buildings

IMPACT FACTOR 3.1

CITESCORE 3.4

Article

Leveraging Machine Learning to Forecast Neighborhood Energy Use in Early Design Stages: A Preliminary Application

Andrea Giuseppe di Stefano , Matteo Ruta, Gabriele Masera and Simi Hoque

Special Issue

Data Analysis and Energy Modeling in Smart and Zero-Energy Buildings and Communities

Edited by

Dr. Marilena De Simone

MDPI

https://doi.org/10.3390/buildings14123866

buildings

MDPI

Article

Leveraging Machine Learning to Forecast Neighborhood Energy Use in Early Design Stages: A Preliminary Application

Andrea Giuseppe di Stefano [1,*], Matteo Ruta [1], Gabriele Masera [1] and Simi Hoque [2]

[1] Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano, Via Ponzio 31, 20133 Milan, Italy; matteo.ruta@polimi.it (M.R.); gabriele.masera@polimi.it (G.M.)
[2] Department of Civil, Architectural and Environmental Engineering, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA; sth55@drexel.edu
* Correspondence: andreagiuseppe.distefano@polimi.it

Abstract: The need for energy efficiency in neighborhood-scale architectural design is driven by environmental imperatives and escalating energy costs. This study identifies three key phases in a design process framework where machine learning can be applied to optimize energy consumption in early design stages. The overall framework integrates machine learning tools into the design workflow, enhancing design exploration from concept level and enabling targeted energy assessments. This paper focuses on the first phase (Phase 1) of the framework, which employs machine learning for building energy forecasting using only the few inputs available in a business-as-usual early-stage design workflow. The CatBoost model was selected for its high accuracy in predicting energy consumption using minimal input data. A preliminary application to a case study in New York City showed high predictive accuracy while reducing the input needed, with R^2 scores of 0.88 for both cross-validation and test datasets. Shapely additive explanation analysis validated the selection of key influencing parameters such as building area, principal building activity, and climate zones. The test demonstrated discrepancies between the test data-driven model and a physics-based energy model values ranging from −8.69% to 11.04%, which can be considered an acceptable result in early-stage design. The remaining two phases, though outside the scope of this study, are introduced at a conceptual level to provide an overview of the full framework. Phase 2 will analyze building shape and elevation, assessing the total energy use intensity, while Phase 3 will apply district-level energy optimization across interconnected buildings. The findings from Phase 1 underscore the potential of machine learning to integrate energy efficiency considerations into neighborhood-scale design from the earliest stages, providing reliable predictions that can inform sustainable design.

Keywords: predictive analysis; energy efficiency strategies; data-driven neighborhood design; design process framework; urban building energy modeling

Citation: di Stefano, A.G.; Ruta, M.; Masera, G.; Hoque, S. Leveraging Machine Learning to Forecast Neighborhood Energy Use in Early Design Stages: A Preliminary Application. *Buildings* **2024**, *14*, 3866. https://doi.org/10.3390/buildings14123866

Academic Editor: Marilena De Simone

Received: 22 October 2024
Revised: 19 November 2024
Accepted: 26 November 2024
Published: 30 November 2024

1. Introduction

The transformation of urban landscapes since the industrial revolution has positioned cities as major hubs of energy consumption and carbon emissions [1]. Today, urban areas are home to over half of the global population and are responsible for approximately 75% of global carbon emissions [2]. With urban populations expected to double by 2050, this proportion is likely to rise further [3]. Within these urban settings, buildings play a critical role, accounting for around 40% of total energy consumption and 38% of CO_2 emissions across the European Union [4]. These figures highlight the critical role buildings play in the global carbon footprint. However, the path to significantly reduce carbon emissions in urban buildings is complex and fraught with challenges. It requires a detailed understanding of urban energy consumption patterns and the development of strategies to reduce this consumption. Collaboration among building designers, policymakers, administrators, and tenants is essential to create and implement effective, cost-efficient measures [1].

Implementing comprehensive strategies for the heating, cooling, and electricity requirements of communities at a district and city level can lead to substantial reductions in energy consumption, lower emissions, and enhanced energy reliability. Although there is a wealth of information on energy-efficient design for neighborhoods and districts [5–11], current resources primarily focus on building-scale and technology-driven strategies, such as renewables integration, passive building strategies and dynamic systems integration, neglecting the specific challenges—and opportunities—of neighborhood-scale sustainable design. Furthermore, the vast array of tools and instruments available to design practitioners is often limiting, due to the lack of a cohesive, streamlined framework to effectively implement energy reduction strategies at the masterplan and district level. Post-carbon master planning is a process designed to identify the best combination of energy sources to fulfill the energy needs of a community over its lifetime. It begins with the setting of objectives in the initial strategic planning stages and extends into the operational phase, which includes measurement and verification tasks. However, the holistic approach to post-carbon design for neighborhoods and districts is complex due to its requirement to balance both quantitative factors, such as economic and technical considerations, and qualitative factors, like environmental and social impacts.

Over the past few decades, building performance simulations (BPSs) have become essential tools for designers, helping them explore a wide range of design options. These simulations involve many variable parameters, creating a complex, multi-dimensional design space. Navigating this space requires substantial modeling and computational efforts, which can increase costs and introduce uncertainties [12]. As a result, the practicality of large-scale applications—such as design space exploration, uncertainty analysis, sensitivity analysis, and optimization—remains limited. Nonetheless, rapid technological advancements, particularly in artificial intelligence (AI) and machine learning (ML)-driven BPS, offer significant new opportunities [13].

AI systems are emerging as key tools for addressing the diverse challenges associated with urbanization. The European Green Deal [14] highlights the urgent need for comprehensive policy initiatives to address the climate crisis, emphasizing improvements in health, quality of life, resilience, and economic competitiveness. It identifies digital innovations, including AI, as pivotal tools for achieving sustainability targets across various sectors [14]. In parallel, the European sustainable investment framework [15] and the Energy Performance Building Directive recast [16] prioritize energy efficiency within the built environment. These strategic initiatives aim to lower greenhouse gas emissions and energy consumption in buildings by 2030, ultimately striving for climate neutrality across Europe by 2050.

In this scenario, ML techniques have been increasingly applied to enhance energy forecasting at the building scale, enabling faster and efficient predictions. Recent studies have demonstrated the potential of ML models in optimizing energy systems [17,18], addressing challenges such as the predictive maintenance of heating, ventilation and air conditioning (HVAC) systems [19] and improving the integration of renewable energy sources [20].

Despite these advancements, there remains a need for integrating ML tools into existing early-stage design workflows effectively, especially at the neighborhood scale. While several studies have explored ML models for energy forecasting [21–26], they often do not address how these tools can be incorporated into the business-as-usual (BAU) workflows used by practitioners. This highlights the necessity for frameworks that bridge advanced computational methods and design practices.

To address this gap a novel framework that integrates ML tools into BAU workflows for neighborhood-scale energy forecasting during early design stages is introduced. The main objective of this paper is to test the first phase of this framework, demonstrating its effectiveness through a case study application, and illustrate how it can be integrated into existing design workflows to enhance energy efficiency considerations from the outset of neighborhood-scale projects.

The novelty of this work lies in three main contributions: first, the integration of machine learning models into early-stage design workflows to facilitate rapid and accurate energy consumption predictions with minimal input data; second, an innovative use of CatBoost for early-stage energy use intensity (EUI) predictions; third, a data-efficient approach that facilitates neighborhood-level energy assessments. By focusing on minimal input requirements, this approach enables practitioners to make informed decisions throughout the project's earliest stages, aligning with carbon reduction pathways and sustainability goals. This framework's uniqueness lies in (1) its seamless integration into BAU workflows, ensuring that it adapts to typical input data and operational structures, and (2) its potential for application across varied design stages.

The remainder of this paper is organized as follows: Section 2 provides a literature review, identifying current gaps and recent advancements in neighborhood-scale energy forecasting using machine learning. Section 3 outlines the methodology, including the rationale for selecting specific variables, machine learning models, and evaluation methods. Section 4 presents the development of the proposed framework and its integration into BAU workflows. Section 5 discusses the development of the framework, while Section 6 discusses its application and the testing of Phase 1 through a case study. Section 7 discusses the results obtained, and Section 8 concludes the paper by summarizing the key findings, contributions, and potential directions for future research.

2. Literature Review

The appeal of data-driven models lies in their ability to provide fast and reliable energy consumption information. However, surpassing traditional BPS methods remains a complex challenge. Despite their promise, these models face significant obstacles in fully replacing conventional BPS approaches due to entrenched practices and various field complexities. As data-driven BPSs continue to advance, they still face challenges such as high computational demands and the risk of overestimating energy savings [27]. Issues like underlying biases and ambiguous data processing also persist. However, recent research highlights the promise of data-driven tools [28,29], including models like artificial neural networks [30,31], support vector machines [32,33], and decision trees [34], which are increasingly used in BPSs [35]. While various researchers have explored different tools [28,36], simulation techniques, algorithms, and evaluation indicators [27,37,38], a gap remains in integrating these advanced machine learning models into practical workflows for neighborhood-scale energy forecasting.

In BPS applications, various methods such as polynomial regression, multivariate adaptive regression splines, Gaussian processes, support vector machines, and artificial neural networks are commonly employed. For instance, Romani et al. [39] applied polynomial models to optimize the heating and cooling energy requirements for a low-energy building in Morocco. Similarly, Cheng and Cao [40] enhanced the prediction of building energy performance by developing a method that leverages evolutionary multivariate adaptive regression splines. To assist with design guidance and performance labeling for passive commercial buildings in hot climates, Rackes et al. [41] utilized support vector machines. Additionally, Yuan et al. [42] introduced a technique based on Gaussian processes that simultaneously calibrates and ranks parameters within building energy models.

The integration of AI and ML models into urban BPSs has gained significant attention recently. Nutkiewicz and Jain [43] examined how physics-based building simulation methods can be integrated with machine learning techniques, specifically using transfer learning to assess the impact of retrofit policies on urban structures. Their integrated approach, known as the Data-driven Urban Energy Simulation, demonstrated its effectiveness in identifying the energy implications of retrofitting urban buildings. In a similar study, Neumann et al. [44] explored the feasibility of creating Positive Energy Districts across different urban typologies in Vienna. Their research emphasized the necessity for comprehensive energy efficiency measures, electrification, and the incorporation of renewable energy sources to transform existing buildings. Focusing on large-scale building analysis, Dai et al. [45]

introduced an innovative methodology that automatically measures building dimensions from remote sensing data using unsupervised machine learning algorithms. Hey et al. [46] highlighted the importance of modeling energy retrofits in urban residential buildings and proposed a concept where carbon valuations inform optimal retrofit solutions. Their approach combined surrogate models, optimization procedures, and neural networks to evaluate building performance, offering valuable insights into policy decisions. Additionally, AI-driven models provide detailed analyses of building energy demand [47], clarify energy dynamics in urban microclimates [48], and categorize unique energy consumption patterns [49], thereby enabling more informed decision-making. At the same time, as urban areas increasingly embrace renewable energy sources, data-driven models become indispensable for forecasting, planning, and optimizing energy consumption and distribution. By utilizing diverse datasets to generate predictions, these models offer valuable insights into emission levels and consumption patterns within complex urban environments.

A longstanding issue in BPSs (both at building and neighborhood scale) is addressing the performance gap—the often significant divergence, sometimes up to 30% [50]—between the anticipated energy performance and the actual energy consumption of buildings. This gap can arise from several factors, including the use of unsuitable modeling tools [51]. Therefore, selecting appropriate modeling tools and methodologies that are validated for energy performance modeling is essential, especially with AI-driven tools. Discrepancies between predicted and actual energy consumption and savings are often linked to challenges in accurately representing occupant behavior [52], interactions between building systems [53], uncertainty in model parameters [54], and operational inefficiencies resulting from low maintenance and issues in the management of building systems [55].

Historically, the limited availability of empirical energy data has posed significant barriers to validating engineering estimates of potential energy savings in buildings. However, the growing adoption of energy benchmarking practices [56] has recently led to a notable increase in accessible building energy data [57]. These datasets offer a real-world basis for analysis, which is indispensable for validating and refining simulation tools used in the design and retrofitting of buildings. By utilizing metered data, the analysis gains a level of precision that hypothetical or averaged datasets cannot provide. This is crucial for the development of reliable energy performance benchmarks and the establishment of energy-saving strategies that are both effective and practical. Moreover, the breadth and depth of those publicly available datasets allow for optimization in the selection of parameters. With numerous instances—each instance representing individual buildings with unique characteristics—and a wide array of parameters capturing various aspects of energy usage and building features, it is possible to tailor a specific data selection process.

In summary, while significant progress has been made in applying machine learning techniques to building and urban energy modeling, there is a research gap in developing and integrating ML-based frameworks into existing design workflows for neighborhood-scale energy forecasting during early design stages. Most existing studies focus on specific algorithms or applications without addressing how these tools can be seamlessly incorporated into the BAU workflows, thus considering the available input data related to the design stage. Additionally, the potential of leveraging large-scale datasets to enhance ML models for practical application has not been fully realized. Addressing this gap is crucial for enabling practitioners to make informed decisions that align with sustainability goals and carbon reduction pathways.

3. Methodology

The methodology for this study was structured into four distinct phases.

The first step involved the identification of a BAU workflow. This baseline workflow was developed based on the authors' experiences and informal interviews with design firms and professionals in both the European Union and the United States, encompassing diverse disciplines within neighborhood-scale architectural design.

In the second step, the research team identified potential areas for improvement and the opportunities to integrate data-driven tools within the established workflow. This process was driven by pinpointing stages where data-driven tools could enhance efficiency, outcome predictability, and design exploration. The integration areas were selected based on their potential to provide substantial improvements over the traditional methods, particularly in terms of data processing, simulation capabilities, and information exchange between different actors.

The third step involved the development of a framework—here introduced at a conceptual level—detailing how data-driven tools, and, in particular, ML algorithms, can be applied to the workflow identified in the first phase.

The fourth and final step tested the feasibility of the initial phase (Phase 1) of this framework with the development of a ML-based tool to optimize massing and functional program. This involved the development of the necessary code and its application to a theoretical case study based in New York City. The new ML-driven approach and the traditional BAU, focusing on traditional model-based simulations, were then directly compared.

3.1. Data Collection and Preprocessing

To build a robust predictive model for energy consumption, a comprehensive dataset that captures a wide range of building characteristics was required. In the realm of building energy consumption analysis, two datasets stand out for their comprehensive coverage and public availability: the Residential Energy Consumption Survey (RECS) in the United States and its commercial counterpart, the Commercial Buildings Energy Consumption Survey (CBECS). These datasets offer an extensive collection of real-world, metered data on energy use across diverse residential and commercial building typologies. The RECS dataset, a product of the United States Energy Information Administration (EIA), offers a detailed account of energy consumption within the residential sector, providing insights into the energy expenditures and equipment usage of homes across the United States. Similarly, CBECS yields granular information regarding the energy-related characteristics of commercial buildings, including their energy usage, equipment types, and operational practices.

In order to develop a comprehensive and flexible database for testing the framework, the two datasets were cleaned and merged into a single one, representing most building typologies from a geometrical and operational energy perspective. Starting from a deep analysis of all the parameters, a selection of the most appropriate and relevant parameters was conducted. The selection of variables was guided by their relevance to energy consumption patterns and their availability across both datasets. Parameters such as building geometry (e.g., square meters, number of floors), building typology, climate zones, and energy consumption data were included to ensure that the model captured the essential factors influencing EUI. The objective of the selection was to represent the buildings' energy patterns while reducing the number of inputs needed, thereby enhancing the model's applicability in early design stages where detailed data may not be available.

3.2. Machine Learning Models

To identify the most effective ML model for predicting energy consumption, several models were tested, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and CatBoost Regressor. Deep learning networks, such as Long Short-Term Memory and Recurrent Neural Network models, were not included in this study because they are primarily designed for time-related predictions and sequential data analysis. Given the focus on static predictions of EUI based on building characteristics rather than temporal patterns, traditional regression and ensemble methods were deemed more suitable for this task.

The rationale behind testing these models is based on their unique strengths and capabilities:

- *Linear Regression*: This model serves as a baseline due to its simplicity and interpretability. It helps to understand the linear relationships between input features and energy consumption.

- *Decision Tree Regressor*: This model is useful for capturing non-linear relationships by splitting the data into subsets based on feature values, providing a straightforward interpretation of decision rules.
- *Random Forest Regressor*: As an ensemble method, it combines multiple decision trees to improve prediction accuracy and reduce overfitting. It is robust to noise and captures complex interactions between features.
- *Gradient Boosting Regressor*: This model builds an ensemble of weak prediction models, typically decision trees, in a sequential manner. It optimizes model performance by correcting errors from previous iterations, making it highly effective for complex datasets.
- *CatBoost Regressor*: This advanced gradient boosting algorithm is specifically designed to handle categorical features without extensive preprocessing. It offers high accuracy, efficient training, and robustness to overfitting, making it well suited for energy consumption prediction tasks.

3.3. Model Training and Evaluation

All the models were tested on the full dataset, with a train–test split of 80%-20%, meaning the models were trained on 80% of the data and tested on the remaining 20%. Hyperparameter tuning was conducted for each model using grid search and cross-validation techniques.

For example, for the CatBoost Regressor, parameters such as the number of iterations, learning rate, depth, and L2 regularization coefficient were optimized. The final selected parameters are as follows:

- *CatBoost Regressor*: iterations = 1000, learning_rate = 0.03, depth = 6, l2_leaf_reg = 3.0.
- *Gradient Boosting Regressor*: n_estimators = 100, learning_rate = 0.1, max_depth = 3, alpha = 0.9.
- *Random Forest Regressor*: n_estimators = 100, max_depth = None.
- *Decision Tree Regressor*: max_depth = None.
- *Linear Regression*: default parameters.

After evaluating the performance of these models using cross-validation and multiple error metrics—including R-squared (R^2), Mean Absolute Error, and Root Mean Squared Error—it has been observed that all metrics consistently identified the CatBoost Regressor as the most accurate and reliable model. For simplicity and ease of interpretability, and since all metrics yielded the same conclusion regarding the best-performing model, it has been opted to present only the R^2 scores in the results. The CatBoost Regressor achieved the highest R^2 scores for both the cross-validation and test datasets, demonstrating its ability to capture the complexities of urban energy consumption patterns. Specifically, it achieved an R^2 score of 0.88, a Mean Absolute Error of 81,683 kWh, and a Root Mean Squared Error of 315,960 kWh.

Recognizing the necessity for computationally efficient modeling, the computational time required by the CatBoost Regressor has been evaluated. While a detailed time comparison with traditional physics-based modeling tools is complex and beyond the scope of this paper, it is reported that the CatBoost Regressor achieved notable computational efficiency with a training time of 2.61 s. This demonstrates the model's capability to provide rapid predictions, highlighting its suitability for early-stage design processes where a quick turnaround is essential.

In summary, this study employs several advanced machine learning and simulation techniques to analyze and optimize neighborhood-scale energy consumption. Specifically, SHapley Additive exPlanations (SHAP), CatBoost Regressor, Grasshopper, and Honeybee were utilized, each serving distinct roles within the framework. SHAP provides interpretability to the machine learning models by attributing the output predictions to input features, thus helping us understand the impact of each feature on the energy consumption predictions [58]. CatBoost Regressor, a gradient boosting algorithm that handles categorical features efficiently, was chosen for its high predictive accuracy and robustness, making it suitable for energy consumption modeling [59,60]. Grasshopper and Honeybee, which

are plugins for the Rhino environments, were employed for their powerful capabilities in parametric design and energy modeling [61–64]. Grasshopper facilitated the creation of complex geometrical configurations and parametric variations, which were then analyzed for their energy performance using Honeybee. Honeybee leveraged an EnergyPlus engine to perform detailed energy simulations, providing us with granular insights into the energy dynamics of different building designs.

The results from this case study provided critical insights into the improvements offered by the phase, highlighting the framework's potential to integrate energy efficiency in urban planning from the earliest design stages.

4. Workflow Identification

The effective integration of advanced digital technologies into architectural and urban design practices depends on their seamless incorporation into the established workflows of designers operating at the district and neighborhood levels. Notably, the growing preference for data-driven methodologies [65] aligns with the broader adoption of such technologies across diverse industries. This widespread uptake is largely due to their intuitive usability and ability to complement existing practices without disrupting traditional workflows.

A standardized workflow was identified based on expert judgment and interviews with professionals, as illustrated in Figure 1 and detailed in Table 1. Data-driven analysis played a pivotal role throughout the process, from initial planning stages to construction, with its impact being most pronounced during the early-stage design phase.

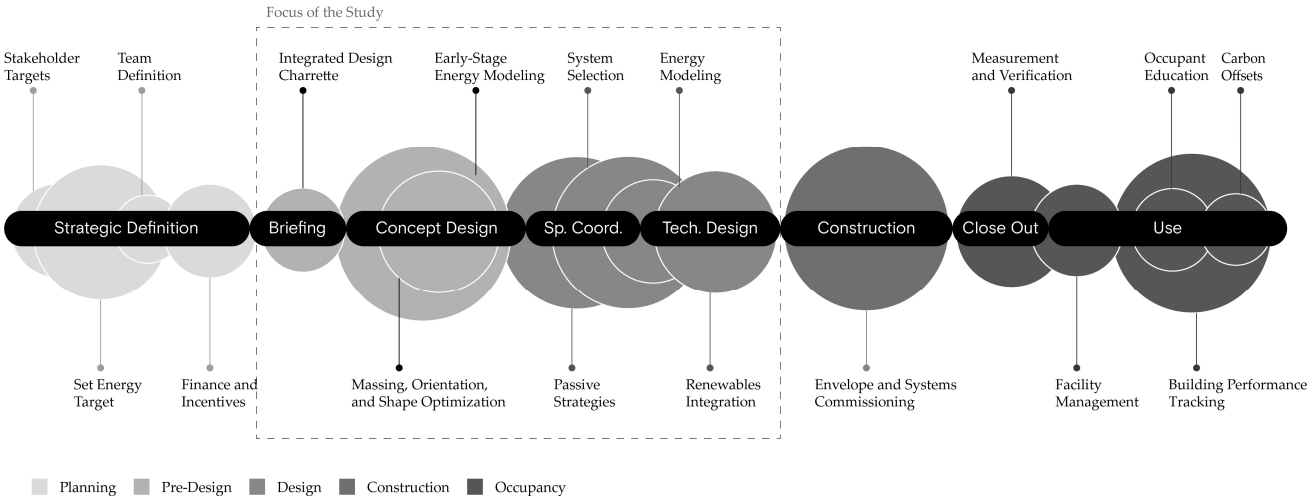

Figure 1. "Phases of the building design lifecycle—Each phase is depicted with a distinct color to denote the corresponding phase, while the circle dimension indicates the level of influence on the project's energy efficiency and sustainability outcomes" (adapted from [66]).

The BAU workflow [66] highlights specific stages, particularly during the early design phases, where the integration of data-driven tools can offer significant benefits. These early stages often involve frequent iterations among various teams and professionals, increasing the risk of errors and information loss. Such challenges present key opportunities for embedding digital tools into the process. The proposed framework promotes a holistic approach, positioning digital tools as integral to the design workflow. It underscores the value of iterative feedback loops, collaborative engagement with stakeholders, and adaptability to diverse project scales and typologies.

Table 1 shows in more detail the highlighted portion of the workflow (Briefing to Technical Design), defining the areas in which the proposed framework could be integrated.

Table 1. Business-as-usual detailed workflow.

Design Phase	Client and Other Stakeholder	Other Disciplines	Sustainability Expert	Energy Modeler (Envelope)	Energy Modeler (Systems)
Briefing	Brief Definition GEA */GFA ** Definition	Set Up Masterplan Concept	Schematic Design Design Basis and High-level Specifications	Define the Regulatory Regime Preliminary Energy Use (Milestone 01)	-
Concept Design	Validate the Building Concept	Define the General Building Envelope Components Run Solar Radiation Analysis	Define Thermal Zones and Conditioned Spaces Collect Data on the Producibility of the PV System	Create Building Massing Model (Milestone 02)	System Definition
Spatial Coordination	Specific Requirements for the PV System	Collect Total Length of Distribution Pipes Extract the Air Flow Rates Air Handling Units Color Plan Run Detailed PV Producibility Analysis	Define Building Schedule and Diversity Rates Collect Information about Systems at Building Scale Collect Information about Air Flow Rates at Building Scale Collect Lighting Loads at Building Scale Collect Information about Urban Scale Systems	Extract PV Production Monthly Data	Set Up the Plant System Configuration Input Lighting Loads Input HVAC Information Input Air Flow Rates Input Producibility of PV System Energy Code Compliance Check (Milestone 03)
Technical Design	-	Mechanical Schedules	Clash and Interference Check (Milestone 04)	Performance Indices Check System Efficiency Check	Heating and Cooling Thermal Performance Index Global Performance Index Renewable Coverage

* Gross External Area; ** Gross Internal Area.

5. Framework Development

In this context, integration techniques refer to the synthesis of data flows and modeling processes, creating a cohesive approach to neighborhood-scale design. The rise in machine learning and data-centric methodologies marks a significant evolution in the BAU workflow, offering notable advancements across the field.

The proposed framework—illustrated in Figure 2 and elaborated in Table 2—introduces the application of three distinct machine learning algorithms to optimize the workflow while preserving the rigor of various assessments. This framework is built on the premise that each project generates unique data streams. Consequently, adopting a layered machine learning strategy enables the selective application of specific components of the model as needed, avoiding the necessity of completely restructuring the existing workflow.

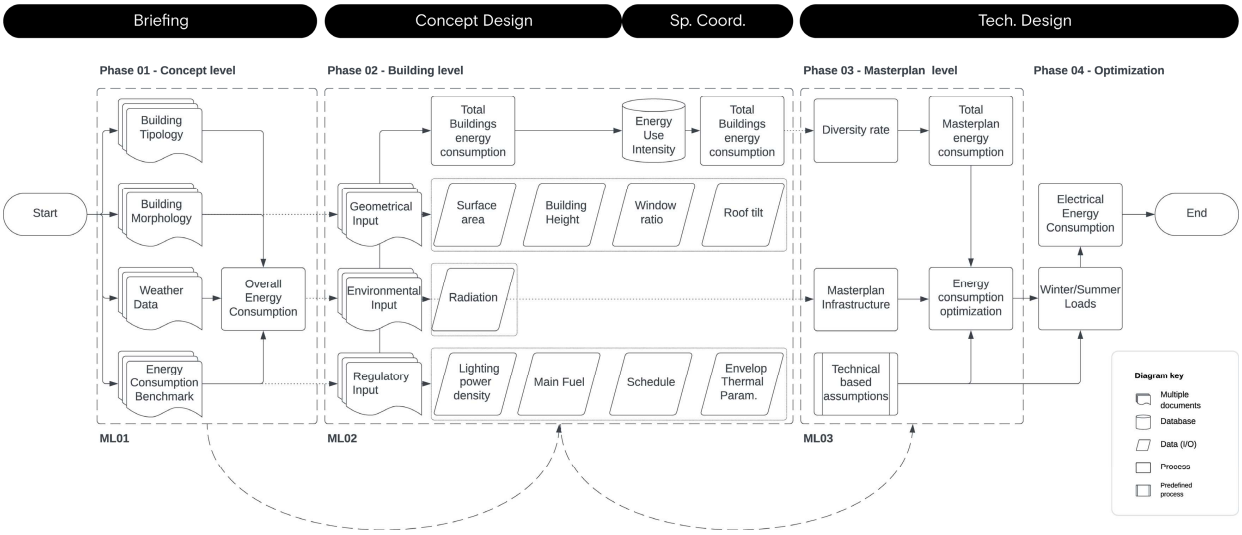

Figure 2. Framework definition.

Table 2. Framework phases and applications.

Phase	Use Case	Inputs	Outputs
Phase 01—Concept Level	Service for policymakers, urban planners, developers and energy providers to understand the overall energy requirements of a masterplan.	Climatic data Building typology Building morphology	Overall energy consumption predictions Identification of critical needs Potential environmental and economic issues
Phase 02—Building Level	Service (or internal use) for architects for early-stage project optimization and for developers for defining building development rules.	Building geometry Environmental inputs	Detailed energy usage predictions Optimization recommendations
Phase 03—Masterplan Level	Service for designers, urban planners, and energy providers to understand urban scale energy consumption and develop active strategies.	Multiple building data Urban layout Energy systems data	Comprehensive energy consumption analysis Urban energy strategies System efficiency recommendations

In the initial phase, related to the concept level (Phase 1), the integration of machine learning can impact the schematic design stage by rapidly synthesizing and interpreting vast datasets to identify optimal design configurations (overall strategy) and energy needs (building scale). This can lead to a reduction in design time and enable a more informed decision-making process regarding the building envelope's performance characteristics.

5.1. Phase 1—Concept Level (ML01): Preliminary Analysis and Data Synthesis

At the outset, Phase 1 serves as the foundational stage, specifically at the Briefing stage, where the initial machine learning algorithm is deployed to analyze building morphology and typology. This phase harnesses architectural principles and standards to set baseline models for various building classes. Integration of weather data aids in understanding the impacts of seasonal and regional variations on energy consumption. Simultaneously, benchmarks from datasets like CBECS and RECS establish industry standards for assessing building performance.

This synthesis of data enables preliminary energy consumption predictions, setting a conceptual benchmark that will be further refined.

5.2. Phase 2—Building Level (ML02): Detailed Building Analysis

Advancing to Phase 2, machine learning refines the massing and establishes the basis for urban models by predicting the effects of design modifications, thereby enhancing sustainability and energy efficiency. This phase introduces specific physical parameters—building footprint and height—crucial for calculating the building's EUI. Further analysis incorporates the surface-to-volume ratio, significantly influencing thermal performance and energy consumption, thus providing an overall understanding of building energy demands. This phase of the framework is not meant to substitute traditional energy modeling tools and, therefore, is not supposed to accurately describe all the typical building properties used in a simulation—that would require many additional inputs.

Rather, it is designed to be integrated into existing workflows, specifically during the Concept Design and Spatial Coordination stages, to provide quick insights into urban energy that inform preliminary design decisions and help streamline the early stages of the project.

5.3. Phase 3—Masterplan Level (ML03): Urban-Scale Energy Modeling

Transitioning to Phase 3, the focus expands to the urban scale, employing data-driven methodologies to evaluate building schedules, loads, and system performance comprehensively. This phase offers real-time insights into energy consumption patterns and identifies potential inefficiencies across multiple buildings. The total EUI, assessed in conjunction with the diversity rate, reflects variations in building function and occupancy, enriching the analysis with technical assumptions that capture a wide array of design and operational factors.

The outcome is a sophisticated model depicting the masterplan's energy profile.

5.4. Phase 4—Optimization: Final Optimization

In the final phase, insights from prior analyses are integrated to establish a robust optimization strategy. This phase incorporates a feedback mechanism attentive to seasonal load variations and refines technical assumptions about building operations. The resulting comprehensive strategy not only enhances energy efficiency but also supports sustainable design and policymaking decisions. The integration of these insights culminates in a pragmatic framework designed to optimize energy consumption and support sustainability objectives across urban developments.

Each phase of this framework builds progressively, ensuring that every level of analysis contributes to a more energy-efficient and sustainable urban environment.

5.5. Integration and Iterative Refinement

Throughout all phases, the framework emphasizes iterative refinement and integration. As each phase feeds into the next, the ML algorithms learn and adapt, continually refining the predictive accuracy of the model. This iterative process ensures that the model remains dynamic and responsive to new data and changing conditions, ultimately culminating in an optimized design that aligns with contemporary sustainability standards and practices.

In order to develop a comprehensive and flexible database for testing the framework, the two datasets have been cleaned and merged into a single one, representing the most building typologies from a geometrical and operational energy perspective. Starting from a deep analysis of all the parameters, a selection of the most appropriate and relevant has been conducted. The objective of the selection is to represent the buildings' energy patterns, while reducing the number of inputs needed. Table 3 represents the merging operation on the cleaned datasets.

Table 3. CBECS and RECS data merging.

CBECS 2018	RECS 2020	Merged Dataset	Description
PUBCLIM	BA-climate	PUBLICM	ASHRAE climate zone
REGION	REGIONC	REGION	Census region
NFLOOR	STORIES	NFLOOR	Number of floors
BTU	KWH	TOTALKWH	Total energy consumption (yearly)
SQFT	SQFTEST	SQM	Total area
WINTYP	TYPEGLASS	WINTYP	Window glass type
RENWIN	ORIGINWIN	RENWIN	Windows upgrade
PBA	TYPEHUQ	PBA	Principal building activity
HDD65	HDD65	HDD65	Heating degree days (base 65)
CDD65	CDD65	CDD65	Cooling degree days (base 65)
YRCONC	YEARMADERANGE	YRCONC	Year of construction
WLCNS	WALLTYPE	WLCNS	Wall construction material
RFCNS	ROOFTYPE	RFCNS	Roof construction material
RENINS	ADQINSUL	RENINS	Insulation upgrade
FUEL	FUELHEAT	FUEL	Main space heating fuel
WKHRSC	-	WKHRSC	Weekly working hours

To merge the datasets, overlapping parameters between CBECS and RECS were first identified and harmonized. Each parameter was matched based on its relevance and representation in both datasets.

The merging process involved standardizing units, resolving discrepancies in parameter definitions, and addressing missing values through statistical imputation techniques. The merged dataset is structured as follows:

- *Location and Geographic Information (ASHRAE climate zone, census region, BA-climate, heating degree days, cooling degree days):* Both RECS and CBECS contain parameters related to the geographic location of the buildings surveyed, including region, state, and sometimes urban or rural classification. This information can be critical for understanding and analyzing energy usage patterns due to climatic, cultural, and infrastructural differences.
- *Building Characteristics (number of floors, total area, principal building activity, year of construction, window glass type, wall construction material, roof construction material):* Key parameters include the age of the building, size in square footage, number of floors, typology, and building characteristics. These factors are vital as they directly influence the building's thermal properties and, consequently, its heating and cooling demands.
- *Heating, Ventilation, and Air Conditioning Systems (main space heating fuel, weekly working hours):* Parameters cover the types of heating and cooling systems present, fuel types used (such as electricity, natural gas, fuel oil), and the working hours of the building.
- *Energy Use (total energy consumption):* This category includes detailed metered data on annual energy consumption. These parameters are critical for assessing energy performance and identifying savings opportunities.
- *Renovation Features (insulation upgrade, windows upgrade):* Parameters that capture the presence of renewed insulation and or windows. These factors, related to the year of construction and the building characteristics, are useful to assess the uptake and impact of these technologies on overall energy consumption and carbon emissions.

Weather-related parameters were selected due to their significant impact on heating and cooling demands, while home characteristics such as the building's age and wall type were chosen for identifying specific energy consumption patterns in similar buildings. Moreover, these parameters have implications on retrofitting potential and energy efficiency improvements. The inclusion of high-level HVAC system information is crucial to understand and potentially optimize energy consumption patterns, thus influencing performance assessment.

By leveraging the selected parameters, designers and stakeholders can gain valuable insights into the energy and carbon implications of their design choices. Despite the rigorous data selection process, challenges such as data gaps, regional discrepancies, and the exclusion of certain parameters were acknowledged. Efforts were made to ensure a balanced representation of various building types and climates to mitigate these issues.

6. Framework Application and Testing

As described in the previous sections, this paragraph is going to test the applicability of Phase 1. This first step of the framework stands on several technical assumptions intrinsic to the dataset and the modeling techniques used. The primary technical assumptions include the following:

- *Data Integrity:* It is presumed that the dataset is representative of the broader population of commercial buildings and that the data collection process was devoid of systematic bias.
- *Conversion Factors:* The conversion of energy units from kBTU to kWh and square footage to square meters is based on standard conversion factors, assuming no loss of information.
- *Statistical Fill:* Missing values within the dataset are filled with zeros, under the assumption that they correspond to non-usage or unreported data, which does not distort the overall energy consumption profile.
- *Variable Selection:* The choice of variables to keep in the dataset is driven by domain knowledge, considering factors such as climate, region, primary building activity, year of construction, and energy consumption across multiple sources.

The simulation process commences with data preprocessing, where the dataset is refined to include only the most pertinent variables in the analysis. The simulation process began by combining the two selected datasets, CBECS and RECS, as described previously. Initial data visualization, using histograms and heatmaps, helped us understand the distribution of energy consumption and identify inter-variable relationships. Outlier detection algorithms were employed to identify and exclude anomalous data points that could negatively impact prediction accuracy. Specifically, the combined data were normalized—removing outliers—based on the EUI metric (measured in kWh/m^2). Any data points that fell outside 1.5 times the interquartile range were excluded. This step mitigated the influence of anomalous observations that could skew the predictions, without oversimplifying the dataset. The original dataset consisted of 24,932 rows, and 22,865 after the initial data cleaning.

Figure 3 illustrates the EUI distribution in the combined dataset. On the left is the original distribution, and on the right is the cleaned one, consisting of 22,865 measurements (against the 24,932 measurements of the original combined dataset). The frequency decreased from approximately 1400 to 1200 due to the removal of outliers, which were data points falling outside 1.5 times the interquartile range. This cleaning process helped ensure that the dataset better represented typical energy use patterns, improving the accuracy of subsequent predictive modeling.

After a first dataset normalization phase, the code analyzes linear correlations between the selected parameters. The heatmap in Figure 4 shows the correlation matrix for the filtered dataset. These attributes reflect commercial and residential building characteristics and their relationship with energy consumption, measured in kilowatt-hours (TOTALKWH) and the total building area in square meters (SQMs). The color scale, ranging from dark blue to dark red, indicates the strength and direction of the correlations. Dark red represents a strong positive correlation, while dark blue indicates a strong negative correlation. Significant correlations are evident between several pairs of variables. The size of the building, represented by SQMs, shows a strong positive correlation with energy consumption (TOTALKWH). Variables like insulation upgrade (RENINS) and window upgrade (RENWIN) show a strong positive correlation with each other, indicating that renovations often occur together as part of comprehensive retrofit projects aimed at im-

proving energy efficiency. However, they do not show a significant correlation with total energy consumption.

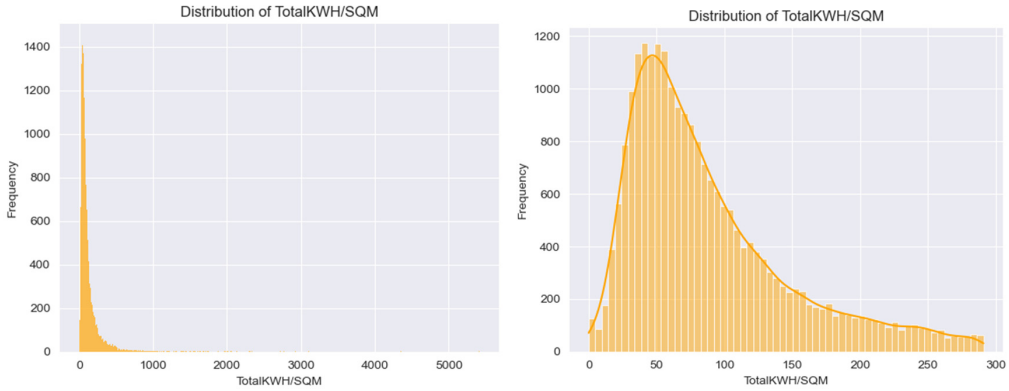

Figure 3. EUI distribution in the combined dataset. On the left is the original distribution, and on the right the cleaned one, consisting of 22,865 measurements.

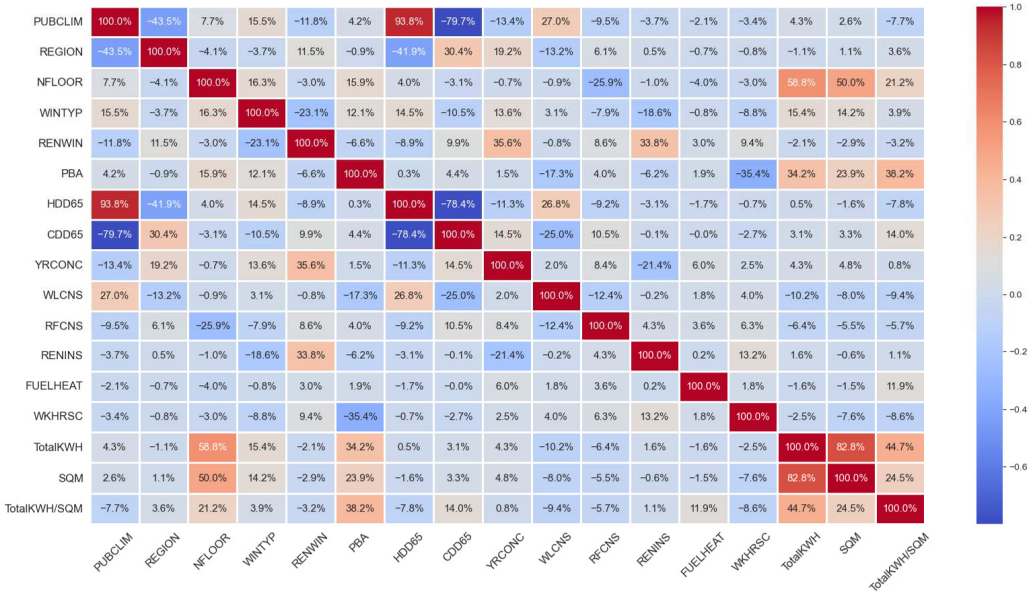

Figure 4. Heatmap showing the correlation matrix for selected building attributes. The color scale ranges from dark blue to dark red, representing the strength and direction of the correlations.

These relationships are critical for both designers and energy modelers as they seek to identify factors that most influence energy use in buildings. Understanding these correlations can lead to more accurate energy models and targeted energy efficiency measures.

After this initial step, various machine learning models, including the Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and CatBoost Regressor, were compared to predict energy consumption. The selection of the most predictive and generalizable model was informed by cross-validation and R-squared metrics. All models were trained on 80% of the dataset and tested on the remaining 20%.

Table 4 provides a comparative evaluation of the performance of various ML models used to predict energy consumption. The models are assessed using the R^2 metric for both cross-validation and test datasets reflecting how well each model captures the variance in the data.

Table 4. ML models' performance comparison.

ML Model	Cross-Validation R^2	Test R^2
Linear Regression	0.57	0.52
Decision Tree Regressor	0.54	0.56
Random Forest Regressor	0.75	0.75
Gradient Boosting Regressor	0.76	0.75
CatBoost Regressor	0.88	0.88

The Linear Regression model shows modest predictive performance, with an R^2 score of 0.57 for cross-validation and 0.52 for the test dataset. This indicates that the model captured just over half of the variance in the energy consumption data but struggled with more complex relationships due to its linear nature.

The Decision Tree model achieved a similar performance to linear regression, with R^2 scores of 0.54 and 0.56 for cross-validation and test datasets, respectively. While decision trees can handle non-linear relationships, this model may have overfitted on certain features, leading to moderate predictive accuracy.

Random forests, a type of ensemble model, demonstrated significant improvement in predictive performance, achieving an R^2 score of 0.75 for both cross-validation and test datasets. This model excels at handling data complexity and generalizes well, leading to better predictions.

Similarly to random forests, gradient boosting also uses an ensemble approach but builds models sequentially to correct previous errors. With R^2 scores of 0.76 and 0.75 for cross-validation and test datasets, respectively, it indicated strong predictive capabilities.

The CatBoost Regressor achieved the highest R^2 scores, 0.88 for both cross-validation and test datasets, demonstrating the best predictive accuracy. CatBoost is particularly effective at handling categorical features and reducing overfitting, making it an effective choice for energy consumption prediction.

The performance comparison suggests that ensemble models, particularly the CatBoost and Gradient Boosting Regressors, provided the most accurate predictions. The results indicate that using advanced ensemble techniques significantly improves predictive performance over simple models like Linear Regression or Decision Trees. However, as shown in Figure 5, it is evident that buildings with lower energy consumption are better represented by the model.

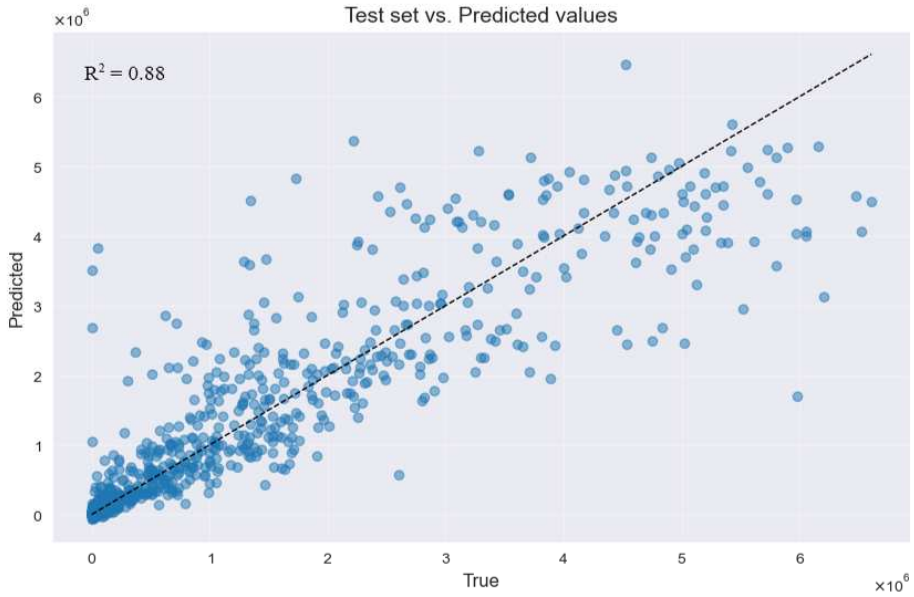

Figure 5. Scatter plot of test set versus predicted values showing the performance of the machine learning model. The diagonal line represents perfect predictions, indicating the model's accuracy.

The scatter plot graph in Figure 5 illustrates the relationship between predicted and actual energy consumption values across a test dataset. The X-Axis (*True*) represents the actual, observed energy consumption values, while the Y-Axis (*Predicted*) represents the predicted energy consumption values using the model developed. The black dashed line running diagonally represents the line of perfect prediction, where each predicted value matches exactly with its corresponding actual value. If the model was perfectly accurate, all data points would fall directly on this line. The individual blue dots represent each building or entity in the dataset. Most of these points cluster closely around the perfect prediction line, indicating that the predictive model performs reasonably well, with many predictions being quite accurate. A few data points deviate significantly from the line, which may be due to various factors such as irregular building characteristics, extreme weather variations, or inaccuracies in the input data.

The spread of points widens as energy consumption increases, suggesting a higher variance in predictions at larger consumption scales. Despite some deviations, the overall trend closely aligns with the perfect prediction line, showing a positive linear relationship between actual and predicted values. This indicates that the predictive model captured the general pattern of energy consumption reasonably well.

In summary, the scatter plot offers a comprehensive view of the model's performance. While it shows some inconsistencies, the proximity of most points to the perfect prediction line indicates that the model is largely effective, providing useful predictions for energy consumption with a reasonable degree of accuracy.

The visual comparison between actual and predicted energy consumption in the bar graph (Figure 6) offers a clear representation of model performance in predicting building energy use. The blue bars represent the actual energy consumption values, while the orange bars depict the model's predictions for each corresponding instance in the test dataset. The model captures the general trend of energy consumption across the dataset; however, there are noticeable disparities between the actual and predicted values in specific instances.

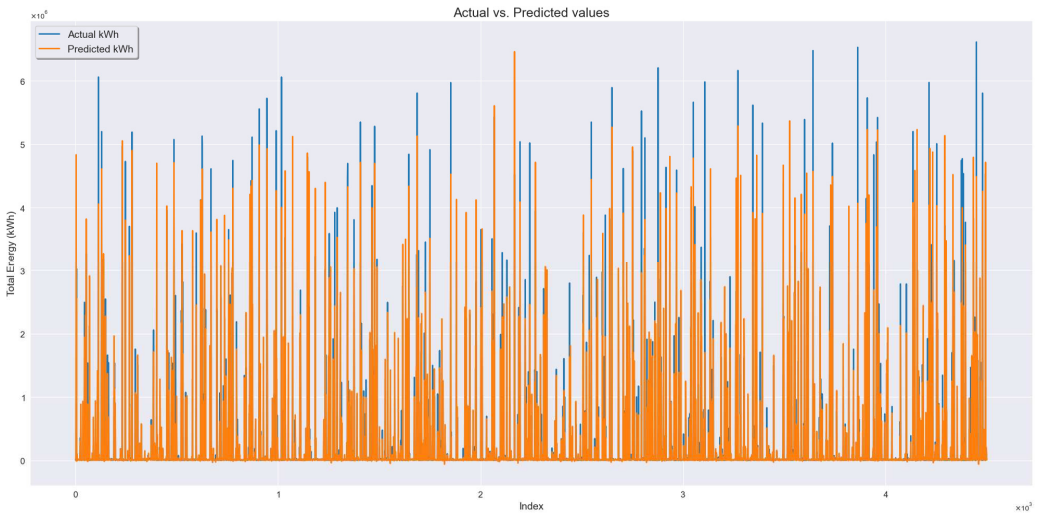

Figure 6. Comparison of actual versus predicted energy consumption values (in kWh) for the test dataset. The blue bars represent actual values, while the orange bars represent predicted values, illustrating the model's performance and accuracy in predicting energy use.

These disparities could be due to the model's inability to capture some of the more complex patterns within the data or to unexpected anomalies that were not represented in the training set. The sharp peaks and troughs in energy usage suggest that the model performs well in lower ranges of energy consumption but may struggle with accurately predicting higher consumption values. This discrepancy could indicate the presence of outliers or extreme values in the test data that were not well represented in the training phase.

The SHAP summary plot provided in Figure 7 visualizes the impact of each feature on the CatBoost model's output. This plot is a tool for interpretability, explaining the prediction of a machine learning model in a more understandable way. In the plot, each point represents a SHAP value for a feature and an instance. The position on the horizontal axis indicates the impact of the feature on the model's prediction. Features are stacked vertically in descending order of importance, with SQM at the top, indicating it as the most impactful feature. The color represents the feature value from low to high (from pink to blue), providing insight into how the value of the feature impacts the prediction. For example, higher values of SQM tend to push the model prediction higher, which can be seen as many blue points have positive SHAP values.

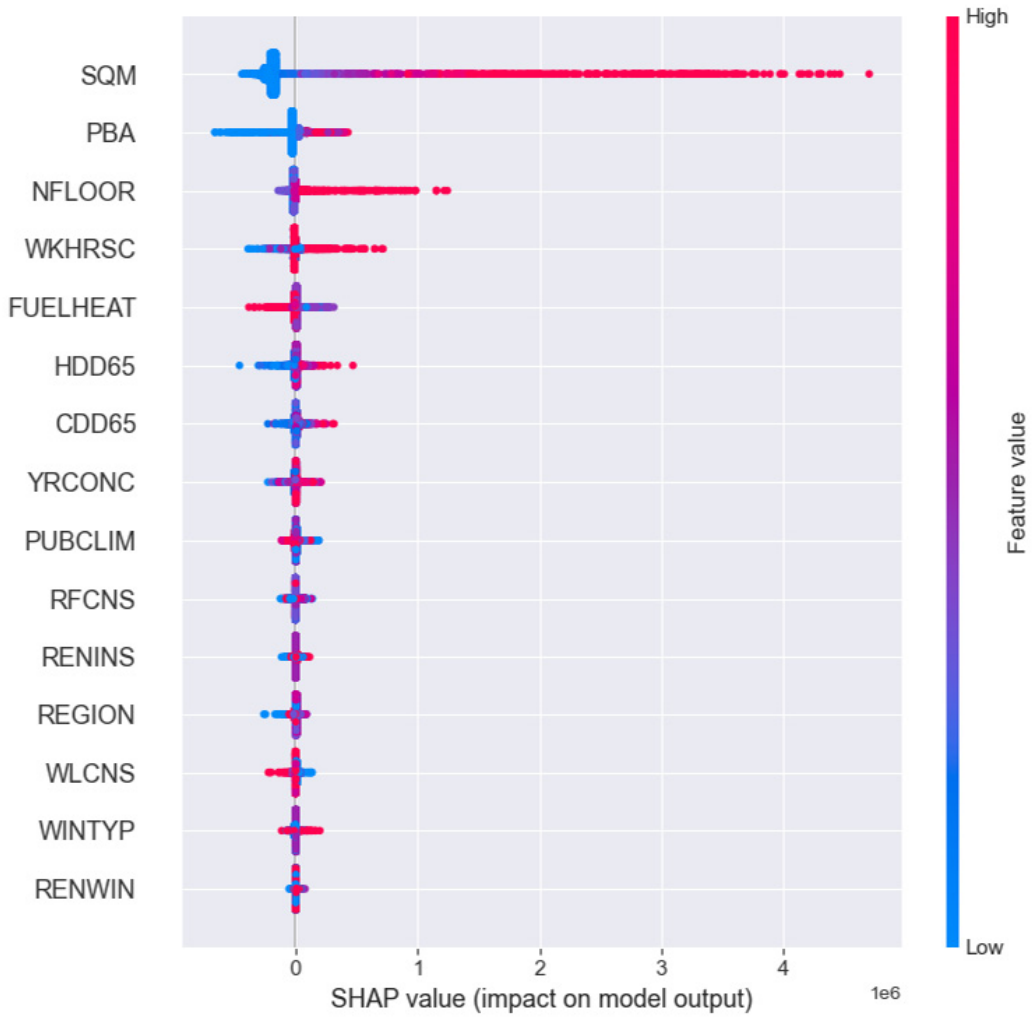

Figure 7. SHAP summary plot illustrating the impact of various features on the model's predictions for energy consumption. The horizontal axis represents the SHAP value, indicating the influence of each feature on the model's output. Features are ranked in descending order of importance.

Several observations can be made from this plot:

- Square Meters (SQM): As the most influential feature, higher values lead to significantly higher SHAP values, suggesting that larger buildings tend to consume more energy.
- Primary Building Activity (PBA): This feature shows a mix of positive and negative impacts, indicating that building use cases (such as office vs. residential) affect energy consumption differently.
- Number of Floors (NFLOOR): Higher building stories tend to have higher SHAP values, indicating greater energy use.

- Working Hours (WKHRSC): A longer working schedule correlates with increased energy consumption.
- Heating Degree Days (HDD65), Cooling Degree Days (CDD65), and ASHRAE Climate Zone (PUBCLIM): These parameters contribute to energy demand, with higher values indicating a greater need for heating or cooling. PUBCLIM is more balanced—closer to the 0 axis—and can be used for both heating and cooling analysis.
- Insulation upgrade (RENINS) and window upgrade (RENWIN) have varying impacts, possibly due to the clustering of renovations in retrofitting projects.

The SHAP analysis results align with the proposed framework, therefore confirming the knowledge-based hypotheses and emphasizing the significant influence of specific features on predicting building energy consumption. This observation correlates with the conceptual structure outlined in Phase 1, which emphasizes these key parameters in understanding the overall energy consumption at concept level.

After validating the initial findings, the same ML model was applied using only the four most influential parameters identified in the framework, as shown in Table 5. This focused approach yielded consistent results, demonstrating the framework's first phase effectiveness in accurately modeling energy consumption.

Table 5. Final dataset with reduced inputs.

	TOTAL KWH	SQM	PBA	NFLOOR	PUBCLIM
1	5.09×10^5	2601.284	3	5	4
2	1.54×10^4	195.0963	6	1	4
3	9.36×10^5	22,296.72	0	1	5
4	1.37×10^6	27,406.385	0	1	3
5	1.16×10^6	8732.882	5	2	3
6	5.88×10^5	3809.023	5	2	5
7	1.14×10^6	11,334.166	0	2	2
8	8.76×10^5	8918.688	4	1	2
12	6.43×10^5	2833.5415	3	3	3
...					
22,503	5.88×10^4	743.224	5	1	4

The reduced dataset, which focuses on the key parameters identified by the framework and the SHAP analysis, includes four primary inputs: SQM, PBA, NFLOOR, and PUBCLIM. These features were selected due to their significant impact on energy consumption as observed in previous analyses.

The results, as illustrated in Figure 8, obtained from using this reduced dataset, demonstrate the effectiveness of prioritizing these parameters. The R^2 value on the validation dataset was 0.8691, and on the test dataset, it was 0.8435. This high predictive accuracy confirms that these four parameters were instrumental in forecasting energy consumption. Notably, the feature importance ranking shows SQM as the dominant factor, accounting for 62.74% of the influence, followed by PBA (16.59%), NFLOOR (11.02%), and PUBCLIM (9.65%).

This simplified yet accurate model demonstrates that carefully selecting key features allows for precise predictions without compromising computational efficiency. Such streamlined analysis supports data-driven decision-making in neighborhood-scale planning and energy optimization, providing an effective tool for architects, engineers, and policymakers.

Lastly, the predictive model was tested against a BAU energy modeling approach using Grasshopper and Honeybee. This test involved a model comprising four mixed-use test case buildings located in New York City. To evaluate the predictive model's accuracy and robustness, various configurations were applied by altering the buildings' dimensions and functions. This process generated a total of nine distinct configurations, as illustrated in Figure 9.

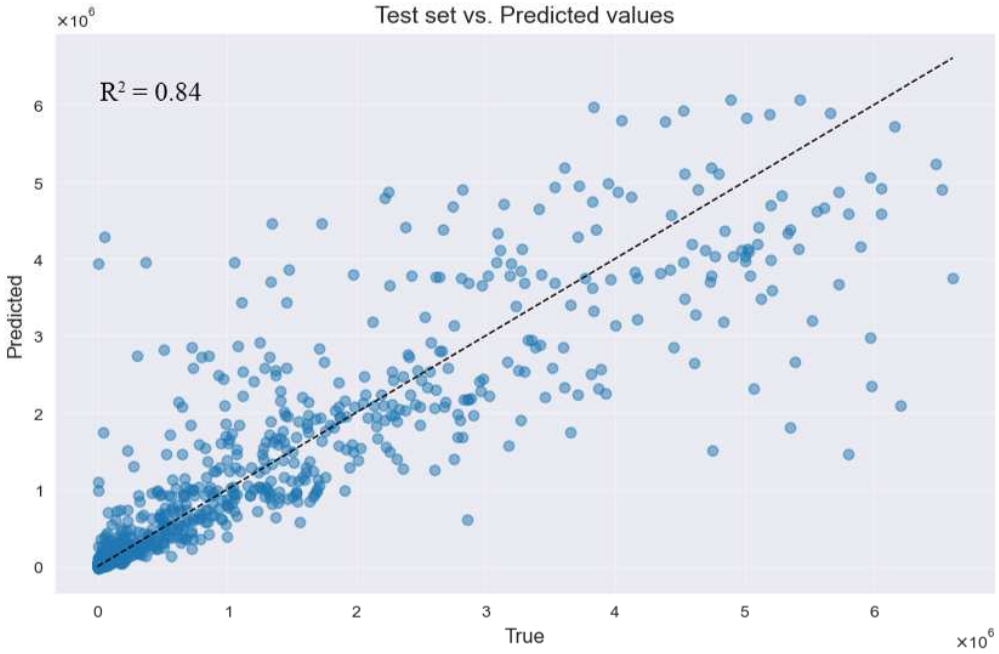

Figure 8. Comparison of actual vs. predicted energy consumption in the reduced inputs dataset.

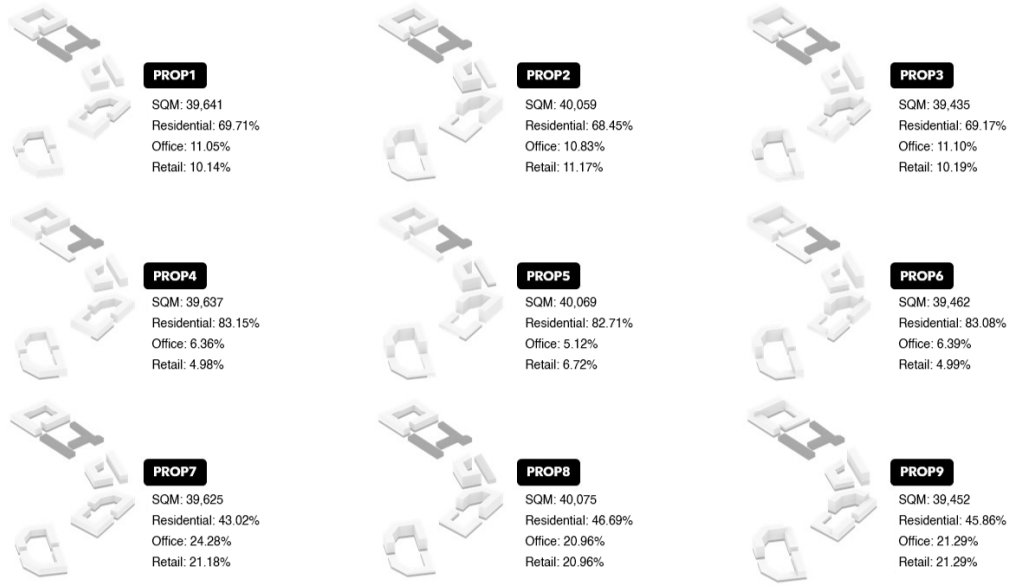

Figure 9. Different buildings cluster configuration. Each color represents a function.

After running the model, the calculated energy consumption values were compared with those predicted by the Grasshopper-based BAU energy model.

The percentage differences between the calculated values and the modeled values are displayed in Table 6. The results show deviations ranging from −8.69% to 11.04%, with the majority of errors staying within a relatively narrow range. These discrepancies highlight the predictive model's performance and its potential to provide energy consumption estimates that align closely with the BAU model.

The findings support a nuanced application of machine learning in energy prediction, highlighting the importance of guarding against overfitting and advocating for an iterative approach to model training and testing. By evaluating multiple models and assessing their performance across various metrics, this study ensures that the selected model not only captures the underlying patterns in the data but also generalizes effectively to new, unseen scenarios.

Table 6. Comparison between BAU energy modeling and ML modeling.

	GH [kWh]	ML01 [kWh]	Error [%]
Prop 1	158.79	152.06	4.24
Prop 2	148.54	161.45	−8.69
Prop 3	180.33	160.42	11.04
Prop 4	160.12	161.26	−0.71
Prop 5	159.56	158.80	0.48
Prop 6	181.74	178.20	1.95
Prop 7	155.19	155.09	0.06
Prop 8	149.55	162.15	−8.43
Prop 9	169.54	152.99	9.76

The results demonstrate the model's potential for practical application in neighborhood-scale energy consumption analysis while underscoring the importance of continual testing and validation against traditional methods. This approach allows designers to incorporate data-driven predictions into their workflows more confidently.

7. Discussion

This study investigated the feasibility of integrating ML tools into early stage neighborhood-scale energy forecasting within business-as-usual, BAU, design workflows. By focusing on the first phase (Phase 1) of the proposed framework, the research focused on predicting building energy consumption using minimal input data typically available during the early design stages.

The results demonstrate that the CatBoost Regressor, an ensemble ML model, effectively predicts energy consumption with high accuracy. Specifically, the model achieved R^2 scores of 0.88 for both cross-validation and test datasets, indicating a strong ability to capture the variance in energy consumption based on limited inputs. This high level of predictive performance confirms the first objective of the study: to facilitate rapid and accurate energy consumption predictions using minimal input data.

The SHAP analysis further validated the framework by identifying the most influential features impacting energy consumption predictions. The key parameters—SQM, PBA, NFLOOR and PUBCLIM—were shown to significantly influence the model's output. By reducing the dataset to these four parameters, the model still maintained a high predictive accuracy (R^2 of 0.84 on the test dataset). This finding underscores the efficacy of the proposed framework in simplifying input requirements without compromising accuracy, aligning with the goal of making the tool practical for early-stage design workflows.

When applied to a hypothetical cluster of buildings in New York City, the model's predictions closely matched those of traditional physics-based energy modeling tools (Grasshopper and Honeybee), with discrepancies ranging from −8.69% to 11.04%. Considering the inherent uncertainties in early-stage design and energy modeling, these discrepancies are acceptable and demonstrate the model's reliability in practical scenarios.

However, the model showed reduced performance with large-scale buildings, likely due to the limited representation of such buildings in the dataset. This limitation highlights the need for incorporating more diverse data, especially for large-scale typologies, to improve model generalizability. Despite this, the overall findings remain robust, as the framework is primarily intended for early design stages where detailed data may not be available.

The ability to predict energy consumption accurately at the early design stages has significant implications for sustainable urban design. It enables architects, designers, and policymakers to make informed decisions that can lead to energy-efficient and sustainable neighborhoods. By integrating ML tools into BAU workflows, practitioners can quickly assess energy performance impacts of design choices, facilitating proactive adjustments that can reduce costs and improve project feasibility.

In summary, the results confirm that the proposed data-driven framework effectively addresses the challenges of early-stage energy forecasting by providing accurate predic-

tions with minimal input data. This contributes to bridging the gap between advanced computational methods and practical design workflows, promoting the integration of energy efficiency considerations from the outset of neighborhood-scale projects.

8. Conclusions

This research developed and tested a novel framework for integrating machine learning tools into early-stage neighborhood-scale energy forecasting within existing design workflows. Utilizing the CatBoost Regressor, the framework achieved accurate energy consumption predictions with minimal input data, specifically targeting static predictions of EUI.

The key findings of the study are as follows:

- *High Predictive Accuracy with Minimal Inputs:* The CatBoost model achieved R^2 scores of 0.88, demonstrating that accurate energy consumption predictions are possible using only a few key parameters available during early design stages.
- *Identification of Influential Parameters:* SHAP analysis confirmed building area, primary building activity, number of floors, and climate zone as the most influential factors affecting energy consumption. Focusing on these parameters allows for streamlined data collection without sacrificing accuracy.
- *Testing Against Traditional Models:* The ML model's predictions closely aligned with those from traditional energy modeling tools, with acceptable discrepancies, confirming its reliability and practical applicability.

The outcome of this research has significant implications for neighborhood-scale and sustainable design. By enabling early-stage predictions of energy consumption, the framework empowers practitioners to incorporate energy efficiency considerations into the design process from the beginning. This proactive approach can lead to more sustainable urban environments, reduced energy costs, and alignment with global sustainability goals.

To enhance the applicability and robustness of the framework, future research should focus on the following:

- *Expansion to Subsequent Phases:* Developing and testing the remaining phases (Phase 2 and Phase 3) of the framework to evaluate its effectiveness across different design stages and with more detailed data inputs.
- *Enhancing Model Generalizability:* Incorporating more diverse datasets, especially for large-scale buildings, to improve model performance and applicability across various building typologies.
- *Integration of Additional Sustainability Metrics:* Expanding the framework to include other sustainability indicators such as embodied carbon, water usage, and indoor environmental quality for a more holistic assessment.
- *Application in Different Contexts:* Conducting case studies in various geographic locations and regulatory environments to validate the framework's adaptability and effectiveness.

In conclusion, this study lays the groundwork for transforming urban energy modeling practices by integrating machine learning into early design workflows. The proposed framework offers a practical, efficient, and accurate tool for predicting energy consumption at the neighborhood scale, contributing to the development of energy-efficient and sustainable cities. By focusing on minimal yet impactful inputs, the framework aligns with the needs of practitioners during the early design stages, enabling data-driven decision-making that supports global sustainability objectives.

Author Contributions: A.G.d.S., Conceptualization, Software, Data curation, Writing—Original draft preparation. M.R., Methodology, Supervision, Writing—Reviewing and Editing. G.M., Validation, Supervision, Writing—Reviewing and Editing. S.H., Conceptualization, Supervision, Writing—Reviewing and Editing, Validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to thank our master thesis students in Politecnico di Milano, Alessandro Aliprandi and Riccardo Viganò, for their contributions to this study, particularly in the development of the test case study massing model, enhancing the quality and robustness of our research.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BAU	Business-as-usual
BPS	Building Performance Simulations
CBECS	Commercial Buildings Energy Consumption Survey
CDD65	Cooling Degree Days (base 65)
EIA	Energy Information Administration
EUI	Energy Use Intensity
HDD65	Heating Degree Days (base 65)
HVAC	Heating Ventilation Air Conditioning
ML	Machine Learning
NFLOOR	Number of Floors
PBA	Principal Building Activity
PUBCLIM	ASHRAE Climate Zone
RECS	Residential Energy Consumption Survey
SHAP	Shapely Additive Explanations
SQM	Total Building Area
TOTALKWH	Total Energy Consumption

References

1. Ang, Y.Q. Using Urban Building Energy Modeling to Develop Carbon Reduction Pathways for Cities. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2022.
2. Intergovernmental Panel on Climate Change Climate Change 2014: Synthesis Report. In *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Geneva, Switzerland, 2015; p. 151.
3. United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*; United Nations: New York, NY, USA, 2019.
4. United Nations Environment Programme Building Sector Emissions Hit Record High, but Low-Carbon Pandemic Recovery Can Help Transform Sector—UN Report. Available online: http://www.unep.org/news-and-stories/press-release/building-sector-emissions-hit-record-high-low-carbon-pandemic (accessed on 1 August 2023).
5. Haase, M.; Baer, D. Implementation of Positive Energy District Concepts and Energy Master Plans for Decarbonization of Districts. *APP* **2022**, *38*, 546–552. [CrossRef]
6. Kapoor, G.; Gupta, J.K. Empowering Cities for Net Zero Carbon: Issues and Options. In Proceedings of the 2nd International Conference on Trends in Architecture and Construction; Varma, A., Chand Sharma, V., Tarsi, E., Eds.; Lecture Notes in Civil Engineering. Springer Nature Singapore: Singapore, 2024; Volume 527, pp. 809–820, ISBN 978-981-9749-87-4.
7. Cai, S.; Gou, Z. Transitioning Positive Energy Buildings towards Positive Energy Communities: Leveraging Performance Indicators for Site Planning Assessments. *Energy Build.* **2024**, *325*, 114976. [CrossRef]
8. Aruta, G.; Ascione, F.; Bianco, N.; Bindi, L.; Iovane, T. Energy Classification of Urban Districts to Map Buildings and Prioritize Energy Retrofit Interventions: A Novel Fast Tool. *Appl. Energy* **2025**, *377*, 124664. [CrossRef]
9. Vikström, L.; Ek, K.; Luciani, A.; Rizzo, A. Co-Designing the Urban Energy Transition: A Resident-Based Approach. *Cities* **2025**, *156*, 105506. [CrossRef]
10. Liu, X.; Yang, D.; Donkers, A.; De Vries, B. Building Sustainable Urban Energy Systems: The Role of Linked Data in Photovoltaic Generation Estimation at Neighbourhood Level. *Appl. Energy* **2025**, *378*, 124749. [CrossRef]
11. Poornima, P.U.; Dhineshkumar, K.; Kiran Kumar, C.; Sumana, S.; Rama Sundari, M.V.; Sivaraman, P.; Shuaib, M.; Rajaram, A. Optimising Rooftop Photovoltaic Adoption in Urban Landscapes: A System Dynamics Approach for Sustainable Energy Transitions. *Biomed. Signal Process. Control.* **2025**, *100*, 107071. [CrossRef]
12. Hong, T.; Langevin, J.; Sun, K. Building Simulation: Ten Challenges. *Build. Simul.* **2018**, *11*, 871–898. [CrossRef]

13. Østergård, T.; Jensen, R.L.; Maagaard, S.E. A Comparison of Six Metamodeling Techniques Applied to Building Performance Simulations. *Appl. Energy* **2018**, *211*, 89–103. [CrossRef]
14. European Commission. *The European Green Deal*; European Commission: Brussels, Belgium, 2019.
15. European Commision. *European Parliament and the Council of European Union Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the Establishment of a Framework to Facilitate Sustainable Investment, and Amending Regulation*; European Commision: Brussels, Belgium, 2020.
16. European Commision. *European Commission Directive (EU) 2024/1275 of the European Parliament and of the Council of 24 April 2024 on the Energy Performance of Buildings (Recast)*; European Commission: Brusells, Belgium, 2024.
17. Li, F.; Li, L.; You, F. Enhancing Energy Efficiency in Cooling Systems through Advanced Machine Learning and Meta-Heuristic Algorithms for Precise Cooling Load Prediction. *J. Appl. Sci. Eng.* **2024**, *28*, 1275–1286. [CrossRef]
18. Hsu, P.-C.; Gao, L.; Hwang, Y. Comparative Study of LSTM and ANN Models for Power Consumption Prediction of Variable Refrigerant Flow (VRF) Systems in Buildings. *Int. J. Refrig.* **2025**, *169*, 55–68. [CrossRef]
19. Es-Sakali, N.; Zoubir, Z.; Idrissi Kaitouni, S.; Mghazli, M.O.; Cherkaoui, M.; Pfafferott, J. Advanced Predictive Maintenance and Fault Diagnosis Strategy for Enhanced HVAC Efficiency in Buildings. *Appl. Therm. Eng.* **2024**, *254*, 123910. [CrossRef]
20. Tharushi Imalka, S.; Yang, R.J.; Zhao, Y. Machine Learning Driven Building Integrated Photovoltaic (BIPV) Envelope Design Optimization. *Energy Build.* **2024**, *324*, 114882. [CrossRef]
21. Joshi, D.S.V.; Patil, R.V.; Tarambale, D.M.; Patil, B.K.; Gandhi, Y. Stochastic Processes in the Analysis of Electrical Load Forecasting. *Adv. Nonlinear Var. Inequalities* **2025**, *28*.
22. Sun, X. Forecasting Residential Building Heating Load With an Innovative Gaussian Process Regression Method. *J. Appl. Sci. Eng.* **2024**, *28*, 1219–1231.
23. Zhang, X.; Pei, L. Cooling Load Forecasting Based on Hybrid Machine-Learning Application with Integration of Meta-Heuristic Algorithm. *J. Appl. Sci. Eng.* **2024**, *28*, 601–614. [CrossRef]
24. Shen, Y. Load Estimation Models for the Heat Demand of Buildings: Application of Optimized Gaussian Process Regression. *J. Appl. Sci. Eng.* **2024**, *28*, 527–541. [CrossRef]
25. Guan, Y. Enhancing Residential Building Heating Load Prediction with Hybrid MLP Models. *J. Appl. Sci. Eng.* **2024**, *28*, 979–994. [CrossRef]
26. Lv, D.; Liang, C.; Lu, X. Improving Heating Load Prediction With LSSVR: Comparative Analysis Of Optimized Models. *J. Appl. Sci. Eng.* **2024**, *28*, 1131–1145. [CrossRef]
27. Wei, Y.; Zhang, X.; Shi, Y.; Xia, L.; Pan, S.; Wu, J.; Han, M.; Zhao, X. A Review of Data-Driven Approaches for Prediction and Classification of Building Energy Consumption. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1027–1047. [CrossRef]
28. Sun, Y.; Haghighat, F.; Fung, B.C.M. A Review of The-State-of-the-Art in Data-Driven Approaches for Building Energy Prediction. *Energy Build.* **2020**, *221*, 23. [CrossRef]
29. Ali, U.; Shamsi, M.H.; Hoare, C.; Mangina, E.; O'Donnell, J. Review of Urban Building Energy Modeling (UBEM) Approaches, Methods and Tools Using Qualitative and Quantitative Analysis. *Energy Build.* **2021**, *246*, 111073. [CrossRef]
30. Javeed Nizami, S.; Al-Garni, A.Z. Forecasting Electric Energy Consumption Using Neural Networks. *Energy Policy* **1995**, *23*, 1097–1104. [CrossRef]
31. González, P.A.; Zamarreño, J.M. Prediction of Hourly Energy Consumption in Buildings Based on a Feedback Artificial Neural Network. *Energy Build.* **2005**, *37*, 595–601. [CrossRef]
32. Dong, B.; Cao, C.; Lee, S.E. Applying Support Vector Machines to Predict Building Energy Consumption in Tropical Region. *Energy Build.* **2005**, *37*, 545–553. [CrossRef]
33. Li, Q.; Meng, Q.; Cai, J.; Yoshino, H.; Mochida, A. Applying Support Vector Machine to Predict Hourly Cooling Load in the Building. *Appl. Energy* **2009**, *86*, 2249–2256. [CrossRef]
34. Touzani, S.; Granderson, J.; Fernandes, S. Gradient Boosting Machine for Modeling the Energy Consumption of Commercial Buildings. *Energy Build.* **2018**, *158*, 1533–1543. [CrossRef]
35. Wu, J.; Nguyen, S.; Alahakoon, D.; Silva, D.D.; Mills, N.; Rathnayaka, P.; Moraliyage, H.; Jennings, A. A Computational Study to Evaluate and Compare Machine Learning-Based Energy Baseline Models Across Multiple Building Types. *Energies* **2024**, *17*, 1285.
36. Cordeiro-Costas, M.; Villanueva, D.; Eguía-Oller, P.; Martínez-Comesaña, M.; Ramos, S. Load Forecasting with Machine Learning and Deep Learning Methods. *Appl. Sci.* **2023**, *13*, 7933. [CrossRef]
37. Runge, J.; Zmeureanu, R. A Review of Deep Learning Techniques for Forecasting Energy Use in Buildings. *Energies* **2021**, *14*, 608. [CrossRef]
38. Farzaneh, H.; Malehmirchegini, L.; Bejan, A.; Afolabi, T.; Mulumba, A.; Daka, P.P. Artificial Intelligence Evolution in Smart Buildings for Energy Efficiency. *Appl. Sci.* **2021**, *11*, 763. [CrossRef]
39. Romani, Z.; Draoui, A.; Allard, F. Metamodeling the Heating and Cooling Energy Needs and Simultaneous Building Envelope Optimization for Low Energy Building Design in Morocco. *Energy Build.* **2015**, *102*, 139–148. [CrossRef]
40. Cheng, M.-Y.; Cao, M.-T. Accurately Predicting Building Energy Performance Using Evolutionary Multivariate Adaptive Regression Splines. *Appl. Soft Comput.* **2014**, *22*, 178–188. [CrossRef]
41. Rackes, A.; Melo, A.P.; Lamberts, R. Naturally Comfortable and Sustainable: Informed Design Guidance and Performance Labeling for Passive Commercial Buildings in Hot Climates. *Appl. Energy* **2016**, *174*, 256–274. [CrossRef]

42. Yuan, J.; Nian, V.; Su, B.; Meng, Q. A Simultaneous Calibration and Parameter Ranking Method for Building Energy Models. *Appl. Energy* **2017**, *206*, 657–666. [CrossRef]
43. Nutkiewicz, A.; Jain, R.K. Exploring the Integration of Simulation and Deep Learning Models for Urban Building Energy Modeling and Retrofit Analysis. In Proceedings of the Building Simulation 2019: 16th Conference of IBPSA, Rome, Italy, 2–4 September 2019; pp. 3209–3216.
44. Neumann, H.-M.; Hainoun, A.; Stollnberger, R.; Etminan, G.; Schaffler, V. Analysis and Evaluation of the Feasibility of Positive Energy Districts in Selected Urban Typologies in Vienna Using a Bottom-Up District Energy Modelling Approach. *Energies* **2021**, *14*, 4449. [CrossRef]
45. Dai, M.; Ward, W.O.C.; Arbabi, H.; Densley Tingley, D.; Mayfield, M. Scalable Residential Building Geometry Characterisation Using Vehicle-Mounted Camera System. *Energies* **2022**, *15*, 6090. [CrossRef]
46. Hey, J.; Siebers, P.-O.; Nathanail, P.; Ozcan, E.; Robinson, D. Surrogate Optimization of Energy Retrofits in Domestic Building Stocks Using Household Carbon Valuations. *J. Build. Perform. Simul.* **2023**, *16*, 16–37. [CrossRef]
47. Robinson, C.; Dilkina, B.; Hubbs, J.; Zhang, W.; Guhathakurta, S.; Brown, M.A.; Pendyala, R.M. Machine Learning Approaches for Estimating Commercial Building Energy Consumption. *Appl. Energy* **2017**, *208*, 889–904. [CrossRef]
48. Nutkiewicz, A.; Yang, Z.; Jain, R.K. Data-Driven Urban Energy Simulation (DUE-S): A Framework for Integrating Engineering Simulation and Machine Learning Methods in a Multi-Scale Urban Energy Modeling Workflow. *Appl. Energy* **2018**, *225*, 1176–1189. [CrossRef]
49. Zhang, W.; Robinson, C.; Guhathakurta, S.; Garikapati, V.M.; Dilkina, B.; Brown, M.A.; Pendyala, R.M. Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach. *Energy* **2018**, *155*, 162–173. [CrossRef]
50. Ahmad, M.; Culp, C.H. Uncalibrated Building Energy Simulation Modeling Results. *HVACR Res.* **2006**, *12*, 1141–1155. [CrossRef]
51. Yoon, Y.; Jung, S.; Im, P.; Salonvaara, M.; Bhandari, M.; Kunwar, N. Empirical Validation of Building Energy Simulation Model Input Parameter for Multizone Commercial Building during the Cooling Season. *Renew. Sustain. Energy Rev.* **2023**, *188*, 113889. [CrossRef]
52. Ryan, E.M.; Sanquist, T.F. Validation of Building Energy Modeling Tools under Idealized and Realistic Conditions. *Energy Build.* **2012**, *47*, 375–382. [CrossRef]
53. Chidiac, S.E.; Catania, E.J.C.; Morofsky, E.; Foo, S. Effectiveness of Single and Multiple Energy Retrofit Measures on the Energy Consumption of Office Buildings. *Energy* **2011**, *36*, 5037–5052. [CrossRef]
54. Eisenhower, B.; O'Neill, Z.; Fonoberov, V.A.; Mezić, I. Uncertainty and Sensitivity Decomposition of Building Energy Models. *J. Build. Perform. Simul.* **2012**, *5*, 171–184. [CrossRef]
55. O'Neill, Z.; Shashanka, M.; Pang, X.; Bhattacharya, P.; Bailey, T.; Haves, P. Real Time Model-Based Energy Diagnostics in Buildings. In Proceedings of the Building Simulation 2011, Sydney, Australia, 14–16 November 2011.
56. Commercial Buildings | ENERGY STAR. Available online: https://www.energystar.gov/buildings (accessed on 17 April 2024).
57. Mathew, P.A.; Dunn, L.N.; Sohn, M.D.; Mercado, A.; Custudio, C.; Walter, T. Big-Data for Building Energy Performance: Lessons from Assembling a Very Large National Database of Building Energy Use. *Appl. Energy* **2015**, *140*, 85–93. [CrossRef]
58. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
59. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *arXiv* **2017**, arXiv:1706.09516. [CrossRef]
60. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support 2018. *arXiv* **2018**, arXiv:1810.11363.
61. Sadeghipour Roudsari, M.; Pak, M.; Viola, A. Ladybug: A Parametric Environmental Plugin for Grasshopper to Help Designers Create An Environmentally-Conscious Design. In Proceedings of the Building Simulation 2013: 13th Conference of IBPSA, Chambery, France, 25–28 August 2013.
62. Motamedi, S.; Liedl, P. Integrative Algorithm to Optimize Skylights Considering Fully Impacts of Daylight on Energy. *Energy Build.* **2017**, *138*, 655–665. [CrossRef]
63. Toutou, A.; Fikry, M.; Mohamed, W. The Parametric Based Optimization Framework Daylighting and Energy Performance in Residential Buildings in Hot Arid Zone. *Alex. Eng. J.* **2018**, *57*, 3595–3608. [CrossRef]
64. Norouzi, M.; Yeganeh, M.; Yusaf, T. Landscape Framework for the Exploitation of Renewable Energy Resources and Potentials in Urban Scale (Case Study: Iran). *Renew. Energy* **2021**, *163*, 300–319. [CrossRef]
65. Amasyali, K.; El-Gohary, N.M. A Review of Data-Driven Building Energy Consumption Prediction Studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [CrossRef]
66. Di Stefano, A.G.; Ruta, M.; Masera, G.; Hoque, S. Integrated Workflow Development for Data-Driven Neighborhood-Scale Building Performance Simulation. *ASME J. Eng. Sustain. Build. Cities* **2024**, *6*, 014501. [CrossRef]