New Insights into Metric Optimization for Ranking-based Recommendation

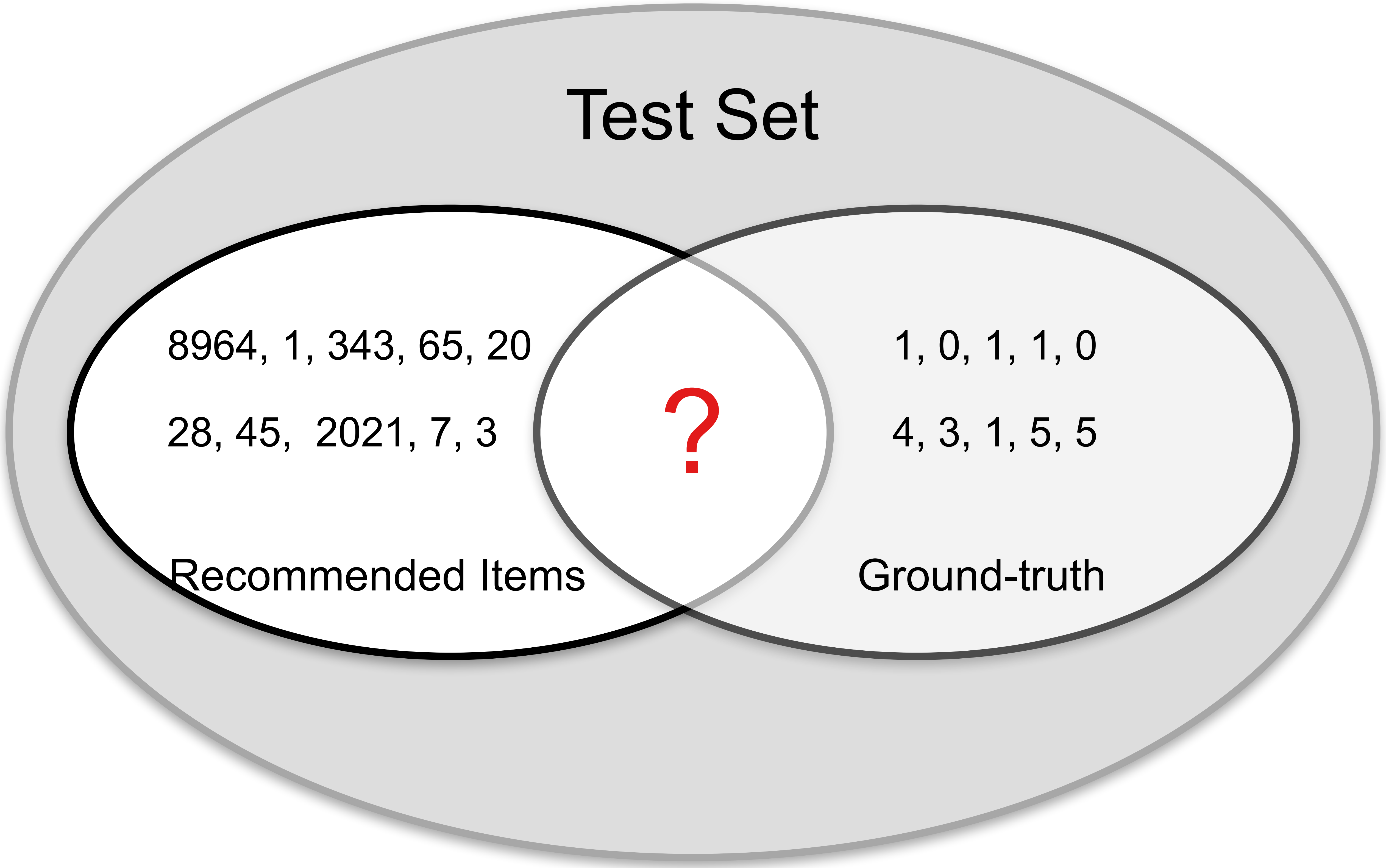
Roger Zhe Li (Presenter), Julián Urbano, Alan Hanjalic

Delft University of Technology, the Netherlands

Email: z.li-9@tudelft.nl

@Zhe_Delft

Offline Evaluation in Recommender Systems



nDCG

AP

RR

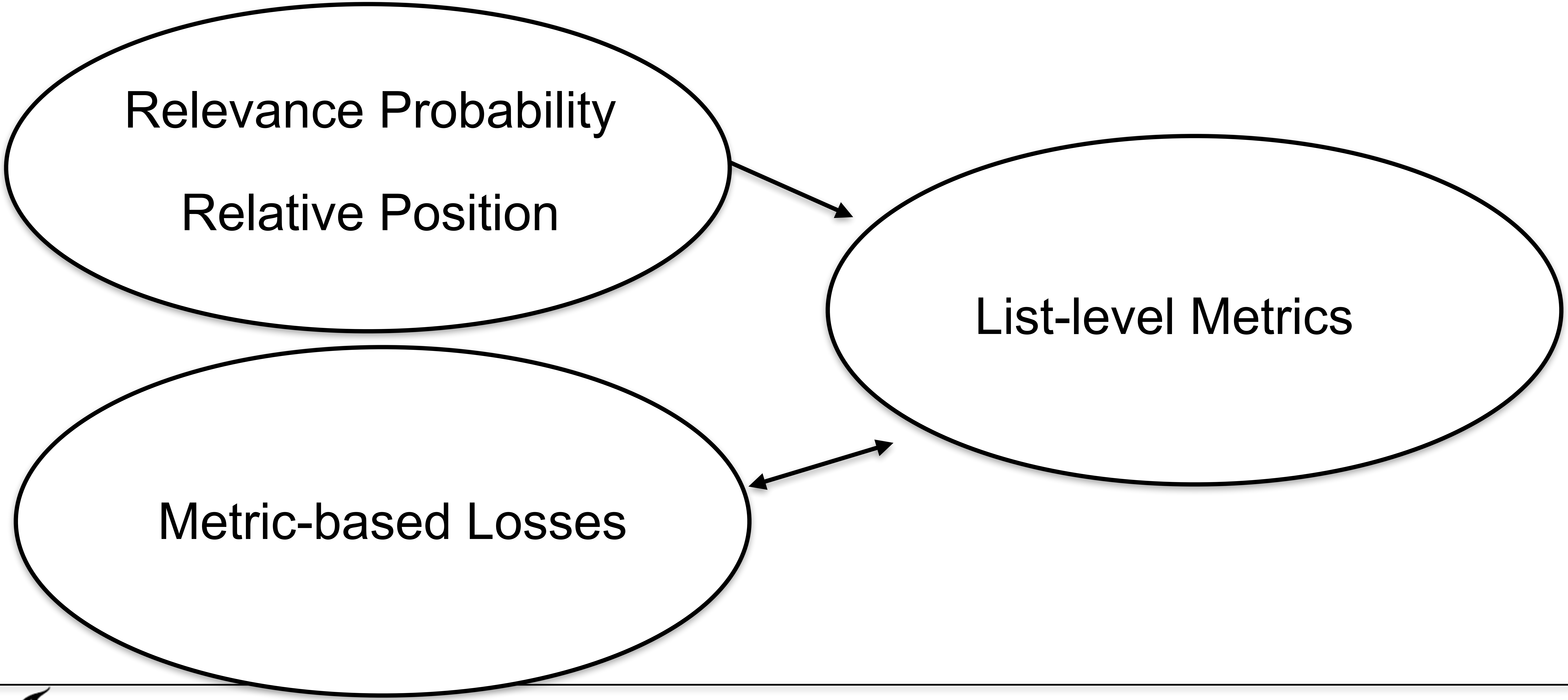
Precision

Recall

…

Optimization

Evaluation



Optimizing for the Same Metric Used for Evaluation?

	Evaluation Metric	Optimization Target
CLiMF (Shi et al, 2012)	RR	RR
TFMAP (Shi et al, 2012)	AP, Precision	AP
Top-N-Rank (Liang et al, 2018)	nDCG	DCG
LambdaRank (Burges et al, 2006)	nDCG	DCG

Is “Optimizing for the Same Metric Used for Evaluation” the BEST Way?

Concerns

- Some metrics are more informative than others;
- Metrics are correlated with each other to a different extent.

Problem

- Goal: investigate the choice of metric to optimize for a recommender.

- Given: {user, item, BINARY relevances}.

- Target: Extensive comparison (effectivess, fairness, etc) on personalized recommendation lists to each user, optimized by different IR metrics.

Strategies

- Pairwise (LambdaRank) and listwise methods for investigation;
- Four metrics: nDCG, AP, RR and RBP(s);
- Different data sparsities for training and testing.

Loss Design for Direct Optimization

Loss: Preliminaries

	nDCG	AP	RR	RBP
LambdaRank	Donmez et al, 2009			?
Listwise	Top-N-Rank, Liang et al, 2018	TFMAP, Shi et al, 2012	CLiMF, Shi et al, 2012	

Optimizing for nRBP

	nDCG	AP	RR	RBP
Range	[0, 1]	[0, 1]	[0, 1]	[0, <1]

DCG -> nDCG

↓

RBP -> nRBP

Optimizing for nRBP: Listwise

$$L_{nRBP}(u) = \sum_{i=1}^{N} y_{ui}(\tilde{R}_{ui} - 1) - \sum_{j=1}^{m_u} (j - 1)$$

- Optimize for an upper bound based on logarithmic transformation and Jensen's inequality;
- Independent of the hyperparameter *p;*
- Lower bound = 0; upper bound not fixed;
- Active users with more items are more important.

Experiments

Datasets

Dataset	#users	#items	#ratings	Density
CiteULike-a	2,465	16,702	157,527	0.383%
Epinions	4,690	32,592	325,154	0.213%
Sports & Outdoors	9,123	119,404	342,311	0.031%
Home & Kitchen	20,531	222,472	795,845	0.017%

Binary

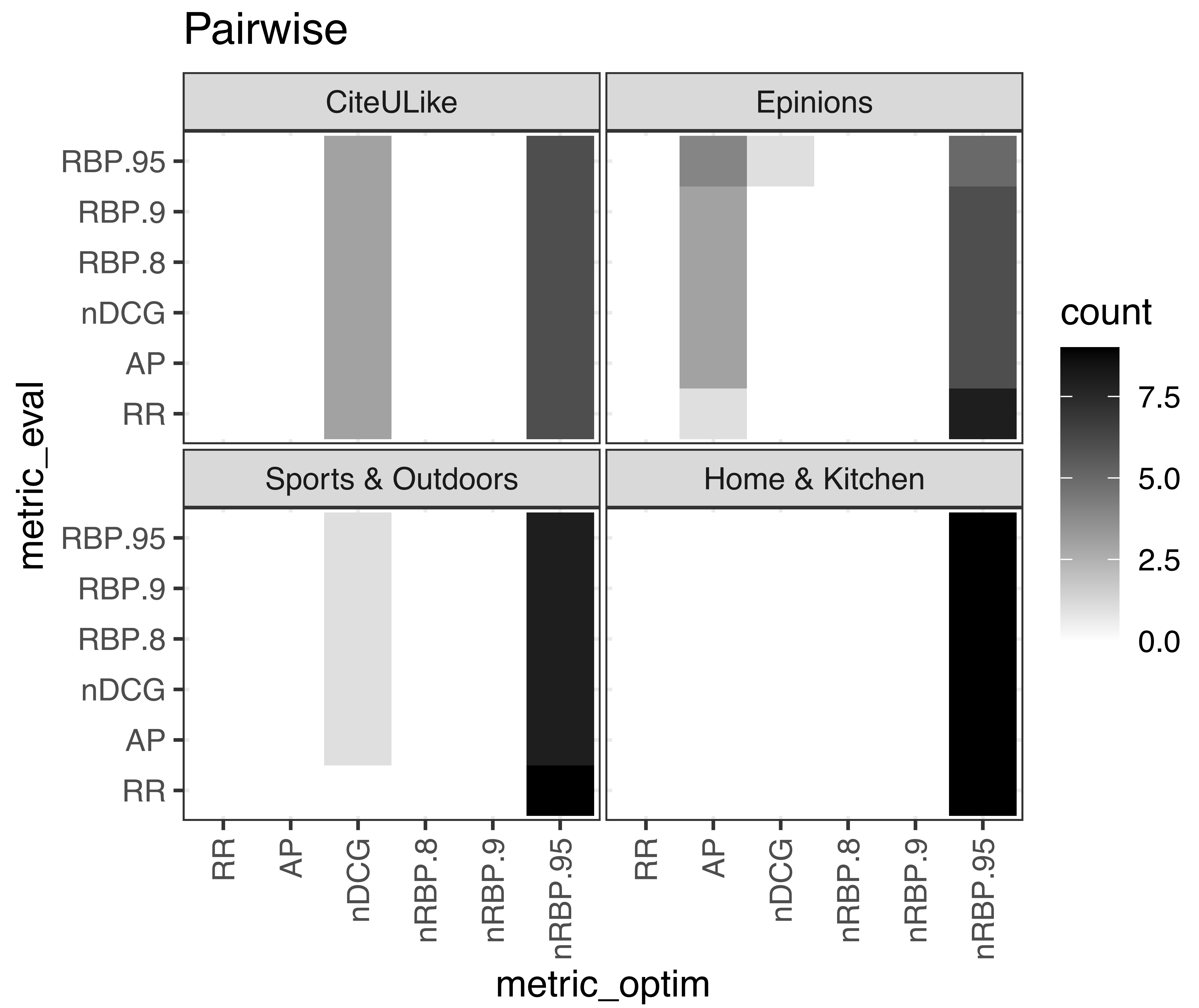
Graded 1-5

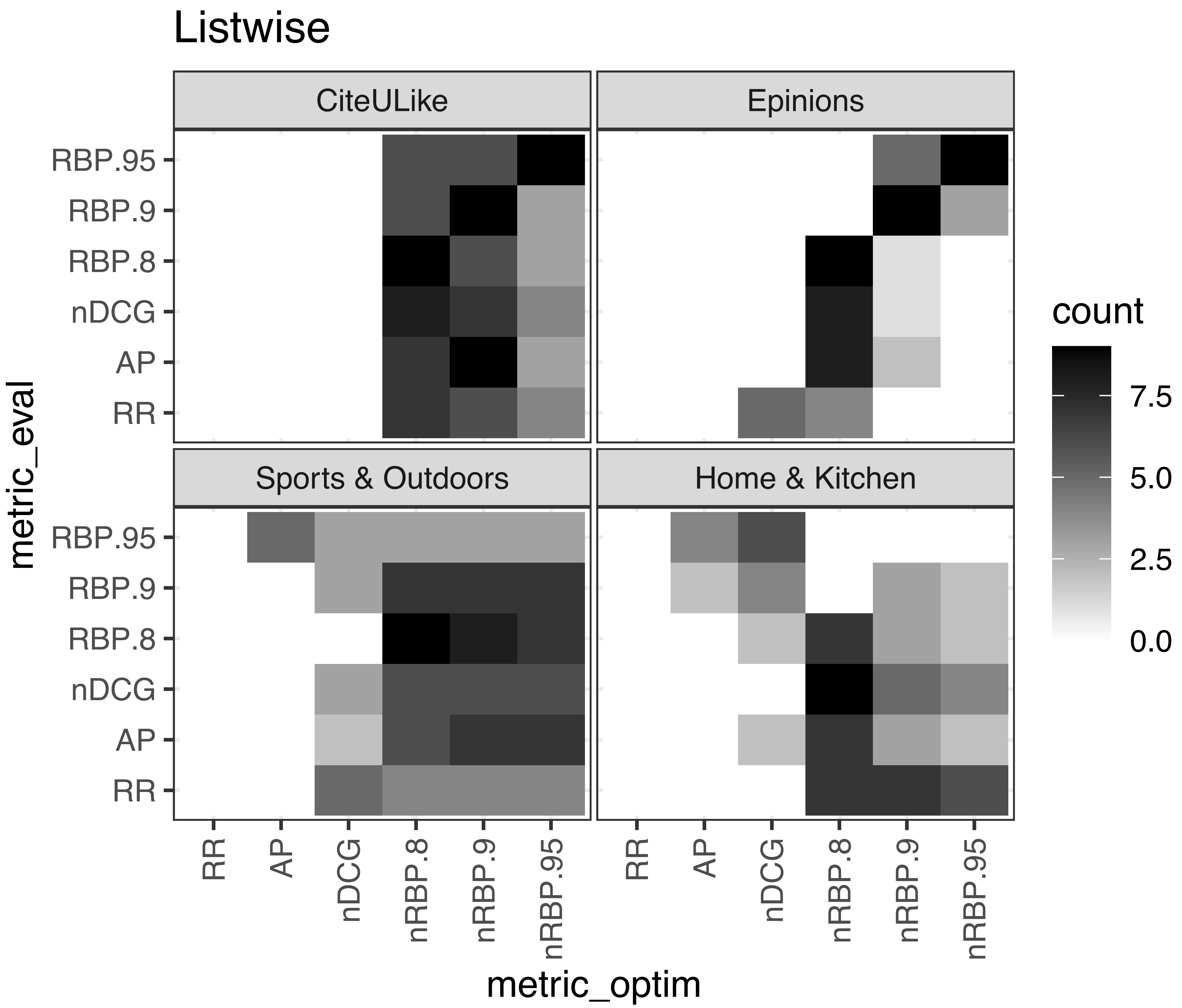
- Binarization: threshold=4 for graded datasets
- 25-core filtering
- User-level split with Train:Test =4:1 (>=5 items per user for testing)

Protocols

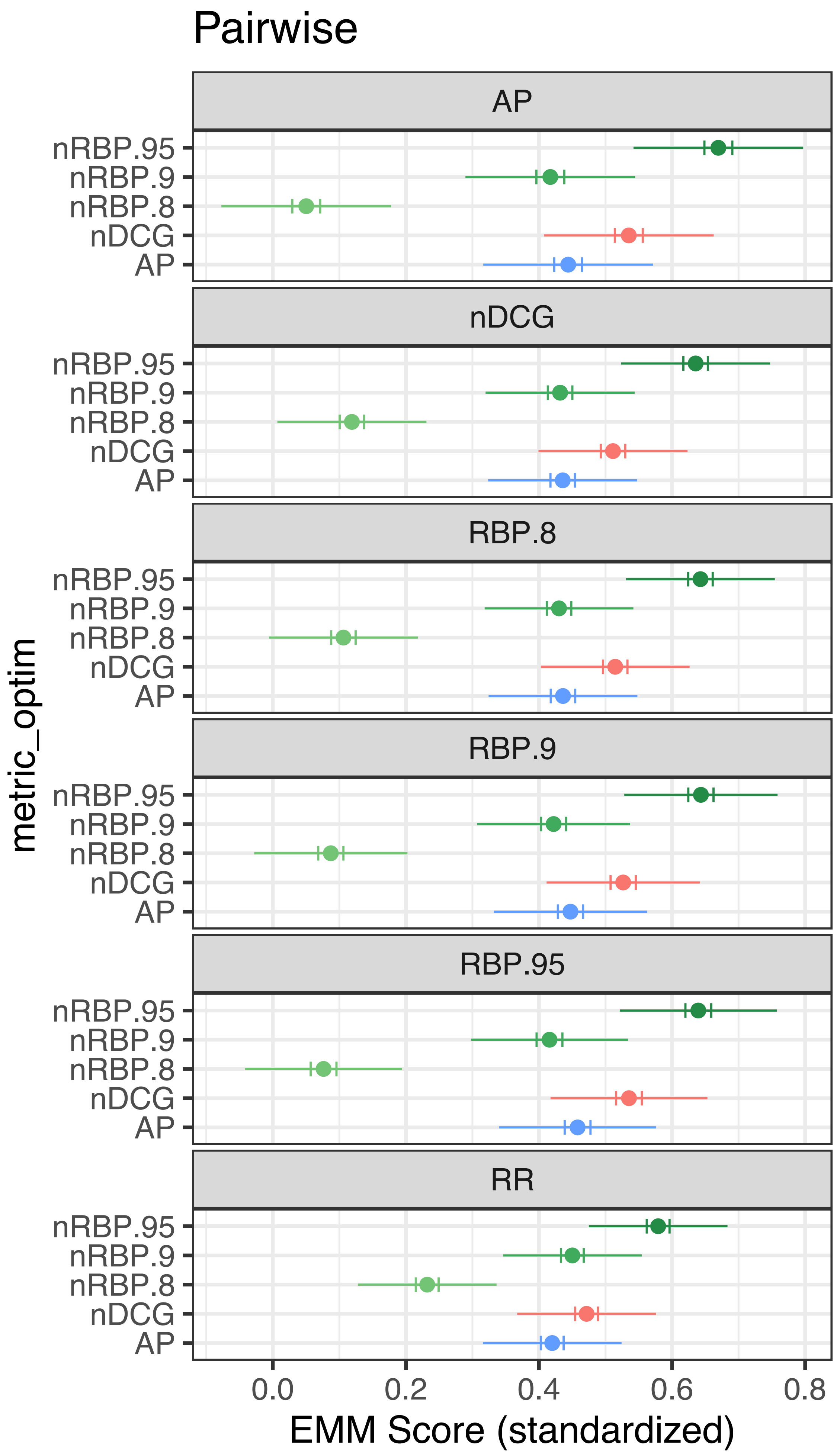
- 3 different splits per dataset
- Evaluation Metric: nDCG, AP, RR, RBP.8, RBP.9, RBP.95
- Recommender: Matrix Factorization
- Negative Sampling Ratio (NSR): 100%, 200%, 500%
- Training Epoch Selection: based on individual *p*'s

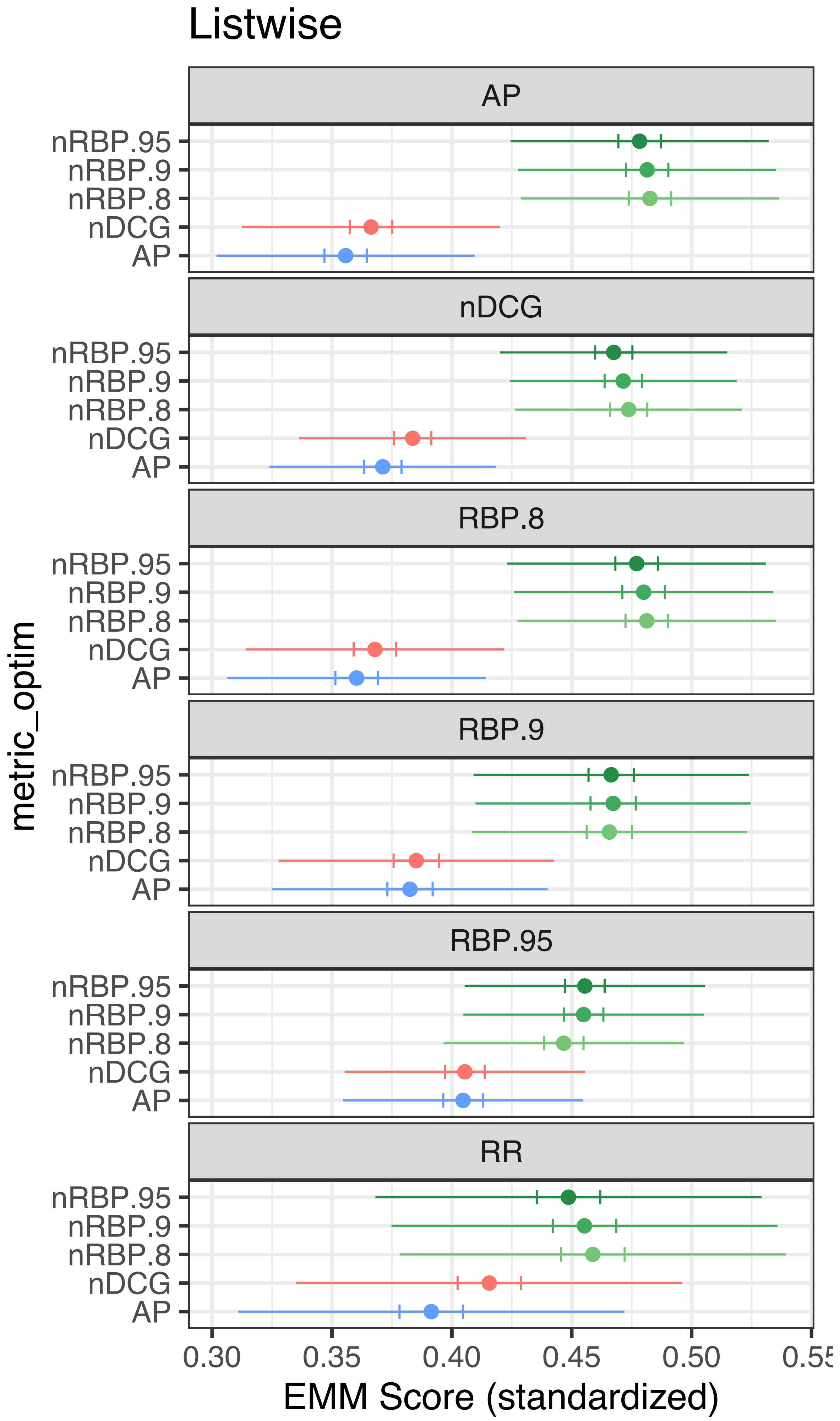
Overall Performance



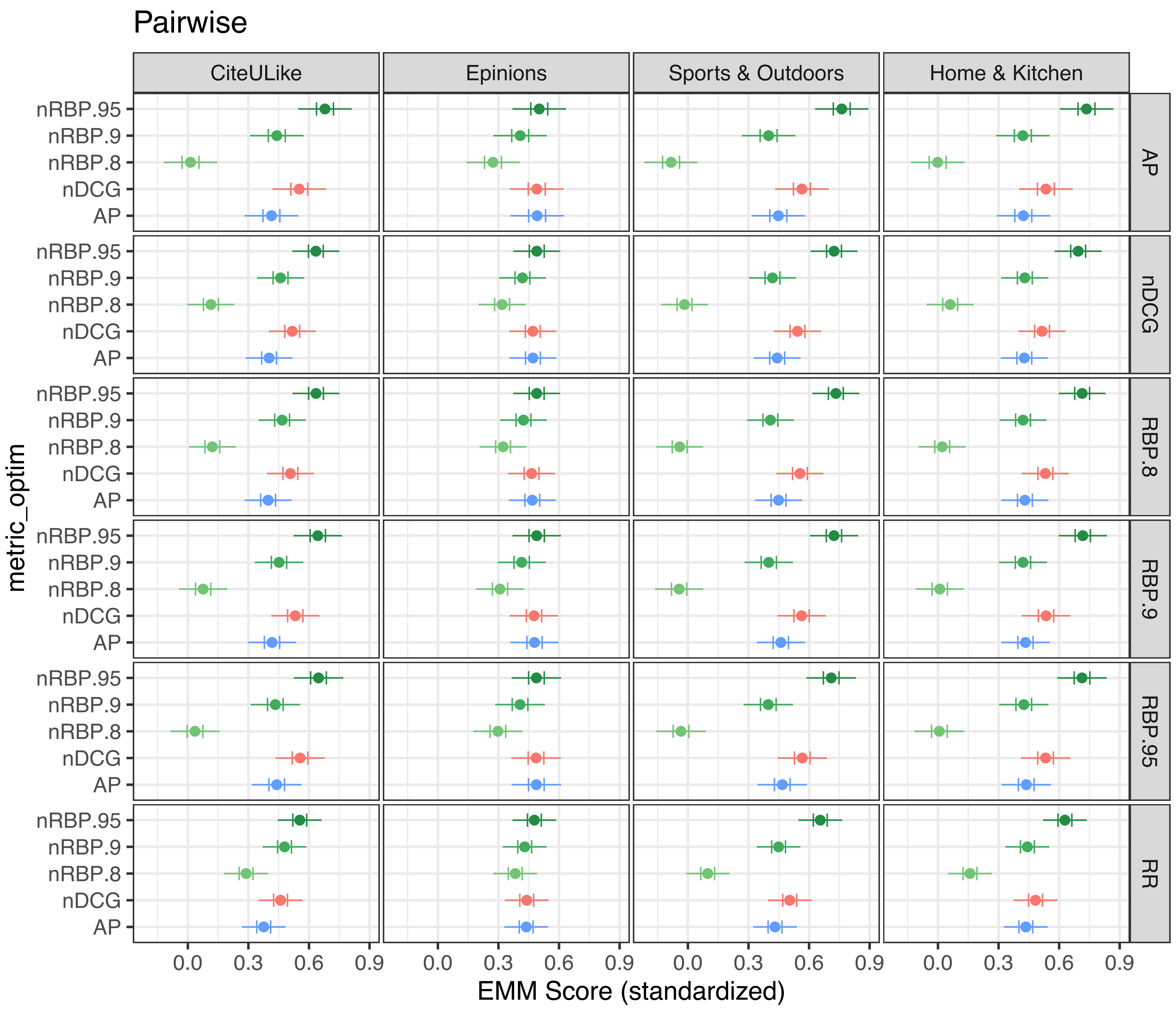


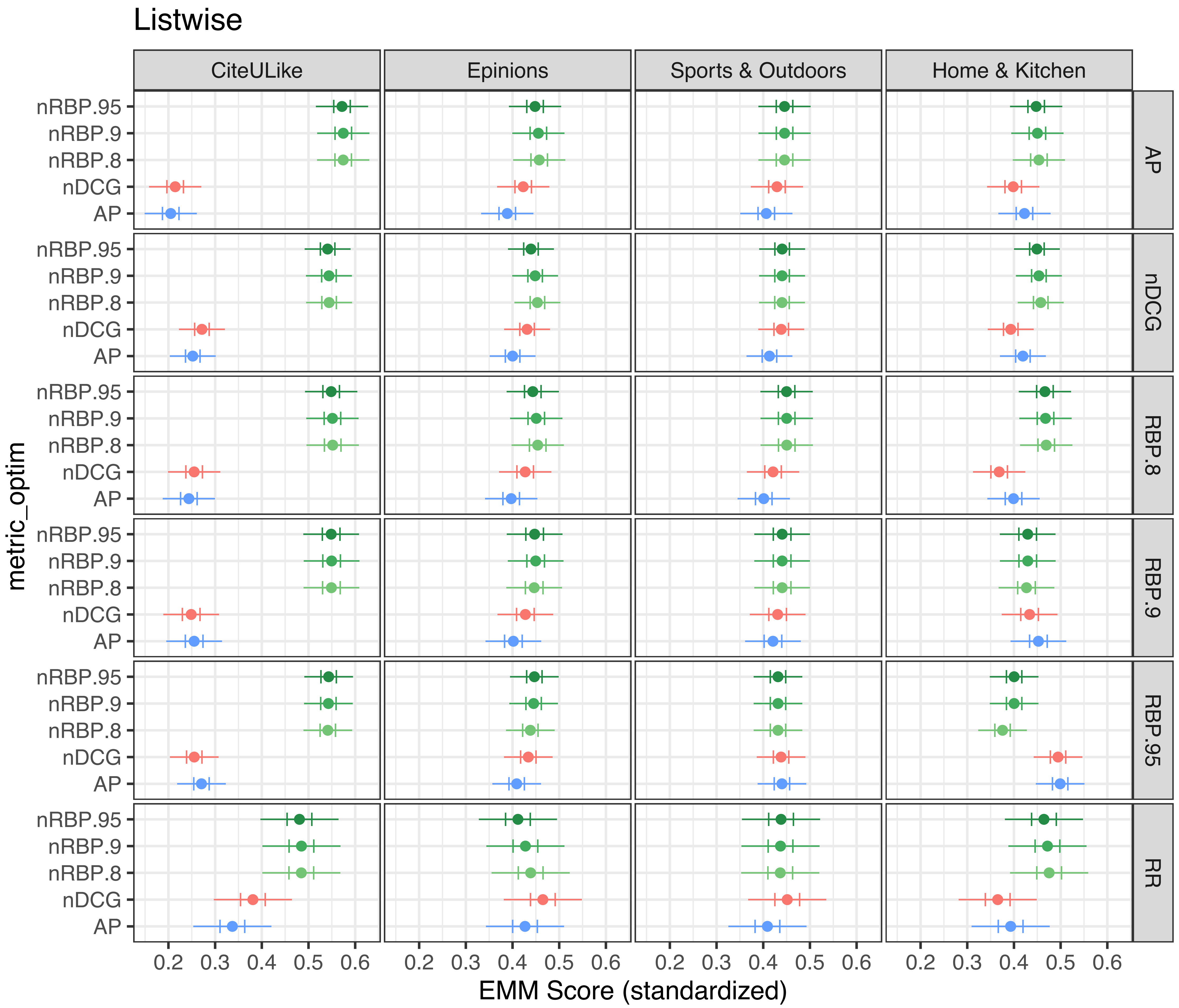
Overall Effectiveness: by Metrics used for Optimization



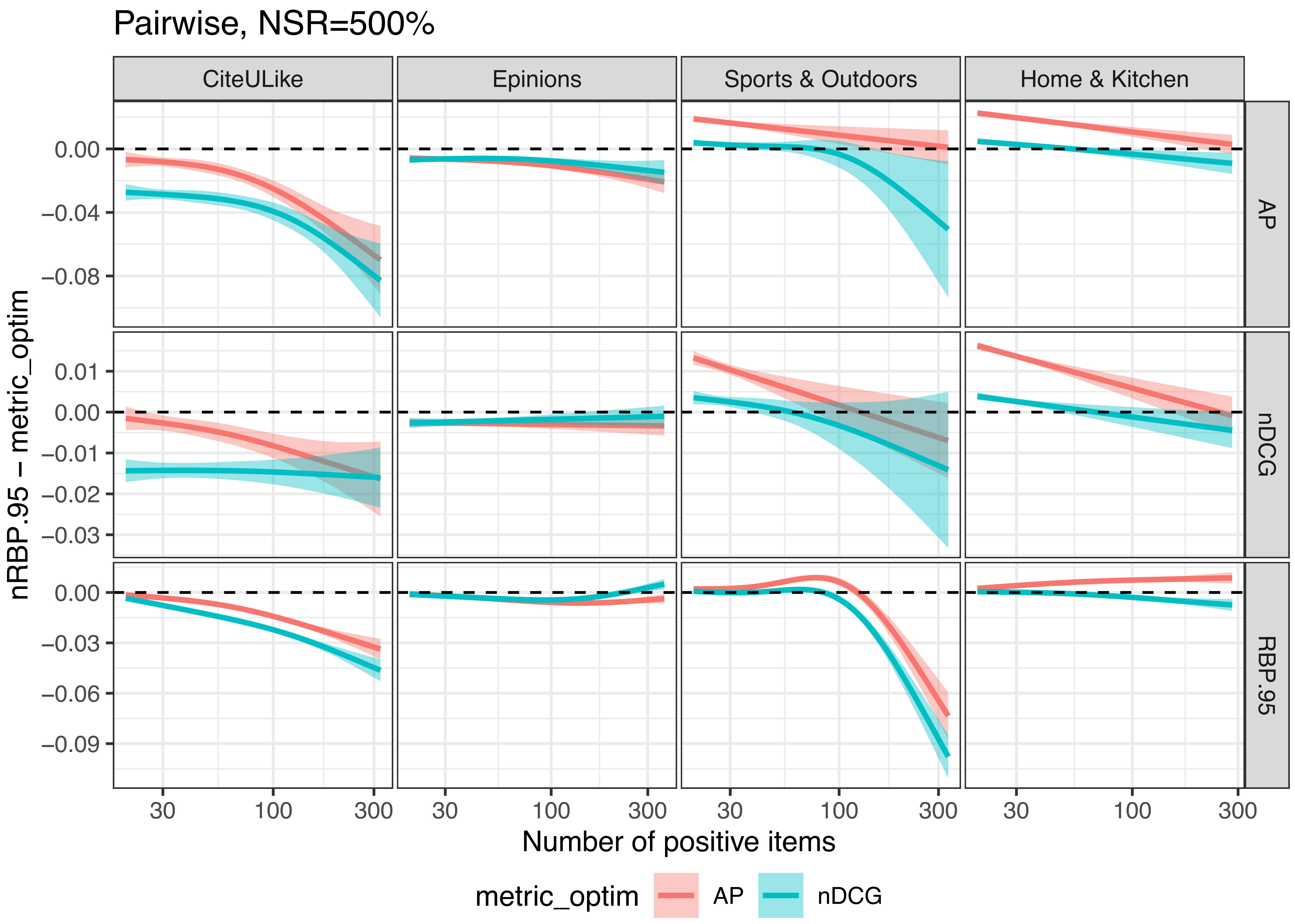


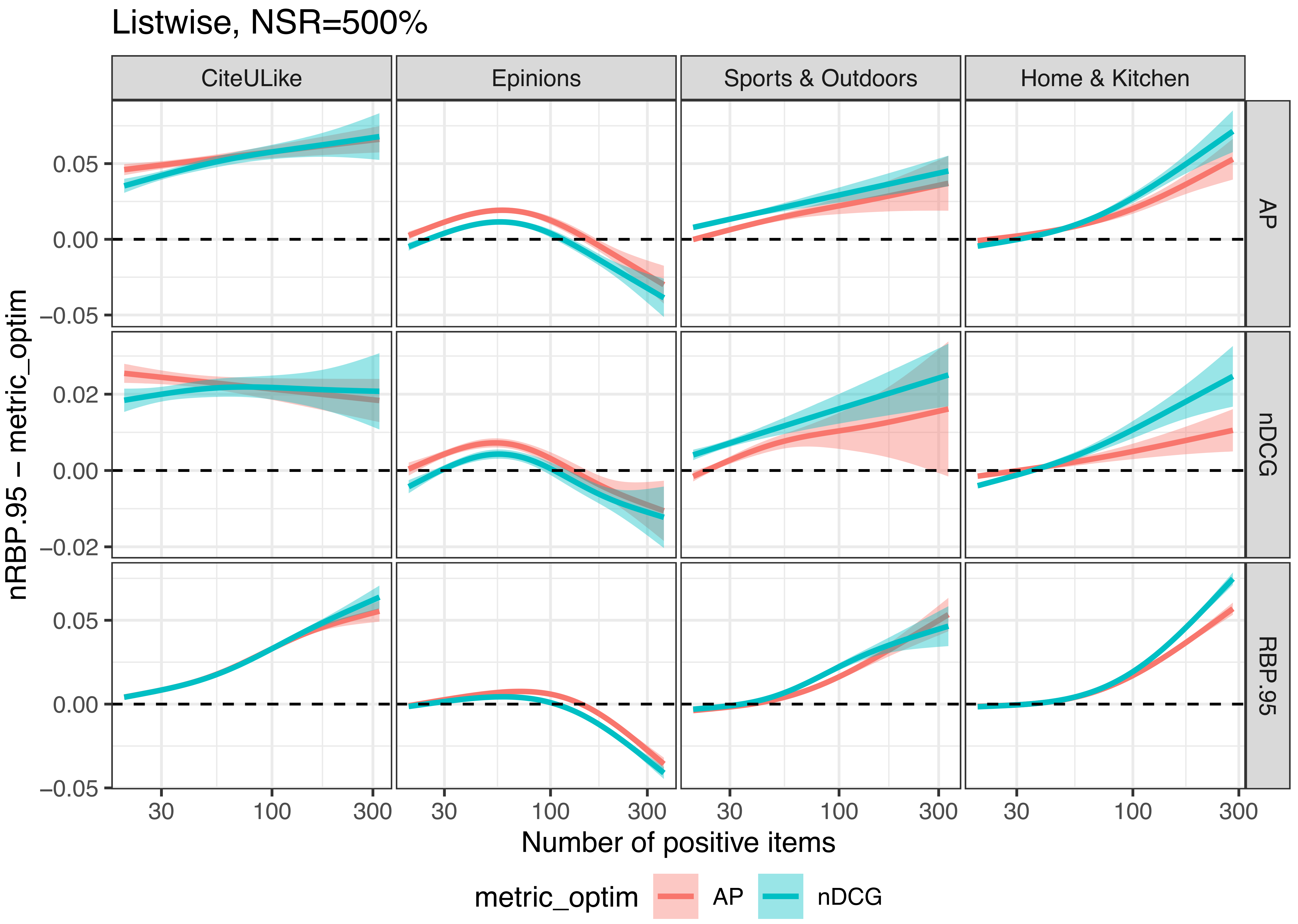
Overall Effectiveness: by Datasets





Individual Analysis on nRBP: Fairness for Effectiveness?





Conclusions

- It is not necessarily the best to optimize for the same metric used for evaluation in ranking-based recommender systems ;

- RBP is a promising alternative to serve as the loss in LTR recommenders.

- RBP-based listwise optimization improves the utility of all users, but favors more on active users.

Code & Data: https://github.com/roger-zhe-li/sigir21-newinsights

Special thanks to SIGIR for providing a travel grant for the first author.