Position: There Is No Ground Truth – Rethinking Evaluation in AI-Driven Channel Prediction

Anonymous Author(s)

Abstract

Machine learning (ML) has rapidly gained traction for wireless channel state information (CSI) prediction, promising improved reliability and reduced overhead for 5G/6G systems. From autoencoder-based CSI compression [1] to large language model-based adaptations [2] today, a plethora of techniques report impressive accuracy in forecasting channel dynamics. **However, this work argues that many of these results are built on flawed evaluation practices**. In particular, current works often assume an idealized "ground truth" provided by synthetic channel models, and thereby overlook key issues: (1) *training–test leakage* when the same generative simulator underpins both training and evaluation; (2) reliance on *synthetic datasets without field validation*; and (3) conflating *memorization with true generalization*. The consequences are inflated performance metrics that may not transfer to operational networks. As a result, there is growing concern that most current works are "overfitting" to the simulation sandboxes – optimizing for a non-existent ground truth rather than solving the real channel prediction problem. We also chart a three-pronged constructive path with concrete guidelines for *benchmark design*, *dataset standards*, and *evaluation protocols*.

1 Introduction and Background

Recent studies have begun to acknowledge the above mentioned issues. For example, [3] notes that models trained on popular channel simulators (QuaDRiGa [4], Sionna [5], MATLAB 3GPP [6]) incur a **7–10%** accuracy drop under even modest deployment mismatches (e.g. a 15° base-station downtilt shift). [7] cautions that digital-twin simulations, while useful, provide only coarse approximations of reality that must be carefully calibrated. Despite these warnings, the majority of ML-for-CSI works still evaluate **solely on synthetic data drawn from the same distribution used for training**, with minimal checks against overfitting or lack of generalization. **In this position paper, these methodological pitfalls are critically examined and a path forward is proposed**. We first dissect how current evaluation practices can be misleading – highlighting leakage issues, unrealistic datasets, and evidence that some reported "gains" may in fact reflect memorization and then chart a constructive path that includes concrete guidelines for *benchmark design*, *dataset standards*, and *evaluation protocols*.

In particular, community-driven benchmarks (including field-collected data and hidden test scenarios) are proposed to **eliminate train–test leakage**, new metrics and stress-tests for **robustness under distribution shifts**, and a roadmap toward **open, field testbeds** for AI-driven channel prediction. Rethinking evaluation now can ensure that the next generation of models improve towards deployment – and not just over-tuned to a convenient fiction of a truly non-existent "ground truth".

1.1 Limitations in current ML-for-CSI evaluation practices:

1) Shared Generative Models and Train–Test Leakage: A prevalent issue is the reuse of the *same channel generative model* for both training and testing ML predictors. Most studies generate CSI data from a single simulator (e.g. a 3GPP geometry-based model or ray-tracer) and then randomly split into train/test sets. While this avoids sample overlap, it **does not guarantee distributional**

Submitted to AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG) Workshop at NeurIPS 2025.

independence – the neural network can implicitly learn this simulator's idiosyncrasies. If the channel simulator's assumptions (urban macro cell with fixed antenna downtilt, specific propagation parameters, etc.) hold during testing, even a memorizing model will appear to perform well. *For instance, a network may overfit to the fixed AoA/AoD statistics of the simulated environment rather than learning a truly general predictive skill.* [3] explicitly demonstrates this effect: when a CNN-LSTM predictor trained on a nominal simulation was evaluated on a *slightly perturbed* version of the same simulator (15° antenna downtilt change), **NMSE degraded by up to 10%**. This gap reveals that some reported accuracies are possibly overestimates, boosted by a tacit train–test overlap in environment assumptions. This phenomenon is referred to as *"digital twin leakage"*, since the digital twin (simulator) essentially "leaks" into both training and evaluation.

2) Synthetic Data without Real-World Validation: Another concern is the heavy reliance on completely synthetic datasets *without any verification on empirical data*. Various "state-of-the-art" AI-based CSI prediction baselines ([1, 8–13]) train and test on (typically the same) simulated channel model (COST 2100, TDL-C models, etc.), reporting tremendous gains (say -20 dB NMSE) without attempting to verify if these gains persist on field measurements. Since real CSI time-series data for ML is challenging to collect at scale – the community has gravitated toward ever more sophisticated simulators (QuaDRiGa [4], Sionna [5], DeepRay [14]) as proxies for "ground truth". Yet even minor deployment perturbations (e.g., antenna-tilt/layout changes) can degrade these headline numbers by 7–10%, underscoring poor out-of-distribution generalization. Notable exceptions include: CSILaBS [15] evaluates on a Nokia Bell Labs field dataset, and PEACH [16] validates with extensive indoor measurements (up to 6 dB NMSE improvement over pilot baselines). These efforts are still rare – most "state-of-the-art" baselines include no real-world testing. Absent at least some empirical or domain-shift testing, reported gains are likely over-optimistic.

3) Memorization vs. True Generalization: A high-capacity neural network can achieve near-perfect accuracy on simulated data by overfitting to a generator's random seed or structural biases—effectively memorizing recurring spatial or temporal patterns. Such models often act as interpolators of the simulator itself, yielding deceptively low test error when evaluated on data from the same generator. In reality, the model may not have learned any fundamental representation of channel dynamics – it has merely indexed the simulator's outcomes. True generalization requires robust performance under *unseen* environments or perturbed conditions. In contrast, the framework in [3] explicitly trains for domain-generalization by mining diverse "hard negative" channel examples across LOS/NLOS, mobility levels, and antenna configs. By using contrastive pre-training on a family of simulated domains, representations that are **robust to scenario variations** are learned, and a 12.5% throughput gain is reported alongside significantly higher multi-step prediction accuracy versus baselines in mixed scenarios.

Few works test generalization under held-out environments or perturbed conditions, making reported results potentially misleading—especially as model sizes scale into the millions of parameters. Their capacity to memorize simulator-specific artifacts grows accordingly. Recent literature acknowledges this concern, noting that *"most CSI prediction models face poor generalization under deployment shifts, primarily due to dependence on synthetic data and overparameterized architectures."* In summary, current evaluation practices often blur the line between learning and overfitting. Without new methodologies to expose memorization, we risk deploying models that fail in the field despite excellent reported metrics.

Coda: The net effect of the above limitations is that reported CSI-ML results likely overestimate real-world performance, as evaluation is often confined to the same synthetic domain used for training. Without empirical or shift-aware testing, we cannot assess whether models generalize or merely exploit simulator structure. In practice, simpler model-based baselines may outperform deep networks when deployed, but current benchmarks fail to reveal this gap. This calls for an urgent reevaluation: the community must stop treating simulation outputs as ground truth, and instead design tests that approximate the uncertainty and variability of real channels. The next section outlines how to achieve this through better benchmarks, standards, and practices.

Table 1 in the Appendix surveys recent CSI prediction studies and highlights how widespread evaluation pitfalls remain—particularly single-domain testing and lack of field validation. Many papers reporting strong gains exhibit signs of train/test overlap or simulator overfitting. Our aim is not to critique individual works, but to drive home the need for systemic change in how we assess AI for channel prediction. Without stronger evaluation protocols, we risk favoring models that perform well only within the closed world of a particular simulator.

2 Toward Robust and Transferable Evaluation in CSI Prediction

It is clear that **incremental fixes are not enough** –evaluation methodology must be rethought from the ground up. This section outlines concrete steps and design principles to ensure that future AI-driven channel prediction research produces meaningful, transferable results. Our recommendations encompass new benchmarking protocols, standards for datasets, and guidelines for rigorous evaluation. Underpinning all these proposals is a simple ethos: *treat the field deployment as the target, and design all evaluations to approximate that reality as closely as possible*. We propose three actionable components: benchmark protocols, dataset standards, and robustness evaluation, all aimed at aligning simulation with deployment.

A. Benchmarking Protocols to Eliminate Leakage

Independent Training and Test Scenario Design: To prevent leakage via shared channel generators, benchmarks must ensure training and test data come from *distinct* scenarios. For instance, a model trained on a 3GPP Urban Macro cell should be evaluated on a perturbed setting—e.g., different antenna configs, carrier frequencies, or layouts. Models that memorize scenario-specific patterns will fail under such shifts. Open challenges can enforce this separation by withholding test scenario details, ensuring models generalize rather than overfit. Benchmarks should withhold test scenarios (e.g., via secret seeds in QuaDRiGa/Sionna) to prevent tuning to test conditions—mirroring ImageNet's [17] protocol to ensure test channels are from a different "distribution family" than training. Over time, scenario diversity (e.g., mmWave, factory, vehicular) can promote generalist models (instead of niche specialists). Enforcing scenario separation in benchmarks will incentivize architectures that truly learn underlying propagation features (e.g. mobility-induced temporal correlation) that transfer, instead of learning the quirks of one environment.

Standardized Data Splits and Cross-Validation: To measure generalization, benchmarks should adopt k-fold cross-validation over distinct environments. For example, a ray-tracing dataset of 10 city maps can be split into 5 folds, training on 4 and testing on the held-out one—rotated across runs. This would catch scenario-specific overfitting (e.g., to a particular antenna tilt). Open leaderboards on such standardized splits, akin to the GLUE benchmark [18], can track progress and discourage selective reporting. To facilitate this, academic and industry groups should cooperate to **release benchmark datasets and split definitions** (we discuss dataset creation next). In summary, robust benchmarking protocols – featuring scenario alternation, hidden tests, and cross-validation – are our first pillar for leakage-free evaluation.

B. Dataset Standards: Synthetic vs. Empirical Data

Releasing Diverse Public Datasets: To move beyond simulator-overfitting, the community needs open *hybrid datasets* combining both synthetic and measured CSI. Synthetic training sets should be generated via open-source tools like QuaDRiGa or Sionna with diverse parameters (e.g., frequency, mobility, LOS/NLOS). Empirical test sets—such as CSI traces collected from testbeds or field trials—can validate real-world performance. One example would pair (a) ray-traced CSI from five cities for training with (b) real indoor/outdoor traces from labs or vendors as held-out test. This approach, already demonstrated in [15], and pursued by initiatives like RISE-6G, would shift benchmarks toward deployment relevance. By federating these efforts into common datasets, the field can shift from *simulator-only benchmarks* to **hybrid benchmarks that include field channels**.

Until such datasets are widespread, synthetic results must be clearly contextualized. Published works should report simulator name, channel model and parameters used to generate data (e.g., "QuaDRiGa 2.6, UMa scenario, etc."), and whether the same generator was used for both train and test. Whenever possible, results should also be reported on a second (different) scenario to test OOD generalization. Metrics like *relative performance drop* (e.g., $-25\,\mathrm{dB}$ in-sim $\rightarrow$ $-15\,\mathrm{dB}$ in-field) can quantify robustness. Real-world evaluation should be highlighted, and statistical significance considered. Ultimately, synthetic metrics should be seen as *preliminary*, while real or shift-tested results must become the standard.

C. Robustness Testing and Field Transferability

A pathway towards stress-testing under distribution shifts is outlined in Appendix B.

Field Data Benchmarking "Bot-Bird" Experiments: There is no substitute for testing ideas in the real world. The community needs **field benchmarking experiments** that go beyond simulation entirely. One promising concept is a *"bot-bird" UAV experiment suite*: a mobile robot or drone (the "bot") equipped with a channel sounder or transceiver (the UE), and optionally a second drone or fixed base station (the "bird") as the transmitter. As the bot moves, it collects CSI indexed by location, enabling spatially rich datasets. Inspired by recent UAV-based channel studies, such systems can generate repeatable trajectories in indoor or outdoor settings, producing *living benchmarks*. ML models can then be evaluated by predicting channel evolution from partial observations (e.g., pilots), with the measured CSI serving as the ground truth for comparison. Periodic open challenges could evaluate models on held-out drone traces, ensuring participants cannot tune to test conditions and that real-world generalization is measured directly. Just as the vision community has "in-the-wild" evaluation for robust models, wireless AI can have on-site evaluation as the ultimate standard

Separation of Generative Model from Evaluation. In traditional wireless research, those who build channel models are distinct from those who design algorithms—a separation that helps prevent overfitting to known structure. We advocate the same in ML-driven evaluation: the process that generates test data should be *independent* from model development. This can be enforced via standardized simulation libraries, third-party evaluation scripts, or blind testing protocols, where trained models are submitted and evaluated on held-out datasets. This mirrors best practices in ML competitions and strengthens reproducibility. For emerging setups like digital twins, it is critical to avoid training and testing on the same environment instance. Instead, models should adapt to unseen twin variants to mimic deployment calibration, which [7] already does implicitly. Similarly, if learned channel models (e.g., GANs) are used in training, test data should come from a distinct source to prevent subtle leakage. In short, structurally separating generation and evaluation is essential for measuring true generalization, not just simulator memorization. Based on the above, we outline a *roadmap* for the community to establish lasting best practices for evaluation of AI-based channel prediction in **Appendix B**.

3 Conclusion

Current AI-driven CSI prediction research is at an inflection point. Exciting breakthroughs are tempered by the realization that *"there is no ground truth" in the absolute sense – only a succession of models and measurements that approximate an ever-changing reality*. We have highlighted how overly idealized evaluations (same-simulator testing, no real-world checks, etc.) can mislead us into overestimating model performance. The encouraging news is that the community is well-equipped to improve: by adopting leakage-free benchmarks, embracing heterogeneous datasets, and stress-testing generalization, we can ensure that progress in the literature translates to progress in the field. We have outlined concrete steps and a vision for making evaluation a first-class citizen in ML-for-wireless research, rather than an afterthought. These changes will not only expose ideas that work best but also drive the development of more resilient models. In essence, a more realistic evaluation regime will **shift the optimization target** – from doing well on a single synthetic metric to doing well *across a spectrum of real-world conditions*. This is the shift needed to move from academic demos to deployed AI in next-generation networks.

In closing, we issue a call to action to the community: **let us redefine "ground truth" to mean the truth on the ground, not just in silico**. By pooling efforts to create open benchmarks and testbeds, by scrutinizing each other's results under varied conditions, and by championing evaluation in the publication process, we can build a foundation of rigorous evidence. This foundation will support credible scientific conclusions and accelerate the adoption of ML for wireless systems. The ultimate reward will be AI models that genuinely earn their accolades – ones that operators find deliver reliable gains not just in papers but in real deployments. Only by acknowledging and overcoming the current evaluation pitfalls can we unlock the full potential of AI-driven channel prediction in 6G and beyond. The time to act is now: the future "AI-native" wireless network will be built on algorithms we validate today. *Let's make sure we validate them right!*

References

[1] Chao-Kai Wen, Wan-Ting Shih, and Shi Jin. Deep learning for massive mimo csi feedback, 2018. URL `https://arxiv.org/abs/1712.08919`.

[2] Shilong Fan, Zhenyu Liu, Xinyu Gu, and Haozhen Li. Csi-llm: A novel downlink channel prediction method aligned with llm pre-training, 2024. URL `https://arxiv.org/abs/2409.00005`.

[3] Sagnik Bhattacharya, Abhiram Rao Gorle, and John M. Cioffi. Contwin: Contrastive learning for robust digital twin csi prediction. *Preprint*, 2025. Stanford University, Department of Electrical Engineering.

[4] Stephan Jaeckel, Leszek Raschkowski, Kai Börner, and Lars Thiele. QuaDRiGa: A 3-D Multi-Cell Channel Model with Time Evolution for Enabling Virtual Field Trials. *IEEE Transactions on Antennas and Propagation*, 62(6):3242–3256, 2014. doi: 10.1109/TAP.2014.2310220.

[5] Jakob Hoydis and et. al. Sionna RT: Differentiable Ray Tracing for Radio Propagation Modeling. *arXiv preprint arXiv:2303.11103*, 2023. URL `https://arxiv.org/abs/2303.11103`.

[6] 3rd Generation Partnership Project (3GPP). Study on channel model for frequencies from 0.5 to 100 GHz (Release 16). Technical Report TR 38.901, ETSI, 2020. URL `https://www.3gpp.org/DynaReport/38901.htm`. Version 16.1.0.

[7] Sadjad Alikhani and Ahmed Alkhateeb. Digital twin aided channel estimation: Zone-specific subspace prediction and calibration, 2025. URL `https://arxiv.org/abs/2501.02758`.

[8] Sijie Ji and Mo Li. Clnet: Complex input lightweight neural network designed for massive mimo csi feedback. *IEEE Wireless Communications Letters*, 10(10):2318–2322, October 2021. ISSN 2162-2345. doi: 10.1109/lwc.2021.3100493. URL `http://dx.doi.org/10.1109/LWC.2021.3100493`.

[9] Jiaming Cheng, Wei Chen, Jialong Xu, Yiran Guo, Lun Li, and Bo Ai. Swin transformer-based csi feedback for massive mimo, 2024. URL `https://arxiv.org/abs/2401.06435`.

[10] Boyuan Zhang, Haozhen Li, Xin Liang, Xinyu Gu, and Lin Zhang. Multi-task deep neural networks for massive mimo csi feedback, 2022. URL `https://arxiv.org/abs/2204.12442`.

[11] Zheng Cao, Wan-Ting Shih, Jiajia Guo, Chao-Kai Wen, and Shi Jin. Lightweight convolutional neural networks for csi feedback in massive mimo, 2020. URL `https://arxiv.org/abs/2005.00438`.

[12] Shunpu Tang, Junjuan Xia, Lisheng Fan, Xianfu Lei, Wei Xu, and Arumugam Nallanathan. Dilated convolution based csi feedback compression for massive mimo systems, 2021. URL `https://arxiv.org/abs/2106.04043`.

[13] Tianqi Wang, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li. Deep learning-based csi feedback approach for time-varying massive mimo channels, 2018. URL `https://arxiv.org/abs/1807.11673`.

[14] Stefanos Bakirtzis, Kehai Qiu, Jie Zhang, and Ian Wassell. Deepray: Deep learning meets ray-tracing. In *2022 16th European Conference on Antennas and Propagation (EuCAP)*, pages 1–5, 2022. doi: 10.23919/EuCAP53622.2022.9769203.

[15] M. Karam Shehzad, Luca Rose, and Mohamad Assaad. Massive mimo csi feedback using channel prediction: How to avoid machine learning at ue? *Trans. Wireless. Comm.*, 23: 10850–10863, September 2024. ISSN 1536-1276. doi: 10.1109/TWC.2024.3376633.

[16] Serkut Ayvasik, Fidan Mehmeti, Edwin Babaians, and Wolfgang Kellerer. Peach: Proactive and environment-aware channel state information prediction with depth images. *Proc. ACM Meas. Anal. Comput. Syst.*, 7(1), March 2023. doi: 10.1145/3579450. URL `https://doi.org/10.1145/3579450`.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

[18] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

[19] Zhilin Lu, Jintao Wang, and Jian Song. Multi-resolution csi feedback with deep learning in massive mimo system, 2021. URL https://arxiv.org/abs/1910.14322.

[20] Yaodong Cui, Aihuang Guo, and Chunlin Song. Transnet: Full attention network for csi feedback in fdd massive mimo system. *IEEE Wireless Communications Letters*, 11(5):903–907, 2022. doi: 10.1109/LWC.2022.3149416.

[21] Zhilin Lu, Jintao Wang, and Jian Song. Binary neural network aided csi feedback in massive mimo system, 2020. URL https://arxiv.org/abs/2011.02692.

[22] Ziang Liu, Tianyu Song, Ruohan Zhao, Jiyu Jin, and Guiyue Jin. An efficient parallel self-attention transformer for csi feedback. *Phys. Commun.*, 66(C), October 2024. ISSN 1874-4907. doi: 10.1016/j.phycom.2024.102483. URL https://doi.org/10.1016/j.phycom.2024.102483.

[23] Juseong Park, Foad Sohrabi, Jinfeng Du, and Jeffrey G. Andrews. Self-nomination: Deep learning for decentralized csi feedback reduction in mu-mimo systems, 2025. URL https://arxiv.org/abs/2504.16351.

Appendix

Appendix A: Table 1

Baseline	Data Source	Train/Test Overlap?	Real-World Validation?	Notable Observations
[1] CsiNet	Synthetic (COST2100)	Yes: same channel model	No	Introduced autoencoder for CSI compression; achieved large NMSE gains on one synthetic indoor dataset, but generalization not verified.
[15] CSILaBS	Synthetic + Real	Partial (sim for training, real for eval)	**Yes**: tested on field data	ML at BS yields 11–43% NMSE improvement in simulation and retains precoding gains on a Nokia Bell Labs OTA dataset, showing feasibility of real-world use.
[7] Dig. Twin	Synthetic (Ray-tracing DT)	Yes: twin model for both	No	Uses a ray-traced digital twin as prior for channel estimation; near-optimal sim performance achieved after RL subspace calibration. Assumes twin $\approx$ reality, but tested only in sim.
[3] ConTwin	Synthetic (QuaDRiGa + variations)	No: multiple scenario domains used	No (varied sim only)	Contrastive learning across LOS/NLOS and mobility domains; improved 10-step CSI prediction NMSE by 24.3% and beam selection by 17.8%. Demonstrated $\sim$7–10% performance drop with minor scenario shift, highlighting generalization gap.
[16] PEACH	**Real** measurements	N/A (field data only)	**Yes**: fully experimental	Predicts CSI via depth camera environment data; experimentally achieves comparable error to pilot-based CSI in a lab, and up to 6 dB NMSE gain under interference. Exemplifies leveraging field measurements for ML.

Table 1: Evaluation setups in representative CSI prediction studies.

More recent works using synthetic evaluation solely are summarized below in Table 2

Baseline (Year)	Dataset / Channel Model
CsiNet (2018) [1]	COST 2100 indoor/outdoor
CRNet [19] (2019)	COST 2100
TransNet [20] (2022)	COST 2100
BCsiNet/Binary-Net [21] (2021)	COST 2100
SwinCFNet [9] (2024)	3GPP-style synthetic channels
EPAformer [22] (2024)	3GPP CDL-A/B, no OTA data
Self-Nomination Beam Prediction [23](2025)	3GPP TR 38.901 (QuaDRiGa)

Table 2: Recent CSI feedback works evaluated solely on synthetic datasets.

Appendix B: Additional Notes on Testbed Evaluation and ML-Generated Channel Models

1. Stress-Testing Under Distribution Shifts: Taking inspiration from robust machine learning, we propose that CSI prediction models be subjected to a battery of **stress-tests that simulate plausible real-world distribution shifts**. These could include, for example:

- *Temporal shift* – test the model on channel sequences that are $2\times$ longer future horizon than it was trained on (does performance gracefully degrade or catastrophically fail?);
- *Spatial shift* – Trained on one base-station location, test on a new location with different blockage layout;
- *Hardware impairment* – add realistic noise or quantization error to the input CSI (does the model cope with imperfect data?);

Many of these shifts can be emulated in simulation by varying parameters. Crucially, models should be evaluated on these *without any further fine-tuning*, to assess innate robustness. A concrete example: for a predictor trained on a pedestrian-mobility dataset (0–5 km/h), test it on vehicular speeds (50–100 km/h) and report the drop in accuracy. Or take a model trained on a certain SNR range and test it when SNR is 10 dB lower. We recommend summarizing such tests perhaps in a **“robustness radar” plot or table** that highlights where a model’s performance starts to break down. This way, even if a method shows excellent baseline results, the community can recognize limitations (e.g. “Model A works well up to moderate mobility, but fails for high Doppler – whereas Model B is more consistent”). Over time, this encourages designing models with built-in generalization techniques [3] to pass these tough tests. We note that other fields (like computer vision) have begun similar evaluations (e.g. ImageNet-C for common corruptions); wireless AI should do the same by developing CSI-C (CSI under Corruptions) and CSI-D (CSI under Distribution shifts) benchmarks. Ultimately, a model that succeeds across varied stress conditions will earn trust for deployment.

2. On Testbed-Based Evaluation: While organizing large-scale field tests is non-trivial, a viable starting point is a shared data collection effort. For example, one lab could host a UAV-based channel measurement experiment and distribute the resulting CSI traces to others for evaluation. Alternatively, multiple groups could contribute measurement sets under varied conditions (urban, suburban, rural), building a collaborative benchmark suite.

3. On AI-Generated Channel Models: If ML-generated or learned models (e.g., GANs, VAEs) are used to synthesize channel data, care must be taken to avoid subtle leakage. If a model is trained on data generated by a known neural network, it may partially invert that generator rather than learn true channel structure. To avoid this, any ML-based channel model used in training should be disclosed, and a different (ideally unknown) generator should be used to produce test data. Such separation protects against "reverse engineering" the simulator and helps ensure meaningful generalization is tested.

4. Roadmap:

Based on the above, we outline a *roadmap* for the community to establish lasting best practices in evaluating AI for channel prediction:

- **Near-Term (Next 1–2 Years)**: Form an *evaluation working group* under workshops like AI4NextG or IEEE ComSoc. This group can define a preliminary *benchmark suite*: e.g., release a multi-scenario simulation dataset (with defined train/test splits to prevent leakage) and a small curated real-world test dataset. The group would also publish an *evaluation checklist* for authors (covering points like declaring data origins, performing one form of shift test, etc.). Workshops and special sessions can encourage submissions to follow these guidelines, perhaps even featuring a *leaderboard track* where papers are ranked on a common test set. During this phase, research could report, for example, “Model A achieves X% NMSE in-distribution, and Y% on the held-out scenario, with a Z% drop” – a level of detail largely missing today.
- **Mid-Term (3–5 Years)**: Develop *community-driven open testbeds*. This could involve expanding existing wireless test facilities (such as POWDER, COSMOS, or foreign equivalents) with standardized CSI collection and making those accessible. This is analogous to “Open CSI Prediction Testbed” where researchers can upload code or models, which then run on a real-time channel data stream from the testbed (could be a fixed set of replayed real traces or live channels). In parallel, work with standards bodies (3GPP, ITU) to incorporate ML evaluation considerations – e.g., a 3GPP study item could specify that any ML-based channel predictor for Release-20 must be evaluated on a common reference model plus at least one independent model. This would formalize what is currently an academic concern into industry practice. Another mid-term goal is compiling a *large-scale real-world dataset* (or a collection of datasets) through contributions from many companies/universities, so that by 5 years out, having real data in the loop is routine.
- **Long-Term (5+ Years)**: Establish *open standards or benchmark competitions* akin to ImageNet or KITTI (for autonomous driving) specifically for wireless channel prediction and estimation. This could take the form of an *IEEE Standard for Wireless AI Model Evaluation* that codifies the principles (no train/test generator overlap, required robustness tests, etc.). Additionally, by this time, we hope an *open repository of channel models and measurements* exists – a “ChannelNet” where anyone can submit a new environment’s data and enrich the evaluation set. With widespread adoption, publishing results on multiple standardized benchmarks (synthetic and real) will be as expected as reporting FLOPs or parameter counts. Finally, we foresee *continuous evaluation* platforms: much like how

MLPerf continuously evaluates hardware on ML tasks, a platform could continuously ingest new CSI traces from, say, pilot signals in live networks (with privacy preserved), and evaluate deployed models in real time. This would truly close the loop between lab and field, ensuring that "ground truth" is always tied to the ground (reality) and not just an assumption.

By following this roadmap, the community can transition to evaluation practices that yield **robust, trustworthy models**. The end result will be a set of AI tools for channel prediction that have proven themselves under rigorous scrutiny – models that operators can deploy with confidence because they've been tested in conditions as harsh as reality itself.