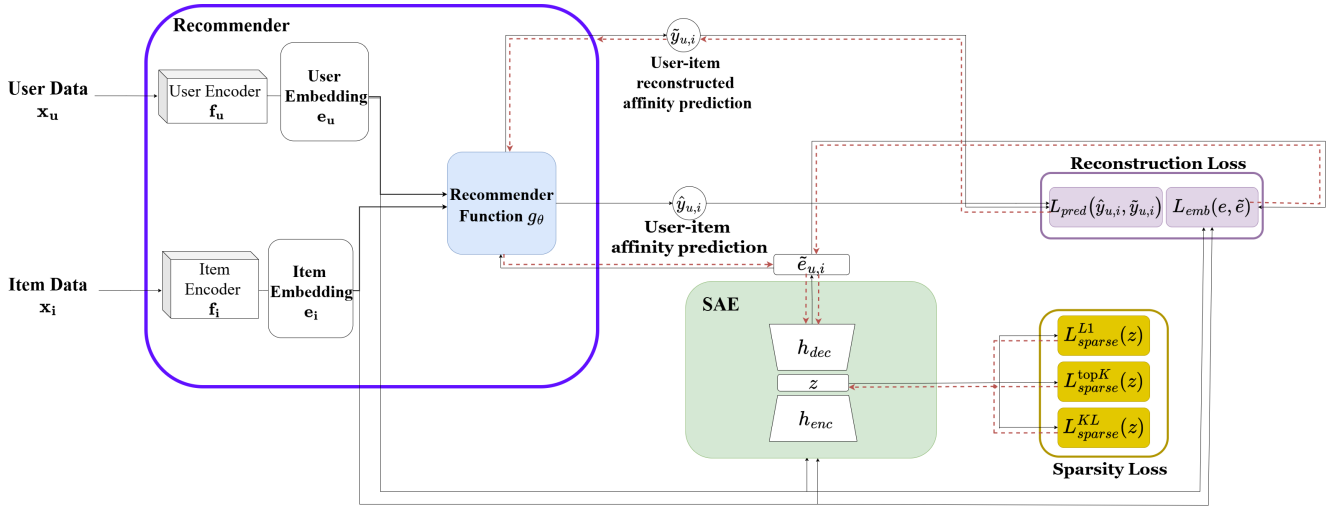


Figure 1: Framework architecture. Solid black arrows indicate forward pass; dashed red arrows show gradient flow. Our innovation adapts monosemanticity to recommender systems by backpropagating a novel prediction-level loss through a frozen recommender, thus preserving user-item interaction semantics.

This term encourages the SAE to preserve interaction semantics and ranking consistency, which are more critical to recommendation quality than geometric proximity alone. Because $\mathcal{L}_{\text{pred}}$ depends on the frozen recommender’s gradients, training requires backpropagation through g , as explained below.

The final reconstruction loss is:

$$\mathcal{L}_{\text{rec}} = \alpha \mathcal{L}_{\text{emb}} + \beta \mathcal{L}_{\text{pred}}. \quad (8)$$

Sparsity Loss ($\mathcal{L}_{\text{sparse}}$) We encourage sparsity using a combination of ℓ_1 regularization and a KL divergence penalty applied to the activations of the bottleneck neurons. The KL term is computed across a batch of input samples and encourages each neuron to be active only a small fraction of the time. For neuron j , let $p_j = \frac{1}{B} \sum_{i=1}^B h_{ij}$ denote its empirical activation rate over a batch of B inputs. The sparsity penalty is then

$$\text{KL}(p_j \parallel \rho) = \rho \log \frac{\rho}{p_j} + (1 - \rho) \log \frac{1 - \rho}{1 - p_j},$$

which penalizes deviations between p_j and the target rate ρ , encouraging compact and disentangled representations (Bank, Koenigstein, and Giryes 2023). The full sparsity loss is given by a weighted sum of the ℓ_1 and KL components, with coefficients λ_1 and λ_2 . Unlike prior work on LLMs, we omit Top K sparsity, which has been observed to lead to unstable dynamics and inactive neurons (Templeton et al. 2024).

Training Procedure

We train the SAE post hoc, keeping the base recommender fixed ($\theta_u, \theta_i, \theta_g$ frozen). At each step, user–item pairs are sampled and the total loss $\mathcal{L}_{\text{total}}$ is computed by passing reconstructed embeddings through the frozen model. Crucially, gradients from $\mathcal{L}_{\text{pred}}$ are backpropagated *through the frozen recommender*, enabling the SAE to align its latent representation with the recommender’s behavioral outputs. This

mechanism is illustrated in Figure 1, which shows how the prediction loss induces a gradient path through the recommender’s scoring function $g(\cdot)$.

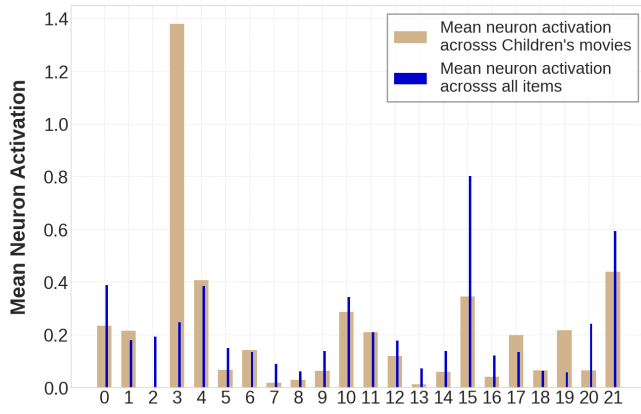
Results

Experimental Setup

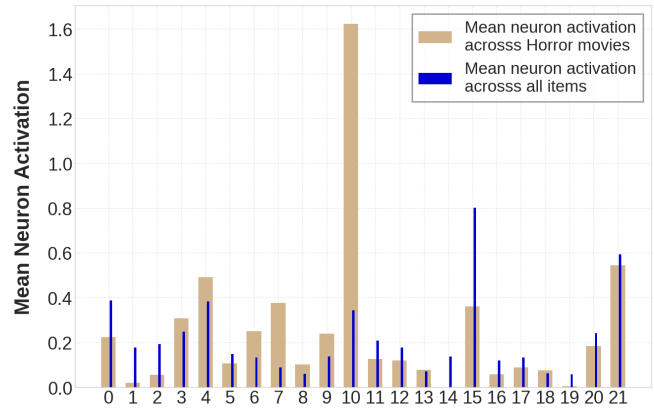
We conduct experiments using two representative recommender models, Matrix Factorization (MF) (Koren, Bell, and Volinsky 2009) and Neural Collaborative Filtering (NCF) (He et al. 2017), on two diverse datasets: MovieLens 1M (ML1M) (Harper and Konstan 2015) and Last.FM (Bertin-Mahieux et al. 2011). These domains differ in content type (movies vs. music), density, and the semantic coherence of item space. In ML1M, we binarize ratings by treating all user-rated items as implicit positive feedback. In Last.FM, where interactions are inherently implicit, we aggregate user–track events at the artist level to enable more meaningful neuron interpretation.

Recommendation Models. Both MF and NCF use user and item embeddings of dimensionality d , with $d = 20$ for ML1M and $d = 100$ for Last.FM. MF computes affinity via inner product, while NCF applies a two-layer MLP (with 64, 32 and 16 hidden units) followed by a sigmoid activation. When training the recommenders on implicit feedback, positive user–item interactions are trained jointly with negative samples, dynamically drawn per epoch with probability proportional to item popularity, following established implicit feedback practices (Hu, Koren, and Volinsky 2008; Paquet and Koenigstein 2013; Rendle 2021).

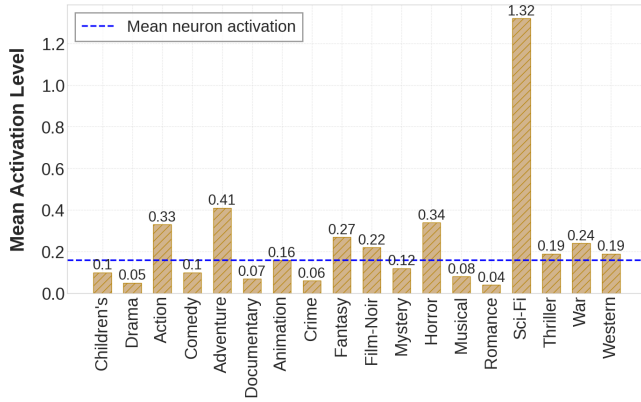
We formed the test set by holding out, for each user, five positive items, using all remaining interactions for training. A validation subset was carved from the training set to monitor Mean Percentile Rank (MPR) for hyperparameter tuning and early stopping. The recommenders were trained using binary cross-entropy loss and optimized with Adam (Kinga,



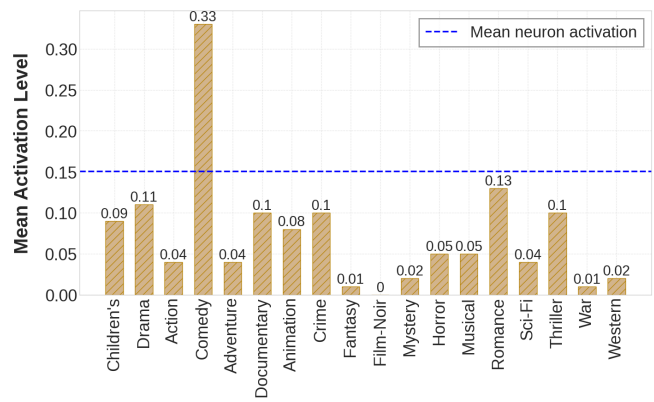
(a) Children (NCF, ML1M)



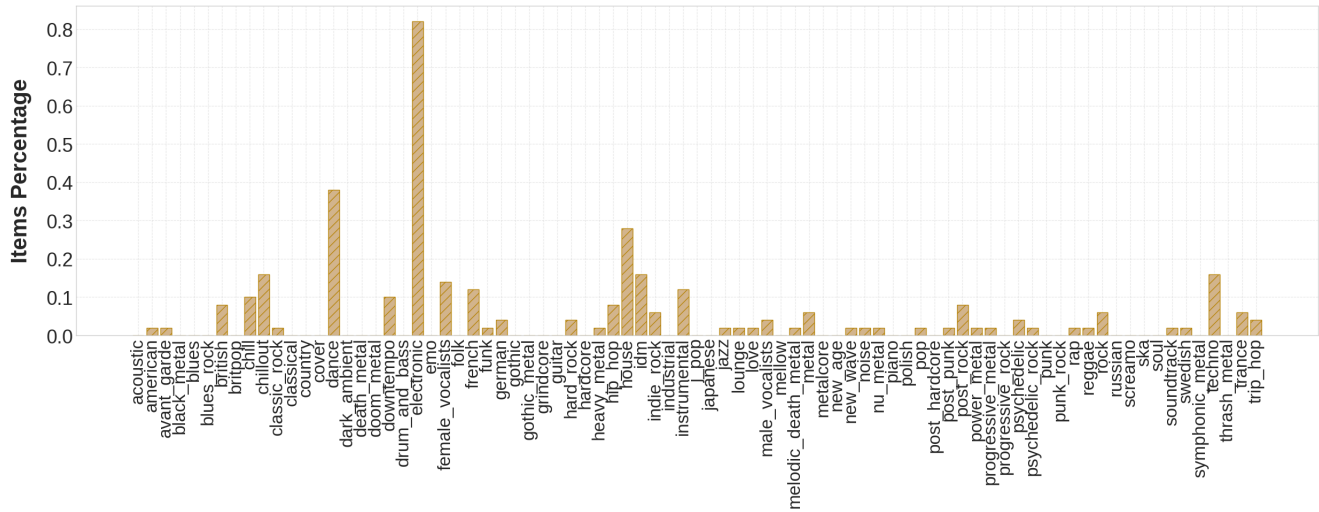
(b) Horror (NCF, ML1M)



(c) Sci-Fi (MF, ML1M)



(d) 90s Comedy (MF, ML1M)



(e) Electronic/Dance (MF, Last.FM)

Figure 2: Representative monosemantic neurons extracted from our SAE bottleneck. **Top row:** Mean activation of all neurons for items from two genres (Children, Horror), revealing sharp peaks at genre-aligned units. **Middle row:** Mean activation of two genre-selective neurons (Sci-Fi, Comedy) across items from various genres, showing strong intra-neuron selectivity. **Bottom row:** Tag distribution among the top-50 activating artists for a neuron in the Last.FM dataset, highlighting alignment with *electronic* music (e.g., *dance*, *house*, *techno*). Together, these examples illustrate the emergence of interpretable, concept-specific neurons across domains.

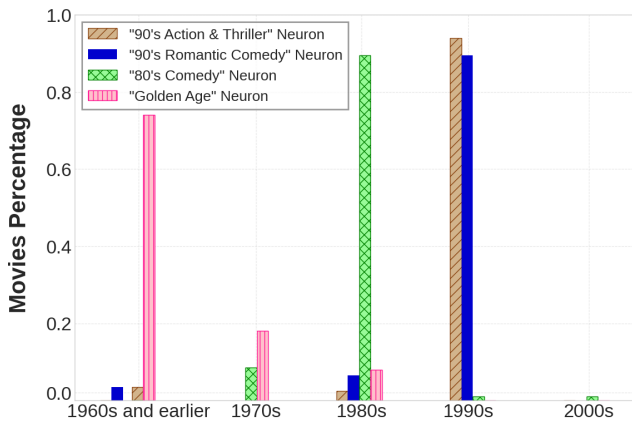


Figure 3: Temporal specialization of four NCF neurons. Each plot shows the decade-wise distribution of top-activating movies, revealing sharp alignment with stylistic eras, e.g., 1990s Thrillers, 1980s Comedies, and Golden Age films).

Adam et al. 2015). We used the following hyper-parameters: *MLIM*: learning rate 0.05, batch size 256, trained for 10 epochs. For *Last.FM*: learning rate 0.05, batch size 8, trained for 30 epochs.

Sparse Autoencoder. For each model–dataset pair, we train a separate SAE consisting of a linear encoder with ReLU activation, a sparse bottleneck of m neurons, and a tied linear decoder (Vincent et al. 2008). We use $m = 22$ for *MLIM*, which allows for compact representations aligned with discrete genres and ratings. For *Last.FM*, where musical preferences are more diverse and less neatly clustered, we adopt a larger bottleneck of $m = 70$ to better accommodate the higher conceptual complexity. We used the Adam optimizer (Kinga, Adam et al. 2015). Hyperparameters were tuned following Pach et al. (2025) by maximizing a monosemanticity score based on the weighted pairwise similarity among each neuron’s top thirty activating items. The full set of final hyperparameters for all models and datasets is provided in our public repository.

Neuron Identification Protocol. To assess interpretability, we employed an automatic labeling pipeline using GPT-4.5. For each latent neuron in \mathbf{z} , we extracted the top-activating items to the model, prompting it to evaluate semantic coherence. When a consistent theme emerged, GPT-4.5 assigned a concise label (e.g., *Comedy*, *Electronic*, *1990s Action*). This procedure was repeated independently across models and datasets, enabling scalable, high-level interpretation without manual annotation.

Qualitative Analysis

We qualitatively analyze the monosemantic neurons learned by our SAE and show that they capture coherent semantic concepts, such as genres, popularity, and stylistic eras, directly from user-item embeddings without requiring external supervision or metadata.

Genre-Aligned Neuron Identification. Figure 2 (top row) shows mean \mathbf{z} activations from the SAE trained on NCF

embeddings (MLIM) for items labeled as *Children* and *Horror*. For each genre, we plot the average activation for items belonging to this genre across all neurons (brown bars), alongside the mean activation over the entire item catalog (blue line). In both cases, a single neuron exhibits a sharp peak, strongly activating for the target genre while others remain near baseline, demonstrating robust monosemantic alignment with semantic categories.

Genre Selectivity Within Neurons. Figure 2 (middle row) shows the activation profiles of two MF neurons aligned with *Sci-Fi* and *Comedy*. For each neuron, we compute the mean activation across items (movies) of different genres. In both cases, a sharp peak emerges for the neuron’s associated genre, while activations for other genres remain near baseline. This illustrates strong intra-neuron selectivity, where a single latent unit reliably distinguishes its semantic target from unrelated content.

Figure 2 (bottom) presents a representative neuron from MF embeddings on the Last.FM dataset. The top-50 most activating artists are annotated with tags prominently associated with *electronic* music such as *dance*, *house*, and *techno*. Despite the greater noise and stylistic diversity in music data, the neuron exhibits strong alignment with electronic music subgenres, demonstrating monosemanticity extraction even in more complex domains.

Popularity Bias and Temporal Trends. Our SAE reveals latent dimensions that capture broader behavioral signals beyond genre. In both datasets, we identify a “popularity neuron” that activates consistently for highly popular items at the head of the long-tail distribution, indicating sensitivity to mainstream appeal. This aligns with well-established popularity bias in recommender systems (Klimashevskaja et al. 2024; Zhu et al. 2021). Quantitative results of this example are discussed below (Table 1).

Figure 3 further highlights neurons that align with specific cinematic periods. Each neuron activates for movies concentrated in a distinct era (e.g., 1990s Action, 1980s Comedies, or pre-1970 classics), demonstrating that temporal and stylistic patterns can also be disentangled directly from user-item behavior.

Takeaway. These results demonstrate that monosemantic neurons emerge naturally from recommender embeddings, encoding genre, style, popularity, and temporal alignment, without requiring supervision or metadata. Such neurons support interpretable, controllable systems, and open the door to fine-grained user modeling and content auditing in recommendation.

Quantitative Analysis

We now assess monosemanticity quantitatively, leveraging genre annotations from MovieLens (MLIM) and Last.FM to evaluate the semantic precision and fidelity of the extracted neurons. We quantify this precision using a simple *semantic purity* measure, defined as the percentage of items within the top K activating items that match the neuron’s assigned label. All purity values are computed using the metadata provided in each dataset.

Movies (ML1M)						
Concept S	MF			NCF		
	K=10	K=20	K=50	K=10	K=20	K=50
Children's	0.90	0.90	0.88	1.00	1.00	0.98
Drama	1.00	1.00	0.98	0.80	0.70	0.65
Action	0.90	0.65	0.60	–	–	–
Adventure	0.80	0.80	0.70	–	–	–
Animation	0.40	0.38	0.48	1.00	0.95	0.76
Comedy	1.00	1.00	1.00	0.90	0.75	0.66
Horror	1.00	1.00	1.00	0.90	0.95	0.96
Romance	1.00	0.85	0.74	0.90	0.85	0.72
Sci-Fi	1.00	1.00	1.00	0.70	0.75	0.64
Thriller	0.80	0.90	0.80	0.80	0.80	0.62
Crime	0.80	0.75	0.52	0.50	0.45	0.48
Golden-Age	1.00	1.00	0.98	0.80	0.80	0.80
90's Drama	0.90	0.85	0.76	1.00	0.95	0.86
90's Comedy	1.00	0.95	0.78	1.00	1.00	0.92
90's Action	1.00	1.00	0.94	1.00	1.00	0.98
80's Comedy	–	–	–	1.00	0.95	0.86
Popularity	4.35%	6.30%	6.34%	1.04%	3.20%	3.73%

Music (Last.FM)						
Concept	MF			NCF		
	K=10	K=20	K=50	K=10	K=20	K=50
Electronic	1.00	0.90	0.82	1.00	1.00	0.76
Metal	1.00	1.00	0.92	1.00	0.95	0.82
Folk	0.90	0.85	0.68	1.00	0.95	0.76
Trance	1.00	0.95	0.72	–	–	–
Country	1.00	0.90	0.56	1.00	0.75	0.40
Hardcore	0.95	0.95	0.96	0.90	0.85	0.68
Reggae	1.00	0.90	0.60	–	–	–
Pop	0.80	0.70	0.68	0.90	0.90	0.70
Rock	1.00	0.95	0.92	1.00	0.80	0.78
Emo	–	–	–	0.95	0.85	0.64
Popularity	2.49%	3.80%	5.49%	0.46%	0.74%	4.40%

Table 1: Semantic purity (%) of concept-aligned neurons in MF and NCF models on MovieLens (ML1M) and Last.FM. Entries show the fraction of top- K activating items ($K = 10, 20, 50$) matching the neuron’s concept. Dashes indicate no neuron was found to match the concept. The final row for each dataset reports the average popularity percentile for a “popularity” neuron.

Semantic Alignment of Neurons. Table 1 reports the semantic purity of monosemantic neurons extracted from MF and NCF models on MovieLens (ML1M) and Last.FM. For each neuron, we compute the fraction of top- K activating items ($K = 10, 20, 50$) that match a target concept such as a movie genre or music style. This evaluates how consistently each latent unit aligns with its category.

Despite fully unsupervised training, high-purity neurons emerge consistently across domains. Purity naturally declines with increasing K , yet many neurons maintain strong alignment. For example, *Comedy* and *Horror* in MF achieve 100% purity across all values of K , and music concepts such as *Country*, *Reggae*, and *Metal* also show near-perfect alignment.

Some neurons exhibit reduced purity at higher K , often due to labeling noise, genre ambiguity, or missing annotations. For instance, the *Crime* movies neuron in NCF is acti-

vated by thematically consistent but unlabeled titles. Similar challenges arise in music, where cross-genre artists or mainstream hits may blur concept boundaries.

The final row in each dataset reports the “popularity neuron” discussed earlier, presenting the average popularity percentile of the top- K activating items. Across all K values, the percentiles remain consistently high, indicating that these neurons capture a strong affinity for widely consumed content. In ML1M, the neuron consistently ranks blockbuster titles near the top of the long-tail distribution, while in Last.FM it activates for hit singles and highly streamed artists. Unlike genre- or style-based neurons, this dimension reflects behavioral appeal rather than semantic coherence.

Overall, these results demonstrate the emergence of semantically meaningful neurons across domains, validating the potential of our approach for unsupervised concept discovery in recommender systems.

Ablation and Fidelity Trade-Offs Figure 4 presents ablation results on the *MovieLens* dataset for both the MF and NCF recommenders. The figure shows how increasing the weight β of our prediction level loss (Equation 8) improves recommendation fidelity, measured via Rank Biased Overlap (RBO) and Kendall Tau correlation between the original and reconstructed top thirty lists. As expected, higher β yields stronger alignment with the original model. Importantly, $\beta=0$ serves as an ablation baseline, revealing that omitting $\mathcal{L}_{\text{pred}}$ leads to poor fidelity. However, fidelity gains taper off at large β values while bottleneck sparsity, and therefore interpretability, declines. The rightmost panel tracks the monosemanticity score (Pach et al. 2025), which reaches its maximum at intermediate β values. These results highlight the need to balance predictive faithfulness with semantic clarity.

Beyond Interpretability: Practical Use Cases

Our method enables more than interpretability: it allows targeted interventions in the recommendation pipeline via neuron-level edits in latent space. These interventions, enabled by our prediction-aware loss, can modify model behavior post-hoc (without retraining) by adjusting neuron activations to promote, suppress, or diversify recommendations.

Targeted Item Promotion. As a central example, we demonstrate how to expose niche audiences to selected items via direct intervention in item representations. Using an MF recommender trained on the *Last.FM* dataset, Figure 5 shows that by increasing the activation of a specific neuron in Bob Dylan’s latent vector \mathbf{z} , we can promote him to users who primarily listen to Metal, Contemporary Pop, and Electronic music. Despite no prior affinity, Dylan is ranked increasingly higher in these users’ top thirty recommendations. This illustrates a lightweight, behaviorally grounded strategy for personalized content promotion that bypasses retraining or reliance on global popularity signals.

Additional Interventions. Additional examples on the ML1M dataset further illustrate this capability. For instance, we show how nudging under eighteen users toward Children’s content and suppressing Horror content for sensitive users can be achieved by adjusting the corresponding neurons. We also demonstrate how a non aligned movie (*Speed*) can be

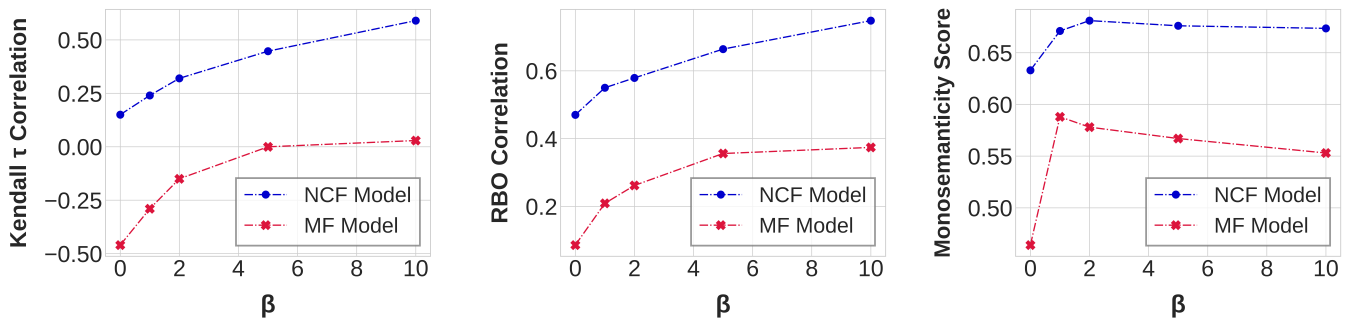


Figure 4: Effect of the prediction-aware loss \mathcal{L}_{pred} on recommendation fidelity and interpretability. Left and center: Rank-Biased Overlap (RBO) and Kendall-Tau correlation between the original and reconstructed top-30 recommendation lists improve with increasing weight β . Right: Monosemanticity score (Pach et al. 2025) peaks at intermediate values, highlighting a trade-off between fidelity and sparsity. Notably, $\beta=0$ corresponds to an ablation without \mathcal{L}_{pred} , underscoring its importance for alignment.

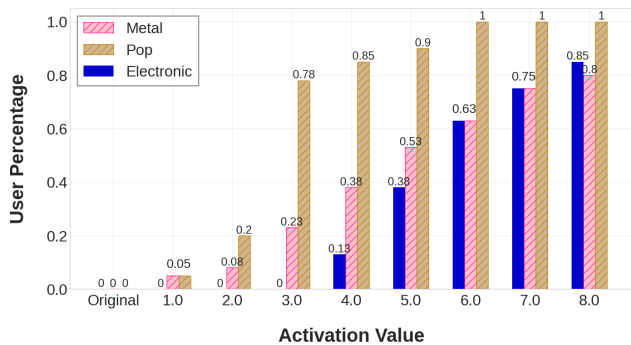


Figure 5: Targeted item promotion in Last.FM via neuron-level intervention. By increasing the activation of a genre-aligned neuron in Bob Dylan’s embedding vector \mathbf{z} (x-axis), the artist becomes relevant to users who prefer Metal, Contemporary Pop, and Electronic music, appearing in their top-30 recommendations despite no prior affinity.

promoted to audience segments with unrelated preferences. Complete results and visualizations are available in our public repository.

Takeaway. Monosemantic neurons provide actionable handles for post hoc control over recommender behavior. They enable precise, interpretable interventions aligned with user or item semantics, supporting applications in fairness, discovery, exposure-driven personalization, and content governance.

Hierarchical Structure with Matryoshka

To examine whether structured bottlenecks enhance interpretability, we integrated Matryoshka Sparse Autoencoders (SAEs) (Bussmann et al. 2025) into our framework. These models train nested dictionaries of increasing size, encouraging early neurons to capture broad features and later ones to specialize. We used a four-level hierarchy, with each level comprising one-quarter of the bottleneck neurons.

In MovieLens (ML1M), this structure had limited impact, likely due to modest genre diversity and a flat semantic space. In contrast, Last.FM revealed an interesting hierarchy: early neurons activated for broad audiences (e.g., mainstream electronic or metal listeners), while later neurons captured

narrower micro-genres (e.g., Nordic melodic death metal, early-2010s dream pop) and hybrid scenes (e.g., jazz-lounge electronica). Unlike LLMs, where early neurons encode general semantic categories, recommender hierarchies reflect both semantic granularity and audience scale. Popular tastes emerge earlier, while niche preferences appear deeper, an effect likely driven by our prediction-aware loss, which optimizes for real user-item preferences. Though not our core focus, these findings show that Matryoshka SAEs may potentially reveal layered audience structure, supporting more controllable and personalized mechanistic interventions.

Conclusion

We introduced a Sparse Autoencoder framework that extracts *monosemantic neurons*, interpreted as latent dimensions aligned with coherent and meaningful concepts within recommender embeddings. A central idea in our approach is a prediction aware reconstruction objective that preserves the user-item interaction patterns underlying recommendation behavior. Across two common recommender models and two datasets, these neurons consistently capture high level properties such as genres, popularity, and temporal trends, all discovered without supervision. While our primary focus is concept extraction, we also present preliminary neuron level interventions that lightly adjust recommendations, indicating promising directions for future controllability. More broadly, uncovering semantically grounded latent factors may help enable faithful explainability methods for recommender systems (Baklanov et al. 2025; Koenigstein 2025). As recommenders continue to shape information access and cultural exposure, developing methods that enhance transparency, interpretability, and accountability is increasingly important. We hope this work contributes to that direction by offering a foundation for more semantically aligned and trustworthy recommender technologies.

Acknowledgments

This work was supported by the Ministry of Innovation, Science & Technology, Israel.

References

- Abdollahi, B.; and Nasraoui, O. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the eleventh ACM conference on recommender systems*, 79–83.
- Baklanov, M.; Bogina, V.; Elisha, Y.; Schein, Y.; Allerhand, L.; Barkan, O.; and Koenigstein, N. 2025. Refining Fidelity Metrics for Explainable Recommendations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2967–2971.
- Bank, D.; Koenigstein, N.; and Giryas, R. 2023. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, 353–374.
- Barkan, O.; Bogina, V.; Gurevitch, L.; Asher, Y.; and Koenigstein, N. 2024. A Counterfactual Framework for Learning and Evaluating Explanations for Recommender Systems. In *Proceedings of the ACM on Web Conference 2024 (WWW '24)*, 3723–3733.
- Barkan, O.; Caciularu, A.; Rejwan, I.; Katz, O.; Weill, J.; Malkiel, I.; and Koenigstein, N. 2021a. Representation learning via variational bayesian networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 78–88.
- Barkan, O.; Hirsch, R.; Katz, O.; Caciularu, A.; and Koenigstein, N. 2021b. Anchor-based collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2877–2881.
- Barkan, O.; Hirsch, R.; Katz, O.; Caciularu, A.; Weill, J.; and Koenigstein, N. 2021c. Cold Item Integration in Deep Hybrid Recommenders via Tunable Stochastic Gates. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE.
- Barkan, O.; Katz, O.; and Koenigstein, N. 2020. Neural Attentive Multiview Machines. *IEEE ICASSP 2020*.
- Barkan, O.; Koenigstein, N.; Yogev, E.; and Katz, O. 2019. CB2CF: A Neural Multiview Content-to-Collaborative Filtering Model for Completely Cold Item Recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 228–236. ACM.
- Barkan, O.; Shaked, T.; Fuchs, Y.; and Koenigstein, N. 2023. Modeling users' heterogeneous taste with diversified attentive user profiles. *User Modeling and User-Adapted Interaction*, 1–31.
- Bertin-Mahieux, T.; Ellis, D. P.; Whitman, B.; and Lamere, P. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.
- Bussmann, B.; Nabeshima, N.; Karvonen, A.; and Nanda, N. 2025. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*.
- Chanin, D.; Dulka, T.; and Garriga-Alonso, A. 2025. Feature Hedging: Correlated Features Break Narrow Sparse Autoencoders. *arXiv preprint arXiv:2505.11756*.
- Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 191–198.
- Dang, Y.; Huang, K.; Huo, J.; Yan, Y.; Huang, S.; Liu, D.; Gao, M.; Zhang, J.; Qian, C.; Wang, K.; et al. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.
- Gaiger, K.; Barkan, O.; Tsipory-Samuel, S.; and Koenigstein, N. 2023. Not All Memories Created Equal: Dynamic User Representations for Collaborative Filtering. *IEEE Access*, 11: 34746–34763.
- Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.
- Ghazimatin, A. e. a. 2020. PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 196–204.
- Guo, S.; Zhang, S.; Sun, W.; Ren, P.; Chen, Z.; and Ren, Z. 2023. Towards explainable conversational recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2786–2795.
- Härle, R.; Friedrich, F.; Brack, M.; Wäldchen, S.; Deiseroth, B.; Schramowski, P.; and Kersting, K. 2025. Measuring and Guiding Monosemanticity. *arXiv preprint arXiv:2506.19382*.
- Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.
- He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. 639–648. Association for Computing Machinery.
- He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, 173–182.
- Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, 263–272. Ieee.
- Kinga, D.; Adam, J. B.; et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California,;
- Klimashevskaja, A.; Jannach, D.; Elahi, M.; and Trattner, C. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5): 1777–1834.
- Koenigstein, N. 2025. Without Fidelity, Explanations Are Just Stories: Rethinking Evaluation in Explainable Recommender Systems. *SSRN (September 26, 2025)*.
- Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.
- Makelev, A.; Lange, G.; and Nanda, N. 2024. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*.
- Melchiorre, A. B.; Rekabsaz, N.; Ganhör, C.; and Schedl, M. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *Proceedings*

- of the 16th ACM Conference on Recommender Systems, 246–256.
- O’Brien, K.; Majercak, D.; Fernandes, X.; Edgar, R.; Chen, J.; Nori, H.; Carignan, D.; Horvitz, E.; and Poursabzi-Sangde, F. 2024. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*.
- Pach, M.; Karthik, S.; Bouniot, Q.; Belongie, S.; and Akata, Z. 2025. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*.
- Paquet, U.; and Koenigstein, N. 2013. One-class collaborative filtering with random graphs. In *Proceedings of the 22nd international conference on World Wide Web*, 999–1008.
- Paulo, G.; Mallen, A.; Juang, C.; and Belrose, N. 2024. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*.
- Qin, P.; Huang, C.; Deng, Y.; Lei, W.; and Chua, T.-S. 2024. Beyond Persuasion: Towards Conversational Recommender System with Credible Explanations. *arXiv preprint arXiv:2409.14399*.
- Rendle, S. 2021. Item recommendation from implicit feedback. In *Recommender Systems Handbook*, 143–171. Springer.
- Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; Cunningham, H.; Turner, N. L.; McDougall, C.; MacDiarmid, M.; Freeman, C. D.; Sumers, T. R.; Rees, E.; Batson, J.; Jermyn, A.; Carter, S.; Olah, C.; and Henighan, T. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.
- Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103.
- Wang, Y.; Xiong, F.; Han, Z.; Song, Q.; Zhan, K.; and Wang, B. 2025. Unleashing the Potential of Two-Tower Models: Diffusion-Based Cross-Interaction for Large-Scale Matching. *arXiv preprint arXiv:2502.20687*.
- Yang, J.; Yi, X.; Zhiyuan Cheng, D.; Hong, L.; Li, Y.; Xiaoming Wang, S.; Xu, T.; and Chi, E. H. 2020. Mixed negative sampling for learning two-tower neural networks in recommendations. In *Companion proceedings of the web conference 2020*, 441–447.
- Yu, Y.; Wang, W.; Feng, Z.; and Xue, D. 2021. A dual augmented two-tower model for online large-scale recommendation. *DLP-KDD*.
- Zajrajew, V.; Baniecki, H.; and Biecek, P. 2025. Interpreting CLIP with Hierarchical Sparse Autoencoders. *arXiv preprint arXiv:2502.20578*.
- Zhang, Q.; Wang, Y.; Cui, J.; Pan, X.; Lei, Q.; Jegelka, S.; and Wang, Y. 2024a. Beyond Interpretability: The Gains of Feature Monosemanticity on Model Robustness. *arXiv preprint arXiv:2410.21331*.
- Zhang, Y.; Zhang, J.; Xu, F.; Chen, L.; Li, B.; Guo, L.; and Yin, H. 2024b. Preference Prototype-Aware Learning for Universal Cross-Domain Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3290–3299.
- Zhu, Z.; He, Y.; Zhao, X.; Zhang, Y.; Wang, J.; and Caverlee, J. 2021. Popularity-opportunity bias in collaborative filtering. In *Proceedings of the 14th ACM international conference on web search and data mining*, 85–93.