

# When Does Conformal Prediction Help Beam Management? A Cross-Scenario Analysis

Johan Eliasson  
<https://github.com/elitan>

**Abstract**—ML-based mmWave beam prediction lacks reliability guarantees. We evaluate conformal prediction (CP) on two DeepMIMO scenarios at 28 GHz, finding: (1) CP benefit is scenario-dependent—set size 1.19 on boston5g\_28 (83.5% accuracy) vs. 1.00 on O1\_28 (92.2%); (2) standard CP exhibits conditional coverage gaps (79% at short distances vs. 90% target); (3) group-conditional CP closes this gap with modest overhead (set size 1.29). Results show CP value depends on base model accuracy and channel complexity.

## I. INTRODUCTION

Fifth-generation systems exploit mmWave bands for multi-gigabit throughput [1]. The 3GPP beam management framework requires sweeping over a codebook of candidate beams, consuming up to  $N$  time slots for an  $N$ -beam codebook [2]. ML-based beam prediction reduces this overhead [3], [4], but treats beam prediction as standard classification without reliability guarantees.

Conformal prediction (CP) provides distribution-free coverage guarantees [7], [8]. Recent work applies CP to beam selection [10], [11], demonstrating feasibility. However, *when does CP actually help?* If the base model is already highly accurate, CP adds little value. If conditional coverage varies across subpopulations, the marginal guarantee may be violated locally.

**Contributions.** We address this gap with three contributions:

- 1) **Cross-scenario CP analysis.** We evaluate CP on two ray-traced channels (boston5g\_28 vs. O1\_28) with different base accuracies, showing that CP benefit is scenario-dependent.
- 2) **Conditional coverage gap analysis.** We expose a distance-dependent coverage gap on boston5g\_28 where short-distance users receive only 79% coverage despite a 90% marginal target.
- 3) **Group-conditional CP.** We apply the group-conditional method of Romano et al. [9] to close the coverage gap, achieving  $\geq 90\%$  coverage in all distance bins.

## II. RELATED WORK

**Conformal prediction.** Split CP [7], [8] provides distribution-free coverage guarantees for exchangeable data. Romano et al. [9] extend CP to provide valid coverage conditionally on groups.

**CP for wireless.** Cohen et al. [12] apply CP to wireless communication problems. Hegde et al. [10] apply CP to beam selection in D-MIMO systems but do not analyze

conditional coverage. Deng et al. [11] propose SCAN-BEST using conformal risk control with different objectives (risk minimization vs. coverage). Neither work analyzes when CP provides practical value or addresses conditional coverage gaps across user subpopulations.

**Our contribution.** We provide the first conditional coverage gap analysis for beam prediction, showing that standard CP can undercover certain user groups by 18 percentage points. We are also the first to apply group-conditional CP to beam selection, demonstrating that it closes this gap at modest cost.

**ML beam prediction.** Deep learning for beam prediction from wide-beam or sub-6 GHz measurements [3]–[5] reduces overhead but lacks reliability guarantees.

## III. SYSTEM MODEL

### A. Antenna and Signal Model

We consider a single-user MISO downlink at  $f_c = 28$  GHz. The base station employs a uniform linear array (ULA) with  $M = 64$  elements at half-wavelength spacing. The array response vector for angle  $\theta$  is

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}} \left[ 1, e^{j2\pi \frac{d}{\lambda} \sin \theta}, \dots, e^{j2\pi \frac{d}{\lambda} (M-1) \sin \theta} \right]^T. \quad (1)$$

### B. DeepMIMO Scenarios

We use two DeepMIMO ray-traced scenarios [6]:

**boston5g\_28:** Urban canyon in downtown Boston. 102,762 valid users, distance range 13–247 m. Dense multipath from building reflections creates a challenging prediction environment.

**O1\_28:** Outdoor scenario with 106,000 valid users, distance range 10–200 m. Cleaner propagation with stronger LOS components yields higher prediction accuracy.

Both use 64-beam DFT narrow codebook and 16-beam wide codebook. Data splits: 80K train, 8K calibration, 8K validation, remainder test.

### C. Wide-Beam Features

For each channel  $\mathbf{h}$ , we measure 16 wide-beam powers  $p_i = |\mathbf{h}^H \mathbf{w}_i^{(w)}|^2$  and form feature vector  $\mathbf{x} = [p_0^{\text{dB}}, \dots, p_{15}^{\text{dB}}]^T$ , z-score normalized.

## IV. CONFORMAL BEAM PREDICTION

### A. Split Conformal Prediction

Given calibration set of  $n$  samples, split CP [7] computes nonconformity scores  $s_i = 1 - \hat{p}(y_i | \mathbf{x}_i)$  and threshold  $\hat{q} = \text{Quantile}(\{s_i\}, \lceil (n+1)(1-\alpha) \rceil / n)$ . At test time, beam  $j$  is

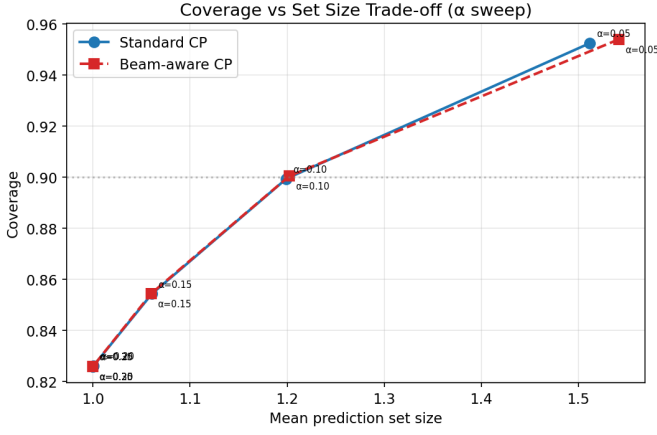


Fig. 1. Coverage vs. average set size for different  $\alpha$  values. CP achieves the target coverage (dashed line) while minimizing set size.

included if  $\hat{p}(j|\mathbf{x}) \geq 1 - \hat{q}$ . This guarantees marginal coverage  $\Pr[y \in \mathcal{C}(\mathbf{x})] \geq 1 - \alpha$ .

### B. Group-Conditional CP

Standard CP provides marginal coverage but may under-cover subpopulations. Group-conditional CP [9] partitions the calibration set by a grouping variable (here: distance quartile) and computes a separate threshold per group:

$$\hat{q}_g = \text{Quantile} \left( \{s_i : i \in \mathcal{G}_g\}, \frac{(|\mathcal{G}_g| + 1)(1 - \alpha)}{|\mathcal{G}_g|} \right) \quad (2)$$

At test time, apply threshold  $\hat{q}_{g(x)}$  based on the test point's group membership. This guarantees  $\Pr[y \in \mathcal{C}(\mathbf{x}) | G = g] \geq 1 - \alpha$  for each group  $g$ .

Fig. 1 validates CP calibration by sweeping  $\alpha$  from 0.05 to 0.30. Empirical coverage closely tracks the target  $1 - \alpha$ , confirming proper calibration. Smaller  $\alpha$  (higher coverage target) requires larger prediction sets.

### C. Adaptive Fallback Protocol

#### Algorithm 1 Adaptive Beam Management

**Require:** Channel  $\mathbf{h}$ , threshold  $K$ , calibrated model  $f$

- 1: Sweep 16 wide beams, compute  $\mathbf{x}$
- 2:  $\mathcal{C}(\mathbf{x}) \leftarrow$  conformal prediction set
- 3: **if**  $|\mathcal{C}(\mathbf{x})| \leq K$  **then**
- 4:   Sweep only beams in  $\mathcal{C}(\mathbf{x})$ ; overhead:  $16 + |\mathcal{C}|$
- 5: **else**
- 6:   Exhaustive fallback; overhead: 64 slots
- 7: **end if**

## V. EXPERIMENTAL SETUP

We evaluate four architectures: MLP (143K params), ResNet-MLP (419K), CNN (23K), and Transformer (72K). All trained with mixup ( $\alpha = 0.4$ ), cosine LR schedule, early stopping. Results: mean  $\pm$  std over 3 seeds.

TABLE I  
SIMULATION PARAMETERS

Parameter	boston5g_28	O1_28
Valid users	102,762	106,000
Distance range	13–247 m	10–200 m
Train / Cal / Val / Test	80K / 8K / 8K / rest	
Narrow / Wide beams	64 / 16	
Conformal $\alpha$	0.1 (90% target)	
Seeds	42, 123, 456	

TABLE II  
TOP-1 ACCURACY COMPARISON (MEAN  $\pm$  STD)

Model	boston5g_28	O1_28
MLP	0.811 $\pm$ 0.006	0.904 $\pm$ 0.018
ResNet-MLP	<b>0.835 <math>\pm</math> 0.007</b>	0.910 $\pm$ 0.003
CNN	0.775 $\pm$ 0.010	0.866 $\pm$ 0.010
Transformer	0.831 $\pm$ 0.003	<b>0.922 <math>\pm</math> 0.001</b>

TABLE III  
TOP-K VS CONFORMAL PREDICTION

Scenario	Method	Coverage	Avg Size	Adaptive
boston5g_28	Top-1	0.833	1	No
	Top-2	0.935	2	No
	Top-3	0.965	3	No
	CP ( $\alpha=0.1$ )	0.899	<b>1.19</b>	Yes
O1_28	Top-1	0.922	1	No
	Top-3	0.995	3	No
	Top-5	0.999	5	No
	CP ( $\alpha=0.1$ )	0.923	<b>1.00</b>	Yes

## VI. RESULTS

### A. Model Accuracy

Table II shows substantial accuracy differences between scenarios. O1\_28 achieves 92.2% top-1 accuracy (Transformer) vs. 83.5% (ResNet-MLP) on boston5g\_28. The 9-point gap reflects O1\_28's cleaner propagation environment.

### B. Comparison with Top-K Baselines

A natural question is whether CP provides value over simply selecting the top- $K$  predicted beams. Table III reveals two contrasting scenarios. On boston5g\_28 (83.3% base accuracy), Top-3 achieves 96.5% coverage but requires 3 beams; CP achieves 90% with only 1.19 beams—a 60% reduction. On O1\_28 (92.2% base accuracy), CP matches Top-1 exactly because the model rarely needs fallback beams. The key insight: **CP provides adaptive set sizes based on per-sample uncertainty**, yielding significant savings when models are less confident.

### C. Conformal Prediction: Scenario Dependence

Table IV reveals the key finding: **CP benefit is scenario-dependent**. On boston5g\_28, standard CP yields set size 1.19 beams—meaningful uncertainty quantification. On O1\_28, set size is exactly 1.00—the model is so confident that CP adds no additional beams. This demonstrates that CP provides value primarily when base accuracy is imperfect.

TABLE IV  
CONFORMAL PREDICTION RESULTS ( $\alpha = 0.1$ )

Scenario	Method	Coverage	Set Size
boston5g_28	Standard CP	$0.900 \pm 0.002$	$1.19 \pm 0.01$
	Group-CP	$0.911 \pm 0.000$	$1.29 \pm 0.02$
O1_28	Standard CP	$0.923 \pm 0.002$	<b><math>1.00 \pm 0.00</math></b>
	Group-CP	$0.932 \pm 0.005$	$1.02 \pm 0.01$

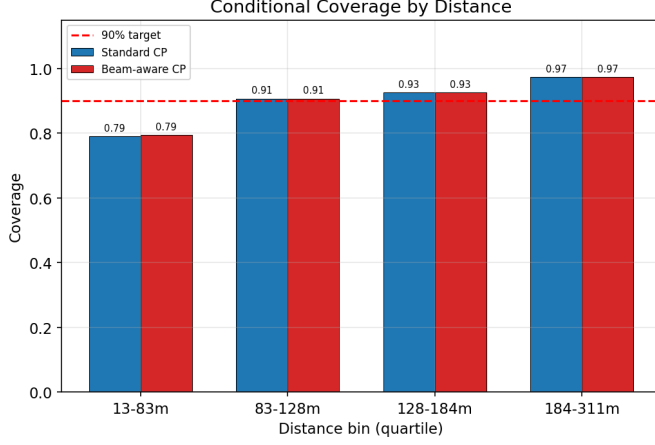


Fig. 2. Conditional coverage by distance quartile on boston5g\_28. Standard CP achieves only 79% coverage at short distances despite 90% marginal target.

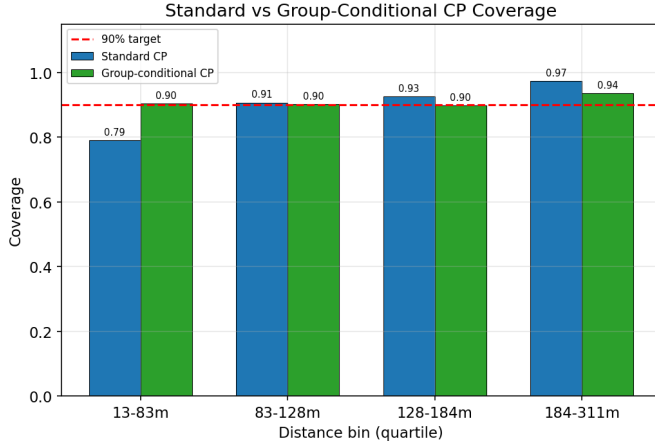


Fig. 3. Group-conditional CP achieves  $\geq 90\%$  coverage in all distance bins on boston5g\_28, at cost of slightly larger prediction sets.

#### D. Conditional Coverage Gap

Fig. 2 exposes a conditional coverage gap on boston5g\_28. Standard CP achieves only 78.8% coverage at short distances ( $< 83$  m) despite 90% marginal target, while far-distance users achieve 97%. This gap arises from dense urban multipath near the BS making prediction harder for short-distance users.

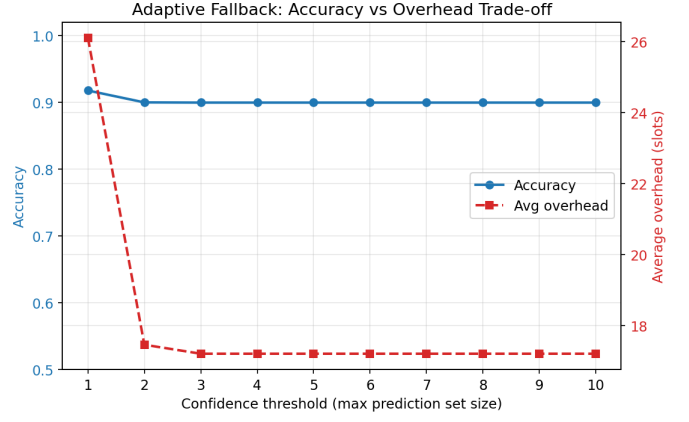


Fig. 4. Adaptive fallback on boston5g\_28. At  $K = 2$ , average overhead is 17 slots (73% reduction vs. exhaustive) with near-perfect effective accuracy.

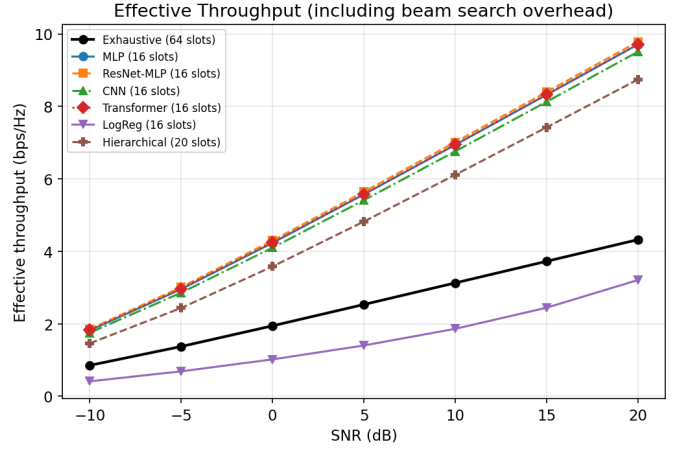


Fig. 5. Effective throughput accounting for beam measurement overhead. The overhead reduction from CP translates to throughput gain at moderate SNR.

#### E. Group-Conditional CP Closes the Gap

Group-conditional CP (Fig. 3) closes the coverage gap by calibrating separately per distance quartile. Near-distance coverage improves from 78.8% to 90.0%, while mean set size increases modestly from 1.19 to 1.29 beams. On O1\_28, group-CP has minimal effect since there is no significant conditional gap to close.

#### F. Adaptive Fallback

The adaptive protocol (Fig. 4) reduces overhead by up to 73% on boston5g\_28 while maintaining high effective accuracy. On O1\_28, with set size 1.00, the adaptive protocol achieves 17 slots overhead with 92% accuracy—strictly dominating both pure ML and exhaustive search.

Fig. 5 shows effective throughput accounting for beam measurement overhead. At moderate SNR, the overhead reduction from adaptive beam management translates directly to throughput gain compared to exhaustive search.

## VII. DISCUSSION

**When does CP help?** CP provides meaningful benefit when base model accuracy is imperfect ( $<90\%$ ). On boston5g\_28 (83.5% accuracy), CP yields set size 1.19 with useful uncertainty information. On O1\_28 (92.2% accuracy), CP adds nothing—the model is already confident enough that all prediction sets are singletons.

**Conditional coverage matters.** Even when marginal coverage meets the target, subpopulations may be under-covered. The 18-point gap between near and far users on boston5g\_28 would cause systematic service degradation for users near the BS. Group-conditional CP addresses this at modest cost.

**Scenario dependence.** The dramatic difference between scenarios (set size 1.19 vs. 1.00) demonstrates that CP benefits cannot be assumed universal. Practitioners should evaluate on their target deployment environment.

**Limitations.** (i) Two scenarios may not generalize to all deployments. (ii) Group-CP requires distance information at test time. (iii) Evaluation is static; temporal dynamics not considered.

## VIII. CONCLUSION

We evaluated conformal prediction for beam management across two ray-traced scenarios. CP benefit is scenario-dependent: meaningful on boston5g\_28 (set size 1.19) but negligible on O1\_28 (set size 1.00). Standard CP exhibits conditional coverage gaps that group-conditional CP closes. These findings caution against assuming universal CP benefit and motivate scenario-specific evaluation.

## REFERENCES

- [1] T. S. Rappaport *et al.*, “Millimeter wave mobile communications for 5G cellular,” *IEEE Access*, vol. 1, pp. 335–349, 2013.
- [2] 3GPP TS 38.214, “NR; Physical layer procedures for data,” v17.0.0, 2022.
- [3] A. Ali *et al.*, “Millimeter wave beam-selection using out-of-band spatial information,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, 2018.
- [4] M. Alrabeiah and A. Alkhateeb, “Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels,” *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, 2020.
- [5] A. Alkhateeb *et al.*, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014.
- [6] A. Alkhateeb, “DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications,” in *Proc. ITA*, 2019.
- [7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.
- [8] A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv:2107.07511*, 2021.
- [9] Y. Romano, R. F. Barber, C. Sabatti, and E. J. Candès, “With malice toward none: Assessing uncertainty via equalized coverage,” *Harvard Data Science Review*, vol. 2, no. 2, 2020.
- [10] D. N. Hegde *et al.*, “Reliable and efficient beam selection using conformal prediction in 6G systems,” in *Proc. IEEE VTC-Spring*, 2025.
- [11] W. Deng *et al.*, “SCAN-BEST: Sub-6GHz-aided near-field beam selection with formal reliability guarantees,” *arXiv:2503.13801*, 2025.
- [12] K. M. Cohen, S. Park, O. Simeone, and S. Shamai, “Calibrating AI models for wireless communications via conformal prediction,” *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 296–312, 2022.