A Appendix

A.1 AMCF Evaluation Metrics:

AMCF also evaluate their general and specific preference. For general preference, similar to our evaluation metrics, their primary metric is *Top M recall at K*, i.e., the intersection between the top M ground truth user preference and top K predicted user preferences. To realize this metric, they need the ground truth user preference – which is not present in the dataset – and hence similar to our technique, they need to simulate. The way they compute the ground truth preference is unjustified and inexplicable. The process involves computing the weight of each item by removing the user and item bias terms, and then the preference of an attribute is just the sum of the weights of the items it occurs in. The latter part is still reasonable; however, we are uncertain about the weight calculation part. For evaluating specific preferences, they use the sorted order of general preference for the attributes present in that specific item – which does not resolve the concerns mentioned above.
We also use proxies to simulate the users' ground truth attribute preference; however there are key differences in our evaluation when compared to AMCF:

- The proxies are more generalizable and reasonable, like the conditional probability of liking and odds of liking.
- We use multiple proxies and report results on all of them to avoid cherry-picking.

A.2 AMCF Post-Hoc Adaptation:

For adapting AMCF [29] to be post-hoc, we made a minor modification to the architecture mentioned in the paper (see Figure 4). We froze the left-hand side of the architecture and only trained the right-hand side, which consists of the attention network that aims to reconstruct the item's embedding from its attributes. Since the user and item embeddings were not trained, this made the technique post-hoc.

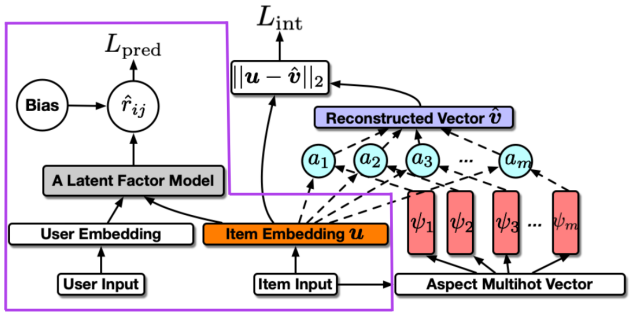

Figure 4: The architecture of AMCF Post-Hoc (AMCF-PH). The part within the blue colored region was not trained, thereby making the technique post-hoc.

A.3 Attribute Distribution Plots

Here we plot the distribution of genres for MOVIELENS-100*K* (Figure 5)and HETREC (Figure 6) datasets. There is a large skew in the distribution, with the top-3 genres and top-4 genres occurring in more than 88% and 95% of their items respectively.

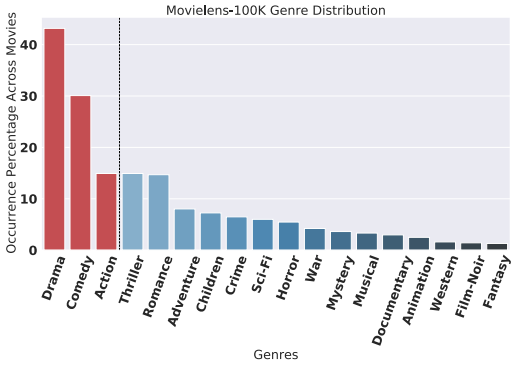

Figure 5: The genre distribution for MOVIELENS-100*K* dataset. The top-3 most popular genres: *Action*, *Comedy*, and *Drama* occur in over 88% of the movies. And therefore, global popularity can achieve very high test set coverage while providing uninformative and unpersonalized explanations.

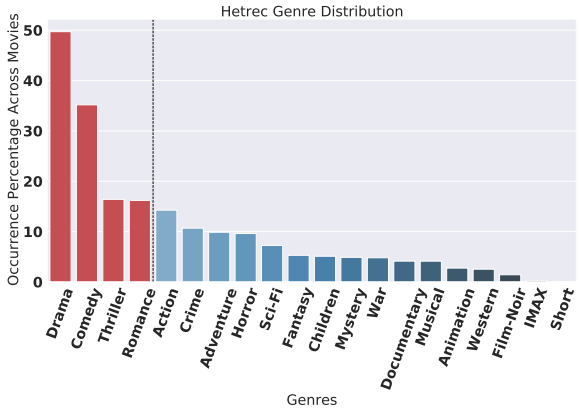

Figure 6: The genre distribution for HETREC dataset. The top-4 most popular genres: *Comedy*, *Drama*, *Romance*, and *Thriller* occur in over 95% of the movies. And therefore, global popularity can achieve very high test set coverage while providing uninformative and non-personalized explanations.

A.4 Choice of removal-based explanation technique:

Covert et al. [13] developed a framework to categorize such methods along three dimensions:

- Attribute removal: how the approach removes attributes from the model,
- Model behavior: what model behavior is it observing, and
- Summary technique: how does it summarize an attribute's impact?

`RecXplainer`, when instantiated in this framework: removes attributes by setting them to zero, analyzes prediction as the model behavior, and summarizes an attribute's impact by removing them individually.
Since our attribute input vector is binary and indicates whether an attribute is present or not, simulating an attribute's removal by setting it to zero is a natural choice. Previous removal-based explanation methods have used *prediction* or *prediction loss* or *dataset*

Table 8: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the MovieLens-100K dataset trained using Matrix Factorization when the feature importances are computed using one feature removal and using SHAP. The mean and standard deviation value for each metric is computed over 5 random seeds. The best results are highlighted in bold. One feature removal achives similar results to SHAP, but is 49 times faster.

Metrics	RX-Linear	RX-MLP	RX-GBDT	RX-Linear-SHAP	RX-MLP-SHAP	RX-GBDT-SHAP
Testset Coverage	60.74 ± 0.21	**67.16 ± 0.59**	63.01 ± 0.3	60.74 ± 0.21	**67.26 ± 0.43**	62.67 ± 0.53
Recommendations Coverage	**69.3 ± 2.4**	**65.23 ± 2.28**	64.27 ± 2.09	**69.3 ± 2.4**	**67.64 ± 2.28**	**65.73 ± 2.45**
CondProb Generalpref Coverage	64.77 ± 1.75	60.78 ± 1.59	71.24 ± 1.12	64.77 ± 1.75	62.76 ± 1.0	**73.17 ± 0.39**
CondProb Generalpref Ranking	15.8 ± 0.34	15.68 ± 0.17	20.28 ± 0.3	15.8 ± 0.34	16.47 ± 0.28	**20.95 ± 0.34**
CondProb Specificpref Coverage	49.78 ± 0.48	49.06 ± 0.22	**51.23 ± 0.27**	49.78 ± 0.48	50.43 ± 0.16	**51.83 ± 0.36**
CondProb Specificpref Ranking	51.62 ± 0.36	52.59 ± 0.27	55.98 ± 0.22	51.62 ± 0.36	53.67 ± 0.37	**56.69 ± 0.3**
Odds Generalpref Coverage	74.42 ± 1.05	72.47 ± 1.16	81.15 ± 0.55	74.42 ± 1.05	74.02 ± 0.93	**83.18 ± 0.27**
Odds Generalpref Ranking	25.87 ± 0.33	29.3 ± 0.83	**33.57 ± 0.39**	25.87 ± 0.33	30.8 ± 0.84	**34.2 ± 0.34**
Odds Specificpref Coverage	56.17 ± 0.5	55.56 ± 0.6	**57.74 ± 0.46**	56.17 ± 0.5	**57.8 ± 0.55**	**58.41 ± 0.46**
Odds Specificpref Ranking	50.34 ± 0.12	52.54 ± 0.28	55.33 ± 0.12	50.34 ± 0.12	53.14 ± 0.28	**55.69 ± 0.12**

Table 9: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the Hetrec dataset trained using Matrix Factorization when the feature importances are computed using one feature removal and using SHAP. The mean and standard deviation value for each metric is computed over 5 random seeds. The best results are highlighted in bold. One feature removal achives similar results to SHAP, but is 90 times faster.

Metrics	RX-Linear	RX-MLP	RX-GBDT	RX-Linear-SHAP	RX-MLP-SHAP	RX-GBDT-SHAP
Testset Coverage	75.46 ± 0.34	**80.98 ± 1.18**	78.77 ± 0.18	75.45 ± 0.33	**82.13 ± 1.43**	78.16 ± 0.24
Recommendations Coverage	**81.4 ± 1.22**	**79.65 ± 2.59**	78.98 ± 0.88	**81.4 ± 1.22**	**82.55 ± 2.38**	**81.25 ± 0.95**
CondProb Generalpref Coverage	84.61 ± 0.42	78.63 ± 2.16	84.56 ± 0.45	84.63 ± 0.44	76.71 ± 1.3	**85.75 ± 0.63**
CondProb Generalpref Ranking	17.72 ± 0.19	16.4 ± 0.62	**19.55 ± 0.37**	17.73 ± 0.18	15.98 ± 0.63	**20.06 ± 0.26**
CondProb Specificpref Coverage	**63.52 ± 0.45**	59.87 ± 0.45	61.86 ± 0.47	**63.52 ± 0.45**	60.19 ± 0.83	**63.06 ± 0.46**
CondProb Specificpref Ranking	54.26 ± 0.06	54.13 ± 0.47	57.46 ± 0.09	54.26 ± 0.06	55.76 ± 0.5	**58.69 ± 0.1**
Odds Generalpref Coverage	**85.15 ± 0.5**	79.01 ± 2.28	84.79 ± 0.53	**85.16 ± 0.49**	76.7 ± 0.81	**85.85 ± 0.49**
Odds Generalpref Ranking	23.32 ± 0.28	23.08 ± 0.48	26.0 ± 0.22	23.32 ± 0.28	23.64 ± 0.36	**26.72 ± 0.3**
Odds Specificpref Coverage	**65.78 ± 0.42**	63.05 ± 0.38	64.86 ± 0.29	**65.78 ± 0.42**	63.79 ± 0.13	**65.83 ± 0.32**
Odds Specificpref Ranking	53.49 ± 0.14	53.36 ± 0.35	56.35 ± 0.04	53.49 ± 0.14	54.91 ± 0.29	**57.68 ± 0.08**

loss for model behavior analysis. We chose *prediction* instead of *prediction loss* for our analysis because we wanted to get specific preference (analog of local interpretability) without requiring the original rating. Previous removal-based explanation methods have used *removing individual attributes* or *Shapley values* or *trained additive models* to get attribute impact value. Removing individual attributes accesses the impact of an attribute by measuring the loss in prediction when that one attribute is removed, and this is what we choose. On the other hand, Shapley value takes all subsets of attributes and then use the cooperative game theoretic formulation to assign impact value to each attribute. It has two disadvantages:

- It creates all subsets of attributes – which is exponential in the number of attributes, making the process very expensive.
- For creating all the subsets, it simulates removing many features that can potentially create attribute vectors that the auxiliary model has not seen, and thereby its prediction can not be trusted in that part of the data manifold.

`RecXplainer` is by design agnostic to the specific feature removal technique. Table 8 and Table 9 show the comparison of using one feature removal and SHAP for feature attribution for MovieLens-100K and Hetrec datasets trained using Matrix Factorization models. In both the cases, we find that SHAP attains slightly better values for most metrics, however, it is much more expensive than using one feature removal technique (**49X** and **90X** respectively). For this reason, we choose the method of removing one feature as the default choice for feature attribution in `RecXplainer`.

A.5 Justification of using Rank-Biased Overlap (RBO):

Comparing ranked lists can be achieved using various rank correlation metrics like Kendall's Tau and Spearman's correlation [46]. These metrics have restrictions that the ranked lists must be conjoint, i.e., they both must contain all the items that the entire universe of items contains. This is not suitable for our metrics because the top-k attributes from either the conditional probability

or the odds liking proxy might not have any common element with the ranking produced by any technique. Hence we choose to compare the ranked lists using rank-biased overlap (RBO) metric [45], which was recently proposed to overcome the limitation posed by correlation-based metrics. Specifically, RBO is better than correlation-based metrics as:

- RBO does not require the two ranked lists to be conjoint, i.e., a ranked list having items that do not occur in the other list is acceptable, e.g., the similarity between [1, 3, 7] and [1, 5, 8] can be measured using RBO.
- RBO does not require the two ranked lists to be of the same length.
- RBO provides weighted comparison, i.e., discordance at higher ranks is penalized more than discordance at lower ranks.

A.6 Ablation Experiment Plots

We plot the change in four metrics as the users with ratings less than a threshold are removed. We report the trends in these metrics in Figure 7 for MovieLens-100*K* and Figure 8 for Hetrec dataset respectively. For `RecXplainer`, we show the metrics using the linear layer as its auxiliary model.

A.7 Metrics for all architectures for MovieLens-100K dataset

In this section, we experiment with all five architectures for the MovieLens-100K dataset. The goal is to demonstrate the effectiveness of `RecXplainer` across a wide variety of collaborative filtering architectures. For each architecture, we train the best model using an extensive hyperparameter search. Across the five models, our conclusions regarding the comparison of `RecXplainer` and the baselines are similar to the ones presented in the evaluation section:

(1) `RecXplainer` performs better than LIME-RS across all five models and for all metrics
(2) `RecXplainer` performs better than AMCF-PH across all five models and for all metrics
(3) Global popularity performs better than `RecXplainer` for test set coverage and recommendation coverage, however `RecXplainer` performs better than it for all eight personalization metrics across all five models.
(4) User specific popularity performs second best for test set and recommendations coverage, however `RecXplainer` performs better than it for all eight personalization metrics across all five models.

A.8 Case Studies

In this section we dive into two case studies demonstrating the use of `RecXplainer`.

A. A case study where the popularity of a genre is insufficient to understand user preferences:
A user has liked 4 items (rated 4 or higher) and disliked 23 items (rated 3 or lower) in the Movielens-100K dataset. The liked movies and their respective genres are:

(1) Good Will Hunting: Drama
(2) Apt Pupil: (Drama, Thriller)
(3) Fast, Cheap & Out of Control: Documentary
(4) Welcome To Sarajevo: (Drama, War)

RecXplainer identifies that the top-3 genre preferences of this user are: Documentary, Drama, and War, the user popularity based method ranks Drama higher than Documentary. RecXplainer captures that this user prefers Documentary more than Drama, since 7 among 23 items this user disliked belonged to the Drama genre.

B. A case study where RecXplainer better explains top recommendations:
A user has liked 4 items and disliked 10 items in the Movielens-100K dataset, and their liked movies are:

(1) Ulee's Gold: Drama
(2) Everyone Says I Love You: (Comedy, Musical, Romance)
(3) Wag the Dog: (Comedy, Drama)
(4) Kundun: Drama

RecXplainer identifies that the top-3 genre preferences of this user are: Comedy, Musical, and Drama. Here are the top-3 recommendations to the user:

(1) Casablanca: (Drama, Romance, War)
(2) A Close Shave: (Animation, Comedy, Thriller)
(3) The Wrong Trousers: (Animation, Comedy),

Since 2 out of the top-3 movies recommended to this uses belong to the Comedy genre, this order of attribute preferences effectively explains the top-3 recommendations to this user.

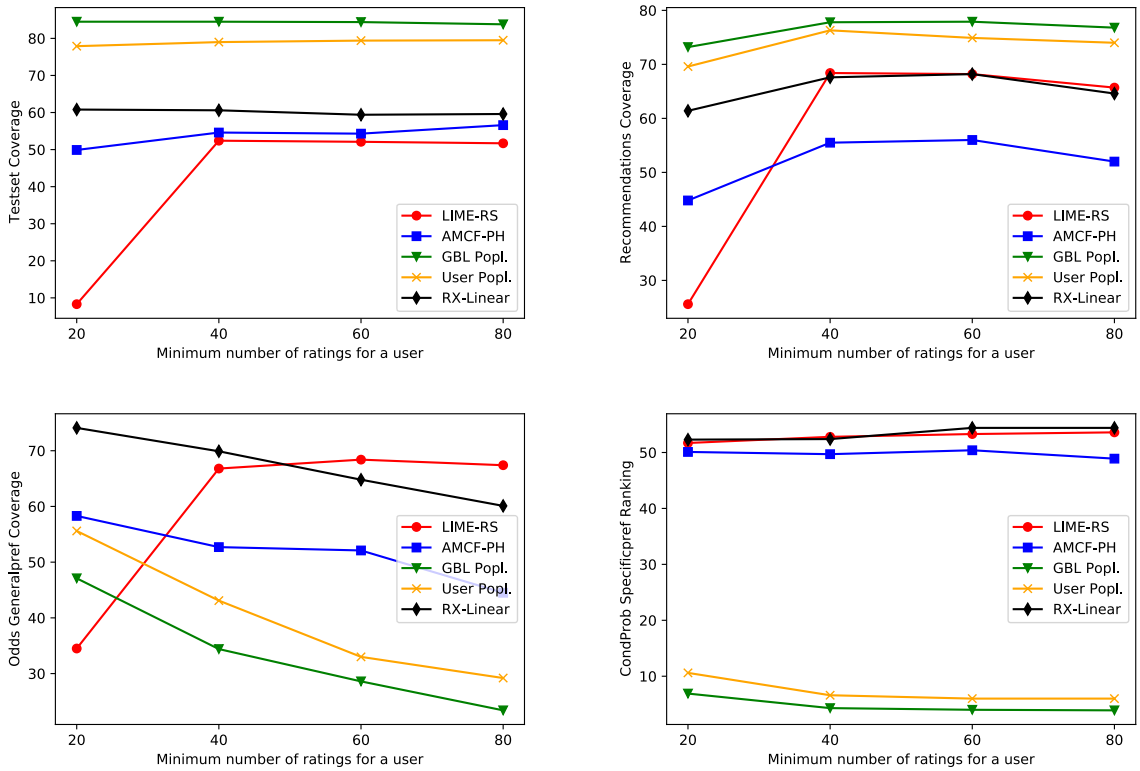

Figure 7: Ablation experiments on Movielens 100k dataset. We only retain users that have rated a minimum of 20, 40, 60, or 80 ratings. The first row shows the test set coverage and recommendations coverage. `RecXplainer` performs consistently ranks the third best across all the minimum user ratings. The second row shows the general preference coverage and ranking of specific preferences. `RecXplainer` ranks first for specific preference ranking across all minimum user ratings. For general preference coverage, `RecXplainer` ranks the best at lower minimum user ratings and second best afterwards.

Table 10: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the MovieLens-100k dataset trained using Matrix Factorization. The mean and standard deviation value for each metric is computed over 5 random seeds. We compare 3 architecture choices for auxiliary modelling of RecXplainer. The best results are highlighted in bold.

Metrics	LIME-RS	AMCF-PH	GBL Popl.	User Popl.	Random	RX-Linear	RX-MLP	RX-GBDT
Testset Coverage	57.42 ± 0.47	49.62 ± 4.59	**84.69 ± 0.18**	77.6 ± 0.59	32.83 ± 0.57	60.74 ± 0.21	67.16 ± 0.59	63.01 ± 0.3
Recommendations Coverage	67.08 ± 2.68	44.2 ± 2.58	**79.12 ± 2.83**	70.12 ± 2.05	26.72 ± 0.89	69.3 ± 2.4	65.23 ± 2.28	64.27 ± 2.09
CondProb Generalpref Coverage	59.85 ± 0.56	54.57 ± 1.73	25.07 ± 0.58	41.51 ± 0.44	45.07 ± 0.91	64.77 ± 1.75	60.78 ± 1.59	**71.24 ± 1.12**
CondProb Generalpref Ranking	13.99 ± 0.41	13.74 ± 0.39	5.04 ± 0.16	9.76 ± 0.23	11.64 ± 0.29	15.8 ± 0.34	15.68 ± 0.17	**20.28 ± 0.3**
CondProb Specificpref Coverage	48.87 ± 0.25	**49.97 ± 1.14**	25.1 ± 0.58	41.44 ± 0.44	43.94 ± 0.5	49.78 ± 0.48	49.06 ± 0.22	**51.23 ± 0.27**
CondProb Specificpref Ranking	51.24 ± 0.15	49.89 ± 0.32	6.29 ± 0.11	10.29 ± 0.25	11.71 ± 0.14	51.62 ± 0.36	52.59 ± 0.27	**55.98 ± 0.22**
Odds Generalpref Coverage	69.31 ± 1.43	60.98 ± 1.46	47.42 ± 0.55	55.78 ± 0.84	44.39 ± 0.49	74.42 ± 1.05	72.47 ± 1.16	**81.15 ± 0.55**
Odds Generalpref Ranking	23.25 ± 0.58	18.94 ± 0.8	17.72 ± 0.16	27.63 ± 0.4	11.25 ± 0.32	25.87 ± 0.33	29.3 ± 0.83	**33.57 ± 0.39**
Odds Specificpref Coverage	54.96 ± 0.62	51.8 ± 0.93	47.47 ± 0.55	55.73 ± 0.84	44.08 ± 0.24	56.17 ± 0.5	55.56 ± 0.6	**57.74 ± 0.46**
Odds Specificpref Ranking	50.67 ± 0.16	50.07 ± 0.44	19.04 ± 0.32	28.04 ± 0.4	11.68 ± 0.1	50.34 ± 0.12	52.54 ± 0.28	**55.33 ± 0.12**

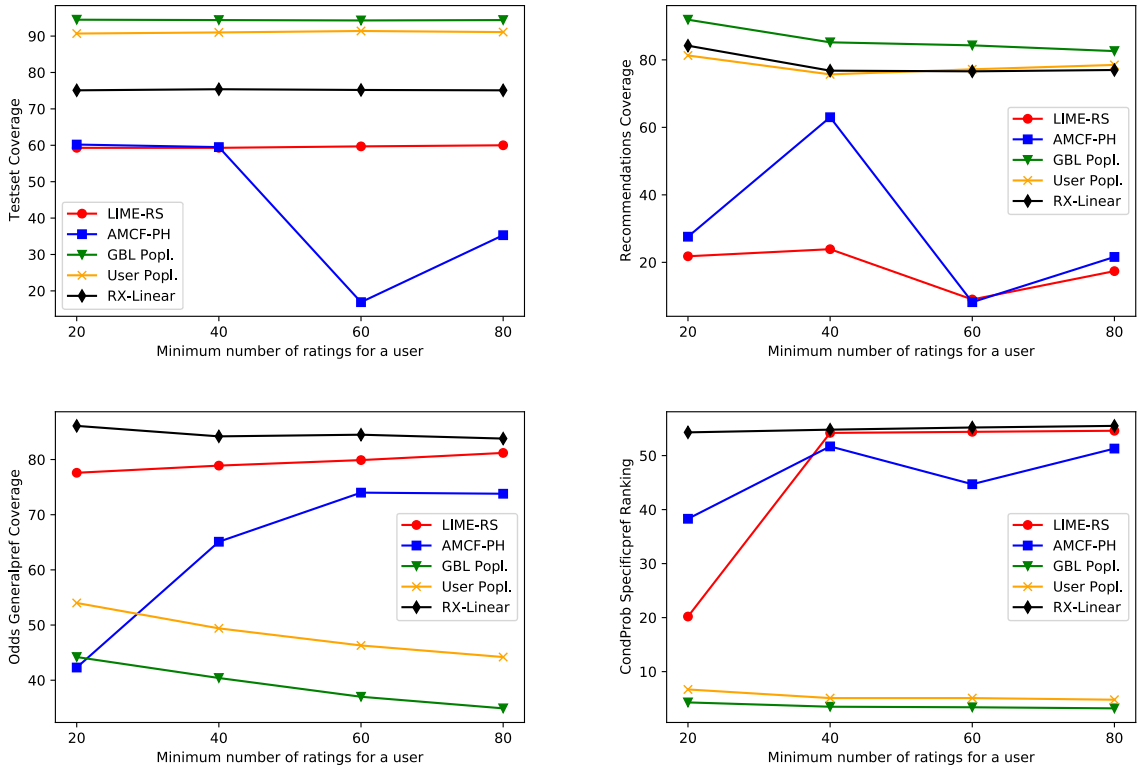

Figure 8: Ablation experiments on Hetrec Movielens dataset. We only retain users that have rated a minimum of 20, 40, 60, or 80 ratings. The first row shows the test set coverage and recommendations coverage. `RecXplainer` performs consistently ranks the third best across all the minimum user ratings for test set coverage. For recommendations coverage, it is the second best for settings 20 and 40, and the third best for settings 60 and 80. The second row shows the general preference coverage and ranking of specific preferences. `RecXplainer` ranks first consistently across all settings.

Table 11: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the MovieLens-100k dataset trained using Autoencoder. The mean and standard deviation value for each metric is computed over 5 random seeds. We compare 3 architecture choices for auxiliary modelling of RecXplainer. The best results are highlighted in bold.

Metrics	LIME-RS	AMCF-PH	GBL Popl.	User Popl.	Random	RX-Linear	RX-MLP	RX-GBDT
Testset Coverage	58.09 ± 0.55	47.91 ± 15.0	**84.69 ± 0.18**	77.88 ± 0.23	32.83 ± 0.57	60.52 ± 0.24	66.66 ± 1.16	64.22 ± 0.66
Recommendations Coverage	64.06 ± 3.71	44.71 ± 21.94	**79.07 ± 3.84**	69.49 ± 2.87	25.24 ± 0.72	68.73 ± 3.95	65.58 ± 2.13	63.76 ± 2.68
CondProb Generalpref Coverage	57.37 ± 1.76	44.52 ± 8.15	27.59 ± 0.69	42.42 ± 0.32	44.28 ± 1.26	65.51 ± 0.55	60.06 ± 0.42	**70.56 ± 1.28**
CondProb Generalpref Ranking	13.47 ± 0.43	10.55 ± 3.35	5.55 ± 0.17	10.13 ± 0.2	11.21 ± 0.54	16.03 ± 0.35	14.67 ± 0.56	**19.71 ± 0.7**
CondProb Specificpref Coverage	48.53 ± 0.37	44.84 ± 0.88	27.52 ± 0.69	42.36 ± 0.32	43.9 ± 0.28	49.96 ± 0.34	48.01 ± 0.29	**50.98 ± 0.43**
CondProb Specificpref Ranking	50.85 ± 0.14	47.11 ± 1.77	6.8 ± 0.17	10.57 ± 0.19	11.65 ± 0.13	51.58 ± 0.3	51.6 ± 0.28	**55.4 ± 0.27**
Odds Generalpref Coverage	67.95 ± 1.29	48.46 ± 4.32	47.38 ± 0.83	56.06 ± 0.98	44.05 ± 0.63	74.93 ± 0.67	70.97 ± 0.92	**80.93 ± 1.83**
Odds Generalpref Ranking	22.25 ± 0.4	12.94 ± 2.23	18.2 ± 0.28	27.37 ± 0.48	11.2 ± 0.18	26.09 ± 0.41	27.33 ± 0.78	**32.65 ± 0.61**
Odds Specificpref Coverage	55.08 ± 0.59	54.34 ± 0.65	47.43 ± 0.83	56.01 ± 0.98	44.01 ± 0.19	56.28 ± 0.55	54.84 ± 0.88	**57.81 ± 0.63**
Odds Specificpref Ranking	50.39 ± 0.07	48.01 ± 1.13	19.76 ± 0.43	27.8 ± 0.5	11.66 ± 0.09	50.46 ± 0.17	51.89 ± 0.1	**55.07 ± 0.14**

Table 12: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the MovieLens-100K dataset trained using Neural CF. The mean and standard deviation value for each metric is computed over 5 random seeds. We compare 3 architecture choices for auxiliary modelling of RecXplainer. The best results are highlighted in bold.

Metrics	LIME-RS	AMCF-PH	GBL Popl.	User Popl.	Random	RX-Linear	RX-MLP	RX-GBDT
Testset Coverage	23.26 ± 6.14	39.11 ± 10.17	**84.69 ± 0.18**	77.88 ± 0.23	32.83 ± 0.57	59.44 ± 0.92	66.6 ± 1.79	63.0 ± 0.46
Recommendations Coverage	19.05 ± 4.81	44.14 ± 9.77	**85.0 ± 0.0**	76.31 ± 0.37	30.72 ± 0.41	64.4 ± 1.6	72.81 ± 2.36	67.25 ± 0.65
CondProb Generalpref Coverage	43.9 ± 7.77	40.25 ± 4.08	27.59 ± 0.69	42.42 ± 0.32	44.28 ± 1.26	**67.47 ± 0.99**	53.36 ± 1.37	**67.91 ± 1.21**
CondProb Generalpref Ranking	11.44 ± 3.02	9.63 ± 1.04	5.55 ± 0.17	10.13 ± 0.2	11.21 ± 0.54	16.63 ± 0.41	12.94 ± 0.39	**18.25 ± 0.3**
CondProb Specificpref Coverage	44.08 ± 0.87	43.19 ± 1.77	27.52 ± 0.69	42.36 ± 0.32	43.9 ± 0.28	**50.0 ± 0.4**	46.6 ± 0.47	**49.75 ± 0.35**
CondProb Specificpref Ranking	47.79 ± 0.46	47.52 ± 0.38	6.8 ± 0.17	10.57 ± 0.19	11.65 ± 0.13	51.83 ± 0.57	49.65 ± 0.32	**53.47 ± 0.17**
Odds Generalpref Coverage	37.09 ± 2.15	42.95 ± 7.04	47.38 ± 0.83	56.06 ± 0.98	44.05 ± 0.63	74.89 ± 0.76	65.54 ± 1.78	**77.45 ± 1.73**
Odds Generalpref Ranking	8.71 ± 0.71	11.3 ± 3.62	18.2 ± 0.28	27.37 ± 0.48	11.2 ± 0.18	25.62 ± 0.31	23.51 ± 1.07	**29.33 ± 0.49**
Odds Specificpref Coverage	45.95 ± 1.0	48.66 ± 4.02	47.43 ± 0.83	**56.01 ± 0.98**	44.01 ± 0.19	**55.03 ± 1.12**	51.89 ± 1.04	**55.48 ± 0.5**
Odds Specificpref Ranking	47.9 ± 0.2	48.19 ± 0.64	19.76 ± 0.43	27.8 ± 0.5	11.66 ± 0.09	50.45 ± 0.09	49.58 ± 0.18	**52.64 ± 0.22**

Table 13: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the MovieLens-100K dataset trained using Factorization Machine. The mean and standard deviation value for each metric is computed over 5 random seeds. We compare 3 architecture choices for auxiliary modelling of RecXplainer. The best results are highlighted in bold.

Metrics	LIME-RS	AMCF-PH	GBL Popl.	User Popl.	Random	RX-Linear	RX-MLP	RX-GBDT
Testset Coverage	56.6 ± 1.04	75.69 ± 1.0	**84.69 ± 0.18**	77.88 ± 0.23	32.83 ± 0.57	60.87 ± 0.53	67.61 ± 1.8	63.51 ± 0.4
Recommendations Coverage	62.8 ± 3.03	**74.4 ± 2.39**	**79.12 ± 4.05**	72.66 ± 2.11	31.88 ± 1.2	64.97 ± 3.29	68.41 ± 1.92	66.02 ± 2.78
CondProb Generalpref Coverage	61.61 ± 1.14	38.58 ± 2.0	27.59 ± 0.69	42.42 ± 0.32	44.28 ± 1.26	64.5 ± 1.11	61.04 ± 1.36	**71.37 ± 0.78**
CondProb Generalpref Ranking	15.68 ± 0.41	8.33 ± 0.42	5.55 ± 0.17	10.13 ± 0.2	11.21 ± 0.54	16.01 ± 0.39	15.21 ± 0.67	**20.16 ± 0.47**
CondProb Specificpref Coverage	48.74 ± 0.3	47.55 ± 0.73	27.52 ± 0.69	42.36 ± 0.32	43.9 ± 0.28	49.95 ± 0.33	48.53 ± 0.24	**51.15 ± 0.22**
CondProb Specificpref Ranking	51.41 ± 0.24	47.06 ± 0.48	6.8 ± 0.17	10.57 ± 0.19	11.65 ± 0.13	52.16 ± 0.29	51.75 ± 0.46	**55.85 ± 0.13**
Odds Generalpref Coverage	71.54 ± 1.43	55.06 ± 1.42	47.38 ± 0.83	56.06 ± 0.98	44.05 ± 0.63	74.66 ± 1.19	71.9 ± 1.37	**80.85 ± 0.77**
Odds Generalpref Ranking	23.73 ± 0.83	20.65 ± 0.55	18.2 ± 0.28	27.37 ± 0.48	11.2 ± 0.18	26.16 ± 0.33	28.17 ± 0.79	**33.51 ± 0.51**
Odds Specificpref Coverage	55.15 ± 0.65	56.58 ± 0.55	47.43 ± 0.83	56.01 ± 0.98	44.01 ± 0.19	56.25 ± 0.52	55.09 ± 0.92	**57.95 ± 0.63**
Odds Specificpref Ranking	50.77 ± 0.22	48.96 ± 0.28	19.76 ± 0.43	27.8 ± 0.5	11.66 ± 0.09	50.5 ± 0.13	52.08 ± 0.18	**55.26 ± 0.09**

Table 14: The test set coverage, top-k recommendations coverage, and the 8 explanation personalization metrics are reported here (averaged over all users) for the MovieLens-100K dataset trained using Deep Factorization Machine. The mean and standard deviation value for each metric is computed over 5 random seeds. We compare 3 architecture choices for auxiliary modelling of RecXplainer. The best results are highlighted in bold.

Metrics	LIME-RS	AMCF-PH	GBL Popl.	User Popl.	Random	RX-Linear	RX-MLP	RX-GBDT
Testset Coverage	54.82 ± 0.86	50.58 ± 0.77	**84.69 ± 0.18**	77.88 ± 0.23	32.83 ± 0.57	60.35 ± 2.11	68.37 ± 2.82	62.92 ± 0.53
Recommendations Coverage	66.78 ± 1.17	51.93 ± 1.17	**79.7 ± 2.13**	73.2 ± 1.34	34.2 ± 0.7	67.51 ± 0.93	70.48 ± 1.57	65.93 ± 0.81
CondProb Generalpref Coverage	53.76 ± 1.24	43.39 ± 2.34	27.59 ± 0.69	42.42 ± 0.32	44.28 ± 1.26	63.16 ± 1.19	53.72 ± 1.79	**69.52 ± 1.25**
CondProb Generalpref Ranking	12.31 ± 0.55	9.33 ± 0.77	5.55 ± 0.17	10.13 ± 0.2	11.21 ± 0.54	15.1 ± 0.55	12.85 ± 0.68	**19.11 ± 0.41**
CondProb Specificpref Coverage	47.21 ± 0.24	44.64 ± 0.89	27.52 ± 0.69	42.36 ± 0.32	43.9 ± 0.28	**49.59 ± 0.68**	47.13 ± 0.54	**50.19 ± 0.26**
CondProb Specificpref Ranking	50.29 ± 0.11	48.69 ± 0.19	6.8 ± 0.17	10.57 ± 0.19	11.65 ± 0.13	51.69 ± 0.61	49.96 ± 0.26	**54.08 ± 0.13**
Odds Generalpref Coverage	62.74 ± 1.56	52.81 ± 2.74	47.38 ± 0.83	56.06 ± 0.98	44.05 ± 0.63	73.23 ± 0.75	66.21 ± 1.2	**80.11 ± 1.69**
Odds Generalpref Ranking	19.56 ± 0.39	15.96 ± 1.32	18.2 ± 0.28	27.37 ± 0.48	11.2 ± 0.18	25.37 ± 0.39	24.72 ± 0.36	**30.97 ± 0.3**
Odds Specificpref Coverage	53.27 ± 0.62	50.5 ± 1.18	47.43 ± 0.83	**56.01 ± 0.98**	44.01 ± 0.19	**55.47 ± 0.76**	53.32 ± 0.74	**56.42 ± 0.61**
Odds Specificpref Ranking	49.67 ± 0.12	49.36 ± 0.34	19.76 ± 0.43	27.8 ± 0.5	11.66 ± 0.09	50.37 ± 0.14	50.13 ± 0.19	**53.79 ± 0.21**