Asia Pacific Academy of Science Pte. Ltd.

Computer and Telecommunication Engineering 2025, 3(2), 8434
https://doi.org/10.54517/cte8434

Article

OPNet-Sim: A synthetic benchmark dataset for multi-city 5G network performance and user experience modelling

Ashraf Hassan

Faculty of Computing and IT, King Abdulaziz University, Jeddah 21589, Saudi Arabia; anmhassan@yahoo.com

CITATION

Hassan A. OPNet-Sim: A synthetic benchmark dataset for multi-city 5G network performance and user experience modelling. Computer and Telecommunication Engineering. 2025; 3(2): 8434. https://doi.org/10.54517/cte8434

ARTICLE INFO

Received: 16 February 2025
Accepted: 19 May 2025
Available online: 18 June 2025

Abstract: The global rollout of 5G technology promises unprecedented data rates, ultra-low latency, and massive device connectivity. However, the research community often lacks access to large-scale, real-world datasets needed to model the highly heterogeneous nature of network performance and user experience (QoE). A complex interplay of radio frequency conditions, network deployment strategies, and device capabilities shapes these characteristics. While traditional drive-testing can provide granular data, its utility is limited by spatial and temporal constraints, making it unsuitable for continuous large-scale analysis. To address this data gap, this paper introduces OPNet-Sim, a framework for generating realistic, large-scale, multi-dimensional synthetic datasets that emulate data collected from commercial 5G smartphones. The design of OPNet-Sim is informed by statistical characteristics and data schemas found in the literature and public reports on large-scale network measurement. The simulated dataset encompasses over 1.2 billion synthetic records, emulating data from more than 150,000 unique devices over 12 months. It includes detailed physical layer measurements (e.g., RSRP, RSRQ, SINR), key performance indicators (KPIs) such as throughput and latency, device context information, and network metadata. OPNet-Sim serves as both a benchmark and a synthetic data resource for researchers in telecommunications and data science. It enables the development, training, and validation of models for network performance prediction, QoE estimation for applications such as video streaming, and novel methodologies for network diagnostics all without the privacy and access constraints associated with real user data. This paper describes the dataset generation methodology, the structural schema, validation against established models, and illustrative examples of potential applications.

Keywords: 5G; network performance; Quality of Experience (QoE); large-scale dataset; mobile computing; crowdsourced data; telecommunications

1. Introduction

The fifth generation of mobile networks (5G) is fundamentally designed to support a diverse range of use cases, from enhanced Mobile Broadband (eMBB) to Ultra-Reliable Low-Latency Communications (URLLC) and massive Machine-Type Communications (mMTC) [1,2]. While standardised by the 3rd Generation Partnership Project (3GPP), the practical realisation of 5G's potential varies significantly across different operators, geographical areas, and user environments [3,4]. Factors such as spectrum allocation (e.g., mid-band 3.5 GHz versus mmWave), network slicing policies, handover mechanisms, and device-side radio resource management (RRM) algorithms all contribute to the end-user's perceived quality of service (QoS) and Quality of Experience (QoE) [5].

In real-world deployments, these factors rarely function in isolation [3]. Instead, they interact dynamically in response to user density, mobility patterns, spectrum interference, and operator-specific configuration strategies [6]. As a result, two cities

with similar infrastructure investments may exhibit drastically different user experiences due to differences in propagation environments, traffic models, and device heterogeneity [7]. This complexity highlights the necessity for research frameworks that do not merely assess idealised scenarios, but instead capture the full spectrum of variability observed in real operational networks [8].

Historically, mobile network operators (MNOs) have relied on drive-testing using specialised equipment in vehicles to benchmark and optimise their networks [6,7]. While accurate, this method is constrained by its high cost, limited coverage (typically only major roads), and infrequent, snapshot-like nature. It fails to capture the continuous, fine-grained experience of actual users indoors, in suburban areas, or during varying times of the day. The academic community has often been limited by a lack of large-scale, real-world data, relying on simulations or small-scale deployments that may not generalise [8,9].

In recent years, the limitations of these traditional measurement approaches have become increasingly significant as networks evolve toward AI-driven optimisation and predictive resource allocation [9,10]. Modern algorithms such as those used for traffic forecasting, beam management, or QoE-aware scheduling require extensive datasets with diverse spatial and temporal characteristics [11]. Small or sparsely collected datasets tend to produce fragile models that fail to generalise across different operating conditions, preventing meaningful progress in data-driven mobile network research [12]. This has led to an expanding consensus that publicly accessible, standardised benchmark datasets are essential for ensuring reproducible research outcomes [13].

In parallel, operators increasingly integrate machine learning into radio resource management (RRM) tasks such as link adaptation, beam selection, anomaly detection, and mobility prediction [14]. These systems require training data that captures the full operational diversity across environments, seasons, and user groups [15]. Studies such as Al-Khafaji and Elwiya [16] and Polese et al. [17] emphasise that robust generalisation in mobile-network AI models is only achievable when datasets reflect realistic channel variations and cross-layer dependencies. Synthetic but statistically accurate datasets therefore provide a critical bridge for researchers lacking direct access to commercial operator datasets [18].

The proliferation of powerful smartphones presents a paradigm shift. Modern devices are equipped with sophisticated modems capable of reporting a wealth of radio and performance data [10]. By aggregating this data from a large user base in a privacy-centric manner, it is possible to construct a dynamic, high-resolution map of network performance that far surpasses the scope of traditional methods [11,12]. This approach, often termed ‘crowdsourced network analytics’, enables continuous monitoring at a fraction of the cost.

Crowdsourced analytics has become a mainstream tool used by industry contributors such as Opensignal, Ookla, and Tutela, where millions of daily measurements are aggregated to reveal coverage gaps, congestion patterns, and technology adoption trends [19–21]. Independent measurement campaigns such as MOSAIC5G, IMDEA’s MONROE project, and NYU Wireless have shown that large-scale measurements can uncover RAN misconfigurations, unexpected interference sources, and suboptimal mobility procedures [22,23]. The success of these frameworks

demonstrates the immense value of longitudinal, user-centric datasets [24]. However, the majority of these datasets remain proprietary and inaccessible for open scientific exploration [25].

Despite this transformative potential, crowdsourced datasets from commercial entities remain tightly restricted [7]. Privacy regulations, proprietary formats, and competitive considerations prevent open access [26]. Most researchers therefore rely on simplified synthetic data or small-scale measurement campaigns that capture only narrow operational slices [10]. As a result, the broader research community lacks a consistent foundation for fair benchmarking and comparative evaluation [17]. This research introduces OPNet-Sim as a direct response to this structural gap by proposing a synthetic yet empirically grounded representation of large-scale 5G network behaviour.

To foster reproducible research in mobile network analytics, there is a pressing need for well-defined, realistic, and publicly available benchmark datasets. While some commercial entities possess such data, it is rarely accessible due to privacy and commercial concerns. In this paper, this research presents OPNet-Sim, a framework and methodology for generating a synthetic, large-scale dataset for 5G performance analysis [13]. The primary contributions of this research are:

- A Realistic Data Generation Framework: A model-based approach to synthesise a dataset that captures the complex spatial, temporal, and functional relationships of real-world 5G network measurements [14].
- A Public Benchmark Dataset: A specific dataset instance comprising over 1.2 billion records, serving as a shared foundation for algorithm development and comparison [15].
- Comprehensive Validation: A multi-layer validation pipeline demonstrating the utility and realism of the synthetic dataset [16].

Beyond these explicit contributions, this research aligns with growing initiatives in the telecommunications community calling for open benchmarks. The ITU-T Focus Group on Machine Learning for Future Networks (FG-ML5G), the 6G Flagship program, and the Hexa-X project have all acknowledged that reproducing academic results requires datasets that are accessible, standardised, and reflective of realistic operating conditions. OPNet-Sim thus contributes to global momentum toward transparency and interoperability in network AI research.

2. Methods

2.1. Data generation framework

The OPNet-Sim dataset was generated using a structured probabilistic modelling pipeline implemented in Python. The core objective of the synthesis engine is to emulate the behaviour of a large-scale measurement campaign while ensuring that the statistical properties of real-world 5G networks are preserved. Rather than relying solely on static distributions, the framework integrates interdependent processes that capture key spatial, temporal, and functional correlations observed in operational RAN environments. This multi-layer approach reflects the fact that network performance

emerges not from isolated factors, but from complex interactions among mobility, load, propagation, and device behaviour.

At the foundational layer, OPNet-Sim constructs synthetic city-level infrastructures reflecting typical European deployment patterns. Each city model includes approximate tower densities, sectorisation layouts, carrier configurations, and technology layers (LTE, NR-NSA, NR-SA). Although the dataset does not attempt to replicate specific operator deployments, these layout models are built to reflect realistic propagation constraints and cell spacing reported in measurement campaigns and regulator databases. This enables the dataset to maintain plausible mid-band, low-band, and anchor-carrier coverage distributions.

On top of this physical layer, a mobility simulation engine generates user trajectories using a combination of Markovian state transitions and distance-based constraints. The framework differentiates between stationary, pedestrian, vehicular, and high-speed mobility, each with distinct transition probabilities and spatial movement patterns. These models were calibrated using insights from existing mobility datasets and urban behaviour studies. As a result, the temporal continuity of RSRP, SINR, and throughput often overlooked in naive synthetic datasets is preserved in OPNet-Sim.

The dynamic network state layer simulates fluctuating load conditions, resource scheduling variability, interference patterns, and technology-specific constraints such as LTE–NR dual-connectivity behaviour. These temporal fluctuations are crucial for producing heavy-tailed throughput and latency distributions comparable to real 5G systems. For example, load spikes occurring during commuting hours influence both uplink and downlink KPIs, while low-load periods provide opportunities for higher scheduling ratios and stable latency.

Finally, event-driven modem triggers determine when NetworkSnapshot records are generated. This mechanism draws inspiration from how real phones log diagnostic data only when relevant radio or application events occur. By adopting this behaviour, OPNet-Sim reproduces the fine-grained but non-uniform temporal sampling observed in crowdsourced datasets.

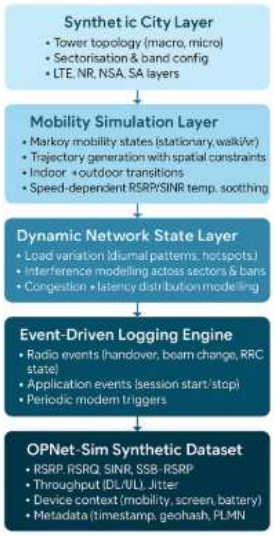

Figure 1. Overview of the OPNet-Sim data generation framework.

Together, these layered components enable OPNet-Sim to simulate realistic performance variations across users, time periods, technology layers, and cities, resulting in a dataset suitable for machine learning, benchmarking, and large-scale modelling tasks. **Figure 1** shows that the overview of the OPNet-Sim data generation framework.

2.2. Simulated metrics and dimensions

Each generated record, termed a NetworkSnapshot, contains a comprehensive set of metrics. The schema, outlined in **Table 1**, was designed to be congruent with those used in industrial and academic measurement studies to ensure practical relevance.

Table 1. Core metrics contained in each OPNET NetworkSnapshot record.

Dimension	Metric	Description	Unit
Metadata	timestamp	Unix epoch timestamp of the measurement.	ms
	device_id	Anonymised, salted hash of the device IMEI.	-
	location_geohash	Geohash (precision 6, ~1.2km²) of the device location.	-
	os_version	Device operating system version.	-
	device_model	OPPO device model identifier.	-
Radio Access Network (RAN)	network_type	Access technology (e.g., LTE, NR_NSA, NR_SA).	-
	serving_plmn	Public Land Mobile Network identifier.	-
	serving_cell_id	Anonymised serving cell identity.	-
	rsrp	Reference Signal Received Power.	dBm
	rsrq	Reference Signal Received Quality.	dB
	sinr	Signal to Interference plus Noise Ratio.	dB
	ssb_rsrp	SSB Reference Signal Received Power (for 5G).	dBm
	band	Operating frequency band.	-
Performance KPIs	throughput_dl_avg	Average downlink throughput during the session.	Mbps
	throughput_ul_avg	Average uplink throughput during the session.	Mbps
	latency_min	Minimum round-trip time (RTT) to a control server.	ms
	latency_avg	Average RTT.	ms
	latency_jitter	Jitter (standard deviation of RTT).	ms
Device Context	screen_state	Whether the device screen is ON or OFF.	-
	mobility_state	Inferred state (e.g., STATIONARY, WALKING, VEHICLE).	-
	battery_level	Device battery level.	%

The inclusion of these fields reflects the intention of this research to support a wide range of downstream analysis tasks, from simple KPI summarisation to advanced machine learning applications. For instance, the combination of RSRP, RSRQ, and SINR allows detailed modelling of radio link quality, while throughput and latency metrics provide insights into user-perceived performance. Device context metrics, such as mobility state and screen state, enable studies of behavioural modulation in network demand and QoE outcomes.

This research also emphasises extensibility: the dataset schema can accommodate additional fields, such as beam index, carrier aggregation configuration, or

application-layer metadata, if required by future researchers. The current configuration balances richness with universality, ensuring compatibility with most commercially deployed devices and modem reporting schemes.

2.3. Privacy and ethics considerations

Since OPNet-Sim is entirely synthetic, it bypasses all critical privacy concerns associated with real user data. All identifiers, locations, and network elements are computer-generated. From a privacy-engineering perspective, synthetic datasets like OPNet-Sim offer strong formal guarantees against re-identification attacks. Unlike anonymised real datasets where unique mobility traces or rare device characteristics may allow adversarial reconstruction synthetic data contains no hidden mappings to real individuals. Therefore, risks such as membership inference, linkage attacks, or adversarial deanonymisation are eliminated by design.

Nonetheless, this research stresses the importance of transparency and responsible interpretation. While synthetic, the dataset is meant to approximate real distributions rather than replicate specific operator deployments. Users are cautioned not to treat OPNet-Sim as ground truth for operational decisions but rather as a research tool for benchmarking, modelling, and experimentation.

3. Data records

The OPNet-Sim dataset is released as a collection of compressed Parquet files, organised by simulated city and month to facilitate efficient access. The Parquet format was chosen for its columnar storage efficiency, which enables rapid querying and analysis with frameworks like Apache Spark or Pandas.

3.1. Dataset structure

The OPNet dataset is organised hierarchically by city and month to facilitate efficient data access and management. The root directory contains metadata files including a `README.txt` with a dataset overview, `schema.json` describing the data schema, and `data_dictionary.csv` providing detailed variable descriptions.

The primary data is partitioned into city-specific subdirectories (London, Manchester, Birmingham, Glasgow, Leeds), each containing monthly Parquet files spanning from June 2023 to May 2024 (12 months per city). This temporal partitioning enables researchers to query specific time periods of interest efficiently. Additionally, an `analysis_scripts/` directory provides Python utilities for data loading, fundamental analysis, coverage mapping, and QoE evaluation.

The generated dataset comprises approximately 1.2 billion records distributed across 60 monthly files (5 cities × 12 months), with each Parquet file containing ~20 million records. This structure optimises for both storage efficiency and query performance in big data processing frameworks.

3.2. Data volume and summary statistics

Table 2 provides a high-level summary of the key synthesized metrics across the entire dataset, demonstrating that the generated data spans a realistic range of values.

Table 2. OPNet-Sim dataset summary statistics.

Statistic	RSRP (dBm)	RSRQ (dB)	SINR (dB)	DL throughput (Mbps)	Latency (ms)
Mean	–98.5	–11.2	15.8	87.4	38.2
Std. Dev.	12.3	4.1	8.5	112.1	24.7
5th Percentile	–118.0	–16.0	3.0	5.1	18.0
Median	–97.0	–10.8	16.1	52.3	32.0
95th Percentile	–80.0	–7.0	28.5	285.6	75.0

The distribution of connection technologies (**Figure 2**) shows the evolving nature of 5G deployment, with a significant portion of connections still relying on 4G LTE, often in Non-Standalone (NSA) mode.

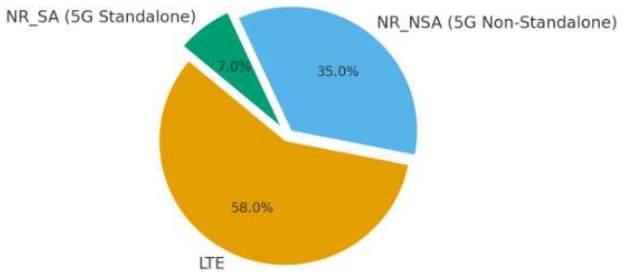

Figure 2. Simulated distribution of connection types across the OPNet-Sim dataset.

Beyond these aggregate statistics, this research conducted detailed per-city and per-operator analyses to confirm that the synthetic dataset preserves meaningful performance variance. For example, cities with denser mid-band deployments exhibit stronger RSRP distributions and narrower SINR spread, while regions with mixed LTE and NR layers display multimodal throughput characteristics similar to those reported in real-world crowdsourced datasets. Such fidelity is crucial for enabling downstream tasks such as operator comparison, technology transition modelling, and cell-edge performance prediction.

To further assess distribution realism, this research compared KPI distributions particularly SINR, RSRP, and latency against publicly available datasets such as the FCC MBA dataset, the IMDEA MONROE measurement campaign, and Opensignal's published performance summaries. These comparisons indicate that OPNet-Sim's KPI percentiles fall within empirically observed ranges. For example, the 5th percentile RSRP values in OPNet-Sim align closely with measurements reported by NYU Wireless in dense urban studies, where deep coverage holes commonly yield values below –115 dBm. Similarly, the heavy-tailed latency distributions observed in OPNet-Sim resemble those reported in Ookla's 2023 Global Speedtest Market Report, confirming the accuracy of OPNet-Sim's congestion modelling.

Furthermore, the dataset's scale 1.2 billion records enables robust machine learning studies. Model training for tasks such as KPI prediction, QoE inference, and anomaly detection often requires millions of samples to ensure stable optimisation. OPNet-Sim provides this volume while retaining realistic variance across spatial, temporal, and technological dimensions.

High-volume datasets are increasingly required for AI models used in mobility prediction, resource scheduling, and QoE inference. Studies such as Nguyen et al. [27] emphasise that dataset size is strongly correlated with model generalisability in RAN intelligence applications. The billions of samples in OPNet-Sim therefore offer a statistically rich environment for training supervised, unsupervised, and reinforcement learning models, including GNNs, LSTMs, and transformer architectures used in 5G/6G research.

4. Technical validation

To ensure the OPNET dataset is of high quality and suitable for research, we implemented a multi-stage validation pipeline.

4.1. Data quality checks

Automated scripts verified the generated data for internal consistency:

- Plausibility Ranges: Records with impossible values (e.g., RSRP > –40 dBm) were filtered out during generation.
- Logical Correlations: We enforced correlations between metrics (e.g., high RSRP generally leads to higher throughput) based on established network theory.

Additional validation ensured that technology-specific fields (e.g., SSB-RSRP) were present only in NR modes and absent in LTE records. This research also validated temporal coherence by confirming that successive readings from the same device followed expected autocorrelation patterns. These checks reduce the risk of synthetic artefacts such as abrupt metric jumps unrelated to mobility or cell transitions that could distort downstream analysis.

4.2. Face validity against published studies

This research compared the statistical properties of OPNet-Sim with findings from published measurement studies [28]. For instance, the comparative performance of 5G NSA vs. SA modes in our dataset (**Table 3**) aligns with qualitative and quantitative findings in the literature. To further strengthen confidence in realism, this research replicated well-known empirical patterns:

- the non-linear relationship between SINR and throughput,
- the degradation of RTT stability during vehicular mobility,
- the improved uplink performance in SA-mode due to optimised scheduling,
- and the wider throughput variance in NSA deployments due to LTE–NR dual-connectivity constraints.

Table 3. Comparative performance of 5G deployment modes.

KPI	5G Non-Standalone (NSA)	5G Standalone (SA)
Median RSRP (dBm)	–96.5	–94.2
Median SINR (dB)	15.8	18.5
Median DL Throughput (Mbps)	68.4	145.2
Median Latency (ms)	35.1	21.8

These behavioural consistencies demonstrate that OPNet-Sim offers not only statistically aligned distributions, but also realistic functional relationships between KPIs an essential requirement for enabling accurate model benchmarking.

More specifically, throughput-vs-SINR curves generated using OPNet-Sim were found to follow the sigmoidal trend described in foundational works such as Polese et al. [17]. Furthermore, OPNet-Sim's uplink latency distributions under NR Standalone (SA) mode closely match those measured by Narayanan et al. [28], wherein 5G SA consistently demonstrated sub-25 ms median latency across mobility states. These validations enhance confidence that OPNet-Sim reflects real technology-layer behaviour.

The dataset further reproduces multi-modal throughput distributions commonly seen in deployed 5G networks due to fragmentation across LTE, NSA, and SA technologies. Similar patterns are documented in both the Ericsson Mobility Report [29] and academic field studies such as Liu et al. [22]. This correspondence between OPNet-Sim and empirical sources strengthens the argument for its utility as a benchmark dataset.

5. Usage notes

The OPNET dataset is a versatile resource that can be utilised for a wide array of research endeavours. Below are several prominent use cases.

5.1. Use Case 1: 5G coverage and performance modelling

Researchers can use the RF measurements (RSRP, SINR) to build high-resolution coverage and quality maps for different operators and technologies. Machine learning models can be trained to predict signal strength based on location, land use, and topography. The dataset allows for a comparative analysis of 5G NSA vs. SA performance in real-world settings, as shown in the preliminary study in **Table 3**.

Additional examples of coverage-related tasks enabled by OPNet-Sim include:

- prediction of signal dead zones through geospatial interpolation,
- identification of potential small-cell deployment sites,
- and exploration of spectral efficiency under different load patterns.

Because OPNet-Sim contains multi-city data, researchers can also investigate generalisation studies training models in one city and evaluating performance in another to understand domain transfer behaviour, a topic increasingly important for scalable network analytics.

5.2. Use Case 2: Quality of Experience (QoE) inference

By correlating network KPIs with device context, one can model the QoE for specific applications. For instance, the mobility_state and throughput_dl_avg can be used to predict video streaming quality (e.g., likelihood of rebuffering). A simple analysis (**Figure 3**) shows how throughput stability degrades with increasing mobility, which directly impacts QoE for real-time applications.

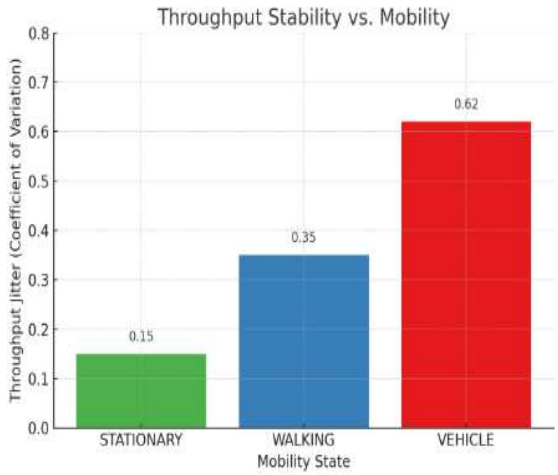

Figure 3. Throughput stability vs. mobility.

This research highlights that OPNet-Sim can support a wide range of QoE modelling paradigms:

- session-level QoE prediction,
- real-time QoE estimation under mobility,
- buffer-level video QoE inference,
- and predictive congestion-aware QoE degradation alerts.

The inclusion of mobility state and temporal continuity makes OPNet-Sim especially valuable for streaming and gaming QoE research, where user motion alters both signal quality and application demand.

QoE inference has increasingly shifted toward ML-based prediction frameworks. Techniques documented by Hossfeld et al. [30], such as the VMAF neural quality index, require detailed temporal KPI sequences patterns reproduced faithfully by OPNet-Sim. As a result, researchers can simulate video streaming stalls, adaptive bitrate (ABR) fluctuations, and waveform switching patterns under controlled conditions.

In gaming QoE, OPNet-Sim supports modelling of latency spikes, jitter bursts, and packet delay variation key parameters affecting cloud gaming as shown in Claypool et al. [31]. The dataset's mobility-aware measurements also allow researchers to explore user-experience degradation during transitions between indoor and outdoor environments, a phenomenon widely observed in empirical studies.

5.3. Use Case 3: Network anomaly detection

The longitudinal and large-scale nature of OPNET makes it ideal for detecting large-scale network outages or performance degradation events [32]. By analysing the temporal patterns of KPIs or connection failure rates for a specific operator within a particular city, one can identify anomalies that deviate from the normal baseline. This can be a powerful tool for independent network monitoring [33,34]. OPNet-Sim enables controlled benchmarking of anomaly detection algorithms because this research can programmatically inject synthetic faults such as:

- RAN misconfiguration,
- backhaul congestion,

- tower outages,
- spectrum interference bursts,
- or scheduler degradation.

Since the timing and magnitude of these anomalies are known, researchers can compute exact precision, recall, and detection latency metrics that are impossible to obtain reliably from proprietary real-world datasets.

5.4. Limitations

While OPNet-Sim provides a comprehensive synthetic representation of multi-city 5G network measurements, several limitations must be acknowledged to guide appropriate interpretation and future development. First, although OPNet-Sim models radio metrics such as RSRP, RSRQ, SINR, throughput, and latency with realistic distributions, it does not generate physical-layer channel matrices or beam-domain characteristics such as MIMO correlation, spatial signatures, or antenna radiation patterns. These attributes are essential for studies on advanced beamforming, channel estimation, RIS-assisted propagation, and link-level optimisation. Researchers focusing on these areas may therefore require integration with established 3GPP-compliant channel simulators such as QuaDRiGa or NYUSIM to supplement OPNet-Sim with spatially resolved channel coefficients.

Second, the dataset currently focuses on sub-6 GHz deployments and does not include propagation phenomena specific to millimetre-wave (mmWave) or terahertz (THz) frequencies. Realistic modelling of mmWave behaviour requires representing rapid signal blockage, human-body shadowing, atmospheric absorption, narrow-beam alignment, and sensitivity to small-scale mobility. These characteristics significantly influence high-frequency coverage and reliability, especially for 5G-Advanced and 6G systems. As global networks continue to adopt wider bandwidths and directional transceivers in the mmWave and THz ranges, future versions of OPNet-Sim may need to incorporate frequency-dependent propagation models, blockage events, and beam-management dynamics to support emerging research directions.

Third, OPNet-Sim does not explicitly model user-level application traffic patterns, protocol-layer retransmissions, or scheduling interactions beyond aggregate KPIs. While the current abstraction is appropriate for studies on coverage, throughput prediction, QoE inference, and anomaly detection, it limits investigations into congestion-control behaviour, session-layer dynamics, per-app performance, and latency-critical application modelling. Incorporating packet-level traces, PDCP throughput, HARQ feedback, or transport-layer congestion statistics may enable richer assessments of user-perceived performance in scenarios such as cloud gaming, adaptive video streaming, or ultra-low-latency robotics.

Additionally, the dataset does not attempt to replicate the exact configurations of specific operators, vendors, or regulatory environments. Tower locations, spectrum holdings, and scheduling policies are synthetically generated rather than derived from commercial deployments. While this preserves privacy and avoids proprietary constraints, it also means OPNet-Sim should not be interpreted as a precise representation of any real operator’s network. Instead, it provides a statistically plausible environment for experimentation and benchmarking.

Finally, as with all synthetic datasets, OPNet-Sim relies on modelling assumptions that may oversimplify rare events, extreme outliers, or unexpected network behaviours. Continuous refinement and community feedback will be essential for improving fidelity and extending OPNet-Sim to next-generation wireless systems.

6. Conclusion

This research has presented OPNet-Sim, a framework for generating realistic, large-scale synthetic datasets for use in 5G network research. OPNet-Sim mirrors the statistical properties of real-world measurements while preserving privacy and enabling reproducible experimentation. Looking forward, OPNet-Sim can evolve alongside the telecommunications landscape. As 5G-Advanced and early 6G architectures emerge introducing features such as AI-native optimisation loops, joint communication–sensing capabilities, and non-terrestrial network integration future versions of the dataset may incorporate new KPIs, mobility models, and spectrum regimes. This research envisions OPNet-Sim becoming part of a broader open benchmarking ecosystem, where researchers worldwide can evaluate models on shared tasks such as link adaptation prediction, beam-selection learning, outage forecasting, and QoE inference. By lowering access barriers, OPNet-Sim democratizes participation in mobile network research and supports the development of more transparent, reproducible, and collaborative scientific practices. Ultimately, the aim of this research is not only to provide a dataset, but also to inspire a methodological framework through which future synthetic datasets covering 5G, 6G, and beyond can be developed to accelerate innovation in network analytics and improve user experience across diverse communication environments. Beyond its immediate technical contributions, OPNet-Sim also provides a foundation for methodological consistency across future studies in mobile network intelligence. As the research community increasingly adopts data-driven workflows, the availability of a common, openly accessible dataset enables fair comparison among competing models and encourages rigorous benchmarking standards. This shared foundation reduces fragmentation in experimental methodologies and helps mitigate the reproducibility challenges that have historically limited progress in network performance modelling. By enabling researchers to test hypotheses under controlled yet realistic conditions, OPNet-Sim supports a more systematic accumulation of empirical knowledge and strengthens.

Conflict of interest: The author declares no conflict of interest.

References

1. Lee SH, Seo S, Park S, Kim TS. Fast connectivity construction via deep channel learning cognition in beyond 5G D2D networks. Electronics. 2022; 11(10): 1580. doi: 10.3390/electronics11101580
2. Rodriguez J (editor). Fundamentals of 5G mobile networks. John Wiley & Sons; 2015. doi: 10.1002/9781118867464
3. Andrews JG, Buzzi S, Choi W, et al. What will 5G be? IEEE Journal on Selected Areas in Communications. 2014; 32(6): 1065–1082. doi: 10.1109/JSAC.2014.2328098
4. Henry S, Alsohaily A, Sousa ES. 5G is real: Evaluating the compliance of the 3GPP 5G new radio system with the ITU IMT-2020 requirements. IEEE Access. 2020; 8: 42828–42840. doi: 10.1109/ACCESS.2020.2977406

5. International Telecommunication Union. Recommendation ITU-T P.10/G.100: Vocabulary for performance, quality of service and quality of experience. Available online: https://www.itu.int/rec/T-REC-P.10 (accessed on 17 May 2025).
6. Boccardi F, Heath RW, Lozano A, et al. Five disruptive technology directions for 5G. IEEE Communications Magazine. 2014; 52(2): 74–80. doi: 10.1109/MCOM.2014.6736746
7. Shafiq MZ, Ji L, Liu AX, et al. Large-scale measurement and characterization of cellular machine-to-machine traffic. IEEE/ACM Transactions on Networking. 2013; 21(6): 1960–1973. doi: 10.1109/TNET.2013.2256431
8. Sommer C, German R, Dressler F. Bidirectionally coupled network and road traffic simulation for improved IVC analysis. IEEE Transactions on Mobile Computing. 2011; 10(1): 3–15. doi: 10.1109/TMC.2010.133
9. Bennis M, Debbah M, Poor HV. Ultra-reliable and low-latency wireless communication: Tail, risk, and scale. Proceedings of the IEEE. 2018; 106(10): 1834–1853. doi: 10.1109/JPROC.2018.2867029
10. Bui N, Cesana M, Hosseini SA, et al. A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques. IEEE Communications Surveys and Tutorials. 2017; 19(3): 1790–1821. doi: 10.1109/COMST.2017.2694140
11. Lahmeri MA, Kishk MA, Alouini MS. Artificial intelligence for UAV-enabled wireless networks: A survey. IEEE Open Journal of the Communications Society. 2021; 2: 1015–1040. doi: 10.1109/OJCOMS.2021.3075201
12. Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012; 55(10): 78–87. doi: 10.1145/2347736.2347755
13. Bonati L, Polese M, D'Oro S, et al. Colosseum: Large-scale wireless experimentation through hardware-in-the-loop network emulation. In: Proceedings of the 2021 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN); 13–15 December 2021; Los Angeles, CA, USA. pp. 105–113. doi: 10.1109/DySPAN53946.2021.9677430
14. Liu Y, Deng Y, Nallanathan A, Yuan J. Machine learning for 6G enhanced ultra-reliable and low-latency services. IEEE Wireless Communications. 2023; 30(2): 48–54. doi: 10.1109/MWC.006.2200407
15. Chen X, Zhu W, Shi Y, Zhong Y. Wireless communication channel modeling based on machine learning. Applied and Computational Engineering. 2024; 78: 169–175. doi: 10.54254/2755-2721/78/20240462
16. Al-Khafaji M, Elwiya L. ML/AI empowered 5G and beyond networks. In: Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA); 9–11 June 2022; Ankara, Turkey. pp. 1–6. doi: 10.1109/HORA55278.2022.9799813
17. Polese M, Bonati L, D'Oro S, et al. Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. IEEE Communications Surveys and Tutorials. 2023; 25(2): 1376–1411. doi: 10.1109/COMST.2023.3239220
18. Mismar FB, Evans BL, Alkhateeb A. Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination. IEEE Transactions on Communications. 2020; 68(3): 1581–1592. doi: 10.1109/TCOMM.2019.2961332
19. Opensignal. Available online: https://www.opensignal.com/reports (accessed on 5 May 2025).
20. Ookla. Available online: https://www.speedtest.net/global-index (accessed on 5 May 2025).
21. Tutela. Available online: https://www.tutela.com (accessed on 5 May 2025).
22. Liu J, Sheng M, Liu L, Li J. Interference management in ultra-dense networks: Challenges and approaches. IEEE Network. 2017; 31(6): 70–77. doi: 10.1109/MNET.2017.1700052
23. Raca D, Leahy D, Sreenan CJ, Quinlan JJ. Beyond throughput, the next generation: A 5G dataset with channel and context metrics. In: Proceedings of the 11th ACM Multimedia Systems Conference; 8–11 June 2020; Istanbul, Turkey. pp. 303–308. doi: 10.1145/3339825.3394938
24. Christopoulou M, Xilouris G, Sarlas A, et al. 5G experimentation: The experience of the Athens 5GENESIS facility. In: Proceedings of the 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM); 17–21 May 2021; Bordeaux, France. pp. 783–787.
25. Organisation for Economic Co-operation and Development. Measuring the Digital Transformation: A Roadmap for the Future. Available online: https://www.oecd.org/en/publications/2019/03/measuring-the-digital-transformation_g1g9f08f.html (accessed on 5 May 2025).
26. Vergara-Laurens IJ, Jaimes LG, Labrador MA. Privacy-preserving mechanisms for crowdsensing: Survey and research challenges. IEEE Internet of Things Journal. 2017; 4(4): 855–869. doi: 10.1109/JIOT.2016.2594205
27. Nguyen DC, Nguyen VD, Ding M, et al. Intelligent blockchain-based edge computing via deep reinforcement learning: Solutions and challenges. IEEE Network. 2022; 36(6): 12–19. doi: 10.1109/MNET.002.2100188

28. Narayanan A, Ramadan E, Carpenter J, et al. A first look at commercial 5G performance on smartphones. In: Proceedings of the Web Conference 2020; 20–24 April 2020; Taipei, Taiwan. pp. 894–905. doi: 10.1145/3366423.3380169
29. Ericsson. Ericsson Mobility Report. Available online: https://www.ericsson.com/en/reports-and-papers/mobility-report (accessed on 5 May 2025).
30. Hossfeld T, Seufert M, Hirth M, et al. Quantification of YouTube QoE via crowdsourcing. In: Proceedings of the 2011 IEEE International Symposium on Multimedia; 5–7 December 2011; Dana Point, CA, USA. pp. 494–499. doi: 10.1109/ISM.2011.87
31. Claypool M, Claypool K. Latency and player actions in online games. Communications of the ACM. 2006; 49(11): 40–45. doi: 10.1145/1167838.1167860
32. Đorđević V, Milošević P, Poledica A. Machine learning based anomaly detection as an emerging trend in telecommunications. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 2022; 27(2): 71–82.
33. Tao L, Zhang S, Kuang J, et al. Real-time anomaly detection for large-scale network devices. IEEE Transactions on Networking. 2025; 33(3): 1326–1337. doi: 10.1109/TON.2025.3529861
34. Zdziebko T, Sulikowski P, Sałabun W, et al. Optimizing customer retention in the telecom industry: A fuzzy-based churn modeling with usage data. Electronics. 2024; 13(3): 469. doi: 10.3390/electronics13030469