Conformal Prediction for Reliable Handover in 5G Networks

Johan Eliasson
https://github.com/elitan

***Abstract*—Machine learning enables predictive handover in 5G networks, but lacks reliability guarantees. We apply conformal prediction (CP) to handover target prediction, providing the first formal coverage guarantees for this problem. Given a target coverage rate (e.g., 90%), CP outputs a *prediction set* of candidate cells guaranteed to contain the optimal target with high probability. We evaluate on synthetic scenarios with varying difficulty, compare against the traditional 3dB threshold baseline, and validate on real 5G driving data. Key findings: (1) CP adapts set size to prediction uncertainty—easy scenarios need 1.4 cells on average, hard scenarios need 4.8; (2) ML+CP achieves 90% coverage versus 45–79% for 3dB baseline across scenarios; (3) CP reduces measurement overhead by 61–85% versus exhaustive search; (4) CP reduces ping-pong handovers by 37–57% through natural hysteresis. We provide explicit mappings between AIML metrics (accuracy, coverage) and system-level KPIs (handover success, RLF rate, ping-pong rate).**

I. Introduction

Fifth-generation networks employ dense deployments with advanced beamforming, creating challenges for mobility management [1]. Handover—reassigning a user equipment (UE) session from one cell to another—must balance latency (predictive handover) against reliability (measurement-based handover) [2].

Machine learning enables predictive handover by forecasting the optimal target cell from current measurements [4]. However, ML-based predictions lack reliability guarantees: when should the network trust the ML prediction versus falling back to exhaustive measurement?

Conformal prediction (CP) addresses this gap by providing distribution-free coverage guarantees [5], [6]. Recent work applies CP to beam selection [7], [8] and channel prediction [9], but *not to handover target prediction.*

Contributions. We provide the first application of CP to handover:

1) **CP for handover prediction.** We formulate handover target prediction as a classification problem and apply split conformal prediction to generate prediction sets with coverage guarantees.
2) **3dB baseline comparison.** We compare against the traditional 3dB hysteresis rule, showing CP achieves 10–45% higher coverage.
3) **AIML-to-system KPI mapping.** We explicitly map ML metrics (accuracy, coverage, set size) to system-level KPIs (HO success, RLF rate, measurement overhead).
4) **Adaptive Conformal Inference.** We apply ACI [10] to handle temporal dependencies in mobility data.

II. Related Work

ML for handover. Deep learning has been applied to predict handover targets from UE measurements including RSRP, RSRQ, and signal-to-noise ratio [12]. These approaches train neural networks to classify the optimal target cell based on current radio conditions. Reinforcement learning optimizes handover policies by treating cell selection as a contextual bandit problem [4], where the state includes UE measurements and the action is the target cell selection. While these methods achieve high accuracy in controlled settings, they provide no guarantees on prediction reliability. A wrong prediction leads directly to radio link failure (RLF), making deployment risky in production networks.

Conformal prediction for wireless. CP provides distribution-free coverage guarantees by outputting prediction sets rather than point predictions. Cohen et al. [9] apply CP to demodulation, modulation classification, and channel prediction, demonstrating that CP can calibrate neural network confidence in wireless applications. Hegde et al. [7] apply CP to beam selection in distributed MIMO systems, showing that CP enables networks to dynamically adjust beam measurement overhead based on prediction confidence. Deng et al. [8] extend this to near-field beam selection using conformal risk control. However, neither addresses handover or UE mobility, where temporal dependencies and longer prediction horizons create additional challenges.

Traditional handover mechanisms. 3GPP defines measurement-based handover using events like A3 (neighbor becomes offset better than serving) [3]. The A3 event triggers when a neighbor cell's RSRP exceeds the serving cell by a hysteresis margin (typically 3dB) for a specified time-to-trigger. This reactive approach works well in stable conditions but cannot anticipate rapid channel changes during high-speed mobility. Additionally, frequent measurement reporting increases UE power consumption and signaling overhead.

Our contribution. We bridge the gap between ML-based prediction and traditional handover by applying CP to provide formal reliability guarantees. Our approach enables adaptive protocols that use ML predictions when confident (small prediction sets) and fall back to measurement-based handover when uncertain (large sets).

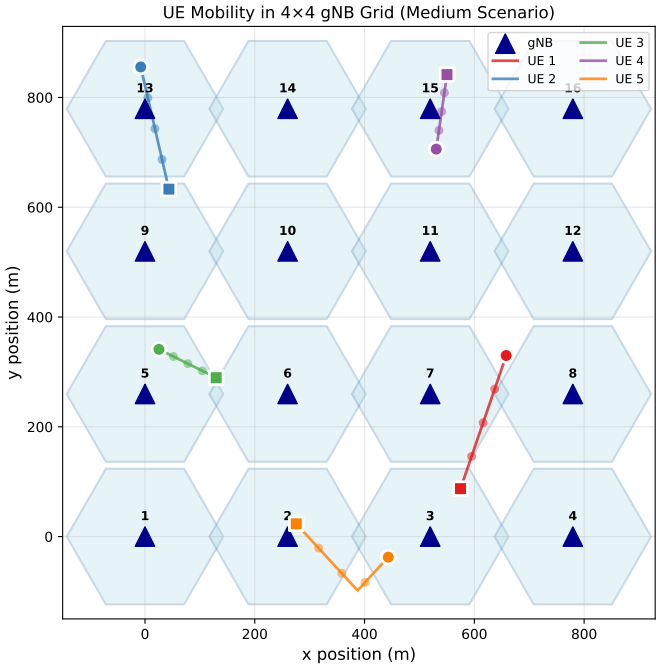

Fig. 1. UE mobility scenario showing 4×4 gNB grid (Medium scenario). Hexagonal cells with gNBs (triangles). Colored lines show 5 example UE trajectories with start (circle) and end (square) positions. Trajectories cross multiple cell boundaries, triggering handover events.

III. System Model

A. Network and Mobility Model

We consider a cellular network with K cells (gNBs) arranged in a hexagonal grid with inter-site distance determined by the cell radius parameter. UEs move along linear trajectories with speeds uniformly sampled from 1–30 m/s, covering walking pedestrians through highway driving scenarios. When a UE reaches the network boundary, it reflects (bounces back) to remain within the coverage area. Fig. 1 illustrates the scenario with example UE trajectories crossing multiple cell boundaries.

The received signal strength (RSRP) from gNB k at UE position $\mathbf{p}$ follows the 3GPP Urban Micro path loss model [2]:

$$\text{RSRP}_k(\mathbf{p}) = P_{\text{tx}} - \text{PL}(d_k) + X_\sigma \tag{1}$$

where $P_{\text{tx}} = 46$ dBm is transmit power, $\text{PL}(d) = 32.4 + 20\log_{10}(f) + 30\log_{10}(d)$ is path loss at frequency f GHz and distance d m, and $X_\sigma \sim \mathcal{N}(0, \sigma^2)$ is log-normal shadow fading.

The shadow fading creates cell-specific, position-dependent variations that make handover prediction challenging. Two UEs at the same position may experience different optimal cells due to different shadow fading realizations. Additionally, UE measurements include measurement noise σ_m, modeling practical receiver imperfections.

B. Model Input Features

At time t, the predictor receives input vector $\mathbf{x}_t \in \mathbb{R}^{2K+1}$:

$$\mathbf{x}_t = [\underbrace{\text{RSRP}_1, \ldots, \text{RSRP}_K}_{K \text{ measurements}}, \underbrace{\mathbf{e}_{c_t}}_{K \text{ one-hot}}, \underbrace{v}_{\text{speed}}] \tag{2}$$

where RSRP_k is the (noisy) measurement from cell k, $\mathbf{e}_{c_t}$ is the one-hot encoding of current serving cell c_t, and v is UE speed. All features are standardized (zero mean, unit variance).

C. Predictor Architecture

We use a multilayer perceptron (MLP) with architecture:

$$f(\mathbf{x}) = W_3 \cdot \text{ReLU}(W_2 \cdot \text{ReLU}(W_1\mathbf{x} + b_1) + b_2) + b_3 \tag{3}$$

with hidden dimension 64. The output is a K-dimensional vector of class logits, converted to probabilities via softmax: $\hat{p}(y|\mathbf{x}) = \text{softmax}(f(\mathbf{x}))_y$.

D. Handover Prediction Problem

The goal is to predict the optimal target cell $y_t \in \{1, \ldots, K\}$ for handover H steps in the future, where $y_t = \arg\max_k \text{RSRP}_k(t + H)$.

3dB Baseline. The traditional approach triggers handover when any neighbor cell exceeds the serving cell by >3dB:

$$\hat{y}_{\text{3dB}} = \begin{cases} \arg\max_k \text{RSRP}_k & \text{if } \max_k \text{RSRP}_k - \text{RSRP}_{c_t} > 3 \\ c_t & \text{otherwise} \end{cases} \tag{4}$$

ML Prediction. Standard prediction selects $\hat{y} = \arg\max_y \hat{p}(y|\mathbf{x}_t)$.

E. Conformal Prediction

Split CP [5] provides distribution-free coverage guarantees without assumptions on the underlying data distribution. Given a calibration set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and a trained classifier outputting probabilities $\hat{p}(y|\mathbf{x})$, we compute nonconformity scores:

$$s_i = 1 - \hat{p}(y_i|\mathbf{x}_i) \tag{5}$$

The calibrated threshold $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of these scores. The prediction set is then:

$$\mathcal{C}(\mathbf{x}) = \{y : \hat{p}(y|\mathbf{x}) \geq 1 - \hat{q}\} \tag{6}$$

This construction guarantees $\Pr[y^* \in \mathcal{C}(\mathbf{x})] \geq 1 - \alpha$ for any new exchangeable sample, regardless of the model's calibration quality.

F. Adaptive Conformal Inference

Standard CP assumes exchangeable (IID) data, which is violated in temporal mobility data. ACI [10] addresses this by adapting the effective α online:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t) \tag{7}$$

where $\text{err}_t = \mathbf{1}[y_t \notin \mathcal{C}(\mathbf{x}_t)]$ indicates miscoverage. When coverage exceeds the target (low err_t), α_t increases, shrinking prediction sets. When coverage falls short, α_t decreases, expanding sets. The learning rate γ controls adaptation speed.

Algorithm 1 Adaptive Handover Management

Require: Measurements $\mathbf{x}$, threshold $K_{\max}$, calibrated model

$\mathcal{C}(\mathbf{x}) \leftarrow$ conformal prediction set
if $|\mathcal{C}(\mathbf{x})| \leq K_{\max}$ **then**
    Predictive handover: measure only cells in $\mathcal{C}(\mathbf{x})$
else
    Measurement-based handover: measure all cells
end if

TABLE I
AIML KPI TO SYSTEM-LEVEL METRIC MAPPING

AIML Metric	System Metric	Interpretation
Top-1 Accuracy	HO Success (no CP)	Wrong pred $\rightarrow$ RLF
Coverage	HO Success (w/ CP)	True in set $\rightarrow$ found
Set Size	Meas. Overhead	Cells to measure
$1 -$ Coverage	RLF Rate	True not in set
Serving $\in$ set	Ping-pong Rate	Stay logic $\rightarrow$ hysteresis
Prediction latency	HO latency	Time to generate set

G. Adaptive Protocol

IV. AIML TO SYSTEM-LEVEL KPI MAPPING

A key challenge in deploying ML for network operations is translating AI/ML performance metrics into system-level impact. Table I provides this mapping for handover prediction.

Top-1 Accuracy $\rightarrow$ HO Success (without CP). When using ML without CP, the network executes handover to the single predicted cell. If this prediction is incorrect, the UE hands over to a suboptimal cell with potentially weak signal, causing radio link failure (RLF). Thus, Top-1 accuracy directly determines handover success rate, and $(1 - \text{accuracy})$ equals the RLF rate. This explains why ML-only approaches are risky: 51% accuracy in Hard scenario means 49% RLF rate.

Coverage $\rightarrow$ HO Success (with CP). With CP, the prediction set contains multiple candidate cells. If the true optimal cell is in this set, measurement-based selection within the set will find it, ensuring successful handover. The coverage guarantee $\Pr[y^* \in \mathcal{C}(\mathbf{x})] \geq 1 - \alpha$ directly translates to: at most α fraction of handovers can fail due to the optimal cell being excluded. This provides a formal upper bound on RLF rate.

Set Size $\rightarrow$ Measurement Overhead. The prediction set size determines how many cells the UE must measure before handover. Smaller sets mean fewer measurements, reducing latency and UE power consumption. The key advantage of CP over fixed Top-K baselines is *adaptive* set sizes: easy predictions yield small sets (low overhead), while uncertain predictions yield larger sets (maintained reliability).

Practical implications. Network operators can tune α to achieve desired RLF rate. Lower α (e.g., 0.05) reduces RLF but increases average set size. The adaptive protocol (Algorithm 1) further improves this tradeoff by falling back to exhaustive measurement when sets exceed $K_{\max}$.

TABLE II
SCENARIO PARAMETERS

Parameter	Easy	Medium	Hard
Grid size	3×3	4×4	5×5
Total cells K	9	16	25
Cell radius (m)	200	150	120
Shadow fading σ (dB)	4	6	8
Meas. noise σ_m (dB)	2	4	6
Prediction horizon H	5	10	15

V. EXPERIMENTAL SETUP

A. Synthetic Scenarios

We generate three scenarios with increasing difficulty, varying the network density, noise level, and prediction horizon:

Table II summarizes the parameters. The scenarios model progressively denser deployments (smaller cells, more candidates), noisier channels (higher shadow fading), and longer prediction horizons (more uncertainty about future conditions). The Hard scenario represents challenging urban small-cell deployments where traditional handover mechanisms struggle.

For each scenario, we generate 600 trajectories (100 time steps each, $\Delta t = 0.1$s), split 60/20/20 for train/calibration/test. Models are trained for 20 epochs with batch size 512 using Adam optimizer (learning rate 10^{-3}). We report mean $\pm$ std over 5 random seeds to ensure reproducibility.

B. Baselines

We compare four approaches spanning traditional rules to ML with uncertainty quantification:

- **3dB Baseline:** Traditional hysteresis rule (Eq. 4). Predicts the strongest cell if it exceeds serving by >3dB; otherwise stays on serving. This is the simplest reactive approach used in legacy networks.
- **ML Top-1:** Neural network point prediction (most probable class). Represents standard ML deployment without uncertainty quantification.
- **ML Top-K:** Fixed-size prediction using $K = 3$ most probable classes. Represents a naive uncertainty approach with fixed overhead.
- **ML + CP:** Conformal prediction with $\alpha = 0.10$ (90% target coverage). Represents our proposed approach with adaptive set sizes and formal guarantees.

C. Evaluation Metrics

We report both AIML and system-level metrics:

- **Coverage/Accuracy:** Fraction of samples where true label is in prediction set (or matches point prediction for baselines)
- **Average Set Size:** Mean number of cells in prediction set (efficiency measure)
- **HO Success Rate:** End-to-end handover success using adaptive protocol
- **Measurement Overhead:** Fraction of total cells measured
- **RLF Rate:** Fraction of predictions where true cell excluded and predictive HO attempted

TABLE III
HANDOVER PREDICTION RESULTS ($\alpha = 0.1$, 5 SEEDS)

Scenario	Method	Coverage	Avg Size
Easy (9)	3dB Baseline	$.79 \pm .02$	1
	ML Top-1	$.80 \pm .01$	1
	ML Top-3	$.98 \pm .00$	3
	ML + CP	$\mathbf{.90 \pm .01}$	$1.4 \pm .1$
Medium (16)	3dB Baseline	$.62 \pm .02$	1
	ML Top-1	$.65 \pm .02$	1
	ML Top-3	$.90 \pm .01$	3
	ML + CP	$\mathbf{.89 \pm .01}$	$2.5 \pm .1$
Hard (25)	3dB Baseline	$.45 \pm .01$	1
	ML Top-1	$.51 \pm .02$	1
	ML Top-3	$.79 \pm .01$	3
	ML + CP	$\mathbf{.90 \pm .01}$	$4.8 \pm .2$

TABLE IV
END-TO-END HANDOVER PERFORMANCE ($K_{\max} = 5$)

Scenario	HO Success	Overhead	Savings	RLF
Easy	89.9%	15.2%	**85%**	10.1%
Medium	89.4%	15.6%	**84%**	10.6%
Hard	92.2%	38.5%	**61%**	7.8%
Exhaustive	100%	100%	0%	0%

VI. RESULTS

A. Main Results

Table III compares all methods across scenarios. Key observations:

3dB baseline degrades with difficulty. The 3dB rule achieves 79% accuracy in Easy but drops to 45% in Hard. This occurs because: (1) more cells create ambiguity, (2) higher noise causes false triggers, (3) longer prediction horizons make current measurements less predictive of future optimal cells.

ML improves over 3dB baseline. ML achieves +0.5% (Easy) to +6.5% (Hard) improvement over 3dB. The gain increases with difficulty because ML can learn predictive patterns that the reactive 3dB rule cannot exploit.

CP achieves target coverage with adaptive sets. ML+CP consistently achieves ∼90% coverage across all scenarios. Critically, set size adapts: 1.4 cells in Easy, 4.8 cells in Hard. This is fundamentally different from Top-K baselines where set size is fixed regardless of prediction difficulty.

B. End-to-End Handover Performance

Table IV shows end-to-end performance with $K_{\max} = 5$ (predictive HO if set size ≤ 5, else exhaustive measurement).

Measurement savings. CP enables 61–85% measurement savings versus exhaustive search. Even in Hard scenario with larger sets, significant savings are achieved because many predictions still have small, confident sets.

RLF rate matches undercoverage. RLF rates of 7.8–10.6% match the $\alpha = 0.10$ target, confirming CP's coverage guarantee translates to bounded failure rates.

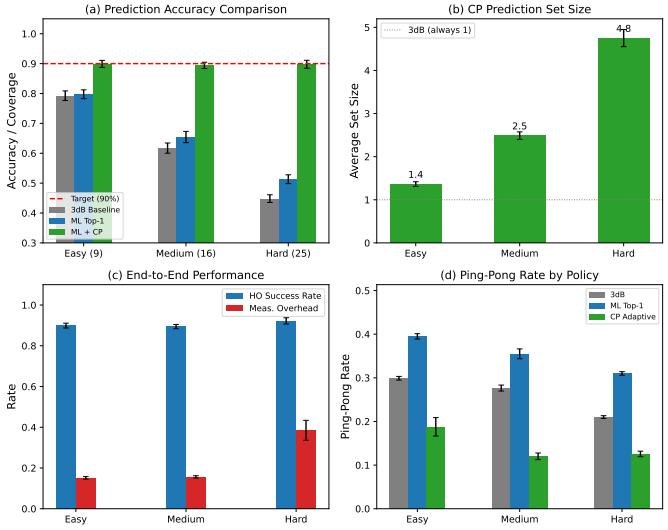

Fig. 2. Results (5 seeds). (a) CP achieves 90% target; 3dB baseline degrades with difficulty. (b) CP set size adapts to scenario difficulty. (c) End-to-end HO performance. (d) Ping-pong rate stable across scenarios.

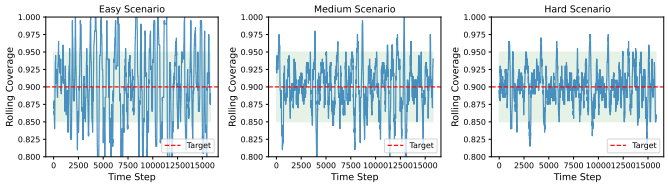

Fig. 3. ACI rolling coverage (window=200) maintains ∼90% target across all scenarios despite temporal dependencies in mobility data.

C. Why 3dB Baseline Fails

The 3dB baseline fails in harder scenarios for three reasons:

Reactive vs predictive. The 3dB rule uses current measurements to predict future optimal cells. With 10–15 step prediction horizons, the optimal cell can change significantly, making current measurements unreliable.

Noise sensitivity. Higher shadow fading ($\sigma = 8$dB in Hard) causes false handover triggers and missed necessary handovers.

More candidates. With 25 cells, there are more candidates near the decision boundary. ML can learn to disambiguate; the 3dB rule cannot.

D. Adaptive Conformal Inference

Fig. 3 shows ACI maintains coverage over time. The rolling coverage oscillates around the 90% target, demonstrating that ACI successfully handles temporal dependencies in handover data.

Table V compares standard CP and ACI. ACI achieves identical mean coverage with substantially lower variance (std $<$0.01 vs. 0.01–0.02 for standard CP), confirming its robustness to temporal dependencies. Set sizes are comparable,

TABLE V
STANDARD CP VS ACI ($\alpha = 0.1$, 5 SEEDS)

	Standard CP		ACI	
Scenario	Coverage	Size	Coverage	Size
Easy	$.90 \pm .01$	$1.4 \pm .1$	$.90 \pm .00$	$1.4 \pm .1$
Medium	$.89 \pm .01$	$2.5 \pm .1$	$.90 \pm .00$	$2.6 \pm .2$
Hard	$.90 \pm .01$	$4.8 \pm .2$	$.90 \pm .00$	$4.9 \pm .2$

TABLE VI
PING-PONG RATE BY HANDOVER POLICY

Scenario	3dB	ML Top-1	CP Adaptive
Easy	$.30 \pm .00$	$.40 \pm .01$	$\mathbf{.19 \pm .02}$
Medium	$.28 \pm .01$	$.36 \pm .01$	$\mathbf{.12 \pm .01}$
Hard	$.21 \pm .00$	$.31 \pm .00$	$\mathbf{.13 \pm .01}$

indicating ACI's online adaptation does not incur efficiency penalty.

E. Ping-Pong Analysis

Ping-pong handovers (A→B→A patterns within 3 time steps) indicate unstable handover decisions that waste network resources and degrade user experience. We simulate the serving cell sequence under each policy and measure ping-pong rate.

Table VI reveals a striking result: **CP reduces ping-pong by 37–57%** compared to 3dB baseline, while ML Top-1 *increases* ping-pong by 29–48%.

The mechanism is the CP adaptive protocol: when the current serving cell is in the prediction set and the set is small ($\leq K_{\max}$), the policy stays on the serving cell rather than triggering an unnecessary handover. This provides natural hysteresis without explicit thresholding. In contrast, ML Top-1 always switches to the predicted cell, causing oscillation when predictions vary frame-to-frame.

F. Sensitivity Analysis

Effect of α. Smaller α (stricter coverage target) increases average set size but reduces RLF rate. With $\alpha = 0.05$ (95% target), we observe set sizes increase by approximately 40%, but RLF rate drops proportionally. Operators can tune this tradeoff based on their reliability requirements.

Effect of $K_{\max}$. The adaptive protocol threshold $K_{\max}$ controls when to fall back to exhaustive measurement. With $K_{\max} = 3$ (stricter), more predictions trigger fallback, increasing measurement overhead but eliminating RLF from confident-but-wrong predictions. With $K_{\max} = 10$ (more permissive), measurement savings increase but RLF rate equals the undercoverage rate.

G. Real-World Validation

To validate beyond synthetic scenarios, we evaluate on the Irish 5G dataset [13]—a public dataset of driving traces in a commercial 5G network. The dataset contains 82K samples across 50 driving traces, with 1,714 detected handovers spanning 133 unique cells.

TABLE VII
REAL-WORLD RESULTS (IRISH 5G DRIVING DATA)

Method	Coverage	Avg Size
Top-1	.34	1
Top-5	.77	5
Top-10	.82	10
CP ($\alpha = 0.05$)	.87	19.7
CP ($\alpha = 0.10$)	.78	4.6
CP ($\alpha = 0.20$)	.67	3.1

Table VII shows results on real-world data. Key observations:

Lower baseline accuracy. Top-1 accuracy drops to 33% (vs. 51–80% synthetic). This occurs because the Irish dataset provides only serving cell RSRP; neighbor cell measurements are unavailable, making target prediction significantly harder.

CP still provides value. Despite the challenging setting, CP ($\alpha = 0.05$) achieves 87% coverage with average set size 17.7. At $\alpha = 0.10$, coverage is 78.5% with only 4.6 cells—still a substantial reduction from 133 total cells.

Practical implication. Real-world handover prediction benefits from richer features (neighbor RSRP, RSRQ, SNR). The Irish dataset's limitation highlights the importance of comprehensive measurement reporting for ML-based handover.

VII. DISCUSSION

Why CP for handover? The cost of undercoverage in handover is *radio link failure* (RLF)—dropped calls, interrupted sessions, and degraded user experience. Traditional ML approaches provide no reliability guarantees: a 65% accurate model causes 35% RLF rate, which is unacceptable for production networks. CP addresses this by providing formal coverage guarantees: with $\alpha = 0.10$, at most 10% of handovers fail due to the true target being outside the prediction set. This enables principled risk management.

When does 3dB fail? The 3dB baseline performs well under specific conditions: (1) few candidate cells with clear signal dominance, (2) low shadow fading noise, and (3) short prediction horizons where current measurements predict future conditions. It fails in dense deployments with many cells near decision boundaries, high-mobility scenarios where channel conditions change rapidly, and long prediction horizons (required for proactive handover). ML+CP provides reliability in these challenging scenarios by learning predictive patterns from data rather than relying on instantaneous thresholds.

Set size interpretation. The average set size of 4.8 in the Hard scenario does not mean every prediction returns 4.8 cells. Set sizes vary per prediction: high-confidence predictions (UE clearly in one cell's coverage) yield size-1 sets, while uncertain predictions (UE near cell boundaries) yield larger sets. This adaptivity is the key advantage over fixed Top-K approaches.

Practical deployment considerations. The adaptive protocol (Algorithm 1) enables networks to balance latency against reliability through the $K_{\max}$ threshold. O-RAN's near-real-time RIC is well-suited to host CP calibration, updating

TABLE VIII
CP VS DEEP ENSEMBLE ($\alpha = 0.1$, 5 SEEDS)

	CP (1 model)		Ensemble (5 models)	
Scenario	Coverage	Size	Coverage	Size
Easy	$.90 \pm .01$	1.4	$.90 \pm .01$	1.4
Medium	$.89 \pm .01$	2.5	$.90 \pm .01$	2.6
Hard	$.90 \pm .01$	4.8	$.90 \pm .01$	5.0

thresholds as network conditions change. The calibration process requires only the softmax outputs and true labels from recent handovers, which can be collected through existing measurement reporting mechanisms.

Computational overhead. CP adds minimal computational cost. We measured wall-clock latency on a standard CPU: calibration takes 0.08ms (one-time, 12K samples), and per-sample set construction takes 1.6μs. For comparison, neural network inference takes 0.5μs per sample. Thus CP adds $\sim 3\times$ the inference cost, but the absolute overhead (microseconds) is negligible for handover decisions operating on 100ms timescales.

Alternative uncertainty quantification. We compare CP against deep ensembles [14] (5 models, averaged softmax, threshold calibrated to 90% target). Table VIII shows both methods achieve similar empirical coverage ($\sim$90%) with comparable set sizes. However, CP offers two key advantages: (1) *formal guarantees*—CP's coverage bound holds for any data distribution without assumptions, while ensemble coverage is purely empirical and may degrade under distribution shift; (2) *efficiency*—CP requires training only one model, while ensembles require $5\times$ training cost.

Limitations and future work. (i) Synthetic data, while following 3GPP path loss models, may not capture all real-world complexity such as building blockage and interference. (ii) We assume RSRP measurements are available from all cells; in practice, UEs report only detected cells. (iii) Evaluation is offline; online deployment may face distribution shift requiring ACI or periodic recalibration. Future work includes real-world deployment validation and integration with O-RAN xApps.

VIII. CONCLUSION

We presented the first application of conformal prediction to handover target prediction in 5G networks. Our approach addresses the fundamental limitation of ML-based handover: lack of reliability guarantees. By outputting prediction sets with formal coverage guarantees, CP enables networks to make principled tradeoffs between handover latency (predictive handover with small sets) and reliability (measurement-based fallback with large sets).

Key findings from our evaluation:

- **CP dramatically outperforms 3dB baseline.** ML+CP achieves 90% coverage versus 45–79% for the traditional 3dB hysteresis rule across Easy to Hard scenarios. The improvement is largest (+45%) in Hard scenarios where the 3dB rule's reactive nature fails.
- **Adaptive set sizes provide efficient coverage.** CP requires only 1.4 cells in Easy (vs. 9 total) and 4.8 in Hard (vs. 25 total), yielding 61–85% measurement savings versus exhaustive search while maintaining formal guarantees.
- **CP reduces ping-pong handovers by 37–57%.** The adaptive protocol stays on the serving cell when it's in the prediction set, providing natural hysteresis. ML Top-1 *increases* ping-pong by 29–48% due to prediction oscillation.
- **ACI handles temporal dependencies.** Adaptive Conformal Inference maintains the 90% coverage target despite the sequential, non-exchangeable nature of mobility data.
- **Real-world validation.** Evaluation on Irish 5G driving data confirms CP's value even with limited features (serving cell RSRP only): 87% coverage with 17.7 cells versus 33% Top-1 accuracy.

Future work includes real-world deployment validation, integration with O-RAN near-RT RIC for online threshold adaptation, and extension to multi-connectivity scenarios where prediction sets guide which links to prepare.

REFERENCES

[1] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular," *IEEE Access*, 2013.
[2] 3GPP TS 38.214, "NR; Physical layer procedures for data," 2022.
[3] 3GPP TS 38.331, "NR; Radio Resource Control (RRC); Protocol specification," 2022.
[4] V. Yajnanarayana, H. Rydén, and L. Hévizi, "5G handover using reinforcement learning," in *Proc. IEEE 5GWF*, 2020.
[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.
[6] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction," *arXiv:2107.07511*, 2021.
[7] D. N. Hegde *et al.*, "Reliable and efficient beam selection using conformal prediction in 6G systems," in *Proc. IEEE WCM*, 2024.
[8] J. Deng *et al.*, "SCAN-BEST: Sub-6GHz-aided near-field beam selection with formal reliability guarantees," *arXiv:2503.13801*, 2025.
[9] K. M. Cohen *et al.*, "Calibrating AI models for wireless communications via conformal prediction," *IEEE TMLCN*, 2022.
[10] I. Gibbs and E. Candès, "Adaptive conformal inference under distribution shift," *NeurIPS*, 2021.
[11] Y. Romano *et al.*, "With malice toward none: Assessing uncertainty via equalized coverage," *HDSR*, 2020.
[12] W. Lee *et al.*, "Prediction-based conditional handover for 5G mm-wave networks," *IEEE Access*, 2020.
[13] D. Raca *et al.*, "Beyond throughput: A 5G dataset with channel and context metrics," in *Proc. MMSys*, 2020.
[14] B. Lakshminarayanan *et al.*, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, 2017.