KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-026-00904-4

TECHNICAL CONTRIBUTION

Correlations Matter in Explanations for Energy Systems

Alexandra Nikoltchovska[1] · Sebastian Pütz[1] · Markus Götz[2,3] · Benjamin Schäfer[1]

Received: 5 September 2025 / Accepted: 12 March 2026

Abstract
Machine learning models are increasingly deployed in critical energy infrastructure, where domain experts require transparent explanations for decision-making. SHAP (SHapley Additive exPlanations) has become a popular method for energy systems applications. However, energy data exhibit inherent correlations due to physical constraints, operational relationships, and market dynamics, posing challenges for interpreting SHAP-based explanations. This work investigates how feature correlations influence SHAP-based explanations using controlled synthetic experiments and real-world power grid data. Our analysis shows that only correlation-aware methods can attribute importance to economically linked features, such as solar generation in predicting fossil fuels, which may reflect genuine systemic interdependencies that are valuable for prediction and scientific understanding. Our findings highlight the tradeoff between *true to the model* explanations that reflect model behavior and *true to the data* approaches that consider real-world dependencies. In complex energy systems with circular dependencies, temporal dynamics, and hidden constraints, explanation validity cannot be universally defined. We emphasize the need for practitioners' awareness of the trade-offs between model analysis, scientific discovery, and operational understanding.

Keywords Machine learning · Explainable artificial intelligence · SHAP · Energy systems · Power grid · Feature correlations

Mathematics Subject Classification 68T01 · 62M10

1 Introduction

Machine Learning (ML) is becoming increasingly important across critical domains such as healthcare, finance, and energy systems, where domain experts require support for informed decisions. The complexity of ML models continues to grow with advancements in model architectures, computation power, and data availability [18]. This enables the analysis of complex relationships but makes models harder to understand.

Energy systems are essential for enabling economic development, technological progress, and societal well-being [36]. As the global demand for clean, reliable, and affordable energy grows, these systems must evolve to accommodate complex and often competing objectives [16]. While energy systems broadly include heating, mobility, and industrial processes, this work specifically considers electrical power systems–focusing on the infrastructure and operations involved in generating, transmitting, and distributing electricity [25].

✉ Alexandra Nikoltchovska
alexandra.nikoltchovska@kit.edu

Sebastian Pütz
sebastian.puetz@kit.edu

Markus Götz
markus.goetz@kit.edu

Benjamin Schäfer
benjamin.schaefer@kit.edu

[1] Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[2] Scientific Computing Center (SCC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[3] Helmholtz AI, Munich, Germany

Published online: 30 March 2026

In power systems specifically, ML applications include, for example, grid stability analysis, as well as accurate forecasting and demand-side management [3]. However, a fundamental challenge is that models are typically developed by ML experts with limited domain expertise. These models are then deployed by domain experts who often lack a detailed understanding of how the models make decisions.

1.1 Motivation and Problem Statement

As the complexity of energy systems is growing, incorporating dependencies and interactions between demand, generation, weather, and other factors, the gap between the expertise of ML developers and domain specialists continues to grow [40]. Deploying ML models in critical infrastructures, such as energy systems, requires the trust of domain experts, as these systems directly impact grid reliability and human lives [14]. For this reason, jurisdictions like the European Union (EU) now require explainability and human oversight measures for high-risk artificial intelligence (AI) applications under regulations such as the EU AI Act [38]. Unfortunately, energy system applications often lack transparency, which makes it difficult to adopt them [31, 32].

The need to bridge this gap has given rise to *Explainable Artificial Intelligence (XAI)*, a field dedicated to making model decisions interpretable and transparent to human users. XAI encompasses both interpretable models by design–such as logistic regression–and post-hoc methods that provide explanations for black-box model decisions after training [24].

SHAP (Shapley Additive Explanations) [22], based on Shapley values from game theory [35], has gained popularity in the research community and has been successfully applied to power grid frequency stability analysis [19, 28], renewable energy forecasting [26], and demand response analysis [2].

However, a critical challenge emerges: most SHAP implementations rely on the independence assumption, stating that a feature's value can be changed without affecting the values of other features [22]. In contrast, real-world energy systems data are inherently correlated. Generation types are coupled through dispatch economics, weather variables are seasonally dependent, and load patterns correlate across geographic regions. These relationships are fundamental to grid operation and cannot be discarded without losing important system understanding. This creates a fundamental trade-off: comprehensive data is necessary for reliable model performance, yet correlated data may compromise explanation reliability when using standard approximation approaches.

This work investigates how feature correlations influence SHAP-based explanations in energy systems. Using controlled synthetic experiments and real power grid data, we demonstrate that correlations can influence feature importance rankings and explanation interpretations across multiple SHAP explainers and discuss implications for deploying XAI in critical energy infrastructure. Data and code are available at [27].

Our analysis addresses three key questions:

1. How do different SHAP methods behave when increasing feature correlation?
2. Which methods remain stable when the independence assumption is violated?
3. What are the practical implications for energy system applications where correlations are unavoidable?

The remainder of this article is structured as follows. Section 2 provides background on SHAP fundamentals and examines the nature of correlations in energy systems data. Section 3 presents our experimental methodology using both synthetic and real-world datasets and reports our findings on how correlations affect different SHAP variants. Section 4 discusses the practical implications for energy system applications and provides recommendations for practitioners. Finally, Section 5 concludes with a summary of key insights and directions for future research.

2 Background and Related Work

To understand how feature correlations challenge explainability methods in energy systems, we first examine the inherent correlations present in energy data. Then, we provide the theoretical foundations of SHAP, and finally explore how SHAP methods handle (or fail to handle) these correlated features.

2.1 Inherent Correlations in Energy Systems Data

Real-world energy systems generate inherently correlated data due to fundamental physical constraints, operational relationships, and market dynamics. These correlations exist at multiple levels, ranging from obvious physical relationships to latent, e.g., economic, interdependencies that arise from market operations. In this work, we use hourly aggregated data from the Continental European synchronous area accessed through the ENTSO-E transparency platform [9] to analyze correlations representative of large interconnected power systems. For specific illustrative examples, we utilize higher-resolution German data (15-min intervals)

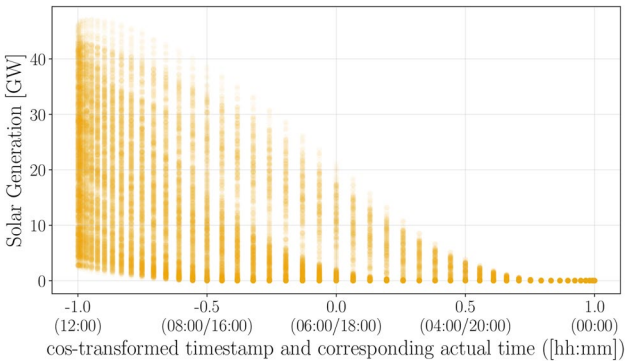
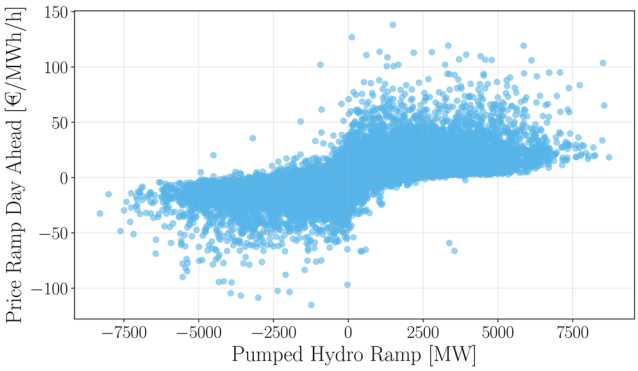


Fig. 1 Solar generation versus cosine-transformed time of day using 15-min resolution German data. Each point represents a 15-min interval of solar generation data plotted against $cos(2\pi t)$, where t is the normalized time of day. The cosine transformation captures the bell curve nature of solar irradiance throughout the day, with maximum values at midnight ($\cos = 1$) and minimum values at noon ($\cos = -1$). Data for 2024 from the ENTSO-E Transparency Platform [8]. Correlation coefficient: $\rho = -0.69$

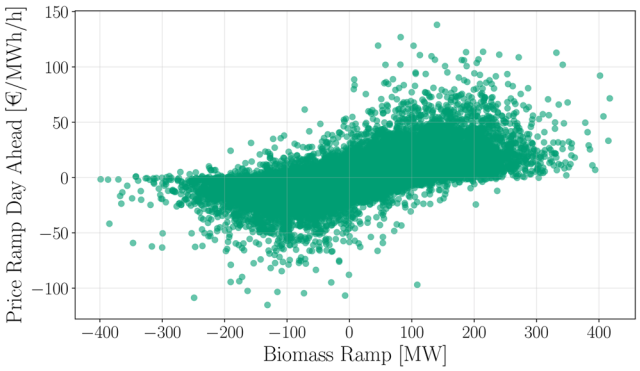
from the same platform to more clearly demonstrate correlation patterns.

Some correlations in energy data are highly intuitive. Figure 1 shows the natural relationship between solar generation and hour of day. When we apply cosine encoding to the hour of day to capture the bell curve nature of solar irradiance, we observe a strong correlation ($\rho = -0.69$) with solar generation. This relationship reflects the fundamental physics of solar energy: generation follows the sun's daily cycle, peaking at midday and dropping to zero at night.

More subtle correlations appear when examining the economic and operational dynamics of electricity markets. Figure 2 shows the relationships of day-ahead electricity price ramps with pumped hydro generation ramps and with biomass generation ramps, respectively.



(a) Pumped hydro ramp vs. price ramp day-ahead ($\rho = 0.63$)



(b) Biomass ramp vs. price ramp day-ahead ($\rho = 0.61$)

Fig. 2 Correlations in Continental European energy markets driven by economic dispatch principles. **a** Pumped hydro storage operators respond to price signals, creating a correlation between generation ramps and price movements as they arbitrage between low and high-price periods. **b** Biomass generation, as a dispatchable but expensive renewable source, correlates with price ramps as it typically ramps up when cheaper sources are unavailable and prices increase

These correlations reflect the complex dynamics of the market. Pumped hydro storage operators respond to price signals by pumping water uphill during periods of low prices and generating electricity during periods of high prices. On the other hand, as a relatively expensive but dispatchable renewable source, biomass generation often increases when prices rise and cheaper sources are unavailable, creating a correlation with price movements. The presence of these correlations in energy data is precisely why we examine their influence on SHAP explanations. While we cannot discard correlated features, as they represent real physical and economic relationships, we must consider their interdependencies when interpreting feature importance.

These correlations in energy data make us rethink what SHAP explanations are really telling us—and what purpose we want to use them for. But first, let's establish the theoretical foundations of SHAP.

2.2 SHAP Fundamentals and the Correlation Problem

SHAP explains machine learning predictions by computing how much each input feature contributes to moving the prediction away from a baseline (typically the average prediction across all training data). The method originates from Shapley values from cooperative game theory, where we can think of features as players in a game, and the payoff is the model's prediction. In game theory, Shapley values fairly distribute the total payoff among players based on their marginal contributions across all possible coalitions. For machine learning, this translates to: given a prediction of 0.8 and a baseline of 0.6, SHAP distributes the difference (0.2) among all input features such that their contributions sum exactly to this difference. The key challenge in adapting Shapley values to machine learning lies in defining what

happens when we "remove" a feature to compute its marginal contribution. Consider a model predicting solar power output using both "solar irradiance" and "time of day" as inputs. To calculate the importance of solar irradiance, we need to evaluate the model with and without this feature. But what value do we use when the feature is "missing"? Two approaches exist [5, 17]:

- Marginal sampling (interventional): Replace missing features with values sampled from their marginal distribution, assuming features are independent.
- Conditional sampling (observational): Replace missing features with values sampled conditional on the observed features, preserving dependencies in the data.

Most practical SHAP implementations use marginal sampling since conditional sampling means estimating high-dimensional conditional distributions, which is computationally expensive and often infeasible for complex datasets [1]. However, this creates our core problem: marginal sampling assumes feature independence, which is fundamentally violated in energy systems data.

2.3 True to the Model or True to the Data?

The choice between marginal and conditional sampling connects to a fundamental question raised by Chen et. al: should explanations be *true to the model* or *true to the data* [6]? If the goal is explaining model behavior, we want to understand how the trained model makes decisions using interventional conditional expectations (marginal sampling). This approach answers: "What has the model actually learned?" It may create unrealistic feature combinations that lie outside of the underlying data distribution. For instance, one might evaluate the model by asking how its prediction would change if there were 10GW of solar generation at 4 o'clock in the morning—a physically impossible scenario, since solar panels do not produce electricity without sunlight. If the goal is knowledge discovery, explanations should reflect true causal or physical relationships in the data using observational conditional expectations (conditional sampling). This approach answers: "What do the relationships in the real system tell us?" It respects natural correlations, providing explanations aligned with physical and economic realities.

In energy systems, both approaches have merit depending on the application. For system analysis and operator decision support, *true to the data* interpretations may be more valuable. For model analysis and validation, *true to the model* explanations might be necessary. However, the computational challenges of conditional sampling mean that most practitioners default to marginal sampling methods, potentially obtaining explanations that don't align with their intended use case.

2.4 Related Work

Several studies have investigated the behavior of SHAP explanations under different conditions. Fryer et al. [10] examined the theoretical foundations of Shapley values in machine learning contexts, highlighting potential pitfalls when assumptions are violated. Sundararajan et al. analyzed various attribution methods, including SHAP, and their sensitivity to different baseline choices and implementation details in [37]. Most relevant to our work, Olsen et al. provided a comprehensive comparison of different SHAP approaches in [29], particularly focusing on how conditional versus marginal sampling affects explanation quality across various datasets and model types. Their analysis demonstrated substantial differences in explanation outcomes depending on the chosen approach, with conditional methods generally providing more faithful explanations for correlated data at the cost of computational complexity. However, to our knowledge, no prior work has investigated how feature correlations specifically affect SHAP explanations in the context of energy systems applications.

3 Empirical Analysis of Correlation Impact

To systematically investigate the effect of feature correlations on SHAP explanations in energy systems, we design a controlled synthetic environment and then move on to real-world energy applications.

We structure our empirical analysis in three parts. First, we use synthetic data with precisely controlled correlation structures to isolate the effects of feature dependencies on SHAP explanations across multiple methods. This controlled environment enables us to establish ground truth and systematically vary correlation strength while observing the behavior of explanations. Second, we extend our analysis to real-world energy data from the Continental European synchronous area (preprocessed and aggregated hourly via the ENTSO-E transparency platform [9]) to demonstrate how these effects manifest in practice.

Finally, we assess the implications of these differences for energy system applications.

3.1 Experimental Design

Our synthetic data experiments are designed to isolate correlation effects while maintaining clear ground truth about feature relevance (Table 1). We generate synthetic datasets with three features where we precisely control correlation

Table 1 Overview of parameters for synthetic data generation experiments designed to study correlation effects on feature importance methods. The experimental setup generates datasets with three features: two independent predictive features (x_1, x_2) and one feature (x_3) that is systematically correlated with x_2 at varying strengths ($\rho = 0.1$ to 0.9)

Parameter	Value/Range	Description
Data set size	{1000, 10 000, 100 000}	Total number of samples per experiment
Feature 1 (x_1)	$\mathcal{N}(0,1)$	Predictive feature, independent
Feature 2 (x_2)	$\mathcal{N}(0,1)$	Predictive feature, independent
Correlation (ρ)	{0.1, 0.2, ..., 0.9}	Systematic correlation variation
Feature 3 (x_3)	$\rho \cdot x_2 + \sqrt{1-\rho^2} \cdot \varepsilon_3$	Irrelevant but correlated with x_2
Noise (ε_3)	$\mathcal{N}(0,1)$	Independent noise for x_3

structures and know exactly which features should receive importance attributions.

The data generation process creates independent predictive features x_1 and x_2, both drawn from standard normal distributions $\mathcal{N}(0,1)$. To examine correlation effects, we introduce a third feature x_3 that is correlated with x_2 but irrelevant to the prediction target. This correlation probe is constructed as shown in Table 1. This construction ensures that x_3 maintains unit variance while achieving precise correlation levels with x_2.

We implement two distinct target functions to capture both linear and non-linear relationships. The linear target function y_{lin} is given by Eq. (1), while the non-linear variant y_{nonlin} includes an interaction term as shown in Eq. (2).

$$\begin{aligned} y_{\text{lin}} &= x_1 + 2x_2 + \varepsilon_{\text{lin}}, \\ \varepsilon_{\text{lin}} &\sim \mathcal{N}(0, 0.1) \end{aligned} \tag{1}$$

$$\begin{aligned} y_{\text{nonlin}} &= x_1 + 2x_2 + x_1 \cdot x_2 + \varepsilon_{\text{nonlin}}, \\ \varepsilon_{\text{nonlin}} &\sim \mathcal{N}(0, 0.1) \end{aligned} \tag{2}$$

In both cases, only x_1 and x_2 contribute to the output while x_3 serves as a correlated but irrelevant feature. The interaction term in y_{nonlin} introduces non-linear dependencies to test SHAP behavior under more complex feature relationships commonly encountered in energy systems applications.

We train two model types on each target function to examine how model complexity interacts with correlation effects in SHAP explanations. Linear Regression models represent simple parametric approaches with known ground truth coefficients, while XGBoost [7] models represent complex non-parametric methods commonly used in practice. For each model-target combination, we generate datasets of varying sizes (see Table 1) samples for training). All models achieved high predictive performance ($R^2 > 0.95$) across all experimental conditions

We evaluate following explanation approaches to assess their behavior under varying correlation structures:

1. *Kernel SHAP* [22]: The original model-agnostic method that uses coalition sampling to approximate Shapley values through specially-weighted least squares regression.
2. *Permutation SHAP* [22]: The current default model-agnostic method that iterates over complete permutations of features to minimize model evaluations while maintaining efficiency. It assumes feature independence and uses marginal sampling for background distribution estimation, which can lead to unreliable results when features are correlated.
3. *Linear SHAP, interventional* [22]: Assumes feature independence when computing Shapley values for linear models.
4. *Linear SHAP, correlation dependent* [22]: Accounts for feature correlations when computing Shapley values for linear models; this distinction between interventional and observational expectations is discussed in [6].
5. *Tree, interventional* [21]: Uses marginal (interventional) conditional expectation, effectively breaking dependencies and assuming feature independence during computation.
6. *Tree SHAP, path-dependent* [21]: Follows decision paths through the model and is theoretically designed for exact computation of feature contributions.
7. *Partition SHAP* [22]: A model-agnostic method that computes Shapley values recursively using hierarchical clustering to group correlated features, resulting in Owen values from cooperative game theory [30]. Owen values extend Shapley values by considering coalition structures, treating feature groups as single players and distributing coalition Shapley values among individual features within each group.
8. *Radical SHAP:* An implementation approaching the original Shapley value concept by training separate models on each of the 2^n possible feature subsets, where *n* is the number of features in the dataset [11]. While all other SHAP explainers define a feature's contribution relative to its value being unknown for a given prediction—simulating this by marginalizing over absent features, either through sampling or analytical

integration—Radical SHAP instead measures each feature's contribution relative to its complete absence from the model during training. It therefore does not belong to either the independence-assuming or correlation-aware category, as that distinction applies to methods that define coalition values through marginalization. Nevertheless, its retrain-per-coalition procedure can produce correlation-sensitive attributions: when the predictive feature x_2 is absent from a coalition, a correlated feature x_3 can partially substitute for it during retraining, absorbing some importance.

9. *Normalized regression coefficients:* As baseline *true to the model* feature importance for the linear regression models.

This selection provides comprehensive coverage of correlation handling approaches (independence-assuming vs. correlation dependent), spans multiple model types (model-agnostic, linear-specific, tree-specific), and includes both current defaults and theoretically superior alternatives, as well as a retraining-based method, enabling systematic analysis across varying correlation structures.

For quantitative comparison, we calculate the relative feature importance as the absolute SHAP value normalized by the sum of all absolute SHAP values:

$$\text{Relative SHAP Value}_i = \frac{\frac{1}{M}\sum_{k=1}^{M}|\phi_{i,k}|}{\sum_{j=1}^{N}\frac{1}{M}\sum_{k=1}^{M}|\phi_{j,k}|} \tag{3}$$

where $\phi_{i,k}$ represents the SHAP value for feature *i* in sample *k*, *M* is the total number of samples, and *N* is the total number of features.

3.2 Results: Linear Target Function

Analyzing SHAP methods under different correlation structures we observe distinct patterns across the explanation methods.

Figure 3 presents the relative SHAP values for each feature as the correlation ρ between x_2 and x_3 increases from 0.1 to 0.9 for the linear target function. The results show a clear split in how the methods behave. Independence-assuming methods (Kernel SHAP, Permutation SHAP, Linear SHAP interventional) maintain stable attribution patterns across all correlation levels. In contrast, the correlation-aware Linear SHAP method and the Radical Explainer show systematic changes in feature importance as correlation increases.

Independence-assuming methods maintain relatively consistent relative importance around 0.33–0.335 for the uncorrelated feature x_1 throughout the correlation range, with slight variations of about 1%, which closely matches the theoretical model coefficient. The slight U-shaped variation visible for x_1 under correlation dependent Linear SHAP and Radical SHAP is a normalization artifact: the sum of absolute SHAP values in the denominator of Equation (3) varies non-monotonically as importance is redistributed between x_2 and x_3. This pattern persists across sample sizes $n \in \{1000, 10000, 100000\}$ but remains negligible (variation < 0.03) and does not affect our conclusions.

The correlated predictive feature x_2, which is the first feature we manipulate in our experimental design, shows a high sensitivity to correlation level. Independence-assuming methods maintain stable relative importance around 0.66, consistent with the feature's doubled coefficient in the linear target function. However, correlation-aware methods show substantial decreases in attributed importance, declining linearly from approximately 0.66 to 0.39 as correlation increases—a reduction exceeding 35%. This shift occurs despite no change in the feature's actual predictive contribution to the target function and reflects almost exactly the data generation process. Similar to Feature 2, feature x_3 shows contrasting behavior between method types, but with an inverse pattern. Independence-assuming methods assign minimal importance (≤ 0.01) across all correlation levels, while correlation-aware methods increase from ca. 0.03 to 0.28–0.29 as correlation increases. The near-zero importance attribution follows directly from how the independence-assuming methods compute feature contributions. These replace x_3 with values sampled from its marginal distribution independently of all other features. Since the model learned a coefficient of approximately zero for x_3, perturbing it produces no meaningful change in model output regardless of ρ. The correlation between x_2 and x_3 is irrelevant because all feature dependencies are broken by construction. Correlation-aware methods, by contrast, respect the joint distribution of x_2 and x_3, causing importance to flow between them proportionally to ρ — reflecting the data generation process. In summary, the correlation-aware Linear Explainer, as well as the Radical Explainer, systematically redistribute the importance between the correlated features x_2 and x_3. As correlation increases, importance flows from the predictive x_2 to x_3, with their combined importance remaining approximately constant. This behavior is consistent with the data generation process.

3.3 Results: Non-linear Target Function

The nonlinear target function trained with XGBoost shows different correlation sensitivity patterns compared to the

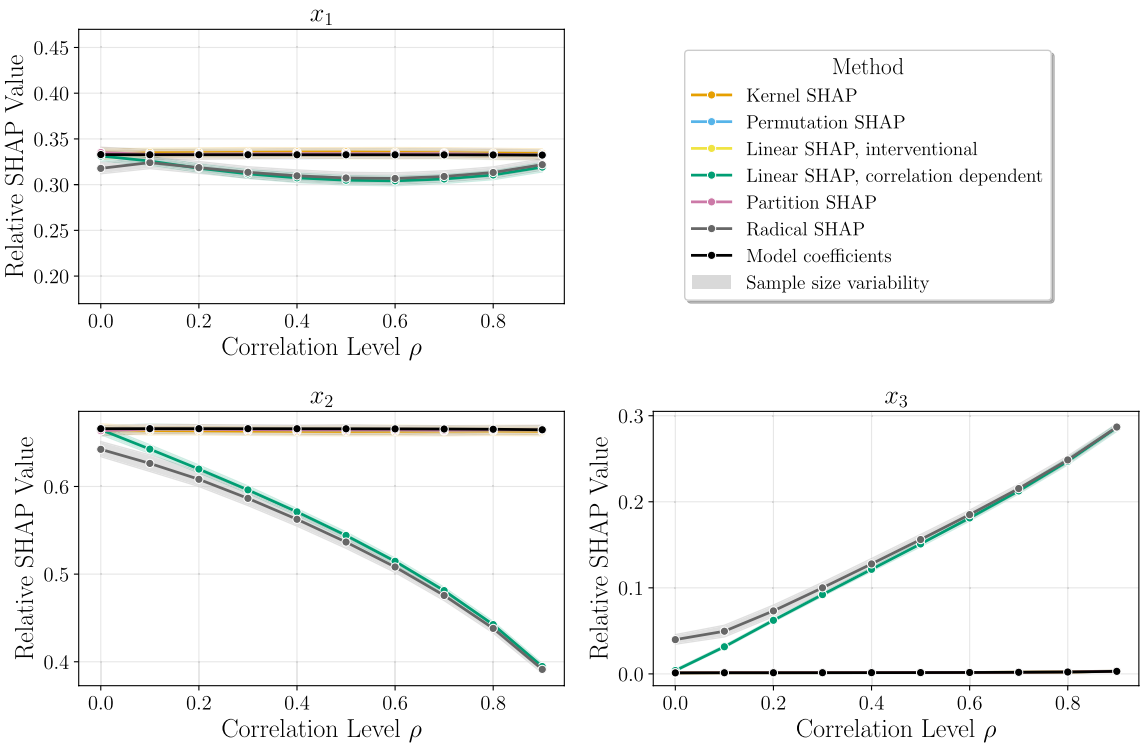


Fig. 3 Relative SHAP values for three features (x_1, x_2, x_3) across different SHAP methods as correlation ρ between features x_2 and x_3 increases from 0.1 to 0.9 in the linear target function. The results show that method behavior fundamentally differs based on correlation assumptions. Independence-assuming methods (Kernel, Permutation, Linear interventional) maintain stable feature attributions across all correlation levels. On the other hand, correlation-aware methods redistribute attribution based on data structure rather than predictive importance - Linear correlation systematically transfers importance from the truly predictive feature x_2 to the correlated feature x_3 as correlation increases. This attribution redistribution follows the *true to the data* paradigm, where explanations preserve the observed correlational dependencies by considering the joint feature distribution rather than manipulating features independently. Radical SHAP, which defines coalition values through retraining rather than marginalization, shows a similar pattern through a distinct mechanism: when x_2 is absent from a training coalition, the correlated x_3 can partially substitute for it during retraining, producing elevated attribution at high correlation levels

linear case. Figure 4 presents the relative SHAP values for each feature as the correlation between x_2 and x_3 increases from 0.1 to 0.9 for the XGBoost model.

According to SHAP documentation, the tree-path-dependent version of the Tree Explainer should exhibit correlation awareness for tree-based models, while the interventional version assumes independence.

For the correlated predictive feature x_2, all explainers show some degree of importance redistribution as correlation increases, but the magnitude is substantially reduced compared to the linear case. The tree-path-dependent version of the Tree Explainer maintains the most stable attribution, while other methods show gradual decreases in x_2 's importance. Similarly, feature x_3, i.e., non-predictive but correlated, shows increases in attributed importance across most explainers, but again with much smaller magnitude changes than observed in the linear target function. All explanation methods demonstrate substantially lower correlation sensitivity for the XGBoost model than for the linear model. The redistribution of importance between the correlated features x_2 and x_3 is present but much smaller in magnitude. Radical SHAP shows qualitatively different behavior from all other methods in the non-linear case. Unlike in the linear setting, it assigns substantial importance to x_3 even at $\rho = 0$, with only a slight upward trend as correlation increases. This occurs because when x_3 is included in a coalition without x_2, the retrained model learns to exploit x_3's independent variance, assigning it importance regardless of its correlation with x_2. As this behavior does not closely align with either the *true to the model* or *true to the data* paradigm in the non-linear setting, we do not consider Radical SHAP further in our analysis. The reduced redistribution in the non-linear case has an important implication: none of the tested methods act as a genuine *true to the data* explainer for non-linear models. The correlation-aware Linear SHAP is only applicable to linear models by construction. For tree-based models, the path-dependent Tree SHAP is often described as correlation-aware, yet our results show it behaves similarly to independence-assuming methods. As Molnar [23] explains, this method pushes all possible feature subsets down the tree simultaneously, meaning attributions are entirely determined by the tree's split structure.

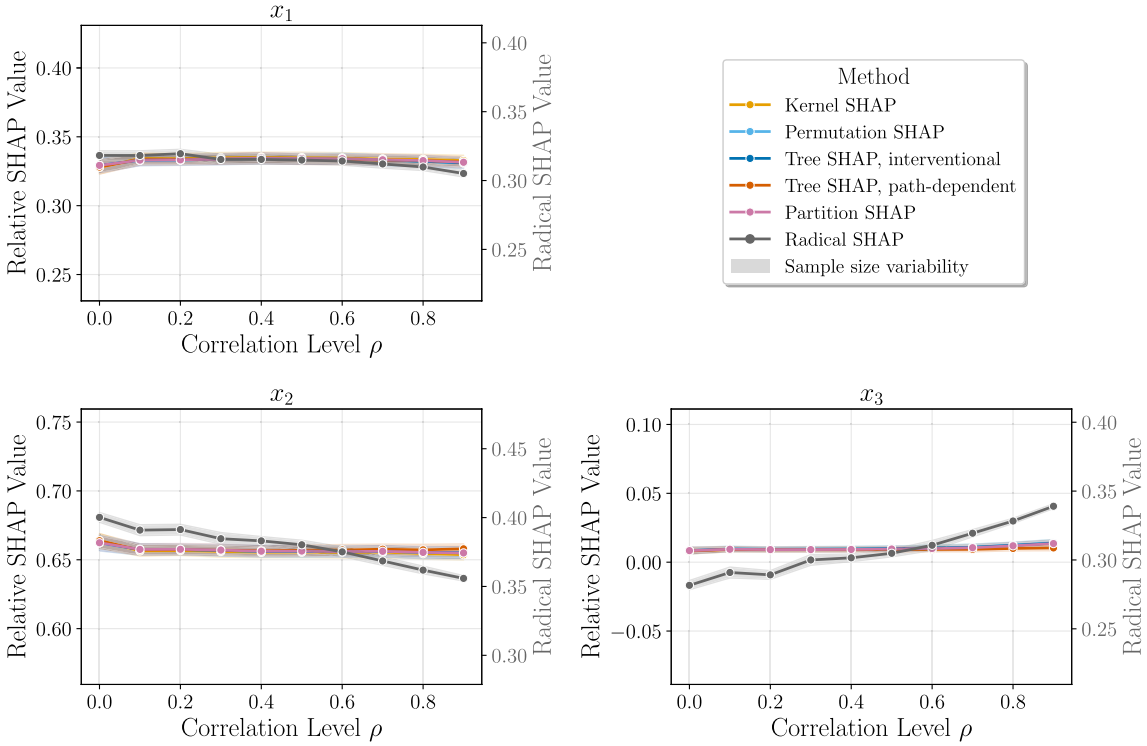


Fig. 4 Relative SHAP values for three features (x_1, x_2, x_3) across different SHAP methods as correlation ρ between features x_2 and x_3 increases from 0.1 to 0.9 in the non-linear target function trained with XGBoost. Contrary to expectations, the tree-path-dependent Tree SHAP method behaves similarly to other independence-assuming methods (Kernel SHAP, Permutation SHAP, Tree SHAP interventional) maintaining stable feature attributions across correlation levels rather than following the *true to the data* paradigm. Radical SHAP uses a separate right-hand axis as its values differ from the other methods

When two features are correlated during training, XGBoost assigns splits to whichever reduces impurity more—here x_2—so x_3 rarely appears in decision paths and receives only a small attribution regardless of its correlation. The method therefore reflects the model's internal feature selection rather than the statistical dependencies in the data. A genuine *true to the data* approach for non-linear models would require proper conditional sampling, which remains computationally intractable [1, 29].

3.4 Results: Application to Real Energy Data

To validate our findings in an energy systems context, we apply our correlation analysis to real energy data from the ENTSO-E transparency platform [9] for the Continental European synchronous area covering the period 2020–2024. The feature set includes hourly measurements of seven variables: hard coal, lignite, gas and coal gas generation, solar generation, and load. We construct our target variable as the total fossil generation, calculated as the sum of gas generation, hard coal generation, lignite generation, and coal gas generation. We add Gaussian noise with standard deviation equal to 5% of the target variable's standard deviation to simulate realistic measurement uncertainty commonly observed in energy system data acquisition and aggregation processes. This level is consistent with documented uncertainties in power generation measurements, where similar error ranges of 5–8% are commonly reported in the literature [13, 15]. We apply standard scaling to normalize all features and the target variable before model training to enable meaningful comparison of SHAP values with model coefficients across features with different units and scales. We use both XGBoost and Linear Regression algorithms to predict the total fossil generation for each hour. Both models achieve a strong performance with $R^2 > 0.95$ due to the straightforward linear relationship between individual generation sources and their sum. This experimental setup creates a controlled scenario where the data-generating process is known: each fossil fuel generation type contributes linearly to the target, while load and solar generation have no direct mathematical relationship to the target variable. However, this mathematical simplicity masks the complex reality of energy systems. In practice, dispatchable fossil fuel generation (natural gas power plants in particular) responds to residual load—the electricity demand remaining after accounting for must-run capacity and renewable generation [4]. Thermal power plants face operational constraints due to startup times and must-run requirements,

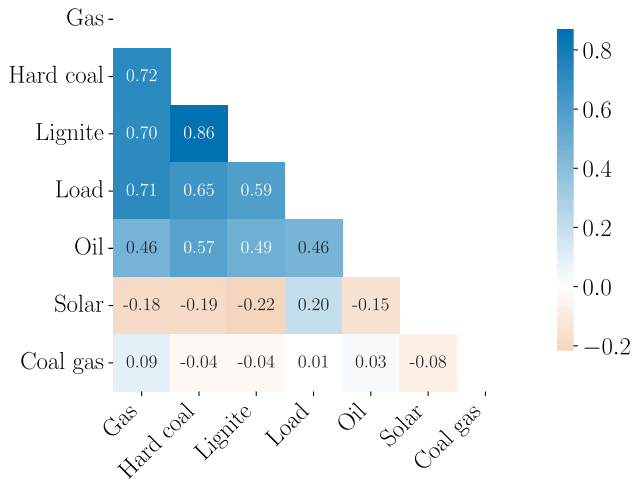


Fig. 5 Correlation heatmap showing Pearson correlation coefficients ρ between the features in our data set

with coal plants often operating under longer-term scheduling to avoid frequent cycling [20]. Solar generation can also influence dispatch decisions through merit order effects, displacing conventional energy sources when renewable output is high. This creates a system of circular dependencies, where features that do not directly contribute to the target variable are nevertheless (causally) linked to those that do. Unlike typical machine learning problems where true feature importance is unknown, we can assess whether SHAP explanations capture the mathematical relationship or the broader (causal) interdependencies within the energy system.

Figure 5 shows the correlation heatmap for our feature set. Multiple feature pairs show strong linear correlations ($\rho > 0.7$). Lignite and hard coal generation show the strongest correlation ($\rho = 0.86$), reflecting closely aligned dispatch patterns. Both are also positively correlated with gas generation ($\rho \approx 0.7$), due to economic merit order dispatch–these fossil fuel sources are typically called upon under similar market conditions when cheaper base load or renewable resources are insufficient [34].

We apply the same SHAP methods from our synthetic analysis to evaluate how the correlations in our data affect explanations in this controlled energy context. We compare Tree SHAP (interventional and path-dependent), Permutation SHAP, Partition SHAP, and Linear SHAP(correlation dependent and interventional) for both prediction models. Figure 6 shows the relative SHAP values for each feature across the different explainers.

True to the model Explainers (Tree SHAP interventional, Permutation SHAP, Linear interventional) produce results that align with the mathematical ground truth. Gas, hard coal, and lignite generation receive the highest importance, which reflects their direct contribution to the target variable. Coal gas and oil generation also receive near-zero attributions, consistent with the model coefficients of the linear model. Note that while all features contribute equally to the raw sum, the normalized model coefficients appear smaller because they are scaled to match the target variable's magnitude. Solar generation and load receive near-zero attributions as expected, since they do not contribute to the target variable. Interestingly, although the path-dependent Tree SHAP method should theoretically be *true to the data*, it produces almost identical results to the interventional Tree SHAP version and the other *true to the model* explainers. This shows that the tree structure may not be greatly affected by the correlation patterns in our generation data. The attributions of Partition SHAP also closely align with the latter models.

The *true to the data* correlation dependent Linear SHAP explainer shows a different behavior from the *true to the model* methods. Individual fossil generation sources receive reduced importance compared to *true to the model* methods, with especially gas and hard coal receiving smaller attributions. Oil and solar generation gain some importance despite their minimal or zero mathematical contribution to the target sum. This reflects the interdependent nature of energy systems where features are correlated due to grid balancing and operational constraints.

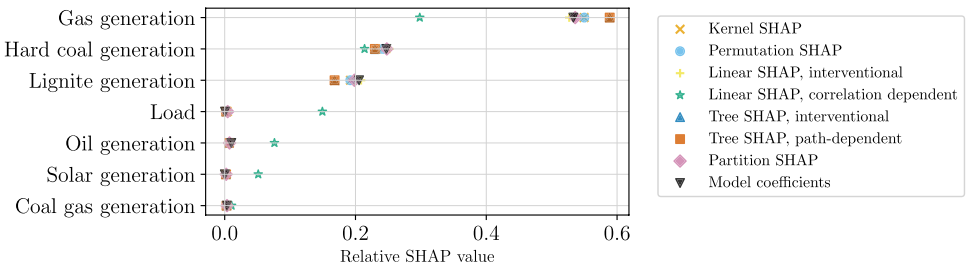


Fig. 6 Relative SHAP values across different explainers for total fossil generation prediction. *True to the model* explainers (Tree SHAP interventional, Permutation SHAP, Linear interventional) identify gas, hard coal, and lignite as primary contributors while assigning near-zero importance to solar generation and load. The correlation dependent Linear SHAP method captures energy system interdependencies, redistributing importance from major fossil sources to correlated features like load and solar generation, reflecting operational constraints and grid balancing requirements

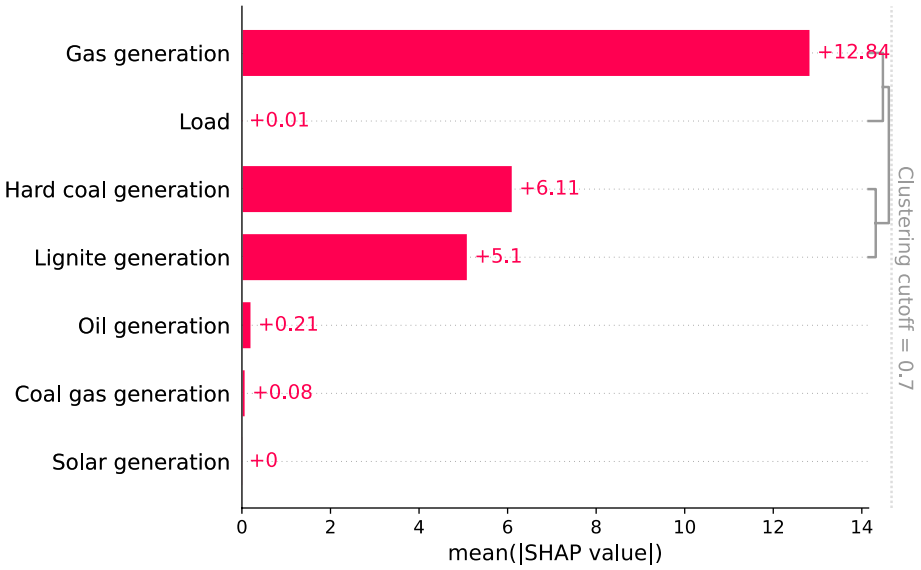


Fig. 7 Partition SHAP feature importance with hierarchical clustering. Partition SHAP provides *true to the model* SHAP values (horizontal bars) while showing data correlations through hierarchical clustering (dendrogram on the right). Despite being a *true to the model* approach, the clustering shows that load and gas generation are grouped together, providing additional context about the interdependent relationships within the energy system data

Partition SHAP occupies a unique position between *true to the model* and *true to the data* approaches. While it is based on a *true to the model* foundation, it explicitly incorporates the correlation structure of the data through hierarchical clustering of features.

Figure 7 shows the mean absolute SHAP values calculated with Partition SHAP along with the underlying feature clustering structure. Individual features maintain importance levels similar to *true to the model* explainers, correctly identifying the mathematical contributors to the target variable. However, the hierarchical clustering shows how correlated generation sources are grouped together, providing additional context about the interdependent relationships within the energy system data. This hybrid approach allows analysts to understand both the direct mathematical contributions and the correlation structure simultaneously, offering a more nuanced view of feature relationships in complex energy systems.

4 Discussion

Our experiments demonstrate a clear split between two types of SHAP methods. Independence-based methods—like Permutation SHAP, Kernel SHAP, and Tree SHAP—maintain stable attributions across all correlation levels. Correlation-aware methods—like Linear SHAP, correlation dependent—redistribute importance as feature correlations strengthen. This difference reflects two ways of thinking about explanations, as described by [17]: *True to the model* methods explain what the model actually learned, simulating feature removal through marginal sampling. With their interventional approach, correlations between features are intentionally broken, allowing attribution of importance to each feature based on its direct effect on the model, rather than effects inherited through correlated features. They assume features are independent and simulate removing them by sampling replacement values from the marginal distribution of the background dataset. While independence-based methods may generate off-manifold samples when features are correlated, this can also be viewed as a feature rather than a limitation. By intentionally breaking correlations, these methods reveal what the model has actually learned about each feature's individual contribution, independent of confounding relationships in the training data. This interventional perspective provides insights into the model's causal structure that remain valid even when features are dependent—they simply answer a different question about model behavior. *True to the data* methods, on the other hand, consider the real-world relationships between features. They respect the correlational structure present in the observed dataset, keeping the natural connections between features when calculating attributions. This creates a gap: When correlations exist in data, *true to the model* explanations may not reflect reality since the model is evaluated on impossible feature combinations. When features are correlated, *true to the data* methods distribute importance across all dependent variables, potentially making each appear less important individually. *True to the model* methods concentrate attribution on whichever feature the model actually uses, even when others carry equivalent information. This challenge becomes clear in our real-world energy data study. Independence-based methods show that fossil fuel sources directly affect the target and reflect the mathematical relationship. Correlation-aware methods shift some importance to solar power and energy demand. These don't directly affect the target, but they're still linked

through economic rules (like dispatch order) that aren't visible in the data.

4.1 The Challenge of Explaining ML Models for Energy Systems

Energy systems present unique challenges for XAI, with effects that extend far beyond the linear correlations demonstrated in our analysis. These systems are characterized by complex interdependencies that contradict the independence assumptions on which many SHAP methods are based. Bidirectional causal relationships create circular dependencies, beyond the scope of traditional feature attribution.

In our correlation-aware methods, solar generation appears to drive total generation, yet solar output is actually dispatched in response to load demand. At the same time, load patterns are influenced by factors such as pricing, system reliability and available generation. This results in a feedback dynamic where supply and demand change together rather than a simple cause-and-effect relationship.

Temporal dynamics add another layer of complexity. Energy systems operate across multiple time scales at the same time: seasonal weather patterns affect how much renewable energy is available, daily demand cycles create predictable levels of energy use, and real-time market mechanisms balance supply and demand at any given moment. The importance of a feature can vary depending on the time period and the time of year. Moreover, the correlations within these systems extend beyond simple linear relationships and are themselves subject to temporal variation. These relationships can evolve gradually and naturally through underlying drift processes, or change abruptly due to external interventions such as regulatory changes. For example, recent work has investigated [33] how regulatory changes in energy markets can instantly influence the fundamental relationships between system variables and thus fundamentally change model parameters. This web of interdependence, combined with the temporal variability of and inter-feature correlations, makes it conceptually problematic to isolate the global 'contribution' of any single feature.

Furthermore, economic and regulatory constraints can create dependencies not captured by the training data. When SHAP methods redistribute attribution based on these observed correlations, they might be capturing these hidden economic rules instead of direct causal effects. These challenges show a fundamental problem with XAI for complex systems. Independence-assuming methods may provide mathematically clean attributions that don't reflect real-world constraints, but correlation-aware methods risk attributing importance to what may be simply spurious statistical associations.

Speaking of causal effects, methods like Causal SHAP [12] and Shapley Flow [39] offer the possibility of incorporating causal information into feature attribution. However, these approaches require practitioners to provide an existing causal graph—a challenging task that itself involves making strong assumptions about the system's causal structure. As Chen et al. point out [5], all SHAP methods inherently assume some form of causal graph structure, whether explicitly acknowledged or not. Independence-based methods correspond to assuming no causal edges between features, while correlation-aware methods assume connectivity between all variables. This means that the choice of SHAP method is fundamentally a choice about the assumed causal structure of the system. Importantly, while each approach has distinct theoretical advantages and limitations that practitioners should be aware of, they also differ in their computational requirements—a consideration that is often limiting in practice [5]. Independence-based methods are computationally straightforward, requiring only sampling from marginal distributions, while correlation-aware methods face exponential complexity when modeling conditional dependencies. This means that the theoretical superiority of a method may be irrelevant if it cannot be implemented efficiently for a given problem size. A notable asymmetry exists between linear and non-linear settings: while correlation dependent explanations are achievable for linear models via Linear SHAP, no practical *true to the data* method is available within the widely-used SHAP package for non-linear models. While research has proposed conditional sampling approaches for model-agnostic explanations [1, 29], these remain outside the standard tooling that most practitioners in the energy domain rely on. The path-dependent Tree SHAP, despite its theoretical orientation toward conditional expectations, is constrained by the tree's split structure [23] and behaves similarly to independence-assuming methods in practice, as our results confirm. The choice between *true to the model* and *true to the data* explanations is therefore in practice often not a free choice—for non-linear models, which dominate practical energy systems applications, practitioners are effectively limited to *true to the model* explanations regardless of their intended use case.

4.2 Revisiting Motivation for XAI

Having put out results into context, it is worth revisiting why we want explanations in the first place. Traditional motivations for XAI include scientific discovery of system behavior and dependencies, increasing trust in AI models and support for decision-making in grid operations. However, without a comprehensive understanding of the underlying causal mechanisms, explanations may give us false confidence rather than useful scientific discovery. This creates

a circular dependency between the quality of explanations and domain knowledge. The trustworthiness of explanations relies on sufficient domain expertise to validate them and put them into context. However, if we know all the mechanisms in our data, the necessity for explanations is reduced. The challenge intensifies when domain knowledge is incomplete or when there are patterns not captured in historical data. With increasing usage of black-box inverters in energy systems, even grid operators might not have complete domain knowledge of their system. Similarly, consumers with decentralized PV, EV, or battery storage systems introduce components, which we will always have incomplete knowledge of. The question then becomes: how do we validate explanations when we have a limited understanding of the domain or when complexity makes it difficult to put our knowledge into context? Therefore, XAI methods can be used to generate ideas for new hypotheses, rather than viewing them as a definitive answer to our questions. While explainable AI holds promise for supporting scientific discovery, domain experts must be cautious: explanations can foster confirmation bias, leading researchers to validate preconceptions and neglect patterns that challenge established beliefs. They can highlight patterns and relationships worthy of further investigation, but should not be treated as authoritative explanations of system behavior without additional validation through domain expertise, controlled experiments, or causal analysis.

5 Conclusion

This work investigated how feature correlations influence SHAP-based explanations in energy systems through controlled synthetic experiments and real Continental European power grid data. Our analysis confirmed that SHAP methods fundamentally split into two categories: independence-assuming methods that maintain stable attributions across correlation levels, and correlation-aware methods that redistribute importance based on statistical dependencies. In synthetic experiments, correlation-aware methods shifted feature importance by over 35% for correlated features, while independence-based methods changed less than 0.2%. This split corresponds directly to the *true to the model* versus *true to the data* paradigm. Using real-world energy data from the Continental European synchronous area, exhibiting strong correlations ($\rho > 0.7$) between variables, we showed that these effects manifest in practice. Our analysis illustrates a principal challenge in applying XAI to complex energy systems: the discrepancy between model behavior, what the data represent, and how humans interpret the results.

No single approach captures all aspects—mathematical accuracy, physical relationships, and operational understanding. Rather than viewing this as a limitation to overcome, acknowledging these tensions explicitly leads to more honest and ultimately more useful explanations. When using XAI for scientific discovery, this implies a need for careful interpretation and validation of insights. Explanations may reinforce confirmation bias rather than reveal new insights, create false confidence when domain knowledge is incomplete, and potentially mislead conclusions about system behavior and causality. Distinguishing between correlation-based patterns and genuine causal relationships requires domain knowledge, highlighting the essential need for collaboration with experts throughout XAI method development and deployment. Practitioners must choose methods based on specific use cases while acknowledging existing limitations. Specifically, most non-linear models will only derive *true to the model* explanations with the analyzed algorithms.

Future work could investigate causal discovery integration to combine XAI with causal inference methods for better distinguishing between correlation-based and causal feature relationships in energy systems. One further promising direction could involve the development of domain-informed explanation methods and systematic approaches for energy domain experts to assess and validate machine learning explanations against known physical and economic principles. While we cannot eliminate the tensions between model fidelity, data reality, and human interpretation, explicit acknowledgment of these trade-offs enables more responsible deployment of XAI in critical energy infrastructure.

Acknowledgements We gratefully acknowledge funding from the Helmholtz Association under grant no. VH-NG-1727, the Networking Fund through Helmholtz AI and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—556503410.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data and Code Availability The code and data used in this study are publicly available at https://github.com/KIT-IAI-DRACOS/correlations-xai-energy. The SHAP variants used in this study are implemented in the open-source SHAP package https://shap-community.readthedocs.io/en/latest/index.html. The Radical SHAP implementation is available at https://github.com/edden-gerber/radical-shapley-values. The energy generation data is sourced from the ENTSO-E Transparency Platform at https://transparency.entsoe.eu/.

References

1. Aas K, Jullum M, Løland A (2020) Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. https://doi.org/10.48550/arXiv.1903.10464. ArXiv:1903.10464 [stat]
2. Antonopoulos I, Robu V, Couraud B, Flynn D (2021) Data-driven modelling of energy demand response behaviour based on a large-scale residential trial. Energy AI 4:100071. https://doi.org/10.1016/j.egyai.2021.100071https://www.sciencedirect.com/science/article/pii/S2666546821000252
3. Aslam S, Aung PP, Rafsanjani AS, Majeed APPA (2025) Machine learning applications in energy systems: current trends, challenges, and research directions. Energy Inform 8(1):62. https://doi.org/10.1186/s42162-025-00524-6
4. Berlin DIW DIW Berlin: Residual load, renewable surplus generation and storage requirements in Germany. https://www.diw.de/de/diw_01.c.458121.de/publikationen/diskussionspapiere/2013_1316/residual_load__renewable_surplus_generation_and_storage_requirements_in_germany.html
5. Chen H, Covert IC, Lundberg SM, Lee SI (2023) Algorithms to estimate Shapley value feature attributions. Nat Mach Intell 5(6):590–601. https://doi.org/10.1038/s42256-023-00657-x (https://www.nature.com/articles/s42256-023-00657-x)
6. Chen H, Janizek JD, Lundberg S, Lee SI (2020) True to the model or true to the data?. https://doi.org/10.48550/arXiv.2006.16234. ArXiv:2006.16234 [cs, stat]
7. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794. https://doi.org/10.1145/2939672.2939785. ArXiv:1603.02754 [cs]
8. ENTSO-E (2025) Entso-e transparency platform. https://transparency.entsoe.eu/. Accessed 12 Aug 2025
9. ENTSO-E (2025) Transparency platform. https://transparency.entsoe.eu
10. Fryer D, Strümke I, Nguyen H (2021) Shapley values for feature selection: the good, the bad, and the axioms. IEEE Access 9:144352–144360
11. Gerber E (2021) Radical Shapley values. https://edden-gerber.github.io/shapley-part-2/. GitHub: https://github.com/edden-gerber/radical-shapley-values
12. Heskes T, Sijben E, Bucur IG, Claassen T (2020) Causal Shapley values: exploiting causal knowledge to explain individual predictions of complex models. https://doi.org/10.48550/arXiv.2011.01625. ArXiv:2011.01625 [cs]
13. Hirth L, Mühlenpfordt J, Bulkeley M (2018) The ENTSO-E transparency platform—a review of Europe's most ambitious electricity data platform. Appl Energy 225:1054–1067. https://doi.org/10.1016/j.apenergy.2018.04.048 (https://www.sciencedirect.com/science/article/pii/S0306261918306068)
14. Hossain E, Khan I, Un-Noor F, Sikander SS, Sunny MSH (2019) Application of big data and machine learning in smart grid, and associated security concerns: a review. IEEE Access 7:13960–13988. https://doi.org/10.1109/ACCESS.2019.2894819 (https://ieeexplore.ieee.org/document/8625421)
15. Huxley OT, Taylor J, Everard A, Briggs J, Tilley K, Harwood J, Buckley A (2022) The uncertainties involved in measuring national solar photovoltaic electricity generation. Renew Sustain Energy Rev 156:112000. https://doi.org/10.1016/j.rser.2021.112000 (https://www.sciencedirect.com/science/article/pii/S1364032121012636)
16. (IEA) IEA (2024) World energy outlook 2024—analysis. https://www.iea.org/reports/world-energy-outlook-2024
17. Janzing D, Minorics L, Blöbaum P (2019) Feature relevance quantification in explainable AI: a causal problem. https://doi.org/10.48550/arXiv.1910.13413. ArXiv:1910.13413 [stat]
18. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. https://doi.org/10.48550/arXiv.2001.08361. ArXiv:2001.08361 [cs]
19. Kruse J, Schäfer B, Witthaut D (2021) Revealing drivers and risks for power grid frequency stability with explainable AI. Patterns 2(11):100365. https://doi.org/10.1016/j.patter.2021.100365 (https://linkinghub.elsevier.com/retrieve/pii/S2666389921002270)
20. Liu P, Trieb F (2023) German atlas of thermal storage power plants (TSPP) - a first approach. J Ener Storage 72:108603. https://doi.org/10.1016/j.est.2023.108603 (https://www.sciencedirect.com/science/article/pii/S2352152X23020005)
21. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2019) Explainable AI for trees: from local explanations to global understanding. https://doi.org/10.48550/arXiv.1905.04610. ArXiv:1905.04610 [cs, stat]
22. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp 4768–4777. https://doi.org/10.48550/arXiv.1705.07874
23. Molnar C (2023) Interpreting machine learning models with SHAP. A guide with python examples and theory on Shapley values, 1 edn. https://christophmolnar.com/books/shap
24. Molnar C (2025) Interpretable machine learning. A guide for making black box models explainable, 3 edn. https://christophm.github.io/interpretable-ml-book
25. Monti A, Ponci F (2015) Electric power systems. In: Kyriakides E, Polycarpou M (eds) Intelligent monitoring, control, and security of critical infrastructure systems. Springer, Berlin, Heidelberg, pp 31–65. https://doi.org/10.1007/978-3-662-44160-2_2
26. Nguyen HN, Tran QT, Ngo CT, Nguyen DD, Tran VQ (2025) Solar energy prediction through machine learning models: a comparative analysis of regressor algorithms. PLoS ONE 20(1):e0315955. https://doi.org/10.1371/journal.pone.0315955
27. Nikoltchovska A, Pütz S, Götz M, Schäfer B (2025) Correlations-xai-energy. https://github.com/KIT-IAI-DRACOS/correlations-xai-energy
28. Nikoltchovska A, Pütz S, Li X, Hagenmeyer V, Schäfer B (2025) Probabilistic and explainable machine learning for tabular power grid data. In: Proceedings of the 16th ACM international conference on future and sustainable energy systems, e-energy '25. Association for Computing Machinery, New York, NY, USA, pp 213–231. https://doi.org/10.1145/3679240.3734623
29. Olsen LHB, Glad IK, Jullum M, Aas K (2024) A comparative study of methods for estimating model-agnostic Shapley value explanations. Data Min Knowl Disc 38(4):1782–1829. https://doi.org/10.1007/s10618-024-01016-z
30. Owen G (1977) Values of games with a priori unions. In: Henn R, Moeschlin O (eds) Mathematical economics and game theory. Springer, Berlin, Heidelberg, pp 76–88. https://doi.org/10.1007/978-3-642-45494-3_7
31. Pfenninger S (2017) Energy scientists must show their workings. Nature 542(7642):393–393. https://doi.org/10.1038/542393a (https://www.nature.com/articles/542393a)

32. Pfenninger S, DeCarolis J, Hirth L, Quoilin S, Staffell I (2017) The importance of open data and software: is energy research lagging behind? Energy Policy 101:211–215. https://doi.org/10.1016/j.enpol.2016.11.046 (https://www.sciencedirect.com/science/article/pii/S0301421516306516)
33. Pütz S, Kruse J, Witthaut D, Hagenmeyer V, Schäfer B (2023) Regulatory changes in German and Austrian power systems explored with explainable artificial intelligence. In: Companion proceedings of the 14th ACM international conference on future energy systems, e-Energy '23 Companion. Association for Computing Machinery, New York, NY, USA, pp 26–31. https://doi.org/10.1145/3599733.3600247
34. Sensfuß F, Ragwitz M, Genoese M (2008) The merit-order effect: a detailed analysis of the price effect of renewable electricity generation on spot market prices in germany. Ener Policy 36(8):3086–3094. https://doi.org/10.1016/j.enpol.2008.03.035 (https://www.sciencedirect.com/science/article/pii/S0301421508001717)
35. Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games, annals of mathematics studies, vol II. Princeton University Press, Princeton, pp 307–317
36. Sovacool BK (2016) How long will it take? Conceptualizing the temporal dynamics of energy transitions. Energy Res Soc Sci 13:202–215. https://doi.org/10.1016/j.erss.2015.12.020 (https://www.sciencedirect.com/science/article/pii/S2214629615300827)
37. Sundararajan M, Najmi A (2020) The many Shapley values for model explanation. https://doi.org/10.48550/arXiv.1908.08474. ArXiv:1908.08474 [cs]
38. Veale M, Zuiderveen Borgesius F (2021) Demystifying the draft EU artificial intelligence act. https://papers.ssrn.com/abstract=3896852
39. Wang J, Wiens J, Lundberg S (2021) Shapley flow: a graph-based approach to interpreting model predictions. https://doi.org/10.48550/arXiv.2010.14592. ArXiv:2010.14592 [cs]
40. Zytek A, Liu D, Vaithianathan R, Veeramachaneni K (2022) Sibyl: understanding and addressing the usability challenges of machine learning in high-stakes decision making. IEEE Trans Visual Comput Graphics 28(1):1161–1171. https://doi.org/10.1109/TVCG.2021.3114864 (https://ieeexplore.ieee.org/document/9552849)