

# Conformal Prediction for Reliable Handover Under Distribution Shift

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

**Abstract**—ML handover prediction is accurate in-distribution but fragile under shift. We evaluate conformal prediction (CP) for 5G handover under synthetic and real-world drift. We compare static CP, Adaptive Conformal Inference (ACI), weighted CP, and a confidence-gated triggered ACI against Top- $k$  and a 3dB hysteresis baseline. On in-distribution synthetic data, static CP reaches 90.3% coverage with set size 2.46. Under speed and noise shifts, weighted CP reaches 91.1% in the speed-shift case. Under severe shadow shift, static CP drops to 69.2% while ACI restores 88.8% at larger sets (5.98). In regime-switch streams, ACI stabilizes rolling coverage near the 90% target. ACI step-size sweep shows  $\gamma = 0.002$ – $0.005$  gives the best reliability-overhead tradeoff in our setup. On Irish 5G driving traces with speed-split drift, static CP reaches 76.2% coverage (size 4.59), weighted CP 73.8% (size 4.19), and ACI 88.0% (size 15.1). Results show reliability under shift needs adaptive conformal control, not static calibration alone.

## I. INTRODUCTION

Predictive handover reduces latency but can fail badly when radio conditions shift. Traditional 3GPP handover logic relies on hysteresis events [2], [3]. ML-based handover prediction improves point accuracy but provides no risk control when distribution changes [4], [12]. In production, that gap maps directly to radio link failure risk.

Conformal prediction (CP) gives finite-sample marginal coverage guarantees [5], [6]. Recent wireless CP work focuses on demodulation, channel tasks, and beam selection [7]–[9], while handover under distribution shift is still underexplored. This paper targets that gap.

### Contributions.

- 1) We benchmark handover reliability under four synthetic shifts plus a regime-switch stream.
- 2) We compare static CP, ACI [10], weighted CP, and confidence-gated triggered ACI under the same base predictor and KPI mapping.
- 3) We validate on Irish 5G driving traces with source-target speed split.
- 4) We provide budget-aware reproducible runs (local-first, capped overflow policy) and release all generated artifacts.

## II. RELATED WORK

ML handover methods span supervised and reinforcement-learning policies [4], [12]. CP in wireless has shown value for calibration and reliability [9]. CP for beam selection shows strong reliability-efficiency tradeoffs [7], [8]. The missing piece is handover under shift: static calibration can fail as mobility, shadowing, and measurement noise drift over time.

## III. SYSTEM AND METHODS

### A. Handover Prediction Setup

At time  $t$ , the model predicts future best cell  $y_t = \arg \max_k \text{RSRP}_k(t + H)$  using input

$$\mathbf{x}_t = [\text{RSRP}_1, \dots, \text{RSRP}_K, \mathbf{e}_{c_t}, v_t]. \quad (1)$$

We use an MLP classifier and softmax scores  $\hat{p}(y | \mathbf{x})$ .

### B. Baselines and Conformal Variants

**3dB baseline:** handover if best neighbor exceeds serving by 3dB.

**Static CP:**

$$\mathcal{C}(\mathbf{x}) = \{y : \hat{p}(y | \mathbf{x}) \geq 1 - \hat{q}\}, \quad (2)$$

with  $\hat{q}$  calibrated on held-out source calibration data.

**ACI:** online update of effective misscoverage level to track sequential drift [10].

**Weighted CP:** source calibration scores reweighted by estimated density ratio  $w(\mathbf{x}) \propto p_T(\mathbf{x})/p_S(\mathbf{x})$  using a source-vs-target logistic discriminator.

**Triggered ACI:** confidence-gated mixture that uses static CP on high-confidence samples and ACI sets on low-confidence samples (threshold from source calibration confidence quantile).

### C. System KPI Mapping

Coverage maps to handover success with bounded miss risk. Set size maps to measurement overhead. Undercoverage maps to RLF proxy. Serving-cell retention in small sets acts as implicit hysteresis and affects ping-pong rate.

## IV. EXPERIMENTAL SETUP

### A. Synthetic Shift Benchmark

Source setting: medium scenario (4×4 cells,  $\sigma = 6$  dB shadowing, measurement noise 4 dB, speed 1–30 m/s, horizon  $H = 10$ ). We train on source and calibrate on source only.

Target shifts:

- 1) IID (same as source)
- 2) Speed shift (20–50 m/s)
- 3) Measurement-noise shift (8 dB)
- 4) Shadow shift ( $\sigma = 10$  dB)
- 5) Regime switch (source-like first half, harsh second half)

Each result is mean±std across 5 seeds (42,123,456,789,1011), 600 trajectories/seed, 20 epochs.

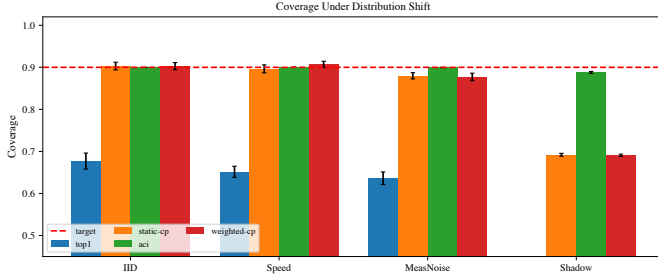


Fig. 1. Coverage across synthetic shifts. Static CP degrades under hard shift, ACI is most robust, weighted CP helps in moderate covariate shift.

TABLE I  
SYNTHETIC SHIFT COVERAGE (MEAN  $\pm$  STD)

Shift	3dB	Top-1	Top-3	Static CP	ACI
IID	.64 $\pm$ .02	.68 $\pm$ .02	.91 $\pm$ .01	.90 $\pm$ .01	.90 $\pm$ .00
Speed	.62 $\pm$ .01	.65 $\pm$ .02	.90 $\pm$ .01	.90 $\pm$ .01	.90 $\pm$ .00
MeasNoise	.55 $\pm$ .02	.64 $\pm$ .02	.89 $\pm$ .01	.88 $\pm$ .01	.90 $\pm$ .00
Shadow	.40 $\pm$ .01	.43 $\pm$ .01	.72 $\pm$ .01	.69 $\pm$ .01	.89 $\pm$ .00
Regime	.49 $\pm$ .02	.53 $\pm$ .02	.79 $\pm$ .01	.78 $\pm$ .02	.89 $\pm$ .00

## B. Real-World Drift Benchmark

Dataset: Irish 5G driving traces [13]. We split traces by average speed: lower-speed traces as source, higher-speed traces as target. Model trains and calibrates on source only, then evaluates on target.

## V. RESULTS

### A. Coverage Under Shift

Table I shows the main trend: static CP is reliable near source but degrades under strong shift (shadow, regime). ACI keeps coverage close to target by expanding sets online.

### B. Tradeoff in Hard Shifts

Paired seed deltas confirm hard-shift reliability gains: ACI vs static is +19.6pp coverage on shadow shift (95% CI [19.3, 19.8]) and +11.2pp on regime switch (95% CI [10.4, 12.2]). Triggered ACI keeps most of that gain (+15.1pp shadow, +8.8pp regime) while reducing overhead versus full ACI by 5.87pp and 2.91pp, respectively.

### C. Regime-Switch Stability

In regime-switch streams, static and weighted thresholds lag after the phase boundary. ACI adapts online and recovers target-level coverage.

### D. ACI Step-Size Sensitivity

Figure 3 quantifies ACI sensitivity. In our regime-switch benchmark, small step sizes ( $\gamma = 0.002$ – $0.005$ ) achieve the highest coverage ( $\approx 89.6\%$ ) with moderate set inflation, while larger values ( $\gamma = 0.05$ ) reduce coverage to 87.3% but lower overhead. This supports tuning  $\gamma$  as a direct reliability-overhead control knob.

TABLE II  
HARD-SHIFT KPI TRADEOFF (MEAN OVER 5 SEEDS)

Shift	Method	Coverage	Set Size	RLF Proxy	Overhead
Shadow	Static CP	.692	2.62	.306	.170
	ACI	<b>.888</b>	5.98	<b>.073</b>	.600
	Triggered ACI	.843	5.40	.121	.542
	Weighted CP	.691	2.60	.307	.168
Regime	Static CP	.780	2.60	.218	.167
	ACI	<b>.892</b>	4.48	<b>.084</b>	.413
	Triggered ACI	.868	4.20	.109	.384
	Weighted CP	.787	2.69	.210	.174



Fig. 2. Rolling coverage in regime-switch stream (window=200). ACI tracks the 90% target more closely than static and weighted CP.

### E. Real-World Drift Results

Trace-bootstrap on Irish confirms the reliability-cost pattern: ACI vs static gives +11.8pp coverage (95% CI [5.4, 20.0]) with +10.53 set size (95% CI [3.98, 19.61]), while triggered ACI gives +5.8pp coverage (95% CI [2.1, 10.7]) with +4.89 set size (95% CI [1.18, 10.32]). Triggered ACI versus full ACI loses 6.1pp coverage (95% CI [−11.5, −2.0]) but cuts set size by 5.65 (95% CI [−12.13, −0.99]). A speed-bin breakdown shows high-speed traces are hardest (static 68.9%, weighted 67.0%) while ACI remains highest at 83.8% and triggered ACI reaches 77.5%.

## VI. DISCUSSION

**When to use which method.** Static CP is a strong default in stable conditions. Weighted CP helps moderate covariate shift when target feature support overlaps source. ACI is the robust choice for severe sequential drift, with  $\gamma$  tuning used to set reliability-overhead preference. Triggered ACI is a middle point when overhead budget is tight.

**System implications.** Reliability gains translate to lower RLF proxy but require explicit overhead budgeting. Table IV gives a simple control policy: low-budget mode prefers static or weighted CP, medium-budget mode uses triggered ACI, and high-budget mode uses full ACI.

**Limitations.** Synthetic channels still simplify real deployments. Irish traces have limited feature richness versus full network measurement reports. We evaluate offline; online deployment requires streaming integration and control-plane constraints.

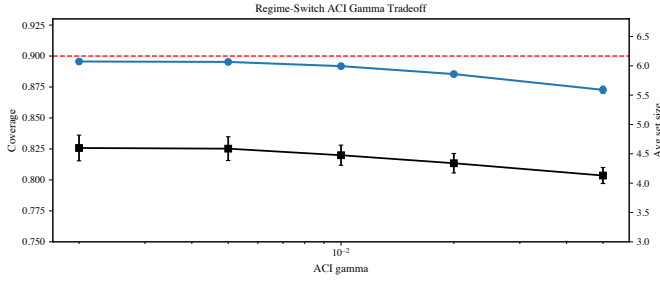


Fig. 3. Regime-switch ACI tradeoff over  $\gamma$ . Smaller  $\gamma$  improves coverage but increases set size and overhead.

TABLE III  
IRISH 5G SPEED-SPLIT DRIFT RESULTS

Method	Coverage	Avg Size
Top-1	.309	1.00
Top-3	.626	3.00
Static CP	.762	4.59
ACI	<b>.880</b>	15.13
Triggered ACI	.819	9.48
Weighted CP	.738	4.19

## VII. CONCLUSION

We presented a shift-focused handover reliability study with conformal prediction. Static CP works well in-distribution but degrades in severe shift. ACI restores near-target coverage under shadow and regime-switch drift. Weighted CP improves moderate shifts with smaller set inflation than ACI. A confidence-gated triggered ACI recovers most hard-shift reliability gains with lower overhead than full ACI. On Irish real-world speed-split drift, ACI achieves the highest reliability. The core practical result is clear: robust handover reliability needs adaptive conformal control, not static calibration alone.

## APPENDIX A

### APPENDIX: ENSEMBLE AND LATENCY

Measured medium-scenario latency: calibration 0.08 ms, CP set construction 1.59  $\mu$ s/sample, NN inference 0.50  $\mu$ s/sample.

## 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. Hévízi, “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.

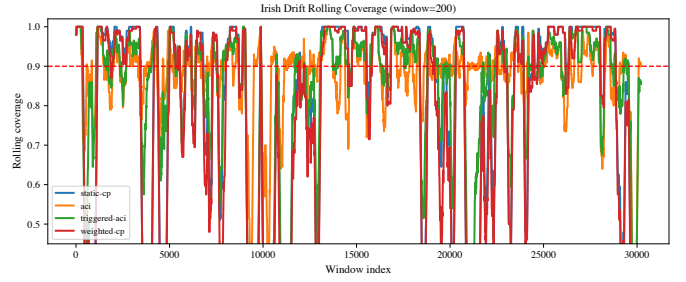


Fig. 4. Irish drift rolling coverage (window=200). ACI is most stable under source-target speed split.

TABLE IV  
BEST METHOD UNDER OVERHEAD CAPS (SYNTHETIC HARD SHIFTS)

Overhead Cap	Shadow Shift	Regime Switch
$\leq .20$	Static CP (.692)	Weighted CP (.787)
$\leq .40$	Static CP (.692)	Triggered ACI (.868)
$\leq .60$	ACI (.888)	ACI (.892)

- [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.

TABLE A1  
APPENDIX: CP VS ENSEMBLE (FROM V5 RUNS, 5 SEEDS)

Scenario	CP (1 model)		Ensemble (5 models)	
	Coverage	Size	Coverage	Size
Easy	.900	1.37	.900	1.38
Medium	.893	2.47	.898	2.57
Hard	.898	4.80	.903	5.05