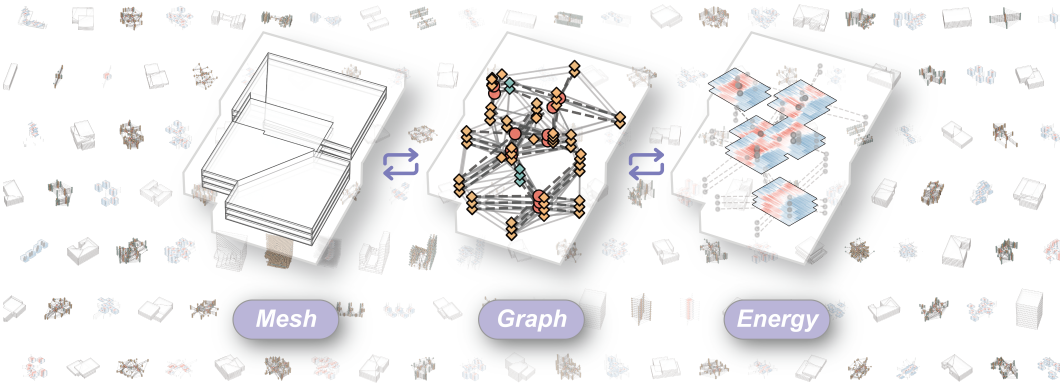
ARCHEGRAPH: A LARGE-SCALE GRAPH DATASET FOR GEOMETRY–TOPOLOGY–PHYSICS ALIGNED BUILDING ENERGY MODELING

Anonymous authors
Paper under double-blind review



Dataset Code

ABSTRACT

Learning transferable surrogates for building performance requires data that expose how geometry and spatial topology interact with environmental forcing, yet existing datasets rarely align these modalities at scale. We present **ArchEGraph**, a large-scale graph dataset aligning building geometry, topology, weather, and zone-level physical responses across 5,481 buildings and 49,326 validated building–weather simulations. The target is the EnergyPlus-derived external thermal load under fixed constructions and internal schedules, isolating variation attributable to geometry, topology, and weather rather than operational energy use. ArchEGraph supports two structured learning problems: (i) Mesh-to-Graph reconstruction, which infers local adjacency and globally consistent latent space partitions from unordered geometric primitives; and (ii) Graph-to-Energy prediction, which maps a static heterogeneous graph and an 8,760-step weather sequence to node-aligned response trajectories. Standardized protocols evaluate in-distribution performance and generalization across building and climate shifts. Results expose substantial open challenges: current set encoders recover local face adjacency reliably but struggle with global face–space incidence, particularly on irregular manually designed buildings. Moreover, shuffling only face–space assignments within the same architecture increases prediction error, directly demonstrating that correct topology provides useful information beyond model capacity. ArchEGraph thus provides a controlled benchmark for studying structured physical learning and transferable geometric inductive biases.

1 INTRODUCTION

Buildings account for a substantial share of global energy consumption and carbon emissions, making accurate and scalable building energy modeling (BEM) critical to sustainable design (Dyson et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2023). Physics-based Building Performance Simulation (BPS) remains the de facto standard (New et al., 2012), but annual simulations require solving coupled, geometry-dependent heat transfer, radiative exchange, and shading processes, limiting iterative design and large-scale analysis (Clarke, 2001; Zhang et al., 2021; Wang et al., 2023; Reinhart & Cerezo Davila, 2016).

In this paper, *external disturbance load* means the net heating or cooling imposed on each thermal zone by outdoor weather acting through walls, windows, roofs, and floors. We derive this signal from EnergyPlus after removing standardized internal gains from occupants, lighting, and equipment; it is a simulated thermal response to geometry and weather, not metered electricity or fuel consumption.

To mitigate this limitation, machine learning-based surrogate models have been developed as fast approximations of BPS (Roman et al., 2020; Von Krannichfeldt et al., 2025). However, most existing approaches rely on global feature vectors or sequence-based representations that ignore explicit spatial structure and physical connectivity (Emami & Graf, 2018; Kelly & Knottenbelt, 2015; Miller et al., 2020; ASHRAE, 2020). This results in two core generalization failures: (i) *geometric shift*, where models fail to extrapolate to unseen architectural layouts due to missing structural inductive bias (Dou & Zhang, 2025; Ribeiro et al., 2018); and (ii) *climatic shift*, where learned mappings degrade under differing weather distributions (Peng et al., 2025). These issues restrict current surrogate models to narrow design distributions and controlled benchmarks.

We argue that building energy prediction should be formulated as a structured physical learning problem, where energy transfer emerges from interactions between spatial zones and envelope components (New et al., 2012; ASHRAE, 2017). This naturally induces a graph representation in which nodes correspond to thermal zones and envelope elements, and edges encode geometric adjacency and physical coupling (Alymani et al., 2023; Wu et al., 2025; Weber et al., 2024; Li et al., 2026). Graph Neural Networks provide a principled framework for modeling such structured interactions (Scarselli et al., 2009; Jia et al., 2024); however, their effectiveness is limited by the absence of large-scale benchmarks that jointly capture geometry, topology, weather forcing, and time-resolved physical responses (Skeie et al., 2025; Emami et al., 2023; Krapf et al., 2023).

To address this gap, we introduce **ArchEGraph**, a large-scale benchmark for geometry–topology–physics coupled learning in building energy modeling. The dataset contains 5,481 building designs and 49,326 valid simulation instances, with over 1.33×10^5 spatial nodes and 1.44×10^6 envelope elements. ArchEGraph supports two tasks: (i) *M2G*, reconstructing building topology (the internal zone configuration that is needed for performance prediction) from geometric inputs; and (ii) *G2E*, predicting zone-level load dynamics under temporal weather conditions. All data is generated via physics-based simulation to ensure consistency across geometry, topology and physics, which is not jointly available in current real-world datasets at scale (Miller et al., 2020; Krapf et al., 2023; Zhu et al., 2025). Together with standardized evaluation protocols and controlled cross-building and cross-climate splits, this establishes a unified benchmark for structured physical learning in buildings.

2 Related Work

Geometry and graph representations for buildings. Buildings are heterogeneous physical systems whose energy behavior depends on both geometric form and spatial connectivity among spaces, envelope elements, openings, and boundary conditions. Existing building information models and simulation formats support building-element representation and physical-parameter storage (buildlingSMART International, 2024; León-Sánchez et al., 2022; New et al., 2012), while semantic schemas improve interoperability across building domains (Pauwels et al., 2017; Rasmussen et al., 2021; Balaji et al., 2018). Geometric graph abstractions support layout reasoning and spatial organization (van Engelenburg et al., 2025; Nauata et al., 2020; Alymani et al., 2023), BIM/geometry processing (Gan, 2022), and thermal-zone conversion (Mediavilla et al., 2023; Xiao et al., 2024). These developments provide important foundations for representing buildings as structured physical systems, especially when geometry, spatial relations, and boundary elements need to be modeled jointly.

Data-driven surrogates for energy prediction. Building performance can be estimated through physics-based simulation or data-driven surrogates. Physics-based simulation provides interpretable performance estimates, while data-driven surrogates enable faster prediction for large-scale design exploration and optimization. Surrogate models based on regressors, tree ensembles, MLPs, recurrent

networks, and hybrid approaches have been widely studied for building energy prediction (Roman et al., 2020; Von Krannichfeldt et al., 2025). More recently, graph neural networks have provided a natural way to model irregular building topologies through message passing (Scarselli et al., 2009), and have been applied to thermal-load and energy prediction tasks (Jia et al., 2024; Wu et al., 2025). Together, these studies show the potential of structure-aware learning methods for connecting building geometry, operational context, and energy performance.

Datasets for building geometry and energy modeling. Existing datasets only partially support geometry-aware energy learning, as summarized in Table 1. Recent 3D building geometry and layout datasets provide increasingly large and detailed geometric resources, covering exterior building models, point clouds, meshes, indoor layouts, and floor-plan annotations across different spatial scales (Peters et al., 2022; Zhu et al., 2025; Selvaraju et al., 2021; Wang et al., 2023; Cruz et al., 2021; Zheng et al., 2020). These datasets support reconstruction, semantic labeling, and layout estimation, but are rarely coupled with the physical-response data needed for energy modeling. Conversely, energy/load datasets provide measured or simulated time-series observations for forecasting, disaggregation, and benchmarking (Emami et al., 2023; Kelly & Knottenbelt, 2015; Murray et al., 2017; Makonin et al., 2016; Pecan Street Inc., 2018; Miller et al., 2020; ASHRAE, 2020), but are typically geometry-agnostic. Existing geometry-energy resources link the two only partially through reference templates, urban exterior models, static certificate labels, or synthetic operation data (Deru et al., 2011; León-Sánchez et al., 2022; Krapf et al., 2023; Li et al., 2021). This motivates datasets that jointly align 3D geometry, space-surface topology, weather/context variables, and time-resolved energy/load responses.

Table 1: Comparison of representative building geometry and energy/load datasets. *3D Geo* denotes building-level geometric representation (e.g., massing or envelope). *Topology* indicates space- or zone-level relationships between interior regions and their boundary surfaces. *Energy/Load* refers to energy or other physical response labels, and *Weather/Context* denotes climate or simulation boundary conditions.

Dataset	Scale	3D Geo	Topology	Energy/Load	Weather/Context	Temporal Resolution
3D building geometry and layout datasets						
3DBAG (Peters et al., 2022)	10M bldgs	✓	✗	–	–	Static
GlobalBuildingAtlas (Zhu et al., 2025)	2.7B bldgs	✓	✗	–	–	Static
BuildingNet (Selvaraju et al., 2021)	2k bldgs	✓	✗	–	–	Static
Building3D (Wang et al., 2023)	160k bldgs	✓	✗	–	–	Static
ZInD (Cruz et al., 2021)	1.5k homes	✓	✓	–	–	Static
Structured3D (Zheng et al., 2020)	3.5K houses	✓	✓	–	–	Static
Building energy and load datasets						
BuildingsBench (Emami et al., 2023)	900k bldgs	–	–	✓	✓	1h
UK-DALE (Kelly & Knottenbelt, 2015)	5 homes	–	–	✓	✗	6s
REFIT (Murray et al., 2017)	20 homes	–	–	✓	✗	8s
AMPds (Makonin et al., 2016)	1 home	–	–	✓	✓	1min
BDG2 (Miller et al., 2020)	1.6k bldgs	–	–	✓	✓	1h
ASHRAE GEP III (ASHRAE, 2020)	1.5k bldgs	–	–	✓	✓	1h
Datasets with both building geometry and energy labels						
PointER (Krapf et al., 2023)	1M bldgs	✓	✗	✓	✗	Static
DOE Ref. Buildings (Deru et al., 2011)	16 archetypes	✓	✗	✓	✓	1 h
AlphaBuilding (Li et al., 2021)	16 archetypes	✓	✗	✓	✓	10min
ArchEGraph (Ours)	**5.5k bldgs & 49k cases**	✓	✓	✓	✓	**1h**

3 ARCHEGRAPH DATASET

In this section, we first provide an overview of the dataset composition, followed by its formal data schema. Finally, we analyze the statistical characteristics of the dataset to demonstrate its diversity in terms of geometry, climate, and physical response. Detailed sample cases and the comprehensive construction pipeline are provided in Appendix A.

3.1 OVERVIEW

ArchEGraph is a graph-centered dataset that represents buildings as heterogeneous graphs, where geometric structures are encoded as graph topology and physical responses are aligned with graph

nodes under varying weather conditions. This unified representation enables learning-based modeling of geometry–topology–physics coupling.

The dataset comprises two complementary sources of building geometry: a curated subset (**ArchEGraph-M**, 580 manually designed cases) and a synthetic subset (**ArchEGraph-P**, 4,901 parametric-generated cases), resulting in a total of $N = 5{,}481$ unique building geometries. By pairing these buildings with up to $M = 64$ distinct global weather sequences across climate regions and retaining only quality-controlled outputs, we obtain $K = 49{,}326$ high-fidelity energy simulation cases.

Constructing ArchEGraph required robust mesh normalization, scalable simulation across building–weather combinations, and strict alignment of zone-level responses with graph nodes; implementation details are provided in Appendix A.2.

Figure 1 illustrates the overall data construction process. Building geometries are first converted into structured graph representations, which are then coupled with weather inputs and processed through energy simulation to generate zone-level physical responses. More detailed data generation procedures are described in the Appendix A.2.

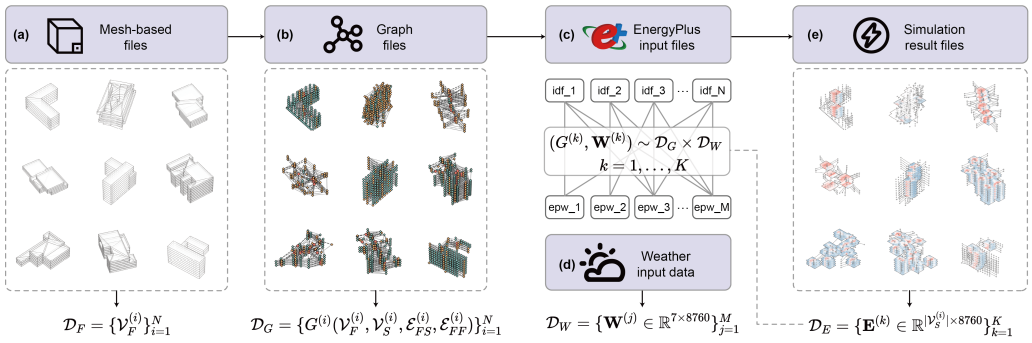


Figure 1: Overview of the ArchEGraph construction workflow. (a) Polygonal mesh files are converted into (b) heterogeneous graph representations, which are then coupled with (c) EnergyPlus input templates and (d) weather files for batch simulation. (e) The resulting aligned modalities form $\mathcal{D}_F$ (geometry), $\mathcal{D}_G$ (graph structure), $\mathcal{D}_W$ (weather), and $\mathcal{D}_E$ (zone-level simulation responses). Each energy simulation instance corresponds to a paired weather–building configuration $(G^{(k)}, \mathbf{W}^{(k)})$.

3.2 Data schema

ArchEGraph adopts a graph-centric abstraction in which the building topology serves as the structural backbone for aligning geometry, weather conditions, and energy responses. Building upon prior graph-based building representations (Li et al., 2026; Yu et al., 2026), we further extend the framework to support weather-aware analysis while preserving the original geometric–spatial abstraction.

Building representation ($\mathcal{D}_G$) The building dataset is represented as a collection of heterogeneous graphs, denoted by $\mathcal{D}_G = \{G^{(i)}\}_{i=1}^N$. Each graph $G^{(i)} = (\mathcal{V}_F^{(i)}, \mathcal{V}_S^{(i)}, \mathcal{E}_{FF}^{(i)}, \mathcal{E}_{FS}^{(i)})$ consists of face-level nodes $\mathcal{V}_F$, spatial nodes $\mathcal{V}_S$, face–face relations $\mathcal{E}_{FF}$, and face–space relations $\mathcal{E}_{FS}$. This representation jointly encodes geometric decomposition and spatial semantics, while remaining compatible with weather-conditioned physical modeling.

Geometry source ($\mathcal{D}_F$) Each graph is derived from a polygonal mesh list stored in the geometry dataset, forming the geometric source representation $\mathcal{D}_F = \{\mathcal{V}_F^{(i)}\}_{i=1}^N$, where the segmented mesh faces $\mathcal{V}_F$ provide the raw geometric basis from which graph nodes and connectivity are constructed.

Weather conditions ($\mathcal{D}_W$) We incorporate $M = 64$ global hourly weather sequences $\mathcal{D}_W = \{\mathbf{W}^{(j)} \in \mathbb{R}^{7\times 8760}\}_{j=1}^M$, each sequence includes seven critical climatic variables (e.g., dry-bulb temperature, global horizontal radiation, wind speed) that drive the thermal dynamics of the buildings.

Energy responses ($\mathcal{D}_E$) The energy dataset aligns simulation outputs with the spatial nodes of the graph. For each building-weather pair $(G^{(k)}, \mathbf{W}^{(k)})$, we record hourly zone-level load responses in $\mathcal{D}_E = \{\mathbf{E}^{(k)} \in \mathbb{R}^{|\mathcal{V}_S^{(i)}| \times 8760}\}_{k=1}^{K}$.

This alignment enables supervised learning on graphs, where the model learns to predict the temporal evolution of physical states based on static graph topology and dynamic external forcing.

3.3 Statistical characteristics

We performed a comprehensive statistical analysis focusing on three dimensions: geometric & topological complexity, climatic coverage, and physical response variance. Detailed results can be found in Appendix A.4.

Geometric complexity We evaluate the structural diversity of ArchEGraph through the lens of graph scale, connectivity, and hierarchical organization. As summarized in Table 2, the dataset encompasses over 1.4M face nodes and 133k spatial nodes, establishing a high-density representation of architectural environments. The subset ArchEGraph-P provides dense coverage of parametric variations, while ArchEGraph-M introduces long-tail instances with complex, irregular topologies (see Figure 2).

Table 2: Summary of ArchEGraph scale and metrics.

Graph element	P	M	All(P+M)
Face nodes $\lvert\mathcal{V}_F\rvert$	1,242,324	203,853	1,446,177
Space nodes $\lvert\mathcal{V}_S\rvert$	119,265	14,209	133,474
FF edges $\lvert\mathcal{E}_{FF}\rvert$	3,846,552	553,239	4,399,791
FS edges $\lvert\mathcal{E}_{FS}\rvert$	1,422,826	140,943	1,563,769

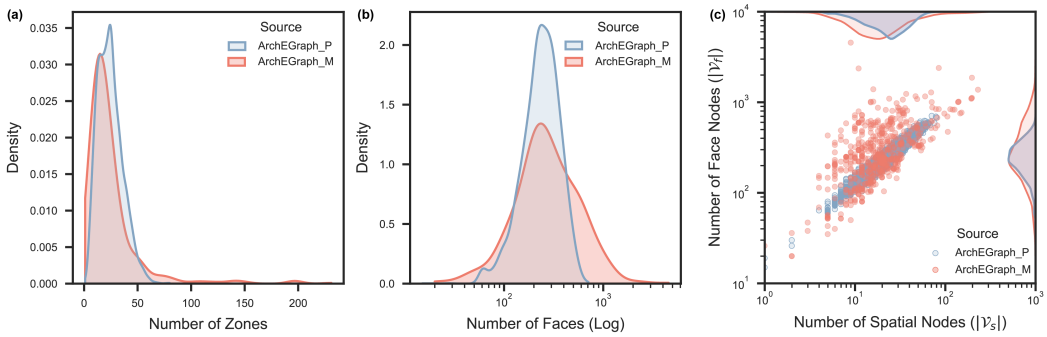


Figure 2: Cross-split geometric statistics of ArchEGraph-P and ArchEGraph-M. (a) Kernel density of the number of spatial nodes (zones) per building. (b) Kernel density of face-node counts (log-scaled axis). (c) Joint distribution between spatial-node count $|\mathcal{V}_S|$ and face-node count $|\mathcal{V}_F|$ with marginal densities, showing scale coupling and diversity across the two subsets.

Building scales span nearly three orders of magnitude, with average $|\mathcal{V}_F| = 263.9$ and $|\mathcal{V}_S| = 24.5$. The joint distribution in Figure 2c reveals a strong coupling between spatial and face-node counts alongside substantial variance. This multi-modal distribution challenges models to generalize across varying mesh resolutions and spatial partitions, ensuring robustness to both synthetic regularities and real-world architectural idiosyncrasies; additional discussion is provided in Appendix A.4.1.

Climatic coverage ArchEGraph contains hourly weather sequences for 64 cities spanning all inhabited continents and ASHRAE climate regions, represented by seven meteorological variables that capture thermal, moisture, solar, and wind conditions (Figure 3). Their distinct profiles provide broad environmental forcing for climate-conditioned learning and cross-climate evaluation. Additional climate-coverage and climate–load analyses are provided in Appendix A.4.2.

Physical response variance The simulated building energy responses exhibit substantial variability across time and space, associated with coupled interactions between climate conditions, building geometry, and dynamic thermal processes. Based on EnergyPlus simulations across 64 cities covering all ASHRAE climate regions (ASHRAE, 2021) (shown in Figure 4a), the results reveal distinct differences in energy response profiles in terms of magnitude and temporal characteristics, highlighting the joint influence of external weather and intrinsic building properties.

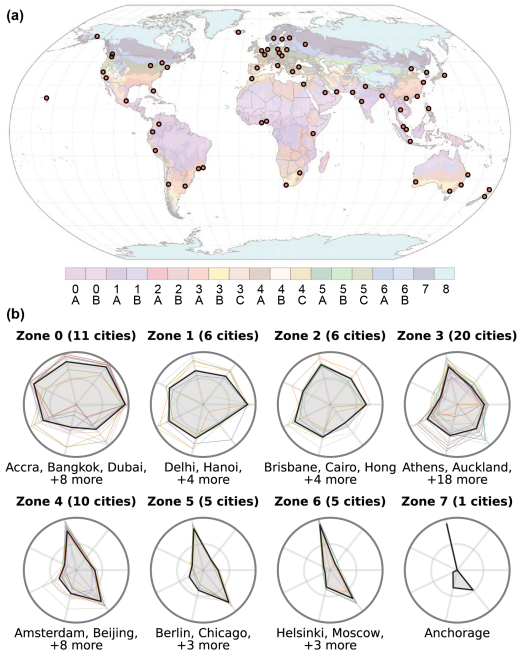


Figure 3: Weather profiles across 64 global cities in the dataset. (a) Geographic location of the 64 cities in the ASHRAE climate region map (ASHRAE, 2021). (b) Radar map of 7 weather attributes clustering across climate regions.

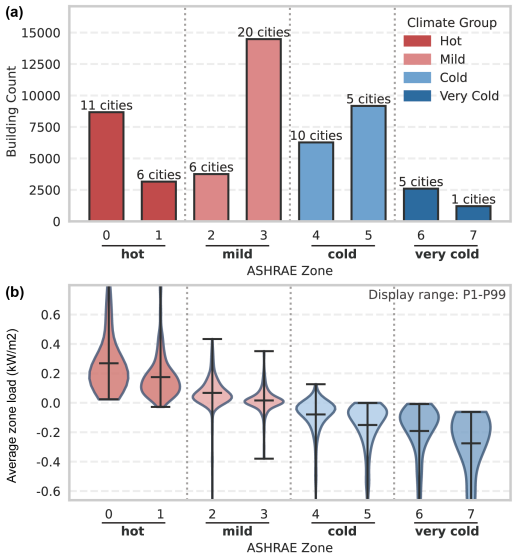


Figure 4: Building samples and energy load across climate regions. (a) Distribution of building samples, showing the total building count and the number of unique cities represented in each climate region. (b) Average zone load intensity, illustrating the distribution of average zone loads ($\mathrm{kW/m^2}$) across climate classifications ranging from hot to very cold.

In Figure 4b, the distribution of annual average building loads, defined here as external thermal load, exhibits clear climate-dependent patterns; lower values correspond to higher outward heat transfer. The systematically varying load levels and variances across climate regions indicate consistent coupling between climatic conditions and simulated thermal responses. More details are provided in Appendix A.4.2.

Overall, these results suggest that the dataset captures physically consistent climate–energy relationships while preserving substantial intra-class variability. Such properties provide a reliable basis for evaluating geometry-aware and climate-generalizable energy prediction models.

4 Tasks

We define two tasks on the proposed aligned dataset (Figure 5): Mesh-to-Graph reconstruction (M2G) and Graph-to-Energy prediction (G2E).

4.1 Mesh-to-Graph reconstruction (M2G)

We formulate Mesh-to-Graph reconstruction as a structured prediction task that infers a heterogeneous building graph from geometric primitives. Given a building composed of a set of faces $\mathcal{V}_F$, each face $f \in \mathcal{V}_F^{(i)}$ is represented by geometric features (c, R, s), where $c \in \mathbb{R}^3$ denotes the centroid, $R \in \mathbb{R}^{3\times3}$ denotes the rotation matrix, and $s \in \mathbb{R}^3$ denotes the oriented bounding box size. No topological or semantic information is provided as input. We aim to learn a function:

$$f_\theta : \mathcal{V}_F^{(i)} \rightarrow G^{(i)} = (\mathcal{V}_S^{(i)}, \mathcal{V}_F^{(i)}, \mathcal{E}_{FF}^{(i)}, \mathcal{E}_{FS}^{(i)}), \tag{1}$$

where $\mathcal{V}_S^{(i)}$ denotes space nodes, $\mathcal{E}_{FF}^{(i)}$ represents face–face adjacency relations, and $\mathcal{E}_{FS}^{(i)}$ encodes face–space incidence relations. Supervision is provided by ground-truth building decompositions, where each face is assigned to a unique space, and both $\mathcal{E}_{FF}^{(i)}$ and $\mathcal{E}_{FS}^{(i)}$ are derived from annotated topological relationships.

4.2 GRAPH-TO-ENERGY PREDICTION (G2E)

We formulate Graph-to-Energy prediction as a supervised spatio-temporal learning task over the aligned dataset $\mathcal{D}_G, \mathcal{D}_W, \mathcal{D}_E$. Given a building graph $G^{(k)} \in \mathcal{D}_G$ and its corresponding weather sequence $\mathbf{W}^{(k)} \in \mathcal{D}_W$, the objective is to learn a mapping function:

$$f_\theta : (G^{(k)}, \mathbf{W}^{(k)}) \rightarrow \mathbf{E}^{(k)}, \tag{2}$$

where $\mathbf{E}^{(k)} \in \mathbb{R}^{|\mathcal{V}_S^{(k)}| \times 8760}$ denotes the zone-level hourly load responses aligned with spatial nodes in the graph. This formulation explicitly models the coupling between heterogeneous graph structure, external climatic forcing, and temporal physical responses, enabling supervised learning of building energy dynamics under varying environmental conditions.

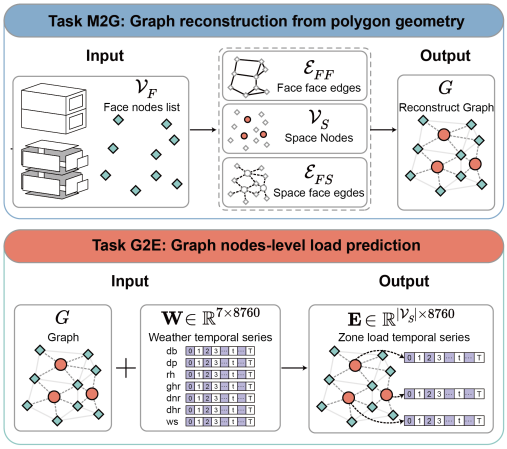


Figure 5: Task definitions in ArchEGraph. Top: M2G Task, predicting $\mathcal{D}_G$ from polygon face inputs $\mathcal{D}_F$; Bottom: G2E Task, predicting zone-level hourly load $\mathcal{D}_E$ by using graph $\mathcal{D}_G$ and weather condition $\mathcal{D}_W$

5 EVALUATION

We evaluate the proposed dataset along three dimensions: (i) task-wise performance on structured reconstruction and prediction, and (ii) generalization under cross-building and cross-climate distribution shifts. For more details, see Appendix B.

5.1 BASELINES

We define the models and evaluation metrics for both benchmark tasks here; detailed full model descriptions and metric definitions are provided in Appendix B.2.

M2G Baselines For M2G, we group baselines into two categories aligned with the topology reconstruction objective: (1) **Neural set baselines**: MLP, which applies pointwise linear transformations with no token interaction, and DeepSets (Zaheer et al., 2017), which aggregates set-level context via a sum-pooling encoder; (2) **Attention-based set baselines**: Perceiver (Jaegle et al., 2021) and SetTransformer (Lee et al., 2019), which leverage cross-attention or inducing-point attention to model set interactions, and TopoTransformer, which applies a standard Transformer encoder directly over face tokens. All five models share identical prediction heads for topology reconstruction and differ only in the face encoder design. We evaluate all methods on three topology components, namely face-face adjacency ($\mathcal{E}_{FF}$), space-node reconstruction ($\mathcal{V}_S$), and space-face relations ($\mathcal{E}_{FS}$), using F1 score and accuracy (Acc).

G2E Baselines For G2E, we compare three baseline categories: (1) the weather-only **WeatherMLP**; (2) heterogeneous face-to-space models **F2S** and **F2S-Attr**; and (3) spatial-operator variants **F2S-GATv2** (Brody et al., 2022), **F2S-TransConv** (Shi et al., 2021), and **F2S-GPS** (Rampášek et al., 2022). The operator variants use identical input modalities, weather encoder, and temporal GRU decoder, differing only in their spatial operator; full architectures and metrics are provided in Appendix B.2.

5.2 MAIN RESULTS

M2G Results: Graph Reconstruction We evaluate the ability of different methods to reconstruct heterogeneous building graphs from geometric inputs. The results demonstrate how effectively each method captures geometric and topological relationships.

Results in Table 3 show three consistent trends. First, ArchEGraph-P consistently outperforms ArchEGraph-M across all metrics, indicating a clear distribution shift from parametric to manually designed geometries. Second, task difficulty is ordered as $\mathcal{E}_{FF}$ (easiest), $\mathcal{V}_S$ (intermediate), and $\mathcal{E}_{FS}$

Table 3: Graph reconstruction performance results on ArchEGraph-P and ArchEGraph-M.

ArchEGraph-P						
Method	$\mathcal{E}_{FF}$ (F1)	$\mathcal{E}_{FF}$ (Acc)	$\mathcal{V}_S$ (F1)	$\mathcal{V}_S$ (Acc)	$\mathcal{E}_{FS}$ (F1)	$\mathcal{E}_{FS}$ (Acc)
MLP	0.9711 ±0.0003	0.9703 ±0.0003	0.9263 ±0.0067	0.9501 ±0.0039	0.3442 ±0.0119	0.8385 ±0.0062
DeepSets	0.9726 ±0.0008	0.9719 ±0.0008	0.9282 ±0.0057	0.9520 ±0.0028	0.3648 ±0.0086	0.8410 ±0.0053
Perceiver	0.9665 ±0.0099	0.9654 ±0.0107	0.9049 ±0.0193	0.9335 ±0.0135	0.3042 ±0.0262	0.8429 ±0.0192
SetTransformer	**0.9758** ±0.0007	**0.9753** ±0.0007	0.9422 ±0.0007	0.9599 ±0.0007	**0.4317** ±0.0035	**0.8810** ±0.0021
TopoTransformer	0.9754 ±0.0005	0.9748 ±0.0005	**0.9443** ±0.0140	**0.9616** ±0.0110	0.4308 ±0.0022	0.8796 ±0.0020
ArchEGraph-M						
Method	$\mathcal{E}_{FF}$ (F1)	$\mathcal{E}_{FF}$ (Acc)	$\mathcal{V}_S$ (F1)	$\mathcal{V}_S$ (Acc)	$\mathcal{E}_{FS}$ (F1)	$\mathcal{E}_{FS}$ (Acc)
MLP	0.8750 ±0.0116	0.8742 ±0.0106	0.7956 ±0.0166	0.8795 ±0.0087	0.1955 ±0.0043	0.8363 ±0.0167
DeepSets	0.8930 ±0.0083	0.8882 ±0.0037	0.8012 ±0.0169	0.8886 ±0.0186	0.2200 ±0.0099	0.8404 ±0.0164
Perceiver	0.9009 ±0.0027	0.8827 ±0.0043	0.7533 ±0.0101	0.8599 ±0.0072	0.1883 ±0.0018	0.8467 ±0.0084
SetTransformer	0.9006 ±0.0086	**0.8951** ±0.0115	**0.8018** ±0.0126	**0.8970** ±0.0107	**0.2508** ±0.0007	**0.8589** ±0.0019
TopoTransformer	**0.9015** ±0.0087	0.8949 ±0.0114	0.7829 ±0.0072	0.8912 ±0.0036	0.2360 ±0.0089	0.8578 ±0.0007

(hardest), with the gap most pronounced on ArchEGraph-M. Third, **SetTransformer** and **TopoTransformer** achieve the best overall performance; **SetTransformer** is more robust on ArchEGraph-M, while **TopoTransformer** remains competitive across components.

The difficulty of $\mathcal{E}_{FS}$ stems from its combinatorial nature: correctly assigning each face to a unique space requires resolving global partition consistency across all face–space pairs simultaneously, whereas $\mathcal{E}_{FF}$ and $\mathcal{V}_S$ are comparatively local or count-level predictions. In irregular manually designed geometries (ArchEGraph-M), ambiguous boundary faces shared between zones further increase the error rate, as local geometric features alone are insufficient to determine zone membership without global context.

G2E Results: Energy Prediction We evaluate spatial-temporal energy prediction performance on reconstructed building graphs.

Table 4: Energy prediction and controlled topology ablation on ArchEGraph-P and ArchEGraph-M. Shuffled uses the same F2S-Attr architecture but randomly reassigns face–space incidence; uncertainty is shown where multiple training seeds are available.

Method	**ArchEGraph-P**			**ArchEGraph-M**		
	MAE(W/m^2)	MSE(W^2/m^4)	RMSE(W/m^2)	MAE(W/m^2)	MSE(W^2/m^4)	RMSE(W/m^2)
WeatherMLP	8.845±0.548	4.481±0.717	6.694±0.884	7.540±0.867	2.257±0.963	4.750±1.460
F2S	5.781±0.332	3.910±0.108	6.253±0.781	5.159±0.365	2.038±0.122	4.515±0.811
F2S-Attr	3.965±0.138	1.289±0.043	3.590±0.610	4.440±0.360	1.926±0.056	4.389±0.628
F2S-Attr (Shuffled)	4.470±0.224	1.514±0.087	3.917±0.677	5.889±0.421	2.722±0.390	5.341±1.005
F2S-GATv2	4.331±0.221	1.369±0.052	3.699±0.337	4.761±0.274	1.941±0.061	4.405±0.298
F2S-TransConv	**3.124**±0.148	**1.100**±0.010	**3.317**±0.154	4.465±0.181	1.958±0.015	4.425±0.165
F2S-GPS	3.290±0.163	1.141±0.018	3.378±0.192	**3.425**±0.147	**1.687**±0.026	**4.108**±0.173

Table 4 shows that graph-aware models consistently outperform the weather-only baseline on both subsets. Among the spatial operators, F2S-TransConv performs best on ArchEGraph-P, whereas F2S-GPS achieves the lowest errors on the more irregular ArchEGraph-M. To isolate the value of topology from model capacity, we shuffle only the destination space assignments of face–space edges in F2S-Attr while keeping its architecture, features, parameters, and training procedure unchanged. This intervention increases MAE by 12.74% on ArchEGraph-P and 32.64% on ArchEGraph-M, demonstrating that correct face–space incidence provides predictive information beyond weather and model capacity.

5.3 Generalization

We evaluate out-of-distribution generalization under two structured shifts in ArchEGraph: cross-building transfer and cross-climate transfer. Detailed subset splits are provided in Appendix A.3. For a fair comparison, all metrics are reported on a normalized scale, which differs from the baseline setting.

Cross-building generalization We train on ArchEGraph-P and test on ArchEGraph-M, isolating structural distribution shift. As shown in Fig. 6(a), all graph-aware models outperform **WeatherMLP**,

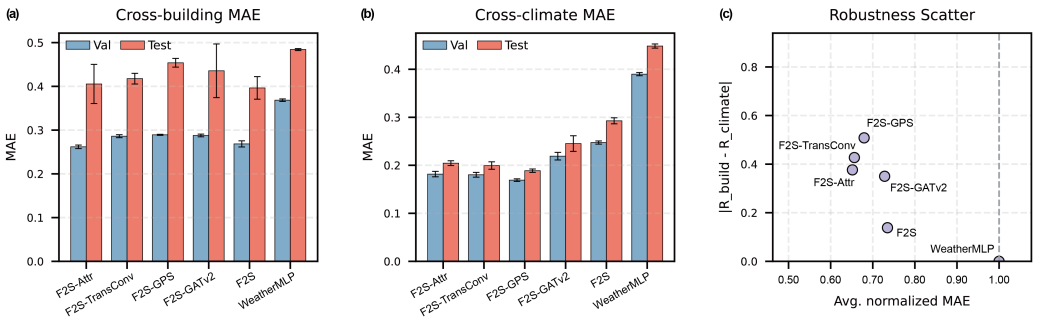


Figure 6: Cross-task generalization of energy prediction models. (a) Cross-building MAE (validation vs. test). (b) Cross-climate MAE (validation vs. test). (c) Accuracy-robustness trade-off. The x-axis shows the average normalized test MAE across both tasks; the y-axis shows cross-task inconsistency $|R_{\text{building}} - R_{\text{weather}}|$, where $R_{\text{task}} = \text{MAE}^{\text{task}}_{\text{model}}/\text{MAE}^{\text{task}}_{\text{WeatherMLP}}$ is the test MAE of each model normalized by the WeatherMLP baseline under the same task. Lower is better in all panels.

confirming that structural inductive bias remains beneficial under geometry shift. **F2S** achieves the best test MAE, while **F2S-GPS** degrades most severely despite its strong in-distribution performance.

Cross-climate generalization We train on a subset of ASHRAE climate regions and test on disjoint zones. Fig. 6(b) shows graph-aware models consistently outperform **WeatherMLP**, with **F2S-GPS** achieving the best test MAE. The gap against **WeatherMLP** is markedly larger than in the cross-building setting, indicating that graph structure is especially beneficial for climate generalization.

As shown in Fig. 6(c), the two shifts reveal a robustness trade-off: **F2S-GPS** achieves the strongest average accuracy but the highest cross-task inconsistency, while **F2S** offers the best balance. Graph topology consistently improves generalization across both shifts, with cross-building transfer remaining the harder regime.

6 Limitations and Future Work

The dataset of ArchEGraph is limited in its coverage of architectural typologies; it only incorporates office buildings, intentionally tailoring the benchmark to a single well-defined typology for controlled evaluation. The varied building geometries in the current dataset, show the robustness of the graph-based representation method and the open-source tools will allow researchers to extends the current dataset to incorporate residential and industrial buildings. Furthermore, energy simulations follow standardized construction templates and are fully simulation-based rather than measured—a deliberate choice, as building performance is inherently a forward-simulation process and consistent real-world observations at scale are unavailable at early design stages. Future work could incorporate material property variation and calibrate the data with real-world measurement to reduce simulation bias. Finally, the fixed computational budget may constrain evaluation of more expressive models; scaling to larger architectures and developing efficient graph learning methods for irregular building geometries remain open challenges.

7 Conclusion

We introduced ArchEGraph, a large-scale benchmark for geometry–topology–physics aligned learning for building energy modeling. By combining graph-structured building representations, aligned weather and load responses, and standardized M2G/G2E evaluation tasks, ArchEGraph supports the systematic study of how building geometry relates to energy use. The research reveals fundamental challenges in generalizing across irregular geometries and diverse climate conditions, showing the limitations of existing methods in capturing the coupling between spatial configuration and performance. We expect ArchEGraph to serve as a foundation for evaluating physically grounded learning approaches and to drive future research toward scalable, interpretable, and cross-domain models for design–simulation integration.

8 AI USE STATEMENT

In this work, we used ChatGPT (GPT-5.6 Sol) to edit the manuscript for clarity, grammar, and readability. We also used this tool to generate the implementation of the WeatherMLP baseline; the authors reviewed, tested, and validated the generated code. Generative AI tools were not used to formulate research questions or hypotheses, design the methodology or experiments, generate, clean, or analyze data, develop theoretical claims or proofs, interpret results, or draw conclusions. The authors reviewed all AI-assisted text and code and take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REFERENCES

Ali Alymani, Wassim Jabi, and Padraig Corcoran. Graph machine learning classification using architectural 3d topological models. *Simulation*, 99(11):1117–1131, 2023. doi: 10.1177/00375497221105894.

ASHRAE. *1993 ASHRAE Handbook: Fundamentals*. ASHRAE, Atlanta, GA, USA, si ed. edition, 1993. ISBN 978-0-910110-97-6.

ASHRAE. *ASHRAE Handbook - Fundamentals*. ASHRAE, Atlanta, GA, 2017. ISBN 978-1-939200-58-7.

ASHRAE. Ashrae great energy predictor iii dataset. Kaggle Competition, 2020. URL https://www.kaggle.com/competitions/ashrae-energy-prediction.

ASHRAE. Climatic data for building design standards. In *ASHRAE Handbook—Fundamentals*, chapter 14. ASHRAE, Atlanta, GA, 2021.

Bharathan Balaji, Arka Bhattacharya, Gabe Fierro, Jingkun Gao, Joshua Gluck, Dezhi Hong, Aslak Johansen, Jason Koh, Joern Ploennigs, Yuvraj Agarwal, Mario Bergés, David Culler, Rajesh K. Gupta, Mikkel Baun Kjærgaard, Mani Srivastava, and Kamin Whitehouse. Brick: Metadata schema for portable smart building applications. *Applied Energy*, 226:1273–1292, 2018. doi: 10.1016/j.apenergy.2018.02.091.

Mahabir S. Bhandari, Som S. Shrestha, and Joshua Ryan New. Evaluation of weather datasets for building energy simulation. *Energy and Buildings*, 49:109–118, Jun 2012. doi: 10.1016/j.enbuild.2012.01.033.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*. OpenReview.net, April 2022. doi: 10.48550/arXiv.2105.14491. URL https://openreview.net/forum?id=F72ximsx7C1. Virtual Event.

buildlingSMART International. Industry foundation classes (ifc) for data sharing in the construction and facility management industries – part 1: Data schema, 2024. URL https://www.iso.org/standard/84123.html.

Joseph A. Clarke. *Energy Simulation in Building Design*. Butterworth-Heinemann, Oxford, UK, 2 edition, 2001. ISBN 978-0-7506-5082-3.

Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2133–2143, 2021. doi: 10.1109/CVPR46437.2021.00217.

Michael Deru, Kristin Field, Daniel Studer, Kyle Benne, Brent Griffith, Paul Torcellini, Bing Liu, Mark Halverson, Dave Winiarski, Michael Rosenberg, Mehry Yazdanian, Joe Huang, and Drury Crawley. U.S. Department of Energy Commercial Reference Building Models of the National Building Stock. Technical report, National Renewable Energy Laboratory, 2011.

Hongwen Dou and Kun Zhang. Transfer learning for cross-building forecasting of building energy and indoor air temperature in model predictive control applications. *Journal of Building Engineering*, 111:113341, October 2025. ISSN 2352-7102. doi: 10.1016/j.jobe.2025.113341.

Anna Dyson, Naomi Keena, Mae-ling Lokko, Barbara Reck, and Christina Ciardullo. Building Materials and the Climate: Constructing a New Future. Technical Report 978-92-807-4064-6, United Nations Environment Programme, 2023.

Patrick Emami and Peter Graf. Buildingsbench: A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. Open Energy Data Initiative (OEDI), National Renewable Energy Laboratory, 2018.

Patrick Emami, Abhijeet Sahu, and Peter Graf. BuildingsBench: A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) – Datasets and Benchmarks Track*, pp. 19823–19857, New Orleans, LA, USA, December 2023. Neural Information Processing Systems Foundation, Inc. doi: 10.52202/075280-0871. Datasets & Benchmarks Track.

Vincent J. L. Gan. Bim-based graph data model for automatic generative design of modular buildings. *Automation in Construction*, 134:104062, 2022. doi: 10.1016/j.autcon.2021.104062.

Andrew Jaegle, Felix Gimeno, Andy Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664, Virtual Event, July 2021. PMLR. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Yilong Jia, Jun Wang, M. Reza Hosseini, Wenchi Shou, Peng Wu, and Chao Mao. Temporal graph attention network for building thermal load prediction. *Energy and Buildings*, 321:113507, October 2024. ISSN 0378-7788. doi: 10.1016/j.enbuild.2023.113507.

Jack Kelly and William J. Knottenbelt. The uk-dale dataset: Domestic appliance-level electricity demand and whole-house demand from five uk homes. *Scientific Data*, 2:150007, 2015. doi: 10.1038/sdata.2015.7.

Sebastian Krapf, Kevin Mayer, and Martin Fischer. Points for energy renovation (PointER): A point cloud dataset of a million buildings linked to energy features. *Scientific Data*, 10(1):639, September 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02544-x.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, Long Beach, CA, USA, June 2019. PMLR. URL https://proceedings.mlr.press/v97/lee19d.html.

Camilo León-Sánchez, Diego Giannelli, Giorgio Agugiaro, and Jantien Stoter. Creation of a CityGML-Based 3D City Model Testbed for Energy-Related Applications. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLVIII-4/W5-2022, pp. 97–104, 2022. doi: 10.5194/isprs-archives-XLVIII-4-W5-2022-97-2022.

Han Li, Zhe Wang, and Tianzhen Hong. A synthetic building operation dataset. *Scientific Data*, 8(1): 213, August 2021. doi: 10.1038/s41597-021-00989-6.

Yihui Li, Jun Xiao, Hao Zhou, and Borong Lin. From geometry to graph: Automation of building performance modeling via convex graph encoding. *Automation in Construction*, 183:106815, March 2026. ISSN 0926-5805. doi: 10.1016/j.autcon.2026.106815.

Stephen Makonin, Bradley Ellert, Ivan V. Bajić, Fred Popowich, Bram Gill, and Linda Bartram. Electricity, water, and natural gas consumption of a residential house in canada from 2012 to 2014. *Scientific Data*, 3:160037, 2016. doi: 10.1038/sdata.2016.37.

Asier Mediavilla, Peru Elguezabal, and Natalia Lasarte. Graph-based methodology for multi-scale generation of energy analysis models from IFC. *Energy and Buildings*, 282:112795, March 2023. ISSN 0378-7788. doi: 10.1016/j.enbuild.2023.112795.

Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition. *Scientific Data*, 7(1):368, 2020. doi: 10.1038/s41597-020-00712-x.

David Murray, Lina Stankovic, and Vladimir Stankovic. An electrical load measurements dataset of united kingdom households from a two-year longitudinal study. *Scientific Data*, 4:160122, January 2017. doi: 10.1038/sdata.2016.122.

Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. House-gan: Relational generative adversarial networks for graph-constrained house layout generation. In *Computer Vision – ECCV 2020*, pp. 162–177. Springer, 2020. doi: 10.1007/978-3-030-58452-8_10.

Joshua R. New, Mark B. Adams, Brian A. Fricke, Bo Shen, and Therese K. Stovall. Energyplus. Oak Ridge National Laboratory, June 2012. URL https://www.osti.gov/biblio/1395245.

OneBuilding.org. Energyplus weather data (epw) repository, 2024. URL https://climate.onebuilding.org. Accessed: 2026-05-03.

Pieter Pauwels, Thomas Krijnen, Walter Terkaj, and Jakob Beetz. Enhancing the ifcowl ontology with an alternative representation for geometric data. *Automation in Construction*, 80:77–94, 2017. doi: 10.1016/j.autcon.2017.03.001.

Pecan Street Inc. Pecan street high-resolution residential electricity dataset, 2018. URL https://www.pecanstreet.org. 40 homes, 1-second resolution.

Fang Peng, Tao Su, Qing Zeng, and Xiaojuan Han. Climate-adaptive energy forecasting in green buildings via attention-enhanced Seq2Seq transfer learning. *Scientific Reports*, 15(1):31829, August 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-16953-y.

Ravi Peters, Balázs Dukai, Stelios Vitalis, Jordi van Liempt, and Jantien Stoter. Automated 3D Reconstruction of LoD2 and LoD1 Models for All 10 Million Buildings of the Netherlands. *Photogrammetric Engineering & Remote Sensing*, 88(3):165–170, 2022. doi: 10.14358/PERS.21-00032R2.

Robert G. Quayle and Henry F. Diaz. Heating degree day data applied to residential heating energy consumption. *Journal of Applied Meteorology and Climatology*, 19(3):241–246, 1980. doi: 10.1175/1520-0450(1980)019<0241:HDDDAT>2.0.CO;2.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Mads Holten Rasmussen, Maxime Lefrançois, Georg Ferdinand Schneider, and Pieter Pauwels. BOT: The building topology ontology of the W3C linked building data group. *Semantic Web*, 12(1):143–161, January 2021. ISSN 1570-0844. doi: 10.3233/SW-200385.

Christoph F. Reinhart and Carlos Cerezo Davila. Urban building energy modeling – A review of a nascent field. *Building and Environment*, 97:196–202, February 2016. ISSN 0360-1323. doi: 10.1016/j.buildenv.2015.12.001.

Mauro Ribeiro, Katarina Grolinger, Hany F. ElYamany, Wilson A. Higashino, and Miriam A. M. Capretz. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. *Energy and Buildings*, 165:352–363, April 2018. ISSN 0378-7788, 1872-6178. doi: 10.1016/j.enbuild.2018.01.034.

Nadia D. Roman, Facundo Bre, Víctor D. Fachinotti, and Roberto Lamberts. Application and characterization of metamodels based on artificial neural networks for building performance simulation: A systematic review. *Energy and Buildings*, 217:109972, June 2020. ISSN 0378-7788. doi: 10.1016/j.enbuild.2020.109972.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Pratheba Selvaraju, Mohamed Nabail, Marios Loizou, Maria Maslioukova, Melinos Averkiou, Andreas Andreou, Siddhartha Chaudhuri, and Evangelos Kalogerakis. BuildingNet: Learning to Label 3D Buildings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10397–10407, 2021. doi: 10.1109/ICCV48922.2021.01023.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pp. 1548–1554. International Joint Conferences on Artificial Intelligence Organization, August 2021. doi: 10.24963/ijcai.2021/214. Main Track.

Kristian Stenerud Skeie, Lillian Sve Rokseth, Carine Lausselet, and Arild Gustavsen. Characterizing and structuring open datasets for assessment of residential building energy use on the neighborhood and urban scale. *Energy and Buildings*, 341, 2025. doi: 10.1016/j.enbuild.2025.115842.

Herbert C. S. Thom. The rational relationship between heating degree days and temperature. *Monthly Weather Review*, 82(1):1–6, 1954. doi: 10.1175/1520-0493(1954)082<0001:TRRBHD>2.0.CO;2.

Casper van Engelenburg, Fatemeh Mostafavi, Emanuel Kuhn, Yuntae Jeon, Michael Franzen, Matthias Standfest, Jan van Gemert, and Seyran Khademi. MSD: A benchmark dataset for floor plan generation of building complexes. In *Computer Vision – ECCV 2024*, volume 15116 of *Lecture Notes in Computer Science*, pp. 60–75. Springer, 2025. ISBN 978-3-031-73636-0. doi: 10.1007/978-3-031-73636-0_4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 5998–6008, Long Beach, CA, USA, December 2017. Curran Associates, Inc. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Lea Von Krannichfeldt, Kristina Orehounig, and Olga Fink. Combining physics-based and data-driven modeling for building energy systems. *Applied Energy*, 391:125853, 2025. doi: 10.1016/j.apenergy.2025.125853.

Likai Wang, Francesco De Luca, Patrick Janssen, Do Phuong Tung Bui, Kian Wee Chen, and Chao Yuan. A cross-platform optimization system for comparative design exploration of competing concepts and strategies. *Journal of Building Engineering*, 115:114413, December 2025. ISSN 2352-7102. doi: 10.1016/j.jobe.2025.114413.

Ruisheng Wang, Shangfeng Huang, and Hongxin Yang. Building3D: An Urban-Scale Dataset and Benchmarks for Learning Roof Structures from Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20076–20086, 2023. doi: 10.48550/arXiv.2307.11914.

Ramon Elias Weber, Caitlin Mueller, and Christoph Reinhart. A hypergraph model shows the carbon reduction potential of effective space use in housing. *Nature Communications*, 15(1):8327, September 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-52506-z.

Stephen Wilcox and William Marion. Users manual for tmy3 data sets. Technical Report NREL/TP-581-43156, National Renewable Energy Laboratory (NREL), 2008. URL https://www.nrel.gov/docs/fy08osti/43156.pdf.

Zhaoji Wu, Mingkai Li, Wenli Liu, Jack C. P. Cheng, Zhe Wang, Helen H. L. Kwok, Cong Huang, and Fangli Hou. Developing surrogate models for the early-stage design of residential blocks using graph neural networks. *Building Simulation*, January 2025. ISSN 1996-8744. doi: 10.1007/s12273-025-1237-7.

Jun Xiao, Hao Zhou, S. Yang, D. Zhang, and Borong Lin. A cad-bem geometry transformation method for face-based primary geometric input based on closed contour recognition. *Building Simulation*, 17(2):335–354, 2024. doi: 10.1007/s12273-023-1081-6.

Jun Xiao, Qiong Wang, Yihui Li, Zhexuan Yu, Hao Zhou, and Borong Lin. A Fully Automated DM-BIM-BEM Pipeline Enabling Graph-Based Intelligence, Interoperability, and Performance-Driven Early Design, January 2026.

Zhexuan Yu, Yihui Li, Jun Xiao, Hao Zhou, and Borong Lin. Physical embedding on building surface spatial relationships shows better performance in graph-based daylight prediction. *Building and Environment*, 290:114–141, February 2026. ISSN 0360-1323. doi: 10.1016/j.buildenv.2025.114141.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 3394–3404, Long Beach, CA, USA, December 2017. Curran Associates, Inc. doi: 10.5555/3294996.3295098.

Liang Zhang, Jin Wen, Yanfei Li, Jianli Chen, Yunyang Ye, Yangyang Fu, and William Livingood. A review of machine learning in building load prediction. *Applied Energy*, 285:116452, March 2021. ISSN 0306-2619. doi: 10.1016/j.apenergy.2021.116452.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision – ECCV 2020*. Springer, 2020. doi: 10.48550/arXiv.1908.00222.

Xiao Xiang Zhu, Sining Chen, Fahong Zhang, Yilei Shi, and Yuanyuan Wang. Globalbuildingatlas: An open global and complete dataset of building polygons, heights and LoD1 3D models. *Earth System Science Data*, 17(12):6647–6668, December 2025. ISSN 1866-3591. doi: 10.5194/essd-17-6647-2025.

APPENDIX

This appendix provides additional details that support the main paper in four aspects: dataset construction and statistics, benchmark implementation and evaluation, and qualitative visualization resources. Section A expands the data pipeline, cleaning strategy, split protocol, and supplementary geometry-weather-energy analyses. Section B reports computational settings, metric definitions, model/training configurations, and extended generalization results for both benchmark tasks. Section C presents interactive and batch visualizations to facilitate qualitative inspection of aligned geometry-topology-physics samples.

A SUPPLEMENTARY DATASET DETAILS

A.1 REPRESENTATIVE DATA CASE

Figure A.1 shows a randomly selected data case.

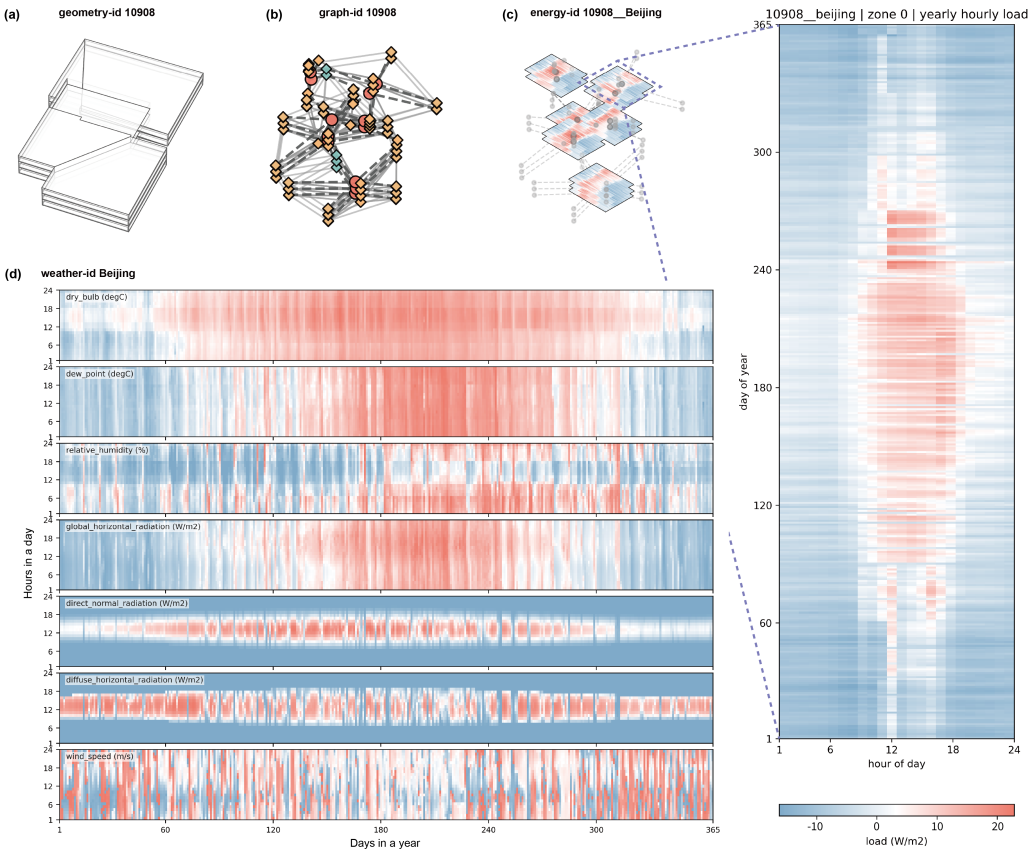


Figure A.1: Example of an ArchEGraph sample showing the aligned representation of building geometry, graph topology, zone-level energy load, and local weather time series. The right panel visualizes the annual hourly load profile of a representative zone, illustrating the spatiotemporal coupling between building structure, weather conditions, and energy demand.

A.2 DATA PREPROCESSING AND CLEANING

The pipeline consists of three sequential stages: (i) geometry normalization and convex decomposition, (ii) geometry-to-graph conversion, and (iii) physics-based simulation integration. An overview is shown in Figure A.2.

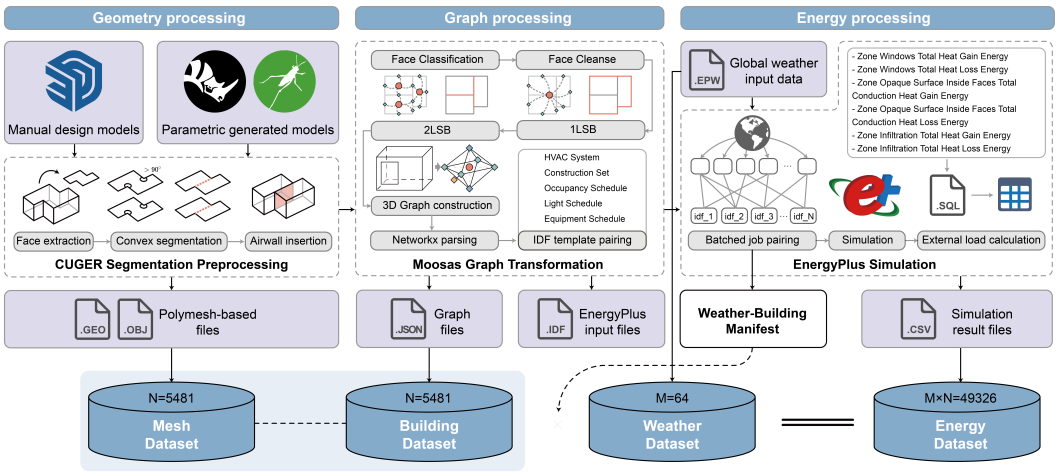


Figure A.2: Dataset construction pipeline

A.2.1 Original Data Sources

The geometric data in ArchEGraph are stored in OBJ format and exported from SketchUp and Rhinoceros 3D. All building instances correspond to office building typologies.

The dataset is constructed from two complementary sources:

(1) Manually designed models. These models are created by students in a building performance simulation and optimization course. Each design is produced under functional and spatial constraints, reflecting realistic architectural decision-making processes.

(2) Parametric generative models. These models are generated in Rhinoceros 3D using the EvoMass plugin (Wang et al., 2025), which enables controllable procedural generation of multi-story building forms. The generative process is driven by key design parameters, including floor height, column spacing, and volumetric operations (e.g., additive and subtractive transformations), allowing systematic exploration of the architectural design space.

In total, the dataset comprises 652 manually designed models and 5,052 parametric models.

Weather data. Climate inputs are obtained from the OneBuilding database (OneBuilding.org, 2024). We select 64 representative cities covering major global climate regions. For each location, the typical meteorological year (TMY) weather files are used, which are standard inputs in the building energy simulation and ensure the comparability between climates (Wilcox & Marion, 2008).

A.2.2 Geometry normalization and convex decomposition

Raw building models harvested from CAD environments (e.g., SketchUp, Rhino) typically suffer from topological defects, including duplicated vertices, inconsistent normal orientations, and non-convex boundaries. We address these through a three-stage normalization process.

(1) Topology cleaning and vertex merging: We first unify the geometric representation by merging vertices within a numerical tolerance ϵ to resolve leak boundaries. All faces are re-indexed to a global vertex list to eliminate redundancy, and inconsistent normals are unified via a breadth-first traversal of face adjacency to ensure a consistent inward/outward orientation.

(2) Minimal convex partitioning: To satisfy the requirements of both EnergyPlus simulation and graph-based spatial reasoning, all non-convex polygons (e.g., L-shaped floors or walls with openings) must be decomposed. We implement a minimal convex decomposition algorithm: complex polygons are first triangulated, followed by a local merging procedure that reconstructs the minimum number of convex sub-polygons. This step ensures that each face primitive is a simple convex polygon, facilitating stable normal calculation and centroid-based spatial indexing.

(3) Air-wall insertion: For models with split polygons, we introduce virtual air-wall connectors to bridge small gaps, ensuring that partially defined spatial volumes are correctly identified as watertight enclosures for simulation.

A.2.3 GEOMETRY-TO-GRAPH CONVERSION

Based on the foundational representation framework proposed in (Li et al., 2026), we convert the normalized geometry G into a heterogeneous graph $G = (\mathcal{V}_F, \mathcal{V}_S, \mathcal{E}_{FF}, \mathcal{E}_{FS})$. The conversion involves three critical alignment steps:

(1) Topological adjacency detection: Adjacency between convexified face nodes $\mathcal{V}_F$ is determined by shared boundary segments. A connectivity edge $e_{ij} \in \mathcal{E}_{FF}$ is defined when the intersection length between two faces exceeds a threshold τ_e, which filters out spurious connections caused by geometric noise or numerical precision errors in CAD exports.

(2) Volumetric enclosure reconstruction: Following the methodology in (Xiao et al., 2024), we reconstruct spatial nodes $\mathcal{V}_S$ (thermal zones) by identifying closed face cycles.

(3) Topological refinement: We apply a rule-based First/Second-Level Space Boundaries procedure, implemented in the open-source software MOOSAS+ (Xiao et al., 2026), to resolve ambiguities between face geometry and space semantics. The algorithm enforces a strict many-to-one mapping from faces to spatial enclosures, except when a face is explicitly classified as an interior partition shared by two spatial nodes.

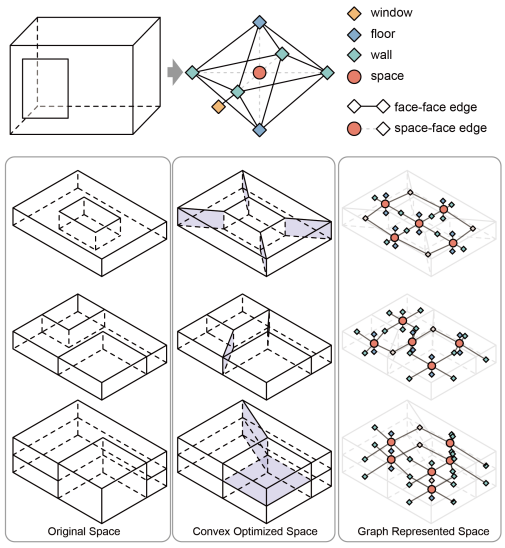


Figure A.3: **Space-to-graph conversion and convex optimization pipeline.** *(Top)* A heterogeneous graph representation: boundary surfaces are mapped to face nodes $\mathcal{V}_F$, and enclosed zones to space nodes $\mathcal{V}_S$. Edges $\mathcal{E}_{FF}$ and $\mathcal{E}_{fs}$ encode adjacency and enclosure relations. *(Bottom)* The full pipeline from raw polygonal meshes to convex-optimized geometry and then to a heterogeneous graph representation, following the framework in (Li et al., 2026).

The final graph integrates semantic attributes (e.g., wall, floor, roof, window) derived from CAD layer metadata and geometric heuristics. Spatial nodes aggregate their constituent faces to store high-level properties such as total volume and envelope exposure. The resulting structure, illustrated in Figure A.3, provides a unified multi-modal representation for learning-based building performance modeling.

A.2.4 ENERGY SIMULATION INTEGRATION

The graph G is deterministically serialized into EnergyPlus-compatible IDF files through a loss-preserving mapping from graph entities to thermal zones, boundary surfaces, and construction assemblies. Each IDF is paired with an EPW weather file selected from 64 global locations that span all ASHRAE climate regions, thereby providing broad variation in external boundary conditions.

Each building model is simulated using EnergyPlus v24.2 with a unified set of physical and operational parameters, following the DOE Medium Office reference building standard (Deru et al., 2011). The simulation covers a full calendar year (Jan. 1–Dec. 31, 2006) at a 15-minute timestep (4 steps/hour), with solar distribution set to `FullExteriorWithReflections` and shading recalculated every 30 timesteps. Internal loads, envelope constructions, and HVAC settings are held constant across all building instances; only geometry and weather inputs vary across cases. Key parameter values are summarized in Table A.1.

After simulation, we derive the external disturbance load, i.e., the thermal response driven by outdoor environmental forcing under fixed internal schedules (occupancy, lighting, and equipment). External disturbance loads represent envelope-mediated heat exchange induced by outdoor climate conditions, including solar radiation, outdoor air temperature, wind pressure, and humidity-related enthalpy

Table A.1: Summary of fixed EnergyPlus simulation parameters applied uniformly to all buildings in the dataset.

Category	Parameter	Value
Simulation	Timestep	4 steps/hour (15 min)
	Run period	Full year (Jan–Dec)
	Solar distribution	FullExteriorWithReflections
Internal loads	Occupancy density	0.0565 people/m^2
	Lighting power density	10.55 W/m^2
	Equipment power density	10.33 W/m^2
Infiltration	Method	Flow/Zone (constant-term only)
	Rate	2.266×10^{-4} m^3/s per m^2 exterior surface
Thermostat	Type	Dual setpoint (heating + cooling)
	Heating setpoint (occupied)	20–21 °C
	Cooling setpoint (occupied)	24–25 °C
	Setback (unoccupied)	15.6 °C / 26.7 °C
HVAC system	Type	Ideal Loads Air System (per zone)
	Heating supply air	$\leq$50 °C, autosized capacity
	Cooling supply air	$\geq$13 °C, autosized flow & capacity
	Economizer	Differential dry-bulb
Envelope	Exterior wall	Brick + LW concrete + 50 mm insulation + gypsum board
	Roof	Membrane + 50 mm insulation + LW concrete
	Window glazing	Simple glazing: U=1.8 W/m^2K, SHGC=0.40, τ_{vis}=0.74
Schedules	Occupancy profile	DOE Medium Office (peak 0.95 on weekdays)
	Lighting / equipment	DOE Medium Office fractional schedules

effects. In our pipeline, these effects are computed implicitly by EnergyPlus from geometry-dependent surface exposure and material properties. We denote the external disturbance load at time t as $Q_{\text{ext}}(t)$.

Internal heat gains arise from occupants, lighting, and equipment:

$$Q_{\text{int}}(t) = Q_{\text{people}}(t) + Q_{\text{lighting}}(t) + Q_{\text{equipment}}(t). \tag{A.1}$$

The total zone thermal load is then:

$$Q_{\text{total}}(t) = Q_{\text{ext}}(t) + Q_{\text{int}}(t). \tag{A.2}$$

However, building energy consumption is not identical to thermal load. HVAC systems respond to the mismatch between thermal demand and comfort targets under operational constraints. A simplified relation is:

$$E_{\text{HVAC}}(t) \propto |Q_{\text{total}}(t) - Q_{\text{set}}(t)|, \tag{A.3}$$

where $Q_{\text{set}}(t)$ denotes the thermal regulation target required to maintain the prescribed comfort range.

In this dataset, we focus on $Q_{\text{ext}}(t)$ as the climate-driven response signal. This choice is consistent with the first two stages of the pipeline, where geometry determines envelope exposure and thereby modulates external thermal forcing. By isolating $Q_{\text{ext}}(t)$, we reduce confounding effects from HVAC control policies and internal behavioral uncertainty, yielding a cleaner setting for physics-consistent geometric representation learning.

A.2.5 Data cleaning and implementation details

We apply a multi-stage pipeline to ensure geometric validity, simulation reliability, and cross-modal consistency between geometry, graph structure, and simulation outputs.

Geometric and simulation filtering. We first remove geometries with more than 10,000 faces, as these typically indicate corrupted models or failures in convex decomposition. We then discard cases that either fail EnergyPlus simulation or exceed a runtime of one hour, which usually corresponds to non-watertight geometry or invalid IDF conversion. Finally, we enforce consistency between simulation outputs and graph representations by computing a spatial matching ratio between detected thermal zones and graph nodes. Cases with a ratio below 80% are excluded. After filtering, the

dataset contains 580 manually designed models and 4,901 parametric models, yielding 5,481 valid graph-structured samples.

Manifest-driven integrity and loading consistency. All samples are indexed via a manifest containing `sample_id`, `building_id`, `weather_id`, and file paths for geometry and simulation outputs. During data loading, each entry is validated as a complete building–weather–energy triplet.

Schema normalization and temporal alignment. Weather inputs stored in `npz` format are converted into a unified tabular schema. Feature aliases (e.g., different naming conventions for dry-bulb temperature) are standardized into a consistent namespace. If temporal indexing is missing, we reconstruct an hourly index to ensure deterministic sequence alignment across all samples.

Graph-energy alignment. Zone-level energy signals are aligned with graph spatial nodes using a structured mapping procedure. When available, we match nodes using `valid_energy_spaces`. If the energy dimension exceeds graph space dimension, redundant channels are truncated; if it is smaller, zero-padding is applied. This guarantees consistent tensor shapes across all samples.

A.3 Dataset Split

Table A.2 summarizes all datasets and evaluation splits used in this work. Unless otherwise specified, all standard splits follow a 70% / 15% / 15% train/validation/test ratio, randomly sampled from the full dataset.

The dataset consists of two large-scale subsets, **ArchEGraph-P** and **ArchEGraph-M**, covering diverse building geometries and weather conditions. For reproducibility and lightweight experimentation, we further construct a reduced subset, **ArchEGraph-Demo**, which contains 75 buildings sampled from both P and M, each paired with 4 representative weather conditions across different climate regions.

In addition to the full datasets, we provide mesh-level splits for the M2G task (**ArchEGraph-P/M/Demo-Mesh**), which focus exclusively on geometric features without climatic variations, ensuring that specific geometries are strictly separated between sets to prevent data leakage.

To evaluate generalization under controlled conditions, we further design two bias-oriented splits:

Building-Bias split isolates geometric variability by fixing a single weather condition (Chicago climate). This results in 3,000 samples (one per building), partitioned into 2,100 / 450 / 450 for training, validation, and testing. Data leakage is prevented by separating building instances across splits.

Weather-Bias split isolates climatic variability by fixing a limited set of buildings while varying weather conditions. We select 150 buildings (75 from ArchEGraph-P and 75 from ArchEGraph-M), each paired with 20 weather scenarios, resulting in 3,000 samples. The dataset is split into 2,100 / 450 / 450, with consistent evaluation scale across both bias settings.

Table A.2: Statistical summary of datasets, mesh-specific splits, and bias-oriented splits.

Dataset / Split	#Cases	#Bldg	#Wea	Train (#Bldg/#Wea)	Val (#Bldg/#Wea)	Test (#Bldg/#Wea)
ArchEGraph-P	30,658	4,901	64	21,460 (4643 / 64)	4,599 (2050 / 64)	4,599 (1979 / 64)
ArchEGraph-P-Mesh	4,901	4,901	–	3,431	735	735
ArchEGraph-M	18,668	580	64	13,067 (580 / 64)	2,800 (572 / 54)	2,801 (573 / 57)
ArchEGraph-M-Mesh	580	580	–	406	87	87
ArchEGraph-Demo	300	75	48	210 (74 / 44)	45 (35 / 26)	45 (35 / 27)
ArchEGraph-Demo-Mesh	300	300	–	210	45	45
Building-Bias	3,000	3,000	1	2,100 (2100 / 1)	450 (450 / 1)	450 (450 / 1)
Weather-Bias	3,000	150	20	2,100 (150 / 14)	450 (150 / 3)	450 (150 / 3)

A.4 Additional dataset statistics

A.4.1 Geometry statistics

Table A.3 summarizes the geometric and topological diversity of ArchEGraph. The mean cyclomatic complexity is 821.05 (std. 520.00), ranging from 22 to 21,121, indicating substantial variation from tree-like layouts to highly loopy configurations. Edge density remains stable (mean 4.19), while the

Table A.3: Cross-split comparison of geometric and topological statistics in ArchEGraph.

Metric	ArchEGraph P (4,901)			ArchEGraph M (580)			ArchEGraph P+M (5,481)		
	mean	std	sum	mean	std	sum	mean	std	sum
Graph Structure									
Space nodes $\lvert\mathcal{V}_S\rvert$	24.33	11.45	119,265	24.50	26.47	14,209	24.35	13.80	133,474
Face nodes $\lvert\mathcal{V}_F\rvert$	253.48	101.80	1,242,324	351.47	325.41	203,853	263.86	146.20	1,446,177
Floor	(46.93%)120.54	52.31	590,786	(27.87%)97.32	204.06	56,448	(44.75%)117.96	77.10	647,234
Wall	(22.81%)55.59	19.43	272,435	(30.92%)101.12	117.46	58,651	(22.89%)60.37	38.80	331,086
Window	(21.75%)53.08	18.77	260,162	(39.36%)144.14	151.43	79,708	(23.52%)62.01	47.60	339,870
Airwall	(8.77%)25.00	18.13	118,868	(5.91%)24.95	77.31	9,032	(8.84%)24.99	30.50	127,900
Edges $\lvert\mathcal{E}_{FF}\rvert$	784.85	381.51	3,846,552	953.86	1,132.97	553,239	802.74	518.35	4,399,791
Edges $\lvert\mathcal{E}_{FS}\rvert$	290.31	138.05	1,422,826	243.01	378.03	140,943	285.31	179.93	1,563,769
Topology									
Density $\lvert\mathcal{E}\rvert/\lvert\mathcal{V}_F\rvert$	4.26	0.87	–	3.62	1.31	–	4.19	0.95	–
Face-to-space ratio $\lvert\mathcal{V}_F\rvert/\lvert\mathcal{V}_S\rvert$	10.89	1.53	–	19.19	25.57	–	11.87	4.20	–
Complexity $\lvert\mathcal{E}\rvert-\lvert\mathcal{V}_F\rvert-\lvert\mathcal{V}_S\rvert$	815.24	403.54	–	869.96	1,229.64	–	821.05	520.00	–

face-to-space ratio averages 11.87 (std. 4.20), capturing the granularity of geometric partitioning relative to spatial zoning.

Element composition is dominated by floor (44.75%), wall (22.89%), and window (23.52%) faces, and airwalls for 8.84%. The two splits exhibit complementary characteristics: Split P (4,901 samples) is statistically compact (e.g., face-to-space ratio 10.89 ± 1.53), while Split M (580 samples) exhibits heavy-tailed distributions with higher structural complexity (e.g., ratio 19.19 ± 25.57) and a markedly larger window proportion (39.36% vs. 21.75%). This combination provides both stable coverage of typical layouts and exposure to rare, high-complexity structures, supporting robust generalization.

A.4.2 Climate and Energy Statistics

Weather coverage. The 64 cities span all inhabited continents and ASHRAE climate regions. Each hourly weather sequence contains dry-bulb temperature, dew point, relative humidity, global horizontal irradiance, direct normal irradiance, diffuse horizontal irradiance, and wind speed, which capture the primary climatic drivers of building thermal response (Bhandari et al., 2012). The profiles in Figure 3 distinguish humid, arid, temperate, and high-latitude regimes through complementary temperature, moisture, and solar patterns, providing diverse forcing conditions for cross-climate evaluation.

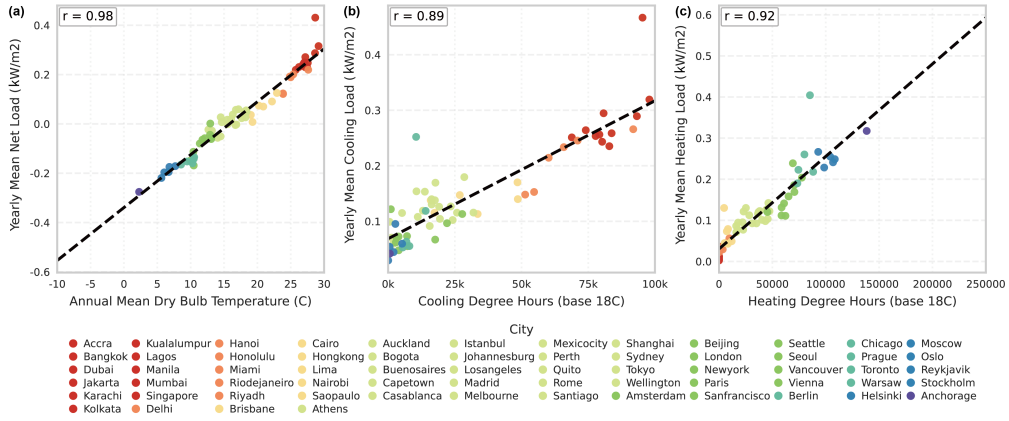


Figure A.4: Climate-load regression across representative cities. From left to right: annual mean dry-bulb temperature versus annual net load, Cooling Degree Hours (base 18 °C) versus annual cooling load, and Heating Degree Hours (base 18 °C) versus annual heating load. Dashed lines indicate linear fits, and color denotes city identity.

To quantify the relationship between climate drivers and energy response, Figure A.4 shows regression results between annual building loads and key meteorological indicators, including outdoor dry-bulb temperature, Cooling Degree Days (CDD18), and Heating Degree Days (HDD18). The degree-day concept was originally introduced by Vernon L. Thom in the mid-20th century as a simple climatic index for estimating heating and cooling demand from temperature data (Thom, 1954). By accumulating deviations of daily mean temperature from a base temperature (typically 18°C), HDD and CDD provide physically interpretable measures of heating and cooling demand over time and are widely used in building energy analysis and climate zoning (e.g., ASHRAE standards (ASHRAE, 1993)).

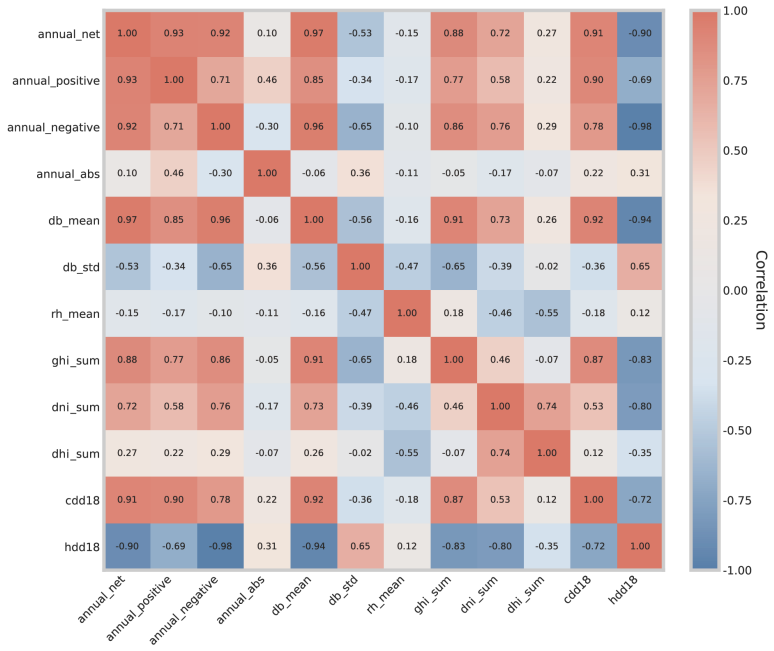


Figure A.5: Correlation matrix of annual energy and weather variables.

The results show strong linear relationships, with correlation coefficients of 0.98 for temperature vs. net load, 0.89 for CDD18 vs. cooling load, and 0.92 for HDD18 vs. heating load. Overall, energy demand is strongly governed by large-scale climatic conditions. Dry-bulb temperature captures overall load trends, while degree-day metrics better isolate heating and cooling effects by aggregating temperature deviations, reducing temporal noise and improving comparability across climates.

Building on these cross-city relationships, the internal dependency structure among climate and energy variables further reveals how different factors jointly shape load characteristics. As summarized in Figure A.5, annual net load is strongly consistent with its decomposed components, indicating stable load aggregation behavior. Temperature-related variables form a tightly coupled group, where mean dry-bulb temperature (db_mean) is strongly correlated with CDD18 (0.92) and anti-correlated with HDD18 (-0.94), reflecting the separation of cooling- and heating-dominated regimes (Quayle & Diaz, 1980). Solar radiation variables (e.g., ghi and dni) also show moderate to strong correlations with temperature and cooling-related metrics, suggesting their contribution to cooling demand. Overall, temperature, radiation, and degree-day metrics dominate the variation in building energy demand, while other climatic factors play a secondary role.

B Supplementary Evaluation Details

B.1 Computational Resources

We conduct all experiments under a unified computational environment to ensure fair comparison across tasks. The overall evaluation protocol consists of four experimental settings: (i) M2G baseline, (ii) G2E baseline, (iii) cross-building and cross-weather generalization, and (iv) benchmark evaluation on the ArchEGraph-Demo dataset.

For each experiment, all models are trained and evaluated on the same hardware configuration. We report the average training time cost, and peak GPU memory use for each experiment. Training time is measured until convergence under early stopping, while inference time is averaged over the full test set. GPU memory usage is recorded as peak allocation during training.

The M2G and G2E experiments constitute the main computational bottleneck due to full-scale graph construction and EnergyPlus-based supervision signals, whereas the generalization and benchmark experiments are comparatively lightweight, involving fixed trained models evaluated under distribution shifts or small-scale datasets.

Overall, this setup allows us to systematically analyze computational efficiency across different stages of the proposed pipeline, ranging from geometry processing to physics-informed prediction tasks.

Table B.4: Computational resource usage across different experimental settings.

Experiment	GPU Type	Case	Batch size	Training Time/h	Peak GPU Memory/GB
M2G (Geometry → Graph)	RTX5080(16GB)	5k	256	0.57-0.78	11.6-15.0
G2E (Graph → Energy)	A100(40GB)	30k	8	5.56-20.14	36.8-39.0
Cross-building Generalization	RTX5080(16GB)	3k	2	1.98-4.21	14.6-15.8
Cross-weather Generalization	RTX5080(16GB)	3k	2	2.24-4.56	14.7-16.0
ArchEGraph-Demo M2G	RTX5080(16GB)	300	8	0.02-0.03	1.2-1.4
ArchEGraph-Demo G2E	RTX5080(16GB)	300	8	0.18-0.22	12.3-15.5

B.2 Model details and metric definitions

B.2.1 M2G details

Model architecture We design TopoTransformer to model the topological relationships between building faces and spaces. Given face features $\mathbf{x}_f \in \mathbb{R}^{N_F \times d_f}$, the model predicts three structured outputs: face-face adjacency $\mathcal{E}_{FF}$, space existence $\mathcal{V}_S$, and space-face incidence $\mathcal{E}_{FS}$.

The model first encodes face tokens using a Transformer encoder (Vaswani et al., 2017):

$$\mathbf{H}_F = \mathrm{Enc}_{\mathrm{Tr}}(\mathbf{x}_f), \quad \mathbf{H}_F \in \mathbb{R}^{N_F \times d_h}. \tag{B.4}$$

Three prediction heads are then applied:

Face-face adjacency. We predict pairwise adjacency using an MLP over concatenated embeddings:

$$\hat{z}_{ij}^{FF} = \phi_{FF}([\mathbf{h}_i \,\|\, \mathbf{h}_j]). \tag{B.5}$$

Space existence. We introduce $S_{\max}$ learnable space queries (default 64). Each query is fused with a global pooled face representation:

$$\mathbf{S} = \mathrm{LN}(\mathbf{Q} + \mathbf{W}_g, \mathrm{mean}(\mathbf{H}_F)), \quad \hat{z}_s^S = \phi_S(\mathbf{s}_s). \tag{B.6}$$

Space-face incidence. We compute incidence via scaled bilinear matching:

$$\hat{z}_{s,i}^{FS} = \frac{(\mathbf{W}_s \mathbf{s}_s)^\top (\mathbf{W}_f \mathbf{h}_i)}{\sqrt{d_h}} + b. \tag{B.7}$$

We compare TopoTransformer against DeepSets (Zaheer et al., 2017), SetTransformer (Lee et al., 2019), Perceiver (Jaegle et al., 2021), and MLP under the same M2G formulation.

Training objective and optimization The M2G training loss is:

$$\mathcal{L}_{\mathrm{M2G}} = \lambda_{FF}\mathcal{L}_{FF} + \lambda_S \mathcal{L}_S + \lambda_{FS}\mathcal{L}_{FS}, \tag{B.8}$$

with weights $(\lambda_{FF}, \lambda_S, \lambda_{FS}) = (1, 1, 2)$.

For $\mathcal{L}_{FF}$ and $\mathcal{L}_{FS}$, we use BCE on sampled positive/negative pairs (negative ratios 1:1 for FF and 2:1 for FS). For $\mathcal{L}_S$, we use a masked BCE with adaptive positive weighting to mitigate class imbalance in space-slot activation. We optimize with AdamW, gradient clipping ($\|g\|_2 \leq 1.0$), early stopping on validation FS-F1, and threshold 0.5 for binary metrics.

Default training setup All M2G baselines are trained under a unified setting for fair comparison. The default setup in the released code uses: Transformer hidden dimension 256, batch size 8, 30 epochs, learning rate 3×10^{-4}, weight decay 10^{-4}, deterministic training, and explicit split CSV files for train/validation/test partitioning.

Metric definitions For each topology component $c \in \{\mathcal{E}_{FF}, \mathcal{V}_S, \mathcal{E}_{FS}\}$, we binarize predicted probabilities by threshold τ:

$$\hat{y}_i^{(c)} = \mathbb{I}\Big(\hat{p}_i^{(c)} \geq \tau\Big). \tag{B.9}$$

Let $\mathrm{TP}_c, \mathrm{FP}_c, \mathrm{TN}_c, \mathrm{FN}_c$ denote true positives, false positives, true negatives, and false negatives, respectively. We report Accuracy and F1 *separately* for each component:

$$\mathrm{Acc}_c = \frac{\mathrm{TP}_c + \mathrm{TN}_c}{\mathrm{TP}_c + \mathrm{TN}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{B.10}$$

$$\mathrm{F1}_c = \frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{B.11}$$

B.2.2 G2E DETAILS

Model architecture The default G2E model (F2S-Attr) is a heterogeneous graph-to-sequence framework that maps building topology and weather conditions to hourly energy loads through spatial message passing and temporal decoding. The architecture operates on a face–space graph augmented with external environmental signals, and consists of three main components: (i) masked heterogeneous message passing, (ii) spatial encoding with attribute-aware edge modeling, and (iii) temporal decoding with a GRU-based sequence model.

Face and space attributes are first embedded via MLP encoders to obtain initial node representations $\mathbf{h}_f$ and $\mathbf{h}_s$. Weather variables are encoded with cyclical temporal features and contextual augmentation (see below) to produce $\mathbf{w}_t$. In F2S-Attr, weather information is injected only into exterior face nodes via a structural mask. Formally, let $\mathcal{V}_F^{\text{out}} \subset \mathcal{V}_F$ denote exterior faces; then the masked weather embedding is defined as

$$\mathbf{h}_f^{(0)} = \begin{cases} \phi_f(\mathbf{x}_f) + \phi_w(\mathbf{w}_t), & f \in \mathcal{V}_F^{\text{out}}, \\ \phi_f(\mathbf{x}_f), & f \notin \mathcal{V}_F^{\text{out}}, \end{cases} \tag{B.12}$$

where $\phi_f(\cdot)$ and $\phi_w(\cdot)$ are MLP encoders for geometric attributes and weather signals, respectively. This design ensures that weather conditions are only injected at exterior boundaries, while interior faces receive information purely through geometric propagation.

Message passing is performed on face-to-space edges with attribute-aware and physics-inspired separation between opaque and transparent interfaces. For each edge e_{fs}, the message is computed as

$$\mathbf{m}_{f\to s} = \psi\big([\mathbf{h}_f \,\|\, \mathbf{e}_{fs} \,\|\, \mathbf{h}_s]\big), \tag{B.13}$$

where $\mathbf{e}_{fs}$ encodes geometric adjacency and material properties. To explicitly model heat-transfer heterogeneity, edges are categorized into transparent and opaque types via a binary indicator $\tau_{fs} \in \{0, 1\}$. We apply a gated modulation:

$$\tilde{\mathbf{m}}_{f\to s} = \alpha_{fs} \cdot \mathbf{m}_{f\to s}, \quad \alpha_{fs} = \sigma\big(\mathbf{a}^\top \mathbf{e}_{fs} + b\big), \tag{B.14}$$

where the attention gate α_{fs} is further conditioned on τ_{fs} to distinguish radiative versus conductive transfer behavior. Aggregated messages update space embeddings via residual stacking:

$$\mathbf{h}_s^{(l+1)} = \mathbf{h}_s^{(l)} + \sum_{f \in \mathcal{N}(s)} \tilde{\mathbf{m}}_{f\to s}. \tag{B.15}$$

After spatial encoding, each space representation is concatenated with time-dependent weather embeddings and passed into a GRU decoder for sequential prediction:

$$\hat{\mathbf{y}}_t = \mathrm{GRU}\big(\mathbf{h}_s \,\|\, \mathbf{w}_t^{\text{enc}}\big). \tag{B.16}$$

To improve temporal expressivity, the hourly index $t \in \{1, \ldots, 8760\}$ is embedded using cyclical positional encoding:

$$\mathbf{x}_t^{\text{time}} = \left[\sin\left(\tfrac{2\pi(t-1)}{24}\right), \quad \cos\left(\tfrac{2\pi(t-1)}{24}\right), \quad \sin\left(\tfrac{2\pi(t-1)}{168}\right), \quad \cos\left(\tfrac{2\pi(t-1)}{168}\right)\right]^{\mathrm{T}}. \tag{B.17}$$

Weather inputs are further augmented with short-term and seasonal context:

$$\mathbf{w}'_t = \left[w_t, \{w_{t\pm h}\}_{h=1}^{3}, \{w_{t\pm 24d}\}_{d=1}^{3}\right], \tag{B.18}$$

capturing both local fluctuations and periodic climate trends.

Finally, to stabilize training, targets are transformed into a signed-log space and normalized:

$$u = \mathrm{sign}(y) \cdot \ln(1 + \alpha|y|), \quad \tilde{u} = \frac{u - \mu_u}{\sigma_u + \epsilon}. \tag{B.19}$$

To evaluate the robustness of our architecture, we benchmark several operator variants by replacing the spatial message-passing module with GATv2 (Brody et al., 2022), TransConv (Shi et al., 2021), and GraphGPS (Rampášek et al., 2022), while keeping all masking, encoding, and decoding components intact. Furthermore, we conduct an ablation study using F2S, which omits the masked encoding mechanism, and a baseline WeatherMLP model that predicts loads using only meteorological conditions.

Training objective and optimization G2E is trained with MSE loss in normalized target space, AdamW optimization, gradient clipping ($\|g\|_2 \leq 1.0$), and early stopping on validation RMSE.

Default training setup All G2E baselines are trained under a unified setting for fair comparison. The default setup uses batch size 8, 100 epochs, learning rate 10^{-4}, and deterministic seeds; evaluation reports both normalized-scale metrics and inverse-transformed original-scale metrics.

Metric definitions For G2E, let $y_{s,t}$ and $\hat{y}_{s,t}$ denote the ground-truth and prediction for space node $v_S \in \mathcal{V}_S$ at time step $t = 1, \ldots, T$. We report node-level MAE, RMSE, and MSE:

$$\mathrm{MAE} = \frac{1}{|\mathcal{V}_S|} \sum_{v_S \in \mathcal{V}_S} \frac{1}{T} \sum_{t=1}^{T} |y_{s,t} - \hat{y}_{s,t}|, \tag{B.20}$$

$$\mathrm{MSE} = \frac{1}{|\mathcal{V}_S|} \sum_{v_S \in \mathcal{V}_S} \frac{1}{T} \sum_{t=1}^{T} (y_{s,t} - \hat{y}_{s,t})^2. \tag{B.21}$$

$$\mathrm{RMSE} = \frac{1}{|\mathcal{V}_S|} \sum_{v_S \in \mathcal{V}_S} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_{s,t} - \hat{y}_{s,t})^2}, \tag{B.22}$$

B.3 Additional generalization results

We report additional generalization results to further support the robustness analysis presented in the main paper. These experiments extend the primary evaluation by testing model behavior under more challenging distribution shifts, including cross-building transfer from parametric to manually designed geometries (Table B.5) and cross-climate transfer across ASHRAE regions (Table B.6). While not included in the main results, these settings provide a more comprehensive view of model stability under changes in spatial structure and environmental conditions.

Table B.5: Cross-building generalization from parametric (P) to manually designed (M) buildings.

	Validation (P)			Test (M)		
Method	MAE	MSE	RMSE	MAE	MSE	RMSE
WeatherMLP	0.3684 ±0.0029	0.4170 ±0.0013	0.6457 ±0.0010	0.4811 ±0.0023	0.6113 ±0.0024	0.7818 ±0.0015
F2S	0.2684 ±0.0070	0.4292 ±0.0091	0.6551 ±0.0070	**0.3964** ±0.0258	**0.4577** ±0.0448	**0.6760** ±0.0328
F2S-Attr	**0.2618** ±0.0040	**0.4145** ±0.0094	**0.6438** ±0.0073	0.4055 ±0.0448	0.5361 ±0.1109	0.7296 ±0.0751
F2S-GATv2	0.2877 ±0.0031	0.4369 ±0.0055	0.6610 ±0.0042	0.4356 ±0.0613	0.5392 ±0.1231	0.7313 ±0.0819
F2S-TransConv	0.2861 ±0.0034	0.4400 ±0.0119	0.6633 ±0.0090	0.4176 ±0.0122	0.5301 ±0.0624	0.7273 ±0.0422
F2S-GPS	0.2892 ±0.0012	0.4450 ±0.0067	0.6671 ±0.0051	0.4539 ±0.0100	0.5553 ±0.0208	0.7451 ±0.0139

B.4 Demo dataset benchmark on two tasks

Table B.6: Cross-climate generalization across ASHRAE climate regions.

	Validation			Test		
Method	MAE	MSE	RMSE	MAE	MSE	RMSE
WeatherMLP	0.3899 ±0.0033	0.6515 ±0.0050	0.8071 ±0.0031	0.4484 ±0.0043	0.8754 ±0.0017	0.9356 ±0.0009
F2S	0.2474 ±0.0033	0.3195 ±0.0065	0.5652 ±0.0058	0.2927 ±0.0063	0.4625 ±0.0124	0.6800 ±0.0091
F2S-Attr	0.1817 ±0.0056	0.1511 ±0.0064	0.3886 ±0.0082	0.2045 ±0.0049	0.2027 ±0.0097	0.4502 ±0.0107
F2S-GATv2	0.2190 ±0.0079	0.2011 ±0.0208	0.4481 ±0.0232	0.2453 ±0.0163	0.2698 ±0.0411	0.5184 ±0.0400
F2S-TransConv	0.1803 ±0.0050	0.1412 ±0.0058	0.3758 ±0.0077	0.1995 ±0.0078	0.1906 ±0.0116	0.4364 ±0.0132
F2S-GPS	**0.1691** ±0.0027	**0.1347** ±0.0035	**0.3670** ±0.0047	**0.1888** ±0.0035	**0.1784** ±0.0032	**0.4223** ±0.0038

To provide a compact benchmark on demo subset, we evaluate representative models on both M2G and G2E tasks in Table B.7. For M2G, results are reported on face–face adjacency ($\mathcal{E}_{FF}$), space-node classification ($\mathcal{V}_S$), and face–space relations ($\mathcal{E}_{FS}$), measured by F1-score and accuracy. For G2E, we compare neural and graph-based models under both normalized and original scales, enabling a consistent assessment across different magnitudes and learning settings. Overall, graph/transformer-based models consistently outperform MLP and set-based baselines across all subtasks, with stronger gains observed on more structurally complex relations such as face–space interactions.

Table B.7: Benchmark results on the demo subset for M2G/G2E tasks.

M2G Task						
	$\mathcal{E}_{FF}$		$\mathcal{V}_S$		$\mathcal{E}_{FS}$	
Method	F1	Acc	F1	Acc	F1	Acc
MLP	0.9165	0.9121	0.8605	0.9160	0.2493	0.8577
DeepSets	0.9546	0.9528	0.8680	0.9167	0.2947	0.8202
Perceiver	0.9244	0.9189	0.8575	0.9094	0.2413	**0.8801**
SetTransformer	**0.9570**	**0.9553**	0.8685	0.9160	**0.3212**	0.8460
TopoTransformer	0.9541	0.9522	**0.8686**	**0.9194**	0.3144	0.8538

G2E Task						
	Normalized			Original scale		
Model	MAE	MSE	RMSE	MAE	MSE	RMSE
WeatherMLP	0.5108	0.5713	0.7558	0.0067	0.000289	0.0170
F2S	0.4169	0.4195	0.6477	0.0056	0.000242	0.0155
F2S-Attr	0.3823	0.3657	0.6047	0.0052	0.000215	0.0147
F2S-GATv2	0.4217	0.4182	0.6467	0.0056	0.000233	0.0153
F2S-TransConv	0.3663	0.3261	0.5710	0.0050	0.000186	0.0136
F2S-GPS	**0.3458**	**0.3023**	**0.5498**	**0.0046**	**0.000151**	**0.0123**

C Additional visualizations

C.1 Online visualization platform for demo dataset

We provide an interactive online visualization platform for the demo dataset via Hugging Face Spaces: ArchEGraph Space, as shown in Figure C.6.

The platform enables browser-based exploration of the **ArchEGraph-demo** dataset without local setup. It supports visualization of (i) building geometry and spatial graph structures, (ii) zone-level energy time series, and (iii) coupled weather–energy relationships. Users can navigate across different samples to inspect variations in building–climate configurations. This interface serves as a lightweight tool for qualitative analysis, helping users understand spatial–temporal patterns and verify model behavior on the dataset.

C.2 Batch visualization of dataset cases

The diversity and consistency of ArchEGraph samples are illustrated through a batch-level visualization (shown in Figure C.7). By aligning multiple cases under a unified representation pipeline, this figure provides an intuitive comparison across geometric configurations, topological structures, and their corresponding physical responses. It highlights how variations in architectural form and spatial organization are reflected in both graph structure and simulated load patterns, offering a qualitative overview of the dataset's richness and cross-case variability.

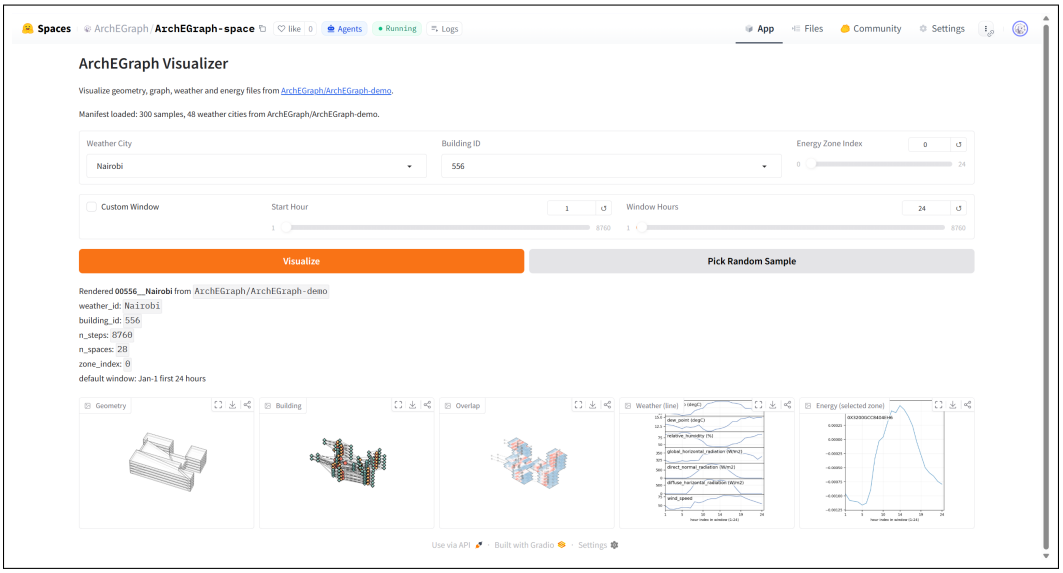


Figure C.6: Screenshot of the visualization platform.

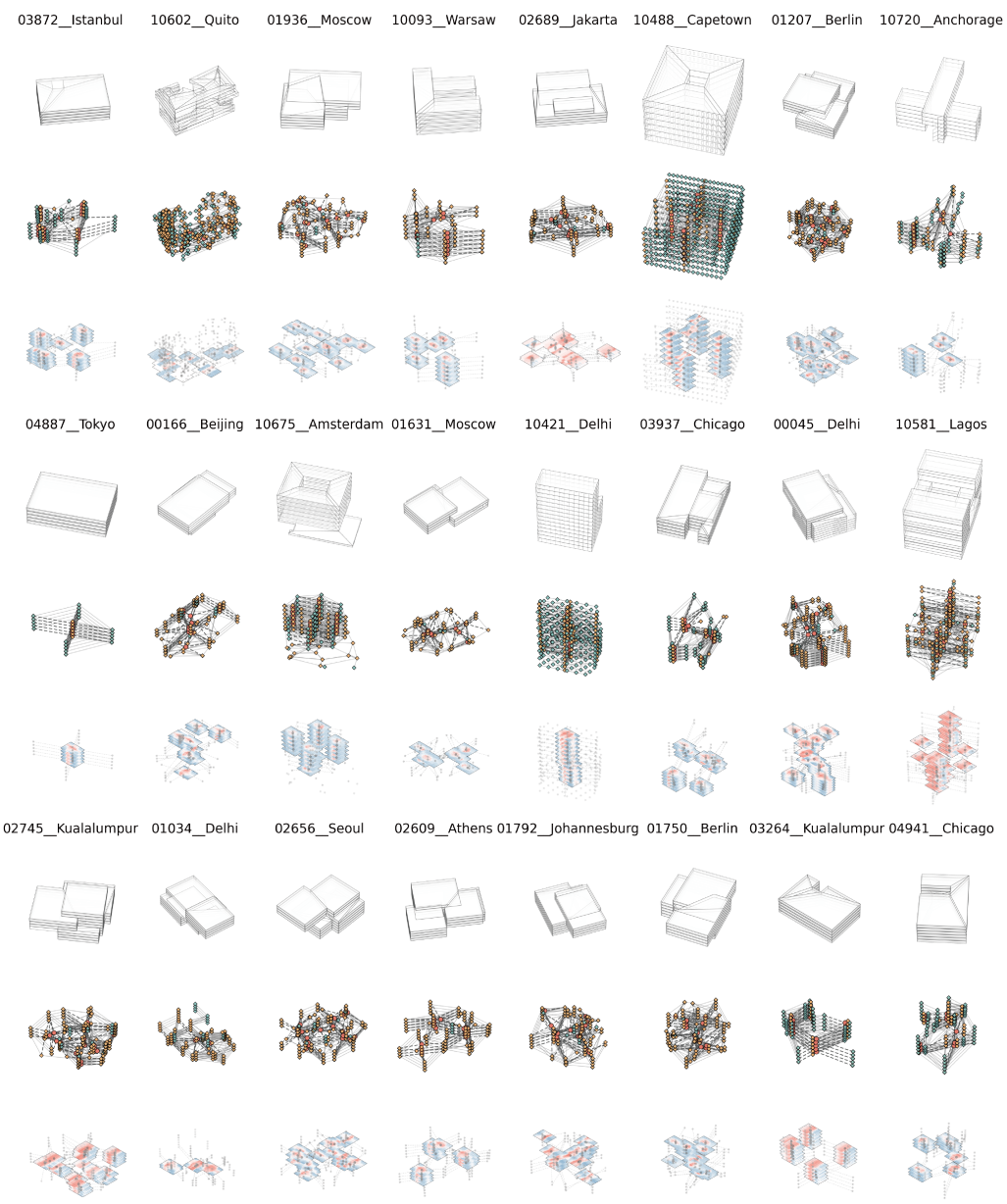


Figure C.7: Batch visualization of aligned samples from ArchEGraph across multiple cities. For each case, the triplet shows geometry mesh (top), corresponding heterogeneous graph representation (middle), and zone-level load-map strip (bottom), highlighting cross-case diversity in form, topology, and physical response patterns.