# Disaggregated Deep Learning via In-Physics Computing at Radio Frequency

Zhihui Gao[1], Sri Krishna Vadlamani[2], Kfir Sulimany[2], Dirk Englund[2], and Tingjun Chen[1]

[1] *Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA.*
[2] *Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.*

**Abstract**

Modern edge devices, such as cameras, drones, and Internet-of-Things nodes, rely on deep learning to enable a wide range of intelligent applications, including object recognition, environment perception, and autonomous navigation. However, deploying deep learning models directly on the often resource-constrained edge devices demands significant memory footprints and computational power for real-time inference using traditional digital computing architectures. In this paper, we present WISE, a novel computing architecture for wireless edge networks designed to overcome energy constraints in deep learning inference. WISE achieves this goal through two key innovations: disaggregated model access via wireless broadcasting and in-physics computation of general complex-valued matrix-vector multiplications directly at radio frequency. Using a software-defined radio platform with wirelessly broadcast model weights over the air, we demonstrate that WISE achieves 95.7% image classification accuracy with ultra-low operation power of 6.0 fJ/MAC per client, corresponding to a computation efficiency of 165.8 TOPS/W. This approach enables energy-efficient deep learning inference on wirelessly connected edge devices, achieving more than two orders of magnitude improvement in efficiency compared to traditional digital computing.

## Introduction

Deep learning (DL) has revolutionized modern computing, enabling breakthroughs across a wide range of applications, including the Internet-of-Things (IoT), computer vision, and large language models (LLMs) [1–5]. As models now scale to billions of parameters [5, 6], the primary energy efficiency bottleneck is no longer just the raw computation efficiency, but also the energy cost of data movement between the local memory and processing units [7]. Moreover, retrieving DL model weights on demand from the cloud requires significant wireless bandwidth, while offloading DL inference to the cloud introduces potential privacy concerns [8]. At the same time, the theoretical lower bound for irreversible computation is set by Landauer's principle at 2.9 zeptojoules (zJ) per bit operation [9–11] at room temperature. In comparison, modern digital computing application-specific integrated circuits (ASICs) operate at energy efficiency in the picojoule range [7]. Bridging this gap calls for fundamentally different computing paradigms, including in-physics computing architectures that perform computing using continuous quantities (e.g., waves) with minimum data movement.

Recent works have explored a variety of in-physics computing approaches to overcome the memory wall, leveraging integrated photonic and optical waveguides [11–20], memristor-based crossbars with analog weight storage [21–26], and reconfigurable metasurfaces [27–31]. While these approaches have demonstrated promising energy efficiency gains [11, 19, 25], they often require specialized photonic or electronic hardware, limiting their scalability and practicality for large-scale deployments. In contrast, radio-frequency (RF) systems [32] present a compelling alternative by enabling wireless broadcast of model weights to edge devices, especially given that modern edge devices rely on wireless connectivity (e.g., cellular or wireless local area networks) for control signaling, data transfer, and Internet access.

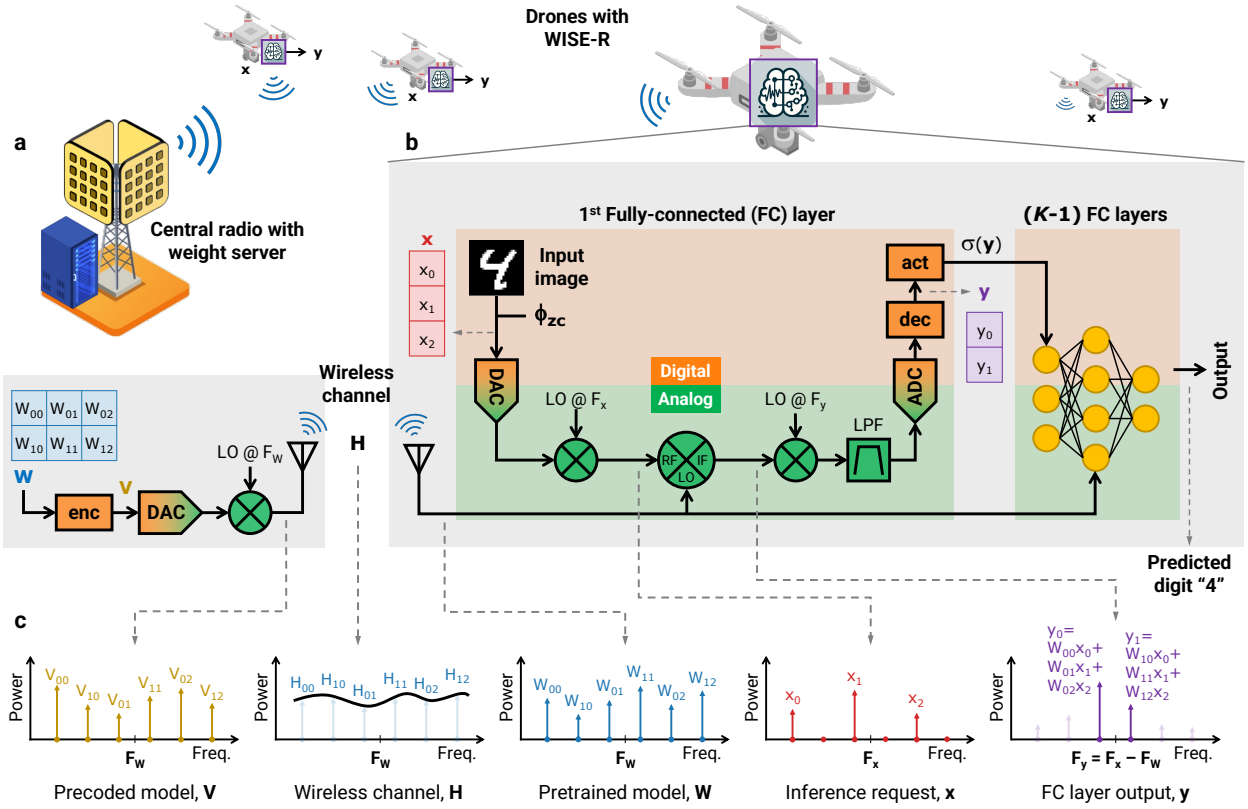


**Fig. 1 The WISE architecture enables disaggregated model access and energy-efficient deep learning (DL) to multiple clients in wireless edge networks. a**, A central radio broadcasts frequency-encoded model weights, $\mathbf{W}$, onto a radio-frequency (RF) signal at the carrier frequency $F_w$, which is precoded to $\mathbf{V}$ to mitigate the distortion introduced during propagation over the wireless channel, $\mathbf{H}$. **b**, Each client equipped with a WISE-R encodes the inference request $\mathbf{x}$ at the carrier frequency $F_x$, and performs local DL inference for $\mathbf{y}$ at the carrier frequency $F_y$, where the matrix-vector multiplications (MVM), or essentially the fully connected (FC) layers, are realized using a passive frequency mixer. **c**, Illustration of the in-physic MVM computation during frequency down-conversion with frequency-encoded $\mathbf{W}$, $\mathbf{x}$, and $\mathbf{y}$.

In this work, we present WISE (WIreless Smart Edge networks), the first edge computing architecture designed for disaggregated and energy-efficient DL via in-physics computing directly at RF (Fig. 1). In WISE, a central radio broadcasts RF signals that encode model weights ($\mathbf{W}$) and leverages the shared wireless channel to provide simultaneous, disaggregated model access to multiple edge clients (Fig. 1a). Each edge client, equipped with a WISE radio (WISE-R), performs inference on local data ($\mathbf{x}$) upon receiving the broadcast RF signals and obtains the matrix-vector multiplication (MVM) result, $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, as part of the DL inference (Fig. 1b). Both model weights and inference requests are frequency-encoded and I/Q modulated to an RF carrier, and in-physics MVM computation is realized using a passive frequency mixer (Fig. 1c). For example, a drone can execute object detection and image classification tasks on its captured images without the need to store the DL model locally. Our analysis shows that this computing paradigm, in the ideal case, achieves an energy efficiency approaching the thermodynamic limit (TDL) of analog hardware as the problem size scales, even exceeding the Landauer bound [9] for irreversible digital computation.

To enable ultra-low-power inference, each client requires minimum active hardware–primarily for analog-to-digital conversion (ADC) and lightweight digital signal processing–while offloading the most computationally intensive MVMs to the analog domain. This is achieved by exploiting RF electronics, such as mixers, that inherently perform signal multiplication and are widely used in modern edge devices. In addition, the encoding of both model weights and inference requests is optimized for spectral efficiency, drawing inspiration from modern wireless communication systems employing orthogonal frequency-division multiplexing (OFDM) and I/Q (de)modulation. A channel estimation and calibration process is integrated to mitigate signal distortions during wireless transmission.

We evaluate the energy efficiency of WISE for general inner-product (IP) computation and DL model inference tasks. Experimental results on a software-defined radio (SDR) platform with over-the-air transmissions demonstrate that WISE achieves an energy efficiency of 6.0 fJ/MAC, measured as energy per multiply-and-accumulate (MAC) operation, for 95.7% classification accuracy on the MNIST dataset [33]. This corresponds to a computation efficiency of 165.8 TOPS/W (Tera MAC operations per second per Watt). The energy efficiency can be further improved to 4.6 fJ/MAC (216.4 TOPS/W) with a slightly reduced classification accuracy of 90%. These represent a two to three orders of magnitude improvement compared to state-of-the-art digital computing ASICs operating at 1 pJ/MAC [7]. Detailed analysis and comprehensive experiments show that WISE has the potential to transform the landscape of wireless edge networks with embedded intelligence and to offer enhanced energy efficiency in a myriad of real-world applications.

# Results

## Central Radio and WISE-R

In WISE, a central radio wirelessly broadcasts model weights to a set of clients for local inference. The complex-valued model parameters and inference requests are encoded in the frequency domain of two RF waveforms via I/Q modulation. These signals are subsequently passed into an RF "computing" mixer, which naturally performs the time-domain multiplication (or frequency-domain convolution) of the two input waveforms during frequency mixing. The resulting output signal carries the desired analog computing results. Essentially, WISE effectively realizes the computation of complex-valued fully-connected (FC) layers, represented by $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, with $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{y} \in \mathbb{C}^M$, and $\mathbf{W} \in \mathbb{C}^{M \times N}$, directly in the analog domain.

Fig. 2a shows the experimental setup of WISE's implementation with three edge clients using an SDR platform (see details in Supplementary Section 12). Specifically in Fig. 2b, the central radio encodes the DL model weights in the $l$-th layer $\mathbf{W}^{(l)}$ onto a complex-valued waveform $w^{(l)}(t)$ with bandwidth $B$, which is then I/Q modulated to a time-domain waveform $r_w^{(l)}(t) = \mathsf{Re}\left\{w^{(l)}(t) \cdot e^{j2\pi F_w t}\right\}$ at frequency $F_w$ for broadcasting. As shown in Fig. 2c, each WISE-R consists of three main components: a transmitter (TX), which encodes the input to the $l$-th layer $\mathbf{x}^{(l)}$ onto a complex-valued waveform $x^{(l)}(t)$, which is then I/Q modulated to $r_x^{(l)}(t) = \mathsf{Re}\left\{x^{(l)}(t) \cdot e^{j2\pi F_x t}\right\}$ at $F_x$; a passive frequency mixer as the analog MVM (or IP) engine for computing $r_y^{(l)}(t) = r_w^{(l)}(t) \cdot r_x^{(l)}(t)$; and a receiver (RX), which I/Q demodulates, filters, and samples the mixed signal $r_y^{(l)}(t)$ at $F_y$ using the minimal required sampling rate, and decodes the output of the $l$-th layer, $\mathbf{y}^{(l)} = \mathbf{W}^{(l)} \cdot \mathbf{x}^{(l)}$. Note that when the computing mixer is used for frequency down-conversion, the carrier frequencies satisfy $F_y = F_x - F_w$, and a spectrum example of $r_x^{(l)}(t)$, $r_w^{(l)}(t)$ and $r_y^{(l)}(t)$ is shown in Fig. 2d. An activation function involving the absolute value function and a Zadoff-Chu (ZC) phase sequence [34] is then applied to $\mathbf{y}^{(l)}$ to generate the input to the next layer, $\mathbf{x}^{(l+1)} = \sigma(\mathbf{y}^{(l)}) = |\mathbf{y}^{(l)}| \cdot \mathbf{\Phi}_{\text{zc}}$. The use of $\mathbf{\Phi}_{\text{zc}}$ converts the real absolute values into complex, which ensures that the power of $x^{(l+1)}(t)$ is evenly distributed across frequency. See Supplementary Section 15 for a detailed workflow example with a three-layer DL model, and Supplementary Section 16 for the single-layer linear regression model without this ZC-phased activation function in digital. WISE also accounts for signal distortion caused by the wireless channel by incorporating channel state information (CSI), $\mathbf{H}$, into the encoding of $\mathbf{W}$ at the central radio, as detailed in Methods section and Supplementary Section 9. Fig. 2e illustrates an example of the WISE's in-physics computation on the DL-based image classification task (MNIST) on the three clients, respectively. Note that the CSI precoding can also be applied on $\mathbf{x}$ on each client (see Supplementary Sections 10 and 13), or eliminated for a wired channel (see Supplementary Sections 8 and 17).

## General IP Computation and Scalability

We benchmark WISE's analog computing performance for the complex-valued IP of two length-$N$ vectors, $c = \langle \mathbf{a}, \mathbf{b} \rangle = \sum_{n=1}^{N} a_n \cdot \overline{b_n}$, where $\overline{b_n}$ denotes the complex conjugate of $b_n$. Compared to MVM, $\mathbf{x}$ is replaced by $\mathbf{a}$ produced by the client, and $\mathbf{W}$ is replaced by $\mathbf{b}$ broadcast by the central radio. This IP computation involves $N$ complex-valued MACs, equivalent to $4N$ real-valued MACs. In particular, the amplitude and phase of $a_n$ and $b_n$ are drawn from independent uniform distributions $\mathcal{U}[0, 1]$ and $\mathcal{U}[0, 2\pi]$, respectively. $N$ subcarriers (excluding padded zero-subcarriers) are placed in the frequency domain when generating $x(t)$ and $w(t)$, and a single subcarrier is captured after the LPF on the $y(t)$ (Fig. 3a–b).

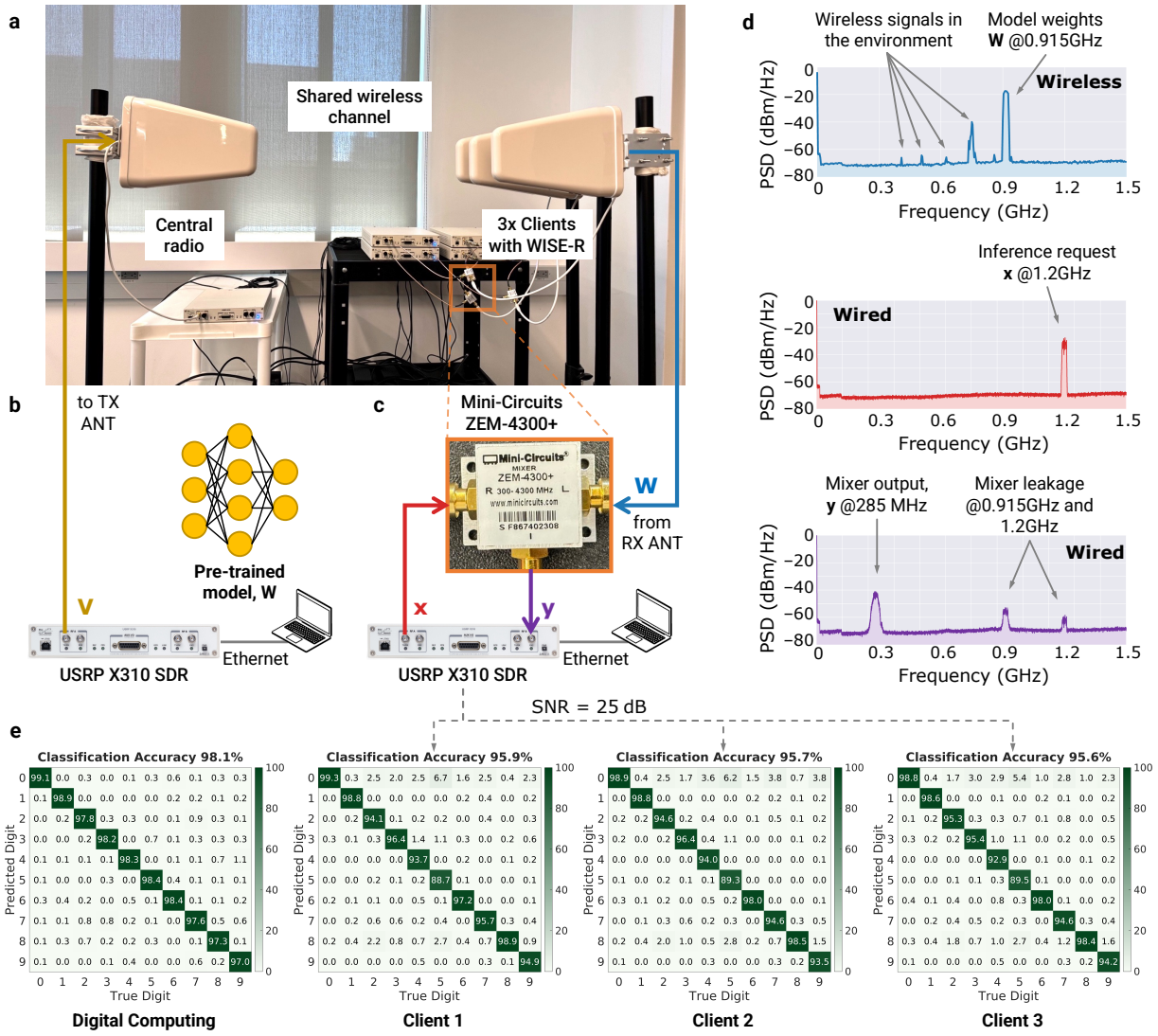


**Fig. 2 WISE's workflow with one central radio and multiple clients. a**, Experimental setup for WISE using a software-defined radio (SDR) platform: **b**, A central radio simultaneously provides disaggregated deep learning (DL) model access to three edge clients, each equipped with a WISE-R. **c**, On each client, the computing mixer performs general matrix-vector multiplications (MVMs) in-physics using the wirelessly received model weights ($\mathbf{W}$) and local inference request ($\mathbf{x}$). **d**, The model weights $\mathbf{W}$ is modulated at $F_w = 0.915\,\text{GHz}$ over a wireless channel, and the inference request $\mathbf{x}$ is modulated at $F_x = 1.2\,\text{GHz}$; after down-conversion, the MVM result $\mathbf{y}$ is located at $0.285\,\text{GHz}$. **e**, WISE achieves classification accuracies of 97.1%–97.4% across the three clients on the MNIST dataset using the LeNet-300-100 model, which is comparable to the accuracy of 98.1% achieved by traditional digital computing but with significantly improved energy efficiency.

Fig. 3c shows the experimental IP computing accuracy, measured by the root mean squared error (RMSE) of the IP obtained by in-physics computing ($\widehat{c}$) compared to the ground truth ($c$), with a normalization factor of $1/\sqrt{N}$ and under varying SNR values (see detailed definition in Supplementary Section 13). The normalization ensures consistent distributions of the IP results across different problem sizes ($N$). WISE achieves an RMSE of 0.055 at 25 dB SNR with $N = 4,096$, equivalent to a computing accuracy of $-\log_2(\text{RMSE}/2) \approx 5$ bit [35, 36], sufficient for various ML inference tasks [37, 38]. Simulation results demonstrate slopes of 6.7 dB/bit computing accuracy for 4,096-point and 32,768-point IP. A similar trend is observed from the experiments in the low SNR regime ($\text{SNR} < 25\,\text{dB}$). In the high SNR regime ($\text{SNR} > 25\,\text{dB}$), the computing accuracy is no longer limited by the thermal noise but by the imperfect channel estimation and computing mixer that inherently operates using on-off switching instead of performing the ideal multiplication.

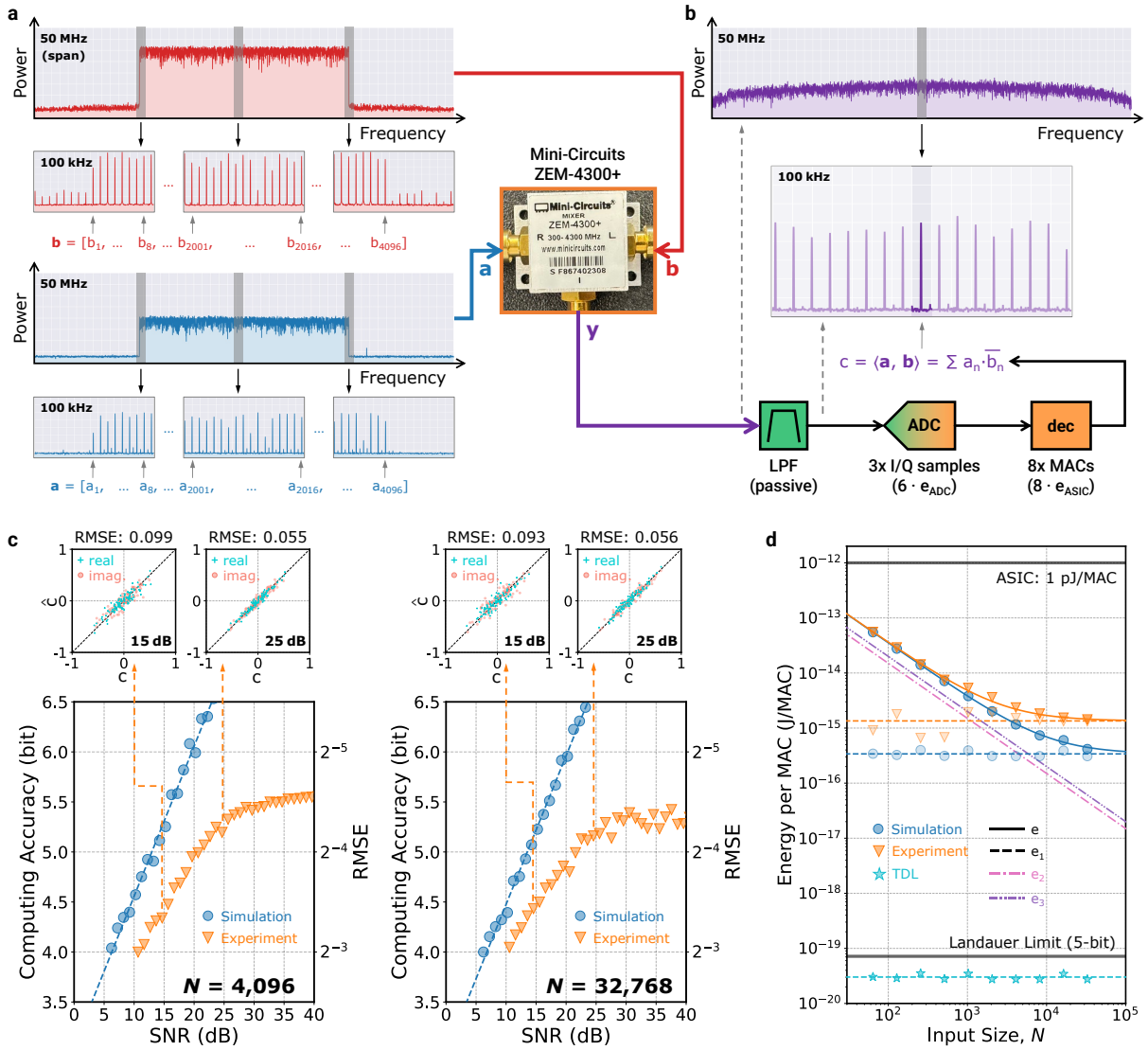


**Fig. 3 Benchmarking general complex-valued inner-product (IP) computation: computing accuracy and energy efficiency. a**, Complex-valued IP computation of two length-$N$ vectors, $c = \langle \mathbf{a}, \mathbf{b} \rangle = \sum_{n=1}^{N} a_n \cdot \overline{b_n}$, where $\mathbf{a}$ and $\mathbf{b}$ are frequency encoded onto $N$ (4,096) subcarriers across a bandwidth of $B$ (25 MHz). **b**, Decoding of the IP result, $c$, after the in-physics IP computation, low-pass filtering, and sampling using an analog-to-digital converter (ADC). **c**, IP computing accuracy achieved by WISE as a function of the signal-to-noise ratio (SNR) for $N = 4,096$ and $N = 32,768$. **d**, Energy efficiency of WISE, $e_{\text{mvm}}$ (J/MAC), required to achieve RMSE $< 0.0625$ (equivalent to 5-bit computing accuracy [35, 36]) as a function of the IP size, $N$.

Fig. 3d plots the energy efficiency of a WISE-R to achieve RMSE $< 0.0625$ (i.e., 5-bit computing accuracy) for IP across varying values of $N$. For $N = 4,096$, WISE-R achieves an energy efficiency of 2.4 fJ/MAC (421.9 TOPS/W) in experiments. As the IP dimension increases, the energy consumption of I/Q sampling and digital FFT is amortized, leading to energy efficiency asymptotically approaching $e_1$ for the waveform generation and I/Q (de)modulation. Experimental results further validate this trend, demonstrating an energy efficiency of 1.4 fJ/MAC (699.3 TOPS/W) with $N = 32,768$, nearly three orders of magnitude lower than the state-of-the-art ASICs operating at 1 pJ/MAC [7, 39, 40]. In simulations with ideal hardware, $e_{\text{tdl}}$ required for achieving 5-bit computing accuracy is projected to be 28.7 zJ/MAC (34.8 ExaOPS/W), surpassing the Landauer limit [9] for MAC computation with 5-bit accuracy.

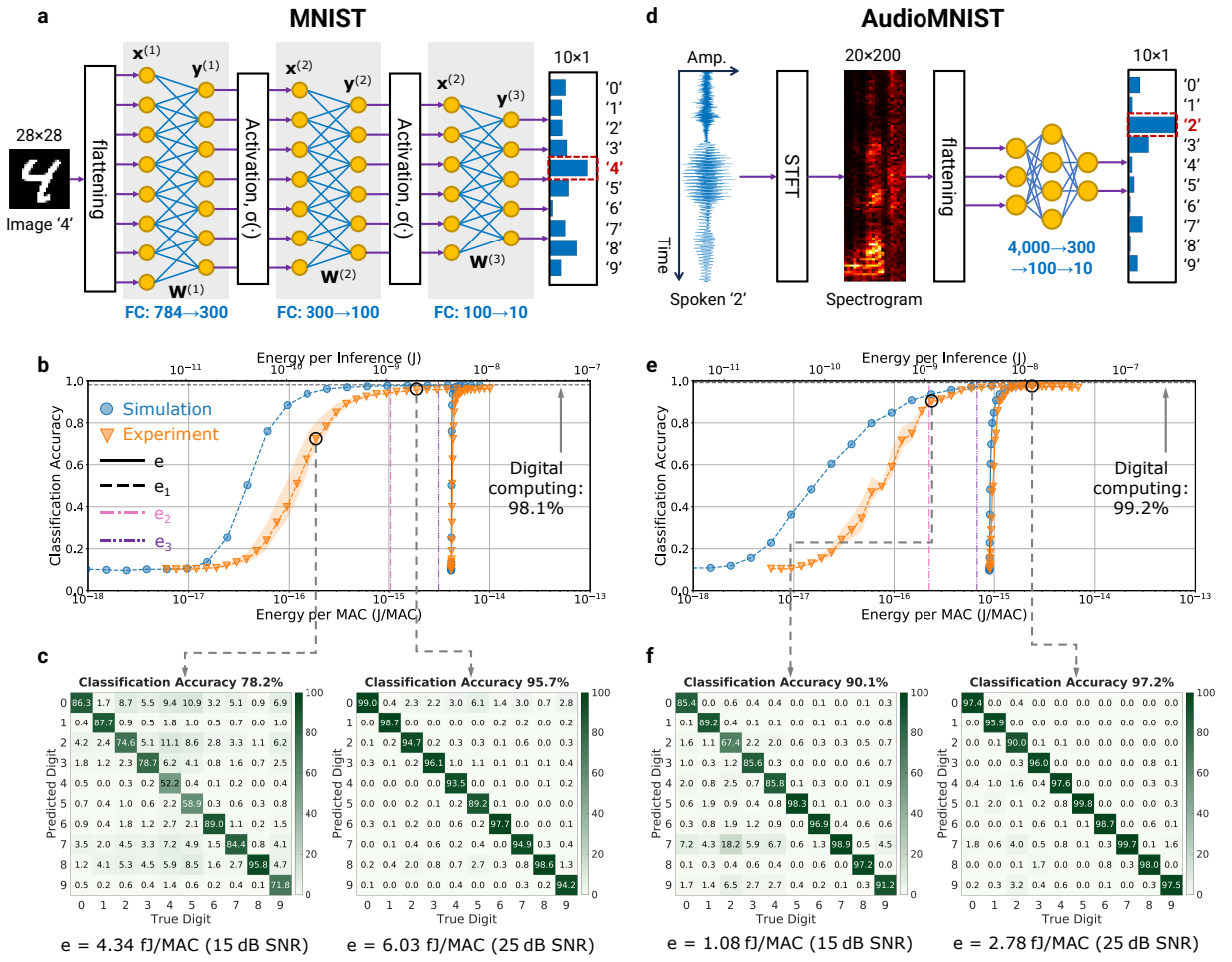


**Fig. 4 WISE for energy-efficiency deep learning (DL) inferences.** **a**/**d**, Deployment of WISE for DL tasks using complex-valued three-layer models: classification of handwritten digits on the MNIST dataset (**a**) and spoken digits on the AudioMNIST dataset (**d**). **b**/**e**, Experimental classification accuracy achieved by WISE on the MNIST (**b**) and AudioMNIST (**e**) datasets over different energy efficiency (J/MAC), and the corresponding energy consumption per inference. **c**/**f**, Confusion matrices for classification accuracy at 10 dB and 20 dB SNR on the MNIST (**b**) and AudioMNIST (**e**) datasets.

## ML for Image/Audio Classification

We deploy WISE for two DL inference tasks, where the central radio wirelessly broadcasts model weights to three clients equipped with WISE-R: image classification on the MNIST dataset [33], and audio signal classification on the AudioMNIST dataset [41]. Both tasks employ a complex-valued model following LeNet-300-100 [33] with three FC layers. The complex-valued model exploits the absolute function combined with a pre-defined Zadoff-Chu phase sequence as the activation function after each of the first two FC layers; for the last layer, only the absolute function is applied before the output layer. The models are trained on an NVIDIA A100 GPU using cross-entropy loss; during the training process, the models with the highest testing accuracy by digital computing are recorded and used.

Each WISE-R performs local inference upon receiving the three-layer model broadcast by the central radio. Each FC layer is formulated as an MVM, which is naturally realized during down-conversion as $x^{(l)}(t)$ and $w^{(l)}(t)$ pass through the computing mixer. The mixer output is subsequently low-pass filtered, digitized, and decoded before applying the activation function. In the experiment, the MVMs corresponding to individual FC layers are divided into smaller MVMs with $M' = 6$, $\alpha = 0.33$, and $\beta = 0.25$. This process is repeated for each layer, and the final classification results are obtained from the output $\mathbf{y}^{(3)}$.

For the MNIST dataset, the images are gray-scaled with dimensions of 28×28 pixels, which are flattened into $\mathbf{x}^{(1)} \in \mathbb{C}^{784}$ as inputs to the DL model. The three-layer FC model has an architecture of 784–300–100–10

following LeNet-300-100 [33] (Fig. 4a), consisting of 0.27 million complex-valued parameters and requiring 1.06 million real-valued MACs. By digital computing, the classification accuracy of the pre-trained three-layer FC model is 98.1% across a testing set of 10,000 images. Fig. 4b shows the averaged classification accuracy across three clients by WISE's in-physics computing at varying different SNR levels. To achieve 90% classification accuracy, the experimental energy efficiency is 4.6 fJ/MAC (216.5 TOPS/W) at 11.8 dB SNR, with a breakdown of 0.5 fJ/MAC, 1.0 fJ/MAC, and 3.1 fJ/MAC for $e_1$, $e_2$, and $e_3$, respectively. Simulations validate this trend, demonstrating an energy efficiency of 4.2 fJ/MAC (236.1 TOPS/W) for achieving 90% classification accuracy. Fig. 4c shows the detailed confusion matrices across three clients at 15 dB and 25 dB SNR, with experimental classification accuracies of 78.2% and 95.7%, respectively.

AudioMNIST is a dataset of audio signals containing spoken digits from '0' to '9'. Each audio clip is converted into a spectrogram using the short-time Fourier transform (STFT), which is then concatenated as a vector with Zadoff-Chu phases, $\mathbf{x}^{(1)} \in \mathbb{C}^{4,000}$. The pre-trained AudioMNIST model consists of 1.23 million complex-valued parameters across three FC layers, involving 4.92 million real-valued MACs (Fig. 4d). Such a three-layer model achieves a digital computing accuracy of 99.2% on the AudioMNIST's testing set with 3,000 audio clips. As shown in Fig. 4e, an experimental classification accuracy of 90% requires the experimental energy consumption of 1.1 fJ/MAC (885.0 TOPS/W), including the energy efficiency breakdown of 0.2 fJ/MAC, 0.2 fJ/MAC, and 0.7 fJ/MAC for $e_1$, $e_2$, and $e_3$, respectively. Simulations under this accuracy level reveal an energy efficiency of 1.0 fJ/MAC (1.0 PetaOPS/W). The discrepancy between experimental and simulation results in both DL tasks is mainly due to imperfect wireless channel estimation and calibration. Fig. 4f shows the confusion matrices with 15 dB and 25 dB SNR. Under 25 dB SNR, the average experimental accuracy across the three clients is 97.2%, with an energy efficiency of 2.8 fJ/MAC (359.7 TOPS/W).

## Discussion

We presented WISE, a novel computing paradigm that enables disaggregated and energy-efficient DL inference simultaneously on multiple edge clients equipped with WISE-R. Leveraging wireless delivery of DL models broadcast by a central radio, each WISE-R utilizes a (passive) frequency mixer to perform IP or MVM computation directly at RF. Through comprehensive theoretical analysis and simulations, we show that WISE achieves energy efficiency approaching the thermodynamic limit as the problem size $N \to +\infty$, surpassing the Landauer bound of conventional digital computing. Extensive experiments demonstrate that WISE achieves over 5-bit computing accuracy for IPs up to $N = 32,768$. For DL tasks involving large-scale MVMs, WISE achieves a classification accuracy of 95.7% and 97.2% using the MNIST and AudioMNIST dataset, respectively, at energy efficiencies of 6.0 fJ/MAC and 2.8 fJ/MAC, corresponding to computation efficiencies of 165.8 TOPS/W and 359.7 TOPS/W. This represents two to three orders of magnitude of energy efficiency improvement compared to digital computing using state-of-the-art ASICs. WISE can also be adapted to various MVM-based DL tasks, including convolutional neural networks [1, 14, 16, 33] and transformers [6, 42].

Taking one step further, the energy efficiency of WISE can be further improved through an all-analog architecture, where energy consumption is primarily attributed to analog waveform generation. Supplementary Section 16 demonstrates the effectiveness of a single-layer analog model, and multi-layer analog models can be realized by integrating electronics that inherently perform non-linear activation functions based on their physical properties, such as transistors or diodes [13, 17, 19, 36]. The gap between practical energy efficiency and the theoretical limit can be further narrowed using advanced hardware and ASICs [43–45]. Beyond outdoor deployments constrained by limited spectrum, WISE is also applicable to indoor compute clusters performing DL inference in a shielded environment, where directional antennas mounted on top of server racks [46] can stream model weights to clients with increased bandwidth. Moreover, a central radio equipped with large-scale antenna arrays [47] can accelerate DL inference for a single broadcast task or serve multiple models to multiple clients, exploiting the spatial multiplexing gain of the wireless channel. The physical separation between the central radio (hosting DL models) and edge clients (generating local inference requests) offers an additional privacy benefit by mitigating the risk of information leakage [8], where the inherent "noisy" nature of the wireless channel can be harnessed for model weight precoding. By integrating pervasive RF signals into the in-physics computing ecosystem, WISE unlocks large-scale DL deployment on ubiquitous edge devices at orders of magnitude lower power consumption and complexity.

# Methods

## Energy Efficiency

The energy consumption of WISE-R is minimized via the wireless broadcast of disaggregated model weights from a central radio. The energy consumed by a WISE-R to perform an MVM consists of three parts: $E_1$ for the generation of $x(t)$ and I/Q modulation, $E_2$ for the I/Q sampling of $\mathbf{y}$ from $y(t)$ using two ADCs after I/Q demodulation, and $E_3$ for the digital FFT operation to decode $\mathbf{y}$, i.e.,

$$\begin{aligned} E_{\text{mvm}} &= E_1 + E_2 + E_3 \\ &= (1+\alpha)(1+\beta) \cdot NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0 + (1+\alpha) \cdot 2M \cdot e_{\text{adc}} + (1+\alpha) \cdot 2M \log_2\left((1+\alpha)M'\right) \cdot e_{\text{dig}}. \end{aligned} \tag{1}$$

Here, $k_B T_0 = -174\,\text{dBm/Hz}$ is the thermal noise power spectrum density at room temperature of $T_0 = 300\,\text{K}$. $\eta \in (0, 1]$ is the overall loss of the WISE-R hardware including the energy efficiency of the TX, insertion loss of the computing mixer, and noise floor of the RX. To ensure robust performance, $\alpha > 0$ is the overhead coefficient of the zero-subcarriers to overcome the LPF's roll-off effect, and $\beta$ is the overhead coefficient of the cyclic prefix for a better timing synchronization tolerance (see Supplementary Section 12). $\mathsf{SNR}$ refers to the signal-to-noise ratio (SNR) measured at the RX. Moreover, $e_{\text{adc}}$ is the energy consumed per sample by an ADC, and $e_{\text{dig}}$ is the energy consumed by an ASIC per real-valued MAC operation in digital computing.

Since each complex-valued MVM involves $4NM$ real-valued MACs, the *energy efficiency* of MVM computation, $e_{\text{mvm}}$, measured by energy per real-valued MAC (J/MAC), is given by

$$\begin{aligned} e_{\text{mvm}} &= \frac{E_{\text{mvm}}}{4NM} = e_1 + e_2 + e_3 \\ &= \frac{(1+\alpha)(1+\beta)}{4} \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0 + \frac{1+\alpha}{2N} \cdot e_{\text{adc}} + \frac{1+\alpha}{2N} \cdot \log_2\left((1+\alpha)M'\right) \cdot e_{\text{dig}}. \end{aligned} \tag{2}$$

It can be seen that $e_{\text{mvm}}$ significantly improves for large values of $N$ since $e_2$ and $e_3$ scale as $\mathcal{O}(1/N)$, e.g., $N = 11,008$ in emerging LLMs such as Llama-2-7b [6]. With $M' = 1$, equation (2) is reduced to $e_{\text{ip}}$ for IP computation (See Supplementary Sections 9 and 14). With ideal hardware ($\eta = 1$, $\alpha = \beta = 0$), $e_{\text{mvm}}$ approaches its thermodynamic limit (TDL) as $N \to +\infty$,

$$e_{\text{tdl}} := \lim_{N \to +\infty} e_{\text{mvm}} = \lim_{N \to +\infty} e_{\text{ip}} = \mathsf{SNR} \cdot k_B T_0 / 4. \tag{3}$$

We define the corresponding *computation efficiency* for each WISE-R as the reciprocal of energy per MAC, $(e_{\text{mvm}})^{-1}$, measured by the number of (real-valued) MAC operations per second per Watt (OPS/W). See Supplementary Sections 8–10 for more details on the energy efficiency and overhead analysis.

## Computation Throughput

The disaggregated setup of WISE treats the shared wireless medium as a channel for the central radio to deliver DL model parameters for energy-efficient inference at each client, which is different than conventional communication systems for data delivery. Hereby, in the context of DL inference, we define the computation throughput as the number of (real-valued) MAC operations per second (OPS). We define the *computation throughput* of this channel over $U$ clients as a function of $B$, $N$, and $M$. Consider the complex-valued MVM, $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, involving $NM$ complex-valued MACs, or $4NM$ real-valued MACs. Waveforms $x(t)$ and $w(t)$ corresponding to $\mathbf{x}$ and $\mathbf{W}$ last for a time duration of $T = (1+\alpha)(1+\beta) \cdot NM/B$. This waveform time $T$ dominates the latency of the disaggregated computation process. Across $U$ clients, a total number of $U \cdot 4NM$ MACs can be completed within the waveform time $T$, corresponding to a computation throughput given by

$$\Lambda = \frac{U \cdot 4NM}{T} = \frac{4 \cdot UB}{(1+\alpha)(1+\beta)}\ [\text{OPS}], \tag{4}$$

which scales as a function of the available bandwidth, $B$, and number of clients, $U$. More details can be found in Supplementary Section 11.

## Wireless Channel Calibration

In the wireless setting, the channel carrying DL model parameters in $\mathbf{W}$ exhibits propagating delay and multi-path effect, therefore requiring a channel estimation and calibration process to guarantee accurate delivery of the model parameters. The channel state information (CSI) from the central radio to a client equipped with WISE-R can be represented by a complex matrix $\mathbf{H} = [H_{m,n}] \in \mathbb{C}^{M\times N}$, which has the same dimension as $\mathbf{W}$. Using a pre-defined signal, $\mathbf{H}$ can be estimated by minimum mean squared error (MMSE) and nearest neighbor interpolation, as described in Supplementary Section 9. The estimated $\mathbf{H}$ is fed back to the central radio, and this process is only performed once as long as the wireless environment does not change significantly. To account for signal distortion introduced by the wireless channel, we apply a precoder on $\mathbf{W}$ to generate the transmitted signal given by $\mathbf{V} = [V_{m,n}] \in \mathbb{C}^{M\times N}$, where $V_{m,n} = \frac{W_{m,n}}{H_{m,n}}$. This precoding on the central radio, termed the $\mathbf{W}$-precoding scheme, ensures that the signal received by the client contains the desired frequency-encoded model weights, $\mathbf{W}$, which can then be used for local inference.

For multiple clients located in proximity, the same estimated $\mathbf{H}$ can be applied to the model weight broadcast to all clients, which does not require modification of the client behavior. One alternative scheme that tolerates diverse CSI across clients with better computing accuracy is to precode $x(t)$ on each client using $\mathbf{H}$ estimated for individual clients. While this client-side precoding scheme incurs extra computing overhead on the client side, it achieves improved computing accuracy. More details about the wireless channel calibration and different precoding schemes can be found in Supplementary Sections 10 and 13.

## MNIST and AudioMNIST Dataset Preparation

Each data sample in MNIST [33] is a gray-scaled image $\mathbf{I} \in [0, 255]^{28\times 28}$ representing a handwritten digit from '0' to '9'. Each image is first reshaped to a 784-point vector and then element-wisely modulated by a 784-point ZC phase sequence $\mathbf{\Phi}_{\text{zc}} = [\phi_{\text{zc}}[m]] \in \mathbb{C}^{784}$ to generate $\mathbf{x} \in \mathbb{C}^{784}$ [34], where

$$\phi_{\text{zc}}[m] = -\frac{m(m+c_f)}{M} \cdot \pi, \text{ where } c_f = M \bmod 2, \ \forall m = 0, 1, \ldots, M-1. \tag{5}$$

Each audio clip in AudioMNIST [41] is a real-valued waveform of English spoken digits from '0' to '9' by 60 people whose native languages are English, German, Chinese, and Spanish. Each waveform, sampled at 48 kHz, lasts for about 0.5 seconds. In our implementation, each waveform is first downsampled to 8 kHz, and the middle 0.5 seconds is truncated to a 4,000-point vector. Then, we perform a short-time Fourier transform (STFT) every 25 ms for 200 non-overlapped time windows to form a spectrogram. Note that we only take the amplitudes of this spectrogram and drop the phase information. Each time window contains 20 samples, which are converted into 20 complex frequency bins by STFT. Finally, the 200 time windows are concatenated into a vector, which is then modulated with the 4,000-point ZC phase sequence $\mathbf{\Phi}_{\text{zc}}$ to generate $\mathbf{x} \in \mathbb{C}^{4,000}$.

## Dataset and Complex Model Architecture

We consider FC layers in our DL model architecture for WISE that employ large-scale MVMs. The complex nature of the RF signals enables FC layers with complex-valued input vector, $\mathbf{x}$, output vector, $\mathbf{y}$, and trainable weight matrix, $\mathbf{W}$. We employ an activation function $\sigma(\cdot)$ that applies an absolute value operation followed by a phase adjustment using the ZC phase sequence. Specifically, for each FC layer except the last, the activation function first computes the element-wise absolute value of $\mathbf{y}$ and then adds a phase based on a $M$-point ZC sequence $\mathbf{\Phi}_{\text{zc}}$,

$$\sigma_m(y_m) = |y_m| \cdot e^{j\phi_{\text{zc}}[m]} = |y_m| \cdot e^{-j\cdot\frac{\pi m(m+c_f)}{M}}, \text{ where } c_f = M \bmod 2, \ \forall m = 0, 1, \ldots, M-1. \tag{6}$$

Hereby, the subscript on $\sigma_m$ indicates the phase shift applied to each element of $y_m$. The reason behind selecting this activation function is twofold. First, it preserves the waveform power by maintaining the

amplitude of each element in $\mathbf{y}$. Second, the use of ZC phase sequence ensures that the power of the input waveform $x(t)$ to the subsequent FC layer is evenly distributed across the spectrum. For the last FC layer, only the absolute function $|y_m|$ is applied, which converts the complex-valued $\mathbf{y}$ to a real-valued vector.

To train the model, we employ the cross-entropy loss on $|\mathbf{y}|$, using Adam optimizer [48] with a learning rate of $10^{-3}$ over 100 epochs. In the testing phase, the predicted class is given by the maximum $|\mathbf{y}|$ after the absolute function. Among the 100 training epochs, we select the model with the highest testing accuracy. All activation functions are computed digitally. When comparing WISE with conventional DL models performed in digital computing, we exclude the energy consumption and latency of the activation functions as they exist in both computing paradigms and are orthogonal to the MVM operations.

### Implementation

To demonstrate the WISE framework, we develop a WISE-R prototype using a USRP X310 SDR and a Mini-Circuits ZEM-4300+ frequency mixer [49], which function as the TX/RX and computing mixer, respectively. Fig. 2 shows the experimental setup, where a central radio broadcasts model weights to three clients over a 25 MHz channel centered at 0.915 GHz, which is limited by the available unlicensed frequency spectrum in the industrial, scientific, and medical (ISM) bands between 902–928 MHz [50]. (See Supplementary Section 17 for the wired experiments with larger bandwidth) The wireless link distance is ≈1 meter, limited by the LO power required for the off-the-shelf diode ring-based mixer. This constraint can be relaxed to support larger link distances using integrated analog computing circuits [45] or beamforming on an antenna array [47]. Each client streams the I/Q modulated waveform $x(t)$ at a carrier frequency of 1.2 GHz to the mixer's RF port over an SMA cable. In the meanwhile, $w(t)$ is received by the WISE-R's antenna and then mixed with the streamed $x(t)$. The output downconverted waveform $y(t)$ from 0.285 GHz is I/Q demodulated, low-pass filtered, and sampled by two ADCs operating at a low sampling rate of 0.2 MHz. For the waveform generation, we place zero padding subcarriers in the frequency domain to mitigate the roll-off effect at the LPF edge, and cyclic prefix in the time domain to improve the timing synchronization tolerance.

We evaluate the energy efficiency of WISE given by equation (2), where the $\mathsf{SNR}$ values are varied by adjusting the transmit power of $x(t)$. The overall loss $\eta = 1.48 \times 10^{-4}$ is the combination of a TX efficiency of 10% [51], insertion loss of the computing mixer (measured at 11.4 dB), and RX noise figure (measured at 16.9 dB). We also conduct simulations for WISE by plugging in the realistic hardware parameters above, and the computing mixer performs analog multiplication. The simulations consider a frequency-flat wireless channel between the central radio and WISE-R with additive Gaussian white noise (AWGN). For both experiments and simulations, we consider ADC energy consumption ($e_{\text{adc}}$) of 1 pJ/sample [52], and digital computing efficiency using ASICs ($e_{\text{dig}}$) of 1 pJ/MAC [7, 39, 40]. See Supplementary Section 12 for more details on the experimental setup and measurements. The TDL of WISE can be simulated based on equation (3) assuming ideal WISE-R hardware with $\eta = 1$ and $\alpha = \beta = 0$.

## Acknowledgments

Z.G. and T.C. acknowledge partial support from the NSF Athena AI Institute for Edge Computing (CNS-2112562). S.K.V. and D.E. acknowledge support from the DARPA NaPSAC program. K.S. acknowledges the support of the Israeli Council for Higher Education and the Zuckerman STEM Leadership Program. D.E. acknowledges partial support from the NSF EAGER program (ECCS-2419204) and the DARPA QuANET program. The authors thank Marc Bacvanski for the useful discussion and for providing feedback on the manuscript.

## Author Contributions

D.E. and T.C. conceived the original concept and system architecture. Z.G. and T.C. designed the experiments using the SDR platform. Z.G. conducted the experiments and analyzed the results. Z.G., D.E., and T.C. wrote the manuscript. All authors contributed to the analysis and refinement of the analog computing schemes, and provided feedback on the manuscript.

## Competing Interests

The authors declare no competing interests.

## Data Availability

The data supporting the claims in this paper is available upon reasonable request.

## Correspondence

Requests for information should be directed to Tingjun Chen (tingjun.chen@duke.edu).

# Supplementary Information: Disaggregated Deep Learning via In-Physics Computing at Radio Frequency

Zhihui Gao[1], Sri Krishna Vadlamani[2], Kfir Sulimany[2], Dirk Englund[2], and Tingjun Chen[1]

[1]*Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA*
[2]*Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

## Contents

# Supplementary Information: Theory

## 1 Notation and Preliminaries

For a complex-valued matrix $\mathbf{A} \in \mathbb{C}^{M\times N}$, let $\mathbf{A}^\top$, $\overline{\mathbf{A}}$, and $\mathbf{A}^*$ denote its transpose, complex conjugate, and conjugate transpose, respectively. For a square matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ that is invertible, let $\mathbf{A}^{-1}$ denote the inverse of $\mathbf{A}$. Let $\mathbf{I}_N$ be the $N \times N$ identity matrix. For two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same dimension, let $\mathbf{A} \odot \mathbf{B}$ denote the element-wise multiplication (Hadamard product), and $\mathbf{A} \oslash \mathbf{B}$ denote the element-wise division.

The discrete Fourier transform (DFT) converts a vector, $\mathbf{x} = [x_n] \in \mathbb{C}^L$, into another vector with equal length, $\mathbf{X} = [X_k] = \mathsf{DFT}(\mathbf{x}) \in \mathbb{C}^L$, where

$$X_k = \sum_{n=0}^{L-1} x_n \cdot e^{-j2\pi \frac{k}{L} n}, \ \forall k = 0, 1, \ldots, L-1. \tag{S1}$$

The DFT operation can be written in the matrix form,

$$\mathbf{X} = \sqrt{L} \cdot \mathbf{D} \cdot \mathbf{x}, \tag{S2}$$

where $\mathbf{D} \in \mathbb{C}^{L\times L}$ is the $L$-point DFT matrix given by

$$\mathbf{D} = \frac{1}{\sqrt{L}} \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & d & d^2 & \ldots & d^{L-1} \\ 1 & d^2 & d^4 & \ldots & d^{2(L-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d^{L-1} & d^{2(L-1)} & \ldots & d^{(L-1)^2} \end{bmatrix}, \tag{S3}$$

where $d = e^{-j2\pi/L}$. Note that the DFT matrix is a unitary matrix satisfying $\mathbf{D}\mathbf{D}^* = \mathbf{I}$, with $\mathbf{D} = \mathbf{D}^\top$ and $\mathbf{D}^{-1} = \mathbf{D}^* = \overline{\mathbf{D}}$. Symmetrically, the inverse DFT (IDFT) operation is given by $\mathbf{x} = \mathsf{IDFT}(\mathbf{X})$, where

$$x_n = \frac{1}{L} \sum_{k=0}^{L-1} X_k \cdot e^{j2\pi \frac{k}{L} n}, \ \forall n = 0, 1, \ldots, L-1. \tag{S4}$$

The IDFT operation can also be written in the matrix form, given by

$$\mathbf{x} = \frac{1}{\sqrt{L}} \cdot \mathbf{D}^{-1} \cdot \mathbf{X}. \tag{S5}$$

The DFT and IDFT operations in equations (S1) and (S4) can be accelerated by the fast Fourier transform (FFT) algorithm. Specifically, given a vector length of $L$ (assuming $L$ is the power of 2), it takes $L/2 \cdot \log_2 L$ complex-valued MACs to conduct DFT or IDFT, which is equivalent to $2L \log_2 L$ real-valued MACs.

We define the circular shift matrix $\mathbf{R}_L \in \mathbf{R}^{L\times L}$ that, for an $L$-point vector, shifts all the elements one position to the right and puts the last element to the first position,

$$\mathbf{R}_L = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \tag{S6}$$

We can then denote a repeated shift operation applied $m$ times by $(\mathbf{R}_L)^m$. We drop the subscript $L$ for brevity when the matrix dimension and context are clear.

## 2 Digital-to-Analog and Analog-to-Digital Conversions

We use $s(t)$ and $s[n]$ to represent the continuous and discrete signal in the time domain, and $S(f)$ and $S[k]$ to represent the continuous and discrete spectrum of the signal in the frequency domain. In this section, we consider real-valued $s(t)$ and $s[n]$ with a single DAC and ADC, and extend the scenario to complex-valued $s(t)$ and $s[n]$ with two DACs and ADCs for I/Q modulation in Supplementary Section 3.

In general, a DAC reconstructs the continuous waveform $s(t)$ from $s[n]$ by a per-sample duration of $T_s = 1/f_s$, or under a sampling rate of $f_s$. Specifically, signal construction using a DAC can be modeled by

$$s(t) = \mathsf{DAC}\left\{s[n]\right\} = \sum_{n=-\infty}^{+\infty} s[n] \cdot h_{\mathrm{DAC}}\left(\frac{t}{T_s} - n\right), \tag{S7}$$

where $h_{\mathrm{DAC}}(\cdot)$ is the reconstruction kernel of the DAC. Note that equation (S7) only constrains the values with integer values of $n$. Hence, there are infinitely many reconstruction kernels from $s[n]$ to $s(t)$ that satisfy such reconstructions. For example, sinc-interpolation limits the bandwidth of the reconstructed waveform within $[-f_s/2, +f_s/2]$, which is

$$s_{\mathrm{sinc}}(t) = \mathsf{DAC}_{\mathrm{sinc}}\left\{s[n]\right\} = \sum_{n} s[n] \cdot \mathrm{sinc}\left(\frac{t}{T_s} - n\right), \text{ where } \mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}. \tag{S8}$$

In practice, a commonly used signal reconstruction for a DAC is zero-order hold (ZOH), given by

$$s_{\mathrm{ZOH}}(t) = \mathsf{DAC}_{\mathrm{ZOH}}\left\{s[n]\right\} = \sum_{n} s[n] \cdot \mathsf{Rect}\left(\frac{t}{T_s} - n - \frac{1}{2}\right), \text{ where } \mathsf{Rect}(x) = \begin{cases} 1, & -\frac{1}{2} \le x \le +\frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{S9}$$

An ADC samples the continuous waveform $s(t)$ and creates the discrete sequence $s[n]$ given by

$$s[n] = \mathsf{ADC}\left\{s(t)\right\} = s\left(nT_s\right), \ \forall n = 0, 1, 2, \dots. \tag{S10}$$

It can be proven that for any reconstruction kernel on DAC, as long as the DAC is synchronized with the ADC, i.e., with the shared $t$, the original discrete sequence can be reconstructed:

$$s[n] \propto \mathsf{ADC}\left\{\mathsf{DAC}\left\{s[n]\right\}\right\}. \tag{S11}$$

For the DAC, we assume a normalized power constraint on the input sequence given by

$$|s[n]|^2 \leq 1, \ \forall n = 0, 1, 2, \ldots, \tag{S12}$$

and $|s[n]|^2 = 1$ corresponds to the DAC's peak output power, $P_{\max}$. The peak-to-average power ratio (PAPR) of the sequence $\mathbf{s} = [s[n]]$ is given by

$$\mathrm{PAPR}[\mathbf{s}] = \frac{\max_n |s[n]|^2}{\mathbb{E}[s^2[n]]}, \tag{S13}$$

where $\max_n |s[n]|^2$ and $\mathbb{E}[s^2[n]]$ represent the peak instant and average power of $\mathbf{s}$, respectively.

For the ADC, we consider the constraint on the input waveform $|s(t)|$ that

$$|s(t)|^2 \leq 1, \ \forall t. \tag{S14}$$

Similarly, the PAPR of the input waveform to the ADC $s(t)$ is given by

$$\mathrm{PAPR}[s(t)] = \frac{\max_t |s(t)|^2}{\frac{1}{T}\int_0^T |s(t)|^2 \ \mathrm{d}t}, \tag{S15}$$

where $\max_t |s(t)|^2$ refers to the peak instant power of the waveform, and $\frac{1}{T}\int_0^T |s(t)|^2 \ \mathrm{d}t$ is the average power of the received waveform over a time period of $T$.

## 3 I/Q Modulation and Demodulation

In wireless communication, I/Q modulation, as illustrated in Fig. S1, is a technique for efficient transmission of information over radio frequencies, which modulates data onto a carrier signal by varying its in-phase (I) and quadrature (Q) components. I/Q modulation enables complex modulation schemes by combining the I and Q components and, therefore, achieves doubled spectral efficiency by modulating information in both the amplitude and phase of the carrier signal. Consider an I/Q modulation system with a sampling rate of $f_s$ (corresponding to the bandwidth $B = f_s$) and a carrier frequency of $F$. The TX tasks input of a digital, complex-valued I/Q sample sequence $s[n]$. The complex-valued baseband waveform with I and Q components, denoted by $I(t)$ and $Q(t)$, can be respectively constructed from the real and imaginary components of $s[n]$ using two DACs. Plugging in equation (S7), we can formulate the waveform construction process as

$$s(t) = \mathsf{DAC}\left\{s[n]\right\} = I(t) - jQ(t), \ \text{where } I(t) = \mathsf{DAC}\left\{\mathsf{Re}\left\{s[n]\right\}\right\} \ \text{and } Q(t) = -\mathsf{DAC}\left\{\mathsf{Im}\left\{s[n]\right\}\right\}. \tag{S16}$$

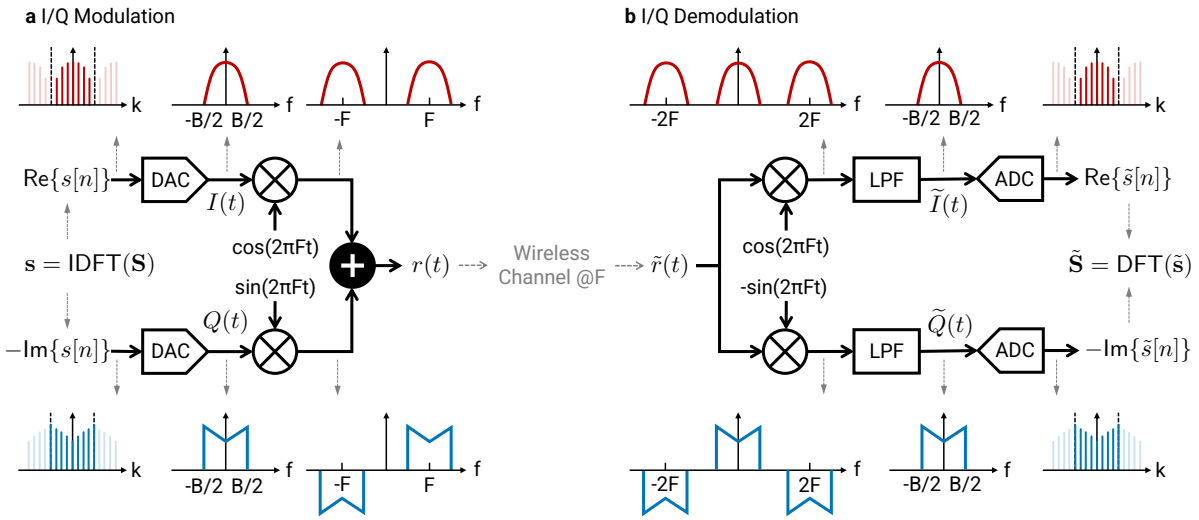


**Fig. S1 The diagram of I/Q modulation and demodulation. a**, The I/Q modulation converts the complex-valued baseband I/Q sequence $\mathbf{s}$ into a real-valued waveform $r(t)$ modulated at carrier frequency $F$. **b**, The I/Q demodulation converts the received real-valued waveform $r(t)$ back to the complex-valued baseband I/Q sequence $\tilde{\mathbf{s}}$, which is proportional to the original $\mathbf{s}$.

Here, the DAC reconstruction kernel can either be sinc interpolation or ZOH. $s(t)$ is then I/Q modulated to carrier frequency $F$ as shown in Fig. S1a, yielding the modulate signal $r(t)$ given by

$$r(t) = I(t) \cdot \cos(2\pi F t) + Q(t) \cdot \sin(2\pi F t) = \mathsf{Re}\left\{ s(t) \cdot e^{j2\pi F t} \right\}. \tag{S17}$$

After over-the-air transmission, the signal received by the RX, $\widetilde{r}(t)$, is I/Q demodulated to recover the baseband waveform, $\widetilde{s}(t)$, using an LO operating at the same carrier frequency $F$ and a low-pass filter (LPF) with a cutoff frequency of $B/2$, as shown in Fig. S1b. This I/Q demodulation process can be written as

$$\widetilde{s}(t) = \mathsf{LPF}\left\{ \widetilde{r}(t) \cdot e^{-j2\pi F t} \right\} = \widetilde{I}(t) - j\widetilde{Q}(t), \text{ where } \widetilde{I}(t) = \mathsf{LPF}\left\{ \widetilde{r}(t) \cos(2\pi F t) \right\} \text{ and } \widetilde{Q}(t) = \mathsf{LPF}\left\{ \widetilde{r}(t) \sin(2\pi F t) \right\}. \tag{S18}$$

Finally, two ADCs operating at the same sampling rate $f_s$ convert the I and Q components of $\widetilde{s}(t)$ into the discrete I/Q samples $\widetilde{s}[n]$, which is

$$\widetilde{s}[n] = \mathsf{ADC}\left\{ \widetilde{s}(t) \right\}, \tag{S19}$$

from which the transmitted I/Q samples $\widetilde{s}[n]$ can be recovered.

## 4 Frequency Mixer as Analog Multiplier

A frequency mixer is a three-port electrical circuit that produces new signals at the sum and difference of the input signal frequencies, which can be used for frequency up-conversion and down-conversion, respectively. The three ports of a frequency mixer are labeled IF (intermediate frequency), RF (radio frequency), and LO (local oscillator). The IF and RF ports can be used as input or output, whereas the LO port always requires an input signal. Essentially, a frequency mixer performs an analog multiplication of the two input waveforms. Without loss of generality, assume two input waveforms $r_1(t)$ and $r_2(t)$ are I/Q modulated to

carrier frequencies at $F_1$ and $F_2$, respectively, given by

$$r_1(t) = I_1(t)\cos(2\pi F_1 t) + Q_1(t)\sin(2\pi F_1 t) \text{ and } r_2(t) = I_2(t)\cos(2\pi F_2 t) + Q_2(t)\sin(2\pi F_2 t). \tag{S20}$$

Let $s_1(t) = I_1(t) - jQ_1(t)$ and $s_2(t) = I_2(t) - jQ_2(t)$ denote the baseband I/Q waveforms corresponding to $r_1(t)$ and $r_2(t)$, respectively. The analog multiplication of the two input waveforms $r_1(t)$ and $r_2(t)$ yields the output waveform $r_o(t)$ at the new frequency $F_o = F_1 \pm F_2$, given by

$$\begin{aligned}
r_o(t) \propto r_1(t)\cdot r_2(t) &= [I_1(t)\cos(2\pi F_1 t) + Q_1(t)\sin(2\pi F_1 t)]\cdot[I_2(t)\cos(2\pi F_2 t) + Q_2(t)\sin(2\pi F_2 t)] \\
&\propto [I_1(t)I_2(t) - Q_1(t)Q_2(t)]\cdot\cos(2\pi(F_1+F_2)t) + [I_1(t)Q_2(t) + Q_1(t)I_2(t)]\cdot\sin(2\pi(F_1+F_2)t) \\
&\quad + [I_1(t)I_2(t) + Q_1(t)Q_2(t)]\cdot\cos(2\pi(F_1-F_2)t) + [-I_1(t)Q_2(t) + Q_1(t)I_2(t)]\cdot\sin(2\pi(F_1-F_2)t) \\
&= \mathsf{Re}\{s_1(t)s_2(t)\}\cdot\cos(2\pi(F_1+F_2)t) - \mathsf{Im}\{s_1(t)s_2(t)\}\cdot\sin(2\pi(F_1+F_2)t) \\
&\quad + \mathsf{Re}\{s_1(t)\overline{s}_2(t)\}\cdot\cos(2\pi(F_1-F_2)t) - \mathsf{Im}\{s_1(t)\overline{s}_2(t)\}\cdot\sin(2\pi(F_1-F_2)t),
\end{aligned} \tag{S21}$$

where $\overline{s}_2(t)$ denotes the conjugate of $s_2(t)$. Note that the output waveform of the mixer $r_o(t)$ at frequency $F_o$ can also be written in the form of

$$r_o(t) = I_o(t)\cos(2\pi F_o t) + Q_o(t)\sin(2\pi F_o t), \tag{S22}$$

where $s_o(t) = I_o(t) - jQ_o(t)$ denotes the corresponding baseband waveform. When the mixer is used for frequency up-conversion, the two input ports are IF and LO, the output port is RF, and their carrier frequency satisfies $F_o = F_1 + F_2$. As long as the carrier frequencies are carefully selected without frequency aliasing, the mismatched frequency component $F_1 - F_2$ will be filtered out by the LPF. Plugging this into equation (S21) yields

$$\begin{cases} I_o(t) = \mathsf{LPF}\{r_o(t)\cdot\cos(2\pi(F_1+F_2)t)\} \propto \mathsf{Re}\{s_1(t)s_2(t)\} \\ Q_o(t) = \mathsf{LPF}\{r_o(t)\cdot\sin(2\pi(F_1+F_2)t)\} \propto -\mathsf{Im}\{s_1(t)s_2(t)\} \end{cases} \Rightarrow s_o(t) = I_o(t) - jQ_o(t) \propto s_1(t)\cdot s_2(t). \tag{S23}$$

When the mixer is used for frequency down-conversion, the RF and LO ports become the input ports for $r_1(t)$ and $r_2(t)$, respectively, and the IF port becomes the output port for $r_o(t)$. In this case, $F_o = F_1 - F_2$, and the frequency component $F_1 + F_2$ is filtered out by the LPF. Similarly, we can derive $s_o(t)$ from equation (S21) as

$$\begin{cases} I_o(t) = \mathsf{LPF}\{r_o(t)\cdot\cos(2\pi(F_1-F_2)t)\} \propto \mathsf{Re}\{s_1(t)\cdot\overline{s}_2(t)\} \\ Q_o(t) = \mathsf{LPF}\{r_o(t)\cdot\sin(2\pi(F_1-F_2)t)\} \propto -\mathsf{Im}\{s_1(t)\cdot\overline{s}_2(t)\} \end{cases} \Rightarrow s_o(t) = I_o(t) - jQ_o(t) \propto s_1(t)\cdot\overline{s}_2(t). \tag{S24}$$

Compared to equation (S23), the only difference in the down-conversion case is the requirement for a conjugate operation on the waveform input to the LO port, $s_2(t)$. In practice, the LO signal supplied to the mixer may exhibit a carrier frequency offset (CFO), $\Delta F$, compared to the desired frequency of the target signal

for up-conversion and down-conversion, this effect can be modeled as

$$\begin{cases} s_o(t) \propto s_1(t) \cdot s_2(t) \cdot e^{j\Delta F t}, \text{ where } F_o = F_1 + F_2 + \Delta F, \text{ for signal up-conversion,} \\ s_o(t) \propto s_1(t) \cdot \overline{s}_2(t) \cdot e^{j\Delta F t}, \text{ where } F_o = F_1 - F_2 + \Delta F, \text{ for signal down-conversion.} \end{cases} \tag{S25}$$

# 5 Orthogonal Frequency-Division Multiplexing (OFDM) System

OFDM is a technique of modulating data symbols onto multiple overlapping but orthogonal subcarriers within a given bandwidth, which is widely used in modern communication systems, including Wi-Fi (e.g., IEEE 802.11n/ac/ax) and cellular (e.g., LTE/5G). Assuming that an OFDM system occupies a bandwidth of $[-B/2, +B/2]$ in the baseband. This bandwidth $B$ is divided into $L$ overlapping but orthogonal subcarriers with a subcarrier spacing of $\Delta f = B/L$. Without loss of generality, $L$ is assumed to be an even number. In the baseband, the $k$-th subcarrier is at frequency

$$f_k = \left(k - \frac{L}{2}\right) \cdot \Delta f = \frac{k - L/2}{L} \cdot B, \ \forall k = 0, \ldots, L-1. \tag{S26}$$

Generally, the OFDM system is structured by OFDM symbols in the time domain. Within one OFDM symbol, the time domain I/Q samples are converted to/from frequency domain data symbols using an $L$-point DFT, so there are $L$ I/Q samples per OFDM symbol in the time domain corresponding to the $L$ subcarriers in the frequency domain. Hereby, we denote the time domain I/Q samples of an OFDM symbol as $\mathbf{s} = [s[n]] \in \mathbb{C}^L$, and its frequency domain data symbols as $\mathbf{S} = [S[k]] \in \mathbb{C}^L$. The data symbols can be derived from the I/Q samples via an $L$-point DFT, i.e.,

$$\mathbf{S} = \mathbf{R}^{L/2} \cdot \mathsf{DFT}(\mathbf{s}) = \sqrt{L} \cdot \mathbf{R}^{L/2} \cdot \mathbf{D} \cdot \mathbf{s}, \tag{S27}$$

where $\mathbf{R}^{L/2}$ is the circular shift matrix that shifts the zero-frequency (DC) subcarrier symbol originally indexed at $k = 0$ to the center of the spectrum at $k = L/2$. As a result, we have

$$S[k] = \sum_{n=0}^{L-1} s[n] \cdot e^{-j2\pi \frac{k-L/2}{L} n}, \ \forall k = 0, \ldots, L-1. \tag{S28}$$

Similarly, the time domain I/Q waveform can be derived from the circularly shifted frequency domain data symbols using an $L$-point IDFT given by

$$\mathbf{s} = \mathsf{IDFT}(\mathbf{R}^{L/2} \cdot \mathbf{S}) = \frac{1}{\sqrt{L}} \cdot \mathbf{D}^{-1} \cdot \mathbf{R}^{L/2} \cdot \mathbf{S}, \tag{S29}$$

where the recovered time domain I/Q samples $\mathbf{s} = [s[n]] \in \mathbb{C}^L$ is given by

$$s[n] = \frac{1}{L} \sum_{k=0}^{L-1} S[k] \cdot e^{j2\pi \frac{k-L/2}{L} n}, \ \forall n = 0, \ldots, L-1. \tag{S30}$$

Based on the Nyquist-Shannon sampling theorem, the minimum sampling required for a system employing I/Q modulation to (re)construct the signal without aliasing is $f_s = B$. Under this sampling rate, an OFDM symbol has a duration of $T = L/f_s = L/B$.

On the TX side, we consider an ideal DAC reconstruction kernel for the OFDM system that substitutes $n$ by $f_s \cdot t$ in equation (S30). Correspondingly, the length-$L$ sequence of I/Q samples becomes a waveform that lasts for a time duration of $L/f_s$. In this case, the transmitted waveform can be written as

$$s(t) = \mathsf{DAC}\left\{s[n]\right\} = \sum_{k=0}^{L-1} S[k] \cdot e^{j2\pi \frac{k-L/2}{T} t}, \ \forall t \in [0, T). \tag{S31}$$

This process can also be described using the Fourier series $\mathcal{F}(\cdot)$, given by

$$s(t) = \mathcal{F}(\mathbf{S}) = \sum_{k=-L/2}^{L/2-1} S[k] \cdot e^{j2\pi \frac{k}{T} t}, \ \forall t \in [0, T). \tag{S32}$$

This baseband waveform is then modulated to carrier frequency $F$ as $r(t) = \mathsf{Re}\left\{s(t) \cdot e^{j2\pi F t}\right\}$.

On the RX side, the received waveform $\widetilde{r}(t)$ at carrier frequency $F$ is I/Q demodulated to $\widetilde{s}(t)$, which is then filtered by an LPF with a cutoff frequency of $f_s/2$, and then sampled by two ADCs at the sampling rate of $f_s$ to acquire the I/Q samples $\widetilde{\mathbf{s}} = [\widetilde{s}[n]]$, i.e.,

$$\widetilde{s}[n] = \mathsf{ADC}\left\{\mathsf{LPF}\left\{\widetilde{s}(t)\right\}\right\}, \ n = 0, 1, \ldots, L-1. \tag{S33}$$

The symbols $\widetilde{\mathbf{S}}$ can then be recovered by equation (S27).

During wireless transmissions, multipath propagation causes delayed copies of the transmitted signal to arrive at the receiver, leading to potential inter-symbol interference (ISI). To mitigate this effect, a replica of the ending I/Q samples in $\mathbf{s}$ is appended to the beginning of $\mathbf{s}$ as the *cyclic prefix*, ensuring that multipath delays do not cause overlap between consecutive OFDM symbols. Assume that a cyclic prefix of $\Delta L$ I/Q samples, extended OFDM symbol with cyclic prefix, $\mathbf{s}' \in \mathbb{C}^{L+\Delta L}$, can be written as

$$s'[n] = \begin{cases} s[L + n - \Delta L], & \text{if } n < \Delta L, \\ s[n - \Delta L], & \text{if } n \geq \Delta L, \end{cases} \quad \forall n = 0, 1, \ldots, L + \Delta L - 1. \tag{S34}$$

Consider a timing delay of $\Delta n$ I/Q samples with $\Delta n \leq \Delta L$, the received $\widetilde{\mathbf{s}} \in \mathbb{C}^L$ after removing the cyclic prefix is

$$\widetilde{s}[n] \propto \begin{cases} s[L + n - \Delta L + \Delta n], & \text{if } n < \Delta L - \Delta n, \\ s[n - \Delta L + \Delta n], & \text{if } n \geq \Delta L - \Delta n, \end{cases} \quad \forall n = 0, 1, \ldots, L-1. \tag{S35}$$

This is equivalent to the original transmitted $\mathbf{s}$ with a circular shift of $(\Delta L - \Delta n)$ I/Q samples. According to the DFT shifting theorem, it holds that

$$\widetilde{S}[k] \propto S[k] \cdot e^{-j2\pi \frac{(\Delta L - \Delta n)}{L} k}, \ \forall k = 0, 1, \ldots, L-1. \tag{S36}$$

This means that the received $\widetilde{\mathbf{S}}$ is proportional to the desired $\mathbf{S}$ with a phase shift of $2\pi(\Delta L - \Delta n)k/L$. Since the timing delay $\Delta n$ is a constant over OFDM symbols, it can be estimated using a reference OFDM symbol and then used to calibrate the remaining OFDM symbols.

# 6 Hybrid Convolution Theorem

The convolution theorem [53] states that the multiplication of two time domain signals equals the convolution of their frequency domain spectrums. Specifically, there are two representations of the convolution theorem in the analog and digital domains: $(i)$ in the analog domain, the multiplication of two continuous waveforms corresponds to the linear convolution of their spectrums, and $(ii)$ in the digital domain, the element-wise multiplication of two discrete I/Q samples corresponds to the circular convolution. We consider a hybrid convolution theorem of these two, which is built on the discrete I/Q samples while it corresponds to the linear convolution in the frequency domain.

Consider an OFDM system with an FFT size of $L$ and subcarrier spacing of $\Delta f$. Let $\mathbf{S}_1 = [S_1[k]] \in \mathbb{C}^L$ and $\mathbf{S}_2 = S_2[k] \in \mathbb{C}^L$ denote two frequency-domain OFDM symbols, whose time-domain waveforms are given by $s_1(t)$ and $s_2(t),\ t \in [0, T)$, where $T = 1/\Delta f$. According to equation (S31),

$$s_1(t) = \mathcal{F}(\mathbf{S}_1) = \sum_{k=0}^{L-1} S_1[k] \cdot e^{j2\pi \frac{k-L/2}{T} t},\ s_2(t) = \mathcal{F}(\mathbf{S}_2) = \sum_{k=0}^{L-1} S_2[k] \cdot e^{j2\pi \frac{k-L/2}{T} t},\ \forall t \in [0, T). \tag{S37}$$

Let '$*$' denote the linear convolution operation that maps $\mathbb{C}^L * \mathbb{C}^L \rightarrow \mathbb{C}^{2L-1}$. Specifically, the linear convolution of two OFDM symbols, denoted by $\mathbf{S}_o = [S_o[k]] \in \mathbb{C}^{2L-1}$, is given by

$$\mathbf{S}_o = \mathbf{S}_1 * \mathbf{S}_2,\ \text{where } S_o[k] = \sum_{\kappa=\max\{0,k-L+1\}}^{\min\{L-1,k\}} S_1[\kappa] \cdot S_2[k-\kappa],\ \forall k = 0, 1, \ldots, 2L-2. \tag{S38}$$

Note that the output symbol $\mathbf{S}_o$ has an extended length of $(2L-1)$ while maintaining the same subcarrier spacing $(\Delta f)$ and waveform time $(T = 1/\Delta f)$. According to equation (S31), its time-domain waveform, $s_o(t)$, is given by

$$s_o(t) = \mathcal{F}(\mathbf{S}_o) = \sum_{k=0}^{2L-2} S_o[k] \cdot e^{j2\pi \frac{k-(2L-1)/2}{T} t} = \sum_{k=0}^{2L-2} \left( \sum_{\kappa} S_1[\kappa] \cdot S_2[k-\kappa] \right) \cdot e^{j2\pi \frac{k-(2L-1)/2}{T} t},\ \forall t \in [0, T). \tag{S39}$$

Notice that in equation (S37), symbol $S_1[\kappa]$ is located at frequency $\frac{\kappa - L/2}{T}$, and symbol $S_2[k-\kappa]$ is located at frequency $\frac{k-\kappa-L/2}{T}$. The multiplication of these two terms results in a symbol located at frequency $\frac{k-L}{T}$, which has a frequency shift of $\Delta f/2$ compared to the symbol $S_o[k]$ located at frequency $\frac{k-(2L-1)/2}{T}$ in equation (S39). Therefore, the following relationship between the time-domain waveforms holds,

$$s_o(t) = s_1(t) \cdot s_2(t) \cdot e^{j2\pi \frac{\Delta f}{2} t} = s_1(t) \cdot s_2(t) \cdot e^{j\pi \Delta f t}. \tag{S40}$$

In summary, the hybrid convolution theorem can be written as

$$s_1(t) \cdot s_2(t) \cdot e^{j\pi\Delta f \cdot t} = \mathcal{F}\big((\mathbf{S}_1 * \mathbf{S}_2)\big), \text{ where } s_1(t) = \mathcal{F}(\mathbf{S}_1) \text{ and } s_2(t) = \mathcal{F}(\mathbf{S}_2). \tag{S41}$$

The bandwidth of $\mathbf{S}_1$ or $\mathbf{S}_2$ is given by $B = L \cdot \Delta f$; as for $\mathbf{S}_o$, the subcarrier spacing remains the same while the FFT size becomes $(2L-1)$, so $\mathbf{S}_o$ occupies a bandwidth of $(2L-1) \cdot \Delta f$. This means that to capture the full spectral information of $s_o(t)$, it is required that the ADC after I/Q demodulation operates at a minimum sampling rate of

$$f_s' = (2L-1) \cdot \Delta f. \tag{S42}$$

Recall from Supplementary Section 4 that a computing mixer performs analog multiplication of two signals, similar to the form of the convolution theorem in equation (S41). Hence, we can configure the computing mixer to calculate the convolution between two discrete signals. In the case of frequency up-conversion (equation (S23)), the carrier frequency of the output signal, $F_o$, satisfies

$$F_o = F_1 + F_2 + \Delta f/2, \tag{S43}$$

which cancels the frequency shift term $e^{j\pi\Delta f \cdot t}$. In the case of frequency down-conversion (equation (S24)), $F_o$ satisfies

$$F_o = F_1 - F_2 + \Delta f/2, \tag{S44}$$

and an extra flipping on $\mathbf{S}_2$ is needed to incorporate the conjugated waveform $\bar{s}_2(t)$ in equation (S24), given by

$$S_2'[k] = S_2[L-1-k], \ \forall k = 0, 1, \ldots, L-1. \tag{S45}$$

# 7 In-Physics MVM Computation Based on Frequency-Encoded OFDM Symbols

In this section, we present a subcarrier mapping algorithm that converts the linear convolution to the in-physics MVM computation. Furthermore, the OFDM system, as discussed in Supplementary Section 5, is employed to efficiently convert the time domain I/Q samples from/to the frequency domain subcarrier symbols by IFFT/FFT. Without loss of generality, we consider the case where the computing mixer performs signal up-conversion.

## A. Subcarrier Mapping Algorithm

Consider the MVM between a complex-valued matrix $\mathbf{W} = [W_{m,n}] \in \mathbb{C}^{M \times N}$ and a complex-valued vector $\mathbf{x} = [x_n] \in \mathbb{C}^N$, and their MVM results is given by $\mathbf{y} = \mathbf{W} \cdot \mathbf{x} = [y_m] \in \mathbb{C}^M$. Consider an OFDM system with a subcarrier spacing of $\Delta f$ and an FFT size of $L = NM$, occupying a signal bandwidth of $B = L \cdot \Delta f = NM \cdot \Delta f$. We encode $\mathbf{W}$ and $\mathbf{x}$ into OFDM symbols $\mathbf{S}_w = [S_w[k]] \in \mathbb{C}^L$ and $\mathbf{S}_x = [S_x[k]] \in \mathbb{C}^L$,

respectively. This encoding process essentially maps elements of $\mathbf{W}$ and $\mathbf{x}$ onto different subcarriers of their respective OFDM symbol. The encoding of $\mathbf{W}$ into $\mathbf{S}_w$ is given by

$$S_w[k] = W_{m,n},\ \forall n = 0, \ldots, N-1,\ \forall m = 0, \ldots, M-1,\ \text{and}\ \forall k = NM - m - nM - 1, \tag{S46}$$

and the encoding of $\mathbf{x}$ into $\mathbf{S}_x$ is given by

$$S_x[k] = \begin{cases} x_n, & \text{if } k = n \cdot M,\ \forall n = 0, \ldots, N-1, \\ 0, & \text{otherwise.} \end{cases} \tag{S47}$$

According to the OFDM system, the I/Q waveforms corresponding to $\mathbf{S}_w$ and $\mathbf{S}_x$ can be derived by equations (S29)–(S30). Specifically, let $\mathbf{s}_w = [s_w[n]] \in \mathbb{C}^L$ denote the I/Q waveform corresponding to $\mathbf{W}$, which is obtained by an $L$-point IDFT given by

$$s_w[n] = \frac{1}{L} \sum_{k=0}^{L-1} S_w[k] \cdot e^{j2\pi \frac{k-L/2}{L} n} = \frac{1}{NM} \sum_{n=0}^{N} \sum_{m=0}^{M} W_{m,n} \cdot e^{-j2\pi \frac{1+m+nM+NM/2}{NM} n},\ \forall n = 0, 1, \ldots, L-1. \tag{S48}$$

Similarly, the I/Q waveform $\mathbf{s}_x = [s_x[n]] \in \mathbb{C}^L$ corresponding to $\mathbf{x}$ is given by

$$s_x[n] = \frac{1}{L} \sum_{k=0}^{L-1} S_x[k] \cdot e^{j2\pi \frac{k-L/2}{L} n} = \frac{1}{NM} \sum_{n=0}^{N} x_n \cdot e^{-j2\pi \frac{n+N/2}{N} n},\ \forall n = 0, 1, \ldots, L-1. \tag{S49}$$

We assume ideal DACs following equation (S31) at the sampling rate of $f_s = L \cdot \Delta f$, waveforms carrying $\mathbf{W}$ and $\mathbf{x}$ have the same duration $T$ given by

$$T = \frac{1}{\Delta f} = \frac{L}{B} = \frac{NM}{B}. \tag{S50}$$

Specifically, the waveform $w(t)$ carrying $\mathbf{W}$ is

$$w(t) = \mathsf{DAC}\left\{\mathbf{s}_w\right\} = \mathcal{F}(\mathbf{S}_w) = \frac{1}{NM} \sum_{n=0}^{N} \sum_{m=0}^{M} W_{m,n} \cdot e^{-j2\pi \frac{1+m+nM+NM/2}{T} t},\ \forall t \in [0, T), \tag{S51}$$

and the waveform $x(t)$ carrying $\mathbf{x}$ is

$$x(t) = \mathsf{DAC}\left\{\mathbf{s}_x\right\} = \mathcal{F}(\mathbf{S}_x) = \frac{1}{NM} \sum_{n=0}^{N} x_n \cdot e^{-j2\pi \frac{n+N/2}{T} Mt},\ \forall t \in [0, T). \tag{S52}$$

Assume that waveforms carrying $\mathbf{W}$, $\mathbf{x}$, and $\mathbf{y}$ are I/Q modulated to carrier frequencies of $F_w$, $F_x$, $F_y$, respectively. For analog computing using the computing mixer for frequency up-conversion (S43), it holds that

$$F_y = F_x + F_w + \Delta f/2 \Rightarrow y(t) = x(t) \cdot w(t) \cdot e^{j\pi \Delta f \cdot t},\ \forall t \in [0, T). \tag{S53}$$

The output waveform $y(t)$ spans a frequency over $(2L-1)\cdot\Delta f$. Therefore, the time-domain I/Q samples $\mathbf{s}_y=[s_y[n]]\in\mathbb{C}^{2L-1}$ can be captured by a pair of ADCs operating at the sampling rate of $(2L-1)\cdot\Delta f$, or a per-sample duration of $T/(2L-1)$,

$$s_y[n]=\mathsf{ADC}\left\{y(t)\right\}=y\left(\frac{n}{2L-1}\cdot T\right),\ \forall n=0,1,\ldots,2L-2. \tag{S54}$$

Finally, the frequency domain subcarrier symbols $\mathbf{s}_y=[s_y[k]]\in\mathbb{C}^{2L-1}$ can be acquired by a $(2L-1)$-point DFT as

$$S_y[k]=\sum_{n=0}^{2L-2}s_y[n]\cdot e^{-j2\pi\frac{k}{2L-1}n},\ \forall k=0,1,\ldots,2L-1. \tag{S55}$$

According to the hybrid convolution theorem (S41) (see Supplementary Section 6), the output symbols $\mathbf{S}_y=[S_y[k]]\in\mathbb{C}^{2L-1}$ as the convolution between $\mathbf{S}_w$ and $\mathbf{S}_x$, with "extended" frequency components is given by

$$\mathbf{S}_y=\mathbf{S}_w*\mathbf{S}_x,\ \text{where}\ S_y[k]=\sum_{\kappa=\max\{0,k-L+1\}}^{\min\{L-1,k\}}S_w[\kappa]\cdot S_x[k-\kappa],\ \forall k=0,1,\ldots,2L-2. \tag{S56}$$

Note that the output symbols carried by the middle $M$ subcarriers indexed at $S_y[NM-M,\ldots,NM-1]$ satisfy

$$\begin{aligned}S_y[NM-1-m]&=\sum_{\kappa=0}^{NM-m}S_x[\kappa]\cdot S_w[NM-1-m-\kappa]=\sum_{n=0}^{N-1}S_x[nM]\cdot S_w[NM-1-m-nM]\\&=\sum_{n=0}^{N-1}x_n\cdot W_{m,n}=y_m,\ \forall m=0,1,\ldots,M-1.\end{aligned} \tag{S57}$$

This means that the desired output vector $\mathbf{y}=\mathbf{W}\cdot\mathbf{x}$ is embedded in the spectrum of $\mathbf{S}_y$. As a result, the output waveform $y(t)$ can then be I/Q demodulated and sampled using an ADC at a sampling rate of $f_s'=(2L-1)\Delta f$ to acquire I/Q samples $\mathbf{s}_y=[s_y[n]]\in\mathbb{C}^{2L-1}$ with no frequency aliasing. Finally, the output symbol $\mathbf{S}_y=[S_y[k]]\in\mathbb{C}^{2L-1}$ can be obtained by a $(2L-1)$-point DFT,

$$S_y[k]=\sum_{n=0}^{2L-1}s_y[n]\cdot e^{-j2\pi\frac{k-L'/2}{L'}n},\ \forall k=0,1,\ldots,2L-2, \tag{S58}$$

from which the MVM result, $\mathbf{y}$, can be extracted as

$$y_m=S_y[NM-1-m],\ \forall m=0,1,\ldots,M-1. \tag{S59}$$

A similar analysis also holds in the case where the computing mixer performs frequency down-conversion based on equation (S45) with $F_y=F_x-F_w+\Delta f/2$.

## B. Energy Efficiency Analysis

We analyze the *energy efficiency* of this "vanilla" in-physics MVM computation on the OFDM system, i.e., energy consumed per MAC operation ($e_{\text{mvm}}$), and the *computation efficiency*, i.e., the number of MAC operations per second per Watt ($e_{\text{mvm}}^{-1}$). Specifically, there are three energy consumption components: $(i)$ $E_1$ for the waveform generation of $x(t)$ and I/Q (de)modulation, $(ii)$ $E_2$ for the I/Q sampling of waveform $y(t)$ using two ADCs after I/Q demodulation, and $(iii)$ $E_3$ for the digital computing based encoding (prior to waveform generation) and decoding (after waveform reception). Note that we only include the energy on the client while excluding that by the central radio broadcasting $w(t)$.

We first derive $E_1$ as follows. Let $P_x$ denote the transmit power of $x(t)$ with a radio hardware efficiency of $\eta_{\text{radio}}$. The total energy required for the client radio to generate the waveform carrying the inference request, $\mathbf{x}$, is given by

$$E_1 = (\eta_{\text{radio}})^{-1} P_x \cdot T = (\eta_{\text{radio}})^{-1} P_x \cdot \frac{NM}{B}. \tag{S60}$$

Let $\eta_{\text{mixer}}$ denote the efficiency of the computing mixer. The received signal power at the RX after the in-physics computing process carried out by the computing mixer is given by

$$P_y = \eta_{\text{mixer}} \cdot P_x. \tag{S61}$$

At radio frequency, the thermal noise power spectrum density is given by $kT_0 = -174\,\text{dBm/Hz}$, where $k = 1.38 \times 10^{-23}\,\text{J/K}$ is the Boltzmann constant and $T_0 = 300\,\text{K}$ is the room temperature. After the mixing, the bandwidth of $\mathbf{S}_y$ is $(2L-1)\cdot\Delta f$, so the total noise power over this bandwidth, $P_{\text{noise}}$, is given by

$$P_n = (\eta_{\text{nf}})^{-1} \cdot k_B T_0 \cdot (2L-1) \cdot \Delta f = (\eta_{\text{nf}})^{-1} \cdot k_B T_0 \cdot (2NM-1) \cdot \Delta f \approx (\eta_{\text{nf}})^{-1} \cdot k_B T_0 \cdot 2B, \tag{S62}$$

where $(\eta_{\text{nf}})^{-1}$ denotes the noise figure of the RX. Combining equations (S61) and (S62) yields

$$\mathsf{SNR} = \frac{P_y}{P_n} \approx \frac{\eta_{\text{mixer}} \cdot P_x}{(\eta_{\text{nf}})^{-1} \cdot k_B T_0 \cdot 2B} \Rightarrow P_x = 2(\eta_{\text{mixer}} \cdot \eta_{\text{nf}})^{-1} \cdot \mathsf{SNR} \cdot k_B T_0 \cdot B. \tag{S63}$$

Plugging in $P_x$ from equation (S60) and denote $\eta = \eta_{\text{radio}} \cdot \eta_{\text{mixer}} \cdot \eta_{\text{nf}}$ as the overall hardware efficiency, we have

$$E_1 = 2NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0. \tag{S64}$$

$E_2$ represents the energy consumption by the ADC, which is proportional to the number of captured I/Q samples. Given a per-sample energy consumption of $e_{\text{adc}}$ (e.g., $e_{\text{adc}} = 1\,\text{pJ/sample}$ [52]), the capturing of $\mathbf{s}_y$ consisting of $(2L-1)$ complex-valued I/Q samples incurs a total energy consumption of

$$E_2 = (2NM-1) \cdot 2e_{\text{adc}} \approx 4NM \cdot e_{\text{adc}}. \tag{S65}$$

Finally, the digital computing energy $E_3$ is proportional to the number of real-valued MACs, where we denote the energy consumption per real-valued MAC in the state-of-the-art ASICs as $e_{\text{dig}}$. This term includes

**Table S1** Comparison of energy consumption and energy efficiency complexity analysis between the vanilla in-physics matrix-vector multiplication (MVM) computation and the three WISE schemes.

| Scheme | $E_1$ | $E_2$ | $E_3$ | | | Energy per MVM* ($E_{mvm}$) | Energy per MAC* | |
|---|---|---|---|---|---|---|---|---|
| | | | Precoding | IFFT Enc. | FFT Dec. | | MVM ($e_{mvm}$) | IP ($e_{ip}$) |
| Vanilla Scheme | $\mathcal{O}(MN)$ (<TDL) | $\mathcal{O}(M)$ | – | $\mathcal{O}(MN\log MN)$ | $\mathcal{O}(MN\log MN)$ | $\mathcal{O}(MN\log MN)$ | $\mathcal{O}(\log MN)$ | $\mathcal{O}(\log N)$ |
| Basic Scheme | | | – | $\mathcal{O}(N\log N)$ | $\mathcal{O}(M)$ | $\mathcal{O}(N\log N+M)$ | $\mathcal{O}\left(\frac{\log N}{M}+\frac{1}{N}\right)$ | $\mathcal{O}(\log N)$ |
| **W**-Precoding | | | – | – | $\mathcal{O}(M)$ | $\mathcal{O}(M)$ | $\mathcal{O}\left(\frac{1}{N}\right)$ | $\mathcal{O}\left(\frac{1}{N}\right)$ |
| **x**-Precoding | | | $\mathcal{O}(N)$ | $\mathcal{O}(N\log N)$ | $\mathcal{O}(M)$ | $\mathcal{O}(N\log N+M)$ | $\mathcal{O}\left(\frac{\log N}{M}+\frac{1}{N}\right)$ | $\mathcal{O}(\log N)$ |

* Excluding $E_1$ or $e_1$, which are lower than the thermodynamic limit (TDL)

the energy consumption associated with an $L$-point IFFT (encoding of $\mathbf{S}_x$ to $\mathbf{s}_x$, equation (S49)) and a $(2L-1)$-point FFT (decoding of $\mathbf{s}_y$ to $\mathbf{S}_y$, equation (S55)), i.e.,

$$\begin{aligned} E_3 &= [2L\cdot\log_2 L + 2(2L-1)\cdot\log_2(2L-1)]\cdot e_{\text{dig}} \\ &= [2NM\cdot\log_2(NM) + 2(2NM-1)\cdot\log_2(2NM-1)]\cdot e_{\text{dig}} \\ &\approx 6NM\cdot\log_2(NM)\cdot e_{\text{dig}}. \end{aligned} \tag{S66}$$

Putting equations (S64)–(S66) together, the total energy consumption $E_{\text{mvm}}$ is given by

$$E_{\text{mvm}} = E_1+E_2+E_3 = \underbrace{2NM\cdot\eta^{-1}\cdot\mathsf{SNR}\cdot k_BT_0}_{E_1} + \underbrace{4NM\cdot e_{\text{adc}}}_{E_2} + \underbrace{6NM\cdot\log_2(NM)\cdot e_{\text{dig}}}_{E_3}. \tag{S67}$$

The corresponding energy efficiency, measured by energy per MAC, $e_{\text{mvm}}$, is

$$e_{\text{mvm}} = \frac{E_{\text{mvm}}}{4NM} = e_1+e_2+e_3 = \underbrace{\frac{1}{2}\cdot\eta^{-1}\cdot\mathsf{SNR}\cdot k_BT_0}_{e_1} + \underbrace{e_{\text{adc}}}_{e_2} + \underbrace{\frac{3}{2}\cdot\log_2(NM)\cdot e_{\text{dig}}}_{e_3}. \tag{S68}$$

It can be seen that $e_{\text{mvm}}$ is dominated by the term $e_3$ and scales as $\mathcal{O}(\log(NM))$ as the problem size grows.

Therefore, WISE incorporates several optimization strategies to significantly reduce the energy consumption for in-physics MVM computation. As shown in Table S1, there are three schemes for WISE: $(i)$ the basic scheme without wireless channel precoding, designed for the wired case or when the CSI is not available; $(ii)$ the **W**-precoding scheme, which precodes the **W** on the central radio without incurring additional energy consumption on the clients while further reducing the energy consumption by time-encoding **x**; $(iii)$ the **x**-precoding scheme, which precodes **x** on the client that supports individual CSI for each client for higher computing accuracy. The **W**-precoding scheme is evaluated in Section 2, and the detailed performance of the other two schemes can be found in Supplementary Section 13.

# 8 WISE's Basic Scheme

We first present a basic scheme of WISE for a wired/cabled channel that significantly optimizes the energy efficiency of the in-physics MVM described in Supplementary Section 7 via three techniques: (*i*) encoding FFT size reduction, (*ii*) ADC sampling rate reduction, and (*iii*) MVM decomposition. We also introduce zero-subcarrier padding and cyclic prefix to overcome two practical issues stemming from the three techniques.

## A. Encoding FFT Size Reduction

The first technique reduces the FFT size from $NM$ to $N$ to save the number of real-valued MACs required for encoding $\mathbf{S}_x$ into $\mathbf{s}_x$. Note from equation (S49) that the generated waveform $\mathbf{s}_x$ is independent of the output size, $M$, and is with a period of $N$ samples, i.e.,

$$\begin{aligned} s_x[n+N] &= \frac{1}{NM}\sum_{n=0}^{N-1} x_n \cdot e^{-j2\pi\frac{n+N/2}{N}(n+N)} \\ &= \frac{1}{NM}\sum_{n=0}^{N-1} x_n \cdot e^{-j2\pi\frac{n+N/2}{N}n} = s_x[n],\ \forall n = 0,1,\ldots,NM-N-1. \end{aligned} \tag{S69}$$

Therefore, we only need to generate the first $N$ I/Q samples in $\mathbf{s}_x$, which can then be repeated for $M$ times to obtain $\mathbf{s}_x$. The generation of the first $N$ I/Q samples has a similar form as an $N$-point IDFT as

$$\mathbf{s}_x = \frac{1}{M\sqrt{N}} \cdot \underbrace{\left[\mathbf{D}^{-1}\cdot\mathbf{R}^{N/2}\cdot\mathbf{x},\ldots,\mathbf{D}^{-1}\cdot\mathbf{R}^{N/2}\cdot\mathbf{x}\right]}_{\text{repeated } M \text{ times}}, \tag{S70}$$

but requires only an $N$-point IFFT involving $2N\log_2 N$ real-valued MACs.

## B. ADC Sampling Rate Reduction

An RX operating at the sampling rate of $(2NM-1)\cdot\Delta f$ is required to capture a total number of $(2NM-1)$ I/Q samples of $\mathbf{s}_y$ in order to recover the full spectral information of $\mathbf{S}_y$. This process incurs significant energy consumption on the waveform reception term $e_2$ on ADC, and the digital computing term $e_3$ for FFT. Fortunately, in equation (S59), we notice that the output vector $\mathbf{y}$ can be demodulated from a set of consecutive subcarriers located in the middle of the spectrum of $\mathbf{S}_y$, indexed from $(NM-M)$ to $(NM-1)$. Let $\mathbf{S}_{y\downarrow} = [S_{y\downarrow}[k]] \in \mathbb{C}^M$, where $S_{y\downarrow}[k] = S_y[NM-M+k], \forall k = 0,1,\ldots,M-1$, equation (S59) can be written as

$$y_m = S_{y\downarrow}[M-1-m],\ \forall m = 0,1,\ldots,M-1. \tag{S71}$$

These $M$ subcarriers in $\mathbf{S}_{y\downarrow}$ only occupy a narrow bandwidth of $B_\downarrow = M\cdot\Delta f$, which is a fraction $1/N$ of the original signal bandwidth of $\mathbf{W}$ and $\mathbf{x}$. Therefore, one can employ an LPF with a cutoff frequency of $B_\downarrow/2$ and an ADC with a sampling rate of as small as $f_{s\downarrow} = B_\downarrow$ to capture the waveform $y(t)$, which yields

$$\mathbf{s}_{y\downarrow} = [s_{y\downarrow}[n]] = \mathsf{ADC}\left\{\mathsf{LPF}\left\{y(t)\right\}\right\},\ \forall t \in [0,T),\ n = 0,1,\ldots,M-1. \tag{S72}$$

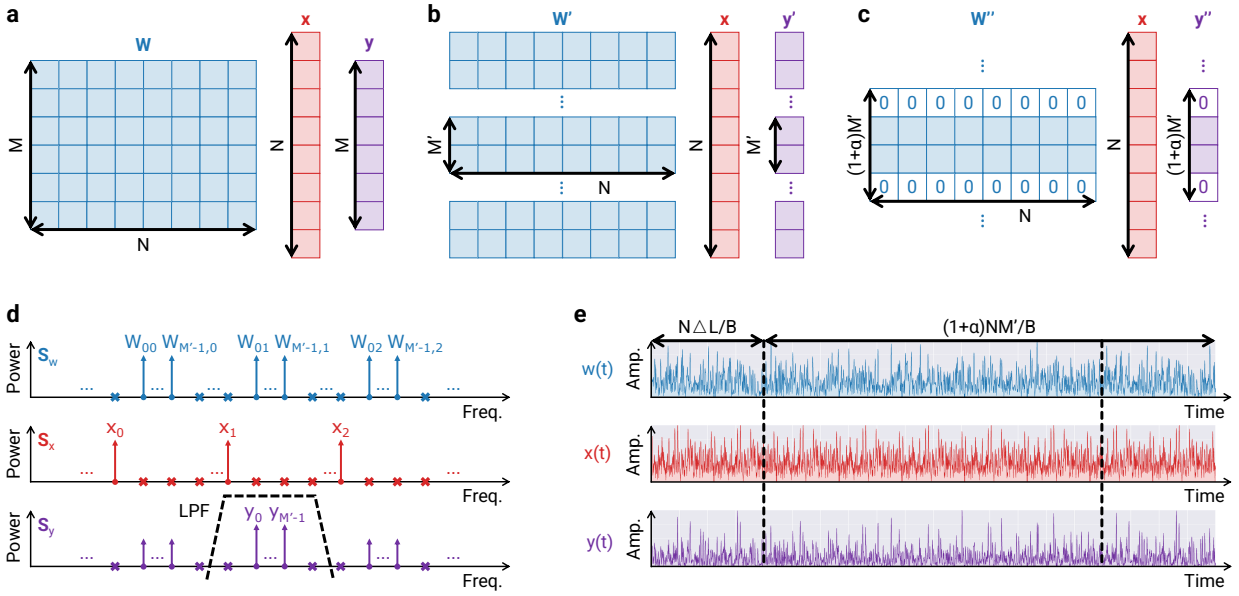


**Fig. S2 The step-by-step waveform generation of WISE's MVM computation. a**, The original matrix-vector multiplication (MVM), $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, with input size $N$ and output size $M$. **b**, Decomposition of the original MVM into small MVMs along the output dimension using a reduced output size, $M'$, which lowers the complexity of DFT-based decoding. **c**, Zero-padding is applied to $\mathbf{W}$ and $\mathbf{y}$, introducing all-zero rows that correspond to zero-subcarriers in the signal spectrum $\mathbf{S}_w$ and $\mathbf{S}_{y\downarrow}$. **d**, The inserted zero-subcarriers in $\mathbf{S}_{y\downarrow}$ effectively mitigate the roll-off effect introduced by the low-pass filter (LPF). **e**, In the time domain, a cyclic prefix is added prior to the signal to mitigate potential synchronization errors.

In this way, only $M$ I/Q samples are captured, and the decoding of $\mathbf{y}$ from $\mathbf{S}_{y\downarrow}$ can be done using an $M$-point FFT following equation (S55) as

$$\mathbf{S}_{y\downarrow} = \sqrt{M} \cdot \mathbf{R}^{M/2} \cdot \mathbf{D} \cdot \mathbf{s}_{y\downarrow}, \text{ where } S_{y\downarrow}[k] = \sum_{n=0}^{M-1} s_y[n] \cdot e^{-j2\pi \frac{k-M/2}{M} n}, \ \forall k = 0, 1, \ldots, M-1. \tag{S73}$$

In this way, only $M$ I/Q samples need to be captured by the ADC on the RX side, and the subsequent decoding of $\mathbf{y}$ involves $2M \log_2 M$ real-valued MACs.

## C. MVM Decomposition

Furthermore, we can decompose the large MVM $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$ in the output dimension $M$ into $M/M'$ smaller MVMs, $\mathbf{y}' = \mathbf{W}' \cdot \mathbf{x}$, as shown in Fig. S2a–b, with $\mathbf{W}' \in \mathbb{C}^{M' \times N}$ and $\mathbf{y}' \in \mathbb{C}^{M'}$. Since each smaller MVM requires only an $M'$-point FFT for decoding that involves $2M' \log_2 M'$ MACs, the total number of MACs required for all $M/M'$ MVMs reduces to $2M \log_2 M'$. Note that the waveform time for a single decomposed MVM is $M'N/B$, and the total computation time for the original MVM remains $\frac{M}{M'} \cdot \frac{M'N}{B} = \frac{MN}{B}$, which matches the waveform duration $T$ of the original MVM computation without decomposition. Therefore, this MVM decomposition incurs no additional overhead in waveform time, and ensures that the energy consumption and computation throughput remain unchanged. This MVM decomposition also proportionally reduces the number of subcarriers in both the TX channel of $\mathbf{S}_w$ and $\mathbf{S}_x$, i.e., $NM'$ subcarriers, and the RX channel of $\mathbf{S}_{y\downarrow}$, i.e., $M'$ subcarriers. As a result, the downsampling ratio from TX to RX remains $N$, and given the same available bandwidth $B$, the TX and RX sampling rates remain unchanged.

## D. Zero-Subcarrier Padding

To mitigate the frequency aliasing effects, the ADC operating at a low sampling rate of $M'\Delta f$ relies on an ideal LPF to filter out frequency components outside of $[-M'\Delta f/2, +M'\Delta f/2]$. An ideal LPF has a brick-wall shape, corresponding to a flat frequency response across the passband, and "zero" frequency response elsewhere. However, such a brick-wall LPF is not practical due to its non-casualty in the time domain; rather, a practical LPF before ADC usually exhibits non-negligible roll-off effects around its cutoff frequency at $M'\Delta f/2$, i.e., there exists a transient frequency range around $M'\Delta f/2$ where the frequency response gradually drops to zero (see Supplementary Section 12 for the detailed measurements).

To overcome the LPF's roll-off effect, zero subcarriers that carry no symbols are padded to the LPF's transient frequency. Consider a padded weight matrix $\mathbf{W}'' \in \mathbb{C}^{(M'+2\Delta M)\times N}$, where $\Delta M$ rows of 0's are padded to the top and bottom of the decomposed weight matrix $\mathbf{W}'$, as shown in Fig. S2c. As a result, the output vector becomes $\mathbf{y}'' \in \mathbb{C}^{M'+2\Delta M}$ with $\Delta M$ 0's padded to the beginning and end of the vector. This process can be written as

$$W''_{m,n} = \begin{cases} W'_{(m-\Delta M),n}, & \text{if } \Delta M \le m < M' + \Delta M, \\ 0, & \text{otherwise,} \end{cases} \text{ and } y''_m = \begin{cases} y'_{m-\Delta M}, & \text{if } \Delta M \le m < M' + \Delta M, \\ 0, & \text{otherwise.} \end{cases} \tag{S74}$$

Based on equation (S71), the received output signal spectrum after LPF and with padded zero subcarriers $\mathbf{S}_{y\downarrow}$ satisfies

$$S_{y\downarrow}[m] = \begin{cases} y'_{M'-1-m+\Delta M}, & \text{if } \Delta M \le m < M' + \Delta M, \\ 0, & \text{otherwise,} \end{cases} \tag{S75}$$

which implies that the output spectrum $\mathbf{S}_{y\downarrow}$ has $\Delta M$ zero subcarriers on both edges, i.e., the transient frequency range of a practical LPF, as shown in Fig. S2d. Let $\alpha = 2\Delta M/M'$ denote the overhead coefficient associated with zero subcarrier padding, the number of subcarriers per decomposed MVM is $(1+\alpha)NM'$ for $\mathbf{S}_x$ and $\mathbf{S}_w$, and $(1+\alpha)M'$ for $\mathbf{S}_{y\downarrow}$. In general, a larger value of $\alpha$ is required if the LPF exhibits a larger transient frequency range, leading to a larger overhead on the waveform duration that is inversely proportional to the subcarrier spacing, $\Delta f$.

## E. Cyclic Prefix

So far, we assume that the client's DACs on the TX side and ADCs on the RX side are perfectly synchronized when generating and capturing $\mathbf{s}_x$ and $\mathbf{s}_{y\downarrow}$. The assumption does not hold in practice, especially when the RX employed an ADC operating at a reduced sampling rate with a downsampling ratio of $N$. Inspired by OFDM-based wireless communication systems, we introduce a cyclic prefix to mitigate the potential delay and timing offset between the DAC and ADC, as shown in Fig. S2e. Different from conventional OFDM-based communication systems, the downsampling ratio of $N$ means that every $N$ I/Q sample in $\mathbf{s}_x$ and $\mathbf{s}_w$ corresponds to a single I/Q sample in $\mathbf{s}_{y\downarrow}$. To ensure an integer number of I/Q samples for the cyclic prefix removal on $\mathbf{s}_{y\downarrow}$, the cyclic prefix length on $\mathbf{s}_x$ and $\mathbf{s}_w$ must be a multiple of $N$. Hereby, we consider a cyclic prefix length of $\Delta L \in \mathbb{N}$ for $\mathbf{s}_{y\downarrow}$, and thus the cyclic prefix length is $N \cdot \Delta L$ on $\mathbf{s}_x$ and $\mathbf{s}_w$. Then, the cyclic

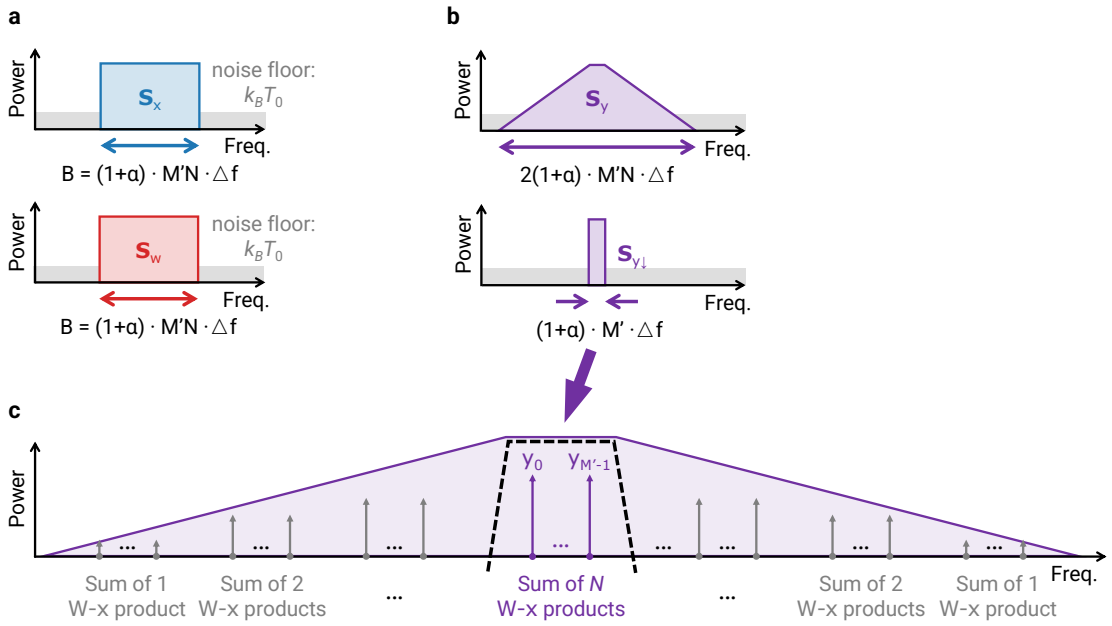


**Fig. S3 The frequency-domain power spectral density comparison of $\mathbf{S}_x$, $\mathbf{S}_w$, $\mathbf{S}_y$ and $\mathbf{S}_{y\downarrow}$. a**, The spectrum of $\mathbf{S}_x$ and $\mathbf{S}_w$ occupies a signal bandwidth of $NM' \cdot \Delta f$. **b**, The spectrum of the computing mixer's output signal, $\mathbf{S}_y$, occupies a signal bandwidth of $(2NM' - 1) \cdot \Delta f$, while the region of interest carrying information about $\mathbf{S}_{y\downarrow}$ is confined to a smaller bandwidth of $M' \cdot \Delta f$. **c**, A zoomed-in view of the output signal spectrum, $\mathbf{S}_y$, which contains $\mathbf{S}_{y\downarrow}$ in the center of the bandwidth.

prefix attachment from $\mathbf{s}_x$ to $\mathbf{s}'_x \in \mathbb{C}^{(1+\alpha)NM'+N\Delta L}$ can be written as

$$s'_x[n] \begin{cases} s_x[(1+\alpha)NM' + n - N\Delta L], & \text{if } n < N\Delta L, \\ s_x[n - N\Delta L], & \text{if } n \geq N\Delta L, \end{cases} \quad \forall n = 0, 1, \ldots, (1+\alpha)NM' + N\Delta L - 1. \tag{S76}$$

Similarly, the cyclic prefix attachment from $\mathbf{s}_w$ to $\mathbf{s}'_w \in \mathbb{C}^{(1+\alpha)NM'+N\Delta L}$ can be written as

$$s'_w[n] = \begin{cases} s_w[(1+\alpha)NM' + n - N\Delta L], & \text{if } n < N\Delta L, \\ s_w[n - N\Delta L], & \text{if } n \geq N\Delta L, \end{cases} \quad \forall n = 0, 1, \ldots, (1+\alpha)NM' + N\Delta L - 1. \tag{S77}$$

On the RX side, only the first $(1+\alpha)M'$ I/Q samples need to be captured. Similarly, we define an overhead coefficient $\beta = \Delta L/(M'+2\Delta M)$ such that the length of $\mathbf{s}_x$ and $\mathbf{s}_w$ per decomposed MVM is $(1+\alpha)(1+\beta)NM'$, and the length of $\mathbf{s}_{y\downarrow}$ is $(1+\alpha)(1+\beta)M'$.

## F. Energy Efficiency Analysis

Similar to the energy efficiency analysis in Supplementary Section 7, the energy consumption for the WISE basic scheme contains three components: $(i)$ $E_1$ for the waveform generation of $x(t)$ and I/Q (de)modulation, $(ii)$ $E_2$ for the I/Q sampling of waveform $y(t)$ using two ADCs operating at reduced sampling rate after I/Q demodulation, and $(iii)$ $E_3$ for the digital computing based encoding and decoding.

First, the new waveform duration for each decomposed MVM including the overhead is $(1+\alpha)(1+\beta)NM'/B$. Given the transmit power of $P_x$, the total energy required for the client radio to generate the

waveforms carrying the total number of $M/M'$ inference requests is given by

$$E_1 = \frac{M}{M'} \cdot (\eta_{\text{radio}})^{-1} P_x \cdot \frac{(1+\alpha)(1+\beta) \cdot NM'}{B} = (\eta_{\text{radio}})^{-1} P_x \cdot \frac{(1+\alpha)(1+\beta) \cdot NM}{B}. \tag{S78}$$

Different from Supplementary Section 7, the SNR in the WISE basic scheme is measured within the captured narrowband of $[-(1+\alpha)M'\Delta f/2, +(1+\alpha)M'\Delta f/2]$. As illustrated in Fig. S3a–b, the original $\mathbf{S}_y$ spans over a bandwidth of $2(1+\alpha)NM' \cdot \Delta f$ with the total power of $\eta_{\text{mixer}} \cdot P_x$, while the bandwidth of interest is only the portion of $(1+\alpha)M' \cdot \Delta f$. Note that the power is not evenly distributed over the subcarriers in $\mathbf{S}_y$; each subcarrier $S_y[k]$ is the sum of multiple products of $x$-$W$ pairs, according to the convolution theorem. Hereby, we assume all the products of the $x$-$W$ pairs are independent and identically distributed (i.i.d.). After the linear convolution based on equation (S38), the first $M'$ elements on the output $\mathbf{S}_y$ (excluding the $\alpha M'$ padded zero subcarriers) only have one such product, the second $M'$ elements are the sum of two of such products, and so on, until the middle $M'$ elements, which is the sum of $N$ of such products as the captured $\mathbf{S}_{y\downarrow}$ or $\mathbf{y}$. The second half of the subcarriers on $\mathbf{S}_y$ follows the same trend but with a reversed symmetry compared to the first half of the subcarriers. According to the law of large numbers and the central limit theorem together with $N \gg 1$, elements of $\mathbf{S}_y$ follow Gaussian distributions, whose variance (or power) is proportional to the number of products to be summed. Therefore, the power density spectrum of $\mathbf{S}_y$ has a "triangle" shape, as shown in Fig. S3c. Moreover, the total power of the middle $M'$ subcarriers of interest, i.e., $\mathbf{S}_{y\downarrow}$, is $1/N$ of the power of $\mathbf{S}_y$. Therefore, the received power $P_y$ within the narrowband can be approximated by

$$P_y \approx \frac{1}{N} \cdot \eta_{\text{mixer}} \cdot P_x. \tag{S79}$$

On the other hand, the noise power only spans the narrow band of $(1+\alpha)M'\Delta f$, so the noise power, $P_n$, is given by

$$P_n = (\eta_{\text{nf}})^{-1} \cdot k_B T_0 \cdot (1+\alpha) \cdot M' \cdot \Delta f. \tag{S80}$$

Combining equations (S79) and (S80), the relationship between $P_x$ and $\mathsf{SNR}$ is given by

$$\mathsf{SNR} = \frac{P_y}{P_n} = \frac{N^{-1} \cdot \eta_{\text{mixer}} \cdot P_x}{(\eta_{\text{nf}})^{-1} \cdot k_B T_0 \cdot (1+\alpha) \cdot M' \cdot \Delta f} = \frac{\eta_{\text{mixer}} \cdot \eta_{\text{nf}} \cdot P_x}{k_B T_0 \cdot B} \Rightarrow P_x = (\eta_{\text{mixer}} \cdot \eta_{\text{nf}})^{-1} \cdot \mathsf{SNR} \cdot k_B T_0 \cdot B. \tag{S81}$$

Plugging $P_x$ in equation (S78) yields

$$E_1 = (1+\alpha)(1+\beta) \cdot NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0. \tag{S82}$$

For each decomposed MVM, only $(1+\alpha)M'$ I/Q samples in $\mathbf{s}_{y\downarrow}$ need to be captured by a pair of ADCs operating at a low sampling rate. Therefore, the total energy consumption $E_2$ across all $M/M'$ decomposed MVMs is given by

$$E_2 = \frac{M}{M'} \cdot (1+\alpha)M' \cdot 2e_{\text{adc}} = 2(1+\alpha) \cdot M \cdot e_{\text{adc}}. \tag{S83}$$

For the encoding energy part of $E_3$, it reduces to an $N$-point IFFT for encoding following equation (S70), including $2N\log_2(N)$ MACs, which needs to be performed only once and can be reused for all the decomposed MVMs. In addition, the decoding energy part of $E_3$ in equation (S73) consumes $(1+\alpha)M'$-point FFT per decomposed MVM. Therefore, $E_3$ including both the encoding and decoding energy is given by

$$E_3 = \left(2N\cdot\log_2 N + \frac{M}{M'}\cdot 2(1+\alpha)M'\log_2((1+\alpha)M')\right)\cdot e_{\text{dig}} = (2N\cdot\log_2 N + 2(1+\alpha)\cdot M\cdot\log_2((1+\alpha)M'))\cdot e_{\text{dig}}. \tag{S84}$$

Putting equations (S82), (S83) and (S84) together, the total energy consumption $E_{\text{mvm}}$ is given by

$$\begin{aligned} E_{\text{mvm}} = E_1+E_2+E_3 = &\underbrace{(1+\alpha)(1+\beta)\cdot NM\cdot\eta^{-1}\cdot\mathsf{SNR}\cdot k_BT_0}_{E_1} + \underbrace{2(1+\alpha)\cdot M\cdot e_{\text{adc}}}_{E_2} \\ &+ \underbrace{(2N\cdot\log_2 N + 2(1+\alpha)\cdot M\cdot\log_2((1+\alpha)M'))\cdot e_{\text{dig}}}_{E_3}. \end{aligned} \tag{S85}$$

The corresponding energy efficiency, measured by energy per MAC, $e_{\text{mvm}}$, is

$$\begin{aligned} e_{\text{mvm}} = \frac{E_{\text{mvm}}}{4NM} = e_1+e_2+e_3 = &\underbrace{\frac{(1+\alpha)(1+\beta)}{4}\cdot\eta^{-1}\cdot\mathsf{SNR}\cdot k_BT_0}_{e_1} + \underbrace{\frac{1+\alpha}{2N}\cdot e_{\text{adc}}}_{e_2} \\ &+ \underbrace{\left(\frac{1}{2M}\cdot\log_2 N + \frac{1+\alpha}{2N}\cdot\log_2((1+\alpha)M')\right)\cdot e_{\text{dig}}}_{e_3}. \end{aligned} \tag{S86}$$

From equation (S86), it can be seen that the energy efficiency term $e_1$ is independent of the MVM dimension, and the terms $e_2$ and $e_3$ respectively scale as $\mathcal{O}(\frac{1}{N})$ and $\mathcal{O}(\frac{\log N}{M}+\frac{1}{N})$. As a result, as the MVM dimensions increase, i.e., $N\to\infty$ and $M\to\infty$, $e_{\text{mvm}}$ approaches $e_1$, which is bottlenecked by the thermal noise and hardware limit, thus achieving significantly enhanced energy efficiency compared to digital computing. Moreover, under an ideal hardware with $\eta=1$ and $\alpha=\beta=0$, we can derive WISE's thermal dynamic limit (TDL) as

$$e_{\text{tdl}} := \lim_{N\to+\infty, M\to+\infty} e_{\text{mvm}} = \mathsf{SNR}\cdot kT_0/4. \tag{S87}$$

This TDL can even exceed the energy efficiency for $b$-bit computation at the Landauer Limit, given by $e_{\text{Landauer}} = b^2\cdot\ln 2\cdot kT_0$, when $\mathsf{SNR}/4 < b^2\ln 2$, or $\mathsf{SNR} < 2.77b^2$.

## G. MVM Decomposition into IPs

According to equation (S86), while a decomposed MVM with a smaller value of $M'$ reduces the energy consumption term $e_3$, it usually requires a large overhead, i.e., larger values of $\alpha$ and/or $\beta$. For the extreme MVM decomposition case with $M'=1$, the original MVM is decomposed into $M$ IPs. We denote the total energy consumption and energy efficiency under this extreme decomposition as $E'_{\text{mvm}}$ and $e'_{\text{mvm}}$, respectively. Since the number of padded zero subcarriers must be a positive integer, we set the minimum $\Delta M = 1$ that yields a frequency-domain overhead coefficient of $\alpha = 2$. In this case, the decoding process in equation

(S73) becomes a three-point FFT ($(1+\alpha)M'=3$), where the IP result is carried by the middle subcarrier, $y_0 = S_{y\downarrow}[1]$. This means the energy consumption term $E_3$ given by equation (S84) does not hold. Instead, the decoding process from $\mathbf{s}_{y\downarrow} \in \mathbb{C}^3$ to $S_{y\downarrow}[1]$ can be rewritten as

$$y_0 = S_{y\downarrow}[1] = \sum_{n=0}^{2} s_y[n] \cdot e^{-j2\pi \frac{1-3/2}{3} n} = s_y[0] + s_y[1] \cdot e^{j\frac{\pi}{3}} + s_y[2] \cdot e^{j\frac{2\pi}{3}}, \tag{S88}$$

which contains two complex-valued MACs or, equivalently, eight real-valued MACs. In addition, the synchronization algorithm in Supplementary Section 12 ensures the sub-sample-level of timing synchronization between the TX and RX sides, and the cyclic prefix with $\Delta L = 1$ is sufficient, corresponding to $\beta = \Delta L/(M' + 2\Delta M) = 1/3$. By plugging $\alpha = 2$ and $\beta = 1/3$ for other two terms $E_1$ and $E_2$, the energy required for the MVM computation, $E'_{\text{mvm}}$, is given by

$$E'_{\text{mvm}} = \underbrace{4NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{E_1} + \underbrace{6M \cdot e_{\text{adc}}}_{E_2} + \underbrace{(2N \cdot \log_2 N + 8M) \cdot e_{\text{dig}}}_{E_3}, \tag{S89}$$

which corresponds to an energy efficiency of

$$e'_{\text{mvm}} = \frac{E'_{\text{mvm}}}{4NM} = \underbrace{\eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{3}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\left(\frac{1}{2M} \cdot \log_2 N + \frac{2}{N}\right) \cdot e_{\text{dig}}}_{e_3}. \tag{S90}$$

In the special case with $M = 1$, the MVM is reduced to a standalone IP computation task given by $c = \langle \mathbf{a}, \mathbf{b} \rangle = \sum_{n=1}^{N} a_n \cdot \overline{b_n}$. The total energy consumption for a standalone IP, denoted by $E_{\text{ip}}$, can be obtained by plugging $M = 1$ into equation (S89), i.e.,

$$E_{\text{ip}} = \underbrace{4N \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{E_1} + \underbrace{6 \cdot e_{\text{adc}}}_{E_2} + \underbrace{(2N \cdot \log_2 N + 8) \cdot e_{\text{dig}}}_{E_3}. \tag{S91}$$

The corresponding energy efficiency, denoted by $e_{\text{ip}}$, can then be derived by plugging $M = 1$ into equation (S90), i.e.,

$$e_{\text{ip}} = \frac{E_{\text{ip}}}{4N} = \underbrace{\eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{3}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\left(\frac{1}{2} \cdot \log_2 N + \frac{2}{N}\right) \cdot e_{\text{dig}}}_{e_3}. \tag{S92}$$

Note that since $\left(\frac{1}{2} \cdot \log_2 N + \frac{2}{N}\right) \cdot e_{\text{dig}} > e_{\text{dig}}$, $e_{\text{ip}}$ is always higher than $e_{\text{dig}}$. This is because the IFFT-based encoding requires $2N \log_2 N$ MACs, which is higher energy consumption than directly computing the IP itself and cannot be amortized compared to the MVM task as $M$ scales. Fortunately, this limitation can be resolved in the $\mathbf{W}$-precoding scheme, described in Supplementary Section 9.

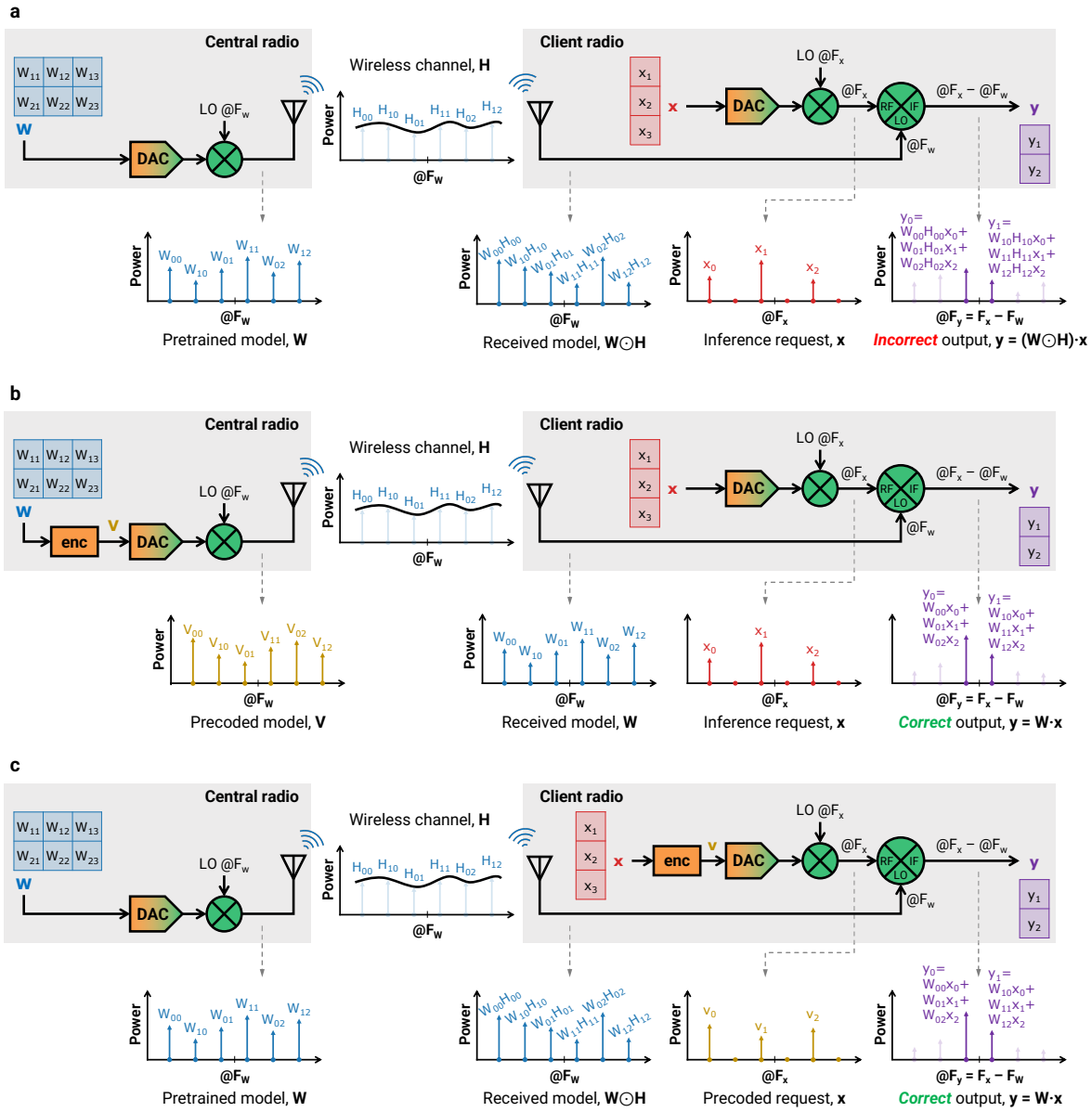


**Fig. S4 The overview of the channel modeling and the two channel calibration schemes of WISE. a**, The wireless channel introduces signal distortion, leading to incorrect matrix-vector multiplication (MVM) results when the channel state information (CSI), **H**, is uncalibrated. **b**, In the **W**-precoding scheme, model weights **W** are precoded into **V** at the central radio to ensure that the correct model weights are received by the client after wireless transmission. **c**, In the **x**-precoding scheme, each inference request **x** is precoded into **v** at the client to compensate for channel effects on the received signal.

# 9 WISE's W-Precoding Scheme: Wireless Channel Calibration at the Central Radio

The basic scheme of WISE, described in Supplementary Section 8, assumes that the $w(t)$ carrying model weights **W** is directly input into the computing mixer. However, WISE leverages the shared wireless medium to broadcast model weights, where signals carrying model weights experience timing delay, multi-path effect, and distortion as they propagate through the wireless channel from the central radio to each client, as

illustrated in Fig. S4a. Therefore, channel state information (CSI) estimation and calibration are required to ensure that the signals received by the clients carry the desired model weights, $\mathbf{W}$, after wireless transmission.

In this section, we consider channel calibration at the central radio, termed the "$\mathbf{W}$-precoding scheme", as shown in Fig. S4b, which precodes the model weights $\mathbf{W}$ into $\mathbf{V} = [V_{m,n}] \in \mathbb{C}^{M\times N}$ before transmission to the clients. This approach does not incur additional computational or energy costs for the client. Moreover, multiple clients in close proximity sharing similar CSI can be served with the same precoded model weights. The $\mathbf{W}$-precoding scheme is derived from the basic scheme described in Supplementary Section 8. For brevity, we use $\mathbf{W}$ and $\mathbf{y}$ to represent the decomposed weight matrix $\mathbf{W}''$ and padded output $\mathbf{y}''$ in Supplementary Section 8, with $M$ referring $(1+\alpha)M'$. By eliminating the IFFT-based encoding on $\mathbf{x}$, the $\mathbf{W}$-precoding scheme further improves the energy efficiency for in-physics MVM computation.

## A. Wireless Channel Modeling

For the wireless channel from the central radio to the client, we define its wireless channel's frequency response $h(f) \in \mathbb{C}$ as a complex-valued function of the frequency $f$. As shown in Fig. S4a, for a general spectrum $\mathbf{S}_{\text{TX}}$, the frequency response on its $k$-th subcarrier at frequency $(F + (k - L/2)\cdot\Delta f)$ can be acquired by $h(F+(k-L/2)\cdot\Delta f)$. Define the channel state information (CSI) as a vector $\mathbf{S}_h = [S_h[k]] \in \mathbb{C}^L$ of the same dimension, which has $S_h[k] = h(F+(k-L/2)\cdot\Delta f)$. Let '$\odot$' and '$\oslash$' denote the element-wise multiplication and division of two vectors with equal length, respectively. The impact of the wireless channel from $\mathbf{S}_{\text{TX}}$ to $\mathbf{S}_{\text{RX}}$ can be formulated as

$$\mathbf{S}_{\text{RX}} = \mathbf{S}_{\text{TX}} \odot \mathbf{S}_h, \text{ where } S_{\text{RX}}[k] = S_{\text{TX}}[k]\cdot S_h[k] = S_{\text{TX}}[k]\cdot h\left(F + \left(k - \frac{L}{2}\right)\cdot\Delta f\right), \ \forall k = 0, 1, \ldots, L. \tag{S93}$$

## B. W-Precoding Scheme: Algorithm

Specifically for the MVM task, the FFT size is $NM$, so that we consider the CSI as $\mathbf{S}_h = [S_h[k]] \in \mathbb{C}^{NM}$. For this $\mathbf{S}_h$, we define its corresponding CSI matrix, $\mathbf{H} = [H_{m,n}] \in \mathbb{C}^{M\times N}$, given by

$$H_{m,n} = S_h[MN - 1 - m - nM], \ \forall n = 0, 1, \ldots, N-1, \text{ and } m = 0, 1, \ldots, M-1. \tag{S94}$$

Based on equations (S46), (S93), and (S94), the MVM including the wireless channel impact can be written as

$$\mathbf{y} = \mathbf{W}\cdot\mathbf{x} = (\mathbf{H}\odot\mathbf{V})\cdot\mathbf{x}, \text{ with } y_m = \sum_{n=0}^{N-1} H_{m,n}\cdot V_{m,n}\cdot x_n, \ \forall m = 0, 1, \ldots, M-1. \tag{S95}$$

To estimate the channel $\mathbf{H}$, we randomize a set of $\mathbf{V}^{(i)}$ and $\mathbf{x}^{(i)}$; then we obtain the MVM results by $(i)$ simulating MVM with wireless channel effect following equation (S95), denoted as $\mathbf{y}^{(i)}$, and $(ii)$ the analog computing by WISE without channel calibration, denoted as $\hat{\mathbf{y}}^{(i)}$. Then, $\mathbf{H}$ can be estimated using the minimum mean squared error (MMSE) method given by

$$\mathbf{H}^\star = \arg\min_{\mathbf{H}} \sum_i \left|\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right|^2 = \arg\min_{\mathbf{H}} \sum_i \left|(\mathbf{H}\odot\mathbf{V}^{(i)})\cdot\mathbf{x}^{(i)} - \hat{\mathbf{y}}^{(i)}\right|^2. \tag{S96}$$

Due to the continuity nature of the channel response $h(f)$, we only need to perform the MMSE optimization in equation (S96) once with a large value of $L$ (e.g., $L = 300$ for the 25 MHz channel used in our experiments) for its $\mathbf{H}^\star$. Then, the corresponding $\mathbf{S}_h^\star$ can be acquired by reversely conducting equation (S94), and the general $h(f)$ can be estimated by nearest neighbor-based interpolation on amplitude and linear interpolation on phases, from which the $\mathbf{S}_h$ of other $L$ and $\mathbf{H}$ of other $N$ and $M$ can be inferred. When there are multiple users, we optimize $\mathbf{H}^\star$ for each user and average them over users to parameterize the overall $\mathbf{H}^\star$. With the optimized $\mathbf{H}^\star$, and the precoded weight matrix $\mathbf{V}$ is given by

$$\mathbf{V} = \mathbf{W} \oslash \mathbf{H}^\star, \text{ where } V_{m,n} = \frac{W_{m,n}}{H^\star_{m,n}}, \ \forall n = 0, 1, \ldots, N-1, \text{ and } \forall m = 0, 1, \ldots, M-1. \tag{S97}$$

Note that the channel estimation process in equation (S96) can be preprocessed, whose cost can be averaged down over an unlimited number of inference requests. Also, the precoding process in equation (S97) is performed on the central radio, which does not impact the energy consumption or the computation throughput of the client.

## C. W-Precoding Scheme: Time Encoding for x

Note that in equation (S70), the $N$-point IFFT applied to $\mathbf{x}$ can be formulated as an MVM given by $(\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2}) \cdot \mathbf{x}$. Therefore, we can detach this IFFT process from $\mathbf{x}$, and attach it as part of $\mathbf{W}$. Specifically, after the detachment, the new input vector $\mathbf{x}' = [x'_n] \in \mathbb{C}^N$ is given by

$$\mathbf{x}' = \left(\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2}\right)^{-1} \cdot \mathbf{x}, \tag{S98}$$

whose corresponding I/Q waveform $\mathbf{s}'_x = [s'_x[n]] \in \mathbb{C}^{NM}$ can be generated by

$$\mathbf{s}'_x = \frac{1}{M\sqrt{N}} \cdot \underbrace{\left[\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2} \cdot \mathbf{x}', \ldots, \mathbf{D}^{-1} \cdot \mathbf{R}^{N/2} \cdot \mathbf{x}'\right]}_{\text{repeated } M \text{ times}} = \frac{1}{M\sqrt{N}} \cdot \underbrace{\left[\mathbf{x}, \ldots, \mathbf{x}\right]}_{\text{repeated } M \text{ times}}. \tag{S99}$$

Therefore, $\mathbf{s}'_x$ can be generated by repeating the original $\mathbf{x}$, i.e., via direct time encoding, and involved no MACs. On the other hand, the new $\mathbf{W}' = [W'_{m,n}] \in \mathbb{C}^{M \times N}$ combined with the IFFT operation is given by

$$\mathbf{W}' = \mathbf{W} \cdot (\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2}). \tag{S100}$$

Plugging this new $\mathbf{W}'$ into the encoding process at the central radio, the corresponding I/Q waveform $\mathbf{s}_w = [s_w[n]] \in \mathbb{C}^{NM}$ is given by

$$s'_w[n] = \frac{1}{NM} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} W'_{m,n} \cdot e^{-j2\pi \frac{1+m+nM+NM/2}{NM} n}, \text{ where } W'_{m,n} = \frac{1}{\sqrt{N}} \sum_{n'=0}^{N-1} W_{m,n'} \cdot e^{j2\pi \frac{n-N/2}{N} n'}, \ \forall n = 0, 1, \ldots, L-1. \tag{S101}$$

Overall, the weight matrix $\mathbf{W}$ is still frequency-encoded and, upon receiving the $\mathbf{W}$-precoded model weights from the central radio, the client performs local MVM computation given by

$$\begin{aligned}
\mathbf{y} &= \mathbf{V}' \odot \mathbf{H} \cdot \mathbf{x}' \\
&= (\mathbf{W}' \oslash \mathbf{H}^\star) \odot \mathbf{H} \cdot \mathbf{x}' = \left[\mathbf{W} \cdot (\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2}) \oslash \mathbf{H}^\star\right] \odot \mathbf{H} \cdot \mathbf{x}' && (\mathbf{W}\text{-precoding at the central radio}) \\
&= \left[\mathbf{W} \cdot (\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2}) \oslash \mathbf{H}^\star\right] \odot \mathbf{H} \cdot \left[(\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2})^{-1} \cdot \mathbf{x}\right] && (\text{time encoding of } \mathbf{x} \text{ at the client}) \\
&= \left[\mathbf{W} \cdot (\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2})\right] \oslash (\mathbf{H}^\star \oslash \mathbf{H}) \cdot \left[(\mathbf{D}^{-1} \cdot \mathbf{R}^{N/2})^{-1} \cdot \mathbf{x}\right] \approx \mathbf{W} \cdot \mathbf{x} && (\text{original MVM}). \qquad \text{(S102)}
\end{aligned}$$

Note that the last approximation is due to the MMSE-based channel estimation, which corresponds to $\mathbf{H}^\star \approx \mathbf{H}$, or $\mathbf{H}^\star \oslash \mathbf{H}$ being approximately an all-ones matrix.

## D. W-Precoding Scheme: Energy Efficiency Analysis

Compared to the basic scheme described in Supplementary Section 8, the $\mathbf{W}$-precoding scheme further improves the energy efficiency by applying "IFFT-less" time encoding on the client. In particular, under this scheme, the energy consumption term $E_3$ becomes

$$E_3 = \frac{M}{M'} \cdot 2(1+\alpha)M' \log_2((1+\alpha)M') \cdot e_{\text{dig}} = 2(1+\alpha) \cdot M \cdot \log_2((1+\alpha)M') \cdot e_{\text{dig}}, \qquad \text{(S103)}$$

where the encoding energy of $(2N \log_2 N \cdot e_{\text{dig}})$ in equation (S84) is eliminated. As a result, the total energy consumption per MVM, $E_{\text{mvm}}$, for the $\mathbf{W}$-precoding scheme is given by

$$\begin{aligned}
E_{\text{mvm}} = E_1 + E_2 + E_3 = &\underbrace{(1+\alpha)(1+\beta) \cdot NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{E_1} + \underbrace{2(1+\alpha) \cdot M \cdot e_{\text{adc}}}_{E_2} \\
&+ \underbrace{2(1+\alpha) \cdot M \cdot \log_2((1+\alpha)M') \cdot e_{\text{dig}}}_{E_3}. \qquad \text{(S104)}
\end{aligned}$$

The corresponding energy efficiency, $e_{\text{mvm}}$, is given by

$$e_{\text{mvm}} = \frac{E_{\text{mvm}}}{4NM} = e_1 + e_2 + e_3 = \underbrace{\frac{(1+\alpha)(1+\beta)}{4} \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{1+\alpha}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\frac{1+\alpha}{2N} \cdot \log_2((1+\alpha)M') \cdot e_{\text{dig}}}_{e_3}. \qquad \text{(S105)}$$

Note that equation (S105) indicates that the energy efficiency of this $\mathbf{W}$-precoding scheme is independent of the output size $M$, due to the elimination of the $N$-point IFFT for encoding. The corresponding TDL under the $\mathbf{W}$-precoding scheme is given by

$$e_{\text{tdl}} := \lim_{N \to +\infty} e_{\text{mvm}} = \mathsf{SNR} \cdot kT_0/4. \qquad \text{(S106)}$$

### E. W-Precoding Scheme: MVM Decomposition into IPs

Similarly, the energy efficiency $E'_{\text{mvm}}$ for the **W**-precoding scheme that decomposes the MVM into $M$ IPs can be rewritten from equation (S89) as

$$E'_{\text{mvm}} = \underbrace{4NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{E_1} + \underbrace{6M \cdot e_{\text{adc}}}_{E_2} + \underbrace{8M \cdot e_{\text{dig}}}_{E_3}, \tag{S107}$$

where the encoding energy consumption of $\left(\frac{1}{2} \cdot \log_2 N\right) \cdot e_{\text{dig}}$ is eliminated due to the time encoding of $\mathbf{x}$. Also, $e'_{\text{mvm}}$ can be derived from equation (S90), which is

$$e'_{\text{mvm}} = \frac{E'_{\text{mvm}}}{4NM} = \underbrace{\eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{3}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\frac{2}{N} \cdot e_{\text{dig}}}_{e_3}. \tag{S108}$$

Here, notice that the energy efficiency for **W**-precoding scheme is independent of $M$. Therefore, for a standalone IP computation with $M = 1$, it holds that

$$e_{\text{ip}} = e'_{\text{mvm}} = \underbrace{\eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{3}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\frac{2}{N} \cdot e_{\text{dig}}}_{e_3}. \tag{S109}$$

Compared to the basic scheme, this **W**-precoding scheme achieves further enhanced energy efficiency for standalone IP computations.

## 10 WISE's x-Precoding Scheme: Wireless Channel Calibration at the Client

One alternative channel calibration scheme is performed for each client by adjusting the input $\mathbf{x}$, which we term as the **x**-precoding scheme, as shown in Fig. S4c. This scheme allows each client to estimate and apply its own CSI, especially when the users are away from each other and the CSIs are diverse among clients. On the other hand, this scheme requires extra computation costs for clients, incurring higher energy consumption. This scheme precodes $\mathbf{x}$ into $\mathbf{v} = [v_n] \in \mathbb{C}^N$.

### A. x-Precoding Scheme: Algorithm

For the same CSI $\mathbf{S}_h$ as considered in Supplementary Section 9, we define the equivalent channel vector $\mathbf{h} = [h_n] \in \mathbb{C}^N$. Due to the smoothness of the channel response $h(f)$ and thus $\mathbf{S}_h$, we let $h_n$ to approximate the channel responses as

$$h_n \approx S_h[MN - 1 - m - nM] = H_{m,n}, \ \forall n = 0, 1, \ldots, N-1, \ m = 0, 1, \ldots, M-1. \tag{S110}$$

This **x**-precoding scheme transmits the $\mathbf{v}$ in substitute of $\mathbf{x}$, which compensates the channel impact on the received $\mathbf{W}$ during the frequency mixing. Then, we can rewrite equation (S95) by this new equivalent channel

vector $\mathbf{h}$ and a precoded $\mathbf{v}$ as

$$\mathbf{y} = \mathbf{W} \cdot (\mathbf{h} \odot \mathbf{v}) \text{ with } y_m = \sum_{n=0}^{N-1} W_{m,n} \cdot h_n \cdot v_n, \ \forall m = 0, 1, \ldots, M-1. \tag{S111}$$

To estimate the equivalent channel vector $\mathbf{h}$, we randomize a series of $\mathbf{v}^{(i)}$ and $\mathbf{W}^{(i)}$, and generate their corresponding $\mathbf{y}^{(i)}$ and $\hat{\mathbf{y}}^{(i)}$. Similarly, $\mathbf{h}$ can be optimized by the MMSE method given by

$$\mathbf{h}^\star = \arg\min_{\mathbf{h}} \sum_i \left| \mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)} \right|^2 = \arg\min_{\mathbf{h}} \sum_i \left| \mathbf{W}^{(i)} \cdot (\mathbf{h} \odot \mathbf{v}^{(i)}) - \hat{\mathbf{y}}^{(i)} \right|^2. \tag{S112}$$

Given the optimized $\mathbf{h}^\star$, the precoded input $\mathbf{v}$ is given by

$$\mathbf{v} = \mathbf{x} \oslash \mathbf{h}^\star \text{ with } v_n = \frac{x_n}{h_n^\star}, \ \forall n = 0, 1, \ldots, N. \tag{S113}$$

Essentially, the MVM computation of this $\mathbf{x}$-precoding scheme can be summarized by

$$\begin{aligned} \mathbf{y} &= \mathbf{W} \odot \mathbf{H} \cdot \mathbf{v} \approx \mathbf{W} \cdot (\mathbf{h} \odot \mathbf{v}) \\ &= \mathbf{W} \cdot (\mathbf{h} \odot \mathbf{x} \oslash \mathbf{h}^\star) && (\mathbf{x}\text{-precoding at the client}) \\ &= \mathbf{W} \cdot [(\mathbf{h} \oslash \mathbf{h}^\star) \odot \mathbf{x}] \approx \mathbf{W} \cdot \mathbf{x} && (\text{original MVM}). \end{aligned} \tag{S114}$$

Here, the last approximation is due to the MMSE-based channel estimation, which corresponds to $\mathbf{h}^\star \approx \mathbf{h}$, or $\mathbf{h}^\star \oslash \mathbf{h}$ being approximately an all-ones vector.

## B. x-Precoding Scheme: Energy Efficiency Analysis

Compared to the basic scheme described in Supplementary Section 8, this $\mathbf{x}$-precoding scheme maintains the same energy consumption terms $E_1$ and $E_2$. On the other hand, the precoding is performed in the frequency domain for $\mathbf{x}$ leveraging the IFFT-based encoding. Moreover, the precoding process based on equation (S113) requires additional $N$ complex-valued MACs (or $4N$ real-valued MACs). Hence, the energy consumption term $E_3$ for the $\mathbf{x}$-precoding scheme is given by

$$E_3 = (4N + 2N \cdot \log_2 N + 2(1+\alpha) \cdot M \cdot \log_2((1+\alpha)M')) \cdot e_{\text{dig}}. \tag{S115}$$

Then, the energy consumption per MVM under this $\mathbf{x}$-precoding scheme is given by

$$\begin{aligned} E_{\text{mvm}} = E_1 + E_2 + E_3 = &\underbrace{(1+\alpha)(1+\beta) \cdot NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{E_1} + \underbrace{2(1+\alpha) \cdot M \cdot e_{\text{adc}}}_{E_2} \\ &+ \underbrace{(4N + 2N \log_2 N + 2(1+\alpha) \cdot M \cdot \log_2((1+\alpha)M')) \cdot e_{\text{dig}}}_{E_3}. \end{aligned} \tag{S116}$$

The corresponding energy efficiency is given by

$$e_{\text{mvm}} = \frac{E_{\text{mvm}}}{4NM} = e_1 + e_2 + e_3$$
$$= \underbrace{\frac{(1+\alpha)(1+\beta)}{4} \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{1+\alpha}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\left(\frac{1}{M} + \frac{\log_2 N}{2M} + \frac{1+\alpha}{2N} \cdot \log_2((1+\alpha)M')\right) \cdot e_{\text{dig}}}_{e_3}. \tag{S117}$$

Similar to the basic scheme, the TDL under this **x**-precoding scheme with $N \to \infty$ and $M \to \infty$ is given by

$$e_{\text{tdl}} := \lim_{N \to +\infty, M \to +\infty} e_{\text{mvm}} = \mathsf{SNR} \cdot kT_0/4. \tag{S118}$$

## C. x-Precoding Scheme: MVM Decomposition into IPs

As for the IP-based MVM decomposition with $M' = 1$, we have the additional precoding energy consumption of $4N \cdot e_{\text{dig}}$. Hence, equation (S89) can be rewritten as

$$E'_{\text{mvm}} = \underbrace{4NM \cdot \eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{E_1} + \underbrace{6M \cdot e_{\text{adc}}}_{E_2} + \underbrace{(4N + 2N \log_2 N + 8M) \cdot e_{\text{dig}}}_{E_3}, \tag{S119}$$

and equation (S90) can be rewritten as

$$e'_{\text{mvm}} = \frac{E'_{\text{mvm}}}{4NM} = \underbrace{\eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{3}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\left(\frac{1}{M} + \frac{\log_2 N}{2M} + \frac{2}{N}\right) \cdot e_{\text{dig}}}_{e_3}. \tag{S120}$$

The energy efficiency of the standalone IP computation, $e_{\text{ip}}$, can be obtained by plugging $M = 1$ in equation (S120),

$$e_{\text{ip}} = \underbrace{\eta^{-1} \cdot \mathsf{SNR} \cdot k_B T_0}_{e_1} + \underbrace{\frac{3}{2N} \cdot e_{\text{adc}}}_{e_2} + \underbrace{\left(1 + \log_2 N + \frac{2}{N}\right) \cdot e_{\text{dig}}}_{e_3}. \tag{S121}$$

Similar to equation (S92) for the based scheme of WISE, it can be seen that this **x**-precoding scheme is not energy efficient for standalone IP computation.

# 11 Computation Throughput Analysis

In this section, we analyze WISE's *computation throughput*, i.e., the number of real-valued MACs per unit time. Without loss of generality, we consider the MVM $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$, where a single central radio wirelessly broadcasts model weights $\mathbf{W}$ to $U$ clients using an OFDM system with FFT size $L = NM$. Each client performs local inference on $\mathbf{x}$ and generates $\mathbf{y}$. In this case, a total number of $4NM$ real-valued MACs is involved per client, and a total number of $4NM \cdot U$ real-valued MACs across all $U$ clients.

To achieve an MVM involving $U \cdot 4NM$ real-valued MACs, the latency is primarily determined by the waveform duration for transmitting $w(t)$ and $x(t)$. The total waveform duration across all $M/M'$ decomposed

MVMs is

$$T_{\text{mvm}} = \frac{M}{M'} \cdot \frac{(1+\alpha)(1+\beta) \cdot NM'}{B} = \frac{(1+\alpha)(1+\beta) \cdot NM}{B}. \tag{S122}$$

Thus, the computation throughput, denoted by $\Lambda$, is given by

$$\Lambda = \frac{U \cdot 4NM}{T_{\text{mvm}}} = \frac{4U \cdot B}{(1+\alpha)(1+\beta)}, \tag{S123}$$

which applies to all three schemes of WISE, described in Supplementary Sections 8, 9, and 10. To conclude, the computation throughput of WISE is proportional to the available bandwidth, $B$, and number of users $U$. With wirelessly broadcast model weights, the computation throughout will be determined by the available wireless bandwidth in the unlicensed (e.g., 25 MHz in the 915 MHz ISM band) or licensed bands.

# Supplementary Information: Experiment

## 12 Experimental Setup


**Fig. S5 The software-defined radio testbed for WISE's experimental implementation. a**, Experimental setup for WISE with model weights broadcast over a wireless channel to three clients, each employing a passive ZEM-4300+ frequency mixer as the computing mixer for in-physics MVM computation. **b**, Experimental setup for WISE with model weights transmitted over a wired channel to a single client.

Fig. S5a shows the detailed experiment setup of WISE, including the Tupavco TP514 Yagi directional antenna, Mini-Circuits ZEM-4300+ [49] as the computing frequency mixer, and USRP X310 software-defined radio (SDR) as the transceiver radio unit. One USRP X310 serves as the central radio that broadcasts $w(t)$ or $v(t)$ (for the $\mathbf{W}$-precoding scheme) over a wireless channel with a bandwidth of $B = 25\,\text{MHz}$ at the carrier frequency of $F_w = 0.915\,\text{GHz}$. Our wireless experiments are conducted in the unlicensed industrial, scientific, and medical (ISM) band centered at 0.915 GHz, which has a limited bandwidth of only 26 MHz between 902–928 MHz [50]. For a client, one RX antenna receives the wirelessly broadcast model weights at the input to the LO port of the computing mixer; and one USRP X310 generates $x(t)$ or $v(t)$ (for the $\mathbf{x}$-precoding scheme) at the carrier frequency of $F_x = 1.2\,\text{GHz}$, which is streamed into the computing mixer's RF port via a cable and a total of 30 dB attenuator. The output signal of the computing mixer from the IF port, $y(t)$, at the carrier frequency of $F_y = 0.285\,\text{MHz}$ is streamed to the RX channel of a USRP X310, which is configured with a low sampling rate of $B_\downarrow = \max\{\frac{25}{N}, 0.20\}\,\text{MHz}$. The wireless link distance is $\approx$1 m, which is limited by the USRP X310's TX power and the required LO input power to the computing mixer.

We also consider a wired setting of WISE with a single client, as shown in Fig. S5b, with the same carrier frequency configuration of $F_w$, $F_x$, and $F_y$. In this case, the TX channel of the central radio directly streams $w(t)$ to the computing mixer's LO port, where no precoding is used. A signal bandwidth of $B = 100\,\text{MHz}$ is employed in this setup with the wired channel, which is limited by the DAC sampling rate of the USRP X310.

### A. Tupavco TP514 Yagi Directional Antenna

In the wireless setup of WISE, we use the Tupavco TP514 Yagi directional antenna as the TX/RX antenna to establish the wireless link between the central radio and each client. In general, a Yagi antenna consists of multiple parallel resonant antenna elements, which focus the transmitted/received RF signal power in a specific direction. The geometry of these antenna elements is determined by the target operating frequency. As a fully passive component, a Yagi antenna offers a higher gain in the intended direction where the RF signal is concentrated, while exhibiting lower gains in other directions when compared to an ideal

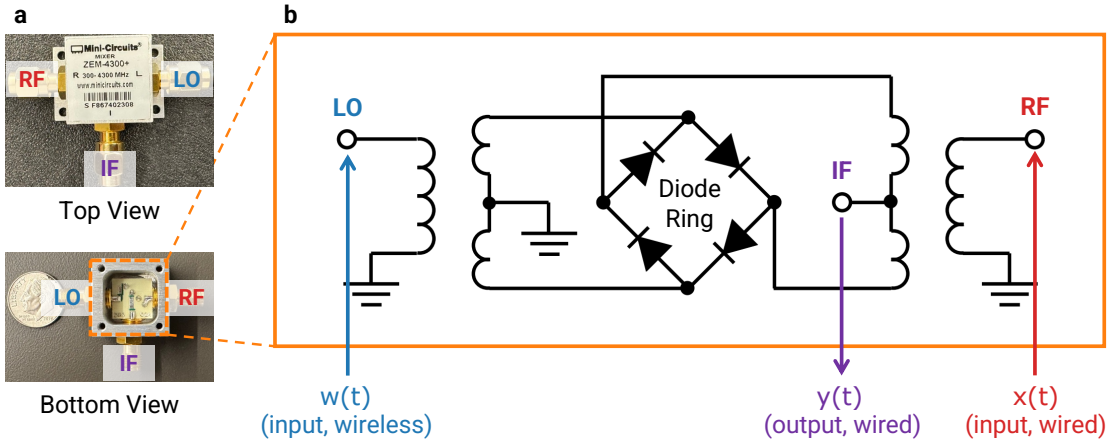


**Fig. S6 The detailed structure of the computing mixer, Mino-Circuits ZEM-4300+ [49]. a**, The external and internal view of the employed frequency mixer, ZEM-4300+. **b**, The schematic of a double-balanced diode mixer composed of a four-diode-bridge, producing an output signal $r_{\rm IF}(t) \propto \mathsf{sgn}(r_{\rm LO}(t)) \cdot r_{\rm RF}(t)$. This mixing process approximates the time-domain multiplication of two RF signals, $r_{\rm LO}(t)$ and $r_{\rm RF}(t)$.

isotropic antenna. Specifically, the Tupavco TP514 Yagi antenna is designed and optimized for dual frequency bands:0.80–0.96 GHz and 1.7–2.5 GHz. This frequency range includes the ISM band at 0.915 GHz utilized in our experiments. The antenna provides an antenna gain of 9 dB in the designated direction, and has horizontal and vertical beamwidths of 65° and 55°, respectively. It is connected to the transceiver radio, a USRP X310 SDR, via an SMA cable.

## B. Computing Frequency Mixer, ZEM-4300+

We exploit the passive double-balanced diode mixer, Mini-Circuits ZEM-4300+ [49], as the computing mixer in the implementation of WISE, as shown in Fig. S6a. The typical schematic of a double-balanced diode mixer is shown in Fig. S6b, whose core component is a four-diode bridge. When performing signal frequency down-conversion, the LO and RF ports serve as the input ports, and the IF port is the output port. Essentially, the waveform input to the LO port, $r_{\rm LO}(t)$, controls which two diodes are on while the other two diodes are off. The on/off status of this four-diode-bridge determines the current direction of the input waveform $r_{\rm RF}(t)$ at the RF port, and thus that of the output waveform $r_{\rm IF}(t)$ on the IF port. Equivalently, this mixer modulates an on-off switching pattern on $r_{\rm RF}(t)$ based on $r_{\rm LO}(t)$; the output waveform on the IF port $r_{\rm IF}(t)$ can thus be formulated as

$$r_{\rm IF}(t) \propto \mathsf{sgn}(r_{\rm LO}(t)) \cdot r_{\rm RF}(t), \text{ where } \mathsf{sgn}(x) = \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ +1, & \text{if } x > 0. \end{cases} \tag{S124}$$

Such on-off switching can be treated as a low-resolution version of the ideal analog multiplication given by equation (S21), where $r_{\rm LO}(t)$ is quantized with an equivalent 1-bit resolution, i.e., "ON" or "OFF". Fortunately, the waveform $r_{\rm LO}(t)$ is a narrowband signal modulated at a carrier frequency such that the quantization noise after the mixing process is mostly distributed across other carrier frequencies, which can be filtered out by an anti-aliasing filter. In our experiments, the residual noise still impacts the computing accuracy of WISE. To mitigate this effect, a sourced analog multiplier, e.g., Gilbert cell [43, 44], can be

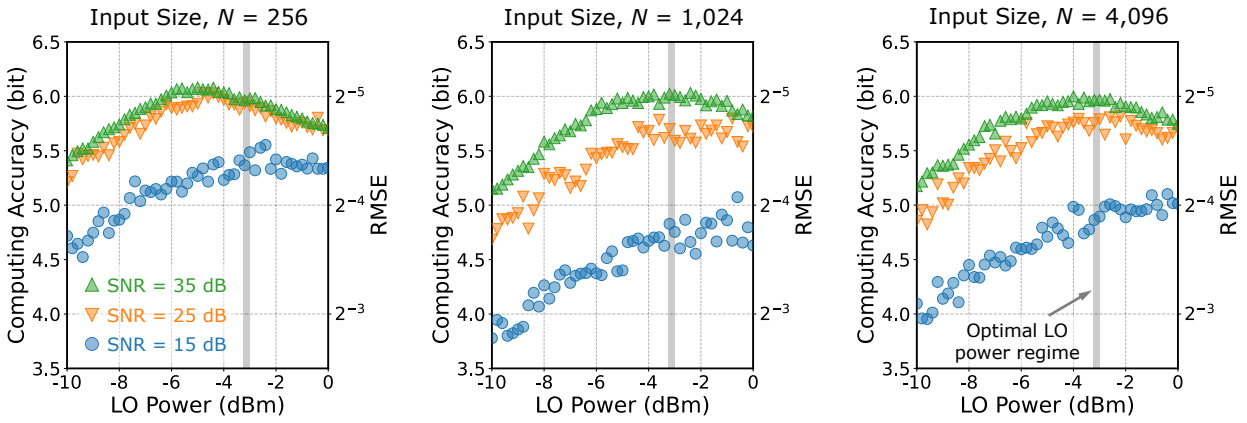


**Fig. S7** Experimental computing accuracy as a function of the input power level to the local oscillator (LO) port of the computing mixer, benchmarked using inner-product (IP) computations across varying input sizes $N = \{256, 1024, 4096\}$. Based on these results, we empirically select an LO power in the range of $-3.00$ dBm to $-3.25$ dBm, which yields optimal computing accuracy.

used for better analog computing performance at the cost of extra energy consumption as it is an active component.

Specifically, the ZEM-4300+ mixer supports an LO and RF frequency range of 0.3–4.3 GHz, where $w(t)$ is modulated to the LO at 0.915 GHz (within the ISM band) and $x(t)$ is modulated to the RF at 1.2 GHz. The mixer supports an IF frequency range of 0–1.0 GHz, which includes the frequency 0.285 GHz to which $y(t)$ modulated. Since the input waveform $w(t)$ to the LO port spans a bandwidth of 25 MHz, different from the typical usage of a frequency mixer, i.e., a single tone signal, the frequency mixer's parameters (e.g., optimal input LO power, insertion loss, etc.) might deviate from that on the datasheet.

We benchmark the optimal LO power for the target in-physics computing tasks, employing a setup with wired transmissions of $w(t)$ to a single client without the channel calibration process, as shown in Fig. S5b. We consider the same inner-product (IP) computation as described in Results section with randomly generated complex-valued vectors $\mathbf{a}$ and $\mathbf{b}$ over the IP dimensions of $N \in \{256, 1024, 4096\}$. We follow the same carrier frequency configuration (0.915 GHz) and bandwidth (25 MHz) setup as described in Methods section, and sweep the signal power of $w(t)$ input to the LO port between $[-10, 0]$ dBm with a step size of 0.2 dB. We also consider three input power levels of $\{-63, -53, -43\}$ dBm into the RF port ($x(t)$), which correspond to three different SNR levels of $\mathsf{SNR} \in \{15, 25, 35\}$ dB. Fig. S7 plots the in-physics computing performance measured by the computing accuracy as a function of the LO power. Overall, the computing accuracy is slightly higher than that in Results section under the wired channel setting. Given the same power of $x(t)$, i.e., the same received SNR of $y(t)$, the LO power of $w(t)$ impacts the computing accuracy. Moreover, the optimal LO power that achieves the best computing accuracy, or the lowest RMSE, reduces as the RF power or SNR increases. For example, with an IP dimension of $N = 4,096$, the optimal LO power is $-4.0$ dBm for $\mathsf{SNR} = 35$ dB, corresponding to an RMSE of 0.031 or ≈6-bit computing accuracy. On the other hand, the optimal LO power is $-0.4$ dBm for $\mathsf{SNR} = 15$ dB, corresponding to an RMSE of 0.058 or ≈5-bit computing accuracy. This trend is plausible since a higher RF power input can compensate for the need for a higher LO power input that activates the frequency mixer into the optimal power regime. To compromise across varying SNR values, we empirically select the LO power in the range of $[-3.25, -3.0]$ dBm in our experiments.

Then, we measure the frequency mixer's insertion loss under the selected LO power, i.e., the power discrepancy of the input power to the RF port and the output power from the IF port. Note that the output

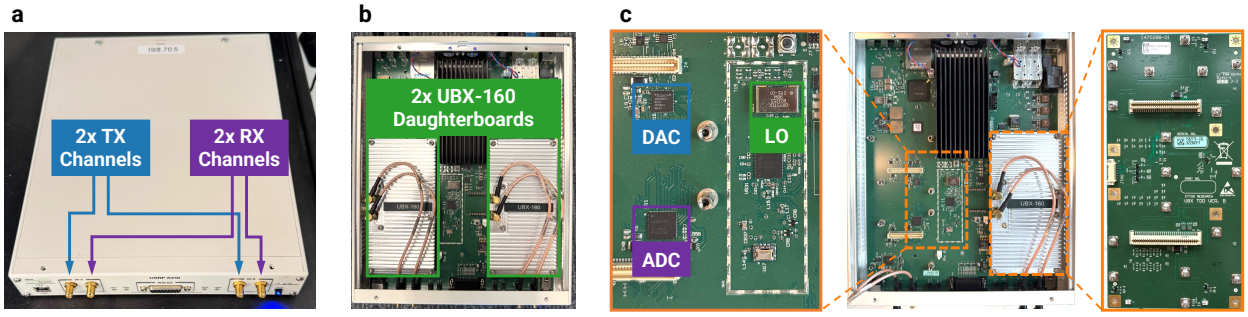


**Fig. S8 A close view of the employed software-defined radio (SDR), USRP X310, in WISE's implementation.** **a–b**, The USRP X310 software-defined radio (SDR) with two UBX-160 daughterboards, supporting two transmit (TX) and two receive (RX) channels. **c**, Close-up view of the USRP X310's internal components, including the digital-to-analog converter (DAC), analog-to-digital converter (ADC), and local oscillator (LO).

power spans over a bandwidth of approximately $2B$ after the convolution, as discussed in Supplementary Section 7. Under this setting, the measured insertion loss of the computing mixer is 11.4 dB, i.e., an efficiency of $\eta_{\text{mixer}} = 0.0724$.

## C. Tranceiver Radio Unit, USRP X310

We use USRP X310 equipped with a UBX-160 daughterboard as the basic transmitter/receiver radio unit, as shown in Fig. S8. Specifically, USRP X310 is a high-performance SDR that supports a carrier frequency of 10 MHz–6 GHz. It is equipped with a 16-bit DAC and a 14-bit ADC per channel, both of which support a sampling rate of 0.196–200 MHz. Note that the TX and RX path includes an LPF as the anti-aliasing filter in the baseband, whose cutoff frequency is default to half of the TX/RX sampling rate or 80 MHz, whichever is smaller. We use a Python-based interface built on top of GNU Radio for radio configuration and data streaming via a 10 Gbps SFP+ interface to a host server.

The TX channel has a gain setting range of 0–31.5 dB at a step size of 0.5 dB, corresponding to a maximum transmitting power of $P_{\max} \approx +23$ dBm with $|s[n]| = 1$, as discussed in Supplementary Section 2. In practice, when transmitting $v(t)$ on the central radio, we consider an average baseband I/Q waveform amplitude of $\sqrt{\mathbb{E}[s_v^2[n]]} = 0.2$, which corresponds to a peak-to-average power ratio (PAPR) of 14 dB without saturation (see Supplementary Section 2). With the TX gain set to 31 dB, the average transmit power of the central radio is $\approx 9$ dBm. For the TX channel that generates $x(t)$, the TX gain is set to 9 dB, and a total of 30 dB attenuators are employed to reduce the transmit power. We sweep the baseband I/Q waveform amplitude for different TX power/energy per MAC settings, whose upper bound is also $\sqrt{\mathbb{E}[s_x^2[n]]} = 0.2$, corresponding to a PAPR of 14 dB. Similarly, the RX channel has the same gain setting range of 0–31.5 dB at the step of 0.5 dB. We configure the RX gain at 20 dB. Together with the frequency mixer, the RX noise figure is measured at 16.9 dB, associated with an energy efficiency $\eta_{\text{nf}} = 2.04 \times 10^{-2}$.

## D. Embedded Anti-Aliasing Filter

We directly use the embedded anti-aliasing filter in USRP X310 as the LPF in WISE, whose cutoff frequency is set to half of the sampling rate, i.e., $f_0 = f_s/2$. To characterize the frequency response of this anti-aliasing filter, we transmit a continuous-wave (CW) signal at the power of $-30$ dBm using the signal generator function within the Keysight N9914B FieldFox Handheld RF Analyzer, and sweep its frequency $f$ around the carrier frequency of $F_y$ =0.285 GHz, at which $y(t)$ is received. Then, the amplitude of the frequency

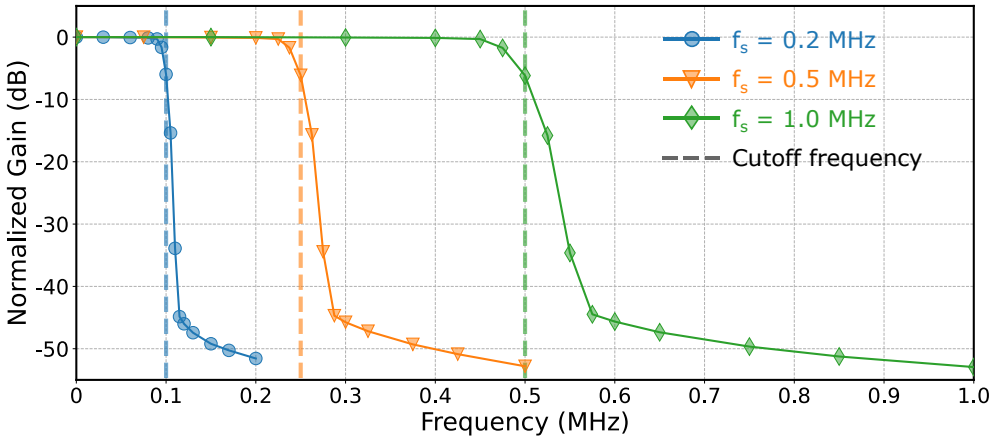


**Fig. S9** Measured normalized frequency response gain of the USRP X310's internal low-pass filter (LPF) at varying sampling rates, $f_s = \{0.2, 0.5, 1.0\}$ MHz, with cutoff frequencies of $\{0.1, 0.25, 0.5\}$ MHz indicated by the dashed lines. A zero-subcarrier padding coefficient of $\alpha > 0.11$ is sufficient to mitigate the roll-off effect of the LPF under these conditions.

response of the anti-aliasing filter is calculated by the power of the received CW signal referred to the power when the CW signal's frequency is swept to exactly 0.285 GHz. Specifically, we consider three low sampling rates employed by the USRP RX $f_s = \{0.2, 0.5, 1.0\}$ MHz, corresponding to the anti-aliasing filter's cutoff frequency as $f_0 = \{0.1, 0.25, 0.5\}$ MHz. The frequency of the CW tone, $f$, is swept with non-uniform step sizes with smaller step sizes around the cutoff frequency of $f_s/2$.

Fig. S9 shows the gain of the frequency response of the anti-aliasing filter as part of the USRP X310 SDR, which is employed at the LPF for WISE. Specifically, the frequency-gain descending slope is proportional to the cutoff frequency $f_0$ or the ADC sampling rate $f_s$. In particular, the gain drops to below $-50$ dB on the stopband, which is sufficient enough to mitigate the frequency aliasing issue for an SNR of 30 dB in our implementation. Also, the gain maintains over $-0.3$ dB at $f = 0.9 \cdot f_0$, which corresponds to the zero-subcarrier padding overhead coefficient of $\alpha = 0.11$. In our MVM implementation, we consider MVM decomposition with $M' = 6$ and $\Delta M = 1$, i.e., $\alpha = 0.33$. This configuration ensures that the padded zero subcarriers are sufficient to compensate for the edge effect of the anti-aliasing filter. In the extreme case of IP computation with $M' = 1$, we still need $\Delta M = 1$, which leads to a large overhead coefficient of $\alpha = 2$.

## E. Wireless Link Distance and Link Budget Analysis

In this section, we examine the wireless link distance between the central radio and client that supports in-physics MVM computation via wireless broadcast of model weights, based on the optimized LO input power and Fig. S7. The link budget equation for a wireless link is given by

$$P_{\mathrm{RX}}[\mathrm{dBm}] = P_{\mathrm{TX}}[\mathrm{dBm}] + G_{\mathrm{TX}}[\mathrm{dBi}] + BF_{\mathrm{TX}}[\mathrm{dB}] - L_{\mathrm{TX}}[\mathrm{dB}] - L_{\mathrm{prop}}[\mathrm{dB}] + G_{\mathrm{RX}}[\mathrm{dBi}] + BF_{\mathrm{RX}}[\mathrm{dB}] - L_{\mathrm{RX}}[\mathrm{dB}], \tag{S125}$$

where $P_{\mathrm{TX}}$ (resp. $P_{\mathrm{RX}}$) denotes the TX (resp. RX) signal power, $G_{\mathrm{TX}}$ (resp. $G_{\mathrm{RX}}$) denotes the TX (resp. RX) antenna gain, $BF_{\mathrm{TX}}$ (resp. $BF_{\mathrm{RX}}$) denotes the TX (resp. RX) beamforming gain if an antenna array is employed for beamforming, $L_{\mathrm{TX}}$ (resp. $L_{\mathrm{RX}}$) denotes the insertion loss on the TX )(resp. RX) due to connectors and cables, etc., and $L_{\mathrm{prop}}$ is the path loss of the wireless link. In particular, we consider the free

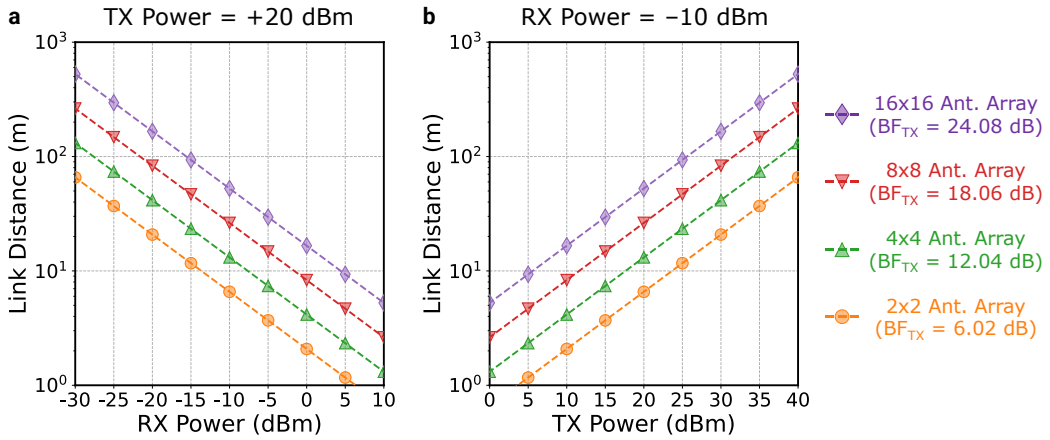


**Fig. S10 The theoretical link distance analysis applying the free-space path loss model. a**, The link distance under varying TX power levels at the central radio, $P_{\mathrm{TX}}$, where the central radio is equipped with an antenna array supporting TX beamforming, and each client is equipped with a signal antenna, and the antenna gain is $G_{\mathrm{TX}} = G_{\mathrm{RX}} = 6\,\mathrm{dBi}$. **b**, The link distance under varying RX power input levels at the LO port of the computing mixer, $P_{\mathrm{RX}}$, under the same setting as **a**.

space path loss [54] given by

$$L_{\mathrm{prop}}[\mathrm{dB}] = 10 \cdot \log_{10}\left(\frac{4\pi dF}{c}\right)^2 = 20 \cdot \log_{10}\left(\frac{4\pi dF}{c}\right), \tag{S126}$$

where $d$ is the link distance between the TX and RX, $F$ is the carrier frequency, and $c$ is the speed of light. Combining equations (S125) and (S126), the link distance, $d$, can be written as

$$d \approx 10^{(P_{\mathrm{TX}} - P_{\mathrm{RX}} + G_{\mathrm{TX}} + BF_{\mathrm{TX}} - L_{\mathrm{TX}} + G_{\mathrm{RX}} + BF_{\mathrm{RX}} - L_{\mathrm{RX}})/20} \cdot \left(\frac{c}{4\pi F}\right). \tag{S127}$$

In our implementation, the USRP X310 supports an average transmit power of $P_{\mathrm{TX}} = +9\,\mathrm{dBm}$ with a PAPR of up to 14 dB, the Yagi antenna provides an antenna gain of $G_{\mathrm{TX}} = G_{\mathrm{RX}} = 9\,\mathrm{dBi}$ with no beamforming ($BF_{\mathrm{TX}} = BF_{\mathrm{RX}} = 0\,\mathrm{dB}$, and the insertion losses $L_{\mathrm{TX}} = L_{\mathrm{RX}} \approx 0\,\mathrm{dB}$ due to short cable length and minimal connections. To feed the LO port with $P_{\mathrm{RX}} = -3\,\mathrm{dBm}$, the path loss $L_{\mathrm{prop}}$ should be 30 dB. Plugging in the carrier frequency of $w(t) = 0.915\,\mathrm{GHz}$, the wireless link distance is recommended to be around 1 meter, as employed in our experiments. This wireless link is determined by the relatively high input power at the computing mixer's LO power required to drive the frequency mixer of our choice. Such a high input power comes from the double-balanced diode architecture of this computing mixer.

To support wireless broadcast of model weight over larger wireless link distances, one can consider using a computing mixer that requires a lower LO input power (i.e., smaller values of $P_{\mathrm{RX}}$). For example, RF mixers integrating an internal LO amplifier (e.g., the PE4152 UltraCMOS quad MOSFET mixer from pSemi) or analog multipliers based on integrated analog correlators [45] can relax the input power constraint on $P_{\mathrm{RX}}$. Another approach is to employ antennas with a higher antenna gain (i.e., larger values of $G_{\mathrm{TX}}$ and/or $G_{\mathrm{RX}}$), or beamforming using an antenna array (i.e., larger values o $BF_{\mathrm{TX}}$ and/or $BF_{\mathrm{RX}}$). Specifically, when referred to a single antenna, a planar antenna array with $N_{\mathrm{ant}} \times N_{\mathrm{ant}}$ antenna elements with half-wavelength spacing between adjacent elements can provide a maximum beamforming gain of

$$BF_{\mathrm{TX}}[\mathrm{dB}] = BF_{\mathrm{RX}}[\mathrm{dB}] = 10\log_{10}\left(N_{\mathrm{ant}}^2\right). \tag{S128}$$

For example, the Argos massive MIMO radio [47] employs a sub-7 GHz antenna array with $N_{\text{ant}} = 8$, which supports a beamforming gain $BF_{\text{TX}}$ of up to 18.06 dB on the central radio. We illustrate the theoretical link distance in Fig. S10, where the central radio employs an antenna array supporting TX beamforming. For example, as shown in Fig. S10a, an 8×8 antenna array, which is commonly employed in modern cellular networks, can support a link distance of 100 meters with an improved TX power of $P_{\text{TX}} = +31.8\,\text{dBm}$.

## F. Time and Frequency Synchronization

Generally, the central radio and each client are not naturally synchronized in the time or frequency domain. Specifically, the client is not aware of the starting point of the transmitted waveform from the central radio, and the LOs on both the central radio and clients may exhibit a carrier frequency offset (CFO), which can lead to inter-subcarrier interference, especially when the subcarrier spacing $\Delta f$ is small. Therefore, we insert preambles into $x(t)$ and $w(t)$, which can be used for time and frequency synchronization between the central radio and each client.

Specifically, given a downsampling ratio $N$, the preamble as an I/Q sample sequence are defined by $\mathbf{s}_{x,\text{pre}} = [s_{x,\text{pre}}[n]] \in \mathbb{C}^{2NL_{\text{pre}}}$ and $\mathbf{s}_{w,\text{pre}} = [s_{w,\text{pre}}[n]] \in \mathbb{C}^{2NL_{\text{pre}}}$ for the baseband I/Q waveforms corresponding to $x(t)$ and $w(t)$, respectively, where $L_{\text{pre}}$ is recommended to be a prime number. These two I/Q waveforms are streamed to the DACs operating at a sampling rate of $f_s$ to generate the analog waveform $x_{\text{pre}}(t)$ and $w_{\text{pre}}(t)$. These two preambles are composed of two identical sequences, each with $NL_{\text{pre}}$ I/Q samples, each of which is generated with a constant amplitude $A$ and randomized phases $\phi_{x,n}$ and $\phi_{w,n}$ for $\mathbf{s}_{x,\text{pre}}$ and $\mathbf{s}_{w,\text{pre}}$, respectively. Specifically, we consider a large amplitude $A$ close to 1 to ensure a high output power and SNR without saturation; the phases are uniformly distributed within $[0, 2\pi]$ to ensure that the signal power is evenly distributed across the frequency band. To sum up, the preamble generation can be written as

$$s_{x,\text{pre}}[n] = s_{x,\text{pre}}[n + NL_{\text{pre}}],\ s_{w,\text{pre}}[n] = s_{w,\text{pre}}[n + NL_{\text{pre}}],\ \forall n = 0, 1, \ldots, NL_{\text{pre}} - 1. \tag{S129}$$

$$s_{x,\text{pre}}[n] = A \cdot e^{j\phi_{x,n}},\ s_{w,\text{pre}}[n] = A \cdot e^{j\phi_{w,n}},\ \forall n = 0, 1, \ldots, NL_{\text{pre}} - 1,\ \text{where } \phi_{x,n}, \phi_{w,n} \sim \mathcal{U}[0, 2\pi]. \tag{S130}$$

Based on equation (S53), the received waveform $y_{\text{pre}}(t)$ experiences a CFO, $\Delta F$, between the TX and RX, i.e.,

$$y_{\text{pre}}(t) \propto x_{\text{pre}}(t) \cdot w_{\text{pre}}(t) \cdot e^{j\Delta F t}. \tag{S131}$$

After the downsampling ratio of $N$, we denote the I/Q waveform corresponding to $y_{\text{pre}}(t)$ as $\mathbf{s}_{y,\text{pre}} = [s_{y,\text{pre}}[n]] \in \mathbb{C}^{2L_{\text{pre}}}$. Assuming that the channel is stable within the transmission time of the preambles, the received $\mathbf{s}_{y,\text{pre}}$ will include two identical sequences given by

$$s_{y,\text{pre}}[n] = s_{y,\text{pre}}[n + L_{\text{pre}}] \cdot e^{j\Delta F \cdot \frac{nN}{f_s}},\ \forall n = 0, 1, \ldots, L_{\text{pre}} - 1. \tag{S132}$$

Therefore, the starting point of $\mathbf{s}_{y,\text{pre}}$ can be detected by performing an auto-correlation with a copy of itself delayed by $L_{\text{pre}}$ I/Q samples [55, 56], i.e.,

$$R_{y,\text{pre}} = \frac{\left|\sum_{n=0}^{L_{\text{pre}}-1} s_{y,\text{pre}}[n] \cdot \bar{s}_{y,\text{pre}}[n+L_{\text{pre}}]\right|}{\sum_{n=0}^{L_{\text{pre}}-1} s_{y,\text{pre}}[n] \cdot \bar{s}_{y,\text{pre}}[n]} \in [0,1]. \tag{S133}$$

In practice, we calculate the auto-correlation, $R_{y,\text{pre}}$, for a sliding window containing $2L_{\text{pre}}$ I/Q samples. When the calculated $R_{y,\text{pre}}$ exceeds a threshold on a given sliding window as a local minimum (e.g., 0.8), a preamble $\mathbf{s}_{y,\text{pre}}$ is considered to be detected, and the starting point of this sliding window is considered as the preamble's starting point. When the starting point of $\mathbf{s}_{y,\text{pre}}$ is detected, we can infer the starting point of the desired waveform $y(t)$ accordingly. Generally, the starting point error of this auto-correlation-based detection algorithm is determined by the sliding window step [55]. As long as we calculate $R_{y,\text{pre}}$ for the sliding windows at the step size of every I/Q sample, a sub-symbol timing offset error can be achieved. Hence, one or two I/Q samples per OFDM symbol for the cyclic prefix is sufficient to ensure the desired synchronization performance.

In addition, the CFO can be estimated by

$$\widehat{\Delta F} = \frac{\mathsf{Angle}\left(\sum_{n=0}^{L_{\text{pre}}-1} s_{y,\text{pre}}[n] \cdot \bar{s}_{y,\text{pre}}[n+L_{\text{pre}}]\right)}{2\pi N L_{\text{pre}}/f_s}. \tag{S134}$$

To calibrate, we can either fine-tune the LO frequency $F_y$, or apply the estimated CFO on the I/Q sample sequence $\mathbf{s}_y$ in the digital domain [55, 56], i.e.,

$$s'_y[n] = s_y[n] \cdot e^{-j\widehat{\Delta F} \cdot \frac{nN}{f_s}}. \tag{S135}$$

To conclude, such a preamble-driven synchronization method allows a short cyclic prefix (e.g., $\Delta L = 1$) and a small subcarrier spacing $\Delta f$ for a large-scale of subcarrier assignment within the accessible bandwidth.

# 13 Channel Calibration Schemes

In this section, we evaluate and compare the performance of WISE across the three schemes: (*i*) the basic scheme (Supplementary Section 8), (*ii*) the **W**-precoding scheme (Supplementary Section 9), and (*iii*) the **x**-precoding scheme (Supplementary Section 10).

## A. General MVM Computation

We first benchmark WISE's three schemes for general complex-valued MVM computation, $\mathbf{y} = \mathbf{W} \cdot \mathbf{x}$. In particular, the IP-based MVM decomposition is considered with randomized $\mathbf{x}$ and squared $\mathbf{W}$ (i.e., $N = M$), where each element $x_n$ and $W_{m,n}$ are independently randomized with uniformly distributed amplitudes $|x|$, $|W| \sim \mathcal{U}[0,1]$ and uniformly distributed phases $\angle x_n$, $\angle W_{m,n} \sim \mathcal{U}[0,2\pi]$. As discussed in Supplementary Section 10, the **W**-precoding scheme is independent of $M$, which means this MVM computation is equivalent to the IP computation in Results section. This MVM computation is also suitable for the basic and **x**-precoding schemes that do not comply with the standalone IP computations. Given the in-phyics MVM

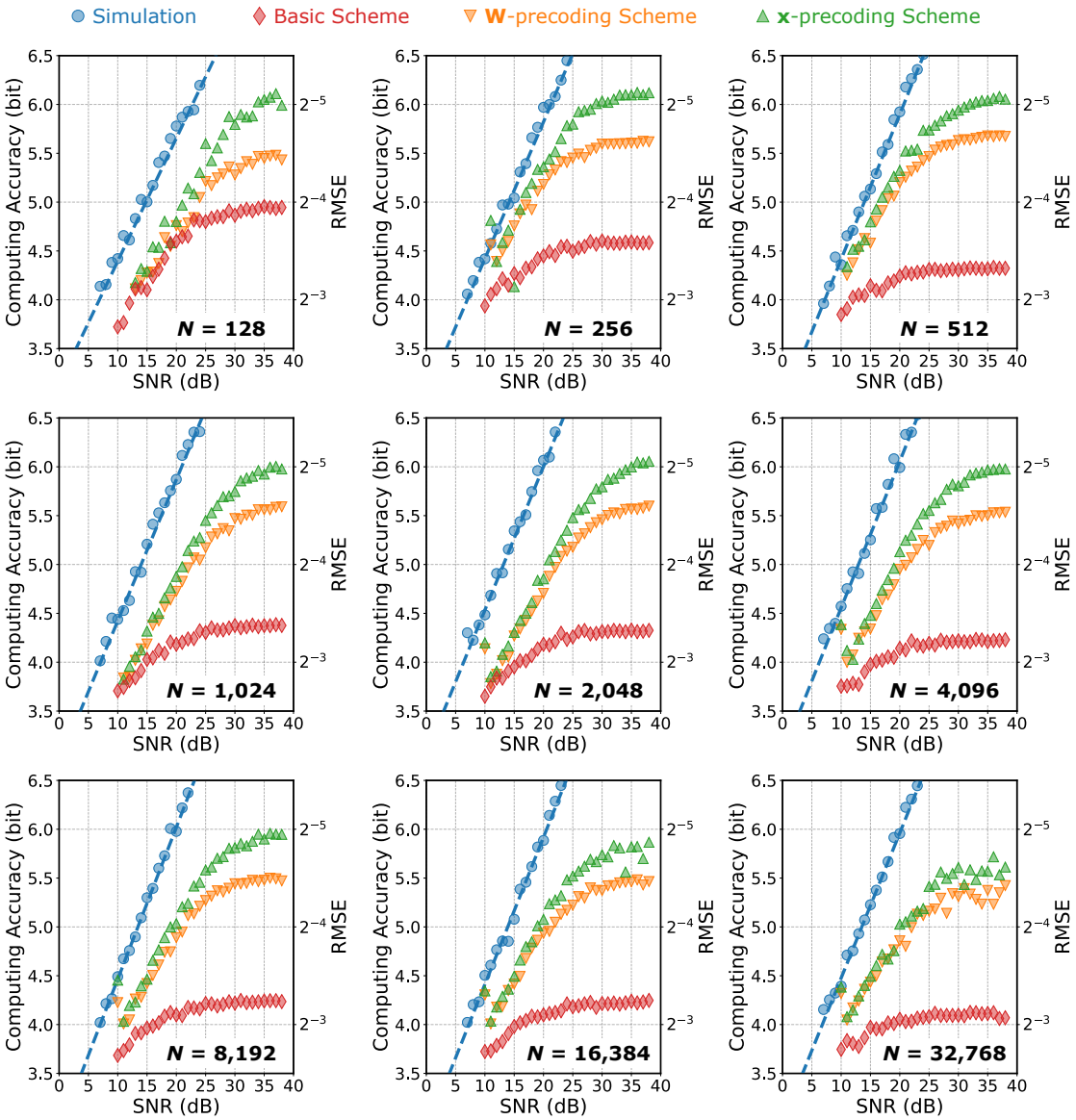


**Fig. S11** Benchmark computing accuracy achieved by WISE's difference schemes compared to simulations for matrix-vector multiplication (MVM) decomposed into inner-products (IPs), with randomized $\mathbf{W}$ and $\mathbf{x}$. Results are shown across varying input/output size, $N = M \in \{2^7, 2^8, \ldots, 2^{15}\}$, and SNR values.

computing result ($\widehat{\mathbf{y}}$) and ground truth ($\mathbf{y}$), we define the RMSE [35, 36] as

$$\mathsf{RMSE} = \sqrt{\mathbb{E}\left[\frac{1}{M} \cdot \sum_{m=0}^{M-1} |\widehat{y}_m - y_m|^2\right]}, \tag{S136}$$

and the computing accuracy can be derived as $-\log_2(\mathsf{RMSE}/2)$ [bit].

Fig. S11 compares the experimental computing accuracy of the three schemes with simulation results under perfect channel calibration and analog multiplication performed by an ideal computing mixer. The comparison is performed under varying MVM dimensions, where $N = M \in \{2^7, 2^8, \ldots, 2^{15}\}$. Overall, the experimental results show that the three schemes require approximately 5 dB higher SNR than the simulations

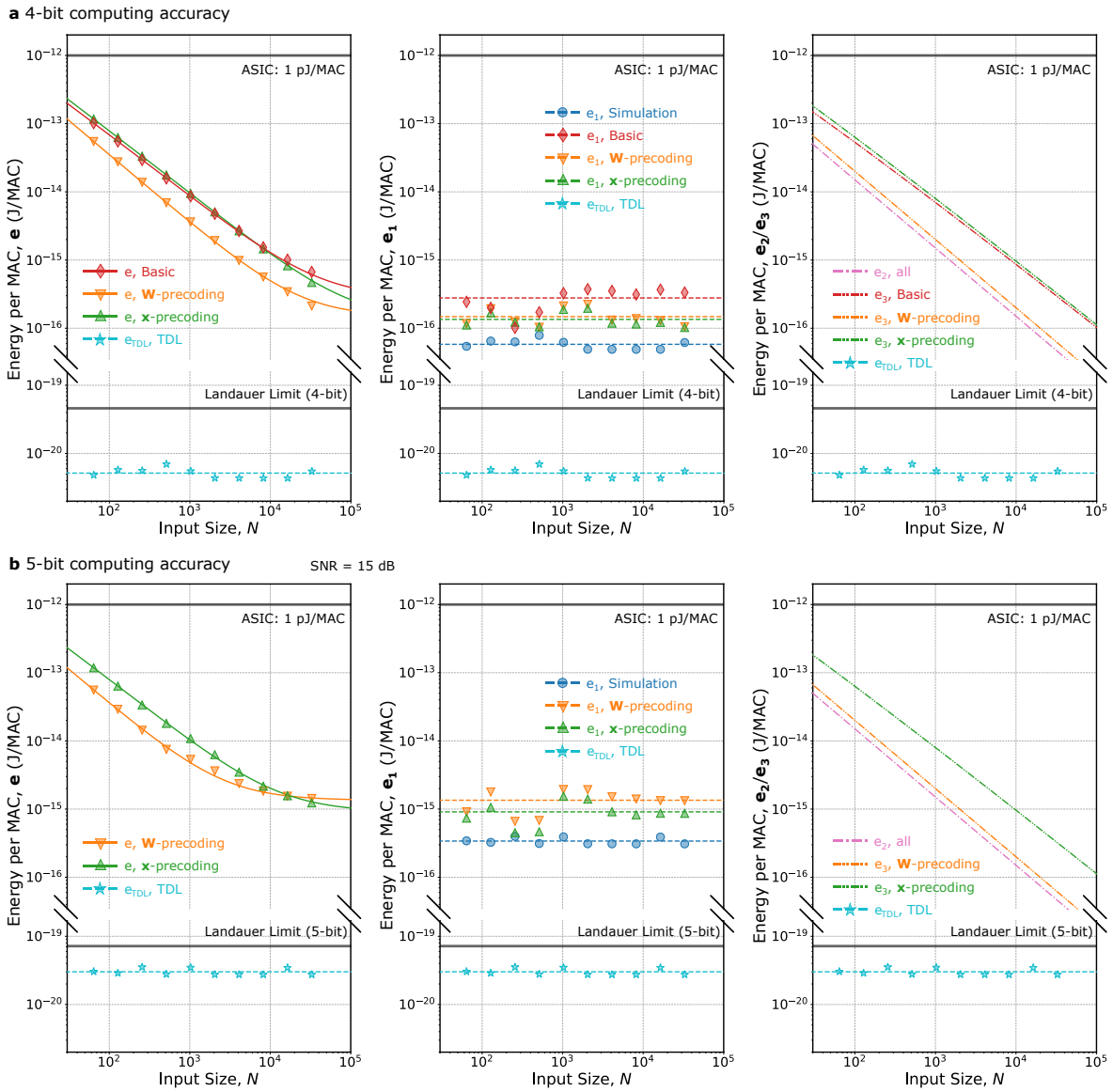


**Fig. S12 The IP-based energy efficiency benchmarking of WISE's basic, W-precoding, and x-precoding schemes.** **a**, The minimum energy per MAC required by the basic, **W**-precoding, and **x**-precoding schemes to achieve RMSE $< 0.125$ (4-bit computing accuracy), with detailed breakdowns. In addition, the energy efficiency corresponding to the thermodynamic limit (TDL) is simulated and compared to the Landauer limit. **b**, The minimum energy per MAC for RMSE $< 0.0625$ (5-bit computing accuracy) of the **W**-precoding and **x**-precoding schemes. Note that the basic scheme cannot achieve the 5-bit computing accuracy and is thus not shown.

to achieve the same computing accuracy. The experimental computing accuracy is limited in the high SNR regime (e.g., >30 dB) due to the on-off switching behavior of the double-balanced diode mixer, as discussed in Supplementary Section 12. Across all values of $N$, the **W**-precoding and **x**-precoding schemes achieve higher computing accuracy than the basic scheme. For example, at 25 dB SNR, the RMSE of the **W**-precoding scheme is 0.055/0.056 and RMSE of the **x**-precoding is 0.043/0.047 with $N = 4,096/32,768$, corresponding to >5-bit computing accuracy. In contrast, the RMSE of the basic scheme is 0.109/0.118, corresponding to ≈4-bit computing accuracy. This performance gap highlights the effectiveness of the wireless channel calibration of the **W**-precoding and **x**-precoding schemes. In addition, at higher SNR levels, the **x**-precoding scheme outperforms the **W**-precoding scheme, achieving an RMSE of 0.032/0.042 at 35 dB SNR

with $N = 4,096/32,768$, equivalent to ≈6-bit computing accuracy. This is because the $\mathbf{x}$-precoding scheme supports CSI estimation and calibration for individual clients. Moreover, as $N$ increases, the computing accuracy achieved by all three schemes degrades, especially for the basic scheme. This degradation occurs since a larger value of $N$ results in reduced subcarrier spacing, $\Delta f$, as more subcarriers are packed within the same bandwidth, $B$. In this case, the impact of inter-subcarrier interference becomes more significant, requiring finer frequency synchronization.

Next, we benchmark the energy efficiency of the three schemes across different input sizes, $N$, based on the energy efficiency analysis for IP-based MVM decomposition given by (S90), (S108), and (S120), respectively. The minimum energy per MAC required for the three schemes to achieve $\mathsf{RMSE} < 0.125$ and $\mathsf{RMSE} < 0.0625$ (4-bit and 5-bit computing accuracy, respectively) is shown in Fig. S12. Note that the energy efficiency of the basic scheme to reach 5-bit computing accuracy is excluded since it cannot achieve this computing accuracy across all considered SNR values. As summarized in Table S1, the energy efficiency ($e_{\text{mvm}}$) of the basic and $\mathbf{x}$-precoding schemes scales as $\mathcal{O}(\frac{1}{N}\log N)$ and the energy efficiency of the $\mathbf{W}$-precoding scheme scales $\mathcal{O}(1/N)$. With large values of $N$, $e_{\text{mvm}}$ converges to the energy efficiency corresponding to the waveform generation, $e_1$. Specifically, with $N = 4,096$, the $\mathbf{W}$-precoding scheme has an energy efficiency of 0.99 fJ/MAC and 2.37 fJ/MAC (1,014.20 TOPS/W and 422.14 TOPS/W) for achieving $\mathsf{RMSE} < 0.125$ (4-bit) and $\mathsf{RMSE} < 0.0625$ (5-bit), respectively, while the energy efficiency for the $\mathbf{x}$-precoding scheme is 3.17 fJ/MAC and 3.96 fJ/MAC (315.44 TOPS/W and 252.26 TOPS/W). The energy saving of the $\mathbf{W}$-precoding scheme comes from the elimination of the FFT-based encoding and precoding for wireless channel calibration, as discussed in Supplementary Section 9, and thus a lower digital computing cost from $e_3 = 2.69$ fJ/MAC for the $\mathbf{x}$-precoding scheme to $e_3 = 0.49$ fJ/MAC for the $\mathbf{W}$-precoding scheme. On the other hand, the extra energy for digital computing is averaged down as the $N$ and/or $M$ increases. For example, at $N = 32,768$, the energy efficiency becomes 0.21 fJ/MAC and 1.43 fJ/MAC (4,710.54 TOPS/W and 697.01 TOPS/W) for the $\mathbf{W}$-precoding scheme to reach $\mathsf{RMSE} < 0.125$ (4-bit) and $\mathsf{RMSE} < 0.0625$ (5-bit), respectively, which is 0.53 fJ/MAC and 1.29 fJ/MAC (1,888.44 TOPS/W and 774.47 TOPS/W) for the $\mathbf{x}$-precoding scheme. By assuming an ideal channel calibration and perfect hardware with no overhead, we also simulate the TDL energy efficiency given by $e_{\text{tdl}}$ in equations (S87), (S106), and (S118), following the same forms over the three schemes. The TDL energy efficiency averaging across all input sizes, $N$, is 5.15 zJ/MAC and 30.15 zJ/MAC (194.17 EOPS/W and 33.17 EOPS/W) for the 4-bit and 5-bit computing accuracy, which is 9.1× and 2.4× lower than the corresponding 4-bit and 5-bit Landauer limit of 45.9 zJ/MAC and 71.8 zJ/MAC, respectively.

## B. Image Classification on the MNIST Dataset

We implement WISE on a complex-valued model with three FC layers based on LeNet-300-100 [33] for handwritten digit image classification on the MNIST dataset. Fig. S13a shows the energy efficiency of WISE on the MNIST dataset. For the MNIST dataset, the maximum input size is $N = 784$, where the digital computing energy efficiency term $e_3$ dominates the total energy efficiency $e_{\text{mvm}}$. The energy efficiency required by the $\mathbf{W}$-precoding scheme to achieve a classification accuracy of 90% is 4.62 fJ/MAC (216.35 TOPS/W), which include $e_1 = 0.47$ fJ/MAC for waveform generation and I/Q (de)modulation with 18.3 dB SNR, $e_2 = 1.04$ fJ/MAC for I/Q sampling, and $e_3 = 3.11$ fJ/MAC for decoding performed in digital computing. For the $\mathbf{x}$-precoding scheme, the energy efficiency to achieve 90% classification accuracy is 28.94 fJ/MAC (35.55 TOPS/W), with a breakdown of $e_1 = 0.30$ fJ/MAC (at 16.3 dB SNR), $e_2 = 1.04$ fJ/MAC, and $e_3 = 27.60$ fJ/MAC. Despite the degraded energy efficiency of the $\mathbf{x}$-precoding scheme, it is still significantly better than that of digital

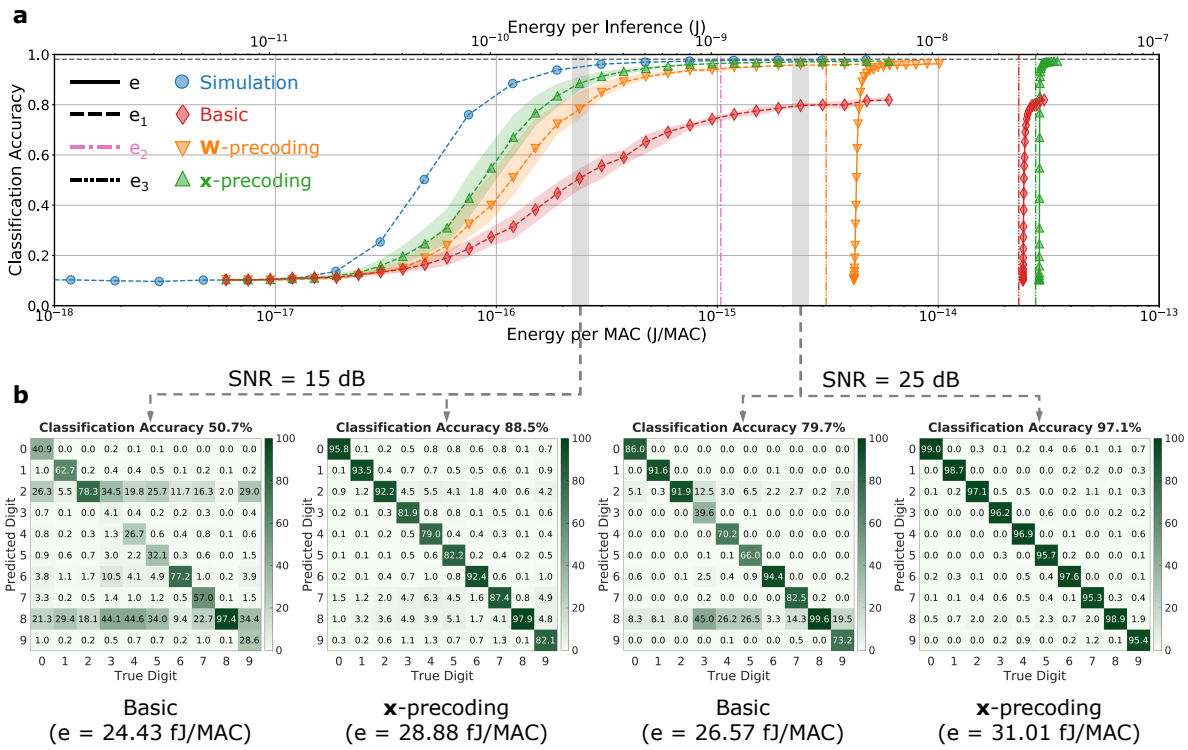


**Fig. S13 The basic, W-precoding and x-precoding scheme comparison on the image classification task on the MNIST dataset. a**, Classification accuracy achieved by WISE's three schemes and simulation, shown as a function of the energy efficiency, $e_{\text{mvm}}$, with a detailed breakdown into $e_1$, $e_2$, and $e_3$. The shaded area of $e_1$ indicates the accuracy variance across three clients under the same SNR. **b**, Confusion matrices of the basic scheme without channel calibration and the **x**-precoding scheme with per-client channel calibration at 15 dB and 25 dB SNR.

computing using state-of-the-art ASICs at 1 pJ/MAC [7, 39, 40]. Note that the basic scheme, however, can only achieve a classification accuracy of up to 81.9% on the MNIST dataset.

Detailed confusion matrices of the classification accuracy achieved by the basic and **x**-precoding schemes at 15 dB and 25 dB are shown in Fig. S13b. Due to the lack of CSI estimation and calibration, the basic scheme achieves the classification accuracy of only 50.7% and 79.7% under 15 dB and 25 dB SNR, respectively. In contrast, with proper CSI estimation and calibration, the **W**-precoding scheme achieves a classification accuracy of 78.2% and 95.7%, and the **x**-precoding scheme achieves a classification accuracy of 88.5% and 97.1% under the same SNR values. With an increased SNR of 29.3 dB, the **x**-precoding scheme achieves a maximum classification accuracy of 97.4%, which is only 0.7% lower than the classification accuracy of 98.1% based on digital computing.

## C. Audio Signal Classification on the AudioMNIST Dataset

We also evaluate the performance of WISE on the AudioMNIST dataset [41] with spoken digits for audio signal classification, using a complex-valued model with three FC layers based on LeNet-300-100, with an input size of $N = 4,000$. Fig. S14a shows the energy efficiency of WISE achieved by the three schemes. Overall, the energy efficiency of WISE is lower on the AudioMNIST dataset compared to that on the MNIST dataset due to the large input size of $N = 4,000$. Specifically, achieving an accuracy of 90% by the **W**-precoding scheme requires a minimum SNR of 15.3 dB and an energy efficiency of 1.13 fJ/MAC (882.10 TOPS/W), which includes $e_1 = 0.24$ fJ/MAC, $e_2 = 0.22$ fJ/MAC, and $e_3 = 0.67$ fJ/MAC. Moreover, the **x**-precoding

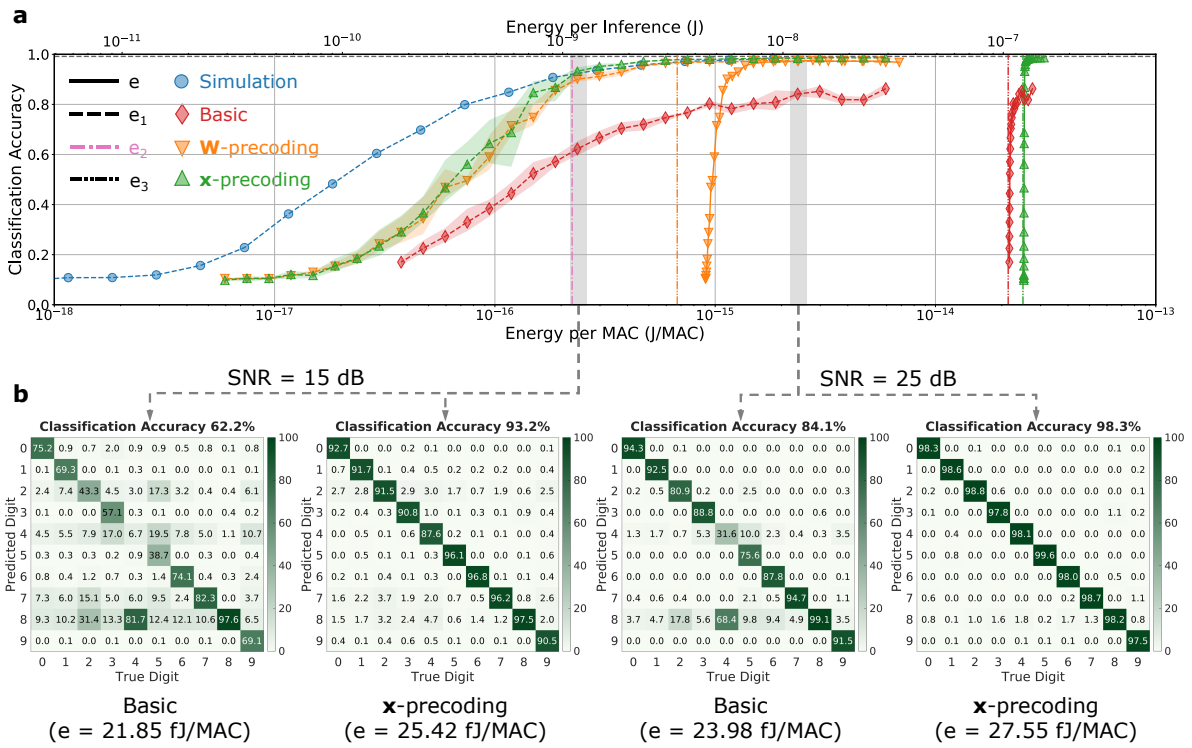


**Fig. S14 The basic, W-precoding and x-precoding scheme comparison on the audio signal classification task on the MNIST dataset. a**, Classification accuracy achieved by WISE's three schemes and simulation, shown as a function of the energy efficiency with detailed breakdowns. The shaded area of $e_1$ indicates the accuracy variance across three clients under the same SNR. **b**, Confusion matrices of classification accuracy achieved by the basic scheme without channel calibration and the **x**-precoding scheme with per-client channel calibration at 15 dB and 25 dB SNR.

scheme requires a minimum SNR of 15.3 dB and an energy efficiency of 25.42 fJ/MAC (33.34 TOPS/W), which can be decomposed into $e_1 = 0.24$ fJ/MAC, $e_2 = 0.22$ fJ/MAC, and $e_3 = 24.96$ fJ/MAC. Similarly, the degraded energy efficiency of the **x**-precoding scheme results from the encoding and precoding performed in digital computing, which dominates the total energy efficiency, given the problem size of the AudioMNIST dataset. In this case, the **x**-precoding scheme achieves an energy efficiency gain of approximately 40× compared to the 1 pJ/MAC energy efficiency by the state-of-the-art ASICs. Beyond, as the MVM scales up on the future DL tasks, e.g., Llama-2-7b [6] with $N = 11,008$, the energy efficiency of both the **W**-precoding and **x**-precoding schemes can be further improved.

Detailed confusion matrices of the classification accuracy achieved by the basic and **x**-precoding schemes at 15 dB and 25 dB are shown in Fig. S14b. It can be seen that the basic scheme achieves a classification accuracy of 62.2% and 84.1% under 15 dB and 25 dB SNR, respectively, and is bounded by 86.3% with further increased SNR values. On the other hand, the **x**-precoding scheme that exploits the per-client CSI estimation and calibration archives a classification accuracy of 93.2% and 98.3% under 15 dB and 25 dB SNR, outperforming the **W**-precoding scheme that achieves a classification accuracy of 90.1% and 97.2%. Further, it achieves a maximum accuracy of 98.6%, only 0.6% lower compared to the classification accuracy of 99.2% based on digital computing.

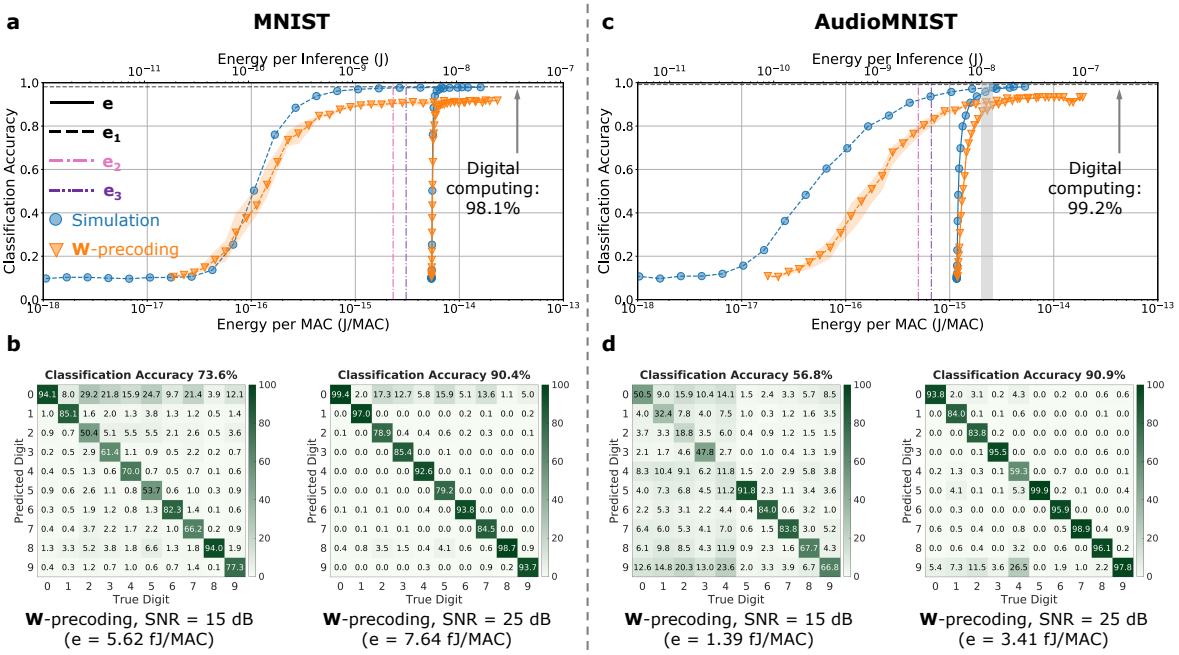


**Fig. S15 The DL inference performance on the MNIST and AudioMNIST by WISE's W-precoding scheme with IP-based MVM decomposition ($M' = 1$). a**, Classification accuracy on the MNIST and AudioMNIST datasets as a function of the energy efficiency with detailed breakdowns. The shaded area of $e_1$ indicates the accuracy variance across three clients under the same SNR. **b**, Confusion matrices of classification accuracy achieved by the **W**-precoding scheme at 15 dB and 25 dB SNR.

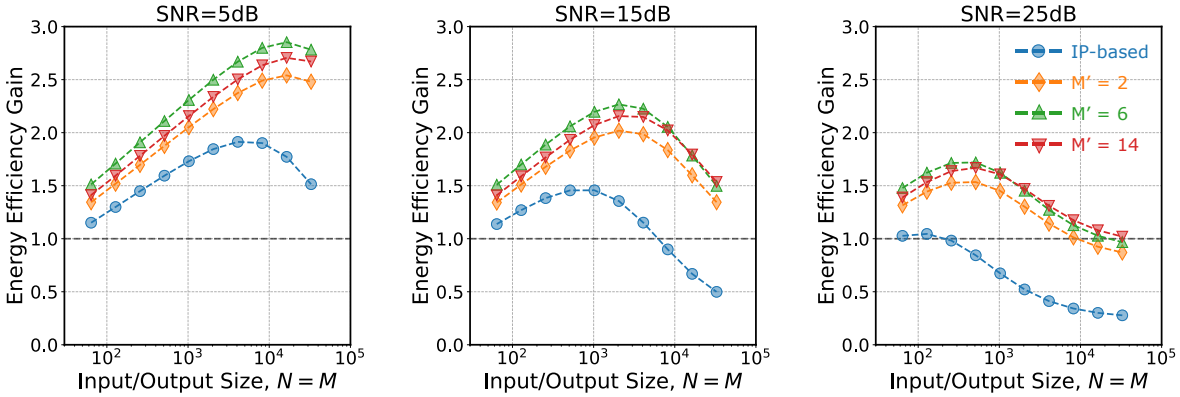


**Fig. S16** Energy efficiency gain achieved by WISE's **W**-precoding scheme across varying MVM decompositions, $M' = \{1, 2, 6, 14\}$, compared to the baseline without MVM decomposition. The zero padding overhead is set to $\Delta M = 1$ ($\alpha = \frac{2}{M'}$, and the cyclic prefix overhead $\beta$ is selected assuming a single I/Q sample $\Delta L = 1$ as the cyclic prefix, i.e., $\beta = \frac{1}{M'+2}$.

# 14 MVM Decomposition into IPs

For the **W**-precoding scheme, we now consider the case where each MVM is decomposed into $M$ IPs, i.e., $M' = 1$, with minimal FFT size at the cost of a slightly higher overhead of $\alpha = 2$ and $\beta = 0.33$, as discussed in Supplementary Section 9. This IP-based MVM decomposition has a lower computation throughput of $\Lambda = 75$ MOPS across three clients, and its energy consumption $e'_{\text{mvm}}$ is given by equation (S108). Fig. S15a shows the confusion matrices on the MNIST dataset under SNR $= 15/25$ dB, whose classification accuracies are 73.6%/90.4% for MNIST, lower than the performance with $M' = 6$ as WISE's default choice in Results section. This performance degradation comes from the potentially higher PAPR on

$x(t)$, $w(t)$, and $y(t)$, which incurs a relatively higher saturation level on the DACs/ADCs. Fig. S15b shows the energy efficiency of this IP-based MVM decomposition. Specifically, to achieve a classification accuracy of 90% on MNIST, the energy efficiency is $e'_{\text{mvm}} = 7.64\,\text{fJ/MAC}$ (130.85 TOP/W), including the breakdown of $e_1 = 2.25\,\text{fJ/MAC}$ for an SNR of 25.1 dB, $e_2 = 2.31\,\text{fJ/MAC}$, and $e_3 = 3.08\,\text{fJ/MAC}$. The same experiments are repeated on the AudioMNIST. Fig. S15c shows the classification accuracy of 56.8% and 90.9% under the SNR of 15 dB and 25 dB. The energy efficiency analysis is further shown in Fig. S15d, where a minimum energy efficiency $e'_{\text{mvm}} = 3.41\,\text{fJ/MAC}$ (293.17 TOPS/W) is needed to achieve a classification accuracy of 90%. This energy efficiency corresponds to $e_1 = 2.25\,\text{fJ/MAC}$, $e_2 = 0.50\,\text{fJ/MAC}$, and $e_3 = 0.67\,\text{fJ/MAC}$.

We further investigate the optimal value of $M'$ for the MVM decomposition based on energy efficiency. Specifically, the energy efficiency is simulated based on equations (S105) and (S108), which is normalized by the energy efficiency without the MVM decomposition, as shown in Fig. S16. For the IP-based decomposition, despite the smallest energy for the term $e_3$, the overhead of $\alpha$ and $\beta$ results in degraded energy efficiency on the term $e_1$, especially in the high SNR regime. As a result, for $N = 4,096$, the energy efficiency gain is 1.92×, 1.16×, and 0.41× to under an SNR value of 5 dB, 15 dB and 25 dB, respectively. We consider three levels of MVM decompositions with $M' = \{2, 6, 14\}$, which correspond to the FFT sizes of $\{4, 8, 16\}$ after attaching the padded zero-subcarriers with $\Delta M = 1$. Among these three decomposition levels, $M' = 6$ achieves the highest energy efficiency gain of 2.70×, 2.22×, and 1.27× for the three SNR levels for $N = 4,096$. This is due to a smaller overhead of $\alpha$ and $\beta$ compared to $M' = 2$, and a smaller FFT size compared to $M' = 14$. To conclude, we empirically select $M' = 6$ for the MVM decomposition used by the **W**-precoding scheme, as described in Methods section. Similar conclusions can also be extended to the basic scheme and the **x**-precoding scheme.

## 15 A Case Study of WISE on a Three-Layer DL Model

We show a detailed workflow of how WISE performs inference on an image of a handwritten digit '4' in MNIST on a 3-FC layer DL model, shown in Fig. S17. In particular, the 28×28-pixel image of digit '4' is formed as a 28×28 real-valued matrix, and is then flattened into a 784-element vector. This vector is then modulated with a 784-point Zadoff-Chu (ZC) phase sequence $\mathbf{\Phi}_{\text{zc}}$ as defined in Methods section, which yields a 784-element complex-valued vector, $\mathbf{x}^{(1)} \in \mathbb{C}^{784}$, as the input to the first FC layer.

The first FC layer has an input size of $N^{(1)} = 784$ and an output size of $M^{(1)} = 300$. The MVM decomposition technique described in Supplementary Section 8 first decomposes the entire MVM with $\mathbf{W}^{(1)} \in \mathbb{C}^{300\times 784}$ into 50 smaller MVMs, each of which has a smaller output size of $M' = 6$. We employ a zero-subcarrier padding overhead of $\alpha = 0.33$ with $\Delta M = 1$, which extends the output dimension per decomposed MVM to $(1 + \alpha)M' = 8$. For each decomposed MVM, the time-encoded input sequence $\mathbf{s}_x$ contains eight duplicated $\mathbf{x}^{(1)}$ in the time domain, with a total number of 6,272 I/Q samples. Finally, we employ a cyclic prefix overhead coefficient of $\beta = 0.25$, which further appends two copies of $\mathbf{x}^{(1)}$ to the front of $\mathbf{s}_x$. The resulting I/Q waveform streamed to the DACs has 7,840 I/Q samples. With a DAC sampling rate of $f_s = 25\,\text{MHz}$, the generated waveform $x(t)$ has a duration of 0.314 ms. The waveform $x(t)$ is then I/Q modulated to the carrier frequency of $F_x = 1.2\,\text{GHz}$. Similarly, at the central radio, the model weights for each decomposed MVM are encoded into I/Q waveform $w(t)$, which is then I/Q modulated to the carrier frequency of $F_w = 0.915\,\text{GHz}$. The model weights are then broadcast wirelessly to the client.

On the client side, $x(t)$ is mixed with the received waveform $w(t)$ and filtered by the LPF with a cutoff frequency of $f_0 = 15.9\,\text{kHz}$, the output waveform $\mathsf{LPF}\{y(t)\}$ is sampled by two I/Q ADCs operating at

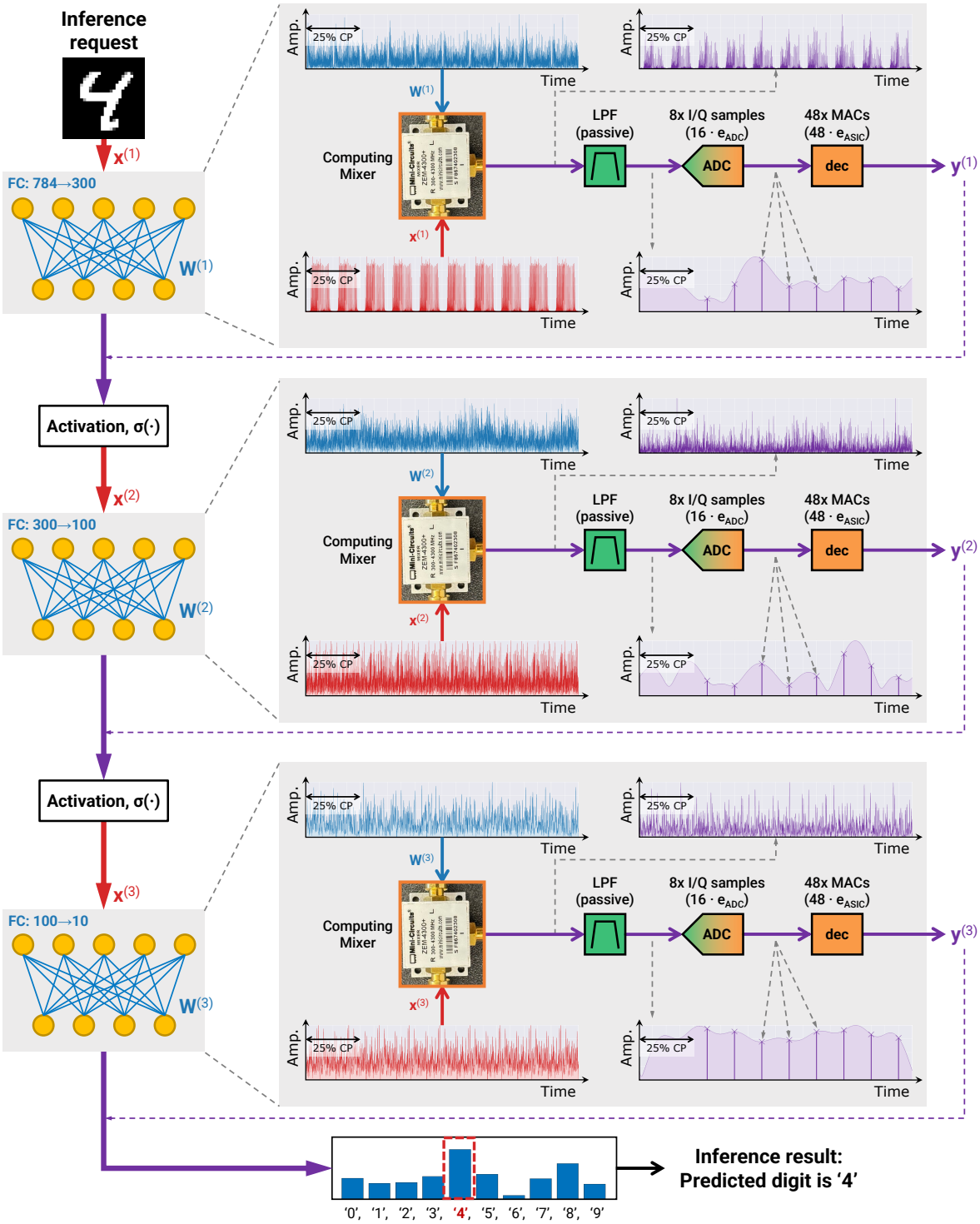


**Fig. S17** Example workflow of WISE: A complex-valued model with three FC layers processes an inference request to predict the digit '4' from a handwritten image. For each FC layer, we show the amplitudes of the time domain waveform $x(t)$, $w(t)$, $y(t)$ before/after the LPF, the sampled I/Q samples after ADC $\mathbf{s}_{y\downarrow}$, of a single decomposed MVM.

31.9 kHz to obtain $\mathbf{s}_{y\downarrow} \in \mathbb{C}^8$, where the first 25% waveform is excluded as the cyclic prefix. Then, an 8-point FFT is performed on $\mathbf{s}_{y\downarrow}$ via digital computing, which yields the subcarrier symbols $\mathbf{S}_{y\downarrow} \in \mathbb{C}^8$ of eight complex-valued symbols. Finally, the middle six symbols in $\mathbf{S}_{y\downarrow}$ are considered as the output $\mathbf{y}' \in \mathbb{C}^6$ of

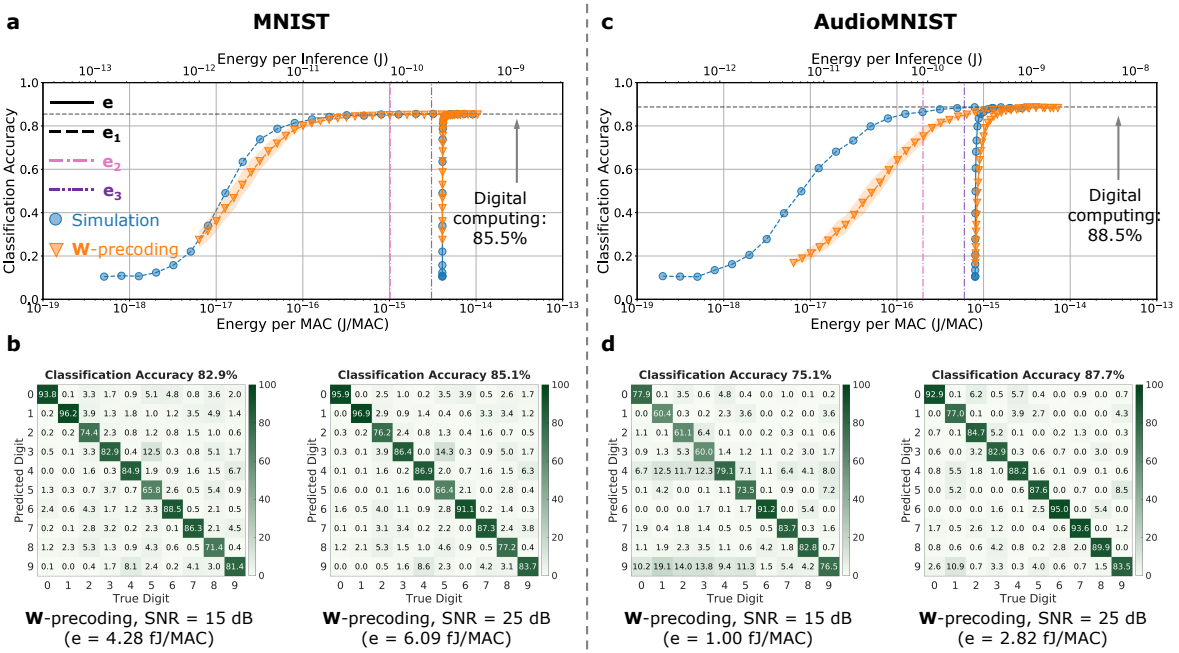


**Fig. S18 The DL inference performance on the MNIST and AudioMNIST by a fully analog linear regression model, which skips the digital computing-based activation functions in the middle layers, and the W-precoding scheme is applied a**, The classification accuracy on MNIST under different energy efficiency $e_{\text{mvm}}$ of the fully analog linear regression model. The shadow area indicates the accuracy variance over the three users. **b**, Under 15/25 dB SNR, the confusion matrices on the MNIST dataset by fully analog linear regression model. **c**–**d**, The energy efficiency analysis and the confusion matrices on the AudioMNIST dataset, respectively.

the decomposed MVM; concatenating $\mathbf{y}$ from all $M/M' = 50$ decomposed MVMs yields the final output of the first FC layer, $\mathbf{y}^{(1)} \in \mathbb{C}^{300}$. This output is passed through the activation function, $\sigma_{300}(\cdot)$, including the absolute function and phase modulation with a 300-point Zadoff-Chu sequence $\mathbf{\Phi}_{\text{zc}}$, which generates the input to the second FC layer, $\mathbf{x}^{(2)}$.

This process is repeated three times, one for each layer, to obtain the output of the last FC layer, $\mathbf{y}^{(3)} \in \mathbb{C}^{10}$. The amplitude of the final output, $\left|\mathbf{y}^{(3)}\right|$ represents the probability of the input image being one of the ten digits '0' to '9'. In this example, the $5^{\text{th}}$ element has the highest amplitude, corresponding to the classification result of digit '4'.

## 16 A Fully Analog Linear Regression Model

We also consider a small linear regression model [57] with a single complex-valued FC layer. For the MNIST and AudioMNIST datasets, only one FC layer of 784/4,000×10 complex-valued parameters transfers the 784/4,000-element input $\mathbf{x}$ into the 10-element output $\mathbf{y}$ for the likelihood of the input being one of the ten digits. Since there is only one FC layer, no nonlinear activation function with absolute function and Zadoff-Chu phase sequence is applied. The absolute function after the FC layer can be realized by directly measuring the absolute power of each subcarrier in $\mathbf{S}_{y\downarrow}$. Therefore, this linear regression model only requires a single transmission without digitally performing absolute functions. On the other hand, the one-time-FFT-based decoding is still required to extract $\mathbf{y}$ from the time-domain waveform $y(t)$, which can be done either in digital computing or internally when being received by a spectrum analyzer. Similar to the LeNet-300-100 models, we train the linear regression model using the Adam optimizer [48] with a learning rate of $1.0 \times 10^{-3}$ over 100 epochs and cross-entropy as the loss function.

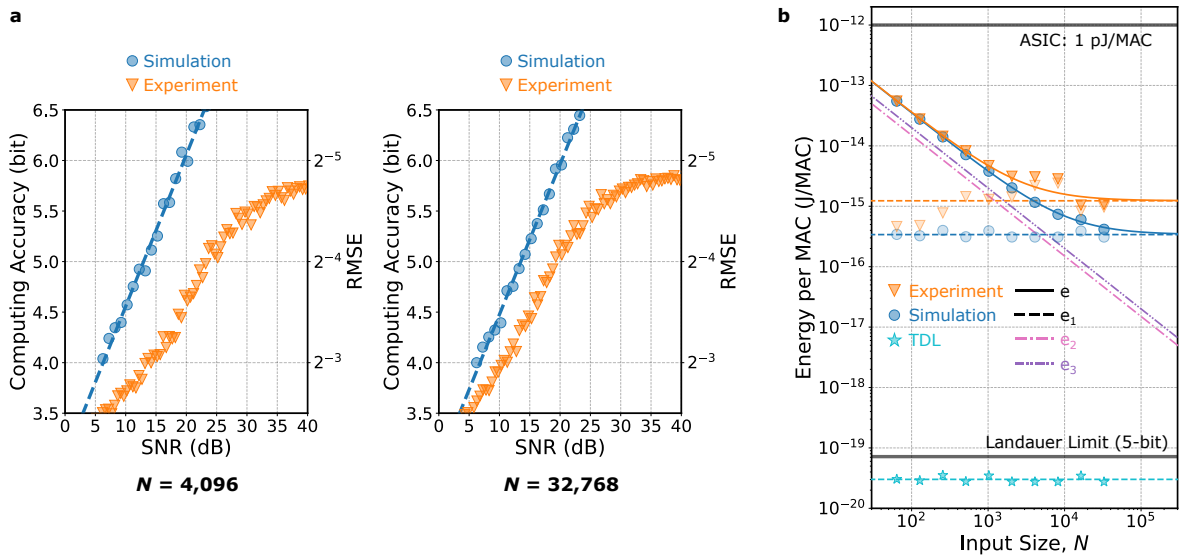


**Fig. S19 General MVM computing accuracy achieved by WISE's basic scheme compared to simulations over a 100 MHz wired channel. a**, Computing accuracy across varying SNR levels with input sizes $N = \{4096, 32768\}$. **b**, Energy efficiency required to achieve RMSE $< 0.0625$ (5-bit computing accuracy) across varying input sizes, $N$, with its breakdown.

Using digital computing, this linear regression model achieves a classification accuracy of 85.5% on the MNIST dataset. Fig. S18a shows the energy efficiency of the linear regression model on the MNIST dataset. Specifically, this linear regression model achieves a classification accuracy of 80% at energy efficiency of $e_{\text{mvm}} = 4.18\,\text{fJ/MAC}$ (239.23 TOP/W), with a breakdown of $e_1 = 0.10\,\text{fJ/MAC}$, $e_2 = 1.02\,\text{fJ/MAC}$, and $e_3 = 3.06\,\text{fJ/MAC}$. This linear regression model only involves 31,360 MACs, which corresponds to the total energy consumption of 131.08 pJ per inference. Compared to the LeNet-300-100 model at a classification accuracy of 80%, this linear regression model consumes 36.4× less energy per inference. As shown in Fig. S18b, WISE achieves a classification accuracy of 82.9% and 85.1% under 15 dB and 25 dB, respectively. The accuracy gap between digital computing and the in-physics computing of WISE of is only 0.4%, which is smaller compared to that of the LeNet-300-100 model described in Results section. This is due to the shallow model architecture, where errors introduced during the in-physics computing process do not accumulate across layers.

On the AudioMNIST dataset, the classification accuracy of this linear regression model with full-precision digital computing is 88.5%. Fig. S18c shows the energy efficiency of the in-physics computing by WISE and the corresponding classification accuracy. To achieve 80% classification accuracy, the energy efficiency of WISE is $e_{\text{mvm}} = 1.12\,\text{fJ/MAC}$ (892.86 TOPS/W), which includes $e_1 = 0.32\,\text{fJ/MAC}$, $e_2 = 0.22\,\text{fJ/MAC}$, and $e_3 = 0.67\,\text{fJ/MAC}$. Given the total number of 160,000 MACs involved in the model, this energy efficiency corresponds to an energy consumption of 179.35 pJ/MAC per inference, which is only 29.8× lower compared to that of the LeNet-300-100 model at the same classification accuracy. Moreover, as shown in Fig. S18d, the linear regression model achieves a classification accuracy of 75.1% and 87.7% at 15 dB and 25 dB SNR.

## 17 WISE over Wired Channels

To comply with the ISM band regulations, our wireless experiments are conducted using a bandwidth of 25 MHz, which limits the computation throughput $\Lambda$. On the other hand, a larger bandwidth $B$ can proportionally increase $\Lambda$ without increasing the energy consumption $e_{\text{mvm}}$. In this section, we consider a wired

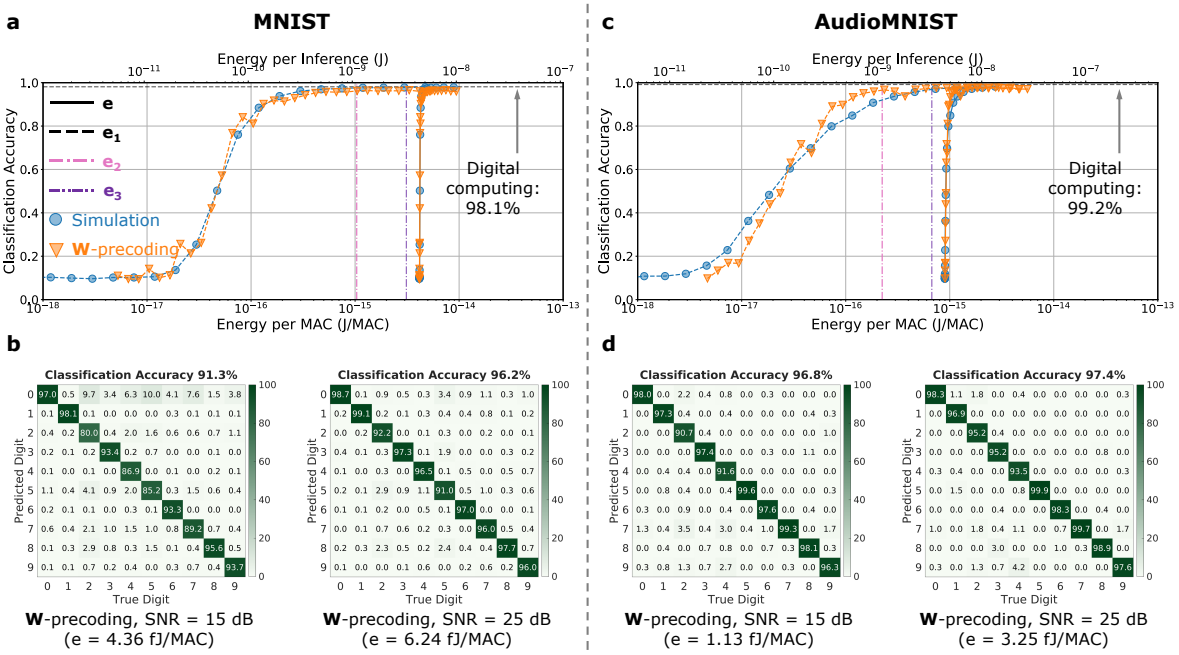


**Fig. S20 The DL-inference performance on the MNIST and AudioMNIST by the WISE's basic scheme over a 100 MHz wired channel. a**, The energy efficiency analysis on the MNIST dataset. **b**, The confusion matrices on the MNIST dataset of the wired channel. **c**–**d**, The energy efficiency analysis and the confusion matrices on the AudioMNIST dataset, respectively.

channel as a substitute for the wireless transmission of $w(t)$, where the central radio's TX port is connected to the computing mixer's LO port using an SMA cable, as shown in Fig. S5b. Such a wired setting ensures the flat channel response so that the precoding process can be skipped, while the time-encoded $\mathbf{x}$ to waive the encoding energy can still be applied. On the other hand, only a single client is supported at a time. In the wired experiment, we employ a bandwidth of as 100 MHz for $w(t)$ and $x(t)$, which is the maximum sampling rate at which the USRP X310 can maintain stable data streaming. According to equation (S123), the per-client computation throughtput is increased from 60 MOPS to 240 MOPS, due to the $4\times$ increase in the signal bandwidth.

Under the 100 MHz wired channel, we first showcase the performance of general MVM computation under the same settings as described in Supplementary Section 13. As shown in Fig. S19a, this wired channel enables slightly higher computing accuracy than the wireless WISE under higher SNRs. For example, under 30 dB SNR, the basic scheme without channel calibration achieves an RMSE of 0.045 and 0.038 with $N = 4,096$ and $32,768$. In addition, Fig. S19b shows the minimum energy efficiency required to achieve $\mathsf{RMSE} < 0.0625$, which is similar to the **W**-precoding scheme presented in Results section. For example, given $N = 4,096$ and $N = 32,768$, the energy efficiency required for WISE is 3.07 fJ/MAC and 1.07 fJ/MAC (325.73 TOPS/W and 934.58 TOPS/W), respectively.

On the MNIST dataset, Fig. S20a presents the energy efficiency achieved by WISE in the wired setup. Note that the flat channel response of the wired setup reduces the gap between the simulation and experimental results. To achieve 90% classification accuracy on the MNIST dataset, the energy efficiency of WISE is $e_{\text{mvm}} = 4.28$ fJ/MAC (233.64 TOPS/W), with a breakdown of $e_1 = 0.13$ fJ/MAC, $e_2 = 1.04$ fJ/MAC, and $e_3 = 3.11$ fJ/MAC. As shown in Fig. S20b, at SNR values of 15 dB and 25 dB, the experimental classification accuracy on MNIST is 91.3% and 96.2%, respectively, which closely matches the simulation results of 93.8% and 97.7%. As for the AudioMNIST dataset, Fig. S20c shows that to achieve 90% classification

accuracy, the energy efficiency of WISE is $e_{\mathrm{mvm}} = 1.01\,\mathrm{fJ/MAC}$ (985.61 TOPS/W), with a breakdown of $e_1 = 0.12\,\mathrm{fJ/MAC}$, $e_2 = 0.22\,\mathrm{fJ/MAC}$, and $e_3 = 0.67\,\mathrm{fJ/MAC}$. Moreover, in Fig. S20d, the experimental classification accuracies are 96.8% and 97.4% under 15 dB and 25 dB SNR, respectively, compared to simulation results of 90.8% and 98.2%. These results showcase the promising performance of WISE's basic scheme in the wired setup, demonstrating its scalability toward higher computation throughput.

# References


[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature **521**(7553), 436–444 (2015)

[2] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: A review. Neurocomputing **187**, 27–48 (2016)

[3] O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)

[4] S. Rokhsaritalemi, A. Sadeghi-Niaraki, S.M. Choi, A review on mixed reality: Current trends, challenges and prospects. Applied Sciences **10**(2), 636 (2020)

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., *Language models are few-shot learners*, in *Proc. Advances in Neural Information Processing Systems (NeurIPS)* (2020)

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[7] M. Horowitz, *Computing's energy problem (and what we can do about it)*, in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)* (2014)

[8] K. Sulimany, S.K. Vadlamani, R. Hamerly, P. Iyengar, D. Englund, Quantum-secure multiparty deep learning. arXiv preprint arXiv:2408.05629 (2024)

[9] R. Landauer, Irreversibility and heat generation in the computing process. IBM Journal of Research and Development **5**(3), 183–191 (1961)

[10] D.A. Miller, Attojoule optoelectronics for low-energy information processing and communications. IEEE Journal of Lightwave Technology **35**(3), 346–396 (2017)

[11] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, D. Englund, Large-scale optical neural networks based on photoelectric multiplication. Physical Review X **9**(2), 021032 (2019)

[12] Y. Shen, N.C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, et al., Deep learning with coherent nanophotonic circuits. Nature Photonics **11**(7), 441–446 (2017)

[13] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.C. Chen, P. Chen, G.B. Jo, J. Liu, S. Du, All-optical neural network with nonlinear activation functions. Optica **6**(9), 1132–1137 (2019)

[14] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T.G. Nguyen, S.T. Chu, B.E. Little, D.G. Hicks, R. Morandotti, et al., 11 TOPS photonic convolutional accelerator for optical neural networks. Nature **589**(7840), 44–51 (2021)

[15] H. Zhang, M. Gu, X. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M.H. Yung, et al., An optical neural chip for implementing complex-valued neural network. Nature Communications **12**(1), 457 (2021)

[16] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A.S. Raja, et al., Parallel convolutional processing using an integrated photonic tensor core. Nature **589**(7840), 52–58 (2021)

[17] L.G. Wright, T. Onodera, M.M. Stein, T. Wang, D.T. Schachter, Z. Hu, P.L. McMahon, Deep physical neural networks trained with backpropagation. Nature **601**(7894), 549–555 (2022)

[18] T. Wang, M.M. Sohoni, L.G. Wright, M.M. Stein, S.Y. Ma, T. Onodera, M.G. Anderson, P.L. McMahon, Image sensing with multilayer nonlinear optical neural networks. Nature Photonics **17**(5), 408–415 (2023)

[19] Y. Chen, M. Nazhamaiti, H. Xu, Y. Meng, T. Zhou, G. Li, J. Fan, Q. Wei, J. Wu, F. Qiao, et al., All-analog photoelectronic chip for high-speed vision tasks. Nature **623**(7985), 48–57 (2023)

[20] S.Y. Ma, T. Wang, J. Laydevant, L.G. Wright, P.L. McMahon, Quantum-limited stochastic optical neural networks operating at a few quanta per activation. Nature Communications **16**(1), 359 (2025)

[21] S. Agarwal, T.T. Quach, O. Parekh, A.H. Hsia, E.P. DeBenedictis, C.D. James, M.J. Marinella, J.B. Aimone, Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding. Frontiers in Neuroscience **9**, 484 (2016)

[22] M.J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S.J. Plimpton, E. Ipek, C.D. James, Multiscale co-design analysis of energy, latency, area, and accuracy of a reram analog neural training accelerator. IEEE Journal on Emerging and Selected Topics in Circuits and Systems **8**(1), 86–101 (2018)

[23] A. Ankit, I.E. Hajj, S.R. Chalamalasetti, G. Ndu, M. Foltin, R.S. Williams, P. Faraboschi, W.m.W. Hwu, J.P. Strachan, K. Roy, et al., *PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference*, in *Proc. ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* (2019)

[24] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, Memory devices and applications for in-memory computing. Nature Nanotechnology **15**(7), 529–544 (2020)

[25] S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin, K. Hosokawa, A. Nomura, T. Yasuda, A. Chen, A. Friz, et al., An analog-AI chip for energy-efficient speech recognition and transcription.

Nature **620**(7975), 768–775 (2023)

[26] S. Jung, H. Lee, S. Myung, H. Kim, S.K. Yoon, S.W. Kwon, Y. Ju, M. Kim, W. Yi, S. Han, et al., A crossbar array of magnetoresistive memory devices for in-memory computing. Nature **601**(7892), 211–216 (2022)

[27] C. Liu, Q. Ma, Z.J. Luo, Q.R. Hong, Q. Xiao, H.C. Zhang, L. Miao, W.M. Yu, Q. Cheng, L. Li, et al., A programmable diffractive deep neural network based on a digital-coding metasurface array. Nature Electronics **5**(2), 113–122 (2022)

[28] S.G. Sanchez, G. Reus-Muns, C. Bocanegra, Y. Li, U. Muncuk, Y. Naderi, Y. Wang, S. Ioannidis, K.R. Chowdhury, AirNN: Over-the-air computation for neural networks via reconfigurable intelligent surfaces. IEEE/ACM Transactions on Networking **31**(6), 2470–2482 (2022)

[29] G. Reus-Muns, K. Alemdar, S.G. Sanchez, D. Roy, K.R. Chowdhury, *AirFC: Designing fully connected layers for neural networks with wireless signals*, in *Proc. ACM International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)* (2023)

[30] J. Tong, Z. An, X. Zhao, S. Liao, L. Yang, *In-sensor machine learning: Radio frequency neural networks for wireless sensing*, in *Proc. ACM International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)* (2024)

[31] M. Cotrufo, S.B. Sulejman, L. Wesemann, M.A. Rahman, M. Bhaskaran, A. Roberts, A. Alù, Reconfigurable image processing metasurfaces with phase-change materials. Nature Communications **15**(1), 4483 (2024)

[32] A. Ross, N. Leroux, A. De Riz, D. Marković, D. Sanz-Hernández, J. Trastoy, P. Bortolotti, D. Querlioz, L. Martins, L. Benetti, et al., Multilayer spintronic neural networks with radiofrequency connections. Nature Nanotechnology **18**(11), 1273–1280 (2023)

[33] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[34] D. Chu, Polyphase codes with good periodic correlation properties. IEEE Transactions on Information Theory **18**(4), 531–532 (1972)

[35] A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong, J. Cochrane, L. Bernstein, D. Bunandar, P.B. Dixon, S.A. Hamilton, M. Streshinsky, et al., Delocalized photonic deep learning on the internet's edge. Science **378**(6617), 270–276 (2022)

[36] R. Davis III, Z. Chen, R. Hamerly, D. Englund, RF-photonic deep learning processor with shannon-limited data movement. arXiv preprint arXiv:2207.06883v2 (2024)

[37] J. Choi, Z. Wang, S. Venkataramani, P.I.J. Chuang, V. Srinivasan, K. Gopalakrishnan, PACT: Parameterized clipping activation for quantized neural networks. arXiv preprint arXiv:1805.06085 (2018)

[38] S. Garg, J. Lou, A. Jain, Z. Guo, B.J. Shastri, M. Nahmias, Dynamic precision analog computing for neural networks. IEEE Journal of Selected Topics in Quantum Electronics **29**(2: Optical Computing), 1–12 (2022)

[39] O. Abari, E. Hamed, H. Hassanieh, A. Agarwal, D. Katabi, A.P. Chandrakasan, V. Stojanovic, *A 0.75-million-point fourier-transform chip for frequency-sparse signals*, in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)* (2014)

[40] N.P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, et al., *In-datacenter performance analysis of a tensor processing unit*, in *Proc. IEEE/ACM International Symposium on Computer Architecture (ISCA)* (2017)

[41] S. Becker, J. Vielhaben, M. Ackermann, K.R. Müller, S. Lapuschkin, W. Samek, AudioMNIST: exploring explainable artificial intelligence for audio analysis on a simple benchmark. Journal of the Franklin Institute **361**(1), 418–428 (2024)

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[43] H. Zhang, A.T. Narayanan, H. Herdian, B. Liu, Y. Wang, A. Shirane, K. Okada, *0.2 mW 70 Fs RMS-jitter injection-locked PLL using de-sensitized SSPD-based injecting-time self-alignment achieving 270 dB FoM and −66 dBc reference spur*, in *Proc. Symposium on VLSI Technology and Circuits* (2019)

[44] H. Choi, S. Cho, *A 7.5 GHz subharmonic injection-locked clock multiplier with a 62.5 MHz reference, −259.7 dB FoMJ, and −56.6 dBc reference spur*, in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)* (2024)

[45] K. Rashed, A. Undavalli, S. Chakrabartty, A. Nagulu, A. Natarajan, A scalable and instantaneously wideband RF correlator based on margin computing. IEEE Journal of Solid-State Circuits **59**(11), 3612–3626 (2024)

[46] M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.A. Blanche, H. Rastegarfar, M. Glick, D. Kilper, *Projector: Agile reconfigurable data center interconnect*, in *Proc. ACM SIGCOMM Conference (SIGCOMM)* (2016)

[47] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, L. Zhong, *Argos: Practical many-antenna base stations*, in *Proc. ACM International Conference on Mobile Computing and Networking (MobiCom)* (2012)

[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[49] Mini-Circuits. Coaxial frequency mixer, 300–4300 MHz. https://www.minicircuits.com/pdfs/ZEM-4300+.pdf

[50] Radio regulations (edition of 2020). https://search.itu.int/history/HistoryDigitalCollectionDocLibrary/1.44.48.en.101.pdf (2020)

[51] S. Imai, H. Sato, K. Mukai, H. Okabe, *A load-variation-tolerant Doherty power amplifier with dual-adaptive-bias scheme for 5G handsets*, in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)* (2024)

[52] B. Murmann. ADC performance survey (1997-2024). [Online]. Available: https://github.com/bmurmann/ADC-survey

[53] A.V. Oppenheim, *Discrete-time signal processing* (Pearson Education India, 1999)

[54] H.T. Friis, A note on a simple transmission formula. Proceedings of the IRE **34**(5), 254–256 (1946)

[55] B. Bloessl, M. Segata, C. Sommer, F. Dressler, *An IEEE 802.11a/g/p OFDM receiver for GNU Radio*, in *Proc. 2nd Workshop on Software Radio Implementation Forum (SRIF)* (2013)

[56] Z. Gao, Y. Chen, T. Chen, *Swirls: Sniffing Wi-Fi using radios with low sampling rates*, in *Proc. ACM International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)* (2023)

[57] D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to linear regression analysis* (John Wiley & Sons, 2021)