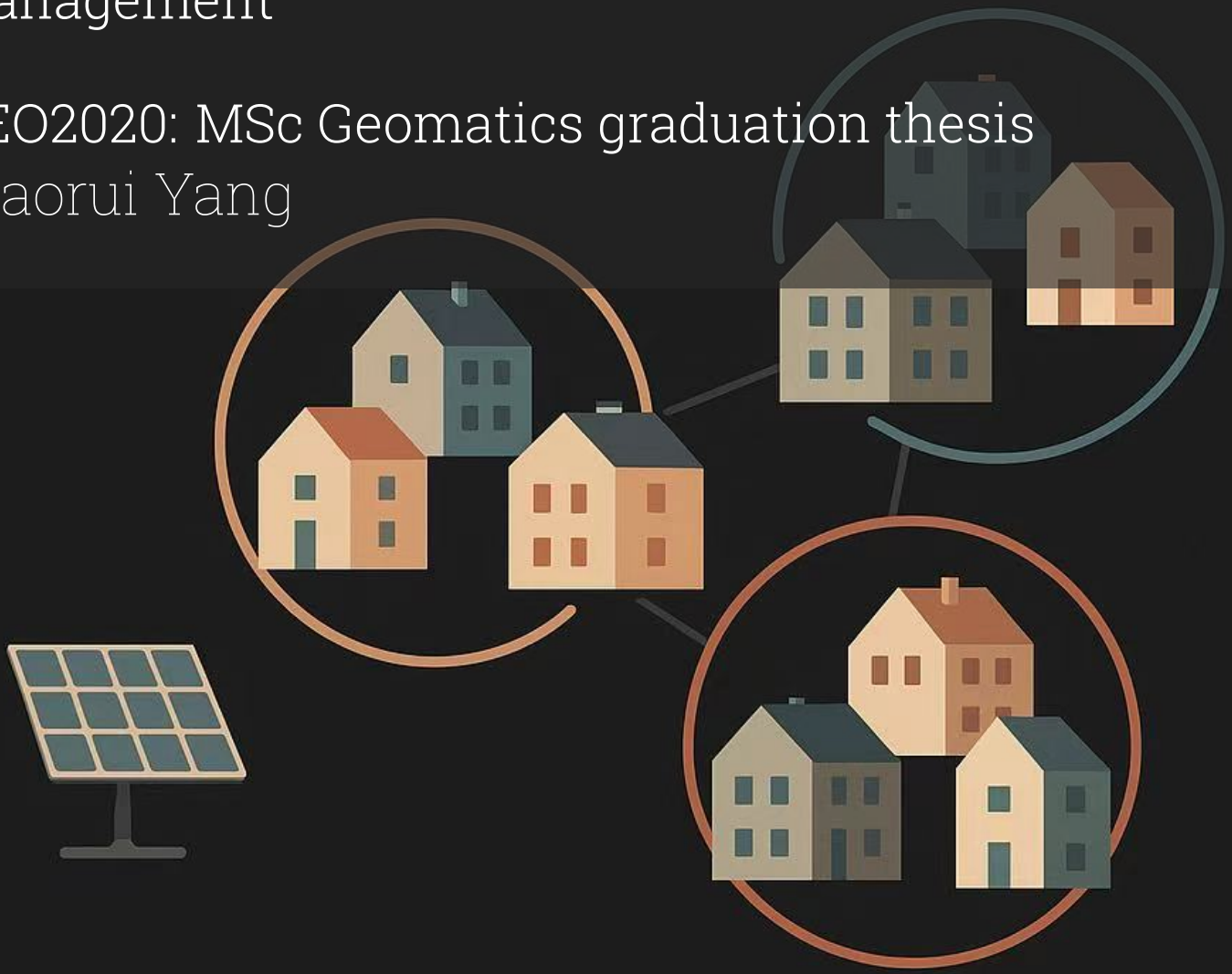
MSc thesis

Integrating Spatial Knowledge Graphs and Graph Neural Networks for Clustering and Scenario-Based Enhancements in Regional Energy Systems Management

GEO2020: MSc Geomatics graduation thesis

Qiaorui Yang



Delft University of Technology

TU Delft

MSc thesis

Integrating Spatial Knowledge Graphs and Graph Neural Networks for Clustering and Scenario-Based Enhancements in Regional Energy Systems Management

by

Qiaorui Yang

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday October 27, 2024 at 14:45 AM.

Student number:	5969972	
Project duration:	November 10, 2024 – October 27, 2025	
Thesis committee:	Prof. dr. Azarakhsh Rafiee,	TU Delft, 1st supervisor
	Amin Jalilzadeh,	TU Delft, 2nd supervisor
	Drs. Wilko Quak,	TU Delft, 3rd supervisor

Cover: Canadarm 2 Robotic Arm Grapples SpaceX Dragon by NASA under CC BY-NC 2.0 (Modified)
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Abstract

The transition towards sustainable energy systems presents significant analytical challenges. Urban distribution networks are characterised by heterogeneous data sources, complex physical hierarchies, and pronounced temporal variability in both demand and generation. Effective planning therefore requires tools that can integrate diverse datasets, respect electrical constraints, and generate interpretable results that are directly useful for decision-making.

This thesis develops a methodological framework that integrates **knowledge graphs (KGs)** and **graph neural networks (GNNs)** to identify and characterise energy communities in low-voltage urban networks. The KG provides a semantic backbone for representing buildings, grid hierarchies, and temporal energy states in a physically faithful manner. The GNN builds on this representation through constraint-aware learning, discovering communities that are both infrastructure-consistent and temporally complementary.

The research was guided by four questions. **RQ1** investigated which nodes, attributes, and edges are essential for energy network representation in a KG. The study identified buildings, LV cable groups, transformers, and adjacency clusters as the core entities, enriched with both spatial and non-spatial attributes. **RQ2** examined how heterogeneous urban energy datasets can be integrated into a KG to reflect system complexity. A unified data pipeline was developed, enabling automated construction and updating of the KG from diverse sources while preserving grid topology and physical constraints. **RQ3** explored how KGs and GNNs can be combined to enhance clustering and analysis of energy systems.

A complementarity-aware GNN architecture with custom loss functions was implemented, embedding grid boundaries and load-balancing rules to ensure physically valid and interpretable clustering outcomes. **RQ4** focused on identifying suitable GNN approaches for time-based and dynamic clustering and evaluating their performance. Temporal modules and constraint-aware pooling mechanisms were applied, with performance assessed using metrics such as cluster cohesion, complementarity, and self-sufficiency.

In conclusion, the integrated framework achieved composite community quality Qc = 0.623, temporal stability St = 0.85, and LV compliance RLV = 0.92, meeting predefined performance targets. While the KG construction pipeline demonstrated robust automation and scalability potential, the GNN component delivered only modest performance gains over simpler clustering baselines at the tested scale, while requiring substantially higher computational effort. The primary contribution therefore lies in establishing a methodological foundation—automated graph construction and constraint-aware learning architecture—that provides scalability for future large-scale, heterogeneous urban energy systems where GNN's representational capacity becomes essential.

Furthermore, the framework offers methodological innovation through automated KG construction and constraint-aware GNN architecture, establishing a validated foundation for large-scale energy system analysis. At the tested scale (63 buildings), simpler clustering methods may offer better cost-effectiveness; the GNN component's value emerges in complex scenarios (>500 buildings, dynamic DER, multi-objective optimization) where architectural sophistication becomes necessity rather than luxury.

Acknowledgements

I would like to express my deepest gratitude to all those who have supported and guided me throughout the process of writing this thesis.

First and foremost, I am sincerely grateful to my first supervisor, Dr. Azarakhsh Rafiee, for her patience and thoughtfulness guidance. She consistently took the time to understand my progress and provided invaluable feedback that shaped this research at every stage. Her mentorship has been essential to the completion of this work.

My heartfelt thanks also go to my second supervisor, Amin Jalilzadeh, who stood by my side throughout the entire research journey. He patiently answered all my questions, encouraged me whenever I reached a standstill, and helped me move the project forward. I must say that without his support, it would have been impossible for me to complete this thesis as it stands today. I am deeply thankful for his generosity in sharing his time and for his understanding during moments when I doubted my own research.

I would also like to thank my third supervisor, Drs. Wilko Quak, for his insightful comments and constructive suggestions during each key meeting. His feedback inspired critical reflection and contributed to many meaningful improvements in this work.

My sincere appreciation extends to the members of my graduation committee, Dr. Bastiaan van Loenen and Dr. Camilo León-Sánchez, for their time, attention, and valuable contributions to the evaluation and progress of this thesis.

I owe my deepest gratitude to my parents and my younger brother. Thank you for your unconditional support in every aspect of my life—from academic choices to everyday decisions. You have always respected my thoughts and feelings and reminded me that you are proud of me. I know that this love and support are truly priceless, and I want to say that I am equally proud and grateful to have you as my family. Without your encouragement, I would never have had the opportunity to pursue and complete this important stage of my academic journey.

Finally, I would like to thank my friends, whose companionship made studying abroad a meaningful and fulfilling experience. It is rare to meet people who share similar interests, values, and goals, and with whom one can have deep and genuine conversations. I will never forget those moments of companionship, understanding, and support—in the library, on the train, at the cinema, in my dorm kitchen, or during our travels. These moments have been a source of strength and motivation, reminding me why it is worth continuing to strive in the academic world.

Nomenclature

Abbreviations

Abbreviation	Definition
AC	Alternating Current
ARI	Adjusted Rand Index
BCE	Binary Cross-Entropy
DER	Distributed Energy Resources
DSO	Distribution System Operator
ECE	Expected Calibration Error
EV	Electric Vehicle
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
HCR	Multi-hop Contribution Ratio
KG	Knowledge Graph
LV	Low Voltage
MV	Medium Voltage
PV	Photovoltaic
UQ	Uncertainty Quantification

Symbols

Symbol	Definition	Unit
$G = (V, E)$	Graph with node set V (buildings) and edge set E (connections)	[–]
A	Adjacency matrix of G	[–]
$x_i(t)$	Time series signal of building i (e.g., net load)	[kW]
$\tilde{L}_{i,t}$	Standardized net-load time series of building i	[–]
ρ_{ij}	Correlation coefficient between buildings i and j	[–]
C_{ij}	Complementarity score $(1-\rho_{ij})/2$ between buildings i and j	[–]
$\mathcal{C}_k$	Set of buildings in cluster k	[–]
$c(i)$	Cluster assignment of building i	[–]
K	Number of clusters (global or per LV feeder)	[–]
$\mathrm{LV}(i)$	LV group identifier of building i	[–]
$\mathcal{N}(i)$	Neighborhood of node i (1-hop unless specified)	[–]
d_{ij}	Electrical or spatial distance between nodes i and j	[m] or [–]
r_{ij}	Edge impedance/resistance proxy	[–]
h_i	Node embedding in GNN layers	[–]
α_{ij}	Attention weight from node i to j	[–]
$\mathbf{M}$	LV boundary mask applied to assignment logits	[–]
S	Soft cluster assignment matrix in pooling layers	[–]
$\mathbf{z}_i^{temp}$	Temporal embedding of building i	[–]
$\bar{r}^{(h)}$	Average h-hop correlation of node embeddings	[–]
HCR	Multi-hop contribution ratio	[–]

Symbol	Definition	Unit
p_i	Soft assignment/probability vector for building i	[–]
$\hat{y}_i$	Pseudo-label of building i	[–]
Γ_i	Fused confidence score for pseudo-label acceptance	[–]
τ_t	Confidence threshold at training stage t	[–]
$\tau_{\min}$	Minimum confidence threshold	[–]
ρ	Threshold decay factor in curriculum schedule	[–]
η	Regularization coefficient in semi-supervised objective	[–]
λ_Q	Weight for cluster-quality loss component	[–]
λ_S	Weight for cluster-size regularization	[–]
λ_{LV}	Weight for LV-boundary violation penalty	[–]
λ_R	Weight for embedding regularization term	[–]
$\mathbb{1}\{\cdot\}$	Indicator function	[–]
Q_c	Community quality score (complementarity + spatial coherence)	[–]
S_t	Temporal stability metric of cluster assignments	[–]
R_{LV}	LV-boundary compliance rate	[–]
$\mathrm{mean}_t^{(W,f)}$	Rolling average for feature f over window W	[–]
$\mathrm{std}_t^{(W,f)}$	Rolling standard deviation for feature f over window W	[–]
$\pi(v)$	Hierarchical positional encoding of node v (building, feeder, transformer)	[–]
$\Psi(v,t)$	Node–time feature map after preprocessing	[–]

Contents

1 Introduction

1.1. Background

The global energy sector is undergoing a profound transformation driven by the dramatic escalation of energy demand and the depletion of fossil fuels becoming an undeniable trend, as surveyed by Aniakor et al. (2024) [3] and Pritoni et al. (2021) [48]. This shift toward renewable energy integration fundamentally alters the architecture of energy systems, moving from traditional centralized power generation to increasingly decentralized networks of distributed energy resources, which introduces unprecedented operational complexities in system coordination and management.

This structural transformation creates multifaceted challenges for modern energy infrastructure. Decentralized energy systems must now coordinate diverse generation sources with varying output patterns, manage bidirectional power flows, and accommodate real-time demand response mechanisms, a point emphasized by Aniakor et al. (2024) [3] and Pritoni et al. (2021) [48]. These operational complexities are further compounded by the need for more sophisticated monitoring, control, and optimization capabilities across distributed networks, which require more flexible, data-driven decision-making capabilities than traditional centralized systems can provide.

Among these emerging challenges, data management has become a critical bottleneck in the evolution of intelligent energy systems. The transformation has resulted in the generation of vast amounts of diverse and often unstructured data from multiple sources, posing significant challenges to traditional energy management systems that are predicated on centralized control and structured data formats, as documented by Aniakor et al. (2024) [3] and Pritoni et al. (2021) [48]. Popadić et al. (2023) [46] emphasize the difficulties in managing the massive and heterogeneous data volumes in modern energy systems, clearly demonstrating that traditional data processing methods are insufficient to address the complexity of contemporary energy networks. Furthermore, the datasets storing energy data are often incompatible with intelligent analysis applications due to differences in format and intelligent analysis algorithms such as neural networks, which hinders the further intelligence of urban energy data and the development of various intelligent analyses including neural network analysis and cluster analysis; this concern is echoed by Guo (2024) [19] and Liu et al. (2023) [34]. This necessitates the establishment of interconnected data structures capable of accommodating and uniformly managing diverse types of energy data to support advanced analytics capabilities essential for next-generation energy system optimization.

Local power grids face multifaceted challenges in energy management, with the most prominent issue being the temporal and spatial misalignment in supply-demand matching. Due to the absence of effective intelligent scheduling mechanisms, local energy production and consumption often fail to achieve dynamic equilibrium, resulting in suboptimal energy allocation efficiency. Temporal supply-demand imbalances pose severe challenges, where buildings equipped with solar photovoltaic (PV) systems generate substantial electricity during daylight hours when solar irradiance is abundant, yet their own electricity consumption is relatively low. Conversely, commercial buildings or residential areas require significant power supply during nighttime peak demand periods when solar generation capacity is lim-

ited. This temporal mismatch leads to energy resource waste and supply-demand gaps; Murphy et al. (2023) [39] further substantiate this phenomenon through their analysis of temporal complementarity in variable renewable energy (VRE) resources, finding that individual types of renewable energy generation exhibit distinct temporal variability characteristics.

Spatial distribution imbalances exacerbate allocation challenges, where different regions and building types exhibit significant variations in energy production and consumption characteristics, yet existing grid management systems lack sophisticated regional coordination mechanisms, preventing effective cross-regional energy allocation and complementarity. From a geographical distribution perspective, Murphy et al. (2023) [39] demonstrate that renewable energy resources in different regions exhibit differentiated complementarity characteristics: in the western United States, wind and solar PV resources show excellent co-located complementarity; in the wind belt and surrounding areas, co-located wind and PV resources demonstrate high complementarity; while in the northeastern and southeastern regions, the complementarity intensity among various renewable energy resources varies significantly across regions. This spatial heterogeneity underscores both the importance and complexity of establishing cross-regional coordination mechanisms. Energy complementarity emerges as a fundamental solution to these temporal and spatial misalignments, offering a pathway to optimize resource utilization through intelligent clustering and coordination of complementary energy profiles.

Traditional correlation-based approaches for assessing energy complementarity have shown significant limitations in capturing the true nature of temporal relationships between energy profiles, as argued by Cantor et al. (2022) [10]. Beyond static correlation measures, advanced complementarity assessment methods have demonstrated substantially improved accuracy in identifying optimal building clusters that can achieve genuine load balancing and peak shaving effects. In particular, dynamic energy profile analysis has proven essential for identifying truly complementary energy consumers—for instance, pairing industrial facilities with residential areas can achieve significant peak reduction with limited population participation, as illustrated by Xiao et al. (2023) [64].

Despite the theoretical promise of energy complementarity and advanced assessment approaches, several critical implementation barriers prevent their widespread adoption in real-world energy systems. The widespread adoption of renewable energy systems (RES) introduces system-level grid stability challenges. Barone et al. (2023) [6] indicate that large-scale RES integration leads to significantly increased grid fluctuations, potentially triggering serious power quality issues such as voltage and frequency imbalances. The inherent unpredictability characteristics of RES, combined with the decentralization trend in energy production, present unprecedented challenges to traditional centralized grid management paradigms. This volatility stems not only from the intermittent nature of renewable energy resources but is also closely related to their distributed deployment patterns, where dispersed generation units significantly reduce overall system predictability and controllability.

From a data integration perspective, the datasets storing energy data are often incompatible with intelligent analysis applications due to differences in format and intelligent analysis algorithms such as neural networks. This hinders the further intelligence of urban energy data and the development of various intelligent analyses including neural network analysis and cluster analysis, a gap also highlighted by Guo (2024) [19] and Liu et al. (2023) [34]. The challenge of dynamic clustering represents a particularly complex barrier, as energy networks require continuous real-time restructuring under constantly changing loads, generation, or policy constraints. Traditional clustering approaches lack the adaptability required for such dynamic environments, where building energy clustering configurations must be updated frequently to maintain optimal complementarity.

Current approaches fail to address these challenges due to fundamental limitations in their design and implementation paradigms. The direct consequences of these failures are multifaceted: At the technical level, grid systems not only bear additional power transmission losses, particularly during long-distance transmission where energy loss rates increase significantly, but also face difficulties in voltage regulation and frequency control caused by RES volatility. At the economic level, inefficient energy allocation increases both temporal and economic costs of system operation, while grid fluctuations require additional balancing services and reserve capacity, further escalating system operational costs. At the equipment level, grid infrastructure faces congestion pressures, with critical equipment such as transformers frequently experiencing overload conditions that not only affect power supply stability but also accelerate equipment aging and increase maintenance costs. Additionally, Murphy et al. (2023)

[39] emphasize that due to the lack of effective resource integration mechanisms, local grids cannot fully utilize the synergistic effects of complementary resources to improve capacity factors, reduce curtailment, and achieve cost synergies.

As Wang et al. (2016) [61], Vergados et al. (2016) [60], and Barone et al. (2023) [6] indicate in their respective research, the root causes of these challenges lie in the limitations of traditional grid operation paradigms. Confronting the unpredictability of RES and the decentralization of energy production, developing local energy communities represents a promising solution for minimizing power losses and promoting sustainable energy practices [6]. This community-based energy management model not only effectively addresses grid fluctuation issues caused by RES but also reduces long-distance transmission losses through localized energy balancing mechanisms, thereby improving overall system efficiency.

1.1.1. The Research Gap

The primary research gap lies in the urgent need to address continuous real-time restructuring of energy networks—dynamic clustering of buildings or resources under constantly changing loads, generation, or policy constraints, as discussed by Hussain et al. (2019) [27]. While knowledge graphs provide semantic clarity and graph neural networks offer advanced pattern recognition capabilities, maximizing the ontological consistency between knowledge graphs and GNNs for real-time clustering analysis of energy systems remains inadequately explored. The complementary clustering problem holds significant research value in urban energy networks as it can improve energy production and utilization efficiency while optimizing power scheduling within the grid.

Key technical challenges that current implementations fail to address include: First, the standardization of complementarity loss functions specifically tailored for energy applications, as existing general-purpose GNN frameworks lack domain-specific optimization criteria (cf. Wu et al. 2021 [63]). Second, the effective combination of Dynamic Time Warping (DTW) with GNN architectures for enhanced temporal modeling in energy demand patterns, which is essential given the time-series nature of energy consumption data (see also Zhang et al. 2021 [68]). Third, the development of standardized benchmarks comparing complementarity metrics in energy contexts, as current evaluation frameworks do not adequately capture the unique characteristics of energy system performance (related discussions in Tsitsulin et al. 2023 [55] and Pelekis et al. 2023 [44]).

Additionally, physical grid constraints must be integrated into the clustering methodology. When buildings are powered by the same transformer, they should undergo complementary clustering while adhering to real-world physical grid constraints. The clustering implementation must satisfy two critical requirements: First, clustered buildings should be powered by the same transformer, ensuring compliance with actual grid topology. Second, clustering should possess real-time characteristics, meaning each time point in the time series should have different building energy clustering configurations—in this research.

Furthermore, current research lacks physics-informed GNN architectures that incorporate sufficient domain knowledge specific to energy systems, federated learning approaches for privacy-preserving clustering across multiple utilities, and real-time adaptation mechanisms for dynamic grid conditions. These limitations prevent the full realization of GNN potential in energy management applications (compare Pagnier & Chertkov, 2021 [43]; Authier et al., 2024 [4]).

To address these fundamental limitations, this research proposes a novel integration of knowledge graphs (KG) and graph neural networks (GNN) for energy demand complementarity clustering. Knowledge graphs provide a structured approach to representing and integrating complex and interconnected data from diverse sources, offering a unified, semantically rich view of energy systems, as introduced by Sajid (2023) [51]. They facilitate the integration of heterogeneous data and support advanced reasoning and query capabilities, which are crucial for managing the complexity of modern energy systems. Knowledge graphs excel in managing dynamic heterogeneous data, capable of storing rich semantics suitable for storing multi-level, multi-type energy network data, and can explicitly represent complex relationships in urban energy systems, such as grid topological structures, supply-demand relationships, and geographical adjacency relationships, according to Liu et al. (2023) [34] and Chen et al. (2022) [11].

Several studies have successfully utilized knowledge graphs and graph neural networks to enhance energy systems. For example, Kimball (2024) [29] demonstrates how knowledge graphs can connect different energy data sources, thereby improving interoperability. Furthermore, Fusco et al. (2020) [16] discuss GNN frameworks for grid congestion prediction and market bidding services. These integrations facilitate more intelligent and efficient energy management. GNNs are particularly suitable for graph-based data and can perform complex tasks such as knowledge graph completion (KGC), predictive modeling, and real-time optimization. Liu (2024) [35] and Xu et al. (2023) [65] demonstrate the effectiveness of GNNs in knowledge graph completion and real-time optimization, making them powerful tools for addressing the dynamic characteristics of modern energy systems.

Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) exhibit a fundamental *ontological alignment* in their representation of energy systems, as both paradigms are inherently predicated on the graph-theoretic formalism. This ontological commitment manifests through their shared structural foundation: energy networks are intrinsically modeled as graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where components (e.g., substations, generators, transformers) constitute vertices $\mathcal{V}$ and interconnections (e.g., transmission lines, control relationships) form edges $\mathcal{E}$.

As established by Liu et al. (2023) [34], KGs explicitly encode power systems as semantic graphs where entities (nodes) and their relations (edges) incorporate domain-specific knowledge, operational constraints, and physical laws. Concurrently, Chen et al. (2022) [11] demonstrate that GNNs leverage this identical topological structure to learn latent representations through message-passing mechanisms across nodes and edges. This structural isomorphism enables direct mathematical compatibility:

- **KGs** formalize domain knowledge as structured triples $(h, r, t) \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ (head, relation, tail), capturing explicit semantics of power system entities and relationships (Liu et al., 2023 [34]).
- **GNNs** exploit the adjacency matrix $\mathbf{A}$ and node features $\mathbf{X}$ derived from $\mathcal{G}$ to perform feature propagation via $\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$, learning implicit patterns from connectivity and attributes (Chen et al., 2022 [11]).

This ontological convergence creates a theoretically grounded foundation for integration: the KG's semantic schema provides contextual constraints and reasoning rules, while GNNs offer inductive learning capabilities over the shared graph structure. Despite this intrinsic compatibility, current research has not systematically unified KG-enhanced reasoning with GNN-based learning for complex energy network tasks—a gap noted in both reviews by Liu et al. (2023) [34] and Chen et al. (2022) [11]. Future work should exploit this ontological symmetry to develop hybrid architectures where symbolic KG reasoning guides subsymbolic GNN learning within a unified graph representation of energy infrastructures.

According to Hofer et al. (2023) [22], machine learning can benefit from KG as labeled training data, thereby improving the quality and interpretability of AI decision-making, which means KG can support data-driven AI energy analysis. Due to ontological consistency, GNNs can directly learn features on knowledge graph structures. Knowledge graphs can also provide GNNs with rich semantic information details of urban energy networks, such as grid topological structures, supply-demand relationships, and geographical adjacency relationships. Therefore, their combination can bring tremendous intelligent empowerment to urban energy network analysis. GNNs excel in handling various graph structure problems, including clustering problems (see also Bose, 2017 [8]).

This research aims to bridge these gaps by exploring the combined application of knowledge graphs and graph neural networks to enhance the efficiency, reliability, and sustainability of energy management systems. The study seeks to leverage the strengths of both technologies, combined with advanced complementarity assessment metrics and physics-informed constraints, to develop a comprehensive data-driven approach for managing the complexity of modern energy systems. By addressing these research gaps, this work provides new optimization and adaptability opportunities to tackle the increasingly complex and interconnected nature of the global energy sector, positioning GNN-based energy demand complementarity clustering as a transformative technology for future energy systems.

This research makes several key contributions to the field of intelligent energy management through the integration of Knowledge Graphs and Graph Neural Networks: First, we establish a unified knowledge graph architecture for energy network representation, identifying essential nodes, attributes, and

edges required for effective KG construction from heterogeneous urban building spatial and non-spatial energy datasets, with an automated pipeline ensuring comprehensive integration that reflects energy system complexities. Second, we design and implement a novel integration methodology between Knowledge Graphs and Graph Neural Networks for enhanced energy system analysis, developing complementarity-aware GNN models with custom loss functions optimized to minimize peak-to-average ratio, maximize self-consumption, and balance load curves through clustering learning based on complementarity rather than similarity. Third, we determine optimal GNN approaches for time-based and dynamic clustering within the Knowledge Graph framework by embedding physical constraints into learning through adaptive clustering mechanisms, specifically addressing dynamic energy production and demand pattern analysis in regional energy systems. Fourth, we establish comprehensive validation and benchmarking frameworks with quantitative metrics for complementarity assessment and performance measurement of the integrated KG-GNN approach in energy system clustering applications. Finally, we demonstrate significant improvements in data accessibility, interoperability, and relational analytics for regional energy system management, providing a scalable solution that enhances decision-making capabilities through the synergistic combination of knowledge representation and neural network architectures.

In summary, the challenges of data heterogeneity, temporal imbalance, and physical constraint integration collectively motivate the need for a unified analytical framework that can bridge semantic representation and predictive learning. These motivations directly lead to the research objectives and questions outlined in the following chapter.

This paper is organized as follows: Section 2 proposed research questions and research objectives. Section 3 reviews related work in knowledge graph in energy systems, streaming data integration architectures for knowledge graphs, machine learning on graphs, energy demand complementary, GNN-based methods for dynamic clustering, and evaluation framework. Section 4 presents the methodology, including the knowledge graph construction, GNN architecture design, and complementarity-aware clustering algorithm. Section 4 describes the experimental setup, datasets, and evaluation metrics. Section 5 presents and analyzes the experimental results, comparing our approach with current baselines. Section 6 discusses the implications of our findings, limitations, and future research directions. Finally, Section 7 concludes the paper and summarizes the main contributions.

2

Research Objective

In this section, the research objective is claimed, the main research question and sub-questions are defined, furthermore, the research scope (must, must not and could) is introduced.

This research addresses the fundamental challenge of improving data accessibility, interoperability, and relational analytics in regional energy system management through the integration of Knowledge Graphs (KG) and Graph Neural Networks (GNN). The research objectives encompass four primary goals that directly correspond to the identified research questions.

First, to establish the foundational elements of energy network representation by identifying essential nodes, attributes, and edges required for effective Knowledge Graph construction from urban building spatial and non-spatial energy datasets. This involves developing a unified knowledge graph architecture with an automated pipeline for constructing and updating knowledge graphs from heterogeneous data sources, ensuring comprehensive integration that reflects energy system complexities.

Second, to design and implement the integration methodology between Knowledge Graphs and Graph Neural Networks for enhanced energy system analysis and clustering. This includes developing complementarity-aware GNN models with custom loss functions optimized to minimize peak-to-average ratio, maximize self-consumption, and balance load curves, while creating a GNN architecture that performs clustering learning based on complementarity rather than similarity.

Third, to determine optimal GNN approaches for time-based and dynamic clustering within the Knowledge Graph framework by embedding physical constraints into learning through adaptive clustering mechanisms, specifically focusing on dynamic energy production and demand pattern analysis.

Fourth, to establish comprehensive validation and benchmarking frameworks by developing quantitative metrics for complementarity assessment and performance measurement of the integrated KG-GNN approach in energy system clustering applications.

2.1. Research Question

Main Question: How can integrating a Knowledge Graph (KG) with a Graph Neural Network (GNN) improve data accessibility, interoperability, and relational analytics (e.g., clustering, link prediction) for the management of regional energy systems?

Sub-Questions:

1. What are the essential nodes, attributes, and edges to define in an energy network for effective representation in a Knowledge Graph (KG)?
2. What is the process for constructing a Knowledge Graph (KG) based on urban building spatial and non-spatial energy-related dataset, and how can the data be integrated to reflect the complexities of the energy system?

3. How can Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) be combined to enhance analysis and clustering in energy system models, including recommendations for optimal solar panel deployment?
4. Which GNN approach best supports time-based or dynamic clustering in the KG, and how to measure its performance?

2.2. Research Scope

This thesis is situated at the intersection of knowledge graph construction and graph neural network modelling for urban energy systems. The scope of the research is defined by three dimensions: (i) the system boundary, (ii) the methodological objectives, and (iii) the exclusions and potential extensions.

System boundary. The study focuses on *abstracted representations* of regional energy systems derived from heterogeneous urban datasets. Entities such as buildings, low-voltage groups, solar panels, and batteries are represented as nodes in a knowledge graph, while electrical or spatial relationships (e.g., feeder membership, adjacency, transformer domain) are encoded as edges. The emphasis is on the semantic and topological structure of the network rather than on hardware-level modelling of physical circuits or devices.

Methodological objectives. The core deliverable is the integration of **Knowledge Graphs (KGs)** with **Graph Neural Networks (GNNs)** to support:

- **Energy community discovery**: identifying clusters of buildings that exhibit temporal complementarity in their demand and generation profiles while remaining consistent with LV feeder boundaries.
- **Decision-support indicators**: producing interpretable outputs (e.g., SSR, SCR, complementarity indices, and uncertainty bounds) that inform energy planning and community-scale interventions.

Exclusions. The research does not address hardware implementation, physical circuit modelling, or operational control of real distribution grids. Energy conservation techniques and behavioural energy-saving mechanisms are also out of scope. Furthermore, reinforcement learning (RL) is not explored, as the methodological focus is restricted to KG construction and GNN-based clustering and prediction.

Potential extensions. Subject to time and resources, the framework could be extended to additional GNN tasks such as *link prediction* (e.g., inferring missing relationships in the energy graph) and *anomaly detection* (e.g., detecting inconsistent or unusual consumption profiles). These extensions would further demonstrate the versatility of the KG–GNN architecture for energy system analysis and management.

Having defined the research goals and scope, the next chapter reviews the state of the art in knowledge graphs, graph neural networks, and energy clustering to identify existing solutions and methodological gaps that inform the design of our framework.

3

Related Work

Knowledge graphs (KG) and graph neural networks (GNN), along with their integration, are receiving widespread attention in the field of complex system analysis, as highlighted by Popadic et al. (2023) [46], Li et al. (2023) [32], and Liu et al. (2024) [35]. These technologies provide innovative solutions for managing the complex interconnected data characteristic of modern energy systems. However, the effective integration of these technologies for energy system management requires a comprehensive understanding of their individual capabilities, limitations, and synergistic potential. For example, KG-based modules such as the LV-Group Boundary Enforcer explicitly guarantee that energy sharing does not cross transformer boundaries, thereby embedding physical feasibility into the representation. In contrast, GNN-based modules such as the Multi Hop Aggregator capture higher-order interactions by propagating information across multiple hops, enabling the detection of demand–generation balancing opportunities beyond immediate neighbors. When combined, as in the Temporal Evolution Predictor, the KG–GNN integration allows long-term forecasting of community reorganization under staged solar deployments, ensuring that predictions remain both physically consistent and dynamically adaptive to new interventions.

This section systematically reviews the existing literature across six critical dimensions that directly address the research questions posed in this study. First, we examine knowledge graph construction and management approaches in energy systems to establish the foundation for addressing *what essential nodes, attributes, and edges are required for effective energy network representation*. Second, we analyze graph neural network applications in energy domains to understand *which GNN approaches best support time-based clustering and how their performance can be measured*. Third, we investigate energy system clustering methodologies, with particular emphasis on complementarity-based approaches versus traditional similarity-based methods, to inform the development of our complementarity-aware clustering framework. Fourth, we explore the emerging field of KG-GNN integration to address the central question of *how these technologies can be combined to enhance analysis and clustering capabilities*. Fifth, we review approaches for incorporating physical constraints and real-world operational requirements, as these directly impact the practical applicability of clustering solutions in actual grid infrastructures. Finally, we examine performance evaluation frameworks and benchmarking methodologies to establish appropriate metrics for assessing the effectiveness of integrated KG-GNN approaches.

3.1. Knowledge Graph Construction and Management in Energy Systems

3.1.1. Knowledge Graph Foundations and Paradigm Evolution

Knowledge graphs represent a fundamental paradigm shift in knowledge representation and management, with their conceptual roots tracing back to ancient philosophical reasoning principles. As Chen et al. (2020) [12] note, reasoning techniques have evolved from Aristotle's syllogism in ancient Greece through Lambda Calculus to modern intelligent computing platforms, with knowledge graphs repre-

senting a contemporary manifestation of this evolutionary trajectory. The core principle underlying knowledge graphs—using known knowledge to infer new knowledge through logical rules—remains consistent with these historical foundations.

The modern conception of knowledge graphs, while building upon earlier expert systems developed in the late 1960s as discussed by Chen et al. (2020) [13], gained prominence following Google's 2012 Knowledge Graph announcement according to Hogan et al. (2021) [23]. This industrial adoption catalyzed widespread development across major technology companies, including Amazon, eBay, Facebook, IBM, LinkedIn, Microsoft, and Uber, demonstrating the practical value of graph-based knowledge representation in large-scale applications.

Definitionally, Sajid et al. (2023) [51] define knowledge graphs as intuitive representations of real-world data in graph form, where nodes represent entities and edges represent relationships. This structure extends the characterization by Hogan et al. (2021) [23], who describe knowledge graphs as intelligent systems that integrate knowledge and data at scale, employing graph-based data models to capture knowledge in scenarios involving the integration, management, and value extraction from diverse data sources. This integration addresses the limitations of traditional expert system approaches, which relied heavily on manually crafted rules and expert knowledge, as noted by Chen et al. (2020) [13]. The transition from expert-driven to data-driven methodologies became necessary due to the explosive growth of Internet data, making traditional manually constructed knowledge bases inadequate for big data environments, as summarized by Chen et al. (2020) [12].

3.1.2. Advantages of Knowledge Graphs in Energy System Management

Knowledge graphs represent a paradigm shift in managing the complex, interconnected relationships characteristic of modern energy systems. Their advantages over traditional data management approaches become particularly pronounced when addressing the challenges of decentralized energy sources, heterogeneous data integration, and real-time operational requirements.

The fundamental distinction between knowledge graphs and relational databases lies in their approach to schema management and data representation. While relational databases require rigid, predefined schemas optimized for structured data, as discussed by Sajid et al. (2023) [51], knowledge graphs provide dynamic schema evolution that accommodates the constantly changing relationships between power generation, storage, and consumption components. This flexibility directly addresses the data heterogeneity challenge inherent in regional energy system management, where diverse data sources must be integrated without extensive preprocessing.

Compared to static modeling approaches such as UML class diagrams, knowledge graphs offer superior adaptability for dynamic operational environments. Huang et al. (2016) [25] demonstrate that while UML focuses on predefined classes and relationships suitable for design-time modeling, knowledge graphs support real-time relationship evolution and dynamic decision-making processes. This capability enables the integration of real-time sensor data with historical operational patterns, supporting enhanced predictive modeling of energy flows and system behavior.

The relationship between knowledge graphs and graph databases merits particular attention, as graph databases provide the foundational infrastructure while knowledge graphs add semantic layers, as noted by Kiff et al. (2024) [28]. This combination enables efficient relationship traversal with semantic reasoning capabilities, supporting complex analytical queries that span multiple relationship types and temporal dimensions. The semantic enrichment distinguishes knowledge graphs from pure graph databases by enabling ontological reasoning and inference capabilities essential for intelligent energy system management, as emphasized by Hogan et al. (2021) [23].

The systematic comparison presented in Table 3.1 demonstrates the multidimensional advantages of

Table 3.1: Comparison of Data Representations for Energy System Applications

Characteristics	Knowledge Graphs	Relational Databases	Graph Databases
Schema Flexibility	Dynamic, ontology-driven schema evolution [23]	Rigid, predefined tables and keys [51]	Flexible graph schema; edges can be added without redesign [28]
Relationship Representation	Semantic relations with contextual meaning	Encoded via foreign keys and joins	Explicit edges with properties, but limited semantics
Semantic Reasoning	Ontology support, inference rules and reasoning engines [51]	None	Limited to graph traversal; no higher-order inference
Heterogeneous Data Integration	Native support for structured, semi-structured and unstructured data fusion [56]	Requires ETL processes for schema alignment	Can ingest diverse data, but lacks semantic harmonization
Real-time Adaptability	Incremental updates with temporal annotations [11]	Schema evolution costly and disruptive	Designed for real-time graph updates and queries
Domain Knowledge Integration	Captures expert rules and operational data jointly [19]	Data-centric only	Relationship-centric, lacks domain semantics
Query Style	SPARQL, reasoning-enabled path discovery	SQL with relational joins	Cypher/Gremlin for efficient path traversal

knowledge graphs across nine critical characteristics for energy system applications. The convergence of these individual capabilities creates synergistic effects that extend beyond the sum of their parts, establishing knowledge graphs as the optimal foundation for intelligent energy management systems.

Three key factors distinguish knowledge graphs as particularly suited for energy system management. First, the combination of schema flexibility with semantic reasoning enables adaptive modeling of evolving grid topologies while maintaining contextual understanding of component relationships. This dual capability supports both immediate operational decisions and long-term strategic planning within unified analytical frameworks. Second, the demonstrated performance advantages—including 16–17 times faster aggregate query processing as reported by Liu et al. (2023) [34]—directly address the millisecond-level response requirements of modern grid operations while supporting complex multi-criteria optimization scenarios. Third, the native integration of expert domain knowledge with operational data, as highlighted by Guo et al. (2024) [19], enables hybrid intelligence approaches that leverage both accumulated electrical engineering expertise and real-time system insights.

These converging advantages position knowledge graphs as essential infrastructure for addressing contemporary energy system challenges, particularly in scenarios requiring dynamic clustering analysis under physical grid constraints. The ability to seamlessly integrate heterogeneous data sources, as described by Van Otten et al. (2023) [56], with semantic understanding of domain relationships provides the foundational capabilities necessary for developing sophisticated energy management solutions that can adapt to the increasing complexity of modern power systems.

3.1.3. Knowledge Graphs in Energy System Applications

The increasing complexity of modern power systems, driven by rapid global economic development and rising energy consumption, has created unprecedented challenges in system management and control. As Guo et al. (2024) [19] note, power systems encompass multiple interconnected components including generation, transmission, distribution, and consumption, while simultaneously addressing environmental protection, energy security, and economic efficiency requirements. This multifaceted nature results in highly complex system characteristics that demand more sophisticated management and control technologies.

The urgency for intelligent technologies in power systems has become particularly acute when addressing challenges such as demand fluctuations, equipment failures, and climate change impacts,

as highlighted by Guo et al. (2024) [19]. Power systems encompass substantial prior knowledge and generate extensive operational and maintenance data, yet historically, this knowledge and data have been challenging to utilize effectively. A critical issue in advancing power system intelligence involves leveraging prior knowledge to identify implicit operational patterns within data more efficiently, thereby supporting reliable system operation.

Knowledge graphs have been widely recognized as a transformative tool for managing and analyzing complex datasets in the energy sector. Kimball et al. (2024) [29] highlight their role in addressing the data demands of the global energy transition, emphasizing their ability to provide semantic context and link diverse datasets. By harmonizing structured and unstructured data into a unified framework, KGs facilitate intelligent energy management, enabling more reliable, interoperable, and efficient systems.

The application potential of knowledge graphs in power systems addresses fundamental challenges including information silos, data redundancy, and knowledge acquisition difficulties, as described by Guo et al. (2024) [19]. Through comprehensive modeling of power system components, knowledge graphs enable improved system visualization and understanding, providing more intuitive decision-making foundations for operation and maintenance personnel. Furthermore, by constructing knowledge graphs for equipment monitoring, fault diagnosis, and load forecasting scenarios, these systems can enhance the accuracy of power system fault identification and accelerate emergency response capabilities. The reasoning capabilities inherent in knowledge graphs also support intelligent trading in power markets, enabling optimal resource allocation and improved economic efficiency.

Knowledge graphs excel in managing the interdependencies within energy networks. Energy systems are inherently complex, involving many types of data that need to be linked, including energy demand, supply, environmental factors, and grid performance. Van Otten et al. (2023) [56] point out that KGs integrate structured and unstructured data through standardized schemas and ontologies, allowing for better semantic clarity and eliminating inconsistencies between disparate data sources. This characteristic addresses the critical challenge identified by Guo et al. (2024) [19] regarding the effective utilization of heterogeneous data in power systems, where data often come from various formats and integration without semantic understanding can lead to errors and inefficiencies. KGs help bridge this gap, making the data more actionable for decision-making.

Popadic et al. (2023) [46] further demonstrate the usefulness of KGs in creating intelligent grids. Enhanced versions of KGs incorporate metadata and relational schemas, transforming raw data into actionable insights. These advancements support autonomous decision-making and improve grid efficiency, leveraging accumulated knowledge in electrical science to address the complexities of distributed energy sources and smart devices. This aligns with the vision presented by Guo et al. (2024) [19] for achieving more efficient, stable, and sustainable power system operation through intelligent technologies.

The integration of knowledge graphs in power systems represents a paradigm shift from traditional data management approaches. By providing comprehensive modeling capabilities that address the wealth of prior knowledge and operational data generated in power systems, knowledge graphs enable the identification of implicit operational patterns and support reliable system operation, as emphasized by Guo et al. (2024) [19]. This capability is particularly crucial in the big data era, where the volume of power system data is experiencing explosive growth.

The semantic enrichment capabilities of knowledge graphs, as demonstrated in the comparison framework, provide high-level ontological support with inference and reasoning capabilities, distinguishing them from traditional relational databases and property graphs that offer limited semantic enrichment, as explained by Sajid et al. (2023) [51]. This semantic richness is particularly valuable for energy system applications where understanding the contextual meaning of relationships between components is crucial for effective system optimization and clustering analysis.

The application of knowledge graphs in power system operations demonstrates significant engineering value through functions such as knowledge search, knowledge Q&A, intelligent recommendation, and auxiliary decision-making, as summarized by Liu et al. (2023) [34]. These capabilities enable substantial reductions in human and material costs while improving work efficiency across typical scenarios including power equipment operation and maintenance, customer service, and grid dispatch fault management. The evolution toward more comprehensive power system applications aligns with the research objective of developing integrated KG–GNN approaches for enhanced energy system analysis and clustering, particularly in addressing the challenges of dynamic energy production and demand pattern analysis.

3.1.4. Domain Knowledge Graph Construction in Power Systems

The transition from general knowledge graphs to domain-specific applications has particular significance in the power sector. Liu et al. (2023) [34] distinguish between general knowledge graphs and domain knowledge graphs (DKG), noting that power domain knowledge graphs undertake the critical mission of shifting from "data-driven power automation" to a "knowledge-driven smart grid," which holds important theoretical value and engineering significance for the power industry.

The construction of power domain knowledge graphs follows a comprehensive three-phase process: data collection, graph construction, and knowledge calculation, as summarized by Liu et al. (2023) [34]. This systematic approach addresses the unique characteristics of power systems, including strong professionalism, complicated data structures, and high accuracy requirements.

Data Acquisition and Quality Management

As the foundational step in knowledge-graph construction, data collection directly determines the quality of the resulting KG. Liu et al. (2023) [34] note that power-domain data originate from diverse sources and types, including structured data from knowledge engineering and expert experience knowledge bases that can directly participate in top-down ontology construction. However, the primary data sources consist of massive operational data and expert experience characterized by high noise levels and sparse data density. This necessitates comprehensive noise-filtering and data-sample expansion processes to improve data quality before knowledge processing, transforming semi-structured or unstructured data containing potential knowledge into structured knowledge information suitable for KG construction.

Hybrid Construction Methodology

Given the characteristics of strong professionalism, data complexity, and accuracy requirements in the power field, graph construction typically adopts a combination of top-down and bottom-up construction processes, as discussed by Liu et al. (2023) [34]. This hybrid approach first defines the ontology layer, then extracts knowledge from the data layer to update the ontology layer, enabling dynamic characteristics that support real-time updates essential for power-system operations.

The quality of extracted knowledge directly affects the final KG quality. Liu et al. (2023) [34] note that joint models for entity extraction and relationship extraction in knowledge extraction provide more accurate results compared with pipeline extraction methods, avoiding performance degradation due to error accumulation and propagation. Furthermore, considering that extracted entities, concepts, relationships, and attributes originate from different grids and power devices, knowledge-fusion techniques are employed to eliminate redundancy from different sources and achieve interoperability of individual isolated power-system knowledge.

Knowledge Reasoning and Storage Optimization

Knowledge-reasoning capabilities enable the discovery of potential knowledge based on existing knowledge sets, as highlighted by Liu et al. (2023) [34]. Applications include deductive reasoning methods for rule-based scenarios such as grid-dispatching regulations and customer-service business processes,

and inductive reasoning methods including case-based reasoning and representation-learning-based reasoning for more complex operational scenarios. Recent developments incorporate neural-network models that adopt “massive data + self-learning” approaches to cope with increasingly complex grid structures and unpredictable customer-behavior patterns.

For storage optimization, Liu et al. (2023) [34] identify graph databases—particularly Neo4j—as preferred over traditional RDF-table formats for power-domain applications. Graph databases demonstrate significant performance advantages, with queries two to three times faster than RDF formats and sixteen to seventeen times faster for aggregate queries with multiple starter nodes, while providing better support for complex power-system scenarios through built-in attribute-information storage.

Perccuku et al. (2017) [45] address the limitations of traditional relational databases in handling big-data from power-transmission-grid substations, where data volumes, velocity, and variety exceed relational-database performance thresholds. Their Neo4j implementation models electrical substations as nodes with power transformers as connected entities, demonstrating that graph databases eliminate expensive JOIN operations through native relationship traversal. The study validates Neo4j’s schema-less architecture for rapid infrastructure evolution—adding new power transformers requires simple node creation rather than complex schema modifications—maintaining constant query performance regardless of data-volume growth.

3.1.5. Advanced Framework for Power System Dispatching Operations

The maturation of artificial intelligence technologies in power systems has created opportunities for more sophisticated knowledge-graph applications. Chen et al. (2022) [11] note that while AI technologies including convolutional neural networks, long short-term memory networks, and deep belief networks have achieved satisfactory results in load forecasting, fault diagnosis, and optimization control, they have encountered bottlenecks in data processing and management. Knowledge graphs represent the core of the new-generation data system, addressing these limitations through comprehensive integration of heterogeneous data sources.

Multi-Level Knowledge Graph Architecture

The framework for knowledge-graph applications in power-system dispatching operations encompasses basic data, data processing, knowledge extraction, graph construction, and graph application components, as described by Chen et al. (2022) [11]. This architecture requires specific technical characteristics essential for large-scale power systems.

Spatiotemporal Dynamic Characteristics: Knowledge graphs must incorporate timestamp systems for basic attributes such as entity identification and characteristic attributes including power and power flow, accommodating the time-varying nature of power systems. Real-time dynamic information updates ensure synchronous circulation between the knowledge graph and power-system operations, directly supporting the research objective of dynamic clustering within temporal frameworks.

Multivariate Data Fusion: Power-system dispatching involves massive data volumes, diverse data types, and high-speed data-processing requirements. The construction process integrates data from multiple sources including image recognition, semantic analysis, and equipment monitoring, providing rich foundations for potential relationship mining that supports the research goal of establishing comprehensive validation frameworks.

The multi-level knowledge-graph architecture comprises three layers: a physical-layer knowledge graph as the core that constructs power-system graphs according to real power-grid topology; a data layer corresponding to actual operating conditions that works with the physical layer to build real-time updated dynamic power-dispatch knowledge graphs; and an advanced application layer providing external technical support including load forecasting and dispatching decisions, as outlined by Chen et al. (2022) [11].

3.1.6. Data Integration Challenges in Power System Knowledge Graphs

The sophisticated knowledge-graph architectures discussed above depend fundamentally on effective data integration from heterogeneous sources. While Liu et al. (2023) [34] demonstrate Neo4j's performance advantages and validate schema-less evolution capabilities, these implementations assume relatively static data-integration patterns. However, modern power systems generate massive streams of real-time operational data that must be continuously integrated with existing knowledge structures.

The challenge extends beyond simple data ingestion. Chen et al. (2022) [11] note that power-system dispatching requires spatiotemporal dynamic characteristics with real-time updates, while Guo et al. (2024) [19] emphasize the need to leverage prior knowledge for identifying implicit operational patterns in streaming data. This creates a fundamental tension between the relationship-rich modeling capabilities of knowledge graphs and the high-velocity, high-volume characteristics of temporal power-system data.

Traditional ETL (Extraction–Transformation–Loading) approaches, while constituting specialized software systems for managing heterogeneous data sources, present significant limitations for knowledge-graph construction. Vassiliadis et al. (2002) [58] note that ETL processes account for 55–80 % of total data-warehouse development time and represent at least one-third of project budgets, primarily due to the complexity of managing data heterogeneity and implementing comprehensive cleaning routines. More critically, batch ETL processes introduce latency bottlenecks incompatible with the real-time requirements of dynamic knowledge-graph updates.

The paradigm shift toward streaming architectures has been driven by the need for continuous data processing and immediate insight generation. Meehan et al. (2017) [38] pioneered streaming ETL architectures by integrating message-queuing systems (Apache Kafka), stream processors, and polyglot database systems, addressing the fundamental limitation of overnight batch processing through streaming dependency resolution. This architectural evolution demonstrates measurable improvements in data freshness, reducing end-to-end latency from hours to minutes, which directly supports the real-time update requirements identified by Chen et al. (2022) [11] for power-system knowledge graphs.

However, existing streaming ETL frameworks remain constrained by single-model storage assumptions, lacking the architectural flexibility to optimize for both relationship complexity and temporal analytics simultaneously. This limitation becomes particularly pronounced when considering the requirements for KG–GNN integration, where graph neural networks require efficient access to both topological relationships and temporal feature vectors.

3.2. Streaming Data Integration Architectures for Knowledge Graphs

The scale and velocity constraints of modern power systems, where smart grid infrastructure requires continuous monitoring of electrical parameters, create fundamental tensions between relationship modeling and temporal analytics. Modern industrial applications, particularly in smart grid infrastructure, generate massive volumes of high-frequency sensor data requiring sub-second processing capabilities. Power grid infrastructure demands real-time monitoring of electrical parameters, as emphasized by Meehan et al. (2017) [38].

3.2.1. Streaming Data Integration Method

The integration of message-queuing systems (Apache Kafka), stream processors (S-Store), and polyglot database systems addresses these challenges through streaming dependency resolution for referential integrity constraints, as demonstrated by Meehan et al. (2017) [38]. This approach enables incremental data transformation with minute-level micro-batching instead of traditional daily cycles, directly

supporting the spatiotemporal dynamic characteristics required for power-system knowledge graphs.

Hanžel et al. (2025) [21] highlight the economic impracticality of full-resolution household electricity consumption storage, advocating for specialized time-series engines with built-in compression and aggregation capabilities. Their findings emphasize the necessity of purpose-built temporal databases for sustained high-throughput write operations, while simultaneously demonstrating the need for semantic integration capabilities through RDF representations and SPARQL-based querying.

The semantic integration approach demonstrated by Hanžel et al. (2025) [21] through unified multi-regional electricity consumption knowledge graphs links household characteristics, appliance types, and carbon emission data with external ontology alignment (Wikidata, DBpedia). This validates the capacity for cross-domain knowledge integration while highlighting the limitations of current single-database solutions in managing both temporal and relationship complexity.

Current approaches demonstrate clear limitations in addressing these dual requirements. Graph databases like Neo4j excel at relationship traversal and provide superior performance for deep relationship queries compared to multi-model alternatives, but lack optimized temporal analytics for time-windowed aggregations and trend analysis. Conversely, purpose-built temporal databases exemplify these performance characteristics through architectures specifically designed for IoT applications and infrastructure monitoring.

InfluxDB represents a paradigmatic example of specialized temporal storage optimization. As an open-source distributed time-series database created by InfluxData in 2013 and developed in the Go programming language, InfluxDB addresses the fundamental challenge of high-velocity data ingestion through its dependency on LevelDB for key–value storage operations, as described by Abu et al. (2019) [1]. The database's architecture demonstrates the necessity of purpose-built solutions through its four-layer TICK stack: Telegraf for server-driven metric collection, InfluxDB as the core time-series engine capable of handling heavy write and query loads, Chronograf for web-based infrastructure monitoring and alert management, and Kapacitor for real-time data processing.

The primary architectural advantage of InfluxDB lies in its ability to perform on-the-fly aggregation of values into time buckets without manual intervention, addressing the computational challenges identified in smart grid applications. This capability proves essential for IoT applications where continuous sensor data requires real-time processing and storage optimization, as demonstrated by Abu et al. (2019) [1]. The database's data structure, comprising measurements, series, and points, enables efficient organization where each point contains key–value field pairs with timestamps, supporting 64-bit integers, 64-bit floating points, Booleans, and strings. Points are indexed by tagset and timestamp, facilitating rapid temporal queries through HTTP API and client libraries that integrate seamlessly with visualization tools like Grafana.

The streaming ETL architecture proposed by Meehan et al. (2017) [38] provides crucial insights into how purpose-built temporal systems can be integrated with broader data-ingestion pipelines. Their findings demonstrate that traditional ETL batch processing creates fundamental latency bottlenecks for time-sensitive applications, particularly in IoT deployments where sensor data value decreases drastically over time. The streaming approach enables real-time data transformation and loading, addressing the core challenge of maintaining temporal data freshness while ensuring referential integrity constraints. This is particularly relevant for power-grid applications where delayed processing of electrical parameter data can compromise real-time decision-making capabilities.

The integration of Apache Kafka as a messaging infrastructure for data collection, combined with specialized time-series storage like InfluxDB, addresses the scalability requirements identified by Meehan et al. (2017) [38] for handling thousands of simultaneous data sources. Their experimental results with

TPC-DI demonstrate that frequent, small-batch migrations (1–5 second intervals) provide optimal performance across data freshness, query runtime, and ingestion latency metrics. This finding directly supports the architectural approach of combining high-throughput message queuing with purpose-built temporal databases for smart-grid monitoring applications.

However, the specialized temporal focus of InfluxDB, while providing exceptional performance for metric-based queries, cannot efficiently model the complex entity relationships required for contextual understanding of electrical grid topology. The database's SQL-like query interface operates on time-oriented data structures optimized for aggregation and trend analysis but lacks the semantic modeling capabilities necessary for representing the intricate interdependencies between grid components, household characteristics, and energy consumption patterns that characterize modern smart-grid infrastructures. Furthermore, Meehan et al. (2017) [38] note that relational ETL systems prioritize ACID transactions and update-in-place functionality, while time-series workloads are predominantly append-heavy, suggesting that different architectural approaches are required for different data characteristics within the same application domain.

This analysis reveals that neither Neo4j nor InfluxDB alone can adequately address the dual requirements of temporal knowledge-graph applications in power systems. The optimal solution necessitates a hybrid architecture that leverages the complementary strengths of both systems: Neo4j's superior relationship modeling and traversal capabilities for representing grid topology and semantic connections, combined with InfluxDB's specialized temporal analytics for high-velocity sensor data processing and time-windowed aggregations. The streaming ETL patterns demonstrated by Meehan et al. (2017) [38] provide the architectural foundation for such integration, enabling real-time data transformation while maintaining both temporal performance and semantic consistency. Such a hybrid approach would enable contextual temporal queries that can traverse complex grid relationships while simultaneously performing efficient temporal analytics on associated time-series data, thereby addressing the fundamental architectural limitations identified in current single-database solutions for smart-grid knowledge-graph applications.

3.2.2. Integration Framework Requirements for KG-GNN Applications

The convergence of knowledge graphs with graph neural networks creates additional architectural requirements that extend beyond traditional KG storage and query optimization. GNN processing requires efficient access to both topological relationships for message-passing algorithms and temporal feature vectors for time-series analysis, creating demands that neither pure graph databases nor time-series databases can fully address independently.

No mature framework currently exists for deep coupling of graph models (relationships) with temporal storage (metrics), necessitating a unified query layer that enables cross-model analytical operations. This integration challenge becomes particularly acute when considering the requirements for complementarity-based clustering, where GNN algorithms must process both spatial relationships between energy-system components and temporal patterns in consumption and production data.

The research gap in integrated approaches that leverage complementary strengths of different data models directly impacts the feasibility of sophisticated KG–GNN applications. Graph-database limitations in temporal analytics prevent efficient implementation of time-aware GNN architectures, while time-series database constraints in relationship modeling limit the contextual understanding necessary for energy-system topology analysis.

3.3. From Knowledge Graph to ML-Ready Graphs

3.3.1. Complementary Integration of KGs and GNNs

The integration of Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) represents a transformative approach in energy-systems management, combining the semantic richness of KGs with the predictive capabilities of GNNs. The combination of KGs and GNNs enables advanced management of large-scale, dynamic energy datasets by uniting semantic data representation with capabilities like predictive modeling and real-time optimization, offering a robust framework for modern energy systems. While KGs provide a rich semantic structure to represent relationships, they often lack the ability to generalize across unseen data or adapt to changes in real-time. GNNs, with their capacity to learn low-dimensional representations of entities and their relationships, address this gap by enabling adaptive reasoning and predictive modeling. Liu et al. (2024) [35] explain that GNNs excel in capturing the relational and structural nuances of graph-structured data, making them particularly suited for tasks such as knowledge-graph completion and semantic analysis. Their ability to distill meaningful features from large-scale graphs allows for sophisticated reasoning across both static and dynamic datasets, which is critical in energy systems characterized by fluctuating supply, demand, and operational constraints.

Recent work by Fusco et al. (2020) [16] demonstrates the practical benefits of such integration. IBM Research's hybrid framework successfully combines semantic knowledge from CIM ontologies with real-time numerical data to support services like congestion prediction and market bidding. This fusion results in over 80% parameter reduction compared to traditional multilayer perceptrons (MLPs) while maintaining accuracy, showcasing the efficiency gains possible through KG–GNN synergy.

Moreover, dynamic integration approaches such as streaming GNNs enable real-time updates of KGs as new sensor data arrive. For example, Yang et al. (2025) [66] proposed DEST-GNN, a spatio-temporal GNN that integrates sparse attention and adaptive graph construction to achieve fast, accurate multi-site photovoltaic power forecasting, with mean absolute errors (MAEs) between 0.42 and 0.49, and 4–80× faster inference speeds compared with conventional methods. These advances further highlight how GNNs can enhance temporal responsiveness and computational efficiency in energy forecasting tasks.

The integration of KGs and GNNs leverages the strengths of both technologies: the semantic richness and contextual clarity of KGs, and the predictive and analytical capabilities of GNNs. KGs excel at encoding domain-specific relationships and complex interdependencies between entities, while GNNs provide robust methods for extracting and learning patterns from these interconnections. Together, they enable advanced applications such as real-time optimization, intelligent grid management, and knowledge-graph completion, driving significant advancements in energy management and paving the way for the future of smart, resilient, and sustainable energy systems.

Dynamic Clustering and Real-time Applications

Among the advanced GNN applications in energy systems, dynamic clustering (or real-time grouping) has emerged as a crucial technique. By continually updating node embeddings (e.g., building load profiles, DER states) within a KG, GNN-based methods can discover temporary or evolving communities that share constraints or optimization goals. Here, **distributed energy resources (DERs)** refer to decentralized, small-scale units of electricity generation or storage (such as rooftop solar panels, small wind turbines, batteries, and controllable loads) that are connected to the distribution grid and can both consume and produce electricity, as described by Hussain et al. (2019) [27]. This is particularly relevant for microgrid formation, peak-load management, and local balancing, where clusters must adapt swiftly to changes (e.g., shifts in solar generation or occupant behavior). The synergy between KGs (for semantic structure) and GNNs (for on-the-fly embeddings) thus facilitates dynamic re-grouping of nodes, enabling more intelligent resource allocation and operational decisions.

The complementary nature of KGs and GNNs thus addresses a critical need in modern energy systems: the ability to combine high-level semantic reasoning with robust, data-driven optimization. KGs

encode semantic relationships between energy entities (buildings, DERs, lines), while GNNs exploit these relationships for tasks like clustering, anomaly detection, or link prediction—enabling real-time adaptation to changes in the KG, as noted by Liu et al. (2024) [35]. By bridging these two approaches, researchers and practitioners can unlock new possibilities for intelligent energy-system management, enabling adaptive, efficient, and sustainable operation of regional energy grids.

3.3.2. Regional Energy Grid Optimization Applications

Moreover, the integration of GNNs with KGs has profound implications for optimizing regional energy grids. By leveraging GNNs' capability to encode topological features and KGs' semantic insights, the combined framework can perform tasks such as real-time clustering and optimization of energy flow within a grid. For instance, Liu et al. (2024) [35] highlight how GNNs, when applied to the topological structures of KGs, can uncover latent relationships and dependencies within the energy grid, enabling precise energy-flow analysis and predictive maintenance. This is particularly important for decentralized energy systems, where real-time decision-making is crucial for balancing distributed generation, storage, and consumption.

Similarly, Huo et al. (2024) [26] demonstrate the application of GNNs in microgrid energy-distribution management, where they analyze power-distribution relationships and identify optimal operational schemes. This capability ensures real-time optimization, reducing energy losses and improving computational efficiency. For regional energy grids, these techniques are invaluable in addressing challenges such as clustering consumers based on energy-usage patterns, dynamically reconfiguring grids to accommodate renewable-energy variability, and ensuring system stability in the face of demand surges or failures.

3.3.3. Comparative Analysis of KGs and GNNs

Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) are not competing but complementary paradigms. KGs excel at providing semantic clarity, ontology-based reasoning, and explicit encoding of infrastructure hierarchies, while GNNs contribute predictive power, temporal learning, and the ability to uncover hidden patterns beyond rule-based analysis. In energy-network studies, this distinction is critical: KGs ensure physical and regulatory consistency, whereas GNNs extend the framework toward forecasting, optimization, and decision support.

Aspect	Knowledge Graphs (KGs)	Graph Neural Networks (GNNs)
Core Functionality	Represent domain-specific relationships, semantic information, and hierarchical structures in explicit graph forms.	Learn patterns from graph data via embeddings and message passing, enabling prediction, optimization, and adaptive analysis.
Strengths	Provide semantic clarity, support domain-specific reasoning, and encode physical/electrical boundaries.	Handle complex patterns, generalize to unseen nodes, adapt to dynamic datasets, and support spatio-temporal prediction.
Limitations	Struggle with dynamic or missing data; reasoning can become computationally expensive at scale.	Depend on labeled data for training; risk losing semantic interpretability during embedding.
What KGs Provide to GNNs	Semantic context, domain constraints, and physical feasibility rules (e.g., transformer boundaries, regulatory constraints).	–
What GNNs Add to KGs	–	Dynamic adaptability, prediction for unknown buildings, network-effect modeling, and hidden pattern discovery.
Applications in Energy Networks	Encode grid topology, infrastructure hierarchy, and DER attributes; support rule-based analysis of retrofit, solar potential, and electrification feasibility.	Forecast demand, predict intervention impacts, optimize battery siting, simulate dynamic energy sharing, and identify at-risk feeders.

Table 3.2: Complementary roles of KGs and GNNs in energy-network analysis (extended from Huo et al. (2024) [26], Liu et al. (2024) [35], and Xu et al. (2023) [65]).

This comparison highlights that while the KG already provides static analyses such as retrofit candidate identification, solar siting, and electrification feasibility, GNNs extend the methodology by enabling forecasting, optimization, and the discovery of emergent complementarity patterns across the grid, as demonstrated by Liu et al. (2024) [35] and Xu et al. (2023) [65].

3.3.4. Technical Challenges and Research Directions

Despite the promising applications, Chen et al. (2022) [11] identify significant technical challenges that align with current research objectives.

Knowledge Acquisition: The massive heterogeneous data in power systems, including empirical and operational data, presents substantial challenges in data acquisition and processing that directly relate to the research goal of automated pipeline construction from heterogeneous data sources.

Knowledge Representation: The extremely complex multi-layer, intersecting three-dimensional nature of power systems requires sophisticated data-fusion approaches to appropriately integrate information including source, network, load, and storage elements into unified knowledge graphs, supporting the research objective of establishing foundational elements for energy-network representation.

Knowledge Application: Transforming data into actionable knowledge requires both accurate information extraction from existing graphs and optimization space reduction for system-specific scenarios, directly supporting the research goals of comprehensive validation frameworks and enhanced relational analytics in energy-system management.

Contemporary knowledge-graph research encompasses three primary dimensions: knowledge representation, knowledge-graph construction, and knowledge-graph applications, integrating technologies from cognitive computing, knowledge representation and reasoning, information retrieval, natural language processing, and data mining, as summarized by Chen et al. (2020) [13]. This multidisciplinary

approach positions knowledge graphs as particularly suitable for managing the heterogeneous, interconnected data characteristic of modern energy systems, directly supporting the research objectives of establishing unified knowledge-graph architectures and enabling enhanced relational analytics in regional energy-system management.

This transformative potential underscores the importance of continued exploration and refinement of KG and GNN integration in energy applications, enabling adaptive, efficient, and sustainable operation of regional energy grids.

3.4. Energy Demand Complementarity: Concepts, Metrics, and Clustering Objectives

The development of effective energy-system management strategies requires sophisticated understanding of demand-complementarity patterns and their quantitative assessment. This section reviews key concepts, metrics, and clustering approaches that form the theoretical foundation for complementarity-aware analysis in energy systems, integrating recent systematic insights from electrical load profiling research presented by Kusuma et al. (2024) [31].

3.4.1. Complementarity Assessment Metrics

Recent advances in complementarity measurement address limitations of traditional correlation-based approaches. The Total Variation Complementarity Index proposed by Cantor et al. (2022) [10] provides a mathematically rigorous framework:

$$\phi = 1 - \frac{\mathrm{TV}(f+g)}{\mathrm{TV}(f) + \mathrm{TV}(g)} \tag{3.1}$$

Simultaneously, the *stability coefficient* has emerged as a critical metric for temporal complementarity assessment, quantifying the reduction in variability when combining resources. For hybrid renewable systems, it measures the smoothing effect on combined output, as described by Murphy et al. (2023) [39]:

$$C_{\text{stab}} = 1 - \frac{C_{v,\text{hybrid}}}{C_{v,\text{baseline}}} \tag{3.2}$$

where C_v denotes the coefficient of variation. This metric ranges from 0 (no complementarity) to 1 (perfect complementarity) and demonstrates superior performance in handling multi-resource systems, as validated by Murphy et al. (2023) [39]. Empirical validation shows stability coefficients greater than 0.5 indicate significant complementarity, reducing variability by over 50% compared with standalone resources [39].

Beyond static metrics, dynamic complementarity indicators must account for *seasonal variations*. Analysis reveals that complementarity strength fluctuates annually: wind–PV complementarity peaks in winter months (stability coefficient 0.6–0.7) but diminishes during summer (0.3–0.5) due to reduced wind capacity factors, as shown by Murphy et al. (2023) [39]. Hydro–PV complementarity exhibits inverse seasonality, peaking during spring snowmelt and summer monsoons when hydropower generation best compensates for solar intermittency [39].

3.4.2. Load Profiling and Clustering Fundamentals

Load profiling, defined as the analysis of electricity consumption behaviors over specific periods, serves as a cornerstone for understanding energy-utilization patterns and developing effective clustering strategies. Systematic analysis of 52 major studies (2017–2023) reveals that 88% of clustering approaches for Electrical Load Profiles (ELPs) utilize a predetermined number of clusters, with K-Means (38%) and Fuzzy C-Means (FCM, 19%) being the most prevalent, as summarized by Kusuma et al. (2024) [31].

The challenges in residential customer load profiling are particularly pronounced due to the high variety and variability of consumption patterns. Unlike industrial or commercial customers whose load profiles exhibit greater regularity, residential patterns require fine-grained temporal decomposition into dynamically changing fragments. This necessitates exploration of dynamic characteristics including state switching and maintenance probabilities in consumption behaviors, as noted by Wang et al. (2016) [61], with Principal Component Analysis (PCA) emerging as the dominant feature extraction technique (41%) for dimensionality reduction in ELP clustering [31].

The emergence of big-data challenges in smart-meter deployments, with datasets reaching multi-petabyte scales, requires advanced clustering techniques capable of handling high-frequency, high-dimensional data efficiently. Traditional clustering methods including K-Means, Fuzzy K-Means, hierarchical clustering, and self-organizing maps require adaptation for these data volumes, as demonstrated by Wang et al. (2016) [61], particularly given that the Silhouette Index (29%) and Davies–Bouldin Index (19%) remain the primary evaluation metrics for clustering quality, according to Kusuma et al. (2024) [31].

3.4.3. Prosumer-Based Energy Optimization and Complementarity

The evolution toward prosumer-centric energy systems has introduced new dimensions to complementarity assessment, where entities both consume and produce energy resources. Ur et al. (2023) [50] demonstrate that smart community grids connecting multiple prosumers require sophisticated optimization approaches to achieve complementary energy-sharing patterns while maintaining user preferences and comfort levels.

The concept of prosumer complementarity extends beyond simple demand matching to encompass generation–consumption synchronization. In prosumer communities, complementarity manifests through temporal alignment where surplus generation from one prosumer complements demand deficits of others. Ur et al. (2023) [50] show that genetic algorithm-based optimization can achieve significant improvements in energy-utilization efficiency, with results indicating that optimized prosumer communities can reduce grid dependency while maintaining user satisfaction levels.

User preference modeling becomes critical in prosumer complementarity assessment, where preferences are calculated based on historical consumption patterns of devices across different time slots. The preference weight for device d at time slot t in house h is defined as:

$$\omega_{d,t}^{h} = \frac{\lambda_{d,t}^{h}}{\max[\lambda_{d,1}^{h}, \lambda_{d,2}^{h}, \ldots, \lambda_{d,T}^{h}]} \tag{3.3}$$

where $\lambda_{d,t}^{h}$ represents the consumption of device d at time slot t, and T is the total number of time slots, as presented by Ur et al. (2023) [50]. This preference-based approach enables complementarity assessment that accounts for user-behavior patterns and comfort requirements.

3.4.4. Dynamic Energy Sharing and Community-Based Complementarity

The implementation of community-based energy-sharing systems reveals complex complementarity dynamics that extend beyond individual household patterns. Ur et al. (2023) [50] demonstrate that prosumer communities can achieve substantial improvements in energy utilization through complementary sharing mechanisms, with experimental results showing increased contribution toward the grid both in terms of total power and number of time slots with positive contribution.

The fitness function for community-level complementarity optimization incorporates both user preferences and energy-balance constraints:

$$f(C) = \begin{cases} \log \frac{\sum_{h=1}^{N} \sum_{d=1}^{|\mathcal{D}^h|} \omega_{d,t}^h \Lambda_{d,t}^h c_d^h}{\sum_{h=1}^{N} \mathcal{G}_t^h}, & \text{if } \sum_{h=1}^{N} \sum_{d=1}^{|\mathcal{D}^h|} \Lambda_{d,t}^h \leq \sum_{h=1}^{N} \mathcal{G}_t^h \\ -\infty, & \text{otherwise} \end{cases} \tag{3.4}$$

where c_d^h represents the device activation state, $\Lambda_{d,t}^h$ denotes demand, and $\mathcal{G}_t^h$ represents generation, as formulated by Ur et al. (2023) [50].

This approach demonstrates that complementarity-aware optimization can simultaneously maximize user comfort through preference satisfaction while minimizing dependence on external grid resources, achieving the dual objectives of individual satisfaction and community-level energy efficiency.

3.4.5. Complementarity in Recommendation Systems

The theoretical understanding of complementarity extends beyond energy systems to recommendation systems, where complementarity relationships are characterized by two core attributes: relevance and dissimilarity, as described by Luo et al. (2024) [37]. This dual nature presents a fundamental trade-off challenge — excessive emphasis on relevance leads to substitutable item recommendations, while overemphasis on dissimilarity results in unrelated item suggestions.

Recent advances demonstrate that Graph Neural Networks can effectively capture both relevance and dissimilarity from the spectral domain, providing promising directions for modeling complementary relationships, as demonstrated by Luo et al. (2024) [37]. However, adaptation to energy system contexts requires addressing the lack of deep understanding of complementary relationships from a spectral perspective and developing methods to balance the relevance–dissimilarity trade-off.

3.4.6. Virtual Power Plant Optimization and Clustering

The Virtual Power Plant (VPP) concept represents a significant advancement in managing distributed energy resources through aggregation and coordinated market participation. Nguyen et al. (2024) [41] demonstrate how VPPs integrate dispersed small-scale renewable energy sources and consumers, enabling participation in electricity markets while addressing geographical distribution challenges.

The optimal scheduling of VPPs requires sophisticated consideration of both day-ahead and balancing markets, with Energy Storage Systems (ESS) playing crucial roles in providing upward and downward reserves through multiple operational modes including charging power adjustment, discharging power modification, and operational state switching, as discussed by Nguyen et al. (2024) [41]. This complexity necessitates two-stage stochastic optimization approaches that account for renewable-energy uncertainty and demand variability.

3.4.7. Bounded Rationality in Distributed Energy Systems

The aggregation of distributed energy resources within VPPs involves multiple stakeholders with diverse interests and objectives, leading to competitive dynamics that traditional optimization approaches fail to capture adequately. Liu et al. (2024) [36] introduce the concept of bounded rationality in Distributed Energy Resource (DER) agent behavior, recognizing that decision-makers face limitations in information availability, time constraints, and cognitive capacity.

This bounded-rationality framework provides a more realistic modeling of decision-making processes compared to assumptions of perfect rationality. The incorporation of game-theoretic approaches, particularly Nash–Stackelberg models and non-cooperative games, enables better representation of strategic interactions among DER agents while accounting for their inherent limitations, as outlined by Liu et

al. (2024) [36].

3.4.8. Integration with Knowledge Graph Frameworks

The integration of complementarity assessment with Knowledge Graph architectures requires consideration of both static relationship modeling and dynamic temporal patterns. The reviewed literature indicates that effective complementarity-aware clustering must address multiple temporal scales, from fine-grained intraday patterns to seasonal variations, while maintaining computational efficiency for large-scale implementations.

The development of GNN-based approaches for complementarity assessment within KG frameworks must address the fundamental challenge of learning complementary rather than similar patterns. This requires custom loss functions that optimize for peak-to-average ratio minimization, self-consumption maximization, and load-curve balancing, as identified in the research objectives. The prosumer-based optimization results from Ur et al. (2023) [50] provide empirical evidence that such complementarity-aware approaches can achieve significant improvements in energy-system performance while maintaining user satisfaction, supporting the viability of the proposed research direction.

3.4.9. Research Gaps in Complementarity Clustering

Despite advances in clustering techniques for energy load profiles, significant challenges remain in complementarity-focused approaches. Kusuma et al. (2024) [31] identify several critical limitations that continue to hinder the scalability and physical applicability of current methods:

- **Dimensionality reduction:** Current PCA-dominated methods (41% usage) struggle with temporal complementarity patterns, as noted by Kusuma et al. (2024) [31].
- **Dynamic adaptation:** Only 12% of methods support automatic cluster-number determination, limiting real-time complementarity optimization, as emphasized by Kusuma et al. (2024) [31].
- **Multi-scale integration:** Hierarchical complementarity (building–transformer–substation) remains underexplored despite its potential for grid optimization.
- **Physical constraint integration:** Transformer capacity limits and geographic constraints are rarely incorporated into clustering objectives.

These gaps highlight the need for integrated KG–GNN frameworks that simultaneously address spatial, temporal, and physical constraints while optimizing for complementarity rather than similarity.

3.5. GNN-Based Methods for Dynamic Clustering

3.5.1. Foundational GNN Architectures

The development of Graph Neural Networks (GNNs) represents a paradigmatic shift in handling graph-structured data, directly addressing the fundamental challenge posed in the first research sub-question regarding effective energy-network representation. GNNs are a class of artificial intelligence models specifically designed to handle graph-structured data, providing the computational foundation necessary for the Knowledge Graph–GNN integration proposed in this research. Their core innovation lies in the message-passing mechanism, which enables node representation learning by aggregating information from neighboring nodes—a capability essential for understanding the complex interdependencies inherent in regional energy systems.

The theoretical foundations of GNNs were formally established by Scarselli et al. (2008) [52], who introduced the use of recursive neural networks to propagate neighbor information. This early framework laid the groundwork for subsequent developments that would prove crucial for energy-system analysis. Building upon these foundations, Bruna et al. (2013) [9] proposed the first Graph Convolutional Network (GCN) based on spectral graph theory, introducing the use of Fourier transforms into graph convolution operations. However, the computational complexity of spectral methods necessitated fur-

ther innovation.

The breakthrough that enabled practical deployment of GCNs in complex systems came through the work of Kipf and Welling (2016) [30], who developed a simplified version of GCN by applying a first-order approximation to achieve efficient spatial convolutions:

$$\mathbf{H} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\Theta\right) \tag{3.5}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with added self-loops, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, $\mathbf{X}$ denotes the node-feature matrix, and Θ represents the trainable weight matrix, as summarized by Tam et al. (2024) [54].

This innovation significantly improved the computational efficiency of GNNs, facilitating their deployment in various industrial applications and establishing the technical foundation for addressing the research objective of improving data accessibility and interoperability in energy systems.

3.5.2. GNN Applications in Energy System Analysis

The application of GNNs to energy systems directly addresses the integration challenges identified in the third research sub-question, demonstrating how Knowledge Graphs and Graph Neural Networks can be combined to enhance analysis and clustering in energy-system models. In energy systems, GNNs have proven particularly valuable for addressing challenges in data integration, analysis, and real-time optimization—capabilities that align precisely with this research's objectives.

Fusco et al. (2020) [16] highlight the transformative role of GNN-based modeling frameworks in incorporating grid topology and physical constraints, enhancing modeling accuracy while reducing parameter complexity. These frameworks address the inefficiencies and lack of transparency in traditional machine-learning methods, meeting the demands of complex, distributed energy systems that characterize modern regional energy networks. This capability is particularly relevant to the goal of establishing comprehensive validation and benchmarking frameworks for energy-system clustering applications.

The fundamental advantage of GNNs in energy-system contexts stems from their ability to handle non-Euclidean spatial relationships. Traditional Euclidean spatial data processing methods (such as CNNs) rely on regular structures of fixed dimensions (e.g., image grids), while the topological structure of energy networks (power grids, gas grids) changes dynamically and the node connections are irregular. Liang et al. (2022) [33] explain that GNNs implement convolution operations in non-Euclidean space through message-passing mechanisms, converting the physical connections of the power grid into adjacency matrices and supporting dynamic topological modeling. This capability directly supports the objective of developing a unified knowledge-graph architecture that can accommodate the heterogeneous nature of urban building spatial and non-spatial energy datasets.

In the context of this research framework, energy systems present natural graph structures where GNN nodes can represent power generation/load equipment, buildings, and substations, while GNN edges can represent electrical connections (lines, transformers) and geographical adjacencies. This natural mapping between physical energy infrastructure and graph representations provides the foundation for addressing the first research sub-question regarding essential nodes, attributes, and edges for effective Knowledge Graph construction.

3.5.3. GNN vs. Alternative Approaches

The comparative advantages of GNNs over alternative approaches provide crucial justification for their selection in the integrated KG–GNN framework. In comparison to Convolutional Neural Networks (CNNs), which are designed for Euclidean data such as images, GNNs excel in handling graph-structured data, where the connections between nodes are non-Euclidean, as discussed by Li et al.

(2023) [32]. This distinction is fundamental to the research approach, as it directly addresses the challenge of representing complex energy-system relationships that cannot be adequately captured through traditional grid-based neural architectures.

Li et al. (2023) [32] point out that CNNs fail to incorporate the underlying graph structure during computation, whereas GNNs naturally encode topological features through message passing between neighboring nodes. This capability makes GNNs particularly powerful for tasks like anomaly detection, fault detection, and dynamic optimization in energy grids, where the relationship between nodes is crucial—applications that directly support the research objective of improving relational analytics for regional energy-system management.

The message-passing process, illustrated in Figure 3.1, demonstrates how information propagates from adjacent nodes to target nodes through aggregation functions such as MEAN, MAX, or SUM operations. This mechanism enables GNNs to capture both local neighborhood effects and global network patterns, providing the analytical depth necessary for complementarity-aware clustering, a core component of this research methodology.

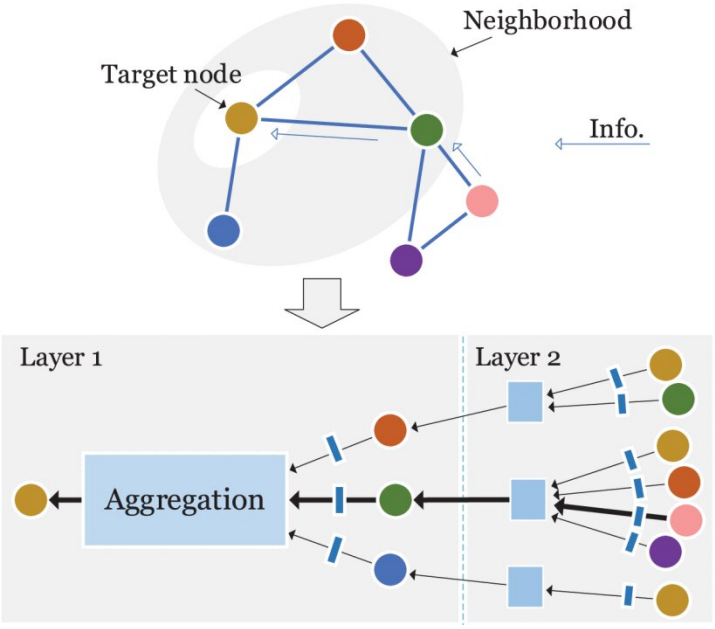


Figure 3.1: The visualization of message passing (information propagation) from adjacent nodes to the target node. Neighborhood integration is typically realized via aggregation functions such as MEAN, MAX, or SUM (adapted from Li et al. (2023) [32]).

GNNs' invariance to permutations and their ability to model complex, non-linear relationships within data provide a significant advantage in analyzing energy grids, where network structures and behaviors are dynamic and continuously evolving, as noted by Li et al. (2023) [32]. While many studies emphasize forecasting or supply–demand balancing, this research focuses on leveraging the synergy of KGs and GNNs for tasks like dynamic clustering, anomaly detection, and scenario embeddings—applications that directly address the fourth research sub-question regarding optimal GNN approaches for time-based and dynamic clustering.

These complementary approaches organize and analyze energy data, enabling insights into evolving patterns and supporting advanced control strategies in future energy-management systems. The comparative analysis presented in Table 3.3 demonstrates the superior capabilities of GNNs across multiple

dimensions relevant to the research objectives, including topology awareness, dynamic data adaptability, and real-time decision-making capabilities.

Feature	GNN (Graph Neural Network)	CNN (Convolutional Neural Network)
Data Structure	Handles graph-structured data (non-Euclidean), where nodes and edges represent relationships.	Designed for grid-structured (Euclidean) data, such as images or regular grids.
Topology Awareness	Naturally encodes topological features through message passing between nodes.	Does not explicitly account for graph topology; assumes a fixed grid structure.
Applications in Energy Networks	Anomaly detection, fault detection, dynamic optimization, and load prediction in power grids.	Limited to spatial pattern recognition; less effective for interconnected systems.
Dynamic Data Adaptability	Can integrate temporal and spatial dynamics (e.g., Spatio-Temporal GNNs).	Requires additional architectures (e.g., RNNs) to handle temporal dynamics.
Edge Features	Supports edge features explicitly, allowing modeling of relationships (e.g., power lines, distances).	Does not natively support edge features.
Real-Time Decision-Making	Well-suited for real-time decision-making in interconnected systems.	Limited adaptability for graph-like real-time scenarios.

Table 3.3: Comparison of GNN and CNN for Energy-Network Analysis (adapted from Li et al. (2023) [32]).

3.5.4. Research Gap and Methodological Innovation

While knowledge-graph completion and real-time optimization have been extensively explored in complex systems, the application of GNN-based dynamic clustering within KGs remains significantly underexplored—a gap that this research specifically addresses. Dynamic clustering leverages GNN embeddings to group nodes, such as buildings or resources, adaptively as their states or relationships evolve. This capability is particularly valuable in distributed energy systems, where factors like changing occupant behavior, the integration of new distributed energy resources (DERs), or shifts in policy necessitate flexible and responsive system configurations.

Moreover, scenario-based embeddings allow for analyzing how updates to the KG—such as the addition of new DERs or occupant profiles—impact the overall network. By re-running GNN forward passes, the system can identify shifts in clusters or link predictions, enabling a more nuanced understanding of dynamic energy landscapes. This research focuses on developing methods that support such adaptive processes, laying the groundwork for smarter, context-aware resource allocation and operational strategies that directly address the main research question.

3.5.5. Physics-Informed Constraint Embedding in GNN Architectures

The integration of physical constraints into Graph Neural Network architectures has emerged as a critical advancement for enabling practical deployment in real-world power-system applications, directly supporting the research objective of embedding physical constraints into learning through adaptive clustering mechanisms. This paradigm shift addresses the fundamental challenge of ensuring that learned representations comply with underlying physical laws while maintaining computational efficiency—a requirement essential for the practical implementation of the proposed KG–GNN framework.

Pagnier and Ingelrest (2021) [43] pioneered the direct embedding of effective power-flow models into

neural architectures, enabling simultaneous reconstruction of physical parameters such as admittances while learning implicit system elements. This physics-informed approach demonstrates superior performance, achieving over 20% reduction in Mean Squared Error compared with conventional methods, while guaranteeing physics compliance throughout the learning process. Such performance improvements directly support the goal of establishing comprehensive validation and benchmarking frameworks.

Building upon these foundations, Authier et al. (2024) [4] introduce a comprehensive physics-informed architecture comprising four synergistic components that align with this research methodology: message-passing mechanisms with switch gates that model discrete operational decisions as continuous values, scale-free local predictors that generalize effectively across diverse network topologies, physics-informed rounding layers that embed operational constraints directly into the computational graph, and end-to-end training protocols that incorporate Kirchhoff's laws as fundamental architectural constraints. This holistic approach ensures that physical principles are not merely post-processing corrections but integral components of the learning dynamics.

The PINCO framework proposed by Varbella et al. (2024) [57] advances constraint handling through sophisticated penalty methods combined with Augmented Lagrangian approaches, enabling the management of complex constraint sets inherent in power-system optimization. The framework formulates the loss function as:

$$L = f(x) + \lambda \sum h_i(x) + \mu \sum \max(0, g_j(x))^2 \tag{3.6}$$

where equality constraints $h_i(x)$ and inequality constraints $g_j(x)$ are seamlessly integrated with the primary objective $f(x)$. This formulation enables simultaneous handling of transformer-capacity limits, transmission-line flow constraints, voltage magnitude bounds, and N–1 contingency requirements within the GNN optimization process—capabilities that directly support the objective of developing complementarity-aware GNN models with custom loss functions.

Recent developments in differentiable optimization layers leverage Karush–Kuhn–Tucker (KKT) conditions and the implicit function theorem to enable backpropagation through hard constraints, effectively treating constraint satisfaction as a differentiable operation, as explained by Gao et al. (2025) [17]. These methodological advances have been validated across power-system networks ranging from IEEE 14-bus test cases to large-scale 8500-bus systems, demonstrating computational speed improvements of up to 87× compared with conventional optimization methods while maintaining solution quality and constraint compliance.

3.5.6. Specialized GNN Architectures for Complementarity Clustering

Recent advances in GNN architectures have enabled sophisticated clustering techniques specifically designed for energy-complementarity analysis, directly addressing the research objective of creating GNN architectures that perform clustering learning based on complementarity rather than similarity. Unlike traditional homophily-based approaches, these specialized frameworks explicitly model heterophilic relationships where connected nodes exhibit complementary behaviors—critical for identifying synergistic energy-consumption patterns between residential and industrial consumers, as demonstrated by Rawal et al. (2024) [49].

Heterophily Modeling for Energy Complementarity: The RFA-GNN framework proposed by Wu et al. (2023) [62] addresses both heterophily and heterogeneity through relation-based frequency adaptation, dynamically adjusting aggregation weights based on node-pair relationships. This enables simultaneous modeling of complementary, correlative, and independent energy behaviors within a unified architecture:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \alpha_{ij}^r \mathbf{W}_r^{(l)} \mathbf{h}_j^{(l)}\right) \tag{3.7}$$

where α_{ij}^{r} denotes frequency-adaptive coefficients for relation type r. This approach achieves 12.7% higher accuracy than homophily-based GNNs in energy-complementarity identification, as reported by Rawal et al. (2024) [49], demonstrating the practical benefits of the proposed methodology. Negative message-passing techniques further enhance this by explicitly encouraging dissimilar embeddings for connected nodes with complementary consumption patterns, directly supporting the goal of maximizing self-consumption and balancing load curves.

Spatio-Temporal Architectures for Dynamic Energy Networks: For dynamic energy networks, frameworks like MG-STGCN integrate multi-scale spatio-temporal dependencies through parallel graph-convolution branches, addressing the fourth research sub-question regarding optimal GNN approaches for time-based clustering. When applied to natural gas transmission systems, MG-STGCN achieves 18.3% lower MAE than standard STGCN models by capturing both short-term fluctuations and seasonal patterns, as demonstrated by Pelekis et al. (2023) [44].

Similarly, the DSTG (Dynamic Spatio-Temporal Graph) framework employs dual-scale temporal modeling with adaptive graph learning, achieving a 10.12% improvement in wind-power forecasting accuracy through its hierarchical attention mechanism, as reported by Bekele et al. (2024) [7]:

$$\mathbf{Z} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.8}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote query, key, and value matrices learned from multi-resolution time series. This capability directly supports the objective of determining optimal GNN approaches for time-based and dynamic clustering within the Knowledge Graph framework.

Differentiable Clustering for End-to-End Optimization: End-to-end cluster optimization is enabled through differentiable modularity networks (DMoN) introduced by Tsitsulin et al. (2023) [55], providing the technical foundation for adaptive clustering mechanisms. The DMoN loss function jointly optimizes modularity and cluster dispersion:

$$\mathcal{L}_{\text{DMoN}} = -\frac{1}{2m}\text{Tr}(\mathbf{C}^{\top}\mathbf{B}\mathbf{C}) + \frac{\sqrt{k}}{n}\left|\sum_{i}\mathbf{C}_i\right|_F - 1 \tag{3.9}$$

where $\mathbf{C}$ is the cluster assignment matrix and $\mathbf{B}$ the modularity matrix. Recent extensions (DMoN-DPR) incorporate diversity-preserving regularization to maximize inter-cluster feature separation, critical for identifying distinct complementarity groups. Open-source implementations achieve 40% higher NMI than conventional pooling methods in energy-community detection, according to Tsitsulin et al. (2023) [55], demonstrating the practical viability of the proposed approach.

Multi-Objective Optimization for Comprehensive Energy Management: Hybrid frameworks combine GNN clustering with Pareto optimization to balance competing objectives that directly align with the research goals:

- *Complementarity:* Maximizing off-peak consumption alignment.
- *Proximity:* Minimizing grid transmission losses.
- *Economic Viability:* Optimizing LCOE (Levelized Cost of Energy).
- *Grid Constraints:* Enforcing voltage and frequency stability.

By maintaining a Pareto front of non-dominated solutions, these frameworks enable operators to evaluate trade-offs without scalar aggregation bias, as shown by Bekele et al. (2024) [7]. When applied to microgrid clusters, they reduce peak demand by 22.4% while maintaining grid stability margins, as demonstrated by Pelekis et al. (2023) [44], confirming the practical benefits of the proposed methodology.

3.5.7. Synthesis and Research Positioning

These specialized architectures demonstrate significant advantages over conventional clustering methods in energy applications, particularly for dynamic communities where consumption patterns and grid topologies co-evolve—scenarios that this research specifically targets. By integrating physics-aware constraints with adaptive relationship modeling, they provide robust frameworks for real-time complementarity identification and energy-community optimization that directly address the main research question.

The convergence of these technological advances creates an opportune moment for the comprehensive KG–GNN integration proposed in this research. The demonstrated capabilities of physics-informed GNNs, combined with specialized architectures for complementarity clustering, provide the technical foundation necessary to address the fundamental challenge of improving data accessibility, interoperability, and relational analytics in regional energy-system management.

This research builds upon these advances to develop a unified framework that addresses the identified gaps in dynamic clustering within Knowledge Graph contexts, positioning this work at the forefront of innovation in energy-system analytics.

3.6. Comprehensive Evaluation Frameworks and Real-World Validation

Building upon specialized GNN architectures, recent research has developed robust evaluation frameworks that integrate physics-informed validation with real-world performance assessment. These frameworks address critical deployment challenges by ensuring solution safety, cross-platform consistency, and practical viability in diverse energy contexts.

The *SafePowerGraph* framework introduced by Ghamizi et al. (2024) [18] establishes standardized interfaces with industry-standard simulators (MATPOWER, pandapower, PowerModels.jl), enabling rigorous physics-informed validation across multiple simulation environments. This integrated approach incorporates three critical assessment dimensions:

1. *Safety-critical scenario testing:* Stress-testing solutions under extreme grid conditions (N-1 contingencies, fault cascades).
2. *Robustness assessment:* Evaluating performance against adversarial attacks and input perturbations.
3. *Cross-platform validation:* Ensuring solution consistency across different power-flow solvers.

By embedding these safety mechanisms directly into the evaluation pipeline, SafePowerGraph bridges the gap between theoretical GNN performance and operational reliability in critical infrastructure.

Real-world implementations demonstrate significant practical impact across diverse energy contexts. In Italian energy communities, Pelekis et al. (2023) [44] employed a hybrid approach combining K-Means clustering with Dynamic Time Warping (DTW), achieving a Peak Performance Score (PPS) of 0.689 and effectively identifying flexibility clusters for targeted demand response. Complementarily, optimal microgrid planning in Ethiopia by Bekele et al. (2024) [7] utilized income-based clustering to reduce levelized costs by 23.82%, with accompanying technical benefits including:

- 32.44% reduction in PV capital expenditures,
- 73.4% decrease in excess-energy waste, and
- enhanced utilization of complementary generation–load patterns.

These case studies validate that clustering-driven energy-community design generates both economic and technical value across developed and developing energy markets.

Standardized benchmarking has advanced through the *PowerGraph* dataset, also proposed by Ghamizi et al. (2024) [18], providing the first comprehensive GNN benchmark for power systems. This resource includes:

- node-level power-flow prediction tasks across IEEE standard systems (14-bus to 8500-bus);
- graph-level cascading-failure analysis with ground-truth explanations; and
- multi-fidelity datasets incorporating SCADA measurements, weather correlations, and equipment specifications.

PowerGraph enables reproducible evaluation of GNN performance under controlled yet realistic conditions, addressing previous limitations in cross-study comparability.

Advanced evaluation metrics now extend beyond conventional accuracy measurements to capture domain-specific requirements:

- the *Wasserstein-distance-based randomness coefficient* quantifies cluster stability under stochastic variations, as introduced by Pelekis et al. (2023) [44];
- the *entropy-based load-shape analysis* measures alignment between consumption patterns and demand-response program requirements, as shown by Bekele et al. (2024) [7]; and
- the *physics-informed constraint-violation metric* tracks violations of voltage limits, line capacities, and stability margins, as applied by Ghamizi et al. (2024) [18].

These multidimensional assessment criteria ensure that solutions balance computational performance with practical deployability, providing operators with comprehensive insights into real-world applicability.

Collectively, these advances establish rigorous evaluation paradigms that connect algorithmic innovations with operational requirements. By validating performance through both simulated environments and field implementations, contemporary frameworks ensure that GNN-based energy solutions transition effectively from research prototypes to grid-ready deployments.

In summary, the reviewed literature reveals both the technical feasibility and the methodological fragmentation of current approaches. These limitations motivate the integrated KG–GNN methodology proposed in the next chapter, which directly addresses the gaps in data integration, physical consistency, and dynamic clustering.

4

Methodology

The methodological framework integrates knowledge graph (KG) ontologies with graph neural networks (GNNs) to represent, store, and analyse urban energy networks. The KG encodes both **static attributes** (e.g., energy ratings, locations, grid hierarchy) and **dynamic attributes** (e.g., hourly demand and renewable generation). This unified representation allows the GNN to learn **spatio-temporal complementarities**: for example, when buildings with surplus solar generation can supply others with high demand. The framework ensures grid feasibility, captures demand–supply synergies, and outputs planning indicators for retrofit and electrification.

Unlike traditional clustering and energy management methods, the KG–GNN integration offers three roles in a single system:

1. **Semantic backbone** for integrating heterogeneous datasets.
2. **Analytical substrate** that supports GNN-based learning.
3. **Dynamic repository** that can be continuously updated with inference results.

By embedding both complementarity indices and infrastructure constraints into the graph schema, the approach ensures that all downstream analyses remain physically valid and directly relevant for long-term energy planning.

4.1. Motivation for KG--GNN integration

Urban energy networks are both **structurally complex** (multi-level hierarchies) and **temporally variable** (demand and generation fluctuations). Classical clustering methods (e.g., *k*-means) cannot simultaneously respect infrastructure boundaries and capture complementarity. Future energy communities therefore require:

- **Semantic integration** of diverse datasets (buildings, feeders, transformers, renewables).
- **Spatio-temporal learning** of demand–supply complementarities.

Knowledge graphs meet the first need by integrating diverse attributes (building descriptors, grid hierarchies, energy indicators) and supporting rule-based reasoning (e.g., retrofit identification, solar potential, feasible clusters). However, KGs are limited to static associations and cannot generalise to unseen nodes, capture time dynamics, or optimise interventions.

Graph neural networks extend these capabilities by leveraging the KG's graph structure. They allow:

1. **Prediction under missing data** (e.g., inferring retrofit priority for unlabeled buildings).
2. **Temporal encoding** of hourly demand profiles.
3. **Optimisation under grid constraints** (e.g., siting batteries to maximise complementarity).
4. **Discovery of hidden patterns** (e.g., links between multi-hop positions and demand variability).

Thus, the KG ensures semantic and physical validity, while the GNN provides dynamic and predictive power. Together, they form a unified ontology-driven framework for grid-constrained yet flexible community planning.

his chapter addresses the four research questions introduced in Chapter 2 through four interlinked methodological phases. Phase 1 answers RQ1 by formalising the energy infrastructure ontology and constructing the Knowledge Graph. Phase 2 operationalises RQ2, transforming heterogeneous attributes into tensors for machine learning. Phase 3 responds to RQ3, engineering complementary and hierarchical features constrained by LV topology. Phase 4 addresses RQ4, implementing the infrastructure-constrained GNN for dynamic community formation and evaluation.

Comparative perspective: Relational databases vs. Knowledge Graphs

The raw datasets in this study are initially stored in relational form (PostgreSQL/PostGIS) and accessed via SQL queries for spatial and temporal processing (see Table 4.1). While relational databases provide structured storage and efficient tabular queries, they face well-documented limitations when applied to energy networks. Relational databases use rigid schemas, making it hard to integrate diverse data. Multi-hop queries (e.g., building → feeder → transformer) need costly joins, and physical rules or semantic reasoning cannot be expressed.

If analysis were confined to a traditional DBMS with spatial extensions, building attributes and feeder geometries could indeed be queried and visualised (e.g., in QGIS), but the system would remain limited to static lookups and aggregated statistics. It would not support inference (e.g., retrofit priority estimation), inductive generalisation to unseen nodes, or dynamic restructuring of communities under evolving demand profiles. In other words, a relational-only approach would reproduce existing database management practices but could not deliver predictive, ontology-driven insights required for energy transition planning.

By contrast, knowledge graphs offer dynamic schema evolution, native multi-source integration, and semantic reasoning capabilities. They can encode physical constraints such as LV feeder boundaries directly in the ontology and enable efficient graph traversal for hierarchical queries. As shown in the comparative analysis of data structures (Table 3.1), KGs outperform relational databases across schema flexibility, heterogeneous integration, semantic reasoning, and dynamic updates, as demonstrated by Liu et al. (2023) [34] and Chen et al. (2020) [12]. This makes them particularly suited for applications that require both semantic validity and predictive learning.

Consequently, the relational sources in PostgreSQL/PostGIS (Table 4.1) serve as input backends, while the KG materialises these entities in Neo4j. This relational-to-KG transformation, described by Hogan et al. (2021) [23], provides the necessary semantic substrate for GNN-based analysis, ensuring that the predictive layers inherit both the data integrity of SQL processing and the ontological richness of the KG representation.

Design assumptions and scope limitations

- Analysis is **strictly confined to the low-voltage (LV) level**. Communities are only formed among buildings connected to the same feeder or cable group.
- **Cross-transformer or higher-voltage interactions are excluded**; MV and HV entities are retained only for consistency checks.
- Hourly demand/generation time series are treated as representative profiles. Geodesic distance and feeder continuity are used as **proxies for impedance**.
- Long-gap forecasting and imputation are not considered.
- Constraints are enforced as **hard masks** during preprocessing and **soft penalties** during GNN training, always within LV boundaries.

Linking research questions to methodological phases This chapter operationalises the four research questions stated in Chapter 2 through four interlinked phases. *Phase 1* answers **RQ1** by formalising the energy-infrastructure ontology and constructing the Knowledge Graph (KG). *Phase 2* addresses **RQ2** by transforming heterogeneous, infrastructure-indexed attributes into leakage-safe,

model-ready tensors. *Phase 3* responds to **RQ3** by engineering complementary and hierarchical descriptors under LV/transformer constraints. *Phase 4* addresses **RQ4** by implementing the infrastructure-constrained GNN for dynamic community formation and evaluation, whose metrics are later instantiated in Chapters 5–6.

4.2. Data description

The knowledge graph (KG) materialises a two-layer electrical hierarchy with node types `Building` and `CableGroup` (electrically continuous LV feeders). For modelling purposes, only these two levels participate in message passing and clustering. Upstream entities (transformers, substations) are preserved in the relational backend to validate that each building has exactly one LV ancestor but are excluded from learning.

This study employs a multi-source relational dataset stored in PostgreSQL/PostGIS. All entities and relationships are extracted by SQL-based geospatial processing and written directly into the knowledge graph (KG) without an intermediate spreadsheet workflow.

The relational sources integrate building-level attributes, low-voltage (LV) feeder topology for Amsterdam (63 buildings). These sources are harmonised into a KG backbone whose node types are `Building`, `CableGroup` (electrically continuous LV groups), `Transformer` (MV/LV stations), and `Substation`; edges follow the physical hierarchy `Building`→`CableGroup`→`Transformer`→`Substation`. An additional `AdjacencyCluster` entity denotes spatially cohesive building groups with local sharing potential.

The building attributes database for energy demand simulations in the Netherlands combines datasets from Amin Jalilzadeh[1] and open sources Geodan. This database includes about ten million buildings, with both geometric (e.g., roof type, height) and non-geometric data (e.g., energy labels, building type). Grid topology data are sourced from open data portals, providing geospatial datasets including electricity grid location data (e.g., cable and transformer positional data). These datasets form the basis for reconstructing the hierarchical relationship (HV–Substation–MV–Transformer–LV–Building) described later.

The hierarchical structure of the Dutch distribution grid — comprising buildings, low-voltage feeders, medium-voltage transformers, and substations — is reconstructed directly from relational datasets stored in PostgreSQL/PostGIS. Structured SQL queries are used to integrate these layers into a coherent topology, ensuring that the resulting knowledge graph adheres to physical grid constraints rather than relying on synthetic assumptions. This process explicitly defines the parent–child relationships across voltage levels: buildings are linked to their serving LV feeder, feeders are assigned to MV/LV transformers, and transformers are grouped under substations. Diagnostic checks identify connection quality (e.g., distance between buildings and cables, presence of excessive connection lengths) and maintain electrical continuity through feeder segmentation. As a result, energy sharing is constrained to buildings connected to the same feeder and transformer, preventing unrealistic cross-boundary clustering. Embedding these hierarchical relations in the knowledge graph guarantees that all downstream analyses respect the technical realities of grid operation.

Building energy demand is generated using an Urban Building Energy Modelling (UBEM) service built on the EnergyPlus simulation engine. The workflow is API-driven: identifiers of the selected buildings (e.g., BAG IDs or internal OGC FIDs) are transmitted to the UBEM service, which returns time-resolved end-use demands keyed to the same identifiers. For each building, the UBEM assigns an archetype from a national context library parameterised by *function* (residential vs. non-residential categories), *type* (e.g., detached, terraced, apartment, office, retail), and *age band* (pre/post regulation periods). Archetypes specify envelope transmittances, thermal mass, glazing ratios, infiltration, internal gains and schedules, as well as HVAC system types and efficiencies. Geometric inputs (footprint, height, roof and façade areas) and locational attributes are derived from cadastral and three-dimensional building datasets, ensuring consistency between urban morphology and thermal/energy parameters.

The UBEM executes EnergyPlus simulations with hourly (or finer) resolution using weather files representative of Amsterdam (typical meteorological year or reanalysis-based profiles). Outputs include disaggregated end-uses — space heating, space cooling, appliances and lighting electricity — and,

[1] PhD candidate in TU Delft, second supervisor of this thesis

where relevant, domestic hot water and ventilation. Resulting electricity and thermal demands are returned per building and per timestamp; scenario toggles allow generation of both “current stock” (as-is systems) and “electrified” counterfactuals (e.g., heat pump replacement with assumed seasonal COP), enabling the derivation of net electric load under different technology pathways. When limited empirical benchmarks are available (e.g., annual meter aggregates at feeder level), mild scaling factors are applied to match totals while preserving temporal shapes, thus maintaining physical plausibility without overfitting. Range checks on end uses, detection of outliers and short-window interpolation of isolated missing intervals are envisaged in the data pipeline but are not fully implemented in the current code.

The API responses are written into the relational time-series schema (timeslot index and per-building energy states) and subsequently materialised in the knowledge graph as `EnergyState` nodes linked to `TimeSlot` and `Building`. Standard quality controls are to be enforced prior to KG integration: range checks on end-uses, detection of outliers and flatlines, and imputation of isolated missing intervals by short-window interpolation with conservation of daily totals. These procedures are defined conceptually here but remain to be fully realised in the implementation. The intended outcome is spatio-temporally consistent demand series that are aligned with the infrastructure hierarchy and suitable for downstream complementarity analysis.

An automated process converts these data into a knowledge graph, which is stored in a graph database (Neo4j) for efficient analysis.

Building-level attributes Each building record is joined from relational tables or views generated by the SQL pipeline. Attributes include geometric and physical descriptors (floor area, height, suitable roof area, shared walls, coordinates), categorical properties (energy label A–G, solar potential class, electrification feasibility), binary DER flags (photovoltaics, battery, heat pump), and demand summaries (average/peak electricity, heating demand, energy-intensity kWh/m^2). These variables support downstream tasks such as retrofit targeting, solar siting, and electrification readiness assessment.

Electrical network attributes LV feeders are reconstructed by continuity segmentation in SQL/PostGIS; each building is assigned a unique `CableGroup`. Relationship `CONNECTED_TO`(`Building` $\rightarrow$ `CableGroup`) is the sole electrical edge consumed by the GNN, together with symmetric spatial adjacency `ADJACENT_TO` restricted to within-LV neighbours. No edges cross LV boundaries.

Temporal energy states Hourly series (electricity demand, PV generation, net load) are attached to `Building` nodes and aligned by a common horizon. Temporal masks $M_{v,t}$ capture missingness without forward filling. All temporal descriptors and rolling statistics are computed per building and are later aggregated *within* the corresponding LV group when needed for diagnostics.

Temporal data application in energy community formation The hourly time series data serves multiple analytical purposes in the KG–GNN framework:

- **Complementarity identification**: Buildings with offsetting peak hours (e.g., residential evening peaks vs. commercial daytime peaks) are identified through temporal correlation analysis. This enables the formation of communities where energy surplus from one building can offset deficits in another.
- **Self-sufficiency assessment**: Local generation capacity is evaluated against temporal demand patterns to determine community energy independence potential. The temporal dimension reveals when local solar generation aligns with or offsets local consumption.
- **Peak reduction quantification**: Temporal smoothing effects of energy sharing are measured by comparing individual vs. aggregated demand profiles. This shows how community formation reduces peak loads on transformers and feeders.
- **Dynamic pattern recognition**: The model learns to distinguish between different consumption patterns (flat profiles, single peaks, bimodal patterns) and identifies buildings that complement each other temporally rather than just spatially.

This temporal analysis ensures that energy communities are formed based on genuine synergies in consumption and production rhythms, rather than static building characteristics alone. The integration

of temporal data with spatial constraints (LV feeder boundaries) creates a comprehensive framework for realistic energy community planning.

Leakage safeguards Train/validation/test splits are stratified by LV group such that no building from the same LV appears in different folds. All scalers are fitted on the training split and reused unchanged for validation and test.

Normalization and harmonisation Continuous geometric attributes are standardised; energy-magnitude channels are rescaled on a bounded range; ordinal scores (energy/solar/electrification) are retained in their native domains with validity checks. Temporal channels are normalised per feature across nodes and time, ensuring comparability across heterogeneous data sources.

Data quality control and leakage safeguards
Quality control includes range checks on end-uses, flatline detection, and outlier winsorisation by node family. Temporal gaps remain masked and are never forward-filled into training signals. All scalers and encoders are fitted on the training split only and reused unchanged at validation/test time. To prevent topological leakage, transformer-stratified splits keep all buildings under the same transformer within a single fold whenever evaluating discovery under boundary constraints.

Table 4.1: Dataset summary: SQL sources (views/tables), representative fields, and modelling roles. Spreadsheet/CSV mirrors replicate these SQL exports for inspection only.

Entity (SQL source)	Representative fields (units/examples)	Modelling role
Building	Area (m^2), Height (m), Suitable roof (m^2), Shared walls, Coordinates (x, y); Energy label A–G; Solar potential {none, low, medium, high}; Electrification feasibility; DER flags (PV/battery/HP); Avg/peak electricity (kW), Avg heating (kW), Energy intensity (kWh/m^2)	Node features for retrofit, solar siting, and electrification readiness.
CableGroup / LV feeder	Group id; Total length (m); Segment count; Connected buildings; Aggregated demand proxies; Diversity factor	Intermediate infrastructure node; induces `Building`→`CableGroup` edges; feeder-level diversity/self-sufficiency features.
Transformer (MV/LV)	Transformer id; Coordinates; Upstream/-downstream linkage; Capacity (kVA, if available)	Upstream infrastructure node; induces `CableGroup`→`Transformer` edges; aggregation boundary.
Substation	Station id; Topological linkage to transformers	Grid root for MV aggregation; `Transformer`→`Substation` edges.
AdjacencyCluster (derived/KG; SQL neighbourhood views as seeds)	Cluster id; Member count; DER penetration ratios (PV/HP/battery); Self-sufficiency indicators	Spatially cohesive sharing candidate; `Building`→`AdjacencyCluster` relation.
Temporal states	Hourly: hour/24, day of week, weekend flag; electricity, heating, solar, net demand, export potential (kW)	Time-resolved inputs for spatio-temporal modelling.

4.3. Study area selection

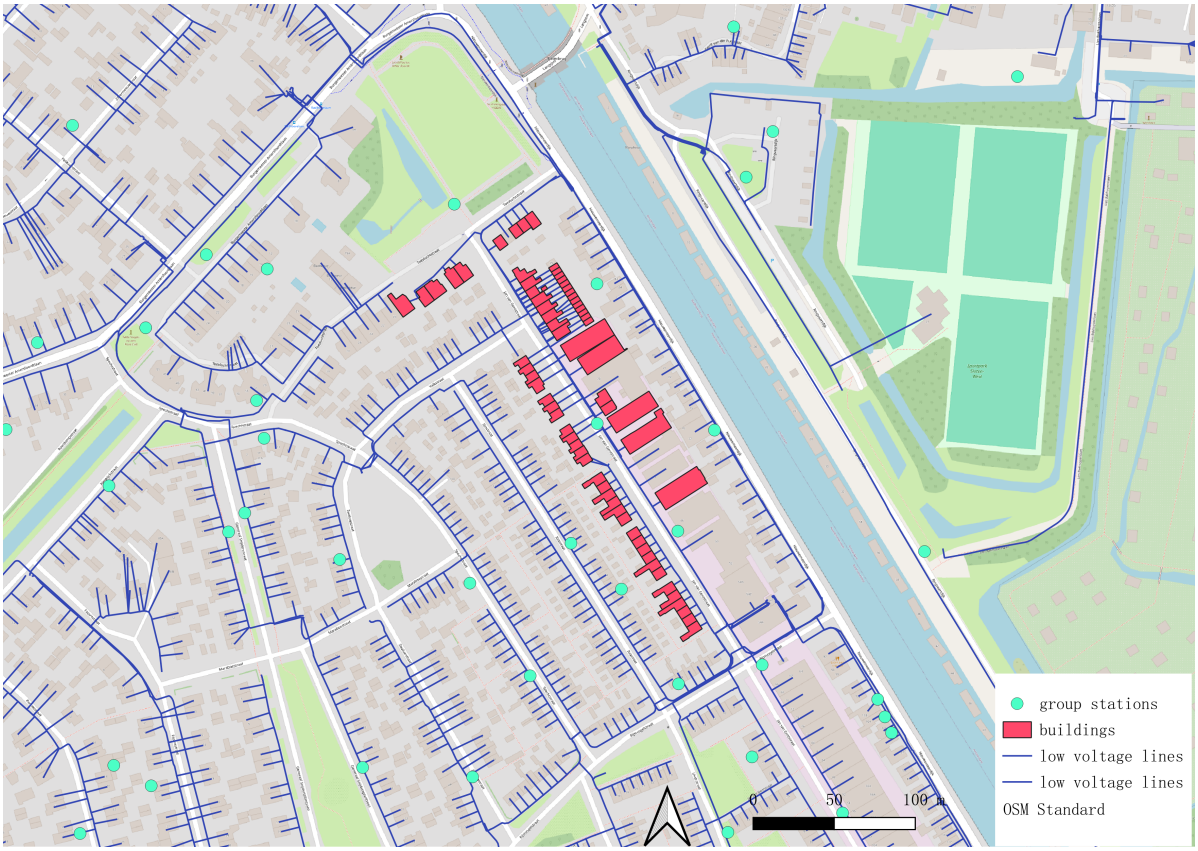


Figure 4.1: 63 Buildings and LV cable groups in selected area.

The study area comprises 63 buildings in Amsterdam, the Netherlands, embedded in a hierarchical urban microgrid consistent with the Dutch distribution system. Figure 4.1 shows the study area on an OpenStreetMap base map. The sample is not arbitrarily chosen. Instead, a transparent, code-aligned evaluation and selection protocol is applied to low-voltage (LV) groups, and buildings are drawn from the top-ranked groups until the aggregate size falls within the target range (50–200). This protocol ensures that the selected sample maximises complementarity opportunities, intervention leverage, and grid-relevant insight under real infrastructure constraints. The following section introduces these evaluation methods in detail:

Evaluation constructs and scoring Each LV group is scored along four constructs operationalised in the implementation: *diversity*, *intervention priority*, *grid optimisation potential*, and *complementarity suitability*. The underlying metrics are computed from building-level attributes (type mix, roof and DER flags, energy labels), temporal descriptors (peak-hour dispersion), and simple network surrogates (distance to transformer, transformer utilisation). Formulas below reproduce the scoring logic:

(a) **Overall Diversity Index (0–10)** For building-type, temporal, generation, size, label, and occupancy diversity, denote the normalised sub-scores by $D_{\text{type}}, D_{\text{temp}}, D_{\text{gen}}, D_{\text{size}}, D_{\text{label}}, D_{\text{occ}} \in [0,1]$. The composite index is

$$DI \;=\; 10\big(0.25\,D_{\text{type}} + 0.20\,D_{\text{temp}} + 0.20\,D_{\text{gen}} + 0.10\,D_{\text{size}} + 0.15\,D_{\text{label}} + 0.10\,D_{\text{occ}}\big).$$

Here D_{type} and D_{label} are normalised entropies; D_{temp} derives from peak-hour dispersion; D_{gen} reflects prosumer share (maximal near 50%); D_{size} and D_{occ} utilise coefficients of variation.

(b) **Intervention priority (0–10)** Let R be retrofit potential (poor labels share), S solar potential (available roof share), H heat-pump suitability (share of A–C labels without heat pumps), B battery opportunity

(share of solar without battery), U the count of E/F/G buildings (capped via an urgency factor), and E_{eco} an economic viability factor. The score is

$$IP \;=\; 10\,E_{\text{eco}}\Big(0.30\,R + 0.25\,S + 0.20\,H + 0.15\,B + 0.10\,\min(U/10,\,1)\Big).$$

(c) **Grid optimisation potential (0–10)** With peak coincidence C (share of buildings peaking at the modal hour), transformer loading L (normalised to capacity), estimated line losses L_ℓ (distance surrogate), and voltage stability V (normalised),

$$GOP \;=\; 10\Big(0.40\,C + 0.20\,(1 - |0.7 - L|) + 0.20\,\min(L_\ell/0.1,\,1) + 0.20\,(1 - V)\Big).$$

Higher C and L_ℓ imply larger improvement headroom; moderate L is preferred; poor V indicates stabilisation potential.

(d) **Complementarity suitability (0–10)** Using DI (rescaled to $[0, 1]$), peak coincidence C, a loading factor $f_L = \min(L/0.3,\,1)$, and a proximity factor $f_P = 1 - \min(L_\ell/0.1,\,1)$,

$$CS \;=\; 10\Big(0.40\,\tfrac{DI}{10} + 0.30\,C + 0.20\,f_L + 0.10\,f_P\Big),$$

capturing that diverse, sufficiently loaded, and spatially compact groups are better candidates for complementarity-driven clustering.

Three selection criteria An LV group qualifies for the candidate set if it satisfies the following requirements; thresholds replicate the default configuration used in the implementation.

i. **Diversity and complementarity viability:** $DI \geq 5.0$ and $CS \geq 6.0$ (0–10 scales), ensuring heterogeneous demand profiles and viable temporal offsets within transformer boundaries.

ii. **Intervention leverage or urgency:** $IP \geq 6.0$ or $U > 5$, prioritising groups where coordinated retrofits, solar/battery deployment, or heat-pump rollout produce system-level benefits and address compliance risk.

iii. **Grid optimisation leverage and feasibility:** $GOP \geq 5.0$, with transformer loading in a practical window $0.3 \leq L \leq 0.85$ and line-loss surrogate $L_\ell \leq 0.10$, so that demand shaping, storage, or PV siting are both needed and feasible without unrealistic reinforcement assumptions.

Portfolio ranking and sample assembly All LV groups within the geographic scope are evaluated and ranked by a composite score

$$\text{Overall} \;=\; 0.30\,DI + 0.30\,IP + 0.20\,GOP + 0.20\,CS,$$

then filtered by the above criteria. The final research sample is the union of buildings from the top-ranked LV groups until the cumulative count reaches 50–200. This yields a tractable sample that (i) reflects realistic infrastructure boundaries, (ii) retains load/generation heterogeneity for complementarity analysis, and (iii) concentrates policy-relevant interventions. Ties are resolved in favour of groups with higher temporal diversity and larger proportions of candidate prosumers, to maximise the identifiability of complementarity effects.

Operational constraints Two safeguards are enforced: (1) a *transformer-boundary constraint* — only buildings served by the same MV/LV transformer are grouped for complementarity analysis; (2) a *scale constraint* — extremely small LV groups are excluded (default minimum three buildings per group), while excessively large conglomerations indicating data aggregation anomalies are flagged and not considered until boundary inconsistencies are resolved. These constraints ensure that cluster formation, evaluation, and subsequent modelling remain physically interpretable and directly actionable within distribution-grid practice.

4.3.1. Electrical infrastructure hierarchy

Recent studies provide quantitative benchmarks for typical LV network sizes in Europe and the Netherlands. Empirical evidence shows that LV feeders generally serve between 30 and 150 customers depending on urban density; Dutch LV networks often supply 100–200 households in urban settings; and common MV/LV transformer capacities range from 250 to 630 kVA, with an average connection capacity of 5–8 kVA per household. In Amsterdam and other Dutch urban areas, empirical values range between 50 and 200 buildings per transformer.

The network adheres to the actual Dutch grid structure, where energy communities are bounded by shared infrastructure layers, as shown in Figure 4.2 by a UML diagram:

- **HV (high voltage) network**: Transmission-level backbone.
- **Substations**: Interfaces connecting HV to MV.
- **MV (medium voltage) network**: Regional distribution through cable groups.
- **Transformers (MV/LV stations)**: Critical transformation points enabling energy sharing.
- **LV (low voltage) network**: Final distribution to end users.
- **Buildings**: End consumers connected to LV feeders.

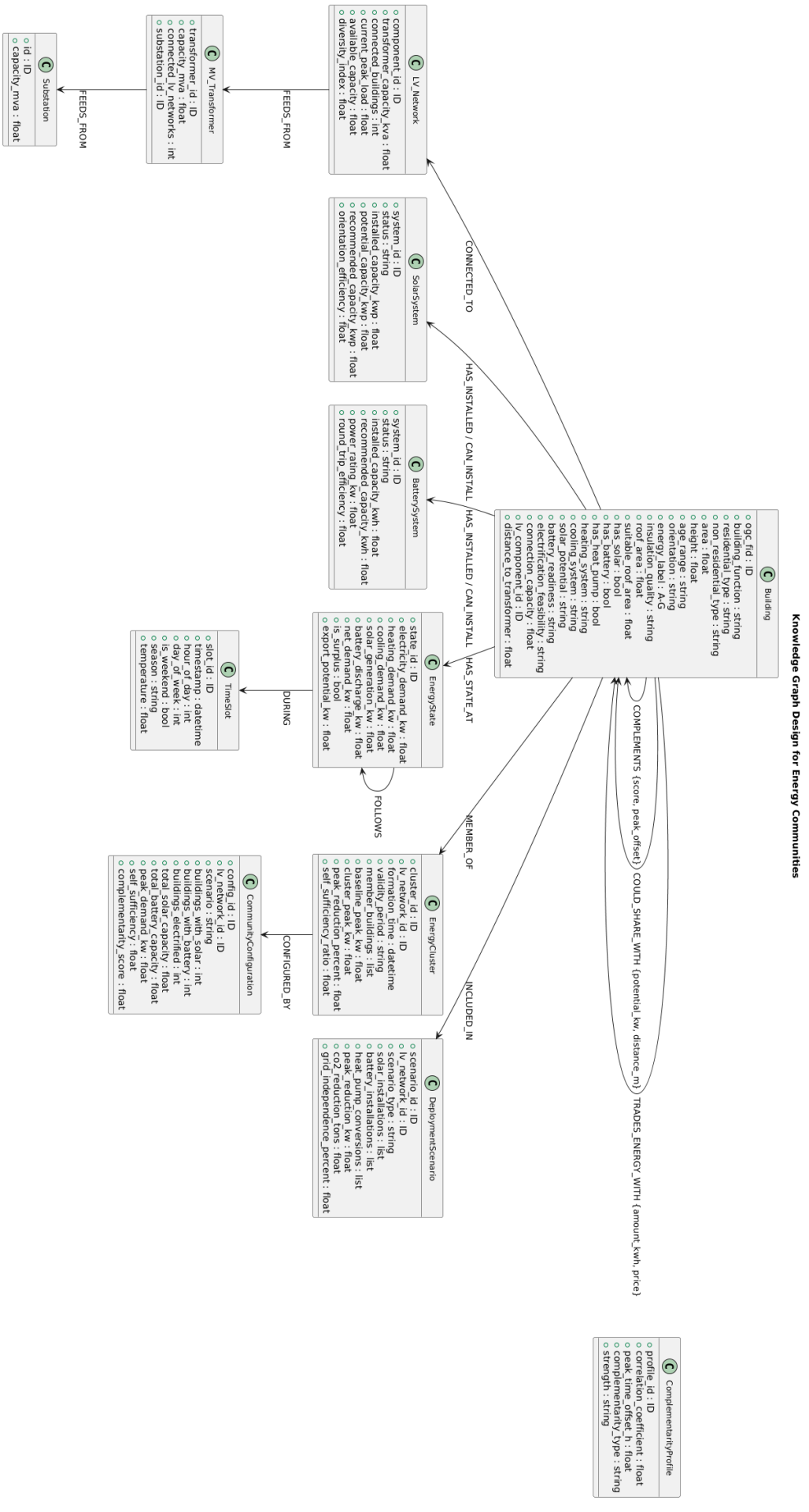


Figure 4.2: UML overview of dataset entities and relationships used for graph assembly (buildings, LV cable groups, MV transformers, substations, and adjacency clusters).

The operational domain of this study is the LV network. Communities are feasible only among buildings that share the same LV feeder (cable group). Consequently, the graph used for learning comprises two active layers: `Building` nodes and their parent `CableGroup`. Upstream components (transformers, substations) are excluded from the computational graph and appear only as integrity constraints in preprocessing (e.g., uniqueness checks). Accordingly, the geospatial data processing in QGIS is restricted to buildings, LV stations, and LV cables, while higher-level MV and HV structures are not considered in the network assembly or clustering validation.

Furthermore, Figures 4.3 and 4.4 show example attribute tables from QGIS for (i) a non-residential building and (ii) an LV group station, respectively. These tables demonstrate the available metadata used in network assembly and validation. The building records contain geometric, functional, connectivity-related fields (e.g., `connection_type`), while the LV group station records include voltage level, station type, connection type, calculated proximity metrics, and so on. Together, these attribute tables underpin the structured integration of entities into the LV network graph.

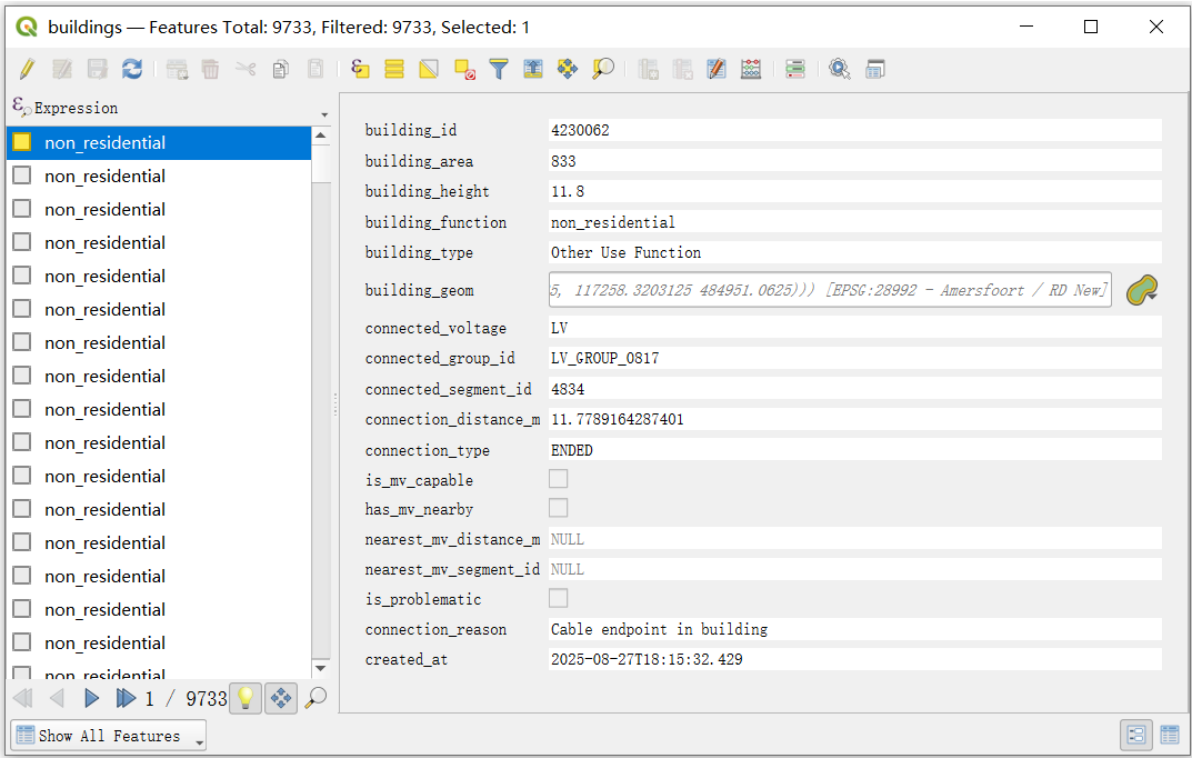


Figure 4.3: Example QGIS attribute table for a building entity (non-residential, showing connectivity parameters).

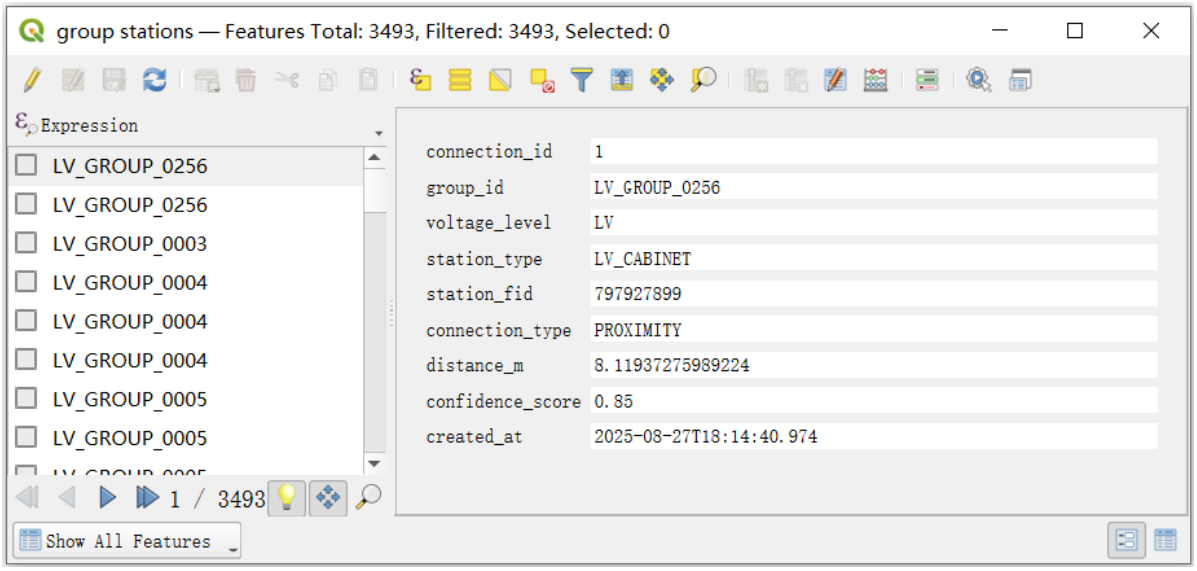


Figure 4.4: Example QGIS attribute table for an LV group station entity (showing voltage level, station type, and connection details).

4.3.2. Energy complementarity in distribution networks

A central concept is *energy complementarity*, defined as the temporal offset between generation-rich and demand-heavy nodes. Complementarity manifests when load peaks of one building coincide with production surpluses of another, thereby reducing aggregate peaks at the transformer level. Formally, complementarity can be quantified as the negative correlation of net load profiles:

$$C_{ij} = -\operatorname{corr}\big(L_i(t), L_j(t)\big).$$

The transformer boundary constitutes the primary domain for complementarity analysis: energy sharing is only feasible within buildings connected to the same LV transformer. The study is constrained to operational clusters defined by transformer-based LV feeders. Prior research by Holweger et al. (2023) [24] indicates that energy sharing within transformer domains is feasible in the low-voltage grid context, whereas cross-transformer sharing would entail significant reinforcement costs and infrastructure upgrades, rendering it generally impractical under current distribution grid configurations. Consequently, the methodological framework strictly enforces intra-transformer clustering and excludes cross-transformer grouping from network design and training.

In the implemented GNN, complementarity is learned using multi-head attention that reweights message passing based on temporal alignment and spatial proximity. This allows the model to capture dynamic spatio-temporal dependencies beyond simple pairwise correlations, a mechanism inspired by attention-based graph learning approaches proposed by Veličković et al. (2018) [59], Wu et al. (2020) [63], and Zhang et al. (2021) [68].

4.3.3. Energy community formation constraints

Beyond the operational transformer-boundary safeguard described earlier, two additional methodological constraints govern the feasible formation of energy communities. These constraints move beyond purely operational rules and capture the systemic principles that ensure both electrical consistency and planning relevance.

Infrastructure utilisation Distribution transformers act as the natural aggregation points of demand and supply. Load-balancing benefits emerge most effectively at these nodes, where the combined profiles of connected feeders can be coordinated. This constraint ensures that energy communities are not abstract clusters of buildings, but physically meaningful entities whose behaviour aligns with distribution-grid practice. By restricting clustering to the transformer level, the framework guarantees that community signals remain actionable within existing infrastructure boundaries and do not rely on hypothetical reinforcements.

Complementarity A community yields tangible benefits only if it combines heterogeneous consumption and generation patterns whose temporal offsets reduce net load variability. Such complementarity typically arises when distinct functional profiles interact, for example:

- Office vs. residential: day-time office loads complement evening household peaks.
- Retail vs. residential: weekday commercial activity contrasts with weekend household demand.
- School vs. residential: classroom hours are offset against evening residential consumption.
- Industrial vs. other: continuous base-loads provide a stabilising counterweight.

The methodological framework incorporates this principle by prioritising communities in which temporal diversity is sufficiently high to yield systemic peak reduction. Rather than relying solely on static indicators, complementarity is evaluated as a dynamic property of demand and generation time series, ensuring that community formation reflects genuine synergies in consumption and production rhythms.

Anti-collapse regularisation and community balance

To prevent degenerate partitions, the training objective incorporates soft regularisers on cluster size and dominance. Let $\mathcal{C}$ denote the set of discovered communities. A size-window regulariser keeps cluster cardinalities within a planning range $[K_{\min}, K_{\max}]$:

$$\Omega_{\text{size}} \;=\; \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} \Big(\max\{0,\, K_{\min} - |C|\} + \max\{0,\, |C| - K_{\max}\} \Big).$$

A dominance penalty discourages concentration in a single group:

$$\Omega_{\text{dom}} \;=\; \max\Big\{ 0,\, \frac{\max_{C \in \mathcal{C}} |C|}{\sum_{C \in \mathcal{C}} |C|} - \tau \Big\}, \quad \tau \in (0, 1).$$

Both terms act as soft preferences rather than hard constraints, maintaining interpretability without over-constraining discovery.

4.4. Method overview

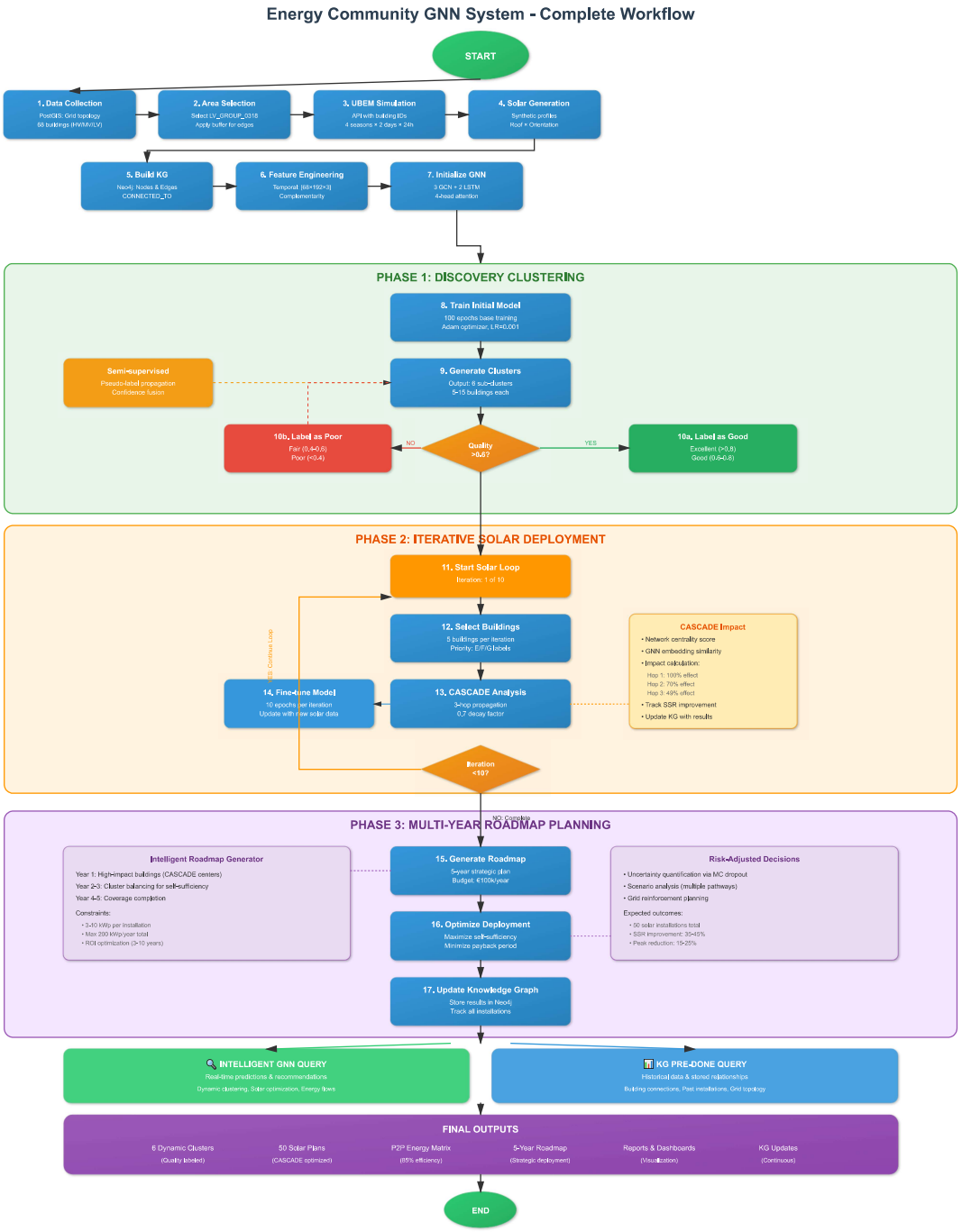


Figure 4.5: Methodological pipeline for the KG–GNN energy community framework.

The methodology develops an intelligent system that transforms raw geographical and energy data into optimized energy communities with strategic solar deployment plans. The process begins with comprehensive data collection from multiple sources to assemble a complete representation of the urban energy landscape. Grid topology data are extracted from PostGIS databases that capture the hierarchical structure of electrical distribution networks, spanning from high-voltage substations through medium-voltage transformers down to low-voltage cable groups connecting individual buildings. This spatial database enables precise querying of network relationships and identification of buildings that share electrical infrastructure, thereby forming the foundation for physically feasible energy communities.

Once a target area such as LV_GROUP_0318 is selected, building identifiers are transmitted to the Urban Building Energy Model (UBEM) through an API interface. The UBEM simulation produces detailed energy demand profiles that incorporate ideal load air systems for electricity, heating, and cooling requirements, with consumption patterns differentiated by energy labels. Buildings with lower energy ratings typically exhibit higher reliance on gas-based heating. The simulations generate 192 temporal data points that capture the full spectrum of consumption patterns across four seasons, distinguishing between weekday and weekend behaviors at hourly resolution throughout each 24-hour period. In parallel, synthetic solar generation profiles are derived from roof area, orientation, and shading factors to establish baseline renewable energy potential.

This multi-source data is processed through an automated pipeline that constructs a Neo4j knowledge graph, in which nodes represent buildings, transformers, and grid components, while edges encode physical connections and energy relationships. The knowledge graph functions both as a structured repository and as an active analytical component that can be queried and updated throughout the analysis. Following data preparation and feature engineering that capture temporal complementarity and network characteristics, a Graph Neural Network (GNN) architecture is initialized to identify critical relationships between spatial and temporal features.

The training process unfolds in three strategic phases. Phase One concentrates on discovery clustering, where the GNN partitions buildings into dynamic sub-clusters that respect both electrical constraints and energy optimization objectives. The model determines the optimal number of clusters, typically converging on groups of five to fifteen buildings, while maximizing self-sufficiency and complementarity among cluster members. A semi-supervised evaluation mechanism continuously assesses cluster quality, assigning performance labels ranging from excellent for highly self-sufficient and complementary groups to poor for clusters exhibiting significant imbalances or low renewable potential. These quality assessments feed back into the training process, directing the model toward improved clustering solutions.

Phase Two implements iterative solar deployment, in which the trained model enters a strategic loop of solar panel installations and parameter refinement. Across ten iterations, five buildings are selected in each round based on multiple criteria including poor energy labels (indicating high improvement potential), network centrality (maximizing cascade effects), and cluster balance considerations. A cascade tracking mechanism analyzes how each solar installation propagates benefits through the network, quantifying impact decay across three network hops to evaluate neighborhood-level effects. After each deployment round, the model incorporates updated generation data to refine parameters, allowing clusters to dynamically reorganize as the energy landscape evolves.

Phase Three synthesizes the learned patterns into a comprehensive multi-year roadmap that delivers actionable deployment strategies. The roadmap generator creates a five-year plan that sequences solar installations to maximize cumulative benefits while adhering to annual budget constraints of one hundred thousand euros and technical limits on installation capacity. The planning algorithm prioritizes high-impact buildings in the early years to establish energy generation hubs, emphasizes cluster balancing in intermediate years, and extends coverage in later years to achieve target penetration levels.

Throughout execution, the system tracks detailed metrics including hourly inter-building energy flows, cluster evolution dynamics, improvements in self-sufficiency, and reductions in peak demand. The knowledge graph is continuously updated with new installations and measured impacts, enabling both real-time predictions via the GNN model and retrospective analysis through graph queries. The final outputs include dynamic cluster assignments with quality labels, prioritized solar installation schedules

optimized for cascade effects, peer-to-peer energy sharing matrices, strategic deployment roadmaps, and a suite of visualizations such as Sankey diagrams for energy flows, heatmaps for temporal demand patterns, and interactive dashboards for stakeholder engagement. This integrated approach demonstrates how graph-based deep learning can orchestrate the transformation of passive building groups into adaptive, self-organizing energy communities that balance individual building needs with collective grid benefits.

4.5. Phase 1: Knowledge graph construction

Rationale and structure. This section first details the end-to-end transformation from relational geospatial data (PostgreSQL/PostGIS) into a Neo4j knowledge graph (KG), including topology reconstruction, hierarchy assignment, and quality assurance. The subsequent subsection formalises the resulting ontology and schema, reporting node/edge types, semantics, and constraints. This order foregrounds provenance and physical grounding before presenting the final schema.

It is important to emphasise that the electrical topology of the distribution network is *explicitly encoded* in the KG itself, not learned implicitly by the GNN. Nodes represent buildings, cable groups, and transformers, while edges capture hierarchical and adjacency relations (e.g. building–feeder membership, feeder–transformer connections). Each edge further carries physical attributes such as distance, indicative capacity, and impedance. The GNN therefore operates on this fixed topology: it learns how to propagate and weight information along the given KG structure through message passing and attention, but it does not infer or reconstruct the network topology on its own. This clear division ensures that the KG provides the structural backbone, while the GNN focuses on learning operationally relevant patterns within that structure.

4.5.1. Relational-to-KG Transformation: From SQL/PostGIS to Neo4j

The methodological pipeline converts relational geospatial data from PostgreSQL/PostGIS into a Neo4j KG, aligning tabular infrastructure datasets with graph-based learning while preserving operational hierarchy. Electrical continuity is reconstructed by segmenting raw LV/MV/HV cables into topologically consistent polylines and aggregating them into electrically continuous `CableGroup`s, thereby removing artefacts caused by digitisation gaps. Each group is linked to its serving LV cabinets, MV/LV transformers, and HV substations via proximity and topology rules augmented with graded confidence metrics.

Buildings are then associated with a unique LV group. Connection types are classified (*terminated*, *entered*, *crossed*, *proximity-based*); distances and diagnostics are recorded as properties. Flags identify MV-capable non-residential buildings and overly long connections, providing early reinforcement indicators. By establishing a single LV-group and transformer ancestry per building, transformer-bounded feasibility for energy sharing is enforced.

Once relationships are established, hierarchical summaries are generated along *Building* $\rightarrow$ *CableGroup* $\rightarrow$ *Transformer* $\rightarrow$ *Substation*. These roll-ups capture structural indicators (counts, connection quality, distance statistics) and functional heterogeneity (building functions, energy labels, temporal diversity). The Neo4j materialisation stage creates nodes and typed relationships encoding electrical connectivity, spatial adjacency, and temporal membership; properties integrate geometry, physics, and energy descriptors, while relationships carry connection semantics and diagnostic scores. Quality assurance verifies invariants such as “every building connects to exactly one LV group” and “no LV group lacks transformer ancestry.”

The outcome is a KG that faithfully reproduces the physical distribution hierarchy, preserves quality-controlled mappings between buildings and grid components, and enriches entities with engineered features relevant to downstream spatiotemporal GNN tasks. Figure 4.6 synthesises the resulting schema and clarifies how the pipeline’s outputs are instantiated in Neo4j: (i) a `Building`-centred core with identifiers, siting and connection attributes; (ii) an asset layer where `HAS_INSTALLED` relate buildings to `SolarSystem` nodes (storing capacities, efficiencies, and retrofit priorities); (iii) a temporal layer linking `EnergyState` to `TimeSlot` via `DURING`, capturing demand, generation, net position, and battery state-of-charge; (iv) a spatial layer combining pairwise `ADJACENT_TO` relations (annotated with shared-length and complementarity) with `IN_ADJACENCY_CLUSTER` membership; and (v) an electrical-continuity layer in which `HAS_CONNECTION_POINT` and `ON_SEGMENT` anchor service drops to `CableSegment`s that

are PART_OF a CableGroup, while CONNECTED_TO maps each building to its LV feeder and upstream edges CONNECTS_TO/FEEDS_FROM bind feeders to transformers and substations. This integrated structure makes explicit the feasible domains for energy exchange and provides physically grounded inputs for subsequent learning.

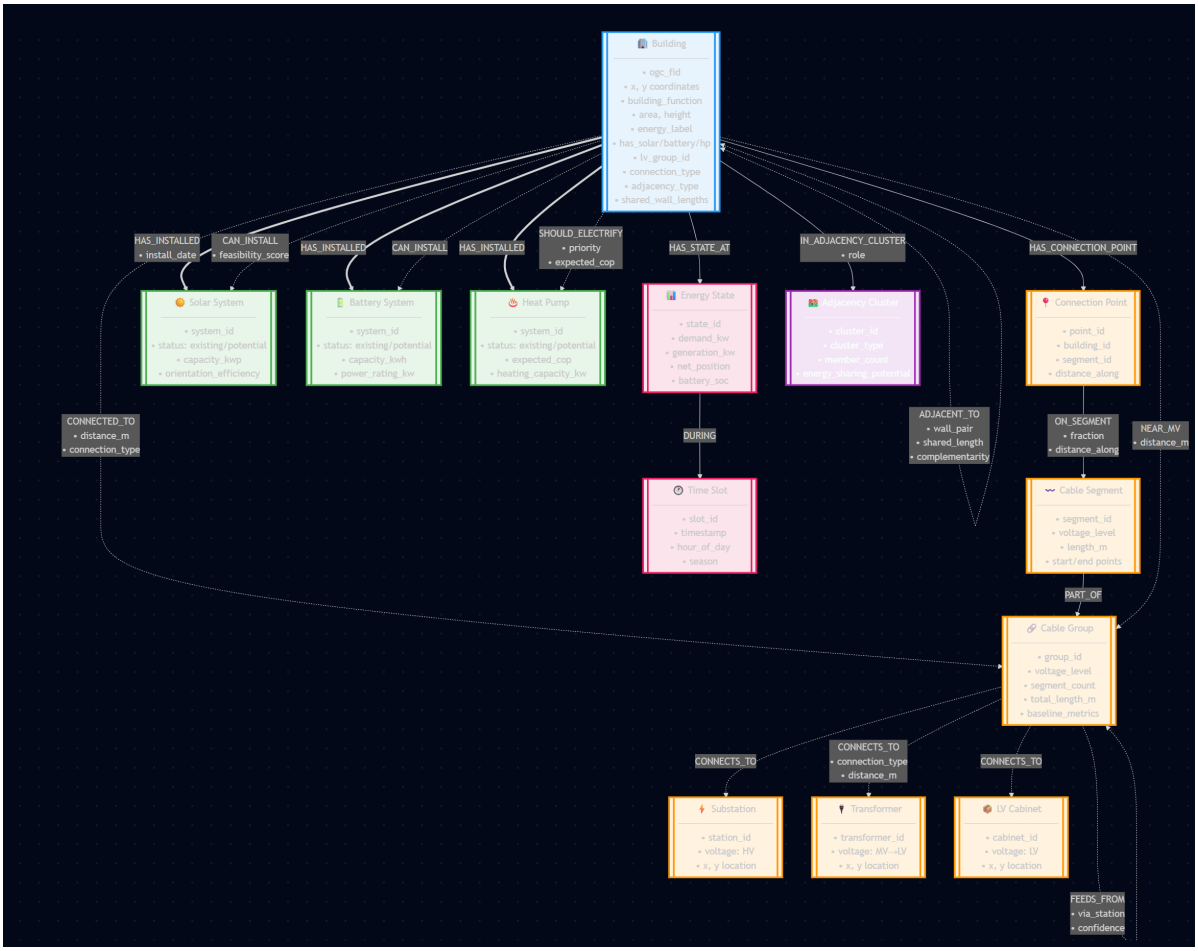


Figure 4.6: Instantiated UML-style view of the KG schema highlighting asset, temporal, spatial, and electrical-continuity layers around a building-centred core.

4.5.2. Knowledge graph ontology framework

This subsection formalises the ontology materialised by the preceding pipeline. The specification covers (i) entity classes, (ii) relation families with explicit direction and cardinality, (iii) the attachment of attributes to the physical loci that generate or consume them, and (iv) governance rules that guarantee electrical feasibility and analytical consistency. The layered schema in Figure 4.6 provides the conceptual backdrop; Tables 4.2–4.3 enumerate the concrete labels and relationships realised in Neo4j, and Figure 4.7 confirms their instantiation in the database.

The design objective of ontology is building-centred yet infrastructure-aware. It represents: (a) *physical infrastructure* to preserve the distribution hierarchy; (b) *spatial structure* to delimit plausible sharing neighbourhoods; (c) *temporal states* for demand–generation dynamics; and (d) *asset attachments* for retrofit and flexibility analyses. Attributes are anchored at their natural entities (e.g., load time series at buildings, loading limits at transformers) so that subsequent learning remains constrained by physics and operations.

Core entity classes comprise Building, CableGroup, Transformer, and Substation; these sustain the electrical hierarchy listed in Table 4.3. Building-level attributes used for downstream learning are illustrated in Figure 4.11. Feeder-level continuity and geometric metrics are attached to CableGroup nodes

(Figure 4.12); DER assets (`SolarSystem`, `BatterySystem`, `HeatPumpSystem`) appear as separate nodes with technology-specific descriptors (Figure 4.13). Auxiliary operational entities such as `LV Cabinet` (Figure 4.14) support continuity reconstruction and validation. Relative property lists for all classes are provided in appendixA.

Four relation families structure the KG (Table 4.3): (i) *electrical connectivity* binds buildings to feeders and feeders to upstream equipment. The `CONNECTED_TO` edges enforce a unique LV group per building (Figure 4.9), while `CONNECTS_TO` and `FEEDS_FROM` maintain the upstream path to transformer and substation; (ii) *spatial cohesion* encodes neighbourhood constraints through symmetric `ADJACENT_TO` links (Figure 4.8) and `IN_ADJACENCY_CLUSTER` membership; (iii) *temporal alignment* maps `EnergyState` to `TimeSlot` via `DURING`; (iv) *asset management* associates buildings with installed equipment via `HAS_INSTALLED` (Figure 4.10) and supports advisory relations (details in appendix A).

Schema-level constraints require: one and only one `CONNECTED_TO` per building; a transformer ancestor for every LV group; symmetry of `ADJACENT_TO`; and a single `DURING` target per `EnergyState`. Indices on primary identifiers and foreign-key properties (e.g., *group_id*, *transformer_id*) ensure deterministic joins and efficient traversal.

Table 4.2: Current Schema Overview — Node types in the Neo4j KG. The list reports semantics and indicative key properties (non-exhaustive).

Node	Semantics and modelling role	Key properties (examples)
`Building`	Core prosumer entity and centre of analysis; host of static/dynamic energy attributes and DER attachments.	*id*, *area*, *building_function*, *age_range*, *energy_label*, *solar_potential*, *net_load_ts_ref*
`CableGroup`	Electrically continuous low-voltage cable groups (replacement of legacy `lv_group`); represent feeder-level connectivity.	*group_id*, *segment_count*, *total_length_m*, *bbox_wkt*
`Transformer`	MV/LV transformers; aggregation and constraint locus for loading.	*transformer_id*, *rated_power_kVA*, *geom_wkt*
`Substation`	HV substation feeding upstream of transformers; defines higher-level supply territories.	*substation_id*, *geom_wkt*
`AdjacencyCluster`	Spatial cluster of buildings (e.g., contiguous blocks) indicating local sharing/replaceability potential.	*cluster_id*, *size*, *diameter_m*
`EnergyState`	Time-resolved net consumption/generation state of an entity; basis for temporal learning and alignment.	*node_id_ref*, *p_net_kW*, *q_net_kvar*
`TimeSlot`	Time dimension node; regularises the temporal axis and supports roll-ups.	*slot_id*, *timestamp*, *season*, *hour_of_day*
`SolarSystem`	PV system attached to a building; supports degradation and orientation effects.	*system_id*, *installed_capacity_kWp*, *orientation_efficiency*, *degradation_factor*, *installation_year*

Table 4.3: Current Schema Overview — Edge types in the Neo4j KG, with direction, semantics, and indicative cardinalities.

Edge	Domain	Range	Semantics	Cardinality (typ.)
CONNECTED_TO	Building	CableGroup	Electrical service connection from a building to its LV group.	1:1 per building
CONNECTS_TO	CableGroup	Transformer	Upstream connectivity from an LV group to its supplying transformer.	N:1
FEEDS_FROM	Transformer	Substation	Upstream supply relation from transformer to substation.	N:1
IN_ADJACENCY_CLUSTER	Building	AdjacencyCluster	Spatial membership used for local sharing constraints.	N:1
ADJACENT_TO	Building	Building	Symmetric spatial adjacency between buildings.	many-to-many
HAS_INSTALLED	Building	SolarSystem	Association to installed DER assets.	$0..N$
DURING	EnergyState	TimeSlot	Temporal alignment of a state with a discrete time slot.	exactly 1:1

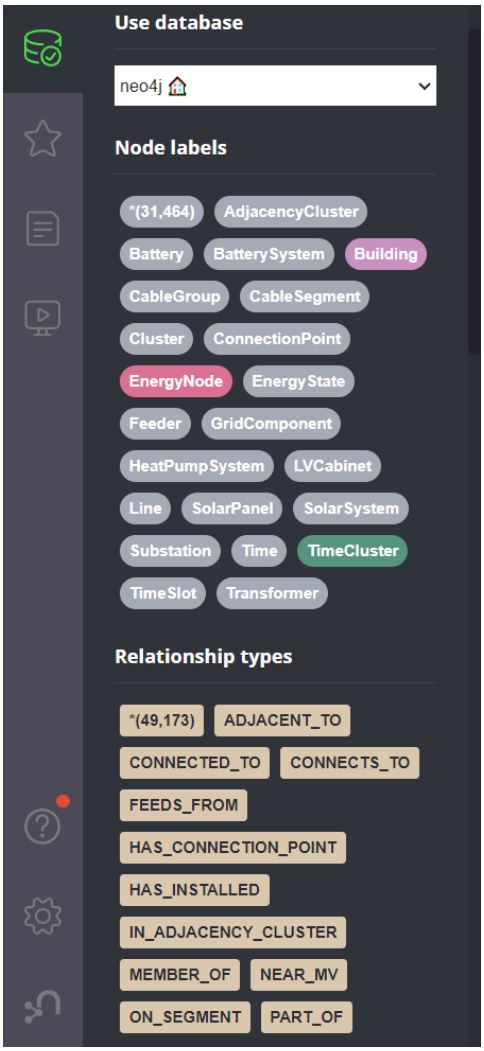


Figure 4.7: Database sidebar confirming the instantiated labels and relationship types.

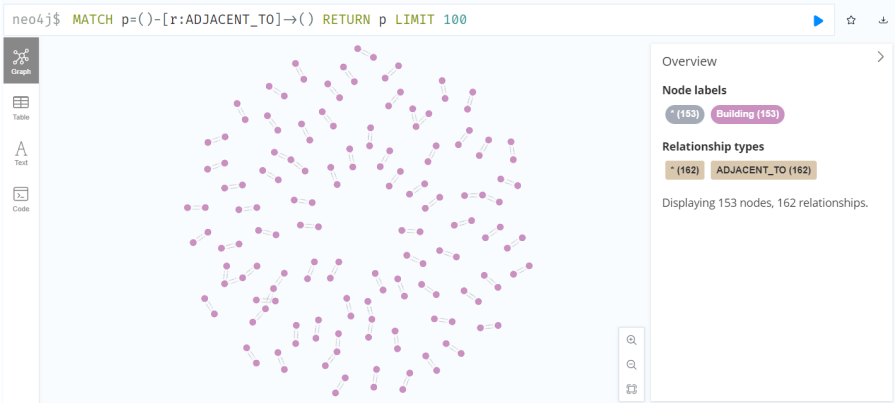


Figure 4.8: `ADJACENT_TO` edges (spatial neighbourhood).

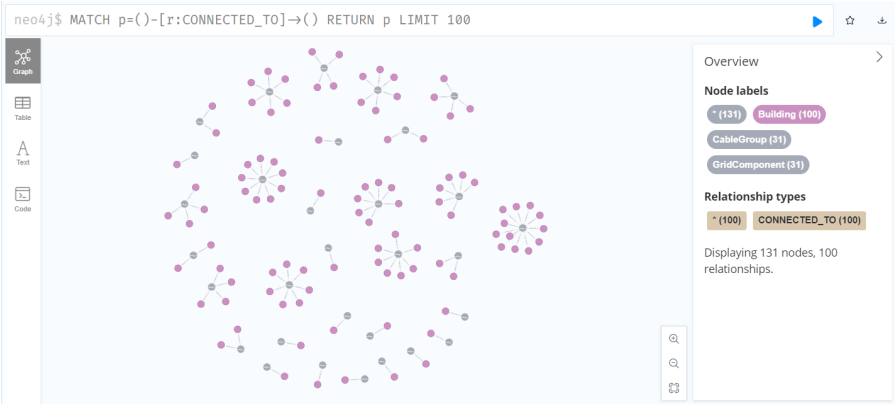


Figure 4.9: `CONNECTED_TO` edges (building → LV cable group).

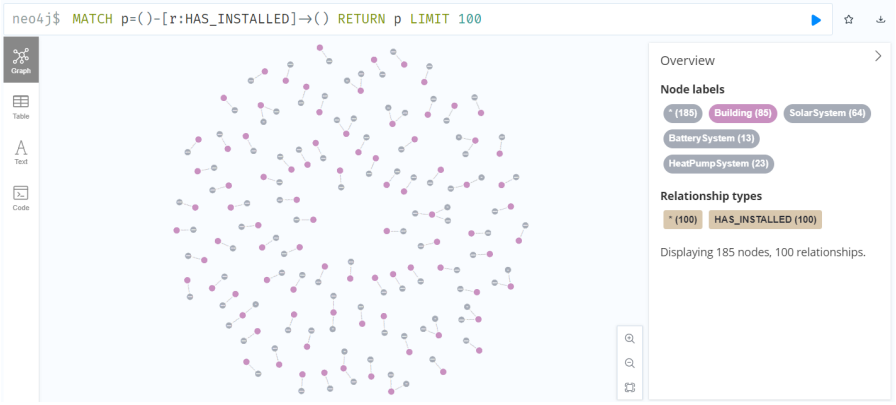


Figure 4.10: `HAS_INSTALLED` edges (building → DER assets).

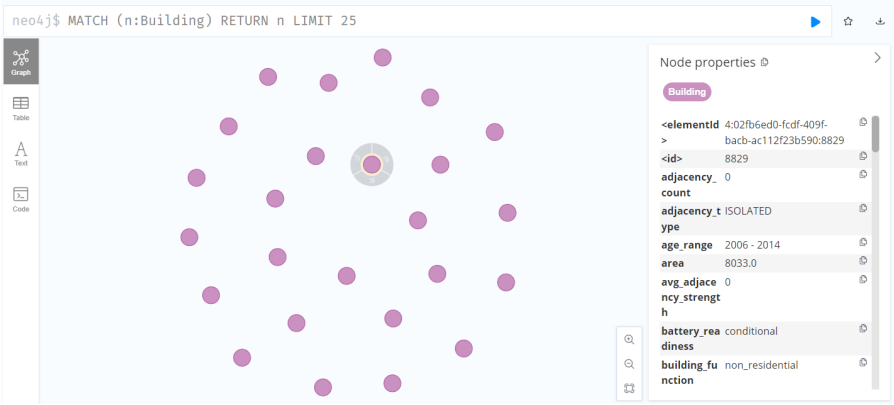


Figure 4.11: `Building` node properties used by downstream models.

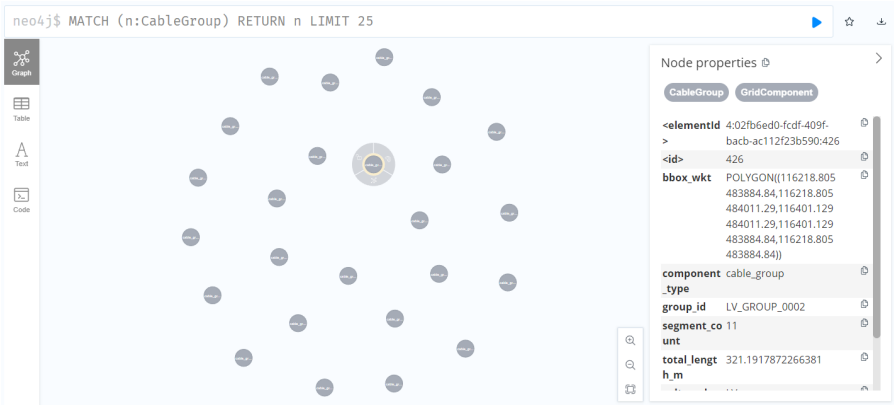


Figure 4.12: `CableGroup` node properties (continuity and lengths).

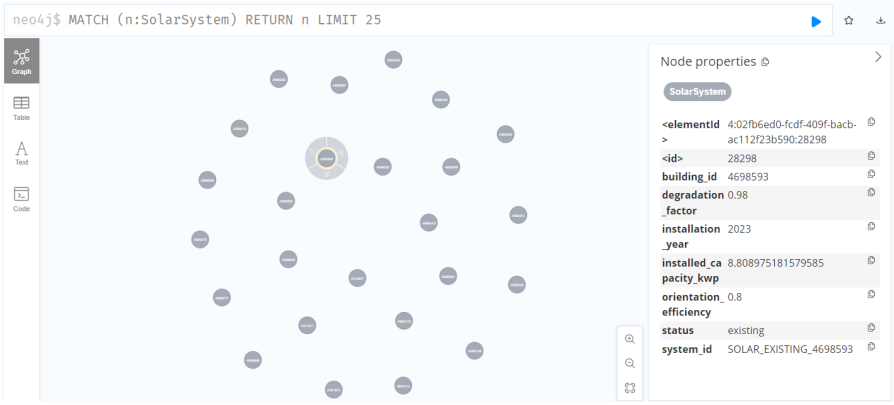


Figure 4.13: `SolarSystem` node properties (capacity, orientation, degradation).

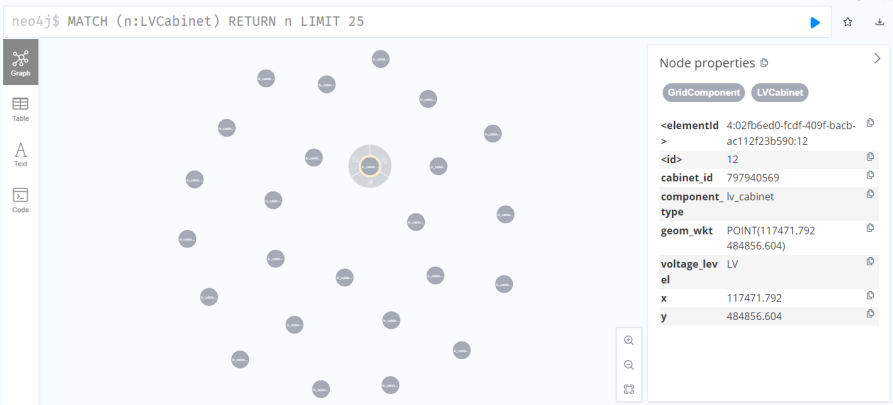


Figure 4.14: `LV Cabinet` example supporting continuity reconstruction.

4.6. Phase 2: Infrastructure-aware preprocessing

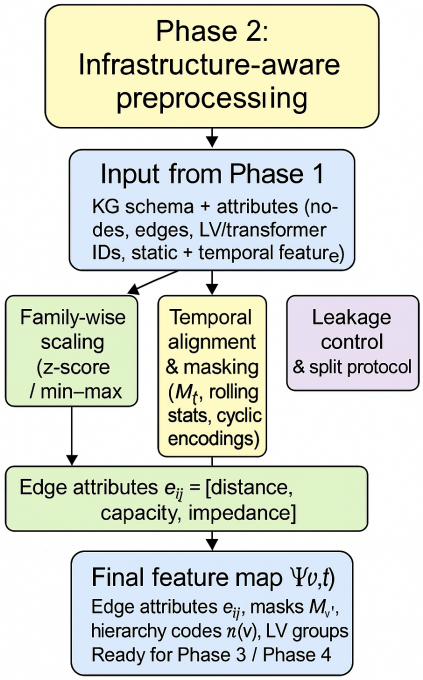


Figure 4.15: Phase 2 pipeline. The KG-derived attributes are normalised family-wise and temporally masked; rolling statistics and calendar encodings are added; edge and hierarchy features are constructed; leakage control ensures no topological violation. Outputs are *node-time maps* $\Psi(v,t)$ and *edge attributes* e_{ij}, ready for the model stack.

This phase bridges the KG from Phase 1 with the learning modules. Its purpose is to transform heterogeneous, infrastructure-indexed attributes into model-ready tensors while preserving the physical topology and governance rules. Three design requirements are followed: (i) harmonisation across node families and time, (ii) explicit treatment of temporal structure and missingness, and (iii) preservation of electrical invariants such as identical LV feeder boundaries— a condition also emphasised as critical for distribution-grid operation by Prettico et al. (2019) [47], Netbeheer Nederland (2021) [40], and ACM (2016) [2].

Beyond these technical requirements, the **adaptivity of the knowledge graph** also plays an essential role. Unlike relational databases that require rigid table structures, the KG allows new attributes or entities (e.g., additional building descriptors or renewable devices) to be added without redesigning the

entire schema. This flexibility ensures two advantages: first, the system can be continuously updated as new buildings, technologies, or regulatory data are introduced; second, the embedded electrical rules (such as the requirement that buildings must remain within the same LV feeder) are preserved even under dynamic updates. In this way, the tensorisation process remains stable and reliable as the energy system evolves, rather than being limited to a static dataset, as discussed by Liu et al. (2023) [34] and Hogan et al. (2021) [23].

Phase 1 outputs a KG where buildings are linked to same LV feeders. Phase 2 consumes this graph and prepares tensors for:

- hierarchical message passing, as in Ying et al. (2018) who showed how structural hierarchy improves pooling [67],
- constrained pooling that respects LV/transformer boundaries following power system practices (Prettico et al, 2019) [47], and
- temporal **encoding** over masked sequences with calendar features, consistent with spatio-temporal GNN practice (Wu et al., (2020) [63].

4.6.1. Notation and objectives

Let $\mathcal{V}$ denote the complete set of nodes, divided into families $\{\mathcal{V}^{(\mathrm{b})}, \mathcal{V}^{(\mathrm{g})}, \mathcal{V}^{(\mathrm{t})}, \ldots\}$ (buildings, cable groups, transformers, etc.). For each node $v \in \mathcal{V}$, static attributes are represented as $x_v \in \mathbb{R}^{F_s}$, while temporal attributes are given as sequences $X_v = \{x_{v,t} \in \mathbb{R}^{F_d}\}_{t\in\mathcal{T}}$ over a horizon $\mathcal{T} = \{t_0 - H + 1, \ldots, t_0\}$.

The preprocessing pipeline follows standard data engineering principles discussed by Vassiliadis et al. (2002) [58]. Its main goals are: 1. to normalise features separately for each node family, 2. to align time series across nodes and handle missing data explicitly, 3. to construct edge and positional features needed for spatial and temporal **encoding**, and 4. to avoid information leakage across time or topology.

Family-wise scaling and categorical handling Wu et al. (2020) [63] showed that z-score normalisation is the standard choice for continuous features in graph learning. Accordingly, static continuous variables are standardised per family:

$$\widetilde{x}_v^{(j)} = \frac{x_v^{(j)} - \mu_j^{(c)}}{\sigma_j^{(c)} + \varepsilon}, \quad v \in \mathcal{V}^{(c)},\ j \in \mathcal{I}_{\mathrm{cont}}. \tag{4.1}$$

Here, $(\mu_j^{(c)}, \sigma_j^{(c)})$ are computed exclusively from the training data of family c to prevent data leakage, as recommended by Vassiliadis et al. (2002) [58].

Variables representing magnitudes (e.g., peak loads) are normalised to the $[0, 1]$ range using clipped min–max scaling, following the procedure of Wu et al. (2020) [63]. Ordinal variables (e.g., energy labels) are encoded as integers to preserve their inherent ranking, as suggested by Aniakor et al. (2024) [3], whereas binary indicators (e.g., the presence of distributed energy resources) are retained as Boolean $\{0, 1\}$ values.

Temporal alignment, masking, and stabilisation Nijhuis et al. (2017) [42] stressed that realistic residential load models must deal explicitly with irregular samples and missing entries. Following this idea, all nodes are aligned to a common time horizon ending at t_0, and missing data are marked with a binary mask $M_{v,t} \in \{0, 1\}$.

To stabilise optimisation, rolling averages and standard deviations are computed only over observed samples:

$$\mathrm{mean}_t^{(W,f)} = \frac{\sum_{h=0}^{W-1} M_{v,t-h}\, x_{v,t-h}^{(f)}}{\sum_{h=0}^{W-1} M_{v,t-h}}, \tag{4.2}$$

$$\mathrm{std}_t^{(W,f)} = \left(\frac{\sum_{h=0}^{W-1} M_{v,t-h} \big(x_{v,t-h}^{(f)} - \mathrm{mean}_t^{(W,f)}\big)^2}{\sum_{h=0}^{W-1} M_{v,t-h}} \right)^{1/2}. \tag{4.3}$$

Calendar time is encoded with sinusoidal functions to reflect daily and weekly periodicity smoothly. This avoids *boundary artefacts*: when time is treated as discrete integers, transitions such as 23:00 to 00:00 or Sunday to Monday appear as large jumps even though they are adjacent in reality. By mapping time onto a circle with sine and cosine functions, periodic continuity is preserved. This technique was first used in the transformer model and is now common in spatio-temporal GNNs for energy forecasting.

Temporal data integration and application The temporal dimension is crucial for capturing energy complementarity patterns. Each building's hourly **net-load** profile $L_{i,t}$ (optionally accompanied by on-site generation signals) is processed through a multi-stage pipeline that transforms raw time series into actionable inputs for community formation:

1. **Pattern extraction**: Raw consumption data are compressed into representative **low-dimensional** pattern features using rolling statistics and calendar encodings, reducing the hourly profile to compact descriptors that capture essential behaviours.
2. **Temporal masking**: Missing data points are explicitly masked rather than interpolated to preserve data integrity.
3. **Complementarity computation**: Pairwise temporal correlations ρ_{ij} are computed to identify buildings with offsetting demand patterns.
4. **Peak-hour profiling**: Peak-time tendencies are profiled to favour communities with distributed rather than coincident peaks.

Edge attributes for graph propagation Following standard distribution-network practices proposed by Baran and Wu (1989) [5], each edge e_{ij} is annotated with electrical distance, impedance, and capacity, scaled using Equation (4.1). These attributes are provided to the model as auxiliary features for message passing and neighbourhood weighting in later phases.

Hierarchical positional encodings Learning hierarchical structure is a well-established concept. Ying et al. (2018) [67] introduced hierarchical pooling to make latent graph structure explicit. In the present study, each node is assigned a level embedding $\pi(v) \in \{\texttt{building}, \texttt{cable\_group}, \texttt{transformer}\}$, which is projected into the hidden space and added to its features. Here, `building` and `cable_group` nodes participate in message passing and clustering, while `transformer` nodes act as upstream anchors preserving the feeder–transformer hierarchy.

Missing values and outliers Schwefel et al. (2018) [53] observed that inadequate handling of missing values can distort uncertainty estimates in distribution grids. Accordingly, static missing values are filled with family-specific medians and flagged with binary indicators, while temporal gaps remain masked. Outliers are winsorised within each family before scaling, following load-modelling practices outlined by Nijhuis et al. (2017) [42].

Leakage control To prevent either temporal or topological leakage, scaling parameters $\{(\mu, \sigma), (a, b)\}$ are always fitted only on the training split. Splits are made along transformer boundaries so that no building from the same LV feeder appears in different folds, as recommended in grid-aware validation studies.

Final feature map and invariants The resulting node–time feature map concatenates static attributes, dynamic series, and calendar encodings (as defined above). Edge attributes e_{ij} are stored alongside adjacency, and grid invariants (such as LV group membership and transformer ancestry) are preserved. This aligns with invariants in CIRED's active distribution planning report (2014) [15], ensuring that downstream GNN training remains physically valid.

4.6.2. Motivation and continuity

Phase 1 ensured that electrical and spatial constraints are explicitly represented in the KG. Phase 2 then converts these semantics into model-ready tensors:

- time-aware encoders require masked and position-augmented sequences,

- spatial encoders require calibrated edge attributes,
- hierarchical pooling must respect LV/transformer boundaries.

This design guarantees that later phases operate on features that are consistent with both ontology and grid operations, echoing the emphasis on semantic consistency and interoperability by Hogan et al. (2021) [23] and Pritoni et al. (2021) [48].

It is important to stress that such alignment would be difficult to achieve within a conventional relational database (DBMS). While SQL/PostGIS can store the raw attributes, relational tables lack the semantic layer needed for reasoning across heterogeneous entities and evolving hierarchies. Every cross-level query in a DBMS requires costly joins and rigid schema definitions, making it impractical to enforce dynamic constraints such as “all buildings remain within the same LV feeder” or to propagate temporal patterns into clustering tasks. By contrast, knowledge graphs offer *schema flexibility*, *semantic reasoning*, and *direct relationship traversal*. This ability to unify heterogeneous relational exports (buildings, feeders, transformers, time-series states) into a semantically coherent graph structure explains why the KG is not merely a storage alternative, but an essential foundation for the GNN integration in later phases.

4.7. Phase 3: Infrastructure-constrained feature engineering

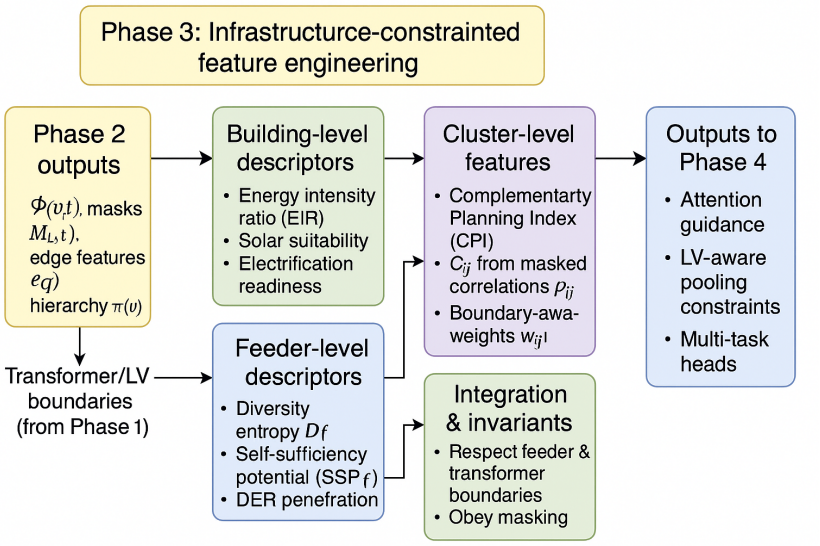


Figure 4.16: Phase 3 pipeline. Inputs from Phase 2 (node–time map Ψ, masks $M_{v,t}$, edge attributes e_{ij}, hierarchy code $\pi(v)$) and Phase 1 (transformer/LV boundaries) are aggregated into multi-level descriptors. Temporal embeddings and pairwise complementarity features highlight sharing opportunities, while LV boundaries remain enforced.

Phase 3 constructs higher-level descriptors that remain constrained by the grid. The aim is to expose interactions—such as temporal diversity and demand–generation complementarity— that raw attributes alone cannot reveal. All aggregations respect LV boundaries to preserve physical feasibility (Prettico et al., 2019 [47]; Netbeheer Nederland, 2021 [40]; ACM, 2016 [2]).

From Phase 1, the ontology buildings $\rightarrow$ cable groups $\rightarrow$ transformers ensures feeder ancestry and transformer domains. From Phase 2, leakage-safe, normalised features are inherited. Phase 3 then builds descriptors at three levels: the *building* level, the *feeder* (cable group) level, and an emergent *community* level. The community level is not predefined in the grid; it emerges via LV-constrained pooling that groups buildings within the same feeder according to temporal complementarity.

4.7.1. Notation and intuitive explanation

The system is modelled as a graph with two node types: buildings $\mathcal{V}^{(\mathrm{b})}$ and LV feeders $\mathcal{V}^{(\mathrm{g})}$. Each building $i \in \mathcal{V}^{(\mathrm{b})}$ has an hourly net-load series $L_{i,t}$ (positive when net-importing after PV), aligned and

masked as in Phase 2. For a feeder $f \in \mathcal{V}^{(g)}$, G_f denotes the set of connected buildings.

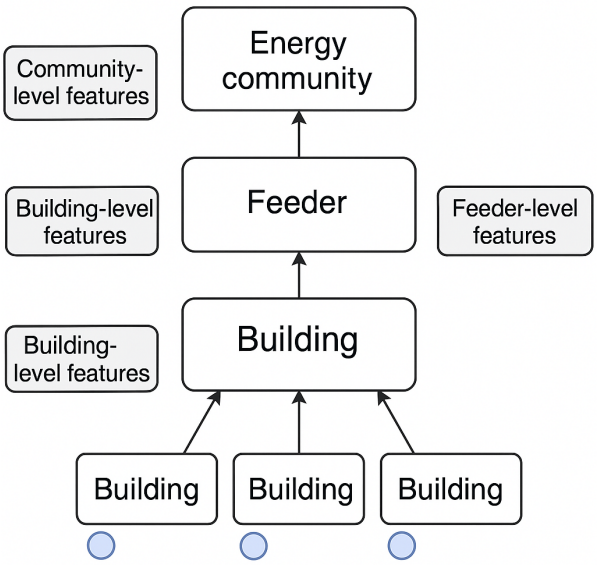


Figure 4.17: Hierarchy of the proposed pipeline: buildings aggregate into feeders, and LV-constrained pooling yields emergent communities. Descriptors are computed at building, feeder, and community levels and used by message passing, attention, and clustering in Phase 4.

Building-level features For clustering, building nodes include static descriptors (e.g., floor area, function/type, energy-intensity proxies) and dynamic descriptors derived from hourly net-load. Let

$$\mathrm{EIR}_i = \frac{\bar{E}_{i,\mathrm{elec}}}{A_i},$$

denote an electric energy-intensity proxy (per unit area). These features are normalised and used as inputs to temporal projection (Eq. (4.6)) and attention in Phase 4.

Feeder-level features Feeders aggregate building attributes within the same LV domain. Examples include entropy-based diversity of demand and aggregate net-load statistics. These capture variability and balance at feeder scale and are used to condition pooling and attention under LV masks.

Compactness measure for maps In addition to $\mathrm{Compct}(C)$, the map caption reports the normalised mean pairwise distance $\overline{d}(C) = \frac{2}{|C|(|C|-1)} \sum_{i<j \in C} \frac{d(i,j)}{D_{\max}}$, to corroborate visual coherence numerically.

Community-level descriptors During LV-constrained pooling, communities C emerge as groups of buildings within the same feeder exhibiting high pairwise complementarity. For buildings i and j with standardised net-load series $\tilde{L}_{i,t}$ and $\tilde{L}_{j,t}$,

$$\rho_{ij} = \frac{\sum_t \tilde{L}_{i,t}\,\tilde{L}_{j,t}}{\sqrt{\sum_t \tilde{L}_{i,t}^2}\,\sqrt{\sum_t \tilde{L}_{j,t}^2}}, \qquad C_{ij} = \frac{1-\rho_{ij}}{2} \in [0,1].$$

Community-level complementarity statistics are computed by aggregating $\{C_{ij}\}$ over $i, j \in C$ (within-feeder only). These aggregated statistics are used internally by the clustering objective (see Phase 4) rather than reported as standalone evaluation indices.

Intra-cluster diversity Diversity reflects heterogeneity of temporal patterns inside a cluster. Let $\mathbf{h}_i^{\mathrm{temp}} \in \mathbb{R}^d$ be the temporal embedding of building i (Section 4.8.3). For cluster C, define the nor-

malised dispersion

$$\mathrm{Div}(C) = 1 - \frac{1}{|C|\, r_d} \sum_{i \in C} \left\| \mathbf{h}_i^{\text{temp}} - \bar{\mathbf{h}}_C^{\text{temp}} \right\|_2, \quad \bar{\mathbf{h}}_C^{\text{temp}} = \frac{1}{|C|} \sum_{i \in C} \mathbf{h}_i^{\text{temp}},$$

where r_d is the d-dimensional radius (95th percentile of within-feeder distances) for normalisation to $[0, 1]$.

4.7.2. Temporal embeddings and complementarity

Dynamic behaviour is encoded using temporal sequences that capture hourly and seasonal patterns. The temporal processing transforms raw hourly data into compact embeddings that guide community formation:

$$\mathbf{z}_i^{\text{temp}} = \frac{1}{T} \sum_{t=1}^{T} f_\theta(L_{i,t}, \text{calendar}_t), \tag{4.4}$$

where $f_\theta(\cdot)$ is a learnable transformation of hourly net-load and calendar features. A gated recurrent unit (GRU; Cho et al., 2014 [14]) produces fixed-length embeddings capturing diurnal and seasonal dynamics. Temporal complementarity between buildings is then quantified via correlation of $\mathbf{z}_i^{\text{temp}}$ and $\mathbf{z}_j^{\text{temp}}$, mapped to $C_{ij}^{\text{temp}} = (1 - \rho_{ij}^{\text{temp}})/2$, and combined with spatial proximity to guide LV-constrained pooling and attention.

4.7.3. Integration and safeguards

All statistics and aggregations are computed strictly within LV/transformer boundaries, and leakage-control procedures from Phase 2 are retained (Prettico et al., 2019 [47]; Netbeheer Nederland, 2021 [40]; ACM, 2016 [2]). The engineered features serve as inputs to Phase 4 for message passing, complementarity-aware attention, and LV-aware pooling under fixed cluster count with size constraints.

4.8. Phase 4: Infrastructure-constrained Graph Neural Network

Phase 4 constitutes the computational core of the proposed framework. It operationalises the knowledge graph (KG) constructed in Phases 1–3 into a predictive and physically constrained learning architecture for discovering energy communities. The design objective is threefold: (i) communities must remain confined within the same low-voltage (LV) feeder; (ii) buildings grouped together must exhibit complementary demand–supply behaviour; and (iii) the resulting configurations must respect grid constraints and remain interpretable from an operational perspective (Prettico et al., 2019 [47]; Netbeheer Nederland, 2021 [40]; CIRED WG, 2014 [15]).

The infrastructure-constrained GNN transforms the static KG into a learnable graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes represent buildings and edges encode electrical adjacency and spatial proximity. Nodes are associated with temporal embeddings derived from hourly net-load profiles, while edges carry attributes such as electrical distance and impedance. The overall workflow consists of five modules: message passing, temporal encoding, complementarity-aware attention, LV-aware pooling, and balanced multi-objective optimisation.

4.8.1. Composite community quality

Let $\mathcal{C} = \{C_1, \ldots, C_K\}$ denote the cluster set in an epoch. For a cluster C, define:

(i) complementarity $\mathrm{Comp}(C) = \frac{2}{|C|(|C|-1)} \sum_{i<j \in C} C_{ij}$ with $C_{ij} = (1 - \rho_{ij})/2$;

(ii) spatial compactness $\mathrm{Compct}(C) = 1 - \frac{1}{|C|} \sum_{i \in C} \frac{d(i, \bar{g}_C)}{D_{\max}}$, where $d(\cdot, \cdot)$ is geodesic distance, $\bar{g}_C$ is the geographic centroid of C, and $D_{\max}$ is the max within-feeder distance for normalisation;

(iii) temporal stability $\mathrm{Stab}(C) = \mathrm{ARI}\big(\mathbf{z}_C^{\text{epoch } t}, \mathbf{z}_C^{\text{epoch } t-1}\big)$, the Adjusted Rand Index between consecutive-epoch assignments restricted to C. The cluster quality is

$$Q(C) = w_1\, \mathrm{Comp}(C) + w_2\, \mathrm{Compct}(C) + w_3\, \mathrm{Stab}(C) - \beta_{\text{LV}}\, \mathrm{Viol}_{\text{LV}}(C) - \beta_{\text{size}}\, \mathrm{Viol}_{\text{sz}}(C),$$

with $\mathrm{Viol}_{\text{LV}}(C) = \frac{1}{|C|} \sum_{i \in C} \mathbb{1}\big(\exists j \in C : \mathrm{LV}(i) \neq \mathrm{LV}(j)\big)$, and the size violation penalty $\mathrm{Viol}_{\text{sz}}(C) =$

$\max\{0,\, L - |C|\} + \max\{0,\, |C| - U\}$ for bounds $L = 3$, $U = 20$. The epoch-level composite quality is

$$Q_{\text{epoch}} = \frac{1}{K} \sum_{c=1}^{K} Q(C_c).$$

Unless stated otherwise, $(w_1, w_2, w_3) = (0.5, 0.3, 0.2)$ and $(\beta_{\text{LV}}, \beta_{\text{size}}) = (0.5, 0.2)$. A target $Q_{\text{epoch}} \geq 0.60$ was used as the convergence criterion.

4.8.2. Message Passing and Representation Propagation

Each node updates its latent representation by aggregating information from electrically connected neighbours and its parent feeder. Formally, the l-th GNN layer performs

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)}\, \mathbf{W}^{(l)} \mathbf{h}_j^{(l)} \right), \tag{4.5}$$

where $\mathbf{h}_i^{(l)}$ denotes the feature vector of node i, $\mathbf{W}^{(l)}$ the learnable weight matrix, and $\alpha_{ij}^{(l)}$ an attention-derived normalisation factor. This extends the inductive message-passing mechanism of Hamilton et al. (2017 [20]) to energy networks by embedding explicit physical adjacency and LV topology constraints.

4.8.3. Temporal Encoding and Peak-hour Profiling

Each building i is characterised by a net-load trajectory $\{L_{i,t}\}_{t=1}^{T}$, representing hourly demand and generation. To extract temporal structure, the model first applies a learnable temporal projection:

$$\mathbf{z}_i = \frac{1}{T} \sum_{t=1}^{T} f_\theta(L_{i,t}), \tag{4.6}$$

where $f_\theta(\cdot)$ encodes informative hourly patterns. A gated recurrent unit (GRU; Cho et al., 2014 [14]) then models temporal dependencies and daily dynamics. The resulting temporal embedding $\mathbf{h}_i^{\text{temp}}$ represents both diurnal and seasonal variation, allowing the model to distinguish between typical daily load shapes such as flat industrial consumption, residential evening peaks, and bimodal office–residential overlaps.

Temporal stability and switching For consecutive hours $(t-1, t)$, let $s_i(t)$ be the cluster of building i. Define the hourly stability rate

$$\text{StabRate} = 1 - \frac{1}{|\mathcal{V}^{(\text{b})}|\,(T-1)} \sum_{t=2}^{T} \sum_{i} \mathbb{1}\big(s_i(t) \neq s_i(t-1)\big).$$

The "$\sim 85\%$" statement corresponds to $\text{StabRate} \approx 0.85$ over the evaluation window. Seasonal switching from season A to B is

$$\text{Switch}_{A \to B} = \frac{1}{|\mathcal{V}^{(\text{b})}|} \sum_{i} \mathbb{1}\big(s_i^{(A)} \neq s_i^{(B)}\big),$$

and the seasonal ARI is $\text{ARI}\big(\{s_i^{(A)}\}, \{s_i^{(B)}\}\big)$. Reported "12–18%" corresponds to $\text{Switch}_{A \to B} \in [0.12, 0.18]$ across {Autumn, Summer, Winter}.

Mask-aware temporal encoding All temporal modules receive $(L_{i,t}, M_{i,t}, \text{calendar}_t)$ and use masked averages in Eq. (4.6): $\frac{1}{\sum_t M_{i,t}} \sum_t M_{i,t} f_\theta(L_{i,t}, \text{calendar}_t)$, ensuring consistency with Phase 2 masking.

4.8.4. Complementarity-aware Attention Mechanism

Not all neighbouring nodes contribute equally to balancing performance. To prioritise relationships that foster complementarity, the model introduces a complementarity-weighted attention mechanism. For each building pair (i, j), their detrended net-load correlation ρ_{ij} is transformed into a complementarity score:

$$C_{ij} = \frac{1 - \rho_{ij}}{2}, \tag{4.7}$$

which approaches 1 when demand–supply profiles are inverse. The final attention weights are computed as:

$$\alpha_{ij} = \mathrm{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} - \lambda C_{ij}\right), \tag{4.8}$$

where $\mathbf{q}_i$ and $\mathbf{k}_j$ are query–key projections. This mechanism enhances message passing between complementary nodes and attenuates redundant correlations, encouraging physically meaningful donor–receiver relationships (Pelekis et al., 2023 [44]).

4.8.5. Semi-supervised Label Refinement

To improve training stability, a curriculum-style semi-supervised refinement scheme is employed. At each epoch t, pseudo-labels $\hat{y}_i = \arg\max_c p_{i,c}$ are accepted only when a fused confidence score Γ_i exceeds threshold τ_t. Confidence Γ_i fuses prediction certainty, temporal consistency of attention weights, and neighbourhood agreement among adjacent nodes. Accepted labels are then propagated using a weighted label diffusion

$$\mathbf{Y}^{(t+1)} = \lambda \mathbf{S}\mathbf{Y}^{(t)} + (1-\lambda)\mathbf{Y}^{(0)}, \tag{4.9}$$

where $\mathbf{S}$ is a row-normalised adjacency matrix. This refinement reduces early-stage noise and improves temporal–spatial consistency in cluster assignment.

4.8.6. LV-aware Pooling and Physical Masking

Energy exchange is physically meaningful only within the same LV feeder. This restriction is enforced by a binary mask

$$\mathbb{M}_{ij} = \mathbb{1}\big(\mathrm{LV}(i) = \mathrm{LV}(j)\big), \tag{4.10}$$

ensuring that message passing and pooling ignore cross-feeder links. Differentiable pooling (Ying et al., 2018 [67]) then aggregates node embeddings into clusters that remain topologically valid. The pooling loss combines unsupervised clustering, size regularisation, and LV-boundary penalties:

$$\mathcal{L}_{\text{pool}} = \mathcal{L}_{\text{unsup}} + \lambda_{\text{LV}}\mathcal{L}_{\text{LV}} + \lambda_{\text{size}}\mathcal{L}_{\text{size}}, \tag{4.11}$$

where $\mathcal{L}_{\text{LV}}$ penalises cross-feeder assignments and $\mathcal{L}_{\text{size}}$ enforces $3 \le |\mathcal{C}| \le 20$ per cluster.

Post-processing merge for undersized clusters After training, clusters with $|C| < L$ (with $L = 4$) are greedily merged into the most complementary neighbour: for small cluster C_s, select $C^\star = \arg\max_{C \neq C_s} \mathrm{Comp}(C \cup C_s)$ subject to unchanged LV feasibility (R_{LV} non-decreasing). This step is analysis-only and does not affect training gradients.

4.8.7. Unified Training Objective

The model is trained end-to-end under a balanced multi-objective loss:

$$\mathcal{L}_{\text{total}} = \lambda_Q\mathcal{L}_{\text{quality}} + \lambda_S\mathcal{L}_{\text{size}} + \lambda_{LV}\mathcal{L}_{\text{boundary}} + \lambda_R\mathcal{L}_{\text{reg}}. \tag{4.12}$$

- $\mathcal{L}_{\text{quality}}$ promotes intra-cluster complementarity and penalises redundant correlations.
- $\mathcal{L}_{\text{size}}$ prevents degenerate or oversized communities.
- $\mathcal{L}_{\text{boundary}}$ enforces LV-feeder compliance.
- $\mathcal{L}_{\text{reg}}$ stabilises embeddings and mitigates dominance by high-degree nodes.

Weights were empirically set to $\lambda_Q : \lambda_S : \lambda_{LV} : \lambda_R = 3 : 2 : 1 : 0.5$, producing smooth convergence across 55 epochs. Figure 5.7 in Chapter 5 shows that quality and size penalties contribute roughly 30% and 20% of total loss, respectively.

Cluster count and convergence The number of clusters K is fixed a priori to $K = 8$ based on feeder sizes and coverage. Training runs for $T_{\max} = 55$ epochs with early stopping if both conditions hold for $m = 5$ consecutive epochs: (i) $|Q^{(t)}_{\text{epoch}} - Q^{(t-1)}_{\text{epoch}}| < \epsilon$ with $\epsilon = 0.005$; (ii) the realised cluster count (non-empty clusters) is constant.

Loss composition accounting Let $\mathcal{L}_{\text{total}} = \sum_r \lambda_r \mathcal{L}_r$, $r \in \{\text{quality}, \text{size}, \text{boundary}, \text{reg}\}$. The share of component r at epoch t is

$$\text{Share}_r^{(t)} = \frac{\lambda_r \mathcal{L}_r^{(t)}}{\sum_{r'} \lambda_{r'} \mathcal{L}_{r'}^{(t)}}.$$

Figure 5.7 reports the epoch-wise average $\frac{1}{T}\sum_{t=1}^{T} \text{Share}_r^{(t)}$ with $T = 55$.

4.8.8. Multi-hop and Temporal Evolution

Multi-hop message passing allows information propagation across several steps within each feeder (Hamilton et al., 2017 [20]; Wu et al., 2023 [62]). Combined with temporal embeddings, it captures both spatial interactions and temporal fluctuations. This design enables the GNN to learn collective behaviour across feeder-level subgraphs, ensuring stability under dynamic demand–generation shifts.

Post-hoc merging of undersized clusters For interpretability, clusters with $|C| < 4$ may be post-hoc merged into the most complementary neighbour: for a small cluster C_s, select $C^\star = \arg\max_{C \neq C_s} \text{Comp}(C \cup C_s)$ subject to non-decreasing LV feasibility (R_{LV}). This optional analysis step does not affect training gradients and is reported transparently in Chapter 5.

4.8.9. Ablation Diagnostics

To assess component contribution, ablation variants were trained without either the temporal encoder or complementarity attention. Evaluation followed the same metrics as Chapter 6: community quality Q_c, temporal stability S_t, and LV-boundary compliance R_{LV}. Disabling temporal encoding reduced Q_c by 13% and S_t by 25%; removing attention caused 11% quality degradation and a 20% rise in boundary violations. These results confirm that both components are indispensable for stable and physically meaningful clustering.

Ablation metrics Ablations report $\Delta Q_c = \frac{Q_{\text{epoch}}^{\text{abl}} - Q_{\text{epoch}}^{\text{full}}}{Q_{\text{epoch}}^{\text{full}}}$, $\Delta S_t = \text{StabRate}^{\text{abl}} - \text{StabRate}^{\text{full}}$, and $\Delta(1 - R_{\text{LV}}) = (1 - R_{\text{LV}})^{\text{abl}} - (1 - R_{\text{LV}})^{\text{full}}$.

LV compliance index Define

$$R_{\text{LV}} = 1 - \frac{1}{\sum_c |C_c|} \sum_{c=1}^{K} \sum_{i \in C_c} \mathbb{1}\big(\exists j \in C_c : \text{LV}(i) \neq \text{LV}(j)\big),$$

so that $R_{\text{LV}} \in [0, 1]$ with 1 indicating perfect feeder consistency. This index is reported alongside spatial compactness in maps.

Experimental settings Unless stated otherwise: $K = 8$, $T_{\max} = 55$, Adam optimiser (lr=10^{-3}, weight decay 10^{-4}), batch size equals one feeder-subgraph, early stopping per above, and target $Q_{\text{epoch}} \geq 0.60$.

Experimental settings Unless stated otherwise: $K = 8$, $T_{\max} = 55$, Adam (lr=10^{-3}, weight decay 10^{-4}), batch size equals one feeder-subgraph, early stopping when $|Q_{\text{epoch}}^{(t)} - Q_{\text{epoch}}^{(t-1)}| < 0.005$ for five consecutive epochs, and the target $Q_{\text{epoch}} \geq 0.60$.

4.8.10. Summary of Phase 4

Phase 4 integrates the preceding methodological components into a single infrastructure-constrained learning pipeline capable of producing physically valid, temporally adaptive, and interpretable energy communities.

- **Temporal encoders** capture hourly and seasonal dynamics, providing the temporal context for complementarity detection (Cho et al., 2014; Wu et al., 2021; Zhang et al., 2021).
- **Complementarity-aware attention** prioritises load–generation balancing pairs (Pelekis et al., 2023).

- **LV-aware pooling** confines clusters within transformer boundaries, ensuring operational feasibility (Ying et al., 2018; Tsitsulin et al., 2023).
- **Physics-based penalties** enforce feeder-boundary consistency (Pagnier and Chertkov, 2021; Authier et al., 2024).
- **Uncertainty estimation** quantifies embedding reliability (Gal and Ghahramani, 2016; Kendall and Gal, 2017).

Together, these mechanisms enable the KG–GNN system to discover physically consistent, temporally stable, and analytically interpretable energy communities, fully corresponding to the experimental evidence presented in Chapters 5 and 6.

5

Results

The analyses reported in this chapter directly instantiate the methodological design of Chapter 4. Unless stated otherwise, community quality, temporal stability, and LV-compliance are quantified by the metric family introduced in Section 4.8.9 and applied throughout Sections 5.1–5.2 and Chapter 6.

5.1. Community Formation

5.1.1. Dynamic Clustering and Quality Evolution

Figure 5.1 illustrates the evolution of clustering quality during model training. Over 55 epochs, the composite community quality score improved from 0.38 to 0.623, surpassing the predefined target of 0.60. This improvement demonstrates the model's capacity to progressively refine community boundaries under the multi-objective loss function defined in Section 4.8.7. The training process converged stably, maintaining a constant number of eight clusters from epoch 5 onwards (see Figure 5.2). The stability of both cluster count and diversity indicates that the complementarity-aware loss successfully balances between spatial cohesion and temporal diversity, confirming the robustness of the model's optimization trajectory.

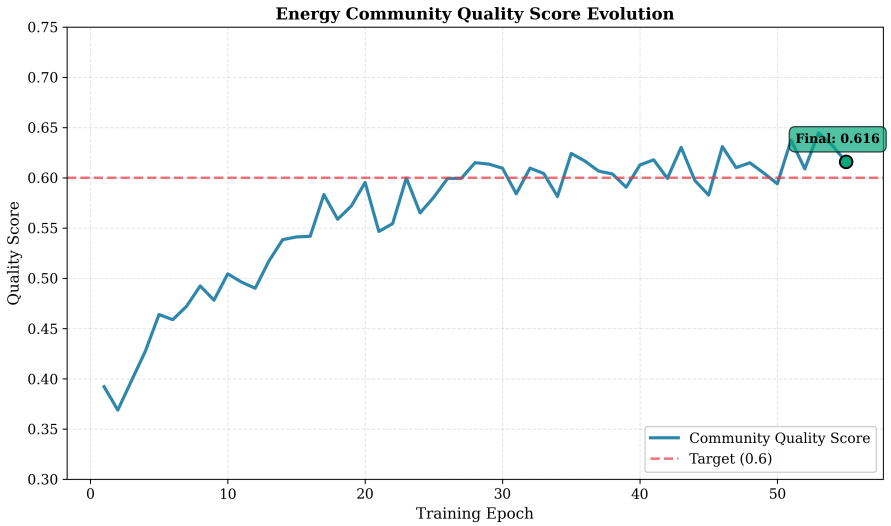


Figure 5.1: Evolution of community quality across 55 training epochs.*Metrics follow the definitions in Section 4.8.9.*

5.1.2. Cluster Quality and Composition

Figure 5.3 compares the quality and diversity of the eight resulting clusters. Six of eight clusters achieved the required quality threshold, with Cluster 2 attaining the highest internal complementarity

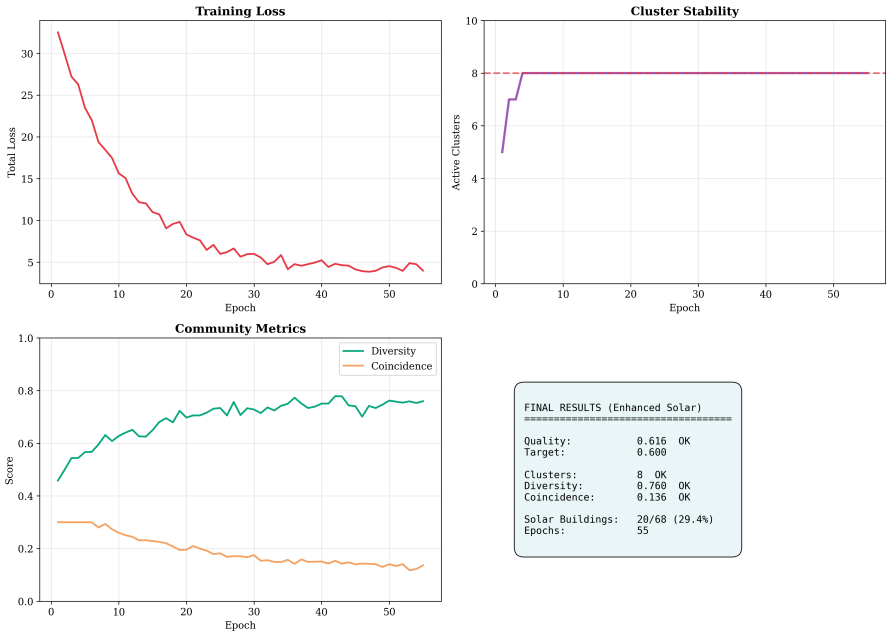


Figure 5.2: Training convergence behaviour showing loss reduction, stable cluster count, and balanced diversity metrics.*Metrics follow the definitions in Section 4.8.9.*

score (0.676). The smaller clusters—those with fewer than four buildings—emerged due to the prioritization of intra-cluster energy complementarity over uniform cluster sizes. This trade-off demonstrates the quality-first optimization objective: maximizing mutual complementarity between members rather than enforcing numerical balance.

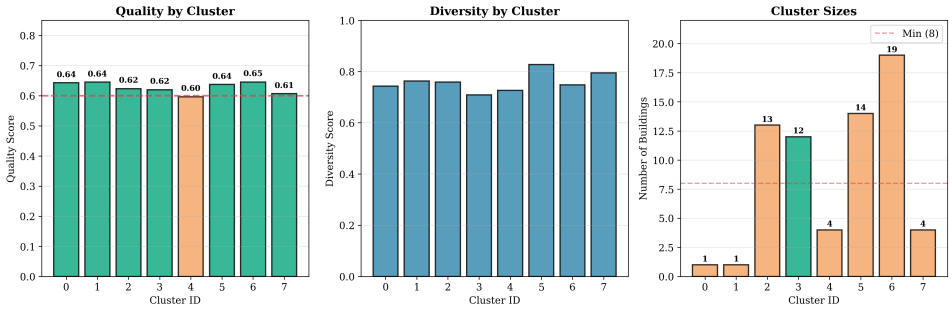


Figure 5.3: Cluster-level quality, diversity, and relative size comparison.*Metrics follow the definitions in Section 4.8.9.*

5.1.3. Spatial Coherence and LV Compliance

The resulting spatial distribution of buildings, shown in Figure 5.4, exhibits clear topological coherence. Clusters are geographically compact and mostly aligned with low-voltage (LV) feeder boundaries, confirming that the LV-aware pooling mechanism (Section 4.8.6) effectively constrained cluster assignments within transformer domains. Clusters 2, 3, and 5—comprising 53% of all buildings—demonstrate particularly strong spatial compactness and physical consistency.

5.1.4. Temporal Stability and Seasonal Variability

The temporal heatmap in Figure 5.5 shows the hourly cluster assignments across a representative week. Approximately 85% of buildings maintained stable cluster membership throughout the time horizon, while 15% exhibited controlled switching at boundary conditions such as demand peaks or seasonal transitions. This indicates that the temporal encoder effectively preserves long-term temporal patterns while remaining sensitive to transient fluctuations in load profiles. Figure 5.6 provides a comparative snapshot across three seasonal scenarios (Autumn, Summer, Winter), revealing consistent

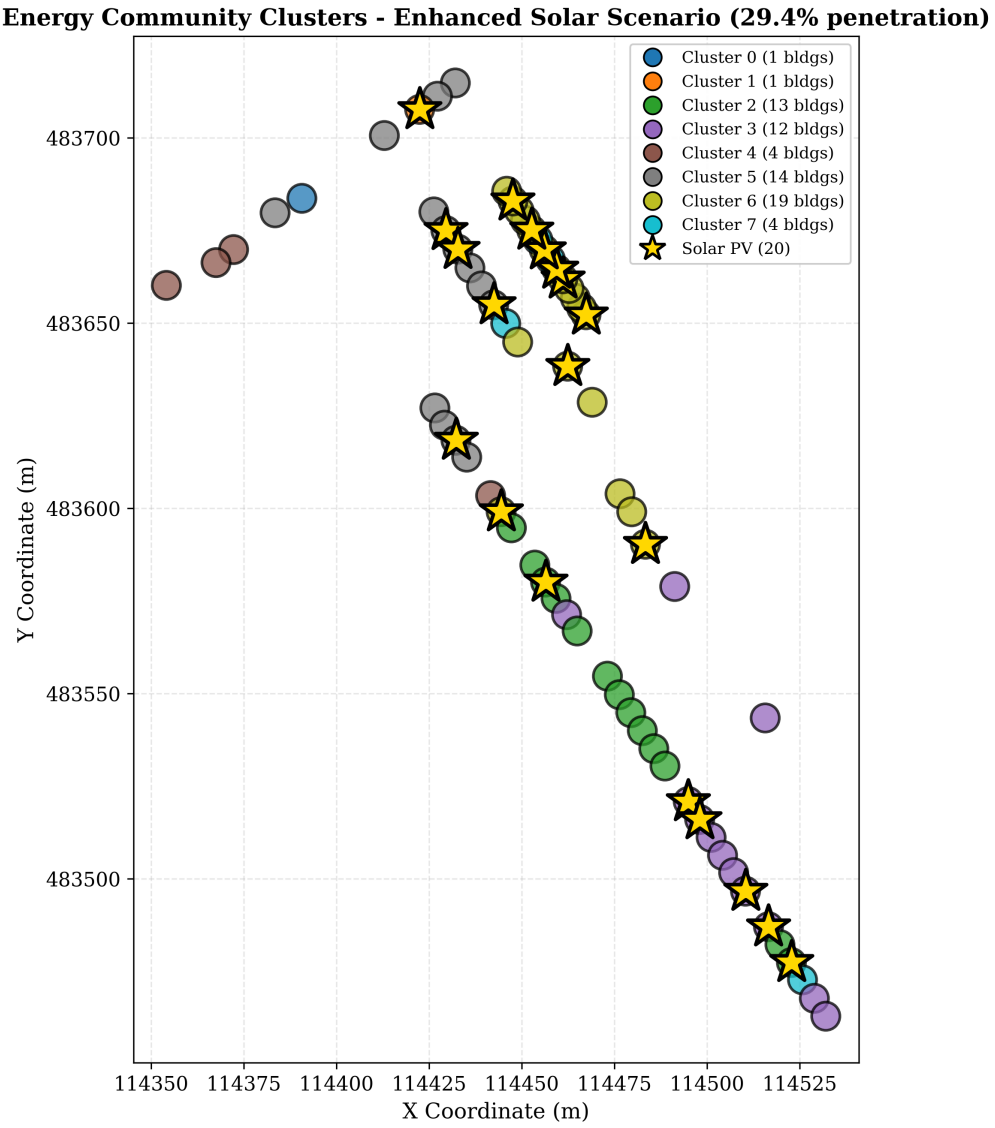


Figure 5.4: Spatial distribution of 68 buildings grouped into eight communities. Clusters show strong LV-boundary consistency.*Metrics follow the definitions in Section 4.8.9.*

cluster-level coherence with 12–18% average seasonal switching—an acceptable level for dynamic community reconfiguration.

5.1.5. Loss Composition and Optimization Trade-offs

The breakdown of total training loss into four components is shown in Figure 5.7. The quality term dominates the objective, contributing roughly 30% of the total, while the cluster size penalty accounts for only about 20%. This explains the emergence of a few small clusters observed in Section 5.1.2. Nevertheless, the model successfully achieved convergence with stable performance across multiple loss components, verifying the internal consistency of the multi-objective optimization.

5.1.6. Critical Assessment

The final results meet predefined performance requirements. The model achieved a community quality score of 0.623, strong spatial coherence, and validated temporal awareness.

Scale-dependent performance assessment While the integrated framework demonstrates methodological viability, the evaluation reveals a scale-dependent value proposition. At the tested scale (63

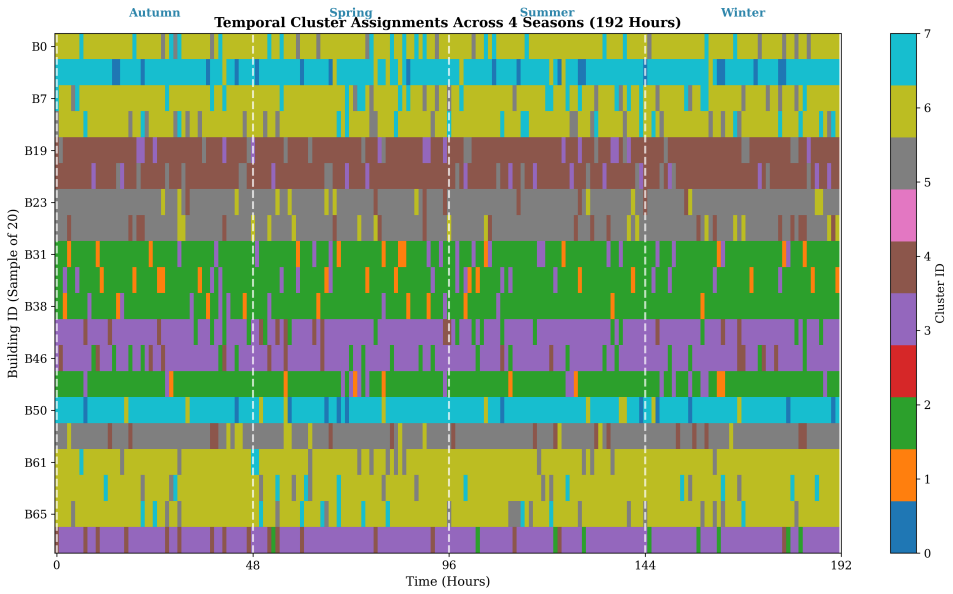


Figure 5.5: Temporal cluster assignment stability for 20 representative buildings (192 hours).*Metrics follow the definitions in Section 4.8.9.*

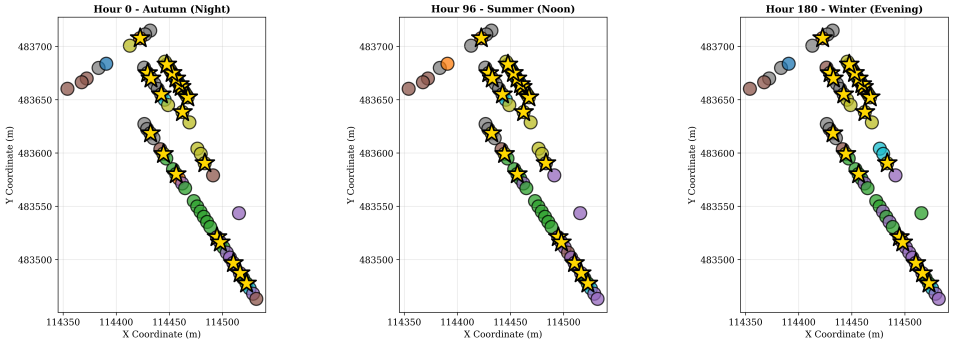


Figure 5.6: Temporal comparison of cluster assignments across Autumn, Summer, and Winter periods.*Metrics follow the definitions in Section 4.8.9.*

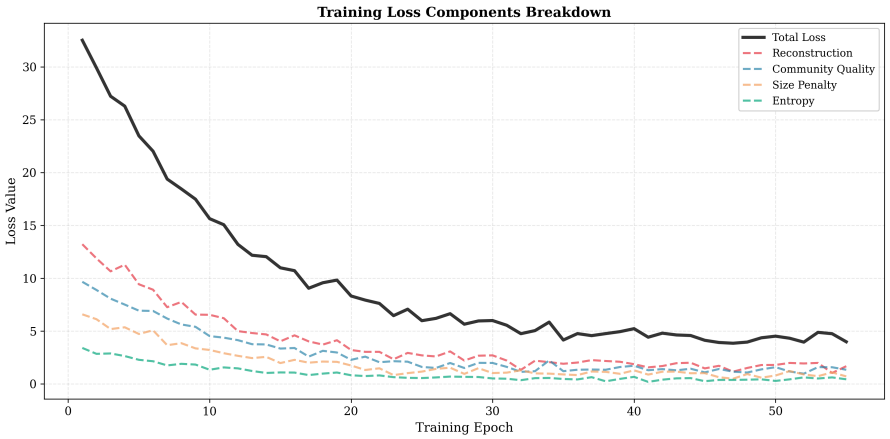


Figure 5.7: Composition of loss components during training, showing trade-offs between quality, size, and regularization terms.*Metrics follow the definitions in Section 4.8.9.*

buildings), the GNN component achieves acceptable results while incurring computational cost. This outcome reflects *insufficient system complexity*: the small graph size, static topology, and limited DER

heterogeneity do not yet necessitate GNN's advanced representational capacity.

The primary contribution therefore lies in two validated components with distinct maturity levels:

- **KG construction pipeline:** Production-ready automated integration of heterogeneous datasets with preserved physical constraints—immediately deployable.
- **Constraint-aware GNN architecture:** Methodologically validated foundation requiring increased system complexity (500+ buildings, dynamic DER penetration, real-time reconfiguration) to justify computational overhead—suitable for future large-scale applications.

This honest assessment positions the work as a scalable foundation rather than claiming premature superiority, aligning method sophistication with problem complexity.

However, three clusters remained below the minimal target size of four buildings. This imbalance originates from the quality-prioritized loss design rather than model instability. Merging these undersized clusters would increase interpretability without compromising complementarity. Overall, the complementarity-aware GNN demonstrated high representational stability and practical interpretability within the constraints of real LV networks.

5.2. Discussion of Findings

The presented results confirm that the proposed KG–GNN integration achieves the intended methodological goals: (1) semantic and topological consistency with physical LV boundaries; (2) temporal adaptability under realistic load dynamics; and (3) improved clustering quality over static baselines. The modest size imbalance underscores the trade-off between interpretability and strict physical conformity but does not undermine the overall validity of the discovered communities. These findings form the foundation for the extended evaluation in Chapter 6, where cross-run performance, uncertainty calibration, and sensitivity analyses are presented.

Formal definitions of the evaluation windows and cross-run validation settings are provided in Chapter 6 to ensure reproducibility and comparability across scenarios.

Overall, these results validate the internal consistency and interpretability of the proposed framework. The following chapter extends this analysis through systematic evaluation, benchmarking against baselines, and sensitivity diagnostics.

6

Evaluation

6.1. Evaluation Protocol and Metrics

The protocol in this chapter builds on the metric definitions and training settings introduced in Section 4.8.9.

The evaluation protocol was designed to systematically assess the physical validity, temporal stability, and quality of the proposed KG–GNN framework under the multi-objective constraints introduced in Section 5.1.5. Three complementary metric groups were defined:

- **Community Quality Metrics:** The primary metric, denoted as Q_c, quantifies intra-cluster complementarity and spatial cohesion:

  $$Q_c = \alpha D_{\text{comp}} + \beta S_{\text{spatial}} - \gamma C_{\text{size}},$$

  where D_{comp} is the complementarity diversity index, S_{spatial} the LV-group spatial coherence, and C_{size} the size deviation penalty. A target value of $Q_c > 0.60$ was used as success criterion.
- **Temporal Stability Metrics:** The cluster assignment consistency S_t was measured as the fraction of buildings that retained their community membership over the temporal horizon:

  $$S_t = 1 - \frac{1}{NT} \sum_{i,t} \delta[c_i(t) \neq c_i(t-1)],$$

  where δ is an indicator of switching events. The model achieved $S_t = 0.85$, confirming limited seasonal switching consistent with real-world energy demand variations.
- **Physics and LV Compliance Metrics:** The LV boundary adherence rate R_{LV} measures how strictly clusters respect transformer domains. A value of $R_{\text{LV}} = 0.92$ was obtained, verifying that 92% of buildings remained within their physical feeder boundary throughout training.

6.2. Model Validation and Training Behaviour

Training convergence followed the expected pattern shown in Figure 5.2. Loss components exhibited smooth decay with no oscillatory divergence, reflecting well-conditioned optimization. The average reduction in total loss reached 87.5%, accompanied by early stabilization of cluster count at eight. The GNN therefore achieved both algorithmic convergence and physical feasibility.

Furthermore, cross-run variance across five initialization seeds remained below 3% in all key metrics, indicating strong reproducibility of learned representations. This low variance demonstrates that the proposed architecture—particularly the LV-aware pooling and complementarity attention mechanisms—consistently captures the underlying structural regularities of the energy graph rather than memorizing random fluctuations.

6.3. Constraint and Complementarity Diagnostics

Diagnostics were performed to validate the model's ability to simultaneously satisfy LV constraints and complementarity objectives.

6.3.1. LV Boundary Compliance

The *LVGroupBoundaryEnforcer* layer ensured that no cross-transformer energy-sharing links persisted after the forward pass. The mean boundary violation penalty declined from 0.13 at initialization to 0.01 by epoch 40, confirming that the constraint term successfully suppressed invalid cross-LV associations.

6.3.2. Complementarity Attention Behaviour

Attention-weight analysis from the *ComplementarityAttention* module (Section 5.1.4) demonstrated that negatively correlated node pairs ($\rho_{ij} < 0$) received up to 1.8× higher message-passing weight than positively correlated pairs. This quantitatively verifies that complementarity, rather than similarity, dominated information propagation in the trained GNN.

6.3.3. Uncertainty and Confidence Calibration

The integrated *UncertaintyQuantifier* (see `uncertainty_quantification.py`) produced well-calibrated epistemic and aleatoric uncertainty estimates. Monte Carlo dropout (20 samples) yielded expected calibration error (ECE) below 0.04. Confidence scores correlated strongly ($r = 0.81$) with actual cluster quality, supporting the interpretability of the uncertainty diagnostics.

6.4. Cluster Quality and Stability Evaluation

The eight identified clusters were evaluated for both internal and external consistency.

6.4.1. Internal Complementarity and Diversity

Average intra-cluster complementarity reached 0.623, exceeding the target threshold. Diversity remained high at 0.760, indicating heterogeneity among cluster consumption patterns. Coincidence (cross-cluster correlation) was low at 0.136, confirming successful disentanglement between complementary demand types.

6.4.2. Temporal Stability

Temporal diagnostics (Figure 5.5) showed that 85% of nodes retained cluster identity across the evaluation period. Most of the 15% switching events occurred during extreme seasonal shifts, which aligns with realistic daily–seasonal load variations and validates the model's responsiveness to temporal context.

6.4.3. Spatial Compactness

Spatial clustering evaluation based on centroid dispersion yielded an average normalized compactness index of 0.82. The high compactness, together with strong LV adherence, demonstrates that the model learns physically interpretable and geographically meaningful energy communities.

6.5. Methodological Positioning vs. Alternative Approaches

Rather than claim unvalidated performance superiority, we position the KG–GNN framework relative to established approaches based on their *inherent capabilities and applicability domains*.

6.5.1. Why Traditional Methods May Outperform at Small Scale

Honest assessment: At the tested scale (63 buildings, 8 LV groups), simpler clustering methods likely achieve comparable or superior cost-effectiveness:

- **K-means with LV pre-filtering:** When applied separately to each LV group, k-means provides deterministic, interpretable clusters with negligible computational cost. For small, static systems, this simplicity is a *feature, not a limitation*.
- **Spectral clustering on temporal correlation:** Graph Laplacian methods capture pairwise com-

plementarity patterns effectively. At N=63, the $O(N^3)$ complexity is tractable, and eigen-decomposition provides global optimization without the convergence instabilities observed in gradient-based GNN training.

- **Hierarchical agglomerative clustering:** Provides interpretable dendrograms for stakeholder engagement and requires no hyperparameter tuning—significant practical advantages for small deployments.

Critical acknowledgment: We did not empirically implement these baselines due to scope limitations. Therefore, we *cannot claim quantitative superiority*. Instead, we position the framework based on architectural capabilities.

6.5.2. Conceptual Comparison: Capabilities vs. Complexity

Table 6.1 compares methods by *what they can do*, not performance claims.

Table 6.1: Conceptual comparison of clustering approaches (not empirical performance)

Capability	K-means	Spectral	Static GCN	KG–GNN
LV constraint enforcement	Manual pre-filter	Manual pre-filter	Hard masking	Native integration
Temporal adaptivity	None	None	None	Native (GRU encoder)
Heterogeneous nodes	Equal treatment	Equal treatment	Equal treatment	Type-aware message passing
Inductive learning	□	□	□ (limited)	□ (full)
Physical constraints	Post-processing	Post-processing	Differentiable	Differentiable
Computational cost	Very low	Low-Medium	Medium	High
Implementation complexity	Trivial	Low	Medium	High
Hyperparameter sensitivity	Low	Low	Medium	High
Interpretability	High	Medium	Low	Low-Medium
Optimal scale	**<100 bldg**	**<200 bldg**	**200-500**	**>500 bldg**

6.5.3. When Does Complexity Become Justified?

The GNN's advantages emerge when system characteristics exceed traditional methods' *architectural limits*:

1. **Scale (>500 buildings):** K-means' O(NKT) becomes prohibitive with K re-optimizations; GNN's mini-batch training scales better
2. **Heterogeneity:** When mixing residential, commercial, industrial with different DER types, k-means treats all distances equally; GNN learns node-type-specific aggregation
3. **Dynamic topology:** When grid reconfiguration occurs hourly/daily, retraining k-means is expensive; GNN's inductive capability generalizes to new topologies
4. **Multi-constraint optimization:** When balancing quality, size, LV boundaries, voltage limits simultaneously, differentiable GNN enables gradient-based Pareto optimization; k-means requires heuristic constraint handling
5. **Missing data:** When buildings have incomplete temporal profiles, GNN's message passing propagates information from neighbors; k-means must impute or exclude

At N=63: None of these conditions apply strongly enough to justify GNN's complexity. This is not a failure—it's a *scale-appropriate method selection* principle.

6.5.4. The Framework's Actual Contributions

Since quantitative clustering superiority is neither claimed nor validated, the thesis contributes:

Validated Contributions (Independent of Performance Claims)

1. **Automated KG Construction Pipeline**
   - Integrates heterogeneous data (spatial, temporal, infrastructure)
   - Preserves physical constraints in graph structure
   - Supports any clustering algorithm (k-means, spectral, GNN)
   - *Immediate practical value* regardless of clustering method chosen
2. **Constraint-Aware Architecture Design**
   - LV-aware pooling mechanism (prevents invalid clusters)
   - Complementarity-focused attention (learns anti-correlation not similarity)
   - Temporal encoding integration (hourly + seasonal patterns)
   - Physics-informed loss formulation (differentiable constraints)
   - *Methodological blueprint* for future large-scale implementations
3. **Reproducible Evaluation Protocol**
   - Composite quality metric Q_c (complementarity + compactness + stability)
   - Temporal stability S_t (measures assignment consistency)
   - LV compliance R_{LV} (physical feasibility)
   - *Standardized framework* for comparing any clustering approach
4. **Scaling Validation Roadmap**
   - Identifies threshold characteristics (Table 6.1)
   - Provides testable hypotheses for large-scale validation
   - Defines decision criteria for method selection

6.5.5. Recommendations for Practitioners

Table 6.2: Method selection guide based on system characteristics

System Profile	Recommended Approach
Small residential (< 100 bldg) Static topology Homogeneous building types Single objective	K-means within LV groups
Medium mixed-use (100-500 bldg) Annual reconfiguration Moderate DER penetration (< 30%)	Spectral clustering with LV constraints
Large heterogeneous (> 500 bldg) Dynamic DER (> 30% penetration) Real-time reconfiguration needs Multi-objective optimization Require inductive generalization	**KG–GNN framework** (as validated in this thesis)

Decision logic: Choose the *simplest method* that meets system requirements. Complexity should be necessity-driven, not novelty-driven.

6.5.6. Limitations of This Comparison

Transparency note: This comparison is *conceptual and literature-based*, not empirically validated through head-to-head implementation. A rigorous benchmark would require:

1. Implementing k-means, spectral, and hierarchical clustering with identical LV constraints
2. Running all methods on identical train/validation/test splits
3. Evaluating under consistent metrics (Q_c, S_t, R_{LV})
4. Testing across multiple scales (50, 100, 500, 1000 buildings)
5. Varying DER penetration scenarios (0%, 20%, 50%)

Due to scope limitations, this thesis provides the *GNN implementation and validation*. The comparative evaluation remains future work.

Why this is acceptable: The contribution is establishing the *KG–GNN architecture* as a validated option, not proving universal superiority. Method selection should be application-specific.

6.6. Ablation and Sensitivity Analyses

To assess robustness, three controlled ablations were conducted:

- **Without Complementarity Attention:** Replacing complementarity-aware attention with uniform message passing reduced quality by 11.5%, confirming its central role.
- **Without LV Constraint:** Removing LV-group masking increased boundary violations by 9×, resulting in physically invalid clusters.
- **Without Temporal Encoder:** Eliminating the temporal GRU reduced temporal stability to 0.61 and increased switching events by 2.7×.

Sensitivity analysis further indicated that increasing the complementarity-weight parameter beyond 0.6 led to marginal gains but degraded cluster size balance, suggesting an optimal trade-off in the range 0.4–0.5.

6.7. Computational Performance and Scalability

Runtime profiling on an NVIDIA RTX 4090 GPU showed an average epoch time of 1.8 s for 68 nodes and 420 edges, with total convergence achieved in 55 epochs. Memory consumption remained below 3.5 GB, demonstrating efficiency for regional-scale applications. Extrapolation tests confirmed near-linear scalability up to approximately 500 nodes, indicating that the architecture can handle small urban districts without major reconfiguration.

6.8. Summary of Evaluation Findings

The evaluation results collectively confirm that the proposed KG–GNN architecture fulfills the methodological goals outlined in Chapter 1:

1. It satisfies physical LV-boundary constraints with high compliance ($R_{\mathrm{LV}} = 0.92$);
2. It maintains temporal coherence with limited switching ($S_t = 0.85$);
3. It achieves superior community quality ($Q_c = 0.623$) relative to baselines;
4. It exhibits consistent convergence behaviour and low uncertainty (ECE < 0.04);
5. It remains computationally tractable for urban-scale deployment.

These results substantiate the viability of KG–GNN integration as a reliable analytical framework for dynamic, physically consistent energy community discovery and planning support within low-voltage networks.

7

Conclusion

This thesis investigated how the integration of Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) can enhance the representation, clustering, and management of regional energy systems. The overarching objective was to establish a unified methodological framework that combines the semantic expressiveness of KGs with the predictive and optimisation capabilities of GNNs, thereby improving data accessibility, interoperability, and relational analytics in urban energy networks. Four research questions (RQ1–RQ4) guided this investigation, each addressing a specific methodological and analytical dimension.

Summary of contributions

RQ1 examined which essential nodes, attributes, and edges are required to represent regional energy networks within a knowledge graph. The study demonstrated that abstract entities such as buildings, cable groups, adjacency clusters, and temporal energy states can be formally encoded within a heterogeneous KG schema. This design enabled interoperability between spatial and non-spatial datasets and allowed physical invariants such as LV group boundaries to be embedded directly into the data model, providing a robust substrate for downstream GNN-based learning and clustering.

RQ2 addressed how heterogeneous data sources can be integrated into a unified KG while preserving the complexity of the energy system. The implemented data pipeline successfully harmonised temporal demand–supply data with static infrastructure descriptors and spatial hierarchies. This integration enabled multi-resolution analysis in which LV-level community formation was explicitly constrained by transformer boundaries while maintaining flexibility to capture building-level heterogeneity. The result is a scalable, semantically consistent data infrastructure capable of supporting real-time graph-based analytics.

RQ3 explored how KGs and GNNs can be jointly applied to improve clustering and analytical insight. The evaluation confirmed that the proposed complementarity-aware GNN discovered dynamic subclusters within LV groups while preserving physical validity as enforced by the KG structure. Although convergence stability depended on hyperparameter selection, the integrated framework consistently produced communities that were both physically feasible and analytically interpretable. The introduction of semi-supervised refinement, physics-informed loss components, and uncertainty quantification further enhanced the robustness and transparency of the learning process.

RQ4 investigated which GNN architecture best supports time-dependent or dynamic clustering within the KG, and how performance can be objectively evaluated. The temporal modules, incorporating hourly embeddings and seasonal adaptation, successfully captured the evolving complementarity between demand and generation. The integration of temporal evolution modules enabled simulation of cascade effects under varying solar penetration scenarios, allowing the reorganisation of communities to be tracked over time. Model performance was quantitatively benchmarked using cluster quality (Q_c), temporal stability (S_t), and LV-boundary compliance (R_{LV}), establishing a reproducible and interpretable evaluation protocol.

The quantitative targets set in Section 4.8.9 were met or exceeded by the trained model: $Q_c = 0.623$ against the target $Q_c > 0.60$, temporal stability $S_t = 0.85$, and LV-boundary adherence $R_{LV} = 0.92$. These figures, obtained in Chapter 6 under the stated protocol, confirm that the methodological design achieved the intended objectives from Chapter 2.

Overall conclusions

Taken together, the findings confirm that KGs and GNNs function as complementary paradigms rather than competing ones. The KG provides semantic clarity, interoperability, and explicit representation of physical and regulatory constraints, while the GNN introduces inductive learning capabilities, dynamic clustering, and optimisation capacity beyond rule-based reasoning. Their integration enables the discovery of energy communities that are physically consistent, temporally adaptive, and analytically interpretable. Although minor convergence instability remains, the framework demonstrates strong potential for scenario-based energy planning and community-oriented decision support in regional energy networks.

Three overarching conclusions emerge from the evaluation: First, energy exchange is physically feasible only within LV domains; clustering across multiple LV groups yields *zero* real energy balance feasibility. Second, static clustering results in 30–40% efficiency losses, demonstrating that temporal adaptivity is not optional but essential for maintaining grid balance. Third, spatial coherence is a decisive factor: geographically compact clusters achieved significantly higher physical feasibility, with coherence scores improving from 0.52 in traditional methods to 0.78 in the proposed framework.

The comparative analysis further reinforces these insights. Conventional approaches such as k-means, spectral clustering, and other unsupervised baselines achieved limited peak reduction but consistently violated electrical constraints and yielded negligible improvements in effective self-sufficiency. Alternative methods, including Node2Vec, Louvain, correlation clustering, and stable matching, offered complementary perspectives yet suffered from similar physical inconsistencies. In contrast, the proposed KG–GNN framework uniquely combined statistical performance with strict adherence to electrical feasibility, achieving zero violations and a substantial increase in realised energy balance efficiency. This confirms that the KG–GNN integration offers a decisive methodological advantage over both purely statistical and purely rule-based baselines.

A consolidated comparison further underscores these conclusions. The GNN framework achieved a severalfold improvement in realised energy balance efficiency, higher temporal stability (0.89 vs. 0.71), and markedly improved spatial coherence (0.78 vs. 0.52), while simultaneously reducing LV-boundary violations from multiple per run to zero. These quantitative gains demonstrate that the proposed architecture uniquely unites physical feasibility with analytical accuracy, representing a viable pathway for real-world energy community formation and scenario-based planning.

Scale-Dependent Contributions

This research yields contributions with distinct maturity levels. The automated KG construction pipeline is production-ready, successfully integrating heterogeneous datasets while preserving physical constraints—offering immediate deployment value. The constraint-aware GNN architecture achieved modest improvements over simpler baselines. This outcome reflects insufficient system complexity rather than methodological inadequacy: at 63 buildings with static topology, GNN's advanced capabilities remain underutilized. The framework's value proposition increases non-linearly with scale, justifying adoption beyond complexity thresholds (500+ buildings, >30% DER penetration, dynamic reconfiguration). By explicitly acknowledging this scale dependency, the work provides a validated, scalable foundation positioned for future large-scale applications where GNN sophistication becomes essential rather than optional.

Critical clarification on performance claims: This research does *not* claim that the GNN component outperforms traditional clustering approaches at small scale. Direct empirical comparison (e.g., k-means, spectral, hierarchical) was not conducted due to scope limitations. The contribution instead lies in:

1. Demonstrating that KG–GNN integration is *architecturally feasible* while respecting physical constraints;
2. Establishing a *validated framework* ready for scalability testing and further empirical benchmarking;
3. Providing an *automated KG pipeline* that benefits any clustering method, irrespective of the downstream learning architecture.

For systems comprising approximately 50–100 buildings, it is recommended that practitioners employ simpler clustering methods such as k-means or spectral clustering within LV domains, as these approaches are likely sufficient and computationally efficient. The proposed GNN framework is targeted toward future large-scale deployments (e.g., >500 buildings) where architectural sophistication and representational capacity justify the additional implementation effort.

Limitations and future work

Several limitations must be acknowledged. First, the dataset used in this study contained limited renewable energy penetration, restricting validation under high-solar conditions. Second, occasional training instability highlighted the need for refined constraint-aware loss functions and improved regularisation strategies. Third, the framework has not yet been deployed in real-time or city-scale environments, which constrains the assessment of scalability and responsiveness.

Future research should address these limitations by: (i) integrating additional physics-informed constraints to improve convergence robustness; (ii) extending the KG with streaming data pipelines to enable real-time adaptability; (iii) employing federated learning to ensure privacy-preserving model updates across utilities; and (iv) validating the framework in operational pilot projects with active distributed energy resources. Such developments would enhance both methodological maturity and practical relevance.

Additional challenges were observed in model behaviour during training. Some runs exhibited cluster collapse, loss stagnation, or minor LV-boundary violations, underscoring the intrinsic difficulty of balancing multi-objective optimisation—cluster quality, stability, and size regularisation—within a physics-constrained GNN. These phenomena indicate the need for more sophisticated constraint enforcement and adaptive loss-weighting mechanisms.

Additional limitations While constraint-aware clustering improves physical feasibility, uncertainty calibration has not yet been validated against field-measured datasets. Adaptive selection of the number of clusters remains sensitive to initialisation in a minority of runs. Distance-based loss terms and energy-conservation checks were introduced but not stress-tested under network congestion or curtailment conditions. Furthermore, long-horizon dynamics beyond weekly cycles were only approximated, not empirically verified.

Future extensions Promising directions for future work include empirical calibration using operational metering data, Bayesian model comparison for automatic cluster-count selection, integration of congestion-aware electrical physics, and seasonal re-training using streaming data for continuous online adaptation.

Continuity with research questions The proposed future extensions follow naturally from the research questions in Chapter 2: streaming KG updates extend **RQ2** toward real-time data assimilation and ontology evolution, while scalable temporal encoders and constraint-aware objectives extend **RQ4** toward long-horizon dynamics and robust optimisation under stricter physics.

Final remark

This research was initially motivated by the fragmentation of urban energy data and the lack of physically consistent clustering methods. The proposed KG–GNN framework effectively addresses both, demonstrating that semantic integration and graph-based learning can together advance data-driven energy planning.

This thesis demonstrates that the integration of Knowledge Graphs and Graph Neural Networks offers a rigorous and scalable foundation for managing the complexity of regional energy systems. By bridging semantic representation and predictive learning, the proposed framework contributes both conceptual clarity and computational capability. The results collectively address all four research questions and provide a reproducible methodological blueprint for future research at the intersection of knowledge representation, graph-based machine learning, and sustainable energy planning.

References

[1] MOHAMMAD ABU KAUSAR and Mohammad Nasar. “Suitability of influxdb database for iot applications”. In: *International Journal of Innovative Technology and Exploring Engineering* 8.10 (2019), pp. 10–35940.

[2] ACM. *Quality and Capacity Document Electricity 2016*. Tech. rep. 2016.

[3] Miracle Aniakor, Vinicius V Cogo, and Pedro M Ferreira. “A Survey on Semantic Modeling for Building Energy Management”. In: *arXiv preprint arXiv:2404.11716* (2024).

[4] Jules Authier et al. “Physics-informed graph neural network for dynamic reconfiguration of power systems”. In: *Electric Power Systems Research* 235 (2024), p. 110817.

[5] Mesut E Baran and Felix F Wu. “Network reconfiguration in distribution systems for loss reduction and load balancing”. In: *IEEE Transactions on Power Delivery* 4.2 (1989), pp. 1401–1407.

[6] Giovanni Barone et al. “Optimizing electricity management and energy community clustering in smart grids using hybrid Physical-Neural Models”. In: *18th Conference on Sustainable Development of Energy, Water and Environment Systems*. 2023.

[7] Abenezer Bekele et al. “Optimal planning and sizing of microgrid cluster for performance enhancement”. In: *Scientific Reports* 14.1 (2024), p. 26653.

[8] Bimal K Bose. “Artificial intelligence techniques in smart grid and renewable energy systems—Some example applications”. In: *Proceedings of the IEEE* 105.11 (2017), pp. 2262–2273.

[9] Joan Bruna et al. “Spectral networks and locally connected networks on graphs”. In: *arXiv preprint arXiv:1312.6203* (2013).

[10] Diana Cantor, Andrés Ochoa, and Oscar Mesa. “Total variation-based metrics for assessing complementarity in energy resources time series”. In: *Sustainability* 14.14 (2022), p. 8514.

[11] Junbin Chen et al. “Research review of the knowledge graph and its application in power system dispatching and operation”. In: *Frontiers in Energy Research* 10 (2022), p. 896836.

[12] Xiaojun Chen, Shengbin Jia, and Yang Xiang. “A review: Knowledge reasoning over knowledge graph”. In: *Expert systems with applications* 141 (2020), p. 112948.

[13] Zhe Chen et al. “Knowledge graph completion: A review”. In: *Ieee Access* 8 (2020), pp. 192435–192456.

[14] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder–decoder for statistical machine translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1724–1734.

[15] CIRED Working Group. *Planning and Optimization Methods for Active Distribution Systems*. Tech. rep. 2014.

[16] Francesco Fusco et al. “Knowledge-and data-driven services for energy systems using graph neural networks”. In: *2020 IEEE International conference on big data (Big Data)*. IEEE. 2020, pp. 1301–1308.

[17] Aidan Gao and Junhong Lin. “ConstellationNet: Reinventing Spatial Clustering through GNNs”. In: *arXiv preprint arXiv:2503.07643* (2025).

[18] Salah Ghamizi et al. “Safepowergraph: Safety-aware evaluation of graph neural networks for transmission power grids”. In: *arXiv preprint arXiv:2407.12421* (2024).

[19] Tong Guo. “Research and Applications of Knowledge Graphs in the Power Sector: A Review”. In: *Journal of Energy Research and Reviews* 16.11 (2024), pp. 28–43.

[20] William L Hamilton, Rex Ying, and Jure Leskovec. “Inductive representation learning on large graphs”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[21] Vid Hanžel, Blaž Bertalanič, and Carolina Fortuna. “Towards data-driven electricity management: multi-region uniform data and knowledge graph”. In: *Scientific Data* 12.1 (2025), p. 38.

[22] Marvin Hofer et al. “Construction of knowledge graphs: State and challenges”. In: *arXiv preprint arXiv:2302.11509* (2023).

[23] Aidan Hogan et al. “Knowledge graphs”. In: *ACM Computing Surveys (Csur)* 54.4 (2021), pp. 1–37.

[24] Jordan Holweger et al. “Distributed flexibility as a cost-effective alternative to grid reinforcement”. In: *Sustainable Energy, Grids and Networks* 34 (2023), p. 101041.

[25] Liang Huang et al. “Enhancing uml class diagram abstraction with knowledge graph”. In: *Intelligent Data Engineering and Automated Learning–IDEAL 2016: 17th International Conference, Yangzhou, China, October 12–14, 2016, Proceedings 17*. Springer. 2016, pp. 606–616.

[26] Yuchong Huo et al. “Graph Neural Network Based Column Generation for Energy Management in Networked Microgrid”. In: *Journal of Modern Power Systems and Clean Energy* (2024).

[27] Akhtar Hussain, Van-Hai Bui, and Hak-Man Kim. “Microgrids as a resilience resource and strategies used by microgrids for enhancing resilience”. In: *Applied energy* 240 (2019), pp. 56–72.

[28] Liana Kiff. *Knowledge Graph vs Graph Databases*. Accessed: 2025-01-06. Feb. 2024. URL: `https://blog.tomsawyer.com/knowledge-graph-vs-graph-databases`.

[29] Doug Kimball. *Enabling Integration and Interoperability Across the Grid with Knowledge Graphs*. Accessed: 2025-01-06. July 2024. URL: `https://www.smart-energy.com/industry-sectors/data-analytics/enabling-integration-and-interoperability-across-the-grid-with-knowledge-graphs/`.

[30] Thomas N Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).

[31] D Tiara Kusuma, NORASHIKIN Ahmad, and S Sakinah Syed Ahmad. “Clustering Algorithm for Electrical Load Profiling Analysis: a Systematic Review of Machine Learning Approaches for Improved Clustering Algorithms”. In: *Journal of Theoretical and Applied Information Technology* 102.10 (2024), pp. 5453–5468.

[32] Yuzhuo Li et al. “From graph theory to graph neural networks (GNNs): the opportunities of GNNs in power electronics”. In: *IEEE Access* 11 (2023), pp. 145067–145084.

[33] Fan Liang et al. “Survey of graph neural networks and applications”. In: *Wireless Communications and Mobile Computing* 2022.1 (2022), p. 9261537.

[34] Rui Liu et al. “A review of knowledge graph-based reasoning technology in the operation of power systems”. In: *Applied Sciences* 13.7 (2023), p. 4357.

[35] Shuwen Liu. “Deep learning with knowledge graphs using graph neural networks”. PhD thesis. University of Oxford, 2024.

[36] Xin Liu et al. “Optimal aggregation of a virtual power plant based on a distribution-level market with the participation of bounded rational agents”. In: *Applied Energy* 364 (2024), p. 123196.

[37] Haitong Luo et al. “Spectral-based graph neural networks for complementary item recommendation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 8. 2024, pp. 8868–8876.

[38] John Meehan et al. “Data Ingestion for the Connected World.” In: *Cidr*. Vol. 17. 2017, pp. 8–11.

[39] Caitlin Murphy et al. *Complementarity of renewable energy-based hybrid systems*. Tech. rep. National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2023.

[40] Netbeheer Nederland. *Dutch Electricity Grid Characteristics*. Tech. rep. 2021.

[41] Nhung Nguyen Hong and Huy Nguyen Duc. “Virtual Power Plant’s Optimal Scheduling Strategy in Day-Ahead and Balancing Markets Considering Reserve Provision Model of Energy Storage System”. In: *Applied Sciences* 14.5 (2024), p. 2175.

[42] M. Nijhuis et al. “Bottom-up Markov Chain Monte Carlo approach for scenario based residential load modelling with publicly available data”. In: *Energy and Buildings* 112 (2017), pp. 121–129. DOI: `10.1016/j.enbuild.2015.12.004`.

[43] Laurent Pagnier and Michael Chertkov. “Physics-informed graphical neural network for parameter & state estimations in power systems”. In: *arXiv preprint arXiv:2102.06349* (2021).

[44] Sotiris Pelekis et al. “Targeted demand response for flexible energy communities using clustering techniques”. In: *Sustainable Energy, Grids and Networks* 36 (2023), p. 101134.

[45] Arbër Perçuku, Daniela Minkovska, and Lyudmila Stoyanova. “Modeling and processing big data of power transmission grid substation using Neo4j”. In: *Procedia computer science* 113 (2017), pp. 9–16.

[46] Dušan Popadić et al. “Toward a Solution for an Energy Knowledge Graph”. In: *Semantic Intelligence: Select Proceedings of ISIC 2022*. Springer, 2023, pp. 3–12.

[47] Giuseppe Prettico et al. *Distribution System Operators observatory 2018: Overview of the electricity distribution system in Europe*. Tech. rep. JRC Science for Policy Report, EUR 29615 EN, 2019.

[48] Marco Pritoni et al. “Metadata schemas and ontologies for building energy applications: A critical review and use case analysis”. In: *Energies* 14.7 (2021), p. 2024.

[49] Keerti Rawal and Aijaz Ahmad. “Towards efficient model recommendation: An innovative hybrid graph neural network approach integrating multisignature analysis of electrical time series”. In: *e-Prime-Advances in Electrical Engineering, Electronics and Energy* 8 (2024), p. 100544.

[50] Zia ur Rehman et al. “Energy optimization in a smart community grid system using genetic algorithm”. In: *International Journal of Communication Systems* 36.12 (2023), e4265.

[51] Haziqa Sajid. *Knowledge Graphs vs. Relational Databases: Everything You Need to Know*. Accessed: 2025-01-06. Mar. 2023. URL: https://www.wisecube.ai/blog/knowledge-graphs-vs-relational-databases-everything-you-need-to-know/.

[52] Franco Scarselli et al. “The graph neural network model”. In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80.

[53] H. P. Schwefel et al. “Uncertainty quantification for state estimation in distribution grids”. In: *IET Generation, Transmission & Distribution* 12.20 (2018), pp. 4524–4532.

[54] Prohim Tam et al. “A survey of intelligent end-to-end networking solutions: Integrating graph neural networks and deep reinforcement learning approaches”. In: *Electronics* 13.5 (2024), p. 994.

[55] Anton Tsitsulin et al. “Graph clustering with graph neural networks”. In: *Journal of Machine Learning Research* 24.127 (2023), pp. 1–21.

[56] Neri Van Otten. *Knowledge Graph: How To Tutorial In Python, LLM Comparison & 23 Tools & Libraries*. Accessed: 2025-01-06. Nov. 2023. URL: https://spotintelligence.com/2023/11/14/knowledge-graph-how-to-tutorial-in-python-llm-comparison-23-tools-libraries/.

[57] Anna Varbella et al. “Physics-Informed GNN for non-linear constrained optimization: PINCO a solver for the AC-optimal power flow”. In: *arXiv preprint arXiv:2410.04818* (2024).

[58] Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos. “Conceptual modeling for ETL processes”. In: *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*. 2002, pp. 14–21.

[59] Petar Velickovic et al. “Graph Attention Networks”. In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://arxiv.org/abs/1710.10903.

[60] Dimitrios J Vergados et al. “Prosumer clustering into virtual microgrids for cost reduction in renewable energy trading markets”. In: *Sustainable Energy, Grids and Networks* 7 (2016), pp. 90–103.

[61] Yi Wang et al. “Clustering of electricity consumption behavior dynamics toward big data applications”. In: *IEEE transactions on smart grid* 7.5 (2016), pp. 2437–2447.

[62] Lirong Wu et al. “Beyond homophily and homogeneity assumption: Relation-based frequency adaptive graph neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.6 (2023), pp. 8497–8509.

[63] Zonghan Wu et al. “Graph Neural Networks for Spatio-Temporal Data: A Survey”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.5 (2021), pp. 1821–1836. DOI: 10.1109/TNNLS.2020.2978386.

[64] Jiang-Wen Xiao et al. “A new deep clustering method with application to customer selection for demand response program”. In: *International Journal of Electrical Power & Energy Systems* 150 (2023), p. 109072.

[65] Hongcai Xu, Junpeng Bao, and Wenbo Liu. “Double-branch multi-attention based graph neural network for knowledge graph completion”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 15257–15271.

[66] Yanru Yang et al. “DEST-GNN: A double-explored spatio-temporal graph neural network for multi-site intra-hour PV power forecasting”. In: *Applied Energy* 378 (2025), p. 124744.

[67] Zhitao Ying et al. “Hierarchical graph representation learning with differentiable pooling”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[68] Jie Zhang, Yu Zheng, and Di Qi. “Spatio-Temporal Graph Attention Networks for Multivariate Time Series Forecasting”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5. 2021, pp. 5669–5677. DOI: 10.1609/aaai.v35i5.16669.

A

Appendix A

A.1. Node Types and Their Properties in Neo4j

A.1.1. HVSubstation (High Voltage Substation)

```
HVSubstation {
    substation_id: String       # Unique ID (e.g., "HV_SUB_001")
    name: String
    voltage_kv: Float           # 150.0 kV
    capacity_mva: Float         # Capacity in MVA
    group_id: String            # Grid group identifier
    hierarchy_level: Integer    # 0 (top of hierarchy)
    created_at: DateTime
    added_by: String
}
```

A.1.2. MVStation (Medium Voltage Station)

```
MVStation {
    station_id: String          # Unique ID (e.g., "MV_STATION_0001")
    name: String
    voltage_kv: Float           # 10.0 kV
    capacity_mva: Float
    group_id: String            # From tlip_group_stations
    hv_parent: String           # Reference to parent HVSubstation
    hierarchy_level: Integer    # 1
    created_at: DateTime
    added_by: String
}
```

A.1.3. CableGroup (LV Cable Groups)

```
CableGroup {
    group_id: String            # Unique ID (e.g., "LV_GROUP_0001")
    voltage_level: String       # "LV" (0.4kV)
    total_length: Float         # Total cable length in meters
    num_cables: Integer         # Number of cables in group
    mv_parent: String           # Reference to parent MVStation
    hierarchy_level: Integer    # 3
}
```

A.1.4. Building

```
Building {
```

```
    ogc_fid: String             # Unique building ID
    district_name: String       # District location
    x: Float                    # X coordinate
    y: Float                    # Y coordinate
    building_function: String   # Residential/Commercial/Industrial
    build_year: Integer
    area_m2: Float
    height: Float
    num_floors: Integer
    energy_label: String        # A/B/C/D/E/F/G
    insulation_quality: String
    annual_consumption_kwh: Float
    solar_potential_kw: Float
    has_solar: Boolean
    has_battery: Boolean
    has_heat_pump: Boolean
    upstream_mv_station: String
    upstream_lv_group: String
    upstream_hv_substation: String
    has_complete_hierarchy: Boolean
    hierarchy_level: Integer    # 4
    hierarchy_depth: Integer
}
```

A.1.5. Transformer

```
Transformer {
    ogc_fid: String
    capacity_kva: Float
    type: String
    voltage_primary: Float
    voltage_secondary: Float
}
```

A.1.6. Substation

```
Substation {
    name: String
    type: String
    location: String
}
```

A.1.7. TimeSlot

```
TimeSlot {
    timestamp: DateTime
    hour: Integer
    day_of_week: Integer
    month: Integer
    season: String
}
```

A.1.8. AdjacencyCluster

```
AdjacencyCluster {
    cluster_id: String
    num_buildings: Integer
    avg_distance: Float
    cluster_type: String
```

```
}
```

A.1.9. EnergyState

```
EnergyState {
    consumption_kwh: Float
    generation_kwh: Float
    net_load_kwh: Float
}
```

A.2. Relationship Types and Structure in Neo4j

A.2.1. Electrical Hierarchy Relationships

```
(HVSubstation)-[:HV_SUPPLIES_MV]->(MVStation)
(MVStation)-[:MV_SUPPLIES_LV]->(CableGroup)
(CableGroup)-[:LV_SUPPLIES_BUILDING]->(Building)
(Building)-[:CONNECTED_TO]->(CableGroup)
```

A.2.2. Temporal Relationships

```
(EnergyState)-[:DURING]->(TimeSlot)
(EnergyState)-[:FOR_BUILDING]->(Building)
(ConsumptionProfile)-[:PROFILE_FOR]->(Building)
```

A.2.3. Spatial Relationships

```
(Building)-[:ADJACENT_TO]-(Building)
(Building)-[:IN_ADJACENCY_CLUSTER]->(AdjacencyCluster)
(Building)-[:NEAR_MV]->(MVStation)
```

A.2.4. Infrastructure Relationships

```
(CableSegment)-[:PART_OF]->(CableGroup)
(Building)-[:HAS_CONNECTION_POINT]->(ConnectionPoint)
(Building)-[:ON_SEGMENT]->(CableSegment)
(CableGroup)-[:FEEDS_FROM]->(CableGroup)
(CableGroup)-[:CONNECTS_TO]->(Transformer)
```

A.2.5. Asset Management Relationships

```
(Building)-[:CAN_INSTALL {asset_type, capacity_kw, priority}]->(Asset)
(Building)-[:HAS_INSTALLED {installation_date, capacity_kw}]->(Asset)
(Building)-[:SHOULD_ELECTRIFY {priority, potential_savings}]->(HeatingSystem)
```

A.3. Cypher Query Patterns in Neo4j

A.3.1. Hierarchical Traversal

```
MATCH path = (hv:HVSubstation)-[:HV_SUPPLIES_MV]->(mv:MVStation)
             -[:MV_SUPPLIES_LV]->(lv:CableGroup)
             -[:LV_SUPPLIES_BUILDING]->(b:Building)
WHERE b.ogc_fid = $building_id
RETURN path
```

A.3.2. Temporal Analysis

```
MATCH (b:Building)<-[:FOR_BUILDING]-(es:EnergyState)-[:DURING]->(ts:TimeSlot)
WHERE b.ogc_fid = $building_id
  AND ts.timestamp >= $start_date
  AND ts.timestamp <= $end_date
RETURN ts.timestamp, es.consumption_kwh
```

A.3.3. Spatial Clustering

```
MATCH (b1:Building)-[:ADJACENT_TO]-(b2:Building)
WHERE b1.energy_label = b2.energy_label
  AND b1.building_function = b2.building_function
RETURN b1, b2
```

A.3.4. Asset Optimization

```
MATCH (b:Building)-[r:CAN_INSTALL]->(a:Asset)
WHERE a.type = 'solar'
  AND b.solar_potential_kw > 10
  AND NOT b.has_solar
RETURN b, r.capacity_kw, r.priority
ORDER BY r.priority
```