Developments in the Built Environment 17 (2024) 100296

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

Using causal inference to avoid fallouts in data-driven parametric analysis: A case study in the architecture, engineering, and construction industry

Xia Chen [a,b,*], Ruiji Sun [b], Ueli Saluz [a], Stefano Schiavon [b], Philipp Geyer [a]

[a] *Sustainable Building Systems, Leibniz University Hannover, Germany*
[b] *Center for the Built Environment, University of California, Berkeley, United States*

A R T I C L E I N F O

Keywords:
Data-driven models
Causal inference
First-principles simulations
Feature selection
Architecture, engineering, and construction industry
Cognitive biases
Biased outcomes
Machine learning methods
Building energy performance
Domain knowledge

A B S T R A C T

The decision-making process in real-world implementations has been affected by a growing reliance on data-driven models. Recognizing the limitations of isolated methodologies - namely, the lack of domain understanding in data-driven models, the subjective nature of empirical knowledge, and the idealized assumptions in first-principles simulations, we explore their synergetic integration. We showed the potential risk of biased results when using data-driven models without causal analysis. Through a case study on energy consumption in building design, we demonstrate how causal analysis significantly enhances the modeling process, mitigating biases and spurious correlations. We concluded that: (a) Sole data-driven models' accuracy assessment or domain knowledge screening may not rule out biased and spurious results; (b) Data-driven models' feature selection should involve careful consideration of causal relationships, especially colliders; (c) Integrating causal analysis results aid to first-principles simulation design and parameter checking to avoid cognitive biases. We advocate for the routine integration of causal inference within data-driven models in engineering practices, emphasizing its critical role in ensuring the models' reliability and real-world applicability.

1. Introduction

In recent decades, successful implementations of machine learning (ML) methods, with the momentum of growing data volume, have brought the data-driven approach into various engineering domains. Together with empirical domain knowledge analysis and first-principles simulations, ML methods have become a handy tool for both academic research and industrial application (LeCun et al., 2015; Raschka et al., 2020; Bertolini et al., 2021). Due to their end-to-end learning behavior, good generalization performance, and fast prediction response, they are favored by researchers and engineers, and are gradually being integrated as a decision-making or analysis assistance tool in the architecture, engineering, and construction (AEC) industry (Seyedzadeh et al., 2018; Marcher et al., 2020; Dimiduk et al., 2018).

The wide adaptability of ML stems from its ability to uncover hidden patterns in data by minimizing error during training, rather than relying on explicit physical process modeling with domain knowledge. However, a critical prerequisite for ML's effective performance is the assumption that all input variables are independent, or even *independent and identically distributed* (i.i.d.) (Schölkopf, 2019) by default. This assumption implies that the probability distribution of each variable should be independent of others. Yet, in engineering domains, a case usually requires considering different factors in an interdisciplinary manner. Consider the following important cases:

- Independence Challenge: A model predicting building energy consumption may incorrectly isolate temperature as a key factor, missing the interdependent increase in occupancy and energy use during colder days. This oversight can lead to misattributed energy consumption drivers.
- Identical Distribution Challenge: A model developed for predicting mechanical failures in a plant, trained mostly on winter data, might erroneously link failure rates to lower temperatures, overlooking the actual cause - increased winter production.

Furthermore, engineers and researchers often gather extensive features to improve model accuracy, based on the intuitive belief that more data leads to better predictions. However, this approach can overlook the critical intercorrelations of features in real-world scenarios. For instance, during the building design or construction phase, the objectives commonly involve building energy performance, environmental impact, cost, occupant's comfort, etc., simultaneously. The well-known

* Corresponding author.
E-mail address: xia.chen@iek.uni-hannover.de (X. Chen).

https://doi.org/10.1016/j.dibe.2023.100296
Received 23 September 2023; Received in revised form 12 November 2023; Accepted 3 December 2023
Available online 12 December 2023

mantra in statistics: "*Correlation does not imply causation*" (Pearl and Mackenzie, 2018; Aldrich, 1995), is not sufficiently considered in engineering scenarios (Hegde and Rokseth, 2020; Chakraborty and Elzarka, 2019) when employing ML methods. Unlike first-principles simulations, which encode causal relationships between variables in explicit physical equations, data-driven processes do not inherently include this information. This gap in process understanding might lead to false implementation and reliability issues for engineers and domain experts. This false implementation situation raises the risk of biased results and spurious conclusions because ML methods rely heavily on the information carried from the distribution of observed data and large predefined sets (Schölkopf et al., 2021).

In this study, we introduce a synergetic framework that integrates empirical domain knowledge, simulations, and data-driven methods, aiming to enhance general engineering analysis. This framework is not only effective in reducing prediction bias in specific engineering contexts, as demonstrated through a building engineering scenario, but also presents a generalized workflow adaptable across various tasks. **Central to our study is the advocacy for conducting causal analysis of input features as a standard practice in any data-driven modeling process for engineering tasks**, where causal dependency result guides us whether we should involve/control some of the input feature(s). This is crucial for two reasons: firstly, to prevent biased estimates and spurious conclusions, which are inherent limitations in data-driven methods regardless of model accuracy; and secondly, to foster a robust link between ML methods and human reasoning by cross-validating data with domain knowledge and examining potential cognitive biases in simulations.

To visualize this integration, Fig. 1 differentiates between two distinct pathways in the modeling process: the original, depicted by black arrows, and the causal integrated process, shown with additional red arrows. The black arrows represent the traditional steps in data-driven modeling, which may miss causal nuances. In contrast, the red arrows indicate an enhanced process where causal analysis informs each step, integrating user's domain knowledge and decision-making. Our approach underscores the importance of recognizing causal dependencies and constructing a causal skeleton, tools vital for knowledge discovery and decision-making in engineering analyses.

2. Framework and methodologies

2.1. Synergetic framework between knowledge, simulation, and data-driven methods

In engineering, the tools we use for modeling and decision-making can be classified into three main categories: *empirical domain knowledge*, *first-principles simulation*, and *data-driven models*:

- **Empirical domain knowledge** is a carrier of individual and past professional experience, providing a fundamental drive to understand, interact, and make decisions in a system. This includes heuristic rules or "rules of thumb" – quick, intuitive information set. However, it is limited by personal competence and often lacks reproducibility.
- **First-principles simulation** is a process based on abstract symbolic abstraction, using mathematical equations and physical/chemical laws to govern the behavior of a system. By starting from basic principles and building up to an understanding of complex phenomena, first-principles simulations are also referred to as "white-box models".
- **Data-driven method** is a computational process based on available data rather than theoretical principles or physical laws. These processes employ ML algorithms, statistical models, and data analysis techniques to extract patterns and relationships from datasets. These patterns are then used to make predictions or generate insights about the system, functioning as "black-box models".

Table 1 illustrates the main advantages and disadvantages of these three major categories we rely on in engineering.

In engineering scenarios, we possess, reuse, and iterate on invariant patterns that can be applied to many cases. These patterns form what is known as knowledge and experience (Chen et al., 2022a). For instance: the case of sinking library[1] updates our consideration of the relationship between building type/usage and building structural engineering. In first-principles simulations, the relationships between these variables are naturally embedded into symbolic formulas and numerical modeling processes as knowledge. However, this type of information input is absent in the data-driven process.

To bridge this gap, we propose to integrate a crucial, transferable element into data-driven methods: the causal dependencies among variables. Fig. 1 graphically conveys how these causal dependencies, once discovered from the data, synergize with empirical domain knowledge and first-principles simulations, as well as with data-driven approaches. Causal dependencies thus become a vital medium for communication between raw data and refined methodologies, offering a framework for users to cross-validate model outputs with their domain expertise. This interplay is inherently non-linear and dynamic, positioning the user at the center of the process. With an aggregate of outputs and information, the user navigates through adjustments and decisions, which may either feed back into the ongoing loop of refinement as depicted in the figure or conclude in a definitive output. This iterative cycle underscores the essence of our approach: a user-centric model that leverages the synergy of methodologies to achieve robust, informed decision-making in engineering.

In Fig. 1, red arrows indicate how causal relationships interact with other engineering modeling approaches. Causality is commonly confused with correlation, but the former presents a different interpretation from observational data: It analyzes the asymmetric change and response between cause and effect, aids in analyzing interventional scenarios, counterfactuals, and answers "what-if" questions. This reasoning ability is essential for informative and sequential decision-making support. Additionally, the extracted causality information provides a feedback loop for users to validate and update their domain knowledge, fostering unbiased modeling.

2.2. Causality

Causality research has become a critical topic and has made substantial contributions across various fields with the widespread adoption of data-driven methods in the past decade (Schölkopf, 2019; Spirtes, 2010). Causal inference examines parameters or properties, considering cause-effect logical sequences to avoid unrealistic conclusions. For a systematic discussion of causal inference research, we refer to research to the works of Pearl (2009), Spirtes et al. (Spirtes, 2010; Spirtes et al., 2000), and Peters et al. (2017).

Our previous research (Chen et al., 2022a) introduced causal inference into the energy-efficient building design process, using a four-step framework that combined causal structure finding and causal effect estimation. In this study, we aim to demonstrate the importance of checking causal dependencies in the context of the general AEC domain. This section briefly clarifies foundational ideas related to causal analysis.

Causal finding algorithms are methods for identifying and returning equivalence classes of proper causal structure based on observational data in an unsupervised, data-driven manner. Essentially, they distinguish asymmetries in sampling distributions to identify feature dependencies and causal directions. Typical causal structure finding algorithms based on observational data fall into three categories: constraint-based, score-based, and hybrid (Kalisch and Bühlmann,

[1] The sinking library: A famed college library is gradually sinking into the ground because its architect failed to take the weight of the books into account.

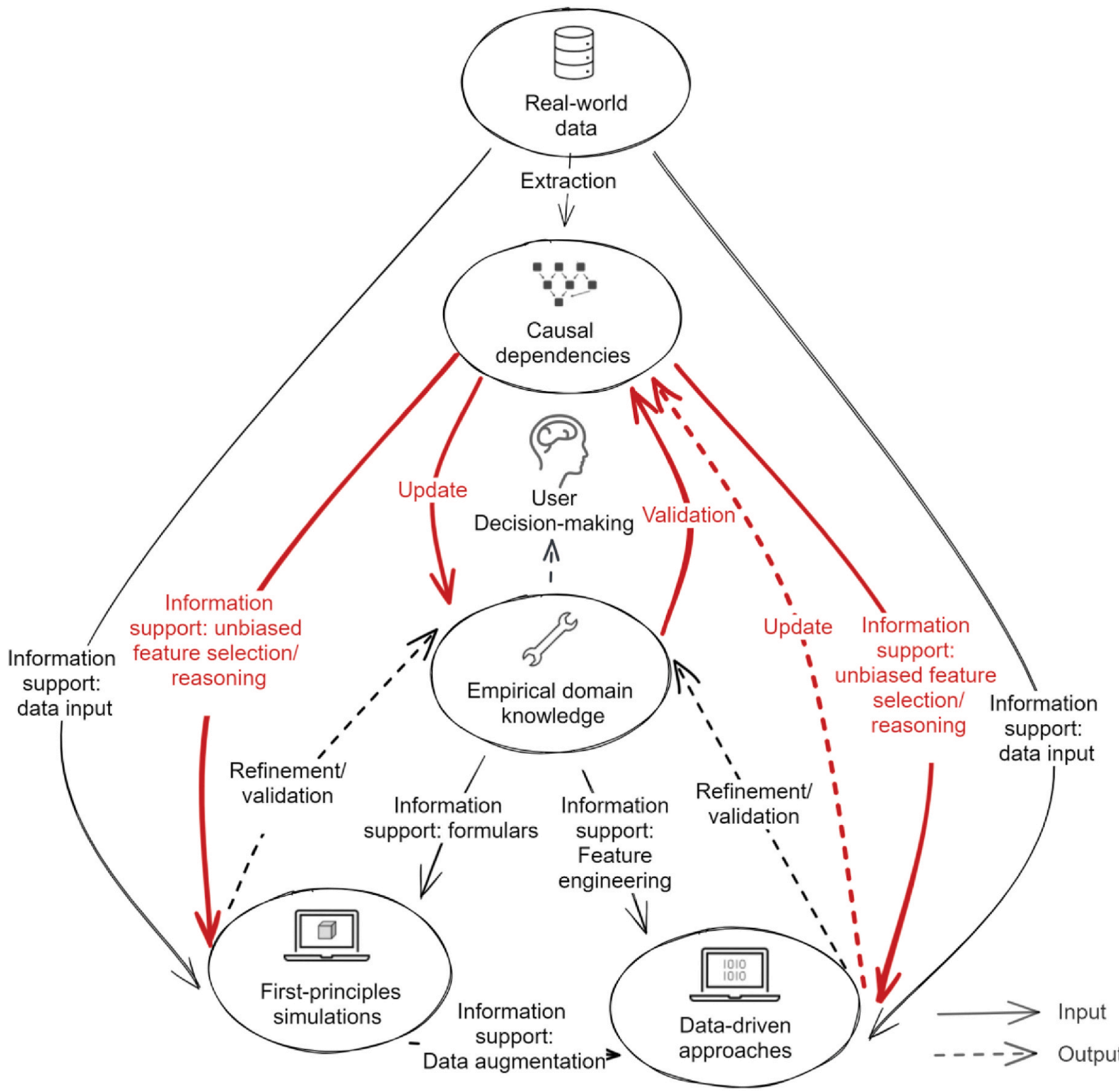

Fig. 1. Illustration of the potentially synergetic nature of the three main engineering modeling processes. Causal dependencies extracted from data represent a type of invariant, transferable knowledge, which play a vital role in offering a feedback loop and interacting with the user's empirical domain knowledge. Except for the data and domain knowledge input, the causal dependencies information support contributes to validation for first-principles simulation, and unbiased estimation/ reasoning for data-driven methods. The red arrows indicate how the causal relationships interact with other engineering modeling approaches. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
The characteristics of relying on empirical domain knowledge, first-principles simulation, and data-driven approach for engineering modeling.

	Advantages	Disadvantages
Empirical domain knowledge	• No extra efforts needed for modeling; • Foundation for scientific inquiry and hypothesis testing;	• Rule of thumb, heavily relies on personal ability; • Limited extent and reliability in non-standard cases;
First-principles simulation	• Good interpretability; • Flexible in modeling details; • Large amount of output variable;	• Time-consuming in detailed simulation; • Modeling efforts required in each new scenario;
Data-driven method	• Fast response in prediction; • Universal approximator; • End-to-end learning behavior;	• Black-box, trustfulness issues; • Data-hungry for training;

2014). In this study, we chose one of the typical score-based methods with a greedy mechanism (DeVore and Temlyakov, 1996), Greedy-Equivalent-Search (GES) (Chickering, 2002a, 2002b).

Directed Acyclic Graphs (DAGs) are graph diagrams composed of variables (nodes) connected via unidirectional arrows (paths) to depict hypothesized causal relationships (Judea, 2010). A causal skeleton DAG with a fixed structure embeds the causal dependencies of given data. To better understand the concept in a domain context, a DAG demonstration with simple cases in the building engineering domain is presented in Fig. 2. Major terms and types in DAG structure combinations are:

- **Directed path** denotes a directed edge $x \rightarrow y$ of x (cause) on y (effect). Intuitively, it means that y is directly influenced by the status of x, altering x by external intervention would also alter y.
- **Confounding structure** (Fig. 2, left) occurs when two variables are linked through a common cause (Confounder) that is not accounted for, potentially introducing bias.

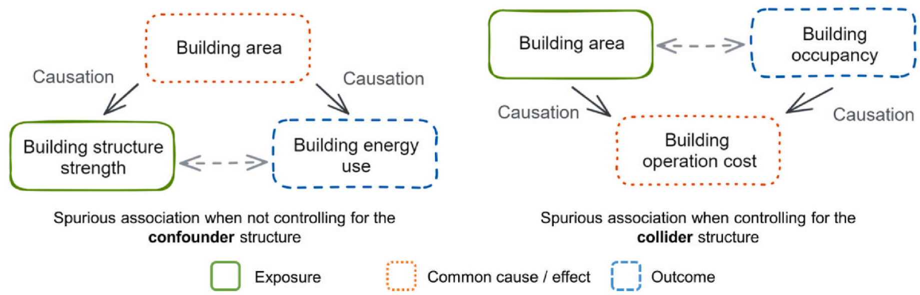

Fig. 2. Causal confounder and collider examples in the context of architectural engineering domain. Failing to identify the causal relationship causes spurious association (backdoor path) and biased results. Left: confounder bias when the common cause is not controlled (Not controlling the 'Building area' would make the data-driven model consider adjusting Building structure strength to change Building energy use); Right: collider bias when the common effect is controlled (Controlling the 'Building operation cost' would make the data-driven model to adjust Building area for making changes in Building occupancy).

- **Collider structure** (Fig. 2, right) exists when two variables affect a common outcome. Controlling for this outcome induces a spurious association between the variables, it is also known as backdoor path.
- **Backdoor path** exists in two variables in a confounding structure where the common cause is not controlled, or two variables in a collider structure where the common effect is controlled, effect variables connected by this backdoor path have a non-causal association and would lead to potential bias with distorting association.
- **Closed path** exists in collider structures where two variables have the same outcome. Unlike directed and backdoor paths, this path is causal-wise irrelevant: there is no causal path between the two cause variables via the collider structure, unless the common outcome is controlled.

DAG rules are principled structural guidelines that enable users to investigate cases for identifying appropriate sets of covariates in complex DAGs and for removing structural bias through adjustments, e.g., d-Separation, backdoor criterion (Pearl, 2000), and their extensions.

DAGs, often defined by prior knowledge, could be incomplete (Guo et al., 2020). In the development of a causal diagram, users utilize their best available prior knowledge to set up the most plausible causal diagram. Subsequently, they adhere to strict DAG rules to identify the causal dependencies between given exposure inputs and the target outcome from the case. In the remaining content, all DAGs are generated and modified by DAGitty (Textor et al., 2016; Textor, 2015).

A "fallout" situation in this context refers to an instance in causal analysis where an indirect, biased relationship exists between an exposure (or cause) and an outcome (or effect), primarily due to the presence of the backdoor path or the opening of the closing path.

2.3. Machine learning

In this study, we focus on ML methods applied to supervised learning tasks, which typically involve addressing a classification or regression problem with labeled data. To ensure that the fallout is irrelevant to the type of data-driven methods used, we examined three mainstream ML methodologies (Singh et al., 2016): tree-based models (Clark and Pregibon, 2017), kernel machines (Hofmann et al., 2008), and neural networks (LeCun et al., 2015), which are mechanistically different and widely applied in engineering domains (Seyedzadeh et al., 2018; Hegde and Rokseth, 2020; Chakraborty and Elzarka, 2019). Brief introductions to their mechanisms are given in Appendix. Beyond these methods, the evaluation of uncertainties is critical for supporting the decision-making process (Chicco et al., 2021; Balestriero et al., 2021), leading us to include a probabilistic, tree-based, gradient-boosting surrogate model – NGBoost (Chen et al., 2022c) in our case study. Instead of generating output as a point prediction, the design of NGBoost incorporates a predictive uncertainties quantification process, offering insights into the output range within the set of feature input descriptions in a data-driven manner.

3. Case study

3.1. Scenario setup

We studied the effect of different designs on energy use for heating (Energy Usage Intensity of heating, *EUI Heating*) by varying I*nsulation Standards* and H*eating Systems*.

To prepare our dataset, we utilized a parametric office building simulation model. This model represents a realistic design space by incorporating a wide range of configurations for building components and zones to train our ML models (training data). The causal reasoning within space is validated by a real-world design project from our previous research (Chen et al., 2022b) (test case): a mixed-usage, four-floor building known as Building.Lab, located on a tech campus in Regensburg, Germany.

We simulated three sets of thermal characteristics to explore design variations in insulation values. These were based on existing standards: the 2020 German Energy Act for Buildings (*GEG*), Net Zero Energy Building (*NZEB*), and *Passive House*. These standards, from baseline to high, have different requirements for components' thermal conductivity (U-values), with a higher standard indicating better building thermal behavior and less energy loss. We also configured three typical building heating systems: *boiler*, air-source heat pump (*ASHP*), and district heating (*DH*). For the modeling tool, we used Grasshopper (Patil et al., 1985), with Honeybee (Rakitta and Wernery, 2021) serving as a high-level simulation interface for EnergyPlus.

In terms of data-driven modeling approaches, as discussed in Section 2.3, we applied Decision Tree (DT), Support Vector Machine for Regression (SVR), Artificial Neural Network (ANN, with Multi-Layer Perception chosen as a basic variation), and NGBoost across all scenarios.

We applied three metrics to facilitate performance comparison across different numerical scales of results: Normalized Root Mean Square Error (NRMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Coefficient of determination (R-squared or R^2). We chose R^2 as our primary reference. The reasoning behind this choice and detailed interpretations of these three metrics are available in (Chicco et al., 2021).

Table 2 lists the input features from the simulation, their ranges, and the corresponding test case setting. To avoid the extrapolation problem (which arises when the test case sample falls outside of the given training dataset's convex hull (Balestriero et al., 2021)), all feature values in the test case are within the range of training data. We fitted and fine-tuned ML models with the training data to achieve well-generalization performance, and used them later to predict different scenarios in the test case, in which all values are extracted from the Building.Lab project in a real-world context.

Table 2
Ranges in training data features and value extracted from the test case. All values in the test case are extracted from the Building.Lab project for the case study.

Building feature/Variable	Training data range	Test case setting
Orientation [°]	[0, 180]	12.5
Number of Floors	[1, 10]	4
Floor Height [m]	[2.8, 4.5]	3.48
Open Office: Heating Setpoint [°C]	[21, 24]	22
Open Office: Air Change Rate (ACH) [1/h]	[4, 6]	4
Open Office: People Per Area (PPA) [people/ m^2]	[0.05, 0.2]	0.15
Volume [m^3]	[4400, 146,000]	6807
Area[a] [m^2]	[1300, 36,000]	1956
Construction Area[b] [%]	[3, 11.5]	6
Window to Wall Ratio North [%]	[0, 0.7]	0.5
Window to Wall Ratio East [%]	[0, 0.7]	0.45
Window to Wall Ratio South [%]	[0, 0.7]	0.34
Window to Wall Ratio West [%]	[0, 0.7]	0.23
Insultation Standard	base, medium, high	Unknown
Heating System	Boiler, ASHP[c], DH[d]	Unknown
Energy Usage Intensity (EUI) Heating [kWh/ m^2a]	[14.6, 327.1]	Unknown

[a] Floor area gross.
[b] Areas covered by walls, columns, or any structural elements.
[c] ASHP: air-source heat pump.
[d] DH: district heating.

Further information regarding modeling configuration, data generation process, and training strategy of data-driven models are available in Appendix.

With the set training data and test case, we first set up two scenarios:

- ***Scenario I***: Full-scale modeling with all input features for EUI heating prediction as the benchmark.
- ***Scenario II***: Masked input features, which represent common situations in real-world engineering scenarios - feature selection by domain knowledge, or only some features are observable/available during data collection.

Scenario I presents an ideal case in research or engineering, demonstrating how the data-driven process helps to provide analytical insights into potential design scenarios. However, in real-world cases, data is rarely as complete as in an ideal scenario due to the presence of unobserved factors, the need for simplification because of the expensive data collection and computation efforts, or subjective manual filtering by end-users using their own domain knowledge or analytical tools. In Scenario II, we illustrate the potential risks of introducing subjective bias associated with such incomplete data: We selected the following input features that are typically cared for by architects or engineers in the building design phase for energy performance evaluation (Marcher et al., 2020; Chen et al., 2022c; Roman et al., 2020): *Open Office: Heating Setpoint, Open Office: ACH, Open Office: PPA, Volume, Area,* and *Window to Wall Ratios*.

In both scenarios, ML models are fitted and evaluated using the training data, then used to predict the output with test case inputs plus different insulation standard and heating system combinations.

3.2. Benchmark and fallout

Table 3 presents the prediction results of different models fitted with the training data in the setting of both scenarios. The results demonstrate the model capabilities in this training case; all ML methods trained by full input features show acceptable performance. The R^2 of all models is above 0.85, while ANN and NGBoost reach an accuracy above 0.95. With the masked feature setting but the same training process as in Scenario I, the result shows **only a minor performance difference between Scenario I & II by monitoring their MLs' accuracy.** We even observed a slight performance improvement for SVR in Scenario II. NRMSE and SMAPE results also align with this interpretation (see Appendix).

Table 3
5-fold cross-validation performance result comparison of different models: Scenario I & II.

	R^2 (Scenario I)	R^2 (Scenario II)
Decision Tree	0.86	0.81
SVR	0.87	0.87
ANN	0.96	0.94
NGBoost	0.95	0.88

Next, the test case is fed with variations for insulation standard and energy system into trained models for both scenarios. We illustrate the corresponding results from different variation combinations in Fig. 3.

Based on the result of Scenario I (Fig. 3a, right), we concluded the following insights:

1. The test case prediction results from ANN and NGBoost are more similar; they also achieve better accuracy in the training process evaluation.
2. The choice of the energy system is the factor that affects the EUI heating the most, with the air-source heat pump (ASHP) system requiring the least energy consumption, and the boiler system the most.
3. Regardless of heating system variation, higher building component thermal standards contribute to reducing total energy consumption, as expected.

With almost the same accuracy performance, the test case prediction result in Scenario II displays unusual patterns that contradict domain intuition, as shown in Fig. 3b. Although the choice of the heating system still shows the deterministic impact on EUI heating, the trend acts oppositely in insulation standard variation: The difference between the building insulation standards is either barely noticeable or even presents an inversed trend. Within the same heating system choice, a higher insulation standard results in more energy consumption in heating. This opposing trend even shows in the ANN, which achieves 0.94 in R^2 during performance evaluation. Furthermore, we observed a drastic increase in the uncertainty range in the output of NGBoost compared to Scenario I (see orange scatter distributions in Fig. 3).

Based on the result from Scenario II, **wrong conclusions** could easily be drawn, potentially misguiding decision-making process in real-world projects or research, e.g.:

> *"In this case, insulation standard choices are unimportant, or adapting a lower insulation standard could help to reduce the energy usage of the building."*

This conclusion drawn from Scenario II clearly conflicts with the result from Scenario I and with common knowledge. We refer to Scenario II as a case of biased estimation or fallout. This fallout is directly linked to potential economic and energy loss, as well as risks if implemented in real-world engineering construction scenarios. Given that the cost of implementing higher insulation standards in buildings is typically an important factor, this misleading conclusion could lead to the decision of investment reduction or underestimation.

Such uncertain performance in the analysis could cause severe trust issues when adopting data-driven methods in engineering scenarios and decision-making processes. This is because real-world scenarios are less likely to provide complete data without hidden variables. It is less relevant to the modeling approach and cannot be ruled out by performance evaluation. As the only difference between the two scenarios is the feature selection, a closer examination of the input analysis, more specifically, the causal dependency analysis, is necessary.

Fig. 3. Test case prediction result based on: (a) Scenario I trained with full-scale features; (b) Scenario II trained with masked features selected manually based on domain knowledge. It results in a biased outcome from predictive ML models: a higher energy standard leads to higher energy consumption! In both subgraphs, the left part shows the selected features with set exposures (treatment inputs we want to vary) and outcome based on the scenario, while the right part is the prediction result on the test case: The y-axis lists different combinations of insulation standard and heating system setting, while the x-axis gives the EUI Heating prediction result from different models (by different markers).

3.3. *Causal dependencies analysis*

From a causal inference analysis perspective, the hidden relationships among input features cause the biased outcomes observed in Scenario II. Similar cases have been discussed in medical statistic research (Patil et al., 1985). In this section, we demonstrate that for the AEC domain, causal discovery can aid designers and engineers in comprehensively examining whether hidden relationships have been neglected and, by controlling them accordingly, avoid subjective bias and biased estimation. For a more intuitive engineering interpretation and evaluation, we expand upon Fig. 3 and present a coherent causal dependencies analysis process to demonstrate that the analysis help avoid the fallout situation, as shown in Fig. 4.

The first step of causal dependencies analysis is causal discovery, which is responsible for extracting a causal skeleton from training data in an unsupervised manner. The skeleton and process itself bring a critical nexus for connecting data-driven results with domain knowledge validation through causal skeleton pruning. In our case study, the pruning process is relatively straightforward, as demonstrated in Fig. 4b; only minor adjustments (marked in orange) are made based on the

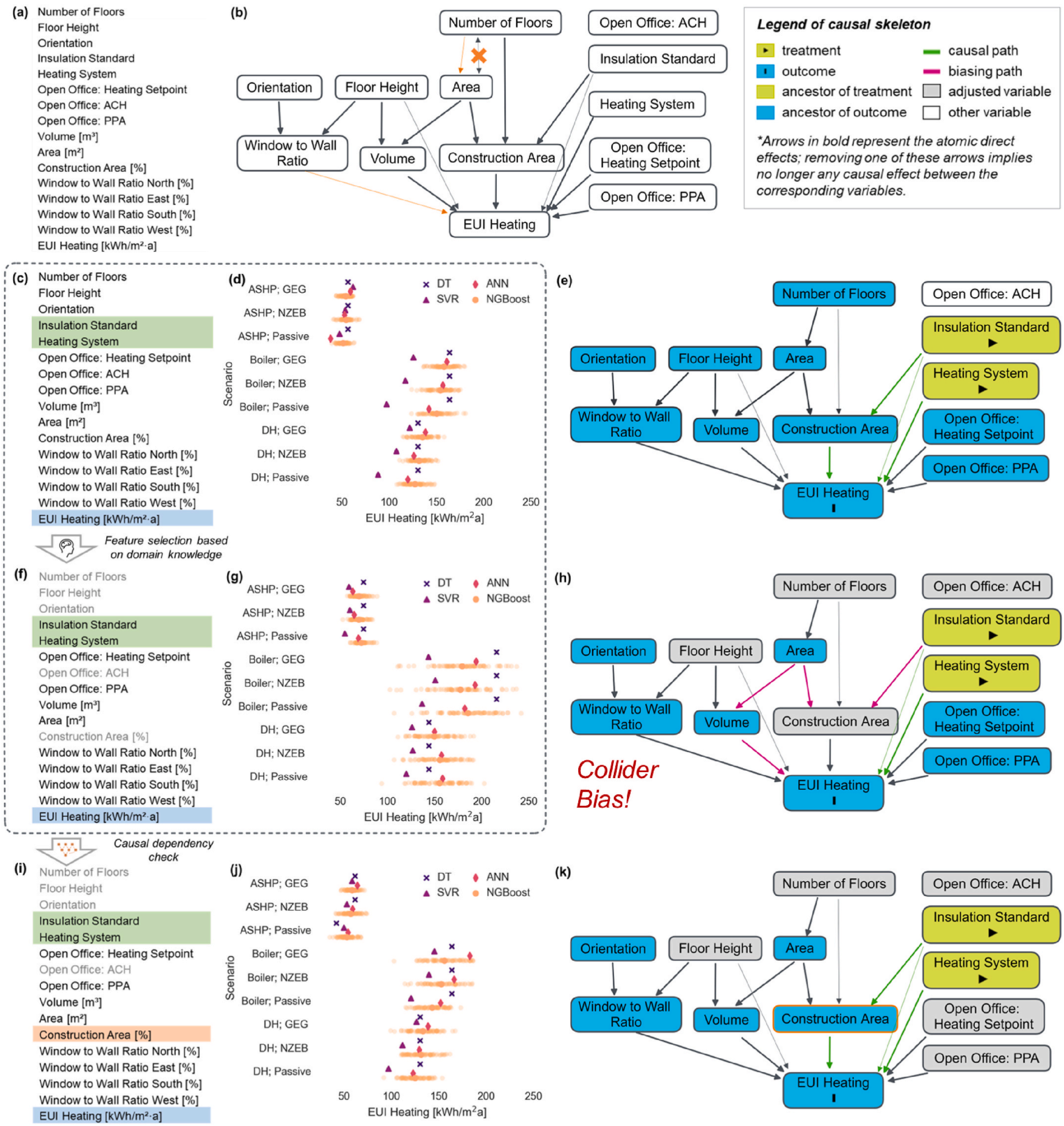

Fig. 4. Causal dependencies analysis process, the dotted box is the content of Fig. 3. (a), (b): Causal structure finding via GES: knowledge extraction based on the training dataset. Minor skeleton adjustments via domain knowledge are marked in orange; (c), (d), and (e): Scenario I; (f), (g) and (h): Scenario II: Blocking Construction Area leads to collider bias because it closes the direct causal path from Insulation Standard → Construction Area → EUI Heating, and open a biasing path from Insulation Standard → Area → Volume→ EUI Heating, which leads to spurious conclusion; (i), (j) and (k): Corrected Scenario II with no biasing path. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

original skeleton generated by GES:

1. Adding a causal dependency (arrow) from *Window to Wall Ratio (WWR)* to *EUI Heating,* since the causal connection between these two variables is slightly indirect. This is due to us manually merging all WWRs into one for a more simplified illustration.
2. Replacing the bidirectional arrow between *Number of Floors* and *Area* with a unidirectional arrow, as the number of floors is typically a variable given based on urban regulations determining the feasible floor area on a specific site.

Subsequent to the setup of the causal skeleton, the exposure inputs (*Insulation Standard* and *Heating System*) and the target outcome (*EUI*

Heating) are integrated into the skeleton, thereby establishing the causal flow, as illustrated in Fig. 4e. Based on the skeleton and scenario setting, we identified three crucial intermediate features*: Window to Wall Ratio, Volume,* and *Construction Area.* These features demonstrate direct causal effect connections to the target outcome and simultaneously carry causal dependencies with other features within the model.

Among these three features, *Construction Area* is at most important: It is the only feature that shares a common cause with the outcome (*EUI Heating*), and the common cause being one of the exposure inputs (*Insulation Standard*). This is expected given that the construction area is an input in the EUI estimation. In fact, the message/knowledge from causal analysis gives that: **As a common cause of the outcome, blocking/controlling the *Construction Area* leads to a biased result (Collider bias)**, because it closes the causal path from: *Insulation Standard → Construction Area → EUI Heating,* and open a biasing path (a detour connection from exposure to the outcome) as: *Insulation Standard → Area → Volume → EUI Heating* (Fig. 4h). This explains the unusual prediction results in Scenario II with variations in *Insulation Standard.* To correctly estimate the direct effect of *Insulation Standard* on *EUI heating,* we should either involve the feature *Construction Area* in the model to keep the causal path open, or we need to exclude *Construction Area, Area,* and *Volume* together to avoid the biasing path. In other words, causal dependencies exist between the building insulation standard, construction area, building area, and volume; controlling the intermediate one and varying the rest leads to a biased sampling situation.

Derivable interpretation from an engineering domain perspective: this causal finding conclusion mentioned above is derivable and can withstand cross-validation of domain knowledge, as the construction area serves as a common effect reflecting the configuration of the building area and building insulation standards: It is important to note that a larger building area and volume do not necessarily result in a proportional increase in the construction area. For instance, the thickness of building internal walls (non-loadbearing) and facades within the same insulation standard remains unchanged. Consequently, as the total building area expands, the building construction area proportion correspondingly shrinks. Meanwhile, higher building insulation standards correlate with better thermal isolation behavior for building façades. Better isolation typically equates to a thicker structure installation, hence the increase in construction area. Although we consider the *Construction Area* not directly affecting the *EUI Heating* since we vary the insulation standards, removing this feature from the model means the model samples through possible ranges from training data (refer to Table 2) and hence cancels out the consequential changes of *Insulation Standard*, while building *Area* and *Volume* are fixed, leading to more biases samples.

3.4. Validation

Building upon the conclusion from the causal dependencies analysis above, we can state:

> *"To properly investigate the causal effect from the Insulation Standard to EUI, the Construction Area should not be ignored for an unbiased effect estimation."*

With the same features selected as in Scenario II, ***Construction Area*** is additionally included. The corresponding performance with the updated feature set is given in Table 4:, while the test case prediction result is illustrated in Fig. 4j. Notably, with only a slight decrease in accuracy compared to the performance in Scenario I (Table 3), the prediction trend and uncertainty ranges of the *EUI Heating* align with the output in Scenario I again.

Table 4
5-fold cross-validation performance result comparison of different models: Validation Scenario.

	R^2
Decision Tree	0.81
SVR	0.90
ANN	0.96
NGBoost	0.90

3.5. Occam's razor for knowledge discovery: identifying the minimal sufficient adjustment set

Causal discovery analysis could also contribute to determining the minimal number of required variables thanks to the concept of "minimal sufficient adjustment sets". A causal DAG helps to answer the following common question in the data-driven process:

> *"Which variables (features) should we include for in our model to get an unbiased estimate of the effect?"*

A "minimal sufficient adjustment set" refers to the smallest set of variables that need to be adjusted to reliably estimate a causal effect, which provides crucial information to help the user collect the minimal but necessary features for unbiased prediction. These sets can be identified manually (Zheng et al., 2017; LeCun et al., 1988) or with a computer package (Textor et al., 2016). In this context, the well-known concept of Occam's razor is appropriate for the causal model preference (Pearl, 2000).

Take our case as an example, one minimal sufficient adjustment set would include: *Construction Area*, *Floor Height,* and *Volume*. A skeleton illustration is given in Fig. 5. As a result, we observe a similar unbiased trend in the case prediction as in Scenario I (Fig. 3a). Combined with the prediction result, we recognize the potential for knowledge discovery in engineering scenarios by interpreting features present in the minimal sufficient adjustment set.

Finally, it is essential to point out that DAGs and the minimal sufficient adjustment set solely provide identification information to ensure unbiased estimation, rather than addressing estimation performance. In engineering contexts, this data-driven process needs to relate to domain knowledge and thus be given context by the task-specific scenario for further analysis.

4. Discussion

We utilize a fallout case to demonstrate an easily identifiable error when using data-driven models. However, identifying such errors could be much more challenging for designers in many cases, potentially leading to a distrust in data-driven methods. While easily identifiable errors are prominent in data-driven methods, similar risks of biased information exist when using first-principles simulations. First-principles simulations, extensively developed by numerous engineers and experts, carry their own biases (Rakitta and Wernery, 2021; Klotz, 2011; Zalewski et al., 2017), often hidden due to their established nature. Cognitive biases (Minsky, 1991) can also cause such fallout situations: An example relevant is the confirmation bias, where engineers might favor information (e.g., a familiar type of design pattern, system deployment, or validation method) that confirms their preexisting beliefs or hypotheses while ignoring or downplaying contrary evidence. This bias leads to a skewed acquisition or utilization of personal domain knowledge. By taking this potential bias into account, simulation results also bear fallout risks and often lack an appropriate adjustment mechanism.

In this context, causal analysis serves as a useful tool for identifying potential biases in prior data, bridging links to reinforce domain knowledge with data-driven methods. We argue that data-driven methods and first-principles simulations are not inherently conflicting. Rather, combining them may offer a practical solution to manage and mitigate the risk of biased outcomes.

While managing cognitive biases is crucial, another significant

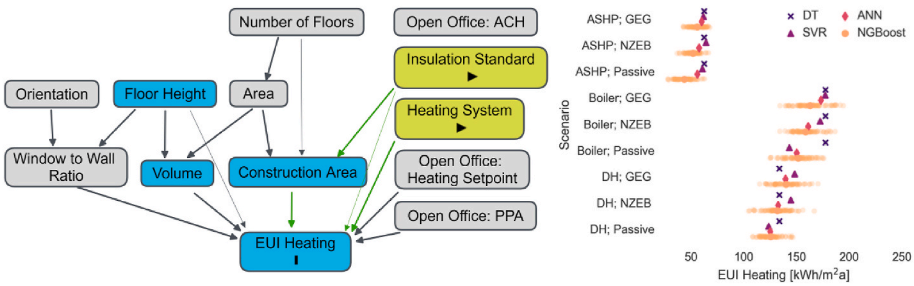

Fig. 5. Minimal sufficient adjustment set based on the case: With Floor Height, Volume and Construction Area as extra inputs, the model generates unbiased estimation with sufficient information from the dataset.

aspect to consider is the process of feature selection. It may seem that the more features (input variables) involved in the modeling process, the more comprehensive the causal skeleton should be. Simply feeding more features into the modeling process doesn't necessarily contribute to the accuracy improvement. We perceive this as a trade-off between precision and accuracy in describing the case:

- More features formalize a good representation of the target case, reducing uncertainty with a more accurate description, but also raising the risk of biased variation analysis.
- Using fewer features reduces the risk of biased result analysis; however, a too simple feature representation might overlook important factors, leading to incorrect conclusions.

To address this, we see huge potential in the future for integrating the causal inference check and knowledge discovery mechanisms directly into the data-driven methodology itself. This would mirror the concept of physics-informed ML (Chen and Geyer, 2023; Karniadakis et al., 2021), where the emphasis is placed on seamlessly integrating process knowledge into data-driven models. Such an integration would ensure that causal insights are more immediately and effectively incorporated into model updates, leading to a more dynamic and responsive framework.

While our current study provides valuable insights into the application of our methodology in engineering scenarios, we recognize the need for further validation using larger and more diverse datasets. Future research could benefit from applying our framework to datasets with higher dimensions to further validate its robustness and generalizability.

In the broader context of our study, the generalizability of our causal analysis approach is of paramount importance. Initially applied in the AEC industry, our methodology demonstrates potential across various engineering domains. It embodies a symbiotic interaction that combines data-driven causal knowledge discovery with cross-validation of prior knowledge, while simultaneously addressing potential biases in modeling tools. This approach is akin to the perception-action-feedback loop in cybernetics but with a distinct emphasis. Here, causal discovery through data-driven models, domain knowledge, and first-principles simulations focuses more on information-theoretic machine assistance or augmented intelligence (Zheng et al., 2017). Engineers and researchers are thus empowered to not only derive predictive conclusions from data but also discover and address gaps in knowledge. This transcends individual limitations in addressing engineering problems, offering a universally applicable, enhanced model of understanding and application.

5. Conclusion

The evolution of engineering analysis methodologies has fostered synergetic interaction among data, domain knowledge, simulations, and data-driven methods. Our case study highlights the potential pitfalls of relying solely on data-driven methods without incorporating causal analysis. We proved that it is critical to examine causal relationships when performing a data-driven analysis to avoid misleading results. Consequently, we advocate for more attention and involvement in causal inference analysis in the engineering community. Moreover, we believe that extracting invariant and transferable information from data is crucial in bridging the gap between domain knowledge, simulations, and data-driven methods in engineering and transcending individual capabilities' limitations.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

We gratefully acknowledge the German Research Foundation (DFG) support for funding the project under grant GE 1652/3-2 in the Researcher Unit FOR 2363 and under grant GE 1652/4-1 as a Heisenberg professorship.

Appendix

i. *Mechanism Introduction of Machine Learning Methods*

Tree-based models seek to identify optimal split points in the data to enhance prediction accuracy. The term "tree" refers to a decision tree, which forms the foundation of tree-based models. The decision tree algorithm identifies which data feature to split on and when to cease splitting based on information gain criteria (i.e., minimizing entropy in data split). While straightforward to interpret, decision trees are generally weak predictors. Enhanced ensemble methods such as bagging, random forest, boosting (Spirtes, 2010), and gradient boosting (Pearl, 2009) have been adapted to improve performance but lead to less interpretable behavior.

Kernel machines utilize a linear classifier to address non-linear problems by defining a separating hyperplane to fit in data and make predictions. A kernel corresponds to a dot product in a typically high-dimensional feature space (Schölkopf et al., 2021). In this space, estimation methods are linear, and all formulations are made in terms of kernel evaluations, thereby avoiding explicit computation in the high-dimensional feature space.

Neural networks comprise input, hidden, and output layers, where each layer is a group of neurons, loosely modeling the neurons in a biological brain. The connections between neurons (also called nodes) carry associated weights/biases. The data is fed into the network and passes through all neurons with activation functions (which add non-linearity to the output) in the forward propagation to produce output. The backpropagation mechanism (LeCun et al., 1988) updates neuron weights/biases according to the difference between prediction and output (loss function evaluation).

ii. Modeling Configuration for Generating Training Data

The test case is a mixed-usage 4-floor building named Building.Lab on a tech campus in Regensburg, Germany (Chen et al., 2022b). The function of this 1956 m^2 building is office and seminar use as well as housing, which consists of four above-ground stories and one underground level with a concrete skeleton structure. For supporting decision-making in energy-efficient building design, we developed a parametric model of an office building in a generic H-shape that covers a wide configuration variety of building components and zones. We varied this model to generate a representative training dataset for well-generalizing models on the target scenarios covering the design space characteristics of the case and similar buildings for performance evaluation. An illustration of the data generation process is given in

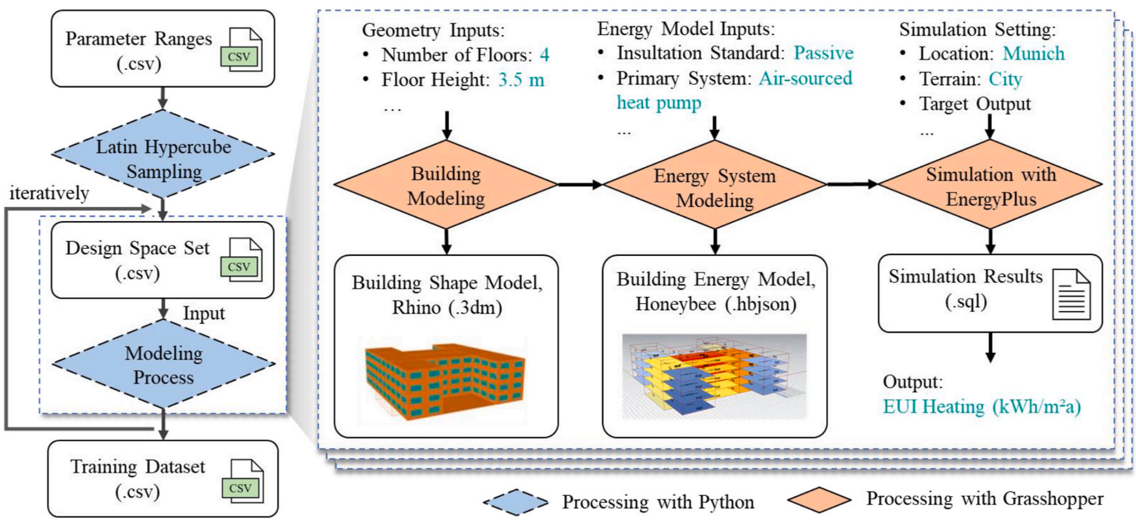

Figure 6. Automatic data generation process with parametric modeling: a generic H-shape office building. The parameter ranges are determined with the consideration of covering the test case scenario and densely sampled with variations. Each sample is fed iteratively into the energy simulation pipeline composited by Grasshopper, Python, and intermediate models. 918 samples were generated as the training dataset.

For the variation of building insulation standards, we simulated three component thermal characteristic sets based on real-world building energy standards and, from low to high: 2020 German Energy Act for Buildings (GEG), Net Zero Energy Building (NZEB), and Passive House. The standards have different requirements for components' thermal conductivity (U-values), as presented in Table 5.

Table 5
Different insulation standard requirements for building component thermal characteristics [$W/m^2 \cdot K$].

Insulation standard of U-Values in building components	Base: GEG (2020 German Energy Act for Buildings)	Medium: NZEB (Net Zero Energy Building)	High: Passive House
Base plate	0.2625	0.206	0.15
Roof	0.15	0.135	0.12
Exterior wall, bearing, above ground	0.21	0.18	0.15
Exterior wall, bearing, under ground	0.2625	0.206	0.15
Window	0.975	0.888	0.8

As for heating systems, three typical building energy systems are simulated: boiler, air-source heat pump (ASHP), and district heating (DH). All systems have been modeled with convective hot water baseboards as their secondary energy system. The hot water loop temperature was 50 °C for the ASHP system variant and 80 °C for the boiler and district heating system variants. The piping system was modeled as adiabatic. The heating setpoint scales a typical office hour schedule to a new target setpoint. During off-work hours (starting from 6 p.m.), only 75% of the setpoint is set. Starting at 6 a.m., setpoints are increased hourly to 85%, 95%, and 100%. The minimum heating temperature is set to 21 °C as we referred to the national standard DIN EN 16798-1, and we intend to find sustainable and high-performing solutions (all options to be inside category I with PPD<6%). As the comfort temperature is 22 °C ± 2K for environments below 16 °C, we chose 21–24 °C. In this simulation model, no cooling system and mechanical ventilation were modeled. The zone ventilation was only set by the air change rate per hour based on exterior air volume demands set from DIN EN 16798-1.

To validate the simulation result, we sampled the generated data (Training data) by different insulation standards and heating systems, as presented in Tables 6 and 7, respectively.

Table 6
Energy Usage Intensity (EUI) Heating distribution, sampled by heating system choice

Energy Usage Intensity (EUI) Heating [kWh/m^{2}a]	All		ASHP		Boiler		DHWB	
	mean	std	mean	std	mean	std	mean	std
	84.6	50.1	45.4	13.5	143.0	44.6	106.3	31.6

Table 7
Energy Usage Intensity (EUI) Heating distribution, sampled by insulation standard

Energy Usage Intensity (EUI) Heating [kWh/m^{2}a]	All		GEG		NZEB		Passive	
	mean	std	mean	std	mean	std	mean	std
	84.6	50.1	90.8	56.5	85.1	49.0	78.0	45.1

iii. Training Process and Result Validation

During the model training process, a hyperparameter grid-search strategy with 5-fold cross-validation (Refaeilzadeh et al., 2009) is applied for fitting data scheme changes in each scenario for all ML models. From an intuitive understanding, it means the same model with all hyperparameter setting combinations are cross evaluated within the 80/20 split training data, to compare and ensure the models' best performance for test case validation. The results analysis by three evaluation metrics in all scenarios is presented in Table 8. A short context of different scenarios is given as follows:

- **Scenario I**: Represents full-scale modeling using all input features for EUI heating prediction, serving as a benchmark.
- **Scenario II**: Involves masked input features to simulate common real-world engineering situations where feature selection is guided by domain knowledge or limited by available data.
- **Validation Scenario**: Retains the same feature selection as Scenario II, with the addition of Construction Area, based on causal analysis. This inclusion aims to mitigate biased outcomes.

Table 8
5-fold cross-validation performance result comparison of different models, all scenarios. Solely validating data-driven model accuracy does not eliminate the risk of biased result.

		Decision Tree	SVR	ANN	NGBoost
Scenario I (Full features)	*NRMSE*	8.22	7.85	4.04	4.51
	SMAPE	0.15	0.14	0.10	0.09
	R^2	**0.86**	**0.87**	**0.96**	**0.95**
Scenario II (Part features)	*NRMSE*	9.70	7.81	5.35	7.48
	SMAPE	0.18	0.14	0.10	0.14
	R^2	**0.81**	**0.87**	**0.94**	**0.88**
Validation (Part features, causal-informed)	*NRMSE*	9.58	6.81	4.43	7.05
	SMAPE	0.18	0.11	0.09	0.14
	R^2	**0.81**	**0.90**	**0.96**	**0.90**

The intention behind Table 8 is to illustrate that solely validating data-driven model accuracy does not eliminate the risk of biased results. Scenario II, though demonstrating decent performance, leads to misleading conclusions, as discussed in Section 3.2. For instance, disregarding insulation standard choices could falsely suggest their insignificance or promote lower standards to reduce energy usage. Scenario II stands for typical real-world situations: where only part of features is available, and the causal relationships between features are often overlooked during data collection. Therefore, causal analysis and dependency checks, as in our validation scenario, are essential to ensure unbiased results under limited feature availability. Choosing Scenario I exclusively is impractical in complex real-world systems due to the inherent challenge of capturing all underlying factors. The comparative accuracy of the three scenarios underscores that relying solely on accuracy metrics can obscure potential biases, strengthening the necessity of integrating causal analysis in the data-driven modeling process.

References

Aldrich, J., 1995. Correlations genuine and spurious in pearson and yule. Stat. Sci. 10 (4).

Balestriero, R., Pesenti, J., LeCun, Y., 2021. Learning in High Dimension Always Amounts to Extrapolation.

Bertolini, M., Mezzogori, D., Neroni, M., Zammori, F., 2021. Machine Learning for industrial applications: a comprehensive literature review. Expert Syst. Appl. 175, 114820.

Chakraborty, D., Elzarka, H., 2019. Advanced machine learning techniques for building performance simulation: a comparative analysis. J. Build. Perform. Simul. 12 (2), 193–207.

Chen, X., Geyer, P., 2023. Pathway toward Prior Knowledge-Integrated Machine Learning in Engineering arXiv preprint arXiv:2307.06950.

Chen, Xia, Abualdenien, Jimmy, Singh, Manav Mahan, Borrmann, André, Geyer, Philipp, 2022a. Introducing causal inference in the energy-efficient building design process. Energy Build. 277, 112583.

Chen, Xia, Saluz, Ueli, Staudt, Johannes, Margesin, Manuel, Lang, Werner, Geyer, Philipp, 2022b. Integrated Data-Driven and Knowledge-Based Performance Evaluation for Machine Assistance in Building Design Decision Support: the 29th EG-ICE International Workshop on Intelligent Computing in Engineering. Aarhus, Denmark.

Chen, X., Guo, T., Kriegel, M., Geyer, P., 2022c. A hybrid-model forecasting framework for reducing the building energy performance gap. Adv. Eng. Inf. 52, 101627.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comp. Sci. 7, e623.

Chickering, D.M., 2002a. Optimal structure identification with greedy search. J. Mach. Learn. Res. 3 (Nov), 507–554.

Chickering, D.M., 2002b. Learning equivalence classes of Bayesian-network structures. J. Mach. Learn. Res. 2, 445–498.
Clark, L.A., Pregibon, D., 2017. Tree-based models. In: Statistical Models in S. Routledge, pp. 377–419.
DeVore, R.A., Temlyakov, V.N., 1996. Some remarks on greedy algorithms. Adv. Comput. Math. 5 (1), 173–187.
Dimiduk, D.M., Holm, E.A., Niezgoda, S.R., 2018. Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. Integr. Mater. Manufact. Innov. 7, 157–172.
DIN EN 16798-1:2022-03, Energetische Bewertung von Gebäuden_- Lüftung von Gebäuden_- Teil_1: Eingangsparameter für das Innenraumklima zur Auslegung und Bewertung der Energieeffizienz von Gebäuden bezüglich Raumluftqualität, Temperatur, Licht und Akustik_- Modul M1-6; Deutsche Fassung EN_16798-1:2019. Berlin: Beuth Verlag GmbH. doi:10.31030/3327351.
Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H., 2020. A survey of learning causality with data. ACM Comput. Surv. 53 (4), 1–37.
Hegde, J., Rokseth, B., 2020. Applications of machine learning methods for engineering risk assessment-A review. Saf. Sci. 122, 104492.
Hofmann, T., Schölkopf, B., Smola, A.J., 2008. Kernel methods in machine learning. Ann. Stat. 36 (3), 1171–1220.
Judea, P., 2010. An introduction to causal inference. Int. J. Biostat. 6 (2), 1–62.
Kalisch, M., Bühlmann, P., 2014. Causal structure learning and inference: a selective review. Qual. Technol. Quant. Manag. 11 (1), 3–21.
Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nat. Rev. Phys. 3 (6), 422–440.
Klotz, L., 2011. Cognitive biases in energy decisions during the planning, design, and construction of commercial buildings in the United States: an analytical framework and research needs. Energy Eff. 4, 271–284.
LeCun, Y., Touresky, D., Hinton, G., Sejnowski, T., 1988. A theoretical framework for back-propagation. In: Proceedings of the 1988 Connectionist Models Summer School, pp. 21–28.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Marcher, C., Giusti, A., Matt, D.T., 2020. Decision support in building construction: a systematic review of methods and application areas. Buildings 10 (10), 170.
Minsky, M.L., 1991. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. AI Mag. 12 (2), 34.
Patil, R.S., Szolovits, P., Schwartz, W.B., 1985. Causal understanding of patient illness in medical diagnosis. In: Computers and Medicine. Springer New York, New York, NY, pp. 272–292.
Pearl, J., 2000. Causality: Models, Reasoning, and Inference/Judea Pearl, second ed. Cambridge University Press, Cambridge.
Pearl, J., 2009. Causal inference in statistics: an overview. Stat. Surv. 3 (0), 96–146.
Pearl, J., Mackenzie, D., 2018. The Book of Why: the New Science of Cause and Effect. Basic Books.
Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press.
Rakitta, M., Wernery, J., 2021. Cognitive biases in building energy decisions. Sustainability 13 (17), 9960.
Raschka, S., Patterson, J., Nolet, C., 2020. Machine learning in python: main developments and technology trends in data science, machine learning, and artificial intelligence. Information 11 (4), 193.
Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. Encyclop. Datab. Syst. 5, 532–538.
Roman, N.D., Bre, F., Fachinotti, V.D., Lamberts, R., 2020. Application and characterization of metamodels based on artificial neural networks for building performance simulation: a systematic review. Energy Build. 217, 109972.
Schölkopf, B., 2019. Causality for Machine Learning.
Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., et al., 2021. Toward causal representation learning. Proc. IEEE 109 (5), 612–634.
Seyedzadeh, S., Rahimian, F.P., Glesk, I., Roper, M., 2018. Machine learning for estimation of building energy consumption and performance: a review. Vis. Eng. 6 (1).
Singh, A., Thakur, N., Sharma, A., 2016. A review of supervised machine learning algorithms. In: 2016 3rd International Conference on Computing for Sustainable Global Development. INDIACom, pp. 1310–1315.
Spirtes, P., 2010. Introduction to causal inference. J. Mach. Learn. Res. 11 (5).
Spirtes, P., Glymour, C.N., Scheines, R., 2000. Causation, Prediction, and Search, second ed. MIT Press, Cambridge, MA.
Textor, J., 2015. Drawing and Analyzing Causal DAGs with DAGitty arXiv preprint arXiv: 1508.04633.
Textor, J., van der Zander, B., Gilthorpe, M.S., Liskiewicz, M., Ellison, G.T., 2016. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int. J. Epidemiol. 45 (6), 1887–1894.
Zalewski, A., Borowa, K., Ratkowski, A., 2017. On cognitive biases in architecture decision making. In: Software Architecture: 11th European Conference. ECSA, Canterbury, UK, pp. 123–137. September 11-15, 2017, Proceedings 11; 2017.
Zheng, N., Liu, Z., Ren, P., Ma, Y., Chen, S., Yu, S., et al., 2017. Hybrid-augmented intelligence: collaboration and cognition. Front. Inform. Technol. Electr. Eng. 18 (2), 153–179.