# Supervised learning-based model of a multi-dwelling heat pump under normal and faulty operating conditions

***Michael Khairallah[a], Assaad Zoughaib[a], Egoi Ortego[a]***

*[a] Center of Energy, Environment, Processes (CEEP) ; Mines Paris - PSL, Versailles, France.*

**Abstract:**

The widespread electrification of heating in collective housing through reversible heat pumps requires robust, scalable solutions for Fault Detection and Diagnosis. While high-fidelity physical digital twins provide the detail necessary to characterize complex failure modes, their computational cost makes them unsuitable for real-time cloud-based supervisory platforms. This work proposes a supervised learning–based reduced-order model of an air-to-water heat pump trained on data generated by a validated Modelica/Dymola digital twin. A structured data generation methodology combining empirical correlations for simulation initialization with a full-factorial design of experiments is developed, producing 190 usable steady-state operating points across the heating-mode envelope. Five surrogate architectures are compared: Random Forest, Gradient Boosting, Gaussian Process, and a feedforward neural network along with a physics-informed neural network variant embedding thermodynamic constraints into the loss function. The Gradient Boosting and Gaussian Process models achieve sub-0.5% MAPE on heating capacity and electrical power against the physical model, with $R^2$ exceeding 0.999. Validation against 25 experimental test bench measurements confirms that the Gaussian Process generalizes best to real-world conditions, predicting heating capacity with 3.4% MAPE and 92% of points within ±10%. The physics-informed variant reduces the COP thermodynamic discrepancy by 64% compared to the standard network. All surrogates achieve speedups of five to six orders of magnitude over the physical model, enabling real-time monitoring of large heat pump fleets.

**Keywords:**

Heat pump, Fault Detection and Diagnosis, Reduced-Order Modeling, Supervised Learning, Digital Twin.

## 1. Introduction

The decarbonization of the building sector is driving a rapid transition from fossil fuel–based heating to electrically driven heat pump systems, particularly in multi-dwelling residential buildings. In France, the RE2020 regulation mandates significant reductions in greenhouse gas emissions from new constructions, positioning air-to-water heat pumps as a key technology for collective housing [1]. As heat pump deployments scale from individual units to fleets serving hundreds of dwellings, ensuring reliable long-term performance becomes a critical operational challenge. Field studies have shown that faults such as refrigerant leakage, heat exchanger fouling, and compressor degradation can reduce seasonal performance by 10–30% if left undetected [2]. This motivates the development of robust, scalable solutions for Fault Detection and Diagnosis (FDD) that can be deployed across large building stocks.

Physics-based digital twins of vapor compression systems provide the detailed thermodynamic modeling necessary to characterize complex failure modes and predict system behavior under diverse operating conditions. High-fidelity dynamic models built in environments such as Modelica/Dymola can accurately reproduce the heat pump cycle including two-phase heat exchange, compressor performance, and expansion device behavior [3]. However, the computational cost of these models, typically requiring seconds to minutes per steady-state evaluation, renders them unsuitable for direct integration into real-time cloud-based supervisory platforms that must monitor dozens or hundreds of units simultaneously. This creates a fundamental tension between model fidelity and operational deployability.

Accordingly, this work develops and benchmarks data-driven surrogate models derived from a calibrated Dymola heat pump model, evaluates their predictive accuracy, thermodynamic consistency, and computational efficiency, and validates the best-performing approach against independent experimental measurements for scalable cloud-based FDD deployment.

## 2. Literature Review

This section provides an overview of the most relevant studies on digital-twin modeling, model reduction, and data-driven fault detection for vapor-compression heat pump systems. It first reviews dynamic modeling strategies and reduction approaches, then examines surrogate-model development and FDD methodologies. The objective is to position the present work with respect to current advances and to highlight the remaining methodological gaps addressed in this study.

### 2.1. Dynamic modeling and model reduction for heat pumps

Dynamic modeling of vapor compression systems has been extensively reviewed by Rasmussen [3] and Li et al. [4], who classify heat exchanger formulations into lumped-parameter, moving-boundary (MB) , and finite-volume (FV) paradigms. The MB method, formalized by Willatzen et al. [5] and extended to control-oriented models by McKinley and Alleyne [6], divides heat exchangers into variable-length zones and produces models with 5–9 state variables, compared to 30–90+ for FV discretizations. Bendapudi et al. [7] showed that MB formulations execute approximately three times faster than FV for centrifugal chiller heat exchangers with comparable thermal accuracy, establishing MB as the preferred approach for real-time and hardware-in-the-loop applications.

The Modelica language has become the dominant platform for component-level heat pump modeling, with several mature libraries available. The TIL library [8] provides grey-box component models interfaced with REFPROP-compatible fluid properties. The open-source ThermoCycle library [9] emphasizes numerical robustness for two-phase flow using finite-volume discretization. More recently, the AixLib library [10] integrates both simplified and detailed heat pump models with Python-based calibration tools, while VCLib [11] provides a modular, documented architecture for education and research. These libraries enable the construction of detailed digital twins but at a computational cost that limits their use to offline analysis or small-scale simulation campaigns.

Formal model order reduction techniques have been applied to vapor compression heat exchangers. Ma et al. [12] developed a POD-DEIM framework for centrifugal chiller models, achieving 80% computation time reduction compared to standard FV discretization with negligible prediction errors. However, these projection-based methods are intrusive, since they require access to the model's internal equations, and have not been widely applied to complete heat pump system models in the context of FDD.

The concept of a digital twin applied to heat pumps has emerged primarily since 2020. Seifert et al. [13] presented a holistic digital twin framework combining Python, TRNSYS, and Modelica within a cloud-based IoT platform, classifying models by lifecycle phase: detailed models for development and simplified models for field deployment. Aguilera et al. [14] demonstrated a digital twin framework for a large-scale ammonia heat pump, where online model calibration decreased performance estimation errors by 3–17 percentage points compared to single-calibration baselines. Despite this progress, applications to multi-dwelling collective housing heat pumps remain scarce.

### 2.2. Surrogate models and FDD strategies for heat pumps

Supervised machine learning models trained on physics-based simulation data have shown strong potential for replacing computationally expensive heat pump models. Zhao et al. [15] trained feedforward neural networks on TRNSYS-generated data for ground-coupled heat pumps, achieving errors below 5% with a 100× speedup over the original simulation. Ablanque et al. [16] developed Gaussian process and MLP surrogates for aircraft vapor compression systems modeled in Dymola, predicting pressures, temperatures, and cooling capacity with substantial CPU time reduction. Bortoff et al. [17] proposed physics-augmented neural networks that learn residual dynamics between a low-fidelity physics model and high-fidelity simulation data, achieving accurate prediction of detailed finite-volume states from partial observations. A physics-constrained deep learning framework for dynamic VCS modeling [18] achieved COP errors of 1.5% with 87–95% computational time reduction by integrating modular component-level deep learning models via physical conservation laws.

FDD for vapor compression systems has evolved from rule-based methods to modern data-driven approaches. The foundational work of Rossi and Braun [19] introduced statistical rule-based detection using residuals between measured and predicted states, capable of detecting approximately 5% refrigerant loss. Li and Braun [20] extended this with virtual sensor decoupling to handle multiple simultaneous faults. More recently, supervised learning methods have achieved high classification accuracy: support vector machines applied to the ASHRAE RP-1043 benchmark dataset reached 99.5% correct diagnosis rates, while Sun et al. [21] applied deep learning to early-stage gradual fault detection in air-source heat pumps.

A critical limitation of data-driven FDD is the transferability gap between laboratory or simulation environments and real-world installations. Bode et al. [22] demonstrated that ML fault detectors trained on NIST experimental data performed near-randomly when applied to a real-world building heat pump, underscoring the need for training data representative of actual deployment conditions. To address data scarcity, several au-

thors have used physical models to generate synthetic fault data. Cheung and Braun [23] developed empirical fault impact models for EnergyPlus across 40+ scenarios. Llopis-Mengual et al. [24] trained an SVM classifier on simulation-generated data with individually and simultaneously imposed faults, achieving 82% balanced accuracy on experimental validation. Navarro-Peris and collaborators [25] further studied feature selection for detecting multiple simultaneous soft faults using simulation data from a detailed vapor compression cycle model.

Despite the significant advances reviewed above, several gaps remain at the intersection of digital twin modeling, surrogate construction, and FDD for heat pumps. First, most surrogate models for heat pump performance are trained on experimental data alone, which limits coverage of the operating envelope and makes it difficult to systematically explore fault conditions. The few works that train on simulation-generated data [16] do not target FDD applications or collective housing heat pumps. Second, the transferability gap identified by Bode et al. [22] highlights the need for system-specific training data that captures the particular characteristics of the target installation — a role naturally filled by a calibrated digital twin. Third, while Llopis-Mengual et al. [24] have demonstrated simulation-trained FDD for residential air conditioning, no existing work integrates all elements of the pipeline proposed here: a validated Modelica digital twin serving as a data generator, empirical correlations ensuring robust simulation convergence, structured DoE for efficient exploration of the operating envelope, and a reduced-order surrogate model comparison under data-limited conditions.

The present work addresses these gaps by developing a surrogate modeling framework for an air-to-water heat pump used in collective housing, leveraging a calibrated Dymola digital twin as the sole data source. This work proposes a supervised learning–based reduced-order model of an air-to-water heat pump trained entirely on data generated by a validated Dymola digital twin. The main contributions are:

- A structured data generation methodology combining empirical correlations for simulation initialization with a full-factorial design of experiments covering the heating-mode operating envelope of a multi-dwelling heat pump
- A comparative evaluation of surrogate architectures for predicting key thermodynamic outputs (pressures, capacity, power, temperatures) from measurable boundary conditions, validated against both the physical model and experimental test bench measurements;
- A surrogate modeling framework designed to support fault detection and diagnosis by enabling the integration of fault parameters as additional input features, with preliminary fault scenarios demonstrating the approach's potential for real-time FDD in cloud-connected supervisory platforms.

# 3. Methodology

The methodology is structured in four stages. First, the physical model and its validation are summarized. Second, empirical correlations derived from experimental data are introduced to initialize the simulations robustly. Third, the design of experiments used to generate training data is described. Finally, four surrogate modeling approaches are presented and compared.

## 3.1. Physical model of the heat pump

The system under study is a reversible air-to-water heat pump designed for collective housing applications. The unit uses propane (R-290) as the working fluid in a conventional vapor compression cycle. The main components are: a scroll compressor with one or two stages (one or two active compressors), a brazed plate heat exchanger serving as the condenser (propane-to-water), a round-tube finned-coil heat exchanger serving as the evaporator (propane-to-air), and a thermostatic expansion valve (TXV) . Figure 1 shows the Dymola model layout of the heat pump cycle

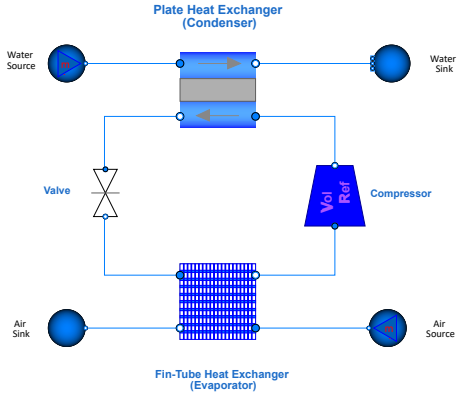


***Figure 1:*** *Heat pump schematic.*

A detailed dynamic model of the heat pump cycle was developed in Modelica and simulated using Dymola 2025x. The condenser and evaporator models [26] use simplified correlations for heat transfer coefficients and neglect pressure drops on both sides, a simplification that substantially reduces computational cost while preserving the dominant thermodynamic behavior. The compressor model relies on manufacturer-provided volumetric and isentropic efficiency maps. The expansion valve is modeled as an isenthalpic device.

The simulation employs a fixed-step explicit Euler solver with a time step of 0.05 s. To follow the behavior of the TXV, the model does not impose the superheat and subcooling as fixed boundary conditions in the classical sense. Instead, a proportional feedback controller adjusts the cycle pressures dynamically until the computed superheat and subcooling match prescribed target values. Formally, the low-side and high-side pressures evolve as:

$$\frac{dP_l}{dt} = K_{sh}\left(\Delta T_{sh,sim} - \Delta T_{sh,target}\right) \tag{1}$$

$$\frac{dP_h}{dt} = K_{sc}\left(\Delta T_{sc,target} - \Delta T_{sc,sim}\right) \tag{2}$$

where $P_h$, $P_l$ are the high (condenser) and low (evaporator) pressures respectively, $\Delta T_{sh}$ and $\Delta T_{sc}$ are the superheat and subcooling levels (targeted or and simulated), $K_{sh}$ & $K_{sc}$ are proportional gains scaled by compressor speed, and the derivatives are clamped to $\pm 20$ kPa/s to prevent numerical instability. This formulation allows the model to converge toward a physically consistent steady-state operating point for any combination of boundary conditions, provided that reasonable initial guesses for the pressures are supplied.

The calibrated model was validated against $28$ operating points for which simulation convergence was achieved, covering air inlet temperatures from $-7$ to $+7\ ^\circ C$, water setpoints from $35$ to $65^\circ C$ , and both single- and dual-compressor configurations. The model reproduces the experimental heating capacity with a MAPE of 5.0% and the compressor electrical power within a MAPE of $4.1\%$. Condensing and evaporating temperatures are reproduced within 2 K and 1 K on average, respectively. Overall, 88% of points fall within $\pm 10\%$ for heating capacity and 96% for compressor power.

## 3.2. Empirical correlations for simulation initialization

A practical challenge in generating large datasets from the physical model is that each simulation requires initial values for the high and low pressures, as well as target values for the superheat and subcooling. Poor initial guesses can lead to convergence failure or excessively long simulation times. To address this, four linear regression correlations were fitted on the 25 validated experimental operating points.

***Table 1:*** *Empirical correlations used for simulation initialization.*

| Output | Correlation | $R^2$ | MAE |
|---|---|---|---|
| $P_{h,initial}$ [bar] | $-0.255\,T_{w,sp} + 0.596\,T_{w,in} + 0.562\,N + 2.723$ | 0.983 | 0.40 |
| $P_{l,initial}$ [bar] | $0.104\,T_{air} - 0.356\,N + 4.026$ | 0.939 | 0.08 |
| $\Delta T_{sh,target}$ [K] | $2.281\,N - 0.050\,T_{air} + 6.890$ | 0.878 | 0.37 |
| $\Delta T_{sc,target}$ [K] | $1.901\,N - 0.150\,T_{w,sp} + 0.176\,T_{w,in} - 4.22 \times 10^{-4}\,\dot{V}_w + 2.561$ | 0.812 | 0.20 |

The high-pressure correlation is dominated by the water inlet temperature ($+0.60$ bar/$^\circ C$), which directly determines the required condensing temperature. The low-pressure correlation depends primarily on the air temperature ($+0.10$ bar/$^\circ C$), reflecting the evaporator thermal source. The superheat correlation captures the effect of compressor staging ($N$): activating a second compressor increases the refrigerant mass flow rate, which tends to starve the evaporator and raise the superheat by approximately $2.3$ K per additional compressor. The subcooling is influenced by a combination of compressor count, water temperatures, and flow rate.

These correlations serve two distinct roles. The pressure correlations ($P_{h,initial}$ and $P_{l,initial}$) provide initial guesses that place the solver close to the expected steady-state solution, improving convergence speed. The superheat and subcooling correlations define the operating targets that the model's internal pressure controller drives toward, following the real TXV behavior across the operating envelope. It is worth noting that these correlations are not inputs to the surrogate model. They are artifacts of the simulation procedure, necessary to generate the training data efficiently, but absent from the final reduced-order model which learns the direct mapping from boundary conditions to thermodynamic outputs.

## 3.3. Design of experiments and data generation

A full-factorial design was adopted to systematically explore the heating-mode operating envelope. The air temperature range was extended beyond the experimental validation domain ($-7$ to $+7^\circ C$) to cover $-10$ to $+15^\circ C$, representing conditions typical of the French metropolitan climate. The relative humidity was varied

across three levels involving dry to near-saturated conditions. The water inlet temperature covers floor heating ($30°C$), low-temperature radiators ($40°C$), and standard radiators ($50°C$). Higher water temperatures ($60°C$) were attempted but resulted in systematic numerical divergence of the simplified physical model and were therefore excluded from the dataset.

***Table 2:*** *Design of experiments: input variables and ranges.*

| Variable | Symbol | Unit | Levels | Count |
|---|---|---|---|---|
| Air temperature | $T_{air}$ | $°C$ | $-10, -5, 0, 5, 10, 15$ | 6 |
| Relative humidity | $\varphi$ | % | $60, 80, 95$ | 3 |
| Air flow rate | $\dot{V}_{air}$ | $m^3/h$ | 8000 for $N = 1$<br>16000 for $N = 2$ | 2 |
| Water temperature pair | $T_{w,in}/T_{w,setpoint}$ | $°C$ | $(30/35), (40/45), (50/55), (60/65)$ | 4 |
| Water flow rate | $\dot{V}_w$ | $L/h$ | $4500, 5500$ for $N = 1$<br>$7000, 8000$ for $N = 2$ | 2 |

The $288$ $(6 \times 4 \times 3 \times 2 \times 2)$ operating points were organized into six batches of approximately 50 simulations each, prioritized by an estimated convergence risk based on the temperature lift. An automated Python pipeline interfacing with Dymola via its scripting API managed the batch execution.

Of the 288 simulations, 61 reached steady state within the maximum simulation time of 80 seconds and 129 produced a trajectory that ran to the time limit without triggering the steady-state termination criterion. The remaining points either failed during initialization or corresponded to the numerically unstable $T_{w,in} = 60°C$ conditions. The 190 usable points were retained after verifying that the "partial" simulations had achieved their superheat targets within 0.2 K on average, indicating near-steady-state conditions despite the subcooling controller not having fully converged, as shown in the figure below.

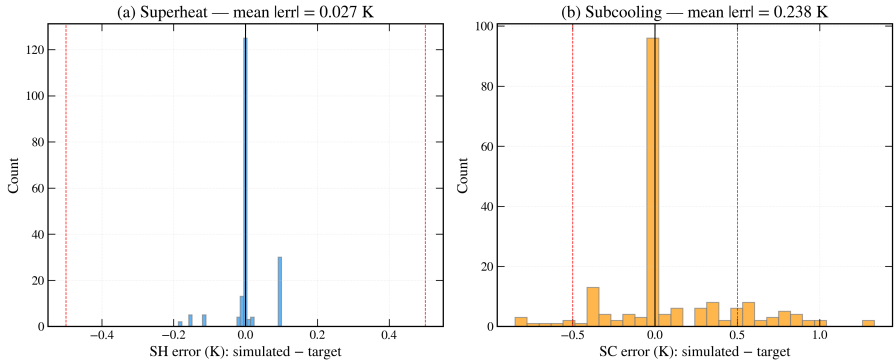


***Figure 2:*** *Distribution of convergence errors for superheat and subcooling. The $\pm 0.5\,K$ bands indicate acceptable tolerance.*

## 3.4. Surrogate modeling approaches

The objective of the surrogate is to replace the physical model with a fast, lightweight function that maps measurable boundary conditions to key thermodynamic outputs of the heat pump cycle. The input vector is defined as

$$\mathbf{x} = \left[T_{air},\ \varphi,\ \dot{V}_w,\ T_{w,in},\ N\right]^\top \in \mathbb{R}^5, \tag{3}$$

and the output vector as

$$\mathbf{y} = \left[\dot{Q}_{heat},\ \dot{W},\ COP,\ P_h,\ P_l,\ \Delta T_{sh},\ \Delta T_{sc}\right]^\top \in \mathbb{R}^8. \tag{4}$$

The surrogate model approximates the physical mapping according to

$$\hat{f}(\mathbf{x}) \approx \mathbf{y} \tag{5}$$

Four modeling approaches were evaluated to span a range of complexity, data efficiency, and interpretability. But first, to ensure an unbiased estimate of generalization performance on unseen operating conditions, we will split the dataset ($\mathcal{D}$) into 80% for training and 20% for validation/testing:

$$\mathcal{D}_{190} = \mathcal{D}^{152}_{train} \cup \mathcal{D}^{38}_{test} \tag{6}$$

All models were trained on the same 152-point training set and evaluated on the same 38-point held-out test set, with stratification by compressor count and water temperature to preserve balanced coverage in both subsets. Before training, all input features were standardized to zero mean and unit variance prior to training.

#### 3.4.1. Random Forest regression

The Random Forest (RF) is an ensemble method that constructs a large collection of independent decision trees, each trained on a random bootstrap sample of the data using a random subset of features. The final prediction is the average across all trees. This averaging reduces the variance inherent in individual trees and produces a robust predictor that handles nonlinear relationships without explicit feature engineering. For the present application, **300** trees were grown with a maximum depth of 15 and a minimum of 3 samples per leaf node. The depth limitation and leaf-size constraint serve as regularization mechanisms, preventing individual trees from memorizing the training data because we have a relatively small dataset. The RF natively supports multi-output regression, predicting all eight targets simultaneously. A useful byproduct of the RF is the feature importance ranking, computed as the mean reduction in prediction variance attributable to each input variable across all splits in all trees. This provides a general indicator of which boundary conditions most strongly influence the heat pump outputs.

#### 3.4.2. Gradient Boosting regression

Gradient Boosting (GB) constructs an ensemble of shallow decision trees in a sequential manner. Unlike RF, where trees are trained independently, each new tree is fitted to the residual errors left by the previous ensemble. At iteration $m$, the model is updated as

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\, h_m(\mathbf{x}) \tag{7}$$

Here, $F_m(\mathbf{x})$ is the prediction after iteration $m$, $F_{m-1}(\mathbf{x})$ is the prediction from the previous iteration, $h_m(\mathbf{x})$ is the new regression tree added at step $m$, and $\eta$ is the learning rate (shrinkage factor), which controls how strongly each new tree contributes. In practice, $h_m$ is fitted to the negative gradient of the loss with respect to $F_{m-1}$, so each iteration corrects the remaining errors.

This iterative correction mechanism often improves accuracy compared with RF, but it requires sequential training and is more sensitive to hyperparameter choices. Trees are kept shallow (maximum depth of 5), because each tree only needs to learn a local correction rather than the full input-output mapping.

For this work, the GB configuration uses **200** boosting stages and a learning rate of 0.1. Since the scikit-learn implementation does not natively support multi-output GB, one independent GB model is trained per output variable, resulting in eight separate models with shared hyperparameters.

#### 3.4.3. Gaussian Process regression

Gaussian Process (GP) regression takes a fundamentally different approach from tree-based methods. Instead of fitting one explicit function, it defines a probability distribution over possible functions and updates this distribution using the observed data. For a new input $\mathbf{x}^*$, GP returns a predictive distribution with both a mean prediction and an uncertainty estimate:

$$f(\mathbf{x}_*) \sim \mathcal{N}\big(\mu(\mathbf{x}^*),\, \sigma^2(\mathbf{x}^*)\big) \tag{8}$$

The predictive mean is given by

$$\mu(\mathbf{x}^*) = k(\mathbf{x}^*, X)\left[K(X,X) + \sigma_n^2 I\right]^{-1}\mathbf{y} \tag{9}$$

and the predictive variance by

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, X)\left[K(X,X) + \sigma_n^2 I\right]^{-1} k(X, \mathbf{x}^*) \tag{10}$$

Here, $k(\cdot,\cdot)$ is the kernel function, $X$ is the matrix of training inputs, $\mathbf{y}$ is the training output vector, $K(X,X)$ is the kernel matrix evaluated on the training set, $\sigma_n^2$ is the observation-noise variance, and $I$ is the identity matrix.

The behavior of the GP is controlled by the kernel, which encodes assumptions on smoothness and correlation structure. In this work, a Matérn kernel with smoothness parameter $\nu = 5/2$ is used, corresponding to a twice-differentiable prior that is consistent with smoothly varying thermodynamic variables. Automatic relevance determination (ARD) is included so that each input dimension has its own optimized length scale.

A separate GP is trained for each of the eight output variables. Kernel hyperparameters (length scales, output variance, and noise variance) are estimated by maximizing the log-marginal likelihood, with three random

restarts to reduce the risk of local optima.

This uncertainty estimate is particularly useful for fault detection. Under healthy operation, measurements are expected to remain within the 95% confidence interval $\mu \pm 2\sigma$. Values outside this interval indicate statistically significant deviations from expected behavior and can be used as an adaptive, threshold-free fault indicator.

#### 3.4.4. Feedforward neural network with physics-informed loss

A feedforward neural network (multi-layer perceptron, MLP) is used to learn the surrogate mapping through stacked affine transformations followed by nonlinear activations. The architecture adopted in this work contains three hidden layers with 128, 128, and 64 neurons. Each hidden layer is followed by batch normalization, a rectified linear unit (ReLU) activation, and dropout regularization:

$$\begin{aligned} \mathbf{z}^{(l)} &= W^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}, \\ \mathbf{a}^{(l)} &= \mathrm{Dropout}_{p=0.15}\Big(\mathrm{ReLU}\Big(\mathrm{BatchNorm}(\mathbf{z}^{(l)})\Big)\Big) \end{aligned} \tag{11}$$

Here, $l$ is the layer index, $W^{(l)}$ and $\mathbf{b}^{(l)}$ are the trainable weights and biases of layer $l$, $\mathbf{a}^{(l-1)}$ is the previous-layer activation vector, and $\mathbf{z}^{(l)}$ is the pre-activation vector. Batch normalization improves training stability, while dropout ($p = 0.15$) reduces overfitting, which is important given the relatively small training set (152 points).

The network is trained by minimizing the mean squared error between predicted and target outputs (both standardized to zero mean and unit variance), using the Adam optimizer with initial learning rate $10^{-3}$ and $L_2$ weight decay $10^{-4}$. A scheduler halves the learning rate after 30 epochs without validation improvement, and early stopping is applied if validation loss does not improve for 100 consecutive epochs.

In addition to the purely data-driven model, a physics-informed variant is trained with a composite loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{total}} &= \mathcal{L}_{\mathrm{data}} + \lambda\,\mathcal{L}_{\mathrm{physics}}, \\ \mathcal{L}_{\mathrm{data}} &= \frac{1}{n_s}\sum_{i=1}^{n_s} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2\,, \\ \mathcal{L}_{\mathrm{physics}} &= \mathcal{L}_{\mathrm{COP}} + 0.5\,\mathcal{L}_{\mathrm{press}} + 0.3\,\mathcal{L}_{\mathrm{bound}} + 0.2\,\mathcal{L}_{\mathrm{pos}} \end{aligned} \tag{12}$$

where $n_s$ is the number of samples, $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$ are predicted and target output vectors, and $\lambda$ controls the strength of the physics constraints.

The individual physics terms are defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{COP}} &= \frac{1}{n_s}\sum_{i=1}^{n_s}\left(COP_i - \frac{\dot{Q}_i}{\dot{W}_i}\right)^2, & \mathcal{L}_{\mathrm{pressure}} &= \frac{1}{n_s}\sum_{i=1}^{n_s}\left[\max\left(0, P_{l,i} - P_{h,i}\right)\right]^2, \\ \mathcal{L}_{\mathrm{bound}} &= \frac{1}{n_s}\sum_{i=1}^{n_s}\left[\max\left(0, 1 - COP_i\right)\right]^2, & \mathcal{L}_{\mathrm{pos}} &= \frac{1}{n_s}\sum_{i=1}^{n_s}\left(\left[\max\left(0, -\dot{Q}_i\right)\right]^2 + \left[\max\left(0, -\dot{W}_i\right)\right]^2\right) \end{aligned} \tag{13}$$

The COP-consistency term is the most informative physically: it penalizes predictions where the COP output is inconsistent with $\dot{Q}/\dot{W}$. For example, predicting $\dot{Q} = 40$ kW and $\dot{W} = 10$ kW implies $COP = 4$, so a predicted $COP = 3.5$ is penalized. The pressure-ordering term enforces $P_h > P_l$, the COP bound enforces heating-mode feasibility ($COP \geq 1$), and the positivity term discourages nonphysical negative power/capacity outputs. In this work, $\lambda = 0.1$ is selected so that the physics term contributes approximately 10% of the initial total loss. This keeps data fitting dominant in early epochs while progressively guiding the model toward thermodynamically consistent predictions.

### 3.5. Evaluation metrics and model comparison protocol

All surrogate models are evaluated on the same held-out test set of 38 operating points, stratified by compressor count and water-temperature pair to preserve representative operating-condition coverage. For each output variable, the following metrics are reported:

$$\mathrm{MAPE} = \frac{100}{n_s}\sum_{i=1}^{n_s}\frac{|\hat{y}_i - y_i|}{|y_i|},\; R^2 = 1 - \frac{\sum_{i=1}^{n_s}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n_s}\left(y_i - \bar{y}\right)^2} \tag{14}$$

MAPE is used as the primary comparison metric because it provides a unit-independent percentage error across outputs of different scales, while $R^2$ quantifies explained variance (values close to 1 indicate strong

predictive performance). For Gaussian Process models, calibration is additionally assessed using 95% interval coverage, i.e., the fraction of test points whose true value lies within the predicted 95% confidence interval. A well-calibrated model should provide coverage close to 95%; higher values indicate conservative uncertainty, whereas lower values indicate overconfidence. Inference speed is evaluated from 5000 repeated single-point predictions and reported as average prediction time. Computational gain is expressed as speedup relative to the physical model, using approximately 10 minutes per steady-state Dymola simulation as baseline.

# 4. Results and Discussion

This section reports surrogate-model performance in terms of accuracy, physical consistency, and computational efficiency, and then presents preliminary fault-detection results.

## 4.1. Surrogate model accuracy

***Table 3:*** *Model comparison — MAPE (%) and $R^2$ on the 38-point test set.*

| | | MAPE (%) | | | | $R^2$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Unit | RF | GB | MLP | GP | RF | GB | MLP | GP |
| $\dot{Q}_{heat}$ | kW | 3.58 | **0.40** | 3.74 | **0.36** | 0.989 | **0.999** | 0.985 | **0.999** |
| $\dot{W}$ | kW | 2.27 | **0.40** | 2.56 | **0.34** | 0.995 | **0.999** | 0.993 | **0.999** |
| $COP$ | — | 3.86 | 0.59 | 1.48 | **0.31** | 0.967 | 0.996 | 0.990 | **0.999** |
| $P_h$ | bar | 1.94 | 0.19 | 0.98 | **0.08** | 0.971 | 0.999 | 0.994 | **0.999** |
| $P_l$ | bar | 3.79 | 0.33 | 1.40 | **0.28** | 0.967 | 0.999 | 0.994 | **0.999** |
| $\Delta T_{sc}$ | K | 10.98 | 6.53 | 7.66 | 6.62 | 0.529 | 0.837 | 0.743 | 0.818 |

*Bold: best performance per output.*

Table 3 summarizes the results. The Gradient Boosting (GB) and Gaussian Process (GP) models consistently outperform the Random Forest and the standard MLP, achieving sub-1% MAPE on all primary outputs. On heating capacity and electrical power, both GB and GP reach 0.4% MAPE with $R^2 > 0.999$; well below the ±10% target stated in the study objectives. The RF achieves 2–4% across most outputs, while the MLP, constrained by the limited training set of 152 points, reaches 1.5–3.7%. The subcooling is the hardest output for all models (MAPE 6.5–11%), a direct consequence of the incomplete convergence of the subcooling controller in the "partial" simulations.

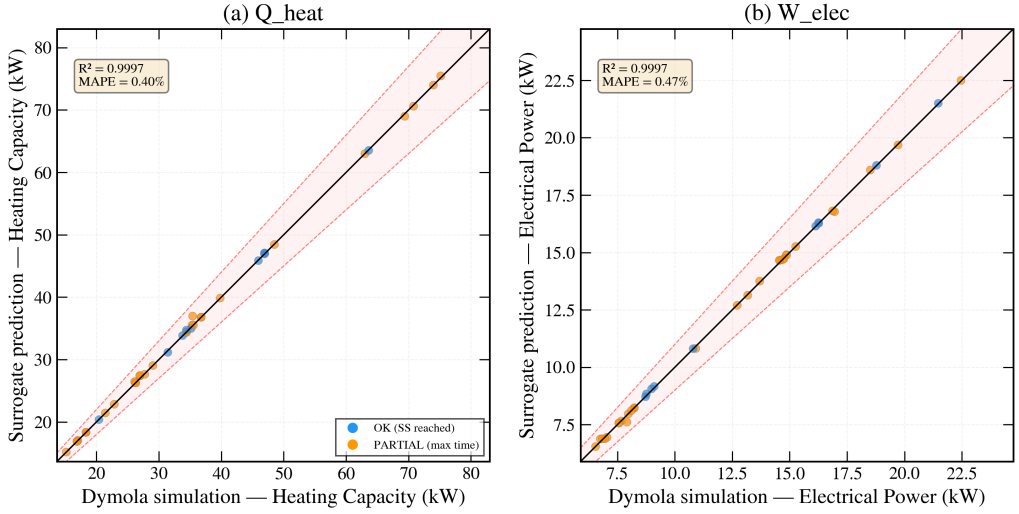


***Figure 3:*** *Parity plots for the Gradient Boosting surrogate on the held-out test set*

Figure 3 presents parity plots for the GB model on the four most operationally relevant outputs. All test points fall within the ±10% bands, with partial simulations (orange) distributing evenly alongside the fully converged points (blue), confirming that the near-steady-state data does not introduce systematic bias into the surrogate predictions.

To validate the surrogate against independent measurements, the GP model, which was trained on all acceptable Dymola simulation points, was evaluated on 25 experimental operating points from the studied heat pump test bench. These tests cover air temperatures of -7, 2, and 7 °C, water inlet temperatures from 30 to 40 °C, and both single- and dual-compressor configurations. Crucially, the experimental inputs are continuous

sensor measurements (e.g., water flow rates of 4892, 5058, or 7335 L/h) that fall between the discrete levels of the training grid (4500, 5500, 7000, 8000 L/h), testing the model's ability to interpolate in unseen conditions.

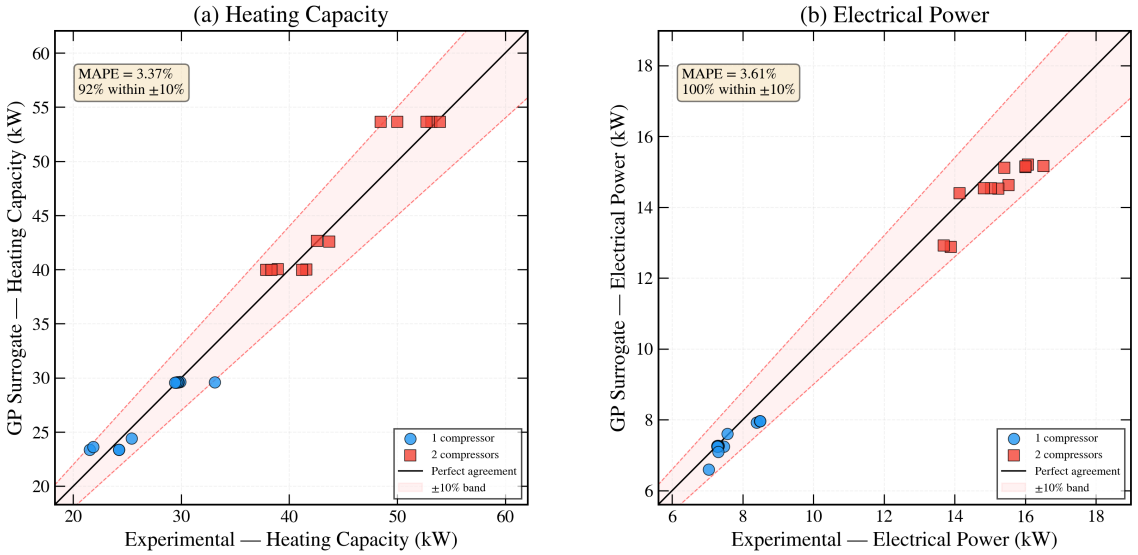


***Figure 4:*** *Comparison of the best surrogate model with the experimental data.*

Figure 4 presents the GP predictions against the experimental measurements. The heating capacity is predicted with a MAPE of 3.4%, with 92% of test points falling within the ±10% error bands. The electrical power achieves 3.6% MAPE with 100% of points within ±10%. The GP outperforms the other surrogates on experimental data, despite the GB and GP having comparable accuracy on the Dymola test set. This is because the GP's smooth kernel interpolation naturally handles continuous inputs between training grid points, whereas the tree-based models effectively snap predictions to the nearest grid level.

## 4.2. Physics-informed neural network

In terms of raw prediction accuracy, the PINN achieves a MAPE of 5.23% on heating capacity and 2.85% on COP, compared to 3.32% and 1.79% for the standard MLP. This moderate increase in prediction error is the expected cost of the physics regularization: the additional loss terms constrain the optimization landscape, preventing the network from fitting the data as tightly as the unconstrained MLP. However, both models remain well within the ±10% target, and the accuracy difference is secondary to the consistency improvement. The key advantage of the PINN emerges in the thermodynamic consistency of its predictions. Figure 5 compares the COP predicted directly by the network against the ratio $\dot{Q}/\dot{W}$ computed from the network's own capacity and power outputs. For a thermodynamically consistent model, all points should lie on the diagonal. The standard MLP exhibits a mean COP discrepancy of 0.074, with visible scatter away from the diagonal at higher COP values. The PINN reduces this discrepancy to 0.027 (a 64% improvement) and the points cluster tightly along the diagonal across the entire operating range. This improvement matters for fault detection

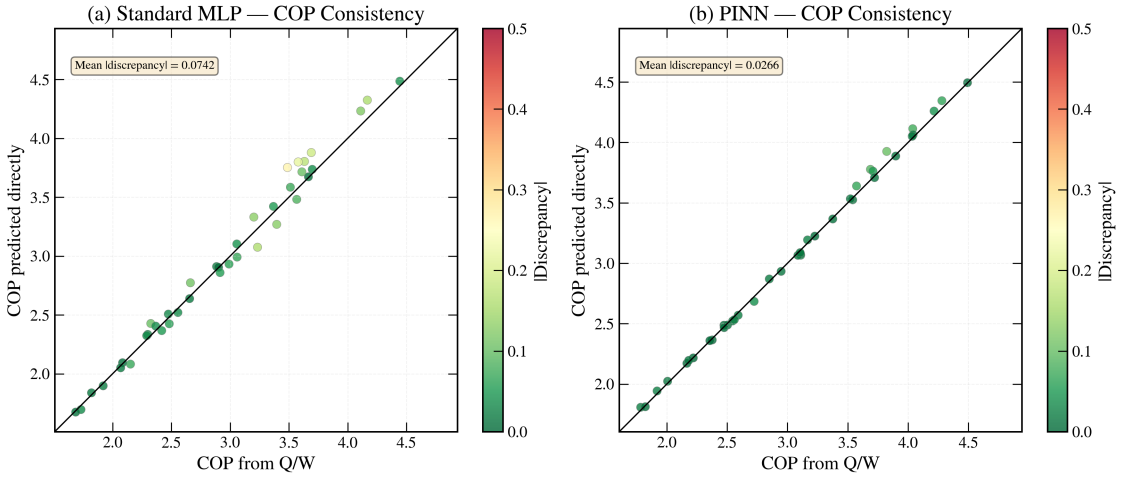


***Figure 5:*** *COP thermodynamic consistency: predicted COP vs. $\dot{Q}/\dot{W}$ for (a) standard MLP and (b) PINN.*

applications. A standard MLP that predicts $\dot{Q} = 40$ kW, $\dot{W} = 10$ kW, and $COP = 3.5$ simultaneously would signal a spurious anomaly in any monitoring system that checks these relationships for consistency. The PINN's internal coherence eliminates such false alarms, making it a more reliable component in a supervisory

platform even if its point-wise accuracy is slightly lower than the unconstrained network.

### 4.3. Computational performance

All surrogates achieve speedups of five to six orders of magnitude over Dymola's 10-minute simulation time. The GB model predicts a single operating point in approximately $0.3$ ms ($2 \times 10^6$ speedup), while the GP requires about $2$ ms ($3 \times 10^5$ speedup) due to the kernel matrix operations. Even the slowest surrogate is entirely adequate for cloud-based monitoring with update intervals of seconds. A single server running the GB model could evaluate over 3000 heat pump units per second, enabling real-time anomaly detection across an entire building stock.

### 4.4. Preliminary fault detection demonstration

To illustrate how the surrogate framework can support fault detection, a water-side fouling scenario was simulated using the Dymola physical model. Fouling of the condenser water filter was emulated by reducing the water volumetric flow rate to 90%, 80%, 70%, and 50% of its nominal value while keeping all other boundary conditions unchanged. The resulting "faulty" outputs were then compared against the predictions of the GP surrogate, which was trained exclusively on healthy-operation data at nominal flow rates. In this configuration, the GP serves as a reference model of expected healthy behavior. For a given set of measured boundary conditions (air temperature, humidity, water inlet temperature, compressor count), the surrogate is queried with the nominal water flow rate and returns a predicted heating capacity, electrical power, and COP along with their associated confidence intervals. When the actual system operates under fouling conditions, the measured outputs deviate from the GP predictions: the reduced water flow rate lowers the heat transfer rate in the condenser, decreasing the heating capacity and shifting the condensing pressure. If the normalized residual, which is defined as the difference between the measured and predicted value, divided by the GP's predicted standard deviation, exceeds the $\pm 2\sigma$ threshold, the system flags an anomaly. The Dymola fault simulations confirm that a 20% reduction in water flow rate produces deviations in heating capacity that exceed the GP's 95% confidence interval for the majority of operating conditions tested, indicating that faults of this magnitude are detectable. At 10% reduction, the deviations remain within the confidence band for most conditions, suggesting that this level of fouling falls below the detection threshold of the current model — a limitation that could be addressed by increasing the training data density in the relevant input region. This approach has two practical advantages. First, it requires no fault-specific training data: the GP detects anomalies purely by identifying deviations from learned healthy behavior, which sidesteps the data scarcity problem that limits most supervised FDD methods. Second, the surrogate evaluates in approximately 2 ms per operating point, making continuous monitoring feasible on a cloud platform, whereas running the same comparison with the Dymola digital twin would require approximately 10 minutes per evaluation, restricting its use to offline diagnostics. The limitations should be noted. This residual-based approach can detect that something is abnormal but cannot, by itself, identify the specific fault type or quantify its severity. Distinguishing between water-side fouling, refrigerant-side fouling, refrigerant leakage, or compressor degradation would require either fault-specific training data or a diagnostic layer that interprets the pattern of residuals across multiple outputs simultaneously. Extending the surrogate with fault intensity parameters as additional inputs, where the model learns the mapping from boundary conditions and fault severities to system outputs, is the subject of ongoing work.

## 5. Conclusion

This work presented a surrogate modeling framework for an air-to-water heat pump used in collective housing, leveraging a validated Dymola digital twin as the sole data source. A structured data generation methodology was developed, combining four empirical correlations for simulation initialization with a full-factorial design of experiments covering 288 operating points. Of these, 190 produced usable results spanning air temperatures from -10 to +14 °C, water inlet temperatures from 30 to 50 °C, and both single- and dual-compressor configurations. Five surrogate architectures were compared on a held-out test set of 38 points. The Gradient Boosting and Gaussian Process models achieved the highest accuracy against the physical model, with MAPE below 0.5% and $R^2$ exceeding 0.999 on heating capacity and electrical power. When evaluated against 25 independent experimental measurements, the Gaussian Process, after it was retrained on all the 190 data points, proved most accurate with 3.4% MAPE on heating capacity and 100% of electrical power predictions within $\pm$10%, owing to its smooth kernel interpolation between the discrete training grid points. A physics-informed neural network variant was shown to reduce the mean COP discrepancy from 0.074 to 0.027 compared to the standard MLP (a 64% improvement in thermodynamic consistency) at a modest cost in raw prediction accuracy. All surrogates achieve computational speedups exceeding $10^5$ relative to the physical model's 10 minute simulation time, enabling real-time evaluation on cloud-connected platforms. A preliminary fault detection demonstration showed that water-side fouling at 20% severity is detectable through the GP's confidence intervals without fault-specific training data.

Future work will extend this framework by introducing fault parameters as continuous input features to the surrogate, enabling quantitative fault severity estimation. Experimental validation under faulty operating conditions and deployment in a cloud-connected supervisory platform for a pilot fleet of collective housing heat pumps are planned as next steps.

# Nomenclature

**Roman symbols**

| | |
|---|---|
| $\dot{Q}$ | Heating Load (W) |
| $\dot{V}$ | Volume Flow Rate ($L/h$ or $m^3/h$) |
| $\dot{W}$ | Electrical Power (W) |
| $N$ | Number of compressors |
| $P$ | Pressure (bar or Pa) |
| $T$ | Temperature ($^\circ C$ or K) |

**Subscripts**

| | |
|---|---|
| $h$ | High |
| $l$ | Low |
| $sc$ | Subcooling |
| $sh$ | Superheat |
| $w$ | Water |

**Abbreviations**

| | |
|---|---|
| FDD | Fault Detection and Diagnosis |
| FV | finite-volume |
| GB | Gradient Boosting |
| GP | Gaussian Process |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| MB | Moving Boundary |
| MLP | Multi-layer perceptron |
| PINN | Physics-informed neural network |
| RF | Random Forest |
| TXV | Thermostatic expansion valve |

# References


[1] Réglementation environnementale RE2020 | Ministères Aménagement du territoire Transition écologique.

[2] I. Bellanco, E. Fuentes, M. Vallès, and J. Salom. A review of the fault behavior of heat pumps and measurements, detection and diagnosis methods including virtual sensors. *Journal of Building Engineering*, 39:102254, July 2021.

[3] Bryan P. Rasmussen. Dynamic modeling for vapor compression systems—Part I: Literature review. *HVAC&R Research*, 18(5):934–955, October 2012.

[4] Pengfei Li, Hongtao Qiao, Yaoyu Li, John E. Seem, Jon Winkler, and Xiao Li. Recent advances in dynamic modeling of HVAC equipment. Part 1: Equipment modeling. *HVAC&R Research*, 20:136–149, January 2014. ADS Bibcode: 2014HVACR..20..136L.

[5] M. Willatzen, N. B. O. L. Pettit, and L. Ploug-Sørensen. A general dynamic simulation model for evaporators and condensers in refrigeration. Part I: moving-boundary formulation of two-phase flows with heat exchange. *International Journal of Refrigeration*, 21(5):398–403, August 1998.

[6] Thomas L. McKinley and Andrew G. Alleyne. An advanced nonlinear switched heat exchanger model for vapor compression cycles using the moving-boundary method. *International Journal of Refrigeration*, 31(7):1253–1264, November 2008.

[7] Satyam Bendapudi, James E. Braun, and Eckhard A. Groll. A comparison of moving-boundary and finite-volume formulations for transients in centrifugal chillers. *International Journal of Refrigeration*, 31(8):1437–1452, December 2008.

[8] M. Gräber, K. Kosowski, C. Richter, and W. Tegethoff. Modelling of heat pumps with an object-oriented model library for thermodynamic systems. *Mathematical and Computer Modelling of Dynamical Systems*, 16(3):195–209, October 2010. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13873954.2010.506799.

[9] Sylvain Quoilin, Adriano Desideri, Jorrit Wronski, Ian Bell, and Vincent Lemort. ThermoCycle: A Modelica library for the simulation of thermodynamic systems. pages 683–692, March 2014.

[10] Laura Maier, David Jansen, Fabian Wüllhorst, Martin Kremer, Alexander Kümpel, Tobias Blacha, and Dirk Müller. AixLib: an open-source Modelica library for compound building energy systems from component to district level with automated quality management. *Journal of Building Performance Simulation*, 17(2):196–219, March 2024. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19401493.2023.2250521.

[11] Christian Vering, Mirko Engelpracht, Stephan Göbel, Sina Hoseinpoori, Fabian Wüllhorst, Christian Schwenzer, Matti Rademacher, Sven Hinrichs, Friederike Chandra, Philipp Mehrfeld, and Dirk Müller. Open-Source vapor compression library (VCLib): Heat pump modeling for education and research. *Computer Applications in Engineering Education*, 30(5):1498–1509, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cae.22540.

[12] Jiacheng Ma, Donghun Kim, and James E. Braun. Proper orthogonal decomposition for reduced order dynamic modeling of vapor compression systems. December 2021.

[13] Joachim Seifert, Lars Haupt, Lars Schinke, Alf Perschk, Thomas Hackensellner, Stephan Wiemann, and Martin Knorr. Digital Twin for Heat Pump Systems: Description of a holistic approach consisting of numerical models and system platform. *CLIMA 2022 conference*, May 2022.

[14] José Joaquín Aguilera, Wiebke Meesenburg, Wiebke Brix Markussen, Benjamin Zühlsdorf, and Brian Elmegaard. Real-time monitoring and optimization of a large-scale heat pump prone to fouling - towards a digital twin framework. *Applied Energy*, 365:123274, July 2024.

[15] Yang Zhao, Tingting Li, Xuejun Zhang, and Chaobo Zhang. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, 109:85–101, July 2019.

[16] Nicolás Ablanque, Nasrulloh Loka, Santiago Torras, Sriram Gurumurthy, Joaquim Rigola, Carles Oliet, Ivo Couckuyt, Tom Dhaene, and Antonello Monti. Vapor compression system data-driven surrogate models for aircraft Environmental Control Systems. *International Journal of Refrigeration*, 178:336–346, October 2025.

[17] Raphael Chinchilla, Vedang M. Deshpande, Ankush Chakrabarty, and Christopher R. Laughman. Learning Residual Dynamics via Physics-Augmented Neural Networks: Application to Vapor Compression Cycles. In *2023 American Control Conference (ACC)*, pages 4069–4076, May 2023. ISSN: 2378-5861.

[18] Jiacheng Ma, Yiyun Dong, Hongtao Qiao, and Christopher R. Laughman. A physics-constrained deep learning framework for dynamic modeling of vapor compression systems. *Applied Thermal Engineering*, 254:123734, October 2024.

[19] T. M. Rossi and J. E. Braun. A statistical, rule-based fault detection and diagnostic method for vapor compression air conditioners. December 1996.

[20] Haorong Li and James E. Braun. Decoupling features and virtual sensors for diagnosis of faults in vapor compression air conditioners. *International Journal of Refrigeration*, 30(3):546–564, May 2007.

[21] Zhe Sun, Huaqiang Jin, Jiangping Gu, Yuejin Huang, Xinlei Wang, and Xi Shen. Gradual fault early stage diagnosis for air source heat pump system using deep learning techniques. *International Journal of Refrigeration*, 107:63–72, November 2019.

[22] Gerrit Bode, Simon Thul, Marc Baranski, and Dirk Müller. Real-world application of machine-learning-based fault detection trained with experimental data. *Energy*, 198:117323, May 2020.

[23] Howard Cheung and James E. Braun. Development of Fault Models for Hybrid Fault Detection and Diagnostics Algorithm: October 1, 2014 – May 5, 2015. Technical Report NREL/SR–5500-65030, 1235409, December 2015.

[24] Belén Llopis-Mengual, David P. Yuill, and Emilio Navarro-Peris. Fault detection and diagnosis algorithm for multiple simultaneous faults in residential air-conditioning systems: Development, validation study and critical analysis. *Applied Thermal Engineering*, 269:125975, June 2025.

[25] Belén Llopis-Mengual and Emilio Navarro-Peris. Selection of relevant features to detect and diagnose single and multiple simultaneous soft faults in air-source heat pumps. *Applied Thermal Engineering*, 238:121922, February 2024.

[26] Etienne Haddad. *Création d'un jumeau numérique d'un système de réfrigération et validation expérimentale : application à la détection de fuites*. thesis, Université Paris sciences et lettres, December 2021.