Orientation Matters: An Empirical Study on Ranging Performance of Bluetooth 6.0 Channel Sounding

Dzenita Dzafic, Jesús Pestana, Markus Schuß, Carlo Alberto Boano, and Kay Römer
Institute of Technical Informatics
Graz University of Technology
Graz, Austria
{dzenita.dzafic; pestana; markus.schuss; cboano; roemer}@tugraz.at

Abstract

Bluetooth 6.0 introduced Channel Sounding (CS), an enhancement to Bluetooth Low Energy (BLE) that enables precise, low-power distance estimation with promised sub-meter ranging accuracy, enabling localization for battery-constrained devices. In this paper, we experimentally evaluate the ranging performance of off-the-shelf Bluetooth 6.0 development kits from two vendors. We collect over 190 thousand ranging samples across three indoor/outdoor environments, single- and dual-antenna ranging modes, relative orientation pairs between initiator and reflector, and full 360° in-plane rotations of the reflector. We show that the often-overlooked relative orientation between devices strongly affects accuracy: certain configurations introduce systematic biases, increasing mean absolute error by up to 88 cm. We find that bias correction can be substantially improved by accounting for relative orientation. When the relative 3D bearing and orientation are approximately known, configuration-specific calibration yields markedly different correction offsets—varying by up to 138 cm on the same platform. When orientation is unknown, deriving a single global offset across all orientations improves ranging accuracy by up to 12% compared to median-based calibration performed at a fixed relative-orientation pair. We further observe a large fraction of outliers, with up to 18% of ranging estimates exceeding 4.31 m error. We therefore investigate online outlier detection using radio-provided indicators such as the tone quality indicator (TQI) on Nordic's nRF54, and find that it may only flag 5.18–33.38% of the outliers, substantially limiting its usefulness for outlier filtering. Finally, we show that dual-antenna configurations improve ranging accuracy by 47.5% and reduce outlier rates from 10.81–17.98% to 1.08–3.71% compared to single-antenna setups. Overall, our study provides practical guidelines for calibration, antenna selection, and online quality assessment to improve real-world Bluetooth ranging systems.

CCS Concepts

• **Computer systems organization** → **Embedded systems**; **Sensor networks**; • **Hardware** → *Sensor devices and platforms*; *Wireless devices*; • **Information systems** → **Location based services**; **Sensor networks**.

Keywords

Bluetooth Low Energy, Calibration, Channel sounding, Device orientation, Dual antenna, Experimental characterization, Indoor, Nordic, Outdoor, Outlier detection, Outlier labeling, Ranging, SiLabs, Tone quality indicator.

1 Introduction

Accurate distance estimation is a key building block for ubiquitous wireless applications, including indoor localization, proximity detection, asset tracking, and spatial awareness. These applications often demand solutions that combine high accuracy with low energy consumption, low cost, and compatibility with widely-deployed hardware platforms [1, 6, 10]. Bluetooth Low Energy (BLE) is one of the most ubiquitous wireless technologies satisfying these requirements. With the release of the Bluetooth Core Specification 6.0 (BLE 6.0) [2], the Bluetooth SIG introduced *Channel Sounding (CS)*, also referred to as BLE Channel Sounding (BLE-CS), as a standardized ranging capability for BLE devices.

Beyond RSSI: BLE-CS for accurate ranging. BLE-CS combines two ranging methods, *phase-based ranging (PBR)* [15, 30] and *round-trip time (RTT)* [7]. Although PBR was an established method, its implementation remained limited to certain vendors prior to BLE 6.0. PBR estimates distance by actively probing the radio channel between two devices: an *initiator*, which starts the procedure, and a *reflector*, which responds. During a ranging exchange, the devices transmit unmodulated tones on multiple frequencies across the 2.4 GHz ISM band to estimate the channel frequency response. IQ samples collected by the initiator and the reflector encode each received tone's in-phase (I) and quadrature (Q) components. Combining the IQ samples of initiator and reflector allows to cancel out phase shifts. By observing how the phase-shift-corrected IQ samples evolve with frequency, the initiator can estimate the signal propagation distance [2] (see Section 3). Compared to traditional Received Signal Strength Indicator (RSSI)-based ranging, CS offers substantially higher accuracy and improved robustness [5, 9, 25]. Current off-the-shelf BLE-CS development kits (DKs) typically incorporate the following techniques for range estimation: RTT [7], or PBR methods such as phase-slope (PS) [30], inverse Fourier transform (IFFT) [15], and multiple signal classification (MUSIC) [3, 14].

Experimental studies on BLE-CS performance. Following standardization, several works have begun to experimentally evaluate BLE-CS on commercial development kits and hardware platforms. Nikodem *et al.* [9] investigated RSSI-, PS-, and IFFT-based ranging using Nordic Semiconductor's nRF52840 (a pre-BLE 6.0 DK) in a multi-anchor office deployment. Ulsamer *et al.* [22] analyzed PS- and IFFT-based ranging with Nordic Semiconductor's nRF54L15 DK outdoor and in a utility-basement environment characterized by strong multipath and non-line-of-sight (NLoS) conditions. Santra *et al.* [13] and Tsemko *et al.* [21] reported IFFT-based ranging results using Infineon's pre-commercial BLE-CS hardware. Wieme *et al.* [26] presented a more comprehensive benchmark of IFFT- and

MUSIC-based ranging with the nRF54L15 DK, including indoor and outdoor settings, and studied effects on ranging of transmit power, PHY configuration, channel selection, frequency-sampling density, and antenna orientation.

Limitations of existing evaluations. Existing experimental studies on BLE-CS remain constrained in several key dimensions. First, most evaluations focus on *one* single-antenna hardware platform [9, 21, 22, 26]. Measurements are also often conducted with *fixed* board orientations [9, 21, 22], implicitly assuming orientation-invariant behavior. Wieme *et al.* [26] analyzed the effect of board orientation, but did not study its effect on the calibration offset and their analysis was limited for mixed vertical and horizontal orientations. Santra *et al.* [13] and Tsemko *et al.* [21] used a dataset including 24 indoor and outdoor environments, but these environments and their effects on calibration and ranging accuracy were not described in any detail. Moreover, reported results typically emphasize aggregate accuracy metrics [9, 21, 22, 26], with limited analysis of systematic effects such as orientation-dependent bias, calibration transferability between environments, and outlier occurrence rates, labeling, and detection. Consequently, the gaps identified above remain insufficiently addressed: (i) benefits of antenna diversity, (ii) environment and fixed-orientation effects on calibration and ranging performance, (iii) calibration transferability, as well as (iv) outlier occurrence rates, labeling, and detection.

Contributions. To address these gaps, this paper presents a systematic evaluation of BLE-CS performance on commercially available BLE 6.0 development kits.

By collecting line-of-sight measurements across three environments, five relative board orientations and eight in-plane reflector rotations, we evaluate how relative board orientations and antenna diversity affect calibration offsets and ranging accuracy. Using their vendor-specific stacks, we benchmark two off-the-shelf DKs supporting BLE-CS according to the Bluetooth 6.0 specification: the Nordic Semiconductor's nRF54L15 DK (`Nordic`), using single-antenna IFFT-based ranging, and the Silicon Labs xG24-DK2606A (`SiLabs`), using its closed-binary single-antenna (`SiLabs 1×1`) and dual-antenna (`SiLabs 2×2`) ranging modes, which likely employ a MUSIC-based algorithm. For `SiLabs 2×2`, we rely on the vendor-supplied calibration rather than performing a separate calibration procedure. Specifically, we make the following contributions:

- We systematically characterize BLE-CS ranging using orientation-diverse datasets, and demonstrate that calibration offsets are substantially more sensitive to orientation diversity than to environment diversity. Specifically, calibration offsets vary across board orientation pairs by up to 70 cm for `Nordic` and 138 cm for `SiLabs 1×1`, whereas environment diversity induces offset variations of only 32 cm and 45 cm, respectively, representing a 2–3× difference in magnitude.
- By comparing results of `SiLabs 1×1` to those of `SiLabs 2×2`, we confirm the superior performance of dual-antenna over single-antenna BLE-CS.
- Dual-antenna BLE-CS achieves substantial gains across three metrics: relative-orientation induced error (maximum biases reduced from 138 cm to 58 cm, 58% improvement), ranging accuracy (`RMSE` improved by 47.5%) and outlier occurrence rates (reduced from 10.81–17.98% to 1.08–3.71%).
- We show that automated outlier filtering based on median absolute deviation only mislabels around ±3.5% of the data compared to manual filtering, and thus propose its use to clean datasets before calibration.
- We explore a built-in confidence metric (`Nordic`'s tone quality indicator) and demonstrate that it may only flag 5.18–33.38% of the outliers, underscoring the necessity for enhanced online outlier detection strategies.
- We release our orientation-diverse BLE-CS ranging datasets and open-source our code[1] to support future research.

Paper outline. After reviewing related work in Sec. 2, we present the ranging principles of BLE-CS in Sec. 3. Our experimental setup and methodology are described in Sec. 4. The evaluation begins with ranging performance without calibration (Sec. 5), which motivates our calibration procedure. We perform outlier filtering (Sec. 6) before calibration, and benchmark two calibration models and present the calibration results in Sec. 7. We then report ranging performance with calibration in Sec. 8 and leverage our outlier labeling to analyze online outlier detection using a built-in confidence metric in Sec. 9. We conclude with a discussion of the limitations of this study and future work in Sec. 10, followed by conclusions in Sec. 11.

2 Related Work

This section reviews BLE-based ranging from RSSI to BLE 6.0 CS, surveys multi-antenna PBR and calibration procedures, and identifies research gaps in orientation-diversity effects on calibration and ranging performance, and evaluation of vendor confidence metrics.

BLE ranging methods. Bluetooth-based ranging has been studied extensively over the past two decades, resulting in a diverse set of approaches with differing assumptions on hardware capabilities and deployment environments [6]. Early BLE ranging approaches predominantly relied on RSSI measurements due to their simplicity and universal availability. However, RSSI-based ranging is highly sensitive to multipath propagation, antenna or device orientation [19], and environmental dynamics, leading to poor accuracy and limited robustness in practical deployments, with an accuracy of 1-4 m [1, 9] and 95$^{\text{th}}$ percentile errors of 7-9 m [9]. The Bluetooth 5.1 core specification (BLE 5.1), introduced angle-of-arrival (AoA) and angle-of-departure (AoD) measurements. AoA and AoD techniques estimate the spatial direction of a BLE signal by exploiting phase differences across a multi-antenna array [27], but require costly and bulky specialized multi-antenna hardware, careful antenna array calibration, and increased signal-processing complexity. Recent work shows that using BLE-based AoA/AoD measurements in single-anchor systems [24, 28] and in multi-anchor or multi-receiver deployments [8, 9] can achieve sub-meter accurate localization, for instance, with median errors of 0.3-0.6 m [8] or 1.05 m [9] and a mean absolute error (`MAE`) of 1.26 m [9]. Round-Trip Time (RTT) based ranging uses time-of-flight (ToF) measurements and was standardized with the BLE 6.0 specification. In BLE-CS, RTT-based ranging is used to support PBR, specifically to resolve distance ambiguities, to extend the measurable range and to enhance security by preventing relay attacks [2, 11]. However, BLE RTT-only ranging achieves meter-scale accuracy (≈ 1–5 m) [7, 11] with BLE 6.0

[1] Datasets and code: http://iti.tugraz.at/BLE-CS-OM

RTT performing better. Prior to BLE 6.0, the link layer and timing primitives were optimized for data transmission rather than for ToF measurements. In contrast, BLE 6.0 [2] standardized RTT as a supporting ranging method, alongside PBR, that relies on the compensation of clock drift and frequency generation inaccuracies via Mode-0 calibration, more precise time-of-departure and -arrival timestamps, and more accurately timed reflector turnaround delay.

BLE-CS on commercial hardware. BLE-CS extends prior BLE ranging techniques by combining PBR and RTT, and by explicitly probing the radio channel across multiple frequencies [2]. Commonly-implemented PBR techniques include phase-slope (PS) [30], inverse Fourier transform (IFFT) [15], and MUSIC [14], a high-resolution spectral method. Multiple antenna paths can be exploited for ranging in BLE-CS, e.g., using MUSIC, which improves ranging performance [2, 3] and reduces orientation-dependent bias [18]. Following its introduction in Bluetooth 6.0, several studies have begun to experimentally evaluate BLE-CS performance on commercial development kits [9, 13, 21, 22, 26], and some of them have reported sub-meter ranging accuracy for outdoor LoS conditions [21, 22].

Multi-antenna Phase-Based Ranging (PBR). Single-antenna PBR is vulnerable to direction-dependent radiation nulls, elevated noise sensitivity, and multipath-induced error, which motivates the exploitation of antenna diversity. Several pre-BLE 6.0 studies have explored PBR using multi-antenna setups. Boer *et al.* [3] recover the one-way channel response (OWCR) instead of the conventional two-way response (TWCR) and feed it to MUSIC, to enable multi-antenna ranging and increased subspace resolution, thereby improving performance over IFFT. Van Marter *et al.* [23] introduce a support-vector-regression (SVR) model that achieves comparable performance to MUSIC for multi-antenna ranging, surpasses it for single-antenna ranging, and reduces computational complexity. Tariq *et al.* [20] propose a data-driven signal-subspace decomposition (SSD) whose eigen-vector spectra are processed by a long short-term memory (LSTM) model, yielding superior robustness to noise and lower computational complexity when compared to MUSIC and SVR [23]. Across these studies, multi-antenna ranging algorithms improve ranging, and mitigate outliers and low signal to noise ratios (SNRs) by leveraging antenna-path diversity to isolate the line-of-sight component. Silicon Labs' AN1493 guideline [18] proposes that, for compact multi-antenna board design, antennas should be spaced at least a quarter-wavelength (≈ 3 cm at 2.4 GHz) and no more than about a half-wavelength (≈ 6 cm) apart. `SiLabs` provides a dual-antenna reference design, and their benchmark [17] on this DK aligns with the performance gains reported in the above-mentioned works for multi-antenna versus single-antenna PBR.

Calibration in BLE-CS. Ranging calibration models for PBR or BLE-CS often implement a single global distance offset [15, 22]. However, Nikodem *et al.* use different calibration offsets for each anchor of a multi-anchor system [9]. Some studies do not even specify their calibration approach, e.g., [26]. In certain cases, algorithms implicitly adapt to per-environment varying calibration offsets, achieving good performance across many environments, through the training of machine-learning high-resolution IFFT [21] or scene-identification [13] models using environment-diverse datasets. In addition, there has only been limited discussion on alternative calibration models, for instance, Nikodem *et al.* [9] discarded fitting a slope parameter in the linear models (w.r.t. the ranging distance) for PS- and IFFT-based ranging, and Santra *et al.* [13] apply correction factors to each IQ-sample depending on its BLE channel. Some studies pre-BLE 6.0 implemented 0.5 MHz frequency sampling steps [15, 16] rather than the standard maximum resolution of 1 MHz of BLE 6.0 [2]. Schröder *et al.* [15] indicate based on their experimental data that the calibration offset may depend upon the implemented ranging algorithm and the deployment environment, and that the origin of such an offset should be further investigated. In three pre-BLE 6.0 works [4, 29, 30], the ranging calibration model consists of: (i) a fixed offset that corrects front-end circuit and antenna delays, as is standard practice in the literature [4, 9, 15, 18, 22], and (ii) a variable offset caused by the crystal-offset difference (COD) between initiator and reflector, which can be pre-calibrated or estimated dynamically during operation [4]. The COD-induced ranging error is linear with respect to the elapsed times between per-BLE-channel exchanged tones, and between probed BLE channels. Compensating the COD-induced error permits to maintain PBR accuracy, while allowing longer elapsed times between tones and fewer exchanged tones through one-to-many or many-to-many ranging approaches [4, 29].

Calibration procedures in BLE-CS. From the works that describe the calibration procedure in some detail, we can extract three calibration methods. Of these, only the first two calculate distance offsets and are therefore relevant to our work:

- *Method 1*, outlier removal and mean error or linear regression calculation [4, 9, 22]. The raw ranging data are first screened to discard obvious outliers (with the outlier identification method unspecified). Then, either the mean error or a linear regression (with the slope forced to 1) is performed, yielding the calibration offset.
- *Method 2*, median-of-medians and linear regression. The Silicon Labs antenna design guidelines [18] recommend acquiring data at five or more different distances and, for each calibration distance, rotating one of the devices through several in-plane orientations (typically eight, in 45° steps). The median of the measurements at each in-plane rotation is taken, then the median of those eight medians is calculated, which filters outliers, producing one clean value per distance. Those per-distance values are fitted with a linear regression (again with the slope forced to 1) to obtain the calibration offset. Their guidelines [18] also emphasize that antenna orientation, and antenna-path and polarization diversity can have a significant impact on channel sounding accuracy and performance.
- *Method 3*, per-channel phase-correction terms (PCTs) [13]. These terms, one per BLE channel, are applied to corresponding IQ samples to compensate channel-dependent antenna and analog front-end phase and magnitude errors. Santra *et al.* [13] estimate these PCTs through zero-distance calibration.

Open challenges. The limitations of existing evaluations (Sec. 1) and the related work highlight the need for empirical studies on BLE 6.0 compliant platforms to assess the effects of relative board orientations and antenna diversity on calibration and ranging performance, which requires orientation-diverse datasets. Since

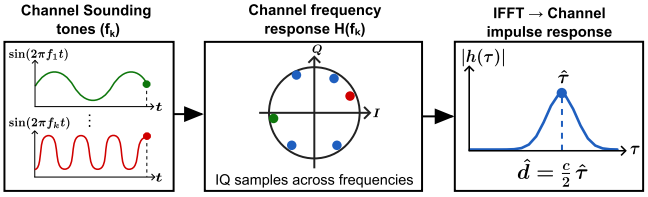


Figure 1: Principle of PBR in BLE-CS: multi-tone channel probing, IFFT to CIR, and peak-delay to distance steps.

orientation-diverse datasets have not yet been considered in relation to calibration, it is relevant to compare both fixed-offset and linear models (w.r.t. ranging distance) as calibration models for standard one-to-one BLE-CS. While vendors provide proprietary confidence metrics for ranging reliability (such as `Nordic`'s tone quality indicators and `SiLabs`' likeliness score), to our knowledge, independent performance evaluations of these metrics for outlier detection are not yet available in the peer-reviewed literature.

3 Ranging Principles with BLE-CS

Instead of inferring distance from received signal strength, BLE-CS estimates distance by probing the wireless channel at multiple discrete frequencies and analyzing the resulting phase response. During a CS procedure, unmodulated tones are exchanged between an initiator and a reflector across a predefined set of channel-sounding frequencies spanning the 2.4 GHz ISM band, with CS synchronization achieved via initial Mode-0 CS_SYNC steps that establish precise timing and agreement on the tone (channel) sequence, and estimate frequency offsets for subsequent compensation [2].

An example processing BLE-CS pipeline implementing IFFT-based ranging is illustrated in Figure 1. As the signal propagates through the channel, each tone experiences a frequency-dependent phase shift. Phase-based ranging (PBR) exploits this effect by relating the phase evolution across frequency to the signal propagation delay. Under line-of-sight conditions, the channel frequency response, $H(f)$, can be approximated as

$$H(f) \approx Ae^{-j2\pi f\tau}, \tag{1}$$

where A denotes the (generally complex) gain, f the frequency, and τ the propagation delay. The associated phase response is

$$\phi(f) = \arg\left(H(f)\right) \approx \phi_0 - 2\pi f\tau, \tag{2}$$

which varies linearly with frequency. The slope of the phase response is $-2\pi\tau$, and reflects the propagation delay.

In practice, CS measurements are available at a finite set of discrete frequencies defined by the CS channel map (Figure 1, left). The resulting complex-valued measurements $H(f_k)$ form a sampled estimate of $H(f)$, the *channel frequency response (CFR)*, shown in Figure 1 (middle). Among other methods, distance estimation can be performed either by estimating the phase slope directly from the CFR [30] or by transforming the frequency-domain measurements into the time domain using an IFFT [13]:

$$h(\tau) = \left|\sum_{k=1}^{N} H(f_k)e^{j2\pi f_k\tau}\right|, \tag{3}$$

where f_k denotes the set of probed CS frequencies. This transformation yields an estimate of the *channel impulse response (CIR)*, which exhibits peaks corresponding to dominant propagation paths (Figure 1, right). In ideal conditions, the location of the strongest peak indicates the estimated signal delay and, consequently, distance. The distance estimate is obtained as $\hat{d} = \frac{c}{2}\hat{\tau}$, where $\hat{\tau}$ denotes the round-trip-time delay associated with the strongest peak and c is the speed of light.

The CS *channel map* determines which frequencies are probed within a CS procedure, while the *channel map repetition* parameter controls whether the same frequency set is reused multiple times within the CS procedure. CS defines 80 channels with 1 MHz spacing; channels 0, 1, 23, 24, 25, 77, and 78 are excluded by specification, while channel 79 is reserved for future use and also disabled [2].

Physical orientation degrades BLE PBR and CS performance by pushing the wireless link into radiation-pattern nulls or cross-polarized conditions, which reduce signal-to-noise ratio (SNR) and induce orientation-dependent group-delay variations across the 2.4 GHz BLE band [18], affecting phase measurements. While constant group delay can be calibrated out through a distance offset, variations in group delay can introduce ranging bias, outliers and reduced precision in PBR estimates. In challenging orientations, RTT is also affected, primarily through reduced SNR and link stability rather than phase-based bias. Silicon Labs therefore recommends in its guidelines [18] antenna designs with flat group delay, avoidance of radiation pattern nulls, and antenna/polarization diversity for stable PBR and CS performance across arbitrary orientations. These represent two possible orientation-induced effects on PBR estimates; further examination of the underlying physical causes of increased PBR error is beyond the scope of this paper.

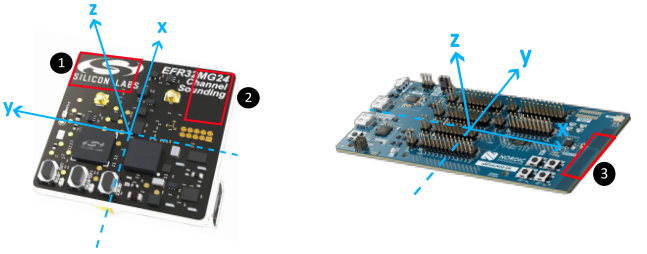
4 Experimental Setup and Methodology

We experimentally evaluate BLE-CS on two commercial off-the-shelf platforms to quantify the impact of relative device orientation and antenna diversity on calibration and ranging performance under controlled conditions. Our experiments are conducted using vendor-provided software stacks.

Hardware platforms. We evaluate BLE-CS on two commercial development kits (DKs) from different vendors: the Silicon Labs xG24-DK2606A (BRD2606A) and the Nordic nRF54L15 DK, hereafter referred to as `SiLabs` and `Nordic`, respectively.

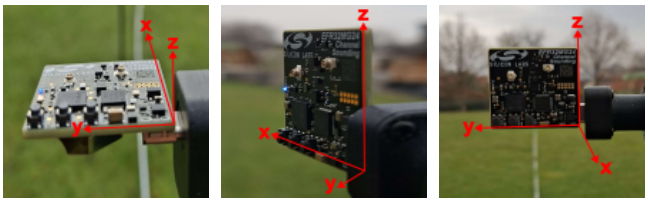
`SiLabs` is evaluated using the vendor-provided Simplicity SDK (v2025.6.2). The board integrates two inverted-F antennas, denoted ANT1 ❶ and ANT2 ❷ (Figure 2a), which can be individually selected to configure 1, 2, or 4 antenna paths for ranging. Depending on the configuration, the radio returns a distance estimate, IQ samples, RSSI, and a likelihood metric, with all internal processing handled by the closed-source vendor stack. Unless otherwise stated, reported ranging results correspond to single-shot measurements obtained per CS procedure.

`Nordic` is evaluated using the nRF Connect SDK (v3.0.2), which is based on the Zephyr real-time operating system. The board features a single PCB monopole antenna ❸ (Figure 2b) and provides access to IQ samples and multiple ranging outputs, including PS-, IFFT-, and RTT-based distance estimates, as well as per-tone quality indicators for individual IQ samples. While the Zephyr-based stack exposes low-level ranging data, the underlying controller implementation remains vendor-specific.



(a) SiLabs (b) Nordic

Figure 2: Local reference frames for the boards.



(a) XY (b) XZ (c) YZ

Figure 3: Fixed coordinate axes that define the three board orientations for the relative orientation pairs, with the x-axis pointing towards the second device.

(a) Grass

(b) Asphalt (c) Indoor

Figure 4: Measurement environments used in this study.

Measurement environment. Experiments are conducted in two outdoor environments and one indoor environment under clear line-of-sight (LoS) conditions, as illustrated in Figure 4. The devices are mounted on non-conductive tripods to avoid coupling effects with conductive supports. All measurements are performed on flat terrain, with devices placed at a height of 1.7 m above ground level. Ground-truth distances are measured using a surveying tape to ensure centimeter-level accuracy.

Outdoor measurements are performed in two settings. The first is a flat grass terrain, where the measurement area is largely free of nearby reflective vertical structures, limiting uncontrolled multipath and isolating the effects of antenna orientation. In this setting (`Grass`, Figure 4a), we evaluate five ground-truth distances spanning 3 m to 20 m. The second outdoor setting is an asphalt terrain with nearby buildings and parked cars, representing a more reflective outdoor scenario. In this setting (`Asphalt`, Figure 4b), we evaluate six ground-truth distances spanning 2 m to 37 m.

Indoor experiments are carried out at our university in a hallway (approx. 3×20 m), featuring reinforced-concrete walls and metallic doors, resulting in a reflective propagation scenario with more multipath interference. In this setting (`Indoor`, Figure 4c), we evaluate five ground-truth distances between 2 m and 10 m.

Relative orientation configuration. For each measurement distance, the *initiator* device is kept fixed, while the *reflector* device is repositioned to the desired distance and rotated using in-plane rotations to acquire orientation-diverse datasets. Three orthogonal orientations of the device coordinate system are considered, denoted as XY, XZ, and YZ, corresponding to different board orientation planes, as illustrated in Figure 3. For each distance and configuration, the reflector performs a full 360° in-plane rotation in 45° increments (which are positioned manually without any aid of additional equipment), resulting in eight discrete orientation angles per position. Based on these orientations, five relative-orientation pairs are evaluated: XY–XY, XY–XZ, XY–YZ, YZ–XZ, and YZ–YZ, where the first term denotes the orientation of the *reflector* and the second term denotes the orientation of the *initiator*. An XZ–XZ relative orientation pair is not considered, because it is similar to the XY–XY pair since both boards are co-planar. That is, the difference between XY–XY and XZ–XZ lies in the relative position of the ground and other surrounding objects. In general, considering the possibilities for co-polarization and cross-polarization, the expectation is for the XY–XY and YZ–YZ pairs to perform best, and for the rest to perform worse.

Measurement procedure and parameters. At each distance, and relative board orientation, 100 CS procedures are executed.

On `Nordic`, CS is operated in a PBR configuration with the initiator role fixed throughout the experiments. Measurements are performed at a transmit power of 0 dBm. The BLE connection operates on the 1M PHY, which also serves as the physical layer for CS synchronization. A full CS channel map is enabled, and a channel map repetition of one is selected.

For `SiLabs`, PBR is selected as the primary operating mode. Measurements are performed at a transmit power of 10 dBm. The BLE connection operates on the 2M PHY, while CS synchronization is performed on the 1M PHY. A full CS channel map, same as on `Nordic`, is employed, enabling all CS-specification-compliant frequencies with 1 MHz spacing, while disallowed and reserved channels are disabled. Each enabled CS frequency is sounded once per CS procedure.

For `Nordic`, ranging estimates are obtained using the vendor-provided IFFT-based PBR. For `SiLabs`, the vendor does not disclose the ranging algorithm. The CS sample applications use the vendor's closed-source RTL (Real-Time Locating) library for distance estimation. An inspection of its binaries and antenna guidelines [18] suggests that it likely implements a MUSIC-based PBR approach.

For `Nordic` and `SiLabs` 1×1, no calibration offset correction is applied during data acquisition. For `SiLabs` 2×2, we rely on the vendor-supplied calibration during data acquisition and in the

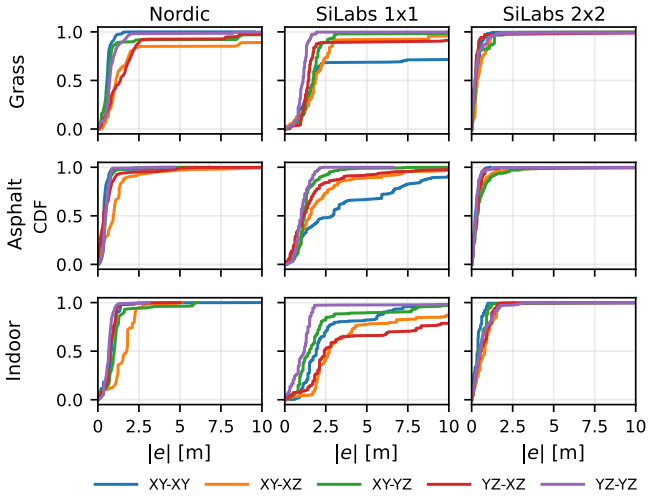


Figure 5: Cumulative distribution functions (CDFs) of the absolute ranging error $|e|$, where each curve corresponds to one relative orientation pair.

analysis of ranging performance, as the effect of the offsets for multi-antenna ranging on the closed-source ranging algorithm is unknown. Therefore, for multi-antenna ranging, the results provided by the implemented method (likely MUSIC) cannot be adjusted in post-processing. Specifically, the software stack provides default calibration offsets of 56 cm (ANT1) and 67 cm (ANT2). We do not perform a calibration procedure for SiLabs 2×2.
Unless otherwise stated, all remaining parameters are kept at vendor-default values. We did not perform any platform-specific tuning.

5 Characterizing BLE-CS Performance

As a first step in our study, we evaluate BLE-CS performance without calibration across the three environments described in Sec. 4.

Metrics. We characterize ranging performance using the absolute error magnitude $|e|$. We characterize the distribution of the absolute errors $|e|$ via its *cumulative distribution function* (CDF). Additionally, we report the percentiles P50, P75, P90, and P95, indicating the error thresholds below which 50%, 75%, 90%, and 95% of measurements fall, respectively. Accuracy is reported as the mean absolute error (MAE) and the root-mean-square error (RMSE), where RMSE emphasizes large deviations and therefore reflects heavy-tail behavior visible in the CDFs. Precision is captured by the inter-quartile range (IQR), i.e., the spread of the central 50% of the $|e|$ distribution.

Performance analysis without calibration. The CDFs of the absolute ranging errors $|e|$ without calibration for Nordic and SiLabs 1×1 and with calibration for SiLabs 2×2 are shown in Figure 5. They reveal clear ranging performance differences between devices and a pronounced dependence on the environment and the relative orientation pairs. The CDFs show that Nordic achieves generally lower errors and tighter dispersion than SiLabs 1×1 across all environments. In the same manner, SiLabs 2×2 outperforms both Nordic and SiLabs 1×1. Specifically, compared to SiLabs 1×1, these results show that dual-antenna BLE-CS significantly improves ranging precision in the tested environments. This is evidenced by the tighter dispersion of the MAEs of SiLabs 2×2 across the CDFs of the relative device orientations.

Table 1: Performance for best vs. worst relative-orientation pairs (per environment and device), in meters [m].

Env.	Device	Best orient. pair MAE	Best orient. pair RMSE	Best orient. pair IQR	Worst orient. pair MAE	Worst orient. pair RMSE	Worst orient. pair IQR
Grass	Nordic	0.94	1.95	0.30	2.75	5.59	1.15
	SiLabs 1×1	1.40	1.71	0.72	4.32	7.99	1.01
	SiLabs 2×2	0.48	1.61	0.35	–	–	–
Asphalt	Nordic	0.53	0.91	0.30	1.24	2.28	0.88
	SiLabs 1×1	1.22	1.42	0.55	2.83	4.61	2.30
	SiLabs 2×2	0.53	1.58	0.35	–	–	–
Indoor	Nordic	0.85	1.09	0.39	1.59	1.76	0.90
	SiLabs 1×1	1.46	3.53	0.75	3.90	6.31	2.40
	SiLabs 2×2	0.64	1.19	0.70	–	–	–

Table 2: Best and worst relative-orientation-pair groups per environment and device.

Dev.	Env.	Best orient. pairs	Worst orient. pairs
Nordic	Grass	XY-XY, YZ-YZ	XY-XZ, YZ-XZ, XY-YZ
	Asphalt	XY-XY, YZ-YZ, XY-YZ, YZ-XZ,	XY-XZ
	Indoor	YZ-YZ, XY-XY, YZ-XZ, XY-YZ	XY-XZ
SiLabs 1×1	Grass	YZ-YZ, XY-YZ	XY-XY, XY-XZ, YZ-XZ
	Asphalt	YZ-YZ, XY-YZ	XY-XY, XY-XZ, YZ-XZ
	Indoor	YZ-YZ	YZ-XZ, XY-XZ, XY-XY, XY-YZ
SiLabs 2×2	Grass	All pairs	–
	Asphalt	All pairs	–
	Indoor	All pairs	–

While indoor environments are often expected to be most challenging due to multipath and RF interference, Indoor does not exhibit the heaviest tails in the CDFs (Figure 5). The performance of SiLabs 2×2 in Indoor is only slightly degraded compared to its performance in both outdoor environments. For Nordic, Grass contains the biggest proportion of outliers, i.e., measurements with unusually high ranging error, for instance, of more than 5 m. For SiLabs 1×1, there are high proportions of outliers in all environments, but with a dependence on the orientation pair. Specifically, the YZ–YZ pair performs best in all environments, followed by the XY—YZ pair. On the other hand, when ranging on dual-antenna mode, SiLabs 2×2 has the least proportion of outliers.
In Table 1 we provide quantitative performance metrics separated by groups of best and worst orientation pairs (orient. pairs) as shown in Table 2, grouped manually for each environment by visual inspection of the CDFs in Figure 5. For Nordic, the consistently worst orientation pair is XY–XZ, whereas YZ–XZ and XY–YZ (long-tailed CDFs) only perform bad in Grass. For SiLabs 1×1, the best orientation pair is YZ–YZ, and XY–XY is surprisingly overall the worst orientation pair; YZ–XZ, and XY–XZ are also in the worst orientation pairs group for all environments (due to their long-tailed CDFs); for the same reason, XY–YZ is also on this group for Indoor. For SiLabs 2×2, all orientations pairs perform reasonably well and are always grouped together in the best orientation pairs group. The rest of the orientations pairs per device and environment are grouped together in the best orientation pairs group.
The most striking information in Table 1 is the good performance of SiLabs 2×2 across all orientation pairs. In contrast, for Nordic and SiLabs 1×1 the RMSE can increase by several meters when

comparing values, per device, for the best and worst orientations pairs. Lastly, ranging precision (measured by IQR) is 0.30–0.75 m for the best orientation pairs across all devices and environments.

Motivation for calibration. As shown in Figure 5, the raw (uncalibrated) estimates of Nordic and SiLabs 1×1 contain both systematic bias and non-negligible fractions of large-error samples that inflate the RMSE and result in heavy CDF tails. Therefore, the next section introduces an outlier labeling procedure to perform outlier filtering before calibration.

6 Outlier Labeling

Outliers with very large ranging errors can arise when the Phase Measurement Units (PMUs) on the initiator and reflector lock onto a reflected (multipath) signal instead of the direct LoS component, causing the measured phase angles to correspond to longer propagation paths [16]. These wrong triggerings of the PMUs lead the ranging algorithm to select an incorrect peak in the IFFT spectrum [15], producing largely incorrect distance estimates [16].

Outlier labeling. Outliers can bias calibration offset calculations: we therefore first focus on labeling outliers to enable calibration using only inliers. We are interested in automatic outlier labeling, so that the overall calibration process becomes more reproducible and does not depend on human labeling.

The outdoor (Grass/Asphalt) and Indoor histograms of the absolute ranging error $|e|$ are shown in Figure 6. Visual inspection reveals a multimodal distribution: a primary approximately Gaussian-distributed error mode, representing the majority of measurements, and a secondary outlier population. The green vertical dashed lines are manually selected cutoffs, τ_{man}, chosen to separate these two components. The red vertical dashed lines are automatically generated outlier thresholds, which are explained in the following.

To automatically separate the inliers from the long outlier tails (up to ≈15–30 m), we first form a robust inlier set using the median absolute deviation (MAD), a robust scale estimator. Assuming a Gaussian distribution for the inliers, one can estimate its standard deviation as $\hat{\sigma} = 1.4826 \cdot \text{MAD}$ [12]. Therefore, we define the MAD inlier cutoff, $\tau_{3\sigma}$, as follows:

$$\begin{aligned} m &= \text{median}(|e|), \\ \text{MAD} &= \text{median}(||e| - m|), \\ \tau_{3\sigma} &= m + 3 \cdot \hat{\sigma} = m + 4.4478 \cdot \text{MAD}. \end{aligned} \tag{4}$$

The so-calculated inlier cutoffs, $\tau_{3\sigma}$, and their manually selected counterparts, τ_{man}, are shown in Table 3, together with their corresponding outlier rates, and the difference Δ between the resulting outlier rates. Six out of nine times, the outlier rates obtained by using the automatically derived $\tau_{3\sigma}$ approximate the ground-truth outlier rates (obtained using τ_{man}) relatively well (maximum difference of ±3.5 %). The other three times they differ by only 5.22–7.29%. That is, the automatic outlier labeling method performs reasonably well for automatic outlier filtering before calibration.

Regarding the outlier rates of the devices in each of the environments: outdoors, SiLabs 1×1 requires larger τ_{man} than Nordic (Grass: 3.29 vs. 2.69 m; Asphalt: 4.46 vs. 1.56 m) and shows higher outlier rates (10.81–11.03% vs. 4.19–6.34%). This aligns with the CDF tails observed in Figure 6. In Indoor, the gap widens: $\tau_{\text{man}} = 4.31$ m and 17.98% outliers for SiLabs 1×1 versus 1.76 m and 9.12% for Nordic. For SiLabs 1×1, this means that 17.98% of ranging estimates in Indoor exceeded 4.31 m error. To evaluate dual-antenna ranging in terms of outlier occurrence rates, SiLabs 2×2 is also included in Table 3 and in Figure 6. Interestingly, SiLabs 2×2 exhibits the smallest outlier rates per environment (1.08–3.71%), much lower than SiLabs 1×1 (10.81–17.98%) and Nordic (4.19–9.12%).

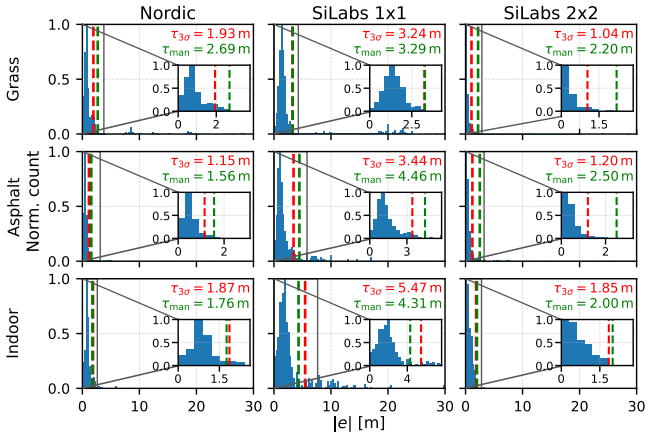


Figure 6: Histograms of the absolute ranging error $|e|$ for Nordic, SiLabs 1×1 and SiLabs 2×2. The vertical dashed lines indicate (green) the manually selected threshold τ_{man}, and (red) the automatically calculated threshold $\tau_{3\sigma}$.

Table 3: Automatically calculated versus manually selected outlier thresholds and rates per device and environment.

Env.	Device	$\tau_{3\sigma}$ [m]	τ_{man} [m]	Outl. (3σ) [%]	Outl. (man.) [%]	Δ [%]
Grass	Nordic	1.93	2.69	12.40	6.34	6.06
	SiLabs 1×1	3.24	3.29	10.81	10.81	0.00
	SiLabs 2×2	1.04	2.20	8.38	1.09	7.29
Asphalt	Nordic	1.15	1.56	7.27	4.19	3.07
	SiLabs 1×1	3.44	4.46	13.33	11.03	2.30
	SiLabs 2×2	1.20	2.50	8.93	3.71	5.22
Indoor	Nordic	1.87	1.76	6.05	9.12	-3.07
	SiLabs 1×1	5.47	4.31	17.13	17.98	-0.85
	SiLabs 2×2	1.85	2.00	1.15	1.08	0.07

Transferability of outlier cutoffs. Table 3 further demonstrates that optimal outlier cutoffs τ_{man} vary significantly across devices and environments. Specifically, the largest variation across devices occurs in Asphalt (4.46 m vs. 1.56 m, i.e., 2.90 m), while the variation across environments for a single device is similar for SiLabs 1×1 (4.46 m vs. 3.29 m, i.e., 1.17 m) and Nordic (2.69 m vs. 1.56 m, i.e., 1.13 m). Comparing SiLabs 1×1 to SiLabs 2×2, indicates that multi-antenna ranging is superior to single-antenna ranging in this regard, as the variation of the outlier cutoffs across environments is significantly smaller for SiLabs 2×2 (2.50 m vs. 2.00 m). Consequently, the presented automated outlier labeling method, which adapts to each device-environment pair, facilitates automatic and reproducible outlier filtering before calibration.

Table 4: Calibration model selection via LOOCV: linear model with fixed unit slope (fix) versus free linear model (lin).

Env.	Device	RMSE_{fix}	RMSE_{lin}	MAE_{fix}	MAE_{lin}
Grass	Nordic	0.453	0.444	0.334	0.332
	SiLabs 1×1	0.662	0.685	0.484	0.502
Asphalt	Nordic	0.294	0.297	0.214	0.217
	SiLabs 1×1	0.771	0.806	0.583	0.630

7 Calibration Procedure

Both the IFFT-based and the MUSIC-based range estimators are expected to be centered (that is, bias-free) under favorable conditions, since they recover the propagation delay from the measured channel response: the dominant IFFT peak corresponds to the physical distance up to a constant hardware delay that can be removed as an additive offset [15, 18], while MUSIC derives distance from the first peak of its pseudo-spectrum and is likewise centered [3].

Calibration model selection. To validate the calibration model, we compare two approaches: (i) a linear model with a fixed unit slope, adhering to the theoretical expectation, and (ii) an unrestricted linear model that fits both the slope and intercept. By evaluating both via leave-one-out cross-validation (LOOCV), we can determine whether the added flexibility of a calibrated slope yields a significant improvement in accuracy or if the unit-slope assumption proves empirically sound. In each LOOCV fold, one entire ground-truth distance is excluded from training; then, the model is fitted on all remaining distances, and then evaluated on the held-out distance across all relative device orientations. The resulting LOOCV RMSE/MAE metrics in Table 4 are estimates of the RMSE/MAE after calibration without outliers, since they were already filtered (using the automatic procedure explained in Sec. 6). These LOOCV results indicate that both models attain similar RMSE and MAE. Therefore, we favor the selection of the simpler model, that is, the linear model with fixed unit slope. Its reduced complexity makes it less prone to overfitting, and thus a more robust choice for calibration.

Calibration procedure. We utilize the linear model with fixed unit slope for calibration and measurement correction. That is, the calibration consists of a single offset, which, throughout this paper, we usually term the calibration offset.

The calibration proceeds as follows. For every sample with measured distance m and ground-truth distance g, we compute the signed error $e = m - g$. We then remove outliers using the automatically calculated cutoff $\tau_{3\sigma}$ (Sec. 6), which is specific to the environment and device. That is, we only keep samples with $|e| \leq \tau_{3\sigma}$. Using the remaining inliers, we estimate the calibration offset β as the mean error of the aggregated measurements, that is, $\beta = \text{mean}(e)$. We apply bias correction by subtraction of the bias β from the raw measurement m_{raw} as follows, $\hat{m} = m_{\text{raw}} - \beta$, with $\hat{m}$ being the bias-corrected distance estimate.

Calibration offsets can be calculated: (i) per device, environment and relative-orientation pair, (ii) per device and environment, (iii) per device, and relative-orientation pair, or (iv) per-device. The difference during the calculations is on the data to be aggregated for the calibration procedure. That is, each β is computed by aggregating all relevant data.

Table 5: Calibration offsets β for Nordic and SiLabs 1×1, and remaining bias γ for SiLabs 2×2. The calibration procedure of rows labeled "YZ–UL" follow Ulsamer *et. al.* [22], while "XY–SL" and "YZ–SL" follow the Silicon Labs guidelines [18].

Env.	Orient. pair	Nordic		SiLabs 1×1		SiLabs 2×2
		β [m]	Δ [m]	β [m]	Δ [m]	γ [m]
Grass	XY–XY	0.58		1.51		−0.25
	XY–XZ	1.01		1.68		−0.37
	XY–YZ	0.56		1.65		−0.21
	YZ–XZ	0.97		1.35		0.00
	YZ–YZ	0.65		1.09		−0.17
	All	0.75		1.38		−0.20
	YZ–UL	0.91	0.17	1.06	−0.32	−0.43
	XY–SL	0.58	−0.16	1.51	0.13	−0.19
	YZ–SL	0.65	−0.09	1.09	−0.29	0.12
Asphalt	XY–XY	0.37		2.42		0.57
	XY–XZ	0.50		1.46		0.23
	XY–YZ	0.48		1.25		0.34
	YZ–XZ	0.31		1.10		0.26
	YZ–YZ	0.44		1.04		0.08
	All	0.43		1.24		0.29
	YZ–UL	0.55	0.12	1.11	−0.14	0.14
	XY–SL	0.37	−0.06	2.42	1.17	0.56
	YZ–SL	0.44	0.01	1.04	−0.21	0.02
Grass+Asphalt	All	0.59		1.31		0.04
	YZ–UL	0.73	0.14	1.08	−0.23	−0.15
	XY–SL	0.41	−0.18	1.51	0.20	0.19
	YZ–SL	0.55	−0.04	1.08	−0.23	0.07
Indoor	XY–XY	0.75		1.96		0.01
	XY–XZ	1.24		2.55		0.56
	XY–YZ	0.88		1.41		−0.02
	YZ–XZ	0.60		1.99		0.46
	YZ–YZ	0.54		1.24		0.15
	All	0.75		1.69		0.23
	YZ–UL	0.64	−0.11	0.97	−0.72	0.12
	XY–SL	0.75	0.00	1.96	0.27	0.20
	YZ–SL	0.54	−0.21	1.24	−0.45	−0.31
All	XY–XY	0.53		1.50		0.11
	XY–XZ	0.92		1.89		0.14
	XY–YZ	0.64		1.44		0.04
	YZ–XZ	0.63		1.48		0.24
	YZ–YZ	0.62		1.03		0.02
	All	0.64		1.44		0.11
	YZ–UL	0.70	0.06	1.04	−0.39	−0.06
	XY–SL	0.53	−0.12	1.50	0.06	0.19
	YZ–SL	0.62	−0.03	1.03	−0.41	−0.06

Calibration results. Following our calibration procedure, we calculate the calibration offsets β for Nordic and SiLabs 1×1 and report them in Table 5. The SiLabs 2×2 stack provides default calibration offsets for ANT1 and ANT2 (Sec. 4); therefore, for SiLabs 2×2 the same procedure yields the *remaining bias* γ for a given configuration w.r.t. the calibration configuration.

The results in Table 5 shown in grayed out blocks are calibration offsets β calculated using state-of-the-art methods. YZ–UL denotes the calibration performed following the procedure used by Ulsamer *et. al.* in [22]. That is, in YZ–UL, the calibration offset is obtained by aggregating data for the relative-orientation pair YZ–YZ and using only the reflector in-plane rotation of 90° (*Method 1*, Sec. 2). XY–SL and YZ–SL denote the calibration procedure, described in Silicon

Labs' antenna design guideline [18], applied to the corresponding relative-orientation pair, XY–XY or YZ–YZ, respectively. This calibration procedure corresponds to a *median-of-medians and linear regression* approach (*Method 2*, Sec. 2).

As shown in Table 5, for both Nordic and SiLabs 1×1, the calibration offsets β vary substantially across relative-orientation pairs. For instance, β ranges 0.56–1.01 m and 0.31–0.50 m for Nordic respectively in Grass and Asphalt. The environment also affects the calibration offsets. Using all rows per environment (pooling all relative-orientation pairs), Nordic shifts from 0.75 m (Grass) to 0.43 m (Asphalt), i.e., a difference of 32 cm, which is smaller than the relative-orientation-induced spread for Grass (45 cm). For SiLabs 1×1, the Grass–Asphalt environment shift is smaller (1.38 m to 1.24 m, i.e., a difference of 14 cm), while the relative-orientation-induced spreads are larger: 0.59 m in Grass (1.68 m to 1.09 m) and 1.38 m in Asphalt (2.42 m to 1.04 m).

For Nordic, the maximum difference between the offsets calculated using the state-of-the-art calibration procedures (rows YZ–UL, XY–SL, YZ–SL) versus our calibration procedure is 21 cm (Δ column). This specifically happens in Indoor, when comparing the procedure YZ–SL ($\beta = 0.54$ m) versus our result for all orientations (row 'All' in the column Orient. pair, $\beta = 0.75$ m). For SiLabs 1×1, these differences reach higher values of 45 cm, 72 cm and even 117 cm. For SiLabs 1×1, the differences of 45 cm and 72 cm happen in Indoor, and can be found in the Δ entries of the SiLabs 1×1 column (YZ–SL: $\Delta = -0.45$ m, YZ–UL: $\Delta = -0.72$ m); and the maximum difference compared to the state-of-the-art procedures happens in Asphalt for the XY–SL procedure ($\Delta = 1.17$ m). These differences in the offsets would affect all rangings during deployment. For state-of-the-art methods, these differences are mainly caused by the aggregation of data from only one relative-orientation pair, which shows that calibration datasets should prioritize orientation diversity. For SiLabs 2×2, the maximum error in the calculation of the remaining bias γ incurred by a state-of-the-art method is 54 cm for the method YZ–SL in Indoor, which results from comparing the values of γ for YZ–SL (−0.31 m) and our procedure (0.23 m).

Generally, the results in Table 5 show higher variability when switching relative-orientations pairs in a given environment than by switching environments. Specifically, the maximum relative-orientation-induced variability per environment happens: for Nordic in Indoor (1.24 m to 0.54 m, i.e., 0.70 m), for SiLabs 1×1 in Asphalt (2.42 m to 1.04 m, i.e., 1.38), and for SiLabs 2×2 in Indoor (0.56 m to −0.02 m, i.e., 0.58). In contrast, the maximum calibration offset shift caused by changing environment happens: for Nordic when changing from either Grass or Indoor to Asphalt (0.43 m to 0.75 m, i.e., 0.32 m), for SiLabs 1×1 when changing from Asphalt to Indoor (1.24 m to 1.69 m, i.e., 0.45 m), and for SiLabs 2×2 when changing from Grass to Asphalt (−0.20 m to 0.29 m, i.e., 0.49).

That is, the maximum error of the bias correction that occurs by using the calibration from another environment is only 32 cm, 45 cm and 49 cm respectively for Nordic, SiLabs 1×1, and SiLabs 2×2. Whereas the maximum error of the bias correction that occurs when calibrating only using one relative-orientation pair (in the same environment) is 70 cm, 138 cm and 58 cm respectively for Nordic, SiLabs 1×1, and SiLabs 2×2. These results highlight the importance of orientation-diverse calibration over calibration procedures using a single relative-orientation pair. Moreover, when the relative 3D bearing and orientation between devices are approximately known, a relative-orientation-specific bias correction can be applied, significantly reducing, or even eliminating, the bias introduced by orientation. However, this strategy requires further validation and is left for future work. While our findings are promising, datasets with finer resolution of relative orientations are needed to further benchmark such bias-correction methods.

Table 6: Ranging performance [m] in Indoor after calibration with global offset (calculated from Grass+Asphalt).

Device	Orient. pair	MAE	MAE*	RMSE	RMSE*	P50	P75	P90	P95
Nordic	XY–XY	0.42	0.34	0.68	0.41	0.20	0.39	0.50	0.77
	XY–XZ	1.12	1.02	1.27	1.11	1.16	1.51	1.69	2.00
	XY–YZ	0.68	0.45	1.21	0.52	0.39	0.55	1.02	2.47
	YZ–XZ	0.41	0.39	0.62	0.55	0.09	0.34	0.54	0.58
	YZ–YZ	0.24	0.21	0.44	0.31	0.05	0.18	0.33	0.44
	All	0.53	0.44	0.88	0.59	0.20	0.49	1.02	1.51
	XY–SL	0.64	0.54	0.95	0.67	0.38	0.67	1.20	1.69
SiLabs 1×1	XY–XY	1.61	1.11	3.09	1.80	0.60	1.38	4.70	5.90
	XY–XZ	3.67	1.41	6.70	1.70	1.40	3.00	9.10	14.80
	XY–YZ	1.89	0.64	5.40	0.88	0.20	0.80	5.62	6.78
	YZ–XZ	3.98	1.16	6.45	1.57	1.10	7.00	10.75	12.60
	YZ–YZ	0.94	0.53	3.26	0.76	-0.12	0.10	0.35	0.50
	All	2.42	0.94	5.21	1.39	0.60	1.55	7.00	10.02
	YZ–UL	2.54	1.02	5.31	1.45	0.83	1.78	7.23	10.25
SiLabs 2×2	XY–XY	0.40	0.37	0.70	0.45	0.06	0.31	0.61	0.71
	XY–XZ	0.80	0.79	0.97	0.93	0.61	1.11	1.51	1.71
	XY–YZ	0.58	0.52	1.11	0.62	0.01	0.47	0.91	0.91
	YZ–XZ	0.71	0.66	1.33	0.80	0.46	0.99	1.21	1.42
	YZ–YZ	0.72	0.60	1.65	0.76	0.16	0.51	1.31	1.71
	All	0.64	0.59	1.19	0.73	0.21	0.71	1.21	1.51

8 BLE-CS Performance After Calibration

Table 6 summarizes the ranging performance in Indoor after calibration, where we use a (constant) device-specific bias β derived from the outdoor datasets (row Grass+Asphalt–All shown in Table 5). The offset is subtracted from each raw ranging estimate. We use $\beta = 0.59$ m for Nordic and $\beta = 1.31$ m for SiLabs 1×1. The table also reports the performance of the dual-antenna SiLabs 2×2 to directly compare it with the single-antenna ranging mode. Ranging performance is shown both with and without outliers; results after outlier filtering (e.g., RMSE*) are marked with an asterisk (*), and the gray-highlighted rows correspond to the specified state-of-the-art methods (namings described in Sec. 7). We are choosing XY–SL for Nordic and YZ–UL for SiLabs 1×1, as they have the biggest Δ compared to our derived global offset (Table 5).

For Nordic, aggregating over all orientation pairs ('All' row) yields an MAE in Indoor of 0.53 m and RMSE of 0.88 m, which improve to 0.44 m and 0.59 m after outlier removal (MAE* and RMSE*). The median absolute error (P50) is around 0.20 m, i.e., close to zero bias, while P95 stays below 1.6 m, indicating that outliers remained rare. Compared to *XY–SL*, which achieves P50 ≈ 0.38 m and P95 ≈ 1.69 m, our global-offset calibration yields lower median and tail errors, and improves RMSE* by 12% (0.67 m reduced to 0.59 m).

For SiLabs 1×1, the global-offset calibration also reduces the overall bias, but substantial orientation-dependent tails remain. Aggregated over all relative-orientation pairs, the percentiles P90 and P95 reach 7.00 m and 10.02 m, respectively, reflecting heavy-tailed error

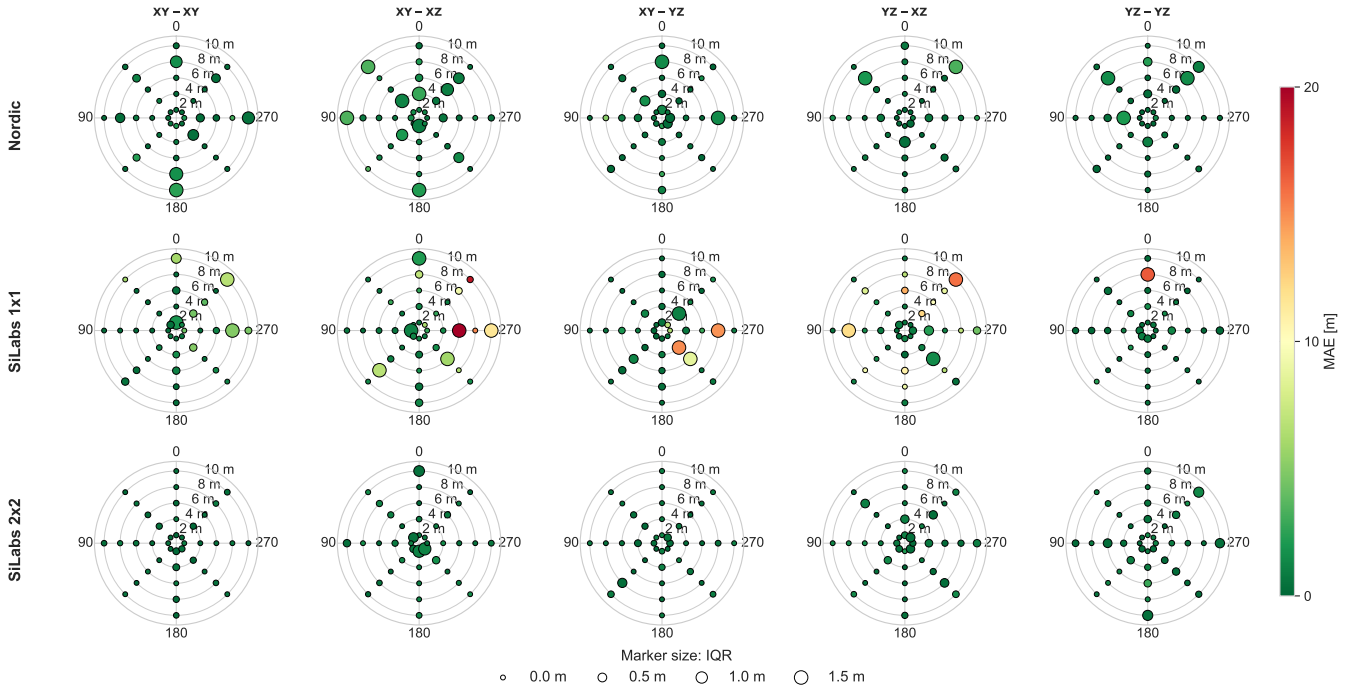


Figure 7: Polar plots of ranging error in `Indoor` after calibration for (rows) devices and (columns) relative orientations. Marker color and size respectively encode `MAE` and `IQR`. Dots and rings indicate in-plane rotations and ground-truth distances.

distributions driven by outliers in certain unfavorable orientation pairs. After outlier filtering, `MAE*` and `RMSE*` improve to 0.94 m and 1.39 m, respectively, but the per-orientation rows in Table 6 show that some configurations (XY–XY, XY–XZ, and YZ–XZ) still exhibit noticeably higher errors (`RMSE*`) than others. The *YZ–UL* calibration applies an offset close to ours (1.08,m vs. 1.31,m; Table 5), so gains are small: our calibration improves `RMSE*` by only 4.1%.

In contrast, `SiLabs 2×2` achieves consistently low error and dispersion across all orientation pairs in `Indoor`, with in comparison very low outlier occurrence rates. Compared to the single-antenna ranging mode (`SiLabs 1×1`), whose aggregated median (`P50`) in `Indoor` is 0.60 m, `SiLabs 2×2` achieves smaller median error (`P50` ≈ 0.21 m) while also presenting substantially fewer outliers (e.g., `P95` ≈ 1.5 m instead of over 10 m for `SiLabs 1×1`). Furthermore, dual-antenna ranging (`SiLabs 2×2`) in comparison to single-antenna ranging (`SiLabs 1×1`) greatly improves `RMSE*` by 47.5% (1.39 m reduced to 0.73 m), but its effect is even greater when considering outliers, that is, `RMSE` improves by 77.6% (5.31 m reduced to 1.19 m); which again confirms the superiority of dual-antenna ranging.

The mismatch between calibration and deployment device orientations can introduce substantial systematic bias and increase ranging error. This is visible in the results shown in Table 6. Specifically, the performance of `SiLabs 1×1` in `Indoor` varies significantly depending on the relative orientation between the devices. For example, the difference in performance between the best-performing orientation pair (YZ–YZ) and the XY–XZ pair is substantial, with a difference of 88 cm in terms of `MAE*` (from 0.53 cm to 1.41 cm) and 94 cm in terms of `RMSE*` (from 0.76 cm to 1.70 cm). This is partly caused by the global offset used for calibration (1.31 cm, Table 5) being much closer to that of the YZ–YZ pair (1.24 cm) than to that of the XY–XZ pair (2.55 cm).

Error patterns in `Indoor`. Figure 7 provides a compact overview of the ranging performance in `Indoor` with additional insights on the in-plane board rotation angles by means of polar plots after applying a single global outdoor offset (`Grass+Asphalt`–All) per device. For the creation of the polar plots outliers are kept in the datasets (that is, no outlier filtering is performed). Each subplot corresponds to one device (rows) and one relative-orientation pair (columns). Each marker represents one configuration defined by a ground-truth distance and a rotation angle: the ring's radius encodes the ground-truth distance (rings at 2, 4, 6, 8, and 10 m), and the marker angle encodes the in-plane board rotation (0°, 45°, 90°, …, and 315°). The marker color indicates *accuracy* via `MAE`, while the marker size indicates *precision* via `IQR`; thus, small green markers correspond to accurate measurements, whereas large orange/red markers indicate degraded accuracy and precision.

For `SiLabs 1×1`, comparing with the CDFs (Figure 5), the distribution of outliers is consistent between both figures. `SiLabs 2×2` achieves consistently low `MAE` and tight `IQR` across all polarization pairs and rotation angles, which is visible in its polar plots. Its visual uniformity aligns with Table 6 and Figure 5 (CDFs). Overall, this confirms that antenna diversity (`SiLabs 2×2`) improves robustness compared to its single-antenna counterpart (`SiLabs 1×1`).

Using polar plots (as in Figure 7) to analyze in-plane rotations, we have attempted to find consistent trends in the occurrence of increased outlier rates and of decreased accuracy and precision. However, we could not find clear patterns with respect to in-plane rotations. As shown in Figure 7, the in-plane rotation angles that

present outliers also present measurements with very good performance. This might be due to our coarse sampling of in-plane rotations (45° steps, positioned manually without technical aids) and relative orientations (only 5), which may not have consistently captured strong dips in antenna radiation patterns.

9 Online Outlier Detection

In this section, we evaluate Nordic's Tone Quality Indicator (TQI), a vendor-specific confidence metric for measurement reliability, to achieve online outlier detection, with the goal of allowing real-time embedded systems to discard unreliable data.

Outlier detection using only the TQI. In Nordic, the controller reports one quality flag per tone, corresponding to an IQ sample at the initiator or reflector, which reflects the expected reliability of the measurement under current radio conditions [2]. Each tone exchange yields a TQI by combining its two quality flags via an AND operation. Consequently, assuming a full channel map and a channel repetition of 1, each ranging procedure produces 72 TQIs (i.e., one per BLE channel). Per ranging, we convert all TQIs into a binary label: a ranging is marked as BAD when the number of BAD TQIs is equal or greater than the bad-tone count threshold, τ_{TQI}.

We therefore use our datasets to evaluate outlier detection performance versus the threshold τ_{TQI}. Table 7 shows the confusion matrix for outlier detection, using the manually selected thresholds for outlier filtering τ_{man} (Sec. 6) as the ground-truth outlier label, for selected bad-tone count thresholds $\tau_{\text{TQI}} \in \{1, 2, 15\}$. The confusion matrix entries TP, FN, FP, and TN denote the true positive, false negative, false positive, and true negative rates, respectively, which are used to derive the true positive rate (TPR), the false positive rate (FPR) and the accuracy. Figure 8 shows how detection performance (TPR, FPR, and accuracy), change with the bad-tone count threshold, τ_{TQI}. Figure 9 shows the absolute error distribution split (OK vs. BAD) by TQI according to the bad-tone count threshold $\tau_{\text{TQI}} = 1$.

For outlier detection, classifying samples as OK vs. BAD using Nordic's TQI is highly effective in Grass, achieving a TPR of 99.37–99.68% (that is, $6.32/(6.32 + 0.02) = 99.68\%$). However, its performance degrades significantly in the other environments: the TPR drops to 51.64–56.81% in Asphalt and further to 5.18–33.38% in Indoor. In addition, outlier detection comes at the cost of false alarms on inliers. For instance, in Grass, around 27.45% of inliers are misclassified as BAD (a FPR of 25.74/(67.92+25.74)).

The value $\tau_{\text{TQI}} = 1$ yields the best possible TPR, but also the worst FPR. The detection performance in Figure 8 show that a good balance between TPR and FPR across environments is achieved with $\tau_{\text{TQI}} = 2$. Increasing τ_{TQI} gradually increases accuracy, with TPR in Grass, Asphalt, and Indoor trending down respectively towards 99%, 50% and slowly to zero. The values $\tau_{\text{TQI}} \in \{4, 15\}$ are the defaults set by Nordic in respectively the nRF Connect SDKs v3.0.2 and v3.2.4. With $\tau_{\text{TQI}} = 15$ the outlier detection achieves around best accuracy but misses almost all outliers in Indoor (TPR= 5.18%). The TQI provides a useful cue for spotting extreme ranging errors, achieving very high TPRs in clean outdoor environments under LoS conditions (≈ 99.5% in Grass). However, in Asphalt and Indoor, all TQIs are respectively OK for 43.19% and 66.62% (1−TPR = 100−33.38) of the outliers, which renders TQI-based outlier detection relatively ineffective in these environments. These differences on outlier detection performance are likely due to: (i) differences in the ranging distances (maximum ranging distances of 20 m, 37 m and 10 m in Grass, Asphalt and Indoor, respectively), and (ii) harsher multipath conditions and RF interference in Asphalt and Indoor, which degrade ranging performance, but not necessarily trigger BAD TQIs, whereas the Grass environment offers cleaner channels where BAD tones and large ranging errors coincide more often. Consequently, relying on Nordic's TQI can lead to substantial rates of misclassification. Given this limitation, we advise against using the current TQI implementation as the sole metric for outlier detection.

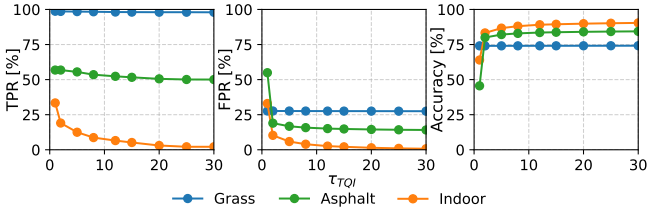


Figure 8: Online outlier detection performance (TPR, FPR, and accuracy) versus bad-tone count threshold τ_{TQI}.

Table 7: TQI-based outlier detection confusion matrix (%), with ground-truth defined using the manual thresholds τ_{man}, for selected bad-tone count thresholds $\tau_{\text{TQI}} \in \{1, 2, 15\}$.

Env.	τ_{TQI}	TP	FN	FP	TN	TPR	FPR	Acc.
Grass	1	6.32	0.02	25.74	67.92	99.68	27.48	74.24
	2	6.31	0.03	25.73	67.93	99.53	27.47	74.24
	15	6.29	0.04	25.70	67.95	99.37	27.45	74.24
Asphalt	1	2.38	1.81	52.62	43.19	56.81	54.92	45.57
	2	2.38	1.81	18.17	77.64	56.81	18.96	80.02
	15	2.17	2.03	14.21	81.60	51.64	14.83	83.76
Indoor	1	3.04	6.07	29.96	60.92	33.38	32.97	63.96
	2	1.74	7.38	9.35	81.53	19.05	10.29	83.27
	15	0.47	8.64	1.92	88.97	5.18	2.11	89.44

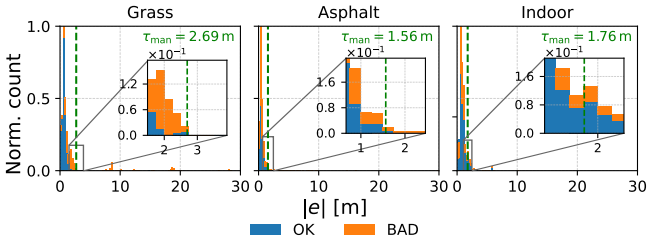


Figure 9: Absolute error $|e|$ histograms across environments split (OK vs. BAD) by Nordic's reliability metric TQI for $\tau_{\text{TQI}} = 1$.

10 Discussion and Future Work

Our findings show that relative board orientation can be major source of bias in BLE-CS ranging, with an impact that can exceed the variation caused by environmental changes. At the same time, there are several limitations of this study that should be considered when interpreting its findings and defining future research directions.

Limited environment diversity. The evaluation covers only three environments: two outdoor settings and one indoor hallway. More indoor and outdoor orientation-diverse datasets are needed to assess the generality of our findings.

Restriction to static LoS conditions. All experiments were conducted under clear line-of-sight conditions with static devices mounted on tripods. While this enables controlled benchmarking, it does not capture important real-world factors such as mobility, blockage by the human body carrying the device or nearby objects, and non-line-of-sight propagation. Future work should investigate how orientation effects interact with such conditions.

Is using two antennas the only way to go? Although dual-antenna ranging is superior to single-antenna ranging, this improvement usually comes at the cost of higher energy consumption, increased hardware complexity, and higher algorithmic overhead. These trade-offs are particularly relevant for battery-constrained devices. Future work should hence investigate whether comparable performance could be achieved with single-antenna systems, for instance, using signal processing techniques.

Cross-platform comparison. The goal of this study was not to compare performance across different platforms, but to characterize BLE 6.0 CS performance in general. A comparison between `Nordic` and `SiLabs` intended to attribute observed performance differences to a single factor would not even be possible due to differences in transmit power, PHY-related settings, firmware stacks, ranging algorithms, and antenna/board design.

Observational nature of the study. This paper primarily quantifies performance, but does not implement mitigation strategies. In particular, we do not evaluate orientation-aware bias correction, ML-based outlier detection, or other system-level compensation methods. Our findings motivate future work on all these aspects.

Focus on ranging rather than localization. The study evaluates link-level ranging accuracy, calibration offsets, and outlier rates, but does not quantify their impact on end-to-end localization performance. In practice, orientation-dependent effects may vary significantly based on node placement and the localization algorithm.

11 Conclusions

Our experiments with commercial BLE 6.0 Channel Sounding platforms show that relative board orientation between initiator and reflector can shift the calibration offset by up to 138 cm (`SiLabs` 1×1) and 70 cm (`Nordic`). This systematic bias dominates the modest environment-induced variations, confirming that orientation-diverse data are essential for BLE-CS calibration. Our experiments confirm that the dual-antenna ranging mode consistently outperformed its single-antenna counterpart, with `RMSE` improved by 47.5%, and outlier rates dropping from 10–18% to 1–4%.

Acknowledgments

This work was partially conducted within the SPIDR^2 project ("Secure, Performant, Intelligent, Dependable, Reliable, and Resilient Wireless Systems") financed by the Technology Innovation Institute. We also thank Firat Bağci for his support in acquiring part of the experimental data used in this study.

References

[1] Safar M. Asaad et al. 2022. A Comprehensive Review of Indoor/Outdoor Localization Solutions in IoT era: Research Challenges and Future Perspectives. *Computer Networks* (2022).
[2] Bluetooth SIG. August, 2024. Bluetooth Core Specification, Version 6.0. Bluetooth® Technology Website. Accessed: Dec. 17, 2025.
[3] Pepijn Boer et al. 2020. Performance of High-Accuracy Phase-Based Ranging in Multipath Environments. In *Proceedings of the IEEE 91st Vehicular Technology Conference (VTC-Spring)*.
[4] Aytürk Düzen. 2019. *Implementation and evaluation of phase-based group ranging and localization with BLE.* Master's thesis. Eindhoven University of Technology.
[5] Ramsey Faragher et al. 2015. Location Fingerprinting With Bluetooth Low Energy Beacons. *IEEE Journal on Selected Areas in Communications* (2015).
[6] Mohammadali Ghaemifar et al. 2025. Bluetooth Low Energy for Indoor Positioning: Challenges, Algorithms and Datasets. *Automation in Construction* (2025).
[7] Supatana Hengyotmark et al. 2017. Pseudo-ranging Based on Round-Trip Time of Bluetooth Low Energy Beacons. In *Proceedings of the 13th International Conference on Computing and Information Technology (IC2IT)*.
[8] Chenglin Huang et al. 2023. Spotlight: A 3-D Indoor Localization System in Wireless Sensor Networks Based on Orientation and RSSI Measurements. *IEEE Sensors Journal* (2023).
[9] Maciej Nikodem et al. 2025. Experimental Evaluation of Multicarrier Phase Difference Localization in Bluetooth Low Energy. *IEEE Sensors Journal* (2025).
[10] Huthaifa Obeidat et al. 2021. A Review of Indoor Localization Techniques and Wireless Technologies. *Wireless Personal Communications* (2021).
[11] Qorvo. 2024. The Future of Positioning: Exploring UWB and Bluetooth Channel Sounding. https://www.qorvo.com/design-hub/blog/the-future-of-positioning-exploring-uwb-and-bluetooth-channel-sounding. Accessed: Jan 2026.
[12] Peter J Rousseeuw et al. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical association* (1993).
[13] Avik Santra et al. 2024. Enhancing Bluetooth Channel Sounding Performance in Complex Indoor Environments. *IEEE Sensors Letters* (2024).
[14] Ralph Schmidt. 1986. Multiple Emitter Location and Signal Parameter Estimation. *IEEE Transactions on Antennas and Propagation* (1986).
[15] Yannic Schröder et al. 2018. Accurate and Precise Distance Estimation from Phase-Based Ranging Data. In *Proceedings of the 9th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*.
[16] Yannic Schröder et al. 2019. Investigation of Multipath Effects on Phase-based Ranging. In *Proceedings of the 10th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*.
[17] Silicon Laboratories Inc. 2024. Channel Sounding Performance Metrics and Benchmark using BRD2606A BLE-CS devices. https://docs.silabs.com/rtl-lib/10.0.2/rtl-lib-channel-sounding-dev-guide/07-channel-sounding-performance-metrics. Version 10.0.2. Accessed: 2026-01.
[18] Silicon Laboratories Inc. 2025. AN1493: Antenna Design Guidelines for BLE Channel Sounding. https://www.silabs.com/documents/public/application-notes/an1493-antenna-design-guidelines-for-ble-channel-sounding.pdf. Revision 0.3. Accessed: 2025-11.
[19] Anna Strzoda et al. 2022. Variability of BLE Advertisement Packets Received Signal Strength and Delivery Probability in the Presence of Interferences. In *Proceedings of the 12th ACM International Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications (DIVANet)*.
[20] Zaid Bin Tariq et al. 2024. A Data-Driven Signal Subspace Approach for Indoor Bluetooth Ranging. *IEEE Journal of Indoor and Seamless Positioning and Navigation* (2024).
[21] Andrii Tsemko et al. 2025. Data-driven Processing using Parametric Neural Network for Improved Bluetooth Channel Sounding Distance Estimation. In *Proceedings of the IEEE 50th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
[22] Tim Ulsamer et al. 2025. Study of BLE6 Ranging Performance in a Utility Basement. In *Proceedings of the Workshop for Computing & Advanced Localization at the 15th International Conf. on Indoor Positioning and Indoor Navigation (IPIN-WCAL 2025)*.
[23] Jayson P. Van Marter et al. 2023. Support Vector Regression for Bluetooth Ranging in Multipath Environments. *IEEE Internet of Things Journal* (2023).
[24] Bowen Wang et al. 2021. BLE Localization With Polarization Sensitive Array. *IEEE Wireless Communications Letters* (2021).
[25] Yapeng Wang et al. 2013. Bluetooth positioning using RSSI and triangulation methods. In *Proceedings of the IEEE 10th Consumer Communications and Networking Conference (CCNC)*.
[26] Jorg Wieme et al. 2025. Performance Evaluation of Bluetooth Channel Sounding on Commercial Hardware. *IEEE Access* (2025).
[27] Martin Woolley. 2021. *Bluetooth Direction Finding — A Technical Overview.* Technical Report. Bluetooth SIG, Kirkland, WA, USA.
[28] Hongyun Ye et al. 2022. A Method of Indoor Positioning by Signal Fitting and PDDA Algorithm Using BLE AOA Device. *IEEE Sensors Journal* (2022).
[29] Pouria Zand et al. 2019. A high-accuracy concurrent phase-based ranging for large-scale dense BLE network. In *Proceedings of the IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*.
[30] Pouria Zand et al. 2019. A high-accuracy phase-based ranging solution with Bluetooth Low Energy (BLE). In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*.