©SHUTTERSTOCK.COM/JOZSEF BAGOTA

ARTIFICIAL INTELLIGENCE-EMPOWERED HYBRID MULTIPLE-INPUT/MULTIPLE-OUTPUT BEAMFORMING

Learning to Optimize for High-Throughput Scalable MIMO

Nir Shlezinger, Mengyuan Ma, Ortal Lavi, Nhan Thanh Nguyen, Yonina C. Eldar, and Markku Juntti

Hybrid beamforming for multiple-input/multiple-output (MIMO) communications is an attractive technology for realizing extremely massive MIMO systems envisioned for future wireless communications in a scalable and power-efficient manner. However, the fact that hybrid MIMO systems implement part of their beamforming in analog and part in digital makes the optimization of their beampattern notably more challenging compared with conventional fully digital MIMO. Consequently, recent years have witnessed growing interest in using data-aided artificial intelligence (AI) tools for hybrid beamforming design. This article reviews candidate strategies to leverage data to improve real-time hybrid beamforming design. We discuss the architectural constraints and characterize the core challenges associated with hybrid beamforming optimization. We then present how these challenges are treated via conventional optimization,

Digital Object Identifier 10.1109/MVT.2024.3396927
Date of publication 20 May 2024; date of current version 18 September 2024.

and identify different AI-aided design approaches. These can be roughly divided into purely data-driven deep learning models and different forms of deep unfolding techniques for combining AI with classical optimization. We provide a systematic comparative study between existing approaches, including both numerical evaluations and qualitative measures. We conclude by presenting future research opportunities associated with the incorporation of AI in hybrid MIMO systems.

Introduction

Massive MIMO systems and high-frequency communications at millimeter wave (mmWave) and subterahertz bands are expected to play a key role in future 6G networks [1]. These technologies are naturally supportive of each other, as massive MIMO using large transmit and receive antenna arrays facilitates generating highly focused beams that are essential for reliable communications at high frequencies, while short wavelength signaling enables packing MIMO configurations with a massive number of elements in a limited aperture. However, implementing such massive MIMO transceivers gives rise to several challenges. One of these core challenges is associated with the notable cost and power consumption of radio-frequency (RF) chains operating at high frequencies, in which conventional fully digital MIMO arrays separately connect each antenna element to a digital signal processing unit.

Hybrid beamforming is considered a leading solution for coping with the above challenge, enabling high-frequency massive MIMO communications with a limited number of RF chains [2]. This is achieved by delegating part of the signal processing to the analog domain, thus dividing the beamforming task into digital and analog counterparts. The possible beampatterns achievable in analog are dictated by the circuitry, with typical implementations based on phase shifters [3], vector modulators [4], and dynamic metasurface antennas (DMAs) [5]. Consequently, hybrid transceivers are inherently constrained in their beamforming capabilities compared with fully digital ones.

While hybrid designs alleviate some of the cost and power issues of massive MIMO systems, their constrained form gives rise to algorithmic and signal processing challenges. Most notably, the beamforming task—i.e., the translation of channel state information (CSI) into a suitable beampattern—involves solving a typically nonconvex constrained optimization problem. Various iterative optimization algorithms have been proposed for tuning hybrid beamformers [6], differing in their considered hardware constraints and objective. A key limitation of these iterative solutions stems from their typically slow convergence, as the beampattern setting must be performed in real time to cope with channel variations.

The emergence of deep learning as an enabler technology for AI has led to the proposal of AI-empowered hybrid beamforming designs. While deep learning typically deals with setting an inference rule based on data, one can also train deep neural networks (DNNs) to tackle challenging optimization problems [7]. Once trained, DNNs infer at fixed latency, dictated by the number of layers, and can thus be used to rapidly map CSI into beampatterns [8]. An alternative approach to leverage data for hybrid beamforming arises from model-based deep learning methodologies [9]. Here, deep learning techniques are used to enhance iterative hybrid beamforming optimizers rather than replacing them, while data are exploited to achieve rapid convergence [10], [11], [12]. The proliferation of different approaches for hybrid MIMO beamforming motivates a unified overview of these methods.

HYBRID BEAMFORMING IS CONSIDERED A LEADING SOLUTION FOR COPING WITH THE ABOVE CHALLENGE, ENABLING HIGH-FREQUENCY MASSIVE MIMO COMMUNICATIONS WITH A LIMITED NUMBER OF RF CHAINS.

In this article, we provide a systematic tutorial of AI-aided methodologies for hybrid MIMO beamforming. While successfully realizing hybrid MIMO transceivers inevitably combines hardware developments with signal processing algorithmic considerations, this work is concerned with the latter, without restricting our attention to a specific implementation. As opposed to previous works reviewing beamforming and AI (e.g., [13] and [14]), here we particularly focus on the design of AI systems for hybrid beamforming and their relationship with optimization methods.

We start by discussing hybrid MIMO systems, reviewing representative architectures and describing how their operation impacts the achievable beampatterns. We pinpoint the design challenges arising from hybrid beamforming, and identify the aspects that motivate incorporating AI. Next, we describe hybrid beamforming design approaches, dividing them into three main families: *optimization-based* methods, which employ iterative optimizers for setting the beampatterns; *DNN-based* schemes, where CSI is mapped into hybrid configurations via a pretrained DNN; and *deep-unfolded* designs, where deep learning techniques are leveraged to facilitate iterative optimization. For the latter, we identify different types of unfolding approaches and discuss how each gives rise to a different design. Based on this division, we provide a comparative study, including both a numerical study and a qualitative comparison, where we identify the interplay between the approaches in terms of several key figures-of-merit. We conclude by discussing research challenges that are left for future exploration, and are expected to pave the way toward harnessing the potential of AI for hybrid MIMO systems.

Hybrid Beamforming

Hybrid MIMO Transceivers

Massive MIMO transceivers are equipped with an antenna array comprised of a large number of elements, denoted M. In the current 5G base stations, M can be on the order of several tens. This number is expected to grow to possibly thousands of antennas in 6G, when evolving from massive MIMO to holographic MIMO [1]. In conventional fully digital MIMO architectures, the signal being fed to each antenna is processed separately digitally, by having a digital processing unit connect to each antenna via a dedicated RF chain.

In hybrid MIMO systems, the number of RF chains, denoted K, is smaller than that of antennas. This is achieved via analog processing that interfaces the RF chains with the antennas, as illustrated in Figure 1. The analog processor achieves different manipulations of the signals. A natural benefit of hybrid MIMO over fully digital architectures stems from the fact that it uses fewer RF chains than antenna elements, which becomes a crucial factor when using large-scale arrays in high frequencies. In addition to reducing RF chains, hybrid designs can also facilitate interference rejection as well as mitigate distortion induced by low-resolution analog-to-digital convertors [4].

Architectural Constraints

Hybrid MIMO systems combine digital and analog signal processing. The processing part carried out in the digital domain is highly flexible, allowing it to effectively apply different mappings to different spectral components. However, analog processing is highly constrained, and the set of different mappings it can realize is dictated by its hardware, with several different hardware architectures proposed in the literature. To exemplify the constraints associated with different designs, we briefly review a few representative analog architectures, focusing on their operation in transmission:

Phase Shifter Networks

The most commonly considered analog hardware employs phase shifters with controllable phases [2]. These are typically divided into fully connected architectures, where a dedicated phase shifter connects each RF chain with each antenna, and to partially connected structures, in which each RF chain is connected to a single antenna via a dedicated phase shifter. Often in practice, the phases applied by each phase shifter cannot be arbitrarily set and must comply to some predefined phase resolution. Furthermore, phase shifters are typically designed to (approximately) preserve the same phase shift over a considered band. Thus, they are often modeled as applying the same mapping to each spectral component.

Discrete Vector Modulators

While phase shifters only affect the phase of the signal, vector modulators are analog circuits that can realize

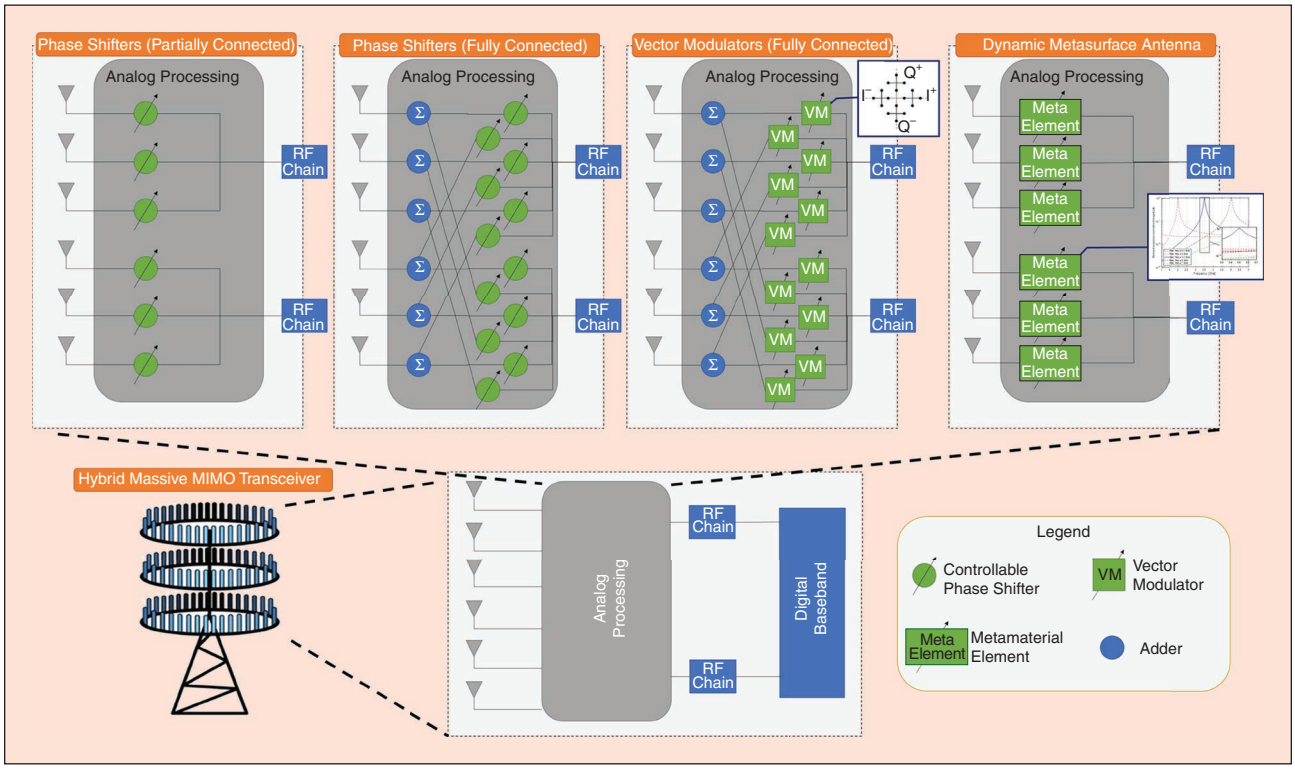

FIGURE 1 Schematic illustration of different hybrid MIMO transceiver architectures and their corresponding analog processing model, including partially and fully connected phase shifter networks, vector modulators, and DMAs.

different combinations of phase shifting and signal attenuation. Such forms of analog circuitry provide additional flexibility compared with phase shifters, due to the ability to also affect the magnitude of the signals in a controllable fashion. Nonetheless, low-power vector modulators are typically constrained to realize only a predefined finite number of different phase-attenuation combinations [4].

DMAs

An emerging technology for realizing holographic MIMO designs antennas using metasurfaces that are planar configurations of controllable metamaterial elements. Unlike the aforementioned architectures, which rely on the incorporation of dedicated analog circuitry, DMAs implement configurable analog processing as an inherent byproduct of their antenna structure [5]. When transmitting, the signal at the output of each RF chain propagates along a waveguide, and is radiated from the elements connected to that waveguide, where each element can realize a form of a frequency-selective Lorentzian filter. Consequently, DMAs inherently implement frequency-selective analog signal processing, which is constrained to take the Lorentzian form.

Hybrid Beamforming Design Challenges

Hybrid beamforming design is concerned with the joint setting of the analog and digital processing to optimize a predefined communication metric for the current channel realization. Typical metrics are the achievable rate or the minimal signal-to-interference-and-noise ratio (SINR) in multiuser communications. Focusing on downlink transmission with the common setting of linear beamforming, the task boils down to designing the precoders applied to the outgoing symbols in digital (where each spectral component can be precoded separately), along with the configuration of the analog processing.

Hybrid beamforming design is associated with multiple core challenges, including:

- *Core challenge 1 (C1):* The resulting optimization problem based on which the digital precoders and the analog configuration are determined is rarely convex. Even when the design objective takes a quadratic form, e.g., the achievable rate of a linear Gaussian channel, the need to divide the processing into digital and analog parts, as well as the hardware constraints imposed on the analog processing, typically results in nonconvex optimization.
- *Core challenge 2 (C2):* Since hybrid beamforming is designed for a given channel realization, it needs to be carried out each time the channel conditions change, i.e., on each coherence duration (which can be as small as 125 μs by 3GPP Release 17). As the coherence duration of wireless communication channels typically decreases with carrier frequency, the design procedure must be performed rapidly to enable reliable communications within each coherence duration.

HYBRID BEAMFORMING DESIGN IS CONCERNED WITH THE JOINT SETTING OF THE ANALOG AND DIGITAL PROCESSING TO OPTIMIZE A PREDEFINED COMMUNICATION METRIC FOR THE CURRENT CHANNEL REALIZATION.

- *Core challenge 3 (C3):* Hybrid beamforming design uses CSI, which is typically obtained from pilot signaling, and is thus likely to be noisy. Consequently, hybrid beamforming design should be able to cope with some level of error in its available CSI.

The above challenges, and particularly C1 and C2, motivate AI-aided designs, as discussed in the following section.

AI-Aided Hybrid Beamforming Design

We next detail leading frameworks for designing hybrid precoders. The first utilizes iterative optimizers that are specific to the problem at hand. The second employs DNNs, i.e., abstract architectures that are tuned from data to map CSI into a hybrid beamformer configuration. The last framework utilizes deep unfolding, which combines iterative optimization with deep learning via different forms of model-based deep learning [9]. The latter constitutes a middle ground between the first two techniques by balancing specificity and data-driven learning capabilities, as illustrated in Figure 2.

Optimization-Based Hybrid Beamforming

As explained earlier, hybrid beamforming design is inherently an optimization problem. As such, it is traditionally tackled using optimization tools, commonly via iterative solvers. Broadly speaking, there are two main approaches to cope with the nonconvexity (C1):

- A leading approach applies convex relaxation, i.e., formulates an alternative problem that is convex. Most commonly, the nonconvex sum-rate objective in the hybrid configuration is often replaced with seeking the hybrid setting that best approximates the fully digital rate-maximizing precoder [3], [15]. Compared to directly maximizing the sum-rate, the relaxed formulation is often simpler to tackle, typically using iterative methods based on alternating optimization. Yet, it may still result in a nonconvex formulation, depending on the hardware constraints. The resulting solution can be shown to approach the rate-maximizing setting in some regimes, and particularly when the number of RF chains K is not smaller than the number of receive antennas [3].
- An alternative approach directly tackles the nonconvex objective, typically by aiming to identify a suitable initial setting of the precoders and refine it using local-convex optimization techniques, e.g., projected gradient ascent (PGA) [10].

While iterative optimizers can often recover useful hybrid beamformers, they tend to require a large number of iterations to converge. As iterations are translated into delay and complexity, this property limits their applicability in time-varying settings by C2. While optimization theory provides techniques for reducing the number of iterations via, e.g., backtracking, such techniques involve additional lengthy computations during inference.

DNN-Based Hybrid Beamforming

Deep learning provides tools for tuning machine learning models parameterized as DNNs to learn a desirable complex mapping from data. DNNs can also be trained to tackle challenging optimization problems, such as those encountered in hybrid beamforming design [7]. Architectures, such as convolutional neural networks (CNNs), were shown to be capable of learning to map MIMO CSI into analog and digital precoders [8].

DNN-based inference rules are typically designed in a supervised manner, i.e., by providing data comprised of inputs and their desired outputs, which the model learns to produce during training. However, for hybrid beamforming, they are often trained unsupervised—namely, by providing a dataset comprised solely of channel realizations—without specifying the desired beamformer for each channel. This is possible because the optimization objective, e.g., sum-rate or SINR, can be evaluated for each selected precoders, while being differentiable with respect to them. Consequently, one may apply conventional gradient-based learning to training DNN-based hybrid beamformers using the (negative) optimization objective as an unsupervised training loss [10].

DNNs are often computationally complex, comprised of a large number of parameters, and their training can be lengthy. Yet, their latency during inference is fixed based on the number of layers, and various software and hardware tools facilitate their parallelization. Consequently, using pretrained DNNs for hybrid beamforming design is often more rapid compared with iterative optimizers. However, the usage of generic highly parameterized models trained from data to replace optimization solvers gives rise to several drawbacks. First, the training of DNNs is often a lengthy task, requiring large volumes of data (i.e., channel realizations) and tedious experimentation to learn a suitable mapping. Furthermore, while their inference latency is fixed, the complexity of applying DNNs in terms of, e.g., floating point operations, is typically large compared with iterative optimizers, being dictated by the number of parameters. Moreover, DNNs are far less flexible compared with optimization methods, and each modification in the task, e.g., the incorporation of an additional user to the network, requires time-consuming retraining. Finally, DNNs are hardly interpretable, in the sense that one can assign operational meaning only to their input and their output, and are typically treated as black boxes.

Deep Unfolded Hybrid Beamforming

Both principled iterative algorithms and data-driven deep learning models possess inherent limitations for hybrid beamforming. This motivates designs that are both model-based and data-driven, capitalizing on the strengths of each approach. Model-based deep learning [9] offers a promising avenue for combining principled

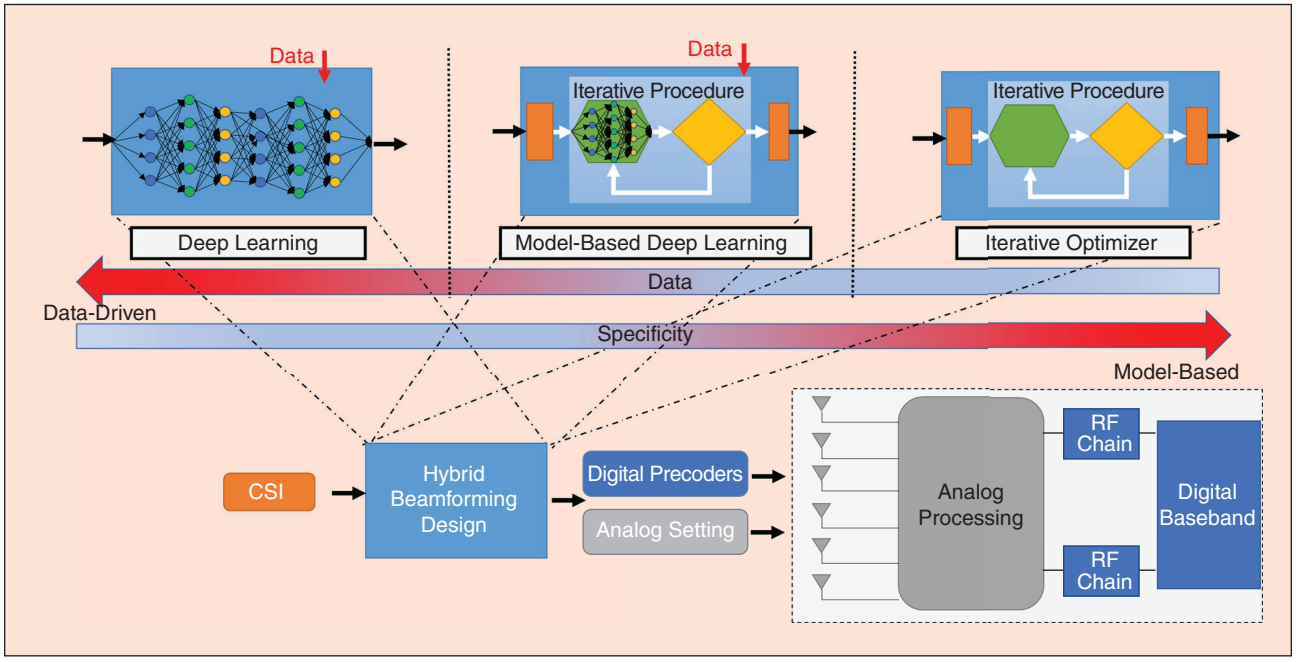

FIGURE 2 Illustration of different approaches for hybrid beamforming design.

mathematical models with deep learning techniques. Among model-based deep learning approaches, deep unfolding is highly suitable for tasks that typically involving iterative methods.

A deep unfolding model is constructed based on the similarity between the sequential operation of an iterative optimizer with L iterations and the forward path of a DNN with L layers. The fundamental idea behind this approach is to treat the iterations of a principled iterative algorithm as an inductive bias of a machine learning model. Consequently, the optimizer with L iterations in transformed into a trainable discriminative model with fixed complexity. This gives rise to three different forms of deep unfolded optimizers [9]:

- *Learned hyperparameters (U1):* The efficiency of iterative optimizers depends on hyperparameters, such as step sizes in projected gradient ascent/descent methods, which are often tuned manually. With a substantial number of iterations, these optimizers can converge to a good objective value and reliable solutions. However, when constrained to a fixed (and limited) number of iterations, the choice of hyperparameters becomes crucial for achieving good performance. Deep unfolding offers a solution by converting the hyperparameters of the iterative solver into trainable parameters and leverage data to train and tune them within a predefined number of iterations.
- *Learned objective (U2):* Iterative optimizers are designed to optimize an objective function, e.g., the system sum-rate, through updating the solution over successive iterations until a convergence criterion is met. Deep unfolding contributes to this procedure via parameterizing the objective function employed in each iteration. This enables learning from data to tune the intermediate solutions based on a different objective, such that the final output is most suitable for the original objective.
- *DNN conversion (U3):* In the third deep unfolding technique, a DNN is designed to facilitate the operation of an iterative optimizer. This is typically accomplished by replacing certain operations, especially those involving computationally intensive tasks like matrix inverse or decomposition, in each iteration with trainable layers. This technique enables varying levels of abstraction. One approach is to maintain the core operation of the iterative optimizer by replacing only specific computations with trainable layers. Alternatively, a highly parameterized DNN can be designed, inspiring by the operation of the original principled optimizer.

In hybrid beamforming, deep unfolded models share the ability of DNNs to train unsupervised. Furthermore, the similarity between the unfolded architecture and iterative optimizers introduces additional factors that can facilitate training. First, the conventional solution acquired through iterative optimizers can serve as the input of the unfolded model. This allows the model to commence with a reliable solution and further improve it over layers though the training process. Moreover, in unfolding models, the output of each layer is associated with the optimization variable, suggesting that the training loss cannot solely be based on the final output after L iterations/layers, as in conventional DNNs, but can also account for the intermediate features in hidden layers. Such training losses, which are not applicable in black-box architectures, encourage the model to produce valid settings at each iteration/layer, and thus constitute a regularization known to facilitate learning.

The above methodologies, especially U1 (e.g., [10] and [11]) and U3 (e.g., [12]), ensure fixed latency hybrid beamforming designs. Specifically, deep unfolding with learned hyperparameters fully preserves the operation of the iterative optimizer, maintaining its flexibility and interpretability. Nonetheless, by learning different step sizes for each iteration [10], and even for the optimized precoder variable [11], notably latency reduction can be attained. Furthermore, the learned hyperparameters can be incorporated into multiobjective designs, such as the robust optimization for coping with CSI uncertainty C3 [10].

Comparative Study

In this section we compare the hybrid beamforming approaches detailed earlier. We present a numerical study comparing representative schemes from each design approach, after which we provide a qualitative comparison.

Numerical Evaluation

To compare the considered hybrid beamforming approaches, we simulate hybrid MIMO systems with fully connected phase shifter network for analog processing. (The source code is available at https://github.com/ortalagiv/AI-Empowered-Hybrid-MIMO-Beamforming.) While hybrid beamforming can be carried out on both the transmit and the receive side, we focus our evaluation on multiuser downlink systems, where a base station equipped with a hybrid antenna array transmits to multiple single antenna users. We compare the following methods for determining the precoders:

- For optimization-based methods, we evaluate the Riemannian manifold optimizer of [3] and the alternating optimizer of [15], which are both based on convex relaxation of the sum-rate objective.
- For DNN-based designs, we use a CNN following the architecture of [8], referred to as *black-box CNN*. This architecture is comprised of three convolutional layers (with 3×3 kernel) followed by three fully connected layers. The CNN was trained to produce both the analog and digital precoders, as well to produce only the analog precoder, while the digital precoder was tuned accordingly to best match the fully digital beamformer. As both implementations yielded similar results, only the latter is reported here.

- For unfolded optimizers, we consider both the ManNet model of [11], that unfolds the convex-relaxed optimization, as well as the unfolded PGA of [10], which augments simple PGA steps applied to the nonconvex sum-rate objective. Both of these unfolded methods use merely 10 iterations while preserving the operation of the iterative optimizers from which they originate following U1.
- To represent an upper bound on the achievable sum-rate, we evaluate that achieved using fully digital beamforming.

In Figures 3–5, the considered MIMO transmitter has $M = 12$ antennas, and serves four single-antenna users by signaling over 16 frequency bins. For training the deep learning models, we generated a dataset of 1,000 mmWave channel realizations with central frequency of 30 GHz using the QuaDRiGa model.

We first set the number of RF chains to $K = 4$. The resulting sum-rates versus signal-to-noise ratio (SNR), depicted in Figure 3, demonstrate that all optimizers based on convex relaxation—i.e., the iterative optimizers of [3] and [15] and the AI-aided ManNet [11]—approach the sum-rate of fully digital beamforming. The black-box CNN and the unfolded PGA are both within a small gap from fully digital beamforming. Nonetheless, the gains of the unfolded designs over purely optimization-based methods are revealed when observing the number of iterations needed to achieve this performance. The sum-rate versus iteration for each iterative method at SNR of 10 dB is reported in Figure 4. There, we observe that the unfolded methods achieve their suitable settings with much less iterations compared with conventional iterative optimizers, indicating the ability of AI-aided designs in notably reducing latency and computational complexity.

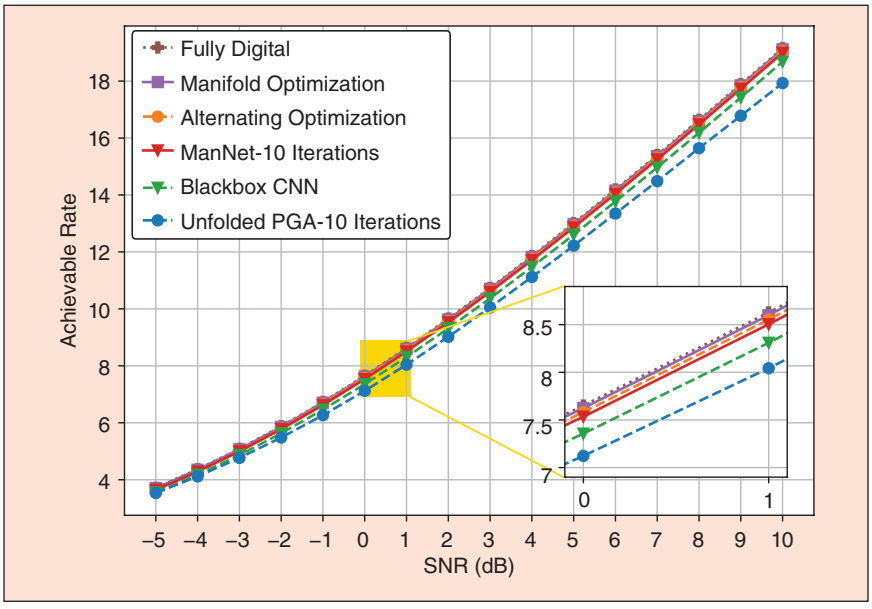

FIGURE 3 Sum-rate versus SNR with 12 antennas and four RF chains.

We next set the number of RF chains to $K = 2$, i.e., less than the number of users, indicating a challenging regime for hybrid beamforming. The results, reported in Figure 5, demonstrate that the gap in this setting between the fully digital precoder and the hybrid beamformers is more dominant compared to that in Figure 3. While both the unfolded and optimization-based methods achieve approximately the same performance, the unfolded schemes do it with much fewer iterations, with an reduction by a factor of 5 to 8× compared with the iterative optimizers of [3] and [15].

While the results in Figures 3–5, consider a MIMO setting without a large number of antennas. In Figure 6 we consider a massive MIMO system with $M = 128$ transmit antennas, $K = 2$ RF chains, and 128 frequency bins, based on the channel model detailed in [11]. It is observed that compared to Figure 3, the hybrid beamforming schemes have more significant performance loss with respect to the fully digital one. Among the compared hybrid beamformers, the deep unfolded optimizers are superior, with ManNet offering the best performance. The black-box DNN, which struggles in learning such complicated tasks from limited datasets, performs far worse than the conventional optimizers and deep unfolding schemes.

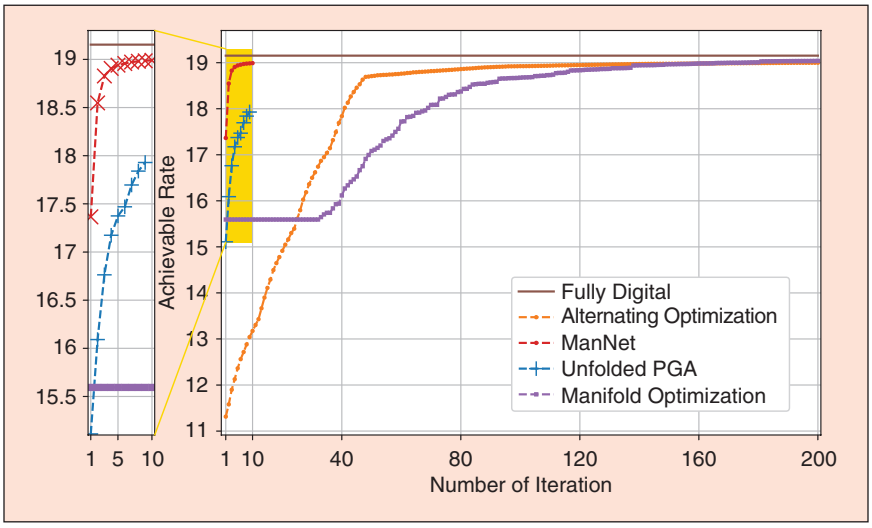

FIGURE 4 Sum-rate per iteration with 12 antennas and four RF chains.

Qualitative Comparison

The approaches detailed earlier for optimizing hybrid beamformers are each suitable for different types of scenarios. The above numerical study allows to compare the approaches in achievable rate. To evaluate additional meaningful comparative aspects, we next discuss five key figures-of-merit: design latency, computational complexity, data requirements, flexibility, and interpretability. The comparison detailed below is summarized in Table 1.

Latency

A core challenge in hybrid beamforming is the need to update the beampattern on each coherence duration C2. Conventional iterative optimizers are typically lengthy, inducing notable latency due to their multiple iterations. This can be mitigated via deep unfolding, particularly via hyperparameter learning U1, as demonstrated in Figure 4. Using DNNs for hybrid beamforming design typically has low latency, as computing the forward pass of a neural network with several layers is of fixed delay, which is reduced with parallelization and hardware accelerators, though not necessarily to the order of the coherence duration of wireless channels.

Complexity

While DNNs often support rapid and fixed-latency hybrid beamforming design, they are computationally complex, being comprised of a large number of parameters, and their limited latency is typically due to parallelization and hardware acceleration. Iterative optimizers are of a much smaller complexity, as each iteration typically involves a small number of operations, yet this complexity is not translated into low latency due to their sequential operation. Deep unfolded designs, particularly with learned hyperparameters U1, share both the low complexity of iterative optimizers while supporting rapid inference due to their inherently fixed number of iterations. We refer readers to [11] for a detailed complexity analysis and comparison of the unfolded PGA, ManNet, black-box CNN, and alternating and manifold optimization methods.

Data

AI-aided hybrid beamforming design leverages data to learn how to map CSI into hybrid precoders. While such learning can be done in an unsupervised manner, training DNNs for such tasks still requires large volumes of data, i.e., channel realizations from the same distribution as that expected at deployment. Deep unfolding balances the dependence on data by imposing an inductive bias on the learned model, trading parameterization for specificity [9], with abstract parameterizations (U3) requiring more data compared with lesser parameterized models (U1).

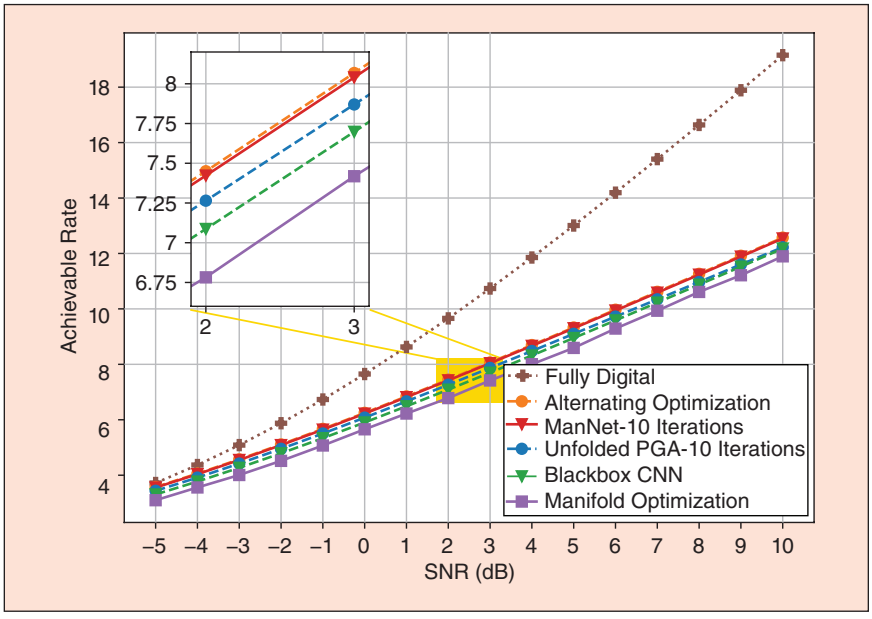

FIGURE 5 Sum-rate versus SNR with 12 antennas and two RF chains.

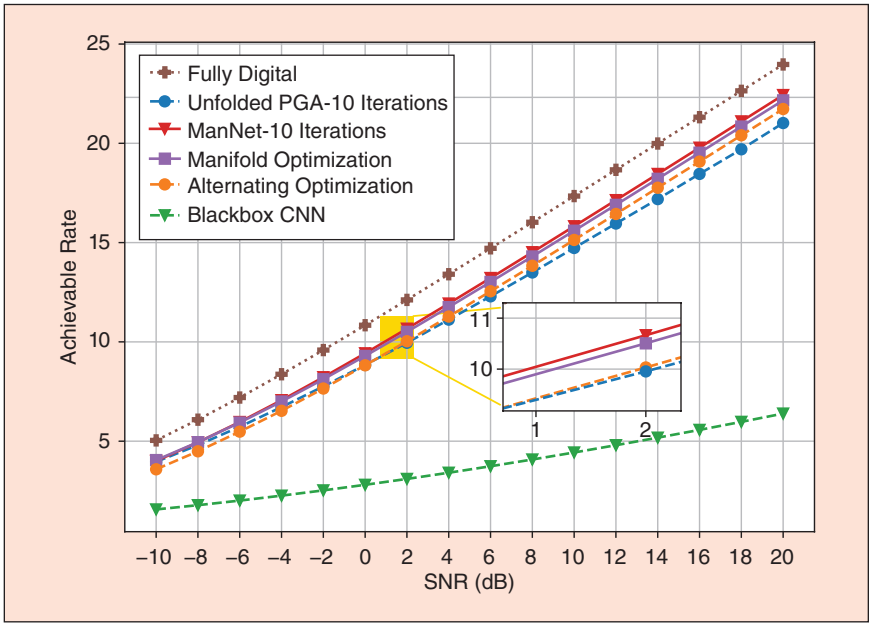

FIGURE 6 Sum-rate versus SNR with 128 antennas and two RF chains.

AN IMPORTANT PROPERTY OF HYBRID BEAMFORMING DESIGN IS THE ABILITY TO UNDERSTAND HOW IT MAPS THE CSI INTO A HYBRID PRECODER, AND TO TRACK ITS PROCESSING CHAIN.

Flexibility

Hybrid beamforming design requires some level of flexibility, as channel configuration, e.g., the number of users, can change over time. Iterative optimizers are extremely flexible, and the same optimizer can be applied in different settings. Similarly, unfolded methods that fully preserve the iterative optimizer (U1) operation also share this flexibility. However, conventional black-box architectures, such as DNNs and CNNs, are trained for a fixed configuration, and are thus highly nonflexible as they have to be retrained when the configuration changes over time.

Interpretability

An important property of hybrid beamforming design is the ability to understand how it maps the CSI into a hybrid precoder, and to track its processing chain. Iterative optimizers are fully interpretable, and so are unfolded optimizers that do not alter their operation (U1). More abstract forms of unfolding that deviate from the optimizer (U3) are less interpretable, yet one can still track their procedure as each iteration is still associated with an operational meaning. For black-box DNNs, only the input and output have an interpretable value.

Summary and Future Research Directions

AI-aided design and model-based deep learning bear the potential of notably facilitating real-time high-throughput hybrid beamforming, which in turn can pave the way toward sustainable and scalable massive MIMO deployments. However, several research directions are to be explored to fully realize the potential of AI-aided beamforming. We next review some candidate topics.

Hybrid MIMO With Integrated Sensing

The 6G networks are envisioned to utilize MIMO transceivers not solely for communications, but also for sensing. Such operation induces various considerations on beamforming design, ranging from coexistence between sensing and communicating spectrum-sharing devices to dual-function signaling. These considerations notably complicate the setting of hybrid beamforming, as the optimization procedure has to account for additional aspects associated with the sensing functionality. This further motivates the exploration of AI-aided techniques for hybrid MIMO with integrated sensing.

Power and Hardware Oriented Designs

While the majority of studies on hybrid MIMO consider phase shifter-based analog circuitry, there are in fact various forms of hybrid architectures, each giving rise to different constraints affecting beamforming design. Furthermore, existing hybrid beamforming methods often overlook the fact that different configurations of the analog circuitry consume different power. For instance, the ability to turn off a subset of the vector modulators in hybrid designs was shown to notably reduce power consumption [4]. This motivates the exploration of hybrid beamforming algorithms that incorporate power and hardware considerations into their optimization procedure, and the associated excessive complexity motivates the usage of the advocated AI-aided strategies.

Distributed Hybrid MIMO

Future wireless communications are expected to deviate from conventional cellular architectures, utilizing multiconnectivity and cell-free topologies [1]. This operation

TABLE 1 *Qualitative comparison between the considered approaches for hybrid beamforming.*

Method	*Latency*	*Complexity*	*Data*	*Flexibility*	*Interpretability*
Iterative optimizers	High: numerous iterations	Low: few operations in numerous iterations	None: no data needed	Fully flexible: applicable with different configurations	Fully interpretable
DNNs	Medium: fixed by forward pass of DNN	High: complex high parameterized models	High: massive datasets needed for training	None: retraining is needed to switch configuration	Not interpretable
Deep unfolded optimizers, U1	Lowest: few predefined iterations of low complexity	Lowest: few operations in few iterations	Low: few parameters trained with small datasets	Flexible: applicable with different configurations though performance may be affected	Fully interpretable: preserve operation as iterative optimizers
Deep unfolded optimizers, U3	Low: few predefined iterations with moderate complexity	Medium: complex parameterized mappings in few iterations	Medium: relatively large number of parameters to train	None: retraining typically is needed to switch configuration	Partially interpretable as one can track intermediate features

extends conventional centralized beamforming into distributed beamforming using a deployment of multiple collaborative MIMO transmitters. The reduced cost of hybrid architectures makes them suitable candidates for massive deployments. The usage of AI in such cases can notably facilitate real-time collaborative hybrid beamforming setting, possibly exploiting distributed machine learning paradigms, such as federated learning and multiagent reinforcement learning.

From Far-Field to Near-Field

An additional consideration impacting beamforming in future wireless communications is the expected transition from far-field communications to near-field. This brings forth new forms of beamforming, as the ability to generate focused beams that can notably mitigate interference. Initial studies have unveiled that focused beams can also be achieved with different forms of hybrid beamforming using lengthy optimization. Future studies are left to explore the ability to simultaneously support far-field and near-field users, and the ability of AI-aided hybrid beamforming in enabling real-time and accurate forming of focused beampatterns for near-field communications.

Author Information

Nir Shlezinger (nirshl@bgu.ac.il) is an assistant professor in the School of Electrical and Computer Engineering in Ben-Gurion University, 84105 Be'er Sheva, Israel. He is a Senior Member of IEEE.

Mengyuan Ma (mengyuan.ma@oulu.fi) is a Ph.D. student in the Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland. He is a Student Member of IEEE.

Ortal Lavi (agivo@post.bgu.ac.il) is a graduate student in the School of Electrical and Computer Engineering in Ben-Gurion University, 84105 Be'er Sheva, Israel. She is a Student Member of IEEE.

Nhan Thanh Nguyen (nhan.nguyen@oulu.fi) is a post-doctoral researcher with the University of Oulu, 90570 Oulu, Finland. He received a Ph.D. degree from Seoul National University of Science and Technology. He is a Member of IEEE.

Yonina C. Eldar (yonina.eldar@weizmann.ac.il) is a professor in the Department of Math and Computer Science, Weizmann Institute of Science, 7610001 Rehovot, Israel, where she heads the Center for Biomedical Engineering and Signal Processing. She is a member of the Israel Academy of Sciences and Humanities, a Fellow of IEEE, and a fellow of European Association for Signal Processing.

Markku Juntti (markku.juntti@oulu.fi) is a professor and head of the Centre for Wireless Communications, Radio Technologies Research Unit, University of Oulu, 90570 Oulu, Finland. He is also an adjunct professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA. He received a Dr.Sc. degree from the University of Oulu, Finland. He is a Fellow of IEEE.

References

[1] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020, doi: 10.1109/MCOM.001.1900411.

[2] A. F. Molisch et al., "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017, doi: 10.1109/MCOM.2017.1600400.

[3] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016, doi: 10.1109/JSTSP.2016.2523903.

[4] T. Zirtiloglu, N. Shlezinger, Y. C. Eldar, and R. T. Yazicigil, "Power-efficient hybrid MIMO receiver with task-specific beamforming using low-resolution ADCs," in *Proc. IEEE IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 5338–5342, doi: 10.1109/ICASSP43922.2022.9746362.

[5] N. Shlezinger, G. C. Alexandropoulos, M. F. Imani, Y. C. Eldar, and D. R. Smith, "Dynamic metasurface antennas for 6G extreme massive MIMO communications," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 106–113, Apr. 2021, doi: 10.1109/MWC.001.2000267.

[6] X. Qiao, Y. Zhang, M. Zhou, and L. Yang, "Alternating optimization based hybrid precoding strategies for millimeter wave MIMO systems," *IEEE Access*, vol. 8, pp. 113,078–113,089, 2020, doi: 10.1109/ACCESS.2020.3002788.

[7] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019, doi: 10.1109/TCOMM.2019.2924010.

[8] A. M. Elbir and A. K. Papazafeiropoulos, "Hybrid precoding for multiuser millimeter wave massive MIMO systems: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 552–563, Jan. 2020, doi: 10.1109/TVT.2019.2951501.

[9] N. Shlezinger, Y. C. Eldar, and S. P. Boyd, "Model-based deep learning: On the intersection of deep learning and optimization," *IEEE Access*, vol. 10, pp. 115,384–115,398, 2022, doi: 10.1109/ACCESS.2022.3218802.

[10] O. Lavi and N. Shlezinger, "Learn to rapidly and robustly optimize hybrid precoding," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 5814–5830, Oct. 2023, doi: 10.1109/TCOMM.2023.3292472.

[11] N. T. Nguyen et al., "Deep unfolding hybrid beamforming designs for THz massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 71, pp. 3788–3804, 2023, doi: 10.1109/TSP.2023.3322852.

[12] E. Balevi and J. G. Andrews, "Unfolded hybrid beamforming with GAN compressed ultra-low feedback overhead," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8381–8392, Dec. 2021, doi: 10.1109/TWC.2021.3092350.

[13] A. M. Elbir, K. V. Mishra, S. A. Vorobyov, and R. W. Heath Jr., "Twenty-five years of advances in beamforming: From convex and nonconvex optimization to learning techniques," *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 118–131, Jun. 2023, doi: 10.1109/MSP.2023.3262366.

[14] H. Al Kassir, Z. D. Zaharis, P. I. Lazaridis, N. V. Kantartzis, T. V. Yioultsis, and T. D. Xenos, "A review of the state of the art and future challenges of deep learning-based beamforming," *IEEE Access*, vol. 10, pp. 80,869–80,882, 2022, doi: 10.1109/ACCESS.2022.3195299.

[15] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016, doi: 10.1109/JSTSP.2016.2520912.

VT