Selective Preference Aggregation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many applications in machine learning and decision making rely on procedures to aggregate human preferences. In such tasks, individuals express ordinal preferences over a set of items by voting, rating, or comparing them. We then aggregate these data into a ranking that reveals their collective preferences. Standard methods for preference aggregation are designed to return rankings that arbitrate conflicting preferences between individuals. In this work, we introduce a paradigm for *selective aggregation* where we abstain from comparison rather than arbitrate dissent. We summarize collective preferences as a *selective ranking* – i.e., a partial order that reflects all collective preferences where at least $100 \cdot (1 - \tau)\%$ of individuals agree. We develop algorithms to build selective rankings that achieve all possible trade-offs between comparability and disagreement, and derive formal guarantees on their recovery and robustness. We conduct an extensive set of experiments on real-world datasets to benchmark our approach and demonstrate its functionality. Selective rankings provide a simple collective lever: set τ to expose disagreement, abstain rather than arbitrate, and constrain downstream algorithms to consensus.

1 Introduction

Many of our most important systems rely on procedures where we elicit and aggregate human preferences. In such systems, we ask a group of individuals to express their preferences over a set of items through votes, ratings, or pairwise comparisons. We then use these data to order items in a way that represents their collective preferences as a group. Over the past century, we have applied this pattern to reap transformative benefits from collective intelligence in elections [1], online search [2], and model alignment [3].

Standard methods for preference aggregation represent collective preferences as a *ranking* – i.e., a total order over n items where we can infer the collective preference between items by comparing their positions. Real-world preference data are noisy, strategic, and shift across populations, making total orders brittle. Rankings reflect an *approximate* summary of collective preferences because it is impossible to define a coherent order when individuals disagree. This impossibility – which is enshrined in foundational results such as Condorcet's Paradox [1] and Arrow's Impossibility Theorem [4] – has cast preference aggregation as an exercise in *arbitration*.

In many use cases for rankings, *we do not need a total order.* Abstaining on contested pairs and keeping only well-supported comparisons yields more robust outcomes. When we aggregate preferences to rank colleges, a total order can strongly influence where students apply and how institutions invest [see e.g., 5–8]. When we aggregate preferences to predict helpfulness [9], a total order can lead us to build models that are aligned with the preferences of a slim majority [10].

Submitted to the NeurIPS 2025 Workshop on Algorithmic Collective Action. Do not distribute.

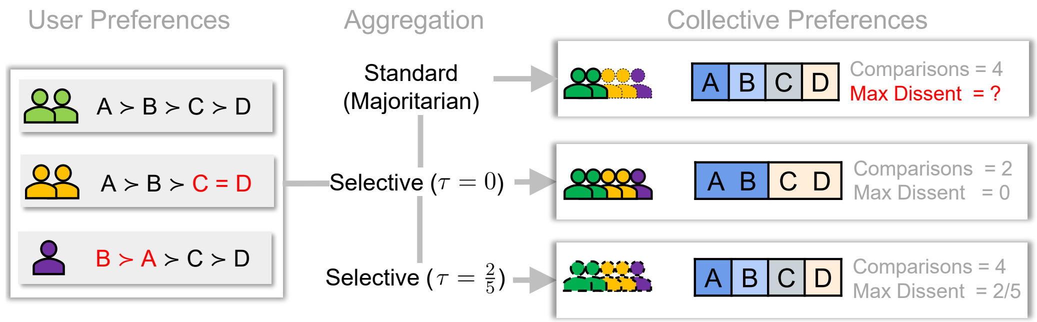


Figure 1: Comparison of collective preferences for 5 users over $n = 4$ items. Standard rankings arbitrate disagreement and hide it. Selective aggregation returns a partial order (tiers): items in different tiers are comparable, and any such comparison overrules at most $100 \cdot \tau\%$ of users. The tiers make disagreement explicit — e.g., $\tau = 0$ gives unanimous $\{A, B\} \succ \{C, D\}$, while $\tau = 2/5$ recovers a total order if one accepts overruling up to 40%.

By fixing τ, we can accept only consensus-backed comparisons, resist gaming, and shape system behavior.

In this work, we propose to address these challenges through *selective aggregation*. In this paradigm, we express collective preferences as a *tiered ranking* – i.e., a partial order where we are only allowed to compare items in different tiers. We view tiers as a simple solution to avoid the impossibility of arbitration: given a pair of items where individuals express conflicting preferences, we can place them in the same tier to abstain from comparison. We capitalize on this structure to develop a new representation for collective preferences that can reveal disagreement, and new algorithms that can allow us to control it.

1. We introduce a paradigm for preference aggregation where we summarize collective preferences as a *selective ranking* – i.e., a partial order where each comparison aligns with the preferences of at least $100(1 - \tau)\%$ of users.

2. We develop algorithms to construct all possible selective rankings for a preference aggregation task. Our algorithms are fast, easy to implement, and behave in ways that are safe and predictable, and we provide an open-source Python library for selective preference aggregation, available on anonymized repository.

3. We conduct a study of preference aggregation in modern use cases and demonstrate how selective aggregation can be used to learn from subjective annotations in a case study in toxicity detection. Our results show how selective rankings can promote transparency and robustness compared to existing approaches.

Related Work

Our work is motivated by applications that must aggregate conflicting preferences. In machine learning, this appears in data annotation and alignment due to ambiguity, subjectivity, or expertise gaps [3, 11–17]. In medicine, conflicts reflect uncertainty about ground truth [18–21]; in content moderation, they reflect differences in opinion [22, 23].

Our approach connects to social choice [24], which develops voting rules and impossibility results [4, 25–27]. Few works consider abstaining from arbitration via partial orders; abstention is often infeasible in settings like elections that require a single winner [28].

We complement rank-aggregation methods [2, 29–31] and coarser representations such as bucket orderings [32–34, and refs.]. Whereas bucket orderings treat within-block items as "equivalent," our tiered ranking treats within-tier items as "incomparable."

2 Framework

We consider a standard preference aggregation task where we wish to order n items in a way that reflects the collective preferences of p users. We start with a dataset where each instance $\pi^k_{i,j}$

represents the pairwise preference of a user $k \in [p] := \{1, \ldots, p\}$ between a pair of items $i, j \in [n]$:

$$\pi_{i,j}^k = \begin{cases} 1 & \text{if user } k \text{ strictly prefers } i \text{ to } j \Leftrightarrow i \overset{k}{\succ} j \\ 0 & \text{if user } k \text{ is indifferent} \qquad\quad \Leftrightarrow i \overset{k}{\sim} j \\ -1 & \text{if user } k \text{ strictly prefers } j \text{ to } i \Leftrightarrow i \overset{k}{\prec} j \end{cases}$$

Pairwise preferences can represent a wide range of ordinal preferences, including labels, ratings, and rankings. In practice, we can convert all of these formats to pairwise preferences as described in Appendix A.2. In what follows, we assume that datasets contain all pairwise preferences from all users for the sake of clarity. We describe how to relax this assumption in Appendix B, and work with datasets with missing preferences in Section 3.

Collective Preferences as Partial Orders Standard approaches express collective preferences as a *ranking* – i.e., a total order over n items where we can compare any pair of items. We consider an alternative approach in which we express collective preferences as a *tiered ranking*:

Definition 2.1. A *tiered ranking* T is a partial ordering of n items into m disjoint *tiers* $T := (T_1, \ldots, T_m)$. Given a tiered ranking, we denote the collective preferences as:

$$\pi_{i,j}(T) := \begin{cases} 1 \text{ if} & i \in T_l, j \in T_{l'} \text{ for } l < l', \\ -1 \text{ if} & i \in T_l, j \in T_{l'} \text{ for } l > l', \\ \perp \text{ if} & i, j \in T_l \text{ for any } l \end{cases}$$

Tiers provide a way to *abstain from arbitration*. Given a pair of items where users disagree, we can place them in the same tier and "agree to disagree." Given a tiered ranking T, we can only make claims about collective preferences by comparing items in different tiers. In what follows, we say that a pairwise comparison between items i, j is *valid* if $\pi_{i,j}(T) \neq \perp$. We refer to a valid pairwise comparison as a *selective comparison*.

Selective Aggregation *Selective ranking* S_τ is a partial order that maximizes the number of comparisons that align with the preferences of at least $100 \cdot (1 - \tau)$ of users. Given a dataset of pairwise preferences over n items from p users, we can express S_τ as the optimal solution to an optimization problem over the space of all tiered rankings $\mathbb{T}$:

$$\begin{aligned} \max_{T \in \mathbb{T}} \quad & \text{Comparisons}(T) \\ \text{s.t.} \quad & \text{Disagreements}(T) \leq \tau p \end{aligned} \tag{SPA$_\tau$}$$

Here, the objective maximizes the number of valid comparisons in a tiered ranking T:

$$\text{Comparisons}(T) := \sum_{i,j \in [n]} \mathbb{I}\left[\pi_{i,j}(T) \neq \perp\right]$$

The constraints restrict the fraction of individual preferences that can be contradicted by any valid comparison in T

$$\text{Disagreements}(T) := \max_{i,j \in [n]} \sum_{k \in [p]} \mathbb{I}\left[\pi_{i,j}(T) = 1, \pi_{i,j}^k \neq 1\right]$$

The *dissent parameter* τ limits the fraction of individual preferences that can be violated by any selective comparison. Given a selective ranking S_τ that places item i in a tier above item j, at most $100 \cdot \tau\%$ of users may have stated $i \not\succ j$.

We restrict $\tau \in [0, 0.5)$ to guarantee that the selective ranking S_τ aligns with a majority of users, and is unique (see Appendix A.2 for a proof). In this regime, we can set τ to trade off coverage for alignment as shown in Fig. 3.

We present an algorithm to construct selective rankings in Algorithm 1.

Algorithm 1 constructs a selective ranking from a dataset of pairwise preferences and a dissent parameter $\tau \in [0, 0.5)$. The procedure first builds a directed graph over items (V_I, A_I). Here, each vertex corresponds to an item, and each arc corresponds to a collective preference that we must not

Algorithm 1 Selective Preference Aggregation

Input: $\{\pi^k_{i,j}\}_{i,j \in [n], k \in [p]}$ *preference dataset*
Input: $\tau \in [0, 0.5)$ *dissent parameter*
1: $w_{i,j} \leftarrow \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$ for all $i, j \in [n]$
2: $V_I \leftarrow [n]$
3: $A_I \leftarrow \{(i \to j) \mid w_{i,j} > \tau p\}$
4: $V_T \leftarrow \mathsf{ConnectedComponents}(V_I, A_I)$
5: $A_T \leftarrow \{(T \to T') \mid \exists i \in T, j \in T' : (i \to j) \in A_I\}$
6: $l_1, \ldots, l_{|T|} \leftarrow \mathsf{TopologicalSort}(V_T, A_T)$
Output: $S_\tau \leftarrow (T_{l_1}, T_{l_2}, \ldots, T_{l_{|T|}})$ *τ-selective ranking*

contradict in a tiered ranking. Given (V_I, A_I), the procedure then builds a directed graph over tiers (V_T, A_T). In Line 4, it calls the ConnectedComponents routine to identify the strongly connected components of (V_I, A_I) which become the set of *supervertices* $V_T = \{T_1, \ldots, T_{|V_T|}\}$, where each supervertex contains items in the same tier. In Line 5, it defines arcs between tiers – drawing an arc from T to T' whose respective elements are connected by an arc in A_I. Given (V_T, A_T), the procedure determines an ordering among tiers by calling the TopologicalSort routine in Line 6. In this case, the graph will admit a topological sort as it is a directed acyclic graph.

We provide guarantees in Appendix B.

3 Experiments

In this section, we present an empirical study of selective aggregation on real-world datasets. Our goal is to benchmark the properties and behavior of selective rankings with respect to existing approaches in terms of transparency, robustness, and versatility. We include additional results in Appendix D, and code to reproduce our results on anonymized repository.

3.1 Setup

We evaluate on 5 preference datasets (Table 1). Each encodes user choices (votes/ratings/rankings); we convert these to pairwise comparisons with ties and build rankings using our method and baselines. We compute solution paths via Algorithm 2 and report three representative points:

- SPA_0: $\tau = 0$ (unanimous comparisons only).
- $\mathsf{SPA}_{\mathrm{min}}$: smallest $\tau > 0$ yielding ≥ 2 tiers (minimal disagreement to state any preference).
- $\mathsf{SPA}_{\mathrm{maj}}$: largest $\tau < 0.5$ (max claims without overruling a majority).

Baselines:

- *Voting rules*: Borda [35] and Copeland [36].
- *Sampling*: MC4 [2], which ranks by the stationary distribution of a Markov chain induced by random walks over user preferences.
- *Median rankings*: Kemeny [37] minimizes collective disagreement; we report an exact ILP (CPLEX v22 [38]) and a heuristic (BioConsert [39]).

3.2 Results

We summarize the specificity, disagreement, and robustness of rankings from all methods and all datasets in Table 1. In what follows, we discuss these results.

On Transparency Standard approaches hide arbitration: a total ranking reveals neither how many users were overruled nor which items are contested. Selective rankings expose both. As shown in Appendix D.1, the dissent parameter quantifies the maximum overruled fraction per comparison, and tiers localize disagreement—only across-tier comparisons are allowed, while same-tier pairs imply at least τ disagreement (e.g., Duke–Columbia).

Dataset	Metrics	Selective			Standard			
		SPA$_0$	SPA$_{\text{min}}$	SPA$_{\text{maj}}$	Borda	Copeland	MC4	Kemeny
nba	Disagreement Rate	0.0%	2.0%	6.4%	8.3%	8.3%	7.9%	8.1%
$n = 7$ items	Abstention Rate	100.0%	42.9%	28.6%	–	–	–	–
$p = 100$ users	# Tiers	1	2	4	7	7	6	7
28.6% missing	# Top Items	7	3	1	1	1	1	1
NBA [40]	Δ Sampling	0.0%	0.0%	0.0%	4.8%	4.8%	0.0%	4.8%
	Δ-Adversarial	0.0%	0.0%	0.0%	19.0%	19.0%	19.0%	14.3%
survivor	Disagreement Rate	0.0%	0.2%	0.2%	6.8%	6.6%	6.4%	6.7%
$n = 39$ items	Abstention Rate	94.9%	42.5%	42.5%	–	–	–	–
$p = 6$ users	# Tiers	2	5	5	39	36	35	39
0.0% missing	# Top Items	1	1	1	1	1	1	1
Purple Rock [41]	Δ Sampling	0.0%	0.0%	0.0%	1.3%	0.8%	0.8%	0.9%
	Δ-Adversarial	0.0%	0.0%	0.0%	2.6%	1.8%	3.1%	1.6%
lawschool	Disagreement Rate	0.0%	0.3%	3.1%	4.7%	4.2%	4.2%	4.1%
$n = 20$ items	Abstention Rate	40.5%	36.8%	4.2%	–	–	–	–
$p = 5$ users	# Tiers	4	6	15	20	20	19	20
0% missing	# Top Items	12	12	2	1	1	1	1
LSData [42]	Δ Sampling	0.0%	0.0%	0.0%	1.6%	1.1%	0.5%	29.5%
	Δ-Adversarial	0.0%	0.0%	0.0%	3.7%	2.6%	2.6%	45.8%
csrankings	Disagreement Rate	0.0%	0.0%	0.1%	12.3%	12.2%	12.2%	13.7%*
$n = 175$ items	Abstention Rate	100.0%	98.9%	95.5%	–	–	–	–
$p = 5$ users	# Tiers	1	2	3	175	168	170	175*
0.0% missing	# Top Items	175	1	1	1	1	1	1*
Berger [43]	Δ Sampling	0.0%	0.0%	0.0%	0.8%	0.8%	0.1%	9.0%*
	Δ-Adversarial	0.0%	0.0%	0.0%	3.1%	1.7%	0.1%	11.1%*
sushi	Disagreement Rate	0.0%	13.6%	42.6%	42.6%	42.6%	42.6%	42.6%
$n = 10$ items	Abstention Rate	100.0%	64.4%	0.0%	–	–	–	–
$p = 5,000$ users	# Tiers	1	2	10	10	10	10	10
0.0% missing	# Top Items	10	8	1	1	1	1	1
Kamishima [44]	Δ Sampling	0.0%	0.0%	0.0%	0.0%	0.0%	2.2%	2.2%
	Δ-Adversarial	0.0%	0.0%	0.0%	2.2%	2.2%	11.1%	11.1%

Table 1: Comparability, disagreement, and robustness of rankings for all methods on all datasets. We report the following metrics for each ranking: *Disagreement Rate*, i.e., the fraction of collective preferences that conflict with user preferences; *Abstention Rate*, i.e., the fraction of collective preferences that abstain from comparison; *# Tiers*, the number of tiers or ranks. *# Top Items*, i.e., the number of items in the top tier or rank. Δ-*Sampling*, the average fraction of collective preferences that are inverted when we drop 10% of individual preferences; and Δ-*Adversarial*, the maximum fraction of collective preferences that are inverted when we flip 10% of individual preferences, respectively.

Unlike traditional methods, a single winner or total order appears only when supported by a majority. In Table 1, we obtain a single winner in 4/5 datasets and a full order in 1/5. On `law`, the most granular solution (SPA$_{\text{maj}}$) yields two "top" schools (Stanford, Yale). On `sushi`, a single winner and total order emerge only at $\tau = 0.48$, indicating substantial contention.

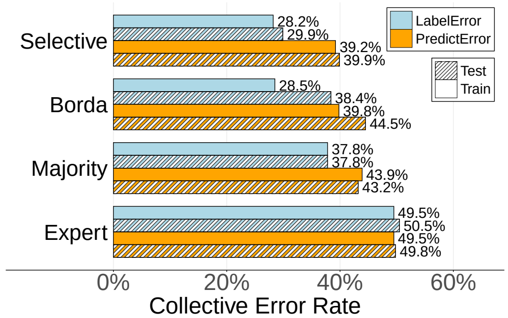
On Robustness Representing collective preferences as a total ranking can change dramatically under small perturbations to individual preferences [45–47]. This sensitivity is structural: in a ranking over n items, any change can affect $\binom{n}{2}$ pairwise preferences. In contrast, selective rankings group items into $m \leq n$ tiers, restricting the number of comparisons that can change and thereby improving robustness. In Table 1, we quantify this via the expected rate of inverted collective preferences under small perturbations: Δ-Sampling and Δ-Adversarial measure inversions from dropping or flipping 10% of individual preferences. For each dataset and method, we repeat these perturbations 100 times and report the mean inversion rate relative to the original ranking.

4 Learning by Agreeing to Disagree

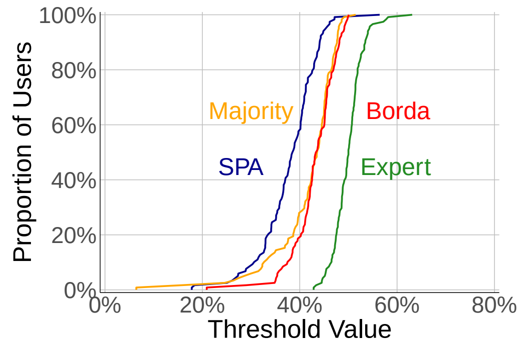
Preference aggregation is used to align models with group preferences (e.g., toxicity or helpfulness). We collect user annotations and aggregate them into training labels [48]. In subjective or ambiguous settings [13, 19, 23], majority vote can skew toward the majority [3, 10], in domains including online platforms. We examine how selective aggregation mitigates this by exposing and controlling disagreement.

Setup We build a toxicity classifier using DICES [49] with n=350 conversations and p=123 users. Labels are $y_i^k \in \{1, -1, 0\}$ for {`toxic`, `benign`, `unsure`}. Users are split into p^{train}=5 (to form training labels) and p^{test}=118 (to evaluate individual-level performance). We construct aggregate training labels (three variants) from the train group, dropping "unsure" (0) and aggregating only $\{-1, 1\}$. We represent these labels as y_i^{Maj}, y_i^{Borda}, y_i^{SPA}, and y_i^{Exp}.

We process the training labels from each method to ensure that we can use a standard training procedure across similar methods. We use the training labels from each method to fine-tune a BERT-Mini model [50] and denote these models as f^{SPA}, f^{Maj}, f^{Borda}, f^{Expert}. We evaluate how each



(a) Group-level errors on DICES. LabelError = avg. disagreement with annotators; PredictError = avg. disagreement of model predictions. Train: $p^{\text{train}}=5$; Test: $p^{\text{test}}=118$. SPA is lowest on both.



(b) CDF of user-level BER on held-out users ($p^{\text{test}}=118$). Higher is better.

method performs with respect to individuals and users in terms of the following measures:

$$\mathsf{BER}_k(f^{\text{all}}) := \tfrac{1}{2}\mathsf{FPR}_k(f^{\text{all}}) + \tfrac{1}{2}\mathsf{FNR}_k(f^{\text{all}})$$

$$\mathsf{LabelError}(y^{\text{all}}) := \tfrac{1}{p}\sum_{k=1}^{p}\mathsf{BER}_k(y^{\text{all}})$$

$$\mathsf{PredictError}(f^{\text{all}}) := \tfrac{1}{p}\sum_{k=1}^{p}\mathsf{BER}_k(f^{\text{all}})$$

We evaluate the performance of each in terms of the balanced error rate for clarity as the data for each user exhibits class imbalance that changes across users. We include additional details on our setup in Appendix D.5.

Results We summarize group- and individual-level results in Section 4. SPA minimizes collective disagreement: label error 28.2% (cf. 37.8% with y^{Maj}). Alignment in labels carries through to predictions: f^{SPA} has prediction error 29.9% (train) and 39.9% (test) vs. 38.4% and 44.5% for f^{Borda}. Across test users $p^{\text{test}}=118$, roughly 60% achieve individual $\mathsf{BER} \leq 40\%$ under y^{SPA}, compared to $\sim 20\%$ for y^{Borda} and y^{Maj}.

Labels that encode collective preferences help: the large label error for y^{Exp} indicates many users disagree with the expert. Results are reported at BER-optimized thresholds; similar patterns hold at other operating points (e.g., TPR $\geq 90\%$), where binary-label baselines such as majority vote may underperform.

5 Concluding Remarks

In many applications where we aggregate human preferences, disagreement is "signal, not noise" [11]. In this work, we developed foundations to aggregate preferences in a way that can reveal disagreement and allow us to control it. Selective aggregation compares only on consensus, resisting adversarial flips and missing data by abstaining on contested pairs. The main limitation of our work stems from algorithm design: the algorithms we have developed in this work are designed to be simple, versatile, and safe. To this end, they behave conservatively in tasks where datasets contain a large number of missing preferences.

In these cases, we can still represent collective preferences as a selective ranking, but the output may collapse into a single tier. This behavior is intentional: it signals that any claim about the collective preferences could be invalidated once the missing preferences are elicited. Looking forward, we can extend our paradigm to such settings by adopting probabilistic assumptions [see e.g., 32] and by developing procedures to streamline preference elicitation.

References

[1] Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, page 1785, 1785.

[2] Cynthia Dwork, Ravi Kumar, Moni Naor, and D Sivakumar. Rank aggregation revisited, 2001.

[3] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social Choice for AI Alignment: Dealing with Diverse Human Feedback. *arXiv preprint 2404.10271*, 2024.

[4] Kenneth J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, New York, 2nd edition, 1951. Revised edition published in 1963.

[5] Max Larkin. How northeastern gamed the college rankings. Boston Magazine, Aug 2014. URL https://www.bostonmagazine.com/news/2014/08/26/how-northeastern-gamed-the-college-rankings/.

[6] Espeland, Wendy Nelson and Sauder, Michael. *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation, 2016.

[7] The Economist. Columbia is the latest university caught in a rankings scandal, 2024.

[8] The Washington Post. Colleges are dropping out of rankings: Here's why yale says it's had enough, Nov 2022.

[9] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01 2022. ISSN 2307-387X.

[10] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint 2402.05070*, 2024.

[11] Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.

[12] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint 2110.05699*, 2021.

[13] Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, 2023.

[14] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. *arXiv preprint 2312.08358*, 2023.

[15] Jessica Dai and Eve Fleisig. Mapping social choice theory to rlhf. *arXiv preprint2404.13038*, 2024.

[16] Jakob Schoeffer, Maria De-Arteaga, and Jonathan Elmer. Perils of label indeterminacy: A case study on prediction of neurological recovery after cardiac arrest. *arXiv preprint 2504.04243*, 2025.

[17] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

[19] David Stutz, Ali Taylan Cemgil, Abhijit Guha Roy, Tatiana Matejovicova, Melih Barsbey, Patricia Strachan, Mike Schaekermann, Jan Freyberg, Rajeev Rikhye, Beverly Freeman, et al. Evaluating AI systems under uncertain ground truth: a case study in dermatology. *arXiv preprint 2307.02191*, 2023.

[20] Mike Schaekermann. *Human-AI Interaction in the Presence of Ambiguity: From Deliberation-based Labeling to Ambiguity-aware AI*. PhD thesis, University of Waterloo, 2020.

[21] Sujay Nagaraj, Yang Liu, Flavio Calmon, and Berk Ustun. Regretful decisions under label noise. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374, 2022.

[23] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[24] Christian List. Social Choice Theory. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Winter 2022 edition, 2022.

[25] Hervé Moulin. *Axioms of cooperative decision making*. Number 15. Cambridge university press, 1991.

[26] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[27] Amartya K Sen. A possibility theorem on majority decisions. *Econometrica: Journal of the Econometric Society*, pages 491–499, 1966.

[28] Roger B. Myerson. Nash equilibrium and the history of economic theory. *European Economic Review*, 43 (4-6):671–697, April 1999.

[29] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

[30] Dorit S Hochbaum and Asaf Levin. Methodologies and algorithms for group-rankings decision. *Management Science*, 52(9):1394–1408, 2006.

[31] Nir Ailon. Learning and optimizing with preferences. In *International Conference on Algorithmic Learning Theory*, pages 13–21. Springer, 2013.

[32] Mastane Achab, Anna Korba, and Stephan Clémençon. Dimensionality reduction and (bucket) ranking: a mass transportation approach. In *Algorithmic Learning Theory*, pages 64–93. PMLR, 2019.

[33] Pierre Andrieu, Bryan Brancotte, Laurent Bulteau, Sarah Cohen-Boulakia, Alain Denise, Adeline Pierrot, and Stéphane Vialette. Efficient, robust and effective rank aggregation for massive biological datasets. *Future Generation Computer Systems*, 124:406–421, 2021.

[34] Aristides Gionis, Heikki Mannila, Kai Puolamäki, and Antti Ukkonen. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 561–566. Association for Computing Machinery, 2006. ISBN 1595933395.

[35] JC de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

[36] Arhur H Copeland. A reasonable social welfare function. seminar on applications of mathematics to social sciences, university of michigan. *Ann Arbor*, 1951.

[37] John G Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

[38] IBM ILOG. *CPLEX User's Manual*. International Business Machines Corporation, 2024. Version 22.1.

[39] Sarah Cohen-Boulakia, Alain Denise, and Sylvie Hamel. Using medians to generate consensus rankings for biological data. In *International Conference on Scientific and Statistical Database Management*, pages 73–90. Springer, 2011.

[40] National Basketball Association. 2020-21 NBA Coach of the Year Voter Selections. https://ak-static.cms.nba.com/wp-content/uploads/sites/46/2021/06/2020-21-NBA-Coach-of-the-Year-Voter-Selections.pdf, 2021.

[41] Purple Rock. Survivor season rankings – spoiler-free summaries. https://www.purplerockpodcast.com/survivor-season-rankings-spoiler-free-summaries/, 2023.

[42] LSData. Law school rankings 2022. https://www.lsd.law/law-school-rankings-2022, 2022.

[43] Emery D. Berger. Csrankings: Computer science rankings. https://csrankings.org/, 2025. Accessed: 2025-08-13.

[44] Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 583–588, 2003.

[45] Abolfazl Asudeh, HV Jagadish, Gerome Miklau, and Julia Stoyanovich. On obtaining stable rankings. *Proceedings of the VLDB Endowment*, 12(3), 2018.

[46] Daniel Halpern, Gregory Kehne, Ariel D Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5657–5664, 2023.

[47] Ruiting Liang, Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Assumption-free stability for ranking problems, 2025.

[48] Michal Lukasik, Lin Chen, Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Felix X. Yu, Sashank J. Reddi, Gang Fu, Mohammadhossein Bateni, and Sanjiv Kumar. Bipartite ranking from multiple labels: On loss versus label aggregation. *arXiv preprint 2504.11284*, April 2025.

[49] Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. *Advances in Neural Information Processing Systems*, 2024.

[50] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint 1908.08962*, 2019.

[51] Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

[52] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. When is it better to compare than to score? *arXiv preprint 1406.6618*, 2014.

[53] Jingyan Wang and Nihar B Shah. Ranking and rating rankings and ratings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13704–13707, 2020.

[54] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *arXiv preprint 1806.05085*, 2018.

[55] Pierre Azoulay and Danielle Li. Scientific Grant Funding. Technical Report 26889, National Bureau of Economic Research, March 2020. Revised June 2021.

[56] John H Smith. Aggregation of preferences with variable electorate. *Econometrica: Journal of the Econometric Society*, pages 1027–1041, 1973.

[57] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Capturing expert arguments from medical adjudication discussions in a machine-readable format. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1131–1137, 2019.

[58] Guangrui Tang and Neng Fan. A survey of solution path algorithms for regression and classification models. *Annals of Data Science*, 9(4):749–789, 2022.

[59] Velocity Test Prep. Top 50 law school rankings & comparisons. https://www.velocitylsat.com/resources/top-law-schools, 2023.

[60] Top Universities. Qs world university rankings for law and legal studies 2024. https://www.topuniversities.com/university-subject-rankings/law-legal-studies, 2024.

[61] Above the Law. Top law schools 2023. https://abovethelaw.com/top-law-schools-2023/, 2023.

[62] TestMax Inc. LSATMax: Comprehensive lsat prep course. https://testmaxprep.com/lsat, 2025.

[63] College Kickstart LLC. U.S. News & World Report Posts 2023 College Rankings. https://www.collegekickstart.com/blog/item/u-s-news-world-report-posts-2023-college-rankings, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract/Section 1 match contributions in Sections 2 and 3 and Appendix B.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 5; assumptions on complete data noted in Section 2.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions with statements; full proofs in Appendix A.2.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: Datasets, preprocessing, baselines, metrics, and dissent settings in Section 3; scripts/configs in anonymized repository.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code?

   Answer: [Yes]

   Justification: Anonymized repo at anonymized repository; all datasets are public and cited.

6. **Experimental Setting/Details**

   Question: Does the paper specify all training/test details needed to understand the results?

   Answer: [Yes]

   Justification: Model/splits and aggregation for demo in Section 4 and Appendix D.5; baselines/metrics in Section 3.

7. **Experiment Statistical Significance**

   Question: Are error bars or equivalent significance indicators reported?

   Answer: [Yes]

   Justification: CDFs of per-user BER (Fig. D.1) and resampling methods reported for robustness metrics.

8. **Experiments Compute Resources**

   Question: Are compute resources and runtimes sufficiently described?

   Answer: [Yes]

   Justification: Compute details in Appendix D.5; exact Kemeny via CPLEX v22 (128 GB CPU); SPA runs in $O(n^2p)$.

9. **Code Of Ethics**

   Question: Does the research conform to the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: Public datasets only; no new human data.

10. **Broader Impacts**

Question: Does the paper discuss potential positive and negative societal impacts?

Answer: [Yes]

Justification: Transparency/robustness benefits in Section 1 and Section 5.

11. **Safeguards**

    Question: Are safeguards described for high-risk releases?

    Answer: [NA]

    Justification: No high-risk models or scraped datasets released.

12. **Licenses for existing assets**

    Question: Are third-party assets credited with licenses/terms?

    Answer: [Yes]

    Justification: Cited in Section 3 and Appendix D.

13. **New Assets**

    Question: If new assets are introduced, are they documented?

    Answer: [Yes]

    Justification: Anonymized library + docs at anonymized repository.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For human-subjects work, are instructions/compensation included?

    Answer: [NA]

    Justification: No new human-subjects or crowdsourcing studies.

15. **Institutional Review Board (IRB) Approvals or Equivalent**

    Question: Are risks/approvals for human-subjects research described?

    Answer: [NA]

    Justification: No new human-subjects research.

Supplementary Materials
Selective Preference Aggregation

A Supplementary Material for Section 2

A.1 Notation

We provide a list of the notation used throughout the paper in Table 2.

Object	Symbol	Description
Items	$i \in [n] := \{1, \dots, n\}$	The objects being ordered, for which users have expressed preferences.
Users	$k \in [p] := \{1, \dots, p\}$	Individuals expressing preferences for given items.
Individual preferences	$\pi^k_{i,j} \in \{-1, 0, 1\}$	Pairwise preference between items i and j for user k.
Tiered ranking	T	A partial ordering of n items into m *tiers*
Collective preference	$\pi_{i,j}(T) \in \{-1, 0, 1\}$	The preference between items i and j in a given ranking.
Selective ranking	S_τ	The partial order returned by solving $\mathsf{SPA}_\tau(\mathcal{D})$.
Dissent parameter	$\tau \in [0, \frac{1}{2})$	The admitted dissent between two items i and j.

Table 2: Notation

A.2 Encoding Individual Preferences as Pairwise Comparisons

Representation	Notation	Conversion
Labels	$y^k_i \in \{0, 1\}$	$\pi^k_{i,j} = \mathbb{I}\left[y^k_i > y^k_j\right] - \mathbb{I}\left[y^k_i < y^k_j\right]$
Ratings	$y^k_i \in [m]$	$\pi^k_{i,j} = \mathbb{I}\left[y^k_i > y^k_j\right] - \mathbb{I}\left[y^k_j > y^k_i\right]$
Rankings	$r^k : [n] \to [n]$	$\pi^k_{i,j} = \mathbb{I}\left[r^k(i) > r^k(j)\right] - \mathbb{I}\left[r^k(i) < r^k(j)\right]$

Table 3: Data structures that capture ordinal preferences over n items. Each representation can be converted into a set of $\binom{n}{2}$ pairwise preferences in a way that ensures (and assumes) transitivity. Item-level representations require fewer queries but may be subject to calibration issues between annotators.

One of the benefits in developing machinery to aggregate preferences is that it can provide practitioners with flexibility in deciding how to elicit and aggregate the preferences. In practice, such choices involve trade-offs that we discuss briefly below. Specifically, eliciting pairwise preferences from users requires more queries than other approaches [51]. However, it may recover a more reliable representation of ordinal preferences than ratings or rankings [i.e., 52]. In tasks where we work with a few items, we can elicit preferences as ratings, rankings, or pairwise comparisons. In tasks where we elicit rankings, we can convert them into pairwise comparisons without a loss of information. In this case, eliciting pairwise comparisons can test implicit assumptions such as transitivity. In tasks where we elicit labels and ratings, the conversion is lossy – since we are converting cardinal preferences to ordinal preferences. In practice, this conversion can resolve issues related to calibration across users [see e.g, 53, 54]. In theory, it may also resolve disagreement [27].

B Theoretical Guarantees

In this section, we present formal guarantees on the stability and recovery of selective rankings.

On the Recovery of Condorcet Winners We often aggregate preferences to identify items that are collectively preferred to all others. Consider, for example, a task where we aggregate votes to select the most valuable player in a sports league or ratings to fund the most promising grant proposal [55]. In Theorem B.1, we show that we can identify these "top" items from a solution path of selective rankings.

Theorem B.1. *Consider a preference aggregation task where a majority of users prefer item i_0 to all other items. There exists a threshold value $\tau_0 \in [0, 0.5)$ such that, for every $\tau > \tau_0$, every selective ranking S_τ will place i_0 as the sole item in its top tier.*

Theorem B.1 provides a formal recovery guarantee that ensures we recover a Condorcet winner or a Smith set [see e.g., 56] when they exist. In practice, the result implies that we can identify such "top items" by constructing and inspecting a solution path of selective rankings.

In tasks where a majority of users prefers an item to all others, the solution path will contain a selective ranking whose top tier consists of a single item. In this case, we can recover the "single winner" and report the threshold value τ_0 as a measure of consensus.

In tasks where such a majority does not exist, every selective ranking S_τ for $\tau \in [0, 0.5)$ will include at least two items in the top tier. In settings where we aggregate preferences to identify a "single winner," we can point to the solution path as evidence that no such winner exists and use it as a signal that further deliberation is required [see e.g., 57].

Stability with Respect to Missing Preferences Standard methods can output rankings that change dramatically once we elicit missing preferences [45–47]. In Proposition B.2, we show that we can build a selective ranking that abstains from unstable comparisons by setting missing preferences to $\pi^k_{i,j} = 0$.

Proposition B.2. *Given a preference dataset with missing preferences $\mathcal{D}^{\text{init}}$, let:*

- *$\mathcal{D}^{\text{true}} \supseteq \mathcal{D}^{\text{init}}$ be a complete dataset where we elicit missing preferences; and*
- *$\mathcal{D}^{\text{safe}} \supseteq \mathcal{D}^{\text{init}}$ be a complete dataset where we set missing preferences to $\pi^k_{i,j} = 0$.*

For any dissent value $\tau \in [0, \frac{1}{2})$, let S^{safe}_τ and S^{true}_τ denote selective rankings for $\mathcal{D}^{\text{safe}}$ and $\mathcal{D}^{\text{true}}$, respectively. Then for any selective comparison $\pi_{i,j}(S^{\text{safe}}_\tau) \in \{-1, 1\}$, we have:

$$\pi_{i,j}(S^{\text{safe}}_\tau) = \pi_{i,j}(S^{\text{true}}_\tau).$$

Proposition B.2 provides a simple way to ensure stability when working with datasets where we are missing preferences from certain users for certain items. In such cases, we can always build a S that is "robust to missingness" in the sense that it will abstain from comparisons that may be invalidated once we elicit missing preferences.

Stability with Respect to New Items In Proposition B.3, we characterize the stability of selective aggregation as we add a new item to our dataset.

Proposition B.3. *Consider a task where we start with a dataset of all pairwise preferences from p users over n items, which we then update to include all pairwise preferences for a new $n+1^{th}$ item. For any $\tau \in [0, \frac{1}{2})$, let S^n_τ and S^{n+1}_τ denote selective rankings over n items and $n+1$ items, respectively. Then for any two items $i, j \in [n]$, we have:*

$$\pi_{i,j}(S^{n+1}_\tau) \in \{-1, 1\}, \pi_{i,j}(S^{n+1}_\tau) \neq -\pi_{i,j}(S^n_\tau)$$

The result shows that adding a new item to a selective ranking will either maintain each comparison or abstain. That is, adding a new item can only collapse items that were in different tiers into a single tier. However, it cannot lead items in the same tier to split. Nor can it lead items in different tiers to invert their ordering.

B.1 Proof of Correctness

Lemma B.4. Consider the graph before running condensation or topological sort, but after pruning edges with weights below τp. Items can be placed in separate tiers without violating Disagreements$(T) \leq \tau p$ if and only if there is no cycle in the graph involving those items.

Proof. We start by connecting the edges in a graph to conditions on the items in a tiered ranking and eventually expand that connection to show the one-to-one correspondence between cycles and tiers. First note that for any items i, j: $w_{i,j} > \tau \iff \sum_{k=1}^{p} 1\left[\pi_{i,j}^{k} \neq 1\right] > \tau p$. This follows trivially from the definition of $w_{i,j} := \sum_{k=1}^{p} 1\left[\pi_{i,j}^{k} \neq 1\right]$. From this, we know that if and only if there exists an arc (i, j) that is not pruned before condensation, we cannot have a tiered ranking with $\pi_{i,j}^{T} = -1$ without violating Disagreements$(T) \geq \tau p$.

If there exists a cycle in this graph, then we know the items in that cycle must be placed in the same tier. To show this, consider some edge i, j in the cycle. We know item j cannot be in a lower tier than i without violating the disagreements property, from the above. So item j must be in the same or a higher tier. But item j has an arrow to another item, k, which must be in the same or a higher tier than both j and i, and so on, until the cycle comes back to item i. This corresponds to the constraint that all items must be in the same tier.

If a set of items is not in a cycle, then these items do not need to be placed in the same tier. If the items are not in a cycle, then there exists a pair of items (i, j) such that there is no path from j to i. Thus i can be placed in a higher tier than j without violating any disagreement constraints. Thus not all items in this set need to be placed in the same tier.

Thus we have shown that for a graph pruned with a given value of τ, items can be placed in separate tiers for a tiered ranking based on that same parameter τ, if and only if there is no cycle in the graph involving all of these items.

□

We draw on this Lemma to prove the main result:

Theorem B.5. Given a preference aggregation task with n items and p users, Algorithm 1 returns the optimal solution to SPA_τ for any dissent parameter $\tau \in [0, \frac{1}{2})$.

Proof of Theorem B.5. Consider that items in our solution are in the same tier if and only if they are part of a cycle in the pruned graph (i.e., if and only if they are in the same strongly connected component). So items are in the same tier if and only if they must be in the same tier for the solution to be feasible. No other feasible tiered ranking could have any of these items in separate tiers. So no other tiered ranking could have any more tiers, or any more comparisons. To do so would require placing some same-tier items in different tiers. Thus, our solution is maximal with respect to the number of tiers, and with respect to the number of comparisons. □

B.2 Proof of Uniqueness

Theorem B.6. The optimal solution to SPA_τ is unique for $\tau \in [0, 0.5)$.

Proof of Theorem B.6. Let T denote an optimal solution to SPA_τ. We will show that the optimality T is fully specified by: (1) the items in each tier and (2) the ordering between tiers. That is, if we were to produce a tiered ranking T' that assigns different items to each tier, or that orders tiers in a different way would be suboptimal or infeasible.

Consider a tiered ranking T that is feasible with respect to SPA_τ for some $\tau \in [0, 0.5)$. Let T' denote a tiered ranking where we swap the order of two tiers in T. We observe that the T' must violate a constraint. To see this, consider any pair of items i, j such that $\pi_{i,j}(T) = 1$ before the swap, but $\pi_{j,i}(T') = 1$ after the swap. One such pair must exist for any swapping of tier orders, because all tiers are non-empty. Because we elicited complete preferences, one of the following conditions must hold:

$$\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \neq 1\right] > \tau p \tag{1}$$

$$\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{j,i} \neq 1\right] > \tau p \tag{2}$$

Assuming that T was an optimal solution to SPA_τ, we observe that the condition in Eq. (1) must be violated because the original optimal solution was valid. Thus, we must have that $\sum_{k\in[p]} \mathbb{I}\left[\pi^k_{j,i} \neq 1\right] > \tau p$. This implies that Disagreements$(T') > \tau p$ for this tiered ranking. Thus, swapping the order of tiers violates constraints because $\tau < 0.5$.

Now note that any separation of items from within the same tier is not possible without violating a constraint. This follows from Lemma B.4, which states that items that are part of a cycle in our graph representation of the problem[1], must be in the same tier for a solution to be valid. And, as specified in our algorithm, we know our optimal solution has tiers only where there are cycles in the graph representation of the problem. So any tiers in the optimal solution cannot be separated.

We can still merge two tiers together without violating constraints, but such an operation reduces the number of comparisons and would no longer be optimal. And after merging two tiers, the only valid separation operation would be simply to undo that merge (since any other partition of the items in that merged tier, would correspond to separating items that were within the same tier in the optimal solution). So we cannot use merges as part of an operation to reach a valid alternative optimal solution.

So we know that for the optimal solution, we cannot separate out any items within the same tier, and we cannot reorder any of the tiers. Merging, meanwhile, sacrifices optimality. Thus, the original optimal solution is unique. $\square$

B.3 Constructing All Possible Selective Rankings

We start with a proof for Proposition B.8.

Proof of Proposition B.8. Recall that in Algorithm 1, an edge (i, j) with weight $w_{i,j}$ is excluded if at least τp users disagree with the preference $j \succ i$. We observe that $w_{i,j} = \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$ corresponds the number of users who disagree with the preference $j \succ i$. Given a dataset, denote the set of dissent values that could lead to different outputs as:

$$\mathcal{W} = \{0\} \cup \left\{ \tau' \mid \exists i, j : \tau' = \left(\frac{1}{p} \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right] \right) < \tfrac{1}{2} \right\}$$

This corresponds to the set of unique $w_{i,j}/p$ for all i, j, with the value 0 included as well. To see this, note $w_{i,j} = \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$. We will now show the following Lemma, which will resolve the original claim.

Lemma B.7. *Given any two adjacent elements $a, b \in \mathcal{W} \cup \{\frac{1}{2}\}$. All dissent values in $\tau \in [a, b)$ lead to the same selective ranking as the selective ranking for $\tau = a$.*

Proof. To show this, note that there exists no edge $i \to j$ such that $ap < w_{i,j} < bp$. If there did exist, then we would have

$$a < \frac{w_{i,j}}{p} < b.$$

This would imply that $\mathcal{W}$ would have to include an additional between a and b. But a and b are adjacent in $\mathcal{W}$. This is a contradiction.

Since there exists no edge $i \to j$ such that $ap < w_{i,j} < bp$, there exists no edge such that the decision to include its arc in the graph changes based on what value of dissent we select in $[a, b)$. Recall that we exclude $i \to j$ iff $w_{ij} \geq \tau p$ $\square$

[1] after pruning edges of weight below τ

Now that we know that for any two adjacent values a, b in $\mathcal{W} \cup \{\frac{1}{2}\}$, all dissent values in $[a, b)$ lead to the same tiered ranking as with dissent value a, we know that for any dissent value $\tau \in [0, \frac{1}{2})$, the largest value of $\tau' \in \mathcal{W}$ that is $\leq \tau$ will lead to the same tiered ranking. Simply substitute τ in for a, and the smallest value above τ in $\mathcal{W} \cup \{\frac{1}{2}\}$ for b (such a value exists, on both sides, because 0 and $\frac{1}{2}$ are both $\in \mathcal{W} \cup \{\frac{1}{2}\}$, and $\tau \in [0.\frac{1}{2})$).

Thus we have shown the required claim.

$\square$

Recovering All Selective Rankings Algorithm 1 is meant to recover a selective ranking in settings where we can set the value of τ a priori (e.g., $\tau = 0\%$ to enforce unanimity). In many applications, we may wish to set τ after seeing the entire path of selective rankings. In a funding task where we only have the resources to fund 3 proposals, for example, we can choose the smallest value of τ from the solution path such that the top tier contains ≤ 3 proposals. In cases where a top three does not exist, this can lead us to save resources or increase our budget. In a prediction task where labels encode collective preferences, we could aggregate annotations with a selective ranking and treat τ as a hyperparameter to control overfitting.

In these situations, we can produce a *solution path* of selective rankings– i.e., a finite set of selective rankings that covers all possible solutions to SPA_τ for $\tau \in [0, \frac{1}{2})$ [c.f. 58]. We observe that a finite solution path must exist as each selective ranking is specified by the arcs in Line 3. In practice, we can compute all selective rankings efficiently by: (1) identifying a smaller subset of dissent parameters to consider as per Proposition B.8; and (2) re-using the graph of strongly connected components across iterations.

Proposition B.8. *Given a dataset of pairwise preferences $\mathcal{D}$, let $\mathcal{S}_\mathcal{W}$ denote a finite set of selective rankings for dissent parameters in the set:*

$$\mathcal{W} = \left\{ \tfrac{w}{p} < \tfrac{1}{2} \mid w = \sum_{k \in [p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right] \text{ for } i, j \in [n] \right\} \cup \{0\}$$

Let S_τ be a selective ranking for an arbitrary dissent value $\tau \in [0, \frac{1}{2})$. Then, $\mathcal{S}_\mathcal{W}$ contains a selective ranking $S_{\tau'}$ such that $S_{\tau'} = S_\tau$ for some dissent value $\tau' \leq \tau$.

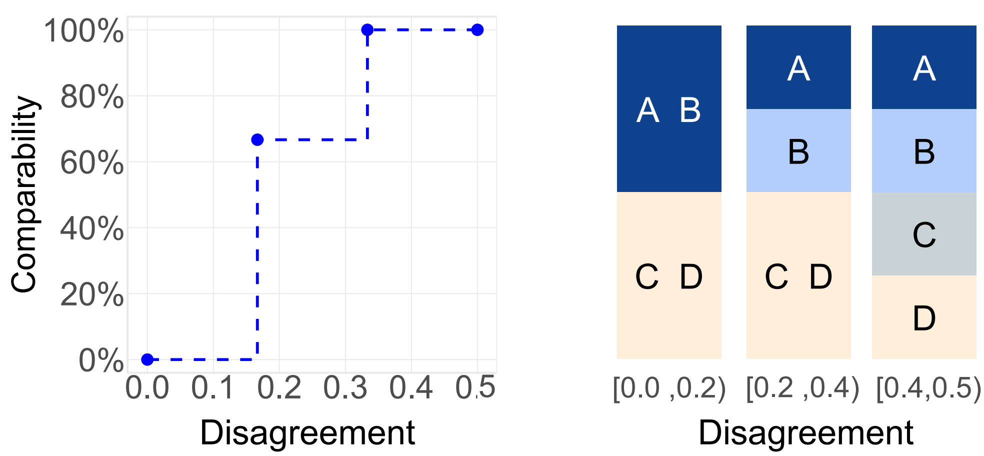


Figure 3: All possible selective rankings for the task in Fig. 1 where we aggregate the preferences of $p = 5$ users over $n = 4$ items $\{A, B, C, D\}$. We show the comparability and disagreement of each solution to SPA_τ on the left, and their selective rankings on the right. Here, the solution for $\tau \in [0, \frac{1}{5}]$ reveals that all users unanimously prefer $\{A, B\}$ to $\{C, D\}$. The solution for $\tau \in (\frac{1}{5}, \frac{2}{5}]$, reveals that we can recover a single winner if we are willing to make claims that overrule at most 1 user, while the solution for $\tau \in (\frac{2}{5}, \frac{1}{2}]$ reveals we can only recover a total order if we are willing to overrule at most 2 users.

We describe this procedure in Algorithm 2. Both Algorithms 1 and 2 run in time $\mathcal{O}(n^2 p)$ – i.e., they are linear in the number of individual pairwise preferences elicited (see Appendix B.4). As we show in Fig. 4, the resulting approach can lead to an improvement in runtime in practice.

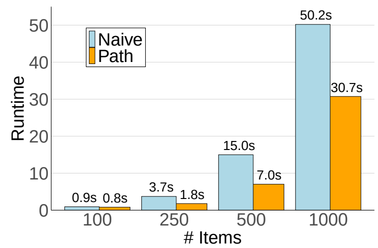


Figure 4: Runtimes to produce all selective rankings for a synthetic task with $p = 10$ users and n items (see Appendix A.2 for details). We show results for a naïve approach where we call Algorithm 1 for all possible dissent values, and the solution path algorithm in Appendix A.2. All results reflect timings on a consumer-grade CPU with 2.3 GHz and 16 GB RAM.

Algorithm We present an algorithm to construct a solution path of selective rankings in Algorithm 2.

Algorithm 2 Solution Path Algorithm

Input: $\mathcal{D} = \{\pi^k_{i,j}\}_{i,j\in[n],k\in[p]}$ *preference dataset*

1: $\mathcal{S} = \{\}$ *initialize solution path*

Construct Initial Preference Graph for $\tau = 0$

2: $w_{i,j} \leftarrow \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} \geq 0\right]$ for all $i, j \in [n]$ *$w_{i,j}$ = # preferences claiming $i \succeq j$*

3: $V_I \leftarrow [n]$ *Vertices represent items*

4: $A_I \leftarrow \{(i \to j) \mid w_{i,j} \geq 0\}$ *Arcs for observed preferences*

Construct Selective Rankings for All Possible Dissent Values

5: $\mathcal{W} \leftarrow \{w_{i,j}$ for all $i, j \in [n] \mid w_{i,j} < \lceil \frac{p}{2} \rceil\} \cup \{0\}$ *Set of dissent parameters (see Proposition **B.8**)*

6: **for** $\tau \in \mathcal{W}$ **do**

7: $\quad A_I \leftarrow A_I / \{(i \to j) \in \mid w_{i,j} \geq \tau p\}$ *Add arcs with support $\geq \tau p$*

8: $\quad V_T \leftarrow \mathsf{ConnectedComponents}((T, A_T))$ *Group items into tiers*

9: $\quad A_T \leftarrow \{(T \to T') \mid \exists i \in T, j \in T' : (i \to j) \in A_I\}$ *Add edges between items to supervertex*

10: $\quad (l_1, \ldots, l_{|V_T|}) \leftarrow \mathsf{TopologicalSort}((V_T, A_T))$ *Sort components based on directed edges*

11: $\quad S_\tau \leftarrow (T_{l_1}, \ldots, T_{l_{|V_T|}})$

12: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{S_\tau\}$

13: **end for**

Output: $\mathcal{S}$ *Selective rankings that cover the comparison-disagreement frontier*

Given a preference dataset Algorithm 2 returns a finite collection of selective rankings $\mathcal{S}$ that achieve all possible trade-offs of comparability and dissent. The procedure improves the scalability by restricting the values of the dissent parameter τ as per Proposition B.8 in Line 2, and by reducing the overhead of computing graph structures. In this case, we construct the preference graph once in Line 4, and progressively add arcs with sufficient support in Line 7.

Algorithm 2 assumes a complete preference dataset – i.e., where we have all pairwise preferences from all users. In practice, we can satisfy this assumption by imputing missing preferences to 0 as described in Proposition B.2. Alternatively, we can also add an additional step after Line 7 to check that the item graph (V_I, A_I) remains connected.

Details on Synthetic Dataset in Fig. 4 We benchmarked Algorithm 2 against Algorithm 1 in Fig. 4 on synthetic preference aggregation tasks where we could vary the number of users and items. We fixed the number of users to $p = 10$ users. For each user $k \in [p]$, we sampled their pairwise preferences as $\pi^k_{i,j} \sim \mathsf{Uniform}(1, 0, -1)$.

B.4 Proofs of Algorithm Runtime

Algorithm 1 Line 1 computes a sum while visiting each pairwise preference for each judge, taking $\mathcal{O}(n^2 p)$ time. All subsequent steps are linear in the graph size: both ConnectedComponents and TopologicalSort are linear in input size, and the other steps are just operations on each edge. So the total runtime is $\mathcal{O}(n^2 p)$.

Algorithm 2 Note that $|\mathcal{W}| = \lceil \frac{p}{2} \rceil$, because w_{ij} only takes integer values and there are $\lceil \frac{p}{2} \rceil$ integers between 0 and $\lceil \frac{p}{2} \rceil$ inclusive of 0 and exclusive of $\lceil \frac{p}{2} \rceil$. so the for loop runs $\lceil \frac{p}{2} \rceil$ times, and everything in the loop runs in time linear in the graph size, so $\mathcal{O}(n^2)$. Thus the whole runtime of the loop is $\mathcal{O}(n^2 p)$. The preprocessing, as before, is $\mathcal{O}(n^2 p)$ time. Note that computing $\mathcal{W}$ can be done in $\mathcal{O}(n^2 p)$ time: just iterate through all w_{ij} for each of the $\lceil \frac{p}{2} \rceil$ possible distinct values, and add the value to $\mathcal{W}$ if it occurs at least once. Thus the total runtime is the sum of a constant number of $\mathcal{O}(n^2 p)$ steps, meaning the total runtime is $\mathcal{O}(n^2 p)$.

C Supplementary Material for Appendix B

This appendix provides proofs and additional results to support the claims in Appendix B.

C.1 On the Top Tier

Theorem C.1. *Consider a preference aggregation task where at most $\alpha < \frac{1}{2}$ of users strictly prefer one item over all other items. Given any $\tau \in [0, \frac{1}{2})$, the tiered ranking from SPA_τ will include at least two items in its top tier.*

Proof. We show the contrapositive: having $> (1-\tau)$ users rank an item first guarantees having only one item in the top tier. Without loss of generality, call an item with $> (1-\tau)$ users rating a specific item first A. Consider WLOG any other item B. No more than τ users claim either of $B \succ A$ or $B \sim A$, because we know $> (1-\tau)$ users claim $A \succ B$. So for any tiered ranking that places some other item B in the same tier as A, we could instead place A above all other items in that tier, and have one more item. Since the result of our algorithm must have the maximal number of tiers, we cannot have a case where A is in the same tier as any other item. $\square$

Lemma C.2. *Consider a preference aggregation task where a majority of users strictly prefer an item i_0 over all items $i \neq i_0$. There exists some threshold dissent $\tau_0 \in [0, \frac{1}{2})$ such that for all $\tau > \tau_0$, every selective ranking we obtain by solving SPA_τ will place i_0 as the sole item in its top tier.*

Proof. Let α denote the fraction of users who strictly prefer i_0 over all items. Since $\alpha > \frac{1}{2}$, we observe that at most $1 - \alpha < 1 - \frac{1}{2}$ users can express a conflicting preference. Given any item $i \neq i_0$, let $\tau_0 = 1 - \alpha$ denote the fraction who users who believe either of $i \succ i_0$ or $i \sim i_0$. For any tiered ranking that places i_0 and i in the same tier, we could instead place i above all other items in that tier, and have one more tier. Since our algorithm returns a tiered ranking with the maximal number of tiers, we cannot have a case where i is in the same tier as any other item. $\square$

C.2 On Missing Preferences

Proof of Proposition B.2. If we are missing preferences, our algorithm's behavior is to assume all missing preferences would be in disagreement with any asserted ordering. This exactly corresponds to the actual disagreement if the true values are all asserted equivalence/indifference, and an upper bound on dissent if the preferences are directional. By doing this, we guarantee that the disagreement property will be satisfied under any possible missingness mechanism, even a worst-case adversarial mechanism. We denote missingness as $\pi_k(i,j) =?$ if the preference is missing. This property is trivial to show. Consider that

$$
\begin{aligned}
\text{Disagreements}(T) &:= \max_{\substack{i,j \in T,T' \\ T \succ T'}} \sum_{k \in [p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right] \\
&\leq \max_{\substack{i,j \in T,T' \\ T \succ T'}} \sum_{k \in [p]} 1\left[\pi_{i,j}^k \in \{0, -1, ?\}\right] \\
&= \max_{\substack{i,j \in T,T' \\ T \succ T'}} \sum_{k \in [p]} \mathbb{I}\left[\pi_{i,j}^k \in \{0, -1\}\right] \text{ if we we set all missing values } \pi_{i,j}^k =? \text{ to } \pi_{i,j}^k = 0
\end{aligned}
$$

Given that overall disagreement when preferences are imputed cannot increase, we have that $\pi_{i,j}(S_\tau^{\text{true}}) = \pi_{i,j}(S_\tau^{\text{safe}})$.

More formally: from the disagreements argument above, we know that $\mathcal{D}^{\text{safe}}$ has the same or more disagreements for any preference than does $\mathcal{D}^{\text{true}}$. Every selective comparison in S_τ^{safe} corresponds to a pair of items in distinct strongly connected components under the constraints from $\mathcal{D}^{\text{safe}}$ (see Lemma B.4). When we relax to only the constraints from $\mathcal{D}^{\text{true}}$, we cannot have more disagreement for any preferences, so those items will remain in distinct strongly connected components. Since they remain in distinct strongly connected components, Lemma B.4 tells us the two items will not be in the same tier.

To show that the two items will have the same ordering in both tiered rankings, note that even under $\mathcal{D}^{\text{true}}$ there must be a constraint on one of the two directions of the preference[2]. And that constraint will still hold under $\mathcal{D}^{\text{safe}}$, which is no less constrained than $\mathcal{D}^{\text{true}}$. Thus, S_τ^{true} cannot have a preference in the opposite direction from S_τ^{safe}

□

C.3 On the Distribution of Dissent

A selective ranking only allows comparisons that violate at most τp of preferences in a dataset. In practice, these violations may be disproportionately distributed across users or items. For example, we may have a task with $\tau = \frac{1}{p}$ where the same user disagrees with all comparisons in a dataset. Alternatively, the violations may be equally distributed across users – so that there is no coalition of users who agrees with all preferences. In Remark C.3, we bound the number of users who can disagree with a selective ranking.

Remark C.3. A τ-selective ranking contradicts the preferences of at most $\frac{p^2}{4} \cdot \tau p$ users.

The result in Remark C.3 only applies in tasks where the number of users exceeds the number of selective comparisons. In other tasks – where the number of selective comparisons exceeds the number of users – the statement is vacuous as we cannot rule out a worst-case where every user disagrees with at least one comparison.

Proof. We observe that a selective ranking with a single tier makes no claims. Thus we can restrict our attention to cases where the τ-selective ranking contains at least two tiers. Given a selective ranking with more than 2 tiers, then any user who disagrees with the ranking of items from non-adjacent tiers, also disagrees with the ranking of two items in adjacent tiers. So every user with a conflict must disagree about the ordering of at least one pair of items in adjacent tiers. This bounds the number of users who disagree as τ times the number of distinct pairs of items in adjacent tiers. This is because no more than τ proportion of users can disagree with any one pairing.

The number of distinct, adjacent-tier pairs is of the form $\sum_{l=1}^{|T|-1} n_l n_{l+1}$ where tier ; contains n_l items, and all the tiers together contain all n items ($\sum_{i=l} |T| n_l = n$). This quantity is maximized when we have $|T| = 2$ tiers that contain $\frac{n}{2}$ items each (rounding if n is odd). In this case, the maximum value is $\frac{n}{4}$ (or slightly below if n is odd). The worst case is tight, achieved with two tiers, each with half the items, and an even number of items. □

C.4 On Stability with Respect to New Items

We start with a simple counterexample to show that selective rankings do not satisfy the "independence of irrelevant alternatives" axiom [4].

Example C.4 (Selective Rankings do not Satisfy IIA)**.** Consider a preference aggregation task where we have pairwise preferences from 2 users for 2 items i and j where both users agree that $i \succ j$.

$$\begin{aligned} \text{User 1}: &\quad i \succ j \\ \text{User 2}: &\quad i \succ j \end{aligned}$$

In this case, every τ-selective ranking would be $\pi_{i,j}(T) = 1$ for any $\tau \in [0, 0.5)$.

Suppose we elicit preferences for a third item z, and discover that each user asserts that z is equivalent to a different item:

$$\begin{aligned} \text{User 1}: &\quad i \sim z \succ j \quad \longleftrightarrow \quad i \succ j \quad z \succ j \quad i \sim z \\ \text{User 2}: &\quad i \succ j \sim z \quad \longleftrightarrow \quad i \succ j \quad j \sim z \quad i \succ z \end{aligned}$$

In this case, every τ-selective ranking would be $\pi_{i,j}(T) = 0$ for all $\tau \in [0, \frac{1}{2})$. This violates IIA because the relative comparison $\pi_{i,j}(T)$ changes depending on the preferences involving z.

[2] Given a dataset of complete preferences and $\tau \in [0, \frac{1}{2})$, at least one of the following must hold: $\sum_{k \in [p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right] > \tau p$ or $\sum_{k \in [p]} \mathbb{I}\left[\pi_{i,j}^k \neq -1\right] > \tau p$. This is because for the former claim to be true, we'd need at least $(1-\tau)p$ preferences to be 1, which then forces the latter claim to be false because we've set $(1-\tau)p > \tau p$ values to be something other than -1.

Proposition C.5. *Consider a preference aggregation task where for a given $\tau \in [0, \frac{1}{2})$ we construct a selective ranking S_n using a dataset $\mathcal{D}$ of complete pairwise preferences from p users over n items in the itemset $[n]$. Say we elicit pairwise preferences from all p users with respect to a new item $n+1$ and construct a selective ranking S_{n+1} for the same τ over the new itemset $[n+1]$. Given any two items $i, j \in [n]$, we have that*

$$(\pi_{i,j}(S_{n+1}) = \pi_{i,j}(S_n)) \vee (\pi_{i,j}(S_{n+1}) = 0).$$

Proof. It is sufficient to show the following:

- When $\pi_{i,j}(S_n) \neq -1$, we never have $\pi_{i,j}(S_{n+1}) = -1$
- When $\pi_{i,j}(S_n) \neq 1$, we never have $\pi_{i,j}(S_{n+1}) = 1$.

Given a dataset of complete pairwise preferences and $\tau \in [0, \frac{1}{2})$, at least one of the following conditions must hold:

$$\text{Condition I:} \qquad \sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right] > \tau p$$

$$\text{Condition II:} \qquad \sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq -1\right] > \tau p$$

This is because for Condition I to be False, we would need at least $(1-\tau)p$ preferences to be 1, which then forces Condition II to be true because we have set $(1-\tau)p > \tau p$ values to be something other than -1.

Consider WLOG that Condition I holds. If $\sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right] > \tau p$, then we know that $\pi_{i,j}(S_n) \neq 1$. Otherwise we would violate the disagreement constraint in SPA_τ. Note that eliciting preferences for a new item does not change $\sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right]$. So we still have $\sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq 1\right] > \tau p$, and we still have $\pi_{i,j}(S_{n+1}) \neq 1$. Thus, we have that both $\pi_{i,j}(S_n) \neq 1$ and $\pi_{i,j}(S_{n+1}) \neq 1$. We can apply a symmetric argument to show Condition II holds. In this case, we would have that $\sum_{k\in[p]} \mathbb{I}\left[\pi_{i,j}^k \neq -1\right] > \tau p$ and see that both $\pi_{i,j}(S_n) \neq -1$ and $\pi_{i,j}(S_{n+1}) \neq -1$.

This guarantees that the claim of Proposition B.3 cannot be violated. When $\pi_{i,j}(S_n) = 0$ so too does $\pi_{i,j}(S_{n+1}) = 0$. When $\pi_{i,j}(S_n) \neq -1$ we never have $\pi_{i,j}(S_{n+1}) = -1$, when $\pi_{i,j}(S_n) \neq 1$ we never have $\pi_{i,j}(S_{n+1}) = 1$. Thus we have proven the claim by cases. $\square$

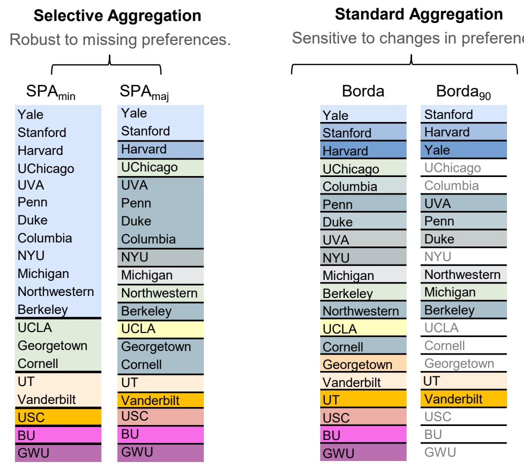


Figure 5: Consensus rankings of U.S. law schools from selective preference aggregation and standard voting rules for the `lawschool` dataset. On the left, we show selective rankings $\mathsf{SPA}_{\text{min}}$ and $\mathsf{SPA}_{\text{maj}}$ for dissent values of $\tau_{\text{min}} = \frac{1}{5}$ and $\tau_{\text{max}} = \frac{2}{5}$. On the right we see Borda on the full dataset, and Borda_{90} after removing 10% of pairwise preferences — illustrating sensitivity to missing data.

D Supplementary Material for Sections 3 and 4

In what follows, we include additional details and results for the experiments in Section 3 and our demonstration in Section 4.

D.1 Descriptions of Datasets

Dataset	p	n	Format	Description
`nba`	7 Coaches	100 Voters	Ballots	2021 NBA Coach of the Year Award, where sports journalists vote for the top 3 coaches
`lawschool`	5 Rankings	26 Schools	Rankings	Top U.S. law schools ranked by 5 organizations based on academic performance, reputation, and other metrics in 2023.
`survivor`	6 Fans	39 Seasons	Rankings	Rankings task where 6 fans of the show Survivor rank seasons 1-40 from best to worst.
`sushi`	5,000 Respondents	10 Sushi Types	Pairwise	Benchmark recommendation dataset collected in Japan, where participants provided pairwise preferences over 10 different types of sushi: ebi (shrimp), anago (sea eel), maguro (tuna), ika (squid), uni (sea urchin), ikura (salmon roe), tamago (egg), toro (fatty tuna), tekka-maki (tuna roll), and kappa-maki (cucumber roll).
`csrankings`	5 Subfields	175 Departments	Rankings	Rankings of computer science departments from csrankings.org based on research output in AI, NLP, Computer Vision, Data Mining, and Web Retrieval.

Table 4: Overview of datasets. We consider five datasets from salient use cases of preference aggregation.

D.2 List of Metrics

In what follows, we provide detailed descriptions of the metrics in Table 1.

Metric	Formula	Description
AbstentionRate(T)	$\frac{1}{n(n-1)} \sum_{i,j \in [n]} \mathbb{I}\left[\pi_{i,j}(T) = \perp\right]$	Given a selective ranking over n items T, the abstention rate represents the fraction of pairwise comparisons where we abstain.
DisagreementRate($T, \mathcal{D}$)	$\frac{1}{n(n-1)p} \sum_{k \in [p]} \sum_{i,j \in [n]} \mathbb{I}\left[\pi^k_{i,j} \neq \pi_{i,j}(T), \pi_{i,j}(T) \neq \perp\right]$	Given a ranking over n items T, the *disagreement rate* represents the fraction of individual preferences in $\mathcal{D}$ that disagree with the collective preferences in T.
#Tiers(S_τ)	$\lvert S_\tau \rvert$	Given a selective ranking S_τ, the number of tiers. For standard methods, each rank is converted to a tier.
#TopItems(S_τ)	$\lvert T_1 \rvert$	Given $S_\tau = (T_1, \ldots, T_m)$, the number of items in the top tier. For standard methods, each rank is converted to a tier.
DisagreementPerUser($T, \mathcal{D}$)	$\operatorname{median}_{k \in [p]} \frac{1}{n(n-1)/2} \sum_{i,j \in [n]} \mathbb{I}\left[\pi^k_{i,j} \neq \pi_{i,j}(T)\right]$	The median fraction of preference violations across users.
Δ Sampling $(T, \mathcal{D})$	$\operatorname{median}_{b \in \{1,\ldots,N_b\}} \left[\frac{\sum_{i,j \in [n]} \mathbb{I}\left[T_{i,j} \neq T^b_{i,j} \wedge T_{i,j} \neq 0 \wedge T^b_{i,j} \neq 0\right]}{\sum_{i,j \in [n]} \mathbb{I}\left[T_{i,j} \neq 0\right]}\right]$	Given the ranking produced on the full dataset T, the median proportion of collective preferences that are inverted when we drop 10% of preferences. We construct a bootstrap estimate by applying the method to N_b datasets where we randomly drop 10% of all preferences and obtain N_b rankings $\{T^1, \ldots, T^{N_b}\}$.
Δ Adversarial $(T, \mathcal{D})$	$\max_{b \in \{1,\ldots,N_b\}} \left[\frac{\sum_{i,j \in [n]} \mathbb{I}\left[T_{i,j} \neq T^b_{i,j} \wedge T_{i,j} \neq 0 \wedge T^b_{i,j}\right] \neq 0}{\sum_{i,j \in [n]} \mathbb{I}\left[T_{i,j} \neq 0\right]}\right]$	Given the original ranking T, the *maximum* proportion of collective preferences inverted when we flip 10% of individual preferences. We construct a bootstrap estimate where we first apply the method to N_b datasets where we randomly flip 10% of all preferences and obtain N_b rankings $\{T^1, T^2, \ldots, T^{N_b}\}$.

Table 5: Metrics used to evaluate comparability, disagreement, and robustness of rankings in Table 1 and Appendix D.4

D.3 Selective Ranking Paths

We present the solution paths of selective rankings for each dataset in Section 3 in Fig. 6 to Fig. 10.

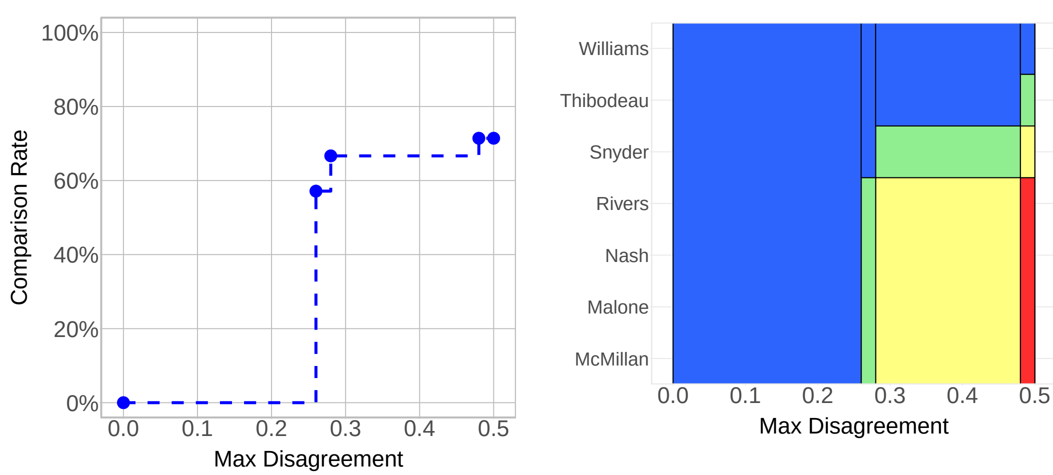


Figure 6: Selective rankings for the `nba` dataset ($n = 7$ items and $p = 100$ users). We show the tradeoff between comparison and disagreement (left) and the unique rankings over the dissent path (right).

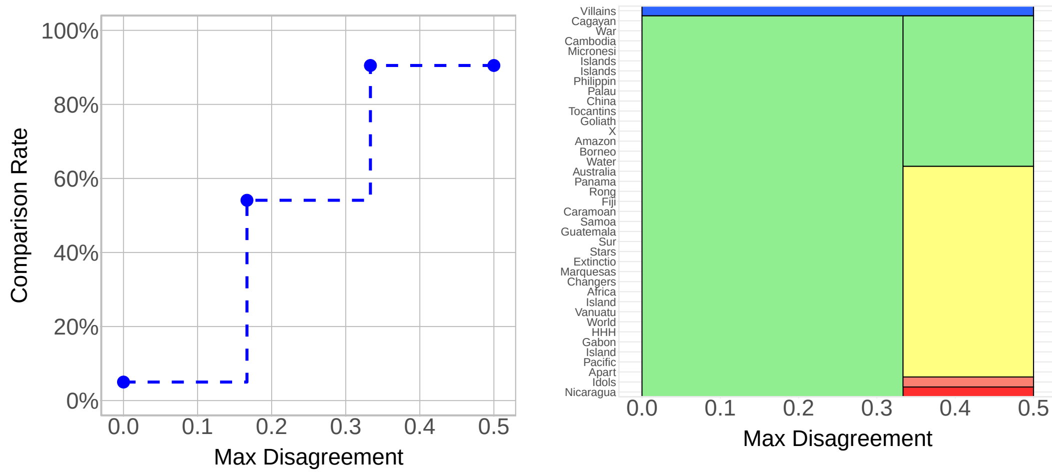


Figure 7: Selective rankings for the `survivor` dataset ($n = 39$ items and $p = 6$ users). We show the tradeoff between comparison and disagreement (left) and the unique rankings over the dissent path (right).

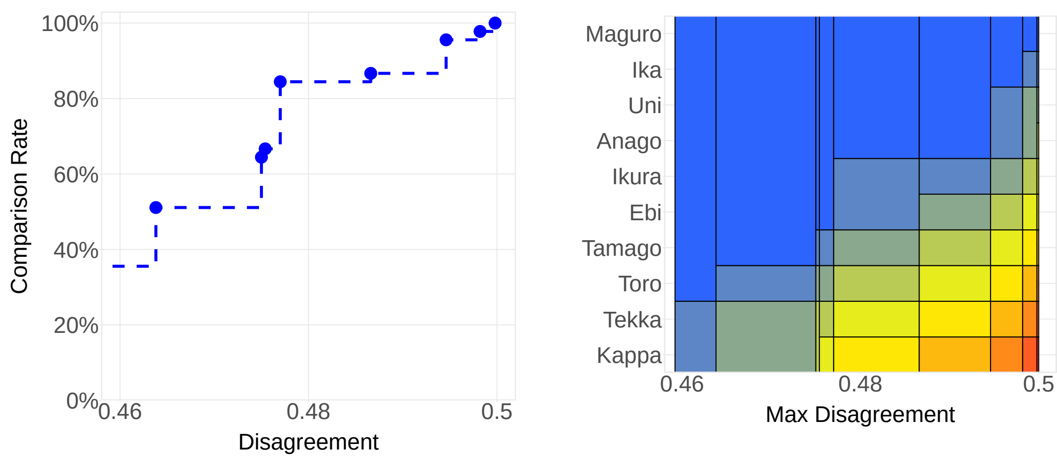


Figure 8: Selective rankings for the `sushi` dataset ($n = 10$ items and $p = 5000$ users). We show the tradeoff between comparison and disagreement (left) and the unique rankings over the dissent path (right). Note that only a subset of dissent values are shown for clarity, focusing on the largest areas of change.

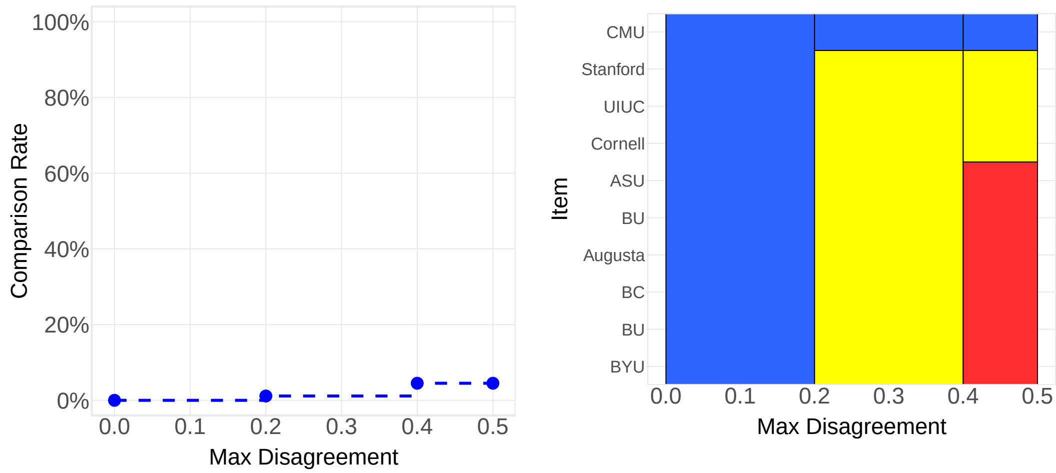


Figure 9: Selective rankings for the `csrankings` dataset ($n = 175$ items and $p = 5$ users). We show the tradeoff between comparison and disagreement (left) and the unique rankings over the dissent path (right). We show the top 10 items, sorted by tier and alphabetically within each tier.

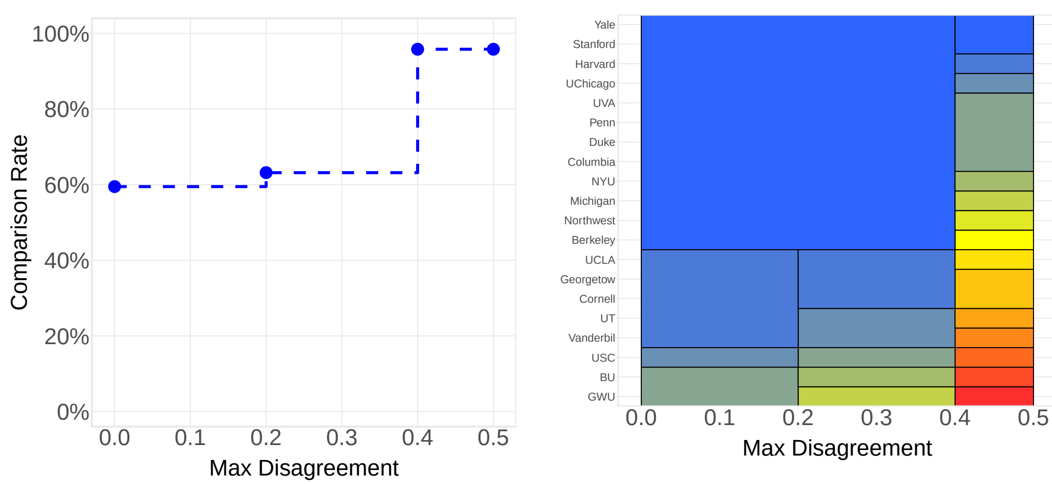


Figure 10: Selective rankings for the `lawschool` dataset ($n = 20$ items and $p = 5$ users). We show the tradeoff between comparison and disagreement (left) and the unique rankings over the dissent path (right).

D.4 Expanded Table of Results

We include an expanded version of our results for all methods and all datasets in Appendix D.4. This table covers the same results as in Table 1, but includes the following additional metrics:

1. Δ *Abstentions [Intervention]*, which measures the proportion of strict collective preferences (e.g., $A \succ B$ or $A \prec B$) that turn into ties or abstentions in the ranking that we obtain after running the method on a modified dataset.
2. Δ *Specifications [Intervention]*, which measures the proportion of ties or abstentions that turn into ties or abstentions in the ranking that we obtain after running the method on a modified dataset.

We report these values for same interventions we consider in Section 3, namely: *Sampling*, where we run the method on a dataset where we randomly omit 10% of individual preferences; and *Adversarial*, where we run the method on a dataset where we randomly flip 10% of individual preferences. Each value corresponds to a bootstrap estimates where we perform the same estimate 100 times. For clarity, we list the Δ − Sampling as Δ − Inversions − −Sampling, and Δ − Adversarial − −Inversions.

		Selective			Standard				
Dataset	**Metrics**	SPA_0	$\mathsf{SPA}_{\min}$	$\mathsf{SPA}_{\mathrm{maj}}$	Borda	Copeland	MC4	KemenyExact	KemenyHeuristic
nba $n = 7$ items $p = 100$ users 28.6% missing NBA [40]	Disagreement Rate	0.0%	2.0%	6.4%	8.3%	8.3%	7.9%	8.1%	8.1%
	Median Disagreement per User	0.0%	0.0%	4.8%	4.8%	4.8%	9.5%	9.5%	9.5%
	Abstention Rate	100.0%	42.9%	28.6%	–	–	–	–	–
	# Tiers	1	2	4	7	7	6	7	7
	# Top Items	7	3	1	1	1	1	1	1
	Dissent	0.0000	0.2600	0.4900	–	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	4.8%	4.8%	0.0%	4.8%	4.8%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	19.0%	19.0%	19.0%	14.3%	14.3%
	Δ Specifications Sampling	0.0%	9.5%	0.0%	0.0%	0.0%	4.8%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	9.5%	0.0%	0.0%	0.0%	4.8%	0.0%	0.0%
	Δ Abstentions Sampling	0.0%	0.0%	28.6%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Adversarial	0.0%	19.0%	28.6%	0.0%	4.8%	33.3%	0.0%	0.0%
survivor $n = 39$ items $p = 6$ users 0.0% missing Purple Rock [41]	Disagreement Rate	0.0%	0.2%	0.2%	6.8%	6.6%	6.4%	6.7%	6.7%
	Median Disagreement per User	0.0%	0.1%	0.1%	7.2%	7.1%	6.8%	7.1%	7.1%
	Abstention Rate	94.9%	42.5%	42.5%	–	–	–	–	–
	# Tiers	2	5	5	39	36	35	39	39
	# Top Items	1	1	1	1	1	1	1	1
	Dissent	0.0000	0.1667	0.3333	–	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	1.3%	0.8%	0.8%	0.9%	0.9%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	2.6%	1.8%	3.1%	1.6%	1.6%
	Δ Specifications Sampling	0.0%	0.0%	0.0%	0.0%	0.4%	0.1%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	5.1%	0.0%	0.0%	0.4%	0.3%	0.0%	0.0%
	Δ Abstentions Sampling	0.0%	52.4%	57.5%	0.0%	0.1%	80.0%	0.0%	0.0%
	Δ Abstentions Adversarial	0.0%	57.5%	57.5%	0.0%	0.4%	89.5%	0.4%	0.4%
lawschool $n = 20$ items $p = 5$ users 0% missing LSData [42]	Disagreement Rate	0.0%	0.3%	3.1%	4.7%	4.2%	4.2%	4.1%	4.1%
	Median Disagreement per User	0.0%	0.0%	1.6%	4.2%	2.6%	2.6%	2.1%	2.1%
	Abstention Rate	40.5%	36.8%	4.2%	–	–	–	–	–
	# Tiers	4	6	15	20	20	19	20	20
	# Top Items	12	12	2	1	1	1	1	1
	Dissent	0.0000	0.2000	0.4000	–	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	1.6%	1.1%	0.5%	29.5%	29.5%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	3.7%	2.6%	2.6%	45.8%	45.8%
	Δ Specifications Sampling	0.0%	11.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Sampling	59.5%	28.2%	95.8%	0.0%	0.0%	55.8%	0.0%	0.0%
	Δ Abstentions Adversarial	59.5%	0.0%	95.8%	0.0%	1.6%	64.2%	0.0%	0.0%
csrankings $n = 175$ items $p = 5$ users 0.0% missing Berger [43]	Disagreement Rate	0.0%	0.0%	0.1%	12.3%	12.2%	12.2%	–	13.7%
	Median Disagreement per User	0.0%	0.0%	0.1%	12.3%	12.6%	12.3%	–	13.5%
	Abstention Rate	100.0%	98.9%	95.5%	–	–	–	–	–
	# Tiers	1	2	3	175	168	170	–	175
	# Top Items	175	1	1	1	1	1	–	1
	Dissent	0.0000	0.2000	0.4000	–	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	0.8%	0.8%	0.1%	–	9.0%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	3.1%	1.7%	0.1%	–	11.1%
	Δ Specifications Sampling	0.0%	0.0%	0.0%	0.0%	0.1%	94.4%	–	0.0%
	Δ Specifications Adversarial	0.0%	0.0%	0.0%	0.0%	0.1%	94.4%	–	0.0%
	Δ Abstentions Sampling	0.0%	1.1%	4.5%	0.0%	0.0%	0.0%	–	0.0%
	Δ Abstentions Adversarial	0.0%	0.0%	4.5%	0.0%	0.1%	0.0%	–	0.0%
sushi $n = 10$ items $p = 5,000$ users 0.0% missing Kamishima [44]	Disagreement Rate	0.0%	13.6%	42.6%	42.6%	42.6%	42.6%	42.6%	42.6%
	Median Disagreement per User	0.0%	13.3%	42.2%	42.2%	42.2%	42.2%	42.2%	42.2%
	Abstention Rate	100.0%	64.4%	0.0%	–	–	–	–	–
	# Tiers	1	2	10	10	10	10	10	10
	# Top Items	10	8	1	1	1	1	1	1
	Dissent	0.0000	0.0020	0.4998	–	–	–	–	–
	Δ Inversions Sampling	0.0%	0.0%	0.0%	0.0%	0.0%	2.2%	2.2%	2.2%
	Δ Inversions Adversarial	0.0%	0.0%	0.0%	2.2%	2.2%	11.1%	11.1%	11.1%
	Δ Specifications Sampling	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Specifications Adversarial	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Sampling	0.0%	35.6%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Δ Abstentions Adversarial	0.0%	0.0%	100.0%	0.0%	0.0%	15.6%	0.0%	0.0%

D.5 Supplementary Material for Section 4

Selective Aggregation with Binary Annotations A key challenge in applying SPA to the DICES dataset is that it elicits categorical labels for each item individually, rather than comparative ratings. This conversion can create unnecessary equivalence, where a pairwise preference is inferred as a tie ($\pi^k_{i,j} = 0$). This is not a reflection of a user's true judgment but an artifact of two limitations: (1) users annotate items individually rather than comparing them, and (2) the annotations are restricted to $\{0, 1\}$ instead of granular ratings. For example, a user may believe item A is significantly more toxic than item B, but the conversion results in a tie if both were labeled "toxic" a distinction that is lost in this setting.

We address this by running a variant of selective aggregation where we construct aggregate labels from users who express a strict preference between items – $i \succ j$ or $j \succ i$. In addition, we assume

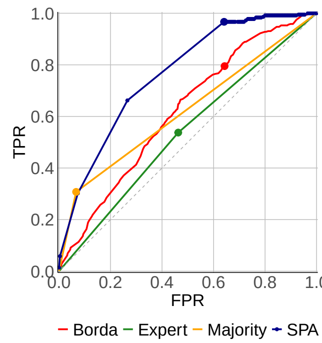


Figure 11: ROC model curves on the training set for all four methods. We highlight the label for each method closest to tpr$>$ 90% on labels with a large dot. f^{SPA} is the only method whose chosen operating point keeps the true-positive rate above 80 % on the model output while controlling FPR.

that users who have not asserted an opinion (because of dataset scope) are "deferring judgment" to those who have.

For each pair of items $i, j \in [n]$, we define:

- $s_{i,j} := \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} = 1\right]$ denote number of users who strictly prefer item i to item j
- $s_{j,i} := \sum_{k\in[p]} \mathbb{I}\left[\pi^k_{i,j} = -1\right]$ denote the number of users who strictly prefer item j to item i.
- The aggregate preference weight $w_{i,j}$ as the proportion of users who strictly prefer i to j among those who expressed a strict preference, scaled to n items. Note that all item pairs had at least 1 preference:

$$w_{i,j} := n \cdot \frac{s_{i,j}}{s_{i,j} + s_{j,i}}$$

In this setup, the dissent parameter τ no longer maintains its standard interpretation because users may not assign a preference to each item, and items may be assigned different weights. As a result, we produce selective rankings for all possible dissent parameters that lead to a connected graph in Algorithm 2. In this case, the maximum dissent value is set to a threshold value where Line 4 returns a disconnected graph.

D.6 Model Training

All experiments used 5-fold cross-validation on the training split. We fine-tuned a BERT-Mini model; all fine-tuning experiments used 5-fold cross-validation on the training split. We optimized with a learning rate of 2×10^{-5} for up to 25 epochs, employing early stopping. We trained in mini-batches of size 16 and enabled oversampling of minority classes in each batch.