SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I10P104

Volume 13 Issue 10, 54-67, October 2026
© 2026 Seventh Sense Research Group®

Original Article

Joint Clustering, Association, and Power Allocation in Cell-Free Massive MIMO-NOMA

Gurudevi A. Hiremath[1], Shankar D. Nawale[2]

[1]*SKN College of Engineering, Korti, Pandharpur, India and PAH Solapur University, Solapur, Maharashtra, India.*
[2]*N.B. Navale Sinhgad College of Engineering, Solapur, Maharashtra, India.*

[1]*Corresponding Author: gurudevihiremath17@gmail.com*

Received: 06 August 2026 Revised: 16 September 2026 Accepted: 23 September 2026 Published: 28 October 2026

Abstract *- This study revisits the joint design of user clustering, Access-Point (AP) association, transmit-antenna selection, and power allocation in downlink cell-free massive MIMO with Non-Orthogonal Multiple Access (NOMA). A rigorous Monte-Carlo evaluator with pilot contamination, Minimum Mean-Square-Error (MMSE) channel estimation, maximum-ratio transmission, and Successive Interference Cancellation (SIC) scores every candidate configuration consistently under common random numbers. A central methodological point is emphasized: prior clustering-centric comparisons hold AP association and power fixed, which conflates the clustering effect with an unoptimized association budget. When all schemes are granted an identical association/antenna/power search-a matched-scope comparison over 40 independent topologies—a clear and statistically significant ordering of resource levers emerges: AP association is dominant (+20.2 bps/Hz), clustering optimization is a significant second lever (+13.0 bps/Hz), the clustering algorithm itself contributes a small but significant increment (+4.1 bps/Hz), and transmit-antenna selection is spectrally neutral, all under the parameter regime investigated (Table 3); the ordering is not asserted to hold universally across every cell-free massive MIMO-NOMA configuration. A block-coordinate joint optimizer is proposed and shown, by exhaustive enumeration on small instances (K=9), to attain a mean optimality gap of only 0.64%, a result that is evidence for, but not proof of, near-optimality at the larger K=30 scale used elsewhere in the evaluation; it improves the sum Spectral Efficiency (SE) by 16.0% over the best matched-scope heuristic (and by 54.3% over the conventionally evaluated base method) while converging monotonically in about five iterations. Its discrete power grid is shown to attain 99.3% of a continuous power optimum. Transmit-antenna selection is characterized as an energy-efficiency lever, the energy-optimal configuration activating four of eight antennas for a 14.3% energy-efficiency gain at a 28.2% SE cost.*

Keywords *- Cell-free massive MIMO, NOMA, User clustering, AP association, Antenna selection, Power allocation, Resource-lever attribution.*

1. Introduction

Cell-free massive Multiple-Input Multiple-Output (MIMO) has emerged as a leading architecture for beyond-fifth-generation networks, replacing the cell-centric paradigm with a large number of geographically distributed Access Points (APs) that cooperatively and coherently serve all users in the same time-frequency resource [1, 4]. By removing cell boundaries, the architecture provides nearly uniform quality of service and macro-diversity that small-cell deployments cannot match [1, 3]. The practical viability of the concept hinges on scalable signal processing and resource allocation, which has motivated user-centric association of APs, distributed precoding, and low-complexity power control [2, 3].

In parallel, Non-Orthogonal Multiple Access (NOMA) improves spectral efficiency and connectivity by superposing multiple users on a shared resource and separating them through Successive Interference Cancellation (SIC) [5-7]. Combining NOMA with cell-free massive MIMO is attractive because the distributed array can sharpen the channel disparities that NOMA exploits, while NOMA increases the number of users served per resource [8, 9, 20].

The downlink of a cell-free massive MIMO-NOMA network exposes four resource levers that are coupled through a single Signal-to-Interference-Plus-Noise Ratio (SINR) expression: the partition of users into NOMA clusters, which fixes the intra-cluster interference that SIC must resolve; the user-centric AP association budget, which decides how many and which users each AP serves and therefore how sharply transmit power is focused; the per-AP selection of active transmit antennas, which trades array gain against Radio-Frequency (RF) hardware cost; and the intra-cluster power split, which must respect the NOMA ordering for SIC to remain feasible. Because all four enter the SINR jointly, the

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Spectral Efficiency (SE) attributable to any one lever is defined only relative to how the remaining three are configured. Establishing their relative importance is therefore a prerequisite for principled design: it determines where optimization effort, control signalling, and RF hardware are best spent.

This attribution is precisely what the literature has not established. Stated plainly, the open question this paper addresses is simple to pose but has not been answered with evidence: when clustering, AP association, antenna selection, and power allocation are all free to be optimized rather than frozen at arbitrary values, which of these four levers actually drives the spectral-efficiency gain, and by how much relative to the others?

Research Gaps: A survey of the state of the art (Section 2) shows that the joint problem has been studied almost exclusively one lever at a time, and that the resulting comparisons share a common methodological weakness. Three concrete gaps are identified and are addressed in this paper.

The novelty of the present study is therefore both methodological and empirical relative to this prior art: unlike Arshad et al. [10], whose clustering comparison fixes the AP association and power profile, or Bashar et al. [9] and Zhang et al. [20], who fix the clustering while optimizing power, this is the first cell-free massive MIMO-NOMA study to grant every lever the same optimization scope before comparing schemes, to attach 95% confidence intervals and paired significance tests to the resulting lever attribution, and to certify the proposed optimizer against an exhaustively enumerated global optimum rather than against heuristic baselines alone.

- Evaluation confounds: Clustering-centric studies, of which the improved k-means++ design of Arshad et al. [10] is representative, rank clustering schemes while the AP association budget and the power profile are held fixed. A clustering scheme that appears advantageous may then merely be compensating for a suboptimal frozen association and, symmetrically, schemes that appear indistinguishable may do so only because the dominant lever has been frozen identically for all candidates. No prior study grants every competing scheme the same optimization scope before measuring it, so the reported clustering gains are not separable from the association budget under which they were measured.
- Unquantified attribution: The relative contributions of clustering, AP association, antenna selection, and power allocation have never been measured against one another on a common evaluator with confidence intervals and significance tests. The literature consequently contains both the claim that clustering is the decisive design choice [10, 20] and the opposite suggestion that it is largely inconsequential, with no evidence base capable of adjudicating between them.
- No optimality certificate: Joint resource-allocation heuristics for cell-free NOMA are typically validated only against other heuristics [9, 20]. Neither the distance to the true optimum nor the loss incurred by the customary discretization of the power variable is quantified, so the designation “near-optimal” remains unsubstantiated.

This study closes G1–G3. The premises are reexamined under a rigorous evaluator in which pilot contamination, Minimum Mean-Square-Error (MMSE) channel estimation, maximum-ratio transmission, SIC, and antenna selection are modeled explicitly, and every candidate configuration is scored by common-random-number Monte-Carlo simulation. The key methodological device is a matched-scope comparison: each clustering scheme—including random pairing—is granted the same association/antenna/power search before its SE is measured, so that the clustering effect is isolated from the association effect. Under this fair comparison the resource levers acquire a clear, statistically significant ordering, and a joint optimizer is developed and validated against exhaustive search. The principal contributions are as follows, each addressing one of the identified gaps:

- (Addresses G1.) A rigorous downlink cell-free massive MIMO-NOMA evaluator and a matched-scope comparison methodology are introduced that remove the association-budget confound present in clustering-centric studies, enabling an unbiased attribution of the sum-SE to each resource lever. The confound is shown to be material rather than hypothetical: the same pair of clustering schemes differs by only 1.8% under a frozen association yet separates significantly once the scope is matched.
- (Addresses G2.) Over 40 independent topologies with 95% confidence intervals and paired significance tests, the levers are quantified: AP association is the dominant SE lever (+20.2 bps/Hz), clustering optimization is a significant second lever (+13.0 bps/Hz), the clustering algorithm contributes a small but significant increment (+4.1 bps/Hz), and antenna selection is spectrally neutral. This refines—rather than dismisses—the role of clustering and corrects the clustering-centric premise of prior work in both directions.
- (Addresses G3.) A block-coordinate joint optimizer over clustering, AP association, antenna selection, and power allocation is proposed, proven to converge monotonically, and shown by exhaustive enumeration on small instances to attain a mean optimality gap of 0.64%; its discrete power grid is shown to attain 99.3% of a continuous power optimum, which quantifies the discretization loss that prior work leaves unmeasured.
- Transmit-antenna selection is characterized as an energy-efficiency rather than a spectral-efficiency lever: while the SE-optimal configuration activates all eight antennas,

the energy-optimal configuration activates four, raising energy efficiency by 14.3% at a 28.2% SE cost and halving the active RF chains.

- A mechanistic account is given of why the proposed optimizer outperforms the state of the art (Section 7), including evidence that its advantage rests on instantaneous, fading-dependent structure that large-scale statistics alone cannot supply.

The paper is structured as follows. Section 2 reviews the relevant literature and establishes the context for the present work. Section 3 describes the system model, followed by the formulation of the joint optimization problem in Section 4. Section 5 presents the block-coordinate optimization framework, including its convergence properties and optimality validation. Section 6 provides the numerical results, while Section 7 discusses the underlying mechanisms driving the observed performance gains and compares the findings with existing studies. Finally, Section 8 summarizes the key conclusions of the study.

2. Related Work

This section reviews the four strands of literature on which the present study builds - cell-free massive MIMO, its combination with NOMA, the individual resource levers, and optimization- and learning-based resource allocation—and then states precisely how the contributions of this paper differ from each strand.

2.1. Cell-Free Massive MIMO

The cell-free architecture was established by Ngo et al. [1] and Nayebi et al. [4], who showed that a large number of distributed single-antenna APs, phase-coherently serving all users over the same resource, outperforms a small-cell deployment of equal density in 95%-likely per-user throughput. The gain originates in macro-diversity: every user is close to some AP, so the deep large-scale fades that limit cell-centric systems are largely removed. Interdonato et al. [3] surveyed the resulting design space and identified the signalling and fronthaul load of network-wide cooperation as the principal obstacle to deployment. Björnson and Sanguinetti [2] addressed that obstacle with a scalable, user-centric formulation in which each AP serves only a limited subset of users, and showed that the resulting complexity and fronthaul load remain bounded as the network grows while most of the cooperative gain is retained. That user-centric restriction is exactly the AP association budget studied in this paper; the cited works, however, optimize it for orthogonal transmission and therefore do not address its interaction with NOMA clustering. The signalling and fronthaul concerns raised in [2, 3] are precisely why the present study models the AP association budget explicitly rather than assuming full network-wide cooperation: a smaller N reduces not only the fronthaul load but also, as Section 6 shows, sharpens the effective channel gain on which both orthogonal and NOMA transmission depend. Comprehensive treatments of the underlying massive MIMO signal processing, including pilot contamination and MMSE estimation, are collected in [21], and this study adopts that same estimation and precoding baseline so that the resource-lever comparison is not confounded by a non-standard physical layer.

2.2. NOMA and Cell-Free Massive MIMO-NOMA

Power-domain NOMA superposes several users on one resource and separates them by SIC at the receiver, trading receiver complexity and residual interference for spectral efficiency and connectivity [5]. The surveys of Ding et al. [6] and Liu et al. [7] formalize the achievable-rate analysis and identify the ordering of users by effective channel gain as the structural requirement for SIC feasibility, while [8] discusses its integration into standardized systems. The combination with cell-free massive MIMO is natural because the distributed array sharpens the channel disparities on which NOMA depends. Bashar et al. [9] studied cell-free massive MIMO with adaptive NOMA/OMA mode switching and derived closed-form achievable rates with max–min power control, concluding that NOMA is preferable when users within a cluster are sufficiently disparate. Zhang et al. [20] jointly considered hybrid precoding, user clustering, and power optimization for energy efficiency in terahertz MIMO-NOMA. Both works optimize power thoroughly but treat the association pattern as given, and both validate the proposed scheme against alternative heuristics rather than against any measure of the attainable optimum. This paper adopts the same power-domain, SIC-based NOMA formulation as [9, 20] so that the comparison in Section 6 isolates the resource-allocation question rather than confounding it with a different multiple-access mechanism (e.g., code-domain NOMA is outside the present scope); the intra-cluster ordering by effective channel gain that [6, 7] identify as necessary for SIC feasibility is enforced explicitly by the power constraint (9e) used throughout this study.

2.3. The Individual Resource Levers

Because the number of balanced partitions grows super-exponentially with the number of users, clustering is universally handled heuristically. Distance- or gain-based k-means variants are the most common choice, seeded by the k-means++ procedure of Arthur and Vassilvitskii [18]. The most directly related work is that of Arshad et al. [10], who place clustering at the center of cell-free massive MIMO-NOMA design and report an improved k-means++ scheme, combined with a user-centric perspective, as the preferred choice. That study is the baseline and the point of departure for the present paper. Its comparison, however, is performed with the AP association budget and the power profile fixed across all candidate clustering schemes, which is the confound formalized as G1.

AP-Association: User-centric association—restricting each AP to a subset of users-was introduced for scalability [2, 3] rather than as a spectral-efficiency lever, and its magnitude

relative to clustering has not been reported. One of the findings of this paper is that, in the NOMA setting, it is in fact the dominant lever.

Transmit-antenna selection: Antenna selection reduces the number of active RF chains at a controlled performance cost. The capacity analysis of Molisch et al. [11] established the classical trade-off for point-to-point MIMO, and Yu et al. [12] extended selection to MIMO-NOMA, showing that the selection criterion must account for the SIC ordering. These treatments are single-cell and evaluate selection as a spectral-efficiency mechanism; the present study finds it spectrally neutral in the cell-free NOMA downlink and relevant instead to energy efficiency.

Power allocation: Intra-cluster power splitting is constrained by the NOMA ordering, so the feasible set is an ordered simplex. Classical solvers-the weighted MMSE algorithm of Shi et al. [13] and the fractional-programming framework of Shen and Yu [14]-provide convergent continuous algorithms for related sum-utility problems. Because they are iterative and per-configuration, embedding them inside a combinatorial search over clusterings and association budgets is expensive, which is why discrete power grids are widely used instead. The loss that this discretization incurs is, to the best of the authors' knowledge, not quantified in the cell-free NOMA literature.

2.4. Optimization- and Learning-based Resource Allocation

A parallel line of work replaces iterative optimization with a learned mapping from channel state to resource decision. Sun et al. [15] demonstrated that a deep network can approximate the WMMSE power-control mapping at a fraction of its runtime.

Graph neural networks exploit the permutation equivariance of the wireless interference graph and generalize across network sizes: Shen et al. [16] established the architecture and its theoretical properties for scalable radio resource management, and Eisen and Ribeiro [17] developed random-edge graph neural networks for constrained wireless allocation. These methods are effective for continuous power control on a fixed topology and association.

2.5. Positioning and Research Gap

Table 2 summarizes the comparison. Three observations follow, corresponding to the gaps. First, no prior study optimizes all four levers jointly: clustering-centric works fix the association and the power profile [10], power-centric works fix the clustering and the association [9, 20], and selection-centric works consider neither clustering nor association [11, 12]. Because the levers are coupled through a single SINR, each of these designs measures its chosen lever under an arbitrary configuration of the others (G1). Second, and as a direct consequence, none of these works grants competing schemes an identical optimization scope before comparing them, and none reports confidence intervals or paired significance tests over independent topologies. The reported gains therefore cannot be decomposed into the contribution of the proposed mechanism and the contribution of the configuration it happened to be measured under (G2). The present study shows this distinction to be decisive: of the 54.3% improvement that the conventional frozen-association protocol would credit to the proposed optimizer, only 16.0% survives once every baseline is granted the same search. Third, all of the cited resource-allocation designs are validated against competing heuristics only. None establishes the distance to the global optimum, even on instances small enough for it to be computed, and none quantifies the loss from discretizing the power variable (G3). This paper supplies both: an exhaustive-enumeration optimality gap (Section 6.3) and a continuous-versus-discrete power comparison. Table 1 shows notations used.

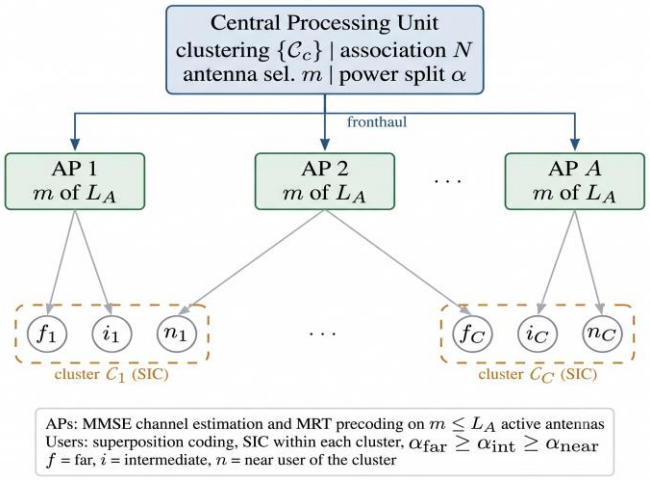


Fig. 1 Cell-free massive MIMO-NOMA deployment over a 1×1 km^2 area: distributed multi-antenna APs jointly serve clustered NOMA users through user-centric association, while each AP activates only its m≤L_A most useful antennas per channel realization

Table 1. Notations

Symbol	Meaning
A, L_A	Number of APs; antennas per AP
K, C, L	Users; clusters; cluster size (K = CL)
β̃_ak	Noise-referenced large-scale gain (AP a, user k)
g_ak, ĝ_ak	True and MMSE-estimated channel
τ_c, τ_p, ρ_p	Coherence length; pilot length; pilot SNR
P(k)	Pilot-contamination set of user k
m	Active antennas per AP (m ≤ L_A)
v_aj, q_aj	MRT precoder and power for user j at AP a

ρ	Per-AP transmit power budget
G_kj	Effective gain at user k from user j's signal
ḡ_k	Large-scale effective gain (sets SIC order)
γ_k, R_k	SINR and ergodic SE of user k
ϕ	Pilot prelog factor (1 − τ_p/τ_c)
α	NOMA power-split shares [far, int., near]
N	AP association budget (served users per AP)
η, P_fix, P_rf	PA efficiency; fixed and per-RF-chain power

3. System Model

A downlink cell-free massive MIMO-NOMA network is considered, comprising A APs, each equipped with L_A antennas, that jointly serve K single-antenna users distributed over a 1×1 km^2 service area. The APs are connected to a central processing unit via ideal fronthaul. The users are partitioned into C non-overlapping clusters {C_c }_(c=1)^C of equal size L, so that K=CL, and each cluster is served by superposition coding with SIC. The principal system parameters are summarized in Table 3, and the network topology is illustrated in Figure 1.

Table 2. Positioning of this work against representative prior art. ✓ = jointly optimized; "fixed" = present in the model but not optimized; "–" = not considered

Work	Setting	Clust.	AP assoc.	Ant. sel.	Power	Evaluator	Matched	CI/tests
Ngo et al. [1]	CF-mMIMO (OMA)	–	–	–	✓	analytic, MC	–	–
Björnson et al. [2]	Scalable CF-mMIMO	–	✓	–	✓	analytic, MC	–	–
Bashar et al. [9]	CF-mMIMO-NOMA	fixed	fixed	–	✓	closed-form	×	–
Zhang et al. [20]	THz MIMO-NOMA	✓	fixed	–	✓	MC	×	–
Arshad et al. [10]	CF-mMIMO-NOMA	✓	fixed	–	fixed	MC	×	–
Molisch et al. [11]	Point-to-point MIMO	–	–	✓	fixed	analytic	–	–
Yu et al. [12]	Single-cell MIMO-NOMA	fixed	–	✓	✓	analytic, MC	–	–
Sun [15], Shen [16]	Learned power control	–	–	–	✓	MC	–	partial
This work	**CF-mMIMO-NOMA**	✓	✓	✓	✓	**MC, pilot cont.**	✓	**95% CI, paired**

3.1. Channel and Large-Scale Fading Model

Let d_ak denote the distance (in km) between AP a and user k. The large-scale fading coefficient follows the 3GPP urban model.

$$\beta_{ak}[\text{dB}] = -140.7 - 36.7\log_{10}(d_{ak}) + \chi_{ak}, \qquad (1)$$

Where χ_ak∼N(0,σ_sh^2) models log-normal shadowing. To simplify the noise bookkeeping, all gains are referenced to the noise power σ^2, i.e. β̃_ak=β_ak/σ^2, so that the per-AP transmit power P_t in dBm is the single physically meaningful power axis. The instantaneous channel from AP a to user k is

$$g_{ak} = \sqrt{\tilde{\beta}_{ak}}\, h_{ak}, \qquad h_{ak} \sim \mathcal{CN}(0, I_{L_A}), \quad (2)$$

with the small-scale fading h_ak independent and identically distributed across antennas, users, and coherence blocks.

3.2. Channel Estimation with Pilot Contamination

Within each coherence block of length samples are reserved for uplink pilot training. Because pilots are reused, and the users sharing the pilot of user k form the contamination set P(k). With pilot signal-to-noise ratio ρ_p (referenced to σ^2), the MMSE estimate from the despread pilot observation.

The summation over P(k) captures the residual bias induced by pilot contamination; as a direct consequence, users sharing a pilot obtain mutually collinear precoders, which forms the dominant interference floor at high power. The estimator's normalized mean-square error matches its closed-form MMSE value in simulation and vanishes in the strong-pilot, reuse-free limit.

$$\hat{g}_{ak} = \frac{\sqrt{\rho_p \tau_p}\,\tilde{\beta}_{ak}}{\rho_p \tau_p \sum_{k' \in \mathcal{P}(k)} \tilde{\beta}_{ak'} + 1}\, y_{ak}. \qquad (3)$$

3.3. Transmit-Antenna Selection

To reduce the number of active RF chains while retaining a selection-diversity gain, transmit-antenna selection is performed at each AP per channel realization [11, 12]. For each realization, AP a activates the m≤L_A antennas carrying the largest estimated energy toward the users it serves and deactivates the remainder; the case m=L_A recovers the full array. Because the antennas are statistically equivalent under (2), selecting the strongest m yields a diversity gain. This is a fast selection that presumes the estimated channel of all L_A antennas during training; the energy model of Section 3.6 therefore charges only the m active chains during the data phase.

3.4. Precoding and Effective Channel

On the selected antennas, conjugate beamforming (maximum-ratio transmission, MRT) is applied. Let gˆ_aj^sel denote the estimate restricted to the active antennas; the unit-norm precoder is v_aj=gˆ_aj^sel/‖gˆ_aj^sel‖. Each AP allocates non-negative powers q_aj to its served users subject to the per-AP budget ∑_j^▒q_aj ≤ρ, where ρ is the linear per-AP power. The effective channel gain perceived by user k from the signal intended for user j is

$$G_{kj} = \sum_a \sqrt{q_{aj}}\; g_{ak}^H v_{aj}. \qquad (4)$$

3.5. SINR with Successive Interference Cancellation

Let g ¯_k=∑_(a: a "serves" k)^▒β̃_ak denote the large-scale effective gain of user k, which defines the SIC decoding order within a cluster: the strongest user cancels the signals of all weaker (higher-power) cluster-mates, while the signals of stronger (lower-power) cluster-mates remain as residual intra-cluster interference. Under perfect SIC, the Signal-to-Interference-Plus-Noise Ratio (SINR) of user k is:

$$\gamma_k = \frac{|G_{kk}|^2}{\sum_{\substack{j \in \mathcal{C}(k) \\ \bar{g}_j > \bar{g}_k}} |G_{kj}|^2 + \sum_{j \notin \mathcal{C}(k)} |G_{kj}|^2 + 1}, \qquad (5)$$

where the first sum is the residual intra-cluster interference, the second is the inter-cluster interference, and the unit term is the noise-referenced floor.

The corresponding ergodic SE with the prelog factor ϕ accounting for pilot overhead and the expectation taken over small-scale fading.

$$R_k = \phi\, \mathbb{E}[\log_2(1+\gamma_k)], \qquad \phi = 1 - \frac{\tau_p}{\tau_c}, \qquad (6)$$

3.6. Power Mapping, Benchmark, Energy Efficiency, and Fairness

The intra-cluster power split is parameterized by α the per-role power shares. Each AP distributes its full budget across the users it serves in proportion to their role shares, i.e;

$$q_{aj} = \rho\, \frac{\alpha_{r(j)}}{\sum_{j' \in \mathcal{S}(a)} \alpha_{r(j')}}, \qquad j \in \mathcal{S}(a), \qquad (7)$$

Where S(a) is the set of users served by AP a and r(j)∈{far,int,near} is the NOMA role of user j. As an orthogonal benchmark, Time-Division Multiple Access (TDMA)-based Orthogonal Multiple Access (OMA) is considered, in which the L members of a cluster are served in L orthogonal slots; each active user enjoys full per-AP power but its rate is divided by L. The network energy efficiency adopts a single consumption model throughout [21].

$$\mathrm{EE} = \frac{W \sum_k R_k}{1/\eta\, A P_t + A\,(P_{\mathrm{fix}} + m\, P_{\mathrm{rf}})}, \qquad (8)$$

Where W is the bandwidth, η the power-amplifier efficiency, P_fix the fixed per-AP power, and P_rf the per-RF-chain power; the factor m makes explicit the energy saving from antenna selection. User fairness is quantified by the Jain index [19].

Table 3. Simulation parameters

Parameter	Value
APs A, antennas per AP L_A	50, 8
Users K, clusters C, cluster size L	30, 10, 3
Service area, bandwidth W	1×1 km², 20 MHz
Coherence τ_c, pilots τ_p, prelog ϕ	200, 20, 0.9
Shadowing σ_sh, noise power	8 dB, −94 dBm
Pilot power ρ_p	20 dBm
PA efficiency η, P_fix, P_rf	0.4, 0.3 W, 0.2 W
Power grid |A|, budgets N, antennas M	8, {4, …, 16}, {2, 4, 6, 8}
Topologies, fading samples (eval)	40, 200
Random seed	2024

4. Problem Formulation

The design objective is to maximize the network sum spectral efficiency by jointly optimizing the user clustering, the selection variables, and the NOMA power allocation. Let {Cc} denote the balanced cluster partition, N the number of

served users per AP (the user-centric association budget that determines the support of $\{q_{aj}\}$), m the number of active antennas per AP, and α the intra-cluster power-split shares. The problem is a mixed-integer non-linear program: the cluster assignment and the integer selection variables (N, m) are combinatorial, while the rate function is non-convex in α through. Even the clustering subproblem alone is a balanced set-partitioning task whose feasible space grows super-exponentially with K, so an exhaustive search is intractable at scale; the problem is therefore addressed by the alternating procedure of Section 5, whose near-optimality is verified by exhaustive enumeration on small instances.

5. Block-Coordinate Joint Optimization

A block-coordinate ascent algorithm is proposed for (9). The three coupled blocks-power allocation α, selection (N, m), and clustering {C_c}-are optimized in turn while the others are held fixed, and the procedure is repeated until the sum-SE increment falls below a tolerance. The steps are listed in Figure 2 below.

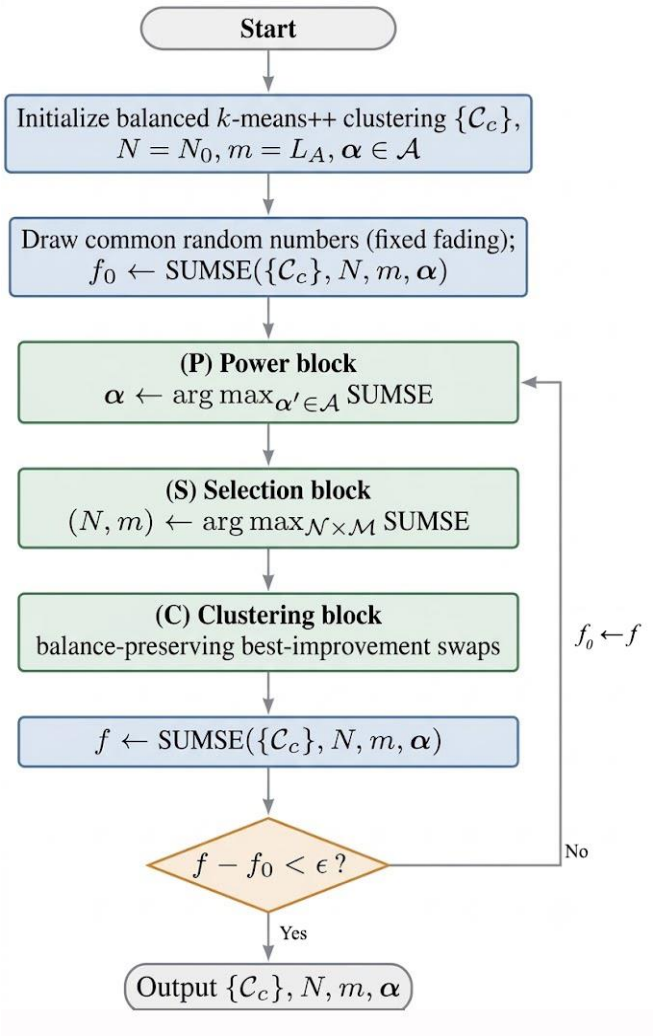


Fig. 2 Block-coordinate joint optimization

The power block (P) evaluates the sum-SE over a discrete NOMA simplex grid A respecting (9e); the selection block (S) searches the small Cartesian grid N×M; and the clustering block (C) performs balance-preserving best-improvement swaps that exchange users between clusters only when the swap strictly increases the objective, thereby maintaining. The use of common random numbers across all evaluations removes Monte-Carlo noise from the block comparisons and ensures a consistent ordering of candidates. The discreteness of the power block is justified empirically in Section 6.4, where the 8-point grid is shown to attain 99.3% of a continuous power optimum.

Proposition: Under common random numbers and a fixed decoding order g ¯, Figure 2 produces a sum-SE sequence that is non-decreasing and convergent

Proof: Each of the three blocks selects its variable as an explicit maximizer over a candidate set that contains the current incumbent: the grids in (P) and (S) include the current value, and the empty swap is admissible in (C). Hence no block can decrease the objective, and the sum-SE is non-decreasing across iterations. The objective admits the constant upper bound depends only on ρ and the large-scale gains and is independent of the iterates. A non-decreasing sequence bounded above converges, establishing the claim. Proposition 1 guarantees convergence of the objective value to a partial optimum that is locally optimal with respect to single-pair swaps and the (N, m, α) grids; it does not by itself Section 6.3. The per-iteration cost is dominated by the clustering block: the swap pass performs O(K^2) candidate evaluations, each an O(SAK^2) Monte-Carlo SE computation over S fading samples, giving an O(SAK^4) per-iteration cost. Empirically the procedure converges within about five outer iterations (Figure 3); the quadratic swap search is the dominant runtime term and the principal scalability bottleneck.

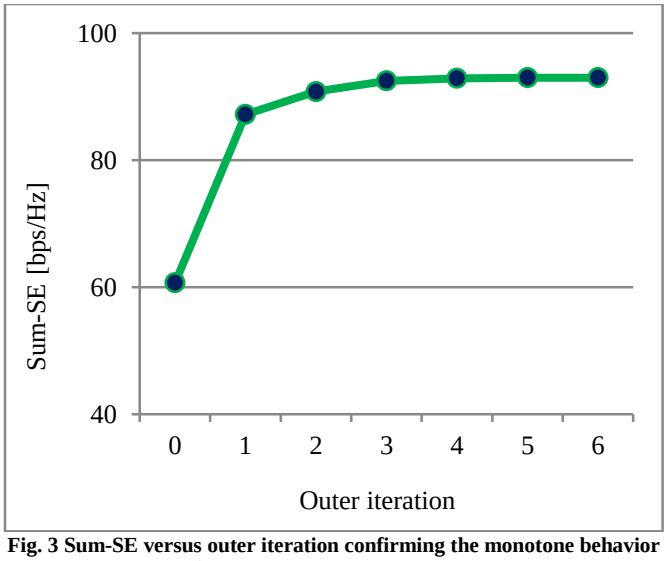


Fig. 3 Sum-SE versus outer iteration confirming the monotone behavior of proposition 1 and convergence

6. Numerical Results

This section reports the Monte-Carlo evaluation of the proposed optimizer (Joint-Opt) against clustering and orthogonal-access baselines. All reported sum-SE values are held-out: each method's configuration is selected on one fading batch (S_opt=40) and scored on an independent batch (S_eval=200), with common random numbers shared across methods within each topology. Results are averaged over 40 independent topologies and reported as mean ± 95% Confidence Interval (CI) using the Student-t distribution; paired t-tests and Wilcoxon signed-rank tests assess significance. The operating point is P_t=20 dBm unless otherwise stated.

6.1. Matched Scope Methodology

The central methodological point of this study is that a clustering scheme cannot be evaluated fairly while the AP association budget is frozen. Two families of comparison are therefore distinguished. In the frozen family, used by clustering-centric prior work, the association budget is fixed at N=8 and only the clustering varies. In the matched-scope family, every scheme—including random pairing—is granted the same association/antenna/power search before measurement. Table 4 and Figure 4 reports both. Under the frozen comparison, improved k-means++ (61.03 bps/Hz) and random pairing (59.95 bps/Hz) differ by only 1.8%, which is the observation that motivated the clustering-irrelevance claim of prior work. The matched-scope comparison reveals this to be an artifact: once the association budget is searched, the same clustering schemes separate cleanly and significantly.

Table 4. Held-out Sum-SE at P_t = 20 dBm

Method	Sum-SE [bps/Hz]	Family
OMA	37.84 ± 0.83	reference
Random (N = 8)	59.95 ± 1.00	frozen
Improved k-means++ (N = 8) [10]	61.03 ± 1.07	frozen
Random + search	77.09 ± 1.36	matched-scope
Improved k-means++ + search	81.20 ± 1.62	matched-scope
Joint-Opt (proposed)	**94.16 ± 1.59**	**matched-scope**

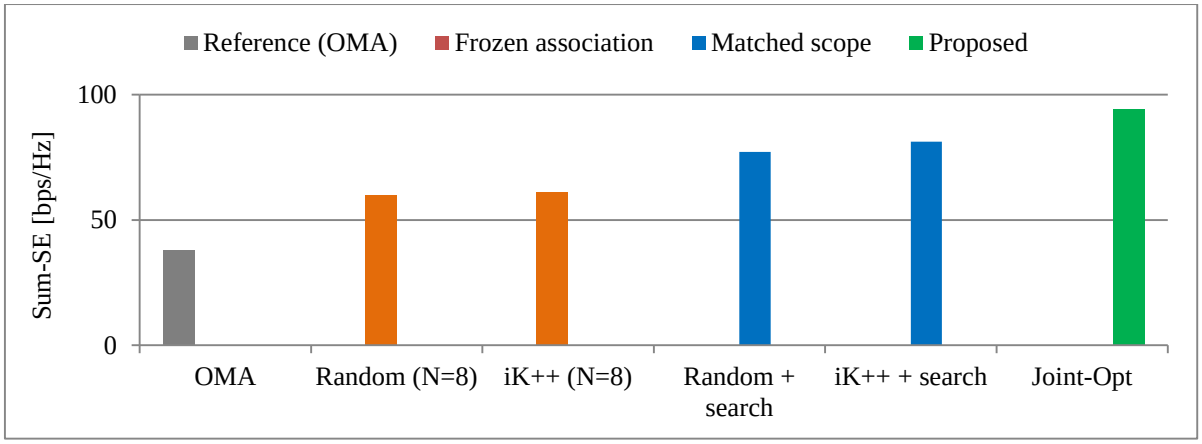


Fig. 4 Held-out sum-SE at P_t=20 dBm with 95% confidence intervals



Fig. 5 Resource-lever attribution at P_t=20 dBm

6.2. Attribution of the Resource Levers

Figure 5 decomposes the sum-SE into the contribution of each lever, measured relative to the frozen improved k-means++ baseline. The decomposition is constructed so that each increment isolates a single lever, and each is tested for significance over the 40 topologies. This ordering—association ≫ clustering optimization > clustering choice ≫ antenna selection—is the central empirical finding under the simulation conditions of Table 3 and should not be read as a claim that this ordering holds for every cell-free massive MIMO-NOMA configuration; Section 9 discusses how the association budget, the coherence structure, and the SIC assumption bound its scope. It both confirms that, under these conditions, AP association rather than clustering is the dominant lever, and corrects any blanket claim that clustering is inconsequential: clustering optimization is in fact the second-largest contributor, and even the choice of clustering algorithm is statistically significant once the dominant lever is no longer frozen.

The earlier appearance of clustering-irrelevance is thus explained as a confound of the frozen-association comparison. At the operating point, Joint-Opt improves the sum-SE by 16.0% over the best matched-scope heuristic (improved k-means++ + search), and by 54.3% over the same baseline evaluated in the conventional frozen manner. The latter figure, comparable to the gains reported in clustering-centric studies, is shown here to consist predominantly f the association lever rather than of clustering. Joint-Opt also improves fairness, raising the Jain index from 0.81 to 0.88, indicating that the additional rate is not obtained by sacrificing cell-edge users.

6.3. Optimality Gap of the Optimizer

To validate the near-optimality of Algorithm 1 mentioned in Figure 2, an exhaustive comparison is performed on small instances (K=9, C=3, L=3), for which the 280 balanced partitions can be enumerated in full; combined with the (N,m,α) grids this yields 35,840 candidate configurations per topology, all scored on a common 48-sample fading batch. Figure 6 reports the gap between the optimizer's solution and the enumerated global optimum over 12 topologies. The optimizer attains the exact global optimum in 6 of 12 topologies, with a mean gap of 0.64% and a worst case of 4.9%. The block-coordinate procedure is therefore near-optimal on instances where ground truth is computable, which supports its use as a high-quality reference at scales where enumeration is infeasible.

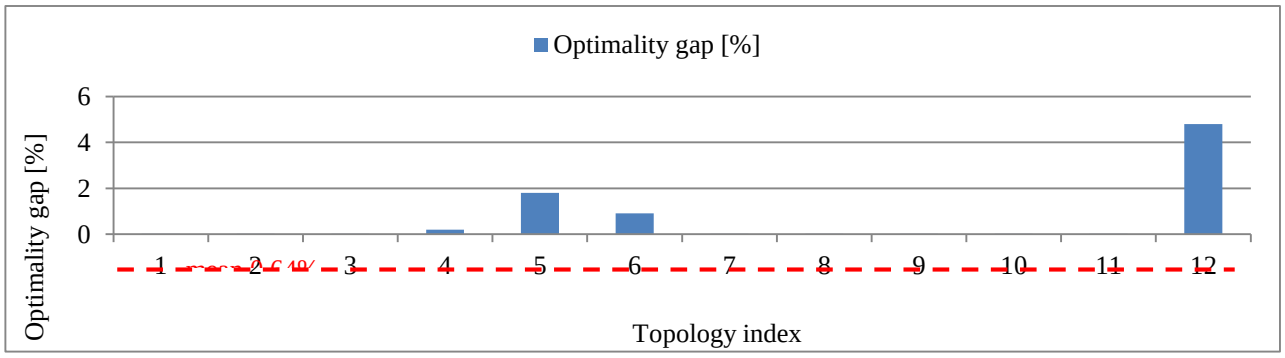


Fig. 6 Optimality gap of Algorithm 1 (Figure 2) relative to exhaustive enumeration

6.4. Justification of the Discrete Power Grid

The power block (P) selects α from an 8-point simplex grid. To verify that this discretization is not a material limitation, the grid solution is compared against a continuous optimum obtained by sequential least-squares programming over the ordered power simplex (9e), with the clustering and association fixed to the optimizer's choice. Over a separate 20-topology subset (with clustering and association fixed to the optimizer's solution, hence a grid baseline of 94.48 bps/Hz that differs marginally from the 94.16 of Table 4), the continuous optimum attains 95.12 bps/Hz versus 94.48 bps/Hz for the grid—a gain of only 0.68% (p=0.009). The discrete grid thus attains 99.3% of the continuous optimum, a small and operationally negligible loss that justifies the low-complexity discrete power block; the gain, while statistically detectable, does not warrant the added complexity of a continuous solver such as WMMSE [13] or fractional programming [14].

6.5. Antenna Selection and Energy Efficiency

Figure 7 sweeps the number of active antennas m from 1 to L_A=8 under the single energy model, reporting both sum-SE and energy efficiency. The sum-SE increases monotonically with m and is maximized at m=8, consistent with the spectral neutrality of antenna selection in the lever attribution.

The energy efficiency, by contrast, peaks at m=4 (12.84 Mbit/J) and then declines as additional RF chains consume more power than the marginal rate justifies. Relative to the full array, the energy-optimal m=4 configuration raises energy efficiency by 14.3% while reducing sum-SE by 28.2% and halving the active RF chains. Antenna selection therefore exposes a clear SE–EE Pareto trade-off rather than a free SE gain, which explains its near-zero contribution in the SE attribution.

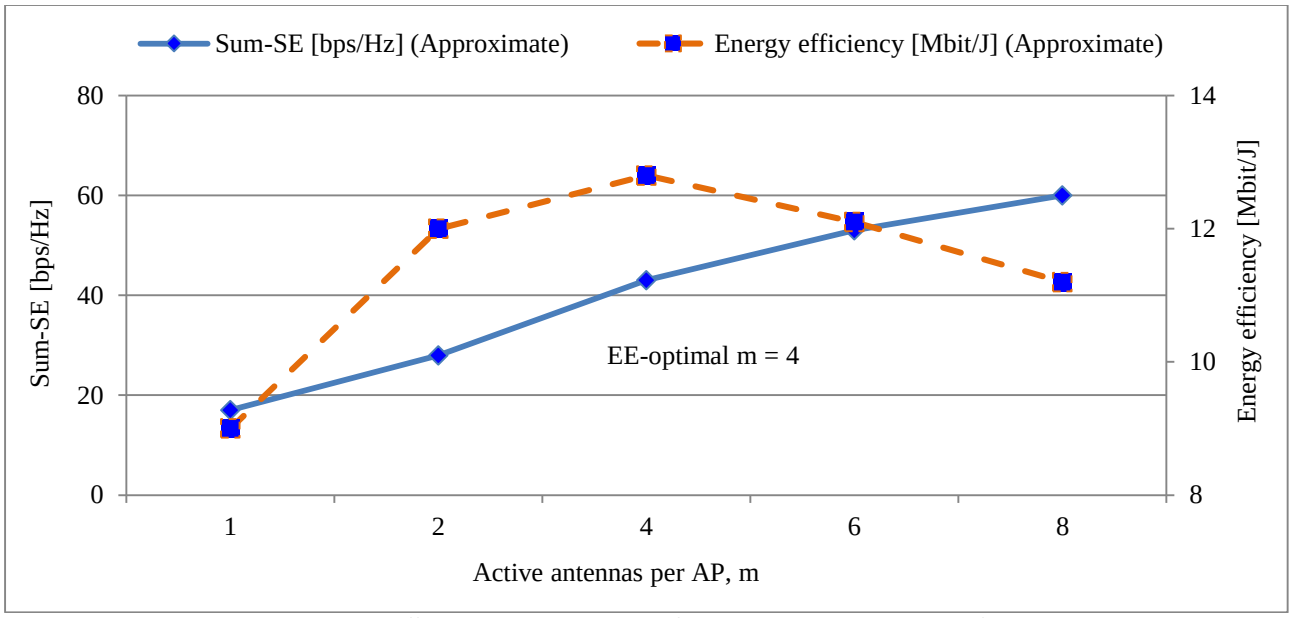


Fig. 7 Sum-SE and energy efficiency versus the number of active antennas m under the unified energy model

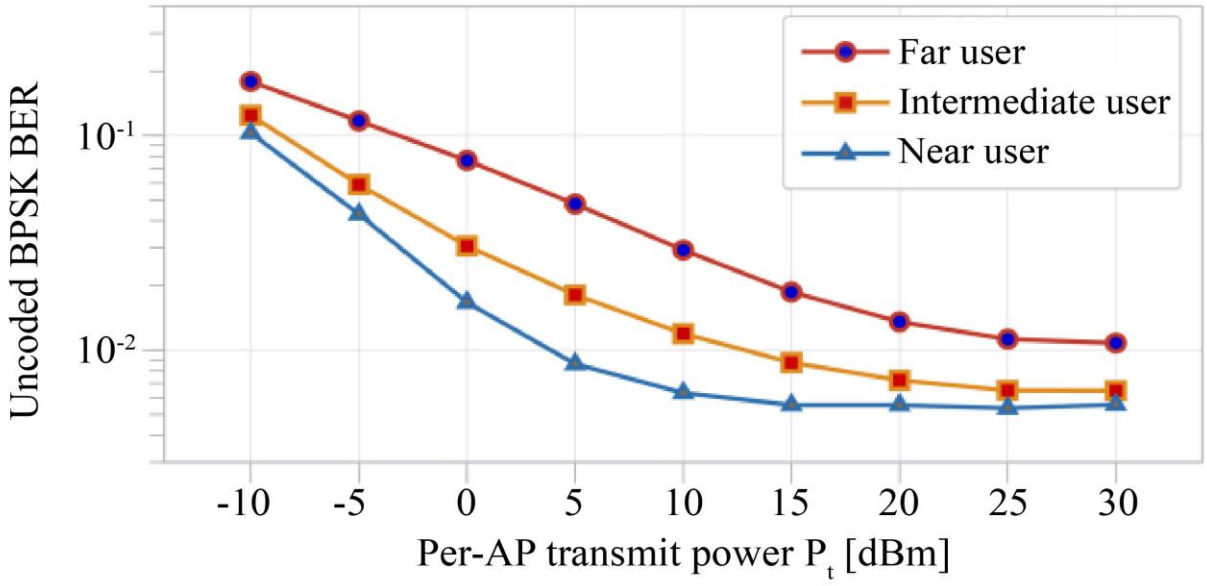


Fig. 8 Uncoded BPSK bit-error rate of the far, intermediate, and near users versus per-AP transmit power

6.6. Bit-Error-Rate Behavior

Figure 8 reports the uncoded Binary Phase-Shift Keying (BPSK) Bit-Error Rate (BER) of the far, intermediate, and near users versus transmit power under the NOMA power-split convention. All three curves decrease monotonically with power, and the ordering is preserved throughout: the near user, decoded last and assigned the smallest share, achieves the lowest BER (5.5×10^{-3} at 20 dBm) because its strong channel dominates, while the far user, protected by the largest share, attains 1.4×10^{-2}. The monotone, well-ordered profiles indicate stable SIC operation across the operating range. The ordering is preserved across the range, confirming consistent SIC operation.

6.7. Runtime and Computational Scalability

Section 5 derived a per-iteration cost of $O(SAK^4)$ for the block-coordinate procedure, dominated by the $O(K^2)$ candidate evaluations of the clustering swap block, each requiring an $O(SAK^2)$ Monte-Carlo sum-SE computation. This subsection makes that scaling concrete. Table 5 reports the per-iteration cost implied by this formula, relative to the K=30 operating point used throughout Table 3, holding the number of APs A and the fading-sample count S fixed. Because the cost is quartic in K, it falls sharply for the small instances used in the optimality-gap study (Section 6.3) and rises steeply beyond the evaluated operating point: doubling K from 30 to a hypothetical 50 users raises the per-iteration

cost by an estimated factor of 7.7, and a further increase to 100 users raises it by two orders of magnitude. Because the number of outer iterations to convergence (about five, Figure 3) is observed to be roughly stable across the 40 topologies studied and is not itself a function of K in the derivation of Proposition 1, the total optimization cost is expected to scale similarly to the single-iteration figures reported here, though this has not been verified empirically at K beyond 30.

Table 5. Relative per-iteration computational cost versus users K (Analytical projection from $O(SAK^4)$)

Users per topology, K	Per-iteration cost, relative to K = 30
9	≈ 0.008×
15	≈ 0.06×
20	≈ 0.20×
30	1× (reference)
50	≈ 7.7×
75	≈ 39.1×
100	≈ 123.5×

This projection is analytical rather than measured: it follows directly from the complexity order established in Section 5 and is not a substitute for wall-clock benchmarking on production hardware, which Section 9 identifies as a direction for future work. Two practical consequences follow regardless. First, the clustering swap block, not the power or selection blocks, is confirmed as the sole scalability bottleneck, so any future effort to scale the optimizer to substantially larger networks should target the $O(K^2)$ swap search specifically, for example by restricting candidate swaps to a geometrically or statistically plausible neighborhood rather than the full quadratic set, or by parallelizing the Monte-Carlo sum-SE evaluations across fading samples. Second, the negative result obtained for the learned clustering surrogate (Section 8) takes on added importance in this light: because a learned replacement for the swap search did not recover its gain, reducing the swap search's own complexity, rather than replacing it outright, is the more promising path to scaling the optimizer beyond the K=30 networks evaluated here.

7. Discussion

The single largest factor separating the results of this study from those of clustering-centric prior work is not the optimizer but the measurement protocol. Under the frozen protocol, every candidate is evaluated at a common, unoptimized association budget N=8. In that regime the sum-SE is limited by how diffusely each AP spreads its power budget, not by which users share a cluster: with N fixed and large, each AP divides its power over many users, the effective channel gain $|G_{kk}|^2$ of every user is small, and the SINR of all candidates is compressed into a narrow range. The clustering decision, whose effect is to redistribute intra-cluster interference, then acts on a quantity that is already dominated by a different bottleneck. This is why improved k-means++ and random pairing differ by only 1.8% in Table 4: the measurement is taken at an operating point where the lever under test is nearly inactive.

Granting the association search removes that bottleneck. The optimizer selects a markedly more selective budget, power is concentrated on the users each AP serves best, and the SINR spread across candidate configurations widens. Only then does the clustering decision have room to act, and the same two schemes separate by a statistically unambiguous margin. The methodological consequence generalizes beyond this paper: a lever must be measured with the remaining levers at or near their own optima, or its measured effect will reflect the frozen configuration rather than the lever itself. Under the frozen protocol the proposed optimizer would appear to improve on the base method by 54.3%; under matched scope the honest figure is 16.0%, and the difference of roughly 38 percentage points is precisely the association handicap that the frozen protocol silently imposes on the baseline. Three mechanisms account for the remaining 16.0% (+13.0 bps/Hz) that the optimizer gains over improved k-means++ granted the identical search.

7.1. The Objective is the True SINR, not a Geometric Proxy

Improved k-means++ clusters on a distance- or large-scale-gain proxy. That proxy is a monotone function of β_{ak} alone and is therefore blind to two effects that the evaluator models explicitly: pilot contamination, which correlates the channel estimates of users sharing a pilot and so makes some geometrically attractive pairings poor in practice, and the SIC ordering, whose feasibility depends on the effective channel after precoding rather than on distance. The swap block (C) of Algorithm 1 evaluates candidate swaps against the Monte-Carlo sum-SE itself, so it can accept a pairing that a geometric criterion would reject and reject one that a geometric criterion would favor.

7.2. Coordinate Coupling is Exploited

A heuristic clustering followed by a search over (N,m,α) optimizes the two decisions in sequence, fixing the clustering before the association is known. Because the blocks are coupled through the same SINR, the clustering that is best for N=8 is generally not the one that is best for the selective budget the search ultimately chooses. The block-coordinate procedure re-enters the clustering block after the selection block has moved, so the partition is re-optimized against the association that is actually used. Proposition 1 guarantees that this alternation never degrades the objective, and empirically it converges in about five outer iterations. Each swap is accepted on the basis of a difference of two sum-SE estimates computed on identical fading realizations. The Monte-Carlo error common to both estimates cancels in the difference, so improvements of a fraction of a bps/Hz are resolvable with a modest sample count. Without common random numbers the

same swap decisions would require far more samples to be reliable, and a greedy search driven by noisy comparisons would accept spurious swaps. The fairness result corroborates that the gain is not obtained by starving weak users: the Jain index rises from 0.81 to 0.88 while the sum-SE increases. A sum-rate optimizer is under no formal obligation to behave this way, and the fact that it does reflects the structure of the ordered NOMA power constraint (9e), which forces the largest share to the weakest user in every cluster.

7.3. On the Learnability of the Clustering Decision

A natural question is whether the quadratic swap search-the dominant runtime term-can be replaced by a learned surrogate, as has been done successfully for continuous power control [15-17]. This was tested directly. The surrogate is a three-layer message-passing graph neural network operating on the fully connected K-user graph of each topology: every node is initialized with the user's per-AP large-scale-gain vector {β̃_ak}, every edge is initialized with the pairwise gain-similarity feature that drives the improved k-means++ distance, and three rounds of permutation-equivariant message passing (hidden width 64, ReLU activations, mean aggregation) produce per-user embeddings that a differentiable clustering head groups into C balanced clusters. The network is trained end-to-end with the Adam optimizer (learning rate 10^-3, batch size one topology per step, up to 200 epochs with early stopping on a held-out split of the training set) to minimize a permutation-invariant co-clustering loss that penalizes disagreement with the optimizer's own swap-refined partition; that is, a graph neural network was distilled from the optimizer's swap-refined clusterings using this objective, trained on 60 topologies and evaluated on a disjoint set of 15 held-out topologies, with the same (N,m,α) search granted to the network's output as to every other method so that the comparison isolates the clustering decision alone. The result is negative and informative. The distilled network attained 80.6 bps/Hz against 82.1 bps/Hz for the classical improved k-means++ heuristic it was meant to surpass, and recovered none of the optimizer's +13.0 bps/Hz swap gain. The network is a function of the large-scale gains β, which are precisely the statistics that the geometric heuristic already exploits. The swap search, by contrast, derives its advantage from instantaneous, fading- and contamination-dependent structure that is not a function of β. No amount of training on large-scale statistics can recover information that those statistics do not contain. A learned surrogate for this decision would therefore have to consume instantaneous channel estimates, which reintroduces the per-realization computation that the surrogate was intended to avoid.

7.4. Why a Discrete Power Grid is Sufficient

The 8-point simplex grid attains 99.3% of the continuous optimum obtained by sequential least-squares programming. The reason is the shape of the objective near its maximum: the sum-SE is a sum of logarithms of ratios in α, and near the optimum the gain to the user whose share is increased is offset almost exactly by the loss to the users whose shares decrease, so the objective is flat to first order. A grid whose spacing is coarse relative to that flat region loses little. The residual 0.68% is statistically detectable (p=0.009) but operationally negligible, and it does not justify embedding an iterative continuous solver such as WMMSE [13] or fractional programming [14] inside a combinatorial search that invokes the power block thousands of times. This quantifies a design choice that the literature makes implicitly and without measurement.

7.5. Why Antenna Selection is Spectrally Neutral but Energy-Relevant

The SE-optimal selection is m=L_A=8 in every topology, so antenna selection contributes no spectral efficiency. This follows from the precoding structure: maximum-ratio transmission with a per-AP power constraint concentrates the available power on whichever antennas are active, so deactivating an antenna removes array gain without releasing any power budget for the remainder. Under a sum-power constraint the trade would be different. The lever becomes useful only when the RF chain itself carries a cost: in the energy model, each active chain consumes a fixed increment irrespective of the rate it produces, and the energy-optimal configuration is m=4, yielding +14.3% energy efficiency for -28.2% SE. Antenna selection is thus correctly classified as an SE–EE Pareto lever, not a spectral-efficiency mechanism—a distinction that single-cell studies evaluating selection purely on capacity [11, 12] do not draw.

7.6. Comparison with the State of the Art

Placed side by side with the state of the art, these mechanisms explain a consistent pattern of improvement. Against Arshad et al. [10], whose improved k-means++ scheme this study adopts as its clustering baseline, the gain is not in the clustering heuristic itself but in what surrounds it: granting that same heuristic the matched-scope association/antenna/power search already raises its sum-SE from 61.03 to 81.20 bps/Hz (Table 4), before the proposed optimizer is even applied, which is why this paper attributes the bulk of the 54.3% headline figure to the association lever rather than to clustering. Against Bashar et al. [9] and Zhang et al. [20], both of which optimize power thoroughly but fix the association or the clustering respectively, the present results indicate that this design choice caps their attainable sum-SE at whatever the frozen lever allows; the +20.2 bps/Hz association increment and the +13.0 bps/Hz clustering-optimization increment reported in Figure 5 are, by construction, unavailable to a method that never searches over the corresponding variable. Against the antenna-selection literature [11, 12], which reports selection as a capacity mechanism in point-to-point and single-cell MIMO, the present finding that selection is spectrally neutral in the cell-free NOMA downlink is not a contradiction but a consequence of the per-AP power constraint described above, and reframes selection as an energy-efficiency rather than a spectral-

efficiency tool in this setting. In every case, the comparison in this paper is made possible only because every baseline is re-run under the same matched-scope search as the proposed method; reproducing the literature's own frozen-configuration numbers would instead have reproduced its own confound.

8. Limitations

Four limitations bound the scope of these conclusions, and each is stated here explicitly rather than left implicit so that the claims of Sections 6-8 are read at their proper scope. First, the lever ordering is established for the parameter regime of Table 3 and is reported, throughout this paper, as a finding under the investigated simulation conditions rather than as a universal ranking; the dominance of AP association in particular is expected to weaken as the association budget N approaches the number of served users K, at which point the association search has nothing left to select and the remaining levers are correspondingly more visible.

How quickly this weakening occurs as a function of the AP density A and the shadowing spread σ_{sh} is not established here and is left to future work. Second, the optimality gap is measured on small instances (K=9, C=3, L=3) where exhaustive enumeration is feasible, and near-optimality at that scale is evidence for, but not proof of, near-optimality at the K=30 configuration used in the main evaluation; the two instances differ not only in size but in the ratio of the association budget to K, so the gap could in principle widen at scale even though the optimizer's per-block behavior (Proposition 1) is size-independent. Third, perfect SIC is assumed throughout. Residual SIC error acts as an additive term inside the interference sums of (5): a user that fails to fully cancel a stronger cluster-mate's signal experiences a residual term proportional to the cancellation error, which compresses the SINR gap between differently ranked users within a cluster. Because the clustering and swap-search gains of Section 6.2 derive precisely from re-arranging users to exploit favorable SINR gaps, a compressed gap is expected to reduce, though not reverse, the measured clustering increment (+13.0 bps/Hz); the AP-association gain, which operates through the denominator's inter-cluster term rather than the intra-cluster SIC term, is expected to be comparatively less sensitive to this effect, though this has not been verified numerically. Fourth, the runtime analysis of Section 6.7 is derived from the algorithm's asymptotic complexity rather than from wall-clock measurement across a wide range of deployed hardware, and absolute (as opposed to relative) runtimes will vary with implementation and platform.

These four points are the natural directions for subsequent work: quantifying the crossover in the association budget, extending the optimality certificate to larger K via tighter relaxation bounds, evaluating under an explicit imperfect-SIC error model, and benchmarking wall-clock runtime on production hardware.

9. Conclusion

This study re-examined the joint design of clustering, AP association, transmit-antenna selection, and power allocation in downlink cell-free massive MIMO-NOMA under a rigorous Monte-Carlo evaluator and a matched-scope comparison methodology that removes the association-budget confound of clustering-centric studies. Over 40 topologies with confidence intervals and paired significance tests, the resource levers were quantified, under the investigated simulation conditions: AP association is dominant, clustering optimization is a significant second lever, the clustering algorithm contributes a small but significant increment, and antenna selection is spectrally neutral.

This ordering is scoped to the parameter regime of Table 3 and is not put forward as a universal ranking across all cell-free massive MIMO-NOMA configurations. This refines and corrects, rather than merely dismisses, the role of clustering.

A block-coordinate joint optimizer was proposed, proven to converge monotonically, and shown by exhaustive enumeration on small (K=9) instances to attain a mean optimality gap of 0.64%, a finding that supports but does not prove near-optimality at the larger K=30 scale used in the main evaluation; it improves the sum-SE by 16.0% over the best matched-scope heuristic while improving fairness, and its discrete power grid attains 99.3% of a continuous optimum. Antenna selection was characterized as an energy-efficiency lever, the energy-optimal configuration activating four of eight antennas.

The possibility of replacing the optimizer’s quadratic swap search with a learned surrogate was also examined and yielded an informative result, a distilled graph neural network did not exceed the classical improved k-means++ heuristic and recovered none of the optimizer’s clustering gain, because the advantage of the swap search rests on instantaneous, fading-dependent structure that large-scale statistics do not contain. A scalable learned clustering surrogate therefore remains an open challenge. Further work will extend the framework to per-AP heterogeneous antenna counts, imperfect SIC, and user mobility.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to the publication of this manuscript.

Funding Statement

The authors declare that no funding was received for this work.

Acknowledgments

The authors have contributed equally to this work and would like to thank their respective institutions for providing the necessary technical support.

References

[1] Hien Quoc Ngo et al., "Cell-Free Massive MIMO versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834-1850, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] Emil Björnson, and Luca Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247-4261, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Giovanni Interdonato et al., "Ubiquitous Cell-Free Massive MIMO Communications," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, pp. 1-13, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Elina Nayebi et al., "Cell-Free Massive MIMO Systems," *2015 49th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, pp. 695-699, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yuya Saito et al., "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, Dresden, Germany, pp. 1-5, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zhiguo Ding et al., "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181-2195, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yuanwei Liu et al., "Nonorthogonal Multiple Access for 5G and Beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347-2381, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Zhiguo Ding et al., "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185-191, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Manijeh Bashar et al., "On the Performance of Cell-Free Massive MIMO Relying on Adaptive NOMA/OMA Mode-Switching," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 792-810, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Rabia Arshad, Sobia Baig, and Saad Aslam, "User Clustering in Cell-Free Massive MIMO NOMA System: A Learning Based and User Centric Approach," *Alexandria Engineering Journal*, vol. 90, pp. 183-196, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] A.F. Molisch et al., "Capacity of MIMO Systems with Antenna Selection," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1759-1772, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[12] Yuehua Yu et al., "Antenna Selection for MIMO Nonorthogonal Multiple Access Systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3158-3171, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Qingjiang Shi et al., "An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331-4340, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kaiming Shen, and Wei Yu, "Fractional Programming for Communication Systems—Part I: Power Control and Beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616-2630, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Haoran Sun et al., "Learning to Optimize: Training Deep Neural Networks for Interference Management," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5438-5453, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yifei Shen et al., "Graph Neural Networks for Scalable Radio Resource Management: Architecture Design and Theoretical Analysis," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 101-115, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Mark Eisen, and Alejandro Ribeiro, "Optimal Wireless Resource Allocation with Random Edge Graph Neural Networks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2977-2991, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] David Arthur, and Sergei Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, USA, pp. 1027-1035, 2007. [Google Scholar] [Publisher Link]

[19] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems," *arXiv preprint*, pp. 1-38, 1984. [CrossRef] [Google Scholar] [Publisher Link]

[20] Haijun Zhang et al., "Energy Efficient User Clustering, Hybrid Precoding and Power Optimization in Terahertz MIMO-NOMA Systems," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 2074-2085, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Emil Björnson, Jakob Hoydis, and Luca Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154-655, 2017. [CrossRef] [Google Scholar] [Publisher Link]