algorithms

MDPI

Article

A Relation-Aware Multi-Driver Pipeline for Interpretable Low-Frequency Load Disaggregation Under Partial Observability

Balázs András Tolnai, Zheng Grace Ma and Bo Nørregaard Jørgensen *

SDU Center for Energy Informatics, Maersk Mc-Kinney Moller Institute, The Faculty of Engineering, University of Southern Denmark, 5230 Odense, Denmark; zma@mmmi.sdu.dk (Z.G.M.)
* Correspondence: bnj@mmmi.sdu.dk

Abstract

Non-intrusive load monitoring (NILM) estimates component-level energy use from aggregate measurements, but low-frequency data limit appliance signatures and make overlapping or weakly observed loads difficult to separate. This paper proposes a relation-aware multi-driver pipeline for interpretable low-frequency load attribution under partial observability. The method does not require supervised component labels or predefined appliance models. It combines semantic feature typing, heterogeneous relation discovery, feature-family construction, mechanism-aware evidence modeling, conservative allocation, event-background separation, and role-based attribution. Only evidence-supported load is assigned to feature families, while unsupported variation is retained as unexplained demand or residual load. The method is evaluated in a simulated EV-focused building case and through measured-building validation on nine ADRENALIN buildings. In the EV case, the selected EV-aligned family achieved a correlation of 0.990 and an NMAE of 0.100 against the withheld EV reference, while heat-pump and base-load recovery was weaker, with NMAE values of 0.565 and 0.895. In the ADRENALIN validation, temperature-associated families achieved median NMAE values of 0.594 using the restricted feature set and 0.576 using the full feature set. Additional comparison, ablation, sensitivity, diagnostic, and runtime analyses show that the pipeline is most effective for dominant event-driven loads, remains limited for smoother or masked lower-magnitude components, and treats unexplained variation explicitly. The results demonstrate a practical framework for interpretable driver-based load attribution when component labels are unavailable or incomplete.

Keywords: non-intrusive load monitoring; low-frequency load disaggregation; relation-aware attribution; feature-family modeling; partial observability; electric vehicle charging; interpretable energy analytics

Check for updates

Academic Editor: Bo Yang

Received: 13 May 2026
Revised: 19 June 2026
Accepted: 24 June 2026
Published: 27 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Energy use in buildings remains a central challenge for decarbonization and power-system planning. Building operations account for approximately 30% of global energy consumption and 26% of global energy-related emissions [1], while buildings in the European Union are responsible for around 40% of energy consumption, with heating, cooling, and domestic hot water representing a major share of household building energy use [2]. This makes building-level energy analysis important for improving efficiency, reducing emissions, and supporting flexible operation in future energy systems. Non-intrusive load monitoring (NILM), also referred to as energy load disaggregation, addresses this need by estimating appliance-level, end-use, or component-level consumption from aggregate

meter data without requiring extensive submetering [3,4]. NILM can support energy feedback, personalized efficiency recommendations, demand-side management, operational diagnostics, and improved understanding of end-use behavior [4–6].

The importance of NILM is increasing as buildings become more electrified and more strongly coupled to transportation and heating systems. Electric vehicles introduce large and temporally concentrated charging loads that can affect local distribution grids [7–9]. Beyond their impact on peak demand, EVs can also provide flexible demand–response and vehicle-to-grid services when charging and discharging are coordinated [10]. Identifying EV-related load components from aggregate measurements is relevant not only for energy feedback, but also for flexibility assessment, grid-aware charging strategies, and demand–response planning. EV charging has consequently become an important target for low-frequency load disaggregation, particularly because charging behavior may remain visible in smart-meter or advanced metering infrastructure data even when high-frequency appliance signatures are unavailable [11–13]. At the same time, heating and cooling loads are often driven by outdoor temperature, building operation, and occupant behavior, making them difficult to separate using conventional appliance-state assumptions alone [14–16].

Low-frequency disaggregation remains challenging because standard smart-meter measurements provide limited temporal detail. High-frequency NILM can exploit transient electrical signatures, switching behavior, and appliance-specific load shapes, but these signatures are largely absent at hourly or sub-hourly resolutions [17–19]. As a result, overlapping loads become harder to distinguish, and classical event-based approaches may become unreliable when multiple processes operate simultaneously [9,20]. These limitations are particularly relevant for EV charging and heating-related demand, where aggregate consumption may reflect user behavior, weather, control logic, calendar effects, and stochastic variation rather than a single identifiable appliance state [21,22].

Contextual variables such as weather, time, electricity price, emissions intensity, and operational indicators are increasingly used in energy analysis and forecasting [22–24]. However, their use in load disaggregation remains less direct. In many applications, contextual features are treated as predictors of aggregate demand rather than as interpretable drivers that can support component-level attribution. This creates a need for methods that can type heterogeneous features, quantify multiple forms of relation between drivers and load, group related features into interpretable families, and convert these relationships into dynamic evidence for decomposition [25–27].

Such feature typing is conceptually related to prior work on semantic data type detection, where column-level properties are used to infer the role of variables in tabular datasets [28]. Existing approaches typically rely on statistical, structural, and pattern-based descriptors to assign predefined types [29], but they do not explicitly capture the structural and temporal behavior required for interpretable load decomposition.

Despite significant methodological progress, existing load disaggregation approaches face persistent limitations. Supervised learning methods can achieve strong performance when labeled appliance or component data are available, but such labels are expensive to obtain and often unavailable outside benchmark datasets [30,31]. Optimization-based, clustering-based, probabilistic, spectral, and deep-learning methods each address parts of the problem, but many remain appliance-specific, event-centered, or dependent on strong assumptions about the observed system [32–34]. Furthermore, many disaggregation formulations implicitly seek to explain the full aggregate signal, even though real-world electricity consumption may include unobserved drivers, weakly measured influences, and residual variation that cannot be reliably attributed [34,35].

To address these challenges, this paper proposes a data-agnostic, relation-aware pipeline for low-frequency aggregate-load attribution. The method uses aggregate load and heterogeneous contextual or operational variables to construct an intermediate representation of the observed system through feature typing, multi-mechanism relation discovery, target-centric relevance analysis, and multi-layer feature-family grouping. These structures are transformed into dynamic evidence signals describing how contextual and operational variable groups relate to the aggregate load. The framework is designed to handle heterogeneous driver types, including continuous weather variables, calendar indicators, event-like signals, and transformed local-context features.

The proposed formulation treats the aggregate load as the target signal and derives interpretable feature-family contributions from relation-supported evidence. Rather than requiring appliance-specific target channels during model construction, the pipeline identifies variable groups that provide evidence for parts of the aggregate-load variation. Component-level signals, when available, are used only for post hoc evaluation of whether inferred family contributions align with physically meaningful load processes. In this sense, the method performs disaggregation through relation-supported attribution: it separates the aggregate signal into interpretable feature-family contributions without requiring the target components to be predefined during model construction.

A key aspect of the framework is its conservative treatment of unexplained variation. The method estimates the evidence-supported portion of the target signal and preserves unsupported variation as unexplained dynamic load or residual demand. The pipeline incorporates mechanism modeling to distinguish between event-driven activation, continuous modulation, and contextual adjustment, enabling structured attribution first at the family level and then within families through role-based attribution.

The proposed method is evaluated using a dataset generated from a multi-agent simulation of decentralized EV charging strategies in a Danish radial distribution network serving 126 households [7]. The analyzed target is a single simulated building-level aggregate-load signal with contextual variables and simulated component-level reference signals for EV charging, heating, and base/inflexible demand. These reference components are excluded from the model inputs and retained only for post hoc evaluation. The primary evaluation examines whether relation-supported feature families inferred from aggregate and contextual data align with known simulated components, rather than performing supervised component reconstruction. Additional validation is performed on nine measured commercial and public buildings from the ADRENALIN dataset, where selected temperature-associated feature-family contributions are compared against measured temperature-dependent reference loads. The evaluation is further complemented by comparison with the earlier MD-ADD formulation, task-aligned Bayesian thermal comparisons on the primary EV dataset and on the ADRENALIN buildings, mechanism ablation, parameter sensitivity, input-perturbation and temporal-resolution sensitivity, failure-mode diagnostics, and runtime analysis.

The novelty of this work lies in integrating semantic feature typing, heterogeneous relation discovery, feature-family construction, mechanism-aware evidence modeling, event-background decomposition, conservative allocation, and role-based attribution into a unified multi-driver attribution framework. Relative to standard unsupervised NILM, the method does not primarily cluster appliance states or events, but instead shifts the analytical unit to relation-supported feature families constructed from contextual and operational variables. Compared with EV-specific NILM, the framework addresses multi-driver attribution without optimizing directly for a predefined target component. Compared with earlier MD-ADD formulations focused mainly on contextual decomposition using

weather and temporal drivers [36], the present framework expands the methodology while applying the conservative attribution philosophy to a different EV-focused setting.

The paper makes the following main contributions:

- A relation-aware multi-driver pipeline that transforms heterogeneous contextual and operational variables into feature-family evidence structures for low-frequency aggregate-load attribution.
- A conservative allocation strategy that assigns load variation only to evidence-supported feature families while preserving unsupported variation as unexplained demand.
- A mechanism-aware attribution framework that integrates heterogeneous relation discovery, event-background decomposition, and role-based within-family attribution into a unified attribution process.
- An empirical evaluation combining a primary EV-focused case study with measured-building validation, comparative evaluation, ablation analysis, sensitivity analysis, failure-mode diagnostics, and runtime analysis.

These contributions clarify both the practical value of the method for dominant event-driven load attribution and its limitations for smoother, lower-magnitude, or weakly observed components.

The remainder of the paper is organized as follows. Section 2 reviews related work on low-frequency NILM, EV load disaggregation, contextual energy modeling, and relation-based feature representation. Section 3 presents the proposed pipeline in detail. Section 4 describes the datasets, experimental setup, evaluation protocols, and implementation characteristics. Section 5 reports the primary case-study results, measured-building validation, comparative analyses, ablation and sensitivity results, failure-mode diagnostics, and runtime analysis. Section 6 discusses the findings, limitations, scalability, deployment implications, and future research directions. Section 7 concludes the paper.

2. Background and Positioning

2.1. Low-Frequency NILM and EV Load Disaggregation

NILM aims to estimate appliance-level or end-use electricity consumption from aggregate measurements collected at a limited number of metering points. In contrast to intrusive monitoring, which relies on submeters or dedicated appliance sensors, NILM seeks to infer disaggregated consumption from measurements already available at aggregation points, such as household or building electricity meters [3,36]. This makes NILM relevant for energy feedback, demand-side management, operational diagnostics, and energy-efficiency applications without requiring extensive submetering infrastructure [4,37].

NILM methods include optimization-based, probabilistic, graph-based, clustering-based, and neural-network approaches [3]. Their suitability depends strongly on the target load, sampling frequency, available contextual information, and assumptions about appliance operation. Temporal resolution is especially important. High-frequency measurements can preserve transient signatures and switching events, whereas low-frequency measurements mainly capture slower changes in active or reactive power. At lower sampling rates, overlapping operations, multi-state devices, and reduced signal detail make event detection and appliance identification more difficult [9,20]. Classical low-frequency approaches often formulate disaggregation as event detection, state identification, or optimization. Integer-programming-based approaches, for example, infer appliance states from step-like aggregate-load representations [32]. Such methods are useful when operating states are identifiable, but become less suitable when aggregate variation reflects overlapping processes, weak signatures, or partially observed drivers. Table 1 summarizes the main NILM approach types discussed in this section and highlights their relevance and limitations for low-frequency attribution.

Table 1. Overview of NILM approaches discussed in relation to low-frequency disaggregation.

Approach	Main Idea in the Reviewed Literature	Relevance to Low-Frequency NILM	Limitation for the Present Attribution Problem
Event- and state-based NILM	Detect load changes and infer appliance operating states from aggregate measurements	Useful when appliance events or stable operating states remain identifiable	Less suitable when multiple processes overlap or when load variation reflects gradual, weak, or partially observed drivers
Optimization-based NILM	Infer appliance states by fitting step-like aggregate-load representations	Can be useful when appliance states and constraints are well defined	Depends on identifiable states and strong assumptions about the observed system
Mean-shift and clustering-based unsupervised NILM	Detect events and group them into appliance profiles without requiring labeled appliance data	Supports unsupervised structure discovery in low-frequency aggregate signals	Usually produces appliance-state groupings rather than explanations of contextual driver contributions
Unsupervised time-series pattern extraction	Identifies subsequences, state changes, events, and recurring structures when labels are unavailable	Useful for finding structure in unlabeled aggregate data	Extracted patterns are not necessarily interpretable as driver-supported load contributions
Dimensionality-reduction approaches	Extract dominant variance structures or reduced representations from consumption curves	Can identify recurring low-frequency structure in aggregate or appliance-level data	Reduced components may be difficult to interpret as relation-supported drivers
EV-focused NILM	Detects or reconstructs EV charging from smart-meter, AMI, or feeder-level measurements	Relevant because EV charging is large, temporally concentrated, and sometimes visible at low frequency	Usually targets EV detection or EV-load reconstruction rather than multi-driver attribution under partial observability
Relation-aware feature-family attribution	Groups contextual and operational variables into driver families and allocates load only when relation-supported evidence exists	Designed for low-frequency aggregate data with heterogeneous drivers and incomplete observability	Cannot recover components that lack sufficient representation in the available feature set

Unsupervised NILM addresses the limited availability of labeled appliance-level data. Mean-shift-based residential disaggregation, for example, detects events in low-frequency aggregate signals and groups them through non-parametric clustering, allowing on/off and multi-state appliance profiles to be inferred without prior knowledge of the number of appliances [36]. More general unsupervised time-series pattern extraction methods identify subsequences, state changes, events, and recurring structures when labeled training data are unavailable or expensive to obtain [38]. Dimensionality-reduction methods provide another unsupervised route by extracting dominant variance structures from appliance consumption curves, including validation on low-frequency REDD measurements [37]. These approaches show that useful structure can be extracted from aggregate data, but they typically produce appliance-state groupings or reduced representations rather than explicit explanations of how observed contextual drivers contribute to aggregate demand.

Electric vehicles add a particularly important low-frequency disaggregation challenge. EV charging introduces large, temporally concentrated, and flexible electricity loads shaped by user behavior, charging strategy, electricity price, and local grid constraints [10]. Pre-

vious work shows that high-power EV charging may remain partially identifiable from low-frequency smart-meter data, although robust reconstruction across households and operating conditions remains difficult [8,9]. EV disaggregation has been proposed for customer feedback and energy-efficiency services [11], as well as for forecasting, distribution planning, and EV-integration analysis using advanced metering infrastructure data [12]. Feeder-level EV detection further extends the problem to aggregated distribution-network measurements, where individual charging events are embedded within larger and more variable background loads [13]. Other EV-focused approaches include signal-separation methods, supervised learning, and deep-learning formulations [39,40]. These studies demonstrate the operational value of EV NILM, but most are designed to detect or reconstruct a specific EV component rather than to attribute aggregate demand across several interacting and partially observed drivers.

2.2. Contextual Drivers and Partial Observability

Many electricity loads are influenced by contextual and operational variables rather than by appliance signatures alone. This is particularly relevant in low-frequency settings, where detailed transient information is unavailable. Weather, calendar effects, electricity price, emissions intensity, control states, and operating conditions can all influence aggregate demand. In energy forecasting, contextual variables are commonly used to improve predictive accuracy. Weather-integrated energy modeling has combined meteorological time series, historical consumption, frequent pattern mining, machine-learning models, and LSTM-based time-series modeling to identify and forecast weather-related energy-use patterns [23]. Building-load forecasting studies have similarly used exogenous variables such as ambient temperature, wind speed, humidity, precipitation, and operating days to represent seasonality and external effects [22]. Contextual drivers are especially important for heating, cooling, and other weather-sensitive loads. Low-frequency temperature-dependent load disaggregation is challenging because the relevant component may vary gradually with weather conditions rather than through sharp switching events [14]. HVAC modeling studies show that meteorological variables and historical load behavior contain information about air-conditioning operation and control [15]. Heating-load analysis has further shown that building energy demand can exhibit distinct seasonal profiles, peak characteristics, and weather-sensitive behavior [41]. Large-scale benchmarking studies have used time-series decomposition, clustering, and change-point regression to derive heating and cooling sensitivity, balance temperatures, and base-load indicators from aggregate energy data [42]. These studies support the need to distinguish persistent baseline demand, weather-sensitive dynamic behavior, and residual variation when detailed end-use metering is unavailable.

Contextual modeling is also relevant beyond weather-sensitive loads. EV charging may be influenced by user behavior, electricity prices, charging strategies, grid constraints, and environmental conditions [10,21]. Multi-source load-forecasting studies have integrated weather, price, and historical load variables, combined temporal modeling with feature-interaction analysis, and used SHAP values to identify influential drivers such as weather patterns and pricing policies [24]. However, most contextual energy-modeling studies use these variables primarily as predictors of aggregate demand. The present work instead treats contextual and operational variables as candidate drivers that can provide relation-supported evidence for attribution within an aggregate-load decomposition framework.

Partial observability remains a central challenge in this setting. In practical low-frequency energy datasets, the observed feature set rarely contains all drivers influencing the target load. Some loads may be unmeasured, weakly represented, or visible only through indirect contextual signals. A disaggregation method that forces the full aggregate

signal into observed components can create unsupported allocations. This motivates formulations that can assign load only when sufficient evidence exists, while retaining unexplained variation as an explicit output.

2.3. Relation Discovery and Feature-Family Representation

Energy datasets contain heterogeneous variables, including continuous sensor signals, binary control states, sparse event indicators, cumulative counters, categorical variables, and time-derived indices. Treating all variables as generic numerical predictors can produce inappropriate relation tests and misleading driver evidence. Semantic data type detection addresses part of this challenge by inferring column roles from statistical, structural, and distributional properties rather than relying only on metadata [28]. More recent approaches emphasize combining structural, statistical, and pattern-based evidence to improve robustness in heterogeneous datasets [29]. For relation-aware disaggregation, such typing is useful because the appropriate dependency test depends on the analytical role and structure of each variable.

A second requirement is the ability to detect multiple forms of dependence. Correlation-based measures are useful for linear or monotonic relationships, but energy systems often contain threshold effects, activation behavior, delayed responses, nonlinear patterns, cyclic modulation, and regime-dependent changes. General dependence-analysis frameworks such as the maximal information coefficient and the broader MINE framework were developed to identify diverse associations in high-dimensional data [43]. Comparative studies show that no single dependence measure is sufficient, because Pearson correlation, rank correlation, mutual information, distance correlation, and related statistics capture different aspects of dependence and differ in sensitivity across relationship types [44]. This supports the use of multi-mechanism relation discovery rather than reliance on a single association score.

Clustering and graph-based representation are also relevant because energy behavior is often structured by groups of related variables rather than isolated predictors. In NILM, spectral clustering and graph-based methods have been used to group load states or signal structures when direct appliance labels are unavailable [26,27]. Voting-based consensus formulations combine multiple clustering outputs and window configurations to improve robustness and reduce dependence on a single clustering configuration [26]. Outside strict NILM applications, shape-based clustering has been used to group load curves according to temporal morphology, while work on imbalanced clustering highlights the difficulty of representing rare but important load behaviors [45–47]. These ideas support the construction of feature families, where related contextual and operational variables are grouped into interpretable driver structures before attribution.

Multivariate energy analytics further motivates explicit modeling of relationships among variables. Graph attention models for energy-consumption anomaly detection, for example, represent sensor streams as interdependent time series and use graph-based attention mechanisms to model relationships among variables [25]. Although such work is not designed for load disaggregation, it reinforces the broader point that energy systems are multivariate and relational. Feature-selection studies further highlight the importance of evaluating relevance, redundancy, and stability before constructing explanatory models, especially in multivariate time-series settings [48]. For the present study, this motivates a pipeline that first characterizes feature types, detects heterogeneous relations, and groups related variables into feature families that can act as candidate driver structures.

2.4. Research Gap and Positioning of the Present Work

The reviewed literature shows that NILM has developed into a diverse field, including neural networks, optimization-based methods, Hidden Markov Models, graph signal processing, clustering, dimensionality reduction, and event-based approaches [3]. However, low-frequency settings remain difficult because detailed appliance signatures are limited and overlapping processes can obscure individual components [9,20]. Existing low-frequency approaches can detect events, infer states, or extract recurring structures when the aggregate signal contains sufficiently informative patterns [32,36,38]. In more heterogeneous settings, however, aggregate demand may reflect several interacting processes, contextual influences, and unobserved drivers. This limits methods that depend on complete component observability or clearly identifiable appliance events.

EV-focused NILM demonstrates that large flexible loads may remain detectable from smart-meter data [8,9]. Prior work has shown the value of EV disaggregation for customer feedback, forecasting, distribution planning, and feeder-level detection [11–13]. However, these approaches usually target EV detection or EV-load reconstruction specifically. They do not generally address the broader question of how aggregate load can be decomposed into multiple relation-supported driver families when only contextual and operational variables are available.

Contextual energy modeling provides another important foundation. Weather, calendar, and operating variables can improve the representation of aggregate demand [22,23], while multi-source forecasting studies have also incorporated price and historical load variables with interpretability mechanisms such as SHAP [24]. Nevertheless, these studies typically use contextual variables as predictors rather than as driver evidence for structured load attribution. Similarly, clustering and graph-based NILM methods demonstrate the value of grouping and consensus representations, but they usually group observations, states, windows, or appliance patterns rather than constructing feature families from heterogeneous contextual relationships [26].

The methodological predecessor to the present work is the multi-driver automatic dependency disaggregation approach for smart buildings [49]. That formulation treated weather and calendar variables as explanatory drivers, separated baseline demand, and retained unexplained energy explicitly instead of forcing all load variation into modeled drivers. The present work extends this direction through semantic feature typing, heterogeneous relation discovery, feature-family construction, relation-aware evidence modeling, mechanism-aware allocation, event-background decomposition, and role-based within-family attribution. It applies the framework to an EV-focused setting containing overlapping charging, heating, base-load, environmental, temporal, and market-related contextual drivers.

The present paper formulates low-frequency disaggregation as a relation-aware dependency decomposition problem under partial observability. Rather than reconstructing predefined appliance channels, the proposed pipeline identifies feature families, derives relation-supported evidence signals, allocates only the explainable portion of the aggregate load, and preserves unsupported variation as residual demand. This positions the method between conventional NILM, contextual energy modeling, and multivariate relation discovery, with the specific aim of supporting interpretable attribution when component labels are unavailable during model construction.

3. A Relation-Aware Multi-Driver Disaggregation Pipeline

3.1. Problem Formulation and Pipeline Overview

The proposed method decomposes an observed aggregate load signal into interpretable components representing persistent baseline demand, relation-supported dynamic

variation, unexplained dynamic load, and residual mismatch. The method is designed for low-frequency settings in which only part of the relevant system behavior is observable. The available features may include sensor measurements, control states, environmental variables, calendar variables, market signals, or derived quantities, and they are treated as incomplete evidence for the processes influencing the aggregate load. The formulation supports incomplete explanations by preserving unsupported target variation as unexplained dynamic load or residual mismatch.

The present pipeline extends the earlier MD-ADD principle that contextual drivers provide incomplete evidence and that unsupported load should remain explicit. The earlier formulation estimated direct contributions from manually categorized weather and calendar variables using out-of-fold predictive modeling and SHAP-based attribution. The present formulation expands the representation and allocation process through automatic feature typing, multiple global and local dependency mechanisms, relation-graph-based feature families, family mechanism roles, validity and explainability control, event-background separation, and role-based feature attribution. The methodological progression concerns the full representation and allocation process rather than a change in model parameters.

Let Y_t denote the observed target load at time t, and let $\mathrm{X_t} = (X_{1,t}, \ldots, X_{N,t})$ denote the observed feature set. In this formulation, the target refers to the aggregate load signal being analyzed, such as main-meter electricity demand. The target load is separated into a baseline component and a dynamic component:

$$\mathrm{Y_t} = \mathrm{B_t} + \mathrm{D_t}, \tag{1}$$

where B_t is the baseline load and D_t is the dynamic load above the baseline. The baseline represents persistent or slowly varying demand that is not treated as directly responsive to the observed driver evidence. The dynamic component is the portion of the target load considered for relation-supported attribution.

The dynamic load is then decomposed into feature-family contributions, unexplained dynamic load, and residual mismatch:

$$\mathrm{D_t} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{C_{k,t}} + \mathrm{U_t} + \mathrm{Res_t} \tag{2}$$

where $C_{k,t}$ is the contribution of feature family k, U_t is the unexplained dynamic load, and Res_t is the remaining reconstruction mismatch. Feature-family contributions are not predefined appliance channels. They are inferred from groups of observed variables that provide relation-supported evidence with the aggregate target. Their physical interpretation is established after decomposition by inspecting the family structure, mechanism roles, and, where available, external reference components used only for post hoc evaluation.

The inferred contributions represent relation-supported attribution within the observed feature space. Their interpretation is evidentiary rather than causal: a family may contain variables that describe the timing, operating context, or statistical signature of a load process without representing its physical cause. Physical interpretation requires inspection of the family structure and, where available, comparison with independent component measurements or operational knowledge.

The unexplained component U_t captures dynamic load that cannot be attributed to the observed features with sufficient support. This may occur because relevant drivers are missing, relationships are weak, processes overlap, or the selected relation mechanisms do not capture the underlying behavior. The residual Res_t captures any remaining mismatch after combining the baseline, feature-family contributions, and unexplained dynamic load.

This establishes a conservative decomposition framework in which attribution is supported by observed evidence, while incomplete explanations remain explicit. Table 2 summarizes the main notation used in the proposed pipeline.

Table 2. Main notation used in the proposed multi-driver disaggregation pipeline. Time, family, role, and feature indices are omitted from some descriptions where they are already explicit in the notation.

Symbol	Description	Symbol	Description
t	Time index.	Y_t^{bg}	Smooth dynamic-load background.
Y_t	Observed target load.	Y_t^{event}	Event-like dynamic load.
X_t	Observed input features.	$R_{k,t}^{\mathrm{event}}$	Event responsibility of family k.
B_t	Baseline load.	$A_{k,t}$	Event-active support of family k.
D_t	Above-baseline dynamic load.	$Co_{k,t}$	Event compatibility term.
k	Feature-family index.	$C_{k,t}^{\mathrm{event}}$	Event contribution of family k.
$C_{k,t}$	Contribution of family k.	β	Event-allocation sharpness.
U_t	Unexplained dynamic load.	$C_{k,t}^{\mathrm{bg}}$	Background contribution of family k.
Res_t	Residual reconstruction mismatch.	$Ma_{k,t}$	Mechanism-aware allocation boost.
η^2	Explained-variance effect size.	p_E	Evidence exponent in allocation.
Q_q	q-quantile operator.	p_V	Validity exponent in allocation.
W_t	Rolling window centered at t.	$We_{k,t}$	Allocation weight of family k.
$E_{k,t}$	Evidence signal of family k.	$As_{k,t}$	Allocation share of family k.
$G_{k,t}$	Gate component.	τ	Family-allocation sharpness.
$I_{k,t}$	Intensity component.	r	Within-family role.
$Mo_{k,t}$	Modulation factor.	$\rho_{k,r,t}$	Budget share for role r.
$L_{k,t}$	Locality factor.	i	Feature index.
γ_k	Family confidence score.	$We_{i,r,t}$	Feature weight within role r.
$V_{k,t}$	Validity score of family k.	Dom_t	Dominance of strongest family.
$\overline{P}_{k,t}$	Average gate/locality support.	$A_{i,r,t}$	Role-specific feature activity.
λ	Confidence-effect parameter.	I_i	Target-importance score.
$V_{\min}$	Minimum validity floor.	p_r	Role-activity exponent.
Z_t	Validity-weighted support.	p_I	Target-importance exponent.
S_t	Explainability score.	$Rs_{i,r,t}$	Within-role feature share.
c	Explainability scale parameter.	$F_{i,t}$	Final feature contribution.
M_t	Explainable dynamic-load magnitude.		

Figure 1 summarizes the full pipeline. The first part constructs an interpretable representation of the input data. Feature typing assigns semantic roles to the observed variables, pairwise relation discovery identifies multiple dependency mechanisms, and multi-layer feature-family construction groups related variables into candidate driver structures. The target load is then reintroduced through target-centric relevance analysis, which identifies the features and families that provide evidence for explaining the aggregate load.



Figure 1. Overview of the proposed relation-aware multi-driver pipeline, showing the transformation from heterogeneous input features and aggregate load into feature families, relation-aware evidence signals, conservative family-level allocation, role-based feature attribution, and residual output.

The second part converts this representation into an attribution model. Target-relevant families are transformed into relation-aware evidence signals and organized into mechanism roles such as gate, intensity, modulation, and local support. Family validity and explainability scores determine how much of the dynamic load is sufficiently supported for attribution. A conservative allocation step distributes only this explainable portion across feature families, while unsupported dynamic variation is retained as unexplained load. Event-background separation improves the handling of short, dominant peaks, and role-based within-family attribution decomposes family-level contributions into feature-level outputs.

The overall design is data-agnostic because it does not rely on predefined appliance models, supervised component labels, or fixed domain-specific decomposition rules. Instead, the pipeline uses semantic feature typing, heterogeneous relation discovery, feature-family grouping, and evidence-constrained allocation to produce an interpretable decomposition under partial observability. The main outputs are baseline load, family-level contributions, feature-level attributions, unexplained dynamic load, and residual demand.

3.2. Feature Typing, Relation Discovery, and Family Construction

The first three stages of the pipeline construct the driver representation used for later attribution. They characterize the semantic structure of each input feature, identify multiple forms of dependency among variables, and group related variables into feature families. This representation is constructed before target-focused attribution so that the later allocation procedure is based on interpretable feature groups rather than isolated variables or generic numerical predictors.

Operational energy datasets contain heterogeneous variables whose numerical format does not necessarily reflect their analytical role. A column may represent a continuous sensor signal, a binary equipment state, a sparse event indicator, a cumulative counter, a cyclic calendar index, a timestamp, or a categorical operating mode. Treating all variables as ordinary continuous predictors would create inappropriate relation tests and could produce misleading driver evidence. The pipeline begins with semantic feature typing. Each input variable receives scores for candidate types, including continuous, bounded continuous, zero-inflated continuous, binary, sparse binary, categorical, ordinal, count, cyclical, angular, datetime, trend-like, and accumulative variables. These scores are computed from structural and statistical evidence, including numeric coverage, cardinality, concentration of repeated values, entropy, integer-likeness, range behavior, temporal smoothness, and monotonic or cyclic structure.

The selected feature type determines which relation tests are appropriate in the next stage. Accumulative variables are analyzed through increments rather than running totals. Zero-inflated and bounded variables are handled separately to avoid confusing inactive periods, measurement limits, and genuinely discrete states. This semantic representation provides the interface between heterogeneous input variables and relation-aware disaggregation. The final feature type is selected through a superseding decision rule, allowing more specific structures such as zero-inflated, bounded, cyclical, angular, trend-like, or accumulative behavior to replace a more general continuous interpretation when the evidence is sufficiently strong. The output of this stage consists of the selected semantic type, confidence information, and diagnostic metadata used by the later relation and evidence-construction stages.

After feature typing, the pipeline evaluates pairwise relationships between variables using dependency tests selected according to the semantic types of the candidate driver and response. Each relationship is evaluated for an ordered pair (a, b), where a is interpreted as the potential driver and b as the response. Before testing, the two series are temporally aligned, compatible variables are converted to numerical form where needed, missing values are removed, and the number of overlapping observations is checked against a minimum support requirement. Optional winsorization may be applied to reduce the effect of extreme values.

The relation discovery stage is designed to avoid the limitations of correlation-only screening. Operational energy data may contain linear trends, threshold effects, sparse activations, cyclic patterns, delayed responses, nonlinear behavior, and regime-dependent changes. The method evaluates several relation families, summarized in Table 3. Each valid test returns the driver, response, relation type, scalar score, primary metric, scope, and diagnostic information required for later evidence construction.

For threshold and band relations, scores are expressed through effect-size measures such as explained variance. Activation relations convert the response into an activation or upper-tail indicator and score the separation of response states across driver conditions. Cyclical relations use group-based or harmonic representations, lagged relations retain the strongest lagged association, and nonlinear or regime-based relations score the improvement obtained by allowing nonlinear, variance-dependent, or time-varying behavior.

Table 3. Relation families used for multi-mechanism dependency discovery.

Relation Family	Eligible Driver or Response Structure	Main Purpose	Primary Score
Linear and monotone	Numeric driver and numeric response	Detect linear or monotone co-variation	$s_{linear} = \max(\lvert r_{Pearson} \rvert, \lvert \rho_{Spearman} \rvert)$
Threshold mean shift	Numeric driver, numeric response	Detect response change across a driver split	η^2
Threshold activation	Numeric driver, activation-eligible response	Detect driver conditions associated with activation	Information gain, activation separation
Tail activation	Numeric driver, nonzero-baseline response	Detect driver conditions associated with high-load states	Tail-probability separation
Band mean shift	Numeric driver, numeric response	Detect elevated or reduced response within a driver interval	η^2
Band activation	Numeric driver, activation-eligible response	Detect activation inside or outside a driver interval	Activation separation
Cyclical modulation	Cyclical or angular driver	Detect periodic response structure	ANOVA effect size or R^2
Multi-level shift	Categorical, ordinal, count, or binary driver	Detect response differences across states or levels	Effect size
Lagged relation	Numeric driver and response	Detect delayed driver-response association	Best lagged R^2
Variance modulation	Numeric driver and response	Detect changes in response variance conditional on driver	Variance-model R^2
Nonlinear smooth relation	Numeric driver and response	Detect quadratic or smooth nonlinear uplift	Incremental R^2
Regime switching	Time-indexed numeric pair	Detect temporal change in mean or slope relation	Improvement over non-regime model

Relation eligibility is constrained by the feature types identified in the previous stage. Linear, nonlinear, threshold, band, lagged, and variance-modulation tests require compatible numeric variables. Cyclical modulation requires cyclic or angular driver semantics. Multi-level relations require a group-like driver. Activation-oriented relations are evaluated only when the response exhibits binary, sparse, zero-inflated, or upper-tail behavior. Constant and datetime variables are excluded from ordinary pairwise relation testing unless they are transformed into analytically meaningful derived variables.

The relation discovery stage can be applied both globally and locally. Global tests evaluate dependencies over the full observation period and identify persistent relationships. Local tests apply the same relation logic over rolling windows and summarize the resulting scores across time. This allows the pipeline to retain dependencies that are intermittent, regime-specific, seasonally active, or masked by stronger full-period effects.

The resulting long-format relation table is used to construct feature families. A feature family is defined as a set of variables repeatedly connected through one or more dependency mechanisms. This step is necessary because operational drivers are rarely represented by

isolated variables. They often appear as groups of related sensors, counters, control signals, calendar variables, or transformed indicators that jointly describe one underlying process.

The relation mechanisms were selected to cover complementary dependency structures expected in low-frequency energy data, including continuous linear or monotonic behavior, nonlinear variation, delayed responses, threshold effects, sparse activation, regime changes, cyclic modulation, and variance-dependent behavior. Each mechanism produces the same standardized relation-record structure, allowing additional relation tests to be incorporated without changing the subsequent family-construction, evidence-modeling, or allocation stages. The mechanism set is extensible rather than fixed to the tests used in the present implementation.

For family construction, relation scores are converted into a common edge-evidence scale in the interval [0, 1]. Correlation-based relations are mapped using absolute association strength, bounded effect-size relations are retained on their native scale, and unbounded activation-strength measures, such as activation risk ratios, are transformed into bounded evidence through an exponential squashing function. A separate graph layer is then constructed for each relation type. Nodes represent features, and weighted edges represent normalized relation evidence between feature pairs. Ordered relations are collapsed into undirected feature-pair evidence by retaining the strongest evidence observed across directions or duplicate rows.

Each relation layer is clustered independently using weighted community detection. If a relation layer contains no retained edges, variables remain as singleton groups for that layer. Final feature families are obtained through consensus grouping across relation layers. For each pair of features, the method computes how often the two features are assigned to the same group across relation layers. These co-assignment values form a consensus matrix, and feature pairs exceeding the required agreement threshold are connected in a consensus graph. Clustering this graph yields the final feature families.

The target variable is excluded from feature-family construction. This prevents the aggregate load from dominating the grouping process and forcing unrelated drivers into the same family merely because they are all associated with the target. The target is reintroduced only in the next stage, where feature and family relevance to the aggregate load are evaluated. For each final family, a representative feature is also identified based on internal edge evidence. This representative is not assumed to be causally dominant, but provides a compact descriptor of the family and supports later interpretation.

The output of this stage consists of semantic feature types, a multi-mechanism relation table, feature-family memberships, representative features, retained relation edges, and the consensus co-assignment matrix. These outputs provide the structured driver representation used by the later attribution stages.

3.3. Target Relevance and Evidence Signal Construction

After feature families have been constructed, the target signal is reintroduced to determine which features and families provide evidence for explaining the aggregate load. This target-centric stage is separated from feature-family construction so that the aggregate load does not dominate the grouping process. Feature-family construction identifies internally coherent driver structures, while target relevance analysis determines which of these structures are useful for attribution.

The analysis starts by extracting relation rows in which a candidate feature acts as the driver and the aggregate load acts as the response. Both global relations and window-summary relations are retained. Global relations describe dependencies that persist over the full observation period, while windowed relations capture dependencies that are local, intermittent, seasonal, or regime-specific. For each feature, the target-facing relation rows

are collapsed into a feature-level summary containing the strongest relation to the target, the corresponding score, the relation type, the qualitative strength label, and summary statistics over valid relation scores. When valid global scores are available, the strongest global relation is used as the primary descriptor. If no valid global score is available, the strongest available score across retained scopes is used.

An overall target-importance score is assigned to each feature by combining the strongest observed target-relation score with the mean relation score across valid target-facing relations, with greater weight placed on the strongest relation. The maximum score captures whether a feature has at least one strong and interpretable dependency with the target, while the mean score rewards repeated support across relation types or time windows. Windowed relation summaries are retained separately through their mean score, maximum score, variability, and valid-window fraction. These quantities indicate whether a feature is consistently related to the target or only relevant under specific operating conditions.

Feature-level target summaries are merged with the feature-family membership table. Family-level relevance is computed by aggregating target-importance scores within each family. For each family, the method records the maximum feature importance, mean feature importance, and strongest target-facing feature. The maximum importance identifies whether the family contains at least one strongly target-relevant feature, while the mean importance indicates whether relevance is distributed across multiple family members. The resulting feature and family summaries guide the selection of families entering the attribution stage and identify which features provide evidence within each family. These scores are not interpreted as causal effects or component labels. They indicate relation-supported target dependency.

Before family evidence is constructed, the target load is separated into a baseline component and a dynamic component. The baseline B_t represents persistent or slowly varying demand that is not treated as directly responsive to the observed driver evidence. The pipeline supports a zero baseline, a constant-minimum baseline, and a rolling-quantile baseline. For the rolling-quantile case, the baseline is defined as

$$B_t = Q_q(Y_\tau : \tau \in W_t), \tag{3}$$

where Q_q denotes the q-quantile and W_t is the rolling window centered around time t. The dynamic load available for attribution is defined as the non-negative above-baseline component:

$$D_t = \max(0, Y_t - B_t). \tag{4}$$

This definition bounds the subsequent attribution problem. Driver families can only claim dynamic load above the selected baseline, and any dynamic load that is not supported by valid family evidence remains available as unexplained load. When the zero-baseline option is used, the same formulation applies to the full aggregate target signal.

Target-relevant families are converted into time-varying evidence signals. For each family, candidate evidence features are selected from the target-centric relation summary. Each candidate is associated with a relation type, target-relation score, target-importance value, and, when available, local-window support. The relation type determines how the feature contributes to the family evidence signal. As summarized in Table 4, threshold, tail-threshold, activation-threshold, and band relations support gate evidence; linear, monotone, nonlinear, lagged, regime-switching, and multi-level relations support intensity evidence; cyclical and state-like relations support modulation evidence; and strong target-facing features that do not match a preferred role may be retained as fallback or local helper evidence.

Table 4. Mechanism roles used for family evidence construction.

Relation Evidence	Mechanism Role	Function in Evidence Construction
Threshold, tail-threshold, activation-threshold, and band relations	Gate	Identifies periods in which a relation-supported condition is active
Linear, monotone, nonlinear, lagged, regime-switching, and multi-level relations	Intensity	Provides a non-negative magnitude signal for the family
Cyclical, angular, calendar, and state-like relations	Modulator	Adjusts evidence according to periodic or contextual conditions
Window-local relation support and augmented target-facing evidence	Local helper	Increases support during periods with stronger local dependency
Strong target-facing feature without a complete role structure	Fallback	Provides evidence when no complete gate or intensity structure is available

When comparable evidence is available, raw physical variables are preferred for intensity roles because they are easier to interpret as magnitude-related drivers. Augmented or locally transformed features are primarily used as local helper evidence when they provide strong target-facing support.

Evidence construction is relation-native. Threshold and band relations are converted into gate signals that are active under driver conditions associated with higher dynamic load. When relation-specific thresholds or band limits are available, these values are used directly; otherwise, data-driven fallback thresholds or intervals are derived from the feature distribution. Gating evidence is represented as positive uplift in D_t during active periods relative to the inactive reference level, rather than only as a binary mask.

Continuous relations are transformed into non-negative intensity signals that represent the time-varying strength of family evidence. Lagged relations use the lag-adjusted feature representation, while regime-switching and multi-level relations use regime- or state-specific representations. Cyclical and state-like features are transformed by estimating the typical dynamic load associated with each state, cycle position, or discretized bin, and assigning positive evidence only to states associated with above-reference dynamic load. This allows calendar-like and state-like variables to contribute without assuming a linear numerical scale.

For each family, the selected evidence features are organized into a mechanism structure. Gate features indicate when the family is active. Intensity features describe the primary magnitude-related evidence. Modulators adjust the strength of the evidence under contextual or cyclic conditions. Local helpers incorporate time-local support when the family-target relation is stronger in specific windows. Families are also assigned coarse mechanism types, such as gated, continuous, cyclical, or regime-based, according to their primary relation evidence. These mechanism types guide later allocation behavior, especially in the event-background separation step. A fallback feature is retained when no complete role structure can be formed. The family confidence score γ_k summarizes the strength of the selected target-facing evidence and the available local support.

The resulting family evidence signal is defined as

$$\mathrm{E_{k,t}} = \mathrm{G_{k,t}} \cdot \mathrm{I_{k,t}} \cdot \mathrm{Mo_{k,t}} \cdot \mathrm{L_{k,t}} \cdot \gamma_\mathrm{k} \tag{5}$$

where $G_{k,t}$ is the gate component, $I_{k,t}$ is the intensity component, $Mo_{k,t}$ is the modulation factor, $L_{k,t}$ is the locality factor, and γ_k is the family confidence score. If no gate is available, the gate defaults to one, meaning that the family is not restricted to specific activation periods. If no locality information is available, the locality factor defaults to one. Modu-

lation and locality are bounded to prevent excessive amplification. All evidence signals are constrained to be non-negative, so they represent support for allocating dynamic load rather than signed prediction error.

The output of this stage is a set of target-facing feature summaries, family relevance summaries, family mechanism descriptions, and one evidence time series per selected family. These evidence signals are not final load contributions. They represent relation-supported claims on the dynamic load and are passed to the subsequent validity, explainability, and allocation stages.

3.4. Validity, Explainability, and Conservative Allocation

The family evidence signals indicate where each family has relation-supported claims on the dynamic load, but the presence of evidence alone is not sufficient for allocation. The pipeline evaluates the validity of each family claim and determines how much of the dynamic load is explainable before distributing load across families. This distinction is central to the conservative design of the method, because it prevents weak, inactive, or locally inconsistent evidence from receiving unsupported load assignments.

For each family, a time-varying validity score is computed from the available mechanism components. Gate activity indicates whether the family is active under a relation-supported condition, while locality support indicates whether the family-target relation is stronger in the corresponding time window. These components are combined into an average support term rather than multiplied, so that one weak component does not fully suppress a family when other evidence remains plausible. The support term is modulated by the family confidence score derived during mechanism construction. The resulting validity score is defined as

$$V_{k,t} = \max\left(V_{\min}, \overline{P}_{k,t} \cdot \max\left(\gamma_k^{\lambda}, V_{\min}\right)\right), \tag{6}$$

where $V_{k,t}$ is the validity of family k at time t, $\overline{P}_{k,t}$ is the average of the available gate and locality support components, γ_k is the family confidence score, λ controls the strength of the confidence effect, and $V_{\min}$ is the validity floor. This formulation applies confidence as a soft adjustment rather than a hard exclusion rule.

The pipeline estimates how much relation-supported evidence is available at each time step. Family evidence signals are weighted by their validity scores and summed across all families:

$$Z_t = \sum_{k=1}^{K} E_{k,t} V_{k,t}, \tag{7}$$

where $E_{k,t}$ is the relation-aware evidence signal of family k. The summed support Z_t is transformed into a bounded explainability score,

$$S_t = \frac{Z_t}{Z_t + c} \tag{8}$$

where c is a scale parameter estimated from a high quantile of the positive support values. This transformation ensures that explainability increases smoothly with available evidence while remaining bounded between zero and one.

The explainability score determines the amount of dynamic load that is eligible for allocation. Let D_t be the dynamic load above the baseline. The explainable magnitude is defined as

$$M_t = \eta(0.15 + 0.85 S_t) D_t \tag{9}$$

where η is a configurable cap on the maximum explainable fraction. The small lower term prevents the explainable magnitude from collapsing abruptly to zero in periods with weak

but non-zero evidence, while the cap ensures that the method does not force all dynamic load into the observed families.

After M_t has been defined, allocation is performed only over the explainable portion of the dynamic load. For each family, a non-negative allocation weight is computed as

$$\mathrm{We}_{k,t} = \mathrm{E}_{k,t}^{p_E} \mathrm{V}_{k,t}^{p_V} \mathrm{Ma}_{k,t} \tag{10}$$

where p_E and p_V control the influence of evidence and validity, and $\mathrm{Ma}_{k,t}$ is a mechanism-aware allocation boost. If total active weight is below a minimum threshold at a given time step, no family allocation is made for that step. All weights are clipped to be non-negative.

To identify cases where one family clearly dominates the available evidence, dominance is computed as

$$Dom_t = \frac{\max_k \mathrm{We}_{k,t}}{\sum_j \mathrm{We}_{j,t} + \epsilon}. \tag{11}$$

If Dom_t exceeds a predefined threshold, the winning family receives a large fixed fraction of the allocation share, while the remaining share is distributed among other active families in proportion to their weights.

The allocation share of each family is obtained through normalized weighting:

$$\mathrm{As}_{k,t} = \frac{\mathrm{We}_{k,t}^{\tau}}{\sum_j \mathrm{We}_{j,t}^{\tau} + \epsilon}, \tag{12}$$

where $As_{k,t}$ is the allocation share and τ controls the sharpness of the allocation.

The family contribution is obtained by multiplying the allocation share by the explainable magnitude:

$$\mathrm{C}_{k,t} = \mathrm{As}_{k,t} \mathrm{M}_t. \tag{13}$$

Because M_t is bounded by the explainability-control stage, the sum of family contributions cannot exceed the amount of dynamic load judged explainable from the available evidence. Families compete only for the explainable magnitude, not for the full observed load.

The unexplained dynamic load is computed as the part of the dynamic load that is not included in the explainable magnitude:

$$\mathrm{U}_t = \max(0, \mathrm{D}_t - \mathrm{M}_t). \tag{14}$$

The final residual is evaluated relative to the original target signal after combining the baseline, family contributions, and unexplained dynamic load. Negative residuals are clipped to preserve the non-negative decomposition semantics. This allocation design produces three distinct outputs: evidence-supported family contributions, unexplained dynamic load, and residual mismatch. The separation prevents forced completeness and preserves the diagnostic value of unresolved demand.

3.5. Event-Background Separation

The conservative allocation stage controls how much dynamic load may be attributed to feature families, but it does not by itself distinguish smooth background variation from short event-like peaks. The pipeline applies an event-background separation step to improve the handling of dominant transient load events. This is particularly relevant when short, high-magnitude events could otherwise be diluted across continuous families or absorbed by smoother background allocation.

The event-background separation is applied to the dynamic load D_t. A smooth dynamic-load background is estimated using a centered rolling quantile:

$$Y_t^{bg} = Q_q(D_\tau : \tau \in W_t), \tag{15}$$

where W_t is a rolling time window and Q_q is a configurable quantile. The resulting background estimate is clipped to remain non-negative, and boundary values are filled to obtain a complete time series. The event-like component is defined as the positive deviation above this background:

$$Y_t^{event} = \max(0, D_t - Y_t^{bg}). \tag{16}$$

This formulation treats events as local increases above the smooth dynamic-load background rather than as all load above the original baseline. The event component captures sharper deviations, while the smoother part of the dynamic load remains available for ordinary background allocation.

Event responsibility is computed separately for each family. The responsibility score combines gate activity, relation-aware evidence, locality support, and family confidence. Gated families receive stronger event responsibility when they are active under a relation-supported condition and have corresponding evidence at the same time step. Continuous or regime-based families may receive reduced event responsibility, depending on the configuration, so that event-like load is preferentially assigned to families with event-compatible evidence. The event responsibility of family k is defined as

$$R_{k,t}^{event} = A_{k,t} \cdot Co_{k,t} \cdot (1 + \lambda \gamma_k), \tag{17}$$

where $A_{k,t}$ is event-active support derived from the gate and evidence signals, $Co_{k,t}$ is an event compatibility term based on gate, evidence, and locality support, γ_k is the family confidence score, and λ controls the influence of confidence. Very small responsibility values are suppressed to avoid numerical noise.

The event contribution of family k is obtained by normalizing family event responsibilities across all active families:

$$C_{k,t}^{event} = Y_t^{event} \frac{\left(R_{k,t}^{event}\right)^\beta}{\sum_j \left(R_{j,t}^{event}\right)^\beta + \epsilon}, \tag{18}$$

where β controls the sharpness of event allocation and ϵ prevents division by zero. If no family has event responsibility at a given time step, no event load is assigned.

The event contribution is combined with the background family contribution obtained from the ordinary allocation stage:

$$C_{k,t} = C_{k,t}^{bg} + C_{k,t}^{event}. \tag{19}$$

After the background and event components are combined, the row-wise sum of all family contributions is capped by the explainable magnitude M_t. If the combined allocation exceeds M_t, all family contributions are scaled down proportionally. This preserves the conservative allocation constraint while allowing event-compatible families to capture short, dominant peaks more specifically.

3.6. Role-Based Attribution and Outputs

After family-level contributions have been estimated, the pipeline decomposes each family contribution into feature-level outputs. This within-family attribution step does not

change the amount of load assigned to each family. Instead, it improves interpretability by distributing family contributions across the features that define each family mechanism.

The attribution uses the mechanism roles introduced in Table 4. For each family, features may be assigned to gate, intensity, modulator, local helper, or fallback roles. Gate features identify activation conditions, intensity features describe the main magnitude-related evidence, modulators adjust the response under contextual or cyclic conditions, and local helpers provide time-local support. A fallback feature is used only when no complete role-based structure is available.

Each role receives a predefined budget share of the family contribution. In the implemented configuration, most of the contribution is assigned to the intensity role, while smaller shares are assigned to gates, modulators, and local helpers. If a family lacks features for a given role, that role's budget is redistributed across the available roles, primarily to the intensity role when present. Let $C_{k,t}$ denote the contribution of family k at time t. The role-level contribution is defined as

$$\mathrm{C_{k,r,t}} = \rho_{\mathrm{k,r,t}} \mathrm{C_{k,t}}, \tag{20}$$

where r denotes the role and $\rho_{k,r,t}$ is the normalized budget assigned to that role.

Within each role, the role contribution is distributed dynamically across the features assigned to that role. The feature share depends on two terms: the feature's time-varying role activity and its target-importance score. Gate activity is obtained from the family gate component when available. Intensity, modulator, and local helper activity are computed from the corresponding relation-native feature signal and normalized to the interval $[0, 1]$. For feature i assigned to role r in family k, the within-role weight is

$$\mathrm{We_{i,r,t}} = \mathrm{A}_{\mathrm{i,r,t}}^{\mathrm{p_r}} (\mathrm{I_i} + \epsilon)^{\mathrm{p_I}}, \tag{21}$$

where $A_{i,r,t}$ is the role-specific activity of the feature, I_i is its target-importance score, p_r controls the activity sharpness for the role, p_I controls the influence of target importance, and ϵ prevents zero weights. The within-role feature share is obtained by normalizing these weights across all features assigned to the same role:

$$\mathrm{Rs_{i,r,t}} = \frac{\mathrm{We_{i,r,t}}}{\sum_{\mathrm{j \in r}} \mathrm{We_{j,r,t}} + \epsilon}. \tag{22}$$

If all weights within a role are zero at a given time step, the method falls back to a uniform split among the features assigned to that role. The final feature-level contribution is computed as

$$\mathrm{F_{i,t}} = \sum_{\mathrm{k,r}} \mathrm{C_{k,t}} \rho_{\mathrm{k,r,t}} \mathrm{Rs_{i,r,t}}. \tag{23}$$

This formulation preserves the family-level energy balance while providing feature-level attribution consistent with the mechanism structure used during allocation. The summation allows a feature to receive contributions through more than one family or role when such overlap exists. Gate features are credited for activation conditions, intensity features receive the main magnitude-related share, and modulator or local helper features receive contributions according to their contextual role. The pipeline stores diagnostic outputs, including role-budget shares and the mean, maximum, and active fraction of each feature's within-role share. These outputs support inspection of whether a family contribution is dominated by one feature, distributed across several mechanism components, or driven mainly by fallback behavior.

The final pipeline outputs are the baseline load, dynamic load, family-level contributions, feature-level attributions, unexplained dynamic load, residual demand, and diagnostic summaries for feature types, relations, families, evidence signals, validity scores, and role-based attribution.

3.7. Computational Complexity and Implementation Characteristics

Let n denote the number of timestamps, p the number of original variables entering pairwise relation discovery, q the number of variables after local-context augmentation, w the number of local temporal windows, h the average number of observations per window, r the number of relation layers, p_r the number of features retained after target-centric pruning, f the number of feature families, m the number of retained relation records, and e the number of retained graph edges.

Relation discovery evaluates $p(p-1) = O(p^2)$ ordered variable pairs. For an aligned pair containing s observations, the tested mechanisms include rank-based association, lag scanning, threshold and band searches, and regime-dependent relations. A conservative per-pair bound is

$$C(s) = O(s\log s + (l + g_t + g_b^2 + g_r)s) \tag{24}$$

where l is the maximum tested lag and g_t, g_b, and g_r are the fixed threshold, band, and regime-search grid sizes. The global and local relation-discovery stage requires

$$O(p^2[C(n) + w(n + C(h))]) \tag{25}$$

The additional wn term reflects the current implementation, in which local windows are extracted through repeated scans of the complete timestamp index. With fixed-duration sliding windows, the number of windows increases with the observation period, giving a conservative implementation-level upper bound of:

$$O(p^2n^2) \tag{26}$$

Precomputing and reusing window boundaries would reduce this overhead and bring the relation-scoring stage closer to:

$$O(p^2 n\log n) \tag{27}$$

for fixed relation grids and lag limits.

Graph construction processes the compact relation table across r relation layers. Consensus construction examines retained feature pairs across the layers and requires

$$O(rp_r^2) \tag{28}$$

time and

$$O(p_r^2) \tag{29}$$

memory in the current dense implementation. Evidence construction, validity and explainability control, event allocation, family allocation, and within-family attribution operate over timestamp by family or timestamp by feature matrices. Since $f \leq p_r$, these stages require approximately

$$O(np_r) \tag{30}$$

time and space.

For the compact-output configuration used in the experiments, the overall space requirement can be summarized as

$$O\left(n(q + p_r) + p^2 r + p_r^2 + e\right) \tag{31}$$

This is a theoretical space bound because peak memory was not instrumented in the runtime experiment. The current implementation is batch-based and does not provide streaming or incremental updates. Global relations, local-window summaries, relation thresholds, graph layers, consensus communities, and allocation matrices are recomputed for each supplied dataset.

All computational analyses were implemented in custom Python code and executed using Python 3.10.11 (Python Software Foundation, Wilmington, DE, USA) in a notebook-based environment through Visual Studio Code 1.120. Data handling, numerical computation, statistical evaluation, machine-learning utilities, and visualization used NumPy 2.0.2, pandas 2.2.3, SciPy 1.15.1, scikit-learn 1.6.1, XGBoost 3.2.0, Matplotlib 3.10.0, IPython 8.32.0, and Permetrics 2.0.0. The custom implementation was used for feature typing, relation discovery, feature-family construction, evidence modeling, attribution, ablation analysis, sensitivity analysis, diagnostic analysis, Bayesian thermal comparison, and runtime analysis.

4. Experimental Setup

4.1. Dataset Origin and Case Context

The experimental evaluation uses an electric-vehicle charging dataset derived from a multi-agent simulation of decentralized EV home-charging strategies [7]. The original simulation model was developed for a Danish residential distribution-grid case in Strib, Denmark, serving 126 residential consumers, and evaluates how EV adoption and decentralized charging strategies affect future distribution-grid loading at hourly resolution. The detailed assumptions of the underlying simulation, including the distribution-grid scenario design and EV charging strategy, are provided in the original source [7]. The present study does not modify or recalibrate that simulation model. Instead, it uses one resulting household-level hourly time series as an evaluation dataset for the proposed attribution pipeline.

The unit of analysis in the present evaluation is a single household-level aggregate-load signal. The method attributes the selected target signal to the driver information represented in the available feature set; it is not designed to separate the consumption of individual households from a combined feeder signal unless household-specific driver information is included in the input data.

The primary EV case uses a simulated two-year analysis horizon from January 2025 to December 2026 at hourly resolution. These calendar dates are assigned timestamps within the simulation scenario and should not be interpreted as historical measurements collected during those calendar years. The year labels do not introduce year-specific real-world information; they only provide a chronological index from which calendar features such as month, day, hour, day of week, and weekend status can be derived. No supervised train/test split is used, because the proposed method is applied as an unsupervised attribution procedure rather than as a predictive model trained on component labels.

The single simulated household-level case provides a structurally rich test environment because the aggregate signal contains overlapping event-driven, weather-sensitive, schedule-related, and market-responsive behavior together with known post hoc component signals. This makes the case suitable for evaluating whether the proposed pipeline can identify heterogeneous dependency structures under partial observability.

The dataset contains an aggregate household-level load signal together with temporal, environmental, market-related, and operational contextual variables. The target variable is Total Load [kW]. In the simulation dataset, this aggregate load is composed of Base/inflexible load [kW], EV Charging Load [kW], and HP Heating Load [kW]. These

component signals are not used as model inputs and are retained only as post hoc reference signals. The dataset's Base/inflexible load [kW] component should be distinguished from the analytical baseline term used in the MD-ADD formulation.

In addition to the primary simulated EV case, measured-data validation was conducted using nine commercial and public buildings from the ADRENALIN project. The ADRENALIN dataset was selected as the measured validation dataset because it is directly aligned with the low-frequency, partial-observability, and contextual-attribution setting considered in this study. It was developed for low-frequency temperature-dependent load disaggregation and combines aggregate energy measurements, weather variables, and validated temperature-dependent reference loads. The dataset origin, building selection, preparation, and validation procedures are documented in detail in [50].

4.2. Dataset Variables and Preprocessing

The primary EV dataset contains hourly timestamp and calendar variables, load variables, market variables, and weather-related contextual variables. For the simulated January 2025 to December 2026 analysis horizon, the full hourly period corresponds to 17,520 time steps before any removal caused by preprocessing. The timestamp column is parsed as a datetime index, after which the data are sorted chronologically. Duplicate timestamps are checked, and when duplicates are present, the first occurrence is retained. The target column is converted to numeric format, and rows with missing target values are removed.

For the primary EV case, the three known component load columns are separated from the dataset and stored for evaluation before the pipeline is applied. The pipeline input excludes Base/inflexible load [kW], EV Charging Load [kW], and HP Heating Load [kW]. The remaining input contains the aggregate target load and contextual variables, including calendar variables, indoor temperature, spot price, CO_2 emission intensity, sunshine duration, outdoor air temperature, humidity, wind speed, precipitation duration, precipitation amount, and solar irradiance. This setup evaluates the method under partial observability because the algorithm receives the aggregate load and contextual drivers, while component loads are introduced only after decomposition for post hoc alignment analysis.

The additional ADRENALIN analysis uses hourly data from the nine retained buildings. For each building, aggregate load is used as the pipeline target, while the measured temperature-dependent load is excluded from the input and retained only for post hoc evaluation. Two feature sets are evaluated. The restricted feature set contains outdoor air temperature and hour of day. The full feature set contains all available meteorological and calendar variables, including outdoor air temperature, relative humidity, direct and diffuse solar radiation, wind speed, wind direction, year, month, day, hour of day, day of week, and a weekend indicator. The prepared building datasets are used without introducing additional component labels or end-use reference signals. Detailed descriptions of the energy-meter processing, weather-data alignment, quality control, and temperature-dependent submeter validation are provided in [50].

4.3. Feature Analysis and Family Construction Setup

The first stage applies automatic feature typing to the pipeline input data. Each variable is assigned a final semantic type and confidence score, which are used to guide relation discovery. Local-context augmented features are generated using rolling windows of 6, 24, and 72 time steps. These augmented variables are used to identify target-facing relationships that may not be visible in the raw feature values.

Pairwise relation discovery is performed using ordered feature pairs with a minimum overlap requirement of 200 observations. Activation relationships are scored using the

log-risk ratio formulation, and lagged relationships are searched up to 48 time steps. Both global relations and rolling-window summaries are computed. The compact relation table is used for feature-family construction.

Feature families are constructed through the multi-layer relation graph procedure described in Section 3.2. Dynamic relation thresholds are computed from the observed score distributions using the 80th percentile of global relation scores, with relation-specific minimum floors. Greedy modularity clustering is applied within relation layers, and a consensus grouping is constructed using a minimum agreement threshold of 0.5. The target load is excluded from grouping, and augmented features are excluded from the family-construction step so that families are formed from original interpretable variables.

After an initial target-centric analysis, features with an overall target importance below 0.20 are removed. This threshold defines the retained target-facing feature set used in the reported pipeline configuration, reducing the influence of weakly target-facing variables before the feature-family construction and MD-ADD allocation stages are repeated. The augmented target-facing relation analysis is merged back into the retained original feature summary so that each retained feature can use either its raw representation or its strongest augmented representation during evidence construction.

4.4. MD-ADD Configuration

For the primary EV case, the MD-ADD pipeline is applied to the augmented input data using Total Load [kW] as the target. In the reported EV experiment, the baseline mode is set to 'none', meaning that the decomposition is performed directly on the aggregate target load without subtracting an estimated analytical baseline. This choice reflects the distinction between the dataset's Base/inflexible load [kW] component and the MD-ADD baseline term. The former is a simulated component of the total load, while the latter is an optional baseline estimate derived from the aggregate target signal. A baseline-mode check showed that the constant-minimum baseline was only 0.02 kW and produced nearly identical EV alignment to the no-baseline configuration, with EV correlation changing from 0.990 to 0.991 and EV NMAE from 0.100 to 0.101. A rolling-quantile baseline was not used because it absorbed 17.6% of the target energy into a local baseline component, thereby changing the attribution objective. The main configuration parameters used in the reported experiment are summarized in Table 5. The pipeline settings are retained across datasets. Within each dataset, the restricted and full runs differ only in the supplied feature set. The following paragraphs highlight the settings most relevant to interpretation.

Role-based within-family attribution is enabled. The configured role budgets assign the largest share to intensity features, followed by modulators, while gate and local helper features receive smaller shares. Family validity and explainability control are enabled, with an explainability cap of 0.95 and a family-validity floor of 0.10. The allocation stage uses evidence and validity powers equal to one, a dominance threshold of 0.85, and a winner fraction of 0.95 for cases where one family clearly dominates the allocation evidence.

The event-background decomposition is enabled to handle sharp EV-related load peaks. The background component is estimated using a 24-step rolling window and a 0.20 quantile. Event responsibility is restricted to gated families, so event-like load is preferentially assigned to families with explicit activation evidence. The allocation excludes Year from relation-based allocation and penalizes calendar features to reduce the risk that broad calendar structure absorbs load that should be attributed to more specific drivers.

The reported values define an interpretable reference configuration rather than a globally optimized parameter set. The minimum overlap of 200 observations provides a support requirement for relation testing, while the maximum lag of 48 h covers delayed responses over a two-day period at hourly resolution. The 80th-percentile relation threshold

retains stronger relation evidence, and the consensus agreement of 0.50 requires support across at least half of the available relation layers. Evidence and validity powers of one preserve proportional weighting without additional nonlinear amplification. The validity floor of 0.10 prevents weak but plausible evidence from being eliminated completely, while the explainability cap of 0.95 preserves an allowance for unsupported variation. The dominance threshold of 0.85 and winner fraction of 0.95 restrict near-exclusive allocation to cases with clearly dominant family evidence. The 24 h event window represents a daily operating cycle, and the 0.20 quantile defines a lower local background against which event-like deviations are identified. The local sensitivity of these settings was evaluated through one-factor-at-a-time variation around the reported configuration.

Table 5. Main configuration parameters used in the reported experiments.

Component	Parameter	Value
Feature filtering	target-importance threshold	0.20
Relation discovery	minimum overlap	200
Relation discovery	lag search maximum	48 time steps
Relation thresholds	threshold quantile	80th percentile
Family grouping	consensus agreement threshold	0.5
Baseline	baseline mode	none
Explainability	explainability cap	0.95
Validity	family-validity floor	0.10
Allocation	evidence power	1
Allocation	validity power	1
Allocation	dominance threshold	0.85
Allocation	winner fraction	0.95
Event-background	rolling window	24 time steps
Event-background	rolling quantile	0.20
Event allocation	eligible families	gated families only

4.5. Evaluation Protocol

The evaluation compares the estimated MD-ADD outputs with the known reference component load signals that were excluded from the pipeline input. Family-level outputs are compared against Base/inflexible load [kW], EV Charging Load [kW], and HP Heating Load [kW]. For each estimated family and each reference component, Pearson correlation, mean absolute error, root mean squared error, and normalized mean absolute error are computed. Normalized mean absolute error is computed as

$$\mathrm{NMAE} = \frac{\sum_t |y_t - \hat{y}_t|}{\sum_t |y_t| + \epsilon} \tag{32}$$

where y_t is the reference component, $\hat{y}_t$ is the estimated family, feature, or proxy signal, and ϵ is a small numerical constant. Correlation and NMAE provide complementary information: correlation describes temporal co-variation, while NMAE describes the magnitude of the absolute error relative to the reference-component energy.

For post hoc component-alignment analysis, the best matching family for each reference component in the primary EV case is identified based on the highest correlation. In the additional ADRENALIN validation, the reported output is instead selected without reference-load information as the feature family containing outdoor air temperature.

Pearson correlation and NMAE are calculated between this family contribution and the measured temperature-dependent load, and the per-building results are summarized using the median across the nine buildings.

The evaluation is designed as a post hoc alignment analysis rather than a supervised component-reconstruction experiment. The model input consists only of the aggregate load and the observed contextual or operational features. Reference component signals are excluded from feature typing, relation discovery, grouping, evidence construction, and allocation, and are introduced only after decomposition to assess alignment with known physical load components. The reported correlation and error metrics are external alignment measures rather than losses optimized during model construction.

Post hoc alignment with a reference component indicates that an inferred contribution follows a similar temporal and magnitude structure. It does not establish causal influence or validate every feature within the corresponding family as a physical driver.

For comparison, Bayesian Disaggregation is included as a task-aligned thermal-load baseline in two settings [51]. On the primary EV dataset, an adapted Bayesian thermal-disaggregation run was performed using aggregate load and outdoor air temperature as inputs and was evaluated against the withheld HP Heating Load reference. It was not evaluated against the EV charging reference because the Bayesian formulation estimates a predefined temperature-dependent thermal component rather than an EV charging component. The same limitation applies to the base/inflexible reference. For the measured-building validation, the previously published Bayesian Disaggregation results for the same nine ADRENALIN buildings are also included [14]. Bayesian Disaggregation is an unsupervised probabilistic method that estimates a predefined temperature-dependent component from aggregate load and outdoor temperature. The configuration without separate weekend treatment is used for the ADRENALIN comparison, and the comparisons are based on NMAE, which is reported for both methods.

A direct EV-specific NILM baseline was not included because the primary EV case is not a standard appliance-level EV NILM benchmark, and EV-specific NILM methods typically target predefined EV detection or EV-load reconstruction rather than component-agnostic multi-driver attribution under partial observability.

To quantify the progression from the earlier MD-ADD formulation, an adapted implementation was applied to the primary EV dataset. The earlier formulation used predefined driver groups consisting of price-derived variables for EV-related attribution, outdoor-temperature and heating-degree-day variables for heating-related attribution, and calendar variables as a separate group. No analytical baseline was applied. Contributions were estimated using five-fold forward-chaining evaluation and out-of-fold XGBoost attribution. The present pipeline used the blindly selected EV- and temperature-associated feature families. Both methods were evaluated over the same 17,517 aligned hourly observations using the correlation and NMAE definitions introduced above. This experiment adapts the earlier formulation to the EV case rather than reproducing its original building-data evaluation.

A feature-level comparison is performed when feature-level outputs are available. The same metrics are computed between each feature-level contribution and each reference component. These comparisons do not imply that the method was trained using the reference components. They are used only to assess whether the unsupervised family and feature contributions align with the known simulated load structure.

The ablation analysis evaluates selected pipeline components using the same input data, target signal, timestamps, post hoc family-selection procedure, and evaluation metrics as the full configuration. The tested variants disable event-background separation, local-context augmentation, the conservative residual cap, validity control, or feature-family grouping. A further variant restricts relation discovery to linear or monotonic

correlation-based relations. Two additional reduced-mechanism configurations use only global relations or exclude activation and threshold mechanisms. All remaining pipeline settings are kept unchanged so that differences reflect the modified component or relation representation.

Parameter sensitivity was evaluated using a one-factor-at-a-time procedure in which one setting was varied while all remaining settings were fixed at the reported values. In addition to the target-importance and relation-threshold analyses, the tested settings were consensus agreement values of 0.40, 0.50, and 0.60; explainability caps of 0.85, 0.95, and 1.00; validity floors of 0.05, 0.10, and 0.20; dominance thresholds of 0.75, 0.85, and 0.95; winner fractions of 0.85, 0.95, and 1.00; evidence and validity powers of 0.5, 1.0, and 2.0; event windows of 12, 24, and 48 h; event quantiles of 0.10, 0.20, and 0.30; maximum lags of 24, 48, and 72 h; and minimum overlap requirements of 100, 200, and 500 observations. Each configuration was evaluated using the same dataset, reference components, family-selection procedure, and evaluation metrics as the reported configuration.

Input-perturbation and temporal-resolution sensitivity were evaluated using the same post hoc component-alignment procedure as the reported EV experiment. Three input-perturbation conditions were tested: 5% zero-mean Gaussian perturbation of the aggregate target load, 5% zero-mean Gaussian perturbation of continuous contextual features, and random removal of 10% of continuous contextual observations. Each perturbation condition was evaluated with three random seeds. Calendar variables, timestamps, and withheld reference component loads were not perturbed. Temporal-resolution sensitivity was evaluated by resampling the primary EV dataset to 2 h, 4 h, and 6 h resolution. Power and continuous contextual variables were resampled using their mean, while calendar variables were reconstructed from the resampled timestamps. The same family-selection procedure, reference components, and evaluation metrics were used across all conditions. These tests are interpreted as input-sensitivity checks rather than calibrated physical sensor-noise models.

Additional diagnostic checks were conducted to examine why the heat-pump and base/inflexible reference components were less successfully recovered than the EV reference component. These checks summarize component magnitude, activity frequency, masking of heat-pump temperature relationships in the aggregate target, base-load distinctiveness relative to the available contextual variables, and a preliminary hierarchical second-pass recovery test. The diagnostics are used only to interpret failure modes and possible improvement directions; they are not introduced as an additional optimized disaggregation configuration.

In addition to numerical metrics, the experiment produces time-series diagnostics for visual inspection, including reconstruction plots, stacked family decompositions, individual family contributions, and temporal comparisons between selected family outputs and reference components. These plots are used to evaluate whether the decomposition is temporally plausible and whether dominant load components are captured consistently in the selected temporal diagnostics.

4.6. Computational Runtime and Scaling Protocol

Computational runtime was evaluated using three executions of the primary EV dataset. The reference case used all 17,517 available timestamps and all 14 contextual input features. The timestamp-scaling case used the first 50% of the observations while retaining all contextual features. The feature-scaling case used the complete observation period and a fixed 50% subset of the contextual features. The resulting configurations contained 210, 210, and 56 candidate ordered variable pairs, respectively.

The algorithm and configuration parameters were unchanged between cases. Each run recorded the total pipeline runtime and the runtime of the main processing stages. Execution

was sequential, optional visualization generation was excluded from the timings, and no memory profiler was used. The reduced cases are empirical scaling checks intended to identify computational bottlenecks rather than accuracy-optimized experimental configurations.

5. Results

5.1. Feature Families and Target-Relevant Evidence

The multi-layer relation grouping procedure produced a compact set of feature families from the retained contextual and operational variables. Figure 2 shows the resulting consensus feature-family graph. The graph indicates that the selected features form distinct relation-supported groups rather than a single undifferentiated feature set. Strongly connected features correspond to variables that are repeatedly co-assigned across relation layers, while weakly connected or isolated features remain separated.

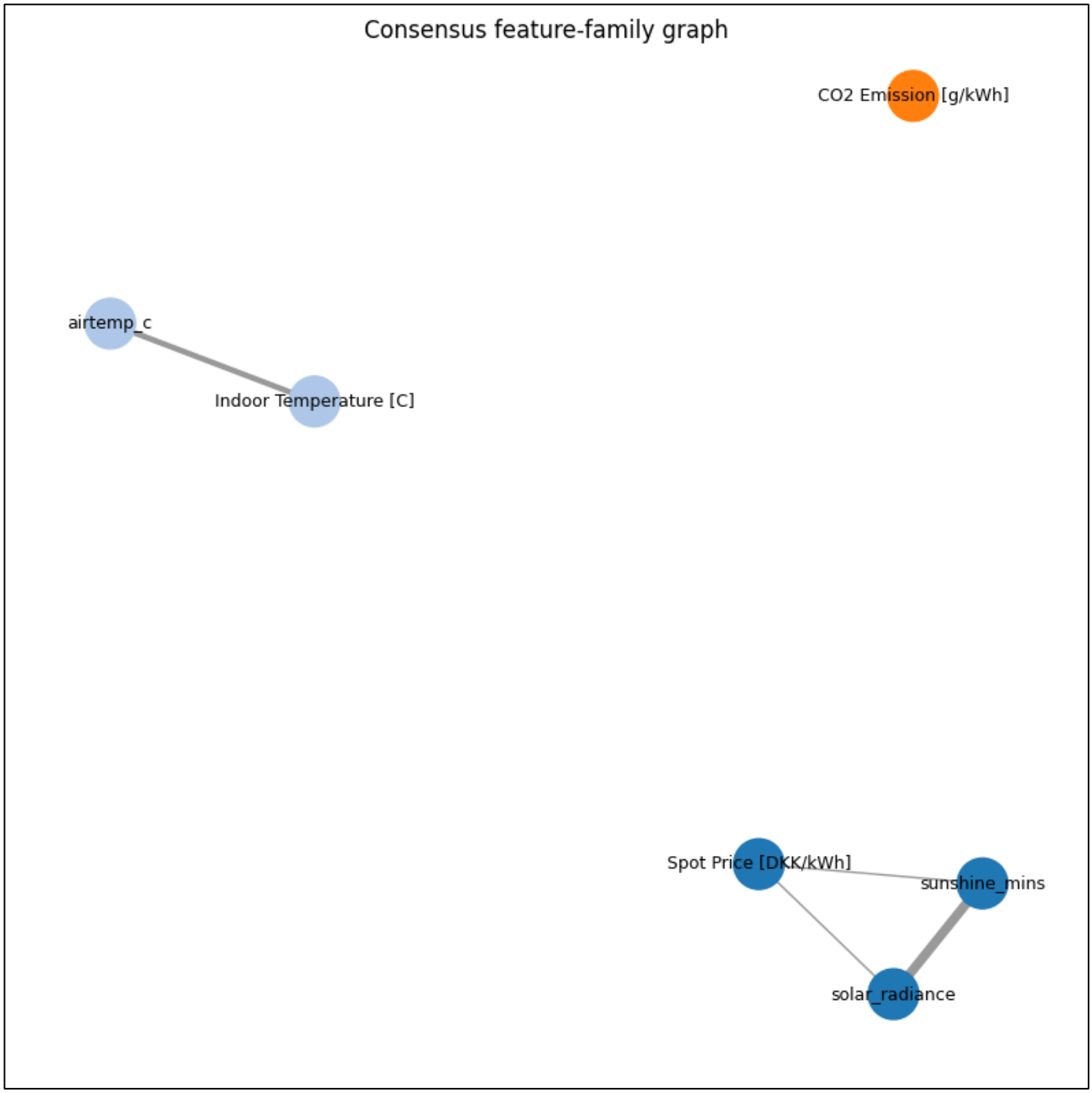


Figure 2. Consensus feature-family graph derived from the multi-layer relation grouping procedure.

The stability of the grouping is further illustrated by the consensus co-assignment matrix in Figure 3. High values along the block-diagonal regions indicate that features within the same family are consistently grouped together across relation layers. Low off-diagonal values indicate limited cross-family ambiguity, suggesting that the final family structure is not dominated by one relation type alone.

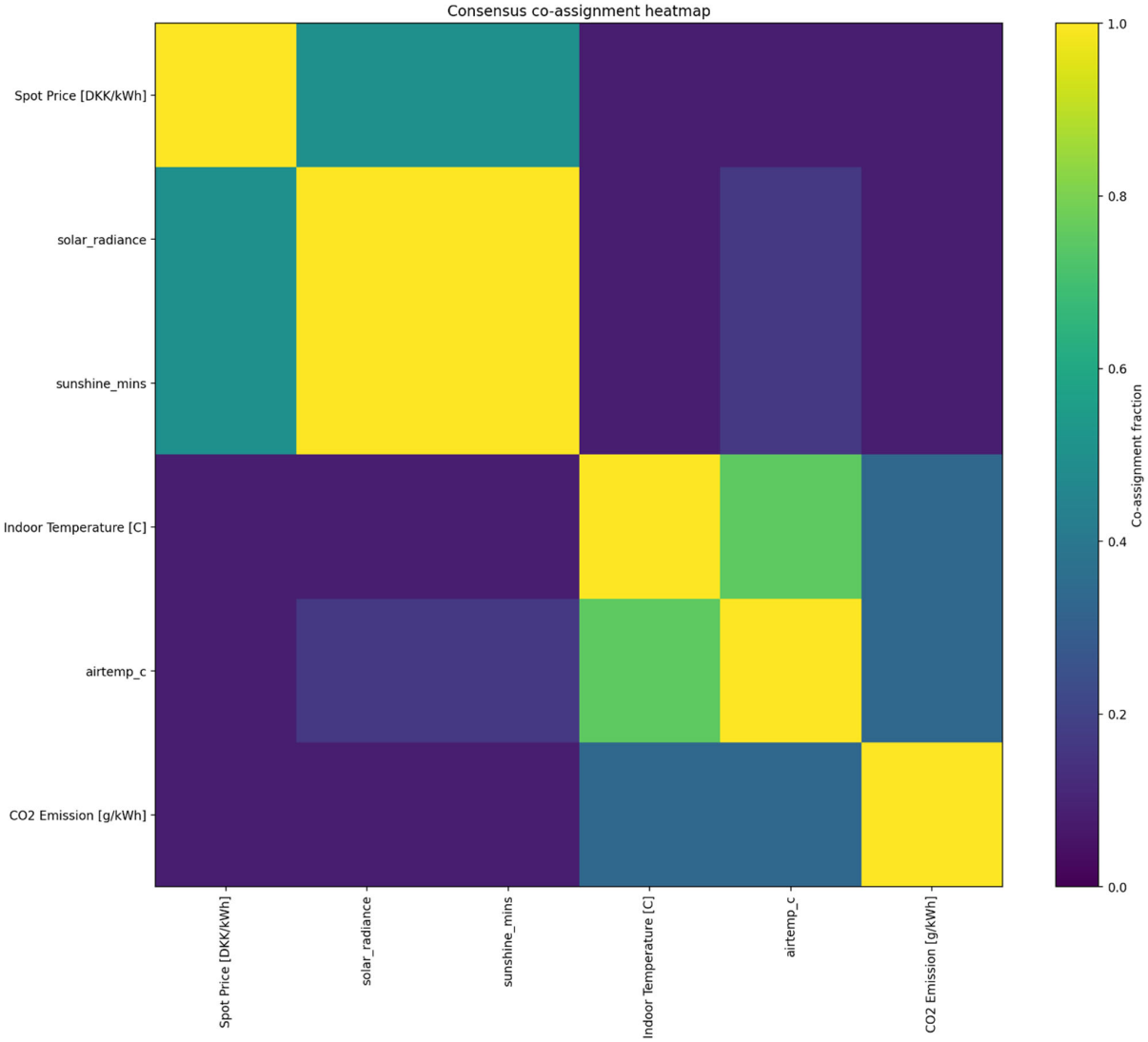


Figure 3. Consensus co-assignment matrix showing the fraction of relation layers in which each pair of features is assigned to the same family.

The final family composition is summarized in Table 6. Three families are retained. Family 0 contains sunshine duration, spot price, and solar radiation, with sunshine duration selected as the representative feature. Family 1 contains indoor and outdoor air temperature variables, with indoor temperature selected as the representative feature. Family 2 contains CO_2 emission intensity as a singleton family. These groups provide the driver structures used in the subsequent target-relevance and attribution stages.

The target-centric relevance analysis evaluates how strongly each family relates to the aggregate load. The family-level summary in Table 7 shows that target relevance is unevenly distributed across the retained families. Family 0 has the highest maximum target importance, with solar radiation as the strongest target-facing feature. Family 2 has moderate target relevance through CO_2 emission intensity, while Family 1 shows slightly lower but still retained relevance through the temperature variables.

Table 6. Consensus feature-family membership with representative features and internal connectivity statistics.

Family ID	Feature	Representative Feature	Representative Score	Number of Links in Family	Mean Weight in Family	Max Weight in Family
0	sunshine_mins	sunshine_mins	6.700	12	0.558	1.00
0	Spot Price [DKK/kWh]	sunshine_mins	1.285	5	0.257	0.84
0	solar_radiance	sunshine_mins	6.305	15	0.420	1.00
1	Indoor Temperature [C]	Indoor Temperature [C]	3.927	9	0.436	0.733
1	airtemp_c	Indoor Temperature [C]	3.922	9	0.436	0.733
2	CO_2 Emission [g/kWh]	CO_2 Emission [g/kWh]	0	0	0	0

Table 7. Family-level target relevance summary, including representative features and importance statistics.

Family ID	Representative Feature	Family Size	Best Feature in Family	Best Feature Score	Mean Family Importance	Max Family Importance
0	sunshine_mins	3	solar_radiance	0.311	0.259	0.311
2	CO_2 Emission [g/kWh]	1	CO_2 Emission [g/kWh]	0.253	0.253	0.253
1	Indoor Temperature [C]	2	Indoor Temperature [C]	0.225	0.224	0.225

At the feature level, Table 8 shows the effect of local-context augmented representations. Several transformed features obtain higher target-facing importance than their original forms, including spot price, indoor temperature, outdoor air temperature, and sunshine duration. The final target-facing importance used by the disaggregation model is defined as the stronger available representation, which allows the attribution stage to use either the raw feature or its most informative augmented form. This step ensures that target-relevant local structure is retained without allowing augmented features to dominate the earlier family-construction stage.

Table 8. Target-feature summary including original and augmented importance and preferred target-facing representations.

Feature	Family ID	Overall Target Importance Original	Overall Target Importance Augmented
solar_radiance	0	0.311181	0.316842
CO_2 Emission [g/kWh]	2	0.253319	0.24574
Spot Price [DKK/kWh]	0	0.245067	0.313284
Indoor Temperature [C]	1	0.225072	0.291161
airtemp_c	1	0.222717	0.270425
sunshine_mins	0	0.221063	0.230131

5.2. Family-Level and Feature-Level Attribution

The MD-ADD decomposition assigns part of the aggregate load to the identified feature families while preserving a residual or unexplained component. The method explains 82.5% of the total load through the identified families, while 17.5% remains as residual or unexplained load. The distribution of target-load contributions across families is shown in Table 9. The largest share is assigned to Family 0, followed by Family 1 and Family 2. The selected target-facing evidence signals for these families are a low-rank electricity-price representation, outdoor air temperature, and CO_2 emission intensity, respectively. These features should be interpreted as relation-derived evidence signals rather than direct physical measurements of individual loads.

Table 9. Family-level contribution shares and associated statistics.

Family	Share of Target	Selected Feature	Selected Relation	Selected Importance	Mean Validity
0	0.665	Spot Price [DKK/kWh]__rank_low_24	threshold_tail	0.491	0.375
1	0.098	airtemp_c	lagged_raw	0.417	0.646
2	0.062	CO_2 Emission [g/kWh]	linear_monotone	0.519	0.721

Temporal behavior of the family-level decomposition is illustrated in Figure 4, which shows the stacked family contributions for January 2025. The figure highlights how different families contribute during specific periods, reflecting both broader background variation and event-driven contributions.

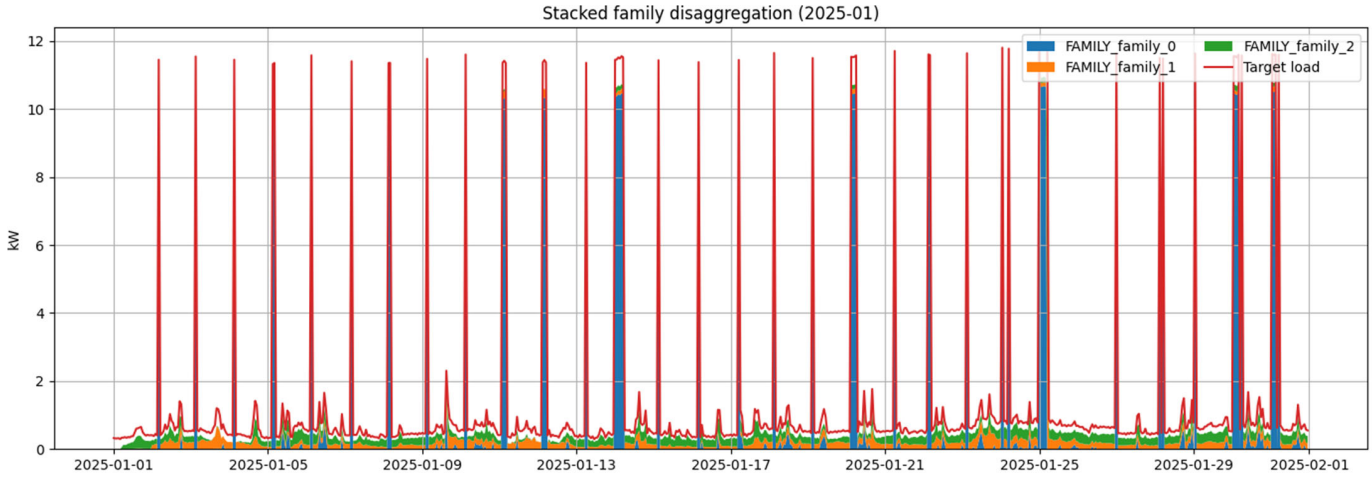


Figure 4. Stacked family-level disaggregation for January 2025.

The attributed reconstruction is shown in Figure 5, where the sum of the baseline and relation-supported family contributions is compared with the observed target load. The gap between the two curves corresponds to load that was not assigned to feature families, including unexplained dynamic variation and any remaining residual mismatch.

The role-based within-family attribution distributes each family-level contribution across its constituent features according to assigned roles, activity patterns, and target-facing relevance. The resulting feature-level contribution shares are summarized in Table 10. The largest contributions are assigned to the low-rank electricity-price representation, sunshine duration, electricity price, and solar radiation, indicating that the final decomposition is mainly driven by features associated with event timing and contextual modulation. Temperature-related and CO_2-related variables receive smaller but visible shares, sug-

gesting that smoother environmental drivers contribute to the decomposition without dominating the attribution.

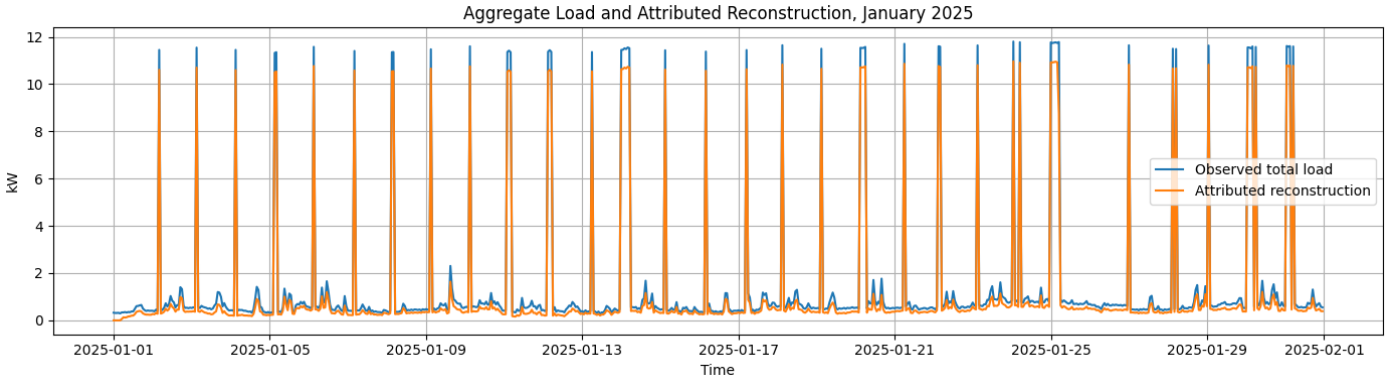


Figure 5. Aggregate load and attributed reconstruction for January 2025.

Table 10. Feature-level contribution shares derived from role-based attribution.

	Share_of_Target
Spot Price [DKK/kWh]__rank_low_24	0.2160
sunshine_mins	0.1666
Spot Price [DKK/kWh]	0.1611
solar_radiance	0.121
airtemp_c	0.0786
CO_2 Emission [g/kWh]	0.0609
Indoor Temperature [C]	0.0147
Indoor Temperature [C]__rank_low_24	0.0025
airtemp_c__rank_low_72	0.0025

Figure 6 shows the temporal evolution of feature-level contributions for January 2025. The figure highlights how individual features dominate during specific periods, reflecting their role in different operating conditions and supporting the interpretation of the family-level decomposition.

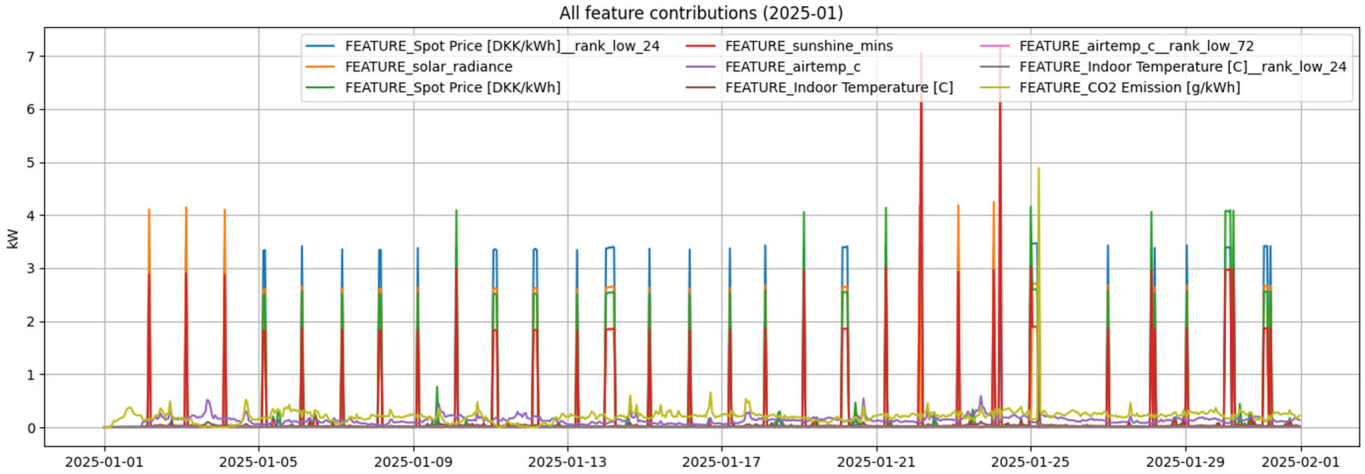


Figure 6. Feature-level contribution time series for January 2025.

5.3. Post Hoc Component Alignment

Although the model operates without direct access to component-level load signals, post hoc comparison provides insight into how the estimated family-level contributions align with the known simulated reference components. The best matching family for each reference component is reported in Table 11. The strongest alignment is observed for EV charging load, where Family 0 achieves a correlation of 0.990 and a normalized mean absolute error of 0.100. The heat-pump and base-load components show weaker alignment, indicating that the decomposition captures the dominant event-driven EV component more clearly than smoother or more continuously overlapping load components.

Table 11. Best matching family for each reference component.

Reference Component	Family ID	Correlation	MAE	RMSE	NMAE
Base/inflexible load [kW]	2	0.159	0.099	0.203	0.895
EV Charging Load [kW]	0	0.990	0.082	0.449	0.100
HP Heating Load [kW]	1	0.349	0.119	0.193	0.565

The full comparison across all families is presented in Table 12. The results show that the decomposition is strongly dominated by the EV-related structure. Family 0 aligns closely with the EV charging reference, whereas the best heat-pump and base-load matches remain substantially weaker. The best heat-pump match reaches only moderate alignment, while the base-load reference remains weakly aligned with all families. This indicates that the present configuration does not recover all simulated components equally. Instead, it is most effective for the dominant, temporally localized EV charging component, while smoother and lower-magnitude heat-pump and background-load variation remains only partially separated under the available contextual information.

Table 12. Comparison between family outputs and reference components using correlation and error metrics.

Family ID	Reference Component	Correlation	MAE	RMSE	NMAE
2	Base/inflexible load [kW]	0.159	0.099	0.203	0.895
1	Base/inflexible load [kW]	0.111	0.089	0.206	0.799
0	Base/inflexible load [kW]	−0.031	0.834	2.742	7.512
0	EV Charging Load [kW]	0.990	0.082	0.449	0.100
1	EV Charging Load [kW]	0.238	0.893	2.938	1.093
2	EV Charging Load [kW]	0.079	0.870	2.972	1.065
1	HP Heating Load [kW]	0.349	0.119	0.193	0.565
2	HP Heating Load [kW]	0.304	0.154	0.220	0.733
0	HP Heating Load [kW]	0.058	0.918	2.707	4.359

The temporal comparison between the family most strongly aligned with EV charging and the reference EV charging load is shown in Figure 7. The estimated family contribution reproduces the timing of the dominant EV charging events with strong consistency across January 2025. Most major charging peaks are aligned with the reference signal, although small differences in peak magnitude and low-level fluctuations remain.

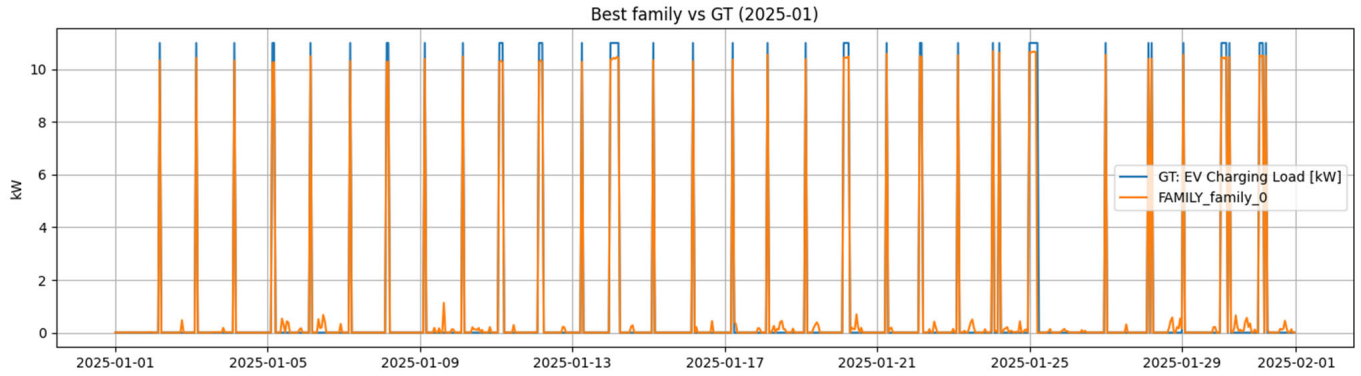


Figure 7. Hourly comparison between the EV-aligned family contribution and the reference EV charging load for January 2025.

Figures 8 and 9 show the corresponding temporal comparisons for the heat-pump and base-load components. The weaker alignment in both cases reflects the lower separability of smoother load components under the available feature set. Family 1 captures some short-term heating-related variation, while Family 2 only partially follows the base-load signal. The large isolated spike visible in both estimated families corresponds to an EV charging event that was not assigned exclusively to the EV-related family, highlighting a remaining allocation ambiguity for individual high-load events.

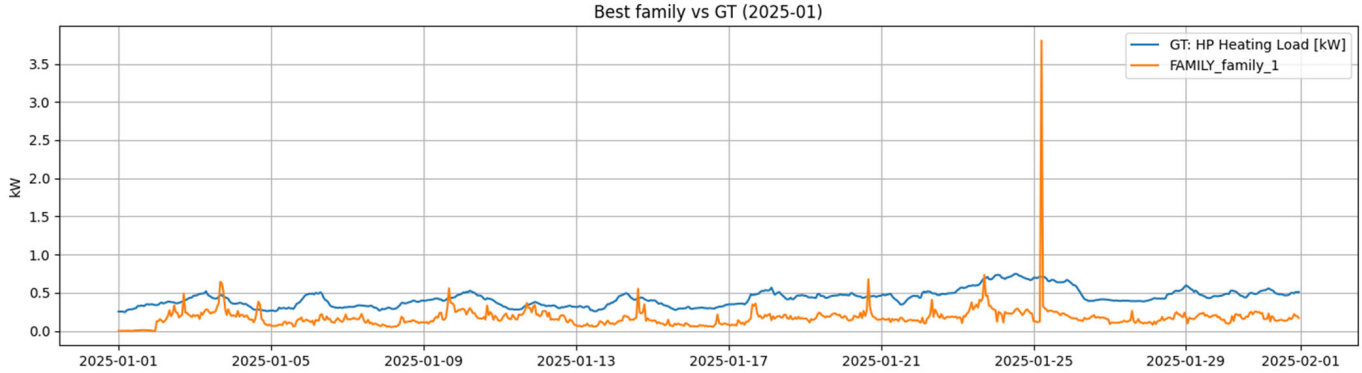


Figure 8. Hourly comparison between the heat-pump-aligned family contribution and the reference heat-pump load.

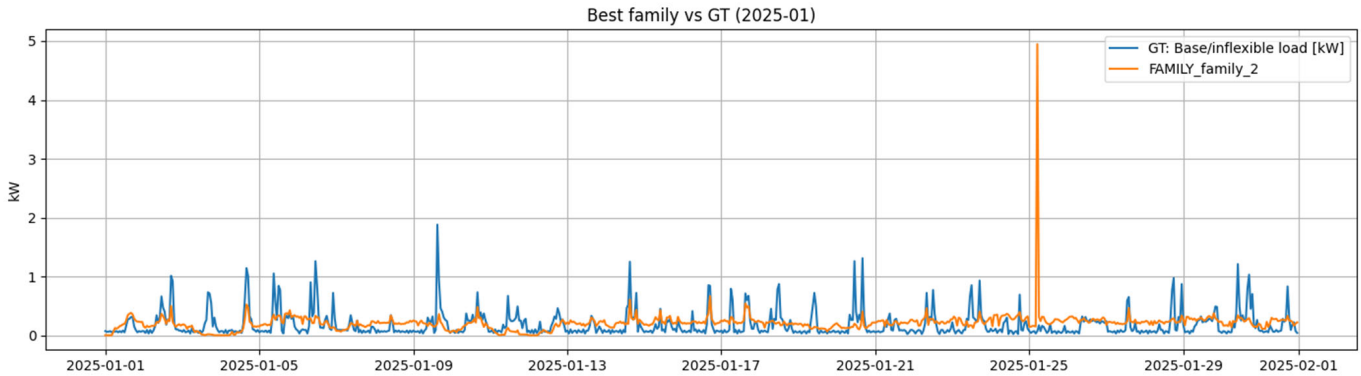


Figure 9. Hourly comparison between the base-load-aligned family contribution and the reference base/inflexible load.

The post hoc reference comparison shows that one estimated family aligns strongly with the event-driven EV charging component at the original temporal resolution. The

family most closely aligned with the heat-pump component partially represents but underestimates that signal, while no estimated family aligns strongly with the base-load component. These results indicate that the proposed relation-aware allocation is most effective for dominant, temporally localized load components and less effective for smoother or lower-magnitude components under the available feature set.

5.4. Additional Measured-Data Validation and Comparative Evaluation

The pipeline was additionally evaluated on the nine measured ADRENALIN buildings using the restricted and full feature sets defined in Section 4.2. For each building, the evaluated output was the feature family containing outdoor air temperature, selected without using the measured temperature-dependent reference load. Table 13 presents the resulting correlation and NMAE values.

Table 13. Alignment of the selected temperature-associated family with the measured temperature-dependent load in the ADRENALIN buildings.

Building	Restricted Feature Set		Full Feature Set	
	Correlation	NMAE	Correlation	NMAE
L03.B02	0.548	0.815	0.561	0.751
L06.B01	0.751	1.143	0.229	0.909
L09.B01	0.579	0.988	0.597	0.877
L10.B01	0.664	0.846	0.106	0.995
L14.B01	0.849	0.554	0.830	0.536
L14.B02	0.889	0.491	0.864	0.555
L14.B03	0.846	0.594	0.828	0.576
L14.B04	0.888	0.515	0.883	0.493
L14.B05	0.931	0.433	0.928	0.441
Median	0.846	0.594	0.828	0.576

A temperature-associated family was retained for all nine buildings under both feature sets. The restricted feature set produced a median correlation of 0.846 and a median NMAE of 0.594. The corresponding values for the full feature set were 0.828 and 0.576. The restricted feature set produced the higher correlation in six buildings, while the full feature set produced the lower NMAE in five buildings. The selected temperature-associated family was also the family with the highest post hoc correlation to the measured reference in all nine restricted feature-set runs and in six of the nine full feature-set runs.

Table 14 compares the median NMAE of the two MD-ADD configurations with the previously published Bayesian Disaggregation results for the same nine buildings. Bayesian Disaggregation achieved the lowest median NMAE of 0.540, followed by the full MD-ADD configuration at 0.576 and the restricted configuration at 0.594.

Table 14. Median temperature-dependent load NMAE across the nine ADRENALIN buildings.

Method and Configuration	Result Source	Median NMAE
Bayesian Disaggregation	Previously published results [14]	0.540
MD-ADD, restricted feature set	Present study	0.594
MD-ADD, full feature set	Present study	0.576

Table 15 compares the adapted earlier MD-ADD formulation, the adapted Bayesian thermal-disaggregation baseline, and the present relation-aware pipeline on the primary EV dataset. The comparison uses the same aligned timestamps and metric implementation. Bayesian thermal disaggregation is reported only for the heat-pump reference because it estimates a predefined temperature-dependent thermal component and is not applicable to EV charging.

Table 15. Comparison of the adapted earlier MD-ADD formulation, Bayesian thermal disaggregation, and the present relation-aware pipeline on the primary EV dataset.

Reference Component	Method	Correlation	NMAE
EV charging	Earlier MD-ADD	0.556	0.974
EV charging	Present MD-ADD	0.990	0.100
Heat-pump heating	Earlier MD-ADD	0.310	0.831
Heat-pump heating	Bayesian thermal disaggregation	0.769	0.278
Heat-pump heating	Present MD-ADD	0.349	0.565

For EV charging, the present pipeline produced a correlation of 0.990 and an NMAE of 0.100, compared with 0.556 and 0.974 for the adapted earlier formulation. No Bayesian result is reported for EV charging because Bayesian Disaggregation is formulated for temperature-dependent thermal-load estimation. For heat-pump heating, the Bayesian thermal baseline achieved a correlation of 0.769 and an NMAE of 0.278, compared with 0.349 and 0.565 for the present MD-ADD heat-pump-associated family and 0.310 and 0.831 for the adapted earlier formulation. This indicates that the heat-pump signal is partly recoverable by a task-specific thermal estimator, while the present pipeline remains broader because it performs component-agnostic feature-family attribution rather than predefined thermal-load reconstruction.

5.5. Proxy, Ablation, and Sensitivity Checks

To contextualize the post hoc component-alignment results, two simple component-specific proxy comparisons were evaluated. These proxies are not direct baselines for the proposed MD-ADD pipeline, because they assume the component of interest and its expected driver structure in advance. The EV proxy is constructed as an event-oriented proxy for EV-like high-load behavior, while the heating proxy is constructed from heating-degree-day information and assumes a temperature-driven heating component. In contrast, the proposed pipeline does not assume in advance that EV charging or heating should be extracted as target components.

Table 16 compares these component-specific proxies with the corresponding MD-ADD family outputs. The EV event proxy achieves a very high correlation with the EV reference component, indicating that the timing of dominant EV events is relatively easy to capture when the component type is assumed in advance. However, its NMAE remains much higher than the MD-ADD result because the proxy is not magnitude-calibrated to the EV load. It captures event timing but assigns inaccurate energy magnitudes. For heating, the HDD proxy achieves high correlation because it directly encodes the expected temperature relationship, but its NMAE remains higher than the MD-ADD heat-pump-aligned family. These results show that component-specific assumptions can recover expected temporal patterns, but they do not solve the component-agnostic relation-supported attribution problem addressed in this paper.

Table 16. Component-specific proxy comparison for EV and heat-pump reference components.

Reference Component	Method	Correlation	NMAE
EV charging	Event proxy	0.999	0.924
EV charging	MD-ADD	0.990	0.100
HP heating	HDD proxy	0.967	0.713
HP heating	MD-ADD	0.349	0.565

Ablation analysis was used to examine the contribution of the main relation, representation, and allocation components. Table 17 compares the full configuration with six successful reduced variants evaluated under the same input data and post hoc evaluation protocol.

Table 17. Ablation results for selected MD-ADD pipeline components.

Configuration	EV Corr.	EV NMAE	HP Corr.	HP NMAE	Base Corr.	Base NMAE	Residual Share
Full MD-ADD	0.990	0.100	0.349	0.565	0.159	0.895	0.175
Without event-background separation	0.991	0.115	0.331	0.590	0.238	0.913	0.175
Without local-context features	0.760	0.450	0.135	1.012	0.067	1.482	0.294
Without conservative residual cap	0.991	0.055	0.349	0.545	0.161	0.913	0.132
Correlation-only relation discovery	0.682	0.851	0.300	1.455	0.021	2.058	0.529
Without validity control	0.990	0.101	0.353	0.549	0.152	0.884	0.173
Without feature-family grouping	0.907	0.561	0.613	0.661	0.222	0.853	0.160

Restricting relation discovery to correlation-based relations produced the largest reduction in EV alignment, with correlation decreasing to 0.682, NMAE increasing to 0.851, and the residual share increasing to 0.529. Removing local-context features reduced EV correlation to 0.760 and increased EV NMAE to 0.450. Without feature-family grouping, EV correlation decreased to 0.907 and EV NMAE increased to 0.561, while HP correlation increased to 0.613 but HP NMAE increased to 0.661.

Removing validity control produced values close to the full configuration. Disabling event-background separation caused a small increase in EV and HP NMAE. Removing the conservative residual cap reduced EV NMAE to 0.055 and the residual share to 0.132. Across the configurations, base-load alignment remained weak.

The global-only relation-discovery variant and the variant without activation and threshold mechanisms did not produce retained final feature families under the unchanged thresholds and did not yield allocation outputs.

A compact sensitivity analysis was conducted because target-importance filtering and relation-threshold selection influence both the retained feature set and the resulting family structure. Table 18 varies the target-importance threshold and relation-threshold quantile while keeping the input data, target variable, baseline configuration, and post hoc evaluation protocol unchanged. The analysis provides a robustness check around the reported configuration by examining whether the main findings remain consistent under nearby target-importance and relation-threshold settings.

Table 18. Sensitivity of post hoc component alignment to target-importance and relation-threshold settings.

Setting	EV Correlation	EV NMAE	HP Correlation	HP NMAE	Base Correlation	Base NMAE	Residual Share	Retained Features	Families
Target importance = 0.15	0.766	0.627	0.140	1.021	0.056	1.488	0.392	8	3
target importance = 0.20	0.990	0.100	0.349	0.565	0.159	0.895	0.175	6	3
target importance = 0.25	0.764	0.456	0.133	1.005	0.066	1.765	0.426	2	2
relation threshold quantile = 0.70	0.991	0.102	0.359	0.614	0.138	1.097	0.227	6	2
relation threshold quantile = 0.80	0.990	0.100	0.349	0.565	0.159	0.895	0.175	6	3
relation threshold quantile = 0.90	0.925	0.377	0.550	0.568	0.240	0.872	0.166	6	4

The sensitivity results show that the selected target-importance threshold of 0.20 provides the strongest overall balance across the evaluated components. Lowering the threshold to 0.15 retains more features but weakens EV alignment and increases the residual share, while increasing the threshold to 0.25 retains only two features and two families and again weakens EV alignment. The relation-threshold results show that EV alignment is comparatively stable at the 0.70 and 0.80 quantiles, with the 0.80 setting giving the lowest EV NMAE and residual share among these two settings. At the 0.90 quantile, the family structure becomes more fragmented and EV NMAE increases. Overall, the qualitative conclusion remains consistent: the method aligns most strongly with the dominant EV charging component, while heat-pump and base/inflexible load alignment remains weaker.

The one-factor-at-a-time analysis was extended to the remaining relation-discovery, family-construction, validity, allocation, and event-background settings. Table 19 summarizes the maximum absolute change from the reported configuration across the alternative values tested for each parameter.

Consensus agreement, explainability cap, and winner fraction produced the largest changes. Increasing consensus agreement from 0.50 to 0.60 increased the family count from three to four and increased EV NMAE from 0.100 to 0.377. Reducing the explainability cap to 0.85 increased EV NMAE to 0.194 and the residual share to 0.262, while increasing the cap to 1.00 reduced EV NMAE to 0.055 and the residual share to 0.132. Reducing the winner fraction to 0.85 increased EV NMAE to 0.150 and HP NMAE to 0.734, while a winner fraction of 1.00 produced corresponding NMAE values of 0.076 and 0.588.

Evidence power, validity power, and dominance threshold produced smaller changes across the tested ranges. Changes in the event window and event quantile remained below 0.001 for both EV and HP NMAE. The tested validity-floor, maximum-lag, and minimum-

overlap settings produced no change in the reported component-alignment metrics for this dataset.

Table 19. One-factor-at-a-time sensitivity of selected pipeline parameters relative to the reported configuration.

Parameter	Tested Values	Maximum EV NMAE Change	Maximum HP NMAE Change	Maximum Residual-Share Change	Maximum Family-Count Change
Consensus agreement	0.40, 0.50, 0.60	0.276	0.003	0.009	1
Explainability cap	0.85, 0.95, 1.00	0.094	0.041	0.087	0
Winner fraction	0.85, 0.95, 1.00	0.050	0.169	0.000	0
Evidence power	0.5, 1.0, 2.0	0.002	0.022	0.000	0
Validity power	0.5, 1.0, 2.0	<0.001	0.011	0.000	0
Dominance threshold	0.75, 0.85, 0.95	<0.001	0.008	0.000	0
Event window	12, 24, 48 h	<0.001	<0.001	0.000	0
Event quantile	0.10, 0.20, 0.30	<0.001	<0.001	0.000	0
Validity floor	0.05, 0.10, 0.20	0.000	0.000	0.000	0
Maximum lag	24, 48, 72 h	0.000	0.000	0.000	0
Minimum overlap	100, 200, 500	0.000	0.000	0.000	0

Input-perturbation and temporal-resolution sensitivity were evaluated to examine whether the reported EV alignment remained identifiable under moderate input corruption and coarser temporal aggregation. Table 20 summarizes the clean 1 h configuration, the three perturbation cases, and the three coarser-resolution cases.

Table 20. Input-perturbation and temporal-resolution sensitivity results.

Condition	EV Corr.	EV NMAE	HP Corr.	HP NMAE	Base NMAE	Residual Share
Clean 1 h baseline	0.990	0.100	0.349	0.565	0.895	0.175
Missing contextual observations, 10%	0.927	0.352	0.390	0.711	0.953	0.198
Contextual-feature noise, 5%	0.983	0.214	0.341	0.632	1.107	0.304
Aggregate-meter noise, 5%	0.920	0.226	0.074	4.008	7.051	0.344
2 h temporal resolution	0.621	0.689	0.150	2.039	3.292	0.476
4 h temporal resolution	0.846	0.499	0.306	1.329	1.259	0.399
6 h temporal resolution	0.947	0.523	0.150	2.340	3.344	0.366

The EV-aligned family remained identifiable under the three perturbation settings, although magnitude accuracy degraded relative to the clean 1 h configuration. Contextual-feature noise produced a high EV correlation of 0.983 but increased EV NMAE from 0.100 to 0.214. Aggregate-meter noise reduced EV correlation to 0.920 and increased EV NMAE to 0.226. Removing 10% of continuous contextual observations produced EV correlation of 0.927 and EV NMAE of 0.352.

Temporal coarsening weakened the decomposition relative to the clean hourly case. EV NMAE increased to 0.689 at 2 h resolution, 0.499 at 4 h resolution, and 0.523 at 6

h resolution, while the residual share increased from 0.175 to between 0.366 and 0.476. Heat-pump and base-load alignment remained fragile, particularly under aggregate-meter perturbation and coarser temporal aggregation. These results indicate that the dominant EV component remains detectable under moderate perturbation, but the method is most reliable for the hourly setting used in the main experiment and for components whose relation structure remains visible in the aggregate load.

Additional diagnostic checks were used to characterize the weaker heat-pump and base/inflexible-load recovery. Table 21 summarizes component-scale, activity, masking, base-load distinctiveness, and hierarchical second-pass indicators. Some diagnostics apply only to one component, for example temperature masking for heat-pump load or contextual distinctiveness for base/inflexible load, so the table intentionally contains several non-applicable entries.

Table 21. Diagnostic checks for weaker heat-pump and base/inflexible-load recovery.

Diagnostic	EV	HP	Base/Inflexible
Share of target energy	0.718	0.185	0.097
Nonzero fraction	0.074	0.944	1.000
Selected-family correlation	0.990	0.349	0.159
Selected-family NMAE	0.100	0.565	0.895
Best temperature or HDD correlation in the true component	n/a	0.967	n/a
Same temperature or HDD correlation in the aggregate target	n/a	0.095	n/a
Same temperature or HDD correlation after estimated EV-contribution removal	n/a	0.279	n/a
Maximum absolute Pearson correlation with contextual variables	n/a	n/a	0.262
Maximum absolute Pearson correlation with selected family outputs	n/a	n/a	0.159
Hierarchical second-pass correlation	n/a	0.546	n/a
Hierarchical second-pass NMAE	n/a	0.525	n/a
Hierarchical second-pass residual share	n/a	0.408	n/a

5.6. Computational Runtime and Empirical Scaling

Table 22 reports the runtime results for the three scaling cases. The full configuration required 2525.0 s. Reducing the observation period by 50% reduced the total runtime to 1158.8 s, while reducing the contextual feature count by 50% reduced it to 977.5 s.

Table 22. Computational runtime and empirical scaling results.

Case	Timestamps	Context Features	Candidate Ordered Pairs	Relation Discovery [s]	Total Runtime [s]	Relation-Discovery Share
Full timestamps, all features	17,517	14	210	2492.3	2525.0	98.7%
First 50% of timestamps, all features	8759	14	210	1142.3	1158.8	98.6%
Full timestamps, 50% of features	17,517	7	56	959.0	977.5	98.1%

Halving the timestamp count reduced total runtime by 54.1%, while halving the contextual feature count reduced it by 61.3%. Relation discovery accounted for more than

98% of total runtime in every case. In the full configuration, original-variable relation discovery required 1328.8 s and augmented target-relation discovery required 1163.5 s, whereas evidence construction, allocation, and attribution required 9.0 s. These results identify relation discovery as the dominant practical computational stage. The reduced runs are empirical scaling observations and should not be interpreted as estimates of a polynomial runtime exponent.

6. Discussion

6.1. Interpretation of Main Findings

The results show that the proposed pipeline can extract structured and interpretable load components from aggregate energy data without using component-level labels as model inputs. The clearest result is obtained for the EV charging component, which shows strong temporal alignment, high correlation, and low normalized error relative to the withheld reference signal. This indicates that the relation-aware and event-focused allocation strategy is effective for dominant, high-intensity, temporally localized load components. In contrast, the heat-pump and base-load components are only partially recovered, showing that smoother, lower-magnitude, or more strongly overlapping components remain harder to separate when their relation structure is masked by stronger aggregate-load behavior.

The additional ADRENALIN results provide context for the weaker heat-pump alignment in the primary EV case. Across the nine measured buildings, the temperature-associated family achieved a median correlation of 0.846 under the restricted feature set, indicating that smoother temperature-responsive structure can be identified when it remains visible in the aggregate load. The restricted feature set produced a slightly higher median correlation and matched the strongest post hoc family in all nine buildings, compared with six buildings under the full feature set, while the full feature set achieved a slightly lower median NMAE. This suggests that the dominant EV and base-load variation masks the heating relationship in the primary case, while broader contextual inputs can improve magnitude alignment but produce more complex family structures.

The comparisons with Bayesian Disaggregation highlight the distinction between component-specific estimation and component-agnostic attribution. On the primary EV dataset, the Bayesian thermal baseline achieved stronger heat-pump alignment than the heat-pump-associated MD-ADD family, with a correlation of 0.769 and an NMAE of 0.278. This result is expected because Bayesian Disaggregation is explicitly formulated for estimating temperature-dependent thermal load from aggregate load and outdoor temperature, and it is not an EV-load disaggregation method. A similar pattern appears in the ADRENALIN comparison, where Bayesian Disaggregation achieved the lower median NMAE of 0.540, compared with 0.576 for the full MD-ADD configuration and 0.594 for the restricted configuration. The specialized Bayesian formulation therefore retains an advantage for predefined thermal-load tasks, while MD-ADD provides broader feature-family attribution under weaker assumptions about which physical components are present.

The comparison with the earlier MD-ADD formulation shows the collective effect of changing from direct attribution over predefined driver groups to automatic relation-aware family construction and allocation. The largest difference occurs for EV charging, where the present pipeline captures the dominant event-driven structure substantially more accurately. Heat-pump correlation changes only modestly, although the lower NMAE indicates improved magnitude alignment. Because the comparison changes several methodological stages simultaneously, the performance difference cannot be assigned to one individual mechanism. It quantifies the overall progression of the framework, while the design-specific effects are interpreted through the ablation and sensitivity results in Section 6.2.

The feature-family structure represents relational evidence rather than direct causal decomposition. The grouping process identifies variables with shared relational behavior and organizes them into families that provide evidence for specific parts of the aggregate-load variation. The representative feature within a family provides a concise descriptor of the group, while the full family structure describes the broader timing and operating context of the estimated contribution. This distinction is especially important for the EV-related family. In the simulated dataset, EV charging is primarily linked to spot-price-driven charging decisions, while other variables may enter the same family because they share temporal structure with charging events. For example, solar-radiation and sunshine-related variables reflect low-solar or night-time conditions that coincide with charging behavior. The family captures the timing and context of EV-related load, with spot price providing the most meaningful interpretation of the simulated charging mechanism.

Mechanism modeling refines this interpretation by distinguishing between activation evidence, intensity-related evidence, modulation, and local contextual support. This prevents the family contribution from being interpreted as a single-feature explanation. Instead, the resulting structure supports a layered interpretation in which raw and augmented variables are grouped by relational similarity, assigned mechanism roles, and translated into family-level and feature-level load contributions. This is important because the final attribution is not optimized directly against known EV, heat-pump, or base-load channels. The reference components are introduced only after decomposition to evaluate whether the inferred family outputs align with physically meaningful load processes.

6.2. Mechanism, Ablation, and Sensitivity Interpretation

The ablation and sensitivity results clarify how the relation-aware framework contributes to the observed attribution behavior. By combining linear, threshold-based, event-driven, lagged, cyclic, and window-local relations, the method can capture interactions that would be missed by a single association measure. This is particularly relevant for EV charging, where the observable structure is not only a continuous linear relationship but also an event-like timing pattern associated with contextual conditions. The ablation analysis shows that local-context augmented features are especially important for the observed EV alignment. Removing these features weakens EV alignment and increases the residual share, indicating that time-localized transformations expose relationships that are less visible in the raw feature values alone.

The ablation analysis clarifies the roles of multi-mechanism relation discovery, feature-family grouping, and validity control. Restricting relation discovery to correlation-based relationships substantially weakened EV alignment and increased the unexplained share, indicating that linear or monotonic co-variation alone did not preserve the event- and context-related structure identified by the full mechanism set. Removing feature-family grouping also weakened EV recovery, although it increased heat-pump correlation while worsening heat-pump NMAE. This mixed result indicates that grouping supports coherent multi-feature attribution but does not necessarily maximize every individual component metric. Validity control had little effect in this case, showing that it was not responsible for the strong EV result; its role is primarily to limit allocation when family evidence becomes locally weak or inconsistent.

The two reduced-mechanism variants that produced no retained families were evaluated under the unchanged thresholds of the main configuration. This outcome indicates an interaction between mechanism coverage and family-retention settings. Separately calibrated global-only or reduced-mechanism formulations may produce different retention behavior.

A defining characteristic of the method is the explicit preservation of unexplained load. Rather than enforcing a complete decomposition of the target signal, the pipeline allocates only the portion of the load supported by sufficient relational evidence. The remaining load is retained as residual or unexplained dynamic variation. This design reduces the risk of assigning energy to spurious or weakly supported drivers and provides an explicit representation of uncertainty under partial observability. In the present experiment, the residual remains non-zero even though the dominant EV charging component is recovered. This indicates that strong recovery of one component does not require forcing all remaining variation into the identified feature families.

The ablation without the conservative residual cap illustrates the trade-off introduced by this design. Removing the cap improves the EV NMAE and increases the explained share, but reduces the residual share. The conservative cap prioritizes residual preservation and reduces the absorption of unsupported or ambiguous variation into estimated families. The residual serves as a diagnostic output that reflects load variation with insufficient feature-family evidence, including variation associated with missing drivers, weak relationships, overlapping components, or allocation ambiguity.

The event-background separation contributes to the characterization of EV charging by improving the handling of short, event-like peaks. The method identifies periods of elevated activity relative to a local background level and assigns responsibility for these events using relation-driven evidence. The ablation analysis shows that disabling this layer does not remove the strong EV alignment, suggesting that the main EV structure is already captured by relation-aware evidence and local-context features. The event-background layer acts as an additional refinement rather than the sole mechanism behind the result. This design is most useful for high-intensity, short-duration loads with clear deviations from background behavior, while smoother components such as heat-pump demand and base load remain less distinctly separated. Event allocation can also remain shared across families when evidence is ambiguous, explaining why isolated EV-related spikes may appear in families that are not the best post hoc match for EV charging.

The sensitivity analysis shows that the pipeline parameters do not have equal influence. Target-importance filtering and consensus agreement affect the retained representation and family structure, while the explainability cap controls the tradeoff between component-level error and preservation of unexplained demand. The winner fraction affects how strongly a dominant family receives the explainable load and produced the largest change in heat-pump NMAE among the allocation settings. Evidence power, validity power, and dominance threshold had smaller effects around the reported configuration, while the tested validity-floor and event-background settings were stable in this case. These values should be interpreted as configurable design controls rather than physical constants or universally optimal settings. Recalibration may be required when the feature set, temporal resolution, or dominant load structure changes.

The input-perturbation analysis should be interpreted as a sensitivity check rather than as a calibrated measurement-uncertainty study. The primary EV dataset already contains background household-demand variation through the simulated base/inflexible component, while the added perturbations impose artificial changes on top of that structure. The results indicate whether the EV-aligned family remains identifiable under moderate corruption of the input signals, not whether the method is universally robust to all realistic metering, sensing, or data-quality errors. The temporal-resolution results further indicate that attribution quality degrades when the hourly signal is aggregated to coarser intervals, especially for weaker heat-pump and base-load components. This supports treating hourly low-frequency data as the evaluated operating regime of the present study.

6.3. Limitations, Generalizability, and Deployment Implications

The method depends strongly on the availability and quality of the observed input features. Only load variations that are reflected in the available contextual or operational variables can be explained. Missing, weakly informative, or indirectly related features limit the recoverable structure and contribute to the residual component. This is an inherent limitation of relation-supported attribution under partial observability, because the method can only allocate load to feature families when sufficient evidence is present in the observed data.

The recovery of smooth or slowly varying components remains limited. Such components may be distributed across several families, underestimated, or retained in the residual when their relation evidence is weaker than that of a dominant event-driven component. This reflects the difficulty of separating overlapping continuous processes using relation-based evidence alone, especially when the aggregate signal contains stronger event-like variation.

Dominant signals can also mask weaker relationships. In the present case, EV charging is large, sparse, and event-driven, accounting for 0.718 of target energy while being nonzero in only 0.074 of timestamps. The heat-pump component is smaller and smoother, accounting for 0.185 of target energy while being nonzero in 0.944 of timestamps. Although the true heat-pump signal is strongly related to outdoor temperature and heating degree days, these relationships are largely hidden in the aggregate target and only partly re-emerge after the estimated EV-aligned contribution is removed. This masking limited the ability of automatic family construction and target-facing evidence selection to isolate temperature-driven heating behavior as a separate family. The preliminary hierarchical second-pass check improved heat-pump correlation from 0.349 to 0.546, reduced heat-pump NMAE from 0.565 to 0.525, and increased the residual share from 0.175 to 0.408. These results suggest that iterative recovery can reveal weaker structure, while reliable magnitude recovery for heat-pump and base-load components would require richer contextual information, such as occupancy indicators, appliance-use proxies, HVAC control states, indoor setpoints, improved baseline modeling, or a component-specific follow-up calibration stage.

The weak alignment with the base/inflexible reference component also reflects the experimental configuration. The dataset's base/inflexible load represents a simulated household-demand component, whereas the analytical baseline in the MD-ADD method represents a decomposition term defined by the selected baseline strategy. With baseline estimation disabled, the full aggregate load is treated as the attribution target, and the base/inflexible component serves as a post hoc reference signal. The resulting alignment reflects agreement with this simulated reference component rather than performance of the analytical baseline estimator.

The evaluation is based on post hoc comparison with withheld reference components. The model operates from aggregate and contextual inputs, and reference signals are used after decomposition to assess alignment. Metrics such as correlation, MAE, RMSE, and NMAE quantify post hoc agreement with withheld reference components in the simulated and measured evaluations. They provide alignment evidence rather than supervised training losses, causal validation, or complete-recovery guarantees. Family membership and feature-level shares should be interpreted as evidence structures within the observed feature space rather than as causal claims.

The pipeline is data-agnostic in the sense that it does not require predefined appliance models or labeled component data as model inputs. The primary empirical evaluation remains the simulated EV-focused household case, which provides the heterogeneous contextual variables and multiple withheld component references needed for detailed multi-driver analysis. The additional ADRENALIN evaluation extends the evidence to

nine measured commercial and public buildings, but it addresses a narrower task because only temperature-dependent reference loads and weather- and calendar-based contextual variables are available consistently across the buildings. The predecessor comparison uses an adapted implementation of the earlier MD-ADD formulation on the EV dataset and is intended to quantify methodological progression under a common case rather than to reproduce the original building-data experiments.

The measured-building results support applicability beyond the simulated case and provide evidence across nine commercial and public buildings. Alignment varied among buildings, and the effect of expanding the feature set was not uniform. These results indicate that successful attribution depends on whether the available contextual variables provide sufficiently distinctive evidence for the underlying load process. Further validation remains necessary to characterize performance across additional component classes, building types, operational conditions, climates, and sampling regimes.

The runtime analysis identifies relation discovery as the main computational bottleneck, accounting for more than 98% of execution time in all three scaling cases. Scalability improvements should prioritize precomputed rolling-window boundaries, cached pair-specific statistics, early pruning of low-information variables, sparse consensus representations, and distributed processing of independent feature pairs. Approximate variants could further reduce cost by evaluating fewer local windows, using coarser threshold and lag grids, or applying an inexpensive relation screen before the more costly mechanisms. The current implementation is sequential and batch-based; streaming or incremental operation would require rolling statistical summaries, bounded buffers, incremental relation updates, and periodic reconstruction of the feature-family structure.

The method is formulated for one aggregate target signal at a time, such as a single building, household, or metered site. Larger portfolios can be handled by running the pipeline separately across multiple meters or buildings. If the input target is a feeder-level aggregate, the method performs feeder-level driver attribution rather than separating individual buildings or households, unless site-specific features or metering information are included. Scaling requires both computational parallelization across sites and careful feature design when many sites are combined into one aggregate signal.

The present implementation supports offline or periodic analysis. Deployment in a BMS or energy-management workflow would require mapping available sensors and contextual variables to the expected input structure, applying data-quality checks before relation discovery, recalibrating thresholds and feature-family structures for the specific site, and validating outputs against available operational knowledge or temporary submetering where possible. The method exposes uncertainty through unexplained dynamic load, residual demand, sensitivity analysis, and failure-mode diagnostics. Formal probabilistic confidence intervals, streaming updates, and site-specific integration procedures remain important engineering extensions.

6.4. Future Work

Future work should prioritize the recovery of masked lower-magnitude components. One promising direction is iterative or hierarchical decomposition, where a dominant component is identified, its contribution is removed or conditioned on, and the relation analysis is repeated on the remaining signal. This may allow weaker structures, such as temperature-driven heating demand, to become more visible after dominant event-driven loads have been accounted for.

Feature representation should be extended. Additional contextual variables and more advanced temporal transformations may improve the ability to capture delayed, seasonal, or regime-dependent relationships. This is particularly relevant for heating and other

smooth components, where the relevant driver signal may not align with the aggregate target through a simple pointwise relationship.

Relation modeling could be improved through adaptive or learned scoring functions. Such extensions may reduce reliance on fixed relation definitions and allow the pipeline to capture dataset-specific dependency patterns more flexibly. Broader validation is needed across measured and simulated datasets with different load structures, sampling resolutions, contextual variables, and dominant component types. Comparisons with established NILM approaches would also help clarify where relation-supported attribution provides the greatest benefit relative to conventional disaggregation methods.

7. Conclusions

This paper presented a relation-aware pipeline for low-frequency load disaggregation under partial observability. The method does not rely on predefined appliance models, supervised component labels, or complete decomposition of the aggregate signal. Instead, it constructs an intermediate representation of the observed system through semantic feature typing, multi-mechanism relation discovery, feature-family grouping, target-centric relevance analysis, mechanism modeling, and conservative allocation. Load is assigned only where sufficient relational evidence is available, while unsupported variation is retained as residual demand.

In the primary EV-focused case study, the method recovered the dominant event-driven component with strong post hoc alignment. The selected EV-aligned family achieved a correlation of 0.990 and an NMAE of 0.100 against the withheld EV charging reference load. The heat-pump and base/inflexible components were less successfully recovered, with correlations of 0.349 and 0.159 and NMAE values of 0.565 and 0.895, respectively. These results show that the proposed relation-aware and event-focused allocation strategy is most effective for dominant, high-intensity, temporally localized load components, while smoother, lower-magnitude, or overlapping components remain more difficult to separate when their relation structure is masked by stronger signals.

The additional measured-building validation on nine ADRENALIN buildings showed that the method can also identify temperature-associated load structure in measured commercial and public building data. The restricted feature set achieved a median correlation of 0.846 and a median NMAE of 0.594, while the full feature set achieved a median correlation of 0.828 and a median NMAE of 0.576. These results support the broader applicability of the attribution logic beyond the simulated EV case, but they also show that task-specific methods can remain advantageous for specialized targets. On the primary EV dataset, the Bayesian thermal baseline achieved lower heat-pump NMAE than the heat-pump-associated MD-ADD family, with NMAE values of 0.278 and 0.565, respectively. In the ADRENALIN comparison, the previously published Bayesian Disaggregation benchmark retained a lower median NMAE of 0.540 for the temperature-dependent reference load. These results indicate that the proposed method should be interpreted as a general relation-aware attribution framework rather than as a replacement for specialized component-specific disaggregation models.

The comparison with the earlier MD-ADD formulation quantified the methodological progression introduced by the present pipeline. On the primary EV dataset, the adapted earlier formulation achieved an EV correlation of 0.556 and an EV NMAE of 0.974, whereas the present relation-aware pipeline achieved an EV correlation of 0.990 and an EV NMAE of 0.100. Heat-pump NMAE also improved from 0.831 to 0.565. These differences indicate that automated feature typing, heterogeneous relation discovery, graph-based family construction, mechanism-aware evidence modeling, conservative allocation, and event-background separation jointly improve the ability to recover dominant relation-supported structure.

The sensitivity analyses further clarified the operating boundaries of the method. EV charging remained identifiable under moderate input perturbations, but magnitude accuracy degraded under missing contextual observations, contextual-feature noise, and aggregate-meter noise. At 10% missing contextual observations, EV correlation remained 0.927, but EV NMAE increased to 0.352. With 5% contextual-feature noise and 5% aggregate-meter noise, EV NMAE increased to 0.214 and 0.226, respectively. Temporal aggregation also weakened the decomposition, with EV NMAE increasing to 0.689, 0.499, and 0.523 at 2 h, 4 h, and 6 h resolution. These findings indicate that the method can preserve dominant event-driven structure under moderate perturbation, but that attribution quality remains sensitive to temporal granularity, aggregate-signal quality, and the availability of informative contextual variables.

The weaker heat-pump and base-load results reflect identifiable failure modes rather than only parameter limitations. The EV component accounted for 0.718 of the target energy while being active in only 0.074 of timestamps, whereas the heat-pump component accounted for 0.185 of the energy while being active in 0.944 of timestamps. The base/inflexible component was active throughout the period and accounted for 0.097 of target energy. Although the true heat-pump load was strongly related to heating degree days, this relationship was largely masked in the aggregate signal and only partly re-emerged after the estimated EV-aligned contribution was removed. A preliminary hierarchical second-pass check improved heat-pump correlation from 0.349 to 0.546 and reduced heat-pump NMAE from 0.565 to 0.525, but increased the residual share from 0.175 to 0.408. This suggests that iterative recovery may help reveal weaker structure, but additional contextual information, improved baseline modeling, and component-specific calibration are likely needed for reliable recovery of smoother and lower-magnitude loads.

The results also reinforce the evidentiary nature of feature-family outputs. The estimated families are formed from statistical and relational structure in the available variables, while reference component signals are used after decomposition for post hoc evaluation. The retained unexplained dynamic load and residual demand are important outputs because they represent load variation with insufficient support from the observed feature families. This conservative treatment of unsupported variation is central to applying load attribution under partial observability.

The runtime analysis identified relation discovery as the main computational bottleneck. In the full runtime configuration, the total runtime was 2525.0 s, with relation discovery accounting for more than 98% of the computation. This indicates that larger-scale use would require implementation-level improvements such as feature pruning, caching, distributed pairwise processing, approximate relation screening, or parallel execution across buildings or meters. The present implementation is best interpreted as an offline or periodic analysis tool, while real-time or large-portfolio deployment would require further engineering development.

The proposed pipeline provides a modular and interpretable framework for driver-based load attribution when labeled component data are unavailable or incomplete. Its main demonstrated strength is the recovery of dominant event-driven load structure through heterogeneous relation discovery, conservative allocation, and structured feature-family interpretation. Its main limitations are weaker recovery of smoother or lower-magnitude components, sensitivity to temporal aggregation and missing driver evidence, and the computational cost of relation discovery. Future work should focus on iterative and hierarchical decomposition, richer contextual and operational feature sets, adaptive relation scoring, formal uncertainty quantification, streaming implementation, and broader validation across diverse measured and simulated datasets.

Author Contributions: Conceptualization, B.N.J., Z.G.M. and B.A.T.; methodology, B.N.J., Z.G.M. and B.A.T.; software, B.A.T.; validation, B.A.T., B.N.J. and Z.G.M.; formal analysis, B.A.T.; investigation, B.A.T.; resources, B.N.J. and Z.G.M.; data curation, B.A.T.; writing, original draft preparation, B.A.T.; writing, review and editing, B.N.J. and Z.G.M.; visualization, B.A.T.; supervision, B.N.J. and Z.G.M.; project administration, B.N.J.; funding acquisition, Z.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is part of the project titled "Automated Data and Machine Learning Pipeline for Cost-Effective Energy Demand Forecasting in Sector Coupling" (jr. Nr. RF-23-0039; Erhvervsfyrtårn Syd Fase 2), The European Regional Development Fund.

Data Availability Statement: Restrictions apply to the availability of the simulated dataset analyzed in this study. The simulation data were obtained from the study described in [7] and are available from the corresponding author upon reasonable request and with permission from the data owner. The ADRENALIN competition dataset used for the measured-building validation is publicly available through the ADRENALIN 2024 Load Disaggregation Competition site: https://codalab.lisn.upsaclay.fr/competitions/19659 (accessed on 19 June 2026). The complete ADRENALIN dataset, including the underlying submetering data from which the competition dataset was derived, is available on Zenodo: https://zenodo.org/records/19553414 (accessed on 19 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMI	Advanced metering infrastructure
ANOVA	Analysis of variance
BiLSTM	Bidirectional long short-term memory
CNN	Convolutional neural network
LSTM	Long short-term memory
EV	Electric vehicle
HDD	Heating degree day
HP	Heat pump
HVAC	Heating, ventilation, and air-conditioning
MAE	Mean absolute error
MD-ADD	Multi-driver automatic dependency disaggregation
MIC	Maximal information coefficient
MINE	Maximal information-based nonparametric exploration
NILM	Non-intrusive load monitoring
NMAE	Normalized mean absolute error
PV	Photovoltaic
REDD	Reference Energy Disaggregation Dataset
RMSE	Root mean squared error
SARIMAX	Seasonal autoregressive integrated moving average with exogenous variables
SHAP	SHapley Additive exPlanations
XGBoost	Extreme Gradient Boosting

References

1. International Energy Agency. *Developing a Global Energy Efficiency Workforce in the Buildings Sector*; International Energy Agency: Paris, France, 2024.
2. European Parliament Council of the European Union. *Directive (EU) 2024/1275 of 24 April 2024 on the Energy Performance of Buildings (Recast)*; European Parliament Council of the European Union: Brussels, Belgium, 2024.
3. Tolnai, B.A.; Ma, Z.; Jørgensen, B.N. A Scoping Review of Energy Load Disaggregation. In *Proceedings of the Progress in Artificial Intelligence*; Springer: Cham, Switzerland, 2023; pp. 209–221.

4. Elixie, A.E.; Alkahtani, A.; Alkawsi, G.; Fatihah, S.; Fazea, Y.; Ekanayake, J. Non-Intrusive Electrical Load Monitoring and Identification: Approaches, Tools and a Case Study. *Appl. Math. Inf. Sci.* **2020**, *14*, 1017–1027. [CrossRef]
5. Ehrhardt-Martinez, K.; Donnelly, K.A.; Laitner, S. *Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities*; American Council for an Energy-Efficient Economy: Washington, DC, USA, 2010.
6. Stankovic, L.; Stankovic, V.; Murray, D.; Liao, J. Energy feedback enabled by load disaggregation. In Proceedings of the 1st Energy Feedback Symposium: "Feedback in Energy Demand Reduction: Examining Evidence and Exploring Opportunities", Edinburgh, UK, 4–5 July 2016.
7. Christensen, K.; Jørgensen, B.N.; Ma, Z.G. Multi-agent Based Simulation for Decentralized Electric Vehicle Charging Strategies and Their Impacts. In *Proceedings of the Progress in Artificial Intelligence*; Springer: Cham, Switzerland, 2025; pp. 220–232.
8. Vavouris, A.; Garside, B.; Stankovic, L.; Stankovic, V. Low-Frequency Non-Intrusive Load Monitoring of Electric Vehicles in Houses with Solar Generation: Generalisability and Transferability. *Energies* **2022**, *15*, 2200. [CrossRef]
9. Mariño, C.; Martino, M.D.; Fernández, A. Non-Intrusive Disaggregation at a Very Low Frequency of Water Heaters and Electric Vehicles. In Proceedings of the 2024 IEEE PES Generation, Transmission and Distribution Latin America Conference and Industrial Exposition (GTDLA), Ixtapa, Mexico, 11–13 November 2024; pp. 1–6.
10. Ma, Z.G.; Jørgensen, B.N. Vehicle-to-Grid Demand Response Methodologies: Challenges, Applications, and Future Directions. In Proceedings of the 10th International Conference on Green Energy Technologies, Nagasaki, Japan, 20–22 July 2025.
11. Mariño, C.; Cossio, G.; Massaferro, P.; Martino, M.D.; Gómez, A.; Fernandez, A. NILMEV: Electric Vehicle disaggregation for residential customer energy efficiency incentives. In Proceedings of the 2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 16–19 January 2023; pp. 1–5.
12. James, A.; Lin, A.Z. Non-intrusive Electric-vehicle Load Disaggregation Algorithm for a Data-driven EV Integration Strategy. In Proceedings of the 2022 IEEE Conference on Technologies for Sustainability (SusTech), Riverside, CA, USA, 21–23 April 2022; pp. 235–242.
13. Martin, C.; Ke, F.; Wang, H. Non-Intrusive Load Monitoring for Feeder-Level EV Charging Detection: Sliding Window-Based Approaches to Offline and Online Detection. In Proceedings of the 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2), Hangzhou, China, 15–18 December 2023; pp. 4958–4963.
14. Tolnai, B.A.; Ma, Z.; Jørgensen, B.N. Comparison of Three Algorithms for Low-Frequency Temperature-Dependent Load Disaggregation in Buildings Without Submetering. In *Proceedings of the Energy Informatics*; Springer: Cham, Switzerland, 2026; pp. 355–370.
15. Chen, J.; Ding, L.; Zhang, K.; Hou, C.; Lai, Z. Deep Learning Based Air Conditioning Set Temperature Prediction with Meteorological Data. In Proceedings of the 2023 IEEE Sustainable Power and Energy Conference (iSPEC), Chongqing, China, 28–30 November 2023; pp. 1–6.
16. Kim, D.W.; Ahn, K.U.; Shin, H.; Lee, S.E. Simplified Weather-Related Building Energy Disaggregation and Change-Point Regression: Heating and Cooling Energy Use Perspective. *Buildings* **2022**, *12*, 1717. [CrossRef]
17. Hart, G.W. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* **1992**, *80*, 1870–1891. [CrossRef]
18. Zoha, A.; Gluhak, A.; Imran, M.A.; Rajasegarar, S. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors* **2012**, *12*, 16838–16866. [CrossRef] [PubMed]
19. Huber, P.; Calatroni, A.; Rumsch, A.; Paice, A. Review on Deep Neural Networks Applied to Low-Frequency NILM. *Energies* **2021**, *14*, 2390. [CrossRef]
20. Luo, Z.; Li, Z.; Zhang, T.; Wang, G.; Deng, W.J.; Li, J.H.; Liang, J.K. A low-frequency residential NILM approach based on adaptive event detection. *J. Build. Eng.* **2025**, *106*, 112543. [CrossRef]
21. Mądziel, M. Impact of Weather Conditions on Energy Consumption Modeling for Electric Vehicles. *Energies* **2025**, *18*, 1994. [CrossRef]
22. Benitez, I.B.; Repedro, K.I.; Aung, T. Forecasting Building Energy Consumption Using Statistical Models Incorporating Operational and Environmental Factors. In Proceedings of the 2024 19th International Conference on Emerging Technologies (ICET), Topi, Pakistan, 19–20 November 2024; pp. 1–6.
23. Mai, T.H.D.; Leung, C.K.; Lu, J.; Giesbrecht, N.; Hnylycia, O.A. A Data Science Solution to Integrate Weather Data for Energy Consumption Analysis. In Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI), San Jose, CA, USA, 7–9 August 2024; pp. 69–75.
24. Cao, H.; Ge, J.; Dong, W.; Guan, X. Application Research of Power Load Forecasting and Real-time Scheduling Based on SHAP Interpretable Combination Model. In Proceedings of the 2025 8th International Conference on Energy, Electrical and Power Engineering (CEEPE), Wuxi, China, 25–27 April 2025; pp. 1022–1026.
25. Zhang, Z.; Chen, Y.; Wang, H.; Wang, Y.; Fu, Q.; Lu, Y. A Multivariate Time Series Anomaly Detection Model Based on Graph Attention Mechanism in Energy Consumption of Intelligent Buildings. In Proceedings of the 2022 Tenth International Conference on Advanced Cloud and Big Data (CBD), Guilin, China, 4–5 November 2022; pp. 122–127.

26. Ghaffar, M.; Sheikh, S.R.; Naseer, N.; Usama, S.A.; Salah, B.; Alkhatib, S.A.K. Accuracy Improvement of Non-Intrusive Load Monitoring Using Voting-Based Consensus Clustering. *IEEE Access* **2023**, *11*, 53165–53175. [CrossRef]
27. Ghaffar, M.; Sheikh, S.R.; Naseer, N.; Din, Z.M.; Rehman, H.Z.; Naved, M. Non-Intrusive Load Monitoring of Buildings Using Spectral Clustering. *Sensors* **2022**, *22*, 4036. [CrossRef] [PubMed]
28. Hulsebos, M.; Hu, K.; Bakker, M.; Zgraggen, E.; Satyanarayan, A.; Kraska, T.; Demiralp, Ç.; Hidalgo, C. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1500–1508.
29. Elmobark, N.; El-Ghareeb, H.; Elhishi, S. BlueEdge neural network approach and its application to automated data type classification in mobile edge computing. *Sci. Rep.* **2025**, *15*, 43823. [CrossRef] [PubMed]
30. Saleem, M.H.; Taha, M.; Rehmani, M.A.A.; Tito, S.R.; Soltic, S.; Nieuwoudt, P.; Pandey, N.; Ahmed, M.D. A comprehensive review of machine learning and deep learning models for non-intrusive load monitoring: Performance, analyses, practical insights, and emerging trends. *Appl. Intell.* **2025**, *55*, 1020. [CrossRef]
31. Rafiq, H.; Manandhar, P.; Rodriguez-Ubinas, E.; Qureshi, O.A.; Palpanas, T. A review of current methods and challenges of advanced deep learning-based non-intrusive load monitoring (NILM) in residential context. *Energy Build.* **2024**, *305*, 113890. [CrossRef]
32. Li, H.; Zhai, Q.; Zhou, Y.; Mao, Y. An Integer Program based Appliance Identification and Load Disaggregation Method. In Proceedings of the 2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), Dubrovnik, Croatia, 14–17 October 2024; pp. 1–5.
33. Kim, H.; Marwah, M.; Arlitt, M.; Lyon, G.; Han, J. Unsupervised Disaggregation of Low Frequency Power Measurements. In Proceedings of the 2011 SIAM International Conference on Data Mining (SDM), Mesa, AZ, USA, 28–30 April 2011; pp. 747–758.
34. Kaselimi, M.; Protopapadakis, E.; Voulodimos, A.; Doulamis, N.; Doulamis, A. Towards Trustworthy Energy Disaggregation: A Review of Challenges, Methods, and Perspectives for Non-Intrusive Load Monitoring. *Sensors* **2022**, *22*, 5872. [CrossRef] [PubMed]
35. Henriet, S.; Simsekli, U.; Richard, G.; Fuentes, B. Energy disaggregation for commercial buildings: A statistical analysis. In Proceedings of the International Workshop on Non-Intrusive Load Monitoring (NILM2018), Austin, TX, USA, 7–8 March 2018; pp. 1–4.
36. Liu, Q.; Nakoty, F.M.; Wu, X.Y.; Anaadumba, R.; Liu, X.D.; Zhang, Y.H.; Qi, L.Y. A secure edge monitoring approach to unsupervised energy disaggregation using mean shift algorithm in residential buildings. *Comput. Commun.* **2020**, *162*, 187–195. [CrossRef]
37. Moradzadeh, A.; Sadeghian, O.; Pourhossein, K.; Mohammadi-Ivatloo, B.; Anvari-Moghaddam, A. Improving Residential Load Disaggregation for Sustainable Development of Energy via Principal Component Analysis. *Sustainability* **2020**, *12*, 3158. [CrossRef]
38. Deveci, Ş.A.; Günay, M. Unsupervised Pattern Extraction of Time Series Data for Energy Disaggregation. In Proceedings of the 2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 26–28 October 2024; pp. 1–6.
39. Wang, Y.; Ma, C.; Zhao, B.; Luan, W. Non-intrusive Electric Vehicle Charging Load Disaggregation based on Independent Component Analysis with Reference. In Proceedings of the 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2), Taiyuan, China, 22–24 October 2021; pp. 490–495.
40. Yang, N.; Wang, Y.Q.; Zhang, Y.N.; Yuan, D. Non-Intrusive Load Monitoring for Energy Management in Smart Grids Incorporating EVs. *IEEE Trans. Consum. Electron.* **2025**, *71*, 1696–1706. [CrossRef]
41. Yang, Y.; Li, R.; Huang, T. Smart Meter Data Analysis of a Building Cluster for Heating Load Profile Quantification and Peak Load Shifting. *Energies* **2020**, *13*, 4343. [CrossRef]
42. Kim, R.; Ko, Y.-D.; Kim, Y.-R. A data-driven framework for building energy benchmarking: Time-series decomposition, clustering and change-point regression. In Proceedings of the Building Simulation 2023: 18th Conference of IBPSA, Shanghai, China, 4–6 September 2023; pp. 750–757.
43. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
44. de Siqueira Santos, S.; Takahashi, D.Y.; Nakata, A.; Fujita, A. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief. Bioinform.* **2014**, *15*, 906–918. [CrossRef] [PubMed]
45. Li, N.W.; Wu, X.; Dong, J.J.; Zhang, D. A shape-based clustering algorithm and its application to load data. *Cogn. Comput. Syst.* **2023**, *5*, 109–117. [CrossRef]
46. Firoozjaei, M.D.; Kim, M.; Alhadidi, D. Time-Series Load Data Analysis for User Power Profiling. In Proceedings of the 2023 25th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 19–22 February 2023; pp. 382–387.

47. Zhu, D.; Sun, Y.; Cui, J.; Hu, Z.; Huang, J.; Zhang, J. Novel Power Load Pattern Recognition Algorithm Based on Unbalanced Data Clustering. In Proceedings of the 2023 3rd International Conference on New Energy and Power Engineering (ICNEPE), Huzhou, China, 24–26 November 2023; pp. 670–673.
48. Bouke, M.A.; Abdullah, A.; Frnda, J.; Cengiz, K.; Salah, B. BukaGini: A Stability-Aware Gini Index Feature Selection Algorithm for Robust Model Performance. *IEEE Access* **2023**, *11*, 59386–59396. [CrossRef]
49. Tolnai, B.A.; Ma, Z.G.; Jørgensen, B.N. Data-Driven Electricity Load Analysis in Smart Buildings: A Multi-Driver Automatic Dependency Disaggregation Approach. *Electronics* **2026**, *15*, 929. [CrossRef]
50. Tolnai, B.A.; Ma, Z.G.; Sartori, I.; Pandiyan, S.V.; Amos, M.; Bengtsson, G.; Lien, S.K.; Walnum, H.T.; Hameed, A.; Rajasekharan, J.; et al. ADRENALIN: Energy Data Preparation and Validation for HVAC Load Disaggregation in Commercial Buildings. In *Proceedings of the Energy Informatics*; Springer: Cham, Switzerland, 2026; pp. 321–337.
51. Massidda, L.; Marrocu, M. A Bayesian Approach to Unsupervised, Non-Intrusive Load Disaggregation. *Sensors* **2022**, *22*, 4481. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.