Sustainability recommendation system for process-oriented building design alternatives under multi-objective scenarios

Chen X., Geyer P.
Institute of Sustainable Building Systems, Leibniz University Hannover, Germany
xia.chen@iek.uni-hannover.de

Abstract. Nowadays, sustainability objective has risen to the most attention in building engineering scenarios. Multi-objective optimization techniques can act as assistance in supporting decision-making in a trade-off of various considerations in an interdisciplinary manner. In this study, we propose a recommendation system to alleviate the difficulty of informed decision-making regarding the rapid potential design space exploration, optimal design solution analysis, and dynamic interaction aligned with ongoing processes. To illustrate how the recommendation system is organized to help designers or engineers approach the general sustainability objective, an early design phase case study based on a real-world, massive energy performance certification dataset is conducted. The generated results conform to interpretations based on domain knowledge, which validate the effectiveness of the system assistance.

1. Introduction

The concept of sustainability within the architecture, engineering, and construction (AEC) domain is inherently complex. It typically incorporates energy performance, environmental impacts, and life cycle costs as interconnected considerations (Gervásio et al., 2014), which naturally composes a multi-objectives scenario. The need to acquire instant, robust, and precise assessments of such indicators in the domain has boosted the development and adaptation of various first-principles methods, as well as data-driven approaches (machine learning, ML) in the recent decade (Kheiri, 2018; Westermann and Evins, 2019). along with the raising interest in Building Sustainability Assessment Systems (BSAS) (Lazar and Chithra, 2020).

Although various sustainability assessment tools exist in the ACE domain (Kumar *et al.*, 2017; Tan *et al.*, 2021), to our best knowledge, three critical characteristics are missing to adapt to current challenges: First, most assessment tools aim to solve multi-criteria decision making (MCDM) problem towards the process weighting, rather than evaluating potential design options, patterns, and consequences. These tools are limited by their dependence on a deterministic set of concise inputs that rely heavily on designers' prior knowledge. Thus, a crucial element is missing: the dynamic potential for design space exploration (DSE) (Østergård, Jensen and Maagaard, 2017) integrated into the design process; Second, current methods are not sufficiently equipped to provide assistance throughout various building development levels (BDLs) (Abualdenien *et al.*, 2020). Such as recommendations as interactive assistance are required to consider qualitative and implicit aspects that are difficult to formalize (Geyer, 2009); Finally, many of these tools are primarily based on pure knowledge-based processes or first-principles simulations. These tools own the computational bottleneck of conducting an exhaustive search in the potential design space to identify optimal solutions.

In this study, we propose a recommendation system for sustainable building design as part of a machine assistance framework. This system recommends alternative optimal solutions considering assumptions and constraints of the design process, enabling a process-oriented, dynamic interactive manner as a dynamic DSE system. By exploring the potential patterns based on optimized results, the generated alternatives assist users' decision-making process in building design and engineering scenarios.

2. Methodology

2.1 Machine Assistance

The sustainability recommendation system extends our previous research: a data-driven, process-based machine assistance framework for decision-making support in energy-efficient building design scenarios (Chen and Geyer, 2022), which consists of three parts: probabilistic surrogate modeling (prediction), ensemble modeling (estimation), and the model interpretation method (inference/ intervention), which gives the framework several unique characteristics:

- Induction under uncertainties: Output distribution evaluation under incomplete inputs with their inherent uncertainties by combining probabilistic surrogate modeling with the ensemble mechanism. In our case, we choose *NGBoost* (Duan *et al.*, 2019);
- Inference: Analyze possible input assumptions' consequences as representative of the potential output value space in the dynamic interactive process by embedding *SHAP* interpretation method (Lundberg and Lee, 2017).
- Feedback loop with consistency: The process shares parametric input representation with different target outputs, ensuring the consistency of the result interpretation. The process is also a feedback loop for building designers to explore potential design space, receive dynamic information, and infer toward lower energy consumption.

Apart from the characteristic mentioned above, the machine assistance framework gives the foundation for aligning sustainability objectives during the design process. In this study, we intend to take a step forward by proposing a sustainability recommendation system that extends the framework with an evolutionary algorithm and clustering result to generate reproducible multi-objective optimized designs.

2.2 Sustainability Recommendation System

The recommendation system consists of five steps with a feedback loop that assist users in conducting informed decision-making for sustainable design at different BDLs:

1. **Objectives setting**: With the updated condition of the design scheme, objectives (Output) selection or scenario (Inputs) adjustment (Deb, 2011) is set by the user based on design conditions, prior knowledge, or extra information feed-in.
2. **Information collection**: The updated objectives and design scheme condition (settings and constraints of the present BDL) are fed to machine assistance, making estimations with model interpretation to update output distribution for each objective, and determine the potential design space.
3. **Optimization**: The information is formalized to an optimizable problem; in this study, the genetic algorithm (GA), *NSGA-II* (Deb *et al.*, 2002) is applied to generate a set of well-performing non-dominated solutions. This algorithm was chosen because it exhibits high robustness, an ability to deal with heterogeneous variables, and no need of weighing a priori. This step delivers the optimal Pareto front of the present BDL's potential design schemes.
4. **Analysis:** Well-performing design solutions are fed into unsupervised clustering to identify common characteristics and patterns; In this study, *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* (Ester *et al.*, 1996; Schubert *et al.*, 2017) is chosen. Clustering results serve to deliver robust configurations against the GA generation randomness.

5. **Assistance**: The analysis results with alternative potential design recommendations are fed back to the designer. This information aids in informed decision-making and allows for necessary adjustments, which in turn update the recommendation system's outputs. This mutual information synchronization pattern gives the dynamic momentum to maximize the expected performance of the design toward sustainability objectives. Hence the system acts as an assistant for sustainable design.

A conceptual illustration is presented in Figure 1.

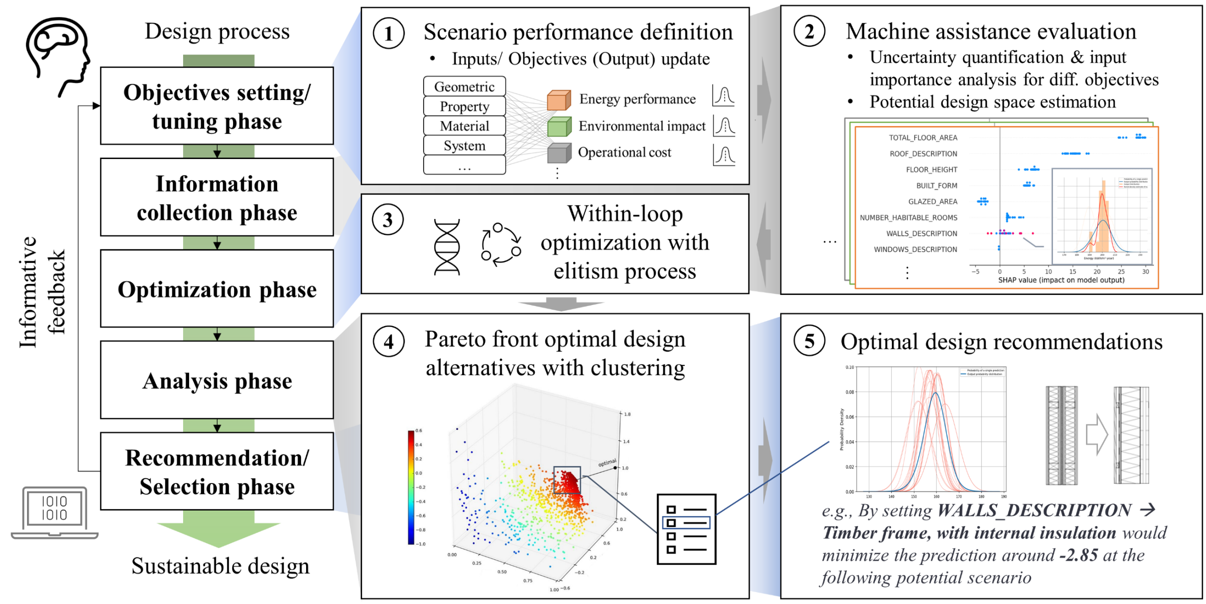


Figure 1: The process illustration of the sustainability recommendation system.

The recommendation system is designed to reveal the following new characteristics that are essential for process-based assistance:

- **Efficient data usage from the real-world and simulation data:** The ML surrogate model owns a solid potential to capture implicit input-output patterns behind the data. It allows the data from real-world collection and synthetic simulation to be fed into the model simultaneously to cover large-scale building cases.
- **Flexibility in applications of building engineering assessments:** Depending on the training inputs definition and objective settings, the recommendation system is suitable to be adapted and applied to building engineering evaluation across the complete life cycle phase (design, construction, operation, retrofitting, etc.)
- **Rapid feedback for process assistance and interaction:** The ML surrogate model is equivalent to encapsulating the corresponding fast feedback function based on the set objective, combined with GA providing multi-objective optimization. This combination removes major repetitive efforts of potential design space exploration and first-principles simulation validation process, making in-time optimal solutions during the design process possible.

2.3 Evaluation Metrics

To facilitate the surrogate modelling performance comparison regardless of the numerical scale of the result in different objectives, the three metrics commonly used in regression task evaluations are selected: Normalized Root Mean Square Error (NRMSE), Symmetric Mean Absolute Percentage Error (sMAPE), and Coefficient of determination (R-squared or R^2). Their mechanism detail and the consideration of metrics selection are referred to in this paper (Chicco, Warrens and Jurman, 2021).

3. Case Study

In the case study, we simulate a scenario in the early building design phase in which the building type, location, and area range are defined; however, precise façade geometry, material, and energy system configuration are unknown.

3.1 Data Description and Pre-processing

To demonstrate a typical multi-objectives optimization case, we selected a scenario in the building's early design phase with the same open data sources used in our previous machine assistance research (Chen and Geyer, 2022): Energy Performance of Buildings Data: England and Wales (epc.opendatacommunities.org, 2020), which is published and maintained by the Ministry of Housing, Communities & Local Government from the UK every half-yearly. The dataset contains dwellings' detail across most UK regions and connects to the *domestic EPC (Energy performance certificate).* The reasons for selecting this data are as follows:

- **Real-world massive dataset with expertise validation**: The data is collected under the EU Directive requirements on the energy performance of buildings. The robustness of the data in relation to buildings is guaranteed by the energy assessor carried out the accreditation scheme based on *Standard Assessment Procedure (SAP) for new dwellings and Reduced SAP* (RdSAP) in the UK (gov.uk, 2012). Each building data corresponds to a certain real-world building with trackable information for validation purposes.
- **Target input/output available**: This EPC dataset is in a fine data condition and contains the necessary information for supporting building early design phase analysis: features in building geometry, component characteristics, and energy systems. Apart from the energy performance data, the dataset also includes each building's environmental impact and cost data.

The dataset contains 19,725,379 building records with various building types and built forms. We applied the same data cleaning process as in machine assistance research to remove the semantic noise and missing data. To specify a design case in this study, we set a scenario to filter and select the sub-data: *a flat, detached building* with records shows built after the year 2007 between 150-250 m^2. Eventually, 7,566 real-world building records remain.

3.2 Inputs/Objectives Definition

Next, we set objectives (i.e., output) based on the given dataset: three indicators are chosen and modified in an annual sum per square meter behaviour: *Energy Consumption* in kWh/m^2/year, *environmental impact* by *CO2 Emission* equivalent in kg/m^2/year, and the *Operational Cost* in £/m^2/year.

For the input parameters, ten features in three major categories are selected as building early design phase parametric representatives; they are: **Geometry**: *Total Floor Area, Floor Height, Building Glazed Area,* and *Number of Heated Rooms*; **Component material property**: *Descriptions of Windows, Walls,* and *Roof;* **Energy system**: *Descriptions of Main Heating Systems, Secondary Heating Systems,* and *Building Ventilation Type.*

In this input feature set from EPC data, only *Total Floor Area, Floor Height,* and *Number of Heated Rooms* are numerical parameters; the rest of the features are composed of semantics descriptions. To ensure the models' performance, we implemented label-encoding on these semantic features into categorical numbers instead of using one-hot encoding to prevent the curse of dimensionality by high-dimensional feature spaces.

Detailed input & output descriptions, ranges, and data types are shown in Table 2. Table 3 presents the labelled encoded semantic categories of input features. Both tables are available in Appendix.

3.3 Surrogate Modelling and Machine Assistance

Once the input features and objectives were determined at one BDL, we fed the data into the surrogate modelling, training corresponding models with a hyperparameter grid-search strategy and 5-fold cross-validation (Refaeilzadeh, Tang and Liu, 2009). We point to our previous study for a detailed tuning of surrogate modeling and machine assistance implementation description (Chen and Geyer, 2022). The result is presented in Table 1. Given the fact that the data is collected from real-world and only ten building parameters representing the early design process as model inputs, all models exhibited a promising performance (sMAPEs are around 10, or 90% accuracy), in which energy consumption prediction being the most accurate, and operational cost prediction being the least.

Table 1: Accuracy result of surrogate models.

Model/Objective	NRMSE	sMAPE	R^2
Energy consumption	8.08	8.78	0.86
CO_2 Emission	5.49	9.35	0.82
Operational Cost	8.45	10.35	0.77

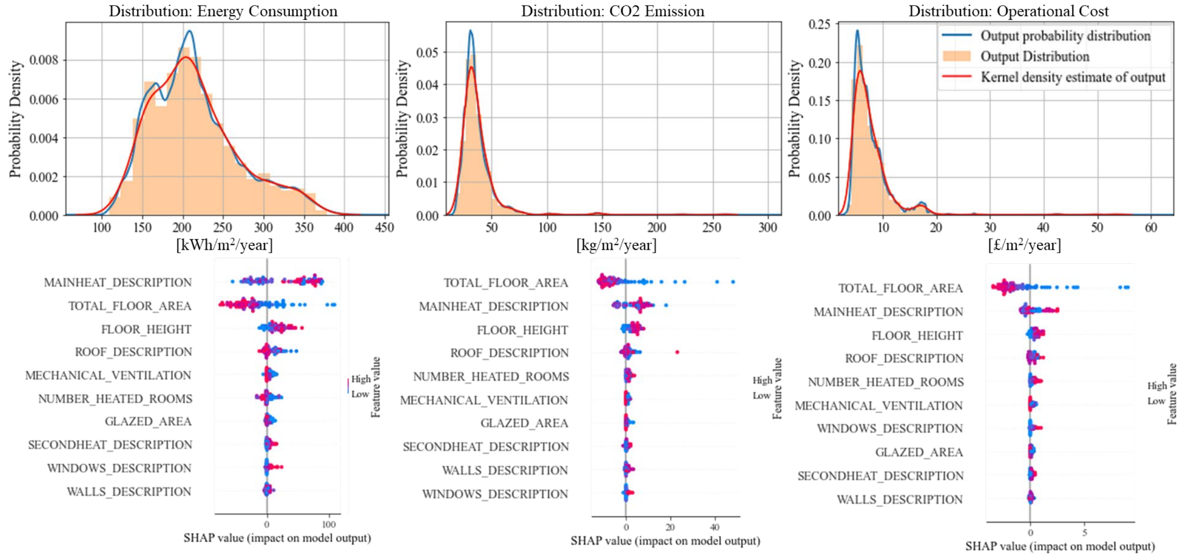
Next, surrogate modelling combined with machine assistance evaluation (Step 2 in Figure 1) gives the estimation result for three set objectives, as illustrated in Figure 2. The estimation results well describe the potential design space within the ranges of input data: For energy consumption, machine assistance estimated the output range between 109.5 and 378.6 kWh/m^2/year, with the top three critical features ranked as main heating system, total floor area, and floor height; For CO2 emission and operational cost, the estimated result shows from 16.3 to 260.3 kg/m^2/year, and 3.2 to 53.6£/m^2/year in a long tail distribution, respectively, with the same top three critical feature listed as total floor area first, then main heating system, and floor height. Besides the result distributions, some primitive information is observable, e.g., For a flat building, a bigger total floor area corresponds to lower energy consumption, CO2 emission, and operational cost in annual average per square meter, while the changes of floor height show opposite trends.

3.4 Pareto Front, Clustering Analysis, and Recommendations

After the machine assistance gives information about the result ranges for all objectives, NSGA-II is then applied with trained surrogated models to find a set of Pareto-optimal solutions in an iterative elitism process. In this test case, we set the problem as minimizing all three objectives and run the GA by the set input ranges with a 1000 population size in 100 generations. Once the Pareto front is determined, we applied DBSCAN for input clustering, and colored outputs with the clustering result. A 3D scatter projection plot is presented in Figure 3.

The axis x, y, and z in the 3d-scatter plot correspond to the energy consumption, CO_2 emission, and operational cost, respectively. Some insightful conclusions are summarized and listed below:

- **The effectiveness of the machine assistance information and GA**: The output ranges of all optimal sample results correctly correspond to the estimation results generated from the machine assistance. All generated sample results from Figure 3 correspond to the minimum end of the objective estimation range in Figure 2.



Input uncertainties	[kWh/m²/year]
Output range	(109.5, 378.58)
mean	214.88
median	207.33
Prob. distribution	
68.0% confidence	(161.24, 266.98)
95.0% confidence	(130.75, 342.62)
99.7% confidence	(110.29, 440.33)

Input uncertainties	[kg/m²/year]
Output range	(16.3, 260.29)
mean	36.44
median	33.24
Prob. distribution	
68.0% confidence	(25.82, 44.11)
95.0% confidence	(20.67, 57.38)
99.7% confidence	(17.31, 74.62)

Input uncertainties	[£/m²/year]
Output range	(3.2, 53.55)
mean	7.58
median	6.68
Prob. distribution	
68.0% confidence	(4.95, 9.72)
95.0% confidence	(3.88, 13.6)
99.7% confidence	(3.34, 18.89)

Figure 2: Estimation result of three objectives within a given potential design space derivative from machine assistance (Chen and Geyer, 2022) . Three columns from left to right present information with regard to Energy consumption, CO2 emission, and Operational cost, individually, while three plots/tables from top to bottom illustrate output distribution, feature importance, and uncertainty estimation, respectively. The feature importance plot is generated by SHAP (Lundberg and Lee, 2017) ; SHAP value samples in each feature row from high to low are marked from red to blue. All semantic features are label encoded; the dictionary is available in Table 3.

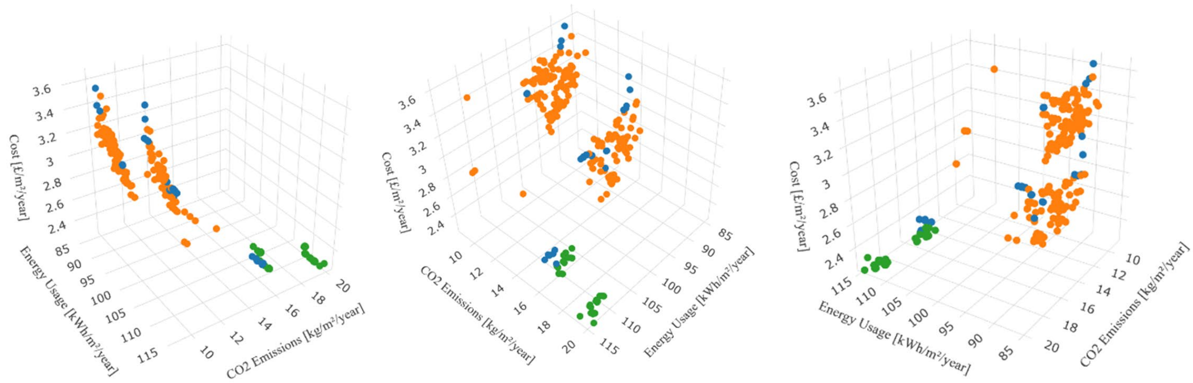


Figure 3: 3d-Scatter plot of Pareto front of building design case in a trade-off between energy performance, environmental impact, and cost, presented in two perspectives. Each scatter dot means a result based on a single optimized design parameter combination, colored by the clustering result from the DBSCAN algorithm learning from design parameter data.

- **The trade-off between objectives is needed:** The defined problem is to minimize all three objectives; however, we noticed that the normal direction of the generated Pareto front point to the global minimum, which means that the trade-off consideration between energy consumption, environmental impact, and operational cost is required in this building design case.

- **Design patterns exist in this sustainable building design case:** We observed a clear grouping behaviour from the input clustering results (orange, blue, and green). Three clusters are identified in the optimal samples: The orange cluster represents the lowest energy consumption, with a steep trade-off between low environmental impact and low operational cost; The green cluster shows differently, with energy consumption and CO_2 emission reaching relatively high positions and the cost staying at low points; The blue cluster plays in a more balanced manner compared to others.

To further investigate the design commonality in these clusters, we use parallel coordinates plots to compare clusters and examine their feature combination patterns, as presented in Figure 4. In our case context, the sustainable design of a detached flat building, the parallel coordinates plot shows clear patterns in optimal design clusters (recommendations) as follows:

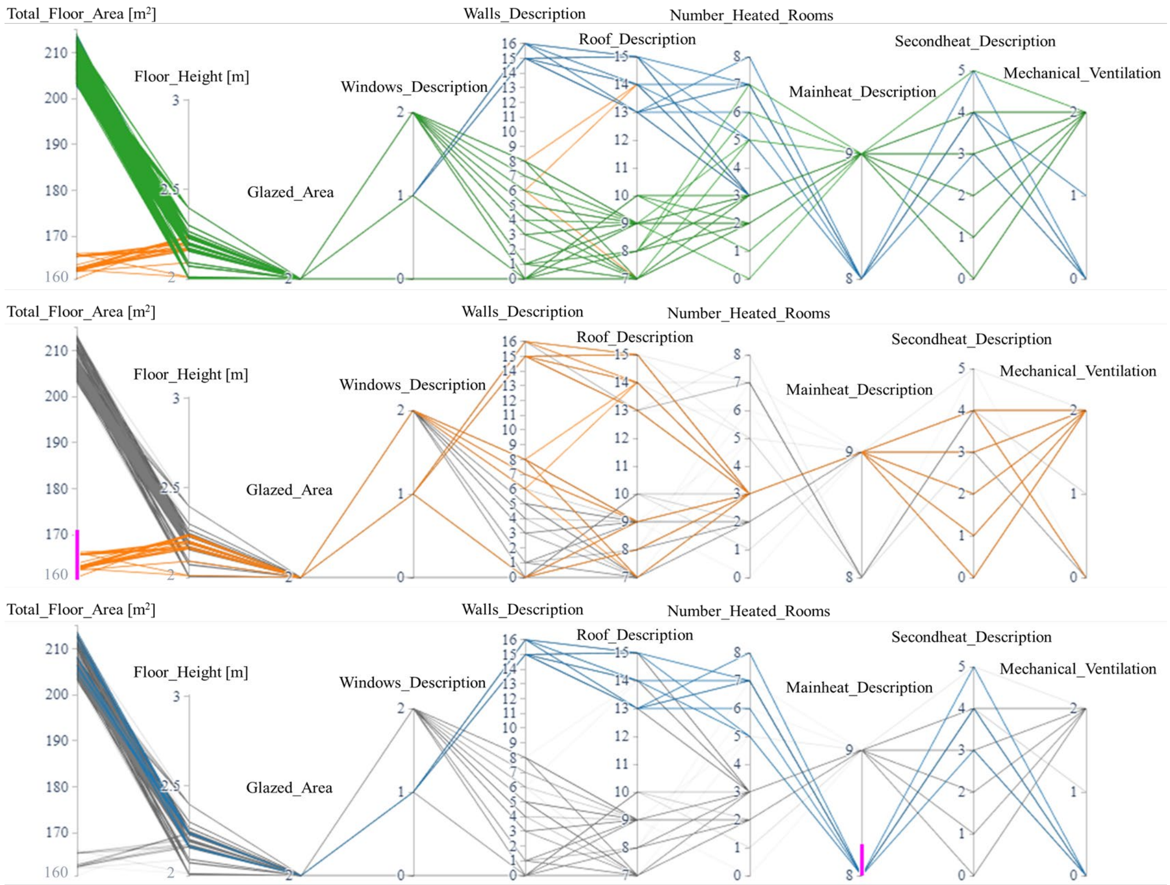


Figure 4: Parallel coordinates plot of optimized design recommendations. Each coordinate represents one input feature with possible values range in a different scale. Features with semantic options are the same label encoded as in Table 3. Each line in the plot stands for a sample. From top to bottom, the first plot shows all three clusters with each sample choice in input features. The colour palette remains the same as in Figure 3. In the second and third rows, only one cluster is coloured to show the cluster options clearly.

- **General patterns:** The generated optimal samples are grouped into two major floor area ranges, around 165 m^2 and 210 m^2. Meanwhile, they have relatively low floor height (around 2.3 m), normal glazed area (10%-20% based on RdSAP), and triple/double glazing windows. The rest of the features are varied by design combinations except the main heating system: only two systems are chosen in optimal designs, community scheme with combined heat and power, or with mains gas.

- **Green cluster:** This cluster has a floor area of around 210 m^2; The wall is well insulated, composed of cavity wall, granite, whinstone, or sandstone; The roof type is pitched with insulation; The main heating system is the community scheme with combined heat and power, and use only nature ventilation in the building.
- **Orange cluster:** This cluster has a smaller three-room-heated building design with an average floor area of around 165 m^2 and well-insulated timber frame walls. These designs have an insulated thatching roof or roof room(s) with an insulated ceiling. The main heating system is the community scheme with main gas, and mechanical ventilation for extract.
- **Blue cluster:** This cluster has a similar floor area range as in the other two clusters with fully triple-glazed windows, timber frame walls, and roof room(s) with an insulated ceiling or thatching roof. These designs have more heated rooms (5-8 rooms) with a heating system of combined heat and power community scheme and natural ventilation.

In fact, these three clusters and the general patterns provide primitive but insightful information as strategies to assist decision-making in the early design phase. In a context of real-world scenario, feeding these recommendations to the designer or engineer helps them narrow down the design variations, and constantly validate their design performance compared with optimal ones to formalize an informative feedback loop. In fact, this feedback loop, corresponding to step 2 to step 5 in Section 2.2 (illustrations from Figure 2 → Figure 3 → Figure 4), creates a dynamic pattern of generating optimal Pareto front based on the growing BDLs. With the new design parameters fixed by designers, the Pareto front updates accordingly and continues the loop in an approaching manner for both ends to meet each other eventually: the ongoing design, and the sustainable objectives.

4. Discussion & Conclusion

In this paper, we construct a sustainability recommendation system that provides an interactive pattern to identify optimal solutions with clusters for a specific design situation. The proposed system enables rapid, informed decision-making aids toward the process in dynamic behaviour throughout potential design space, which is defined from the Building Development Level (BDL) with its set of variables.

Essentially, this system explained and proved only a straightforward mindset: Using MLs to learn and map implicit relationships between architectural design and physical characteristics, while the evolutionary methods are used to eliminate the time and resources wasted in the exhaustive search for optimal solutions. However, the way of representing building designs is not limited to parametric models, as shown in this study. With the charging development of multimodal machine learning (MML) and large-scale language models (LLM), the same mindset can be seamlessly adapted to these models: e.g., using MML to capture information from natural language description and generating corresponding design prototypes, parameterization via the MML and optimized with GA, and feedback in the form of language, image, or other design representations. It contains the potential to cause an impact that reshapes the ACE industry.

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

Table 2 Input & Output features

Feature	Category	Description	Data type	Range
Total floor area	Geometry	Total Useful Floor Area (m^2)	float	[10, 230]
Floor height	Geometry	Average height of the storey in meters	float	[2, 4.2]
Glazed area	Geometry	Ranged estimate of the total glazed area of the Habitable Area.	category	3
Number heated rooms	Geometry	The number of heated rooms in the property.	int	[1, 9]
Windows description	Component characteristics	Overall description of the property feature	category	5
Walls description	Component characteristics	Overall description of the property feature	category	18
Roof Description	Component characteristics	Overall description of the property feature	category	19
Mainheat. description	Energy system	Overall description of the property feature	category	21
Secondheat. description	Energy system	Overall description of the property feature	category	8
Mechanical ventilation	Energy system	Identifies the type of mechanical ventilation the property has.	category	3
Energy consumption current per m^2	Output	Current estimated total energy consumption for the property per year (kWh/m^2).	float	[65, 392]
CO_2 emissions current per m^2	Output	CO_2 emissions per square meter floor area per year in kg/m^2	float	[1.76, 346.75]
Cost operation current per m^2	Output	Current estimated annual energy costs for heating, hot water, and lighting per year in £/m^2	float	[2.84, 64.94]

Table 3: Dictionary of labelled feature.

Feature	Labelled code
Glazed area	[Less Than Typical (less than 10%): 0, More Than Typical (more than 20%): 1, Normal: 2]
Windows description	[Fully double glazing: 0, Fully triple glazing: 1, Mostly double glazing: 2, Partial double glazing: 3, Single glazing: 4]
Walls description	[Cavity wall, insulated: 0, Cavity wall, filled cavity: 1, Cavity wall, ei.: 2, Cavity wall, ii.: 3, Granite or whinstone, insulated: 4, Granite or whinstone, ei.: 5, Granite or whinstone, ii.: 6, Sandstone, insulated: 7, Sandstone, ii.: 8, Solid brick, insulated: 9, Solid brick, no insulation: 10, Solid brick, ei.: 11, Solid brick, ii.: 12, System built, insulated: 13, System built, ei.: 14, System built, ii.: 15, Timber frame, insulated: 16, Timber frame, ii.: 17]
Roof description	[Flat: 0, Flat insulated: 1, Pitched: 2, Pitched 100mm li.: 3, Pitched 12mm li.: 4, Pitched 150mm li.: 5, Pitched 200mm li.: 6, Pitched 250mm li.: 7, Pitched 270mm li.: 8, Pitched 300+mm li.: 9, Pitched 300mm li.: 10, Pitched 50mm li.: 11, Pitched 75mm li.: 12, Pitched insulated: 13, Pitched insulated at rafters: 14, Roof room(s) ceiling insulated: 15, Roof room(s) insulated: 16, Thatched: 17, Thatched with additional insulation: 18]
Mainheat. description	[Air source heat pump, radiators, electric: 0, Boiler and radiators, LPG: 1, Boiler and radiators, electric: 2, Boiler and radiators, mains gas: 3, Boiler and radiators, oil: 4, Boiler and underfloor heating, LPG: 5, Boiler and underfloor heating, electric: 6, Boiler and underfloor heating, mains gas: 7, Community scheme: 8, Community scheme with CHP: 9, Community scheme, mains gas: 10, Electric ceiling heating: 11, Electric storage heaters: 12, Electric underfloor heating: 13, Ground source heat pump, radiators, electric: 14, Ground source heat pump, underfloor, electric: 15, No system present: electric heating assumed: 16, Portable electric heating assumed for most rooms: 17, Room heaters, electric: 18, Warm air, electric: 19, Warm air, mains gas: 20]
Secondheat. description	[None: 0, Portable electric heaters: 1, Room heaters, coal: 2, Room heaters, dual fuel (mineral and wood): 3, Room heaters, electric: 4, Room heaters, mains gas: 5, Room heaters, smokeless fuel: 6, Room heaters, wood logs: 7]
Mechanical ventilation	[mechanical, extract only: 0, mechanical, supply and extract: 1, natural: 2]

ii. with internal insulation; ei. with external insulation; li. loft insulation

References

Abualdenien, J. *et al.* (2020) 'Consistent management and evaluation of building models in the early design stages', *Journal of Information Technology in Construction*, 25, pp. 212–232.

Chen, X. and Geyer, P. (2022) 'Machine assistance in energy-efficient building design: A predictive framework toward dynamic interaction with human decision-making under uncertainty', *Applied Energy*, 307, p. 118240.

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

Deb, K. *et al.* (2002) 'A fast and elitist multiobjective genetic algorithm: NSGA-II', *IEEE transactions on evolutionary computation*, 6(2), pp. 182–197.

Deb, K. (2011) 'Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction', in *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing:* Springer, London, pp. 3–34. Available at: https://link.springer.com/chapter/10.1007/978-0-85729-652-8_1.

Duan, T. *et al.* (2019) *NGBoost: Natural Gradient Boosting for Probabilistic Prediction*. Available at: http://arxiv.org/pdf/1910.03225v4.

https://epc.opendatacommunities.org/ (2020), 17 November (Accessed: 17 November 2020).

Ester, M. *et al.* (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise', *kdd*, pp. 226–231.

Gervásio, H. *et al.* (2014) 'A macro-component approach for the assessment of building sustainability in early stages of design', *Building and Environment*, 73, pp. 256–270. doi: 10.1016/j.buildenv.2013.12.015

Geyer, P. (2009) 'Component-oriented decomposition for multidisciplinary design optimization in building design', *Advanced Engineering Informatics*, 23(1), pp. 12–31. doi: 10.1016/j.aei.2008.06.008

gov.uk (2012) 'Standard Assessment Procedure (SAP) for new dwellings and Reduced SAP', 2012. Available at: https://www.gov.uk/guidance/standard-assessment-procedure.

Kheiri, F. (2018) 'A review on optimization methods applied in energy-efficient building geometry and envelope design', *Renewable and Sustainable Energy Reviews*, 92, pp. 897–920. doi: 10.1016/j.rser.2018.04.080

Kumar, A. *et al.* (2017) 'A review of multi criteria decision making (MCDM) towards sustainable renewable energy development', *Renewable and Sustainable Energy Reviews*, 69, pp. 596–609.

Lazar, N. and Chithra, K. (2020) 'A comprehensive literature review on development of Building Sustainability Assessment Systems', *Journal of Building Engineering*, 32, p. 101450.

Lundberg, S.M. and Lee, S.-I. (2017) 'A Unified Approach to Interpreting Model Predictions', *Advances in Neural Information Processing Systems:* Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Østergård, T., Jensen, R.L. and Maagaard, S.E. (2017) 'Early Building Design: Informed decision-making by exploring multidimensional design space using sensitivity analysis', *Energy and buildings*, 142, pp. 8–22.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009) 'Cross-validation', *Encyclopedia of database systems*, 5, pp. 532–538.

Schubert, E. *et al.* (2017) 'DBSCAN revisited, revisited: why and how you should (still) use DBSCAN', *ACM Transactions on Database Systems (TODS)*, 42(3), pp. 1–21.

Tan, T. *et al.* (2021) 'Combining multi-criteria decision making (MCDM) methods with building information modelling (BIM): A review', *Automation in Construction*, 121, p. 103451.

Westermann, P. and Evins, R. (2019) 'Surrogate modelling for sustainable building design – A review', *Energy and buildings*, 198, pp. 170–186. doi: 10.1016/j.enbuild.2019.05.057