# Physics-Informed Virtual Sensing for Isentropic Efficiency: Enabling Sensor Reduction in Heat Pumps

Savvas Eftychis[1], Sławomir Nowaczyk[2], Klas Berglöf[3], Metkel Yebiyo[4] and Sepideh Pashami[5]

[1,2,5] *Center for Applied Intelligent Systems Research, Halmstad University, Halmstad, 30118, Sweden*
*savvas.eftychis@hh.se*
*slawomir.nowaczyk@hh.se*
*sepideh.pashami@hh.se*

[3] *Schneider Electric, Älvsjö, Sweden*
*klas.berglof@se.com*

[4,5] *RISE Research Institutes of Sweden, Kista, Sweden*
*metkel.yebiyo@ri.se*
*sepideh.pashami@ri.se*

## ABSTRACT

Virtual sensors are increasingly used in Industrial Internet of Things (IIoT) systems to estimate quantities that cannot be directly measured or when physical sensors are unavailable. In heat pump systems, compressor isentropic efficiency is a commonly used thermodynamic performance indicator typically computed from pressure and temperature measurements at both compressor inlet and outlet. This study presents an analysis in virtual sensing of isentropic efficiency under sensor reduction, studying the effect of physics-informed features. A comprehensive feature space was constructed from raw measurements and thermodynamic properties computed via CoolProp software. Feedforward neural networks were trained for all feasible combinations of two to four input features across multiple sensor removal scenarios. Model performance was assessed using structured data splits that allow for evaluation of generalizability from in-distribution training data to out-of-distribution unseen operating conditions. Results show that excluding the suction temperature sensor yields the most favorable trade-off between in-distribution accuracy and out-of-distribution robustness. Analysis across all sensor removal scenarios reveals that feature composition is the primary determinant of out-of-distribution performance, rather than model architecture or hyperparameter tuning. Robust feature sets consistently include discharge entropy together with suction pressure and saturation temperature, reducing out-of-distribution error by up to 70% compared with a raw-sensor baseline, at a modest cost in in-distribution accuracy.



## 1. INTRODUCTION

In industrial and energy systems, the increasing adoption of Industrial Internet of Things (IIoT) technologies has enabled real-time monitoring and data-driven decision-making, forming a key aspect of the Industry 4.0 paradigm. This is supported by the widespread deployment of sensors, which serve as the data backbone for Prognostics and Health Management (PHM) practices, enabling continuous health assessment and predictive maintenance of critical system components. Within this landscape, virtual sensors have emerged as a valuable tool for estimating quantities that cannot be directly measured, by inferring them from available physical measurements (Perera, Ratnaweera, Dasanayaka, & .C, 2023).

In heat pump systems, performance indicators such as compressor isentropic efficiency or coefficient of performance are central to PHM practices and serve as primary indicators for fault detection and condition-based maintenance (Cui & Wang, 2005). These quantities cannot be measured directly but are computed from combinations of temperature and pressure sensor measurements. Deviations from expected values can signal faults or performance degradation (Spitler, Berglöf, Mazzotti Pallard, & Witte, 2021). Although pressure and temperature sensors are common in modern heat pump systems, reducing the number of sensors is beneficial for cost-sensitive applications. However, if a performance indicator is based on a sensor that is not present, it cannot be calculated, directly limiting the scope of PHM.

This motivates an analysis of virtual sensing for thermodynamic performance indicators under sensor reduction, examining which sensors can be removed and which input features best compensate for the missing measurement. A central aspect of this analysis is evaluating out-of-distribution (OOD) robustness, that is, how well models trained on known operating conditions generalize to unseen ones. Compressor isentropic efficiency is used as a representative performance indicator to explore these questions.

Heat pump systems operate across a wide range of conditions, driven by fluctuating ambient temperatures and varying load demands. Although data-driven models can effectively capture nonlinear process relationships, they often struggle to generalize beyond the operating conditions represented in their training data (Guo et al., 2024). Consequently, the selection of input features and model configuration becomes critical for both performance on in-distribution (ID) conditions (regimes represented in the training data) and generalization to OOD conditions, where the model encounters previously unseen operating regimes.

This study systematically evaluates virtual sensor models for compressor isentropic efficiency under sensor reduction scenarios, using measurements from a heat pump system in a commercial building. All feasible feature combinations of two, three, and four inputs, constructed from raw sensor data and physics-derived transformations, are evaluated across multiple sensor removal scenarios, with neural network models optimized through hyperparameter search and tested on both ID and OOD data.

The main contributions of this article are:

1. A comprehensive evaluation of isentropic efficiency approximation with a subset of sensors, **identifying which sensor can be removed** with the least impact on virtual sensing accuracy and OOD robustness.
2. **A systematic search of physics-informed feature combinations** for isentropic efficiency estimation, identifying compact feature sets that achieve strong predictive performance under reduced sensor availability.
3. Empirical evidence that **feature composition** is the primary determinant of OOD robustness, with physics-informed features reducing OOD error by up to 70% compared with a raw-sensor baseline.

The remainder of this work is structured as follows. Section 2 reviews related work on virtual sensing and physics-informed machine learning for heat pumps. Section 3 describes the methodology, covering the theoretical calculation of isentropic efficiency, physics-informed feature engineering, the experimental protocol, and model development. Section 4 presents results, including a comparison of sensor removal scenarios and an analysis of the best-performing feature sets. Section 5 outlines directions for future work, and Section 6 summarizes the main findings.

## 2. RELATED WORK

Virtual sensors estimate hard-to-measure process variables from available measurements, offering advantages over hardware sensors in cost, maintenance, and deployability (Perera et al., 2023; Kadlec & Gabrys, 2009). Data-driven approaches can capture complex nonlinear behavior but are prone to degraded performance in OOD operating conditions (Perera et al., 2023).

Physics-informed machine learning (PIML) addresses this by incorporating prior physical knowledge through embedding governing equations into network architectures (Karniadakis et al., 2021), constraining loss functions, or engineering input representations (Wu, Sicard, & Gadsden, 2024). This work follows the latter strategy: rather than modifying the model itself, we augment the input space with features derived from known thermodynamic relationships. Niresi et al. (Niresi, Bissig, Baumann, & Fink, 2024) propose a related approach for virtual sensing in district heating networks, enriching graph neural network inputs with physics-derived quantities and demonstrate consistent improvements over purely data-driven baselines regardless of model architecture, supporting the generality of physics-informed input augmentation as a strategy for virtual sensing in IIoT systems.

Such physics-informed features are model-agnostic and may yield descriptors that remain more stable across varying operating regimes than raw sensor measurements alone.

Heat pump condition monitoring relies on a range of thermodynamic performance indicators that characterize system and component behavior. Some examples include coefficient of performance, compressor isentropic efficiency, superheat, logarithmic mean temperature difference and heat transfer coefficients for condensers and evaporators (Cui & Wang, 2005; Wang et al., 2024). Residuals between expected and observed values of such indicators, as well as virtual sensor outputs, have demonstrated strong fault-indicative capability (G. Li, Hu, Liu, Fang, & Kang, 2021). These quantities are typically derived from combinations of sensor measurements and thermodynamic property calculations, rather than being measured directly. As a result, they are natural targets for virtual sensing and physics-informed feature engineering, since both approaches seek to improve estimation accuracy, interpretability, and robustness under varying operating conditions.

Wang et al. (2025) further explored this by augmenting raw measurements with the complementary physics-informed features and introduced a Multi-Source Ranking Information Ensemble method to identify an optimal feature subset for fault classification. In their work, they found that 9 of the 13 most discriminative features were physics-informed. While that work targets classification, it supports the premise that thermodynamic features carry predictive signal beyond raw sen-

sor measurements.

## 3. METHODOLOGY

Our objective is to develop and evaluate virtual sensor models for estimating compressor isentropic efficiency using a reduced set of sensors. We first compute reference isentropic efficiency values from a fully instrumented system with four sensors, based on the theoretical formula presented in section 3.1. A comprehensive set of candidate input features is then constructed from available measurements using physics-informed transformations (section 3.2). For a given sensor removal case, we exclude all features that depend on that sensor and select all possible combinations from the remaining features as described in section 3.3. For each feature combination, we train feedforward neural network regression models as described in section 3.4. For the evaluation and comparison of the feature sets we used a structured data split, allowing us to isolate unseen operating conditions. Feature sets are then compared in the analysis section to identify which perform well for ID and OOD. This procedure is repeated for each sensor removal scenario.

### 3.1. Isentropic Efficiency Calculation

The reference isentropic efficiency values used in this study are computed from a fully instrumented compressor using measured thermodynamic states and refrigerant property evaluation using CoolProp (Bell, Wronski, Quoilin, & Lemort, 2014). The available measurements include suction pressure $p_{in}$, suction temperature $T_{in}$, discharge pressure $p_{out}$, and discharge temperature $T_{out}$.

The isentropic efficiency of the compressor is defined as

$$\eta_{is} = \frac{h_{isen} - h_{in}}{h_{out} - h_{in}} \tag{1}$$

where $h_{in}$ is the specific enthalpy at compressor inlet, $h_{out}$ is the specific enthalpy at compressor outlet (actual process), and $h_{isen}$ is the specific enthalpy at compressor outlet under isentropic compression. This quantity serves as the ground-truth target for training and evaluating the virtual sensor models developed in this study.

The calculation proceeds as follows.

1. Inlet State Determination
   The thermodynamic state at the compressor inlet is determined from measured suction pressure and temperature:

$$h_{in} = h(p_{in}, T_{in}) \tag{2}$$

$$s_{in} = s(p_{in}, T_{in}) \tag{3}$$

   where $h(\cdot)$ and $s(\cdot)$ denote CoolProp calls returning specific enthalpy and entropy, respectively, with the enclosed arguments serving as inputs.

2. Isentropic Outlet State
   The isentropic discharge state is defined as the state at discharge pressure with entropy equal to the inlet entropy. Using the measured discharge pressure and the computed inlet entropy:

$$h_{isen} = h(p_{out}, s_{in}) \tag{4}$$

   This represents the specific enthalpy the refrigerant would have if compression occurred isentropically.

3. Actual Outlet State
   The actual discharge enthalpy is computed from measured discharge pressure and temperature:

$$h_{out} = h(p_{out}, T_{out}) \tag{5}$$

   Again, refrigerant properties are evaluated using CoolProp.

The sequence of property evaluations described above highlights how the information from measured pressures and temperatures propagates through thermodynamic property calculations to yield the final efficiency value. This propagation involves nonlinear transformations of raw pressure and temperature measurements into energy-based quantities like enthalpy and entropy, embedding thermodynamic state information that raw sensor readings alone do not directly capture.

### 3.2. Feature Engineering Under Sensor Reduction

The feature pool was constructed to include not only raw measurements, but also quantities that describe the thermodynamic state of the refrigerant and descriptive features that capture the refrigerant's behavior in accordance with our physical understanding. This physics-informed feature engineering process can be distinguished into two categories: (i) thermodynamic property evaluation and (ii) algebraic transformations that encode our physical understanding of the underlying processes.

**Thermodynamic Properties:** Using measured pressures and temperatures, thermodynamic state variables were computed using CoolProp. For each instance, suction and discharge properties were evaluated using the refrigerant state equations. These derived properties embed physical knowledge of refrigerant behavior into the feature space and allow the model to access physically meaningful quantities beyond raw sensor readings.

**Algebraic Transformations:** Beyond direct thermodynamic properties, additional features were generated through algebraic transformations such as differences, normalized differ-

ences, ratios and combined energetic terms. These transformations capture physically relevant relationships such as pressure ratio, pressure rise, and energy balance indicators. By systematically expanding the feature space with structured combinations, we enable the evaluation of whether physically informed representations improve model robustness and generalization performance.

For each sensor removal scenario, all features directly or indirectly dependent on the removed sensor were excluded from the candidate pool. From the remaining features, combinations of predefined cardinalities (two, three, and four inputs) were generated and evaluated.

Table 1 summarizes the features considered in this study.

Table 1. Summary of engineered feature categories

| Category | Representative Features |
|---|---|
| Raw Measurements | $p_{in}, T_{in}, p_{out}, T_{out}, EP$ |
| Thermodynamic Properties | $h_{in} = h(p_{in}, T_{in})$<br>$h_{out} = h(p_{out}, T_{out})$<br>$s_{in} = s(p_{in}, T_{in})$<br>$s_{out} = s(p_{out}, T_{out})$<br>$h_{isen} = h(p_{out}, s_{in})$<br>$d_{in} = d(p_{in}, T_{in})$<br>$d_{out} = d(p_{out}, T_{out})$<br>$z_{in} = z(p_{in}, T_{in})$<br>$z_{out} = z(p_{out}, T_{out})$<br>$T_{sat} = T_{sat}(p_{in}, sat^*)$<br>$T_{super} = T_{in} - T_{sat}$ |
| Ratios & Differences | $p_{ratio} = p_{out}/p_{in}$<br>$T_{ratio} = T_{out}/T_{in}$<br>$\Delta h_{is} = h_{isen} - h_{in}$<br>$\Delta T = T_{out} - T_{in}$<br>$\Delta p = p_{out} - p_{in}$<br>$\Delta T_{norm} = (T_{out} - T_{in})/T_{in}$<br>$\Delta p_{norm} = (p_{out} - p_{in})/p_{in}$ |

### 3.3. Data, Splits and Experimental Protocol

#### 3.3.1. Data

The dataset used in this study originates from heat pump system in a commercial building, provided through PREMA-HEAPS, a collaborative project between Halmstad University, RISE, ClimaCheck, Vasakronan and Enrad. The system operates under variable load conditions and includes continuous monitoring of compressor suction and discharge pressures and temperatures, as well as electrical power consumption. Measurements were recorded for every minute, and the data were collected over a period between January 01 and August 06 2025. After filtering for stable operations, the total number of observations was $\sim$8500.

#### 3.3.2. Data Splits

To evaluate OOD robustness, the dataset was partitioned into four mutually exclusive subsets: training, validation, test, and generalization (gen).

Principal component analysis (PCA) was first applied to raw sensor data and isentropic efficiency. A median split along the first principal component divided the data into a development set (below median) and a generalization set (at/above median). This ensures the model can be evaluated on a distinct, OOD operating region.

The remaining 50% of the data were randomly partitioned into training, validation, and test subsets (80%-10%-10%). The training set was used for parameter learning, the validation set for hyperparameter optimization and early stopping, and the test set for ID performance evaluation.

This four-way split allows simultaneous assessment of:

- ID generalization (test set),
- OOD robustness (gen set),
- Hyperparameter selection bias (validation set).

### 3.4. Model Development and Hyperparameter Optimization

For each sensor removal scenario, all features dependent on the removed sensor were excluded. From the remaining feature pool, all possible feature combinations of size two, three, and four were generated. Each combination defines a candidate feature set.

**Neural Network Architecture:** For every feature set, a multilayer perceptron (MLP) regression model was trained to estimate isentropic efficiency. The architecture consisted of fully connected layers with nonlinear activation functions. The output layer contained a single neuron corresponding to the predicted efficiency.

**Hyperparameter Optimization:** Hyperparameter tuning was performed using Optuna. For each feature set, 100 independent trials were conducted, with the exception of two reference feature sets, one comprising the raw measurements of the three remaining sensors, and one augmenting these with electrical power consumption ($EP$). For the two reference feature sets, 500 trials were conducted. This increased trial budget for the reference sets ensures a more thorough search of the hyperparameter space, providing stronger baselines against which physics-informed feature combinations are compared. To establish the upper bound of MLP performance, two additional reference feature sets were defined using all four raw measurements (with and without EP), each also tuned over 500 trials. The deviation of reduced-sensor models from this upper bound reflects the information loss not recovered through learning.

The hyperparameter search space included:

- Number of hidden layers (1-3),

- Number of neurons per layer (32, 64, 128, 256),
- Dropout (0-0.3 with step: 0.05),
- Learning rate ($10^{-6}$ - $10^{-1}$, logarithmically sampled),
- Batch size (32, 64, 128),
- Optimizers (Adam, AdamW, RMSprop, SGD).

The objective function minimized the root mean squared error (RMSE) on the validation set.

**Model Evaluation:** For every trained model, performance metrics were computed on all four data splits (training, validation, test, and gen). This resulted in 100 trained models per feature set, each associated with its corresponding hyperparameter configuration and performance metrics. To simulate practical deployment, the generalization split is assumed to be unavailable and model selection was based on validation performance (RMSE).

This experimental design enables a systematic evaluation of how feature composition influences model performance under both ID and OOD conditions.

## 4. RESULTS AND DISCUSSION

Section 4.1 presents an overview of performance distributions across all sensor removal cases and identifies the most promising scenario. Section 4.2 evaluates feature sets according to ID and OOD performance, reveals that there is a trade-off between them and identifies key features for OOD robustness.

### 4.1. Performance Overview Across Sensor Removal Scenarios

Removing any sensor reduces the information available to the virtual sensor and is expected to degrade predictive performance relative to the fully instrumented baseline. To quantify this information loss, we compared models trained on all four raw sensor values against those trained on the values available under each sensor removal scenario. Table 2 shows the results of this comparison, capturing the validation, test and generalization RMSE for each case.

The results show that sensor removal leads to a substantial degradation in model performance. For models including EP, the validation RMSE increases from 0.0391 to 0.4629 (about 12×) under the worst sensor-removal scenario ($P_{out}$ excluded) and to 0.2558 (about 6.5×) for the best case scenario ($T_{in}$ excluded). As expected, the generalization RMSE is consistently higher than the validation and test RMSE across all cases, reflecting the difficulty of generalizing to unseen operating regimes. Among the four scenarios, excluding $T_{in}$ results in the smallest information loss, making it the most favorable candidate for virtual sensing.

For each sensor removal scenario, multiple feature sets were constructed as described in previous sections. To assess whether physics-informed features can compensate for the information loss caused by sensor removal, we compared the RMSE distribution of all evaluated feature sets for each scenario, across all splits. These distributions can be seen in Fig. 1, in the form of a violin plot, for validation, test and generalization splits.

We also include the values for the two reference models, with feature sets that include the raw values of the remaining sensors with and w/o EP.

The validation and test distributions for each scenario exhibit a bimodal structure. The lower RMSE mode corresponds to feature sets that incorporate information from all three remaining sensors, while the higher mode corresponds to feature sets drawing from only one or two sensors. Achieving competitive ID performance therefore requires that the feature set retains information from all three sensors.

In all four scenarios we observe that the generalization RMSE exhibits substantially greater variance than the validation and test RMSE. The wide spread in generalization RMSE across feature sets indicates that feature composition is the primary determinant of OOD robustness, rather than model architecture or hyperparameter tuning.

To compare the effect of physics-informed features against the raw-sensor baselines for each sensor removal scenario, we selected the best-performing feature set based on validation RMSE. The corresponding validation, test, and generalization RMSE values are reported in Table 2 ('PI' columns). Compared to the raw-sensor baselines, the PI feature sets alone do not improve model performance across the validation, test, and generalization splits. In the $T_{in}$ exclusion case, the generalization RMSE in fact increases relative to the raw-sensor baseline.

However, there are PI feature sets that can achieve a much lower generalization RMSE. Table 2 also reports the lowest RMSE achieved separately for validation, test and generalization splits (under the 'Pi (min)' column) for the $T_{in}$ exclusion case. This can also be seen in Fig. 1. For all the scenarios studied, various feature sets achieved a lower generalization RMSE than the raw-sensor baselines. Interestingly, the scenario that excludes $T_{in}$ can achieve the lowest generalization RMSE (Fig. 1). This indicates that this scenario also offers the greatest potential for OOD robustness.

Given its lowest validation RMSE (Table 2) and the potential to achieve the lowest generalization RMSE, the $T_{in}$ exclusion case is used for all subsequent analysis.

Table 2. RMSE of full-sensor reference models (all 4 sensors, with/without EP) and of each sensor removal scenario, evaluated with the remaining raw sensors (with/without EP) and the best physics-informed feature set (PI). For the $T_{in}$ exclusion case, we also see the result for the lowest RMSE achieved per validation, test and generalization splits 'PI (min)'.

| | **All sensors** | | **Pout excluded** | | | **Pin excluded** | | | **Tin excluded** | | | | **Tout excluded** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Raw w EP | Raw w/o EP | Raw w EP | Raw w/o EP | PI | Raw w EP | Raw w/o EP | PI | Raw w EP | Raw w/o EP | PI | PI (min) | Raw w EP | Raw w/o EP | PI |
| Val | 0.0391 | 0.0452 | 0.4629 | 0.5301 | 0.4439 | 0.4264 | 0.4341 | 0.4153 | 0.2558 | 0.2530 | 0.2495 | 0.2495 | 0.4373 | 0.4992 | 0.4382 |
| Test | 0.0497 | 0.0726 | 0.5123 | 0.5887 | 0.5232 | 0.5267 | 0.5478 | 0.5329 | 0.3343 | 0.3413 | 0.3489 | 0.3272 | 0.4719 | 0.5389 | 0.4768 |
| Gen | 0.0994 | 0.1131 | 0.9731 | 1.0319 | 1.2079 | 1.5780 | 2.0331 | 1.0269 | 1.0017 | 1.1311 | 1.5978 | **0.3026** | 2.4238 | 2.5899 | 2.0809 |

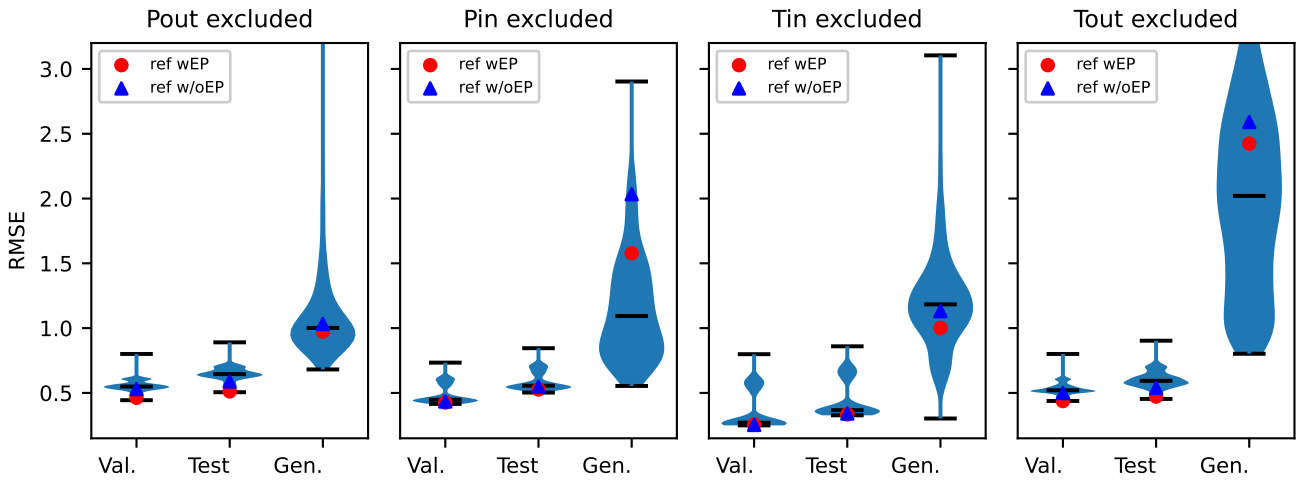


Figure 1. Distribution of RMSE across all evaluated feature sets for the four sensor removal scenarios. Violin plots show validation, test, and generalization performance (RMSE) of the best model obtained for each feature set based on validation error. Reference models (with and w/o compressor power) are included for comparison.

### 4.2. Feature set analysis under $T_{in}$ Exclusion

Prior to analyzing the joint behavior of ID and OOD performance, we examine the characteristics of the feature space under the $T_{in}$ exclusion scenario. Removing the suction temperature eliminates direct access to thermodynamic state information at the compressor inlet, which propagates through multiple derived features, including inlet enthalpy and entropy. This shift constrains the feature space to representations that rely primarily on discharge-side information. As a result, the remaining feature pool becomes inherently biased toward discharge-side properties.

Figure 2 shows a scatter plot of the generalization RMSE as a function of the test RMSE for all 781 evaluated feature sets in the $T_{in}$ exclusion scenario. Each point corresponds to the best model for a given feature set, selected by lowest validation score. Three distinct regions emerge, labeled A, B, and C.

Regions A and B both exhibit low test RMSE, while region C exhibits a higher test error. This separation is the bimodal structure observed in the violin plots of Fig. 1: feature sets in regions A and B incorporate information from all three available sensors, whereas those in region C draw from only one or two. Competitive ID performance therefore requires coverage of all available sensors.

The critical distinction lies between regions A and B. Both achieve comparable test RMSE, yet they differ markedly in generalization performance. Region A contains only six feature sets but achieves a generalization RMSE as low as 0.3026, compared to 1.0017 for the raw-sensor reference located in region B. However, this improvement comes at a modest cost in validation and test performance.

Figure 3 provides further insight by showing the top 10 feature sets ranked according to validation RMSE (row 1), test RMSE (row 2), and generalization RMSE (row 3). Each bar displays the three performance metrics for the corresponding feature set, in comparison with the performance of the reference set (red dashed line).

The feature sets in rows 1 and 2 originate entirely from region B. Although these configurations achieve the lowest validation and test errors, they exhibit poor generalization performance, with generalization RMSE often exceeding that of the raw-sensor reference model. This indicates that optimizing feature sets based on validation or test performance does not reliably identify configurations that generalize well in out-of-distribution samples.

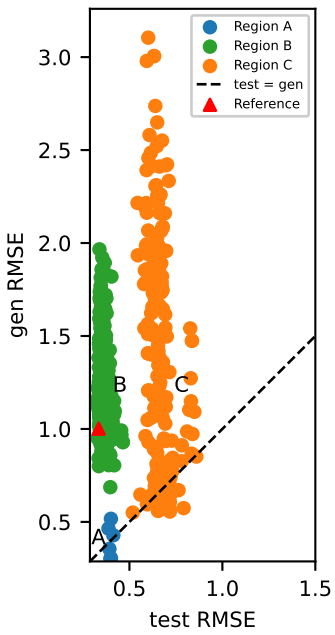


Figure 2. Scatter plot showing how the different feature sets perform on generalization and on test set. Reference point shown is with raw sensor data and includes the power consumption as well. Points can be separated into regions A, B and C.

In contrast, the top feature sets ranked by generalization RMSE (row 3) correspond to region A, the compact cluster in Fig. 2 that simultaneously achieves low test and low generalization error compared to the tested physics-informed feature sets.

The top validation-ranked feature sets (row 1) are dominated by discharge enthalpy ($h_{out}$), which appears in nearly every configuration. However, because the differences between the top feature sets are very small, we cannot make a strong claim about discharge enthalpy contributing to a better validation set.

In contrast, the top generalization-ranked feature sets (row 3) reveal a pattern. The top three feature sets consist of discharge entropy ($s_{out}$), in combination with suction pressure ($p_{in}$) and saturation temperature ($T_{sat}$), while the following three augment these with electrical power consumption ($EP$). These six configurations form a distinct group with validation RMSE between 0.30 and 0.32 and generalization RMSE between 0.30 and 0.52. The remaining feature sets show substantially higher validation errors while achieving only marginal improvements in generalization performance, confirming that the top six configurations are structurally distinct.

To further examine the characteristics of feature sets with strong generalization performance, we repeat the analysis with an alternative model selection criterion: rather than selecting the best model per feature set based on validation RMSE, we select based on generalization RMSE. This draws from the same pool of 100 trained models per feature set that were optimized based on validation RMSE, differing only in the selection criterion applied. These results are shown in Fig. 4.

The same structural pattern reappears among the top six feature sets, matching the combinations identified in the previous analysis. Despite the hyperparameter search being guided by validation RMSE, selecting models based on generalization performance again yields the same core feature group: discharge entropy ($s_{out}$), suction pressure ($p_{in}$), and saturation temperature ($T_{sat}$). The top three feature sets consist of these three features, while the next three extend them by including electrical power consumption ($EP$). Further analysis of the top 10 feature sets (Fig. 5) underscores the role of $s_{out}$ in OOD robustness. While $s_{out}$ exists in 2 of the top 10 feature sets ranked by validation RMSE and 4 of those ranked by test RMSE, it is present in 8 of the top 10 feature sets ranked by generalization RMSE.

The best-performing feature set reduces generalization RMSE from 1.0017 (raw-sensor reference) to 0.3026, representing a 70% improvement. This gain comes at the cost of moderately higher ID errors, with validation RMSE increasing from 0.2558 to 0.3229 and test RMSE from 0.3343 to 0.3975. The trade-off is therefore asymmetric: the improvement in robustness substantially outweighs the reduction in ID accuracy.

These results indicate that the separation between regions A and B in Fig. 2 is driven primarily by feature composition rather than model tuning. Feature sets that achieve strong generalization consistently include discharge entropy ($s_{out}$), often in combination with suction pressure ($p_{in}$) and saturation temperature ($T_{sat}$).

This pattern suggests that certain physics-derived features are more strongly associated with robust OOD performance.

### 4.3. Deployment Considerations and Limitations

While the results demonstrate that physics-informed features can substantially improve OOD robustness, several limitations should be considered when interpreting these findings and assessing their applicability to practical deployment.

**Data period.** The dataset spans January to August 2025, omitting a significant part of winter month operations. Since the PCA-based split defines OOD regions within the observed range, reported generalization performance should be regarded as a lower bound on the seasonal variability.

**Sensor faults.** Virtual sensors inherit the characteristics of the physical sensors they rely on. Removing $T_{in}$ eliminates one potential source of faults but increases dependence on $p_{in}$, $p_{out}$, and $T_{out}$, leaving the virtual sensor exposed to drift and calibration inaccuracies.

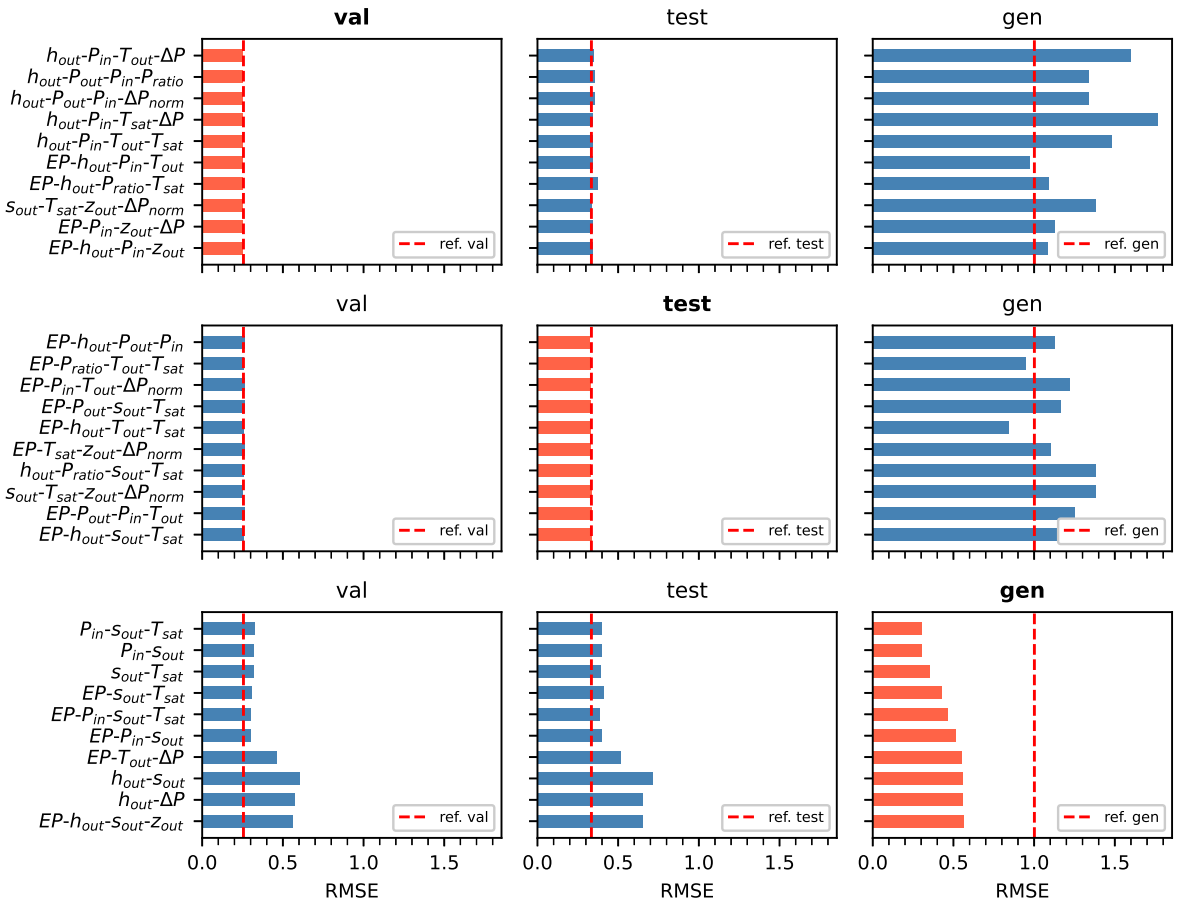


Figure 3. Top 10 feature sets ranked by validation (row 1), test (row 2), and generalization (row 3) RMSE. For each feature set, the displayed model is the one with the lowest validation RMSE among all 100 hyperparameter trials. The dashed reference line corresponds to the raw sensor values including electrical power consumption (EP).

**Single system and healthy-unit assumption.** The analysis is based on a single heat pump unit. The proposed methodology is expected to generalize on units with similar components and refrigerants. Furthermore, the approach assumes the training unit is fault-free, any undetected fault or degradation during data collection would be encoded as normal behavior for the sensor. Domain-expert validation of unit health is required for reliable deployment.

## 5. FUTURE WORK

Several directions for future work emerge from this study.

**Anomaly detection and fault diagnosis.** While isentropic efficiency is already recognized as a primary fault indicator in heat pump PHM (G. Li et al., 2021; Wang et al., 2025), distinguishing genuine faults from efficiency variations due to changing operating conditions remains a challenge. The OOD-robust virtual sensor developed in this study directly addresses this: because the physics-informed feature sets remain stable across operating regimes, residuals between the virtual sensor output and the computed efficiency are less likely to be confounded by legitimate regime changes, potentially yielding cleaner fault signals. Extending the proposed framework toward fault detection and isolation is therefore a natural next step.

**Improving interpretability through symbolic regression.** The systematic feature exploration conducted here identified compact, physics-informed feature sets that achieve strong predictive performance. Symbolic regression could be applied to these feature sets to derive explicit analytical expressions approximating isentropic efficiency from a small number of variables. Such expressions would enhance interpretability and facilitate deployment in industrial monitoring systems (S. Li, Wu, Lin, Song, & Feng, 2025). Moreover, the same methodology could be extended to derive interpretable models for other thermodynamic performance indicators.

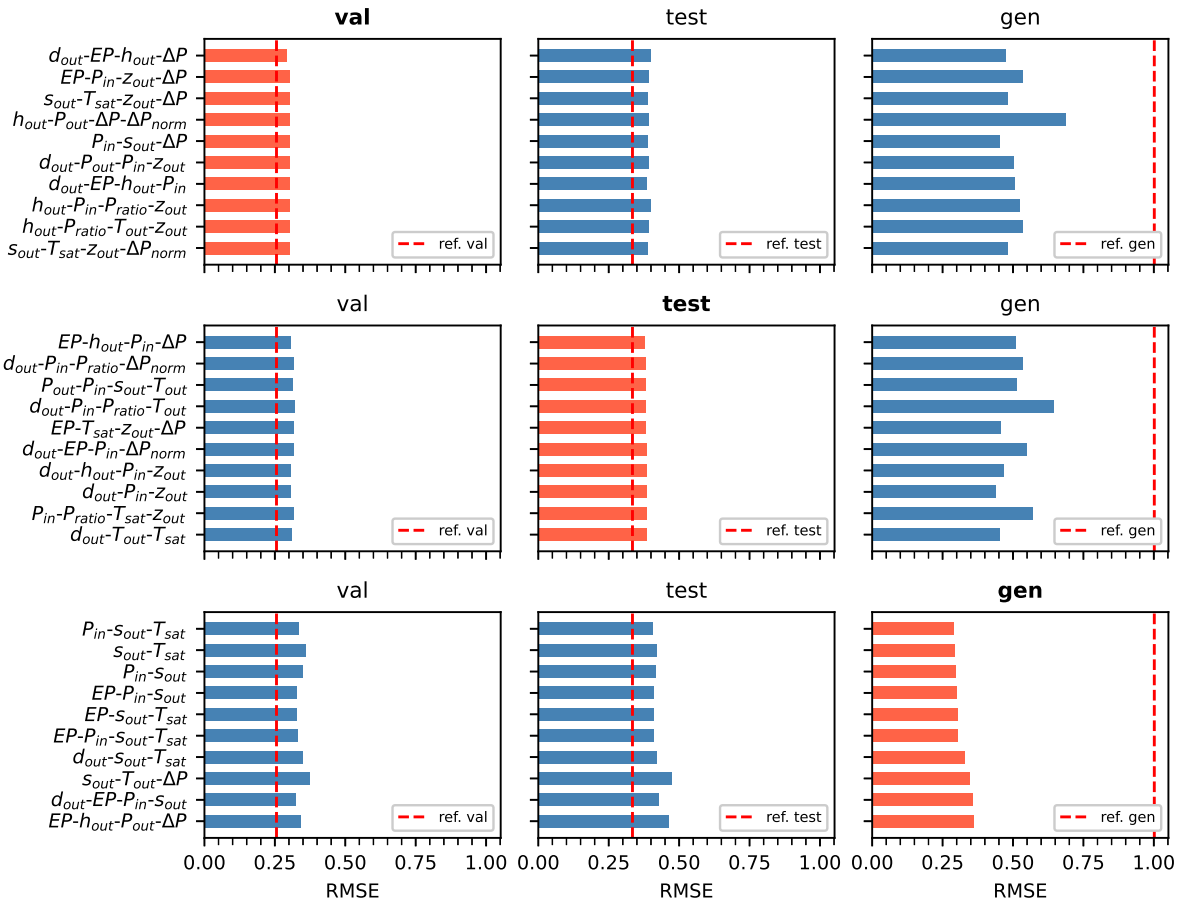


Figure 4. Top 10 feature sets ranked by validation (row 1), test (row 2), and generalization (row 3) RMSE. For each feature set, the displayed model is the one with the lowest generalization RMSE among all 100 hyperparameter trials. The dashed reference line corresponds to the raw-sensor baseline including electrical power consumption (EP).

## 6. CONCLUSIONS

This study investigated the role of physics-informed feature engineering in improving the robustness of virtual sensor models under sensor reduction. Using data from a heat pump system in a commercial building, neural network models were developed to estimate compressor isentropic efficiency when one of the physical sensors is unavailable. A large feature space was constructed from raw measurements and physics-derived transformations, and all feasible feature combinations were systematically evaluated.

For the heat pump system studied, excluding the suction temperature sensor ($T_{in}$) gave the lowest validation and generalization errors among the four sensor removal cases, identifying it as the most suitable candidate for virtual sensing. Under this scenario, the best-performing feature sets consistently combined discharge entropy ($s_{out}$) with suction pressure ($p_{in}$) and saturation temperature ($T_{sat}$). This configuration reduced generalization RMSE from 1.0017 (raw-sensor reference) to 0.3026, a 70% improvement, at the modest cost of an increase in test RMSE from 0.3343 to 0.3975.

Beyond these system-specific findings, the study supports several broader conclusions that may inform virtual sensing efforts in similar IIoT and PHM applications:

- **Physics-informed features play a key role in OOD robustness.** Across all sensor removal scenarios, the variance in generalization error across feature sets was substantially larger than the variance attributable to model architecture or hyperparameter tuning. Thermodynamically derived quantities, particularly entropy-based features, were consistently associated with lower OOD error than raw sensor measurements, suggesting that physics-derived representations remain more stable across operating regimes. Investment in physics-informed feature engineering can therefore yield larger gains in robustness than investment in model optimization.
- **Validation-based model selection is insufficient for OOD deployment.** Feature sets optimized solely on validation or test performance frequently exhibited poor general-

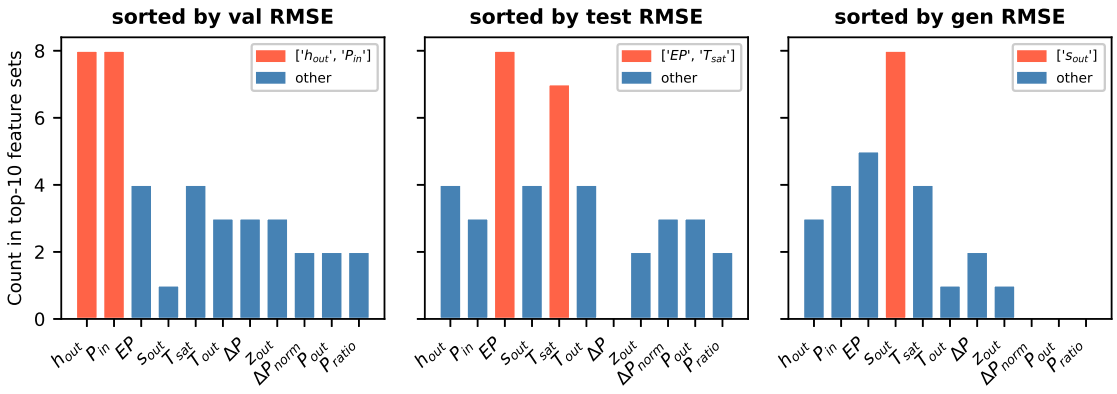


Figure 5. Top features by occurrence in the 10 best-performing feature sets, where sets are ranked separately by validation (row 1), test (row 2), and generalization (row 3) RMSE.

ization, in some cases worse than the raw-sensor baseline. Evaluation protocols that explicitly isolate unseen operating regions are necessary to identify configurations suitable for real-world deployment.

- **OOD robustness can be significantly improved with little cost in ID performance.** The most robust feature sets sacrifice a small amount of in-distribution accuracy in exchange for substantial gains in out-of-distribution performance, a trade-off that is generally favorable for industrial monitoring applications operating under variable conditions.

## ACKNOWLEDGMENTS

This project is funded by the Swedish Energy Agency and conducted within the PREMA-HEAPS project, a collaboration between RISE, Halmstad University, ClimaCheck, Vasakronan, and Enrad, which provided data from from a heat pump system in a commercial building used in this study. The authors would also like to thank Associate professor Janet Lin for providing valuable feedback that helped improve the quality of the manuscript.

## NOMENCLATURE

| | |
|---|---|
| $d$ | density |
| $h$ | specific enthalpy |
| $p$ | pressure |
| $EP$ | electrical power |
| $s$ | specific entropy |
| $T$ | temperature |
| $z$ | compressibility factor |
| $\Delta h_{is}$ | isentropic enthalpy difference |
| $\Delta p$ | pressure difference |
| $\Delta T$ | temperature difference |
| $\eta_{is}$ | isentropic efficiency |
| $sat^*$ | refers to saturated vapor |

### Subscripts and superscripts

| | |
|---|---|
| $in, out$ | inlet, outlet (for sensor values) |
| $is$ | isentropic outlet state |
| $norm$ | normalized quantity |
| $ratio$ | ratio of outlet to inlet |
| $super$ | superheat |

## REFERENCES

Bell, I. H., Wronski, J., Quoilin, S., & Lemort, V. (2014). Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library CoolProp. *Industrial & Engineering Chemistry Research*, *53*(6), 2498-2508.

Cui, J., & Wang, S. (2005). A model-based online fault detection and diagnosis strategy for centrifugal chiller. *International Journal of Thermal Sciences*, *44*(10), 986-999.

Guo, Y., Wang, N., Shao, S., Huang, C., Zhang, Z., Li, X., & Wang, Y. (2024). A review on hybrid physics and data-driven modeling methods applied in air source heat pump systems for energy efficiency improvement. *Renewable and Sustainable Energy Reviews*, *204*, 114804.

Kadlec, P., & Gabrys, B. (2009). Soft sensors: where are we and what are the current and future challenges? *IFAC Proceedings Volumes*, *42*, 572-577.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*, 422-440.

Li, G., Hu, Y., Liu, J., Fang, X., & Kang, J. (2021). Review on fault detection and diagnosis feature engineering in building heating, ventilation, air conditioning and refrigeration systems. *IEEE Access*, *9*, 2153-2187.

Li, S., Wu, F., Lin, W., Song, W., & Feng, Z. (2025). Identification of key parameters and construction of empir-

ical formulas for isentropic and volumetric efficiency of high-temperature heat pumps based on xgboost-mlr algorithm. *Energies*, *18*, 4454.

Niresi, K. F., Bissig, H., Baumann, H., & Fink, O. (2024). Physics-enhanced graph neural networks for soft sensing in industrial internet of things. *IEEE Internet of Things Journal*, *11*(21), 34978-34990.

Perera, Y. S., Ratnaweera, D. A. A. C., Dasanayaka, C. H., & .C, A. (2023). The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review. *Engineering Applications of Artificial Intelligence*, *121*, 105988.

Spitler, J. D., Berglöf, K., Mazzotti Pallard, W., & Witte, H. (2021). *IEA HPT Annex 52 - Long-term performance monitoring of gshp systems for commercial, institutional and multifamily buildings: Guidelines for calculation of uncertainties* (Tech. Rep.). IEA Heat Pump Technology.

Wang, Z., Guo, J., Xia, P., Wang, L., Zhang, C., Leng, Q., & Zheng, K. (2024). Feature selection for chillers fault diagnosis from the perspectives of machine learning and field application. *Energy and Buildings*, *307*, 113937.

Wang, Z., Xia, P., Guo, J., Zhou, S., Wang, L., Wang, Y., & Zhang, C. (2025). Efficient feature selection for enhanced chiller fault diagnosis: A multi-source ranking information-driven ensemble approach. *Building Simulation*, *18*, 141-159.

Wu, Y., Sicard, B., & Gadsden, S. (2024). Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring. *Expert Systems with Applications*, *255*, 124678.