Article:
Abdalla, T. and Peng, C. (2026) Interpretable machine learning for occupant-specific PM2.5 exposure assessment in higher education buildings. Journal of Building Engineering, 121. 115632. ISSN: 2352-7102

https://doi.org/10.1016/j.jobe.2026.115632

Interpretable Machine Learning for Occupant-Specific $PM_{2.5}$ Exposure Assessment in Higher Education Buildings

Abstract

Outdoor-origin fine particulate matter ($PM_{2.5}$) poses significant health risks in Higher Education Institution (HEI) buildings, where occupants spend extended periods across diverse functional spaces. This study develops a scalable framework for indoor $PM_{2.5}$ exposure assessment by coupling CONTAM–EnergyPlus co-simulation with machine learning metamodels and SHapley Additive exPlanations (SHAP)-based interpretability. An Extreme Gradient Boosting (XGBoost) metamodel trained on 2,729 zones across five UK HEI buildings achieved high predictive accuracy ($R \approx 0.95$; $R^2 > 0.90$ on held-out data). SHAP analysis, representing what appears to be the first such application in HEI indoor air quality assessment using a physics-driven metamodel, identified building airtightness (Q_{50}) as the dominant exposure driver, followed by infiltration air change rate (ACH_{INF}) and indoor–outdoor temperature difference (ΔT). Microenvironmental modelling indicated indicative pronounced exposure heterogeneity among occupant groups within the adopted assumption space: offices dominated staff exposure while educational facilities drove student exposure. Improving airtightness from baseline to Q_{50} = 3 $m^3/h{\cdot}m^2$ reduced population-weighted exposure by up to 32.3%; however, approximately 88% of zones still exceeded the WHO 2021 annual guideline of 5 $\mu g/m^3$. These findings demonstrate that envelope improvements alone are insufficient for WHO compliance and must be complemented by integrated mechanical ventilation and filtration strategies, alongside urban-scale policies such as Clean Air Zones and emissions control measures that reduce outdoor $PM_{2.5}$ at source. The framework provides a transparent, physics-grounded basis for screening HEI building stocks and prioritising evidence-based air quality interventions.

Keywords: Indoor air quality, $PM_{2.5}$ exposure assessment, Higher education buildings, Machine learning metamodels, Interpretable machine learning, Building airtightness

1. Introduction

Fine particulate matter comprising particles with aerodynamic diameters smaller than 2.5 microns ($PM_{2.5}$) poses significant health risks due to its ability to penetrate deep into the respiratory and cardiovascular systems, contributing to increased morbidity and premature mortality (COMEAP 2009; Apte et al., 2018; Burnett et al., 2018; Pope et al., 2019; Krittanawong et al., 2023). There is strong and growing evidence that long-term $PM_{2.5}$ exposure increases the risk of dementia, particularly Alzheimer's disease (Huang et al., 2025; Rogowski et al., 2025). The World Health Organization (WHO) recently updated its air quality guidelines, reaffirming the annual recommended $PM_{2.5}$ standard at 5 $\mu g/m^3$, a

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

value set in the 2021 revision and underscored in its 2025 air quality standards database release (WHO, 2025). The stricter guideline is considered necessary to better protect public health, as a substantial proportion of the global and urban population remains exposed to $PM_{2.5}$ concentrations above this threshold, amplifying risks especially among vulnerable groups (Im et al., 2025).

While outdoor $PM_{2.5}$ pollution has been extensively studied, indoor exposure is similarly critical since individuals spend approximately 85-90% of their time indoors (Pruszyński et al., 2023; Hancock 2018). Higher Education Institution (HEI) buildings represent complex indoor environments characterised by diverse spatial layouts, multiple functional zones, and heterogeneous occupant activity patterns (Gaidajis and Angelakoglou, 2009; Sarbu and Pacurar, 2015; Emmerich et al., 2019), presenting unique challenges for accurate indoor air quality (IAQ) assessment (Branco et al., 2024; Lama et al., 2022; Erlandson et al., 2019). Despite this importance, significant research gaps remain. First, in-depth IAQ studies for HEI buildings are sparse (Erlandson et al., 2019; Lama et al., 2022), and existing work rarely combines room-level exposure assessment with interpretable modelling of building-physics drivers (Sherman and Dickerhoff, 1998; Taylor et al., 2014 and 2015; Jones et al., 2015; Gillott et al., 2016) Second, while high-fidelity multizone airflow and thermal simulation tools such as the coupled CONTAM-EnergyPlus framework offer physically robust modelling capabilities (Feustel, 1999; Emmerich, 2001; Hensen and Lamberts, 2011; Dols and Polidoro, 2020), their computational intensity limits scalability across large building stocks. Third, machine learning approaches that could address this scalability challenge typically function as 'black boxes,' lacking the transparency needed for actionable policy guidance.

To address these gaps, this study develops an integrative framework that leverages the high-resolution CONTAM-EnergyPlus co-simulation outputs as training data for advanced machine learning metamodels. The methodology employs Extreme Gradient Boosting (XGBoost; Chen and Guestrin, 2016) integrated with SHapley Additive exPlanations (SHAP; Lundberg and Lee, 2017) to provide rapid, occupant-specific $PM_{2.5}$ exposure indicators across thousands of indoor zones. To the authors' knowledge, this study represents the first application of SHAP-based interpretability to indoor PM2.5 exposure assessment in Higher Education buildings using a physics-driven metamodel framework, directly overcoming the traditional 'Black-box' limitation of complex data-driven IAQ studies. By statistically quantifying the marginal contribution of key explanatory variables (e.g., building envelope airtightness Q_{50}, infiltration air change rate ACH_{INF}, and indoor-outdoor temperature difference, ΔT), this methodology transforms complex predictions into transparent and actionable insights, allowing findings to be directly translated into microenvironmental and population-weighted exposure metrics for distinct HEI cohorts.

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

In this study, the framework is applied to a representative university building stock to systematically quantifying occupant-specific exposure heterogeneity during the heating season. The findings demonstrate the utility of combining physics-driven simulations with interpretable machine learning for advancing environmental modelling of indoor air quality in complex institutional settings, thereby addressing current gaps in room-level, interpretable exposure assessment for HEI buildings. By providing zone-level, STG-specific exposure metrics and transparent attribution of key physical drivers, the approach supports facility managers and policymakers in developing evidence-based, targeted intervention strategies for healthier HEI environments globally.

2. Materials and methods

This section describes the integrative modelling framework developed to provide rapid, interpretable predictions of occupant-specific $PM_{2.5}$ exposure. The methodology begins by using high-resolution CONTAM–EnergyPlus co-simulation outputs from our previous physics-based study (Abdalla & Peng (2025) as the foundational input data for training machine learning metamodels. The focus here is on the scalable predictive modelling and exposure assessment components, including building stock characterisation, co-simulation data processing, metamodel development, and derivation of integrated personal and population-weighted exposure metrics.

2.1 Study setting and foundational input data

The main campuses of University of Sheffield are located in Sheffield city centre, a densely built urban area influenced by traffic-related and urban background air pollution sources. Outdoor $PM_{2.5}$ concentrations were characterised using data from the Sheffield Devonshire Green monitoring station (UKA00575), located approximately 500 m from the central campus, to provide contextually relevant ambient pollutant levels for infiltration modelling (**Figure 1**).

The study examines five HEI buildings selected from the University of Sheffield's building stock. The CONTAM–EnergyPlus co-simulation framework modelled a total of 2,729 thermal zones at room-level resolution, which includes occupied spaces as well as circulation areas, service spaces, and zone subdivisions required for accurate multizone airflow modelling. **Table 1** summarises the distribution of the 1,139 primary occupied space types (offices, educational facilities, shared facilities, circulation, and services) and their areas across all five buildings; Buildings 1–4 provided the training and testing data for the metamodel, while Building 5 was reserved as an independent hold-out for validation.

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Figure 1. Sheffield Devonshire Green Monitoring Station (UKA00575) and Weston Park Weather Station

Table 1. Area schedule of space types within the sampled HEI buildings (N = number of spaces)

Space Type	Building 1		Building 2		Building 3		Building 4		Building 5	
	N	Area (m^2)	N	Area (m^2)	N	Area (m^2)	N	Area (m^2)	N	Area (m^2)
Offices	22	451	19	292	127	5605	203	4119	6	244
Educational Facilities	0	0	7	424	54	4260	23	1897	14	903
Shared Facilities	6	46	11	165	78	2335	31	460	11	123
Circulation	12	145	25	368	199	2669	91	2218	18	576
Services	5	233	9	102	137	1534	25	364	6	101
Total	45	875	71	1351	595	16403	373	9058	55	1947

Data inputs, sourced from the University's Estates and Facilities Management (EFM) records, included architectural drawings, building use and occupancy policies, and envelope-related properties such as U-values. Building envelope airtightness (Q_{50}) was estimated using the UK CIBSE TM23 standard (CIBSE, 2022) and building construction year (**Table 2**), as direct blower-door test data for the non-domestic building stock were unavailable. This approach yielded building-level Q_{50} values ranging from 3 (well-sealed) to 13 $m^3/h{\cdot}m^2$ (leaky) which were assigned to each building and used to inform zone-level infiltration modelling parameters. The theoretical occupancy profiles applied to each space type in the co-simulation are detailed in **Table 5**, while **Table 6** presents the assumed time fractions for each Similar

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Time-Activity Group (STG) across microenvironments. As noted in **Section 2.2**, detailed mechanical ventilation systems (AHUs) were excluded from the models to isolate the $PM_{2.5}$ infiltration pathway; the heating season was therefore modelled as infiltration-dominated with windows assumed closed. As a result, the framework currently applies exclusively to building environments where infiltration is the dominant mechanism for outdoor $PM_{2.5}$ transport. These results should not be extrapolated to buildings equipped with active mechanical ventilation or filtration systems without additional modelling to account for the specific removal and supply dynamics of those systems.

Table 2. CIBSE TM23 UK standard for allowable airtightness in buildings (CIBSE, 2022)

Building Type	Building Airtightness Q_{50} ($m^3/h \cdot m^2$) @ 50 Pa	
	Best Practice	Normal Practice
Offices (Naturally Ventilated)	3.0	7.0
Offices (Mixed Mode Ventilation)	2.5	5.0
Offices (Air Conditioned)	2.0	5.0
Schools	3.0	9.0

Hourly weather and ambient air quality data for the heating season (defined in this study as November-April) were obtained from local monitoring stations (DEFRA, **Figure 2**). During this period, the buildings were modelled as infiltration-dominated, consistent with institutional heating policies that assume windows remain closed. Consequently, the total zone airflow rate (ACH_T) was approximated by the infiltration air change rate (ACH_{INF}), with no dedicated mechanical supply ventilation (i.e., mechanical ventilation contribution of 0%). Only basic local exhaust systems (e.g., kitchens, toilets) were represented. Similarly, CO_2 concentrations, which are primarily occupant-generated and typically controlled by dilution ventilation, were outside the scope of this infiltration-focused analysis.

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

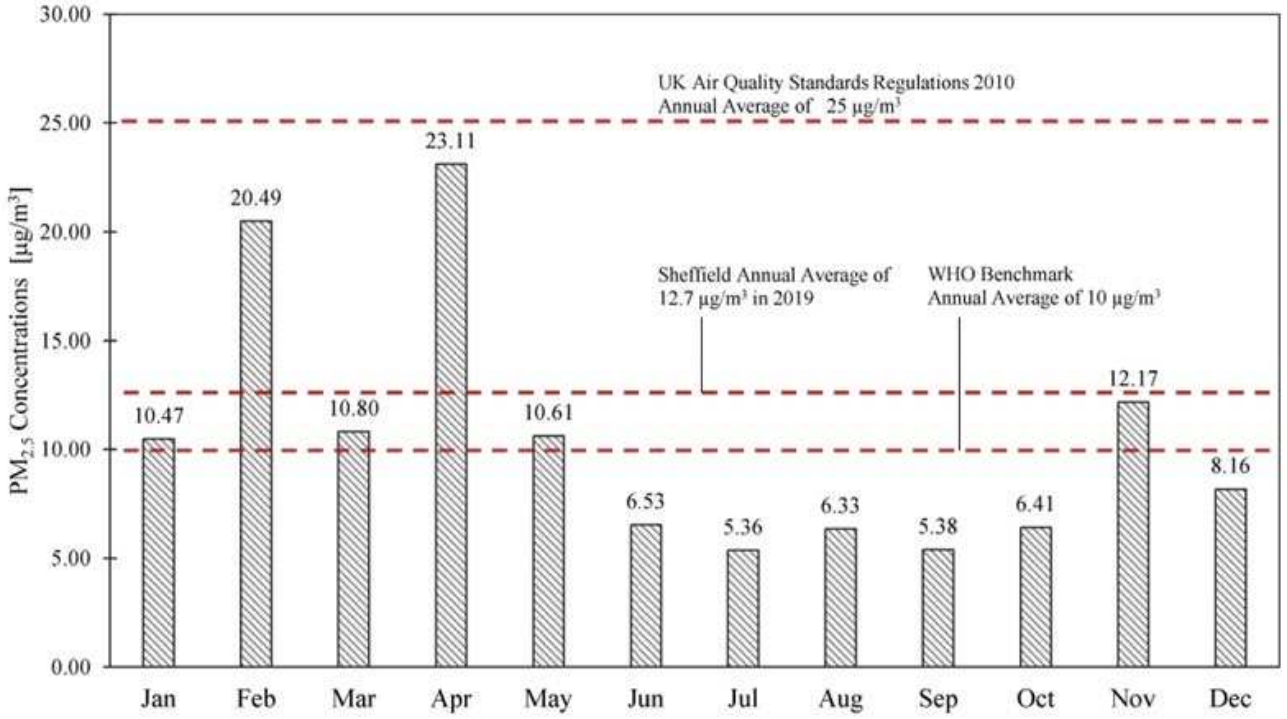


Figure 2. Monthly Average Outdoor PM2.5 Concentrations in Sheffield in 2019 (UKA00575, DEFRA)

2.2 Foundational data generation: High-resolution air-thermal co-simulation

To capture the substantial spatial and temporal variability of $PM_{2.5}$ infiltration within complex HEI buildings, the preparatory phase employed a high-resolution multizone air-thermal co-simulation using coupled CONTAM-EnergyPlus framework. This co-simulation was essential to represent the dynamic interplay between thermal conditions and airflow required for accurate prediction of room-level $PM_{2.5}$ concentrations, and its detailed methodology, parameterisation, assumptions (e.g., simplified window operation, exclusion of detailed AHU models), and validation (via indirect temperature correlation) for are comprehensively documented in our previous publication (Abdalla and Peng, 2025).

The high-resolution CONTAM-EnergyPlus simulations focused specifically on infiltration of outdoor-origin $PM_{2.5}$. To isolate this pollutant pathway, the models intentionally excluded significant indoor $PM_{2.5}$ sources (e.g., combustion, smoking, local combustion appliances) and complex mechanical ventilation systems such as Air Handling Units (AHUs), with only basic exhaust systems represented in kitchens and toilets. This exclusion was justified because UK law prohibits indoor smoking and no fuel combustion for cooking or heating is permitted on the HEI premises; while minor indoor sources such as printers and human activities may contribute, these were excluded due to the absence of comprehensive, zone-level data on their presence and usage patterns across the diverse spaces modelled (Abdalla and Peng, 2025). The physical parameters governing pollutant transport included a particle penetration factor P = 0.8 for general infiltration (P = 1 for natural ventilation) and a particle deposition rate k = 0.39h^{-1}. Airflow and

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

leakage were modelled across the 2,729 zones using a flow exponent n = 0.65, with discharge coefficient of 0.60 for windows and 0.78 for open internal doors. These physical inputs, combined with the building-specific airtightness values Q_{50} ranging from 3 (well-sealed) to 13 $m^3/h \cdot m^2$ (leaky), governed the effective leakage areas and air exchange rates for the heating season simulations.

The outputs of this established co-simulation study provide the foundational dataset for the current work, supplying hourly and season-averaged indoor $PM_{2.5}$ concentration fields for the metamodel training and subsequent exposure assessment. Because the simulations primarily represent infiltration-driven exposure in naturally ventilated or infiltration-dominated scenarios, the omission of detailed AHU and filtration modelling is explicitly acknowledged as a limitation, particularly for high-occupancy spaces such as lecture halls where mechanical ventilation and filtration could substantially alter indoor concentrations.

2.3 Interpretable Machine Learning Roadmap

2.3.1 Metamodel rationale and selection

This work developed machine learning (ML) metamodels as computationally efficient surrogates for the intensive physics-based simulations (Xian, J. & Wang, Z. 2024), enabling rapid and scalable indoor air quality assessment across the HEI building stock. The metamodels were trained on simulated heating season (November- April) indoor $PM_{2.5}$ concentrations generated by the CONTAM–EnergyPlus co-simulation for 2,729 zones. Because the training data were systematically generated by deterministic co-simulation rather than collected from sensors, the dataset did not contain missing values or gaps typical of empirical studies. Preprocessing involved aggregating the high-frequency (15-minute) co-simulation outputs to hourly and seasonal averages, and calculating derived features including indoor–outdoor temperature difference (ΔT) and exposed façade to zone volume ratio (Aef:Vz).

Among the algorithms investigated, including Generalised Additive Models (GAM), Random Forest Regression (RFR), and Extreme Gradient Boosting (XGBoost), XGBoost was selected as the primary metamodel due to its favourable predictive performance and flexibility for capturing non-linear and non-monotonic relationships between key physical inputs and concentrations. The dataset was randomly partitioned into training (70%) and testing (30%) subsets. Hyperparameters were optimised using 3-fold Randomised Search Cross-Validation (RSCV) implemented in scikit-learn (Pedregosa et al., 2011) for RFR and XGB, and Generalised Cross-Validation (GCV) within pyGAM (Servén & Brummitt, 2018) for GAM. The hyperparameter search spaces and optimal configurations for RFR and XGB are summarised in **Table 3**; optimal values were selected based on the lowest cross-validation RMSE. Detailed

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

performance metrics and comparative evaluation of the tuned models are presented in Section 3.2 (**Tables 8–9**, **Figure 4**).

Table 3. Hyperparameter search spaces and optimal values for Random Forest Regressor (RFR) and Extreme Gradient Boosting (XGB) metamodels optimised via 3-fold Randomised Search Cross-Validation (RSCV)

Algorithm	Hyperparameter	Description	Search Range	Optimal Values
RFR	n_estimators	Number of decision trees in the forest	200-2000 (10 values)	800
	max_features	Number of features considered at each split	'auto', 'sqrt'	'sqrt'
	max_depth	Maximum depth of individual trees	10-110 (11 values)	70
	min_samples_split	Minimum samples required to split an internal node	2, 5, 10	2
	min_samples_leaf	Minimum samples required at each leaf node	1, 2, 4	1
	bootstrap	Whether bootstrap samples are used for tree building	True, False	False
XGB	n_estimators	Number of boosting rounds (trees)	200-2000 (10 values)	800
	leaming_rate	Step size shrinkage to prevent overfitting	0.01, 0.1, 0.2, 0.3	0.01
	max_depth	Maximum depth of individual trees	10-110 (11 values)	100
	colsample_bytree	Fraction of columns randomly sampled per tree	0.4-0.9 (6 values)	0.8
	colsample_bylevel	Fraction of columns randomly sampled per tree level	0.4-0.9 (6 values)	0.4
	subsample	Fraction of observations randomly sampled per tree	0.5-0.9 (5 values)	0.8

2.3.2 Feature selection and collinearity mitigation

The inputs for the XGBoost metamodel were first identified through a sensitivity analysis framework that systematically assessed linear, monotonic, and non-monotonic relationships in the co-simulation outputs, supported by non-parametric group comparison tests (Kolmogorov–Smirnov and Kruskal–Wallis) to rank explanatory variables. Building airtightness (Q_{50}) consistently emerged as the dominant factor influencing infiltrated $PM_{2.5}$ concentrations, with infiltration air change rate (ACH_{INF}), indoor–outdoor temperature difference (ΔT), exposed façade to zone volume ratio (Aef:Vz), and scaled wind speed (v) also showing strong and physically interpretable effects.

To address potential multicollinearity, a Variance Inflation Factor (VIF) analysis was applied to all candidate input variables identified by the sensitivity analysis, including Q_{50}, ACH_{INF}, ΔT, zone air permeability rate at 4 Pa (Q_4), zone volume (Vz), Aef:Vz, and v. Variables with VIF values exceeding the

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

selected threshold were removed from the final feature set to ensure that the retained predictors were sufficiently independent for reliable prediction. The resulting VIF values for each tested variable, and the decision to retain or exclude them, are reported in **Table 4**, providing transparent documentation of the multicollinearity mitigation process.

Table 4. Variance Inflation Factor (VIF) values for selected candidate input variables and multicollinearity decisions (target VIF < 5)

Variable	Description	Final VIF	Multicollinearity decision
Q_{50}	Building envelope airtightness @ 50 Pa	1.6	Retained
ACH_{INF}	Infiltration air change rate	1.9	Retained
ΔT	Indoor–outdoor temperature difference	2.0	Retained
Aef:Vz	Exposed façade to zone volume ratio	1.4	Retained
v	Scaled local wind speed	1.5	Retained
Q_4	Zone air permeability @ 4 Pa	14.2	Excluded
Vz	Zone volume	80.6	Excluded
Aef	Area of exposed façade	31.3	Excluded
Az	Zone floor area	41.1	Excluded
Hz	Zone height	5.4	Excluded
φ	Zone orientation	1.6	Retained

During this VIF-based screening, physically meaningful variables such as Q_4 and Vz were removed because they were highly collinear with retained predictors whose physical roles were already well established. Q_4 exhibited strong collinearity with Q_{50} and ACH_{INF}, both of which non-parametric tests had already identified as dominant drivers of infiltrated $PM_{2.5}$, while the effect of zone volume (Vz) on dilution and concentration heterogeneity is implicitly captured by the exposed façade to zone volume ratio (Aef:Vz). Consequently, excluding Q_4 and Vz does not distort the physical meaning of the model; instead, it yields a more parsimonious set of five robust, non-collinear input variables, including Q_{50}, ACH_{INF}, ΔT, Aef:Vz, and v, that preserve the key physical drivers of infiltration and provide a stable basis for subsequent SHAP-based interpretability analysis. Importantly, model development trials confirmed that the qualitative SHAP rankings of the primary drivers remained stable under the alternative feature sets considered during this selection process, ensuring that the exclusion of collinear variables did not alter the fundamental physical interpretations.

2.3.3 SHapley Additive exPlanations (SHAP)

Interpretability was achieved by integrating SHapley Additive exPlanations (SHAP; Lundberg and Lee, 2017) with the XGBoost metamodel. SHAP quantifies the marginal contribution of each predictor to model output, providing both global and local explanations that enhance transparency in data-driven IAQ models.

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

In this study, SHAP values were computed for the final set of five input variables (Q_{50}, ACH_{INF}, ΔT, Aef:Vz, v) across the three high-accuracy regression models (GAM, RFR, XGB). Average absolute SHAP values were used as a measure of global feature importance, with the associated percentage contribution indicating the share of explained variation attributable to each variable. Local SHAP values were also obtained for every individual zone prediction, enabling zone-level attribution of contributions from each input variable. The resulting global rankings and local contribution patterns are presented and interpreted in Section 3.3 (**Table 9**, **Figures 5-6**).

2.4 Exposure assessment methodology

Occupant-specific exposure to outdoor-sourced, infiltration-driven $PM_{2.5}$ was estimated by combining the ML metamodel's predicted zone-level concentrations with the theoretical time-activity patterns (t_{ij}) in a microenvironmental modelling framework, yielding integrated personal exposure (EI_i) and population-weighted exposure (PWE) metrics for distinct similar time-activity groups (STGs).

2.4.1 Similar Time-Activity Groups (STGs)

To account for heterogeneous exposure risks across the HEI community, four categories of Similar Time-Activity Groups (STGs) were established: Academic Staff, Administrative Staff, Undergraduate Students, and Postgraduate Research (PGR) Students. These STGs were defined to reflect typical functional roles in HEI buildings, but the associated time–use patterns are based on secondary data rather than locally measured occupancy behaviour. This approach acknowledges that exposure is largely dependent on an individual's trajectory through different microenvironments.

2.4.2 Occupancy data and microenvironment definition

The microenvironment (ME) approach categorised building zones (e.g., Offices, Lecture Halls, Laboratories) as microenvironments. Due to constraints including pandemic-related restrictions and the absence of a campus-wide sensor network and survey infrastructure, real-time, zone-level occupancy and activity data could not be collected. Instead, STG time fractions were based on theoretical assumptions informed by designed maximum occupancies from Estates and Facilities Management (EFM), Higher Education Statistics Agency (HESA, 2021) classifications of building users, and time–activity fractions from Klepeis et al. (2001) for corresponding subgroups (**Tables 5-6**). This STG framework therefore represent a stylised but policy-relevant approximation of typical HEI time-use patterns rather than an empirically calibrated occupancy schedule.

The reliance on theoretical time-use patterns introduces uncertainty into the exposure estimates; the EI_i and PWE metrics reported in this study should therefore be interpreted as scenario-based, illustrative

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

indicators of potential cohort-level exposure disparities under assumed occupancy behaviour, rather than precise predictions of individual risk.

Table 5. Theoretical occupancy profiles for each space type used in the co-simulation

Space Type		Theoretical Occupancy Profile
General Offices	Administration and Academic Offices	9:00 – 17:00 (1-hour break between 12:00 – 13:00)
Educational Facilities	Lecture Theatres / Labs / Workshops / Studios	2-hour occupancy interval between 10:00 – 17:00 (including a 10-min break at each interval and a 1-hour break between 12:00- 13:00)
	Seminar Rooms	1-hour occupancy interval between 10:00 – 17:00 (including a 10-min break at each interval and 1-hour break between 12:00- 13:00)
	Study / Computer Rooms	10:00 – 17:00 (1-hour break between 12:00-13:00)
Shared Facilities	Common Areas	10:00 – 17:00
	Kitchen/Toilets	N/A
Circulation		N/A
Services		N/A

Table 6. Assumed typical time fractions spent in each microenvironment across HEI building users

Microenvironment	Academic Staff	Administration Staff	Undergraduate Students	Post Graduate Students
Offices	0.4705 [197 min]	0.941 [395 min]	0.00	0.600 [360 min]
Educational Facilities	0.4705 [197 min]	0.00	0.784 [470 min]	0.134 [80 min]
Circulation	0.016 [10 min]	0.016 [10 min]	0.016 [10 min]	0.016 [10 min]
Shared Facilities	0.035 [15 min]	0.035 [15 min]	0.200 [120 min]	0.250 [150 min]

2.4.3 Methods of exposure quantification

The sources underscore that traditional ambient air quality indices, which rely solely on stationary outdoor monitoring, are often insufficient and can lead to substantial errors in estimating personal exposure, as they disregard individual movement and indoor microenvironment dynamics. The exposure assessment in this study is therefore founded on a microenvironmental modelling approach. A microenvironment is defined as a three-dimensional space where pollutant levels are considered uniform or exhibit consistent statistical properties over time. This method is necessary because personal exposure varies significantly with individual's time-activity patterns and trajectories through different microenvironments.

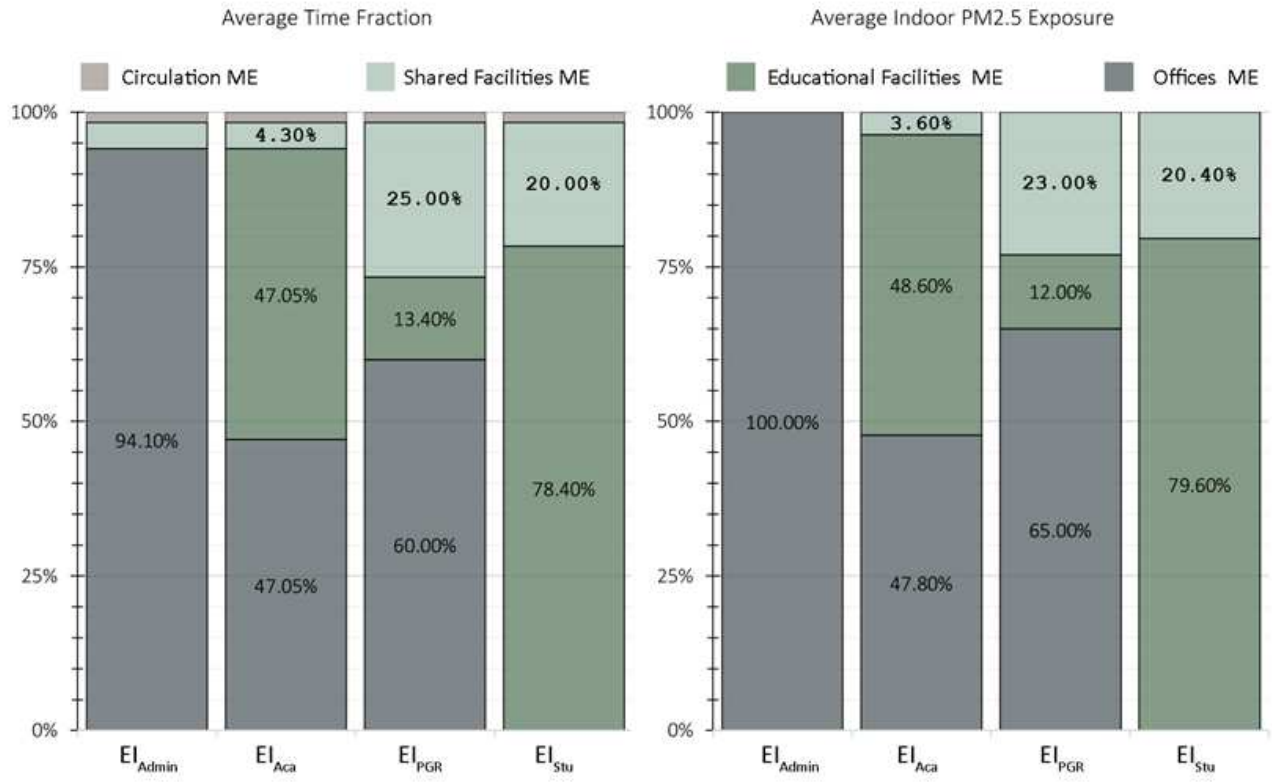
A. Integrated Personal Exposure (EI_i): The time-weighted average

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Time-activity profiles for STGs were based on theoretical assumptions and the aforementioned HESA classifications, reflecting typical occupancy schedules and spatial usage patterns rather than empirically measured trajectories. These STGs enabled estimation of personal exposure to outdoor-sourced indoor $PM_{2.5}$ by time-weighting zone-specific average concentrations with the proportion of time groups spent in each microenvironment (**Figure 3**). Personal exposure is calculated as the time-weighted integrated exposure, combining the simulated pollutant concentration in a given space with the duration an individual spends there. The integrated personal exposure (EI_i) for person *i* is determined by aggregating exposure across all microenvironments (*J*) visited during a specified time period, using a time-weighted average adapted from previous studies (Watson et al.,1988):

$$EI_i = \sum_{j}^{J} C_j \, t_{ij}$$

Where EI_i is the time-weighted integrated personal exposure for person *i*. C_j is the average pollutant concentration (here, indoor $PM_{2.5}$ from outdoor sources) predicted for microenvironment *j* (derived from the machine learning metamodel or co-simulation), and t_{ij} is the aggregate time fraction that person *i* (or STG *i*) spends in microenvironment *j*. Because t_{ij} values are derived from assumed STG time fractions, EI_i estimates are conditional on these stylised schedules and should be interpreted as scenario-based indicators of cohort-level exposure rather than precise predictions of individual risk.



This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Figure 3. Average annual time fractions and ME contributions to STGs' integrated personal exposure (EI_i) for outdoor-sourced indoor $PM_{2.5}$

Calculating personal exposure using this approach necessitates coupling the predicted $PM_{2.5}$ concentrations with the defined STGs. For HEI settings, STGs include cohorts such as Academic Staff, Administrative Staff, Undergraduate Students, and Postgraduate Research (PGR) Students. Although the present analysis focuses on standard exposure metrics, some studies distinguish personal exposure from potential inhaled dose, which reflects individual physiological differences and is calculated from $PM_{2.5}$ mass concentration and inhalation rate; the latter may better reflect health effects but lies beyond the scope of this study.

B. Population-Weighted Exposure (PWE)

To evaluate the overall impact of building characteristics and interventions on the community under outdoor-origin, infiltration-dominated conditions, Population-Weighted Exposure (PWE) is used. PWE provides a robust estimate of population exposure characteristics within a specific area by aggregating microenvironment exposures according to the size and distribution of occupant groups across the building stock. It is an effective metric for identifying microenvironments with higher indoor concentrations of outdoor-sourced $PM_{2.5}$, thereby helping to prioritise targeted mitigation strategies. PWE is determined by weighting concentrations according to the number of people in a given area. The general form of the PWE equation is defined as (Abdul Shakor et al., 2020; Aunan et al., 2018):

$$PWE_i = \frac{1}{P_i} \sum_i C_i P_i$$

where PWE_i is the population-weighted exposure for microenvironment *i*. P_i is the population in microenvironment *i* (based on maximum design occupancy). C_i is the annual average indoor $PM_{2.5}$ concentration in microenvironment *i*. The methodology relies on annual average concentrations (C_i) to align with annual exposure guidelines. Because P_i values are based on design occupancies and C_i values reflect simulated infiltration-only concentrations, the resulting PWE metrics should be interpreted as scenario-based indicators of cohort-level exposure under the assumed boundary conditions, rather than precise predictions of realised population dose. The occupancy and microenvironment characteristic data was obtained from the University's Estates and Facilities Management office (**Table 7)**.

Table 7. Occupancy and area characteristics of microenvironments and subcategories of the HEI buildings

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

	Subcategory	Maximum Occupancy	Area (m^2)	Occupancy Density (m^2/person)
Offices	Administration Offices	859	4184.4	4.5
	Academic Offices	315	2862.5	9
	PGR Offices	245	1490.2	6
Educational Facilities	Lecture Halls	794	1219.6	1.5
	Labs	478	1194.5	2.5
	Studios	861	2156.6	2.5
	Workshops	320	802.2	2.5
Shared Facilities	Study & Computer Rooms	782	1,954.7	2.5

Since the high-accuracy XGBoost metamodel was explicitly trained and tested using data simulated for the high-risk heating season (November to April), the annual average indoor $PM_{2.5}$ concentration (C_i) for each microenvironment (*i*) was calculated by combining the metamodel predictions with the corresponding non-heating season concentrations:

Heating-season concentrations (November–April): indoor $PM_{2.5}$ concentrations were predicted using the validated XGBoost metamodel. Non-heating -season concentrations (May–October): indoor $PM_{2.5}$ concentrations were taken directly from the original CONTAM-EnergyPlus co-simulation outputs, because the assumed window operability (fully open) and resulting natural ventilation during this period fall outside the parameter space used for training the XGBoost model. This hybrid approach avoids extrapolating the metamodel beyond its heating-season training domain and preserves the variance associated with naturally ventilated, lower-concentration months, ensuring that the final annual C_i values used in EI_i and PWE calculations reflect the full 12-month concentration dynamics. Consequently, these annual exposure metrics represent composite indicators derived from these two distinct modelling approaches—metamodelling for the heating season and physics-based simulation for the non-heating season—and should not be interpreted as statistically unified predictions generated by a single model.

3. Results

3.1 Identification of key exposure determinants

The initial stage in developing a predictive machine learning (ML) framework requires defining the most critical physical and environmental parameters influencing the concentrations of outdoor-sourced infiltrated $PM_{2.5}$ (C_i). Given the intricate and non-linear relationships within the HEI building context, a comprehensive sensitivity analysis framework was applied to the high-resolution co-simulation outputs, examining linear, monotonic, and non-monotonic relationships. This analysis identified a reduced

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

subset of five key explanatory variables for model training: building airtightness (Q_{50}), infiltration air change rate (ACH_{INF}), indoor-outdoor temperature difference (ΔT), scaled wind speed (v), and the exposed façade to zone volume ratio (A_{ef}:V_z). These five variables were then subjected to Variance Inflation Factor (VIF) analysis, as described in Section 2.3.2, to ensure that the final feature set used for training was free from problematic multicollinearity.

Across the range of statistical tests conducted, Q_{50} was consistently identified as the dominant factor influencing $PM_{2.5}$ concentrations. While initial correlation analyses sometimes ranked Q_{50} lower (due to its categorical nature across zones), specialised group comparison tests, specifically the Kolmogorov-Smirnov and Kruskal-Wallis quantile tests, definitively established Q_{50} as the parameter with the highest influence (Rank 1) on C*i*. This finding holds significant practical importance, demonstrating that a low Q_{50} value (representing a tight building envelope) is intrinsically linked to low $PM_{2.5}$ concentrations

The remaining variables capture the essential dynamic aspects of particle infiltration. The infiltration air change rate (ACH_{INF}) was identified as a significant factor (Rank 2–3) alongside $Q_{50,}$ confirming its role in modulating airflow and particle transport. Furthermore, the indoor-outdoor temperature difference (ΔT) exhibited a strong negative non-linear relationship with $PM_{2.5}$ levels, highlighting the influence of thermal stack effects on air exchange. Together with scaled wind speed (v) and the exposed façade to zone volume ratio (Aef:Vz), these findings emphasise that both envelope properties and driving forces for airflow are critical determinants of heating-season indoor PM2.5 levels in HEI buildings.

3.2 Metamodel predictive performance and validation

The machine learning (ML) metamodels were strategically developed as computationally efficient surrogates for the intensive physics-based simulations, aiming to maintain accuracy while dramatically reducing computational time across the HEI building stock. The metamodels were trained and tested using heating-season (November–April) indoor PM2.5 concentrations from the CONTAM–EnergyPlus co-simulation for Buildings 1–4, consistent with the hybrid annual exposure framework in which non-heating-season concentrations are taken directly from the physics-based model (Section 2.4). Following cross-validation and hyperparameter optimisation, performance evaluation metrics (R, R^2, RMSE, MAE, and model accuracy) for GAM, RFR, and XGB on the heating-season test dataset are summarised in **Table 8**, with corresponding regression plots shown in **Figure 4**; the XGBpost-HPT model achieved the highest predictive accuracy (95.0%), with R^2 = 0.92, RMSE = 0.37 μg/m^3, and MAE = 0.26 μg/m^3, outperforming RFRpost-HPT (R^2 = 0.88) and GAMpost-HPT (R^2 = 0.815).

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Table 8. Performance evaluation metrics (R^2, RMSE, MAE) for GAM, RFR, and XGB metamodels, assessed on the heating season co-simulation testing dataset (derived from Building 1 through Building 4 co-simulation), following hyperparameter tuning (HPT)

Performance Metric	GAM	RFR	XGB
R	0.900	0.935	0.960
R^2	0.815	0.880	0.920
RMSE	0.560	0.450	0.370
MAE	0.440	0.315	0.260
Model Accuracy	91.70%	93.90%	95.00%

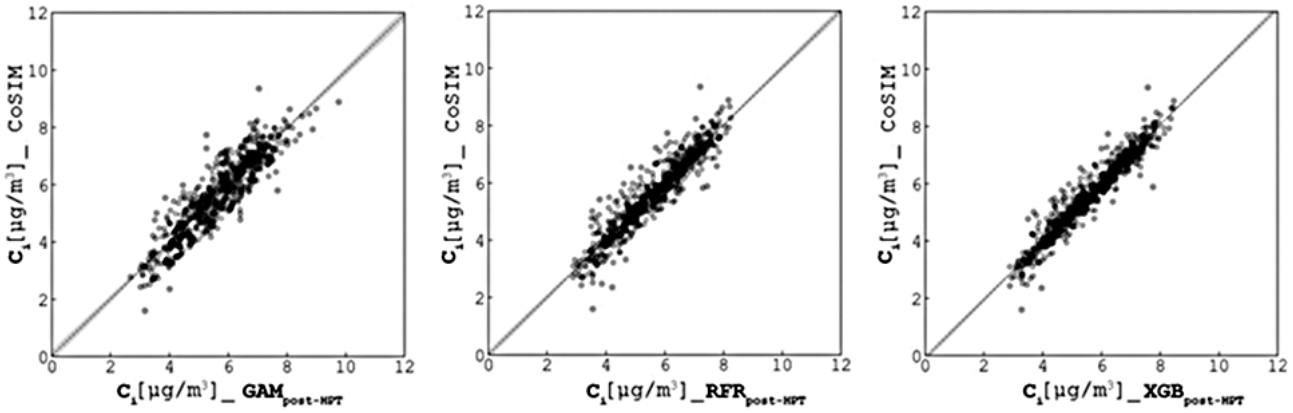


Figure 4. Regression plots of CONTAM-EnergyPlus co-simulation dataset vs. testing dataset for fitted GAM, RFR, and XGB models (heating season, Building 1 – Building 4, following hyperparameter tuning)

To confirm the generalisability beyond the buildings used for training and testing, the fitted models were evaluated on an independent, unseen hold-out dataset derived from the co-simulation of the fifth selected building (Building 5). On this hold-out dataset, the XGBpost-HPT model maintained high predictive performance, achieving R = 0.95, $R^2 \approx 0.91$, RMSE = 0.6 µg/m^3, and MAE = 0.4 µg/m^3, whereas GAMpost-HPT and RFRpost-HPT yielded lower R^2 values of 0.86 and 0.62, respectively (**Table 9**). The stability of XGB's performance across both the primary test dataset and the hold-out building provides confidence that the subsequent SHAP-based interpretability analysis (Section 3.3) and exposure quantification (Section 3.4) are grounded in a robust and transferable predictive model

Table 9. Model evaluation metrics for the GAM, RFR, and XGB models (Building 1-4) fitted for heating season co-simulation dataset vs. testing dataset (Building 5 as the 'hold-out') following hyperparameter tuning (HPT)

Performance Metric	GAM	RFR	XGB
R	0.790	0.935	0.950
R^2	0.620	0.860	0.905
RMSE	1.150	0.618	0.580
MAE	0.955	0.460	0.410

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Model Accuracy	78.59%	88.85%	90.60%

Overall, these results validate XGBoost as the preferred metamodel for predicting heating-season indoor PM2.5 concentrations across the heterogeneous HEI building stock. Its combination of high accuracy, low error, and robust performance on unseen data ensures that the subsequent interpretability (SHAP) and exposure assessment stages build on a reliable surrogate for the underlying physics-based simulations.

3.3 Interpretability of exposure drivers (SHAP analysis)

The application of SHapley Additive exPlanations (SHAP) was essential for enhancing the transparency and interpretability of the high-accuracy XGBoost metamodel, enabling the complex relationships between physical inputs and heating-season indoor $PM_{2.5}$ concentrations to be expressed as physically meaningful exposure drivers. In this subsection, SHAP is used to quantify both the global importance and the local, zone-specific contributions of the five key input variables (Q_{50}, ACH_{INF}, ΔT, v, and Aef:Vz) to the predicted concentrations, providing an interpretable basis for understanding how building fabric and airflow conditions shape infiltration-driven $PM_{2.5}$ in HEI buildings.

The SHAP analysis provided a global interpretation of the metamodels by quantifying the overall influence and ranking of the five key input variables (Q_{50}, ACH_{INF}, ΔT, v, and Aef:Vz) across GAM, RFR, and XGB models. As summarised in **Table 10**, building airtightness (Q_{50}) consistently ranked as the primary driver of heating-season $PM_{2.5}$, explaining approximately 32.5–39.7% of the total variation depending on the algorithm, while ACH_{INF} and ΔT occupied the second and third ranks across all three models. This cross-model agreement in both ranking and percentage contribution demonstrates that the core interpretations of the dominant physical drivers are stable and not an artefact of a particular ML architecture.

Table 10. Average absolute SHAP values and explained variation (%) by variable for heating season $PM_{2.5}$ concentrations (rank in parentheses)

Model	Aef:Vz		ACH_{INF}		ΔT		Q_{50}		$\boldsymbol{v}$	
	SHAP	%	SHAP	%	SHAP	%	SHAP	%	SHAP	%
GAM	+0.17 (4)	11.3	+0.24 (3)	15.9	+0.11 (5)	7.3	+0.60 (1)	39.7	+0.39 (2)	25.8
RFR	+0.15 (5)	9.9	+0.39 (2)	25.7	+0.21 (4)	13.8	+0.51 (1)	33.6	+0.26 (3)	17.1
XGB	+0.19 (5)	12.6	+0.34 (2)	22.5	+0.21 (4)	13.9	+0.49 (1)	32.5	+0.28 (3)	18.5

Figure 5 presents SHAP summary plots for the GAM, RFR, and XGB models, visualising the distribution of SHAP values for each variable and confirming the global rankings reported in Table 10. Low Q_{50} values

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

(tighter envelopes) are associated with strongly negative SHAP values, indicating substantial reductions in predicted $PM_{2.5}$, whereas higher Q_{50} values consistently exert positive contributions. By contrast, higher ACH_{INF} values tend to produce positive SHAP values, reflecting increased infiltration-driven concentrations, while larger indoor–outdoor temperature differences (ΔT) are generally associated with negative SHAP contributions, consistent with reduced infiltration under the simulated heating-season stack conditions. These patterns emphasise that, across models, envelope airtightness and infiltration airflow are the dominant levers for controlling heating-season indoor PM2.5 in HEI buildings, with wind-driven effects (v) and façade-to-volume ratio (Aef:Vz) modulating the impact of these primary drivers.

Figure 6 illustrates the local interpretability of the XGB metamodel using a representative SHAP plot for a single zone, showing how the contributions of Q_{50}, ACH_{INF}, ΔT, Aef:Vz, and *v* shift the predicted concentration *f(x)* away from the dataset mean (base value). Positive SHAP values indicate features that increase the predicted $PM_{2.5}$ relative to the average, whereas negative SHAP values indicate features that reduce it. By examining such plots for individual rooms, facility managers can identify which specific physical factors—such as a leaky envelope (high Q_{50}), elevated infiltration rate (high ACH_{INF}), or an unfavourable façade-to-volume ratio (Aef:Vz)—are driving elevated concentrations in that zone and can prioritise targeted interventions where they will be most effective.

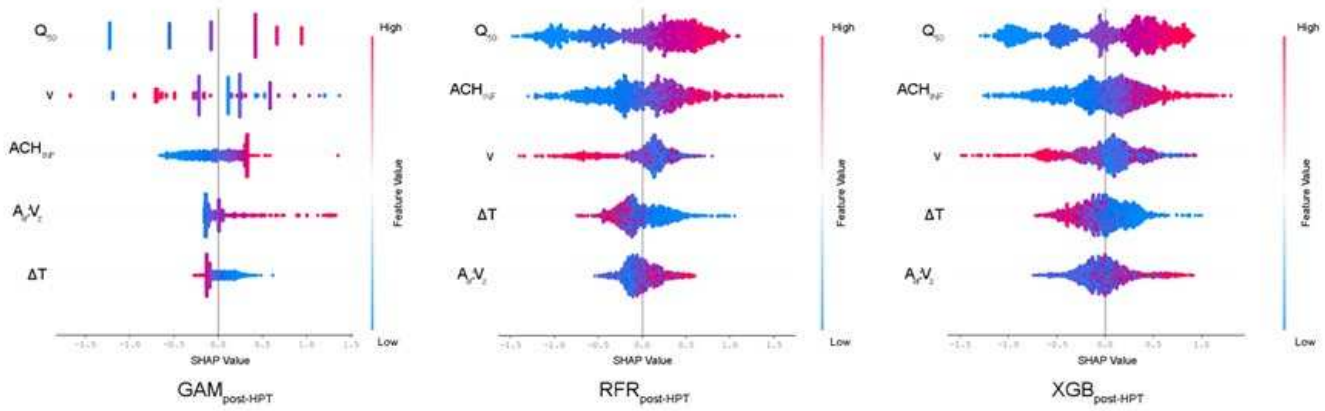


Figure 5. SHAP summary plots for GAM, RFR, and XGB metamodels, illustrating the global importance and signed impact of Q_{50}, ACH_{INF}, ΔT, *v*, and Aef:Vz on predicted heating-season $PM_{2.5}$ concentrations across all zones

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

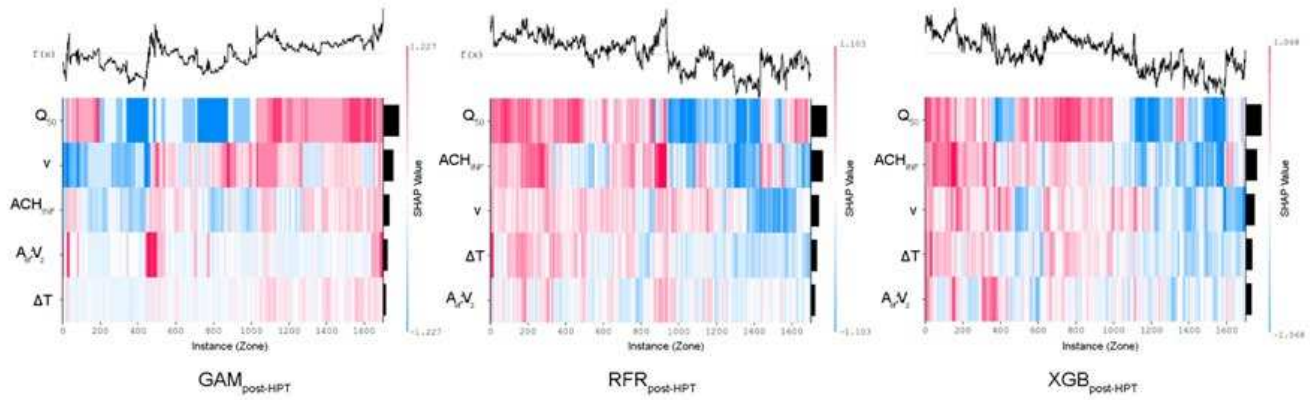


Figure 6. Example SHAP local interpretability plot for the XGB metamodel, showing zone-level contributions of Q_{50}, ACH_{INF}, ΔT, Aef:Vz, and *v* to the deviation of a single zone's predicted heating-season $PM_{2.5}$ concentration from the dataset mean

Taken together, the global and local SHAP results demonstrate that building airtightness (Q_{50}), infiltration air change rate (ACH_{INF}), and temperature difference (ΔT) are the dominant determinants of heating-season indoor $PM_{2.5}$ concentrations, with Aef:Vz and *v* modulating their effects. This interpretable ranking underpins the subsequent exposure analysis (Section 3.4) and supports the policy conclusion that envelope measures, while necessary, must be complemented by ventilation and filtration strategies to achieve compliance with stringent WHO 2021 guidelines.

3.4 Quantification of occupant exposure disparities

Using the annual average indoor $PM_{2.5}$ concentrations (C_i) constructed by combining heating-season metamodel predictions with non-heating-season co-simulation outputs (Section 2.4), and the STG-based time–activity and occupancy assumptions described earlier, this section quantifies disparities in integrated personal exposure (EI_i) and population-weighted exposure (PWE) among occupant cohorts and microenvironments within the HEI building stock. These exposure metrics relate solely to outdoor-origin $PM_{2.5}$ under infiltration-dominated conditions. This represents a clear boundary for the current study; the findings should not be extrapolated to mechanically ventilated or filtered buildings, as such systems would substantially alter the indoor-outdoor concentration relationship (**Table 11**).

Table 11. Average heating-season and annual indoor $PM_{2.5}$ concentrations from outdoor sources in different microenvironments (ME) for the baseline airtightness Q_{50}, Q_{50} = 7 $m^3/h{\cdot}m^2$, and Q_{50} = 3 $m^3/h{\cdot}m^2$ scenarios; these values provide the C_i inputs for the exposure calculations in Section 3.4

Microenvironment (ME)		Baseline Q_{50}		Q_{50} = 7 $m^3/h{\cdot}m^2$		Q_{50} = 3 $m^3/h{\cdot}m^2$	
Type	Sub-Category	Heating Season	Annual	Heating Season	Annual	Heating Season	Annual

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Offices	Admin Offices	6.4 (±0.9)	10.1 (±1.3)	6.1 (±1.0)	9.7 (±1.3)	4.4 (±0.9)	7.4 (±1.2)
	Academic Offices	5.8 (±1.34)	9.2 (±1.7)	4.5 (±1.0)	7.6 (±1.3)	3.4 (±1.4)	6.1 (±1.2)
	PGR Offices	7.0 (±0.6)	10.9 (±1.2)	6.0 (±0.5)	9.6 (±1.0)	4.5 (±0.8)	7.7 (±1.1)
Educational Facilities	Workshops	5.9 (±0.6)	9.2 (±1.0)	5.1 (±0.9)	8.3 (±1.2)	3.8 (±1.2)	6.6 (±1.6)
	Lecture Halls	6.3 (±0.7)	10.0 (±1.0)	5.3 (±0.8)	8.7 (±1.1)	3.7 (±0.7)	6.6 (±1.0)
	Labs	5.9 (±0.2)	9.2 (±0.6)	5.0 (±0.7)	8.0 (±1.34)	3.7 (±0.7)	5.9 (±1.3)
	Studios	6.3 (±1.2)	9.4 (±8.8)	4.9 (±0.7)	8.1 (±0.9)	3.5 (±0.7)	6.2 (±0.9)
Shared Facilities	Computer Rooms	5.8 (±1.4)	9.4 (±1.6)	4.6 (±1.4)	7.9 (±1.7)	3.1 (±1.1)	5.8 (±1.5)

3.4.1 Personal exposure disparities (EI_i) across occupant groups

The calculated integrated personal exposure (EI_i) illustrated pronounced disparities among the different STG categories under the adopted schedules, directly reflecting their assumed time-allocation patterns within the campus microenvironments under the STG schedules described in Section 2.4.1. Under the Baseline Q_{50} scenario, Postgraduate Research (PGR) students experienced the highest total exposure (EI_{PGR} ≈10.2 μg/m^3), followed by Administrative Staff (EI_{Adm} ≈ 9.5 μg/m^3), while Undergraduate Students and Academic Staff recorded lower values consistent with their shorter assumed residence times in offices and long-stay spaces. The analysis of microenvironmental contribution revealed stark differences in exposure drivers:

Administrative Staff exposure was dominated entirely by the Offices microenvironment, contributing 100% of their total integrated exposure (e.g., ≈ 9.5 μg/m^3 at baseline), because they are assumed to spend almost all working hours in office spaces (**Table 7**). This indicates that, under the current STG assumptions, office-focused interventions, such as portable filtration or targeted airtightness and ventilation upgrades, would yield the largest reductions in Administrative Staff exposure. Undergraduate Student exposure (EI_{Stu}) was primarily incurred in Educational Facilities (e.g., lecture halls, labs), which accounted for up to 80.8% of their total integrated exposure. This finding highlights Educational Facilities as key exposure hotspots for the largest student cohort, suggesting that improving IAQ in teaching spaces is critical for reducing student exposures. **Table 12** shows the annual personal exposure to indoor $PM_{2.5}$ using various Q_{50} scenarios.

Table 12. Annual integrated personal exposure (EI_i) to indoor $PM_{2.5}$ from outdoor sources for Similar Time-Activity Groups (STGs) under different Q_{50} scenarios, based on assumed time-activity fractions (t_{ij}) across microenvironments

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Microenvironment	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Sub-Category	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Time Fraction (t_{ij}, %)	Personal exposure (E_i, μg/m^3)
Offices	10.0 (± 1.45)	Admin Offices	10.1 (± 1.3)	0.94	9.5
		Academic Offices	9.2 (± 1.7)	0.47	4.3
		PGRs Offices	10.9 (± 1.2)	0.60	6.6
Educational Facilities	9.4 (± 1.34)	Workshops	9.2 (± 1.0)	0.78 (0.47)	7.2 (4.3)
		Lecture Halls	10.0 (± 1.0)	0.78 (0.47)	7.8 (4.7)
		Labs	9.2 (± 0.6)	0.78 (0.47)	7.2 (4.3)
		Studios	9.4 (± 8.8)	0.78 (0.47)	7.3 (4.4)
Shared Facilities	9.4 (± 1.6)	Study Rooms	9.4 (± 1.6)	0.78 (0.3)	7.4 (2.8)

Baseline Q_{50} = 13 m^3/h·m^2 for Building 1 and Building 2; 10 m^3/h·m^2 for Building 3 and Budling 4

Microenvironment	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Sub-Category	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Time Fraction (t_{ij}, %)	Personal exposure (E_i, μg/m^3)
Offices	9.2 (± 1.6)	Admin Offices	9.7 (± 1.3)	0.9	9.2
		Academic Offices	7.6 (± 1.3)	0.5	3.6
		PGRs Offices	9.6 (± 1.0)	0.6	5.8
Educational Facilities	8.2 (± 1.0)	Workshops	8.3 (± 1.2)	0.78 (0.47)	6.5 (3.9)
		Lecture Halls	8.7 (± 1.1)	0.78 (0.47)	6.8 (4.1)
		Labs	8.0 (± 1.4)	0.78 (0.47)	6.3 (3.8)
		Studios	8.1 (± 0.9)	0.78 (0.47)	6.4 (3.8)
Shared Facilities	7.9 (± 1.7)	Study Rooms	7.9 (± 1.7)	0.78 (0.30)	6.2 (2.4)

Q_{50} = 7 m^3/h·m^2 across all buildings

Microenvironment	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Sub-Category	Annual $PM_{2.5}$ Concentration (C_j, μg/m^3)	Time Fraction (t_{ij}, %)	Personal exposure (E_i, μg/m^3)
Offices	7.0 (± 1.3)	Admin Offices	7.4 (± 1.2)	0.94	7.0
		Academic Offices	6.1 (± 1.2)	0.470	2.9
		PGRs Offices	7.7 (± 1.1)	0.600	4.6
Educational Facilities	6.2 (± 1.0)	Workshops	6.6 (± 1.6)	0.78 (0.47)	5.2 (3.1)
		Lecture Halls	6.6 (± 1.0)	0.78 (0.47)	5.2 (3.1)
		Labs	5.9 (± 1.3)	0.78 (0.47)	4.6 (2.8)
		Studios	6.2 (± 0.9)	0.78 (0.47)	4.9 (2.9)
Shared Facilities	5.8(± 1.5)	Study Rooms	5.8(± 1.5)	0.78 (0.30)	4.6 (1.8)

Q_{50} = 3 m^3/h·m^2 across all buildings

Under the three airtightness scenarios, integrated personal exposure (EI_i) highlights persistent disparities between Similar Time-Activity Groups: PGR and Administrative Staff consistently exhibit the highest annual EI_i due to prolonged time in offices, while Undergraduate Students' exposure is dominated by Educational Facilities. Tightening the envelope substantially reduces EI_i for all STGs, yet

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

the relative ordering between cohorts and the dominant microenvironments contributing to their exposure change little, indicating that differences in time–activity patterns remain the primary driver of cohort-level disparities (**Table 13**).

Table 13. Examples of the relative contributions from Offices, Educational Facilities, and Shared Facilities microenvironments to STG-based annual integrated personal exposure (EI_i) to indoor $PM_{2.5}$ from outdoor sources under varying building airtightness (Q_{50}) scenarios

STG Category	Microenvironment	$PM_{2.5}$ Concentration (C_j, μg/m³)	Time Fraction (t_{ij})	$C_j * t_{ji}$ (EI_j, μg/m³)	Microenvironment Contribution (%)
Admin Staff	Offices	10.1	0.94	9.5	100 %
			$EI_{Adm} = \sum C_j\, t_{ij}$ = 9.5 μg/m³		
Academic Staff	Offices	9.2	0.47	4.3	47.8 %
	Educational Facilities	9.4	0.47	4.4	48.6 %
	Shared Facilities	9.4	0.04	0.3	3.6 %
			$EI_{Aca} = \sum C_j\, t_{ij}$ = 9.0 μg/m³		
PGR Student	Office	10.9	0.60	6.6	65.0 %
	Educational Facilities	9.4	0.13	1.2	12.0 %
	Shared Facilities	9.4	0.25	2.4	23.0 %
			$EI_{Pgr} = \sum C_j\, t_{ij}$ = 10.2 μg/m³		
Undergraduate	Educational Facilities	9.4	0.78	7.3	79.6 %
	Shared Facilities	9.4	0.20	1.9	20.4 %
			$EI_{Stu} = \sum C_j\, t_{ij}$ = 9.2 μg/m³		

Baseline Q_{50} = 13 m³/ h·m² for Building 1 and Building 2; 10 m³/h/m² for Building 3 and Building 4

STG Category	Microenvironment	$PM_{2.5}$ Concentration (C_j, μg/m³)	Time Fraction (t_{ij})	$C_j * t_{ji}$ (EI_j, μg/m³)	Microenvironment Contribution (%)
Admin Staff	Offices	9.7	0.94	9.2	100 %
			$EI_{Adm} = \sum C_j.\, t_{ij}$ = 9.2 μg/m³		
Academic Staff	Offices	7.6	0.47	3.6	46.3 %
	Educational Facilities	8.2	0.47	3.8	50.1 %
	Shared Facilities	7.9	0.04	0.3	3.6 %
			$EI_{Aca} = \sum C_j.\, t_{ij}$ = 7.7 μg/m³		
PGR Student	Office	9.6	0.60	5.8	65.4 %
	Educational Facilities	8.2	0.13	1.1	12.36 %
	Shared Facilities	7.9	0.25	2.0	22.24 %
			$EI_{Pgr} = \sum C_j.\, t_{ij}$ = 8.9 μg/m³		
Undergraduate	Educational Facilities	8.2	0.78	6.4	80.26 %
	Shared Facilities	7.9	0.20	1.6	19.74 %
			$EI_{Stu} = \sum C_j.\, t_{ij}$ = 8.0 μg/m³		

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

491 Q_{50} = 7 m^3/ h·m^2 across all buildings

STG Category	Microenvironment	$PM_{2.5}$ Concentration (C_j, μg/m^3)	Time Fraction (t_{ij})	*Cj * tji* (EI_j, μg/m^3)	Microenvironment Contribution (%)
Admin Staff	Offices	7.4	0.94	7.0	100 %
			$EI_{Adm} = \sum C_j . t_{ij}$ = 7.0 μg/m^3		
Academic Staff	Offices	6.1	0.47	2.9	47.9%
	Educational Facilities	6.2	0.47	2.9	48.7%
	Shared Facilities	5.8	0.04	0.2	3.4%
			$EI_{Aca} = \sum C_j . t_{ij}$ = 6.0 μg/m^3		
PGR Student	Office	7.7	0.60	4.6	66.8%
	Educational Facilities	6.2	0.13	0.8	12.1%
	Shared Facilities	5.8	0.25	1.4	21.1%
			$EI_{Pgr} = \sum C_j . t_{ij}$ = 6.8 μg/m^3		
Undergraduate	Educational Facilities	6.2	0.78	4.9	80.8%
	Shared Facilities	5.8	0.20	1.2	19.2%
			$EI_{Stu} = \sum C_j . t_{ij}$ = 6.1 μg/m^3		

492 Q_{50} = 3 m^3/ h·m^2 across all buildings

493 3.4.2 Population-weighted exposure (PWE) and compliance challenge

494 The PWE metric provided an overall assessment of the impact of infiltration on the entire HEI occupant
495 community, weighted by the maximum design occupancy of each zone and derived from annual average
496 indoor $PM_{2.5}$ concentrations (C_i) that combine XGBoost predictions for the heating season (November–
497 April) with CONTAM–EnergyPlus outputs for the non-heating season (May–October).

498 In the Baseline Q_{50} scenario, the overall annual PWE was 9.6 μg/m^3, nearly equal to the former World
499 Health Organization (WHO) 2005 annual guideline of 10 μg/m^3. Specific high-occupancy areas, namely
500 PGR offices (10.9 μg/m^3), Administrative Offices (10.1 μg/m^3), and Lecture Halls (10.0 μg/m^3), were found
501 to be equal or slightly higher than the 10 μg/m^3 threshold. Improving the airtightness to the Moderate
502 scenario (Q_{50} = 7 m^3/h·m^2) resulted in an 11.5% reduction, bringing the total annual PWE down to
503 8.5 μg/m^3 and moving all measured microenvironments below the WHO 2005 annual guideline. **Figure 7**
504 illustrates that improving Q_{50} to the moderate airtightness scenario reduces PWE to indoor $PM_{2.5}$ from
505 outdoor sources in most microenvironments compared with the Baseline Q_{50} scenario.

506 Further improvements to the well-sealed envelope scenario (Q_{50} = 3 m^3/h·m^2) led to a maximum
507 reduction in total annual PWE of 32.3% compared with the baseline, achieving a final exposure level of
508 6.5 μg/m^3. However, when this result is compared with the more stringent WHO 2021 annual guideline
509 of 5 μg/m^3, a critical limitation emerges: approximately 88% of indoor zones still exceed this tighter
510 benchmark, even in the best-case airtightness scenario (Q_{50} = 3 m^3/h·m^2). While some

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

microenvironments, such as PWE_{Labs} (5.9 μg/m^3) and PWE_{Study} (5.8 μg/m^3), demonstrate levels close to the 5 μg/m^3 guideline, the high failure rate shows that envelope tightening, although effective in lowering infiltration-driven concentrations, is insufficient on its own to meet the WHO 2021 $PM_{2.5}$ challenge. Instead, these results argue for integrated mitigation strategies that combine improved airtightness with controlled mechanical ventilation, advanced filtration, or mixed-mode systems, particularly in naturally ventilated, high-occupancy spaces such as lecture halls and offices.

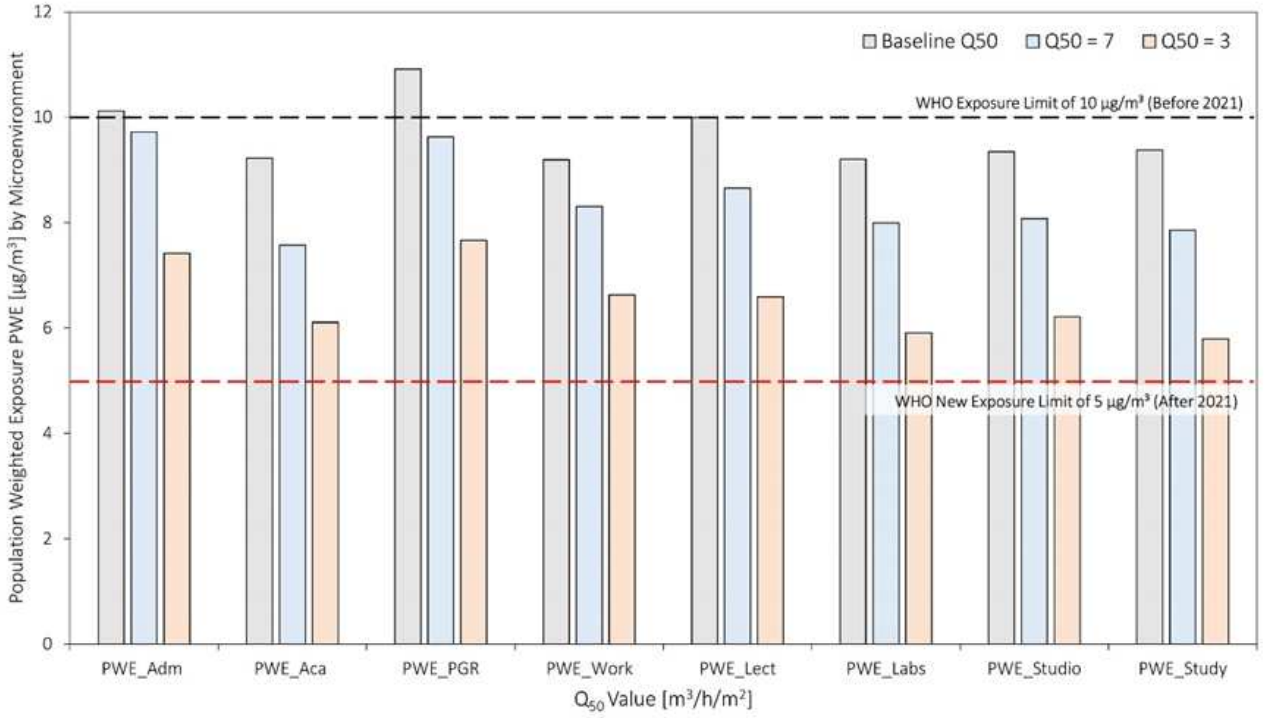


Figure 7. Population-weighted exposure (PWE) to indoor $PM_{2.5}$ from outdoor sources across HEI microenvironments under baseline, moderate (Q_{50} = 7 m³/h·m²), and well-sealed (Q_{50} = 3 m³/h·m²) airtightness scenarios, illustrating the progressive but incomplete reduction in annual exposure relative to WHO 2021 guideline.

4. Discussion

4.1 The utility of sensitivity analysis and ML metamodelling for IAQ assessment

The study addresses the challenge of conducting rapid, large-scale Indoor Air Quality (IAQ) assessments in complex HEI buildings, where traditional physics-based simulations, despite their precision, are computationally intensive and impractical for institution-wide application. To overcome this, a novel integrative framework was developed, combining high-resolution CONTAM-EnergyPlus co-simulations with machine learning (ML) metamodels. These ML metamodels serve as computationally efficient surrogates for the intensive physics-based simulations, reducing assessment time and enabling scalable IAQ evaluations across the building stock. Specifically, the XGB metamodel demonstrated high

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

predictive accuracy ($R \approx 0.95$; $R^2 > 0.90$), proving to be a robust and efficient alternative for conducting IAQ assessments and enabling HEI facility managers to screen thousands of zones for elevated $PM_{2.5}$ exposure without repeating computationally intensive co-simulations.

Furthermore, sensitivity analysis played a crucial role in systematically identifying the most influential input variables affecting infiltrated $PM_{2.5}$ concentrations. This analysis revealed intricate, non-linear, and non-monotonic relationships between parameters and indoor $PM_{2.5}$ levels, offering insights beyond traditional correlation analyses. The five key parameters consistently identified were building airtightness (Q_{50}), infiltration air change rates (ACH_{INF}), indoor-outdoor temperature difference (ΔT), scaled wind speed (v), and the exposed façade-to-zone volume ratio (A_{ef}:V_z). By confirming this reduced, non-collinear feature set before metamodel training, the sensitivity framework ensured that subsequent SHAP interpretations of Q_{50}, ACH_{INF}, and ΔT as dominant exposure determinants rest on physically meaningful and statistically robust inputs. This is further supported by the fact that qualitative SHAP rankings were found to be stable under the alternative feature sets considered during model development. This integrated sensitivity–machine learning approach provides an interpretable and scalable pathway to IAQ assessment in HEIs, empowering facility managers, policymakers, and engineers to promptly evaluate the IAQ impacts of different building designs and intervention scenarios, and to support evidence-informed decisions that enhance both health and sustainability within higher education campuses globally.

4.2 Enhancing interpretability: The role of SHAP values

SHAP values significantly enhance interpretability of ML metamodels for IAQ assessment, complementing recent applications in building energy modelling and control (Chen et al., 2023; Cui et al., 2024; Gao et al., 2025). This research marks a notable contribution as one of the first applications of SHAP-based interpretability to indoor exposure assessment in HEI buildings using a physics-driven metamodeling framework, complementing recent SHAP-enabled interpretability work in building energy and control domains (Gao et al., 2025). SHAP quantifies the marginal contribution of each predictor to individual $PM_{2.5}$ predictions, allowing both global rankings of key drivers and local, zone-level explanations of why specific exposures are high or low.

Across the three algorithms evaluated, Generalised Additive Models (GAM), Random Forest Regression (RFR), and Extreme Gradient Boosting (XGB), building airtightness (Q_{50}) consistently emerged as the primary driver of heating-season indoor $PM_{2.5}$ concentrations. GAM attributed almost 40% of the variation in infiltrated $PM_{2.5}$ to Q_{50}, while XGB and RFR attributed 32.5% and 33.6%, respectively. This consistent ranking across alternative high-accuracy models confirms that the dominance of envelope airtightness is stable under the reduced, non-collinear feature set and is therefore robust to reasonable changes in

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

model architecture. In practical terms, this detailed quantification confirms that improving airtightness is the primary lever for reducing infiltration-driven concentrations within the limits of the infiltration-dominated scenarios studied.

SHAP provides both global interpretability. At the global level (**Figure 5**), lower values of Q_{50} (tighter envelopes) are associated with negative SHAP values, indicating a strong reduction in predicted $PM_{2.5}$, whereas higher infiltration air change rates (ACH_{INF}) correspond to positive SHAP values that increase predicted concentrations. These patterns clarify how envelope sealing and infiltration dynamics jointly shift predicted concentrations towards or away from the dataset average. At the local level (**Figure 6**), SHAP decomposes each zone's prediction into contributions from Q_{50}, ACH_{INF}, ΔT, Aef:Vz, and v, highlighting whether a high concentration in a particular zone is driven more by a loose envelope, elevated infiltration, or unfavourable thermal conditions. This zone-specific decomposition enables facility managers to identify which physical drivers need to be changed, such as targeted sealing, ventilation adjustments, or operational controls, for specific rooms rather than applying uniform mitigation across the entire building stock.

By providing transparent and explainable ML models, SHAP can foster trust among facility managers, policymakers, and researchers. This clarity translates into actionable insights regarding which building characteristics and environmental factors are the strongest determinants of indoor pollution levels, thereby guiding practical decision-making in building design and management and highlighting critical microenvironmental disparities that warrant targeted intervention. In this way, integrating SHAP values with the ML metamodel significantly enhances the understanding of $PM_{2.5}$ exposure determinants in HEI buildings and supports evidence-based interventions for healthier indoor environments.

4.3. Microenvironment modelling for exposure assessment

Microenvironmental modelling was employed as a key approach in the study to estimate occupant exposure to indoor $PM_{2.5}$ from outdoor sources within HEI buildings, based on predictions from the machine learning metamodel. This method defines a microenvironment as a three-dimensional space where pollutant levels are uniform or exhibit consistent statistical properties over time, acknowledging that personal exposures vary significantly with individual time-activity patterns. To account for these varying exposure risks, four categories of Similar Time-Activity Groups (STGs) were established for HEI occupants: Academic Staff, Administrative Staff, Undergraduate Students, and Postgraduate Research (PGR) Students, informed by UK Higher Education Statistics Agency (HESA) data. Time–activity fractions for each STG were derived from HESA classifications and the Klepeis et al. profiles, and should therefore be interpreted as stylised, assumed schedules rather than empirically calibrated behaviour. These groups spend time across four main microenvironment types: offices, educational facilities, shared

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

facilities, and circulation areas, with the study focusing solely on exposure from outdoor-sourced $PM_{2.5}$. As these annual metrics are composite indicators derived from separate heating and non-heating season models, they provide an indicative rather than a statistically unified longitudinal profile.

Personal Exposure (Ei) was calculated using a time-weighted average method, multiplying predicted $PM_{2.5}$ concentrations by the proportion of time an individual spent in each microenvironment. Given that the underlying time–activity fractions are theoretical, these EI_i values represent scenario-based indicators of cohort-level exposure patterns rather than precise estimates of individual risk. The study analysed three scenarios of building airtightness (Q_{50}): Baseline, 7 $m^3/h{\cdot}m^2$, and 3 $m^3/h{\cdot}m^2$. Under the adopted assumptions, results suggested that Administrative Staff and PGR Students consistently experienced the highest indicative personal exposures due to their prolonged time in offices, with the exposure of Undergraduate Students significantly driven by educational facilities. This implies that mitigation measures such as enhanced filtration or targeted ventilation upgrades should prioritise office-dominated spaces used by Administrative and PGR Staff, while lecture halls and other educational facilities are critical intervention targets for the largest student cohort. Improving building airtightness progressively reduced personal exposure across all groups, with the tightest envelope (Q_{50} = 3 $m^3/h{\cdot}m^2$) leading to the most substantial reductions.

Beyond individual exposure, the study also quantified Population-Weighted Exposure (PWE) to indoor $PM_{2.5}$, aggregating individual microenvironment exposures by considering the population size and distribution of occupant groups. Population data, based on maximum design occupancy for each zone, was obtained from the University's Estates and Facilities Management office. At baseline airtightness, the overall annual PWE was 9.6 $\mu g/m^3$, nearing the WHO 2005 guideline of 10 $\mu g/m^3$, with specific microenvironments like PGR offices, administrative offices, and lecture halls exceeding this threshold. A moderate improvement in airtightness (Q_{50} = 7 $m^3/h{\cdot}m^2$) resulted in an 11.5% reduction in total annual PWE to 8.5 $\mu g/m^3$, bringing all microenvironments below the 10 $\mu g/m^3$ guideline. Further tightening to a well-sealed envelope (Q_{50} = 3 $m^3/h{\cdot}m^2$) led to a 32.3% reduction in total annual PWE, reaching 6.5 $\mu g/m^3$. However, even with well-sealed buildings (Q_{50} = 3 $m^3/h{\cdot}m^2$), approximately 88% of zones still exceeded the latest, more stringent WHO 2021 guideline of 5 $\mu g/m^3$. This persistent non-compliance confirms that while tighter envelopes substantially reduce infiltration-driven exposure, airtightness alone may not be sufficient in environments with high outdoor pollution (Afroz, R. et al., 2023).

4.4. Microenvironmental exposure and differentiated policy implications

Microenvironmental modelling played a crucial role in providing detailed insights into occupant exposure to indoor $PM_{2.5}$ from outdoor sources within HEI buildings, ultimately revealing pronounced disparities in exposure across different airtightness scenarios, diverse microenvironments, and distinct

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

occupant categories. This approach leveraged Similar Time-Activity Groups each with unique time-activity patterns and preferential use of specific building spaces. In this study, because these time-activity patterns are based on assumed, stylised schedules rather than measured behaviour, the resulting disparities should be interpreted as scenario-based indicators of relative cohort exposure rather than precise estimates of absolute personal risk. For instance, Administrative Staff, who typically spend extensive time in office settings, consistently exhibited higher absolute exposure levels at baseline and benefited more significantly from envelope tightening measures. This suggests that office-dominated spaces used by Administrative and PGR staff are priority candidates for targeted mitigation measures such as local filtration or upgraded ventilation. Conversely, lecture halls and other educational spaces emerged as significant exposure hotspots for undergraduates, directly reflecting their extended occupancy patterns within these environments. This granular understanding of varying exposure profiles within HEIs contrasts with simpler residential contexts, where different factors might dominate overall exposures, thereby highlighting the necessity for tailored policy responses.

Building upon these findings, the study advocates for integrated, differentiated policy responses to address indoor $PM_{2.5}$ exposure in HEI buildings: rather than treating airtightness (Q_{50}) as a single primary mitigation lever, policies should combine envelope sealing with supplementary measures such as controlled mechanical ventilation, advanced filtration, or mixed-mode systems. Given that building-level measures alone are unlikely to deliver full compliance with the WHO 2021 guideline in polluted urban contexts, they should be implemented in parallel with broader urban air-quality policies, such as local clean air zones and traffic-emission controls, that reduce outdoor $PM_{2.5}$ and thereby lower the infiltration-driven component of indoor exposure across HEI campuses.

4.5 Limitations and future work

This research has several limitations that point to priorities for future work and that should inform interpretation of the results. The analysis relied on a predominantly deterministic modelling framework, using fixed input values rather than propagating full parameter uncertainty, thus future work should employ probabilistic methods to quantify aleatoric and epistemic uncertainties in key variables such as deposition rates, building airtightness (Q_{50}), and wind pressure coefficients, and to express exceedance of the WHO 2021 guideline with explicit confidence bands rather than as exact threshold crossings.

A further limitation is the paucity and inconsistency of high-quality building data in HEI stocks, including envelope airtightness, HVAC specifications, detailed occupancy schedules, and window operation behaviours, which necessitated estimating Q_{50} from standards and construction year rather than direct tests and emphasises the need for estates teams to develop IAQ-relevant data repositories. The coupled CONTAM–EnergyPlus co-simulation focused on infiltration-driven airflow with basic exhaust systems,

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

intentionally excluding detailed Air Handling Units and filtration in high-occupancy spaces such as lecture halls, thus the exposure and PWE estimates are most directly applicable to naturally ventilated or infiltration-dominated HEI buildings and may underrepresent mitigation potential where mechanical ventilation and filtration are present; future studies should explicitly model such systems to quantify the benefits of integrated envelope–ventilation strategies. Consequently, the reported exposure and PWE estimates are bounded to naturally ventilated or infiltration-dominated HEI buildings. These results must not be extrapolated to mechanically ventilated or filtered buildings without additional modelling. Future work is required to integrate these mechanical removal mechanisms into the metamodel to quantify the enhanced mitigation potential they provide.

Model validation was also constrained: indirect validation of indoor temperature did not meet ASTM benchmarks, and direct empirical validation of indoor $PM_{2.5}$ was prevented by COVID-19 restrictions, so future work should prioritise in-situ measurements to validate both concentration fields and exposure metrics, and current exposure estimates should be regarded as model-based scenarios rather than fully validated absolute levels. Microenvironment definitions and Similar Time-Activity Groups were derived from theoretical assumptions and general time-activity patterns, and the underlying occupancy and behavioural profiles (including negligible window opening during the heating season) were stylised rather than measured, implying that cohort-specific exposure differences are illustrative of potential disparities under the assumed schedules rather than definitive estimates of real-world exposure distributions; more advanced statistical classification of $PM_{2.5}$ time series and sensor- or survey-based behaviour data are needed to refine these groupings.

Finally, the metamodels were trained and tested on four buildings (2,729 zones) with one held-out building, which yielded high predictive accuracy but limits the breadth of structural and operational diversity sampled; expanding the number and variety of buildings, and incorporating additional predictors such as detailed mechanical ventilation characteristics and indoor source terms, would improve generalisability. Taken together, these limitations indicate that while the infiltration-driven indoor concentration field and the dominance of Q_{50}, ACH_{INF}, and ΔT are physically robust findings, the cohort-level exposure metrics and policy implications should be interpreted as conditional on current data and assumptions and progressively refined as richer empirical datasets and more comprehensive models become available.

5. Conclusions

This study developed an integrative framework coupling CONTAM–EnergyPlus co-simulation with XGBoost metamodelling and SHAP-based interpretability for indoor PM2.5 exposure assessment in HEI buildings. The principal findings are:

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

1. **Metamodel accuracy**: The XGBoost metamodel achieved high predictive accuracy ($R \approx 0.95$ on testing data; $R^2 > 0.90$ on a held-out building), providing a computationally efficient surrogate for physics-based simulation across 2,729 zones.
2. **Dominant exposure drivers**: SHAP analysis consistently identified building airtightness (Q_{50}), infiltration air change rate (ACHINF), and indoor–outdoor temperature difference (ΔT) as the primary determinants of heating-season indoor $PM_{2.5}$, with transparent global and local interpretations supporting targeted intervention guidance.
3. **Exposure disparities**: Microenvironmental modelling indicated pronounced exposure heterogeneity within the stylised schedules of the occupant groups, with offices dominating staff exposure and educational facilities driving student exposure.
4. **Policy Implications**: While enhancing envelope airtightness to Q_{50}=3 $m^3/h \cdot m^2$ reduces population-weighted annual exposure by up to 32.3%, the study reveals that approximately 88% of indoor zones still exceed the stringent WHO 2021 annual guideline of 5 μg/m3. This persistent non-compliance demonstrates the inherent limitations of using airtightness as a singular primary mitigation lever in environments with high outdoor pollution. Consequently, the findings advocate for integrated, differentiated policy responses that combine envelope sealing with supplementary mechanical measures, such as controlled ventilation and advanced air filtration. Furthermore, since building-level interventions alone fail to achieve universal compliance, these results underscore the necessity of urban-scale interventions, including Clean Air Zones and broader emission control policies, to reduce ambient $PM_{2.5}$ concentrations at the source and lower the baseline outdoor exposure against which all building-level measures operate.

Within acknowledged limitations, including exclusion of detailed AHU/filtration modelling and use of stylised occupancy schedules, this framework offers a physically grounded, interpretable basis for screening HEI building stocks and designing evidence-based IAQ intervention strategies.

References

Abdalla, T. & Peng, C. (2025). Spatiotemporal variability of PM2.5 infiltrations in higher education buildings: A multizone air-thermal co-simulation analysis. Journal of Building Engineering, 113473. https://doi.org/10.1016/j.jobe.2025.113473

Abdul Shakor, A.S., Pahrol, M.A. & Mazeli, M.I. (2020). Effects of Population Weighting on PM10 Concentration Estimation. Journal of Environmental and Public Health, 2020. https://doi.org/10.1155/2020/1561823

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

Afroz, R., Guo, X., Cheng, C.-W., Delorme, A., Duruisseau-Kuntz, R., & Zhao, R. (2023). Investigation of indoor air quality in university residences using low-cost sensors. Environmental Science : Atmospheres, 3(2), 347–362. https://doi.org/10.1039/D2EA00149G

Apte, J. S., Brauer, M., Cohen, A. J., Ezzati, M., & Pope, C. A. (2018). Ambient PM2.5 Reduces Global and Regional Life Expectancy. Environmental Science and Technology Letters, 5(9), 546–551. https://doi.org/10.1021/acs.estlett.8b00360

Aunan, K., Ma, Q., Lund, M.T. & Wang, S. (2018). Population-weighted exposure to PM2.5 pollution in China: An integrated approach. Environment International, 120, pp.111–120. https://doi.org/10.1016/j.envint.2018.07.042

Branco, P., Sousa, S., Dudzińska, M., Ruzgar, D., Mutlu, M., Panaras, G., Papadopoulos, G., Saffell, J., Scutaru, A., Struck, C., & Weersink, A. (2024). A review of relevant parameters for assessing indoor air quality in educational facilities. Environmental research, 119713. https://doi.org/10.1016/j.envres.2024.119713

Burnett R, Chen H, Szyszkowicz M, Fann N, Hubbell B, Pope CA, Apte JS, Brauer M, Cohen A, Weichenthal S, Coggins J, Di Q, Brunekreef B, Frostad J, Lim SS, Kan H, Walker KD, Thurston GD, Hayes RB, Lim CC, Turner MC, Jerrett M, Krewski D, Gapstur SM, Diver WR, Ostro B, Goldberg D, Crouse DL, Martin RV, Peters P, Pinault L, Tjepkema M, van Donkelaar A, Villeneuve PJ, Miller AB, Yin P, Zhou M, Wang L, Janssen NAH, Marra M, Atkinson RW, Tsang H, Quoc Thach T, Cannon JB, Allen RT, Hart JE, Laden F, Cesaroni G, Forastiere F, Weinmayr G, Jaensch A, Nagel G, Concin H, Spadaro JV. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proc Natl Acad Sci U S A 115(38):9592-9597. https://doi.org/10.1073/pnas.1803222115

CIBSE (2022). TM23 Testing buildings for air leakage.

COMEAP (2009). Long-Term Exposure to Air Pollution: Effect on Mortality. A report by the Committee on the Medical Effects of Air Pollutants.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu. pp.785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Xiao, F., Guo, F., & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*. https://doi.org/10.1016/j.adapen.2023.100123.

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

758 Cui, X., Lee, M., Koo, C., & Hong, T. (2024). Energy consumption prediction and household feature
759 analysis for different residential building types using machine learning and SHAP: Toward energy-
760 efficient buildings. *Energy and Buildings*. https://doi.org/10.1016/j.enbuild.2024.113997.

761 Dols, W.S. & Polidoro, B.J. (2020). CONTAM User Guide and Program Documentation - Version 3.4
762 (NIST Technical Note 1887).

763 Emmerich, S. (2001). Validation of Multizone IAQ Modeling of Residential-Scale Buildings: A Review,
764 ASHRAE Transactions, [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=860837

765 Emmerich, S.J., Ng, L.C. & Dols, W.S. (2019). Simulation analysis of potential energy savings from air
766 sealing retrofits of U.S. Commercial buildings. ASTM Special Technical Publication, STP 1615, pp.61–
767 70. https://doi.org/10.1520/STP161520180021.

768 Erlandson, G., Magzamen, S., Carter, E., Sharp, J. L., Reynolds, S. J., & Schaeffer, J. W. (2019).
769 Characterization of Indoor Air Quality on a College Campus: A Pilot Study. International Journal of
770 Environmental Research and Public Health, 16(15), 2721. https://doi.org/10.3390/ijerph16152721

771 Feustel, H.E. (1999). COMIS-an international multizone airflow and contaminant transport model.
772 Energy and Buildings, 30(1), pp.3–18. https://doi.org/10.1016/S0378-7788(98)00043-7

773 Gaidajis, G. & Angelakoglou, K. (2009). Indoor air quality in university classrooms and relative
774 environment in terms of mass concentrations of particulate matter. Journal of Environmental Science
775 and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering, 44(12), pp.1227–1232.
776 https://doi.org/10.1080/10934520903139936.

777 Gao, Y., Hu, Z., Yamate, S., Otomo, J., Chen, W., Liu, M., Xu, T., Ruan, Y., & Shang, J. (2025). Unlocking
778 predictive insights and interpretability in deep reinforcement learning for Building-Integrated
779 Photovoltaic and Battery (BIPVB) systems. *Applied Energy*.
780 https://doi.org/10.1016/j.apenergy.2025.125387.

781 Gillott, M., Loveday, D., White, J. A., Wood, C., Chmutina, K. & Vadodaria, K. (2016). Improving the
782 airtightness in an existing UK dwelling: the challenges, the measures and their effectiveness, Building
783 and Environment, 95, pp. 227–239. doi: 10.1016/j.buildenv.2015.08.017

784 HESA (2021). HESA - Higher Education Statistics Agency. Available at: https://www.hesa.ac.uk/stats

785 Hancock, T. (2018). Creating healthy cities and communities. Canadian Medical Association Journal,
786 190, E206 - E206. https://doi.org/10.1503/cmaj.180102

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

787 Hensen, J. & Lamberts, R. (2011). Building Performance Simulation for Design and Operation. London;
788 New York: Spon Press.

789 Huang X, Steinmetz J, Marsh EK, Shoemaker N, Yip J, Harley CA, Zheng SIH, Kaufman JD, Ward MH,
790 Huang K, Robichaux KE, McKinley A, Pautler RA, Simon JE, Sahle YA, Ma Y, Cathey AL, Yu JH, Thomas
791 CR, Wheeler KA, Kinney PL, McConnell R, Brauer M, Cohen AJ, Brodeur SK, Shakya MC, Hay SI. (2025).
792 A systematic review with a Burden of Proof meta-analysis of health effects of long-term ambient fine
793 particulate matter (PM2.5) exposure on dementia. Nature Aging, 5, pp.897-908.
794 https://doi.org/10.1038/s43587-025-00844-y

795 Im U, Ye Z, Schuhen N, Crippa M, Colette A, Viana M, Terrenoire E, Cuvelier C, Langerock B, Lu Y,
796 Tarrason L, Bessagnet B, Liu X. (2025). Europe will struggle to meet the new WHO Air Quality
797 Guidelines. npj Clean Air 1: Article 13. https://doi.org/10.1038/s44407-025-00013-w

798 Jones, B., Das, P., Chalabi, Z., Davies, M., Hamilton, I., Lowe, R., Mavrogianni, A., Robinson, D. & Taylor,
799 J. (2015). Assessing uncertainty in housing stock infiltration rates and associated heat loss: English and
800 UK case studies, Building and Environment, 92, pp. 644–656.
801 https://doi.org/10.1016/j.buildenv.2015.05.033

802 Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., Behar, J. V., Hern, S.
803 C. & Engelmann, W. H. (2001). The National Human Activity Pattern Survey (NHAPS): a resource for
804 assessing exposure to environmental pollutants, Journal of Exposure Science & Environmental
805 Epidemiology, 11(3), pp. 231–252. https://doi.org/10.1038/sj.jea.7500165

806 Krittanawong, C., Qadeer, Y., Hayes, R., Wang, Z., Virani, S., Thurston, G., & Lavie, C. (2023). PM2.5 and
807 Cardiovascular Health Risks. Current problems in cardiology, 101670.
808 https://doi.org/10.1016/j.cpcardiol.2023.101670

809 Lama, S., Fu, C., & Lee, A. (2022). Indoor Air Quality (IAQ) Evaluation of Higher Education Learning
810 Environments. Journal of Smart Buildings and Construction Technology, 4(1), 1–14.
811 https://doi.org/10.30564/jsbct.v4i1.4042

812 Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions, arXiv, May 22,
813 version 2 (Nov 25, 2017). https://doi.org/10.48550/arXiv.1705.07874.

814 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Muller, A.,
815 Nothman, J., Louppe, G., Prenttenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,
816 Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

817 Python. Journal of Machine Learning, 12, 2825–2830.
818 https://doi.org/https://doi.org/10.48550/arXiv.1201.0490

819 Pope, C., Coleman, N., Pond, Z., & Burnett, R. (2019). Fine particulate air pollution and human
820 mortality: 25+ years of cohort studies. Environmental research, 108924.
821 https://doi.org/10.1016/j.envres.2019.108924

822 Pruszyński, J., Cianciara, D., Włodarczyk-Pruszyńska, I., Górczak, M., & Padzińska-Pruszyńska, I.
823 (2023). Indoor Generation Era. Risks and challenges. Journal of Education, Health and Sport.
824 https://doi.org/10.12775/jehs.2023.48.01.002

825 Rogowski, C.B., Lim, Y.H., de Lange, T., Peters, A., Andersen, Z.J., Brunekreef, B., Chen, J., Hoffmann,
826 B., Katsouyanni, K. & Bellavia, A. (2025). Long-term air pollution exposure and incident dementia: a
827 systematic review and meta-analysis. The Lancet Planetary Health, 9(8).
828 https://doi.org/10.1016/S2542-5196(25)00118-4

829 Sarbu, I. & Pacurar, C., 2015. Experimental and numerical research to assess indoor environment quality and
830 schoolwork performance in university classrooms. Building and Environment, 93, pp.141–154.
831 https://doi.org/10.1016/j.buildenv.2015.06.022.

832 Servén, D., & Brummitt, C. (2018). pyGAM: Generalized Additive Models in Python (0.8.0). Zenodo.
833 https://doi.org/10.5281/zenodo.1208723

834 Sherman, M.H. & Dickerhoff, D. (1998). Airtightness of US dwellings. Transactions- American Society of
835 Heating Refrigerating and Air Conditioning Engineers, 104, pp.1359–1367.

836 Taylor, J., Mavrogianni, A., Davies, M., Das, P., Shrubsole, C., Biddulph, P. & Oikonomou, E. (2015).
837 Understanding and mitigating overheating and indoor PM 2.5 risks using coupled temperature and
838 indoor air quality models, Building Services Engineering Research and Technology, 36(2), pp. 275–289.
839 https://doi.org/10.1177/0143624414566474

840 Taylor, J., Shrubsole, C., Davies, M., Biddulph, P., Das, P., Hamilton, I., Vardoulakis, S., Mavrogianni, A.,
841 Jones, B. & Oikonomou, E. (2014). The modifying effect of the building envelope on population exposure
842 to PM 2.5 from outdoor sources, Indoor Air, 24(6), pp. 639–651. https://doi.org/10.1111/ina.12116

843 Watson, A.Y., Bates, R.R. & Kennedy, D. (1988). Assessment of Human Exposure to Air Pollution:
844 Methods, Measurements, and Models. In: Air Pollution, the Automobile, and Public Health. Available
845 at: https://www.ncbi.nlm.nih.gov/books/NBK218147/ (accessed online 22/10/2025)

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632

World Health Organization. (2025). WHO unveils updated global database of air quality standards. World Health Organization, Geneva. Available at: https://www.who.int/news/item/26-02-2025-who-unveils-updated-global-database-of-air-quality-standards (accessed online 22/10/2025).

Xian, J. & Wang, Z. (2024). A physics and data co-driven surrogate modeling method for high-dimensional rare event simulation. Journal of Computational Physics, 510, 113069. https://doi.org/10.1016/j.jcp.2024.113069

This manuscript was accepted for publication by ***Journal of Building Engineering*** on 15 February 2026. https://doi.org/10.1016/j.jobe.2026.115632