Extending Causal Discovery to Live 5G NR Network With Novel Proportional Fair Scheduler Enhancements

Roopesh Kumar Polaganga, *Graduate Student Member, IEEE*, and Qilian Liang, *Fellow, IEEE*

***Abstract*—Understanding the cause-and-effect connections in intricate systems like telecommunication networks is essential for enhancing and optimizing their performance. This article, in its novelty, extends causal discovery to the context of 5G new radio (NR) Networks by leveraging established technique of the greedy equivalent search (GES) algorithm to real-world 5G network data. Our research uncovers causal relations among several network characteristics represented in the form of a directed acyclic graph (DAG). A specific causal link between latency and downlink throughput is further analyzed unveiling a robust correlation between network utilization and latency, manifesting the factors of packet loss and retransmissions especially when the network infrastructure becomes overwhelmed. Building on these findings, the second significant contribution of our study involves the introduction of a novel enhancement to the NR proportional fair (PF) scheduling algorithm. This enhancement incorporates retransmission considerations to improve network resource utilization. Our experimental results show notable gains in network efficiency and resource allocation, highlighting the potential for real-world enhancements based on the causal insights uncovered in our research. This research broadens the horizons of causal discovery within 5G NR networks and presents a tangible pathway for enhancing network performance and resource allocation, with implications for the broader field of network optimization.**

***Index Terms*—5G new radio (NR) networks, causal discovery, causality, greedy equivalent Search (GES), proportional fair (PF) scheduler.**

I. Introduction

Current machine learning approaches usually tend to exploit the correlations observed between the input data elements [2], [11]. However, there could be spurious correlations within the data that may sometimes lead us to wrong conclusions. For example, one of the data illustrations from [6] suggests a very high (99.79%) correlation between the U.S. spending on science, space, and technology with the suicides happened by hanging, strangulation and suffocation. Such high correlation may suggest some false conclusions of U.S. spending on science and technology may have been the reason for an increase in suicides happened, but there could be no direct dependency on these two data metrics and requires no actions taken. Hence, there is a need to explore causality for the data produced from complex systems like a telco network so that an actionable outcome driven cause-effect analysis can be formulated to improve network utilization and overall performance.

Fig. 1. Example of a DAG.

Causality is the fundamental concept of identifying cause-and-effect relationships in various fields as introduced by Judea Pearl in [17] and [25]. It helps to understand how changes in one factor can lead to the changes in another and thus enabling us to identify and predict the outcome of various events. Intricate systems like a real-world telecommunications network can have numerous interconnected components and variables. Causality in this context can help unravel the complex cause-and-effect relationships among network attributes. The insights gained from causal discovery offer a foundation for practical applications of network optimization and end-user experience enhancement for 5G and beyond. By understanding the impact of changes in network characteristics on the Quality of Service (QoS), causality plays a pivotal role in managing and evolving mobile networks to meet the ever-increasing demands of a connected world.

Directed acyclic graphs (DAGs) are often used to represent a causal relationship between variables [13]. The direction of the edges in a DAG is interpreted as causal or directional relationship between variables. The vertices (circles) in a causal DAG represent variables and edges (arrows) represent causation. An edge from A to B suggests that A influences B as shown in Fig. 1. The most fundamental property of a DAG is that it is acyclic. This means there are no closed loops or cycles in the graph [25].

While causality represents the fundamental idea that one variable influence another, causal inference refers to the process of drawing conclusions about causality from observed data, often involving statistical or experimental techniques.

Received 26 February 2024; revised 21 July 2024 and 4 September 2024; accepted 10 September 2024. Date of publication 13 September 2024; date of current version 25 December 2024. This work was supported in part by the U.S. National Science Foundation under Grant CCF-2219753. *(Corresponding author: Roopesh Kumar Polaganga.)*

Roopesh Kumar Polaganga is with the Access Network Development, T-Mobile US Inc., Bellevue, WA 98012 USA, and also with the Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX 76010 USA (e-mail: RoopeshKumar.Polaganga@T-Mobile.com).

Qilian Liang is with the Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX 76010 USA (e-mail: Liang@Uta.edu).

Digital Object Identifier 10.1109/JIOT.2024.3459798

TABLE I
TYPES OF CAUSAL DISCOVERY ALGORITHMS

Approach	Algorithms
Constraint-Based	PC (*Peter & Clark*) FCI (*Fast Causal Inference*) RFCI (*Reliable FCI*)
Score-Based	Hill-Climbing Search Grow-Shrink Algorithm
Hybrid	GES (*Greedy Equivalent Search*)

Causal discovery (also referred to as causal structure search in [21]) goes a step further, focusing on the automated or algorithmic identification of causal relationships within data, especially when the causal structure is unknown [18]. Analyzing such statistical properties of purely observational data is essential especially when it is difficult or even impossible to conduct interventions to obtain causal relations. Causal discovery has been greatly exploited across the fields like biomedical science fields [4], [9], [14] while some introductory work has been done in [3] to extend the causal discovery algorithms to different IT monitoring databases and to explore its feasibility. Verhelst [20] explored causality via causal inference related to telco network to understand customer churn. However, no work has been done to extend the concept of causality and the causal discovery algorithms to real-world 5G new radio (NR) mobile network data. This is where our novel work contributes and helps to improve system performance with actionable outcomes.

A. Causal Discovery Algorithms

Glymour [21] summarized the computational methods for causal discovery that were developed in the past three decades along with some illustrations. A python-based "causal-learn" extension has been developed by [1] and provided as an intuitive application programming interface (API) for researchers to explore and apply multiple causal discovery algorithms based on their observational data. Table I is an attempt to summarize the causal discovery algorithms mentioned in the literature and categorized by their approach. Nevertheless, it is essential to re-emphasize that these methods are not uniformly effective, and there is no one superior algorithm over others in obtaining causal discovery results but rather depends on the specific data sets under consideration. Understanding the assumptions and limitations for each of these algorithms is needed to identify the right algorithm to be applied to our 5G NR network data.

Constraint-based algorithms like PC and FCI handles various data types using reliable conditional independence tests. However, PC assumes linearity with no confounding and thus providing asymptotically correct results. FCI can accommodate confounders outputs equivalence classes and not complete causal information [28]. RFCI is an extension of FCI that incorporates reliability constraints but still has the equivalent class limitation. They are complemented by score-based algorithms, like hill-climbing search and grow-shrink algorithms that aim to infer causal relationships from data by optimizing a scoring criterion [25], [29]. Hybrid approach like greedy equivalent search (GES) introduced in [27] combines elements of both constraint-based and score-based methods to find a causal graph. It starts with a fully connected graph and iteratively applies operations to add, remove, or reverse edges while optimizing a score, typically the Bayesian information criterion (BIC) score [22], [30], [31]. GES algorithm has been extended to this work to obtain causal discovery due to its hybrid approach and its ability to accommodate nonlinear data since linearity cannot be assumed for 5G NR network data. Additional details on network data used in this work is explained at a greater detail in Section II.

As the GES algorithm is used in the context of Bayesian networks, it seeks to identify the optimal structure of a Bayesian network that best fits the input data. It uses conditional independence tests and searches for the most likely directed edges between variables. While GES does not have a single mathematical formula, its operation per [1], [27] is attempted to be outlined in below steps:

1) *Initialization:* Start with a fully connected graph, where all variables are connected to each other.
2) *Evaluate Candidate Operations:* For each pair of variables A and B in the data set, three possible operations are considered.
   a) $A \rightarrow B$ (directed edge from A to B).
   b) $B \rightarrow A$ (directed edge from B to A).
   c) $A \longleftrightarrow B$ (undirected edge between A and B).
3) *Score the Operations:* Calculate a score for each of the candidate operations using a scoring criterion, such as the BIC which is used for this work. The score quantifies how well the operation fits the data while penalizing for model complexity.
   a) $\text{Score}(A \rightarrow B) = \text{BIC}(\text{DAG} + A \rightarrow B)$.
   b) $\text{Score}(B \rightarrow A) = \text{BIC}(\text{DAG} + B \rightarrow A)$.
   c) $\text{Score}(A \longleftrightarrow B) = \text{BIC}(\text{DAG} + A \longleftrightarrow B)$.

   where BIC(DAG) represents the BIC score for the entire DAG.
4) *Select the Best Operation:* Choose the operation with the highest score among all the candidates. If the highest score is negative, then no operation is performed for this pair.
5) *Update the Graph:* If an operation was selected in Step 4, update the graph G accordingly by adding, removing, or reversing the corresponding edge.
6) *Repeat Steps 2–5:* Iterate through all possible pairs of variables based on the input data, considering operations, scoring, and updating the graphs till no more operations result in further score improvement. The final graph represents the estimated causal structure that maximizes the chosen scoring criterion for given input data.

By following this approach, GES aims to find a DAG that represents the causal structure of the data while adhering to the observed conditional independence relationships.

The remainder of this article is structured as follows. In Section II, we delve into the specifics of the real-world 5G NR network data utilized for causal discovery. Section III presents the results of causal discovery and

TABLE II
NETWORK ATTRIBUTES

Variable	*Network Attributes*	*Unit*
X1	Mean Timing Advance	meters
X2	Session Duration	sec
X3	Mean CQI	#
X4	Mean PUSCH SINR	dB
X5	Average MAC DL Throughput	kbps
X6	Average MAC UL Throughput	kbps
X7	Mean Latency	msec
X8	Mean Jitter	msec

conducts a comprehensive data analysis involving network variables. Section IV introduces enhancements to the NR proportional fair (PF) scheduling algorithm and illustrates the improvements observed through simulations. Section V offers the concluding remarks, while Section VI outlines potential avenues for future research.

II. Real-World (Live) 5G NR Network Data

A. Network Attributes

Mobile network operators (MNOs) often have multiple streams of data that is being generated from various network elements and stored per its utility. For example, all the RAN network data that streams from gNBs can be stored as one data set while the diagnostic data collected from user equipment (UEs) can be stored in a separate data set. Such storage and usage can vary from operator to operator based on their specific need and their usability of such data sets.

In this work, two different data sources that update periodically are explored and consolidated per desired geo-location and timeframe such that they complement each other. First data source is referred to as "Session Records" that store the records of each individual session of every UE while they are connected on 5G NR network. These detailed records are stored for every single UE that's connected to target gNB that is under consideration. This data is directly streamed from gNBs to storing servers where data is parsed and consumed in desired tabular format. Since there could be few hundred sessions being served by gNBs at every given instance, these records are further aggregated for easy handling. Eight different network attributes are chosen such that there is a consistent data availability with no data integrity issues like missing data variables within the data sets. Each of these network attributes are assigned a variable (X1, X2, ... X8) for further analysis purpose and listed in Table II.

Second data source is referred to as "performance monitoring (PM) records" where gNB periodically streams Counters that are rolled up as key performance indicators (KPIs) that operators typically use to monitor their network performance. These predefined metrics are populated for every 15 min interval irrespective of the amount of traffic its carrying and gets stored in operations support systems (OSSs) that further parse and stores the data. Key difference among the two data sources is that while Session Records data gives valuable insights into user sessions, it does not have any visibility on overall network loading conditions experienced by gNBs. While "PM Records" data does provide KPIs that shows the network utilization on radio resource level, but the user experience metrics are consolidated among all users and cannot depict a clear end user experience like Session Records data does. Hence, there is a need to consider both the data sets together to gain a holistic view of causal discovery observed.

B. Data Collection

Input data sets used for this work are collected from a U.S.-based MNO's live 5G NR network. Data is collected directly from gNBs for both Session Records and PM Records and aggregated on hourly basis. Complete data set comprises of user sessions for an entire 24 h duration on a typical weekday in the month of July 2023. Data collection is from multiple sites located in Seattle, WA area with a mix of both densely and lightly loaded scenarios spreading across urban and rural areas. Since NR is implemented as both standalone (SA) as well as non-SA (NSA) in this network, data sets account for both the types of NR implementations. However, since it is the same gNBs that are operating in both SA + NSA mode, same radio resources are shared among both implementations. This network has NR implemented in FR1 with N41 (2500 MHz) as its mid-band NR layer with 100 Mhz bandwidth while N71 (600 MHz) is its low-band NR layer with 15 Mhz bandwidth. Since all the gNBs under consideration belong to a single RAN vendor and have a similar network configuration with no significant differences, all the NR feature sets in place are common and accounts for data consistency. All the entries are further filtered for nonguaranteed bit rate (non-GBR) class of 5G channel quality indicators (5QIs) representing a typical data session (excluding voice) performed by the users. This is to account for the fact that voice traffic is carried by 4G long term evolution (LTE) in case of NSA sessions while it can be carried by 5G NR in case of SA as voice of NR (VoNR) is enabled in this network. Final data set from Session Records has about 176 000 sessions sampled after eliminating missing network attributes and accounting for data consistency.

III. Causal Discovery Results

A. Causal Discovery

Applying the GES algorithm using causal-learn python API provided in [1] to the real-world live 5G NR network data for eight key network attributes mentioned earlier results in the causal discovery DAG as shown in Fig. 2.

Even when the network attributes are given in specific order of X1 to X8, the GES algorithm's output is observed to be in a different order which suggests that all the combinations of network attributes are evaluated and only the ones that has statistically high cause-and-effect relations been formed as depicted in the causal discovery DAG. Duration of the session is observed to be the primary cause among all the network attributes and influences most of the others. This is because a longer session gets influenced by the inherent mobility scenarios in the real-world. Both the uplink and downlink (DL) throughputs are caused by channel quality indicator (CQI) and signal-to-interference-noise-ratio (SINR) seems to

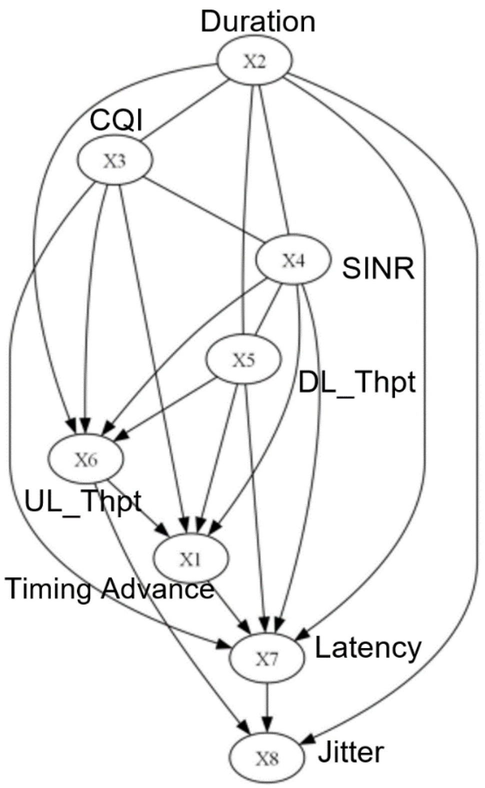


Fig. 2. GES algorithm-based causal discovery DAG.

be inherent and reflects the role played by radio channel and its quality metrics. However, no clear relation is identified between session duration and channel quality metrics of CQI and SINR as they did not meet the statistical criteria. In real world, it is not always going to be direct relations among these attributes since an increase in duration doesn't necessarily cause a change in channel conditions unless the user is in mobility. One key observation among throughput relations is that DL throughput seems to be causing uplink throughput but not the other way around. This is because this network is heavily utilized in DL and such DL throughput seems to require ack/nack in the uplink which in turn contributes to uplink throughput. Also, both uplink and DL throughputs tend to cause timing advance along with channel quality metrics of CQI and SINR. Latency is the key outcome that is being caused by channel conditions of CQI, SINR along with session duration, timing advance and DL throughput. Very end of this causal discovery observations has the network attribute of jitter which caused primarily by latency along with uplink throughput and duration of the session.

This research aims to delve deeper into the direct causal relationship observed between DL throughput and latency in comparison to the indirect relation between uplink throughput and latency via timing advance. This is to help further reinforce the observed causality using additional methods, including Granger causality (GC) and correlation. Such deeper analysis is required for action-based causal analysis to help improve system's performance.

B. Granger Causality for Throughput and Latency

GC is another statistical concept that also helps to determine the causal relationship between two time series data sets. Recent extension of GC to 5G network as been done in [5] to show its utility to optimize mobile networks. It refers to the idea that if the past values of a time series help to predict the future values of another time series, then we can conclude that first time series "Granger Causes" the second time series [23]. In other words, GC is a measure of causal relationship between two time series, and it is used to determine whether one time series can be used to predict the future values of another time series. However, it is important to note that Ganger causality does not imply a causal relationship in the sense of a direct cause-and-effect relationship. Rather, it simply indicates that one time series can be used to predict the future value of another time series, which may or may not be due to a causal relationship between the two. As the Session Records data collected per UE is time series based, GC can be applied to quantify the predictability of one series on another.

GC can be measured using equation where x_t and y_t be the covariance stationary sequences, set up a regression model of x_t for lags of y and x as shown in (1). MATLAB toolbox provided in [24] helps to implement and quantify GC for the input 5G NR network data

$$x_t = c + \sum_{i=1}^{n} h_t y_{t-i} + \sum_{j=1}^{n} a_j x_{t-j} + \varepsilon_t \qquad (1)$$

where c is constant.

Since GC is directional, quantification is performed per relation in both directions. Observed GC results as shown in Fig. 3 illustrates a high causal relation between latency versus DL throughput is in the order of 70 for both directions while its only in the range of 30 for latency and uplink throughput in both directions. Hence, observed results show that each relation of latency toward DL and uplink throughput are equally predictable in both directions. However, the future predictability of latency's time series is higher from DL throughput rather than from uplink throughput. This further confirms the direct causal relations obtained between DL throughput and latency versus the uplink throughput and latency using the GES algorithm earlier.

C. Correlation Among Causal Relations

As mentioned in the introduction, correlation is often assumed for causality which can lead to inaccurate actionable results. However, idea here is to observe and quantify correlation for causally related variables. Based on the causal discovery relations identified, an hourly aggregated trending of 4 network variables – DL throughput, uplink throughput, latency, and jitter are shown in Fig. 4 using a double-axis plot. Left y-axis represents throughput of both uplink and DL in kilobits per second (kbps) while the right y-axis represents latency and jitter in milliseconds (msec). The x-axis represent the hour of the day covering entire 24 h of a day. Bar graph representation of throughput shows that DL throughput is relatively much higher than uplink throughput at any given hour of the day. While jitter seems consistent between 3 and

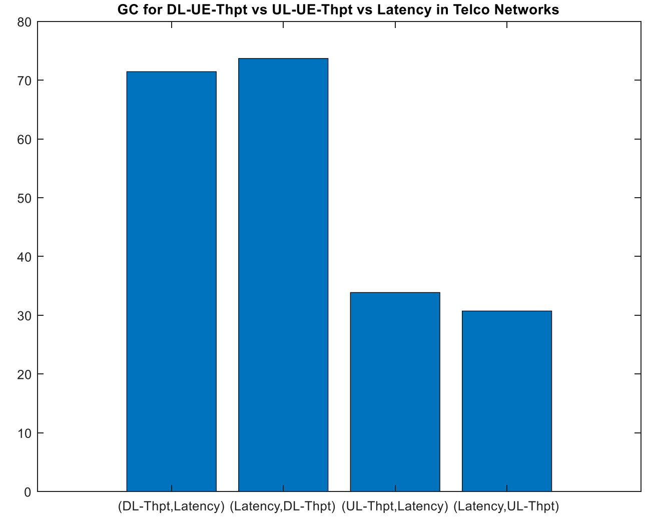


Fig. 3. GC for throughput and latency.

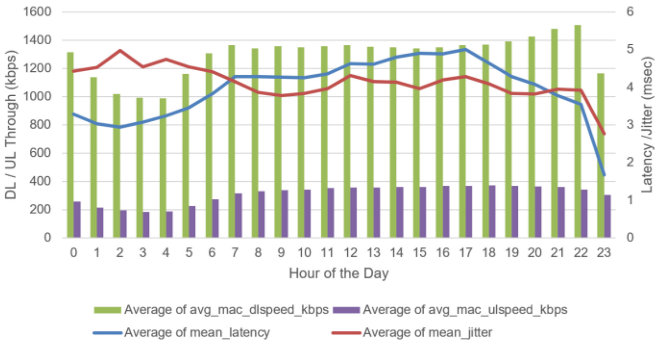


Fig. 4. Hourly trending of four network attributes.

5 ms, latency trend-line seems to be closely following the DL throughput trend throughout the day and illustrating that latency tends to go higher as the DL throughput increase for most hours of the day. When higher data rates are often attributed to lower latency, this seems to be a counter-intuitive observation. Such behavior can be explained by certain network characteristics like network congestion or load utilization.

Higher throughput may attract more users and applications, leading to network congestion. Congestion can result in packet loss and retransmissions, especially when the network infrastructure becomes overwhelmed. Also, higher throughputs often involve more aggressive resource allocation on time, frequency, or spatial domains. If the resources are limited or over-allocated, there may be contention or interference among users, leading to collisions and retransmissions and thus resulting in queuing delays and higher latency.

PM Records provide the average percentage of radio resources being consumed on physical DL shared channel (PDSCH) for non-GBR 5QIs. This data from PM records is obtained for the same day and same geographical area as the initial Session Records data is collected for consistency. Fig. 5 shows both the latency and network utilization plotted together in double-axes view for same hours of the day. A strong correlation can be observed among these network variables.

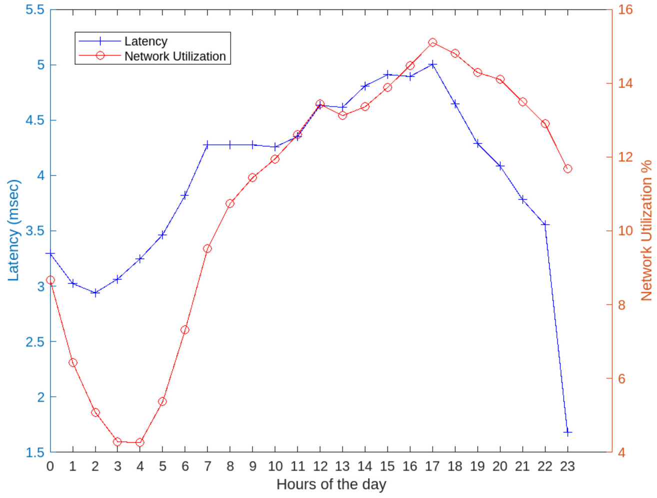


Fig. 5. Hourly trending of latency and network utilization.

A positive correlation coefficient indicates a positive linear relationship, while a negative coefficient indicates a negative linear relationship. A value close to 1 or −1 indicates a strong correlation, while a value close to 0 suggests a weak or no correlation. For the data shown in Fig. 5, correlation coefficient of 0.77 is obtained suggesting a strong correlation among both the network variables and supports the real-world causal discovery analysis so far.

Such in-depth analysis of network metrics propelled by the insights from causal discovery offers a foundation for practical applications in 5G NR network optimization. This pioneering work showcases the extent this causal discovery approach can be applied to real-world mobile networks to gain insights and to optimize it based on the network operator's use case.

IV. Proposed Proportional Fair Scheduler Enhancement

A. NR Scheduler Introduction

Scheduler resides in medium access control (MAC) layer in gNB and it is responsible for managing resource allocation to connected users in real-time across both time and frequency domain. Fig. 6 shows an overview of a typical DL scheduler while a similar approach is applicable to uplink resource allocation as well. Resources are first assigned for the retransmissions, irrespective of the adopted scheduling strategy and so the pending retransmissions feedback is given directly to the scheduler to schedule such UEs first. Once all the retransmissions are scheduled, scheduler is provided with a list of priority UEs to schedule. Such list of priority UEs is obtained as an outcome of the specific scheduler algorithm being used. Choice of a specific scheduling algorithm is usually dependent on the end goal of MNOs, but they generally tend to deploy such that the network resources are optimized while it caters for best end user experience. Scheduler inputs that are fed into scheduler algorithms typically include CQIs, total number of UEs, average data rates, packet delays, queue status, buffer levels, and QoS identifier [19].

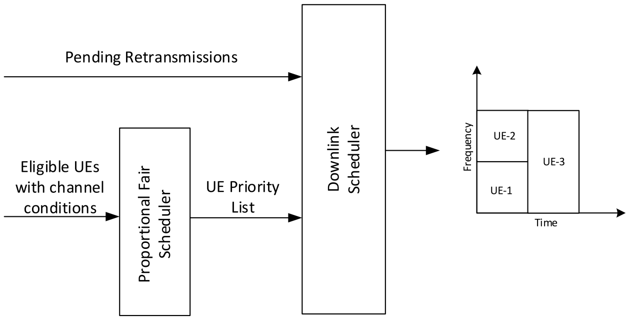


Fig. 6. Overview of DL NR scheduler.

Real-world gNBs from which the 5G NR network data is collected for causal discovery uses PF scheduler that considers channel conditions for each UE in the queue to avoid accumulating data in the buffer. A comparative analysis among four widely used scheduling algorithms, namely round robin (RR), best CQI (BCQI), fractional frequency reuse (FFR), and PF that are applicable to future communication networks has been summarized in [12]. This work highlights that PF scheduler achieves a maximum balance between fairness and maximum cell throughput by illustrating the algorithm's decision-making flow. Enhancements proposed in [15] attempts to overcome a limitation of PF scheduler where it does not consider the QoS for each UE. This enhancement allows to prioritize traffic like ultra reliable low latency communication (URLLC) over enhanced mobile broadband (eMBB) traffic when serving both types.

B. Proposed Scheduler Enhancements

While the current PF scheduler prioritizes retransmissions when scheduling the UEs in buffer and does not account for scheduling weight [10], [26], literature suggests that all the proposed enhancements should consider additional scheduler inputs like QoS requirements when using PF scheduler for new transmissions. Our causal discovery learnings from this work suggests that DL throughput causes latency and scheduling UEs to achieve higher throughput based on their channel conditions and buffer status requirements does not always result in improved latency. To realize an action-based causal approach, we propose a feedback loop of retransmissions per UE back into the scheduler algorithm. This feedback is used as an input for new transmissions to the same UE and influences its scheduling weights to maintain latency requirements.

The specific steps in the proposed PF scheduler are illustrated in the flowchart shown in Fig. 7. At the beginning of each transmission time interval (TTI), the proposed PF scheduler first verifies whether retransmissions (ReTXs) are scheduled. If ReTXs are present, they receive priority for resource allocation based on feedback from prior transmissions. If no ReTXs are scheduled, the scheduler then identifies UEs eligible for new transmissions. CQI inputs are considered for all eligible UEs. The PF scheduler computes scheduling weights by considering CQI, buffer status, and ReTX feedback. UEs are then sorted in descending order according to their scheduling weights. The UE with the highest weight is selected for resource allocation. The scheduler subsequently checks if all UEs have been scheduled. If resource blocks (RBs) remain available, the scheduler continues by allocating these to the next highest-weighted UE. Mathematically, this can be represented in five steps as below.

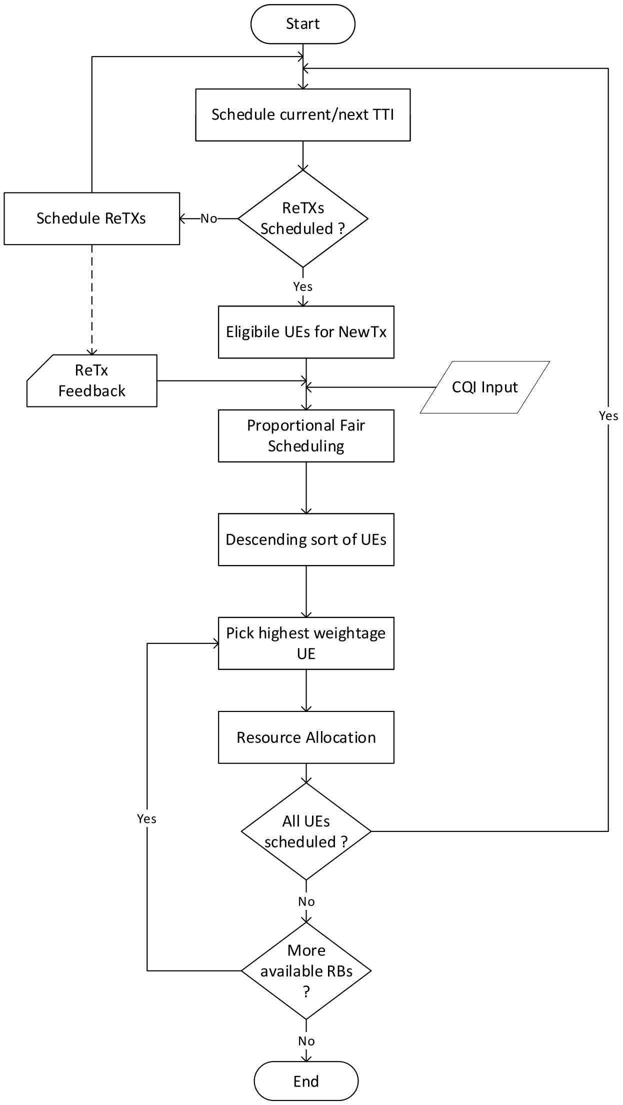


Fig. 7. Proposed PF scheduler enhancements.

1) *Initialization:* For every evaluation, scheduler considers all the remaining resources *(RB)* available after all the pending retransmissions are scheduled. *(UE, RB)* represents the resource allocation matrix with N_{UE} users to schedule.
2) For all the eligible UEs that have pending new transmissions, the newly developed technique computes the weight $w_{i,j}$ corresponding to ith user with jth RB by multiplexing the PF metric to the inverse of retransmission HARQ failure rate (Rx_i) of each user as shown in (2). This allows to consider previously failed transmissions and accounts for it in the current TTI evaluation

$$w_{ij} = \frac{r_{ij}}{R_i} \times \frac{1}{Rx_i} \tag{2}$$

where $r_{i,j}$ is the instantaneous rate and R_i is the average data rate detailed in [26]. Such implementation factors in the buffer requirements along with their past data rates for each UE.

3) All the eligible UEs are ranked according to $w_{i,j}$ weightage metrics saved in the resource allocation matrix *(UE, RB)*. Priority is given to the UE_i having highest metric $w_{i,j}$.
4) For each RB_j, we look at the UE_i with the highest metric across users to assign it the RB_j as shown in

$$RB_j \rightarrow UE\big(\text{argmax}(w_{ij})\big);\ i = 0, 1, 2, \dots, N_{UE}. \quad (3)$$

5) In case there are multiple users that end up having the same weightage, (3) can be updated to (4) to rank UEs according to their priority level by selecting users with lowest priority value

$$RB_j \rightarrow \text{argmin}\big(UE\_\text{priority}(\text{argmax}(w_{ij}))\big)$$
$$i = 0, 1, 2, \dots, N_{UE}. \quad (4)$$

Resource allocation is to be repeated for each symbol until all the available RBs are depleted. By implementing these steps, we realize the benefits of our proposed enhancements to the existing PF scheduler by leveraging the GES algorithm inputs in processing 5G NR network data shows that proposed method can significantly improves latency and throughput due to the adaptive consideration of retransmission feedback and scheduling weights. Further enhancements that can be made to this proposed algorithm based on [8] and [16] are considered out of scope for this work and are covered in Section VI.

Even with the existing PF scheduler, when resources are prioritized for scheduling retransmissions of previously failed transmissions, it can assist in completing these delayed transmissions. However, this approach inadvertently delays the scheduling of other UEs. The PF algorithm does not inherently compensate for UEs that require retransmissions, especially when these UEs have high buffer requirements. Consequently, these UEs may receive disproportionate scheduling priority, impacting their ability to handle both new transmissions and retransmissions. This results in inefficient utilization of network resources and can lead to decreased goodput metrics.

By incorporating specific adjustments to the proposed PF scheduler, such as dynamic buffer-aware prioritization and adaptive resource allocation for retransmissions, we can enhance overall network performance. These adjustments address the bias introduced by high buffer requirements by ensuring that retransmissions are balanced with the needs of other UEs. This results in a more equitable distribution of scheduling resources, leading to improved throughput and better overall goodput metrics. The proposed enhancements optimize resource usage and mitigate delays, thereby improving the efficiency of 5G NR network data scheduling.

C. Simulation Setup

MATLAB 2023b is used for simulation purpose and MathWork's "NR TDD Symbol Based Scheduling Performance Evaluation" example is used to simulate 5G NR network. Every possible effort is made to set the simulation parameters as close as possible (within system limitations) to replicate real-world 5G network from which original data outlined in Section II was obtained.

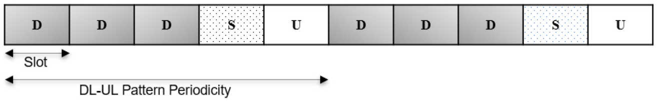


Fig. 8. TDD pattern configuration.

TABLE III
SIMULATION PARAMETERS

Simulation Parameter	***Value***
Number of UEs	15
Application Throughput	1Mbps – 100Mbps
SCS	30kHz
Scheduling Type	Symbol-based
Bandwidth	100 MHz
Carrier Frequency	2.5GHz TDD
Scheduler Type	Proportional Fair & Proposed Enhancement
Modulation	Up to 64QAM
CQI Range	0 - 15

Simulation parameters used are as summarized in Table III, Carrier Frequency of 2.5 GHz is chosen with a bandwidth of 100 MHz and TDD frame patten of three DL slots, one uplink and one special slot as shown in Fig. 8. Special slot has seven DL symbols and five uplink symbols and a guard period of two symbols. Each slot is 1 ms duration while DL-UL Pattern Periodicity is 5ms and subcarrier spacing (SCS) is 30 kHz. A total of 15 UEs are simulated to be active on the gNB with scheduling done at symbol-level granularity. Each UE is randomly located in the varying azimuth and elevation in reference to gNB to simulate a real-world scenario. Also, each UE is set with a varying size application DL throughput ranging between 1 to 100 Mb/s to simulate high loading scenario on the 100 MHz carrier. Simulation settings of other layers are left untouched from Mathwork's original example as they are agnostic to these scheduler changes.

D. Performance Evaluations

The performance evaluation metrics [7] were outlined to compare the resource allocation algorithms while comparing a comprehensive list of all the scheduler algorithms related to mobile networks. Such performance evaluation metrics include delay, throughput, goodput, and spectral efficiency. These metrics are used to quantify the performance of our proposed PF scheduler enhancements in comparison to the current implementation.

As summarized in Table IV, several key takeaways can be highlighted in the comparison of existing PF scheduler with our proposed PF scheduler. The proposed scheduler shows improvements in multiple performance metrics. Specifically, it leads to a 2.86% increase in average cell DL throughput and a more substantial 5.6% boost in average cell DL goodput. Notably, it maintains the same peak DL throughput

TABLE IV
SIMULATION RESULTS

Performance Metrics	***Existing PF Scheduler (Pre)***	***Proposed PF Scheduler (Post)***	***Delta %***
Cell DL Throughput (Mbps)	5.59	5.75	2.86%
Cell DL Goodput (Mbps)	4.99	5.27	5.6%
Peak DL Throughput (Mbps)	15.55	15.55	-
Cell DL spectral efficiency (bits/s/Hz)	1.00	1.06	6%
Peak DL Spectral Efficiency (bits/s/Hz)	3.11	3.11	-

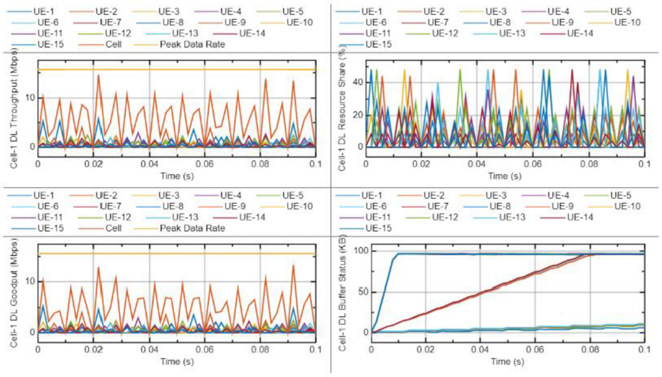


Fig. 9. Simulation output of proposed PF scheduler enhancement showing metrics of DL throughput, DL goodput, resource share, and buffer status for all 15 users over a 100 ms simulation duration.

as the existing scheduler, demonstrating consistency in high-performance capabilities. Additionally, the proposed scheduler achieves a 6% improvement in overall cell DL spectral efficiency. This reflects the efficient utilization of network resources without wasting them on network retransmissions. However, there is no change in peak DL spectral efficiency. Fig. 9 presents the MATLAB simulation output for the proposed PF scheduler, highlighting four key metrics over a 100-ms simulation period. The top-left graph shows the DL throughput in Mbps for individual users (UE-1 to UE-15) and the cell's peak data rate. The peak data rate of the cell remains constant, representing the overall cell capacity, while the average cell throughput is the sum of all 15 users. The data reveals the variability in throughput experienced by different users, indicative of the scheduler's attempt to balance fairness and performance. The top-right graph illustrates the resource share percentage for each user, demonstrating the dynamic allocation of resources among users, which is essential for maintaining fairness in a multiuser environment. This plot showcases the PF aspect of the scheduler by allocating resources without notable bias among a variety of users.

The bottom-left graph displays the DL goodput in Mbps, representing the actual user-perceived data rate after accounting for retransmissions and other losses. The goodput trends closely follow the throughput patterns, with minimal delta between the two, underscoring the efficiency of the proposed scheduler. This minimal delta represents the minimized loss of packets due to retransmissions, which could otherwise impact the end user's perceived experience. Finally, the bottom-right graph depicts the DL buffer status in KB for each user, providing insight into the buffer occupancy and its impact on latency and throughput. The buffer status trends highlight users with higher data requirements and the scheduler's effectiveness in managing buffer occupancy to minimize latency. Certain users have high buffer requirements from the beginning, while another set of users show a gradual rise in buffer requirement, and yet another set have relatively low buffer requirements throughout the simulation. Collectively, these metrics demonstrate the proposed PF scheduler's ability to dynamically allocate resources, maintain fairness, and optimize overall network performance in a 5G NR environment. Since MATLAB does not account for over-the-channel transmission delays, we do not have a way to quantify the impact on latency but it is safe to assume that latency is either maintained or improved with an improved goodput. Overall, these findings suggest that the proposed PF scheduler enhances data throughput and spectral efficiency, which can contribute to improved network performance and user experience.

V. CONCLUSION AND FUTURE WORK

This article introduces a novel extension of causal discovery to real-world 5G NR network data, identifying relationships among key metrics that reflect end-user experience, such as throughput and latency. Such findings allowed to propose a novel enhancement to the PF scheduler algorithm by considering retransmission feedback for calculating scheduling weights. This proposal is demonstrated via simulation to improve both user experience and network resource utilization, including a 2.86% increase in average cell DL throughput and a 5.6% boost in average cell DL goodput, indicating reduced packet loss and retransmissions. It also maintains peak DL throughput and achieves a 6% improvement in overall cell DL spectral efficiency, optimizing both user experience and network resource utilization in a 5G NR environment.

This work can be further evolved to areas like 1) identifying and deriving action-based causal discovery insights among additional network variables; 2) further optimizing PF scheduler for delay sensitive applications beyond eMBB; and 3) causality-based layer management to preserve end user's Quality of Experience (QoE).

ACKNOWLEDGEMENT

Real world network data is from T-Mobile US Inc and adheres to its data governance standards.

REFERENCES

[1] Y. Zheng et al., "Causal-learn: Causal discovery in python," 2023, *arXiv:2307.16405*.

[2] J. Isabona, A. L. Imoize, S. Ojo, D. T. Do, and C. C. Lee, "Machine learning-based GPR with LBFGS kernel parameters selection for optimal throughput mining in 5G wireless networks," *Sustainability*, vol. 15, no. 2, p. 1678, 2023. [Online]. Available: https://doi.org/10.3390/su15021678

[3] A. A. Bachir et al., "Case studies of causal discovery from IT monitoring time series," 2023, *arXiv:2307.15678*.

[4] Y. Wen et al., "Applying causal discovery to single-cell analyses using Causal Cell," *eLife*, vol. 12, May 2023, Art. no. e81464, doi: 10.7554/eLife.81464.

[5] R. K. Polaganga and Q. Liang, "Transfer entropy and granger causality in real-world telecommunication networks," Preprint, 2023. [Online]. Available: https://doi.org/10.21203/rs.3.rs-3444189/v1

[6] T. Vigen. "Spurious correlations." Tylervigen.com. 2023. [Online]. Available: https://tylervigen.com/spurious-correlations

[7] A. Mamane, M. Fattah, M. E. Ghazi, M. E. Bekkali, Y. Balboul, and S. Mazer, "Scheduling algorithms for 5G networks and beyond: Classification and survey," *IEEE Access*, vol. 10, pp. 51643–51661, 2022, doi: 10.1109/ACCESS.2022.3174579.

[8] A. Mamane, M. Fattah, M. El Ghazi, and M. El Bekkali, "Packet delay budget-based scheduling approach for 5G time division duplex," in *Proc. ICDTA*, 2022, pp. 312–321, doi: 10.1007/978-3-031-02447-4_33.

[9] J. Kelly, C. Berzuini, B. Keavney, M. Tomaszewski, and H. Guo, "A review of causal discovery methods for molecular network analysis," *Mol. Genet. Genom. Med.*, vol. 10, no. 10, Oct. 2022, Art. no. e2055, doi: 10.1002/mgg3.2055.

[10] A. Mamane, M. Fattah, M. El Ghazi, Y. Balboul, M. El Bekkali, and S. Mazer, "Proportional fair buffer scheduling algorithm for 5G enhanced mobile broadband," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 5, p. 4165, 2021, doi: 10.11591/ijece.v11i5.pp4165-4173.

[11] J. E. Preciado-Velasco, J. D. Gonzalez-Franco, C. E. Anias-Calderon, J. I. Nieto-Hipolito, and R. Rivera-Rodriguez, "5G/B5G service classification using supervised learning," *Appl. Sci.*, vol. 11, no. 11, p. 4942, 2021, doi: 10.3390/app11114942.

[12] K. Ashfaq, G. Ali Safdar, and M. Ur-Rehman, "Comparative analysis of scheduling algorithms for radio resource allocation in future communication networks," *PeerJ Comput. Sci.*, vol. 7, p. e546, 2021, doi: 10.7717/peerj-cs.546.

[13] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021, doi: 10.1109/JPROC.2021.3058954.

[14] X. Shen, S. Ma, P. Vemuri, and G. Simon, "Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology," *Sci. Rep.*, vol. 10, p. 2975, Feb. 2020, doi: 10.1038/s41598-020-59669-x.

[15] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, 2020, Art. no. 45674–45688. [Online]. Available: https://ieeexplore.ieee.org/document/9020161/

[16] A. Karimi, K. I. Pedersen, and P. Mogensen, "Low complexity centralized multi-cell radio resource allocation for 5G URLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9120469/

[17] J. Pearl and D. Mackenzie, *The Book of Why*. London, U.K.: Penguin Books, 2019.

[18] R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang, "Causal discovery in the presence of missing data," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1762–1770. [Online]. Available: https://proceedings.mlr.press/v89/tu19a.html

[19] A. Mamane, M. E. L. Ghazi, G. R. Barb, and M. Oteşteanu, "5G heterogeneous networks: An overview on radio resource management scheduling schemes," in *Proc. 7th Mediterr. Congr. Telecommun. (CMT)*, 2019, pp. 1–5, doi: 10.1109/CMT.2019.8931369.

[20] T. Verhelst, O. Caelen, J. C. Dewitte, B. Lebichot, and G. Bontempi, "Understanding telecom customer churn with machine learning: From prediction to causal inference," in *Proc. 31st Benelux AI Conf., 28th Belgian-Dutch Mach. Learn. Conf., BENELEARN*, 2020, pp. 182–200, doi: 10.1007/978-3-030-65154-1_11.

[21] C Glymour, K Zhang, and P Spirtes, "Review of causal discovery methods based on graphical models," *Front. Genet.*, vol. 10, p. 54, Jun. 2019, doi: 10.3389/fgene.2019.00524.

[22] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 1551–1560.

[23] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Financial networks based on Granger causality: A case study," *Physica A Stat. Mech. Appl.*, vol. 482, Sep. 2017, pp. 65–73, doi: 10.1016/j.physa.2017.04.046.

[24] A. Seth,"A MATLAB toolbox for Granger causal connectivity analysis," *J. Neurosci. Methods*, vol. 186, pp. 262–273, 2010, doi: 10.1016/j.jneumeth.2009.11.020.

[25] J. Pearl, *Causality*, 2nd ed. Cambridge, MA, USA: Cambridge Univ. Press, doi: 10.1017/CBO9780511803161.

[26] Z. Sun, C. Yin, and G. Yue, "Reduced-complexity proportional fair scheduling for OFDMA systems," in *Proc. Int. Conf. Commun., Circuits Syst. (ICCCAS)*, 2006, pp. 1221–1225. [Online]. Available: http://ieeexplore.ieee.org/document/4064107/

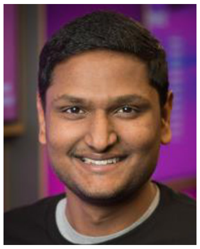
[27] D. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, Mar. 2003, doi: 10.1162/153244303321897717.

[28] P. Spirtes, C. Glymour, and R. Scheines, "Constructing bayesian networks models of gene expression networks from microarray data," in *Proc. Atlant. Symp. Comput. Biol.*, 2000, pp. 1–5.

[29] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 1993.

[30] W. Buntine, "Theory refinement on Bayesian networks," in *Proc. 7th Conf. Uncertain. Artif. Intell.*, 1991, pp. 52–60.

[31] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.



Roopesh Kumar Polaganga (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Pondicherry Engineering College, Puducherry, India, in 2013, the M.S. (Thesis) degree in electrical engineering from the University of Texas at Arlington (UTA), Arlington, TX, USA, in 2015, and the MBA degree from Capella University, Minneapolis, MN, USA, in 2019. He is currently pursuing the Ph.D. degree with UTA.

Since 2015, he has been working as a Principal Systems Architect Engineer with T-Mobile US, Bellevue, WA, USA. While at UTA, he served as a Graduate Research Assistant with the Communication and Networking Lab under the guidance of Dr. Liang, focusing his research on ultra-wideband and LTE technologies. At T-Mobile US Inc., he successfully designed several features and solutions in 5G-NR, LTE/LTE-Advanced, and IoT for everyday customer use. In addition to technology development, he contributed to multiple M&A projects to realize network synergies and improve overall customer experience. He has authored three journal papers and holds over 24 U.S. granted patents. He has received several corporate recognitions within T-Mobile US, including peak nominations within the company. His technical areas of interest include wireless telecommunications, cloud networks, Internet of Things, and AI/ML in telecom networks.

Qilian Liang (Fellow, IEEE) received the B.S. degree in electrical engineering from Wuhan University, Wuhan, China, in 1993, the M.S. degree in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1996, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2000.

He was a Member of Technical Staff with Hughes Network Systems Inc., San Diego, CA. He is a Distinguished University Professor with the Department of Electrical Engineering, University of Texas at Arlington (UTA), Arlington, TX, USA. He has authored or co-authored over 300 journals and conference papers, seven book chapters, and has six U.S. patents pending. His current research interests include radar sensor networks, wireless sensor networks, wireless communications, compressive sensing, smart grids, signal processing for communications, and fuzzy logic systems and applications.

Dr. Liang was a recipient of the 2002 IEEE Transactions on Fuzzy Systems outstanding Paper Award, the 2003 U.S. Office of Naval Research Young Investigator Award, the 2005 UTA College of Engineering Outstanding Young Faculty Award, the 2007, 2009, and 2010 U.S. Air Force Summer Faculty Fellowship Program Award, the 2012 UTA College of Engineering Excellence in Research Award, and the 2013 UTA Outstanding Research Achievement Award. He was inducted into the UTA Academy of Distinguished Scholars in 2015.