Sustainable Cities and Society 148 (2026) 107575

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

Theory-informed and interpretable graph learning for urban commuting flows

Minwei Zhao, Dailuo Zhang, Zhecheng Shi, Cai Wu *

Thrust of Urban Governance and Design, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

ARTICLE INFO

Keywords:
Urban mobility
Spatial interaction
Commuting flows
Graph learning
Scaling laws
Sustainable planning

ABSTRACT

Origin–destination (OD) flow prediction is central to sustainable urban mobility planning and network design. However, contemporary computational approaches increasingly sacrifice theoretical consistency for predictive flexibility. Spatial interaction modelling has long been grounded in gravity theory and empirically observed scaling regularities, yet existing Graph Neural Networks (GNNs) structurally violate these principles by conflating origin–destination roles and ignoring distance–flow scaling laws. We present PIG-GNN (Physics-Informed Gravity Graph Neural Network), which couples a dual-tower origin/destination encoder with an OD interaction graph and embeds soft physics-informed constraints on distance decay, Zipf scaling, and flow distribution. The priors are calibrated from prior literature and empirical analysis of multiple OD systems and are applied as adaptive penalties during training. On England's Home-to-Work (H2W) commuting flows, PIG-GNN outperforms classical gravity and radiation models, tree-based regressors, and recent deep baselines while producing predictions consistent with observed scaling laws. The model preserves a bimodal flow structure that separates intra-urban high-frequency trips from inter-urban gravity-dominated links, enabling regime-aware interpretation relevant to sustainable mobility planning. These results show that combining role-aware graph architecture with physics-informed priors improves accuracy, calibration, and interpretability in OD forecasting, offering a theory-informed and planning-relevant tool for sustainable urban mobility systems.

1. Introduction

Human mobility, expressed as flows between geographic locations, shows stable and reproducible patterns across scales and urban systems (Gonzalez et al., 2008; Noulas et al., 2012). Empirical evidence consistently reports distance-decay and scaling laws, where interaction intensity declines with distance in power-law–like forms (Barbosa et al., 2018; Brockmann et al., 2006; Gonzalez et al., 2008). These regularities reflect fundamental constraints of human activity, transport networks, and urban organization, and recur in commuting, migration, and multipurpose travel systems. In sustainable urban systems, understanding such regularities is increasingly important because mobility flows shape accessibility, infrastructure demand, and the spatial organization of daily activities, thereby influencing planning decisions across urban regions (Altintasi & Kirtiloglu, 2026; Sugar & Kennedy, 2021).

Modelling these flows therefore requires both predictive accuracy and fidelity to macroscopic regularities (Batty, 2013; Fotheringham & O'Kelly, 1989; Oshan, 2021). The gravity model provides the canonical theoretical foundation (Gabaix, 1999; Isard, 1956), positing interaction proportional to origin/destination masses and inversely related to distance, with impedance effects (Fotheringham, 1983; Sen & Smith, 2012). Its interpretability and empirical success across contexts make it a persistent anchor for spatial analysis (Roy & Thill, 2004). Yet fixed functional forms can miss the non-linear, context-dependent interactions in contemporary, polycentric systems (Burger & Meijers, 2012), limiting their usefulness for planning-relevant forecasting in increasingly complex urban mobility systems (Altintasi & Kirtiloglu, 2026).

This limitation has driven the adoption of deep learning (Schabenberger & Gotway, 2017), which can learn flexible functional forms while preserving spatial autocorrelation implied by Tobler's First Law (Tobler, 1970). Graph Neural Networks (GNNs) explicitly encode spatial adjacency, enabling neighbourhood message passing that captures local context effects often ignored by classical models (Berg et al., 2017; Hamilton et al., 2017). As a result, GNNs have become effective tools for geographic applications (Ghanbari et al., 2025; Terroso-Sáenz & Muñoz, 2022), and are increasingly being used in sustainable urban mobility studies for demand prediction and data-driven planning support (Betkier et al., 2025; Wang et al., 2023).

* Corresponding author.
E-mail address: caiwu@hkust-gz.edu.cn (C. Wu).

https://doi.org/10.1016/j.scs.2026.107575
Received 21 March 2026; Received in revised form 19 May 2026; Accepted 6 June 2026
Available online 13 June 2026

However, current GNN-based spatial interaction models exhibit two important theory gaps that also limit their usefulness for sustainable urban mobility planning. First, they often ignore the production–attraction asymmetry central to spatial interaction theory (Biscaia & Mota, 2013), collapsing the generative role of origins and the attractive role of destinations, which weakens interpretability and theoretical coherence (Fotheringham, 1983; Sen & Smith, 2012). Second, purely data-driven training lacks constraints that enforce established scaling laws such as distance decay (Batty, 2008; Gonzalez et al., 2008; Zipf, 1946). This can yield accurate fits that nevertheless violate macro-level geographic principles, generalize poorly, and provide weaker support for scenario-based urban mobility analysis (Xu & Gu, 2025).

These limitations motivate the need for neural architectures that leverage the spatial modelling power of GNNs while remaining consistent with geographic theory and more useful for interpretable, planning-oriented analysis in sustainable urban systems.

To address this, we propose the Physics-Informed Gravity Graph Neural Network (PIG-GNN), which couples theory-aligned structure with physics-informed learning. The framework has two modules. A *Theory-Informed Structural Module* uses a dual-tower architecture that separates "production" and "attraction" encoders, mirroring gravity-model asymmetry (Wilson, 1970; Zipf, 1946), while an OD-level interaction graph captures competition and substitution (Fotheringham, 1983). A *Physics-Informed Learning Module* embeds empirically validated scaling regularities, derived from prior literature and our analysis of eight OD datasets spanning travel scales and purposes (Section 3.2.3), as adaptive soft constraints in the objective (Gupta et al., 2019). The constraints include: (i) a slope prior on the distance-decay exponent $\beta \in [0.2, 3]$; (ii) a Zipf prior with $\alpha \in [0.5, 3]$; and (iii) a distributional prior that keeps predicted flows consistent with theoretical scaling forms. These priors guide learning without suppressing data-driven non-linearities.

The main contributions of this study can be summarized as follows, with an emphasis on theory-informed, interpretable, and planning-relevant urban mobility analysis:

1. We propose PIG-GNN, a theory-consistent GNN framework for spatial interaction modelling that preserves the production–attraction asymmetry via a dual-tower (origin/destination) encoder, and captures OD-level dependence (competition/substitution) through an OD interaction graph.
2. We formulate a physics-informed optimization strategy that embeds mobility scaling laws as adaptive, empirically bounded priors, including constraints on distance decay and Zipf-like scaling, together with a distributional prior that promotes scaling-consistent predictions without suppressing non-linearities.
3. We evaluate PIG-GNN on a real-world commuting OD system, showing improved predictive accuracy and geographic plausibility, with gains in both error-based metrics and scaling-law fidelity, thereby supporting more reliable urban mobility analysis and planning.
4. We further demonstrate an emergent bimodal structure in learned OD patterns, separating gravity-dominated inter-urban links from saturation-prone intra-urban regimes, offering insight into the boundary between macro regularities and urban complexity and informing multi-scale mobility planning.

2. Literature review

2.1. The gravity model: A theoretical cornerstone of spatial interaction

The gravity model is the foundational paradigm for quantifying spatial interaction, positing that flow intensity is proportional to origin/destination masses and inversely related to spatial impedance (Zipf, 1946). A critical distinction is the functional anisotropy between origins and destinations (Haynes et al., 1984). Unlike reciprocal physical forces, spatial interaction is driven by asymmetric mechanisms: the generative capacity of an origin (e.g., residential zones) differs from the attractive potential of a destination (e.g., employment centres) (Barbosa et al., 2018). This separation of production and attraction roles underpins the directed nature of urban flows.

Theoretical rigour was advanced by Wilson's entropy-maximization framework, grounding the heuristic formula in statistical mechanics (Danwakai, 1967). The derivation shows that the most probable flow configuration under constraints assumes the gravity form. Production-constrained, attraction-constrained, and doubly constrained variants further incorporate balancing factors so predicted aggregates match observed boundary conditions (Haynes et al., 1984; Wilson, 2013). Owing to this robustness, the gravity framework and its extensions remain central to contemporary urban analysis (Lv et al., 2021; Ma et al., 2025; Wu et al., 2021), supporting applications from transport planning and epidemic transmission modelling to supply chain logistics (Randrianarisoa & Gillen, 2022; Simini et al., 2012; Zhou et al., 2022).

Despite these strengths, practical use hinges on specifying an impedance function, typically via fixed parametric forms such as power-law or exponential decay (Oshan, 2021; Zhang & Li, 2024). While these forms capture the general friction of distance, their rigidity can limit the model's ability to represent complex, non-linear substitution effects inherent in modern systems (Gu et al., 2024; Simini et al., 2021). Nevertheless, the gravity model's core theoretical insights, particularly the asymmetric production-attraction framework and the fundamental role of distance impedance, remain foundational principles that continue to guide contemporary spatial interaction modelling, providing essential structural guidance for modern data-driven approaches that seek to preserve theoretical interpretability while gaining representational flexibility.

2.2. Geographic scaling laws: Universal constraints on spatial patterns

Urban systems, despite their complexity, exhibit robust regularities known as geographic scaling laws (Batty, 2008; Bettencourt et al., 2007). These laws reveal scale-invariant structures in spatial organization, from macroscopic allometry to the morphology of urban infrastructure networks (Arcaute et al., 2015; Bettencourt, 2013). In spatial interaction, a pervasive regularity is distance decay: interaction intensity decreases with distance and is often well approximated by a power-law form ($P(d) \sim d^{-\beta}$), especially across broad spatial scales (Brockmann et al., 2006; Gonzalez et al., 2008). This matters because exponential decay implies a characteristic interaction scale beyond which flows rapidly vanish, whereas power-law relationships lack a single characteristic scale, suggesting that long-distance connections, though weaker, remain significant. The resulting heavy-tailed distribution reflects the fractal and hierarchical nature of modern transportation networks (Batty, 2013; Chen, 2015), indicating that the regularity is not a statistical artefact but a property of urban systems (Barbosa et al., 2018). Complementing distance-decay scaling, OD flows exhibit Zipf-like rank-frequency distributions, where the rth largest flow scales as $y_r \propto r^{-\alpha}$ (Rong et al., 2024). This pattern reflects hierarchical organization in spatial interactions, with a few dominant flows and a long tail of low-volume links.

Despite the universality of these forms, decay parameters (β) vary across contexts (Boucherie et al., 2025), shaped by infrastructure and interaction types. Inter-urban and interregional flows often show lower friction due to higher-speed transport networks and the aggregation of long-range opportunities (Chen, 2015; Kang et al., 2013; Lambiotte et al., 2008), whereas intra-urban mobility can deviate towards polynomial forms under daily commuting bounds (Šveda & Madajová, 2023). Theoretical advances reconcile this heterogeneity by identifying deeper invariant mechanisms. Power-law patterns emerge from aggregating movements within hierarchical spatial containers, from neighbourhoods to regions, rather than from a single unbounded distribution (Alessandretti et al., 2020). Integrating visiting frequency with travel distance further unifies these patterns under a universal visitation law, revealing a consistent energetic constraint on individual mobility (M Schlapfer et al., 2021). Thus, parameter variation is a signature of the coupling between urban morphology and limits of human activity.

Consequently, the robustness of geographic scaling laws provides a theory-driven criterion for validating spatial models. Unlike purely statistical correlations, these macro-level regularities represent stable equilibria between interaction benefits and distance friction. A geographically valid model must therefore minimize prediction error while reproducing these scaling regimes. This positions scaling laws as ideal constraints for grounding data-driven models and maintaining consistency with real spatial processes.

2.3. The paradigm shift in computational approaches for spatial interaction

The computational evolution of spatial interaction modelling reflects a shift towards balancing representational flexibility with theoretical consistency (Roy & Thill, 2004; Zhang et al., 2019). Machine learning methods, such as random forests, initially mitigated the functional rigidity of classical gravity models by learning complex non-linear relationships. Yet, by treating origin–destination pairs as independent observations (Ou et al., 2019), these methods overlook spatial autocorrelation and heterogeneity, ignoring the local contexts that condition urban flows.

Graph Neural Networks (GNNs) advance this line by explicitly modelling spatial dependence via message passing (Bui et al., 2022), aligning computation with the topology of spatial systems. Still, as primarily data-driven architectures, standard GNNs prioritize fit over domain principles. This leaves room to improve theoretical consistency by integrating governing laws of spatial behaviour directly into the modelling framework.

To address these deficits, knowledge-guided machine learning (KGML) enhances neural networks with scientific principles through multiple integration strategies (Karpatne et al., 2024). Physics-informed neural networks (PINNs), for example, encode physical laws or empirical regularities as loss constraints, ensuring predictions adhere to governing equations (Karniadakis et al., 2021). Complementary approaches embed domain knowledge in network architecture. For spatial interaction modelling, GNNs capture spatial dependence but have yet to fully leverage such knowledge integration. Two aspects remain underexplored.

First, in architecture, standard designs often use symmetric encoders, overlooking the gravity-model asymmetry (Silva & Nelson, 2012). This neglects functional anisotropy, where production and attraction are distinct mechanisms (He et al., 2017). Embedding this distinction as a structural inductive bias is essential for disentangling heterogeneous generative and attractive forces and for aligning learned representations with the functional organization of spatial systems.

Second, in optimization, purely data-driven objectives often ignore geographic scaling laws. Without geo-physical regularization, models optimized solely on error metrics can learn spurious correlations that violate macro-level principles such as distance decay. Integrating these laws as constraints guards against overfitting to local idiosyncrasies and improves generalizability and geographic plausibility across spatial contexts. These gaps highlight a mismatch between computational power and geographical theory that this research seeks to bridge.

Taken together, the literature points to an opportunity at the intersection of spatial interaction theory, geographic scaling laws, and modern representation learning. Classical models offer theoretical structure and physically grounded regularities, while GNNs offer expressive capacity for complex, non-linear spatial dependencies. Yet these strands have evolved largely in parallel. This gap motivates a unified framework that integrates spatial interaction principles and empirically validated scaling laws as inductive biases within flexible deep learning architectures, reconciling interpretability with predictive performance in origin–destination flow modelling.

3. Methodology

In this section, we present the Physics-Informed Gravity Graph Neural Network (PIG-GNN), which integrates spatial interaction theory with neural representation learning. PIG-GNN treats theory as an active design constraint by embedding role asymmetry and scaling consistency, ensuring predictions remain data-adaptive yet theoretically grounded.

The methodological development follows a coherent flow from empirical observation to structural design and algorithmic enforcement. We first synthesize cross-system empirical regularities in origin–destination dynamics from prior literature and our own analysis, establishing stable bounds on distance–decay exponents and consistent distributional and Zipf properties across mobility systems. These empirical priors form the evidential foundation upon which the model is anchored. Building on this basis, we construct a theory-aligned structural representation that differentiates between production and attraction roles through separate encoder pathways, while an interaction mechanism captures competitive and substitutive relationships among OD pairs. Finally, the learning process is guided by physics-informed objectives that softly constrain model behaviour to respect the empirically verified scaling laws established at the outset. Through this formulation, theoretical insight, empirical regularity, and data-driven flexibility converge within a single integrated modelling strategy.

Fig. 1 provides a comprehensive overview of the full modelling workflow, illustrating the integrated process from data preprocessing to model evaluation. The workflow encompasses six key stages: (1) data preprocessing and feature engineering, where raw OD flows and spatial attributes are transformed into normalized inputs suitable for graph-based learning; (2) dual-graph construction, creating both spatial adjacency graphs that encode neighbourhood relationships and OD interaction graphs that capture observed flow patterns; (3) theory-informed dual-tower encoding, where production and attraction features are processed through separate GNN encoders operating on the adjacency graph to generate role-specific spatial embeddings; (4) OD interaction modelling, employing edge-level message passing on the OD graph to capture destination competition and origin substitution effects; (5) physics-informed prediction, integrating learned embeddings with distance features and applying adaptive constraints derived from empirical scaling laws; and (6) model evaluation and analysis, including performance assessment and the identification of bimodal mobility regimes. We begin with the description of the data and the construction of spatial and OD graphs.

3.1. Data and graph construction

3.1.1. Datasets and preprocessing

Our empirical analysis draws on the Home-to-Work (H2W) OD dataset derived from England's Middle Layer Super Output Areas (MSOAs), capturing routine, economically motivated commuting patterns that reflect daily flows between residential origins and employment destinations. The dataset is processed using a unified pipeline to maintain methodological coherence.

Figs. 2 and 3 provide a comprehensive overview of the spatial organization and overall structure of the H2W dataset through multiple complementary visualizations. Together, these visualizations reveal strong directional organization around major employment hubs — particularly London, Manchester, and Birmingham — that reflects the spatial imprint of labour-market hierarchies and urban economic structure. The overall distribution demonstrates the hierarchical nature of mobility networks and the systematic spatial dependencies inherent in employment-residence relationships.

Beyond the flow records themselves, we incorporate socio-spatial descriptors of each MSOA as explanatory attributes that are theoretically linked to the generative and attractive dimensions of spatial interaction. Working-age population (ages 16–64) is used as a proxy

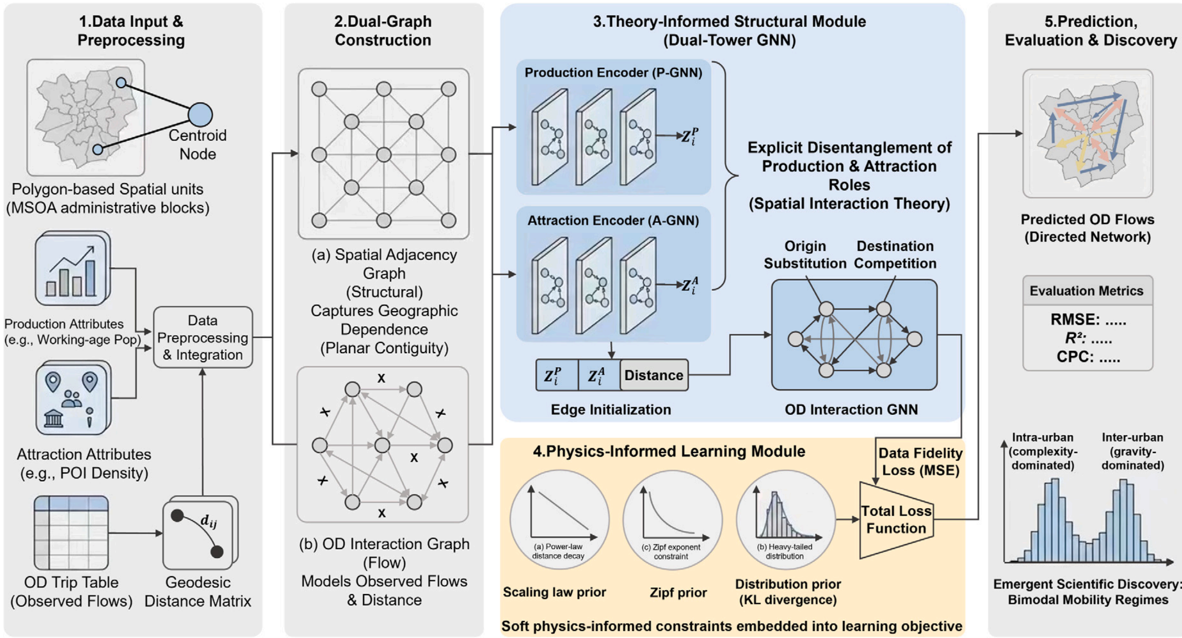

Fig. 1. Technical workflow of the PIG-GNN framework. The schematic illustrates the modelling pipeline from OD data preprocessing and spatial-unit delineation, through dual-graph construction, to the theory-informed dual-tower architecture and physics-informed learning objective.

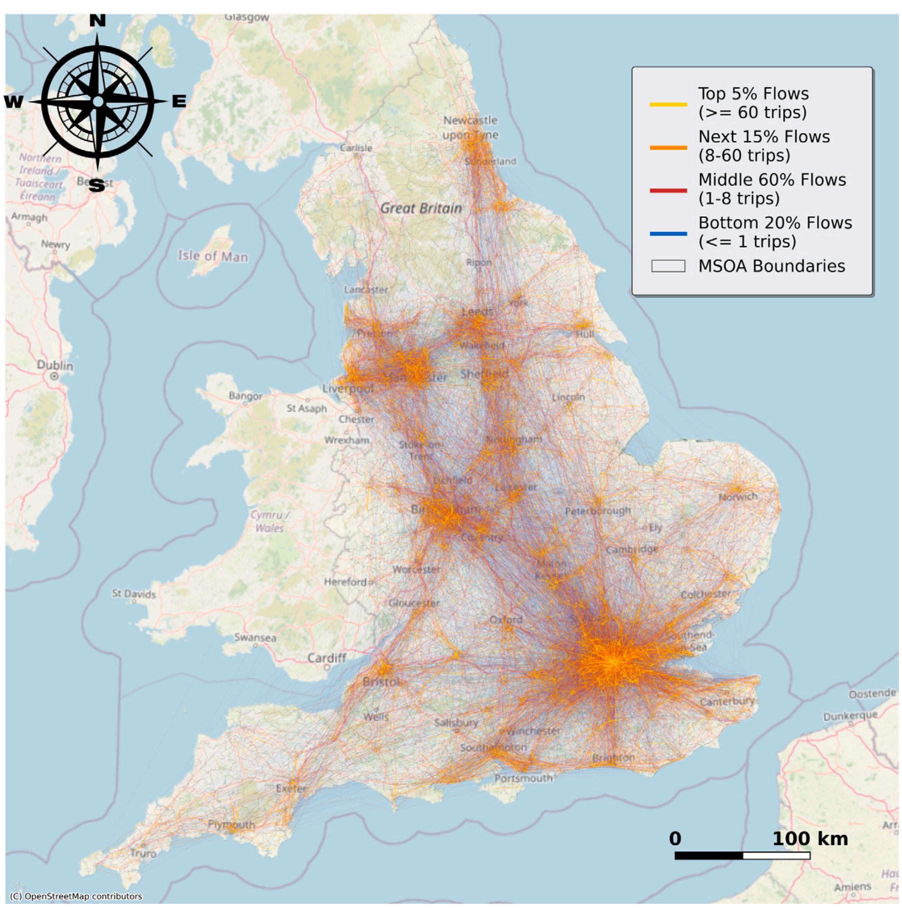

Fig. 2. Overview of the H2W dataset and study area.

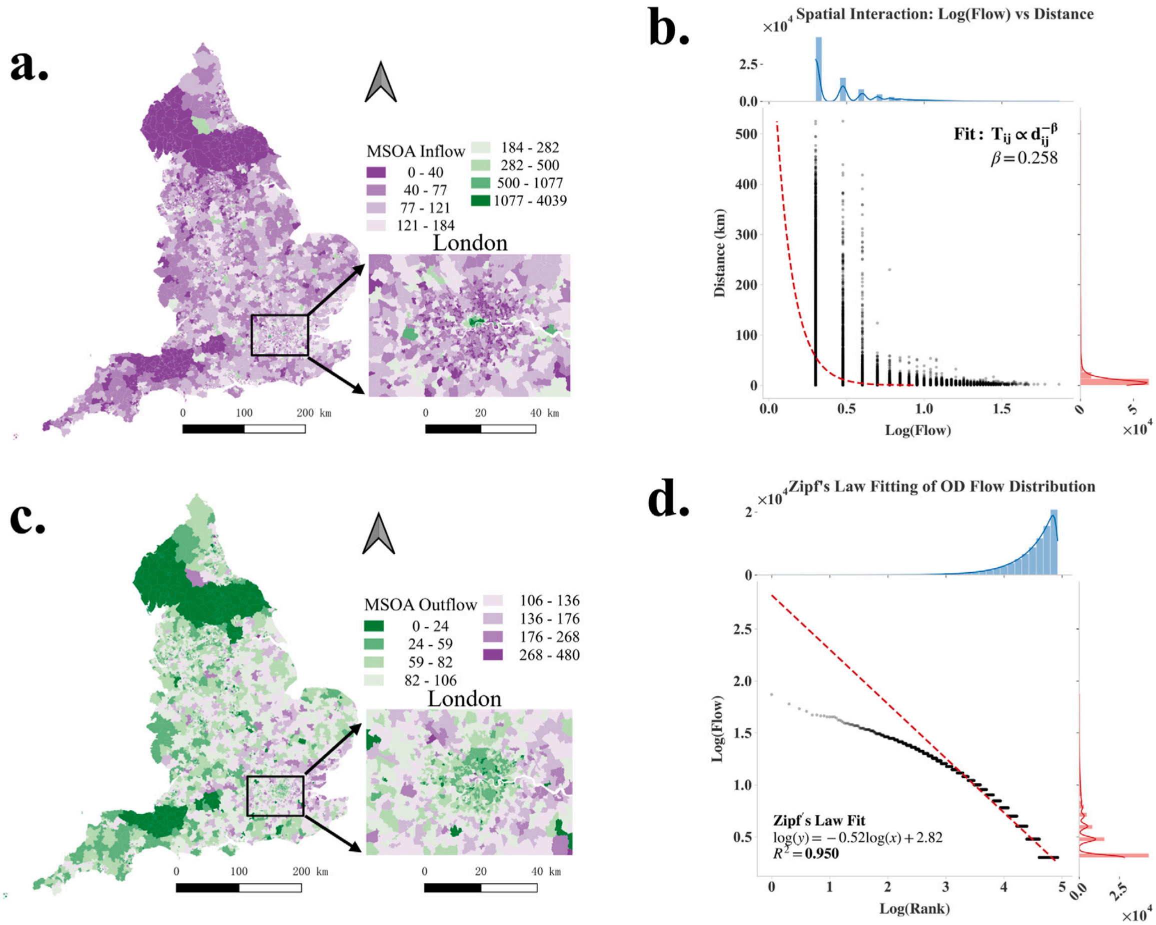

Fig. 3. Flow scaling and spatial structure of the H2W dataset.

for the local capacity to originate movement, consistent with the interpretation of labour-active residents as potential trip generators (Muenz et al., 2007). Likewise, the density of points of interest (POIs) serves as an indicator of localized functional intensity, reflecting concentrations of commercial, service, and amenity destinations that tend to attract inbound mobility. Fig. 4 visualizes the spatial patterns of these two descriptors, revealing distinct geographic distributions that reflect their respective roles in spatial interaction processes. These attributes later form the node-level inputs to the role-specific encoders within our modelling architecture.

We acknowledge that these proxies are coarse. Working-age population imperfectly captures the commuting-relevant labour force, and POI density aggregates heterogeneous land uses without distinguishing employment-oriented destinations from retail, leisure, or residential amenities. Finer-grained employment land-use, workplace-population, or job-sector data would strengthen the physical grounding of the production–attraction inputs. In this study, we retain these publicly reproducible proxies as the baseline input specification, while reporting an enriched-proxy robustness analysis in Appendix A.7.

To reduce skewness in the raw intensity measurements and place all variables onto comparable numeric scales, all count-based quantities are transformed using a $\log(1+x)$ mapping and subsequently standardized. Working-age population and POI density are processed as:

$$x_i^{\mathrm{prod}} = \frac{\log(1+\mathrm{pop}_i) - \mu_{\mathrm{prod}}}{\sigma_{\mathrm{prod}}}, \qquad x_i^{\mathrm{attr}} = \frac{\log(1+\mathrm{poi}_i) - \mu_{\mathrm{attr}}}{\sigma_{\mathrm{attr}}}, \tag{1}$$

where $\mu_{\mathrm{prod}}, \sigma_{\mathrm{prod}}$ and $\mu_{\mathrm{attr}}, \sigma_{\mathrm{attr}}$ are the empirical means and standard deviations computed across MSOAs. Trip counts between MSOA pairs are similarly expressed on a $\log(1+y)$ scale, and this log representation is retained throughout the modelling process. Distances between MSOA centroids are computed geodesically using the Haversine formula and encoded via $\phi(d_{ij}) = \log(1 + d_{ij})$. These transformations ensure that the core quantities used in subsequent modelling—origin potential, attraction potential, distance, and flow—are numerically stable, comparable across orders of magnitude, and compatible with the log-linear relationships assumed in classical spatial interaction formulations.

3.1.2. Graph construction

Having established the underlying variables and their preprocessing, we now formalize the spatial system into a graph representation that supports relational learning. The PIG-GNN framework distinguishes two conceptually different but complementary relational structures: one encoding spatial neighbourhood embeddedness, and another encoding realized movement interactions. These are integrated within a dual-graph architecture that enables the model to leverage both contextual proximity and activity-driven linkage patterns. As shown in Fig. 5, graph construction proceeds from two inputs: MSOA administrative polygons (used to define areal units and centroids) and the OD trip database.

Spatial adjacency graph. The spatial adjacency graph $\mathcal{G}_{\mathrm{adj}} = (\mathcal{V}, \mathcal{E}_{\mathrm{adj}}, \mathbf{W})$ captures spatial contiguity relationships among MSOAs, where $\mathcal{V} = \{1, 2, \ldots, N\}$ denotes the set of N MSOA nodes, and $\mathcal{E}_{\mathrm{adj}}$ represents edges connecting spatially adjacent units. The adjacency matrix $\mathbf{W} \in \{0,1\}^{N \times N}$ is defined as:

$$W_{ij} = \begin{cases} 1 & \text{if MSOAs } i \text{ and } j \text{ share a boundary (planar contiguity)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

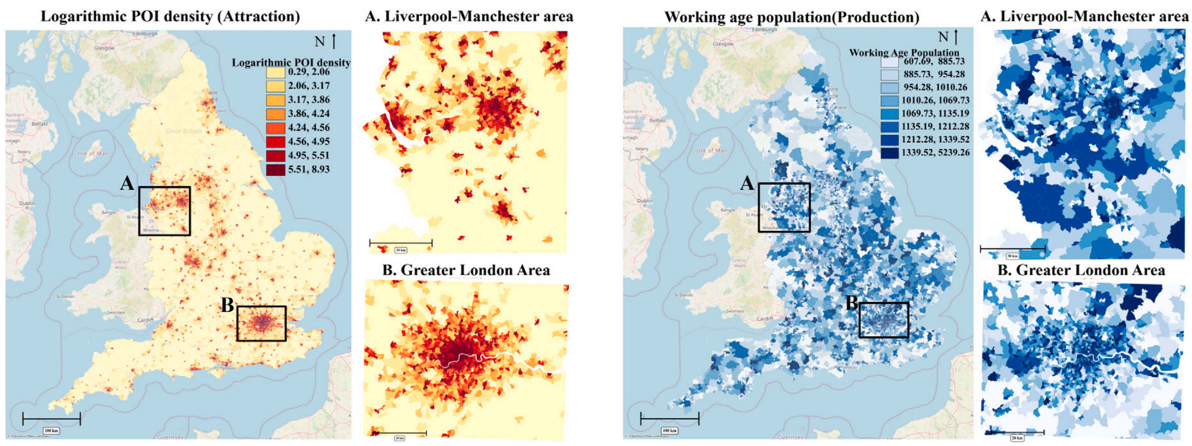

Fig. 4. Spatial distribution of node features across MSOAs: (a) Production feature (working-age population), used as input to the P-GNN (production tower), and (b) Attraction feature (log POI density), used as input to the A-GNN (attraction tower). The visualization reveals distinct spatial patterns: working-age population exhibits broad concentration in major urban regions, while POI density shows sharper clustering around commercial and service centres. Both features are log-transformed and standardized before being passed to the dual-tower encoders, ensuring numerical stability and compatibility with the physics-informed learning framework.

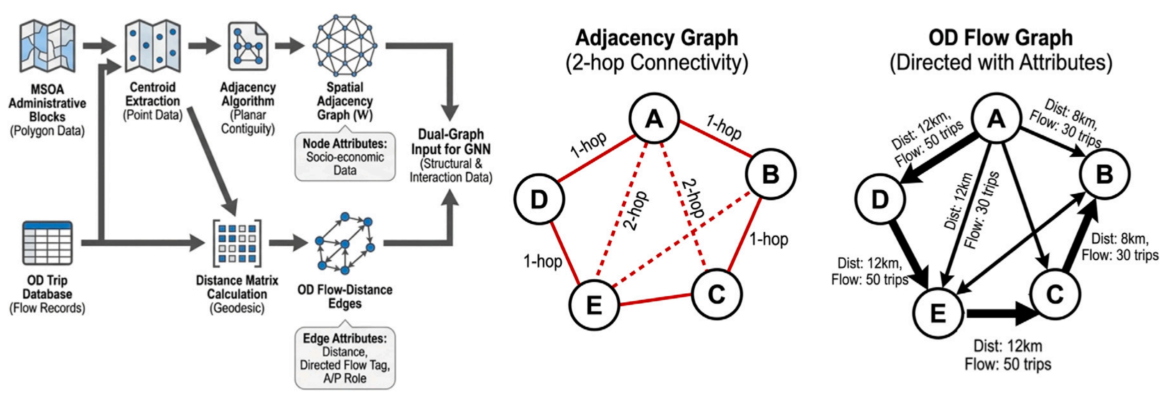

Fig. 5. Dual-graph construction from MSOA contiguity and observed OD flows.

We employ planar contiguity (shared-boundary adjacency) rather than distance-based or k-nearest-neighbour definitions, as it preserves local spatial structure without imposing arbitrary distance thresholds. This choice supports message passing that respects administrative boundaries while empirically preserving neighbourhood effects relevant to both production and attraction mechanisms. Retraining PIG-GNN with alternative spatial adjacency definitions yields almost the same RMSE, MAE, R^2, and CPC as first-order planar contiguity. Specifically, we compare the default shared-boundary graph with 4-nearest-neighbour, 6-nearest-neighbour, 8-nearest-neighbour, and radius-based graphs defined on MSOA centroids. The results in Appendix A.11 and Table A.16 show that all variants remain within a very narrow performance range, supporting the default shared-boundary definition despite variable MSOA footprints.

Each node i in the adjacency graph is associated with feature vectors $\mathbf{x}_i^A$ and $\mathbf{x}_i^P$, representing production and attraction attributes, respectively. These features are passed through separate encoders (A-GNN and P-GNN) operating on the same spatial structure, enabling role-specific message passing while maintaining spatial context.

OD interaction graph. The OD interaction graph $\mathcal{G}_{\mathrm{OD}} = (\mathcal{V}, \mathcal{E}_{\mathrm{OD}}, \mathbf{D}, \mathbf{Y})$ captures observed flow relationships, where $\mathcal{E}_{\mathrm{OD}} \subseteq \mathcal{V} \times \mathcal{V}$ represents directed edges corresponding to non-zero OD pairs, $\mathbf{D} \in \mathbb{R}^{|\mathcal{E}_{\mathrm{OD}}|}$ encodes geodesic distances, and $\mathbf{Y} \in \mathbb{R}^{|\mathcal{E}_{\mathrm{OD}}|}$ represents observed flow volumes. Unlike the adjacency graph, which encodes structural relationships, the OD graph serves as a sparse interaction graph that reduces computational complexity by focusing on observed connections rather than the complete $N \times N$ OD matrix.

For each observed OD pair $(i \to j)$, we construct a directed edge with attributes:

$$\mathbf{e}_{ij} = \begin{bmatrix} \log(1 + y_{ij}) \\ \phi(d_{ij}) \end{bmatrix}, \tag{3}$$

where y_{ij} is the observed trip count and $\phi(d_{ij}) = \log(1 + d_{ij})$ is the log-transformed, standardized distance. Self-loops (edges where $i = j$) are explicitly removed, as they represent intra-MSOA flows that are not the focus of inter-zonal interaction modelling. This removal eliminates 2.11% of edges in the H2W dataset and ensures that the model learns from genuine spatial interactions rather than internal movements.

The OD graph serves multiple purposes: (1) it provides the target structure for flow prediction, with edge attributes serving as both inputs (distance) and outputs (flow); (2) it enables the OD interaction module (Section 3.2.2) to capture competition and substitution effects through edge-level attention; and (3) it reduces the prediction space from $O(N^2)$ to $O(|\mathcal{E}_{\mathrm{OD}}|)$, where $|\mathcal{E}_{\mathrm{OD}}| \ll N^2$ for sparse OD matrices, improving computational efficiency while maintaining predictive accuracy.

The primary training target therefore consists of observed positive OD edges. To examine zero-inflation behaviour without switching to a fully dense $N \times N$ prediction target, Appendix A.8 reports a controlled

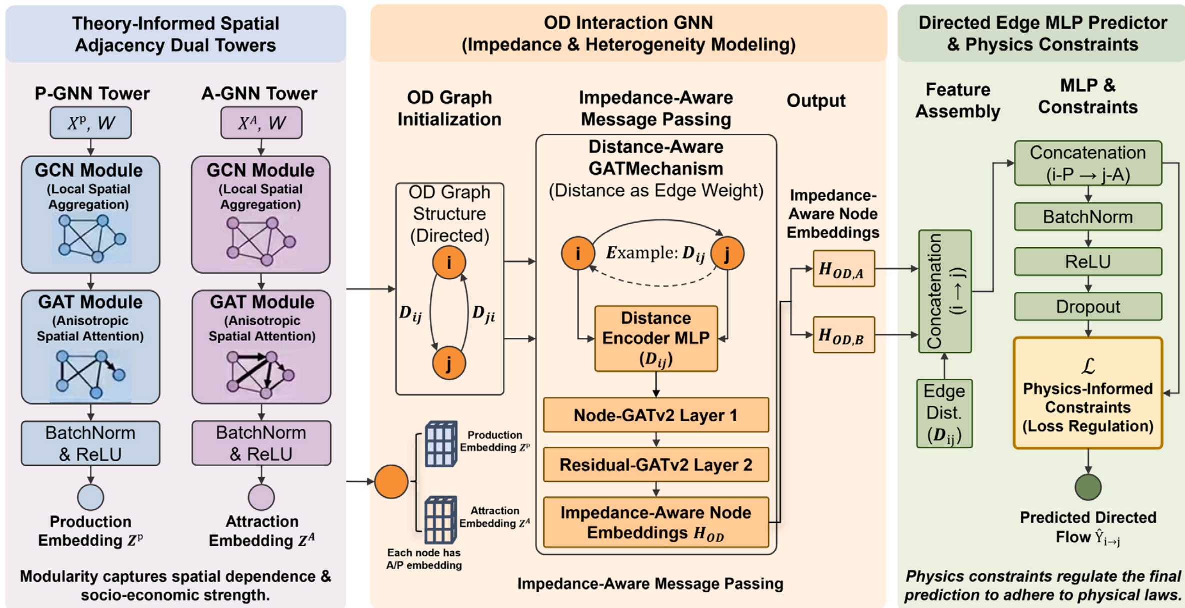

Fig. 6. PIG-GNN architecture with three stages: dual-tower encoding, OD interaction modelling, and physics-informed prediction.

augmentation experiment in which 10% randomly sampled zero-flow pairs are added to the supervision set.

3.2. PIG-GNN architecture

The PIG-GNN architecture is organized into three coordinated stages that transform raw spatial features into flow predictions through role-aware encoding, competitive interaction modelling, and constraint-guided learning. Fig. 6 provides a comprehensive visualization of the complete architecture, illustrating the integrated workflow from raw input features to final flow predictions. The architecture unfolds as follows: (1) a dual-tower encoder that derives role-aware embeddings from the spatial adjacency structure, explicitly separating production and attraction mechanisms; (2) an OD interaction mechanism that models pairwise impedance and competitive relationships between flows through edge-level message passing; and (3) a physics-guided prediction stage that integrates edge embeddings and distance features via an MLP to generate $\hat{y}_{ij}$ under slope, Zipf, and distributional regularization. Rather than layering these components sequentially in an isolated manner, the architecture allows theoretical insights and data-driven representation learning to mutually condition one another, creating a unified framework where each stage informs and constrains the others, as visually represented in the integrated workflow diagram.

3.2.1. Theory-informed dual-tower structure on adjacency graph

Spatial interaction theory, as formalized in gravity models, posits a fundamental asymmetry between origins and destinations: origins produce flows via generative capacity, while destinations attract flows via receptive potential. This is not merely semantic; origins compete for outbound opportunities and destinations for inbound flows, creating asymmetric dynamics that symmetric graph encoders fail to capture.

To preserve this theoretical asymmetry, we design a dual-tower architecture that explicitly disentangles production and attraction roles, as illustrated in the left portion of Fig. 6. The architecture operates on the spatial adjacency graph $\mathcal{G}_{\text{adj}}$ (Section 3.1.2), where each MSOA node is associated with distinct feature vectors: $\mathbf{x}_i^A$ for attraction attributes (POI density) and $\mathbf{x}_i^P$ for production attributes (working-age population). These features are passed through separate, independent encoders that share the same spatial structure but learn role-specific representations:

$$\mathbf{z}_i^A = f_A(\mathbf{x}_i^A, \mathbf{W}), \qquad \mathbf{z}_i^P = f_P(\mathbf{x}_i^P, \mathbf{W}), \tag{4}$$

where f_A and f_P are L-layer GATv2 stacks operating on the adjacency matrix $\mathbf{W}$, and $\mathbf{z}_i^A, \mathbf{z}_i^P \in \mathbb{R}^d$ are d-dimensional embeddings capturing production and attraction capacities, respectively.

The adjacency graph plays two roles. It captures spatial dependence via local message passing and provides a shared spatial scaffold for both towers, ensuring production and attraction embeddings are learned in the same geographic context for meaningful comparison and interaction.

As shown in Fig. 6, the Production Tower (P-GNN) and Attraction Tower (A-GNN) process role-specific features with identical structures on the same adjacency graph $\mathbf{W}$, enabling role-specific message passing while preserving spatial coherence.

Separating production and attraction encoders prevents role leakage. By maintaining distinct parameter spaces, origin-side and destination-side factors propagate through appropriate channels, improving interpretability and reducing over-smoothing—an advantage when distinguishing intra-urban high-flow cores from inter-urban links.

3.2.2. OD graph impedance module and spatial heterogeneity

While the adjacency graph captures spatial dependence through local message passing, it cannot directly represent the *spatial interaction* patterns that govern OD flows. Put differently, adjacency specifies *which locations are close*, but it does not specify *which locations exchange trips* or *the intensity of those exchanges*. To bridge this gap, we design an OD graph impedance module that operates on the sparse interaction graph $\mathcal{G}_{\text{OD}}$ (Section 3.1.2), capturing distance-based impedance and spatial heterogeneity in flow patterns.

The use of a sparse OD graph, rather than a fully connected $N \times N$ matrix, is motivated by both computational and theoretical considerations. Computationally, most OD matrices are highly sparse—in the H2W dataset, only $|\mathcal{E}_{\text{OD}}| = 305,107$ edges exist out of a possible $N^2 = 6,856^2 \approx 47$ million pairs, representing a sparsity of 99.4%. Constructing a fully connected graph would introduce $O(N^2)$ computational complexity and memory requirements, making training infeasible for large-scale systems. Theoretically, the sparse OD graph focuses learning on observed interactions rather than hypothetical connections, reducing the risk of learning spurious patterns from zero-flow pairs that may simply reflect unobserved or non-existent relationships.

The OD graph impedance module captures spatial heterogeneity in flow patterns by learning edge-specific representations that encode distance, competition, and substitution effects, as shown in the centre portion of Fig. 6. For each observed OD pair ($i \to j$), we initialize an edge embedding by concatenating role-specific node embeddings from the dual towers and a distance feature:

$$\mathbf{h}_{ij}^{(0)} = \left[\mathbf{z}_i^P \parallel \mathbf{z}_j^A \parallel \phi(d_{ij})\right], \tag{5}$$

where $\phi(d_{ij}) = \log(1 + d_{ij})$ is the log-transformed, standardized distance, and $\mathbf{z}_i^P, \mathbf{j}_j^A$ are the production embedding of origin i and attraction embedding of destination j learned from the adjacency graph. This initialization encodes the fundamental gravity model relationship: flow depends on origin production, destination attraction, and distance impedance.

We then apply two layers of GATv2 message passing over the OD graph, where each edge attends to its edge-neighbours that share the same origin or destination:

$$\mathbf{h}_{ij}^{(\ell+1)} = g_{\mathrm{GATv2}}\Big(\mathbf{h}_{ij}^{(\ell)},\ \{\mathbf{h}_{ik}^{(\ell)}\}_{k\neq j},\ \{\mathbf{h}_{kj}^{(\ell)}\}_{k\neq i}\Big),\quad \ell = 0, 1. \tag{6}$$

This edge-centric attention mechanism captures two forms of spatial heterogeneity: (1) destination competition, where edges sharing the same destination ($k \to j$) compete for limited attraction capacity, and (2) origin substitution, where edges sharing the same origin ($i \to k$) represent substitutable destination choices. These mechanisms are analogous to the balancing factors in doubly constrained gravity models but are learned adaptively from data rather than imposed through fixed constraints.

3.2.3. Empirical validation of physics-informed priors

A key premise of physics-informed modelling is that constraints should be empirically defensible rather than assumed. Here we consolidate evidence for prior ranges and distributional regularities used in PIG-GNN. We follow a two-stage design: (i) synthesize the mobility literature to establish broad parameter ranges (Table 1), and (ii) validate and refine these ranges on multi-scale OD datasets while checking distributional properties that matter for learning (Table 2). Fig. 7 visualizes the empirical parameter distributions and scale characteristics of the analysed datasets.

Literature synthesis for distance-decay exponent. We review published scaling-law studies and compile reported distance-decay exponents (β) across mobility systems (Table 1). The exponent is estimated by regressing flow intensity on logarithmic distance. Reported values range from about 0.80 to 2.71, with most between 1.0 and 2.0. The variation is scale dependent: intra-city trips tend to show lower decay ($\beta \sim 0.2$–1.5), while interregional and international flows are steeper ($\beta \sim 1.5$–2.7). This synthesis supports a broad yet defensible prior range $\beta \in [0.2, 3.0]$.

Empirical analysis of multi-scale OD datasets. Literature ranges alone do not capture distributional properties critical for learning. We therefore analyse multiple OD datasets spanning commuting, urban mobility, metropolitan systems, and long-range migration (Table 2). For each dataset, we extract four signatures: (1) distance-decay exponent β, (2) Zipf exponent α, (3) KL divergence D_{KL} between observed distributions and theoretical forms implied by β and α, and (4) monotonicity strength of distance decay. Median and mean trip distances (km) summarize scale, from short-range urban trips to interregional migration.

The results shown in Fig. 7 are coherent but scale dependent. Distance-decay exponents range from 0.258 (H2W) to 2.503 (Baidu Migration), and Zipf exponents from 0.539 to 2.819, while monotonicity remains generally high (0.600–0.909). KL divergence values are moderate across scales, indicating consistency with theoretical scaling forms alongside meaningful, dataset-specific deviations. These observations refine the literature ranges into actionable intervals for model constraints.

Distributional consistency constraint. Beyond interval constraints on β and α, we use a distributional consistency constraint based on the Kullback–Leibler divergence D_{KL}. During training we compute $D_{\mathrm{KL}}(\hat{p} \parallel p_{\mathrm{theory}}(\beta))$ and $D_{\mathrm{KL}}(\hat{p} \parallel p_{\mathrm{theory}}(\alpha))$, where theoretical distributions are constructed from the current β and α. This dynamic reference avoids locking the model to a fixed template while still penalizing large departures from empirically supported scaling forms.

Taken together, the evidence shows regularity with flexibility. Aggregate distance–decay behaviour falls within $\beta \in [0.2, 3.0]$ and rank-frequency scaling within $\alpha \in [0.5, 3.0]$, while monotonicity and KL metrics indicate consistent structure without suppressing heterogeneity. These intervals act as transferable priors that anchor learning rather than enforce a single generative mechanism. In PIG-GNN, interval constraints guide β and α, and the dynamic KL term encourages distributional realism, jointly improving interpretability and generalization across spatial scales.

3.2.4. Physics-informed edge prediction with adaptive loss design

The final stage of PIG-GNN (right panel of Fig. 6) combines learned node embeddings and edge representations with physics-informed constraints to produce flow predictions. The edge prediction MLP takes as input the concatenation of the origin production embedding, destination attraction embedding, refined edge embedding, and an explicit distance channel:

$$\hat{y}_{ij} = \mathrm{MLP}\big(\left[\,\mathbf{z}_i^P \parallel \mathbf{z}_j^A \parallel \mathbf{h}_{ij} \parallel \phi(d_{ij})\,\right]\big), \tag{7}$$

where the MLP has three fully connected layers with BatchNorm, Dropout, and ReLU activations. Retaining the explicit distance channel $\phi(d_{ij})$ acts as an inductive bias, allowing departures from rigid distance-decay while preserving impedance sensitivity.

Data-only optimization can yield locally accurate yet geographically implausible patterns (e.g., collapsed tails or distorted rank-frequency structure). We therefore embed physics-informed constraints through adaptive loss terms grounded in the empirical priors of Section 3.2.3. Fig. 8 summarizes the loss design and how empirical intervals are translated into soft constraints via the Soft Interval Constraints Application module.

As illustrated in Fig. 8, the adaptive loss integrates four components. The Data Fidelity term ($\mathcal{L}_{\mathrm{data}}$) minimizes mean squared error between predicted and observed flows. The Soft Interval Constraints module enforces empirically validated intervals: (1) Slope prior ($\mathcal{L}_{\mathrm{slope}}$, $\beta \in [0.2, 3]$); (2) Zipf prior ($\mathcal{L}_{\mathrm{zipf}}$, $\alpha \in [0.5, 3]$) so ranked flows follow $y_r \propto r^{-\alpha}$; and (3) Distributional prior ($\mathcal{L}_{\mathrm{KL}}$), which regularizes predicted distributions against theoretical forms constructed from the current β and α. The Adaptive Learning Process proceeds in three stages: (1) initialization of β and α within empirical intervals (Section 3.2.3); (2) joint optimization under data fit and constraint satisfaction; and (3) soft penalties for interval violations via quadratic costs:

$$\mathrm{penalty}(\theta) = \begin{cases} (\theta_{\min} - \theta)^2 & \text{if } \theta < \theta_{\min} \\ 0 & \text{if } \theta_{\min} \le \theta \le \theta_{\max} \\ (\theta - \theta_{\max})^2 & \text{if } \theta > \theta_{\max} \end{cases} \tag{8}$$

where θ represents a learned parameter (e.g., β, α, or D_{KL}) and $[\theta_{\min}, \theta_{\max}]$ is the corresponding empirical interval.

The overall objective combines data fidelity with three physics-informed regularizers:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\mathrm{data}}}_{\text{fit}} + \lambda_{\mathrm{slope}} \underbrace{\mathcal{L}_{\mathrm{slope}}}_{\text{power-law slope}} + \lambda_{\mathrm{zipf}} \underbrace{\mathcal{L}_{\mathrm{zipf}}}_{\text{rank-frequency scaling}} + \lambda_{\mathrm{KL}} \underbrace{\mathcal{L}_{\mathrm{KL}}}_{\text{distributional match}}, \tag{9}$$

where $\lambda_{\mathrm{slope}}, \lambda_{\mathrm{zipf}}, \lambda_{\mathrm{KL}}$ are adaptive weights.

Adaptive weights are updated based on the ratio of data loss to constraint loss, keeping constraints influential without overwhelming data fit:

$$\lambda_k^{\mathrm{adaptive}} = \lambda_k^{\mathrm{base}} \times \frac{\mathcal{L}_{\mathrm{data}}}{\mathcal{L}_k + \epsilon}, \tag{10}$$

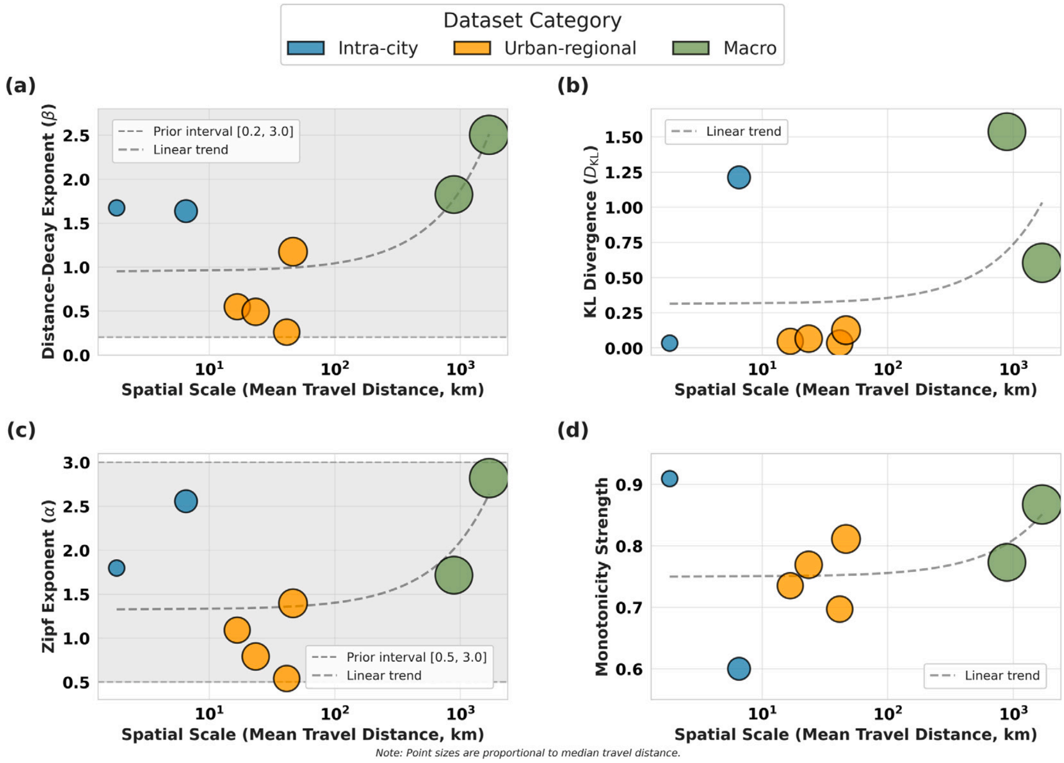

Fig. 7. Empirical parameter distributions and scale characteristics of the OD datasets analysed in this study. Each point represents one dataset, positioned according to its estimated distance-decay exponent β and Zipf exponent α, with marker size/colour indicating dataset scale and flow magnitude statistics.

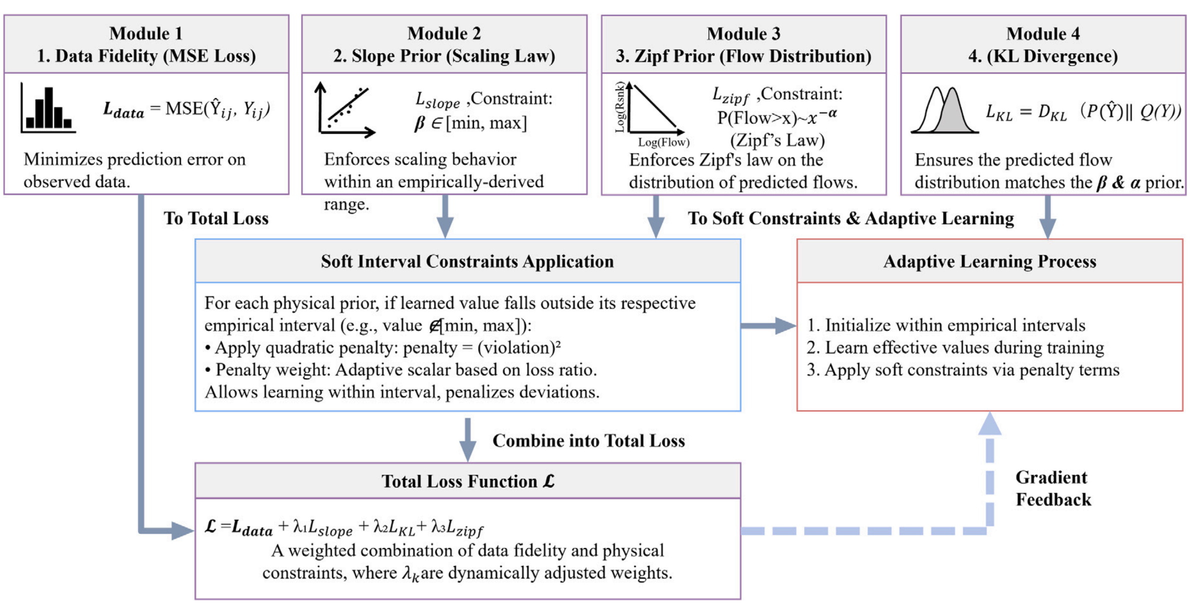

Fig. 8. Physics-informed loss design architecture for OD flow prediction. The figure illustrates the integration of four loss components (Data Fidelity, Slope Prior, Zipf Prior, and Distributional Prior) through the Soft Interval Constraints Application module, which constrains learned parameters (β, α, D_{KL}) within empirically validated intervals. The Adaptive Learning Process manages parameter initialization, optimization, and soft constraint application, with the Total Loss Function combining all components through adaptive weighting.

Table 1
Synthesis of scaling laws from mobility literature. The variation in Beta (β) across datasets validates the need for a flexible prior range rather than a fixed decay parameter.

Reference	Dataset description	N	Beta
Brockmann et al. (2006)	1 million bank note trajectories in the US	1 M	1.59
Gonzalez et al. (2008)	6-month mobile phone records for 100k users	16 M	1.75
Liu et al. (2014)	Social media check-ins, 370 Chinese cities	0.5 M	0.80
Liu et al. (2012)	GPS taxi trips in Shanghai (1 week)	1.5 M	1.20
Hawelka et al. (2014)	Global Twitter records for 13M users	944 M	1.62
Yan et al. (2013)	Travel diaries, Frauenfeld, Switzerland	36 K	1.05
Yao and Lin (2016)	High-frequency taxi GPS, South China	6.8 M	2.71
Deville et al. (2016)	One-month CDR, tower-level locations (Portugal)	1.3 M	1.50
Deville et al. (2016)	6-month mobile logs (W. Europe)	6 M	1.35
Deville et al. (2016)	5 years of mobile activity (Rwanda)	1.5 M	1.57

Table 2
Statistical properties of multi-source OD datasets computed in this study. The wide range of observed Beta (0.258–2.503) and Alpha (0.539–2.819) coefficients informs the broad constraint intervals used in PIG-GNN. Mono.: monotonicity strength of distance decay (0–1, higher means stricter monotonic decrease).

Dataset	N	Median	Mean	Beta	KL Div	Mono.	Alpha (Zipf)
H2W (England)	298,681	14.15	41.54	0.258	0.033	0.697	0.539
ALLOD (England)	2,383,725	28.99	46.64	1.173	0.125	0.811	1.395
Shared Bicycle (Shenzhen)	26,870,820	1.34	1.84	1.672	0.033	0.909	1.796
Didi Taxi (Chengdu)	6,980,323	5.39	6.56	1.635	1.212	0.600	2.556
Boston	126,847	13.25	16.77	0.547	0.046	0.735	1.090
Los Angeles	116,902	19.10	23.52	0.489	0.065	0.769	0.790
Liantong Migration (County)	8,173,881	773.31	890.80	1.824	1.537	0.773	1.715
Baidu Migration (City)	12,569	1516.30	1696.82	2.503	0.604	0.867	2.819

where λ_k^{base} is the base weight and ϵ is a small constant to prevent division by zero. This adaptive scheme ensures that when data fit is poor, constraint weights increase to provide stronger guidance, while when constraints are well-satisfied, their weights decrease to allow more flexible learning.

Complete mathematical formulations are provided in Appendix A.1. Collectively, these constraints inject geographic realism, improve generalization under sparsity, and preserve expressive capacity by balancing theoretical consistency with data-driven flexibility.

3.3. Training and evaluation

We use stratified K-fold cross-validation by distance deciles and origin/destination strength quantiles to preserve the long-tailed OD distribution and avoid leakage. Hyperparameters (including physics-informed weights) are selected via random search with early stopping under three-fold cross-validation, and performance is reported as mean ± standard deviation across folds. Baselines, ablations, and full hyperparameter ranges are documented in Appendix A.3.

4. Results

4.1. Overall model performance and baseline comparison

Overall, PIG-GNN achieves the best predictive accuracy and calibration among deployable baselines, attaining the lowest RMSE (0.297±0.002), one of the highest CPC values (0.791±0.004), and the joint-best MAE (0.187±0.009) (Table 3). Fig. 9 shows the relative magnitude of change of PIG-GNN and other baselines across the four metrics. In terms of R^2, PIG-GNN (0.394 ± 0.014) substantially improves over unconstrained gravity (0.284 ± 0.002) and GravityGNN (0.351 ± 0.056), indicating stronger explanatory power across heterogeneous commuting patterns.

We note that the doubly-constrained gravity model attains a higher R^2 (0.435 ± 0.005) and slightly higher CPC (0.792 ± 0.002). However, this model is not a fully comparable predictive baseline because it enforces the observed origin and destination marginals (row/column sums) as hard constraints, which provides oracle information unavailable in many realistic forecasting and scenario-planning settings where future marginals are unknown. Therefore, we interpret it as an upper-bound reference rather than a deployable competitor.

To examine how predictive performance varies with spatial scale, we further conduct a distance-binning error analysis on the held-out test folds (Appendix Table A.14). PIG-GNN consistently improves over the Gravity and DualTower baselines across all distance quintiles and achieves the highest CPC in every bin. In the short- and medium-distance bins (0–15.4 km), PIG-GNN also achieves the best RMSE, MAE, R^2, and CPC among all compared models. In the two longer-distance bins (>15.4 km), Random Forest attains slightly lower point-wise RMSE/R^2, but PIG-GNN maintains the highest CPC, indicating that it better preserves the distributional structure of long-distance, low-frequency flows. These results suggest that the model's gains are driven not only by improved local commuting prediction but also by stronger scaling-consistent representation of longer-distance OD interactions. Full numerical results are reported in Appendix A.9.

The performance improvements demonstrate robust generalization under varying flow complexity and spatial heterogeneity. Classical models (gravity, radiation) show limited capacity to capture complex spatial dependencies, while machine learning regressors (Random Forest, XGBoost) improve accuracy but lack spatial structure and theoretical grounding. Graph-based baselines capture spatial dependencies but fail to preserve theoretical consistency, leading to calibration degradation. PIG-GNN's integration of theory-informed structure and physics-informed learning enables it to achieve both high predictive accuracy and geographic coherence.

Fig. 10 further indicates stable training under the physics-informed objective, where the adaptive weighting balances data fidelity with the slope/Zipf/KL constraints so that all terms converge without domination or divergence, the learned scaling parameters (β, α) remain within empirically plausible ranges, and the KL component adjusts dynamically to maintain distributional consistency.

Fig. 11 complements the aggregate metrics by visualizing the error structure and calibration behaviour of each baseline. Each row corresponds to one model (from unconstrained gravity to PIG-GNN), where the left panel shows the kernel-density scatter of predicted versus observed flows in log space (with marginals indicating sample density), and the right panel summarizes residual statistics binned by flow-magnitude deciles to reveal systematic bias across the distribution. To

Table 3
Performance comparison of PIG-GNN and baseline models on the H2W dataset. Metrics include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2), and Common Part of Commuters (CPC). Results are reported as mean ± standard deviation across 5-fold stratified cross-validation. Best results among deployable baselines are highlighted in bold; ties at the displayed precision are both highlighted. The doubly-constrained gravity model uses observed row and column marginals as hard constraints, i.e., oracle information, and is reported as an upper-bound reference rather than a deployable competitor; its entries are therefore not counted when selecting bold highlights.

Model	RMSE ↓	MAE ↓	R^2 ↑	CPC ↑
Unconstrained Gravity (Zipf, 1946)	0.343 ±0.002	0.233 ±0.000	0.284 ±0.002	0.754 ±0.001
Gravity (PPML) (Silva & Tenreyro, 2006)	0.347 ±0.034	0.250 ±0.036	0.245 ±0.044	0.756 ±0.021
Hurdle Gravity (Gu & Shen, 2021)	0.323 ±0.004	0.225 ±0.002	0.287 ±0.011	0.770 ±0.002
Doubly-constrained Gravity (Wilson, 1971)	0.304 ±0.002	0.235 ±0.002	0.435 ±0.005	0.792 ±0.002
Radiation Model (Simini et al., 2012)	0.342 ±0.002	0.233 ±0.000	0.284 ±0.002	0.754 ±0.001
Random Forest (Breiman, 2001)	0.336 ±0.002	0.215 ±0.001	0.310 ±0.003	0.769 ±0.002
XGBoost (Chen & Guestrin, 2016)	0.334 ±0.002	0.216 ±0.001	0.317 ±0.002	0.767 ±0.002
MLP (OD graph) (Rumelhart et al., 1986)	0.317 ±0.005	**0.187 ±0.006**	0.310 ±0.018	0.778 ±0.003
Mixture Density Network (Bishop, 1994)	0.361 ±0.006	0.219 ±0.007	0.204 ±0.023	0.748 ±0.008
Dual-tower GCN (multi-graphs) (Wang et al., 2019)	0.322 ±0.042	0.190 ±0.042	0.323 ±0.046	0.780 ±0.028
GMEL (Geo-Contextual) (Liu et al., 2020)	0.323 ±0.029	0.218 ±0.031	0.333 ±0.031	0.779 ±0.018
Sat2flow MLP-Diffusion (X. Wang et al., 2025)	0.320 ±0.008	0.196 ±0.015	0.300 ±0.027	0.774 ±0.008
Gravityformer (no P.I.) (Y. Wang et al., 2025)	0.315 ±0.038	0.196 ±0.040	0.365 ±0.072	0.781 ±0.027
GravityGNN (Our prototype)	0.319 ±0.037	0.209 ±0.040	0.351 ±0.056	0.778 ±0.024
TransFlower (Luo et al., 2024)	0.303 ±0.002	0.204 ±0.004	0.373 ±0.007	0.783 ±0.004
GAGNN (Rong et al., 2023)	0.305 ±0.002	0.188 ±0.003	0.361 ±0.002	0.785 ±0.003
PIG-GNN (Ours)	**0.297** ±0.002	**0.187** ±0.009	**0.394** ±0.014	**0.791** ±0.004

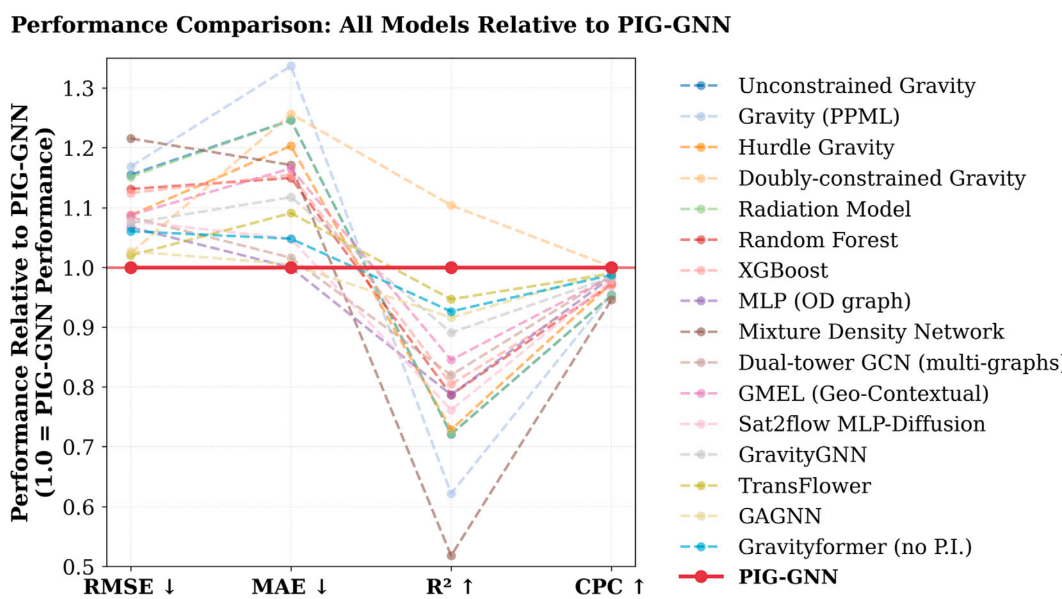

Fig. 9. Comprehensive performance comparison across all baseline models on the H2W dataset.

complement the accuracy comparison, we also report the computational footprint of all compared models, including parameter counts and normalized runtime estimates, in Appendices A.4 and A.5.

Several consistent patterns emerge. Traditional gravity models capture the coarse global trend but exhibit pronounced underestimation for high-flow OD pairs, which appears as a flattening below the diagonal in the scatter plots and increasingly negative residuals in the top deciles. The doubly constrained gravity model shows improved alignment largely because it enforces the observed origin and destination marginals using ground-truth row/column sums; we therefore treat it as a partially oracle-assisted reference rather than a deployable predictor. Tree-based regressors (XGBoost and Random Forest) better fit nonlinearities for mid-range flows, yet retain noticeable variance and bias for larger flows. The naive edge MLP reduces average bias but lacks structural inductive bias, leading to less stable predictions, while the mixture density network captures multimodal uncertainty but displays calibration instability. GravityGNN improves structural consistency compared with non-graph baselines, but it still systematically underestimates high-flow interactions. In contrast, PIG-GNN shows the tightest concentration around the diagonal together with more balanced residual profiles across deciles, indicating stronger calibration and robustness across both frequent short-range flows and rare, high-impact long-tail connections.

4.2. Ablation analysis and component contributions

The ablation results Table 4 reveal consistent patterns. Removing the dual-tower structure (+ Dual-tower) increases prediction errors and weakens calibration, validating the necessity of separating origin and destination representations for role-aware learning. Removing the OD

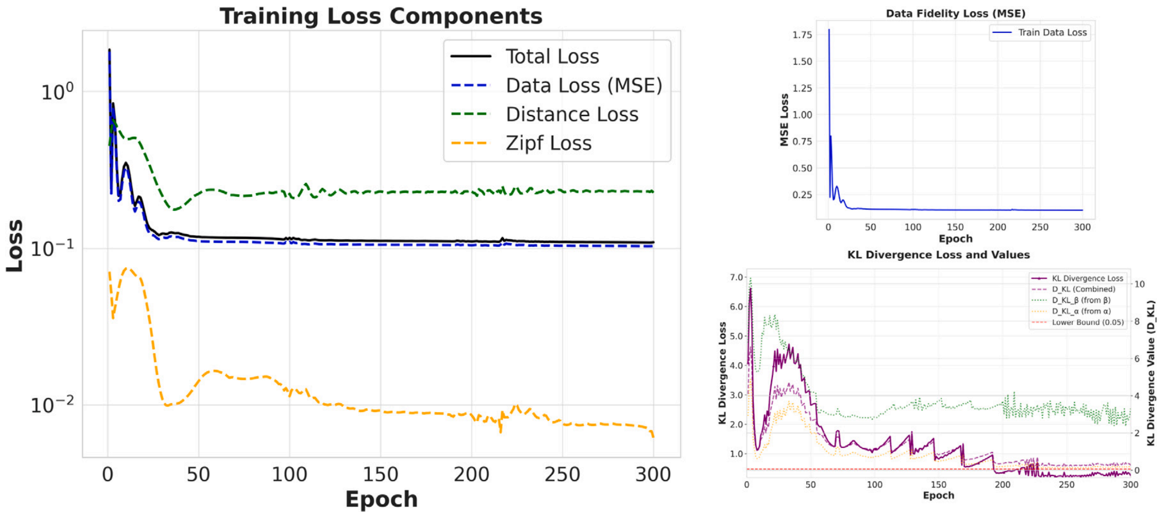

Fig. 10. Training dynamics of PIG-GNN showing the evolution of loss components throughout training. The figure illustrates: (a) individual loss components (data fidelity, slope prior, Zipf prior, and distributional KL divergence constraint) and their convergence patterns; (b) adaptive weight evolution for each constraint, demonstrating how the weighting mechanism balances constraint satisfaction with predictive accuracy; and (c) learned parameter trajectories (β, α) showing that they remain within empirically validated intervals.

Table 4
Component-wise progression from MLP baseline to full PIG-GNN. Metrics include RMSE, MAE, R^2, and CPC, reported as mean ± standard deviation across 5-fold cross-validation.

Model Variant	RMSE ↓	MAE ↓	R^2 ↑	**CPC** ↑
MLP baseline	0.317 ±0.005	0.187 ±0.006	0.310 ±0.018	0.778 ±0.003
+ Adjacency GCN	0.307 ±0.013	0.179 ±0.012	0.353 ±0.055	0.787 ±0.014
+ Adjacency GATv2	0.306 ±0.010	**0.169** ±0.003	0.359 ±0.034	0.790 ±0.005
+ Dual-tower	0.301 ±0.010	0.199 ±0.016	0.378±0.041	0.783 ±0.010
+ OD Message Passing	0.300 ±0.007	0.202 ±0.007	0.384 ±0.025	0.784 ±0.007
+ **Physics-informed losses**	**0.297** ±0.002	0.187 ±0.009	**0.394**±0.014	**0.791** ±0.004

interaction block (+ OD Message Passing) produces similar declines, underscoring its importance for capturing competition and substitution effects among destinations. The adjacency graph GCN (+ Adjacency GCN) contributes significantly to spatial context learning, with its removal leading to substantial performance degradation.

Disabling the physics-informed loss terms (+ Physics-informed losses) yields systematic calibration degradation and flatter distance-decay curves, particularly evident in the CPC metric. On the H2W dataset, removing physics-informed losses reduces CPC from 0.791 to 0.784. Among the three physics-informed constraints, the slope prior contributes most to global scaling alignment, the Zipf prior preserves the rank-frequency structure of flows, and the distributional prior maintains long-tail fidelity. The full PIG-GNN configuration thus achieves the best balance between accuracy and interpretability, producing geographically consistent OD surfaces that adhere to known urban scaling laws.

4.3. Dual-tower interpretability: counterfactual analysis

To validate the theoretical interpretability of the dual-tower architecture and demonstrate its ability to disentangle origin and destination roles, we conduct systematic counterfactual experiments. These experiments examine how modifications to production features (working-age population) and attraction features (POI density) independently and jointly affect predicted flow magnitudes, providing direct evidence for the role-aware learning mechanism.

Table 5 reports counterfactual analysis results across 11 scenarios, systematically varying production and attraction multipliers from 0.8 to 1.5 (representing −20% to +50% changes). The results reveal several key insights. First, production features exhibit strong, non-linear effects on predicted flows: a 20% increase in production capacity leads to a 22.01% increase in predicted flows, while a 50% increase yields a 57.27% increase—demonstrating a super-linear response that aligns with gravity model theory where origin mass is a primary determinant of flow magnitude. The symmetry of effects is well-preserved: a 20% decrease in production results in a −17.88% flow reduction, indicating consistent directional behaviour.

Second, attraction features show relatively modest direct effects on flow predictions: a 20% increase in attraction capacity leads to only a 0.53% increase in predicted flows, while a 50% increase yields a 2.53% increase. This asymmetry between production and attraction effects reflects the fundamental structure of the gravity model, where origin mass typically dominates flow magnitude, while destination attractiveness operates more subtly through competitive and substitutive mechanisms captured by the graph message passing. This pattern may also reflect the coarse proxy nature of POI density for destination attractiveness in commuting systems

Third, combined scenarios reveal that production features dominate flow predictions, with attraction effects providing secondary modulation. When both features are increased by 20% (Both+20%), the resulting 21.56% flow increase closely matches the production-only scenario (22.01%), confirming production's primary role. Notably, increasing production while decreasing attraction (Prod+Attr-) yields a 22.69% increase—slightly higher than production-only—suggesting that reduced attraction may actually enhance flow concentration to high-production origins. Conversely, decreasing production while increasing attraction (Prod-Attr+) results in a −15.50% decrease, demonstrating that enhanced destination attractiveness cannot compensate for reduced origin capacity.

These counterfactual findings provide a supportive diagnostic that the dual-tower architecture responds to production and attraction perturbations in directions consistent with spatial-interaction theory. The

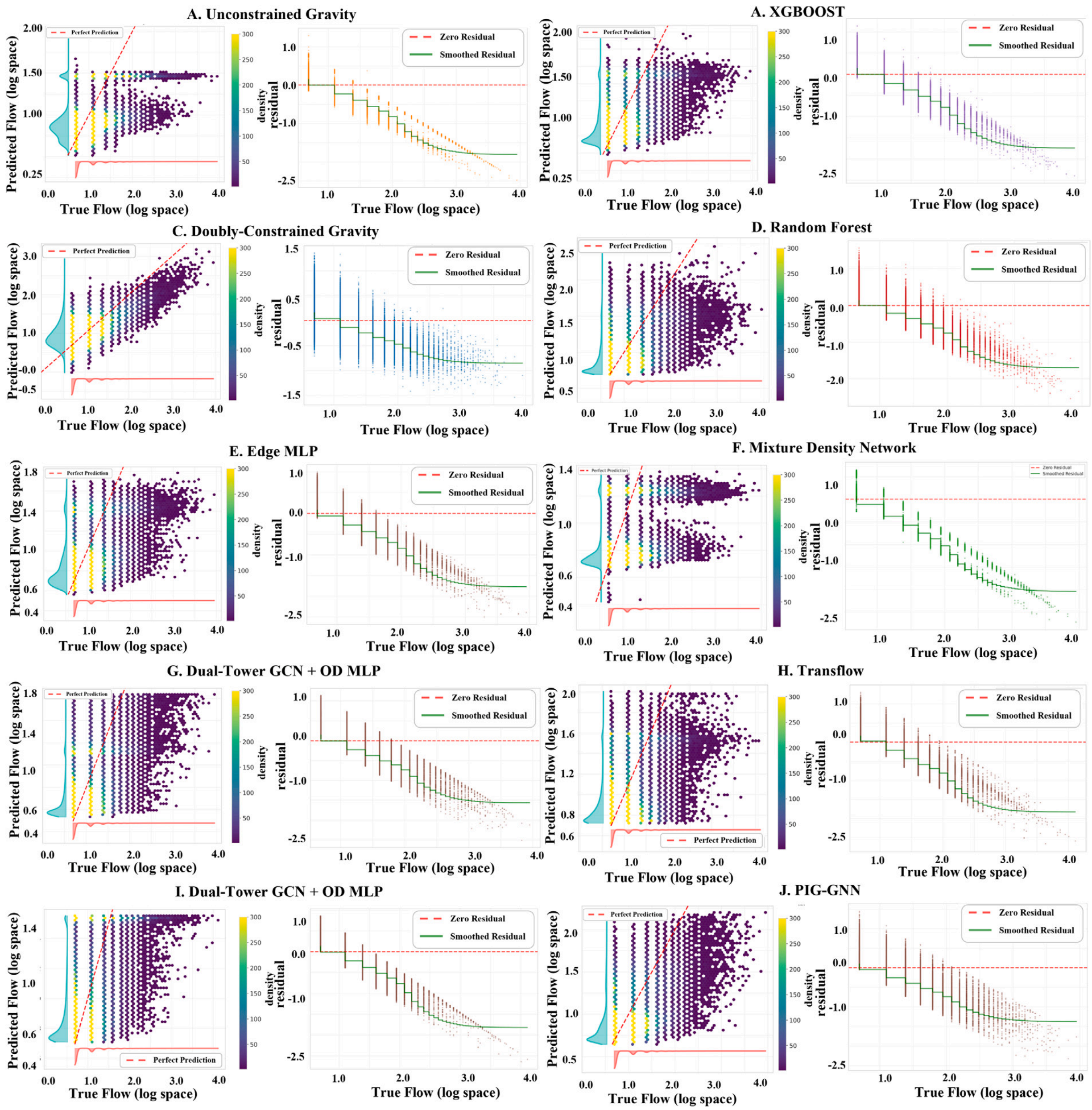

Fig. 11. Prediction–observation alignment and residual patterns across baseline models on H2W.

Table 5
Counterfactual analysis of dual-tower interpretability on the H2W dataset. Each scenario modifies production and/or attraction features by the specified multipliers, and reports the resulting mean change in predicted flows (log space and percentage). The analysis demonstrates the differential impacts of origin (production) versus destination (attraction) features, validating the role-aware learning mechanism of the dual-tower architecture.

Scenario	Production multiplier	Attraction multiplier	Mean change (log)	Mean change (%)	Std change
Baseline	1.0	1.0	0.0000	0.00%	0.0000
Prod+20%	1.2	1.0	0.1878	22.01%	0.0253
Prod+50%	1.5	1.0	0.4857	57.27%	0.0370
Prod-20%	0.8	1.0	−0.1557	−17.88%	0.0396
Attr+20%	1.0	1.2	0.0032	0.53%	0.0146
Attr+50%	1.0	1.5	0.0178	2.53%	0.0374
Attr-20%	1.0	0.8	0.0012	0.04%	0.0111
Both+20%	1.2	1.2	0.1836	21.56%	0.0201
Both−20%	0.8	0.8	−0.1678	−19.54%	0.0232
Prod+Attr-	1.2	0.8	0.1936	22.69%	0.0282
Prod-Attr+	0.8	1.2	−0.1376	−15.50%	0.0546

clear separation of production and attraction effects demonstrates that the model successfully learns role-aware representations, with production features primarily governing flow magnitude and attraction features operating through more subtle competitive mechanisms. This interpretability is crucial for policy applications, where understanding the differential impacts of origin capacity versus destination attractiveness enables targeted interventions in urban planning and transport infrastructure.

We emphasize that the counterfactual perturbations above are intended as a local monotonicity check rather than a causal identification claim. Large linear scalings can push inputs into out-of-distribution regions unseen during training, so the resulting changes should be interpreted as theory-consistency diagnostics rather than as rigorous causal effects. In particular, the multipliers at the extremes of the 0.8–1.5 range may enter OOD input regions. The perturbation results therefore indicate whether the learned dual-tower representations respond in directions consistent with production–attraction theory, but they do not establish causal effects of changing population or attraction attributes. Complementarily, the distance-decay and Zipf regularities enforced through the physics-informed learning module provide distribution-level evidence that macroscopic mass–distance structure is encoded alongside this local sensitivity probe.

5. Discussion

5.1. Why PIG-GNN works: structure, priors, and implications for urban mobility analysis

PIG-GNN's performance gains stem from the combination of structural disentanglement and physics-informed regularization, aligning model capacity with production–attraction asymmetry and empirically established scaling regularities in spatial interaction (Batty, 2008; Biscaia & Mota, 2013; Brockmann et al., 2006; Gonzalez et al., 2008). The dual-tower encoder limits role leakage by letting origin embeddings prioritize supply and outbound competition while destination embeddings encode attraction and inbound congestion; this improves identifiability in counterfactual analysis because shocks propagate through the appropriate role-specific channel (Abrahamsson, 1998; Mohammed & Oke, 2023). The OD edge-attention module contextualizes each flow by learning how competing destinations share demand and how inflows interact through crowding effects, providing a data-adaptive analogue to classical balancing and competition mechanisms in spatial interaction models (Roy & Thill, 2004).

The power-law, Zipf, and distributional priors act as soft constraints that stabilize learning around observed scaling regimes while still permitting local deviations; this is consistent with physics-informed and knowledge-guided machine learning principles (Ectors et al., 2019; Karpatne et al., 2024; Liu et al., 2015; Zhao et al., 2025). Together, the architecture and priors provide geographic realism without hard-coding a rigid functional form, reconciling interpretability with predictive flexibility.

We further assess sensitivity to the *centres* of the soft distance-decay (β) and Zipf (α) prior intervals on the H2W dataset. Rather than sweeping loss weights alone, we fix the training protocol and centre a narrow admissible band $[c - 0.03, c + 0.03]$ on alternative coefficients cAppendix A.10. The results show that RMSE and CPC vary smoothly across the tested prior centres: RMSE is minimized under low-to-moderate exponents, while CPC peaks around empirically plausible β and α values. Even when the interval is shifted far from the preferred region, such as $\beta = 3$ or $\alpha = 3$, RMSE increases only modestly and CPC remains above 0.784. This indicates that the physics-informed priors provide a useful inductive bias when aligned with observed scaling behaviour, but do not make the model catastrophically sensitive to moderate prior misspecification.

From a methodological standpoint, physics-informed graph models can surface latent mobility regimes without pre-segmentation and highlight interpretable departures from theory, offering diagnostic signals for where gravity-like scaling holds or breaks down (Alessandretti et al., 2020; Chen, 2015; Wu et al., 2025). The observed bimodality suggests that hierarchical or mixture-of-experts formulations may better capture intra- and inter-urban processes that operate at different spatial scales (Alessandretti et al., 2020; Liu et al., 2014). The persistent underfit in the high-pattern regime points to the limits of static OD-side features; incorporating transport-network topology and multimodal accessibility metrics may better capture micro-scale mechanisms in dense urban cores (Major et al., 2020; Yang & Zhang, 2025). Future work could extend to cross-city comparisons (e.g., transit-oriented vs. automobile-dependent systems), harmonized OD datasets to test whether bimodality is universal or context-specific, and models for censored or incomplete OD matrices (e.g., zero-inflated formulations) in sparse mobility systems.

In this sense, the model is not only a more accurate OD predictor, but also a more interpretable analytical tool for understanding how mobility demand is organized across urban systems and spatial scales.

5.2. Generalizability

Although the main experiments focus on England's MSOA-level Home-to-Work (H2W) dataset, the PIG-GNN framework is not tied to a specific trip purpose or local context. Its design principles, which separate origin and destination processes while embedding physics-based scaling priors, are consistent with spatial interaction theory and widely observed mobility scaling laws (Alessandretti et al., 2020).

To first assess generalization beyond commuter-specific interactions, we test the same modelling pipeline on the large-scale ALLOD dataset, which covers all trip purposes in England using the same MSOA spatial units and input feature configuration. Detailed dataset statistics and the expanded ALLOD benchmark are reported in Appendix A.6. Compared with H2W, ALLOD is substantially larger and more heterogeneous in edge count and flow distribution, providing a stricter same-country robustness test under a heavier-tailed, multi-purpose flow regime.

Beyond the same-country ALLOD test, we further assess cross-city transferability on two independent 500 m-grid OD datasets from Xi'an and Guangzhou. These cities represent contrasting urban morphologies: Xi'an is an inland mono-centric city with an ancient walled core, whereas Guangzhou is a coastal polycentric mega-city embedded in the Pearl River Delta. In both settings, PIG-GNN outperforms Gravity, Random Forest, XGBoost, TransFlower, and GravTransf across RMSE, R^2, and CPC (Tables A.17 and A.18; Figs. A.16 and A.19). The fitted scaling exponents also remain within the empirically validated prior ranges in both cities, with $\beta = 0.425$ and $\alpha = 0.582$ for Xi'an, and $\beta = 0.327$ and $\alpha = 1.066$ for Guangzhou (Figs. A.17 and A.20).

These cross-city validations deliberately use different zonal systems, spatial resolutions, and locally available production/attraction proxies from the England H2W benchmark. Their absolute metric values should therefore not be interpreted as directly comparable across countries or spatial units. Instead, the intended evidence is whether PIG-GNN maintains strong within-city rankings under fixed local definitions, and whether the fitted distance-decay and Zipf exponents remain within the empirically calibrated prior ranges.

Taken together, these results suggest that PIG-GNN's role-aware architecture and physics-informed priors are not specific to the England H2W benchmark alone. The framework can be adapted to other national and regional systems by changing the available production, attraction, and impedance proxies, and may extend to related domains such as tourism, freight, and logistics, where destination competition and substitution effects also play a central role (Simini et al., 2012). Such transferability is important for sustainable cities, where mobility

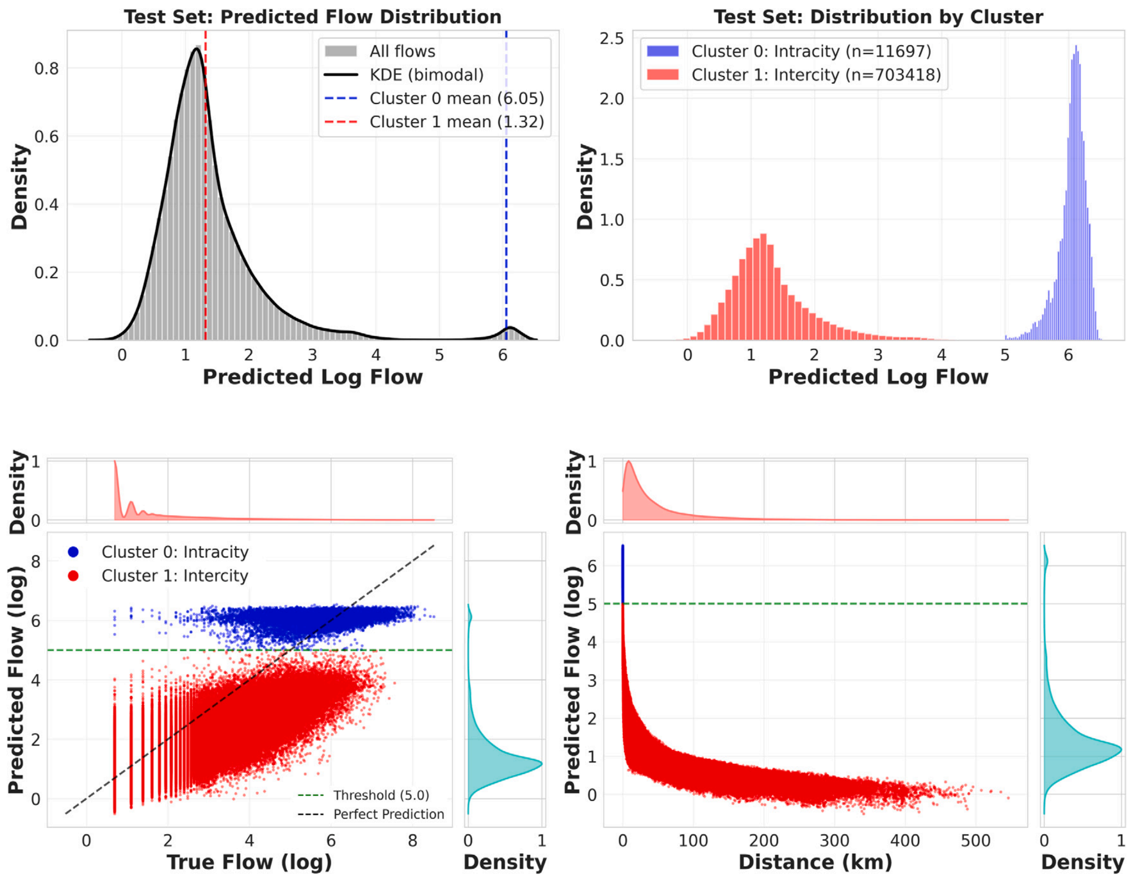

Fig. 12. Bimodality and distance-decay regimes in predicted OD flows on the H2W dataset.

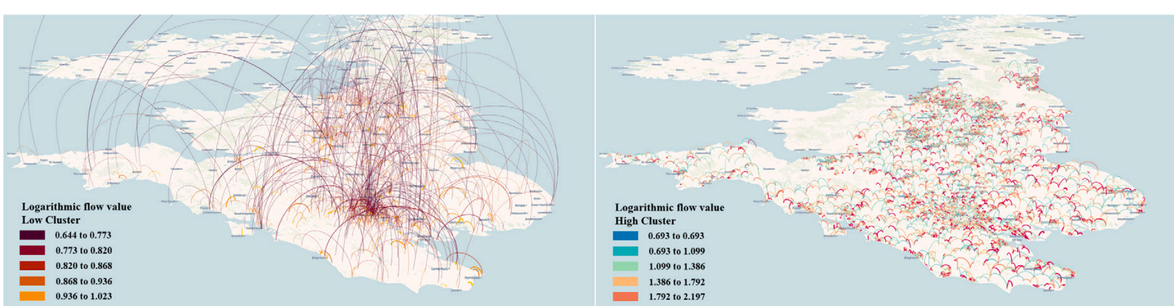

Fig. 13. Sampling visualization of low-value traffic clustering (1000 records) (left) Sampling visualization of high-value traffic clustering (5000 records) (right).

forecasting often needs to support multiple interacting systems rather than a single travel purpose in isolation.

5.3. Interpreting bimodality and multi-scale mobility regimes

On the H2W dataset, PIG-GNN preserves a distinctive bimodal pattern in the predicted OD flow distribution: scatter plots of predicted versus observed flows show a clear separation into two regimes, consistent with prior evidence of multi-scale mobility structure (Chen et al., 2023; Li et al., 2021). Classical gravity models tend to underestimate large flows and compress the high-flow regime towards the low-flow mode (peak collapse), whereas the learned distribution maintains a pronounced gap in log-space. The high-pattern regime corresponds to dense, short-range flows typical of intra-urban mobility, while the low-pattern regime corresponds to sparse, long-range connections typical of inter-urban mobility, a separation that aligns with documented scale-dependent distance decay and interaction heterogeneity (Cadwallader, 2013; Nayka & Sridhar, 2019; Zheng et al., 2016).

To investigate this bimodality, we apply a threshold-based clustering to identify the two flow regimes and examine their predictive characteristics. Figs. 12 and 13 summarize the two-regime structure through complementary diagnostics. The left panel plots predicted versus true log flows, with points coloured by the two clusters and marginal densities highlighting the bimodal separation; the low-flow cluster concentrates near the perfect-prediction line ($y = x$), whereas the high-flow cluster exhibits a systematic tendency towards overprediction. The right panel relates predicted log flow to distance using the same clustering, revealing distinct distance-decay profiles: the low-flow cluster shows a steeper decay consistent with gravity-like inter-urban behaviour, while the high-flow cluster displays substantially weaker distance dependence. The horizontal reference line indicates the threshold used to separate the two regimes. Together, the two panels support the interpretation that the observed bimodality reflects genuine geographic heterogeneity between inter-urban and intra-urban mobility regimes rather than an artefact of model fitting.

The low-flow regime exhibits steep distance decay—consistent with gravity-like inter-urban behaviour where distance deterrence and opportunity competition follow classical mass–distance scaling (Chen, 2015; Ewedairo et al., 2018). In contrast, the high-flow regime is associated with short-range intra-urban movements, where micro-scale mechanisms such as destination crowding, mixed land use, hub accessibility, and transfer frictions can dominate over distance impedance alone (He et al., 2017; Lambiotte et al., 2008). While both regimes exhibit power-law decay with distance, they likely arise from different generative processes: aggregate opportunity competition in the low-flow regime versus local urban structure effects in the high-flow regime (Wang, 2025). Notably, PIG-GNN separates these regimes without explicit regime-specific modelling; the dual-tower encoders and OD edge-attention provide role-specific context and local adaptation to density and congestion, helping the model preserve heavy tails and maintain calibration where physics-informed priors exert stronger guidance.

The bimodal structure can thus serve as a diagnostic signal for the boundary between macro-scale gravity rules and micro-scale urban complexities, supporting multi-scale policy design: micro-level measures (e.g., interchange capacity, last-mile connectivity) for short-range accessibility versus regional investments that reduce impedance for inter-urban corridors (Monzon et al., 2017; Zhu & Rui, 2025).

A deeper examination suggests that bimodality is not unique to PIG-GNN but an inherent feature of OD flow data; different models, however, represent and preserve this structure to different degrees. Classical spatial interaction models (gravity, radiation) exhibit peak collapse—their rigid forms underestimate large flows and collapse the high-pattern regime towards the low-pattern mode. Machine learning regressors (Random Forest, XGBoost) achieve competitive aggregate statistics ($R^2 \approx 0.317$) yet produce smooth, unimodal scatter plots (e.g., Fig. 11) that smooth over the underlying bimodality. Graph-based models, including GravityGNN and PIG-GNN, capture bimodal structure through their spatial architectures; PIG-GNN further delivers interpretable bimodality—physics-informed priors and dual-tower separation help ensure the bimodal split reflects genuine geographic heterogeneity rather than artefacts, with implications for predictive accuracy and interpretability.

5.4. Implications for sustainable urban mobility planning

The regime-aware interpretation of OD flows enabled by PIG-GNN's dual-tower architecture and physics-informed constraints provides a useful basis for sustainable urban mobility planning across spatial scales (Haynes et al., 1984; Sen & Smith, 2012). By preserving production–attraction asymmetry and distance-based scaling regularities, the framework offers a more interpretable account of how daily mobility demand is organized within and between urban areas (Monzon et al., 2017; Zhu & Rui, 2025). This is valuable in planning contexts where decision-makers require not only accurate forecasts, but also transparent evidence on whether predicted changes are associated with residential growth, employment concentration, corridor connectivity, or short-range accessibility constraints. The bimodality identified by the model further implies different intervention logics for inter-urban and intra-urban mobility regimes, supporting more targeted planning responses for regional connectivity, urban accessibility, and infrastructure management (Alessandretti et al., 2020; Jiang & Jia, 2011). More concretely, these properties make PIG-GNN relevant for planning tasks that require scenario-sensitive OD forecasts rather than purely descriptive prediction. For example, the disentangled production–attraction representation can help distinguish whether predicted demand shifts are primarily associated with changes at origins (e.g., residential expansion) or destinations (e.g., employment or service concentration), while the regime-aware structure can support different intervention logics for short-range urban accessibility versus longer-range inter-urban connectivity. This is particularly relevant for evaluating land-use change, transport investment, and corridor-specific accessibility interventions under heterogeneous mobility regimes. In this sense, the framework functions less as a black-box forecasting device and more as an interpretable planning-support tool for evaluating how urban form, accessibility, and mobility interventions may reshape flow patterns under heterogeneous urban regimes.

5.5. Limitations

Several limitations should be acknowledged. First, although we extend the evaluation to ALLOD and cross-city validation in Xi'an and Guangzhou, the core benchmark remains centred on England; broader validation across transit-oriented, automobile-dependent, and lower-data urban contexts would further strengthen external validity. Second, the production and attraction inputs rely on coarse proxies, such as working-age population and POI density. Finer-grained workplace-population, employment-land-use, or job-sector data would improve the physical grounding of the model. Third, the current implementation uses static attributes and Euclidean distance, rather than explicit transport-network topology, travel time, or multimodal accessibility, which may limit its ability to capture corridor effects and micro-scale intra-urban mechanisms. Fourth, the primary H2W training target focuses on observed positive OD edges; although the zero-flow augmentation ablation shows that PIG-GNN can accommodate sampled zero-flow pairs, a full zero-inflated framework with separate link-formation and flow-intensity components remains future work. Fifth, the physics-informed priors are implemented as soft constraints, preserving flexibility but not guaranteeing strict scaling-law adherence for every individual prediction. Finally, the model is predictive rather than causal, so the counterfactual analyses and planning implications should be interpreted as diagnostic and scenario-oriented rather than as evidence of policy effects.

6. Conclusion

This study advances theory-consistent OD flow forecasting by integrating a role-aware graph architecture with physics-informed scaling priors in a unified framework. PIG-GNN explicitly separates production and attraction processes through a dual-tower encoder, represents OD-level dependence (competition/substitution) via an interaction graph, and regularizes learning with empirically grounded scaling constraints without enforcing a fixed functional form. On England's MSOA-level home-to-work flows, these design choices yield predictions that are not only more accurate, but also more geographically coherent, reproducing key macroscopic regularities and supporting regime-aware interpretation of OD patterns, including the observed bimodal structure (Section 5.3). These properties make the framework particularly suitable for urban applications where prediction needs to remain interpretable and structurally meaningful rather than purely data-driven.

More broadly, our results suggest that scaling laws can function as a practical inductive bias for mobility prediction: when embedded as adaptive priors, they improve generalization while preserving the flexibility of neural models. The proposed framework therefore provides a principled route to reconcile data-driven forecasting with spatial interaction theory, which is increasingly important for sustainable urban mobility analysis and planning support. In particular, disentangled production–attraction representations can facilitate diagnosis of where changes originate, while regime-aware predictions can inform differentiated interventions across spatial scales.

As future work, it would be valuable to extend PIG-GNN in four directions: (i) incorporating multimodal transport-network structure and time-varying impedance to better capture within-city mechanisms; (ii) developing uncertainty-aware outputs (e.g., calibrated predictive intervals) to support risk-sensitive and resilience-oriented planning; (iii) strengthening causal validity through counterfactual or quasi-experimental evaluation designs for policy analysis; and (iv) testing transferability across cities, time periods, and travel purposes to assess robustness under distribution shift. Overall, PIG-GNN offers a transparent and transferable foundation for OD forecasting that is simultaneously theory-aligned, planning-relevant, and suitable for sustainable urban mobility analysis across heterogeneous urban systems.

CRediT authorship contribution statement

Minwei Zhao: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dailuo Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Zhecheng Shi:** Writing – review & editing, Visualization. **Cai Wu:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the members of the HKUST(GZ) Urban Morphology Studio for their discussions.

Appendix. Method transparency supplement

A.1. Physics-informed priors: detailed formulations

This appendix reports the full formulations for the slope, Zipf, and distributional priors used in PIG-GNN.

A.1.1. Power-law slope prior

We estimate a distance-decay exponent β by regressing $\log \hat{y}_{ij}$ on $\log d_{ij}$ within distance bins and taking a weighted average across bins. For bin k, the OLS slope is

$$\beta_k = \frac{\sum_{(i,j)\in\mathcal{B}_k}(\log d_{ij} - \overline{\log d}_k)(\log \hat{y}_{ij} - \overline{\log \hat{y}_k})}{\sum_{(i,j)\in\mathcal{B}_k}(\log d_{ij} - \overline{\log d}_k)^2}, \tag{A.1}$$

and the global slope is $\beta = \frac{\sum_k w_k \beta_k}{\sum_k w_k}$ with $w_k = |\mathcal{B}_k|$. The slope prior enforces $\beta \in [0.2, 3]$:

$$\mathcal{L}_{\text{slope}} = \begin{cases} (\beta_{\min} - \beta)^2 & \beta < \beta_{\min} \\ 0 & \beta_{\min} \le \beta \le \beta_{\max} \\ (\beta - \beta_{\max})^2 & \beta > \beta_{\max} \end{cases} \tag{A.2}$$

Gradients follow standard OLS differentiation.

A.1.2. Distributional prior

We compute KL divergence between the predicted log-flow distribution and two theoretical distributions derived from the learned β and α. Theoretical log flows are constructed as

$$\log \hat{y}_{ij}^{\text{theory}}(\beta) = \bar{\log}\, y + \beta \bar{\log}\, d - \beta \log d_{ij}, \quad \log \hat{y}_{(r)}^{\text{theory}}(\alpha) = \bar{\log}\, y + \alpha \bar{\log}\, r - \alpha \log r, \tag{A.3}$$

and converted to histogram distributions for KL computation. The combined KL term uses adaptive weights:

$$D_{\text{KL}} = \lambda_{\text{KL},\beta} D_{\text{KL}}(\hat{p} \parallel p_{\text{theory}}(\beta)) + \lambda_{\text{KL},\alpha} D_{\text{KL}}(\hat{p} \parallel p_{\text{theory}}(\alpha)). \tag{A.4}$$

The loss applies a lower-bound constraint, $\mathcal{L}_{\text{KL}} = (D_{\text{KL,min}} - D_{\text{KL}})^2$ when $D_{\text{KL}} < D_{\text{KL,min}}$.

A.1.3. Zipf rank-frequency prior

We estimate the Zipf exponent α by regressing $\log \hat{y}_{(r)}$ on $\log r$:

$$\alpha = -\frac{\sum_{r=1}^{|\mathcal{E}|}(\log r - \overline{\log r})(\log \hat{y}_{(r)} - \overline{\log \hat{y}})}{\sum_{r=1}^{|\mathcal{E}|}(\log r - \overline{\log r})^2}, \tag{A.5}$$

and enforce $\alpha \in [0.5, 3]$:

$$\mathcal{L}_{\text{zipf}} = \begin{cases} (\alpha_{\min} - \alpha)^2 & \alpha < \alpha_{\min} \\ 0 & \alpha_{\min} \le \alpha \le \alpha_{\max} \\ (\alpha - \alpha_{\max})^2 & \alpha > \alpha_{\max} \end{cases} \tag{A.6}$$

We use a differentiable sorting approximation to compute gradients through the ranking operation.

A.2. Training and hyperparameter search

We adopt stratified K-fold cross-validation by distance deciles and origin/destination strength quantiles to preserve the long-tailed OD distribution and reduce split-induced bias. Within each outer fold, model selection is conducted via random search with early stopping under three-fold cross-validation on the training portion of the fold. Unless otherwise stated, the same split strategy, early-stopping criterion, and selection rule are applied to all baselines and ablation variants to ensure a fair comparison. Final performance is reported as mean ± standard deviation across outer folds.

Early stopping is based on validation performance with a fixed patience and a maximum number of epochs; the best checkpoint on the validation folds is retained for evaluation. Random search is performed for a fixed budget of trials (20 trials for graph models), and the configuration yielding the best average validation performance across the inner folds is used for the corresponding outer-fold training (see Table A.6).

Table A.6
Hyperparameter search space for graph models (including PIG-GNN and graph baselines). Final configurations are selected from 20 random trials under three-fold cross-validation with early stopping. Parameters not listed use standard defaults in the corresponding implementations.

Hyperparameter	Search range or options
Learning rate	$\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}, 5\times10^{-3}\}$
Hidden dimension	{16, 32, 64, 128, 256}
# message-passing layers	{1, 2, 3, 4}
Dropout rate	{0.0, 0.2, 0.3}
Aggregation function	{mean, sum, max}
Activation function	{ReLU, ELU, LeakyReLU}
Weight decay	$\{0, 1\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}, 5\times10^{-4}\}$
Batch size	{32, 64, 128}
Optimizer	{Adam, AdamW}
Physics-informed weight λ_{slope}	{0, 0.1, 0.3, 1, 3, 10}
Physics-informed weight λ_{zipf}	{0, 0.1, 0.3, 1, 3, 10}
Physics-informed weight λ_{KL}	{0, 0.1, 0.3, 1, 3, 10}

Notes on physics-informed weights. The coefficients $\lambda_{\text{slope}}, \lambda_{\text{zipf}}, \lambda_{\text{KL}}$ control the relative strength of the distance-decay slope prior, Zipf-like scaling prior, and distributional consistency prior, respectively. Setting a coefficient to 0 recovers the purely data-driven objective, which is used both for baseline comparisons (e.g., *GravityGNN w/o physics*) and ablation studies. All three weights are tuned under the same inner-fold protocol to avoid biasing any specific prior.

A.3. Baselines

We compare PIG-GNN with representative OD forecasting methods spanning classical spatial interaction theory, non-graph machine learning, graph neural networks, and recent deep models. All baselines use the same input features (when applicable), the same preprocessing and target transformation defined in the main text, and the same training/evaluation protocol described in Appendix A.2. This keeps differences attributable to model design rather than data handling.

Unconstrained gravity model (UGM). UGM (Zipf, 1946) is a canonical spatial-interaction model in which flows are expressed as a function of origin and destination masses with a distance-impedance term. It serves as a theory-driven reference capturing global distance decay through a fixed functional form, and is calibrated on the training folds only.

Gravity model with PPML estimation (gravity PPML). Gravity PPML (Silva & Tenreyro, 2006) estimates the gravity specification using Poisson pseudo-maximum likelihood rather than log-linear regression. This formulation is widely used in flow modelling because it handles zero flows naturally and avoids the bias introduced by log-transforming sparse OD data. It therefore provides a stronger econometric gravity baseline under the same fold-wise training protocol.

Hurdle gravity model. The hurdle gravity model (Gu & Shen, 2021) uses a two-part formulation: it first models whether a flow exists between an OD pair, and then predicts the positive flow magnitude conditional on non-zero interaction. This design is better suited to sparse OD matrices with many zeros, and serves as a stronger classical baseline than single-stage gravity formulations when zero inflation is substantial.

Doubly constrained gravity model (DCGM). DCGM (Wilson, 1971) enforces both production and attraction constraints such that predicted origin outflows and destination inflows match observed marginals. This baseline reflects the classical production–attraction formulation and provides a strong interpretability benchmark, with constraints enforced using the training-fold marginals only.

Radiation model. The radiation model (Simini et al., 2012) predicts flows using intervening opportunities between OD pairs and does not require an explicit distance-decay parameter. It provides a parameter-light classical alternative to gravity-type impedance formulations and is fit under the same fold splits as other baselines.

Random forest (RF). RF (Breiman, 2001) is an ensemble of decision trees trained on engineered OD covariates. It provides a strong non-linear tabular baseline that does not rely on neural representations or graph message passing, and uses the same feature standardization applied to other non-graph models.

XGBoost. XGBoost (Chen & Guestrin, 2016) is a gradient-boosted tree model trained on the same OD covariates as RF. It is included as a competitive non-neural baseline to assess whether boosted non-linear fitting can match graph-based inductive biases under identical data splits.

Multilayer perceptron (MLP). The MLP baseline (Rumelhart et al., 1986) applies a feed-forward network to concatenated origin/destination features (and distance), removing graph structure entirely. This isolates the contribution of graph message passing and explicit OD dependence modelling, while keeping the same optimizer/early-stopping protocol as graph models.

Mixture density network (MDN). MDN (Bishop, 1994) predicts a conditional mixture distribution over flows, enabling likelihood-based learning that can accommodate heteroscedasticity and potential multi-modality. It provides a probabilistic neural comparison to deterministic regressors under the same training schedule.

Single-tower GNN. This baseline uses a conventional GNN with a shared encoder that maps origins and destinations into a single embedding space, followed by an OD predictor. It represents the common role-collapsed architecture that does not explicitly separate production from attraction, and uses the same OD graph construction and message-passing depth as the full model.

Dual-tower GCN. Dual-tower GCN (Wang et al., 2019) adopts separate GCN encoders for origins and destinations, preserving role separation while relying on standard adjacency-based propagation. In our implementation, this model follows the spirit of multi-graph GCN-based OD prediction while factorizing origin and destination representations into two role-specific towers. Compared with PIG-GNN, it excludes OD-level interaction modelling and physics-informed regularization, while keeping identical feature inputs and training protocol.

GMEL (geo-contextual). GMEL (Liu et al., 2020) is a geo-contextual embedding framework for commuting flow prediction that learns latent origin and destination representations from regional attributes and spatial context. It provides a strong graph-based urban-flow baseline by explicitly modelling geographic context and region similarity, while remaining distinct from our gravity-inspired interaction design.

GravityGNN (w/o physics-informed regularization). GravityGNN shares the same gravity-inspired backbone as PIG-GNN but trains with the prediction loss only (i.e., all physics-informed priors are disabled). This baseline isolates the contribution of the proposed scaling constraints beyond architectural choices, and is trained with the same early-stopping and model-selection rules.

Sat2Flow diffusion. Sat2Flow Diffusion (X. Wang et al., 2025) is a diffusion-based generative framework for flow prediction and is included as a representative recent generative baseline, using the same train/validation split and evaluation metrics.

Gravityformer. Gravityformer (Y. Wang et al., 2025) is a transformer-based OD modelling approach that incorporates gravity-related inductive biases via attention-based interaction modelling, serving as a strong deep baseline under the same data splits. In our comparison table, “Gravityformer (no P.I.)” denotes the base predictive variant without the additional physics-informed constraints used in PIG-GNN.

TransFlower. TransFlower (Luo et al., 2024) is an explainable transformer-based model for commuting flow prediction that introduces flow-to-flow attention to capture dependencies among OD interactions. It serves as a competitive recent attention-based baseline with stronger interaction modelling capacity than independent OD regressors, while remaining fully data-driven.

GAGNN. GAGNN (Rong et al., 2023) is a gravity-guided adversarial graph neural network for OD generation. It combines graph representation learning with gravity-inspired priors in a generative framework, making it a relevant recent deep baseline for structured flow prediction. Compared with our approach, GAGNN relies on adversarial/generative learning rather than explicit physics-informed regularization on the predictive objective (see Table A.7).

Implementation notes for non-graph baselines.

Table A.7
Hyperparameter search space for non-graph baselines. Final configurations are selected by random search under the protocol in Appendix A.2.

Model/component	Search range or options
RF: # trees	{200, 500, 1000}
RF: max depth	{None, 10, 20, 40}
RF: min samples split	{2, 5, 10}
RF: min samples leaf	{1, 2, 4}
RF: bootstrap	{True}
XGBoost: # boosting rounds	{500, 1000, 2000} with early stopping
XGBoost: learning rate (η)	{0.01, 0.05, 0.1}
XGBoost: max depth	{3, 5, 7, 10}
XGBoost: row subsample	{0.6, 0.8, 1.0}
XGBoost: column subsample	{0.6, 0.8, 1.0}
XGBoost: min child weight	{1, 5, 10}
XGBoost: ℓ_2 regularization (λ)	{0, 1, 5, 10}
XGBoost: ℓ_1 regularization (α)	{0, 0.5, 1}
MLP/MDN: hidden dimension	{64, 128, 256}
MLP/MDN: # layers	{2, 3, 4}
MLP/MDN: dropout	{0.0, 0.2, 0.3}
MLP/MDN: learning rate	{1×10^{-4}, 5×10^{-4}, 1×10^{-3}}
MLP/MDN: weight decay	{0, 1×10^{-5}, 1×10^{-4}}

A.4. Ablations

We perform build-up ablation studies to quantify the contribution of each component in PIG-GNN. All ablation variants share identical inputs, preprocessing, training protocol, and model-selection procedure described in Appendix A.2, so differences reflect architectural and loss-design choices rather than data handling.

Build-up protocol (reverse ablation). Starting from the MLP baseline, we add components sequentially to mirror the incremental gains reported in the main ablation table: (i) MLP (no graph structure), (ii) + adjacency GCN message passing with a single shared tower, (iii) + OD-level interaction message passing (competition/substitution), (iv) + dual-tower origin/destination separation, and (v) + physics-informed losses. This cumulative design makes the marginal benefit of each module explicit in the direction of model complexity.

Physics-informed losses (final step). The full model adds the three priors with adaptive weights $\lambda_{\text{slope}}, \lambda_{\text{zipf}}, \lambda_{\text{KL}}$. Setting these coefficients to 0 recovers the purely data-driven objective used in earlier build-up stages.

Evaluation consistency. All build-up variants are trained with the same early-stopping criterion and inner-fold random search budget as the full model. Metrics are computed on the same outer-fold test splits, and reported as mean ± standard deviation to match the main experimental tables.

A.5. Calculation details

To facilitate a transparent comparison of computational cost across models, Table A.8 summarizes the theoretical complexity, parameter scale, per-epoch runtime, and projected total training time under a unified setting of 250 epochs. The comparison is intended to provide a normalized estimate of relative computational burden rather than a strict hardware-independent benchmark, as actual runtime may vary with implementation details, batching strategy, and backend optimization. Overall, PIG-GNN belongs to the higher-complexity model family due to its dual-tower graph encoder and explicit edge interaction module, but its runtime remains substantially lower than transformer-based baselines and within a practical range for large-scale OD modelling.

All experiments were conducted under the same computing environment for consistency and reproducibility, as summarized in Table A.9.

A.6. Generalizability on a large-scale ALLOD dataset

To further evaluate whether the proposed PIG-GNN framework generalizes beyond the commuter-specific setting studied in the main paper, we conduct an additional experiment on a large-scale ALLOD dataset. Unlike the main Home-to-Work (H2W) dataset, which only captures commuting flows, the ALLOD dataset includes trips for *all* purposes under the same spatial units and input configuration. This provides a more challenging and realistic test of generalization, as the underlying mobility patterns are substantially more heterogeneous.

As summarized in Table A.10, the H2W and ALLOD datasets share the same node set (6856 MSOAs), but differ markedly in edge density and flow distribution. In particular, ALLOD contains approximately eight times more edges than H2W (2.38M vs. 298K), with a nearly ten times larger mean flow, a 41 times larger flow standard deviation, and a 93 times larger maximum flow. These pronounced distributional shifts make ALLOD a substantially stricter benchmark for assessing whether the model captures generalizable mobility mechanisms rather than overfitting to commuter-specific characteristics.

Evaluation protocol. To ensure fair comparison with the main experiments, we follow the same 5-fold stratified cross-validation protocol and use the same evaluation metrics. All baseline methods use the same hyperparameter settings and training configurations as in the H2W experiments, with only minor adjustments to batch size and training epochs to accommodate the larger graph. PIG-GNN is evaluated with the same architecture and physics-informed loss formulation as described in the main paper.

To address baseline consistency, Table A.11 expands the ALLOD benchmark to better align with the deployable baseline set used in the main H2W experiments. Specifically, we add XGBoost, unconstrained gravity, the radiation model, dual-tower GCN, dual-tower GCN with residual connection, and gravity PPML, alongside the models already reported in the original ALLOD comparison. The doubly-constrained gravity model remains excluded because its iterative balancing procedure is prohibitively slow on the large-scale ALLOD graph with approximately 2.4M edges.

Table A.11 reports the expanded comparison results on the ALLOD dataset. PIG-GNN achieves the best RMSE (0.6006), MAE (0.4388), and R^2 (0.7431). Compared with the next-best deployable model, XGBoost, PIG-GNN reduces RMSE from 0.6412 to 0.6006 (6.3% improvement), reduces MAE from 0.4645 to 0.4388 (5.5% improvement), and improves R^2 from 0.7073 to 0.7431 (5.1% relative improvement). Under this heavy-tailed multi-purpose regime, Random Forest attains the highest CPC (0.6748), followed closely by XGBoost (0.6717), while PIG-GNN remains competitive on CPC (0.6076). Overall, the expanded ALLOD benchmark shows that PIG-GNN retains substantially stronger point-wise predictive accuracy under a much larger and more heterogeneous OD system, while its CPC remains reasonable despite not being the highest metric.

These results indicate that the gains of PIG-GNN are not limited to the H2W commuter-flow scenario. Instead, the proposed inductive biases — including the dual-tower adjacency encoder, the explicit edge interaction module, and the gravity-informed physics loss — remain effective when the model is exposed to a denser graph and much more dispersed flow distributions. Overall, this experiment provides additional evidence that PIG-GNN does not merely fit the specific statistical characteristics of the H2W dataset, but generalizes to larger, more complex, and more heterogeneous origin–destination prediction tasks.

A.7. Production and attraction proxy robustness

This appendix reports a controlled feature-engineering exercise that probes how predictive performance depends on the granularity of production and attraction inputs. Using publicly available spatial attributes

Table A.8
Computational complexity comparison of baseline models and PIG-GNN. Epochs are unified to 250 for normalized runtime projection.

Model	Backbone	Time complexity	Params (K)	Epochs	Complexity score	Time/epoch (s)	Est. total time (s)
Edge MLP	MLP	$\mathcal{O}(E)$	2	250	1	0.108	26.9
Gravity PPML	Linear (3→1)	$\mathcal{O}(E)$	0	250	1	0.018	4.4
Hurdle	Logistic + Poisson	$\mathcal{O}(E)$	0	250	1	0.007	1.8
Radiation Model	LinearRegression (CPU)	$\mathcal{O}(E)$	0	–	1	N/A	0.0
Edge MDN	MLP (MDN)	$\mathcal{O}(E)$	3	250	2	0.026	6.5
GAGNN	GAT + MLP	$\mathcal{O}(N+E\_adj)$	2	250	2	0.346	86.5
Random Forest	RandomForest (CPU)	$\mathcal{O}(E \cdot trees)$	0	–	2	N/A	1.3
XGBoost	XGBRegressor (GPU)	$\mathcal{O}(E \cdot trees)$	0	–	2	N/A	0.3
GCN + Edge MLP	GCN (adj)	$\mathcal{O}(N+E\_adj)$	53	250	3	0.032	7.9
GATv2 + Edge MLP	GATv2 (adj)	$\mathcal{O}(N+E\_adj)$	62	250	3	0.020	4.9
GMEL	GATv2 + GBRT (GPU)	$\mathcal{O}(N+E\_adj)$	85	250	3	0.145	36.1
SIM-GAT	GAT-style	$\mathcal{O}(E)$	59	250	4	0.013	3.3
Dual-Tower GCN	GCN (adj)	$\mathcal{O}(N+E\_adj)$	71	250	4	0.035	8.7
Edge Diffusion	MLP + Diffusion	$\mathcal{O}(E \times T)$	2	250	6	0.003	0.7
Dual-Tower GATv2 (OD)	Dual-Tower GATv2 (OD)	$\mathcal{O}(N+E\_od)$	3	250	7	0.012	3.1
Dual-Tower GCN + OD MLP	GCN + GATv2 (OD)	$\mathcal{O}(N+E\_adj+E\_od)$	105	250	7	0.044	10.9
PIG-GNN (Ours)	**Dual-Tower GATv2 (adj) + Edge Interaction**	$\mathcal{O}(N+E\_adj+E\_od)$	**678**	**250**	**7**	**0.139**	**34.7**
TransFlower	Transformer	$\mathcal{O}(E^2)$	282	250	8	0.372	93.0
Gravity Transformer	Transformer	$\mathcal{O}(E^2)$	836	250	9	2.237	559.2

Table A.9
Computing environment for reproducibility.

Item	Specification
OS/Kernel	Linux 5.15.0-163-generic
CPU	AMD EPYC 7542 32-Core Processor (2 sockets)
Logical CPUs	128
System Memory	503 GiB RAM
GPU	NVIDIA GeForce RTX 4090 × 8
GPU Memory	24,564 MiB per GPU
NVIDIA Driver	570.124.04
Python	3.12.8
PyTorch	2.7.1+cu118
CUDA (PyTorch)	11.8

Table A.10
Dataset statistics for H2W and ALLOD. Although the two datasets share the same spatial units (MSOAs), ALLOD is substantially larger and exhibits much heavier-tailed flow distributions, making it suitable for evaluating model generalizability.

Metric	H2W	ALLOD	Ratio (ALL/H2W)
Nodes	6856	6856	1.00
Edges	298,681	2383,725	7.98
Flow mean	1.73	17.14	9.89
Flow std	2.00	82.79	41.33
Flow max	74	6905	93.31
Distance mean	37 419.11	42 259.64	1.13

aligned to MSOAs, we construct three enriched proxy specifications: *rich_simple*, with 5 production and 5 attraction variables including population density, full-time employment share, commute-mode shares, POI counts, transport POI counts, and night-light intensity; *strict_dir*, with 4 production and 6 attraction variables using stricter residential/employment splits; and *pca*, with two principal components per side derived from the richer variable sets. All variants use the same final PIG-GNN architecture, 5-fold stratified cross-validation, GATv2 dual towers, AdamW optimizer, cosine learning-rate schedule, and 800 training epochs as in the main experiments. Only the production/attraction input dimensionality differs. Table A.12 reports RMSE and CPC for each variant.

The enriched proxy specifications consistently improve over the original coarse-proxy baseline. The best variant, *rich_simple*, reduces RMSE from 0.2941 to 0.2812 and increases CPC from 0.7966 to 0.8080. The *strict_dir* and *pca* variants also improve both metrics. These results support the reviewer's concern that finer-grained production and attraction proxies are beneficial, while also showing that the PIG-GNN architecture can absorb richer physical inputs without changing the model design.

A.8. Zero-flow augmentation ablation

This appendix documents a controlled augmentation of the H2W OD supervision set in which 10% randomly sampled zero-flow OD pairs are appended to each fold's training and test graphs alongside observed positive edges. The goal is not to replace the sparse OD prediction task with a fully dense $N \times N$ zero-inflated model, but to test whether PIG-GNN can accommodate explicit zero-flow supervision. We compare PIG-GNN with unconstrained gravity on overall, non-zero-only, and zero-only evaluation splits under the same 5-fold cross-validation protocol. On the zero-only subset, all ground-truth flows are constant at zero, so R^2 and CPC are not substantively informative and are reported as N/A.

The results in Table A.13 show that PIG-GNN handles sampled zero-flow pairs substantially better than the gravity baseline. On the zero-only split, PIG-GNN achieves RMSE 0.3519 ± 0.0154, compared with 0.7657 ± 0.0007 for Gravity. On non-zero edges, PIG-GNN remains more accurate than Gravity, with RMSE 0.3131 ± 0.0016 versus 0.3384 ± 0.0014. This suggests that the neural architecture can learn near-zero predictions while retaining reasonable accuracy on observed positive OD flows. A full zero-inflated framework with separate link-formation and conditional flow-intensity components remains future work.

A.9. Distance-binning error analysis

Table A.14 reports RMSE, MAE, R^2, and CPC by distance bin on the H2W test set for unconstrained gravity, Random Forest, DualTower, and PIG-GNN. Distance bins are defined from Euclidean inter-MSOA centroid distances in kilometres and partitioned into quintile-based intervals on the pooled test folds. Metrics are reported as mean ± standard deviation across 5-fold cross-validation.

A.10. Physics-informed parameter sensitivity

This appendix complements the component-wise ablation on physics-loss weights by reporting a sensitivity sweep over the centres of the soft β- and α-interval priors. We fix the training protocol using one

Table A.11
Generalization of PIG-GNN on the large-scale ALLOD dataset. Evaluation metrics are RMSE and MAE (lower is better), and R^2 and CPC (higher is better). Results are reported as mean ± standard deviation across 5-fold stratified cross-validation. The best value for each metric is shown in **bold** and the second-best value is underlined. The doubly-constrained gravity model is omitted on this dataset due to prohibitive runtime on approximately 2.4M edges.

Model	RMSE ↓	MAE ↓	R^2 ↑	CPC ↑
PIG-GNN (Ours)	**0.6006±0.0079**	**0.4388±0.0048**	**0.7431±0.0069**	0.6076±0.0070
XGBoost	0.6412±0.0002	0.4645±0.0002	0.7073±0.0004	0.6717±0.0010
Random Forest	0.6439±0.0004	0.4672±0.0002	0.7049±0.0004	**0.6748±0.0010**
GravTransf	0.6457±0.0076	0.4789±0.0082	0.7032±0.0068	0.5666±0.0333
TransFlower	0.6679±0.0259	0.4962±0.0225	0.6820±0.0246	0.5744±0.0773
GMEL+GBRT	0.7057±0.0072	0.5104±0.0050	0.6454±0.0073	0.5861±0.0052
EdgeDiff	0.7540±0.0239	0.5495±0.0211	0.5948±0.0264	0.5221±0.0637
EdgeMDN	0.7711±0.0399	0.5344±0.0385	0.5755±0.0447	0.4520±0.1339
Unconstrained Gravity	0.7732±0.0007	0.5888±0.0003	0.5743±0.0007	0.5546±0.0022
Radiation Model	0.7739±0.0007	0.5888±0.0003	0.5736±0.0008	0.5549±0.0023
DualTower	0.8537±0.0554	0.5825±0.0227	0.4790±0.0684	0.4676±0.0964
Dual-Tower GCN	0.9588±0.0588	0.5883±0.0270	0.3431±0.0792	0.4908±0.0140
Dual-Tower GCN (Residual)	0.9707±0.0148	0.5983±0.0081	0.3290±0.0199	0.4880±0.0102
Hurdle	1.0984±0.0023	0.9881±0.0028	0.1410±0.0034	0.5784±0.0016
Gravity PPML	1.2858±0.2790	1.1594±0.2586	−0.2320±0.5796	0.4967±0.1319

Table A.12
Feature-engineering exploration for production/attraction proxies with the final PIG-GNN architecture. Results are reported over 5-fold cross-validation with 800 training epochs. All variants share the same architecture and hyperparameters; only the production/attraction input dimensions differ.

Variant	P/A dimensions	RMSE ↓	CPC ↑
baseline	1 production + 1 attraction (working-age pop, POI density)	0.2941±0.0004	0.7966±0.0008
rich_simple	5 production + 5 attraction (pop density, full-time share, commute-mode shares, POI counts, transport POI, night-light intensity)	**0.2812±0.0011**	**0.8080±0.0008**
strict_dir	4 production + 6 attraction (stricter residential/employment splits)	0.2816±0.0011	0.8078±0.0009
pca	2 PCA components (production) + 2 PCA components (attraction)	0.2864±0.0007	0.8031±0.0005

Table A.13
Zero-flow ablation on the H2W dataset. Ten percent randomly sampled zero-flow OD pairs are added to the training and test sets. Results are reported as mean ± standard deviation across 5-fold cross-validation. On the zero-only subset, R^2 and CPC are not informative because all ground-truth flows are zero, and are therefore reported as N/A.

Model	Split	RMSE ↓	MAE ↓	R^2 ↑	CPC ↑
	Overall	0.3169±0.0014	0.2148±0.0038	0.5163±0.0055	0.7726±0.0026
PIG-GNN	Non-zero only	0.3131±0.0016	0.2037±0.0040	0.3700±0.0062	0.7821±0.0027
	Zero only	0.3519±0.0154	0.3262±0.0188	N/A[a]	N/A[a]
	Overall	0.3967±0.0010	0.2675±0.0004	0.2419±0.0025	0.7288±0.0008
Gravity	Non-zero only	0.3384±0.0014	0.2177±0.0005	0.2642±0.0028	0.7561±0.0009
	Zero only	0.7657±0.0007	0.7652±0.0007	N/A[a]	N/A[a]

[a] On the zero-only subset, all ground-truth flows are constant at zero, so R^2 and CPC are not informative and are therefore reported as N/A.

validation fold, 400 epochs, and fixed physics-loss weights, and centre a narrow admissible band $[c - 0.03, c + 0.03]$ on alternative candidate coefficients c. This design directly tests whether the model is sensitive to misspecification of the prior interval itself, rather than only to the scalar loss weights. Table A.15 reports RMSE, MAE, R^2, and CPC for each configuration, and Fig. A.14 visualizes the corresponding univariate and joint sensitivity patterns.

The sensitivity results show two main patterns. First, performance changes smoothly rather than discontinuously as the prior centres move across the tested range. RMSE is lowest under low-to-moderate exponent settings, while CPC peaks around empirically plausible distance-decay and Zipf exponents. Second, even relatively large shifts of the admissible interval, such as $\beta = 3.0$ or $\alpha = 3.0$, lead only to modest RMSE degradation and retain CPC values above 0.784. These results indicate that the physics-informed priors act as soft inductive biases rather than brittle hard constraints: they improve alignment when consistent with the data-generating process, but the neural predictor remains robust under moderate prior misspecification.

Table A.14
Distance-binning error analysis on the H2W test set. Distance bins are defined using Euclidean inter-MSOA centroid distances in kilometres and partitioned into quintile-based intervals on the pooled test folds. Metrics are reported as mean ± standard deviation across 5-fold cross-validation. Within each distance bin, the best value for each metric is shown in bold and the second-best value is underlined.

Distance bin	Model	RMSE ↓	MAE ↓	R^2 ↑	CPC ↑
[0.0, 1.8] km	Gravity	$0.5646_{\pm 0.0021}$	$0.4516_{\pm 0.0012}$	$0.0932_{\pm 0.0015}$	$0.6513_{\pm 0.0016}$
	Random Forest	$0.5518_{\pm 0.0026}$	$0.4441_{\pm 0.0017}$	$0.1340_{\pm 0.0045}$	$0.6644_{\pm 0.0017}$
	DualTower	$0.5468_{\pm 0.0050}$	$0.4392_{\pm 0.0045}$	$0.1496_{\pm 0.0153}$	$0.6683_{\pm 0.0030}$
	PIG-GNN	**$0.4988_{\pm 0.0028}$**	**$0.3980_{\pm 0.0026}$**	**$0.2925_{\pm 0.0062}$**	**$0.6988_{\pm 0.0029}$**
[1.8, 6.2] km	Gravity	$0.3405_{\pm 0.0007}$	$0.2795_{\pm 0.0002}$	$0.0193_{\pm 0.0008}$	$0.7449_{\pm 0.0003}$
	Random Forest	$0.3352_{\pm 0.0012}$	$0.2677_{\pm 0.0004}$	$0.0497_{\pm 0.0055}$	$0.7525_{\pm 0.0004}$
	DualTower	$0.3312_{\pm 0.0014}$	$0.2697_{\pm 0.0062}$	$0.0723_{\pm 0.0076}$	$0.7555_{\pm 0.0014}$
	PIG-GNN	**$0.3098_{\pm 0.0011}$**	**$0.2388_{\pm 0.0014}$**	**$0.1883_{\pm 0.0055}$**	**$0.7764_{\pm 0.0004}$**
[6.2, 15.4] km	Gravity	$0.2419_{\pm 0.0007}$	$0.2002_{\pm 0.0002}$	$-0.0291_{\pm 0.0034}$	$0.8087_{\pm 0.0004}$
	Random Forest	$0.2330_{\pm 0.0008}$	$0.1670_{\pm 0.0005}$	$0.0456_{\pm 0.0033}$	$0.8302_{\pm 0.0005}$
	DualTower	$0.2387_{\pm 0.0045}$	$0.1947_{\pm 0.0103}$	$-0.0019_{\pm 0.0375}$	$0.8138_{\pm 0.0061}$
	PIG-GNN	**$0.2225_{\pm 0.0015}$**	**$0.1497_{\pm 0.0008}$**	**$0.1293_{\pm 0.0067}$**	**$0.8442_{\pm 0.0008}$**
[15.4, 50.9] km	Gravity	$0.1672_{\pm 0.0016}$	$0.1223_{\pm 0.0004}$	$-0.0430_{\pm 0.0038}$	$0.8759_{\pm 0.0008}$
	Random Forest	**$0.1545_{\pm 0.0017}$**	$0.0878_{\pm 0.0006}$	**$0.1098_{\pm 0.0071}$**	$0.9048_{\pm 0.0010}$
	DualTower	$0.1734_{\pm 0.0049}$	$0.1319_{\pm 0.0098}$	$-0.1243_{\pm 0.0819}$	$0.8682_{\pm 0.0075}$
	PIG-GNN	$0.1560_{\pm 0.0022}$	**$0.0848_{\pm 0.0017}$**	$0.0920_{\pm 0.0123}$	**$0.9067_{\pm 0.0018}$**
[50.9, 561.2] km	Gravity	$0.1400_{\pm 0.0015}$	$0.0710_{\pm 0.0009}$	$-0.2992_{\pm 0.0084}$	$0.9223_{\pm 0.0010}$
	Random Forest	**$0.1201_{\pm 0.0009}$**	$0.0633_{\pm 0.0005}$	**$0.0433_{\pm 0.0007}$**	$0.9322_{\pm 0.0005}$
	DualTower	$0.1333_{\pm 0.0075}$	$0.0785_{\pm 0.0083}$	$-0.1812_{\pm 0.1302}$	$0.9172_{\pm 0.0075}$
	PIG-GNN	$0.1228_{\pm 0.0011}$	**$0.0583_{\pm 0.0012}$**	$-0.0003_{\pm 0.0031}$	**$0.9363_{\pm 0.0009}$**

Table A.15
Narrow-interval physics-prior sensitivity on H2W (1-fold, 400 epochs, fixed λ). Intervals are $[c-0.03, c+0.03]$ around the listed centre c. Within each parameter group (β and α), the best value for each metric is shown in **bold** and the second-best value is underlined.

Configuration	RMSE ↓	MAE ↓	R^2 ↑	CPC ↑
No physics ($\lambda_s = 0$, $\lambda_z = 0$)	0.2974	0.1871	0.4044	0.7936
β-centred intervals				
$\beta = 0.05 \pm 0.03$	0.2973	0.1868	0.4050	0.7937
$\beta = 0.10 \pm 0.03$	**0.2972**	0.1867	**0.4053**	0.7938
$\beta = 0.15 \pm 0.03$	0.2974	0.1855	0.4046	0.7949
$\beta = 0.20 \pm 0.03$	0.2980	0.1834	0.4022	0.7962
$\beta = 0.25 \pm 0.03$[a]	0.2980	**0.1824**	0.4022	0.7972
$\beta = 0.30 \pm 0.03$[a]	0.2989	0.1825	0.3989	**0.7973**
$\beta = 0.40 \pm 0.03$	0.3005	0.1843	0.3927	0.7965
$\beta = 0.50 \pm 0.03$	0.3026	0.1871	0.3846	0.7949
$\beta = 0.70 \pm 0.03$	0.3068	0.1913	0.3683	0.7887
$\beta = 1.00 \pm 0.03$	0.3066	0.1943	0.3689	0.7865
$\beta = 1.50 \pm 0.03$	0.3070	0.1954	0.3673	0.7862
$\beta = 2.00 \pm 0.03$	0.3073	0.1957	0.3662	0.7870
$\beta = 3.00 \pm 0.03$	0.3080	0.1970	0.3633	0.7848
α-centred intervals				
$\alpha = 0.20 \pm 0.03$	0.2978	0.1900	0.4030	0.7913
$\alpha = 0.30 \pm 0.03$	**0.2971**	0.1875	**0.4057**	0.7936
$\alpha = 0.40 \pm 0.03$	0.2976	0.1829	0.4036	0.7976
$\alpha = 0.50 \pm 0.03$[a]	0.2991	**0.1797**	0.3981	**0.7996**
$\alpha = 0.60 \pm 0.03$[a]	0.3024	0.1828	0.3854	0.7978
$\alpha = 0.70 \pm 0.03$	0.3060	0.1904	0.3714	0.7895
$\alpha = 1.00 \pm 0.03$	0.3064	0.1955	0.3699	0.7866
$\alpha = 1.50 \pm 0.03$	0.3070	0.1947	0.3675	0.7880
$\alpha = 2.00 \pm 0.03$	0.3075	0.1945	0.3655	0.7887
$\alpha = 3.00 \pm 0.03$	0.3074	0.1963	0.3659	0.7866

[a] Indicates best configurations near the empirically observed coefficients.

A.11. Spatial adjacency definition sensitivity

This appendix formalizes the claim in Section 3.1.2 that neighbourhood definition has limited impact on PIG-GNN under our protocol. We retrain the final architecture with identical features, losses, and cross-validation folds, replacing first-order planar contiguity (1HOP) with 4-nearest-neighbour, 6-nearest-neighbour, 8-nearest-neighbour, and radius-based adjacency graphs constructed from MSOA centroids. Table A.16 reports RMSE, MAE, R^2, and CPC for each adjacency definition.

A.12. Xi'an cross-city validation

To assess transferability to an inland mono-centric city with a distinct orthogonal street grid and ancient walled core, we trained PIG-GNN on 500m-grid OD data from Xi'an, China. Using residential population, workplace population, and POI density as production and attraction proxies, we tested PIG-GNN against five strong baselines: Gravity, Random Forest, XGBoost, TransFlower, and GravTransf. Table A.17 reports the 5-fold CV results; PIG-GNN outperforms all competitors across every metric.

Fig. A.15 shows the study area at 500 m grid resolution. Fig. A.16 compares model performance across RMSE, R^2, and CPC. The fitted scaling laws (Fig. A.17) confirm that both priors hold in Xi'an: the distance-decay exponent β=0.425 (R^2=0.822) and the Zipf rank-frequency exponent α=0.582 (R^2=0.900) both fall within the paper's empirically validated prior ranges.

A.13. Guangzhou cross-city validation

To assess transferability to a coastal polycentric mega-city with Pearl River Delta integration and subtropical climate, we trained PIG-GNN on 500m-grid OD data from Guangzhou, China. Across 9768 grid cells and 123K OD edges, PIG-GNN again outperforms all five baselines across every metric (Table A.18).

Fig. A.18 shows the study area. Fig. A.19 compares model performance. The fitted scaling laws (Fig. A.20) confirm that both priors hold in Guangzhou as well: β=0.327 (R^2=0.828) and α=1.066 (R^2=0.824), both within the empirically validated prior ranges.

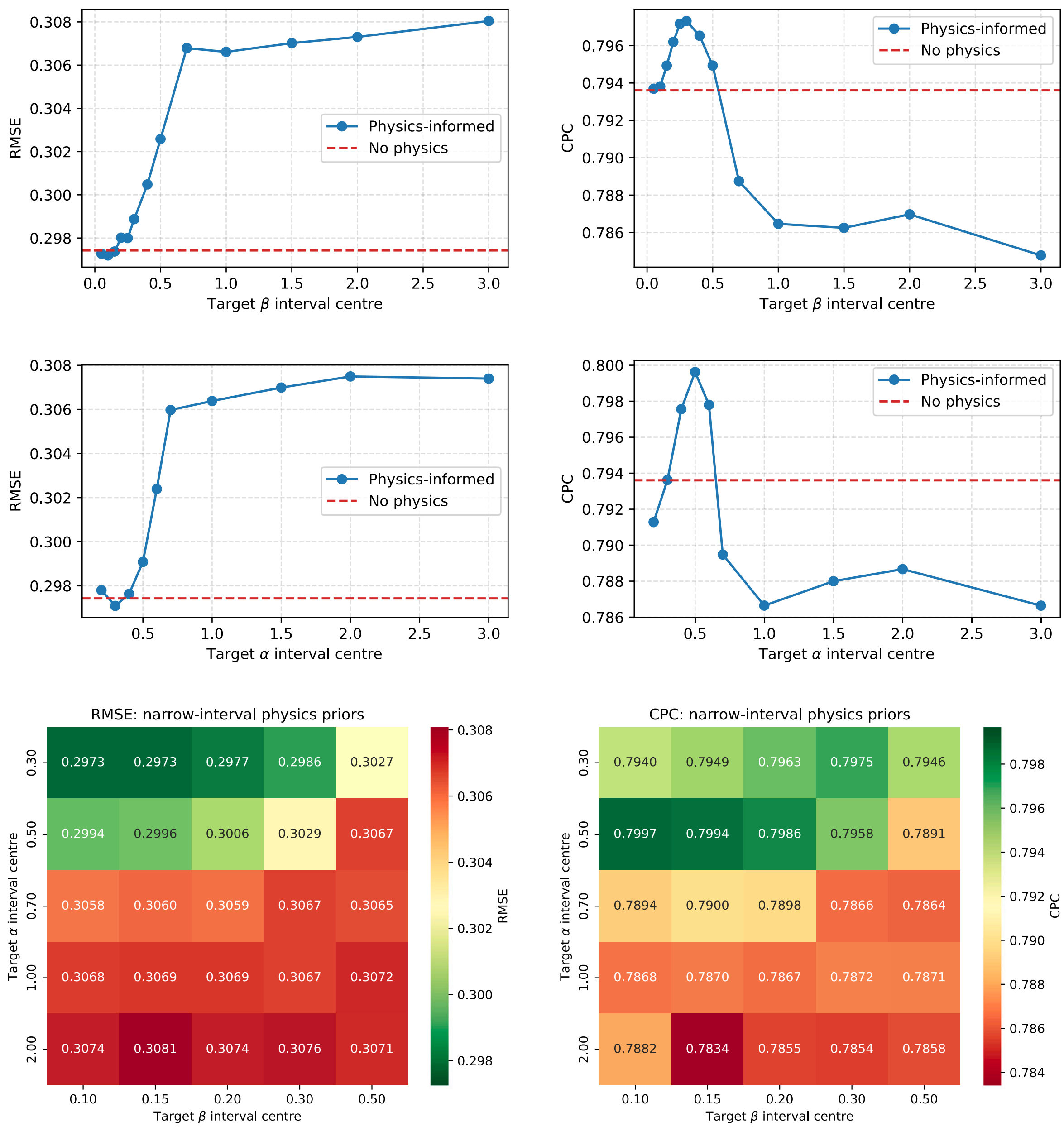

Fig. A.14. Sensitivity to narrow-interval physics priors ($[c - 0.03, c + 0.03]$). Dashed red lines mark the no-physics baseline. The top row shows the β sweep, the middle row the α sweep, and the bottom row the 2D heatmaps over the joint interval grid.

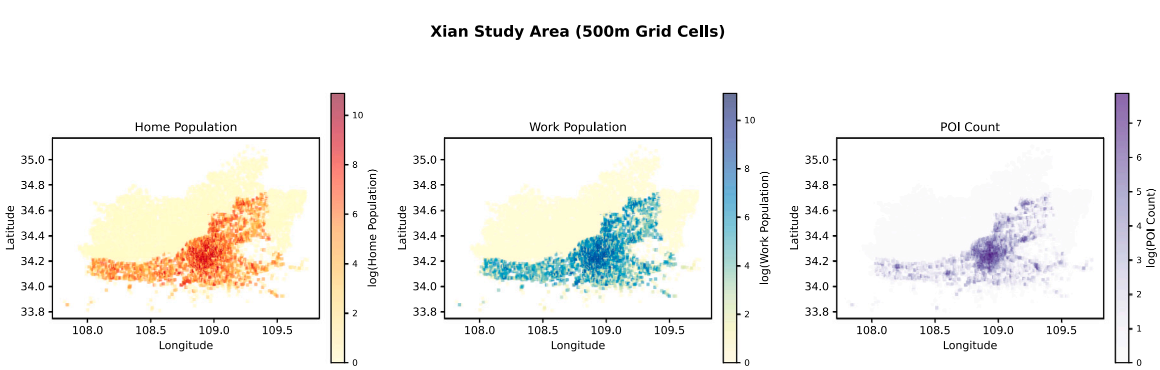

Fig. A.15. Xi'an study area at 500 m grid resolution (from left to right: log residential population, log workplace population, and log POI count).

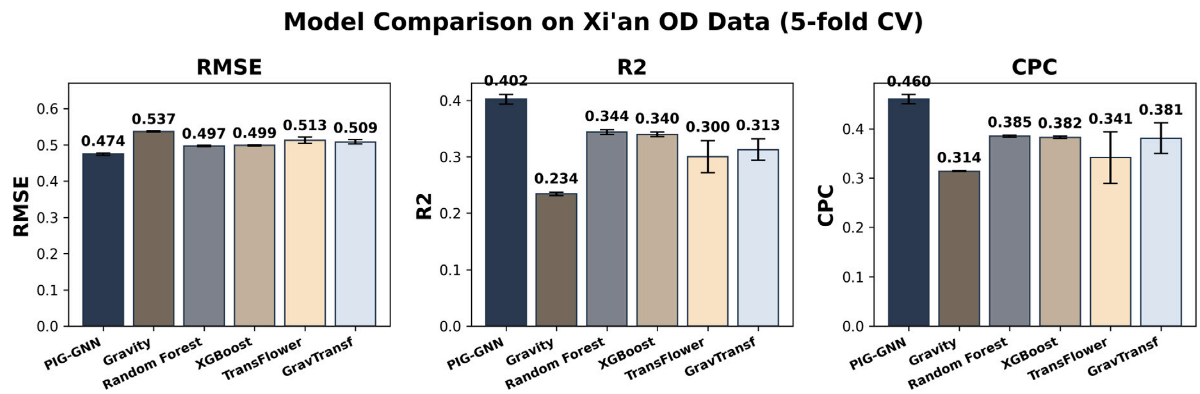

Fig. A.16. Model comparison on Xi'an 500m-grid OD data (5-fold CV).

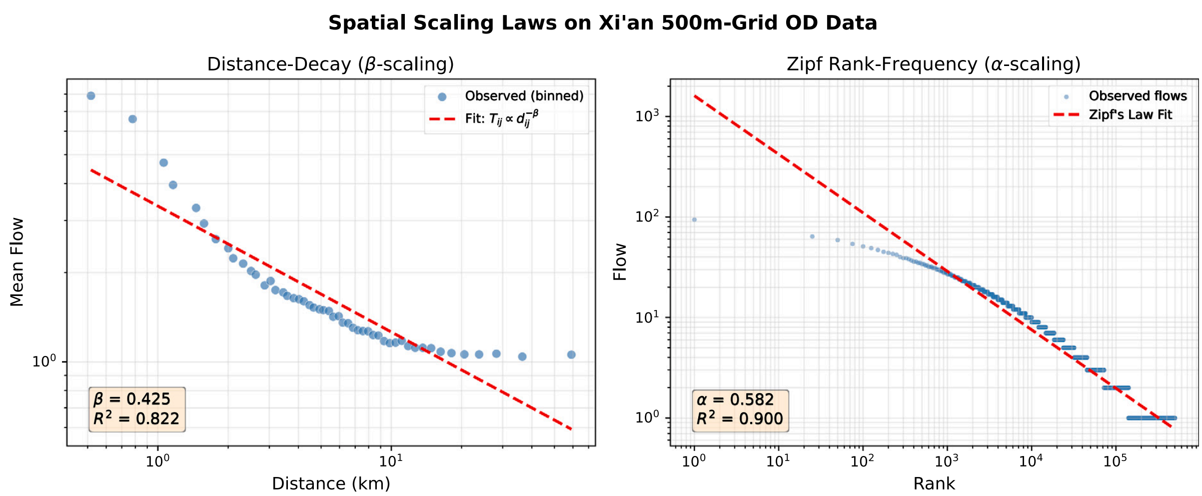

Fig. A.17. Spatial scaling laws on Xi'an 500m-grid OD data. Left: distance-decay (β-scaling, β=0.425, R^2=0.822). Right: Zipf rank-frequency (α-scaling, α=0.582, R^2=0.900).

Table A.16
Adjacency graph sensitivity analysis for PIG-GNN on H2W. Results are reported as mean ± standard deviation across 5-fold cross-validation. The default first-order planar-contiguity graph is compared with 4-NN, 6-NN, 8-NN, and radius-based adjacency graphs constructed from MSOA centroids. Best values are shown in bold and second-best values are underlined. *Note:* This is a controlled adjacency-sensitivity rerun under identical settings within this sweep; values are intended to compare neighbourhood definitions rather than to replace the main benchmark in Table 3.

Adjacency	RMSE ↓	MAE ↓	R^2 ↑	CPC ↑
1HOP	**0.3025$_{\pm 0.0027}$**	0.1916$_{\pm 0.0019}$	**0.3968$_{\pm 0.0114}$**	0.7909$_{\pm 0.0016}$
4-NN	0.3027$_{\pm 0.0038}$	0.1909$_{\pm 0.0029}$	0.3961$_{\pm 0.0152}$	**0.7914$_{\pm 0.0020}$**
6-NN	0.3029$_{\pm 0.0039}$	0.1920$_{\pm 0.0043}$	0.3954$_{\pm 0.0160}$	0.7901$_{\pm 0.0030}$
8-NN	0.3026$_{\pm 0.0029}$	0.1913$_{\pm 0.0029}$	0.3965$_{\pm 0.0120}$	0.7909$_{\pm 0.0020}$
RADIUS	0.3033$_{\pm 0.0028}$	**0.1904$_{\pm 0.0021}$**	0.3937$_{\pm 0.0116}$	0.7902$_{\pm 0.0015}$

Table A.17
Cross-city validation on 500m-grid Xi'an OD data.

Model	RMSE ↓	R^2 ↑	CPC ↑
Gravity	0.5368 ± 0.0017	0.2342 ± 0.0028	0.3139 ± 0.0010
TransFlower	0.5211 ± 0.0224	0.2774 ± 0.0615	0.3110 ± 0.0640
GravTransf	0.5064 ± 0.0068	0.3186 ± 0.0160	0.3832 ± 0.0349
Random Forest	0.4970 ± 0.0019	0.3436 ± 0.0040	0.3849 ± 0.0020
XGBoost	0.4972 ± 0.0018	0.3431 ± 0.0036	0.3907 ± 0.0028
PIG-GNN	**0.4744 ± 0.0033**	**0.4019 ± 0.0084**	**0.4597 ± 0.0095**

Note: Results are reported as mean ± standard deviation over 5-fold CV.

Table A.18
Cross-city validation on 500m-grid Guangzhou OD data.

Model	RMSE ↓	R^2 ↑	CPC ↑
Gravity	1.0441 ± 0.0015	0.2088 ± 0.0049	0.5454 ± 0.0023
Random Forest	0.9858 ± 0.0021	0.2946 ± 0.0042	0.6077 ± 0.0022
TransFlower	0.9632 ± 0.0147	0.3264 ± 0.0207	0.6169 ± 0.0255
XGBoost	0.9812 ± 0.0024	0.3011 ± 0.0046	0.6119 ± 0.0021
GravTransf	0.9316 ± 0.0025	0.3700 ± 0.0053	0.6496 ± 0.0016
PIG-GNN	**0.9089 ± 0.0040**	**0.4004 ± 0.0067**	**0.6511 ± 0.0045**

Note: Results are reported as mean ± standard deviation over 5-fold CV.

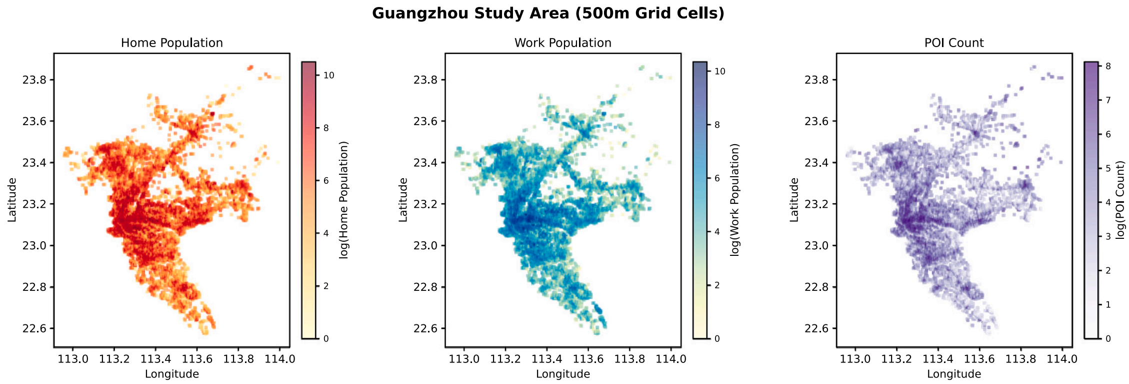

Fig. A.18. Guangzhou study area at 500 m grid resolution (from left to right: log residential population, log workplace population, and log POI count).

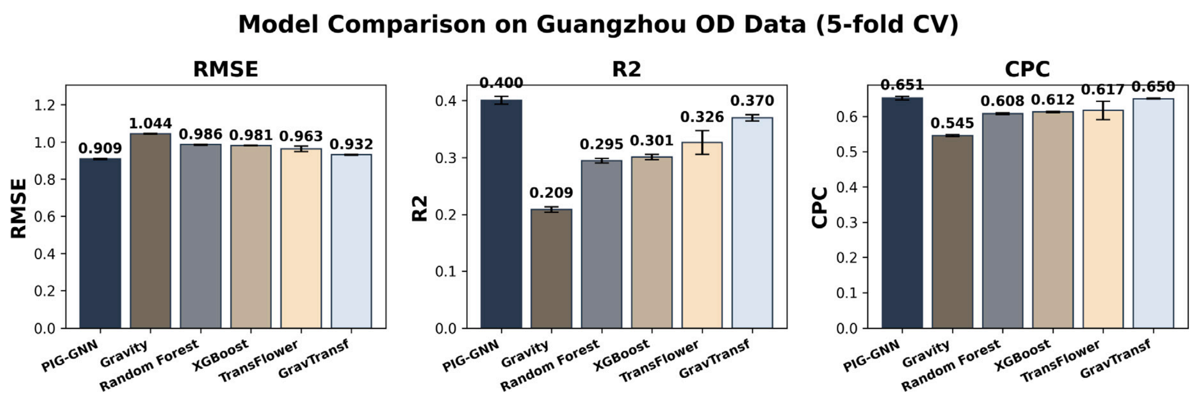

Fig. A.19. Model comparison on Guangzhou 500m-grid OD data (5-fold CV).

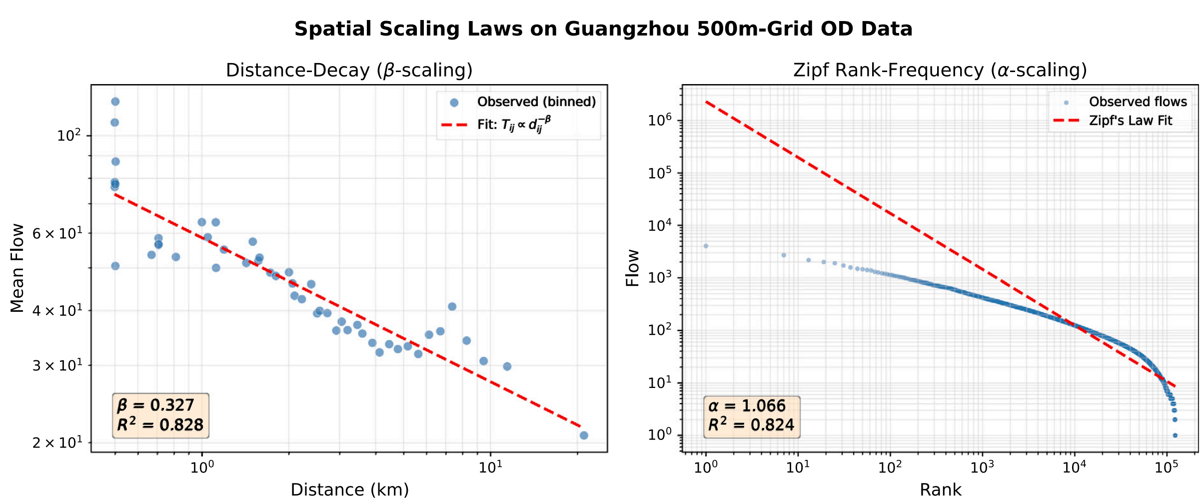

Fig. A.20. Spatial scaling laws on Guangzhou 500m-grid OD data. Left: distance-decay (β-scaling, β=0.327, R^2=0.828). Right: Zipf rank-frequency (α-scaling, α=1.066, R^2=0.824).

Data availability

Data will be made available on request.

References

Abrahamsson, T. (1998). Estimation of origin-destination matrices using traffic counts-a literature survey.

Alessandretti, L., Aslak, U., & Lehmann, S. (2020). The scales of human mobility. *Nature*, *587*, 402–407.

Altintasi, O., & Kirtiloglu, O. S. (2026). Data-driven urban mobility planning: A gis-based spatial modeling approach for shared mobility hub placement. *Sustainable Cities and Society*, Article 107261.

Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., & Batty, M. (2015). Constructing cities, deconstructing scaling laws. *Journal of the Royal Society Interface*, *12*, Article 20140745.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., & Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1–74.

Batty, M. (2008). The size, scale, and shape of cities. *Science*, *319*, 769–771.

Batty, M. (2013). *The New Science of Cities*. MIT Press.

Berg, R. v. d., Kipf, T. N., & Welling, M. (2017). Graph convolutional matrix completion. arXiv preprint arXiv:1706.02263.

Betkier, I., Oszczypała, M., & Dawid, W. (2025). Bike-sharing systems for sustainable cities: Demand prediction using graph neural networks and spatial aggregation. *Sustainable Cities and Society*, *128*, Article 106479.

Bettencourt, L. M. (2013). The origins of scaling in cities. *Science*, *340*, 1438–1441.

Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, *104*, 7301–7306.

Biscaia, R., & Mota, I. (2013). Models of spatial competition: A critical review. *Papers in Regional Science*, *92*, 851–872.

Bishop, C. M. (1994). Mixture density networks.

Boucherie, L., Maier, B. F., & Lehmann, S. (2025). Decoupling geographical constraints from human mobility. *Nature Human Behaviour*, 1–12.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*, 462–465.

Bui, K. H. N., Cho, J., & Yi, H. (2022). Spatial-temporal graph neural network for traffic forecasting: An overview and open research issues. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *52*, 2763–2774.

Burger, M., & Meijers, E. (2012). Form follows function? linking morphological and functional polycentricity. *Urban Studies*, *49*, 1127–1149.

Cadwallader, M. (2013). Migration and intra-urban mobility. In *Population geography: progress & prospect (Routledge Revivals)* (pp. 257–283). Routledge.

Chen, Y. (2015). The distance-decay function of geographical gravity model: Power law or exponential law? *Chaos, Solitons & Fractals*, *77*, 174–189.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, Y., Xu, H., Chen, X. M., & Gao, Z. (2023). A multi-scale unified model of human mobility in urban agglomerations. *Patterns*, *4*.

Danwakai, R. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*, 253–269.

Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A. L., & Wang, D. (2016). Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, *113*, 7047–7052. http://dx.doi.org/10.1073/pnas.1525443113.

Ectors, W., Kochan, B., Janssens, D., Bellemans, T., & Wets, G. (2019). Exploratory analysis of zipf's universal power law in activity schedules. *Transportation*, *46*, 1689–1712.

Ewedairo, K., Chhetri, P., & Jie, F. (2018). Estimating transportation network impedance to last-mile delivery: A case study of maribyrnong city in Melbourne. *The International Journal of Logistics Management*, *29*, 110–130.

Fotheringham, A. S. (1983). A new set of spatial-interaction models: The theory of competing destinations. *Environment and Planning A: Economy and Space*, *15*, 15–36.

Fotheringham, A. S., & O'Kelly, M. E. (1989). *Spatial interaction models: formulations and applications*: *Vol. 1*, Kluwer academic publishers Dordrecht.

Gabaix, X. (1999). Zipf's law and the growth of cities. *American Economic Review*, *89*, 129–132.

Ghanbari, M., Karimi, M., Claramunt, C., & Lagesse, C. (2025). Enhancing urban transportation flow modeling through a graph neural network-based spatially weighted interaction model: A case study of Chicago taxi data. *Journal of Geovisualization and Spatial Analysis*, *9*, 15.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*, 779–782.

Gu, H., & Shen, T. (2021). Modelling skilled and less-skilled internal migrations in China, 2010–2015: Application of an eigenvector spatial filtering hurdle gravity approach. *Population, Space and Place*, *27*, Article e2439.

Gu, X., Tang, X., Chen, T., & Liu, X. (2024). Predicting the network shift of large urban agglomerations in China using the deep-learning gravity model: A perspective of population migration. *Cities*, *145*, Article 104680.

Gupta, A., Shukla, N., Marla, L., Kolbeinsson, A., & Yellepeddi, K. (2019). How to incorporate monotonicity in deep networks while preserving flexibility?. arXiv preprint arXiv:1909.10662.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, *30*.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*, 260–271. http://dx.doi.org/10.1080/15230406.2014.890072.

Haynes, K. E., Fotheringham, A. S., et al. (1984). *Gravity and spatial interaction models*: *Vol. 2*, Sage publications Beverly Hills.

He, J., Li, C., Yu, Y., Liu, Y., & Huang, J. (2017). Measuring urban spatial interaction in wuhan urban agglomeration, central China: A spatially explicit approach. *Sustainable Cities and Society*, *32*, 569–583.

Isard, W. (1956). Regional science, the concept of region, and regional structure. *Papers in Regional Science*, *2*, 13–26.

Jiang, B., & Jia, T. (2011). Exploring human mobility patterns based on location information of us flights. arXiv preprint arXiv:1104.4578.

Kang, C., Zhang, Y., Ma, X., & Liu, Y. (2013). Inferring properties and revealing geographical impacts of intercity mobile communication network of China using a subnet data set. *International Journal of Geographical Information Science*, *27*, 431–448.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*, 422–440.

Karpatne, A., Jia, X., & Kumar, V. (2024). Knowledge-guided machine learning: Current trends and future prospects. arXiv preprint arXiv:2403.15989.

Lambiotte, R., Blondel, V. D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., & Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A. Statistical Mechanics and its Applications*, *387*, 5317–5325.

Li, Z., Huang, X., Hu, T., Ning, H., Ye, X., Huang, B., & Li, X. (2021). Odt flow: Extracting, analyzing, and sharing multi-source multi-scale human mobility. *PLoS One*, *16*, Article e0255259.

Liu, H., Chen, Y. H., & Lih, J. S. (2015). Crossover from exponential to power-law scaling for human mobility pattern in urban, suburban and rural areas. *The European Physical Journal B*, *88*, 117.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, *14*, 463–483. http://dx.doi.org/10.1007/s10109-012-0166-z.

Liu, Z., Miranda, F., Xiong, W., Yang, J., Wang, Q., & Silva, C. (2020). Learning geo-contextual embeddings for commuting flow prediction. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 808–816).

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS One*, *9*, Article e86026.

Luo, Y., Wan, Z., Chen, Y., Mai, G., Chung, F. l., & Larson, K. (2024). Transflower: An explainable transformer-based model with flow-to-flow attention for commuting flow prediction. arXiv preprint arXiv:2402.15398.

Lv, J., Wang, Y., Liang, X., Yao, Y., Ma, T., & Guan, Q. (2021). Simulating urban expansion by incorporating an integrated gravitational field model into a demand-driven random forest-cellular automata model. *Cities*, *109*, Article 103044.

M Schlapfer, L. D., O'Keefe, K., Santi, P., Szell, M., Salat, H., Anklesaria, S., Vazifeh, M., Ratti, C., West, G., et al. (2021). The universal visitation law of urban mobility. *Nature (Lond)*, *593*.

Ma, S., Zhang, D., Zhao, Y., Zhang, X., Wu, L., & Cai, Y. (2025). Exploring future landscape changes for polycentric urbanization using cellular automata calibrated with radiation model. *Environment and Planning B: Urban Analytics and City Science*, *52*, 115–130.

Major, M. D., Tannous, H. O., Al-Thani, S., Hasan, M., Khan, A., & Salaheldin, H. (2020). Macro-and micro-scale modeling of multi-modal transportation spatial networks in the city-state of Doha, Qatar. In *56th ISOCARP virtual world planning congress*.

Mohammed, M., & Oke, J. (2023). Origin-destination inference in public transportation systems: A comprehensive review. *International Journal of Transportation Science and Technology*, *12*, 315–328.

Monzon, A., Alonso, A., & Lopez-Lambas, M. (2017). Joint analysis of intermodal long distance-last mile trips using urban interchanges in EU cities. *Transportation Research Procedia*, *27*, 1074–1079.

Muenz, R., et al. (2007). *Aging and demographic change in European societies: main trends and alternative policy options*: *Vol. 703*, World Bank, Social Protection Advisory Service.

Nayka, S., & Sridhar, K. S. (2019). Determinants of intra urban mobility: A study of Bengaluru.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PLoS One*, *7*, Article e37027.

Oshan, T. M. (2021). The spatial structure debate in spatial interaction modeling: 50 years on. *Progress in Human Geography*, *45*, 925–950. http://dx.doi.org/10.1177/0309132520968134.

Ou, J., Lu, J., Xia, J., An, C., & Lu, Z. (2019). Learn, assign, and search: Real-time estimation of dynamic origin-destination flows using machine learning algorithms. *IEEE Access*, *7*, 26967–26983.

Randrianarisoa, L. M., & Gillen, D. (2022). Reducing emissions in international transport: A supply chain perspective. *Transportation Research Part D: Transport and Environment*, *102*, Article 103074.

Rong, C., Ding, J., & Li, Y. (2024). An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *ACM Computing Surveys*, *57*, 1–49.

Rong, C., Wang, H., & Li, Y. (2023). Origin-destination network generation via gravity-guided gan. arXiv preprint arXiv:2306.03390.

Roy, J. R., & Thill, J. C. (2004). Spatial interaction modelling. *Papers in Regional Science*, *83*, 339–361.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis*. Chapman and Hall/CRC.

Sen, A., & Smith, T. E. (2012). *Gravity models of spatial interaction behavior*. Springer Science & Business Media.

Silva, S. J., & Nelson, D. (2012). Does aid cause trade? evidence from an asymmetric gravity model. *The World Economy*, *35*, 545–577.

Silva, J. S., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 641–658.

Simini, F., Barlacchi, G., Luca, M., & Pappalardo, L. (2021). A deep gravity model for mobility flows generation. *Nature Communications*, *12*, 6576.

Simini, F., González, M. C., Maritan, A., & Barabási, A. L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*, 96–100.

Sugar, L., & Kennedy, C. (2021). Urban scaling and the benefits of living in cities. *Sustainable Cities and Society*, *66*, Article 102617.

Šveda, M., & Madajová, M. S. (2023). Estimating distance decay of intra-urban trips using mobile phone data: The case of Bratislava, Slovakia. *Journal of Transport Geography*, *107*, Article 103552.

Terroso-Sáenz, F., & Muñoz, A. (2022). Nation-wide human mobility prediction based on graph neural networks. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *52*, 4144–4160.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*, 234–240.

Wang, H. (2025). Ranking dynamics of urban mobility. *Journal of Transport Geography*, *128*, Article 104331.

Wang, R., Wang, Q., & Li, N. (2023). Percolation transitions in urban mobility networks in America's 50 largest cities. *Sustainable Cities and Society*, *91*, Article 104435.

Wang, Y., Wang, Z., Zhang, F., Kang, C., Ruan, S., Zhu, D., Tang, C., Ma, Z., Zhang, W., Zheng, Y., et al. (2025). A gravity-informed spatiotemporal transformer for human activity intensity prediction. arXiv preprint arXiv:2506.13678.

Wang, Y., Yin, H., Chen, H., Wo, T., Xu, J., & Zheng, K. (2019). Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1227–1235).

Wang, X., Zhao, T., Tu, W., Zhang, B., Chen, G., & Cao, J. (2025). Sat2flow: A structure-aware diffusion framework for human flow generation from satellite imagery. http://dx.doi.org/10.48550/arXiv.2508.19499, URL https://arxiv.org/abs/2508.19499arXiv:2508.19499.

Wilson, A. G. (1970). The use of the concept of entropy in system modelling. *Journal of the Operational Research Society*, *21*, 247–265.

Wilson, A. G. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning A*, *3*, 1–32.

Wilson, A. (2013). *Entropy in urban and regional modelling (Routledge Revivals)*. Routledge.

Wu, X., Pan, T., & He, Z. (2025). Physics-informed mobility perception networks for origin-destination flow prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Wu, C., Smith, D., & Wang, M. (2021). Simulating the urban spatial structure with spatial interaction: A case study of urban polycentricity under different scenarios. *Computers, Environment and Urban Systems*, *89*, Article 101677.

Xu, W., & Gu, X. (2025). Spatiotemporal impacts of purpose-specific human mobility on air pollution: Evidence from taxi trajectories and interpretable machine learning. *Sustainable Cities and Society*, *126*, Article 106411. http://dx.doi.org/10.1016/j.scs.2025.106411.

Yan, X. Y., Han, X. P., Wang, B. H., & Zhou, T. (2013). Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Scientific Reports*, *3*, 2678. http://dx.doi.org/10.1038/srep02678.

Yang, T., & Zhang, Y. (2025). Urban systems science and cross-scale dynamics: A conceptual framework to advance integrated planning and governance. *Transactions in Planning and Urban Research*, *4*, 3–13.

Yao, C. Z., & Lin, J. N. (2016). A study of human mobility behavior dynamics: A perspective of a single vehicle with taxi. *Transportation Research Part A: Policy and Practice*, *87*, 51–58. http://dx.doi.org/10.1016/j.tra.2016.03.005.

Zhang, L., Cheng, J., & Jin, C. (2019). Spatial interaction modeling of od flow data: Comparing geographically weighted negative binomial regression (GWNBR) and ols (GWOLSR). *ISPRS International Journal of Geo-Information*, *8*, 220.

Zhang, X., & Li, N. (2024). An activity space-based gravity model for intracity human mobility flows. *Sustainable Cities and Society*, *101*, Article 105073.

Zhao, Y., Cheng, S., Gao, S., Wang, P., & Lu, F. (2025). Predicting origin-destination flows by considering heterogeneous mobility patterns. *Sustainable Cities and Society*, *118*, Article 106015.

Zheng, Z., Rasouli, S., & Timmermans, H. (2016). Two-regime pattern in human mobility: evidence from GPS taxi trajectory data. *Geographical Analysis*, *48*, 157–175.

Zhou, S., Zhou, S., Zheng, Z., Lu, J., & Song, T. (2022). Risk assessment for precise intervention of covid-19 epidemic based on available big data and spatio-temporal simulation method: Empirical evidence from different public places in Guangzhou, China. *Applied Geography*, *143*, Article 102702.

Zhu, J., & Rui, X. (2025). Spatiotemporal dynamics and multi-scale equity evaluation of urban rail accessibility: Evidence from Hangzhou. *ISPRS International Journal of Geo-Information*, *14*, 361.

Zipf, G. K. (1946). The p 1 p 2/d hypothesis: on the intercity movement of persons. *American Sociological Review*, *11*, 677–686.