A machine learning-based framework to support performance-based early design of buildings: Data and Methodology

Mohsen Zaker Esteghamati
Assistant Professor, Dept. of Civil & Environmental Engineering, Utah State University, Logan, USA

ABSTRACT: Achieving resilient and sustainable infrastructure urges developing computational tools to explicitly consider performance objectives in all design and construction stages. The majority of critical decisions are made at earlier stages of the design. Early design can be substantially improved by incorporating quantitative methods to evaluate the consequences of these decisions. However, implementing quantitative methods poses several challenges, including imprecision of design variables and time- and effort-intensiveness of such assessments. This paper presents a modular framework to select suitable candidate structural systems, characterize their design parameters range, and communicate their expected hazard and environmental performance during their life cycles. The framework leverages a machine-learning-assisted workflow that performs mapping between crude design- and topology-related parameters and global hazard and environmental performance indicators. Next, a sequence of surrogate models with varying fidelity aids in performing the convergence-divergence cycle of early design. Lastly, a deep learning architecture with a customized loss function maps the result of simpler static analysis to the detailed description of seismic performance, linking early design to the next design stages. A case study is presented to illustrate the application of the framework to evaluate the embodied carbon and seismic-related repair cost of an inventory of 720 multi-story concrete frames with varying topologies in Charleston, South Carolina.

Architectural design is a mental transaction that transforms *concepts* into *representation* (Bueno and Turkienicz, 2014), through a nonlinear process that transitions between different concepts to render the desired product (Smith and Smith, 2014). The sequence of these transitions demarcates a given representation and helps the designer to re-think the relationship between different representations (Bueno and Turkienicz, 2014), particularly at earlier stages of the design. Architectural design encompasses several design phases, starting with schematic design to design development and construction documents (Mehta et al., 2017).

Early design is the most critical stage in improving building performance (Østergård et al., 2016). An unaided early design follows the designer's familiar domain of expertise, and subsequently, cognitive biases (Rezaee et al., 2015). Therefore, this design process will only explore a small set of alternatives using limited criteria, which will most likely miss higher-performing alternatives. Incorporating quantitative methods can address these shortcomings and improve the early design process (Zaker Esteghamati and Flint, 2021).

Performance-based engineering (PBE) is a probabilistic approach to quantify buildings' performance against natural hazards using quantitative metrics such as repair cost and functionality loss (Alvarez et al., 2013; Bertero and Bertero, 2002; Ciampoli et al., 2011). Therefore, PBE can inform the designer regarding the hazard consequences of different decisions. In addition, the modular nature of PBE facilitates incorporating different economic and environment-related metrics, leading to a more holistic design.

Despite the advantages of PBE for early design, its implementation poses several challenges, such as time- and effort-intensiveness of PBE assessments and the necessity of expertise on hazard impact modeling e that might not be available to the design team. Furthermore, a successful early design must explore many alternatives. This design space exploration will exhaust computational resources and pressure the project timeline when compounded by PBE assessment.

Different decision support tools have been proposed to supplement early design, although none are optimal. Perhaps the most popular methods are building information models (BIM)-based tools accompanied by interactive plugins and visual scripting (Cheung et al., 2012; LLatas et al., 2022). Knowledge-based decision support tools are also available that provide preliminary guidance on material selection and sizing (Schneider-Marin et al., 2022; Zabalza Bribián et al., 2011), or complement BIM tools to define preset assemblies (Rezaee et al., 2019).

Recently, machine learning (ML) approaches have been introduced to address building design and assessments (Olu-Ajayi et al., 2022; Sun et al., 2021), albeit their application is still limited in early design (Singh et al., 2022). ML modeling directly learns from the data by exploiting the availability of a large amount of simulation data. Therefore, ML models can detect unforeseen relationships and make predictions without establishing a formal hypothesis.

This paper describes a data-driven framework to conduct a performance-based early design. Throughout this paper, early design refers to a combination of schematic and design development phases, emphasizing the latter. The framework exploits the complex and implicit relationship between geometry, design, hazard, and environmental performance. The framework's primary workflow extracts this relationship by training supervised ML models on performance inventories, combining these models with other surrogate models to perform a sequential appraisal of the design space. The performance inventory compiles PBE assessments consistent with building taxonomy and site, and is supported by a knowledge-based module. The knowledge-based module organizes prior knowledge (such as published PBE assessments) in relational databases to provide data for data-driven models, or to directly aid with deriving knowledge-based surrogate models. The developed surrogate models are then implemented in a sequence to explore the design space. Lastly, a deep learning model maps the result of simplified analysis to detailed performance-based assessments suitable for later stages of the design. A case study illustrates the application of the framework to explore the design of mid-rise concrete office buildings in Charleston, South Carolina, from seismic vulnerability and environmental performance perspectives.

1. FRAMEWORK OBJECTIVES

The underlying concept of the proposed framework is to combine PBE and ML modeling, where ML models overcome the computational challenges and offer a fast means to extend and scale PBEE over the design space. As shown in Figure 1, the resultant data-driven performance-based seismic design framework provides four advantages over traditional heuristics.

First, the proposed framework can compile a large set of design alternatives using sampling-based approaches. The sampling-based approaches also allow space reduction to decrease the sheer effort needed for the high cardinality of real-world design problems. Second, the

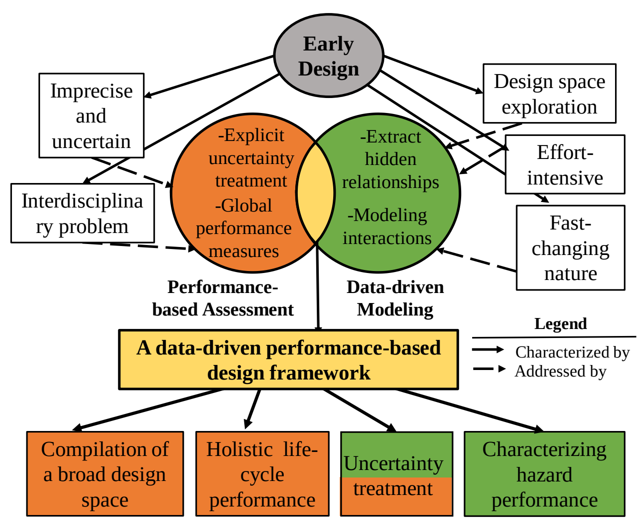

Figure 1: Framework objectives

framework can consider multiple performance objectives, such as environmental, economic, and hazard-related performance, allowing for a holistic design approach. Third, the framework can characterize the range of hazard performance range for different early design decisions by applying PBE to the approximate description of design alternatives. Such extensions provide a significant advantage over the current use of PBE at the later stages of design. Lastly, the framework allows for scrutinizing different sources of uncertainty (e.g., material properties, modeling, hazard characterization, and intensity) and identifying their impact on the performance of candidate designs. This explicit treatment of uncertainty increases the designer's confidence in the design guidance provided by the framework.

2. FRAMEWORK OVERVIEW

The proposed framework provides risk-informed insights on the best possible structural systems and preliminary estimates on system design and configuration (e.g., weight, footprints). Figure 2 shows the schematics of the proposed framework. The underlying notion of the proposed framework is to use *surrogate models* to estimate the performance range based on crude design and geometry information. Here, "surrogate" is referred to all simplified models that can provide a mapping between input (i.e., design and geometry data) and output (i. e., performance metrics) such as knowledge-based or data-driven models. These surrogate models must tolerate imprecise and limited information of early design.

The primary workflow leverages supervised-learning ML algorithms to build surrogate models from PBE assessment data. The PBE data can be generated by alternate pathways (detailed assessment or simplified models) and is supported through a knowledge-based module. The knowledge-based module organizes prior knowledge (e.g., existing PBE assessment) in relational databases. As a result, the framework provides a fast means to explore design space, where the designer only needs to change the statistical surrogate model input accordingly to get an instant assessment.

As shown in Figure 2, the framework also combines different surrogate models to perform convergence-divergence cycles. Typical early design iterates between *convergence* cycles, where the designer removes unfavorable design alternatives, and *divergence* cycles where they expand the design pace by introducing new alternatives A sequence of surrogate models can be applied, where the models with lower fidelity are used at the earlier stage (i.e., larger design space) to remove redundant alternatives. For example, knowledge-based models (e.g., performance range based on available data of similar systems) can be used to screen design alternatives (i.e., convergence), and then low-order dynamic models, such as single-degree-of-freedom (SDOF) systems, can be generated for a

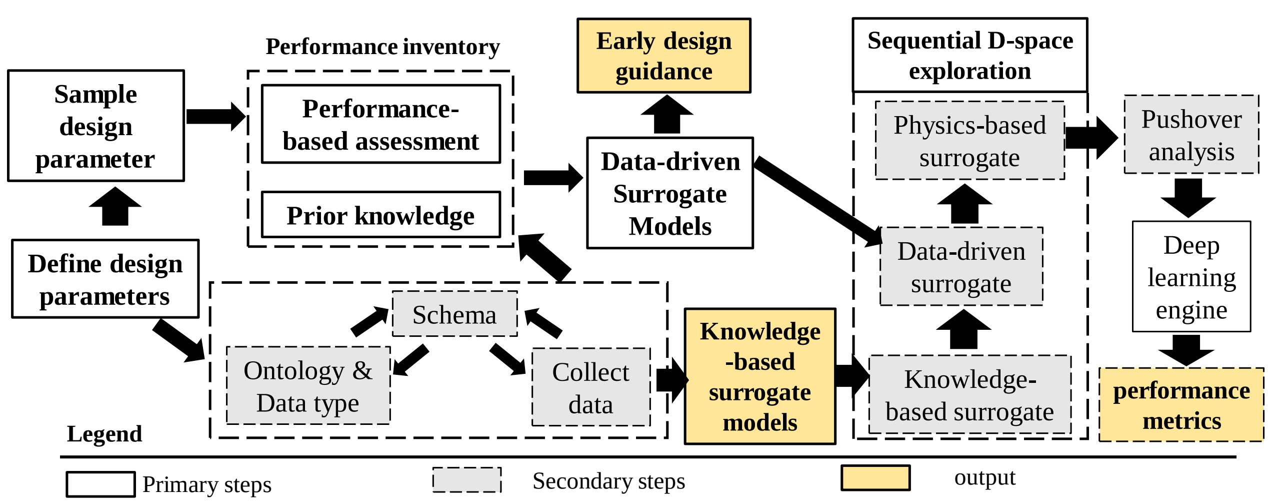

Figure 2: Schematics of the proposed framework

few of selected candidates to perform a parametric study (i.e., divergence).

In the last stage, the framework uses a deep learning engine to estimate the hazard performance of the selected alternatives (from parametric analysis) based on a simplified static analysis denoted as pushover analysis. The deep learning model uses a customized encode-decoder architecture using Long Short-term Memory (LSTM) algorithm (Soleimani-Babakamali and Zaker Esteghamati, 2022). The deep learning engine maps the resulting force-deformation relationship from pushover analysis into probabilistic seismic demand models, which can be readily mapped to seismic damage or loss through pre-defined damage states and vulnerability functions.

3. DATA NEEDS

The proposed framework uses different types of data to support the surrogate models. Preparing data, including acquisition and preprocessing, is the first and perhaps the most critical step of any ML-based framework. Few open databases currently provide adequate data on building taxonomies, hazard types, and performance measures.

Relational databases (RDBs) are a promising platform to present future hazard performance databases. RDBs offer better data sharing and collaboration. In addition, RDBs organize data more efficiently using interrelated data tables and shared fields (i.e., keys). There have been a few attempts to use RDBs to structure performance data. For example, Esteghamati *et al*. compiled an open relational database, INSSEPT (Zaker Esteghamati et al., 2020), to aggregate 222 PBEE case studies from 39 papers. Such a database can provide the seismic performance of the different structural systems over a broad geographical region. Omoya *et al.* provided recovery and damage parameters of 3695 buildings after the 2014 Napa earthquake as a relational database (Omoya et al., 2022).

Besides hazard performance data, there is a severe need to collect data on other aspects of a building's performance, such as environmental impacts. Life cycle assessments (LCAs) provide a systematic approach to evaluating environmental impacts, although LCA studies significantly vary over their definition of scope and system boundary, obstructing their reuse. Esteghamati *et al.* compared six commercial buildings with varying foundation, structural, and envelope systems, and combined the result with similar studies from the literature. Nevertheless, due to small sample size, the classical statistical test did not have adequate power to differentiate between the impacts of different systems (Zaker Esteghamati et al., 2022). Such observation emphasizes the importance of collecting and curating LCA data for future data-driven modeling.

4. ILLUSTRATIVE EXAMPLE

A case study of mid-rise concrete frames was studied to assess the framework's applicability. The investigation aimed to identify the range of geometry (e.g., height, floor area) and design parameters (e.g., average section sizes) that yield the best performance for earthquake hazard and environmental performance. Here, earthquake hazard performance was quantified in terms of life-cycle repair cost, whereas environmental impact was measured as embodied carbon due to the building service life of 50 years.

An inventory of 720 two-dimensional analytical models of concrete frames was generated in OpenSees (McKenna, 2011) using an automated workflow. The workflow uses Latin Hypercube Sampling (LHS) to cover geometric configuration and a pseudo-directional approach to populate the design parameters based on nonlinear dynamic analysis. The analytical models use plastic hinge formulation, where frame members were modeled as elastic elements with two nonlinear hinges at both ends. The parameters of these nonlinear hinges were derived based on regression equations derived from previous experimental studies (Haselton et al., 2008). All frames were checked based on minimum code requirements to ensure they conform to basic safety objectives (e.g., allowable drift, and adequate strength). Additional

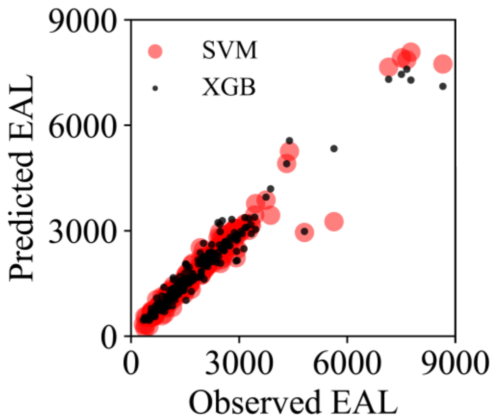

Figure 3: Comparison of SVM and XGB prediction for the test set frames

information on frame modeling and design can be found elsewhere (Zaker Esteghamati and Flint, 2021).

The compiled inventory was then subjected to 80 site-specific synthetic ground motions derived from geologically-realistic hazard assessment (Chapman and Talwani, 2006). The results were post-processed to derive maximum displacement and acceleration responses at story levels. Fragility functions (i.e., the probability of exceeding a damage state) were then calculated through a univariate and logistic regression for non-collapse and collapse cases, respectively. The fragility functions were then used to derive the annual expected repair cost (EAL) based on the HAZUS vulnerability function following an assembly-based approach (Ramirez et al., 2012). Three assemblies of structural, non-structural drift-sensitive, and non-structural acceleration-sensitive were selected.

Supervised ML models were trained on the compiled performance inventory. The input was taken as height, floor area, the weight of lateral systems, the average area of beams over the entire building and at the first floor, and the average reinforcement ratio of beams, whereas the response was taken as annual repair cost. The result shows that support vector machines (SVM) and extreme gradient boosting (XGB) algorithms provide the highest accuracy. Figure 3 compares the prediction of SVM and XGB on the test set. On average, XGB and SVM provide an adjusted

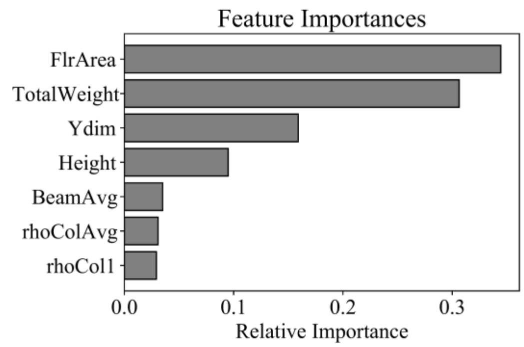

Figure 4: Relative importance of XGB model

R-squared of 0.86 and 0.96, where model root mean squared errors are $283 and $285, respectively. The developed ML models can also measure the sensitivity of estimated performance to the input parameters. For example, Figure 4 shows the importance of different features on EAL predicted by the XGB model. It can be observed that the most important predictors of seismic loss are floor area, building dimension, height, and total weight.

ML prediction error can be decomposed into bias and variance. Here bias shows how the average prediction deviates from the true response values, whereas variance quantifies the variation in an ML model's prediction. Figure 5 shows that the two best models (XGB and SVM) have low bias and variance for collapse loss, and similar variance for structural and non-structural acceleration-sensitive losses. However, SVM shows a slightly lower bias for both structural and non-structural acceleration-sensitive assemblies.

A simple linear regression was found adequate to relate embodied carbon due to initial construction and hazard-related repair to building

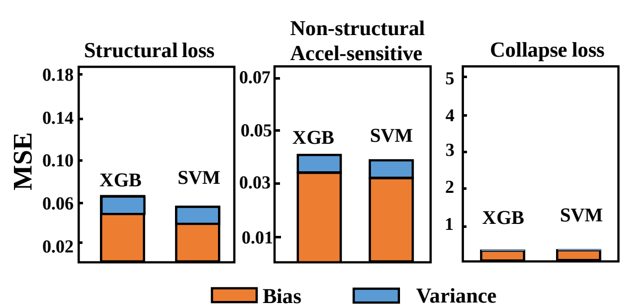

Figure 3: Bias and variance decomposition of ML models for different assemblies

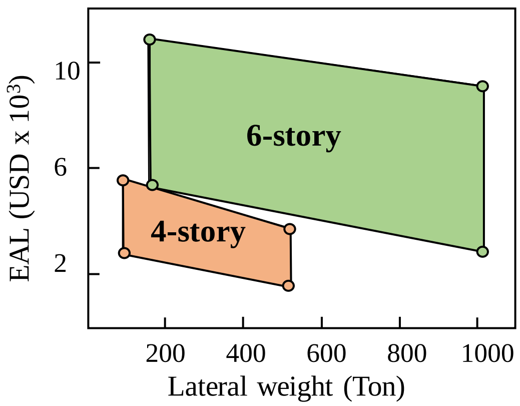

Figure 6: Comparison of earthquake-related repair cost for two building heights using imprecise information fed into SVM models

weight. The regression shows that an additional 0.3 tons of CO_2 will be produced for each additional ton in building weight. This is mainly due to the small value of hazard-related carbon emissions (i.e., about 1% of initial construction values). Therefore, the linear relationship between weight and embodied carbon suggests that the designer should select the alternatives with the lowest weight.

Figure 6 shows how SVM models demarcate the design space by providing a loss range based on a preliminary estimate of the input variable. Here, story height was fixed (6- and 4-story), and upper and lower bound values were provided for other parameters based on previous experience. As shown in Figure 6, SVM with imprecise information can distinguish the median repair cost between the two topologies. In addition, the primary insights from SVM indicate that structural systems with larger lateral weight (larger beams and column sections) led to smaller EAL for both building topologies. It should be noted that while the results are intuitive, they suggest that ML models captured a realistic relationship between hazard performance and design parameters.

Figure 7 shows the sequential application of prior literature, developed SVM model, and equivalent SDOFs to estimate the seismic loss range for a 6-story concrete building. As Figure 7 shows, the sequence can provide an estimate close to one obtained from the detailed analysis of a set

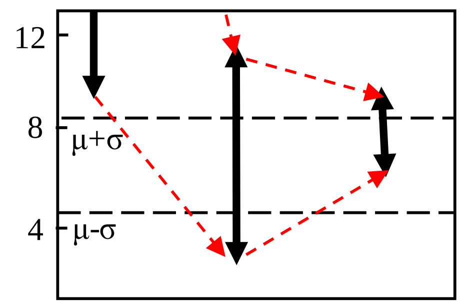

Figure 7: The sequential application of knowledge-based (KBS), SVM (DDS), and equivalent SDOF (PBS) surrogate models to estimate loss of a 6-story building based on imprecise information

of 6-story building models for this site. However, the sequential application could not accurately capture the variation in loss, and over-estimated the median loss for this building typology.

The deep learning engine was then used to estimate the demand model based on the pushover analysis of design alternatives. The deep learning model was trained on pushover data from detailed finite element models. These demand models can then be used to derive fragility and repair values. Figure 8 shows the application of this engine for a frame in the compiled inventory. The R-squared of the estimated fit from the deep learning model is only 1.4% smaller than the one obtained from detailed analysis. This observation shows that deep learning models can successfully estimate performance without performing dynamic analysis. Therefore, the designer can use the deep learning model to readily map the result of the parametric study at the divergence cycle to performance endpoints, bridging to later stages of design.

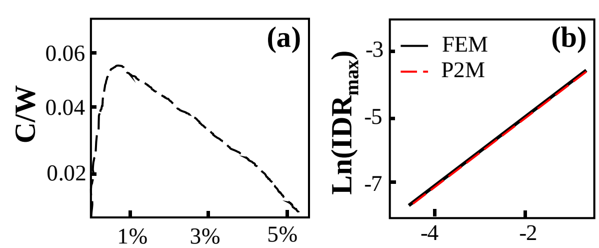

Figure 8: Application of deep learning model to estimate PSDM of a frame based on its pushover curve

5. CONCLUSIONS

This study presented an ML-assisted framework to estimate the hazard and environmental performance of building design alternatives as part of a risk-based early design. The framework consists of four main workflows: (1) a relational open performance data inventory that aggregates prior performance-based assessments and data, (2) a supervised training workflow that develops ML models mapping performance metrics to imprecise descriptors of building design and geometry, (3) a sequential workflow that applies surrogate models with lower fidelity at earlier stages, and (4) a deep-learning based engine that provides a fast means to map the result of simpler static analysis to seismic performance metrics. A preliminary application of the framework to a building inventory in Charleston, SC, resulted in the following findings:

1. ML models could demarcate the seismic hazard performance range (in terms of repair cost) of a building taxonomy at a given site using only imprecise geometry and design information.
2. The best ML models (XGB and SVM) show smaller bias and variance to predict collapse loss and larger bias and variance to predict repair cost due to damage to non-structural acceleration-sensitive assemblies.
3. Floor area, total weight, and height show a relatively higher impact on the XGB prediction of the total repair cost.
4. It is feasible to map the result of simpler static analysis to probabilistic demand models through an encoder-decoder deep learning architecture. Such capability allows to rapidly map the results of parametric studies to detailed assessments suitable at later stages of design.

6. REFERENCES

Alvarez, A., Meacham, B., Dembsey, N., and Thomas, J., 2013. Twenty years of performance-based fire protection design: challenges faced and a look ahead. *Journal of Fire Protection Engineering*, **23**(4), 249–276. SAGE Publications Ltd STM. DOI: 10.1177/1042391513484911

Bertero, R. D., and Bertero, V. V., 2002. Performance-based seismic engineering: the need for a reliable conceptual comprehensive approach. *Earthquake Engineering & Structural Dynamics*, **31**(3), 627–652. DOI: 10.1002/eqe.146

Bueno, E., and Turkienicz, B., 2014. Supporting tools for early stages of architectural design. *International Journal of Architectural Computing*, **12**(4), 495–512. SAGE Publications Sage UK: London, England.

Chapman, M. C., and Talwani, P., 2006. *Seismic Hazard Mapping for Bridge and Highway Design in South Carolina*.

Cheung, F. K. T., Rihan, J., Tah, J., Duce, D., and Kurul, E., 2012. Early stage multi-level cost estimation for schematic BIM models. *Automation in Construction*, **27**, 67–77. DOI: 10.1016/j.autcon.2012.05.008

Ciampoli, M., Petrini, F., and Augusti, G., 2011. Performance-Based Wind Engineering: Towards a general procedure. *Structural Safety*, **33**(6), 367–378. DOI: 10.1016/j.strusafe.2011.07.001

Haselton, C. B., Liel, A. B., Lange, S. T., and Deierlein, G. G., 2008. *Beam-Column Element Model Calibrated for Predicting Flexural Response Leading to Global Collapse of RC Frame Buildings* (p. 152). PEER.

LLatas, C., Soust-Verdaguer, B., Hollberg, A., Palumbo, E., and Quiñones, R., 2022. BIM-based LCSA application in early design stages using IFC. *Automation in Construction*, **138**, 104259. DOI: 10.1016/j.autcon.2022.104259

McKenna, F., 2011. OpenSees: A Framework for Earthquake Engineering Simulation. *Computing in Science and Engineering*, **13**(4), 58–66. DOI: 10.1109/MCSE.2011.66

Mehta, M., Scarborough, W., and Armpriest, D., 2017. *Building Construction: Principles, Materials, and Systems* (3rd edition.). NY, NY: Pearson.

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., and Ajayi, S., 2022. Machine learning for energy performance prediction at the design stage of buildings. *Energy for Sustainable Development*, **66**, 12–25. DOI: 10.1016/j.esd.2021.11.002

Omoya, M., Ero, I., Zaker Esteghamati, M., Burton, H. V., Brandenberg, S., Sun, H., Yi, Z., et al., 2022. A relational database to support post-earthquake building damage and recovery assessment. *Earthquake Spectra*, **38**(2), 1549–1569. SAGE Publications Ltd STM. DOI: 10.1177/87552930211061167

Østergård, T., Jensen, R. L., and Maagaard, S. E., 2016. Building simulations supporting decision making in early design – A review. *Renewable and Sustainable Energy Reviews*, **61**, 187–201. DOI: 10.1016/j.rser.2016.03.045

Ramirez, C. M., Liel, A. B., Mitrani-Reiser, J., Haselton, C. B., Spear, A. D., Steiner, J., Deierlein, G. G., et al., 2012. Expected earthquake damage and repair costs in reinforced concrete frame buildings. *Earthquake Engineering & Structural Dynamics*, **41**(11), 1455–1475. DOI: 10.1002/eqe.2216

Rezaee, R., Brown, J., Augenbroe, G., and Kim, J., 2015. Assessment of uncertainty and confidence in building design exploration. *AI EDAM*, **29**(4), 429–441. Cambridge University Press. DOI: 10.1017/S0890060415000426

Rezaee, R., Brown, J., Haymaker, J., and Augenbroe, G., 2019. A new approach to performance-based building design exploration using linear inverse modeling. *Journal of Building Performance Simulation*, **12**(3), 246–271. Taylor & Francis. DOI: 10.1080/19401493.2018.1507046

Schneider-Marin, P., Stocker, T., Abele, O., Margesin, M., Staudt, J., Abualdenien, J., and Lang, W., 2022. EarlyData knowledge base for material decisions in building design. *Advanced Engineering Informatics*, **54**, 101769. DOI: 10.1016/j.aei.2022.101769

Singh, M. M., Deb, C., and Geyer, P., 2022. Early-stage design support combining machine learning and building information modelling. *Automation in Construction*, **136**, 104147. DOI: 10.1016/j.autcon.2022.104147

Smith, A., and Smith, K. S., 2014. *Developing Your Design Process: Six Key Concepts for Studio* (1st edition.). New York: Routledge.

Soleimani-Babakamali, M. H., and Zaker Esteghamati, M., 2022. Estimating seismic demand models of a building inventory from nonlinear static analysis using deep learning methods. *Engineering Structures*, **266**, 114576. DOI: 10.1016/j.engstruct.2022.114576

Sun, H., Burton, H. V., and Huang, H., 2021. Machine learning applications for building structural design and performance assessment: State-of-the-art review. *Journal of Building Engineering*, **33**, 101816. DOI: 10.1016/j.jobe.2020.101816

Zabalza Bribián, I., Valero Capilla, A., and Aranda Usón, A., 2011. Life cycle assessment of building materials: Comparative analysis of energy and environmental impacts and evaluation of the eco-efficiency improvement potential. *Building and Environment*, **46**(5), 1133–1140. DOI: 10.1016/j.buildenv.2010.12.002

Zaker Esteghamati, M., and Flint, M. M., 2021. Developing data-driven surrogate models for holistic performance-based assessment of mid-rise RC frame buildings at early design. *Engineering Structures*, **245**, 112971. DOI: 10.1016/j.engstruct.2021.112971

Zaker Esteghamati, M., Lee, J., Musetich, M., and Flint, M. M., 2020. INSSEPT: An open-source relational database of seismic performance estimation to aid with early design of buildings. *Earthquake Spectra*, **36**(4), 2177–2197. SAGE Publications Ltd STM. DOI: 10.1177/8755293020919857

Zaker Esteghamati, M., Sharifnia, H., Ton, D., Asiatico, P., Reichard, G., and Flint, M. M., 2022. Sustainable early design exploration of mid-rise office buildings with different subsystems using comparative life cycle assessment. *Journal of Building Engineering*, **48**, 104004. DOI: 10.1016/j.jobe.2022.104004